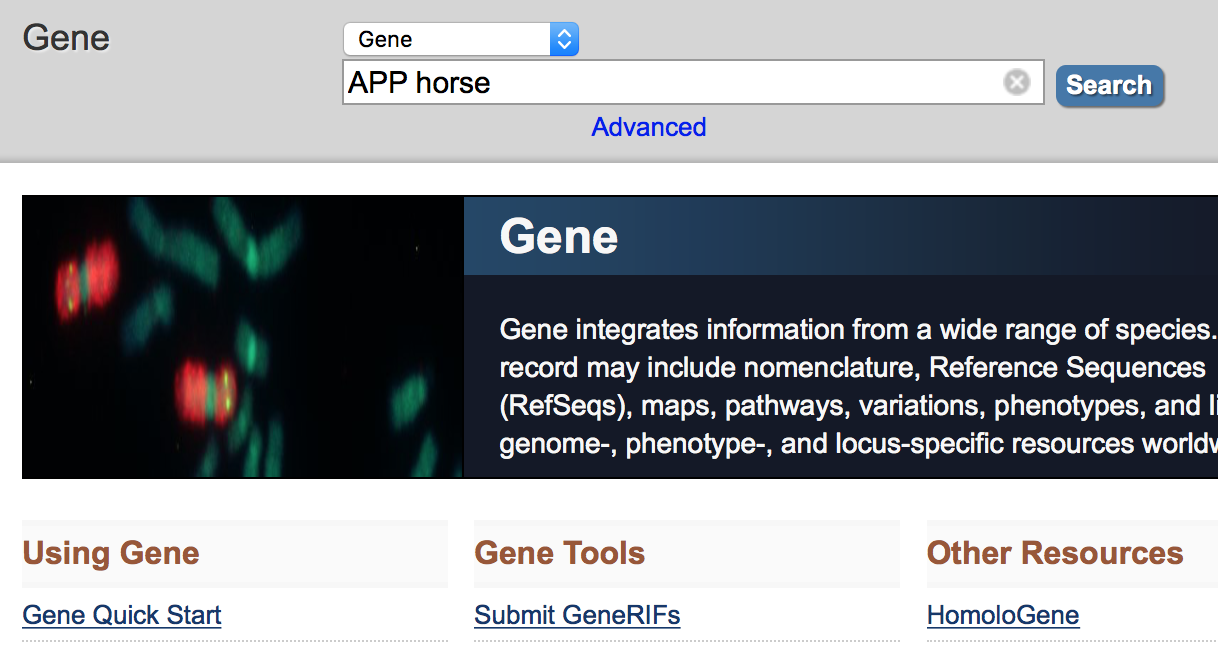
**FINDING EXONS- NCBI & BLAST**

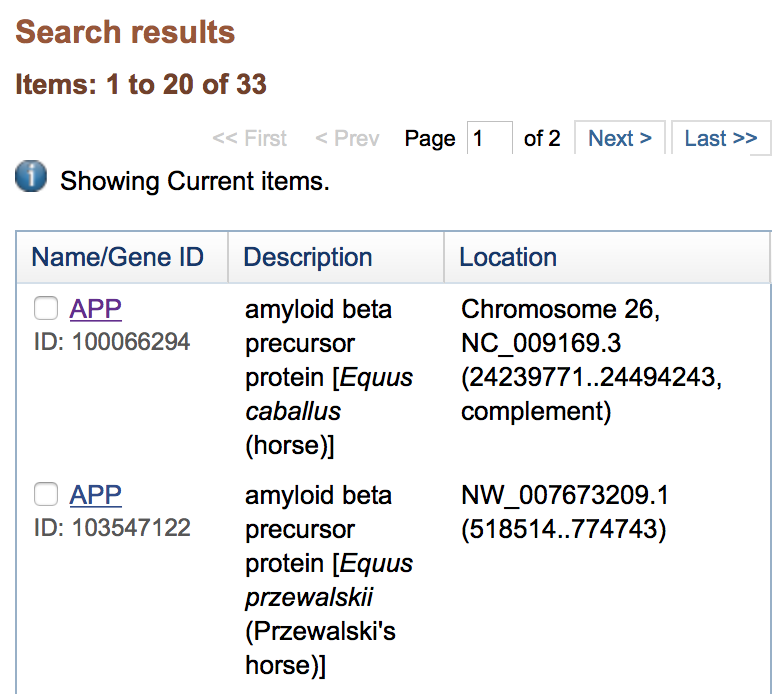
The goal of this tutorial is to get you comfortable with locating gene sequences using NCBI and BLAST for manual annotation.

1. Start by logging onto NCBI gene: <https://www.ncbi.nlm.nih.gov/gene> -Enter the gene name & the species (equus caballus or horse) and hit Search

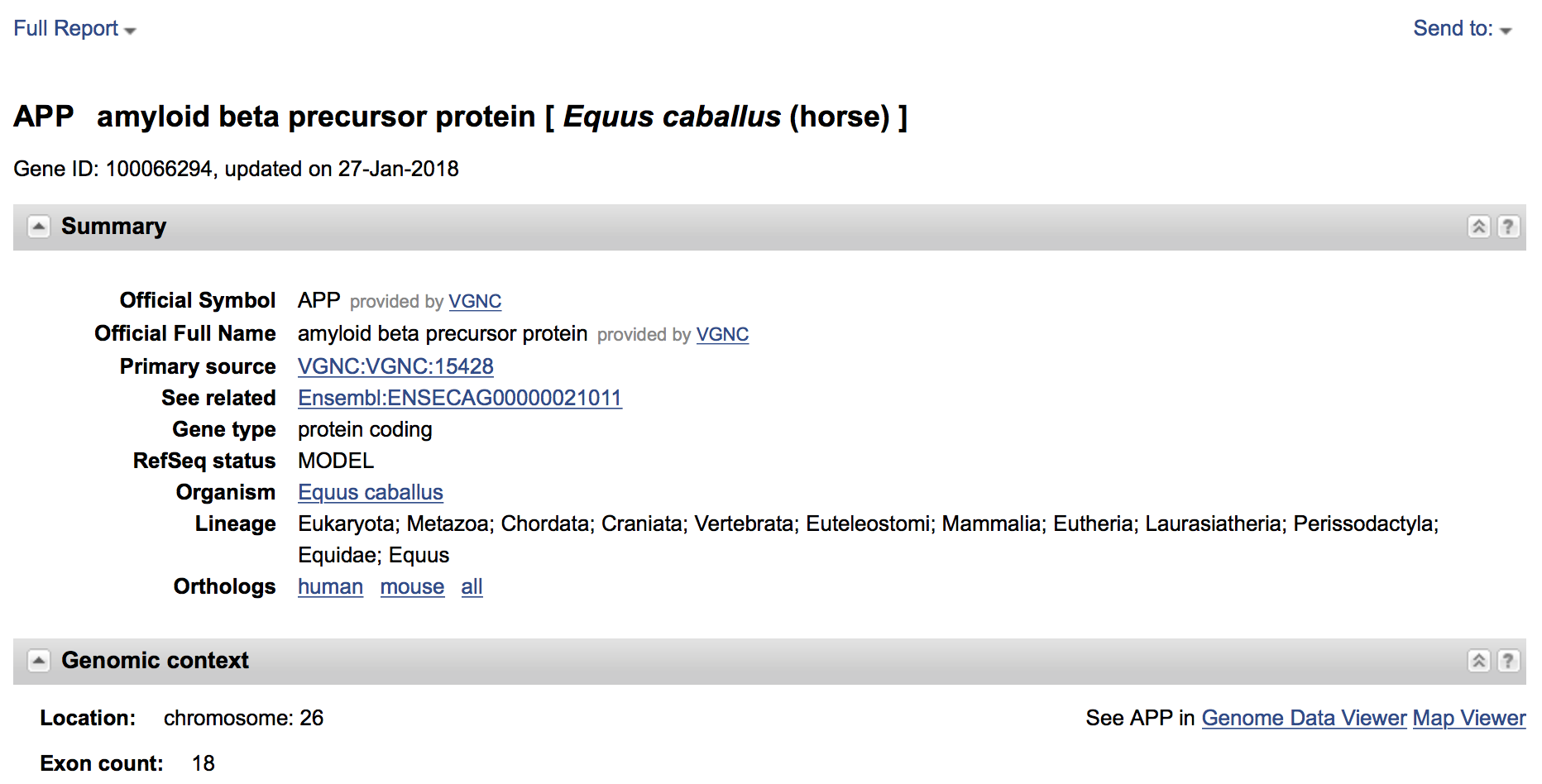
-This pulls up the reference genome (EquCab 3.0) which is a whole genome sequence for one individual horse that we can use for comparison. Because this genome is from one individual there are potential broken genes. Part of the process with manual annotation is determining if the differences between the horse & human genome are normal species variation or if the changes could indicate a potential mutation in a region. The average person also has a bunch of potentially deleterious mutations, such as premature stop codons, frameshifts, missense mutations in highly conserved places that don’t end up leading to clinical disease.

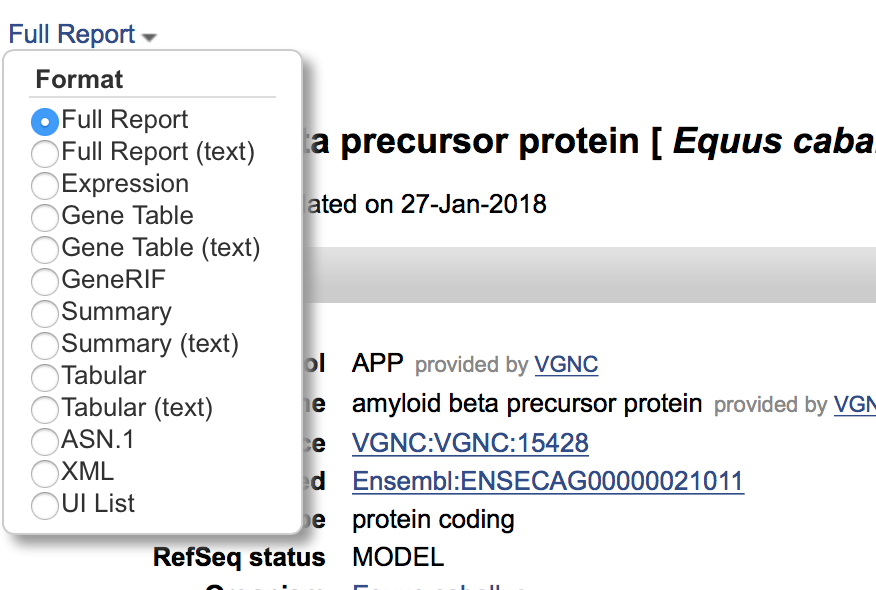


1. Select the equine Gene from the search result list:

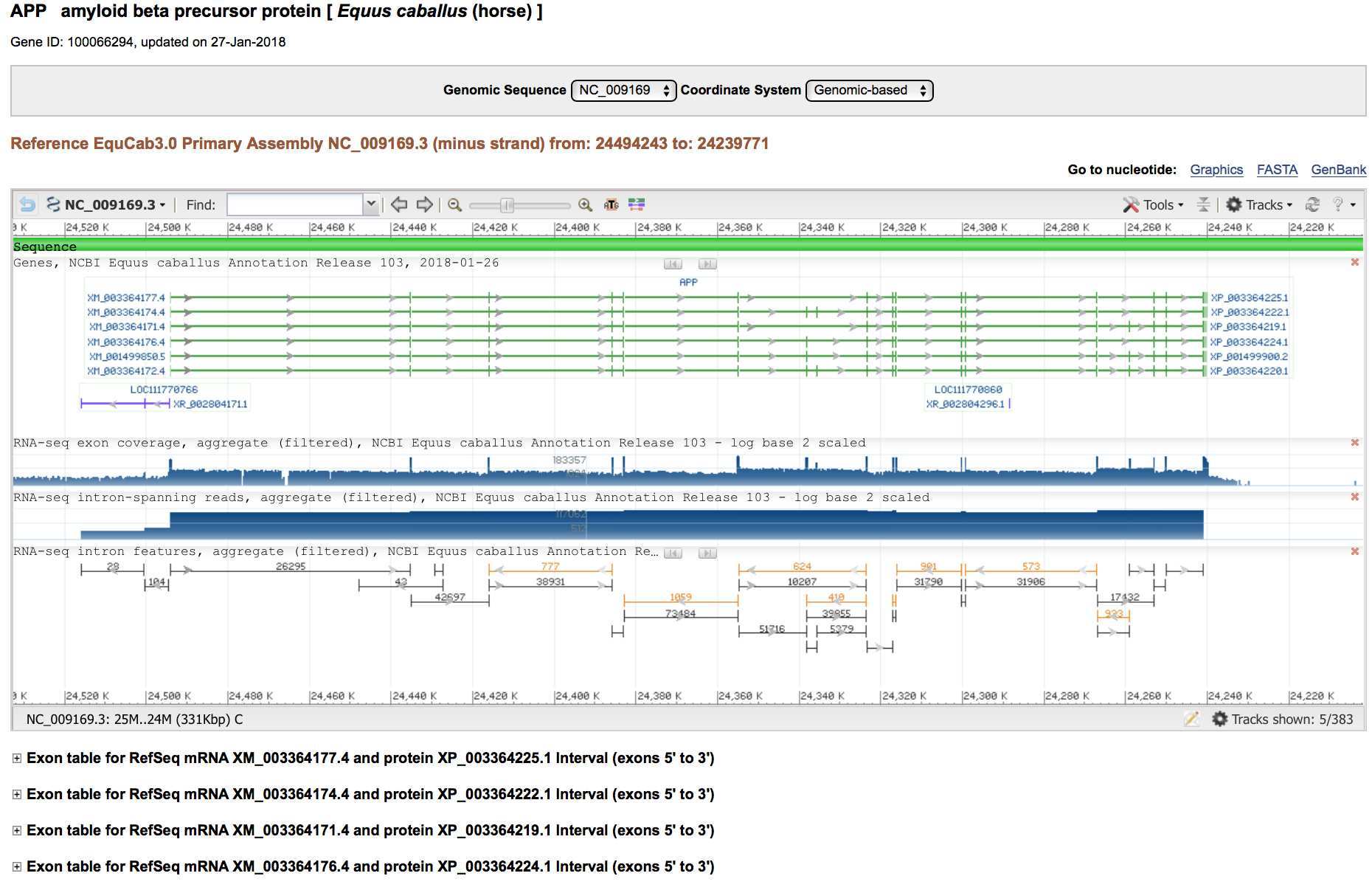


-Information is available on this screen including the description, location and position (Chr 26: 24,239,771-24,494,243).

1. Once the gene is selected a lot more information can be obtained including the description of the gene, gene type, location and exon count. Continuing on this page will show diagrams of the gene, exons etc. 
2. Select the “Full Report” drop down in the upper left corner and change the format to “Gene Table”. This will bring up a screen with the associated transcripts and exons/coding sequence for review.



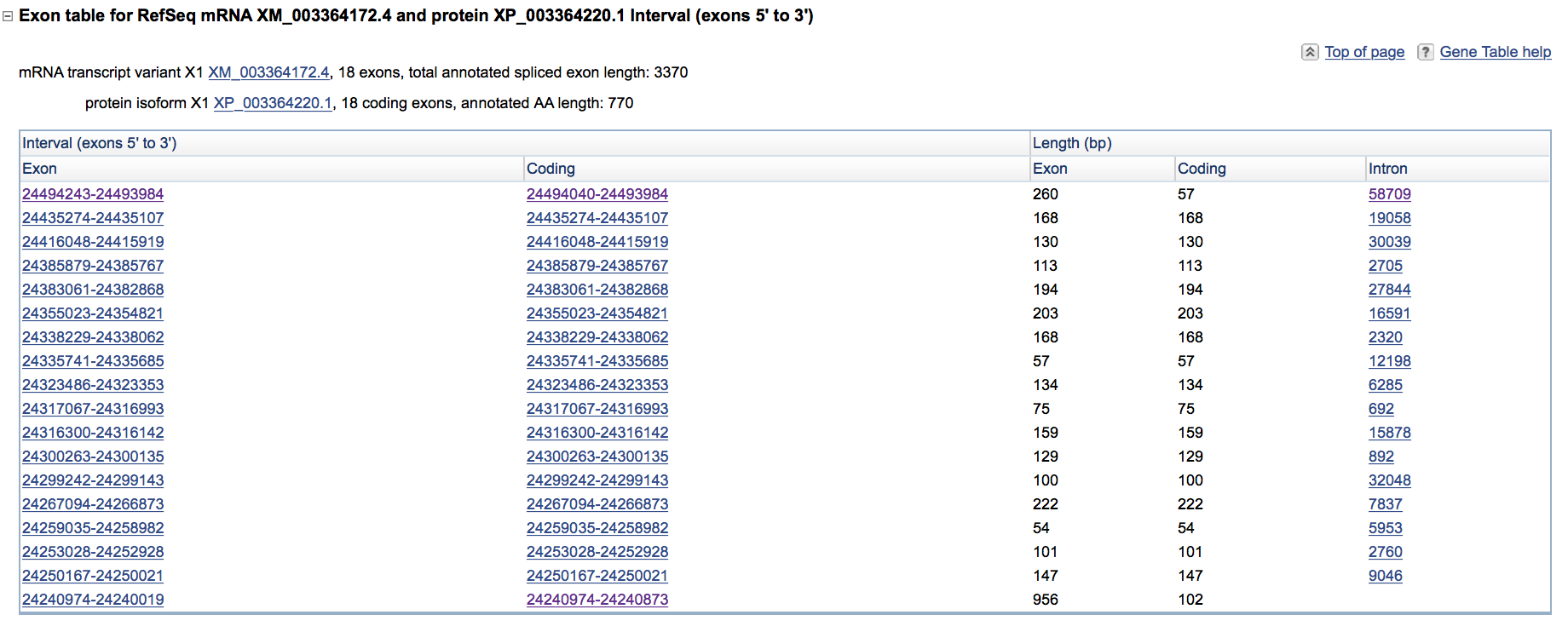
1. The “Gene Table” screen looks like the following with the diagram showing the entire gene above the available transcripts (in bold black ink). Selection of one of the available transcripts will show you the exons/coding sequence/introns of that transcript.



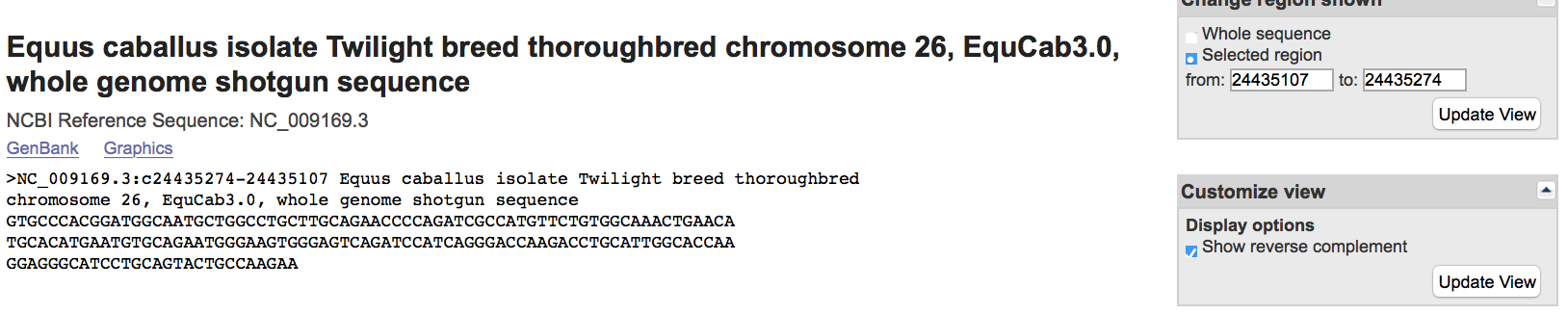
1. The transcript screen shows the number of exons, total annotated exon length and the annotated AA length. It is helpful to look through the available transcripts and pick the largest transcript (by total annotated exon length or annotated AA length) to begin with. Clicking on the individual exon or coding sequence will bring up the translated sequence. The exon & coding lengths will differ for the beginning and ending exons due to the 5’ and 3’ untranslated regions.

**CLARIFY HOW TO PICK THE LARGEST TRANSCRIPT**

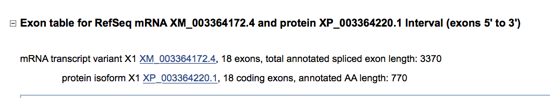
**CLARIFY LARGEST VS. THE ONE THAT MATCHES**



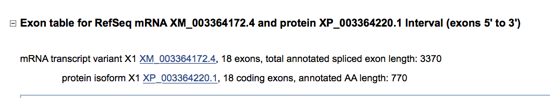
1. This is the individual coding sequence of the exon with the base pair position on the right of the screen.



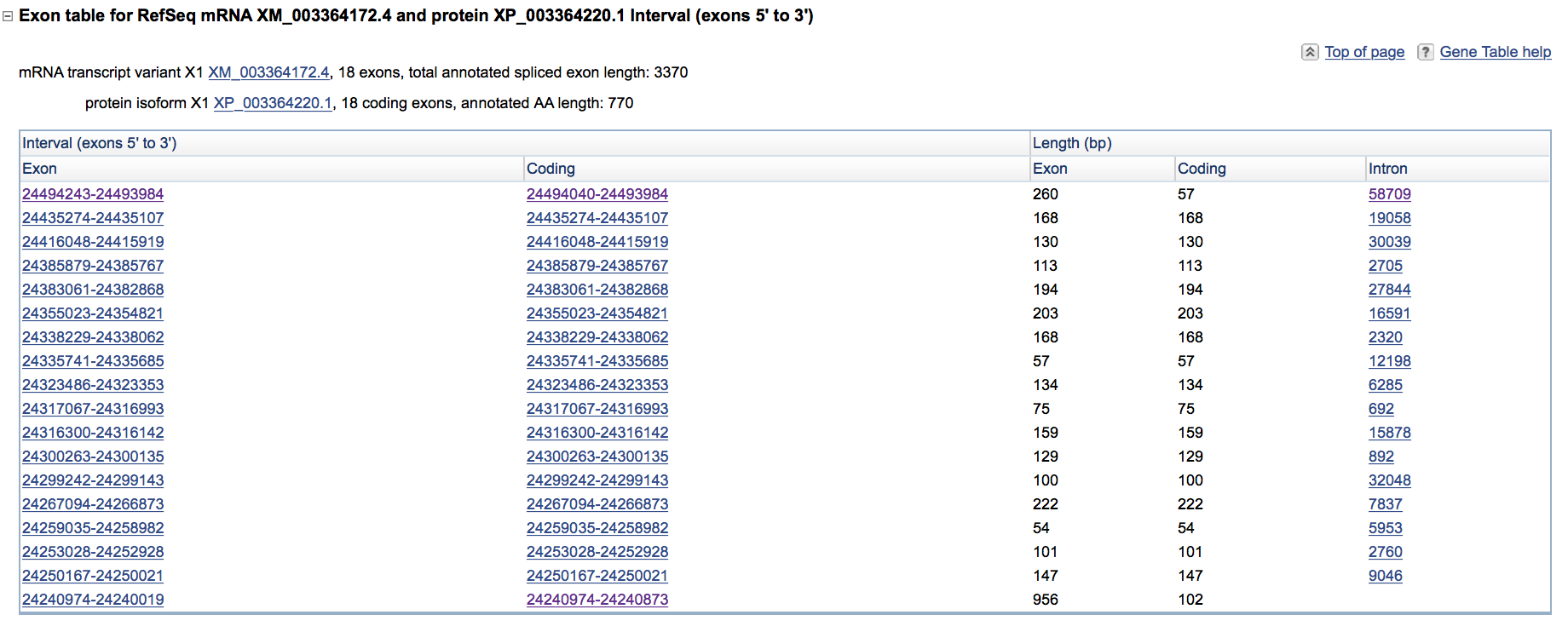
1. Now you repeat steps 1-7 to locate the same gene in the human in a separate window as you will need to be able to go back and forth between the horse & human genome for comparison.
2. Comparison of transcripts begins by looking at the number of the exons and associated base pair length. It is important to ask yourself these questions as you are going through the comparison phase as it will drive the rest of the process. A descriptive of how to find the answers to these questions is detailed below.
   1. Do the horse & human genome have same number of amino acids?
   2. Do the horse and human have the same number of exons?
   3. Are the exons the same length? Compare the exons and base pair lengths between both horse & human genomes.
   4. Is there a 5’ and 3’ untranslated region?
   5. Is there a start codon (ATG)?
   6. Assess the splice junctions. Do all intron regions begin with ga and end with ag?
   7. Is there a stop codon (TAG, TAA, TGA)?
   8. Are there n’s in the sequence?
3. Do the horse & human genome have the same number of amino acids? This is located in the transcript description.



1. Do the horse & human have the same number of exons? This can also be found on the transcript screen.



1. Are the exons the same length? Compare the base pair length of the human and horse genomes by looking at the coding length. The total exon length includes the untranslated regions which can make it longer than the coding length (especially the first few and last few exons as this is where the 5’ and 3’ UTR would be located).

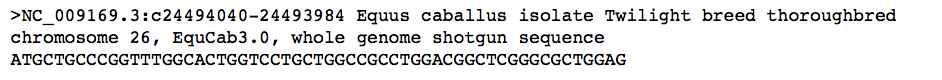


1. Is there a 5’ UTR and 3’ UTR?

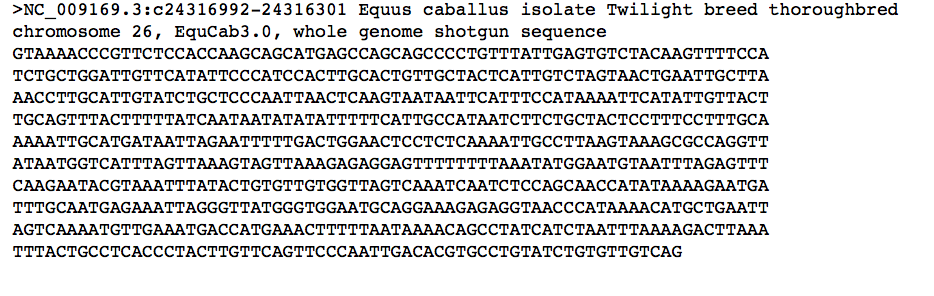
-This can be assessed by looking at the difference between the exon & coding base pair lengths. The 5’ and 3’ UTR would be included in the exon length but not in the coding length which helps determine if they are present at the start and end of the transcription.

1. Is there a start codon (ATG)?

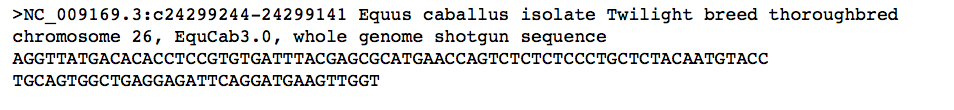
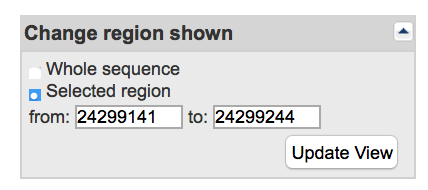
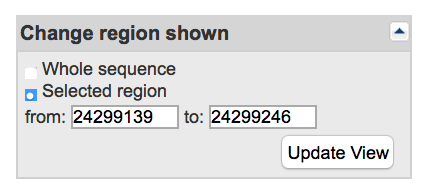
-It is often helpful to first try and locate the start codon within the human genome as if it isn’t until exon 2 or 3 it will probably be in a similar place in the horse genome. Then you can start going through the equine coding sequence and look for the associated start codon. The start codon can also be found by looking for the first exon in which the exon and coding length differ as this indicates that the 5’ UTR and transcription sequence are on the same exon.



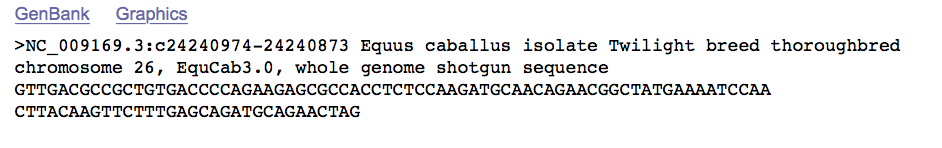
1. Assess the splice junctions. Do all intron regions start with “GT” and end with “AG”. This can be done two ways. First you can select the intron sequence from the main transcript table and view the start and end of the sequence.



OR you can select the coding sequence and extend the base pair length by two both upstream and downstream to include the end and start of the surrounding introns. When doing it this way the first two letters on the sequence should be the ending “AG” of the preceding intron and the last two should be the starting “GT” of the next intron.



1. Does it contain a stop codon (TAG, TAA, TGA) prior to the start of the 3’ untranslated region? This can be determined by looking at the coding sequence of the last few exons and looking for a stop codon at the end. By comparing the coding and exon length you can also get an idea of which exon should contain the 3’ UTR (it can span more than one exon).



1. Are there n’s in the sequence?

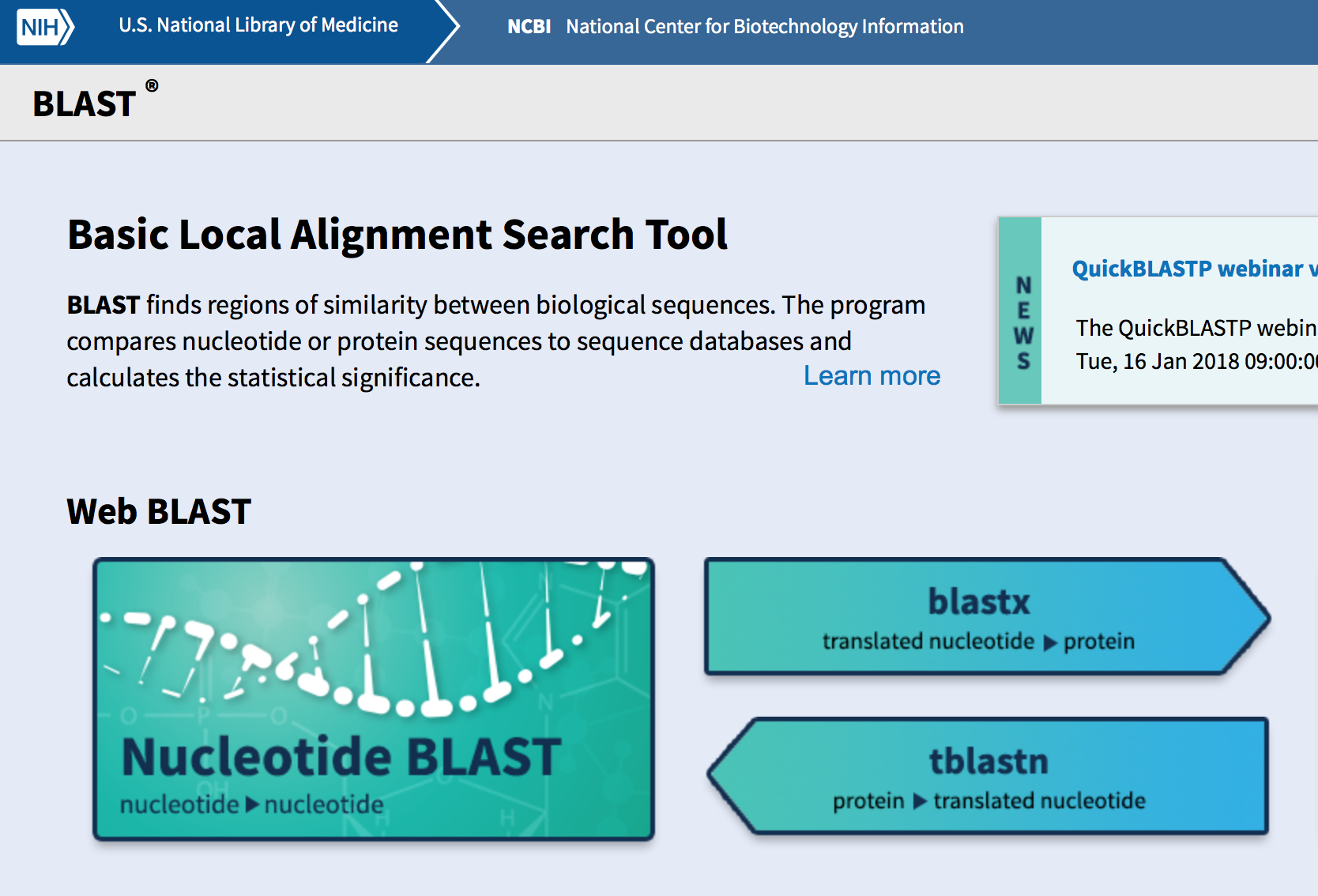
Once you have answered the above questions it is time to closer assess exons via BLAST. Below is a description on how to do this process of comparing coding sequences.

Please also see the associated flowchart and tutorials on how to deal with common problems that may be encountered as you work through the above question

**USING BLAST**

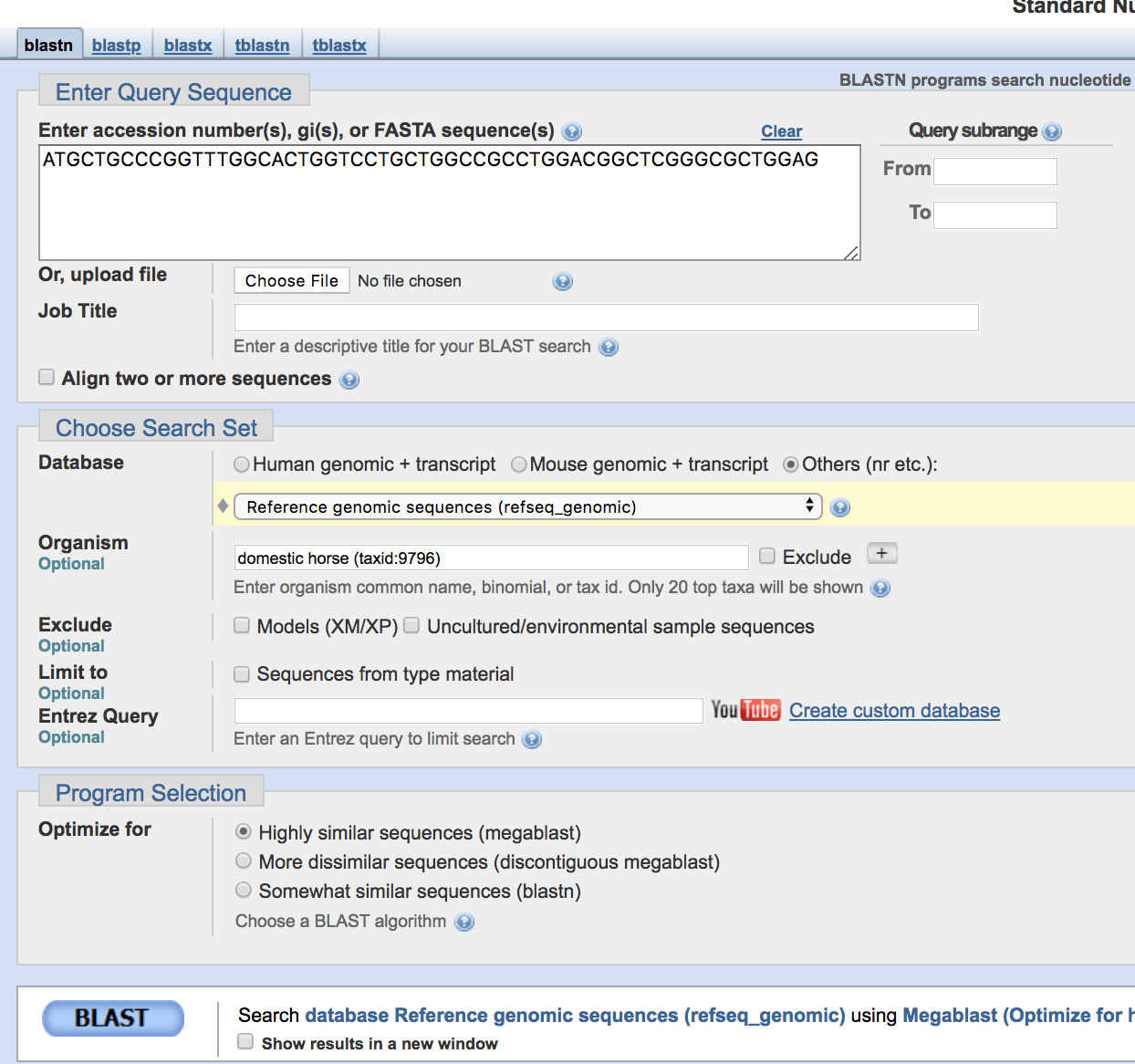
1. Start by opening the BLAST browser at NCBI: <https://blast.ncbi.nlm.nih.gov/Blast.cgi>

-Select nucleotide BLAST



1. Once you are in the nucleotide BLAST screen you will enter the horse sequence in the yellow box, select reference genome sequences (refseq.genomic) under databases and search the human genome for a match. Select BLAST to match the sequences.

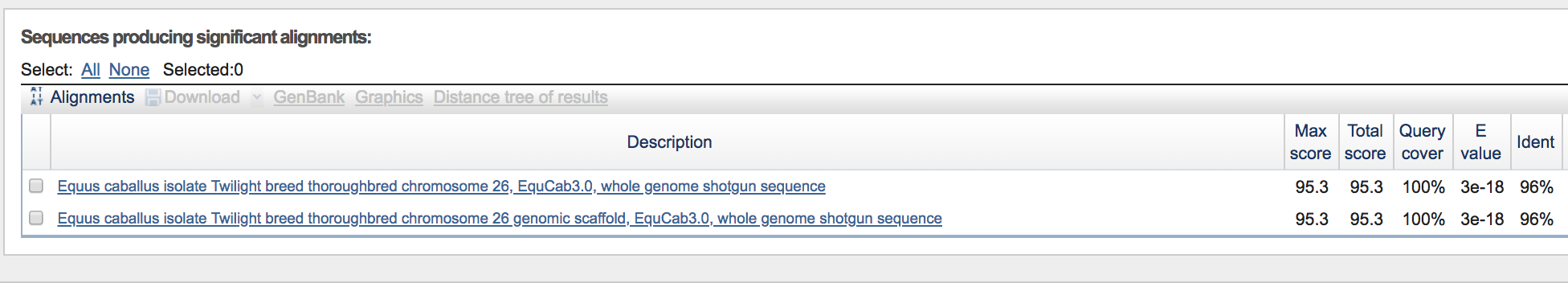
\*You may also need to BLAST the human sequence to the horse genome depending on what is missing and what information you are trying to obtain



-The screen will update multiple times as the BLAST is being performed and then results will be shown on screen. This screen shows the number of base pairs that are in the BLAST sequence.



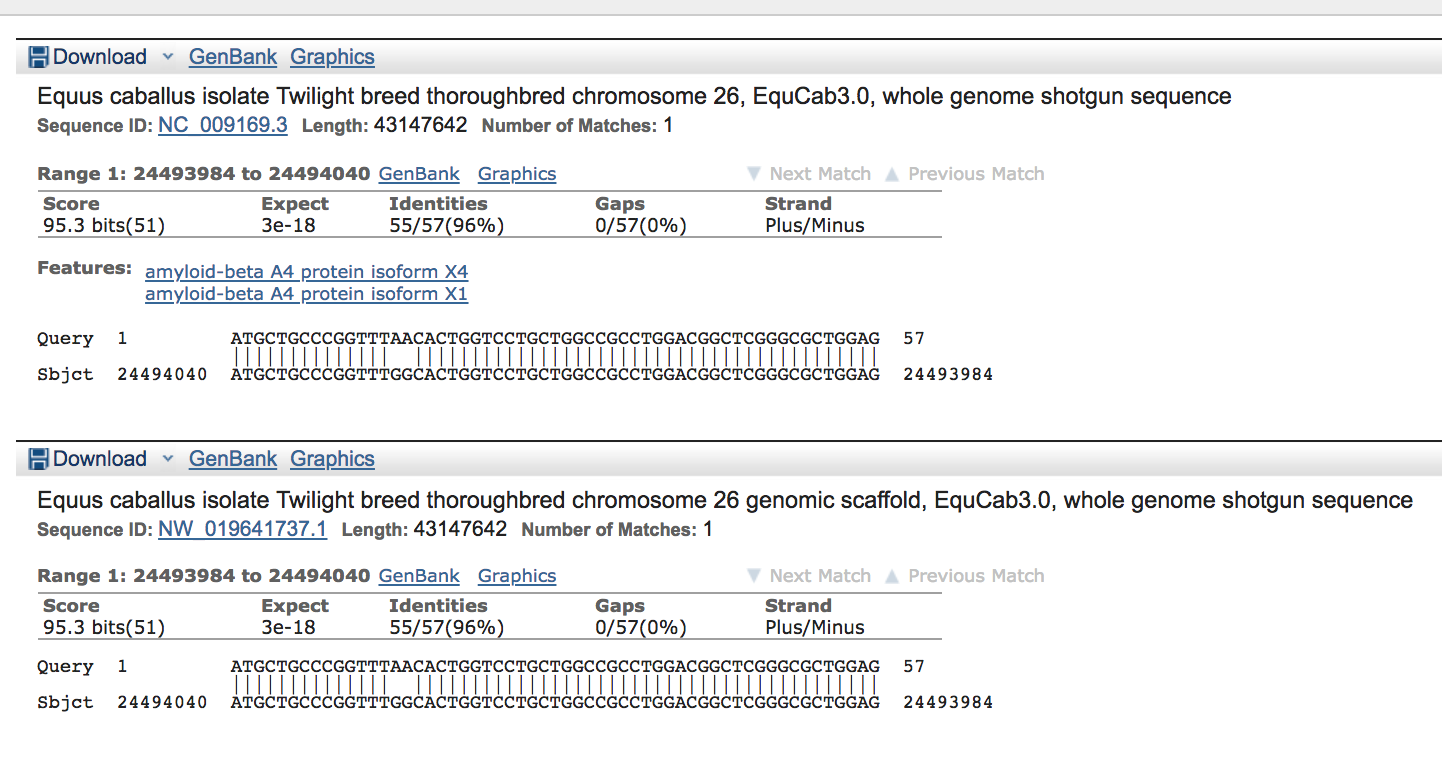
1. Results will include the “Descriptions” (shown below) that show all of the separate transcript alignments and the identity match % (96%).



1. Below the descriptions section is the “Alignments” section which provides the individual transcript matches. The identities section shows how many of the nucleotides match up between the horse & human (in this example: 55/57 or 96%, shown in red box) of the base pairs. The “query” sequence is the human sequence that was entered into BLAST and the “subject” is the equine reference sequence.

On the left and right side of the subject sequence is the base pair position of this sequence. (For example this sequence spans between the base pair positions of 24,494,040 and 24,493,984- shown in the blue boxes).

Any difference can be visualized at the individual base pair level and is shown with the absence of a vertical line joining the two base pairs shown within the green box below.



**MULTIPLE SPECIES ALIGNMENT**

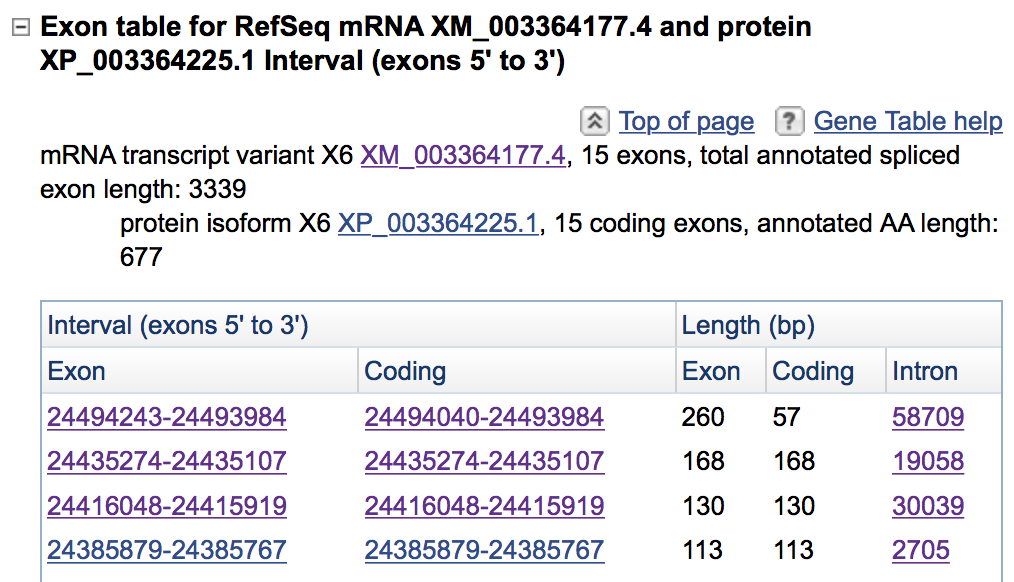
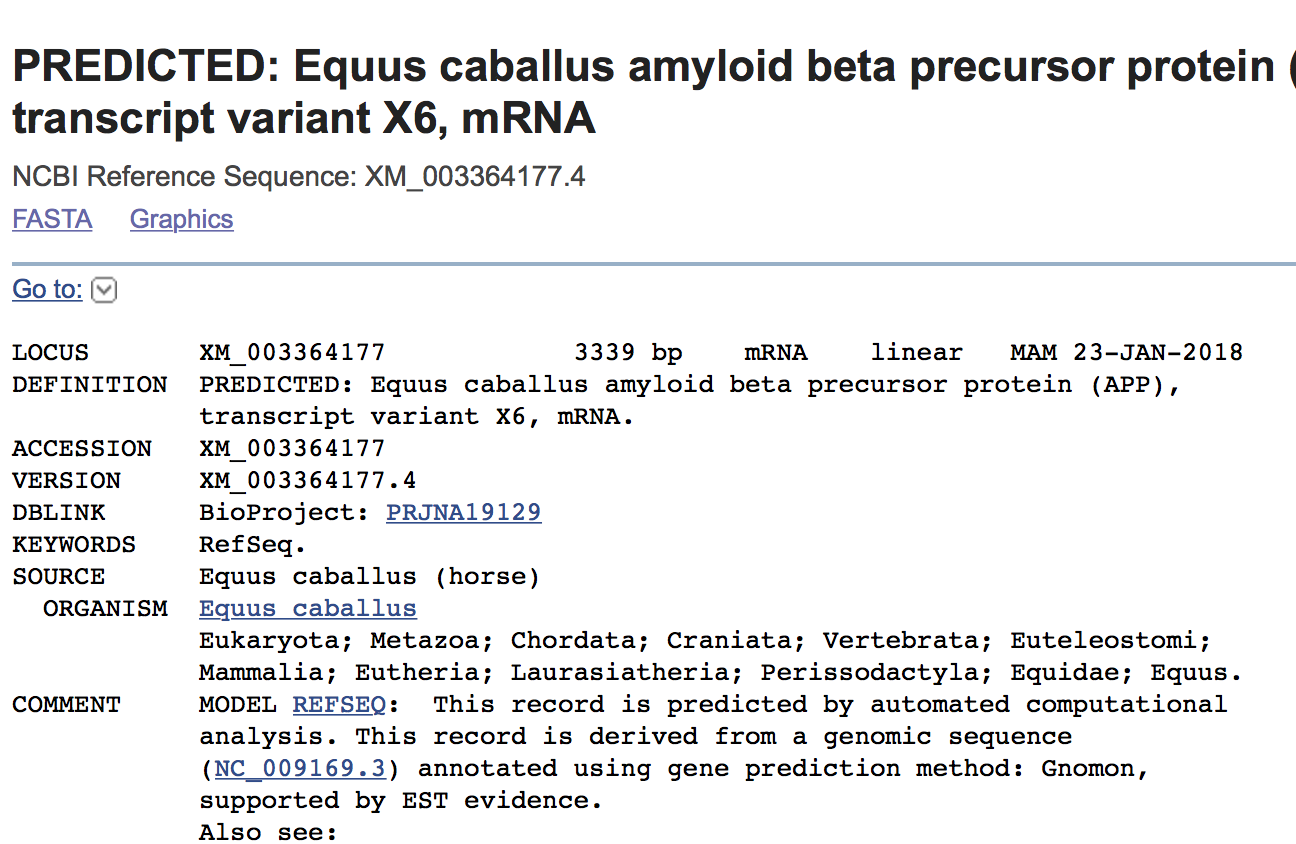
Another component of comparing genomes is to do a multiple species alignment that looks at a gene sequence (either as nucleotides or as amino acids) across many species. This allows us to compare the species of interest to others and look at whether or not variants that are occurring are present in conserved or non-conserved regions of the gene. A conserved region is one in which the sequence is very similar to identical in multiple species which could indicate that this region is important for general function. Non-conserved regions can vary greatly between species which often indicates that they are not as important for general functions, and variants identified in non-conserved regions are often not as significant.

There are a variety of programs available online that can perform multiple species alignment. For this example we will use MultAlin (Multiple sequence alignment by Florence Corpet).

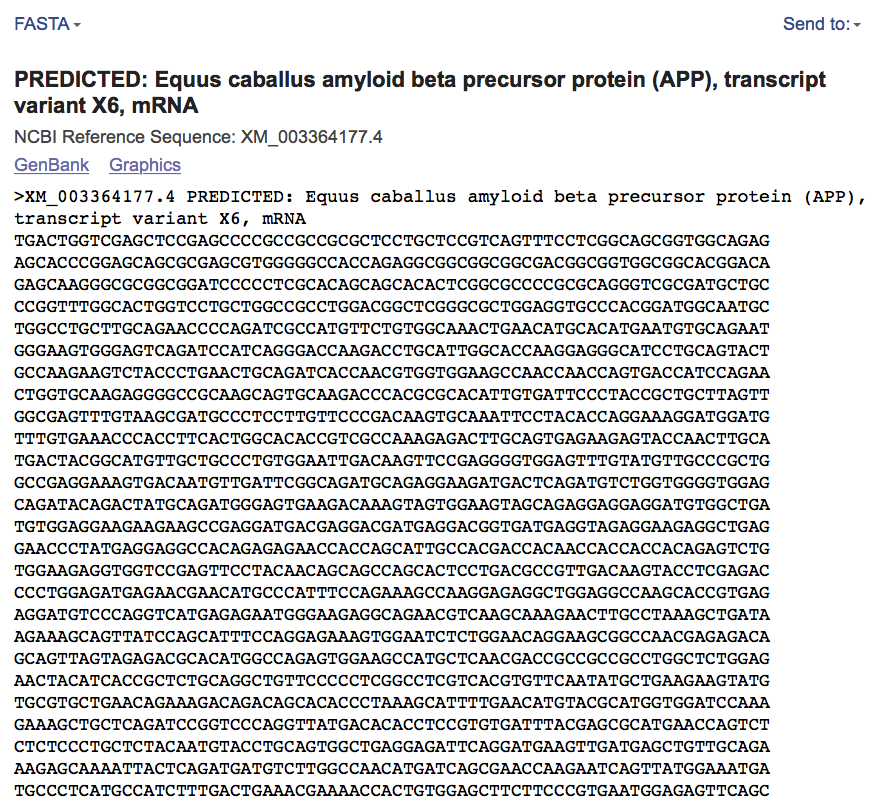
<http://multalin.toulouse.inra.fr/multalin/>

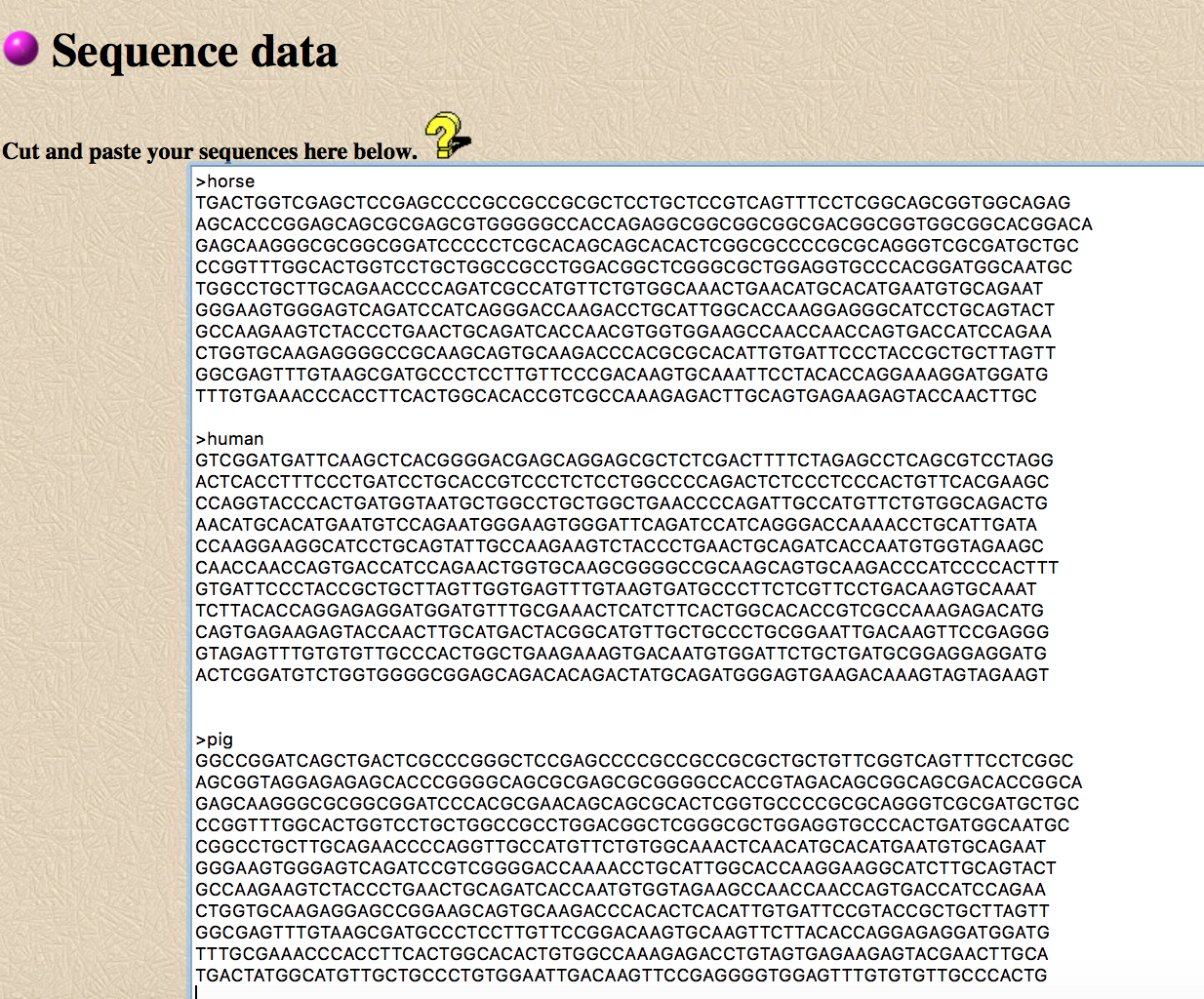
In order to perform multiple alignment you need to pull the FASTA sequence from the desired species as described below and insert it into the MultAlin website.

1. Locate the desired gene (ex: APP horse) and open the gene table screen that shows the exon transcript. Select the name of the mRNA transcript variant which is listed in the description of that transcript (shown below) which will then open another screen where you have the option of selecting the FASTA sequence.

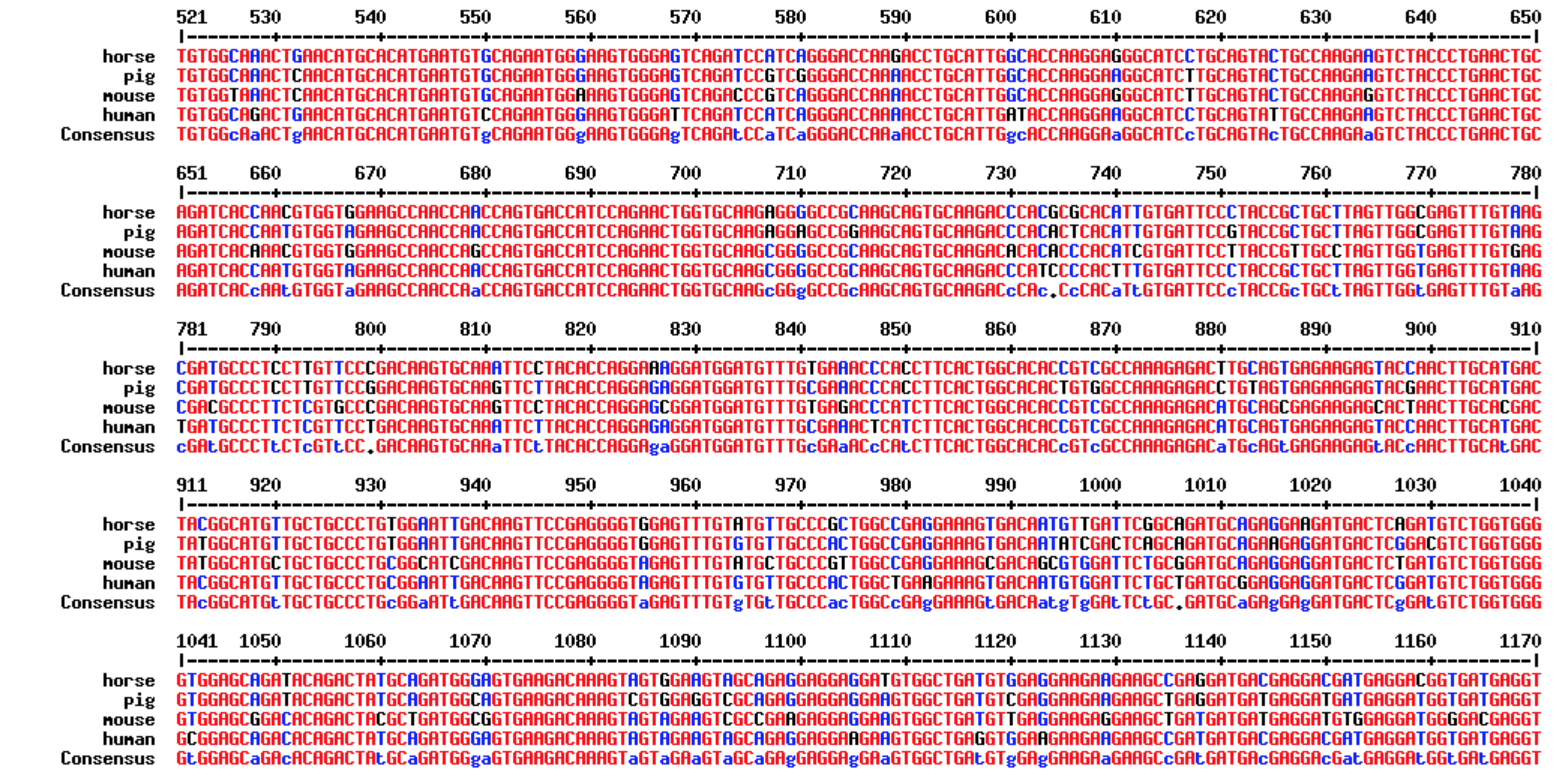


1. Selecting FASTA will open the nucleotide sequence for the entire gene. This entire sequence is then copy and pasted into the MultAlin sequence box. Each individual sequence that is put into the MultAlin search box needs to be led with a >Species (ex: >horse) as this will give a title to the sequence so you can identify which sequence is from which animal.

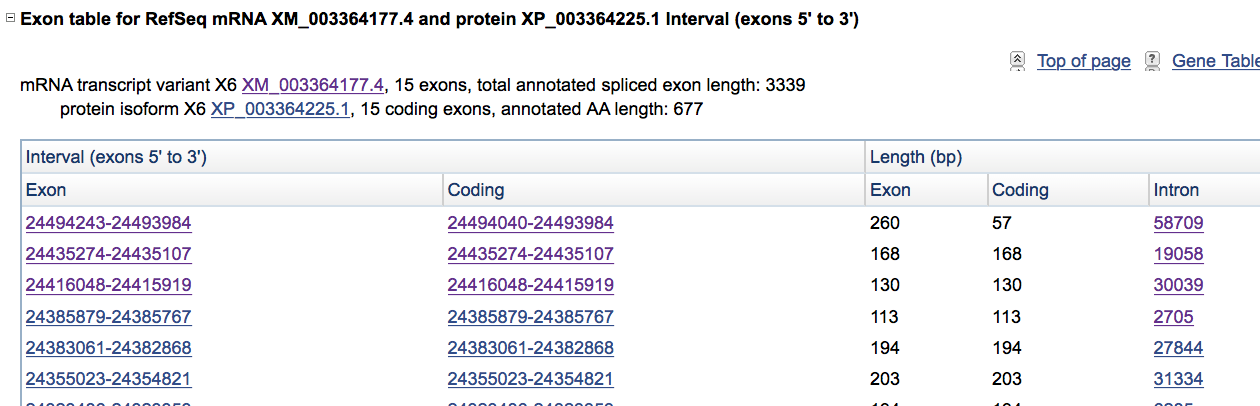


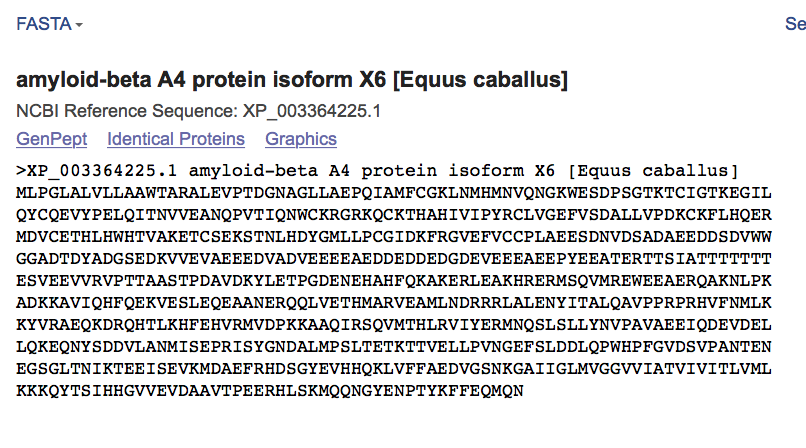


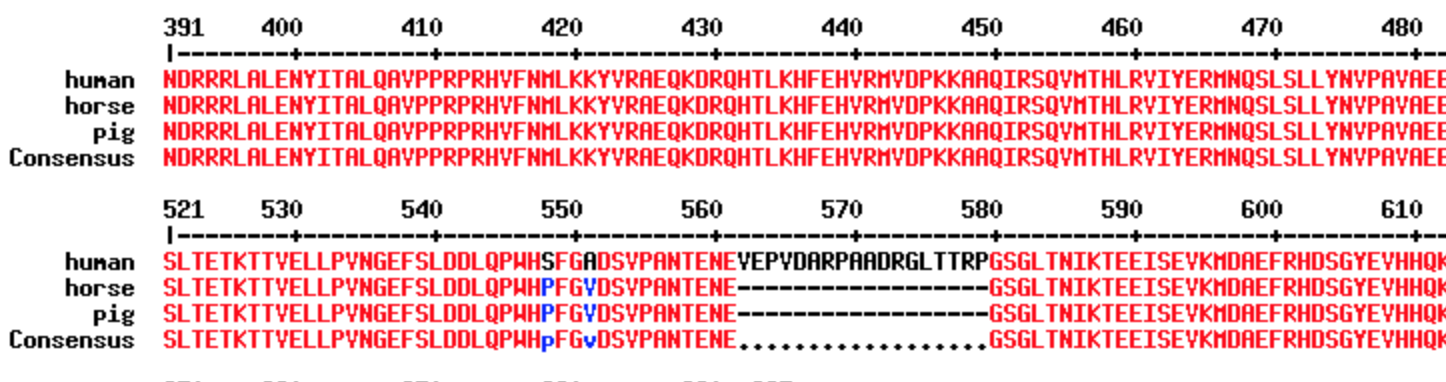
1. Results will look like the image below. This creates a row for each animal transcript (with the title you created by putting the name after the “>” sign) that you are comparing and allows you to look across multiple species at one time for the nucleotide conservation. Everything shown in red is a match across species with differing nucleotides shown in blue.



1. This step is then often repeated to look at the amino acid conservation across species as nucleotide differences that affect the amino acid sequence are likely more significant and have the potential to cause greater effects. This is done by clicking on the protein isoform located in the transcript description and obtaining the FASTA sequence of amino acids.





1. The FASTA amino acid sequence is then pasted into the sequence box for multiple species (same as shown above for amino acids) to generate a multiple species alignment that looks like the one shown below. Amino acids that are the same are shown in red with changes shown in blue and additional amino acids that aren’t present in other species shown in black.