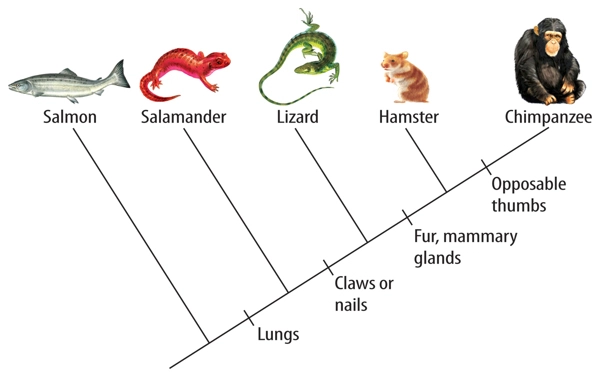
BLAST Lab

Follow the instructions for this lab (adapted from <https://lhsapbiologyclass.wordpress.com/lab-info-resources/blast-lab/>). The directions written in ***bold and italics*** are items that need to be recorded either on a Google Document on your piece of paper.

Introduction: Comparing DNA Sequences to Understand Evolutionary Relationships with BLAST

How can bioinformatics be used as a tool to determine evolutionary relationships and to better understand genetic diseases?

In this laboratory investigation, you will use BLAST to compare several genes, and then use the information to construct a cladogram. A cladogram is treelike, with the endpoints of each branch representing a specific species. The closer two species are located to each other, the more recently they share a common ancestor. In the cladogram below, you will also see the shared derived characters. Note that the placement of the shared derived character corresponds to when (in a general, not a specific sense) that character evolved; every species above the character label possesses that structure.

1. You can do this lab as a hand-written assignment or online. If you do it online, open the document on my website that says “Blast Lab – Student answer sheet”. I have linked other websites on my page that can help you review too!
2. ***Read and answer questions 1, 2a, and 2b on the student answer sheet.***

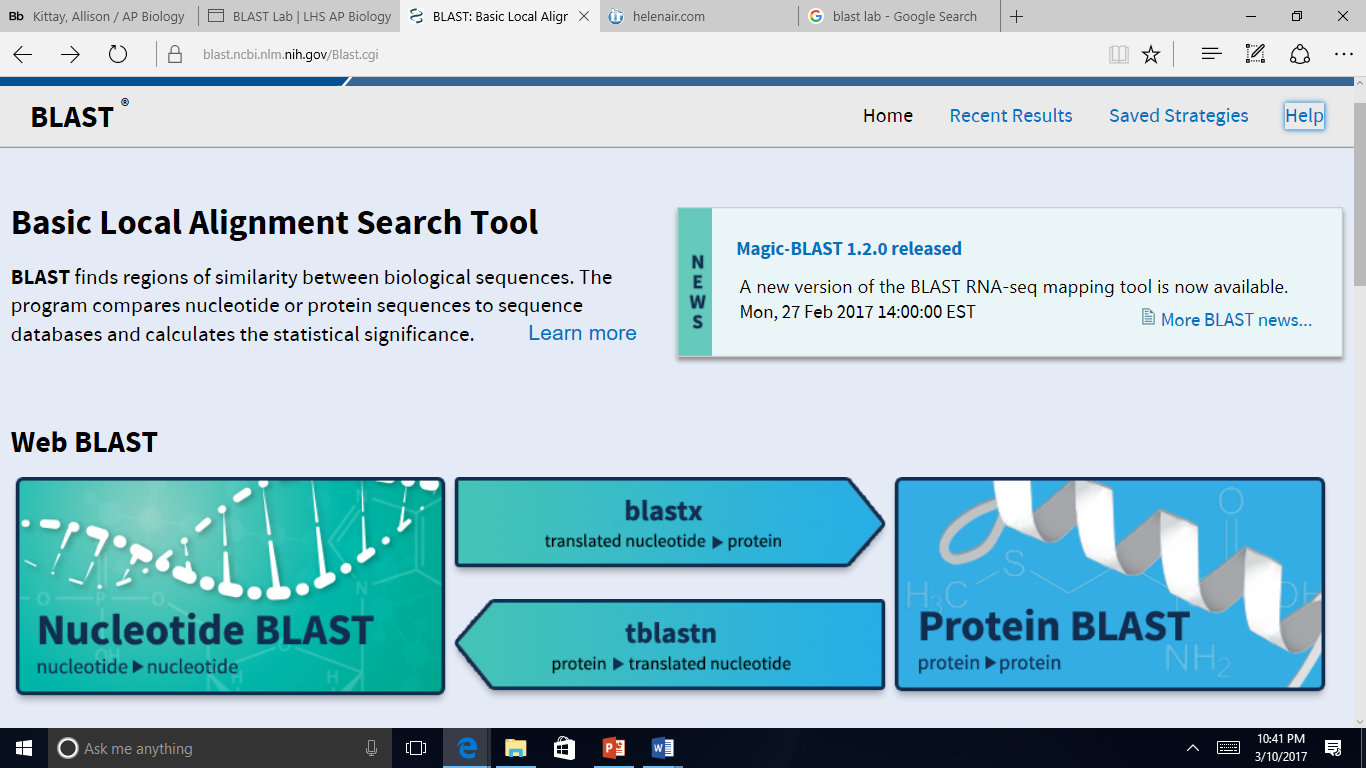
Next, you will be using the BLAST database to gather information about this fossil. Little is known about this fossil, and It appears to be a new species. Upon careful examination of the fossil, small amounts of soft tissue have been discovered. Normally, soft tissue does not survive fossilization; however, rare situations of such preservation do occur. Scientists were able to extract DNA nucleotides from the tissue and use the information to sequence several genes. Your task is to use BLAST to analyze these genes and determine the most likely placement of the fossil species on Figure 4 (on your answer sheet.) A team of scientists has uncovered the fossil specimen in Figure 3 near Liaoning Province, China. Make some general observations about the morphology (physical structure) of the fossil, and then record your observations in your notebook.

1. ***Describe where you think the fossil belongs on the cladogram (remember bold instructions are items you need to answer on the student answer sheet).*** (Drawing the cladogram is optional.)

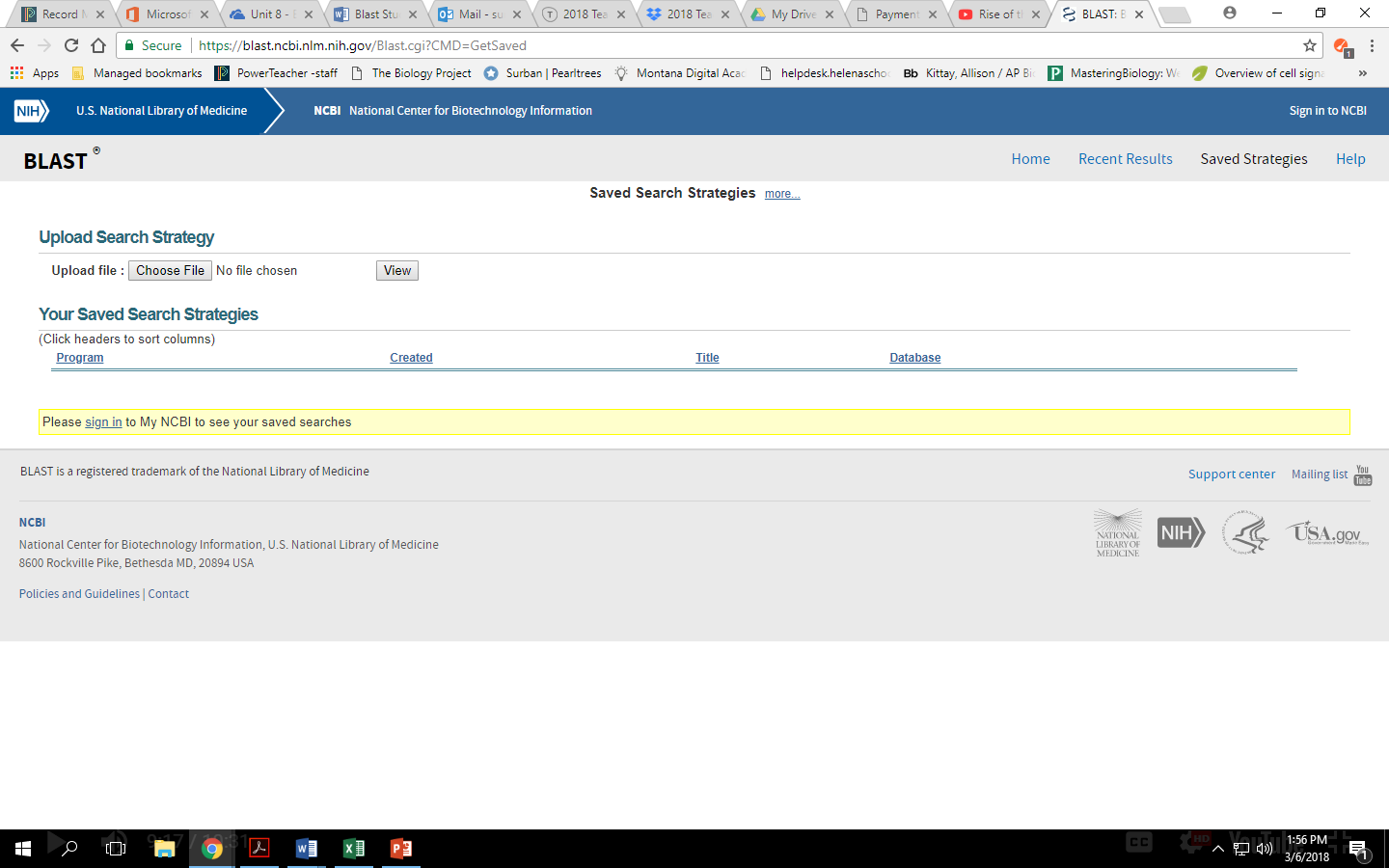
Follow the procedure as outlined below. I updated the screenshots but let me know if you have problems.

1. Open the website labeled **“Gene Files”** on my page (<https://lhsapbiologyclass.wordpress.com/lab-info-resources/blast-lab/>). Scroll down until you see the title “Direct link to files:”. Download the four gene files (they are in orange font). You only need to download the files. **Do not open them.**
2. Go to the BLAST homepage (also linked on my page) and follow all these directions!!! https://blast.ncbi.nlm.nih.gov/Blast.cgi

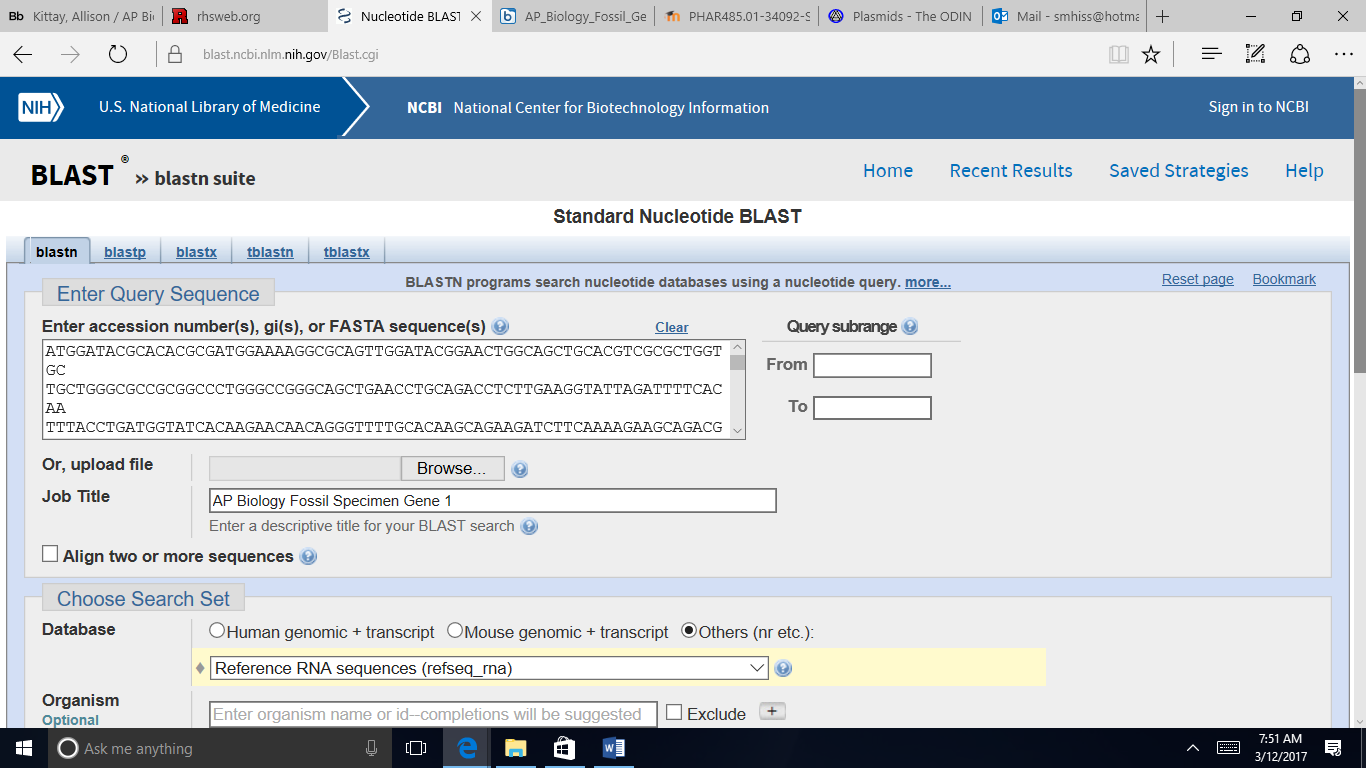
Click on Saved Strategies (at the top). Notice you could use a nucleotide BLAST or a protein BLAST. In this lab we are using a nucleotide BLAST! (See screenshot on the next page.)

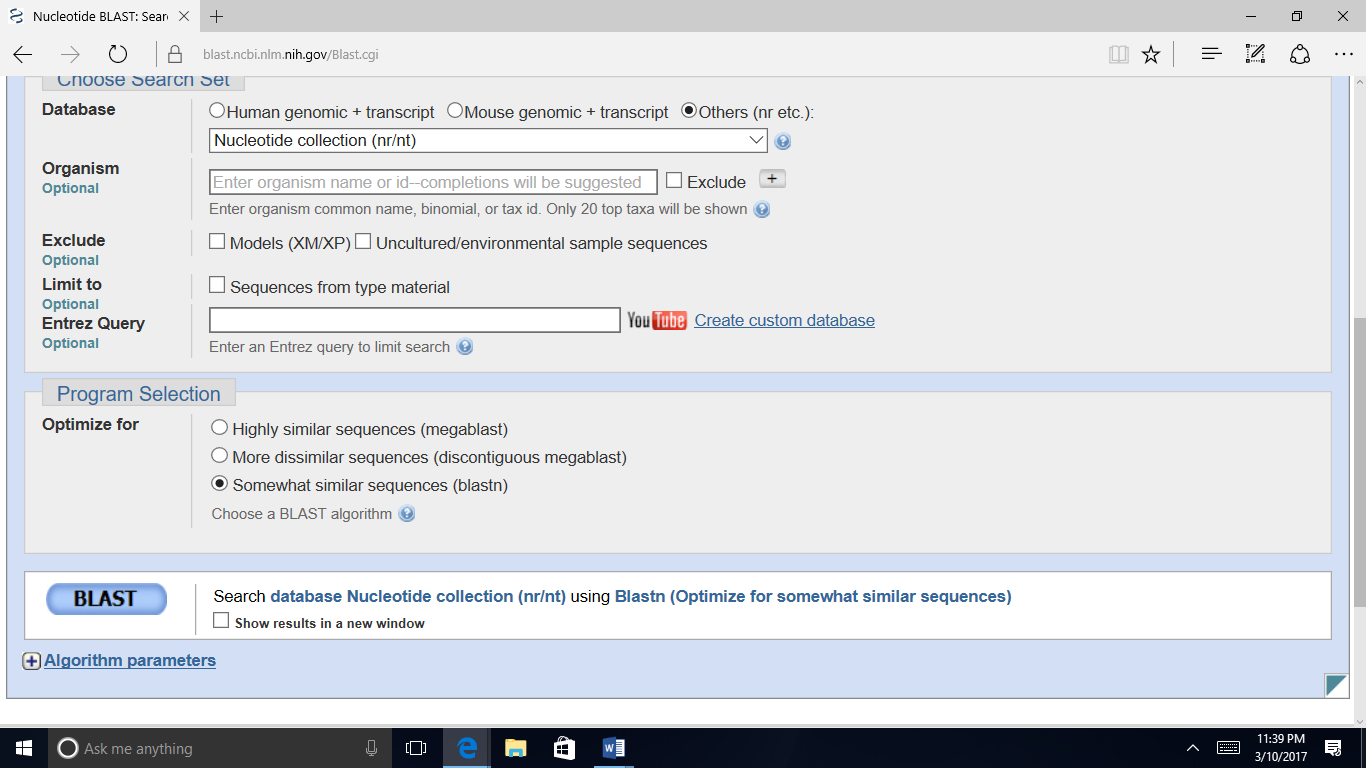


Click “Choose File” and upload your file. (The other settings are fine). Then Click “View”

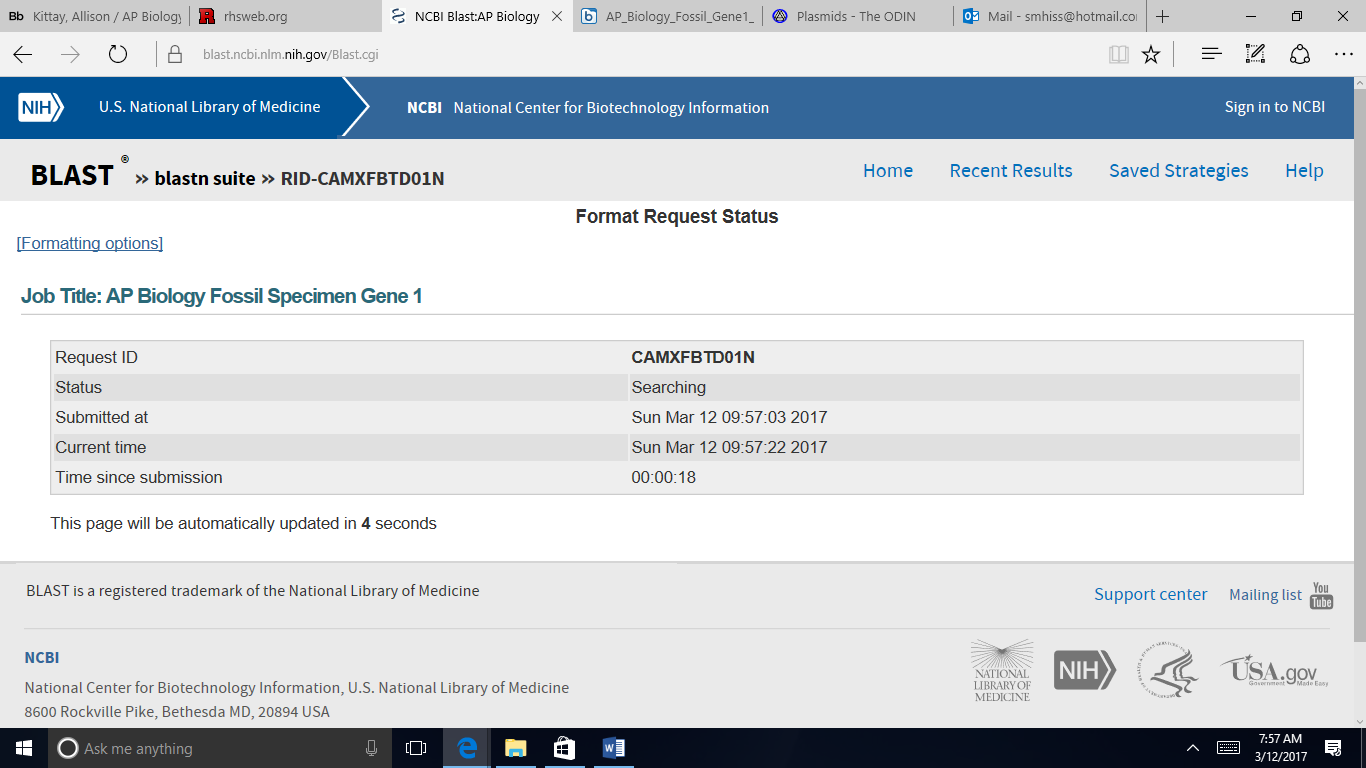


It will then look something like this: Leave all the settings alone and click on “BLAST”





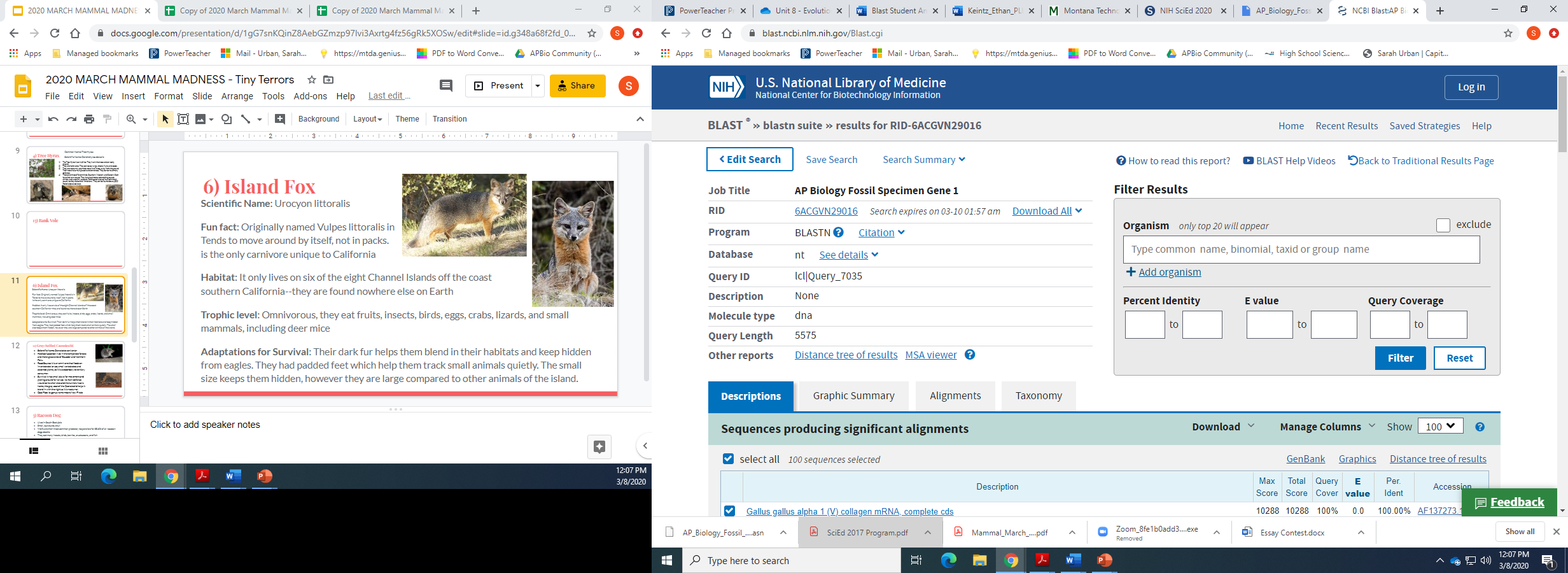
Wait for the database to search for sequences (usually under 30 seconds – but may be over a minute – it took 10 minutes for me once!). It will look like this



Analysis of Sequences – Graphic Summary The chart **below** is a graphical summary of your first sequence.

The first line ( ) under the tab **“Descriptions”** represents the most similar sequence that BLAST was able to pull from the data bases it searched. If you hover over the line, you will see the species of organism that this sequence was derived from. Other, less similar sequences are included from top (most similar) to bottom (less similar).

Descriptive Summary below lists from top to bottom the information about each of the organisms that were represented in the graphic summary. The species in the list are those with sequences identical to or most similar to the gene of interest. NOTE: Species with common ancestry will share similar genes. The more similar genes two species have in common, the more recent their common ancestor and the closer the two species will be located on the cladogram.



“**Descriptions” Tab** - Use this list to see what the numbers in the chart mean!

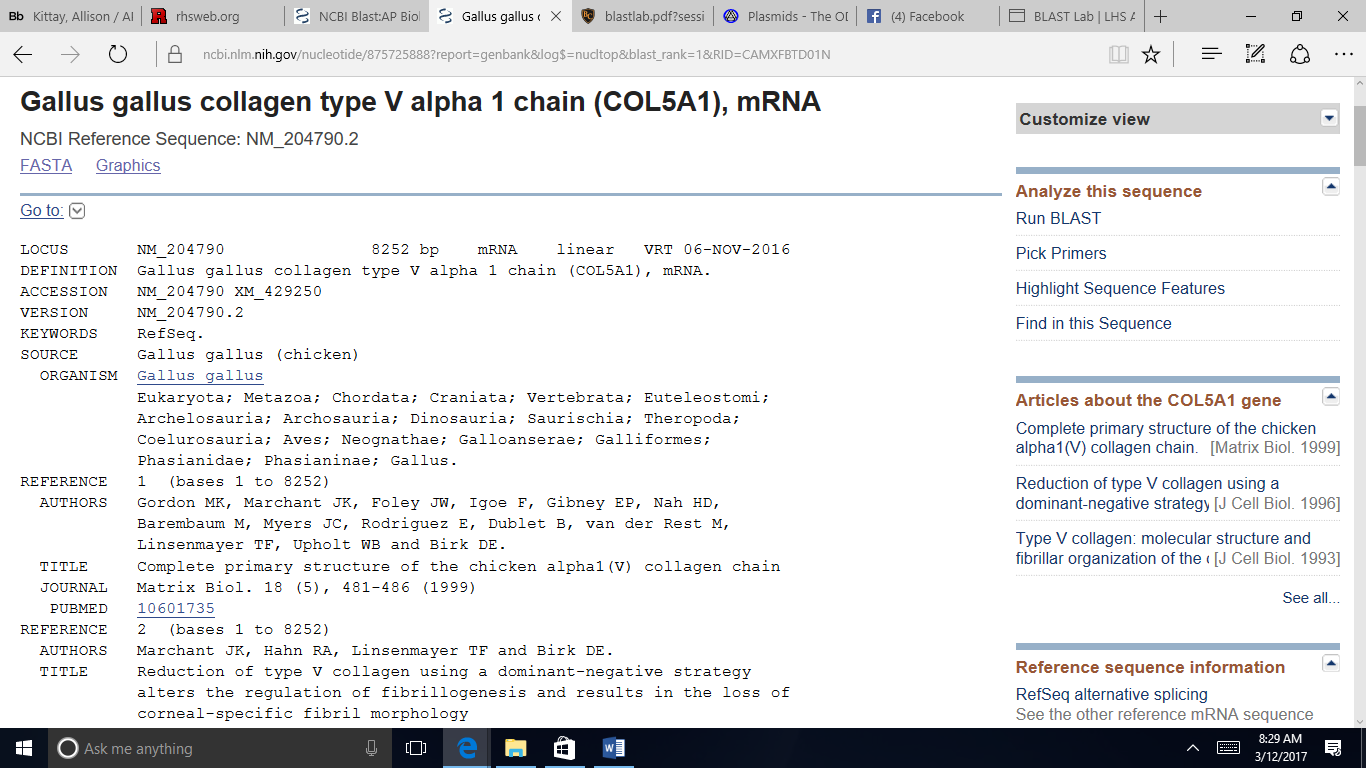
Max(imum) Score: ( ) the highest alignment score of a set of aligned segments from the same subject (database) sequence. This normally gives the same sorting order as the E Value. The higher the max score, the closer the alignment.

E(xpect) Value : ( ) the number of alignments expected by chance with a particular score or better. It is sort of like a control for your hypothesis. The lower the e value, the closer the alignment. Sequences with e values less than 1 e - 04(1x10 - 4) can be considered related with an error rate of less than 0.01%.

Ident(ity) O ( ) The extent to which two (nucleotide or amino acid) sequences have the same residues at the same positions in an [alignment](https://www.ncbi.nlm.nih.gov/books/NBK62051/def-item/alignment/), often expressed as a percentage. The higher the %, the more identical.

Accession : ( )If you click on the Accession number for a particular species listed, you will get a full report that include the classification scheme of the species, the research journal in which the gene was first reported, and the sequence of bases that appear to align with your gene of interest. It will identify the gene, in this case it is collagen, that you are working with and you will also see the common name of the organism if it has one!

If you click on the link under the “Accession” label, (for the first gene it is “[AF137273.1](https://www.ncbi.nlm.nih.gov/nucleotide/AF137273.1?report=genbank&log$=nucltop&blast_rank=1&RID=6ACGVN29016)” it will give you information about the organism!!! Here is a screen shot of the first gene.

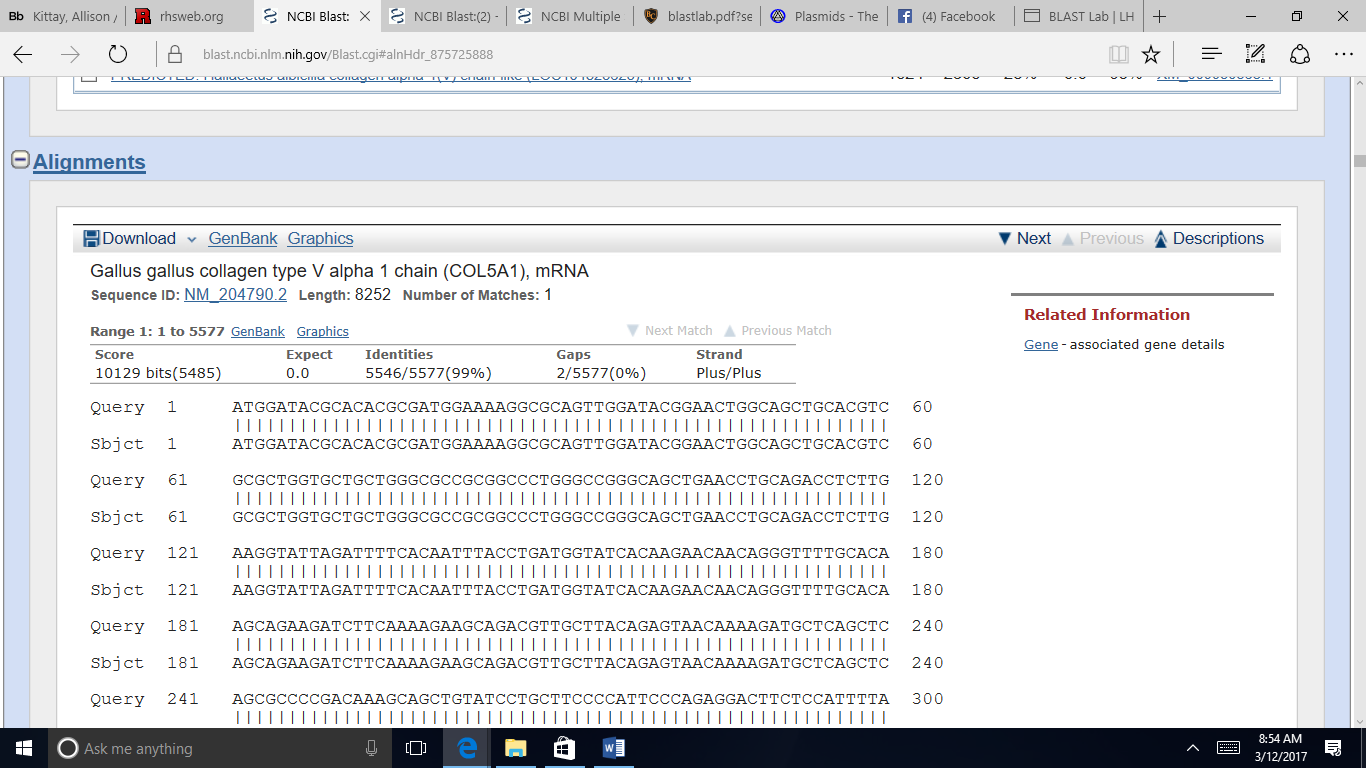


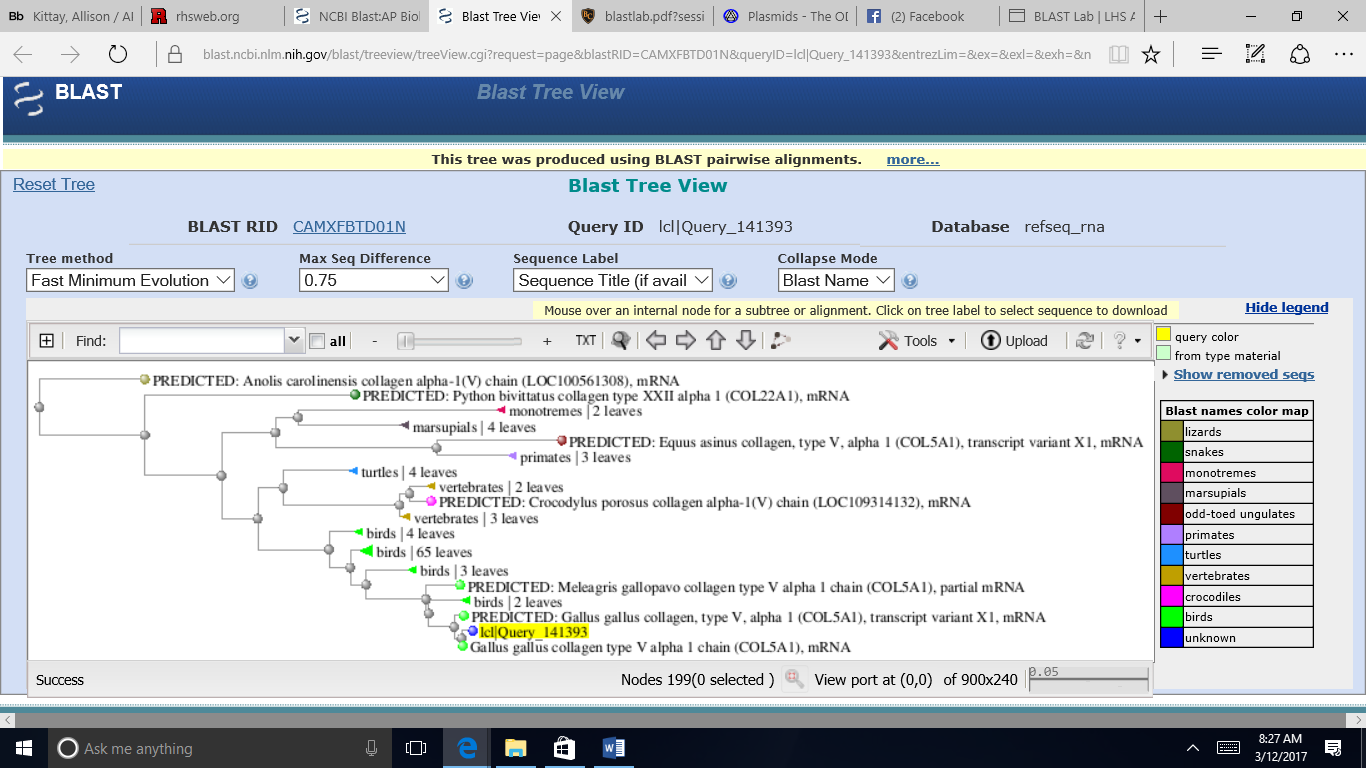
Protein name

Species Name (common name)

Classification (Domain, etc.)

**Alignments Tab:** If you click on the tab “Alignments “ ( ) it shows the actual nucleotide comparisons between your unknown or “query” sequence and the most common sequence belonging to *Gallus gallus*.



The results are tedious at best but there is a quick way for you to build a cladogram for the sequences. Scroll back up to the top of the analysis page. Right above the tabs (specifically the “graphic summary” there is a link that says “Distance tree of results” ( ) to see how this gene aligns with other species. If you want to save a picture of your tree, click on “Tools” then save your document as a PDF file. (This screen shot was from last year….yours might look a bit different)

Recall that species with common ancestry will share similar genes. The more similar genes two species have in common, the more recent their common ancestor and the closer the two species will be located on a cladogram. As you collect information from BLAST for each of the gene files, you should be thinking about your original hypothesis and whether the data support or cause you to reject your original placement of the fossil species on the cladogram. For each BLAST query, consider the following:

* The higher the score, the closer the alignment.
* The lower the e value, the closer the alignment.
* Sequences with e values less than 1e-04 (1 x 10-4) can be considered related with an error rate of less than 0.01%.

***6) For each of the genes, fill out the chart on your answer sheet. Repeat for all four genes.***

***7) After you have BLASTed all of the genes, answer the rest of the lab questions on your document.***

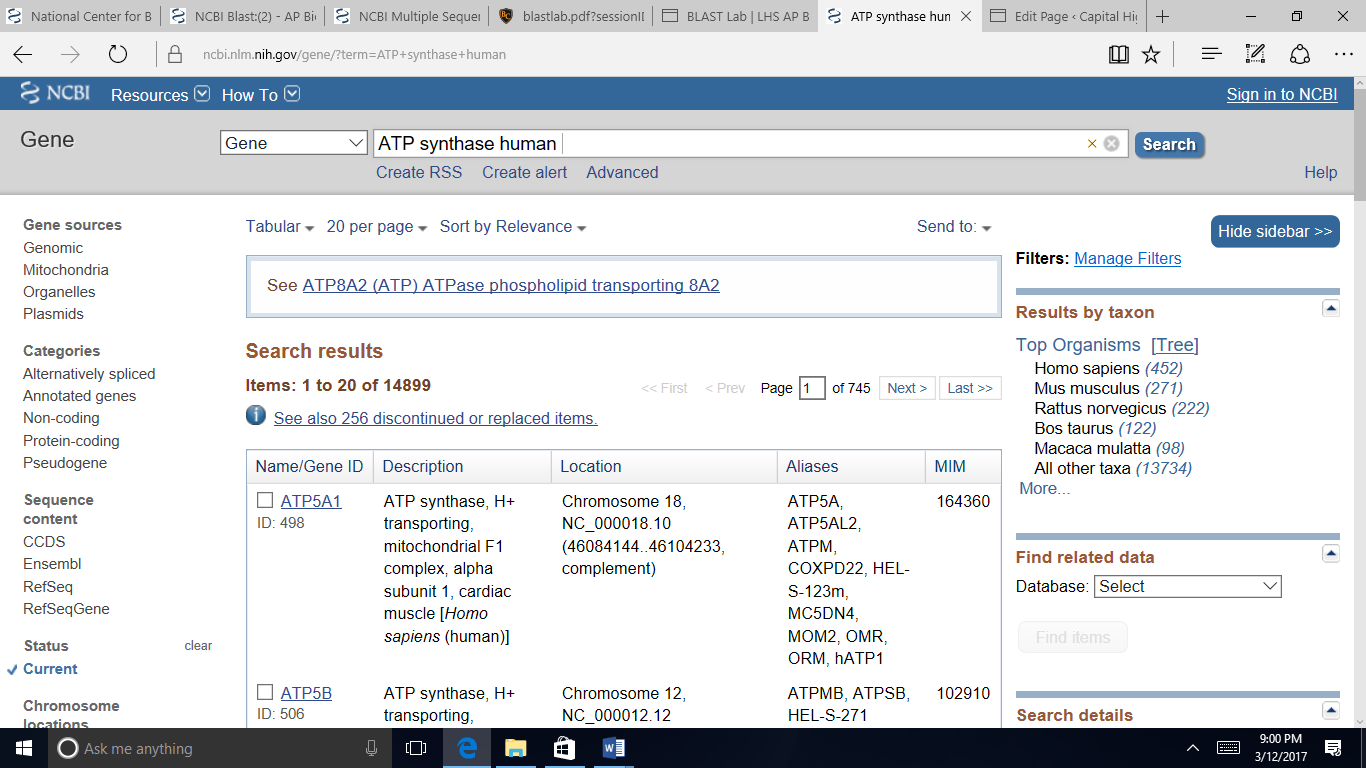
PART II – Designing Your Own Investigation

Now that you have completed this investigation, you should feel more comfortable using BLAST. The next step is to learn how to find and BLAST your own genes of interest. If you select a human gene, BLAST will compare that gene sequence to any similar sequences in the databases. You will be researching the closest match to your human gene.

1. To locate a gene, you will go to the Entrez website https://www.ncbi.nlm.nih.gov/gene/

2. Search for a gene that you know is present in humans. For example, you might use human actin, myosin, catalase, keratin or ubiquitin – or another one that you can pick that we throughout the year. You can use gene names like Pax1, SRY1. Enzymes are also proteins so you might search “human ATP synthase”. You can decide which gene you want to use. Type the name of the human gene in the search bar and click on “search.”

3. In the previous assignment, you were given gene sequences to put into BLAST. In this assignment, you have used Entrez to find a segment you want to search for. The page you are on now, lists the results of your search. You should be able to select a reasonable sequence by clicking on the heading.



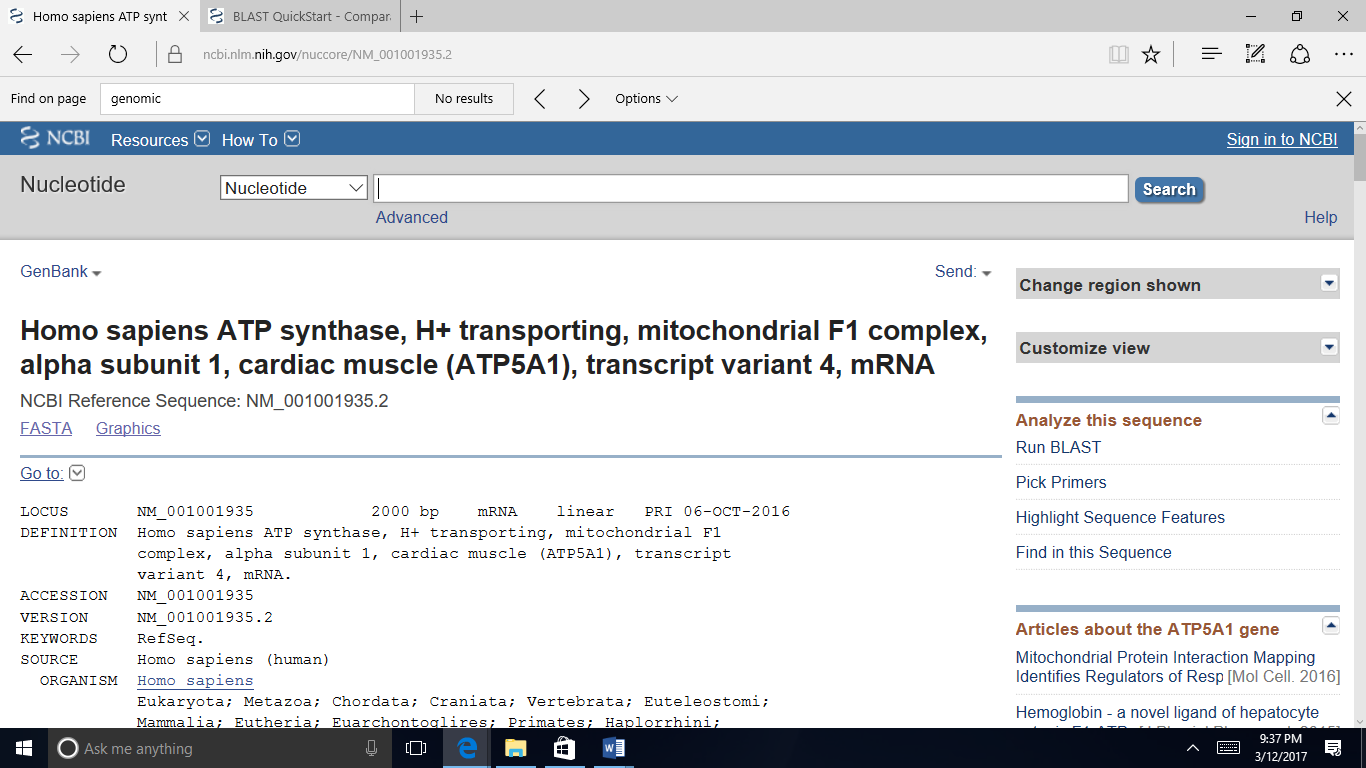
4. This next page has a lot of information about your gene but scroll all the way down to, “NCBI Reference Sequences (RefSeq).” (See the diagram below!)



One of the things you can do is to select a gene sequence FASTA file but, try something new. You can choose a file under “Genomic”( ) or “mRNA and Proteins.” ( ) Click on one of the selections.



In the next page you should either be able to click on “FASTA” or there may be a menu on the right side of the screen that allows you to directly BLAST the sequence. NOTE: FASTA format is a text-based format for representing either nucleotide sequences or peptide sequences, in which nucleotides or amino acids are represented using single-letter codes – ACGT for nucleotides; ACDEFGHIKLMNPQRSTVWY for amino acids.



BLAST the data! If it works, skip to the questions…..if you can’t figure it out, you can copy and paste either the DNA or amino acid sequence by following) the directions below.

a. You will see a sequence of DNA nucleotides complimentary to the mRNA for the protein you have selected. Copy this sequence (only the nucleotides), and return to BLAST http://blast.ncbi.nlm.nih.gov/Blast.cgi

b. Select “Nucleotide BLAST” and, instead of selecting “Saved strategies” just paste your sequences into the rectangle box labeled “Enter Query Sequences.” You will be pasting a FASTA sequence into the box.

c. You can change some of the parameters of your analysis at this point. For example, under “Choose Search Set.” You can select whether you want to search the human genome only, the mouse genome only, or all genomes available. Under “Program Selection,” you can choose whether or not you want highly similar sequences or somewhat similar sequences. Choosing “Somewhat similar sequences” will provide you with more results but your tree will be much more complex.

d. Click on BLAST and you will get the analysis of your gene of interest.

PART II Questions: **(Answer on your answer sheet)**

1) Title the cladogram based on your gene of interest, you can insert it into your Google Doc WORD or copy it on your paper.

2) What is the function in humans of the protein produced from the gene you selected?

3) Would you expect to find the same protein in other organisms? If so, which ones – why? Which other organisms had gene sequences most similar to the human gene you selected?

4) Is it possible to find the same gene in two different kinds of organisms but not find the protein that is produced from that gene? Why might this happen?

5) If you found the same gene in all organisms you test, what does this suggest about the evolution of this gene in the history of life on earth?