**Case Study: What to do when more than one gene**

**appears to be a portal protein**

**Objectives:**

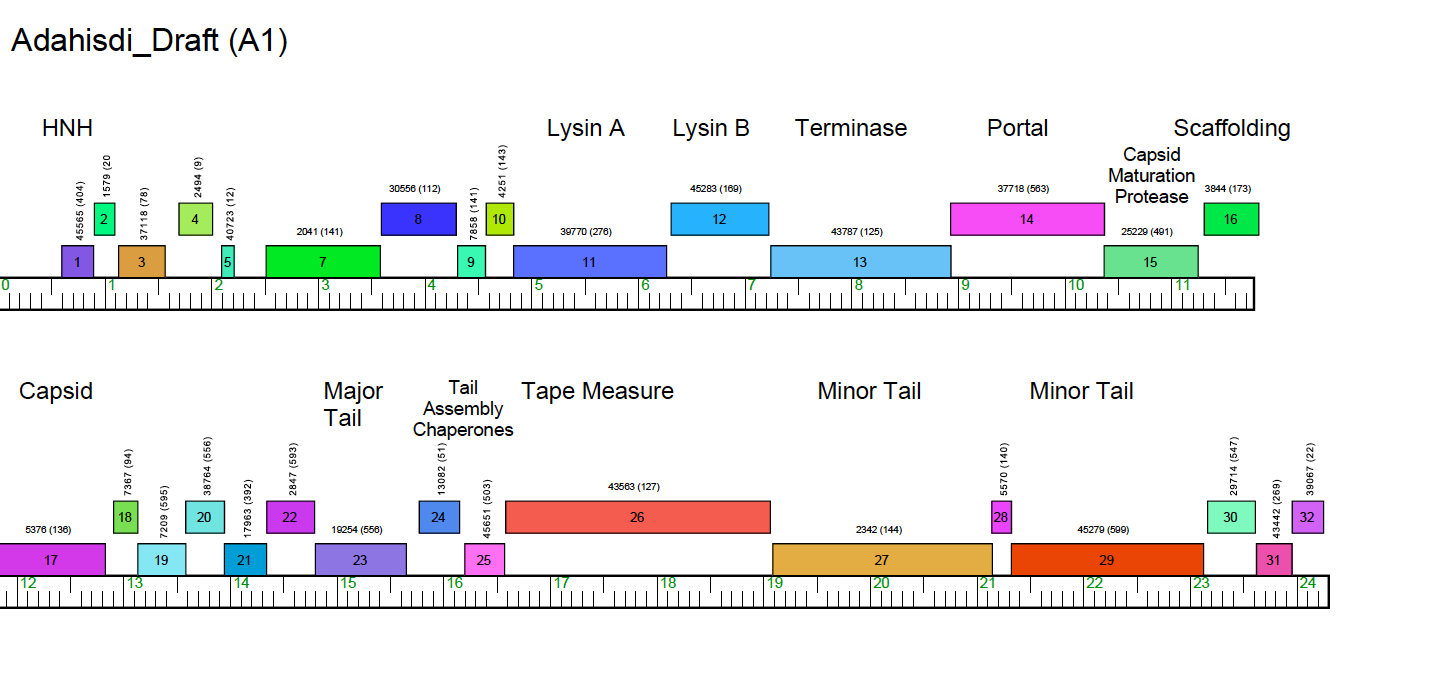
* Evaluate synteny of phage genomes and location of portal protein and head-to-tail connector complex proteins
* Successfully evaluate HHPRED results that mislead annotator to call multiple portal proteins

**Introduction:**

Bacteriophage genomes contain only **ONE** portal protein. This portal protein is commonly located at the beginning of the genomes (in the structural genes) and located directly downstream of the terminase. Portal proteins are larger genes usually spanning ~1400 base pairs.

**Cluster A1 Phage Adahisdi (Synteny)**

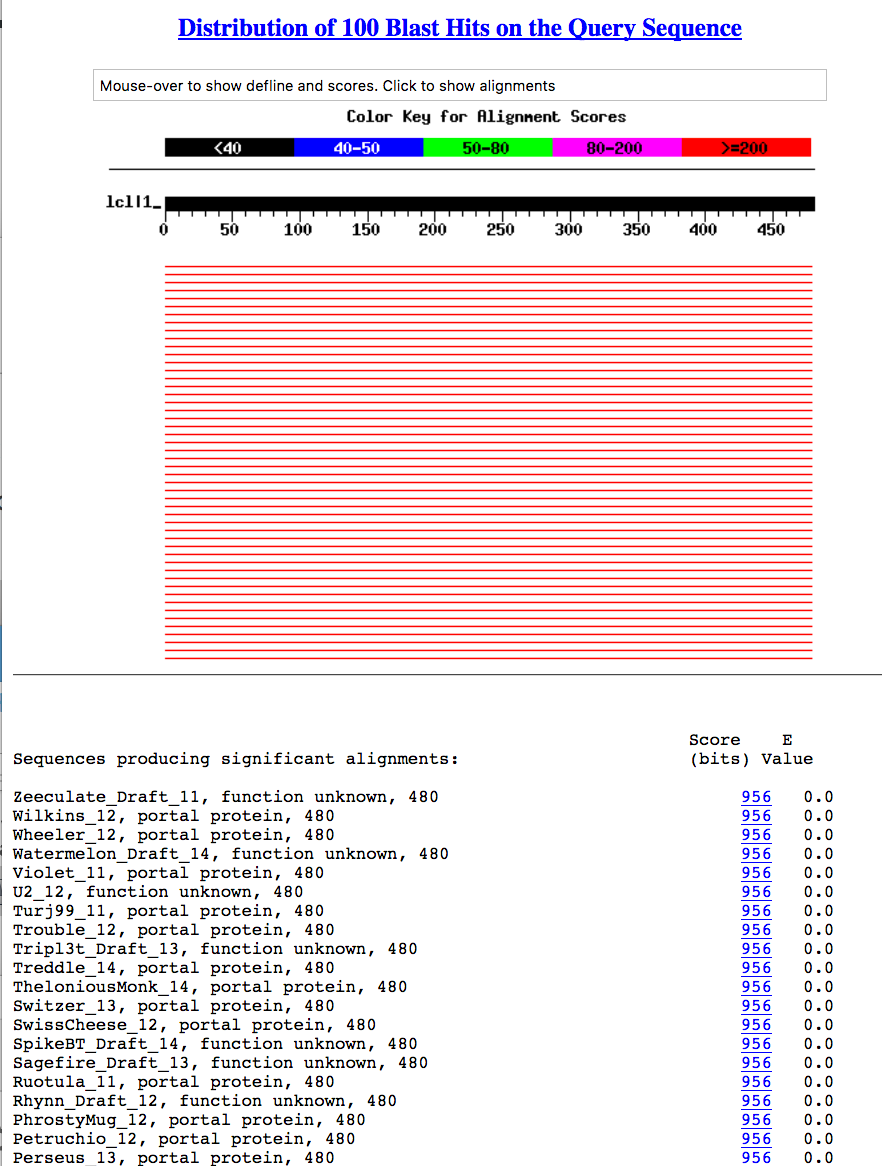
The phamerator map below is the A1 structural proteins. While Lysin A and Lysin B are typically found after tail proteins, they are located upstream of terminase in Cluster A1 genomes. The next 4 genes show conserved order of portal, capsid maturation protease, scaffolding protein and major capsid protein. The genes between the major capsid protein (gp17) and the major tail protein (gp23) are good candidates for head-to-tail connector complex proteins. The evaluation of BLAST and HHPred data for gp14 and for gp19, gp20 and gp22 can be confusing.



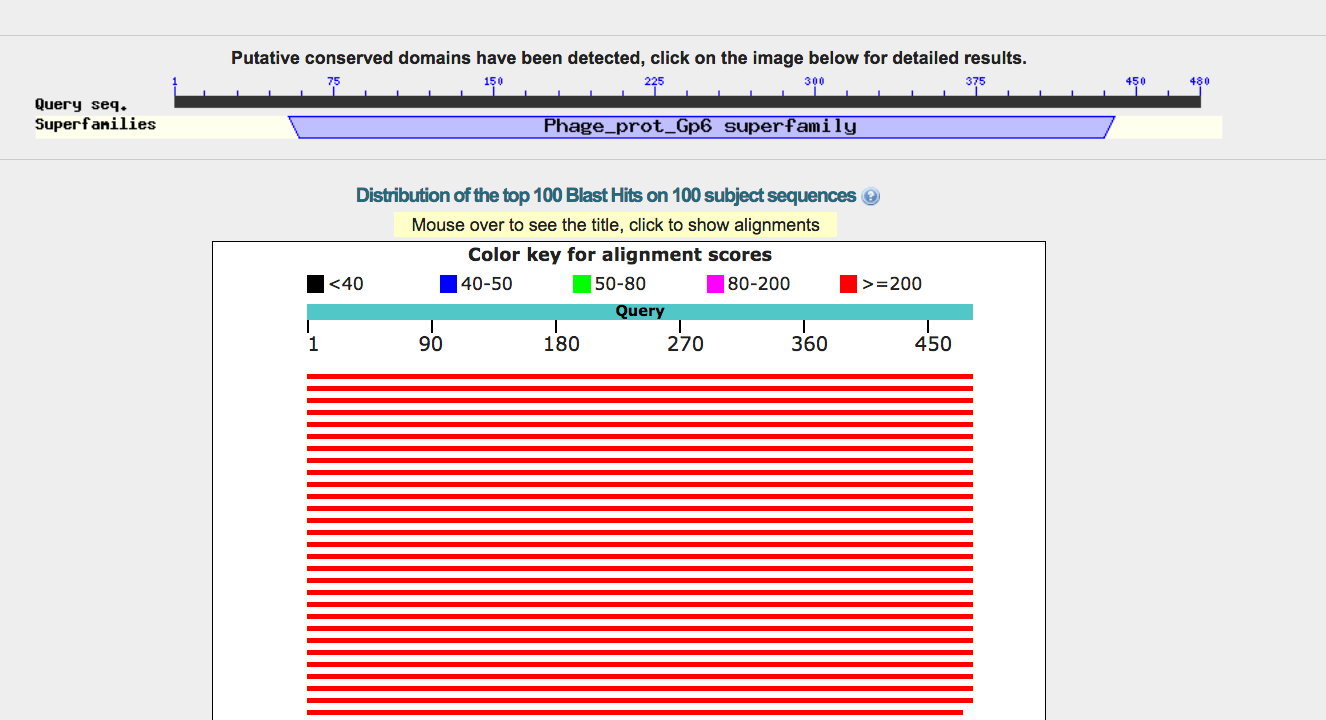
**Functional evaluation of gp14:** In order to assign a function to this protein, you will need to collect the following data:

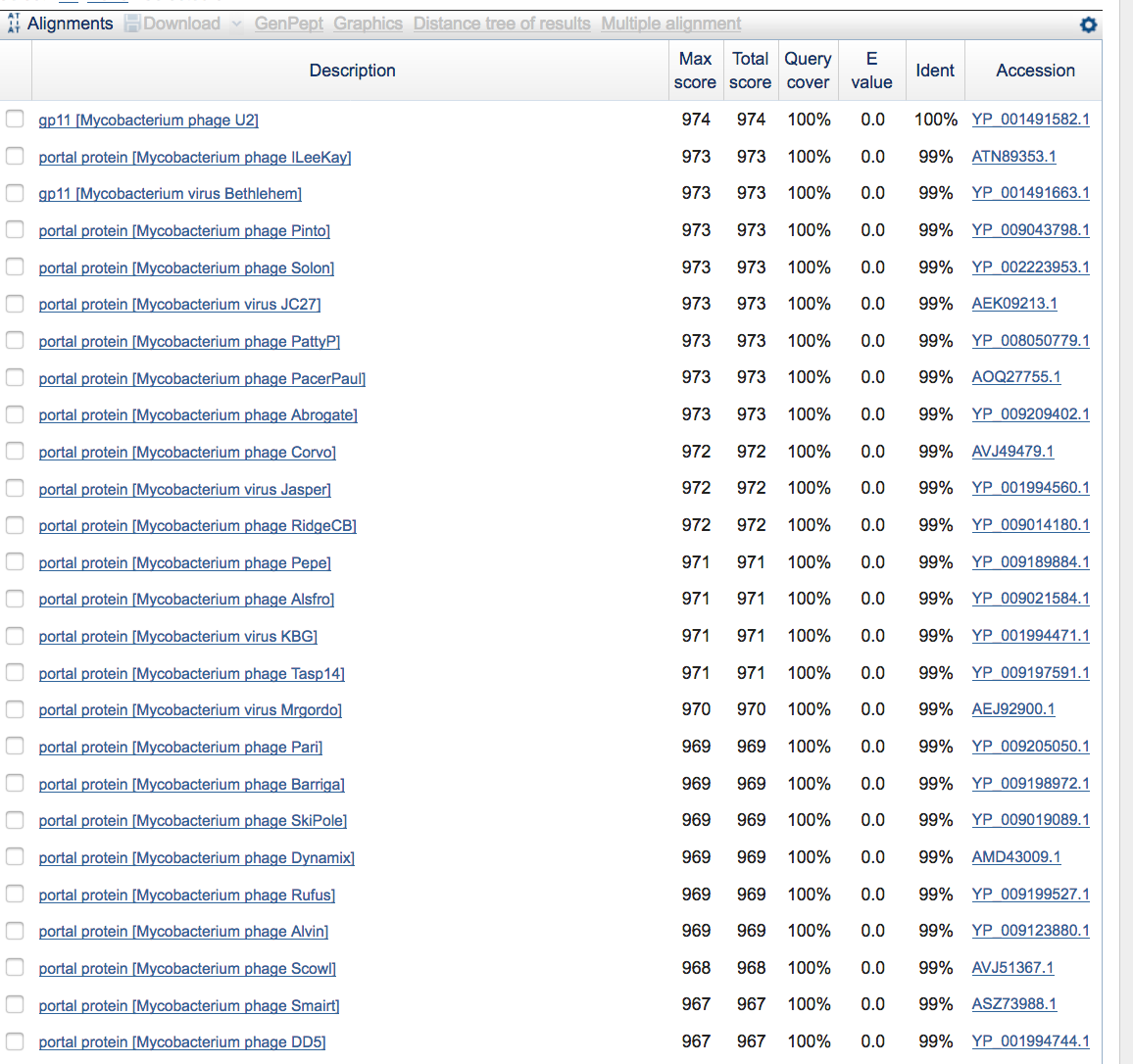
BlastP at PhagesDB, BlastP at NCBI, and HHpred

**BlastP at PhagesDB:**

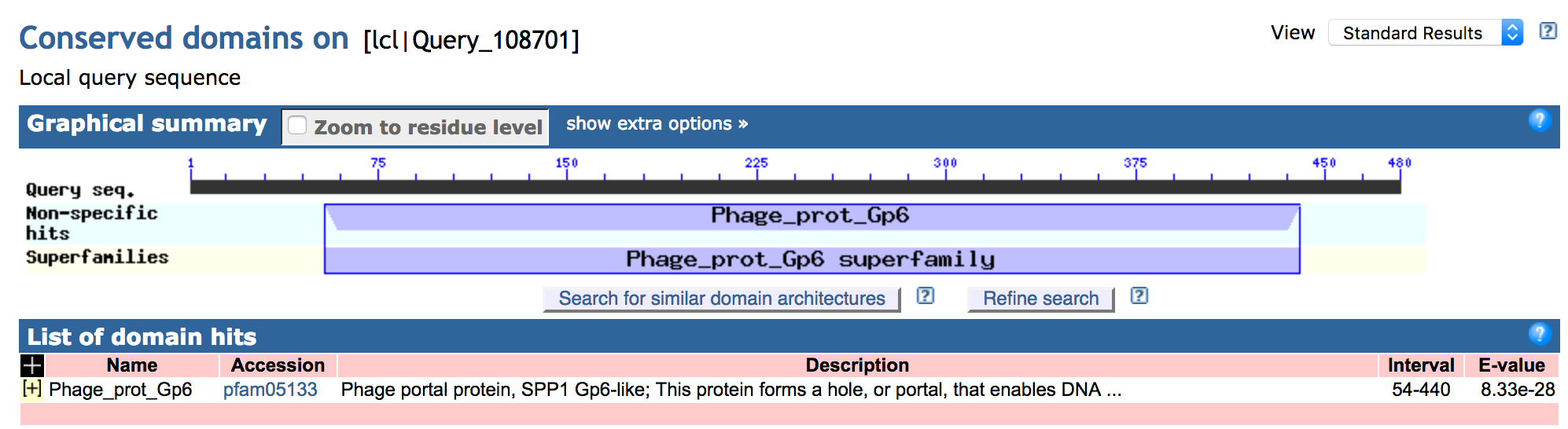
****

**BlastP at NCBI:** I have only included a partial snapshot of the data, but it clearly represents the consensus of the alignments at NCBI.



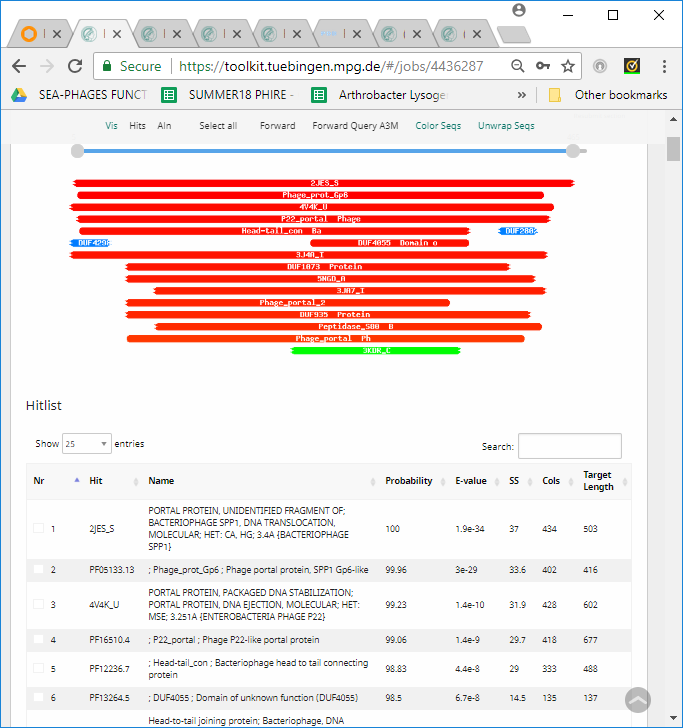


The Conserved Domain Database Results:



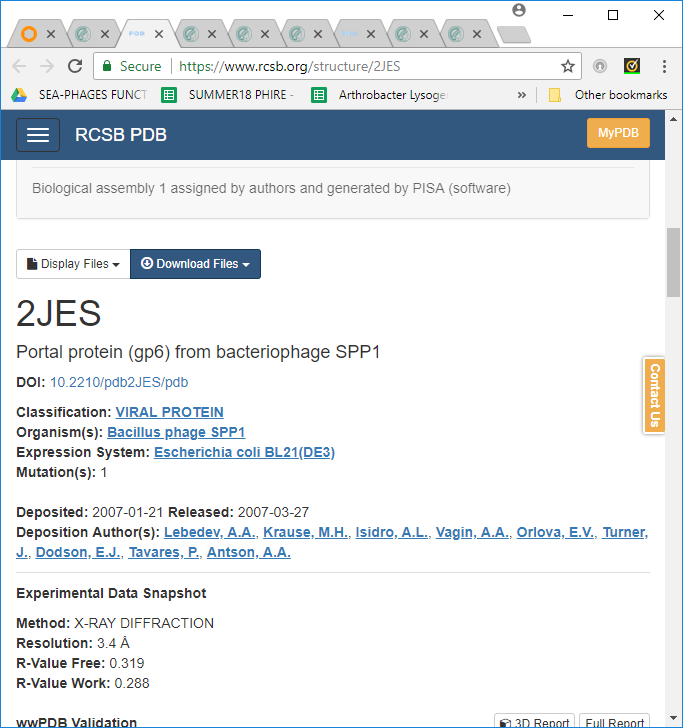
Closer inspection of this data can include alignment length and the source of the assigned function. Upon review, you will determine that the alignment is excellent along the entire length of the protein and that the first 500 alignments are to mycobacteriophages in the PhagesDB collection.

**HHpred Search:** Review the [Bioinformatics Guide](https://seaphagesbioinformatics.helpdocsonline.com/article-19) to determine the limitations of HHPred and assign [functions](https://seaphagesbioinformatics.helpdocsonline.com/article-43). Note that HHpred performs iterative alignments that are then aligned to the various databases. You will want to pay attention to how your gene matches SPP1 proteins in the HHpred results because the evidence in the literature for these SPP1 proteins is well described.



There are 21 matches in the Hitlist for Adahisdi\_gp14, of which 14 are significant (>90% probability). Three of those hits are linked to crystal structures of SPP1, P22, and T7 (hits 1 and 3). Because of these excellent hits to crystal structures of phage proteins, the PFam data would be less needed.

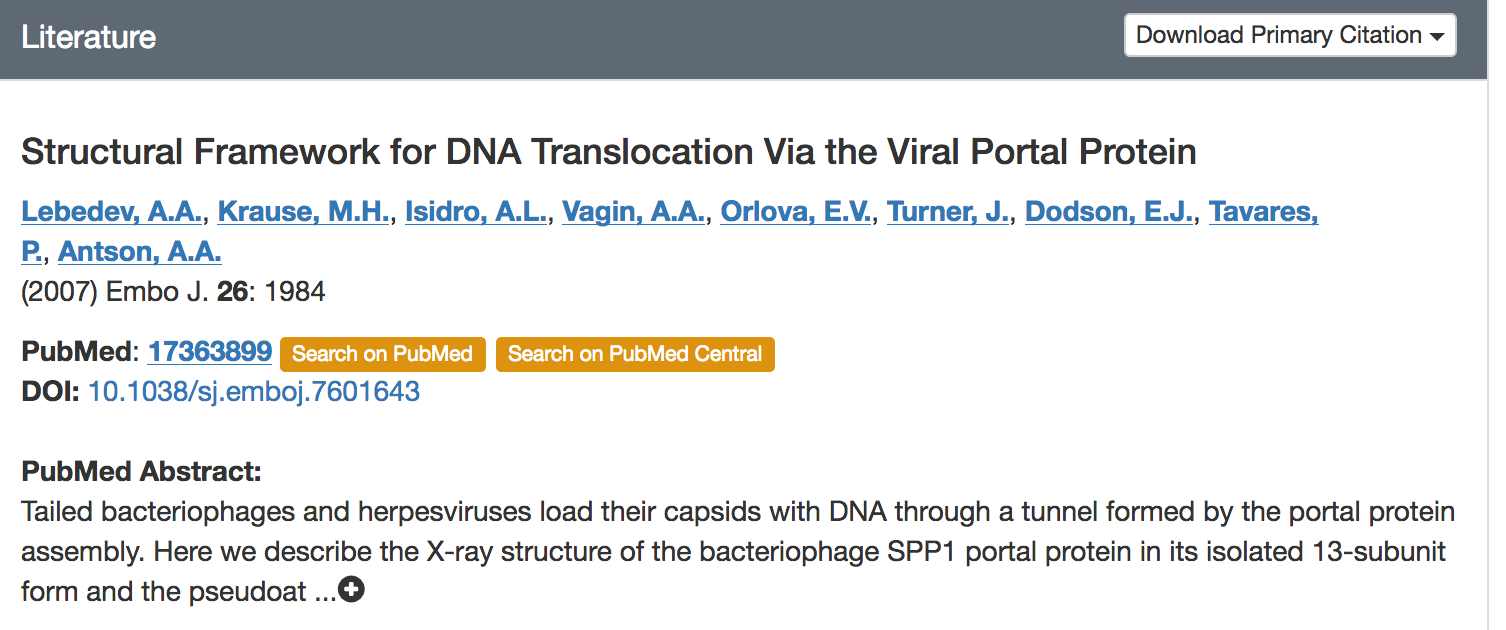
If we click on the first Hit, we are linked to this page:



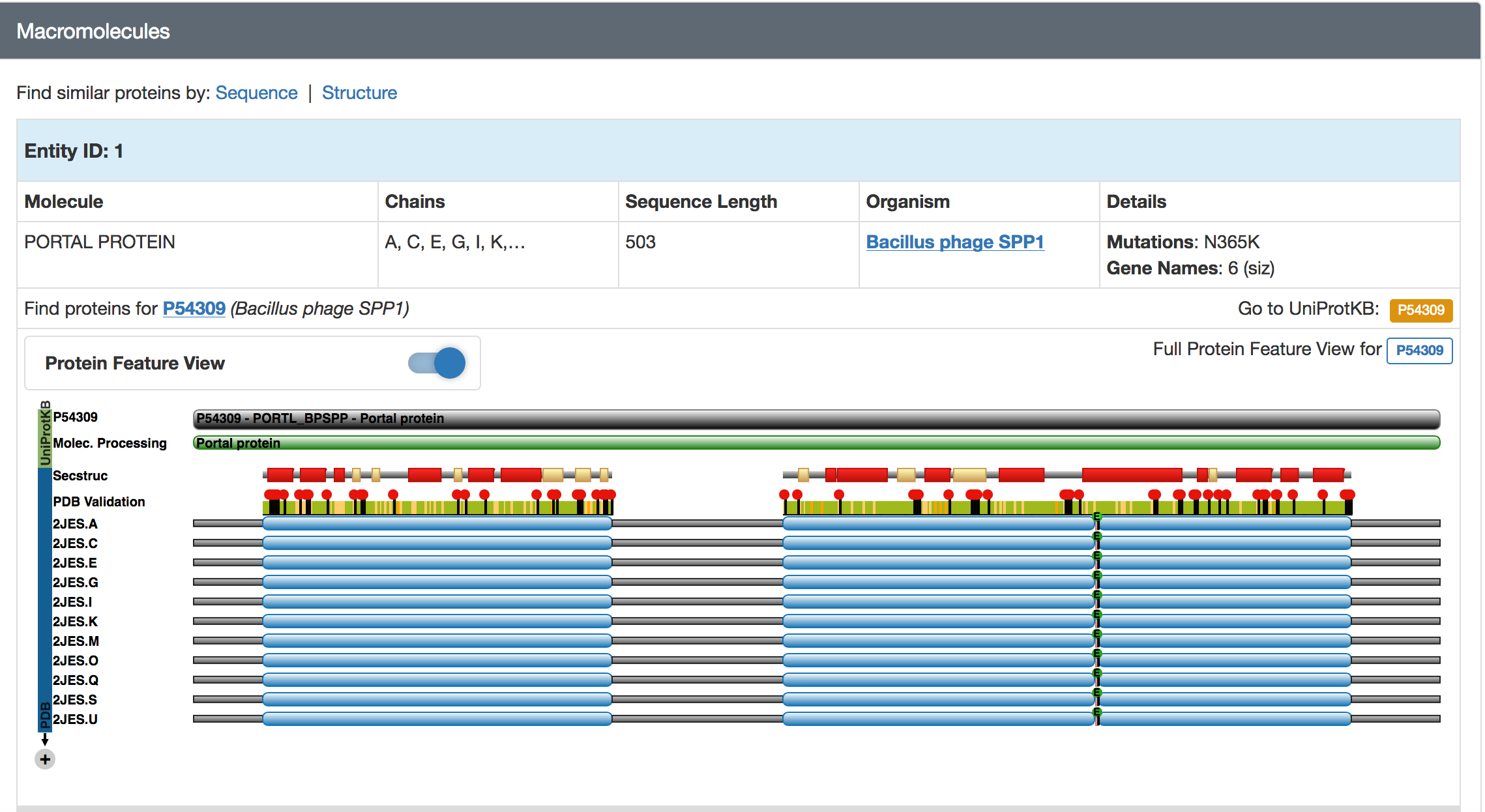
As you scroll down this page, more details into the results are provided. Note that the Hitlist lists the match to 2JES and not just one of its subunits. As you can see from the alignments, Adahisdi\_gp14 matches the entire length of the SPP1\_gp6, a portal protein.

Those details include:

**Reference to the literature:**



**Alignment data:**

****

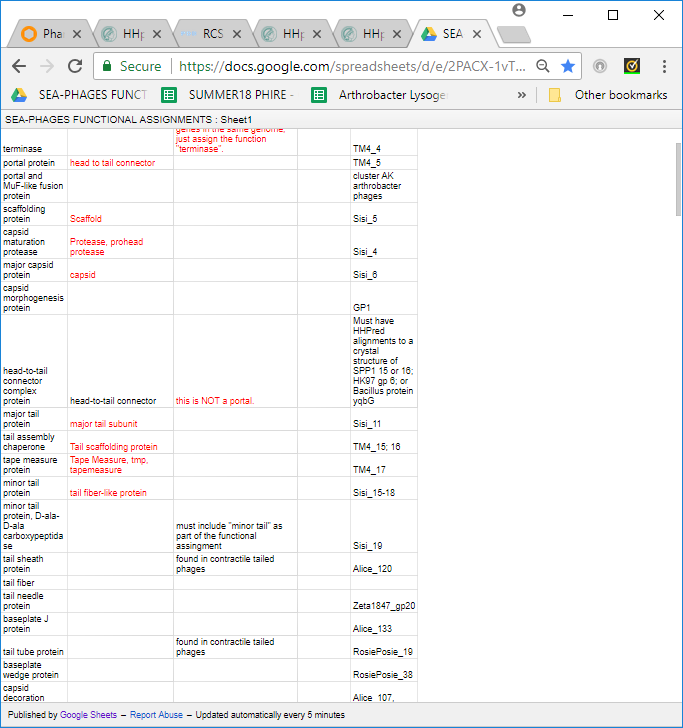
From this data it is easy to see that Adahisdi\_gp14 is the (one and only) portal protein.

**Functional evaluation of Adahisdi\_gp18, gp19, gp20, gp21 and gp22:**

These genes are located between the capsid gene and the major tail gene. Gp23 is the major tail protein, followed by the tail assembly chaperones and tape measure.

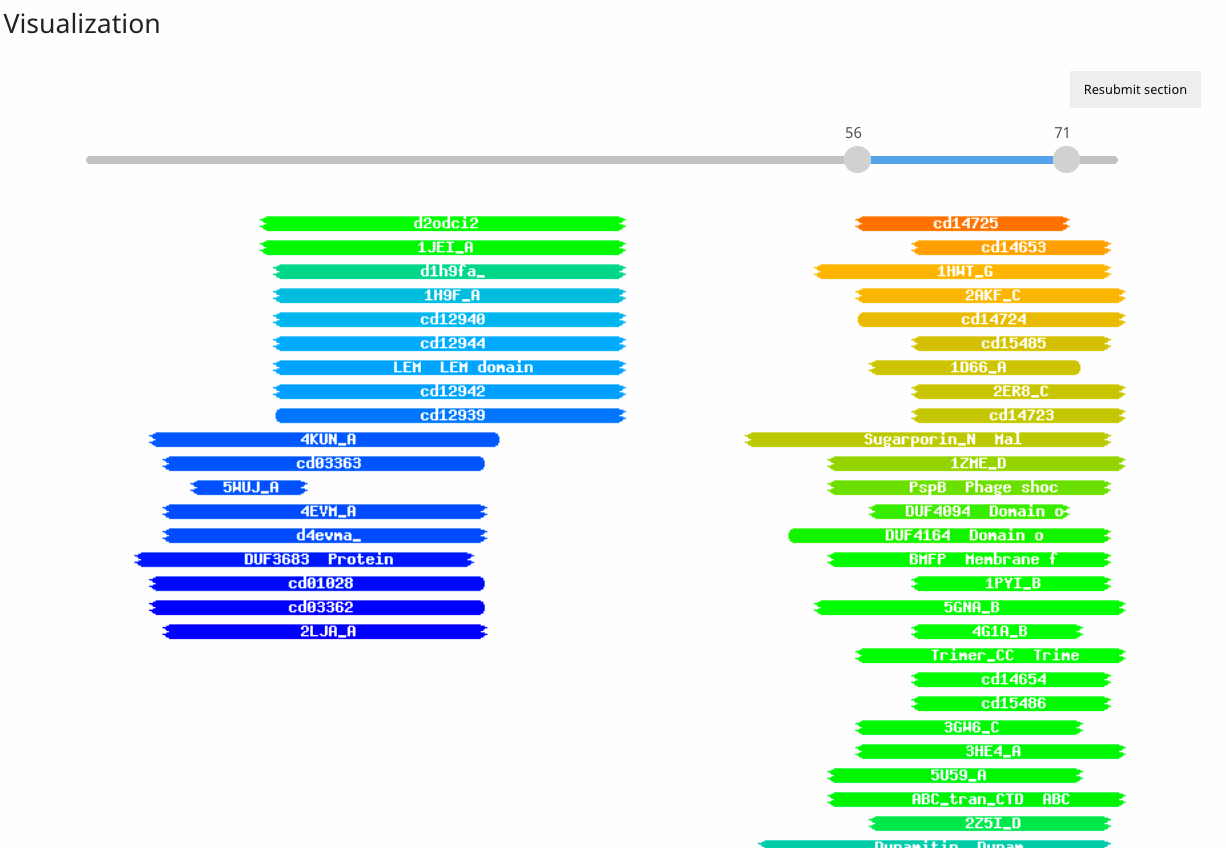
**BlastP at PhagesDB and BlastP at NCBI**

This BlastP data is not shown here, mostly because at this moment in time, the functional assignments are not reliable. Prior to last fall (2017), we were identifying head-to-tail connectors by position in the genome. Currently, we are only assigning the function of “head-to-tail connector complex protein” to those genes that match SPP1 gp15 and 16, HK97 gp6 and the Bacillus protein yqbG as per the [“SEA-PHAGES Functional Assignments”.](https://seaphagesbioinformatics.helpdocsonline.com/article-96)



In order to assign functions, the relevant data is found in the HHpred data. I have included the graphic representation of the data for gp18 – 21.. They will be evaluated in the same way.

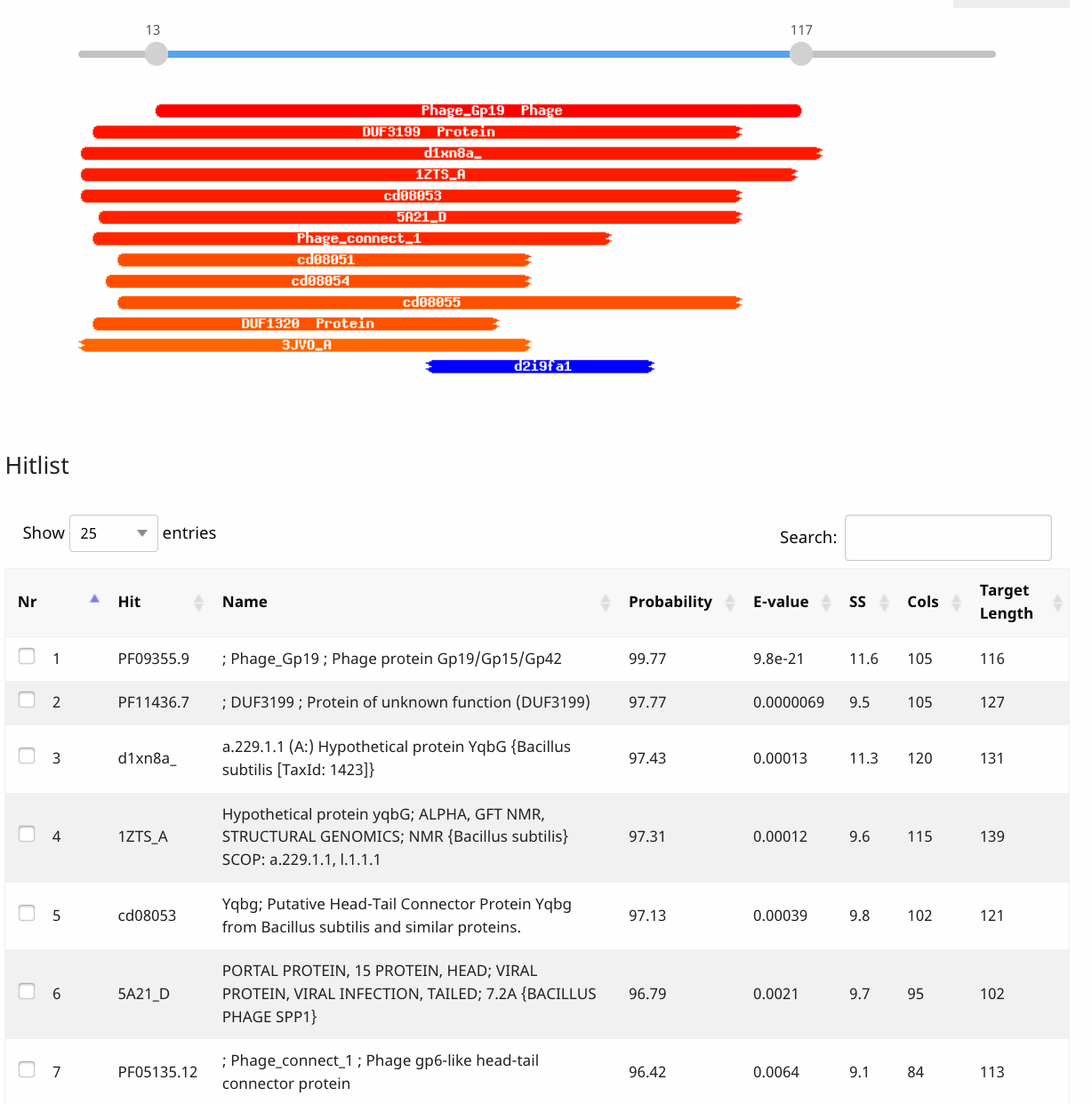
**gp18**



****

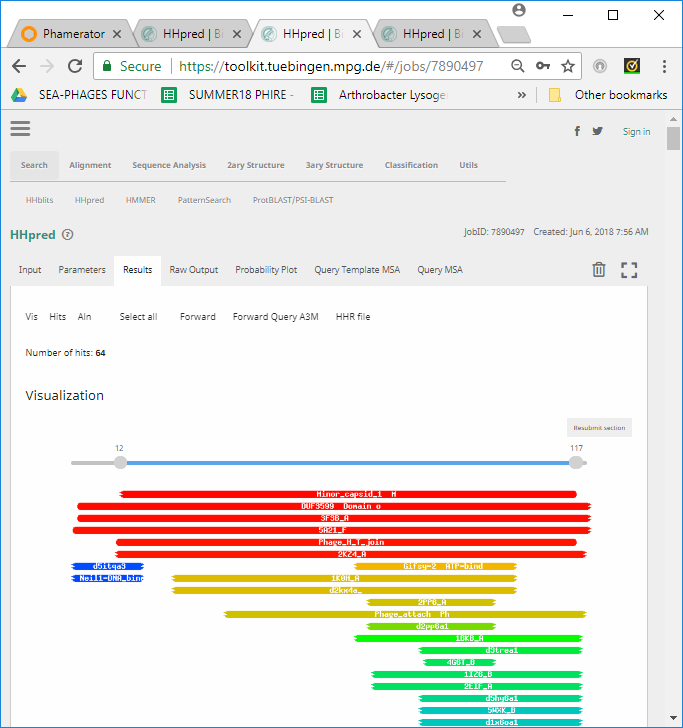
There are 250 “hits” to gp18. Only two hits are shown, mostly because only one hit (the first one) is numerically significant (Probability score of 90.5%). However, this function would not be attributed to a phage, so should be recorded at NKF.

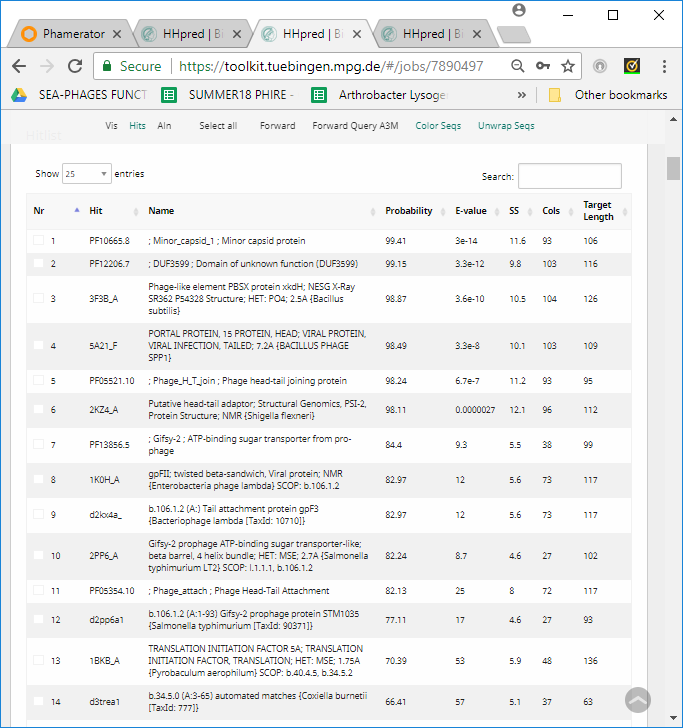
**gp19**



There are 29 hits to gp19. The most convincing hit in this collection is #6. It appears that it may be a portal protein. But remember we have already called the portal, so we must evaluate more closely. See Protein Database information for more details (below). Note that hit#6 is 5A21\_D. The “D” is a significant part of the determination.

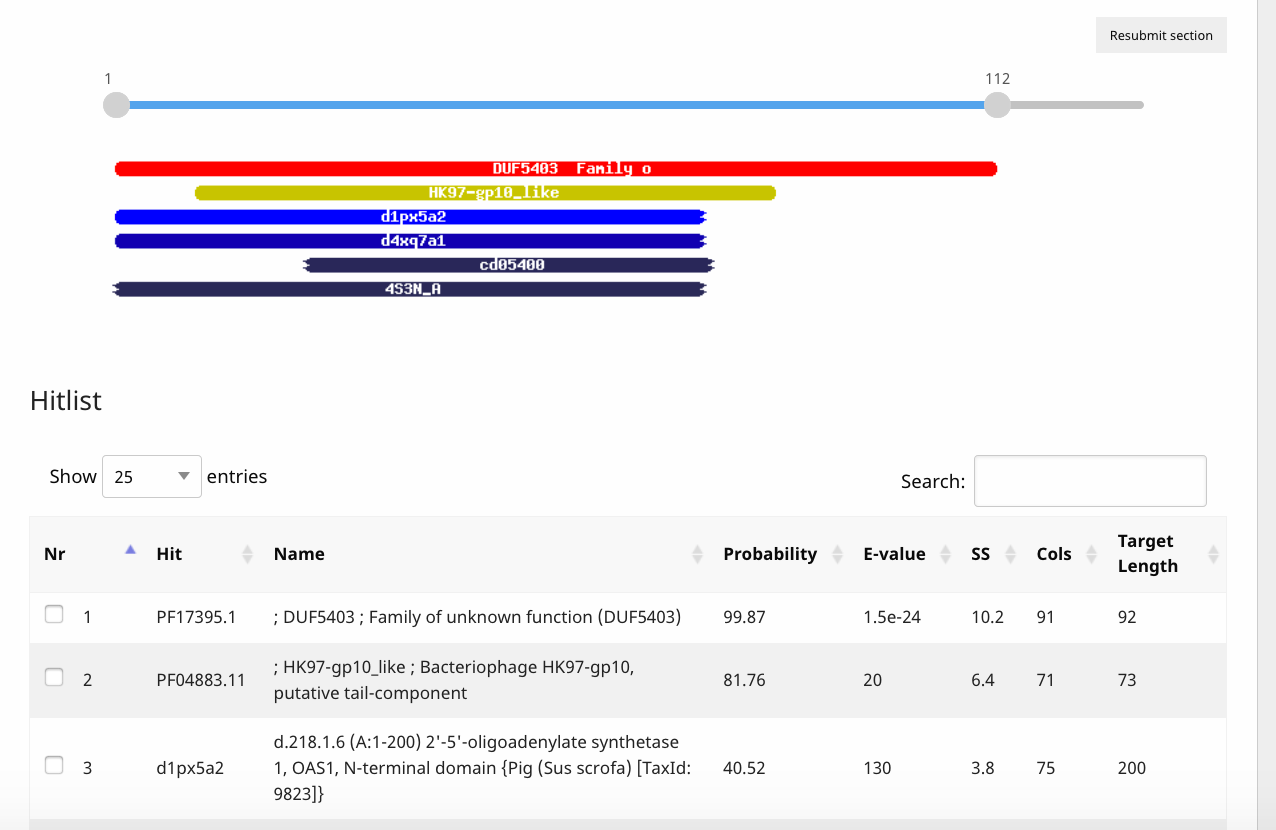
**gp20**





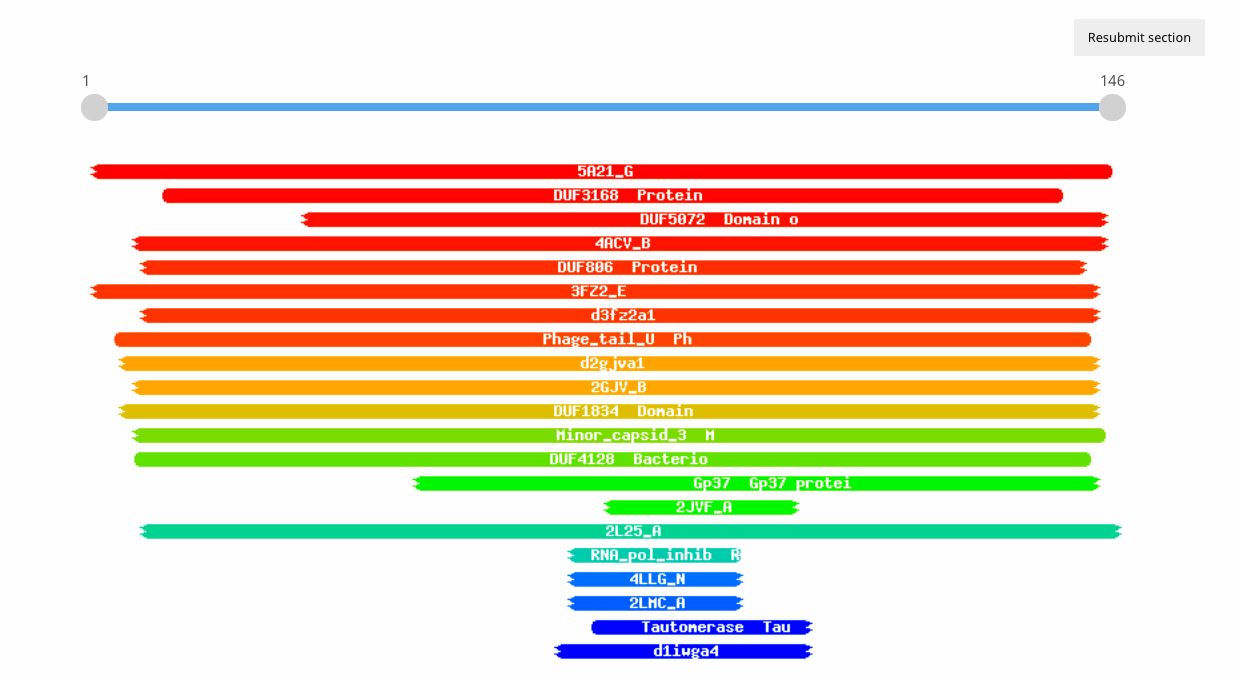
There are 64 “Hits”. There are two hits in the list that I would investigate. The first is #3 “phage-like element PBSX protein xkdH ....” from Bacillus cereus. The Hit link “3F3B\_A” informs you that this a phage protein of unknown function. The second hit that I would investigate is #4, “5A21\_F”. From the list available on this page, it appears that it may be a portal protein. But remember we have already called the portal, so we must evaluate more closely. See Protein Database entry information for more details (below). Note that hit#6 is 5A21\_F. The “F” is a significant part of the determination.

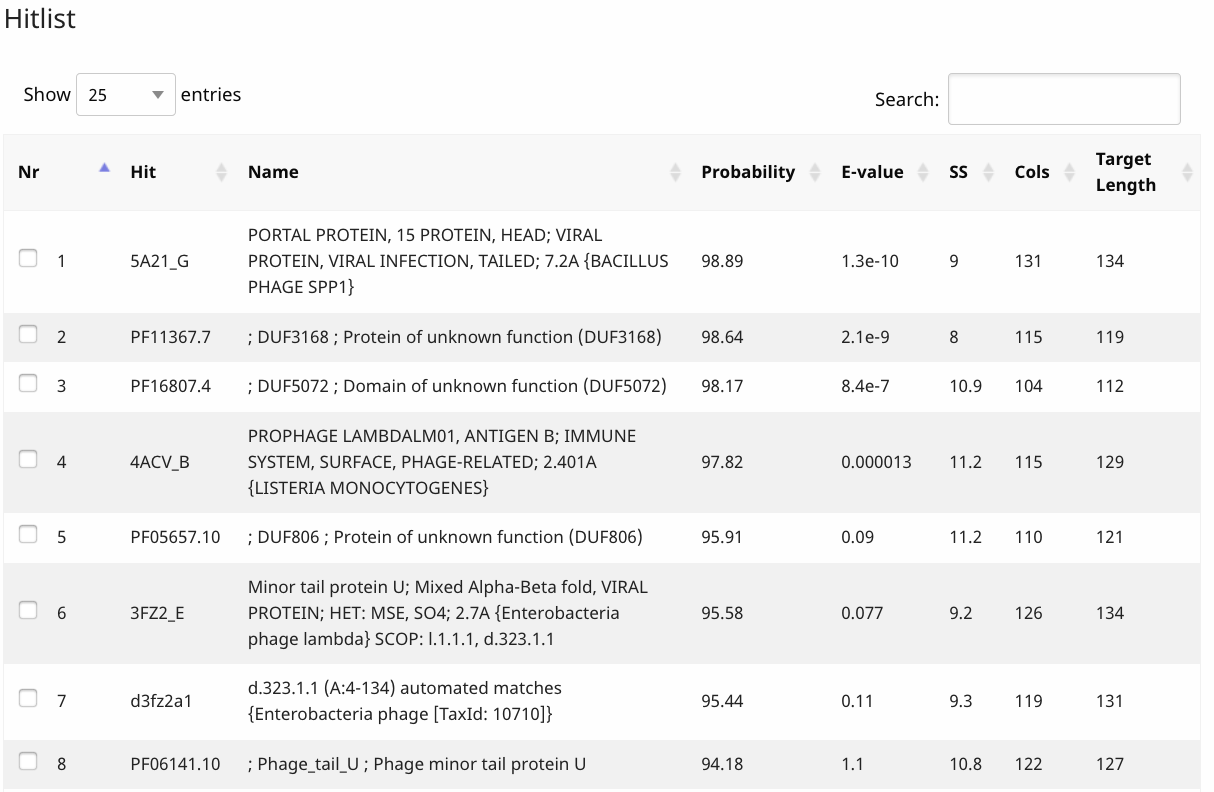
**gp21**

****

The first hit is to a DUF (domain of unknown function). The second hit is to HK97 gp10. The literature described it as “it may be a tail component”, but without further bench work, it is best labeled NKF.

**gp22**

****

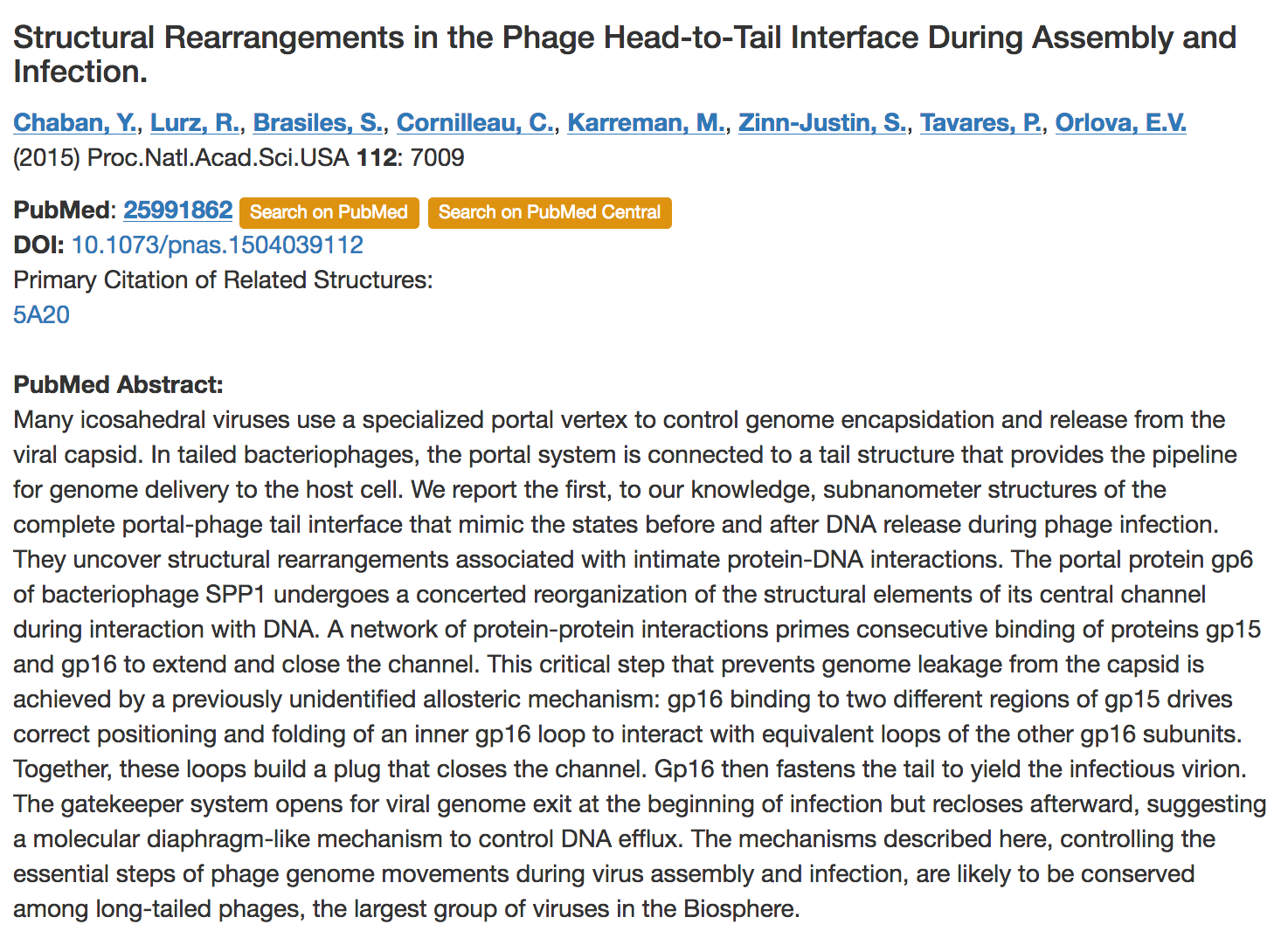
****

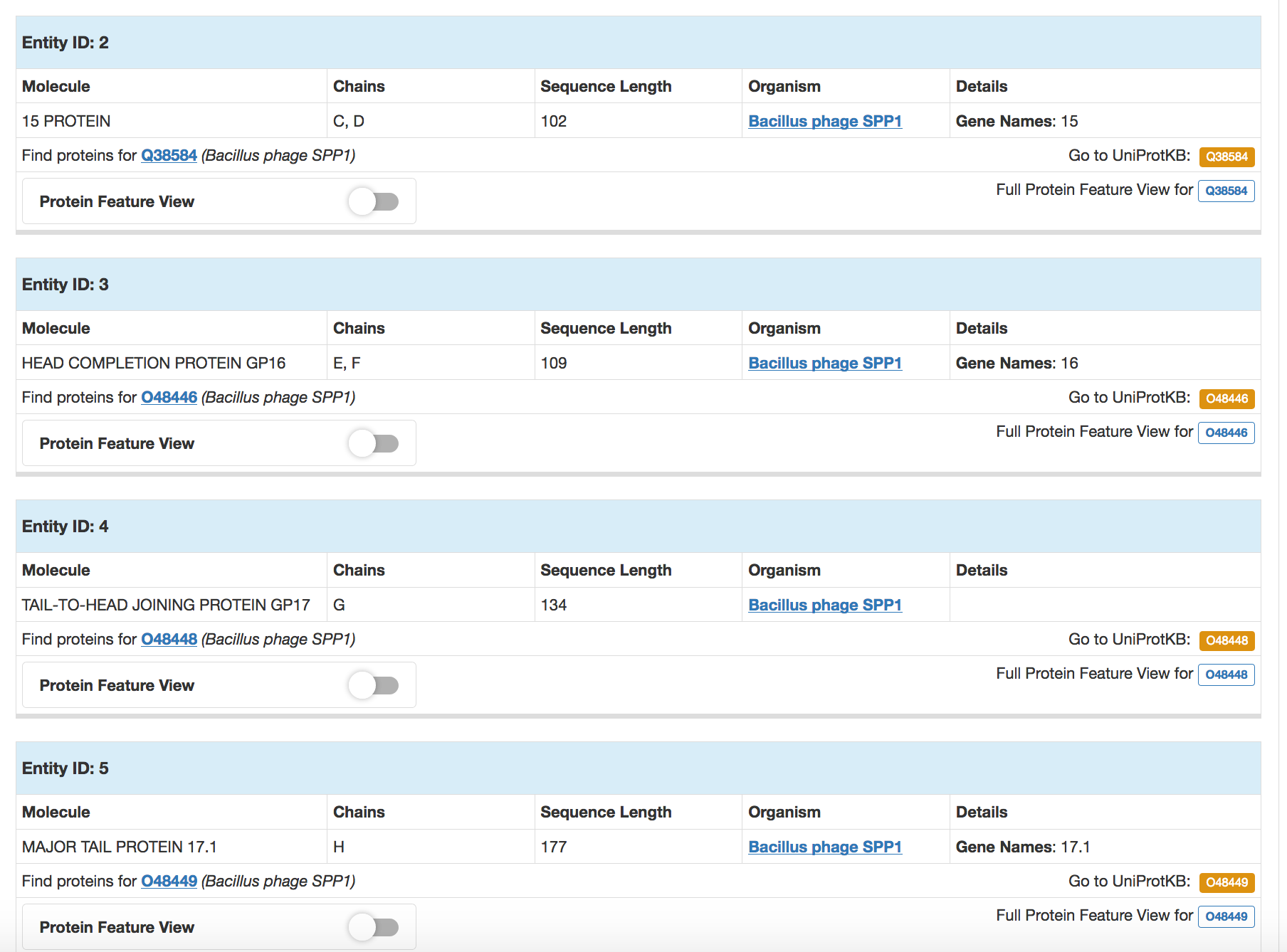
There are 48 “Hits”, 8 of which would be considered significant. I would investigate four of the hits in the list. The first is #1 “5A21\_G”. From the list available on this page, it appears that it may be a portal protein. But remember we have already called the portal, so we must evaluate more closely. See Protein Database entry information for more details (below). Note that hit#6 is 5A21\_G. The “G” is a significant part of the determination. The other hits to investigate include # 4, 6 and 8. **Hit #4:** Without experimental evidence, I am unwilling to assign a non-related host antigen to one of our actinobacteriophages. **Hits #6 and #8:** These two hitspoint to “minor tail protein U of Escherichia phage lambda”. If you continue down the list you will see hits to minor capsid protein and phage structural genes, and DUFs. Because the #1 hit is to a phage protein with a function that fits the synteny of the genes and because that entry is so well described in the literature, it a more likely functional assignment. If the alignments are believable, assigning the function that correlates is also believable.

**Protein Database entry information**

The Hit to look at is the one that takes you to the Protein Database. Regardless of which gene you look at, they link to the same PDB page. It is the one with the 5A21 prefix (For gp19, it is 5A21\_D. For gp20 it is 5A21\_F. For gp22 it is 5A21\_G.

When you click on the link “5A21\_D”, you are linked to the primary citation of related structures and an explanation of the molecules:



****

Looking at these results you can see that chain A & B align to the portal protein, chain, B aligns to a protein with no designated function; C and D align to 15 protein (Though this likely a structural protein, “no known function” is assigned); E and F align “head completion protein gp 16” and can be annotated as Head-to-tail connector complex protein; and G aligns to “head completion protein gp 17” and can be annotated as “head-to-tail connector complex protein”.

The following functions can be assigned:

|  |  |  |
| --- | --- | --- |
|  | PBD “Hit” | Annotated Function |
| gp18 | none | NKF |
| gp19 | 5A21\_D | NKF |
| gp20 | 5A21\_F | Head-to-tail connector complex protein |
| gp21 | HK97\_gp10 | NKF |
| gp22 | 5A21\_G | Head-to-tail connector complex protein |