Created by Paul Yenerall, 2010

**Using BLAST (Basic Local Alignment Search Tool) - blastn**

The purpose of this tutorial is to help teach students how to use and understand NCBI’s BLAST (Basic Local Alignment Search Tool). This tutorial will cover blastn (blast nucleotide) and touch upon the use of other blast searches. Doing this tutorial will help students understand BLAST, which in turn will make subsequent searches (such as blastp) easier.

For questions, corrections, or help with this tutorial, please e-mail [paulyenerall@gmail.com](mailto:paulyenerall@gmail.com)or [lzhou1@pitt.edu](mailto:lzhou1@pitt.edu)

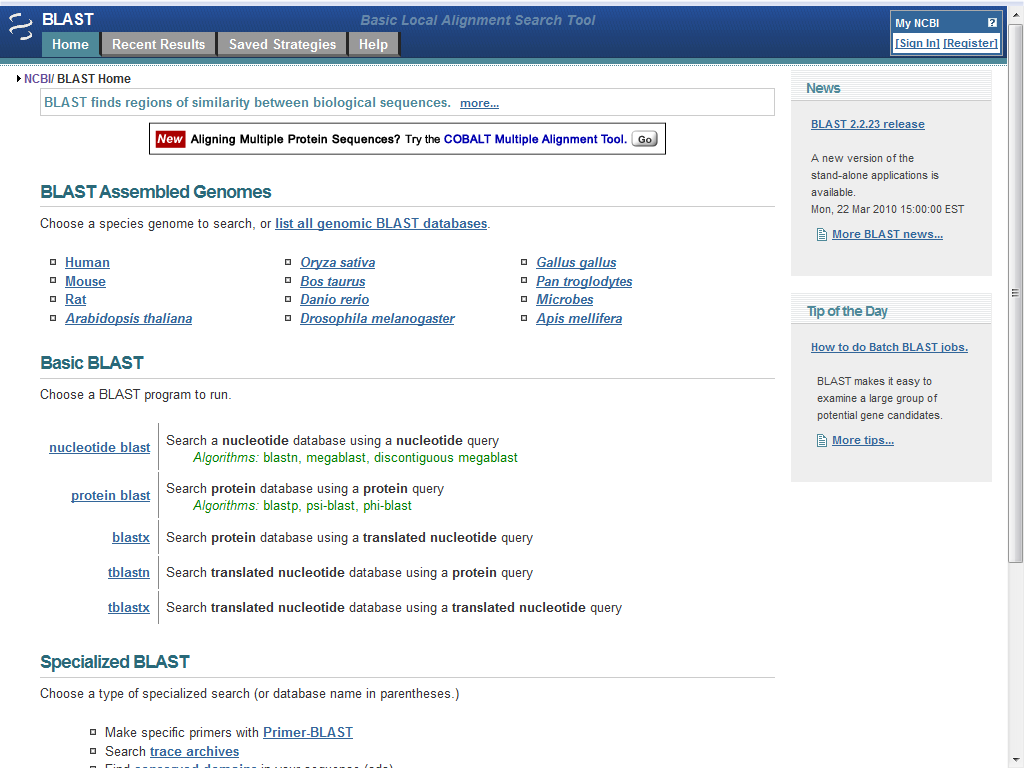
**Using BLAST search**

Before we begin, let’s first review what BLAST is and why we use BLAST. From NCBI’s website:

“The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.”

Not only can you use BLAST to infer functional and evolutionary relationships between sequences and species, but you can also use BLAST to match a nucleotide or protein sequence to its corresponding gene or protein. In this tutorial I will provide you with a nucleotide sequence and by using blastn (BLAST nucleotide) we will find out what gene this nucleotide sequence is for, along with other important facts (such as species, protein that it codes for, and similar genes). We will also review different functions of BLAST as well as BLAST results that aren’t as cut and dry as our example.

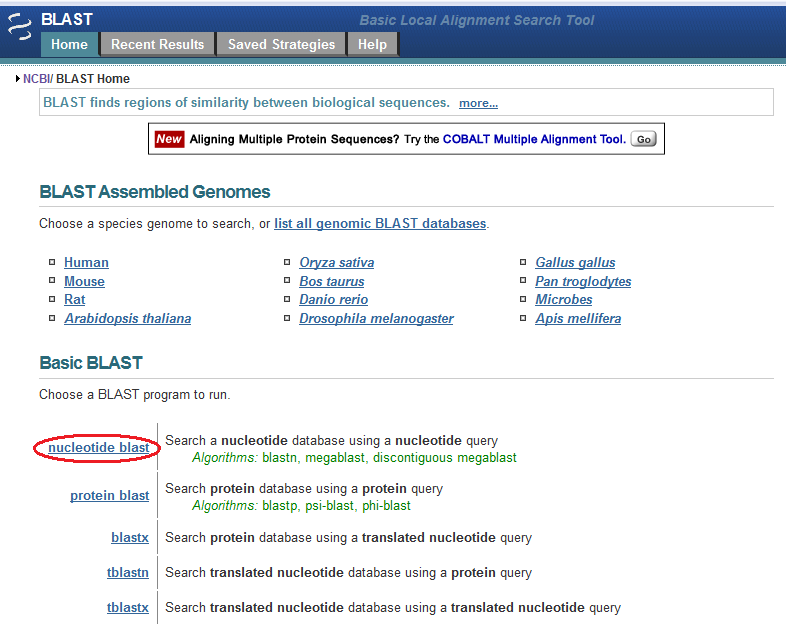
**1)** First, load up an internet browser, preferably Mozilla Firefox, and go to the URL <http://blast.ncbi.nlm.nih.gov/Blast.cgi>. A page like the one listed below should appear (next page):



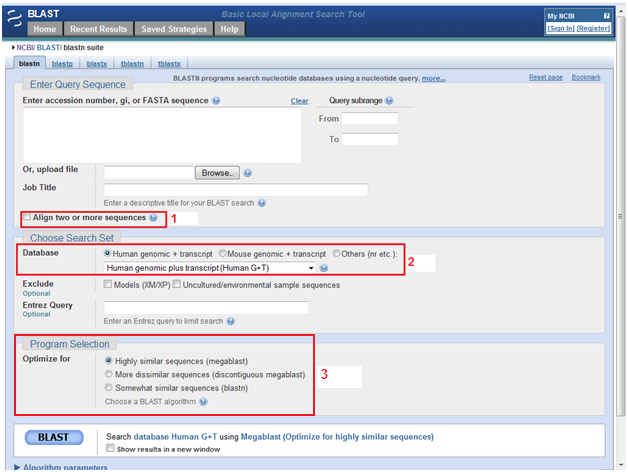
Now, before you begin a BLAST query, you need to determine what type of BLAST query you need to perform. As you can see, BLAST has five basic BLAST programs you can run: nucleotide blast (blastn), protein blast (blastp), blastx, tblastn, and tblastx. Each program can be used in a different scenario and these scenarios will be explained below (information taken from NCBI’s BLAST help section):

Blastn (blast nucleotide) – Use this tool if you are given a nucleotide sequence greater than 20 nucleotides and wish to identify the query sequence or find nucleotide sequences similar to the query sequence.  
Blastp (blast protein) – Use this tool if you are given a protein sequence greater than 15 residues (amino acids) and wish to identify the query sequence or find protein sequences similar to the query sequence.  
Blastx – Use this tool if you are given a nucleotide sequence greater than 20 nucleotides and wish to find similar proteins to a translated query in a protein database (i.e. you have a nucleotide sequence and wish to find it’s corresponding protein sequence or protein sequences similar to what this nucleotide sequence codes for).  
Tblastn – Use this tool if you are given a protein sequence greater than 15 residues (amino acids) and wish to find similar proteins in a translated nucleotide database (i.e. you have a protein sequence and wish to find it’s corresponding nucleotide sequence or nucleotide sequences that code for a similar protein)  
Tblastx – Use this tool if you are given a nucleotide sequence longer than 20 nucleotides long and wish to find similar proteins to a translated nucleotide query in a translated database (i.e. you have a nucleotide sequence and wish to find it’s translated nucleotide sequence or similar translated nucleotide sequences – this differs from blastn because it only looks at the coding region of your nucleotide sequence)

For this part of the tutorial, I will provide you with a nucleotide sequence and we will try to determine what gene or exon this nucleotide sequence comes from. Therefore, we will use a nucleotide BLAST (blastn). Doing nucleotide BLAST first will help explain many features of BLAST, making subsequent searches (such as blastp, blastx etc.,) faster and easier to perform. Click on the link labeled “nucleotide blast” as illustrated below:



The following page should then be displayed. Note the highlighted regions, as these will be discussed later (next page):



1 – For now, just note that you have the option to align two or more sequences – we will explore this in greater detail at the end of the tutorial.  
2 – This is the database your query will be searched against. By default, it searches humans (“Human genomic plus transcript”). For this project, we will be changing this in the next step NOTE: this is NOT selected in this picture but will be changed later  
3 – This is the program selection. By changing the program, you select how closely related you want the results to be. Megablast is the most commonly used search, as it finds the most similar sequences when long query sequences are used. However, because we do not know the source of our nucleotide sequence, we will use “Somewhat similar sequences (blastn)” NOTE: this is NOT selected in this picture but will be changed later

**2)** Now, please copy and paste this nucleotide sequence into the box labeled “Enter accession number, gi, or FASTA sequence” (located on the next page):

CGCTGGTGCAACTCGAAGACCTATCTCCTTCCCGGGGGGGCTTCTCCGGCATTTAGGCCTCGGCGTTTGGAAGTACGGAGGTTTTTCTCGGAAGAAAGTTCACGGAAGTGGAAGAAATGGATTTATCTGCTGTTCGAATTCAAGAAGTACAAAATGTCCTTCATGCTATGCAGAAAATCTTGGAGTGTCCAATCTGTTTGGAACTGATCAAAGAACCGGTTTCCACACAGTGCGACCACATATTTTGCAAATTTTGTATGCTGAAACTCCTTAACCAGAAGAAAGGACCTTCCCAGTGTCCTTTGTGTAAGAATGAGATAACCAAAAGGAGCCTACAAGGAAGTGCAAGGTTTAGTCAACTTGTTGAAGAGCTGCTGAAAATCATTGATGCTTTTGAGCTTGACACCGGAATGCAGTGTGCAAATGGTTTTAGTTTTTCAAAAAAGAAAAATAGTTCTTCTGAGCTTCTGAATGAGGACGCTTCTATCATCCAGAGTGTGGGCTACCGAAACCGCGTCAAAAAACTTCAACAGATTGAGTCTGGAAGTGCCACCTTGAAGGACAGTCTCAGTGTCCAGTTGTCTAACCTTGGGATTGTGAGATCAATGAAGAAAAATAGGCAGACACAACCTCAAAACAAATCTGTCTACATTGCTTTAGAATCTGATTCTTCTGAGGAGAGAGTAAATGCACCCGATGGCTGCAGTGTGAGAGATCAGGAATTGTTTCAGATCGCCCCTGGAGGAGCTGGAGATGAAGGCAAGTTGAACTCTGCAAAAAAGGCTGCTTGTGACTTTTCTGAGGGCATAAGAAACATTGAACATCATCAATGCAGTGATAAAGATTTGAACCCCACTGAGAATCATGCAACTGAAAGGCATCCAGAAAAATGTCCGCGTATTTCTGTTGCAAACGTGCACGTGGAGCCGTGTGGCACGATGCTCGTGCCAGCTCATTACAGCGTGGGACCCGCAGTTTATTGTTCACTGAGGACAGACTGGATGCAGAAAAGGCTGAATTCTGTGATAGAAGCAAACAGTCTGGCGCAGCAGTGAGCCAGCAGAGCAGATGGGCTGACAGTAAAGAAACATGTAATGGCAGGCCGGTTCCCCGCACTGAGGGAAAGGCAGATCCAAATGTGGTTCCCTCTGTGGTAGAAAGCAGTGGAATCATCCGAAAAGCCTGTGCCCTGAGAATTCTGGAGCTACCACTGACGTTCCTTGGATAACACTGAATAGCAGCATTCAGAAAGTGAATGAGTGGTTTTCCAGAACTGGTGAAATGTTAACTTCTGACAATGCATCTGACAGGAGGCCTGCGTCAAATGCAGAAGCTGCTGTTGTGTTAGAAGTTTCAAATGAAGTGGATGGATGTTTCAGTTCTTCAAAGAAAATAGACTTAGTTGCCCCTGATCCCGATAATGCTGTAATGTGTACAAGTGGAAGAGACTTCTCCAAGCCAGTAGAGAATATTATCAACGATAAAATATTTGGGAAAACCTATCAGAGAAAGGGAAGCCGCCCTCACTTGAACCATGTGACTGAAATTATAGGCACATTTACTACAGAACCACAGATTATACAAGAGCAGCCCTTCACAAATAAATTAAAACGCAAAAGAAGTACATGCCTTCATCCTGAGGACTTTATCAAGAAAGCAGATTTAACAGTTGTTCAAAGGATTTCTGAAAATTTAAATCAGGGAACTGACCAAATGGAGCCAAATGACCAAGCAATGAGTATTACCAGTAACGGTCAGGAGAACAGAGCAACAGGTAATGATCTTCAGAGAGGGAGAAATGCTCATCCAATAGAATCATTGAGAAAGGAACCTGCTTTCACAGCTAAAGCCAAATCTATAAGCAACAGTATAAGTGATTTGGAGGTAGAATTAAATGTTCACAGTTCAAAAGCACCTAAGAAAAATAGGCTGAGGAGGAAGTCTACCAGGTGTGTTCTTCCACTCGAACCAATCAGTAGAAATCCGAGCCCACCTACTTGTGCTGAACTTCAGATCGAGAGTTGTGGTAGCAGTGAAGAAACAAAGAAAAACAATTCCAACCAAACCCCAGCCGGGCACATTAGAGAGCCTCAACTCATCGAAGACACAGAACCCGCAGCTGATGCCAAGAAGAACGAGCCAAATGAACACATAAGGAAGAGAAGTGCCAGTGATGCGTTCCCAGAAGAGAAATTAATGAACAAAGCTGGTTTATTAACTAGCTGTTCAAGTCCTAGAAAGCCTCAAGGACCTGTCAATCCTAGCCCTGAGAGAAAAGGAATAGAGCAACTTGAAATGTGCCAGATGCCTGATAATAACAAAGAACTCGGGGATTTGGTCCTGGGAGGAGAGCCCAGTGGGAAACCTACTGAACCATCTGAGGAGAGCACCAGTGTGTCCTTGGTACCCGACACAGACTACGACACCCAGAACAGTGTCTCAATACTGGAAGCGAACACTGTCAGATATGCAAGAACAGGATCAGTTCAGTGTATGACTCAGTTTGTCGCAAGTGAAAACCCCAAGGAACTTGTCCATGGTTCTAACAATGCTGGAAGTGGCTCGGAGTGCTTCAAGCACCCATTGAGACATGAACTTAACCACAATCAAGAGACAATAGAAATGGAAGACAGTGAACTTGATACTCAGTATTTGCAGAATACATTTCAAGTTTCAAAGCGTCAGTCATTTGCTTTATTTTCAAAACTTAGAAGTCCCCAAAAGGACTGTACTCTGGTAGGTGCCCGCTCTGTGCCCTCAAGGGAACCAAGTCCAAAGGTGACTTCTAGAGGTGAACAAAAAGAACGTCAGGGACAAGAAGAGTCTGAAATCAGTCATGTACAGGCAGTCACAGTCACAGTAGGCTTACCTGTGCCCTGTCAGGAAGGTAAGCCAGGTGCTGTTACAATGTGTGCTGATGTTTCTAGGCTTTGTCCGTCATCTCATTATAGAAGCTGTGAGAATGGACTCAACACCACAGATAAATCTGGAATTTCACAAAACTCACATTTTAGACAATCAGTTTCTCCCCTCAGGTCATCTATAAAAACTGACAATAGAAAAACTCTGACAGAGGGNCGATTTGAGAAACATACTGAAAGGGGRATGGGAAATGAGACTGCTGTTCAAAGTACAATACACACAATTAGTCTAAATAACAGAGGAGATGCTTGTCTAGAAGCCAGCTCAGGCAGTGTTATTGAAGTACATTCCACTGGTGAAAACGTCCAGGGGCAACTAGATAGAAACAGAGGGCCTAAGGTAAACACCGTGTCTCTATTAGATAGTACACAGCCTGGTGTCTCTAAGCAGAGTGCTCCTGTAAGTGATAAGTATCTTGAAATAAAGCAGGAGAGTAAGGCTGTCAGTGCAGACTTCTCTCCATGTCTGTTCTCAGATCATCTTGAAAAACCTATGAGAAGTGATAAGACTTTTCAGGTTTGCTCTGAGACACCTGATGACCTGTTGGATGATGTTGAAATACAGGAAAATGCTAGCTTCGGTGAAGGTGGCATAACGGAAAAGTCTGCTATTTTTAATGGAAGTGTCCTGAGAAGAGAGTCCAGTAGGAGCCCTAGCCCTGTAACCCATGCATCGAAGTCGCGGAGTCTCCACAGAGGGTCTAGGAAATTAGAATTCTCAGAAGAGAGCGACTCCACTGAGGATGAAGACCTTCCCTGCTTCCAACACTTGCTGAGCAGAGTAAGCAGCACACCCGAACTTACCAGATGCAGCAGTGTGGTGACACAGCGTGTGCCAGAGAAAGCGAAGGGGACCCAAGCACCCCGGAAGAGTAGCATCAGTGACTGCAATAATGAGGTGATCTTGGGAGAGGCATCTCAGGAATATCAGTTTAGTGAGGATGCAAAATGCTCTGGCAGCATGTTCTCCTCACAGCACAGTGCTGCCCTAGGTTCNCCTGCAAATGCACTCTCCCAGGATCCCGATTTTAATCCACCTTCCAAACAAAGGCGTCACCAGGCTGAGAATGAGGAAGCCTTTCTAAGTGATAAGGAATTGATTTCAGATCATGAGGACATGGCAGCTTGCCTGGAAGAGGCTAGCGATCAAGAAGAGGACAGTATAATCCCGGACTCAGTGGCGTCTGGATATGAGAGTGAAGCAAACCTCTCTGAAGACTGCTCACAGAGTGATATTCTAACCACTCAGCAGAGGGCGACCATGAAGGATAACCTGATAAAGCTCCAGCAGGAAATGGCTCAGCTGGAAGCTGTGCTAGAGCAGCATGGGAGCCAGCCGTCTGGCCACCCCCCTTGCCTCCCAGCTGACCCTTGTGCCCTAGAAGACCTGCCAGATCCAGAACAAAACAGATCAGGAACAGCAATTTTAACTTCAAAGAACATTAATGAGAATCCTGTAAGCCAGAATCCAAAGCGCGCCTGTGATGACAAATCCCAACCACAACCTCCGGACGGTCTCCCCAGTGGAGATAAGGAGTCAGGAATGAGAAGGCCTTCCCCTTTTAAATCTCCACTGACCAGCAGTAGGTGCTCTGCACGAGGCCACTCCAGGAGTCTTCAGAATAGAAACTCCACSTCGCAAGAGGAGCTCCTTCAGCCTGCTGRATTGGAGAAGTCATGTGAACCACACAACTTAACAGGGCGGTCTTGCTTGCCAAGGCAAGATCTCGAAGGAACCCCATACCCGGAATCTGGAATCAGGCTAGTCTCCAGCAGAGACCCTGATTCTGAATCCCCTAAAGTGTCAGCCCTTGTTTGCACTGCACCAGCTTCAACCTCTGCACTGAAAATATCCCAAGGTCAAGTTGCTGGGTCTTGCAGGAGTCCAGCTGCTGGTGGTGCTGATACTGCAGTGGTAGAAATTGTGAGCAAGATAAAGCCGGAAGTGACGTCTCCAAAAGAACGAGCGGAGAGAGATATATCCATGGTGGTGTCAGGCTTGACCCCTAAAGAAGTAATGATTGTCCAAAAGTTTGCTGAAAAATACCGCCTTGCTTTAACTGATGTAATTACTGAGGAAACTACGCATGTAATTATAAAAACAGATGCTGAGTTTGTGTGTGAACGGACACTGAAATATTTTCTGGGCATTGCAGGAGGAAAGTGGATAGTTAGCTATTCTTGGGTGATCAAGTCTATCCAAGAAAGAAAACTTCTGAGTGTGCATGAATTTGAAGTCAAAGGAGACGTTGTGACTGGAAGCAACCACCAAGGTCCAAGGCGATCCAGAGAATCCCAGGAAAAGCTCTTTGAAGGCCTACAGATCTATTGTTGTGAGCCCTTCACCAACATGCCCAAAGATGAGCTGGAGAGGATGCTGCAGCTGTGTGGGGCTTCTGTGGTGAAGGAGCTTCCATTGCTCACCCGTGACACAGGTGCTCATCCAATTGTGCTCGTGCAGCCAAGTGCCTGGACAGAAGACAACGACTGCCCTGATATTGGGCAGCTGTGCAAGGGACGTCTAGTGATGTGGGACTGGGTGTTGGACAGTATATCCGTCTACCGGTGTCGGGATCTGGATGCCTACCTGGTACAGAATATCACCTGTGGCCGTGATGGTAGCGAGCCACAAGACTCCAATGATTAATTTAAGAAATGGTCTTCCCGTCCCGGGAGCATCTCT

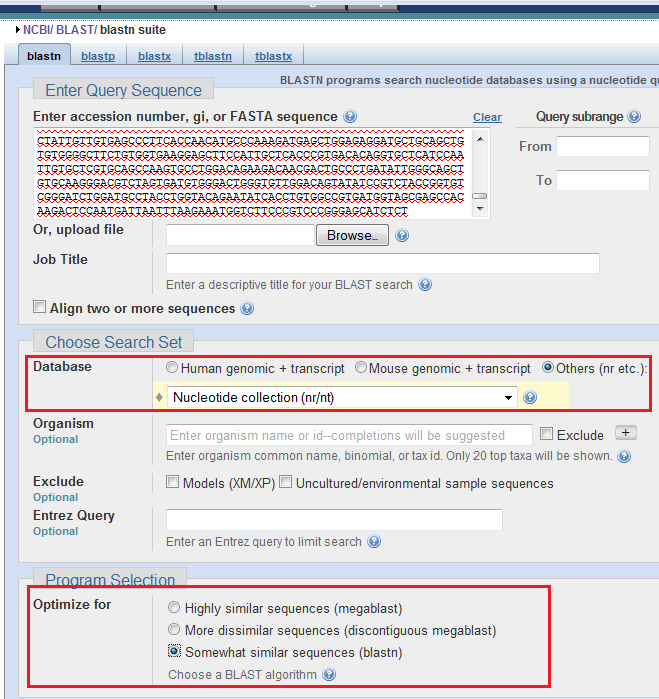
**3)** Now, because we have no header, we do not know which species this nucleotide sequence is from. Because of this, we need to change the database we are searching from “Human genomic plus transcript (Human G+T)” which only searches chromosomes and transcripts from *Homo sapiens* to “Nucleotide collection nr/nt” (illustrated below) which searches nucleotide sequences from all species.

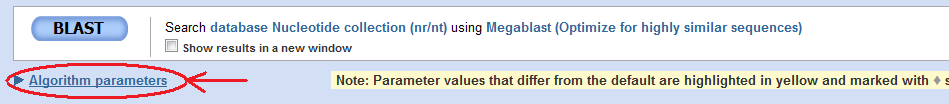
*When you change the database to “Nucleotide collection nr/nt”, you will have the option to include or disclose species using their taxonomy ID or their common name, as illustrated below. Do not do either for this search; however, sometimes this helps weed out species you know will be similar but do not want to search against or only include species you are interested in.*

**

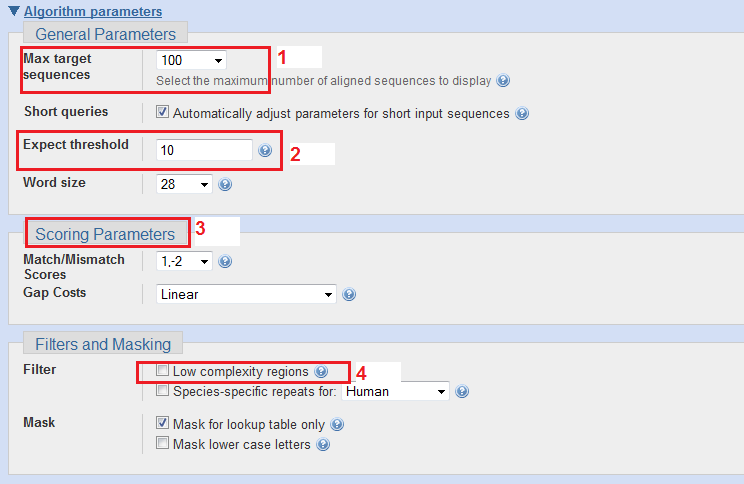
*Remember, we are not specifying specific organisms we want to search against or exclude, however, in future searches you should keep this option in mind as it may prove useful.*

We must also change the type of program used, under “Program Selection” as previously mentioned in this tutorial. When you have done this, your page should look like this (**note: make sure you changed the “Database” and “Program selection” field as illustrated below**):



Before we hit the “BLAST” button and begin the blast search, we first need to review the Algorithm Parameters for this BLAST search. To do this, click the blue link at the bottom of the page, below the BLAST button labeled “Algorithm parameters” as illustrated below:  


This should display a menu that looks like the one displayed below. Pay attention to the highlighted areas below, as they will be briefly discussed:



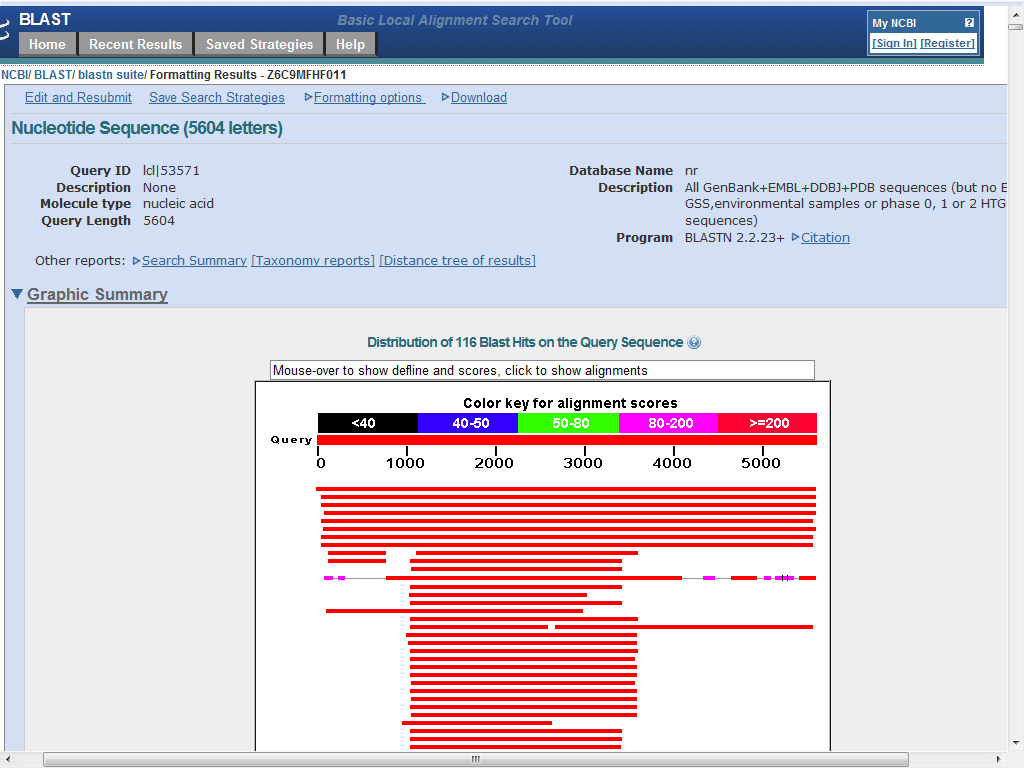
1 – This is the maximum number of sequences that will be displayed as matches. Sometimes you may want to increase or decrease this, depending upon your search. For our purposes, you can leave it at 100.

2 – The expect threshold is basically the sensitivity level; specifically it is a cut off number for the number of chances a random match may occur (e-value cut off). If you are using a small query sequence, you may want to increase the expect threshold (to 100 or 1000, again depending upon size of query), but for our purposes, you can leave it at 10.

3 – The scoring parameters are always set at default. This is how BLAST computes the score for each match. You may change this if you wish to use different Match/Mismatch/Gap scores; however, generally this is kept the same. For our purposes, leave this section as is.

4 – The filter determines if it will filter out certain user defined regions. Filtering out low complexity regions is specifically highlighted here. You should filter low complexity regions if you have a long, known genomic sequence to avoid random matches and to enhance the speed of alignment. Otherwise, do not filter low complexity regions. For our purposes, choose not to (uncheck this box) filter out low complexity regions.

**4)** Remember, before we continue, you need to change the filter (under Algorithm parameters) to not filter low complexity regions (uncheck the box labeled “Filter low complexity regions”. After you do this, click the BLAST button. If the system is busy, it may take awhile to align your sequence; however, it shouldn’t take more than 5 minutes. It generally only takes seconds. The page displayed should look like this:



A brief description of this graphical output is listed below:

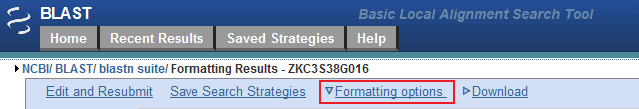
1) The top key where it says “Query” tells you how long the query (the sequence we used) is (5,604 nucleotides in our example). The thick lines below this are possible matches to your query. The more a line matches the length of your query sequence (and the more solid it is throughout), the more this possible match aligns with your query. Areas where there are gaps (no line drawn) indicate that this area in the possible match displayed does not match your query sequence at this location.

2) The color of the thick lines below the top key (the possible matches) indicate how strong of a possible match this specific area is. For instance, a red line indicates that the score is 200 or over for this area, which is generally a strong match. A blue line, however, indicates that this area has a score between 40 and 50, which is generally not a strong match.

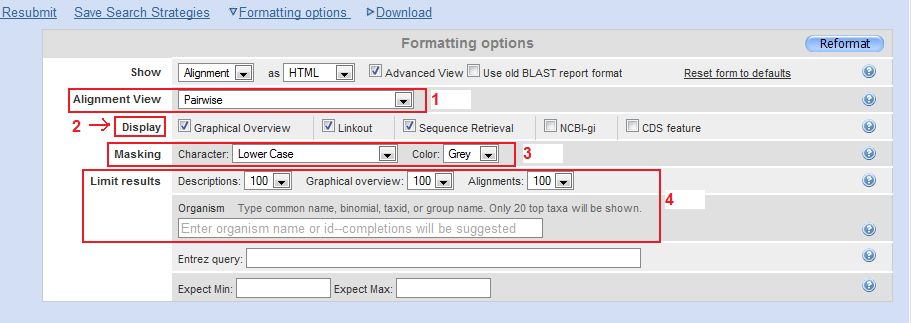
3) By scrolling your mouse pointer over a line, BLAST will provide you a brief description of the gene you scrolled over.

Looking at this graphical output gives us a good idea of how strong the matches are at a glance, but does not tell us enough to determine which possible match is a correct match. For instance, the first match seems to be perfect – it has a solid red line that follows our query from the first nucleotide to the last. To determine this for sure, however, we need to examine the analytical statistics (as well as the actual alignment) that BLAST has generated for us. These analytical statistics (e-value, score, and query coverage) are more important than this graphical output.

Now, before we go any further, let’s look at the formatting option BLAST provides. Click the “Formatting Options” link at the top of the page, as illustrated below:



Clicking this link should provide a drop down box, as illustrated below. Please pay particular attention to the highlighted regions, as they will be briefly described:



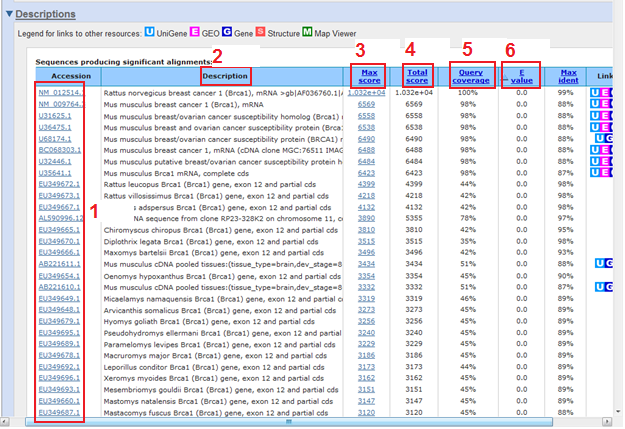
1 – This part is the “Alignment view”. By changing this, we can change how BLAST displays its alignment (which is produced further down this page). One option you can pick from the drop down box, labeled “Pairwise with dots for identities” is particularly useful; this option makes it so that every nucleotide that is matched in the alignment section has a dot. This makes it easier to find where the sequence did not align (noted by the absence of a dot). This is very useful for finding mutations, or polymorphisms.

2 – This part will change what is displayed on the results page from this BLAST search.

3 – This part lets you edit how BLAST displays masked areas. You can change the type of character displayed, as well as the color of the character displayed. By default, any masked area is displayed as a lower case grey letter.

4 – This part lets you change the BLAST results displayed. Much like when you select a database, here you can choose to include or disclose organisms (based off taxonomy ID or common name), choose how many alignments you want to include (which was listed under algorithm parameters) or choose how many matches you want displayed on the graphical output.

**5)** Now, scroll down the result page to the “Descriptions” section. We will take a moment reviewing this before continuing on. Your page should look like the page illustrated below (note the highlighted parts):



A brief description of the parts labeled is listed below:

1 – This is where you can find the accession number of your possible matches. By default, BLAST places the best scoring result (“Max Score”) at the top and the worst at the bottom. Clicking this accession number (blue text) will take you to the GenBank record for this gene.

2 – This is the full name of the match. This may be full genes or individual exons as illustrated here; it can also be other results, such as whole chromosomes or whole assemblies, depending upon the database searched.

3 – The Max score is the bit score for the gene (which is developed computationally using an algorithm that assigns different scores for matches, mismatches, and gaps) and is related to the e-value. The higher the score is the better of a possible match it is.

4 – The Total score is the sum of all high scoring pairs for the match in the same database. Sometimes this number will be different than the Max score. Sorting by Total Score can be useful because it has a better ability to predict exon matches, as opposed to intron pseudogene matches. Just as with the Max score, the higher the Total Score is the better of a possible match it is.

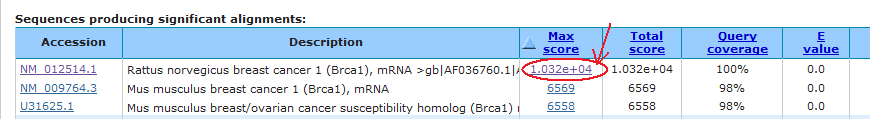
5- The Query coverage is the percent of the possible match that can be linked directly (paired) to the query sequence.

6 – The E-value, or expect value, is a computational prediction of how often your query seqeunce may be randomly matched to a pseudogene or incorrect match. Therefore, the lower this number is the better of a possible match this is.

Take a quick look at our results before we continue. Because I provided you with a very well annotated nucleotide sequence in multiple species, we have a multitude of good possible matches. Normally, your results will not look this good; however, for simplicity in this tutorial we used a BLAST search that resulted in strong possible matches, as weaker possible matches usually need further investigation than just a BLAST search to verify that they are the correct match. With weaker possible matches, however, BLAST is useful as a starting point to begin investigation in certain genes or sequences.

Now the question is which possible match is the correct match? In this case, it is very easy to determine this. While many possible matches have an e-value of 0 (a good indicator of a strong possible match because this means it is not a random match), only one possible match (the first match listed) has 100% query coverage. The query coverage is also a good indicator of a correct match because it tells you how much of the nucleotide sequence in this possible match can be directly linked to your query sequence. Because the first match has a low e-value and perfect query coverage, it has an extremely high score (all possible matches have a high score, however, the correct match – the first possible match listed – has an extremely high score). From the description you can see that the nucleotide sequence I provided you with is our old friend the BRCA1 gene again; this time, however, it is from a rat (*Rattus norvegicus)*. All the other matches generated by BLAST are orthologs to the BRCA1 gene in rats. They are highly similar, with well conserved regions, but are not identical.

**6)** Now while it is obvious which possible match is the correct match, let’s investigate this blastn result a little more. In less obvious BLAST searches, further investigation will be more helpful than just the analytical analysis (“Description” section). Click on the Max Score (blue lettering) for the first possible (and correct) match for our search, as illustrated below:



This should jump you down the page to the “Alignments” section, specifically the alignment of your provided nucleotide sequence against the nucleotide sequence for the BRCA1 rat gene. Your page should look like the following image. Pay special attention to the highlighted regions:



1 – By clicking this link, BLAST will show you a phylogenetic tree that will help show the evolutional relationships between your query sequence (your query sequence will be highlighted in the tree if you click this link) and other possible matches. This can be useful to compare species and estimate the distance of evolution.

2 – This part is the Identities, which is the number of direct matches our query sequence had to this possible match. In our example, we had almost identical matches (the loss of 3 nucleotides was most likely the result of me changing the nucleotide from FASTA format into Microsoft Word). A high Identity is a strong indicator of a good match.

3 – This part tells us what strand (remember, there’s plus and minus strands in DNA) our possible match is located on

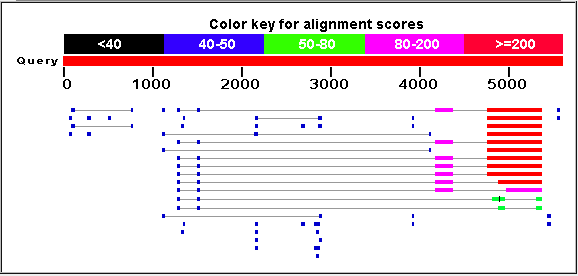
4 – This part tells us how many gaps there are between our query nucleotide sequence and this possible match. Remember, gaps result in large score penalties and are less likely to occur than mismatches, so a significant number of gaps can indicate an incorrect match. Again, the 3 gaps in our example are most likely a result of my error transferring the nucleotide sequence between formats.

5 – This part tells us where exactly on our query sequence (top letters) and where exactly on our possible match (bottom letters), matches were found. This can be very useful, especially when annotating genes. For instance, if we are blasting a gene (or single exon) against a species’ whole nucleotide sequence (genome), we need to know where the match exists in order to investigate the match further on the species’ genome. Also, if only a small part (less than 20%) of our query sequence matches to a large subject sequence, we can assume that this is most likely a random match.

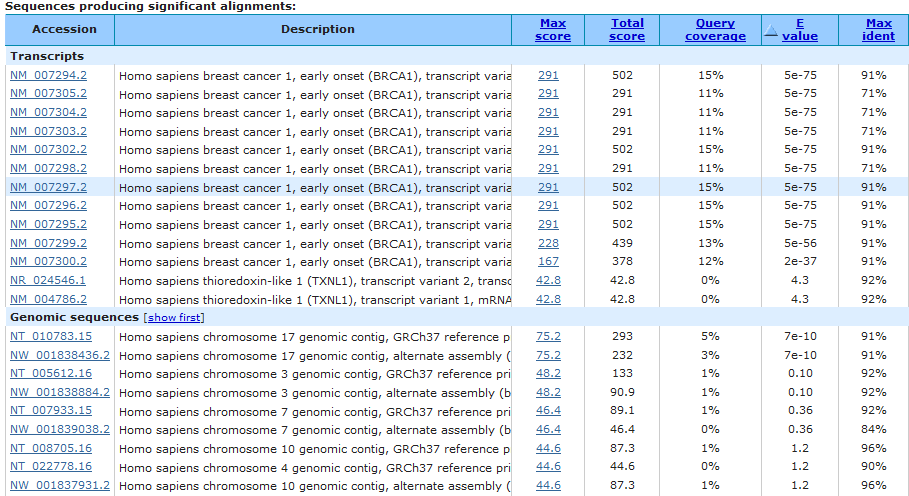
6 – This part tells us how long our subject (possible match) sequence is. In our case, it’s about as long as our query sequence (5604) which is a good sign of a correct match.

Now let’s take a look at a BLAST query that doesn’t produce such strong results. If you are doing research, or comparing genes/proteins/exons of two distantly related species your BLAST results will be much weaker. Below I will show you two examples: one example that illustrates a BLAST search with somewhat weak results, and another BLAST search that provides very weak results.

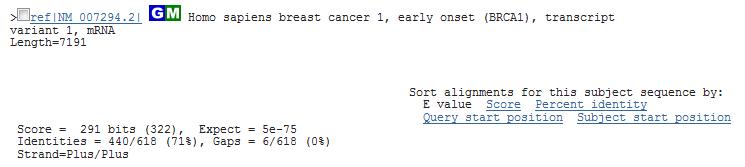
The following results are the graphical output and analytical (“Description”) output from a BLAST search of BRCA1 in the species *Gallus gallus domesticus* (chickens) against the human genomic + transcripts database.



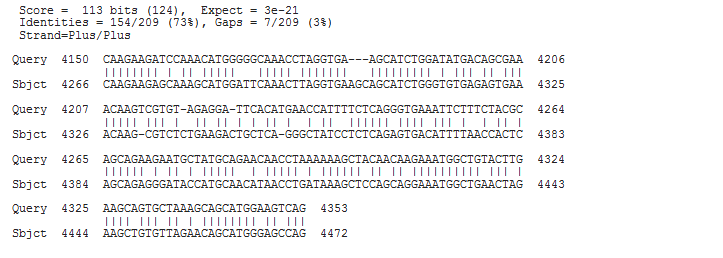
Just from our graphical output we can see that there are no very strong matches between the BRCA1 gene in humans and the BRCA1 gene in chickens (or any other genes for that matter). There is one area in the BRCA1 gene that is highly conserved between species (the solid red line). Remember, as previously discussed, matches are illustrated by a thicker line. The strength of the match is indicated by the color of the thicker line. As you can see, there seems to be good conservation (indicated by a thick red line) between species in the first few possible matches between nucleotides 4750-5400 (approx.) and a mild amount of conservation (indicated by a thick pink line) between species between nucleotides 4200-4300 (approx.). There are other matches indicated by the small, thick blue line; however, remember blue matches are not very strong matches. There’s a good chance that these blue matches are random matches. Let’s look at the analytical (“Description) data generated by BLAST (next page)



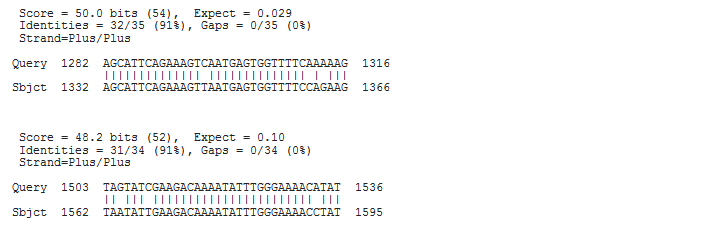
Now remember, the reason these results look different from our previous results is because I used a different database for these results (Human Genomic plus Transcripts) whereas we used the Nucleotide Collection before. That is why we have both genomic sequences and transcripts. From these results we can deduce a few things. By examining how close to zero the e-value is for the Transcripts part of the page, we can tell that these possible matches are not just random matches. By examining the e-value for the Genomic Sequences part of the page, we can tell which chromosome BRCA1 is located on in humans (chromosome 17). By looking at the query coverage, we can also tell that not much of the BRCA1 gene is conserved between chickens and humans (approximately 15% is conserved in transcript variant 1 – the first and best possible match). Furthermore, because no possible match is significantly matched, we can also tell that the evolutionary distance between us (humans) and chickens is large. Let’s take a look at the best transcript alignment. As you will see, the best possible match has four separate alignments to the BRCA1 gene. This means that BLAST found similarities at four separate regions on the BRCA1 gene in humans when compared to the BRCA1 gene in chickens. We will look at each alignment separately (next two pages):



This first alignment has a very low e-value, a good amount of identities (71%) and most importantly, a very small amount of gaps (6/618 for 0%); this means that this section is highly conserved (similar) between species. Remember, this is the region that in the graphical output was shown as a solid red line.



This second alignment still produces good results (a very low e-value, 73% identity and 3% gaps) and contains a fair amount of nucleotides (209) so we can assume that this section is fairly well conserved between species and not a random match. Remember, this was the region in the graphical output that was a solid pink line.



These final two alignments do not have very good results. Both have e-values approaching 1, which means that there is a chance this is a random match. Furthermore, because both alignments use such small sequences (35 and 34 nucleotides, respectively) this shows us that there is only a little conservation in these areas. Because only one area was matched significantly (the second alignment wasn’t a strong match) for only approximately 14-15% of the sequence, we can again conclude that the evolutionary distance between these two species is far.

# Now let’s take a look at a very poor BLAST query – comparing the BRCA1 gene in humans to the BRCA1 gene in *Strongylocentrotus purpuratus* (California purple sea urchin) (next page):

# 

# As you can see from the graphical output, the matches are very poor – this tells us there is little conservation between species and the evolutionary distance is very great. Remember, thick red lines indicate a good match. In this example, the best match is denoted by a short, thick green line. Let’s take a look at the analytical data (“Description” section) for further investigation:

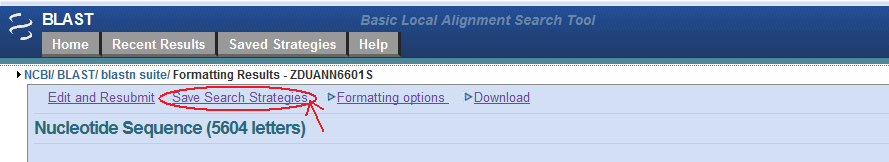
# 

# Taking a quick look at the results shows us that we have one match with some similarity (the first match) because of its low e-value. Notice that the query coverage is very low; however, keep in mind that this is the query coverage for this gene against a whole human chromosome (chromosome 12 for the first match), so we wouldn’t expect the query coverage to be a large number. Notice also that we have no transcripts; this means that the predicted BRCA1 gene in sea urchins isn’t similar at all to the BRCA1 gene (or any gene, for that matter) in humans. Let’s investigate the best possible match’s (first possible match) alignment for further investigation.

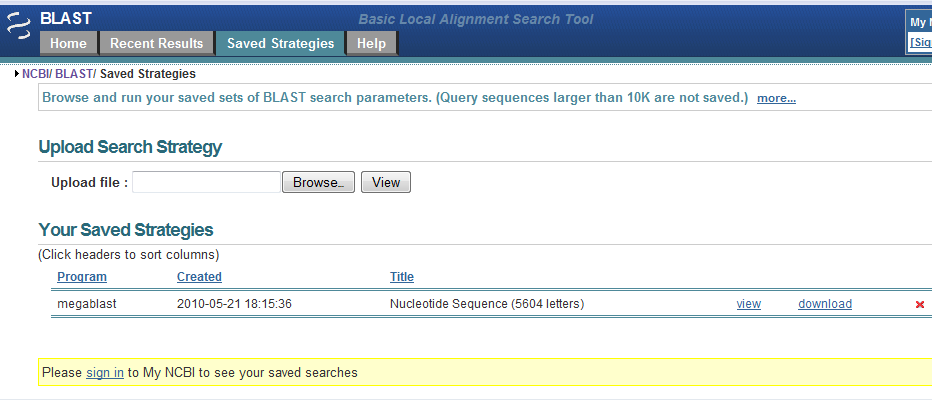
# 

# While our e-value is low, looking at this alignment tells us that this is most likely a random match (note: the lower case letters are masked nucleotides – remember, masked nucleotides are nucleotides deemed to have low compositional complexity and are therefore biologically uninteresting). Our identities percent is good (83%) but it is only out of 55 nucleotides. Furthermore, we have 6/55 gaps (10%). Our query sequence (the predicted BRCA1 gene in sea urchins) was 7,400 nucleotides long and we blasted it against a whole entire human chromosome. Therefore it’s very likely for random matches to occur, especially small random matches. This means that this alignment is more than likely a random match.

**7)** We have completed blasting a nucleotide sequence and matching it to a gene (in our case, to *Rattus norvegicus’* BRCA1 gene) as well as investigating other BLAST queries. Remember, for more information about any possible matches generated by BLAST, simply click on the Accession number in the “Description” section to be linked to the GenBank record for that match. Now that we’ve completed this BLAST search, let’s save the results BLAST provided us with so it can be accessed quickly in the future. Saving a blast result is very easy. Simply click “Save search strategies” located at the top of the page, as illustrated below:



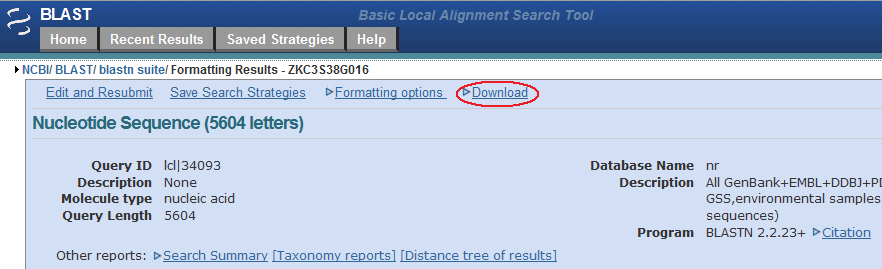
Clicking this link should produce this page (if you have other saves, it will look different) (next page):



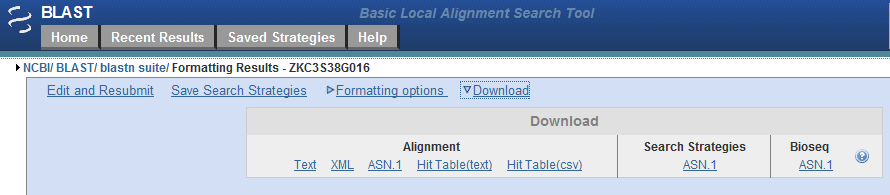
Note: if your query sequence is longer than 10,000 nucleotides/amino acids, it will not be saved.

If you have other saves (using different programs or records from NCBI) you can login to view all of your saves here. In a large project, this can save a lot of time because you will be able to view saved BLAST searches and NCBI records in just one click.

**8)** If you’d prefer to download the results onto your computer from BLAST (as opposed to saving them), BLAST provides you with this option as well. Look at the top of the page for the link labeled “Download” as illustrated below:



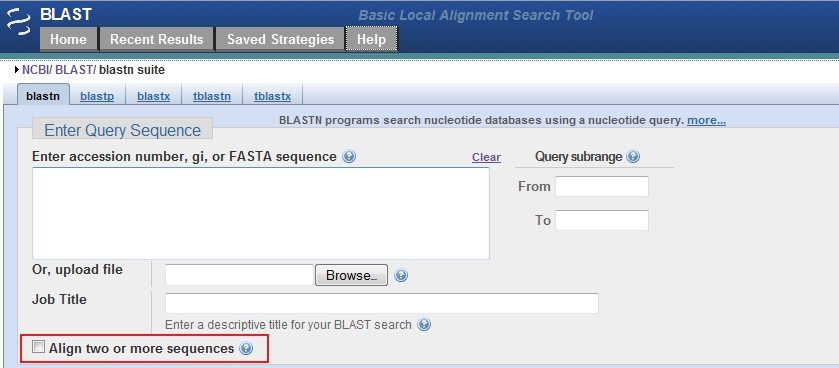
After clicking “Download”, a box should appear as illustrated below:



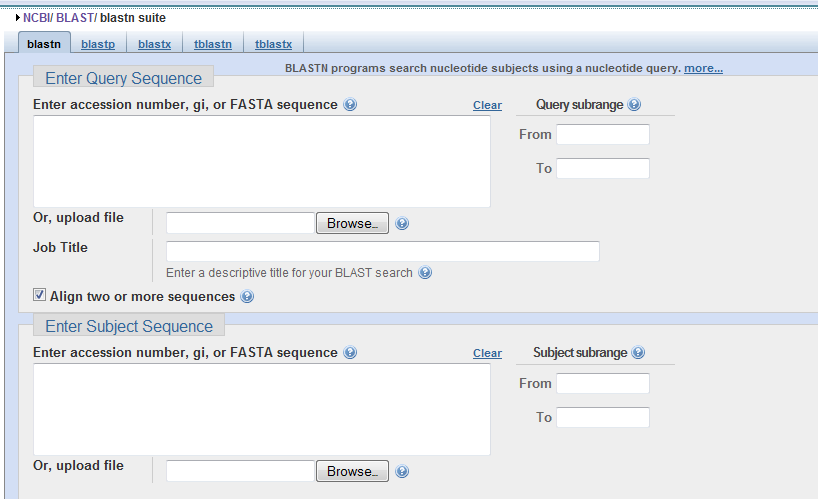
The two most important things you can save are the alignment (which can be useful when annotating genes) or your search strategies. While BLAST can save this online for you, it may be useful to have an off-line version for personal use.

**9)** Let’s take a look at another feature blastn (or any basic BLAST search) can perform for us – aligning two or more sequences. Previously in the tutorial I highlighted where you can choose if you wish to align multiple sequences, but now we will go a little more in-depth into this option in this part of the tutorial. Aligning two or more sequences is helpful to compare sequences (for similarities) you have that may not be already in NCBI’s database, that may be snippets of genes, or just to narrow your search between two specific sequences.

Go back to the BLAST homepage (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>). We will be aligning two nucleotide sequences, so choose “Blast nucleotide” as we have done before. After the page loads, select the checkbox “Align two or more sequences” as illustrated below:

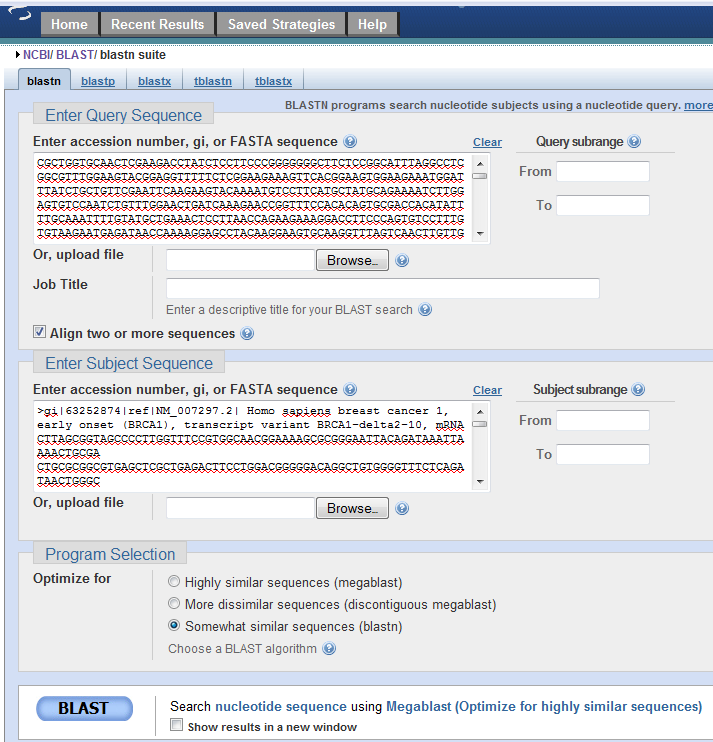


Selecting this should change your page to look like this:



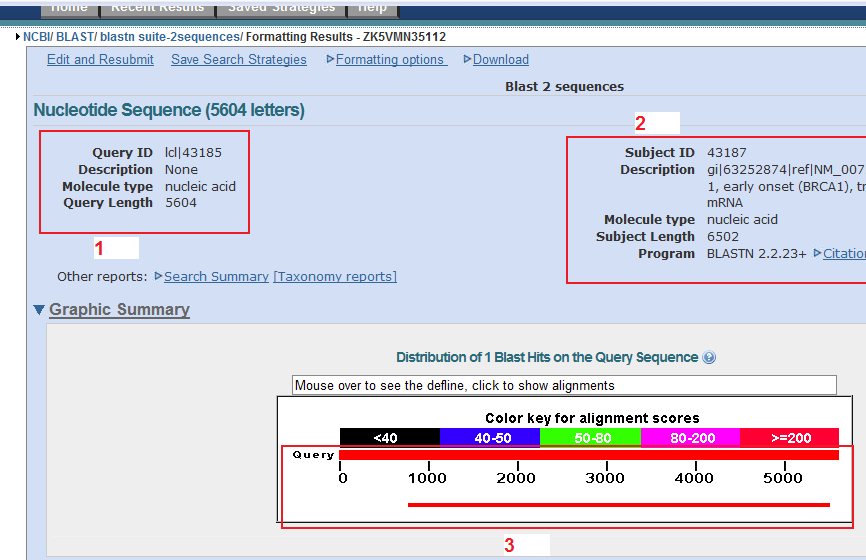
We now have two places to input new sequences. For the first query sequence, enter in the sequence for *Rattus norvegicus*’s BRCA1 gene that I have already provided for you (located on page 5 of this tutorial). For the subeject sequence, we will use the nucleotide sequence for *Homo sapiens*’ BRCA1 isoform 3 (link provided: <http://www.ncbi.nlm.nih.gov/nuccore/63252874?report=fasta>).

Once both of these sequences have been pasted in, your page should look like this:

****

Note: make sure you have selected “Somewhat similar sequences (blastn) and that your algorithm parameters are the same as our last search (do not mask low complexity regions)

Now, hit the BLAST button. Your page should look something like the following page. Pay close attention to the highlighted regions, as they will be briefly discussed:

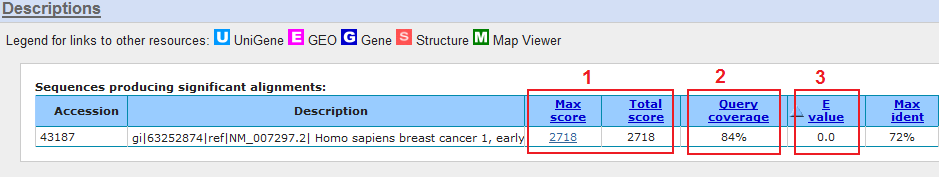


1 - This part is a description of the first sequence we entered, the rat BRCA1 gene. This is referred to as the query sequence.

2 – This part is a description of the second sequence we entered, the human BRCA1 gene. This is referred to as the subject sequence.

3 – Look over this graphical representation. Remember, the first sequence (the rat BRCA1 gene) is the “Query” part attached to the key. The second sequence (the human BRCA1 gene) is the red line below the key. Note that while these are both BRCA1 genes, the matching (pairing) is not 100%, however, where matches are made, there is a large amount of similarity (indicated by a red line, as opposed to a pink, green, or blue line). We would not expect the rat BRCA1 gene to match identically to the human BRCA1 gene, so this is normal. Also, note that our match to the human BRCA1 gene begins somewhere around nucleotide 800 on our query (rat) and extends to somewhere around nucleotide 5600 on our query (rat), however, the human BRCA1 gene is over 6502 nucleotides long (this means we have a ~1700 nucleotides unmatched). Furthermore, there is only one match. This means that the rat BRCA1 gene only matches to the human BRCA1 gene in one part of the sequence. All of this will be discussed later.

Now let’s take a look at a more important aspect, the description (analytical data). The analytical data is shown below. Pay attention to the highlighted parts, as we will talk about them briefly (next page):



1 – Notice we have a fairly high score. This indicates a good match as expected.

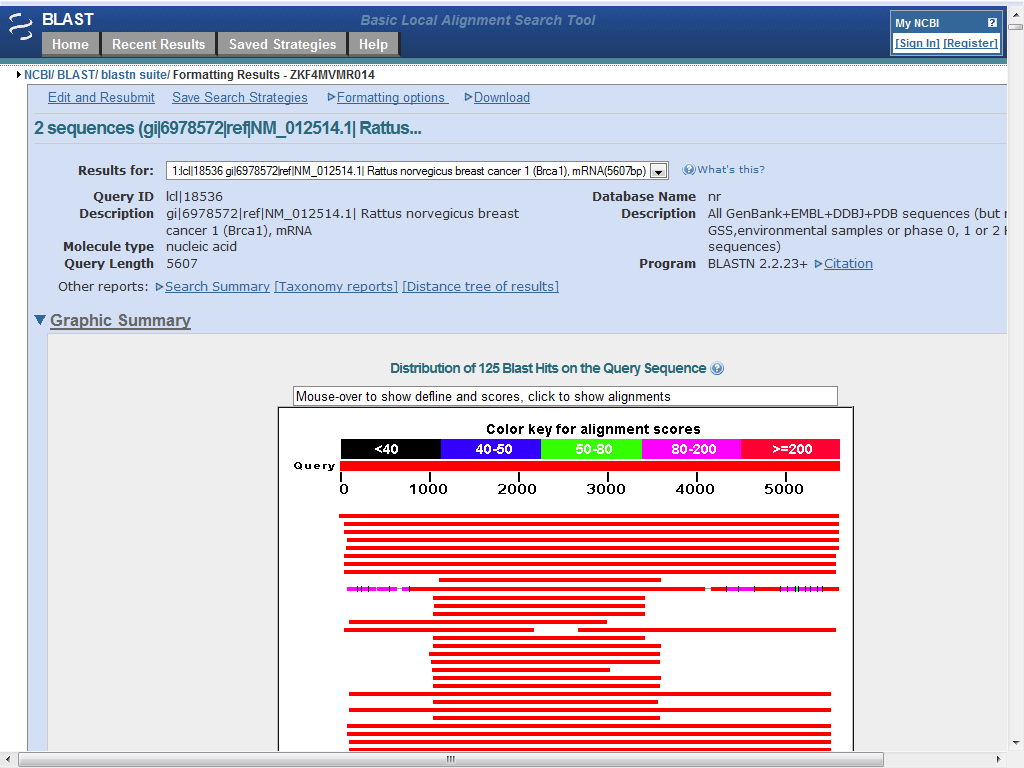
2 – We also have fairly good query coverage. Because this is between two different species (but from somewhat related species) we wouldn’t expect 100% query coverage. This shows us, however, that the rat BRCA1 gene is similar to the human BRCA1 gene.

3 – Our e-value is 0, which is good, as this means it is not just a random match, and that these two genes are similar.

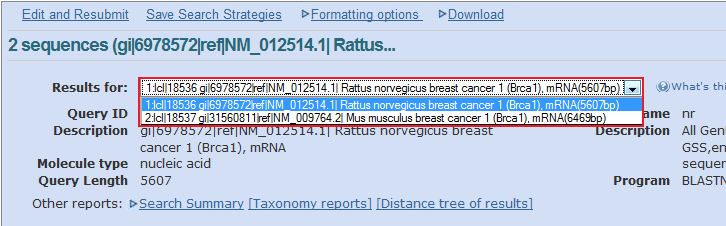
Now scroll down to the “Alignments” part of the page. Take a look at the section labeled “Identities”. Our identities for this query were 3567/4926 (72%) and we had 216 gaps out of 4926 (4%) matched areas. This is yet again a good indicator of a highly conserved region – while both sequences weren’t matched to each other 100% at every location, we had a minimal number of gaps. Remember that gaps are much less likely to occur and are penalized more by BLAST’s scoring parameters.

Furthermore, note that just as we saw in the graphical output, the first alignment is at nucleotide 780 on our query (rat) to nucleotide 178 on our subject (human) and that the last alignment is at nucleotide 5515 on our query (rat – note this isn’t the end of the rat’s BRCA1 gene) to nucleotide 5077 on our subject (human – again, note this isn’t the end of the human’s BRCA1 gene). This lets us know that this region (between 780-5155 on rats and between 178-5077 on humans for the BRCA1 gene) is highly conserved between these species, but the rest of the gene isn’t. We know the rest of the gene isn’t highly conserved between species because BLAST only provided us with one alignment. If there were other areas of BRCA1 that were conserved between species, BLAST would have provided us with a separate alignment for these areas.

**10)** Finally, let’s look at one more feature BLAST has involving multiple sequences. Let’s say you wanted to find orthologs to multiple sequences, but only wanted to perform a BLAST search once. Or perhaps you have a file you want to upload that contains multiple sequences (separated by headers). Blast can handle this. Simply upload your file, or paste multiple sequences (separated only by their header) into the “Query Sequence” box, and BLAST will align both of them separately. I will not show you this step as it is very simplistic, however, try any two sequences (I used BRCA1 in mice and rats in my next example) and blast them. You will notice that blast will work on the first provided sequence first, then the second, and so on. A page that looks like this should be the result of a blastn search with multiple queries (next page):



Notice the dropdown box at the top of the page that wasn’t there before. By clicking that box, as illustrated below, you can view either of the sequences you BLAST’ed individually, on its own page:



While doing this is never necessary, if you have to BLAST and compare multiple sequences (especially if they’re already uploaded in FASTA format in one file) BLAST can save you time by handling multiple sequences, as long as they are separated by a header.