**Genome Annotation QC**

**How does SM\*ART Review Your Genome Annotation?**

Please note: this applies to genomes that have passed preliminary QC. Refer to the Bioinformatics Guide for information on how to pass Preliminary QC if you need to.

During preliminary QC, your file was checked for completeness, formatting of CDs and tRNA features, the frameshift in the tail assembly chaperones, cluster specific oddities, and for obvious violations of the guiding principles, including large gaps or overlaps, genes without a start or stop codon, as well as some phage biology basics.

**Global look:**

1. **BLAST the phage on phagesdb from the phage page**

Take note of the most closely related phages as you will use these to guide your review. Final annotations are likely to be more useful for review, but you should not discount the draft annotation data. Notice if different phages are more similar to your phage in different regions of the genome.

1. **Open up a Phamerator map**

Include the phage you are checking and a few\* similar (finalized) ones; Make sure that you turn on “gene descriptions” on the Phamerator map such that you can view functional assignments from the phagesdb database.

\*few is a variable term---to start, you may want to add a large number of phages to your map to get a sense of how congruent the annotations of the cluster/subcluster are. Once you’ve explored the cluster/subcluster, you may want to refer to only a few finished annotations for the bulk of your review work.

1. **Scroll through the pham pages for each gene.**

Take note of genes that belong to phams that contain genes assigned a variety of conflicting functions, or that only a few genes have been assigned a function, etc. These should be addressed and reconciled in the annotation.

1. **Open both the minimal file and the complete notes file in DNA Master**.

Optional : Run a Genome Comparison with two or three closely related genomes and the file you are checking in DNA Master (add link). Take note of which genes or portions of the genome are likely to be very similar to previously annotated phages, which are not. You can also easily identify phams in which the start is not conserved as the genes will be different lengths. (This can also be done on a Phamerator map).

**Local look:**

1. **Open the frames window of the minimal file in DNA Master.**
2. **Examine the genes one by one in the Description tab, the BLAST tab, and frames window.**

Look at the Complete Notes for the first ten genes (or any random ten genes in the genome). Are they complete and coherent? Does the information recorded in the notes reflect the gene content in the annotation? Can you identify any misconceptions about the bioinformatic tools and what their outputs mean from the notes?

If the notes seem clear, complete, and demonstrate understanding of the tools, move on through the rest of the genome.

Scroll through the BLAST tab in the minimal file and the description tab in the Complete Notes file. Flip back and forth between the Frames window and your description and BLAST tabs so you can see the same genes in the different windows.

Analysis: If the Complete Notes for multiple genes demonstrate that the annotator is interpreting the data the right way, the BLAST data for each gene is 1:1 with the majority of gene alignments, and the genes look like they are nicely filling the space in the frames window, without large gaps or overlaps—preferably showing a 4bp overlap---it is likely genes are called correctly.

If the BLAST alignments are not 1:1, and/or there are alternate start choices that minimize the gap/overlap better than the one selected in the file, you should go back and read the notes about that gene more closely. If I think the annotator weighed all the evidence and made the right choice or made a reasonable choice in a murky situation, I’ll leave the start as called. Otherwise, I’ll change the start--- usually to support a 4bp overlap, to include all coding potential, to support a conserved start in Starterator, or to make sure that I’ve captured an entire functional domain.

1. **Examine suspicious gaps in CDs features**

Any suspicious gaps that you see in the frames window (such as one that contains a good size ORF, an ORF where the start/stop codons overlap those of the surrounding genes, or an ORF with even a hint of coding potential on a GeneMark ouput) you should examine with BLAST and HHPred.

I “add” ORFs like these to my file so that I can more easily run the analyses. If the results come back with well-supported functional data or good BLAST alignments, I’ll leave it in the file, otherwise, I’ll take it back out. Ideally, there will be a mention that this type analysis for this gap was already performed in the cover sheet.

Check your Phamerator map. Do any of the related genomes have a similar gene predicted in the same area?

Make sure to check the Cluster specific annotation forums to see if there is any information about a particular gap. Some clusters have conserved gaps that contains small RNAs or binding sites for example. In that case, the gap should be there.

**4. Check functions.**

Are they all on the official list?

Is one and only one gene assigned the function of:

portal

tape measure

major capsid

terminase

major tail ?

Are all functions well-supported with three lines of evidence?

Check your Phamerator maps. Do the phages in the map have similar functions for similar genes? Do they have additional functions that your annotation doesn’t have? If so, you will have to assess which is correct based on the Notes in the annotation you are reviewing and/or rerunning the analyses.

Check the Pham pages on phagesdb.org. Are there any functions listed for a pham member that are not included in your annotation? ? If so, you will have to assess which is correct based on the Notes in the annotation you are reviewing and/or rerunning the analyses.

**5. Validate and renumber (if necessary).**

You’ve finished! The minimal file as outlined in the Bioinformatics Guide is the starting point for generating the GenBank submission files.