cat RNAfileMatches4.csv | sed 's/sequence10.fasta-align.hmm/Snypr/g' > RNAfileMatches5.csv

cat RNAfileMatches5.csv | sed 's/sequence1.fasta-align.hmm/Yc2/g' > RNAfileMatches6.csv

cat RNAfileMatches6.csv | sed 's/sequence2.fasta-align.hmm/Asc-2/g' > RNAfileMatches7.csv

cat RNAfileMatches7.csv | sed 's/sequence6.fasta-align.hmm/Ptpn5/g' > RNAfileMatches8.csv

cat RNAfileMatches8.csv | sed 's/sequence8.fasta-align.hmm/Atp12A/g' > RNAfileMatches9.csv

cat RNAfileMatches9.csv | sed 's/sequence9.fasta-align.hmm/Lhx2/g’ > RNAfileMatches10.csv

cat RNAfileMatches4.csv | sed 's/sequence10.fasta-align.hmm/Snypr/g' > RNAfileMatches5.csv

cat RNAfileMatches4.csv | sed 's/sequence10.fasta-align.hmm/Snypr/g' | sed 's/sequence1.fasta-align.hmm/Yc2/g' | sed 's/sequence2.fasta-align.hmm/Asc-2/g' | sed 's/sequence6.fasta-align.hmm/Ptpn5/g' | sed 's/sequence8.fasta-align.hmm/Atp12A/g' | sed 's/sequence9.fasta-align.hmm/Lhx2/g' > RNAfileMatchesFinal.csv

Quantitative

-compare percent matches for animal and human – what does each start with

e values

# of matches

The Expect value (E) is a parameter that describes the number of hits one can "expect" to see by chance when searching a database of a particular size. It decreases exponentially as the Score (S) of the match increases. Essentially, the E value describes the random background noise. For example, an E value of 1 assigned to a hit can be interpreted as meaning that in a database of the current size one might expect to see 1 match with a similar score simply by chance.

The lower the E-value, or the closer it is to zero, the more "significant" the match is. However, keep in mind that virtually identical short alignments have relatively high E values. This is because the calculation of the E value takes into account the length of the query sequence. These high E values make sense because shorter sequences have a higher probability of occurring in the database purely by chance. For more details please see the calculations in the [BLAST Course](https://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html).

The Expect value can also be used as a convenient way to create a [significance threshold](https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=BlastHelp#expect) for reporting results. You can change the Expect value threshold on most BLAST search pages. When the Expect value is increased from the default value of 10, a larger list with more low-scoring hits can be reported.

For this problem we selected transcripts 1 and 9 for extended evaluation in BLASTWhen the *Optimize* function of the search was changed from *megablast* (highly similar sequences) to *discontinuous megablast* (more dissimilar sequences) and *blastn* (somewhat similar sequences), there were changes in the number of hits, percent match, and e-value results from both the human and mouse genomic databases. It was interesting to compare sequences 1 and 9, because sequence 9 generally had more matches for both the human and mouse genomic databases at each of the optimization levels. Sequence 9 also had higher percentage identity matches. Lastly, to compare the human and mouse results, the mouse percent matches were higher overall, the e-values were lower, and the number of hits were greater, as would be expected since the original transcripts were from rodents.

Human Database Comparisons:

For sequence 1, moving from megablast to discontinuous megablast to blastn, the number of hits increased from 6 to 18 to 55. So, as the criteria is loosened, there are more possible matches. In terms of e-values, each optimize search had results with e-values of zero, however the range increased significantly to up to 5.1 during the blastn search. The closer the E-value is to zero, the more “significant” the match is. This means it is less likely that matches will be found by chance. Finally, for the percent match in identity of the transcript sequences, the megablast range was 82-86%, the discontinuous megablast range was 15-79%, and the blastn range was 70-93%. The megablast range was largest. For sequence 9, moving from megablast to discontinuous megablast to blastn, the number of hits increased from 7 to 34 to 75. So, as the criteria is loosened, there are more possible matches. For the e-values, the megablast range was from 0 to 1e-79, which is very small compared to the range of 0 to 1.5 for both the discontinuous megablast and blastn. The percent match range was again small and very high for megablast as 82-90%, but broadened to 71-91% and 70-93% for discontinuous megablast and blastn, respectively.

Mouse Database Comparisons:

For sequence 1, moving from megablast to discontinuous megablast to blastn, the number of hits increased from 8 to 10 to 30. The-evalue range grew again from 0 up to 1.2 for the blastn. The percent match in identity of the transcript sequences, for megablast was 78-100%, the discontinuous megablast and blastn ranges were both 69-100%.

For sequence 9, moving from megablast to discontinuous megablast to blastn, the number of hits increased from 8 to 29 to 79. This was a substantial increase in the number of matches as the criteria moved from highly similar to somewhat similar sequences . For the e-values, the all megablast results were 0, while the discontinuous megablast and blastn values were up to 0.35 and 1.2, respectively. The percent match range was again small and very high for megablast as 5 hits were 99% matches and 3 were 100%. The ranges then increased to 70-100% and 66-100% for discontinuous megablast and blastn.