**NAME\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**BMS370 Project – Unit III –**

**Using BLAST for comparative genomics and visualizing HGT**

**Notes and suggestions for future changes**

**(After implementing 1 semester)**

* **Run chlamydia BLAST for both part 2 and part 3**
* **Make a chlamydia 16S species tree and have students fill in Add and PRIP gene presence and % ID to C. pneumoniae on the tree.**
* **Better in-class prep for the tree interpretation and making**
* **If you have more time to introduce it, use MEGA to get better trees**

ANSWERS ARE IN RED WITH HIGHLIGHTING

**Part 1: BLAST and tree analysis**

The purpose of this exercise is to get familiar with the Basic Local Alignment Search Tool (BLAST). BLAST finds regions of local similarity between sequences. The program compares a query sequence (a protein or nucleotide sequence) against a sequence database and calculates significance of matches. BLAST divides the query sequence into shorter words and initially looks for matches of these words only. The tool gives a score based on a scoring system e.g. in blastn, it will give +1 for each match and -2 for each mismatch.

BLAST can be freely accessed at the NCBI website at:

<http://blast.ncbi.nlm.nih.gov/Blast.cgi>

More information on BLAST and the parameters used in the BLAST algorithms can be found at:

<http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs>

Learning Goals:

* Become familiar with BLAST and be able to use it for homology searches
* Be able to interpret BLAST results
* Picking a best match for a query sequence
* Identify an unknown sequence

*This set of questions demonstrates the different matches you get when you run a BLAST search and how to pick the best matches for your search.*

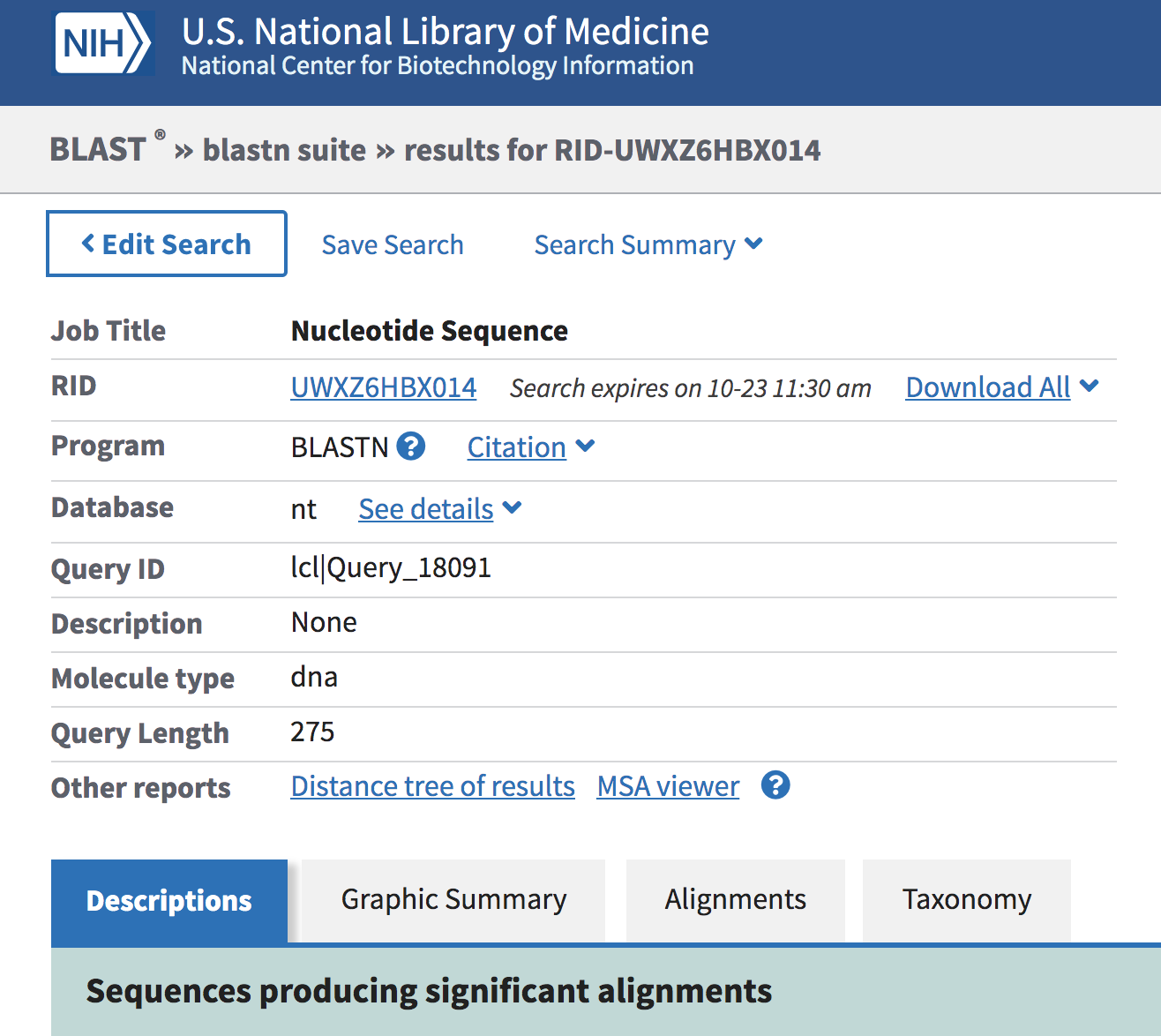
The FASTA sequence for this part of your problem set is given below:

>Unknown 16S

TTGTAAGCACTTTCGCCTGGGAATAAGAGAGATTGGCTAATATCCAATCGATTTGAGCGTACCAGGTAAAGAAGCACCGGCTAACTCCGTGCCAGCAGCTGCGGTAATACGGAGGGTGCTAGCGTTAATCGGATTTATTTGGGCGTAAAGGGCGTGTAGGCGGAAAGGAAAGTTAGATGTTAAATTTTGGGGCTCAACCCCAAGTCAGCATTTAAAACTATCTTTCTAGAGGATAGATGGGGAAAAGGGGAATTCCACGTGTAGCGGTGAAATGC

**Go to** [**http://blast.ncbi.nlm.nih.gov/Blast.cgi**](http://blast.ncbi.nlm.nih.gov/Blast.cgi) **and click on nucleotide blast under BASIC BLAST and copy and paste the FASTA sequence above in the “Query Sequence Box”.**

*Run a blastn and wait for the results to answer the following questions*: your results should look like this



.

**1A)** How many sequences are there in the database that was used for your search? *Click on Search Summary at the top of the page*

If nr: 57,082,062 or more

If 16S (incorrect search string) will be less

**1B**) Look at “Descriptions” – what is your identification of the origin of this sequence? E.g. what species of bacteria is it?

**Chlamydia pneumoniae**

**1C)** What is the bit (Max) score, E-value and query coverage values of the best BLAST result for the matching species? What is the Max score, E-value and percent identity for the next most closely related species?

**Chlamydophila Pneumoniae**

Max score:490

E-value: 1 X 10-134

Percent Identity: 98.91%

**Chlamydia abortus**

Max score: 451

E-value: 7 X 10-123

Percent Identity: 96.38

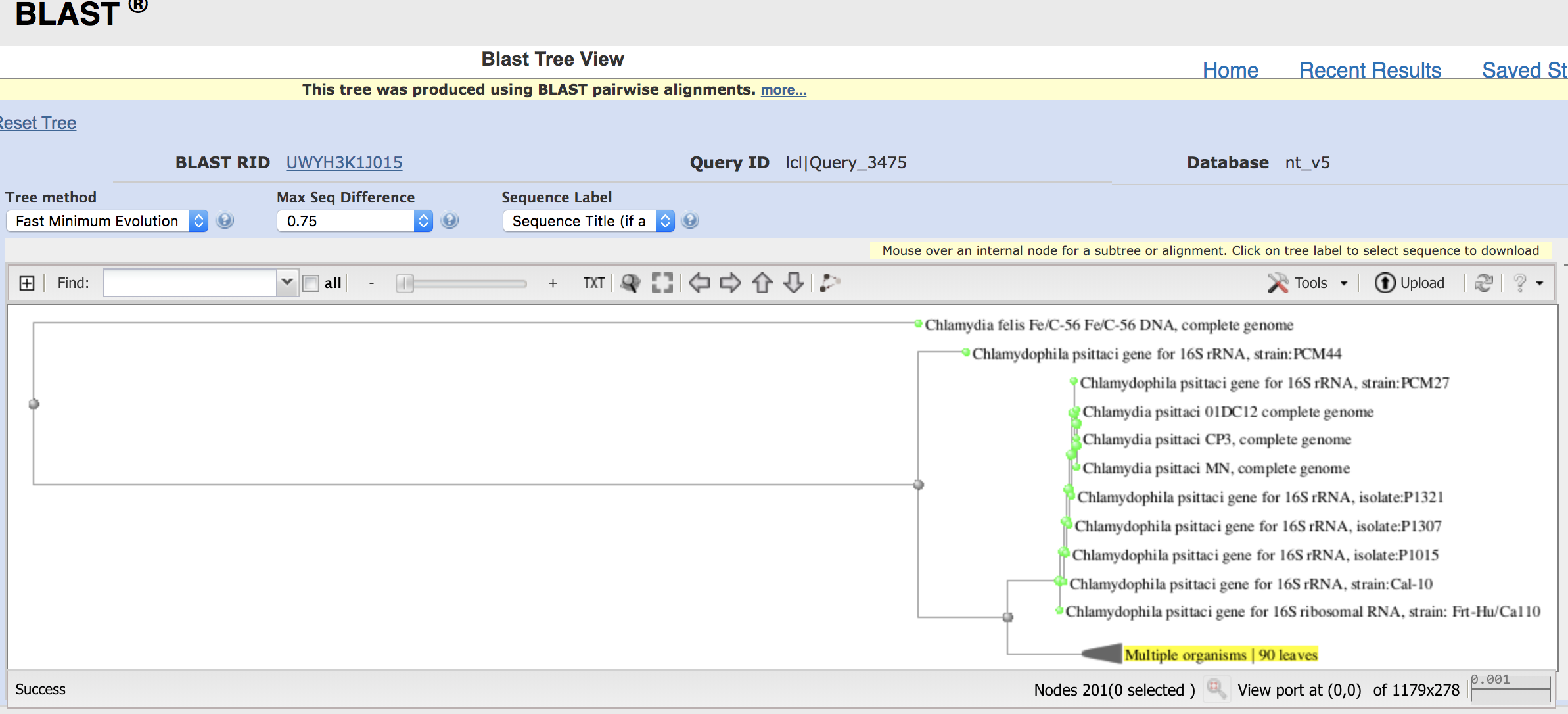
**1D)** Click on the top sequence hit, then “Alignments” then “Graphics” – what is the likely identification of the gene(s) or gene products encoded in your search sequence? *Hint: hover over the annotation marked in purple*

**16S rRNA sequence**

**1E)** Click on Other Reports> Distance Tree of Results. What do you notice about the closely related sequences?

**Take a screenshot of the distance tree and paste it here.**





**1F)** Why do you think the E-value (conservation and similarity score) of the sequences was so close to zero for this gene sequence?

**The E-value was close to zero (high exponent) indicating that the sequences are statistically unlikely to be related by chance. Because the small subunit for 16s ribosomal RNA is functional at the nucleotide level, it is highly conserved within the same or similar species and commonly used to indicate evolutionary relatedness within a genus.**

**Part 2: Identify horizontal gene transfer**

The purpose of this exercise is to use BLAST to identify horizontally transferred housekeeping genes in related species

Learning Goals:

* Use BLAST to compare evolution of other sequences to 16S identity in the same genus of bacteria
* Analyse BLAST results and compare to previous results

Return to BLAST at:

<http://blast.ncbi.nlm.nih.gov/Blast.cgi>

Carry out a **protein BLAST** using Chlamydia pneumoniae Add (adenosine deaminase) and restrict results to the refseq proteins

**>ADI88758.1 adenosine deaminase, partial [Chlamydia pneumoniae]**

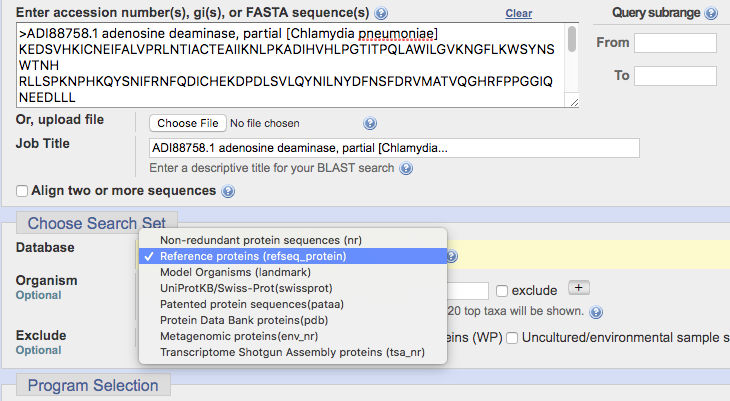
**KEDSVHKICNEIFALVPRLNTIACTEAIIKNLPKADIHVHLPGTITPQLAWILGVKNGFLKWSYNSWTNH**

**RLLSPKNPHKQYSNIFRNFQDICHEKDPDLSVLQYNILNYDFNSFDRVMATVQGHRFPPGGIQNEEDLLL**

**IFNNYLQQCLDDTIVYTEVQQNIRLAHVLYPSLPEKHARMKFYQILYRASQTFSKHGITLRFLNCFNKTF**

**APQINTQEPAQEAVQWLQEVDSTFPGLFVGIQSAGSESAPGACPKRLASGYRNAYDSGFGCEAHAGEGIE**

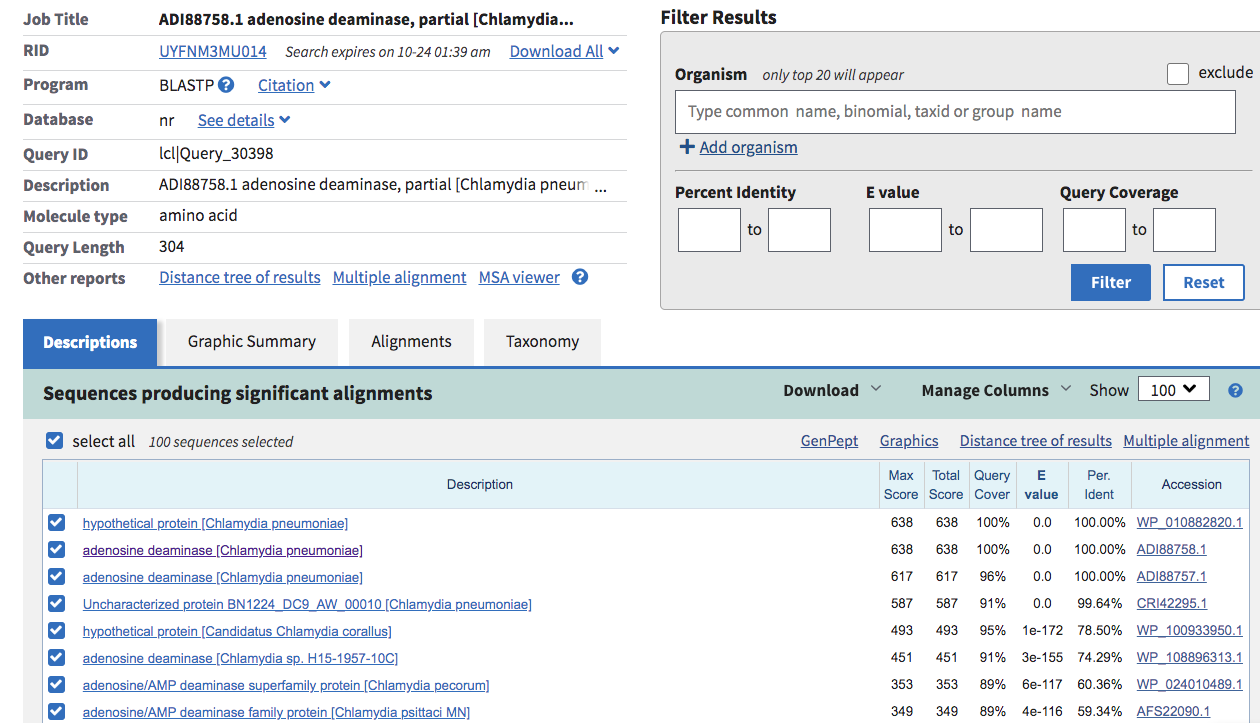
**TRTIFSSAKVNPEGLIEITRVTFS**



i) Go to <http://blast.ncbi.nlm.nih.gov/Blast.cgi> and click on nucleotide blast under BASIC BLAST and copy and paste the FASTA sequence above in the “Query Sequence Box”.

**Entrez query (optional):** This option allows you to restrict your search to a subset of entries.

*Run a blastp and wait for the results to answer the following questions*:

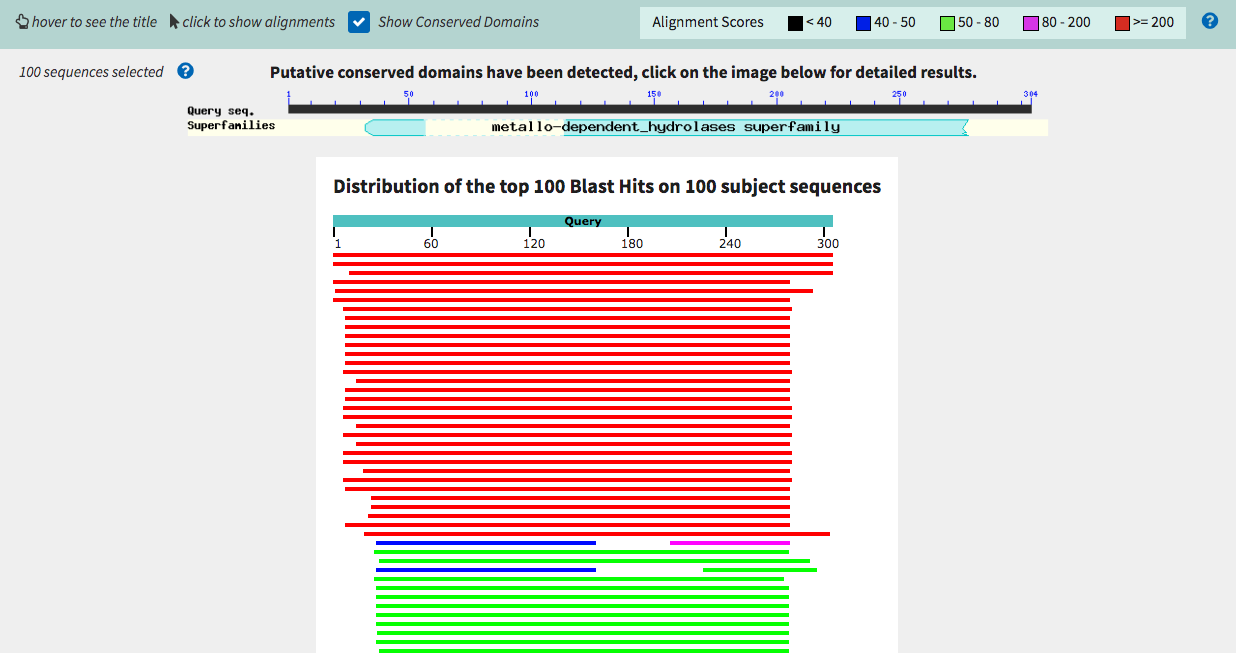


**2A)** What is the E-value for the NEXT closest sequence? Is it from the same species or not? What bacterial phyla are represented through the list (hint – click on the distance tree and look at ‘blast name’ color coding on the upper right)?

Select seq ref|WP\_100933950.1| hypothetical protein [Candidatus Chlamydia corallus] 493 493 95% 1e-172 78.50%

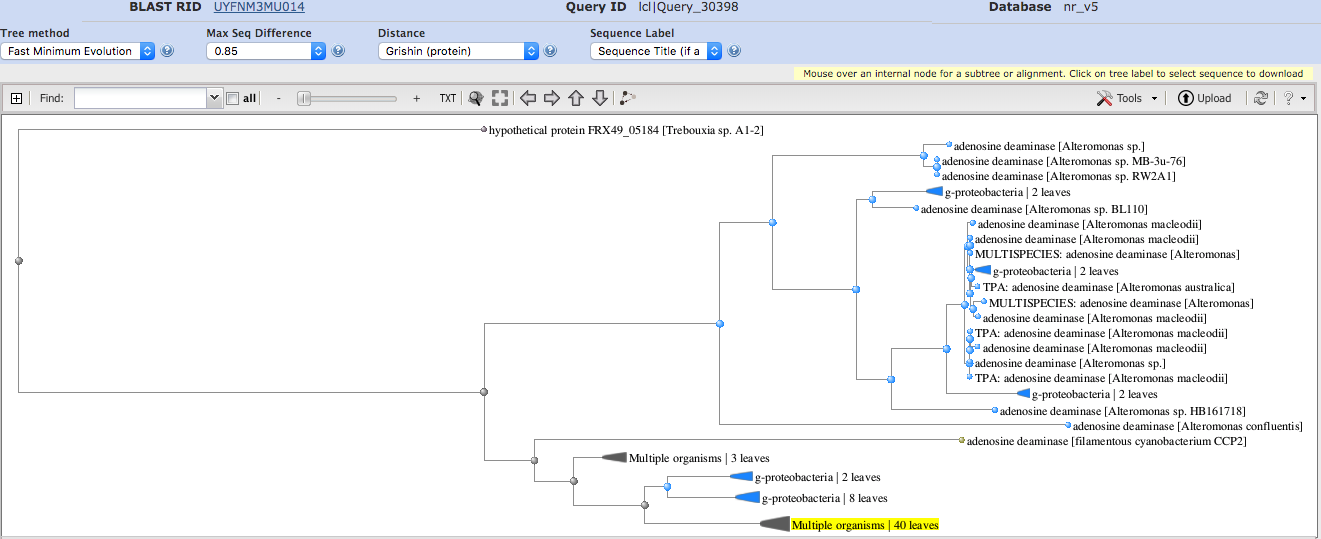
The bacterial are from Alpha, Beta, Gamma and Epsilon-proteobacteria as well as a eukaryotic plant fungus, penicillium and aspergillus and a cyanobacterium. These are not particularly closely related to Chlamydiae suggesting horizontal gene transfer of the *Add* gene. However, the Chlamydia Add genes are all closely clustered meaning they were probably acquired at one time in the lineage and passed on.

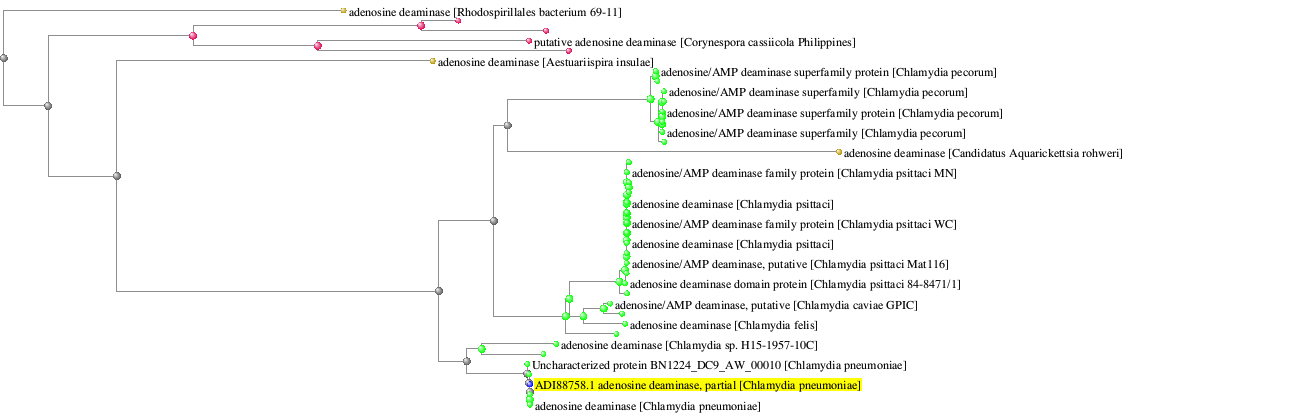
**2B)** Click on “graphic summary” – note that the colored bars represent *regions of alignment* coded by *similarity scores*. As you hover over the green bars, what do you notice about the sequence lengths and the origin of the sequences?



**The less related sequences are of different lengths and some only match part of the sequence – indicating that parts of the protein are lost in divergent species**

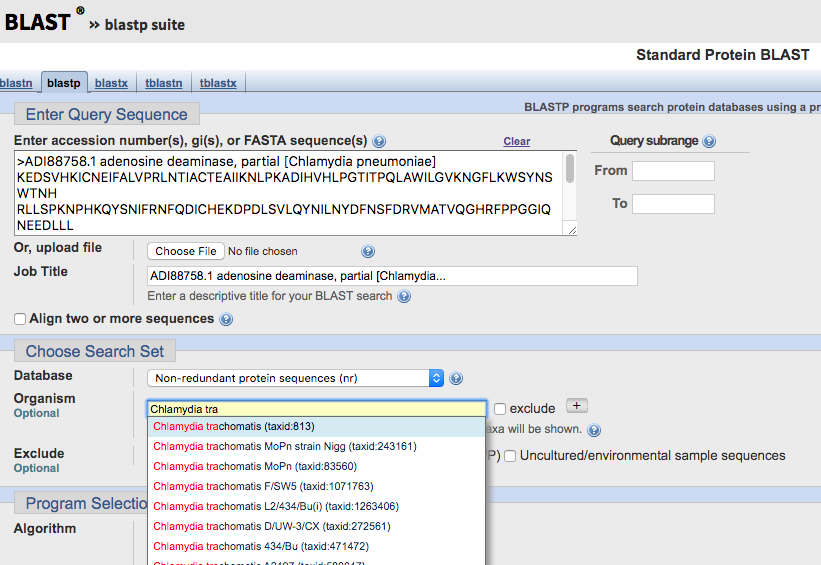
**2C)** Click on Other reports> “distance tree of results” and screen shot the tree. Compare this tree to the 16S tree in part 1. Hover over the tree in yellow and expand the subtree to see the Chlamydia sequences. What do you notice?





**The Add gene is quite similar in a number of species, and the many species in the Chlamydia genus to not appear to share the Add gene.**

**2D)** Redo the protein blast, but this time select only Chlamydia trachomatis (the STI and ocular infection associated Chlamydia) as the search genome(s). \*be sure to select the non redundant protein sequences\* What is your result? **What can you conclude about the origin of the C. pneumoniae and C. trachomatis *Add* genes?**



**Chlamydia trachomatis has a different Add type gene with only a small matching region. Either the C.t. version of the gene has been changing a lot, or the two alleles of the Add gene have different origins.**

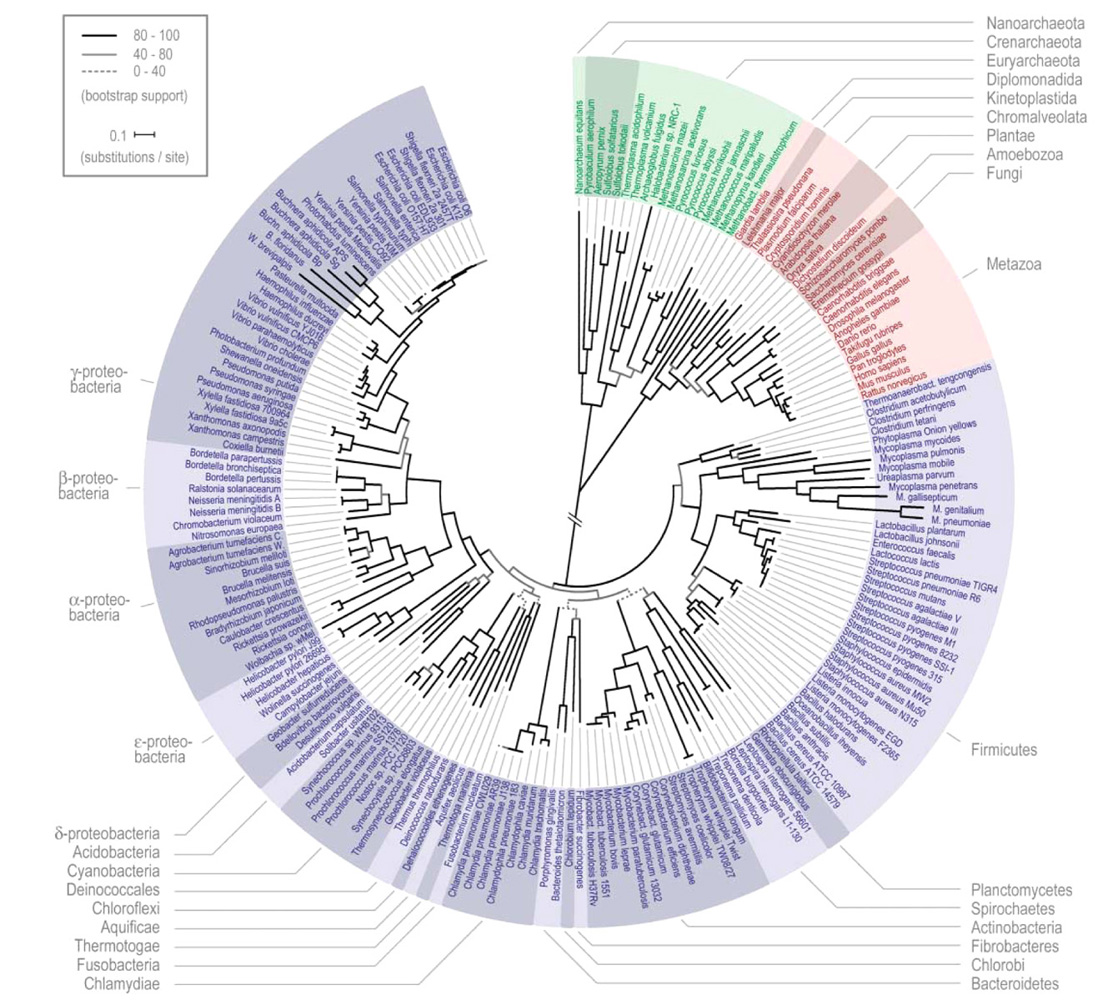
Background:

the Chlamydiae <https://en.wikipedia.org/wiki/Chlamydiae>

Koala chlamydia <https://www.livescience.com/62517-how-koalas-get-chlamydia.html>

And Prescott p875-876

Reference tree: phyla of bacteria by 16S sequences



**Part 3: PRIP sequence and its origin**

The purpose of this exercise is to use BLAST to compare the distribution of unknown genes within the chlamydiae

Learning Goals:

* Use BLAST to find Chlamydia with integrated phage sequences
* Analyse BLAST results to compare phage-derived genes and their distribution

Reference: <https://www.tandfonline.com/doi/full/10.4161/21597073.2014.965076>

Rosenwald AG, Murray B, Toth T, Madupu R, Kyrillos A, Arora G. Evidence for horizontal gene transfer between Chlamydophila pneumoniae and Chlamydia phage. Bacteriophage. 2014 Dec 15;4(4):e965076. eCollection 2014. PubMed PMID: 26713222; PubMed Central PMCID: PMC4589997.

Start with this sequence from Chlamydia pneumoniae;

>Chlamydia pneumoniae\_LPCoLN PRIP-like protein

MRELNAFELTQPEEYRNRWVLMPCLKCRFCRTQHAKVWSYRCVHEASLYEKNCFLTLTYDDKHLPQYGSLVKLHLQLFLKRLRDRISPHKIRYFGCGEYGTKLQRPHYHLLIFNYDSLLDG

Make a protein-protein BLAST search to find other proteins with similar sequences

Under ‘descriptions’ Select the 8-10 most similar sequences as well as 1-3 with less than 60% percent identity

**3A)** Click on ‘distance tree of results’ at the upper right corner of DESCRIPTIONS AFTER selecting a few sequences of interest. Screen shot the tree.

Hint; <https://en.wikipedia.org/wiki/Microviridae>

If BLAST will not let you get a subset distance tree for some reason, you can do this analysis with the tree of all results.



**3B)** What is the evolutionary origin of this PRIP gene? Did it come from Chlamydiae to the phage, or the other way around?

It came from the phage

**3C)** What supports your conclusion?

The ancestral genes at the root of the tree are all viral in nature, in addition the branch of the tree with Chlamydia sequences also has a chlamydia phage gene. Finally, like the Add gene results not all Chlamydia species have this PRIP gene.

**Part 4: Conclusions**

**4A)** Why does the distribution of PRIP, 16S and Add genes differ among the closely related chlamydiae?

Each gene was derived from a different kind of gene transfer; vertical or horizontal and thus they are distributed differently among the species in the same genus.

**4B)** Suggest mechanisms for the evolutionary origin and method of acquisition (e.g. vertical or horizontal and if horizontal what method) of each gene among the Chlamydiae.

**16S**

**Vertically transmitted from the ancestor (LUCA) of chlamydiae**

**Add**

**Horizontally transmitted from another bacterium, probably by natural competence**

**PRIP**

**Horizontally transmitted by phage transduction from the chlamydia phage (microviridae)**