1) Michael Crichton's fantasy about cloning dinosaurs, Jurassic Park contains a putative dinosaur DNA sequence. Use nucleotide-nucleotide BLAST against the default nucleotide database to identify the real source of the following sequence:

>DinoDNA "Dinosaur DNA" from Crichton's JURASSIC PARK p. 103 nt 1-1200

GCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGC

GGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCG

TGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTGGC

TGCTCACGCTGTACCTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTG

CCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGTAA

AGTAGGACAGGTGCCGGCAGCGCTCTGGGTCATTTTCGGCGAGGACCGCTTTCGCTGGAG

ATCGGCCTGTCGCTTGCGGTATTCGGAATCTTGCACGCCCTCGCTCAAGCCTTCGTCACT

CCAAACGTTTCGGCGAGAAGCAGGCCATTATCGCCGGCATGGCGGCCGACGCGCTGGGCT

GGCGTTCGCGACGCGAGGCTGGATGGCCTTCCCCATTATGATTCTTCTCGCTTCCGGCGG

CCCGCGTTGCAGGCCATGCTGTCCAGGCAGGTAGATGACGACCATCAGGGACAGCTTCAA

CGGCTCTTACCAGCCTAACTTCGATCACTGGACCGCTGATCGTCACGGCGATTTATGCCG

CACATGGACGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAA

CAAGTCAGAGGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAA

GCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCTTCGGG

CTTTCTCAATGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTG

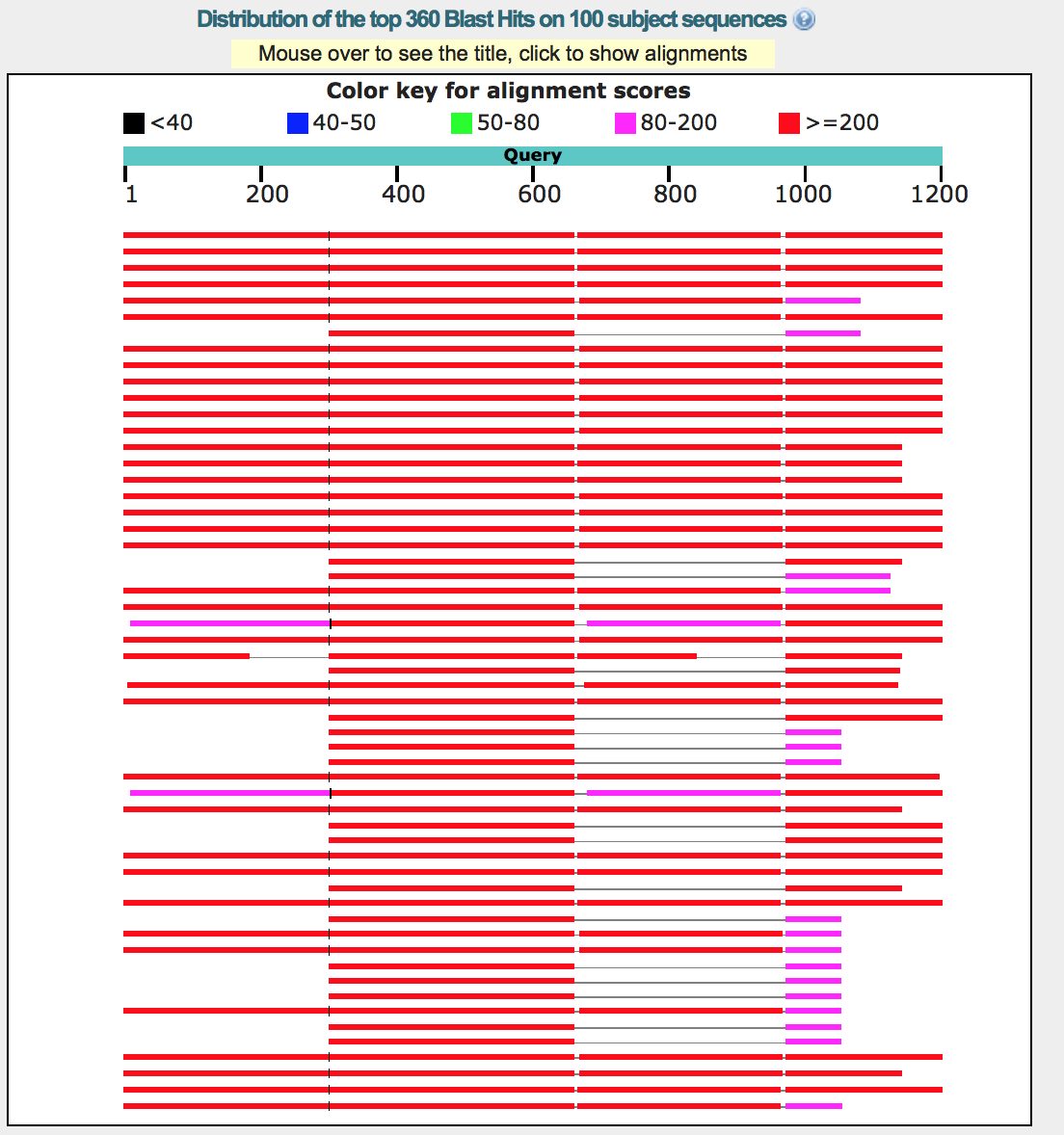
ACGAACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCA

ACACGACTTAACGGGTTGGCATGGATTGTAGGCGCCGCCCTATACCTTGTCTGCCTCCCC

GCGGTGCATGGAGCCGGGCCACCTCGACCTGAATGGAAGCCGGCGGCACCTCGCTAACGG

CCAAGAATTGGAGCCAATCAATTCTTGCGGAGAACTGTGAATGCGCAAACCAACCCTTGG

CCATCGCGTCCGCCATCTCCAGCAGCCGCACGCGGCGCATCTCGGGCAGCGTTGGGTCCT



­­

2) Mark Boguski of the NBCI noticed this and supplied Crichton with a better sequence for the sequel, The Lost World. Identify the most likely source of this sequence using nucleotide-nucleotide BLAST. Mark imbedded his name in the sequence he provided. To see Mark's name use the translating BLAST (blastx) page the sequence below. (Look for MARK WAS HERE NIH).

>DinoDNA "Dinosaur DNA" from Crichton's THE LOST WORLD p. 135

GAATTCCGGAAGCGAGCAAGAGATAAGTCCTGGCATCAGATACAGTTGGAGATAAGGACG

GACGTGTGGCAGCTCCCGCAGAGGATTCACTGGAAGTGCATTACCTATCCCATGGGAGCC

ATGGAGTTCGTGGCGCTGGGGGGGCCGGATGCGGGCTCCCCCACTCCGTTCCCTGATGAA

GCCGGAGCCTTCCTGGGGCTGGGGGGGGGCGAGAGGACGGAGGCGGGGGGGCTGCTGGCC

TCCTACCCCCCCTCAGGCCGCGTGTCCCTGGTGCCGTGGGCAGACACGGGTACTTTGGGG

ACCCCCCAGTGGGTGCCGCCCGCCACCCAAATGGAGCCCCCCCACTACCTGGAGCTGCTG

CAACCCCCCCGGGGCAGCCCCCCCCATCCCTCCTCCGGGCCCCTACTGCCACTCAGCAGC

GGGCCCCCACCCTGCGAGGCCCGTGAGTGCGTCATGGCCAGGAAGAACTGCGGAGCGACG

GCAACGCCGCTGTGGCGCCGGGACGGCACCGGGCATTACCTGTGCAACTGGGCCTCAGCC

TGCGGGCTCTACCACCGCCTCAACGGCCAGAACCGCCCGCTCATCCGCCCCAAAAAGCGC

CTGCTGGTGAGTAAGCGCGCAGGCACAGTGTGCAGCCACGAGCGTGAAAACTGCCAGACA

TCCACCACCACTCTGTGGCGTCGCAGCCCCATGGGGGACCCCGTCTGCAACAACATTCAC

GCCTGCGGCCTCTACTACAAACTGCACCAAGTGAACCGCCCCCTCACGATGCGCAAAGAC

GGAATCCAAACCCGAAACCGCAAAGTTTCCTCCAAGGGTAAAAAGCGGCGCCCCCCGGGG

GGGGGAAACCCCTCCGCCACCGCGGGAGGGGGCGCTCCTATGGGGGGAGGGGGGGACCCC

TCTATGCCCCCCCCGCCGCCCCCCCCGGCCGCCGCCCCCCCTCAAAGCGACGCTCTGTAC

GCTCTCGGCCCCGTGGTCCTTTCGGGCCATTTTCTGCCCTTTGGAAACTCCGGAGGGTTT

TTTGGGGGGGGGGCGGGGGGTTACACGGCCCCCCCGGGGCTGAGCCCGCAGATTTAAATA

ATAACTCTGACGTGGGCAAGTGGGCCTTGCTGAGAAGACAGTGTAACATAATAATTTGCA

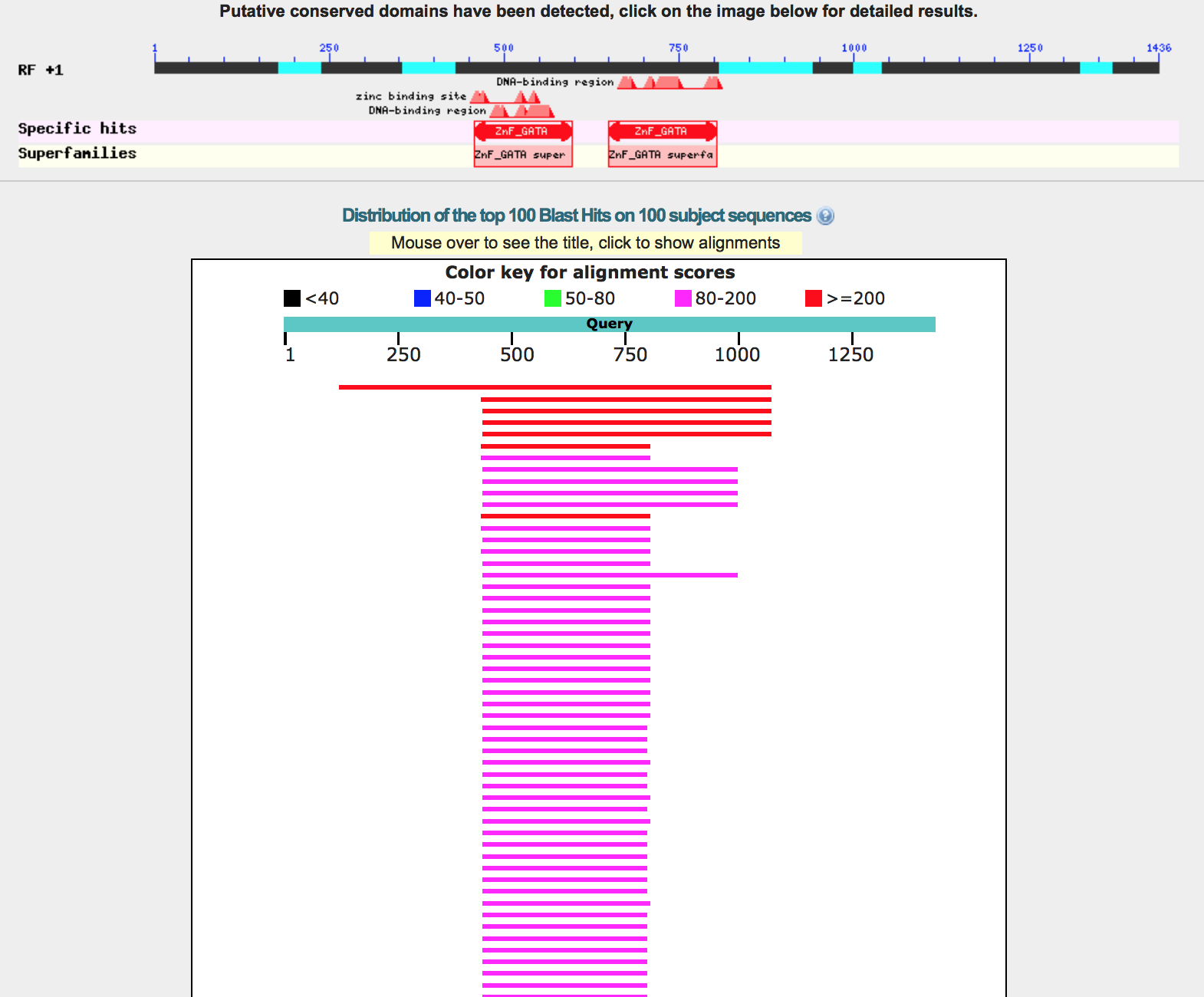
CCTCGGCAATTGCAGAGGGTCGATCTCCACTTTGGACACAACAGGGCTACTCGGTAGGAC

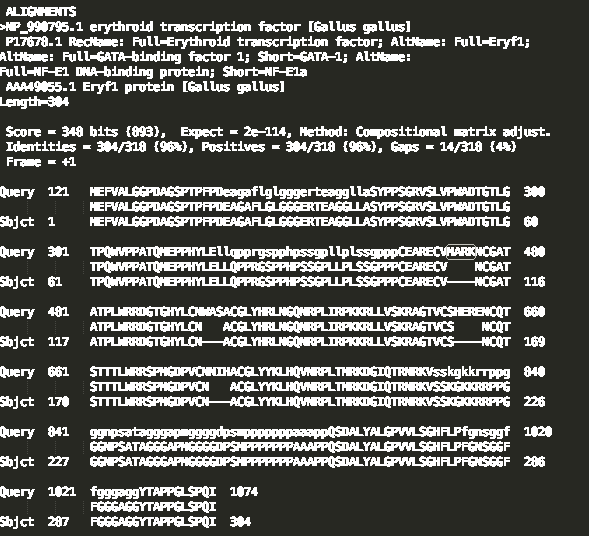
CAGATAAGCACTTTGCTCCCTGGACTGAAAAAGAAAGGATTTATCTGTTTGCTTCTTGCT

GACAAATCCCTGTGAAAGGTAAAAGTCGGACACAGCAATCGATTATTTCTCGCCTGTGTG

AAATTACTGTGAATATTGTAAATATATATATATATATATATATATCTGTATAGAACAGCC

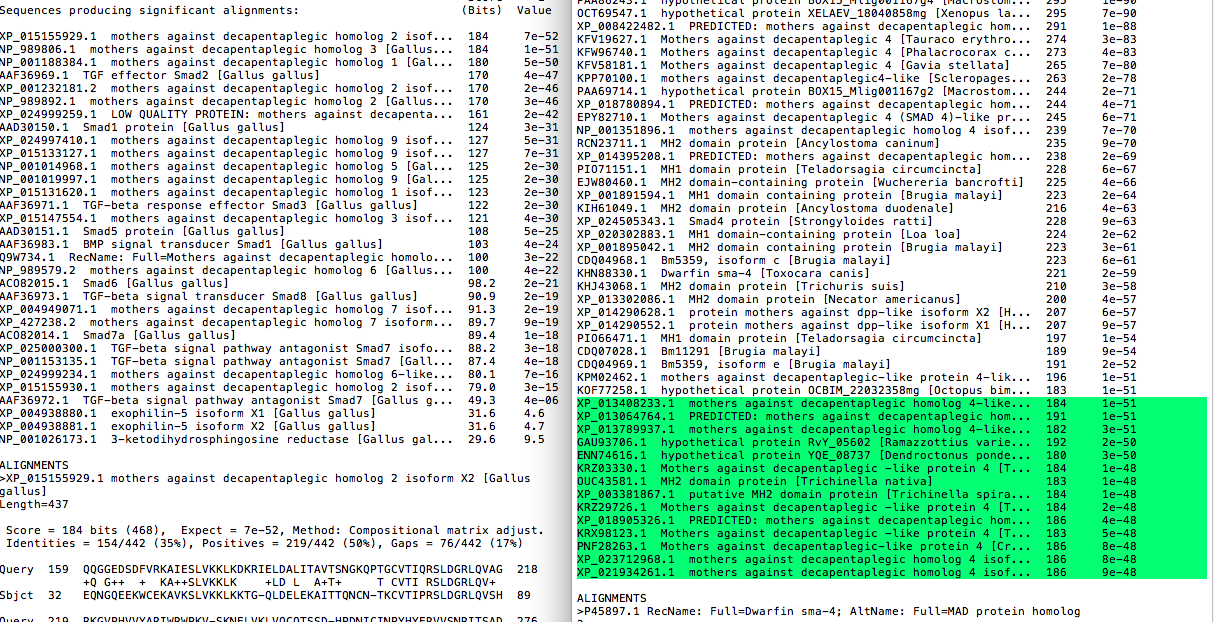
TCGGAGGCGGCATGGACCCAGCGTAGATCATGCTGGATTTGTACTGCCGGAATTC





**In the GAPS**

3) The C. elegans gene SMA-4 is a member of the dwarfins gene family, which play a role in TGF-mediated signal transduction. In order to identify potential homologs in other species, use the protein-protein blast page to perform search against the non redundant protein database (nr) using SMA-4 (accession number P45897) as the query sequence. Find all chicken (Gallus gallus) proteins that are similar to SMA-4. (Use the Tax Blast link at the upper left of the graphic to help in finding the chicken proteins.) Now run the search again and restrict to chicken proteins through the Entrez query advanced option. What proteins are found? Compare the Expectation values of these hits to the same hits found against nr with no organism restriction. Why are the E values different for the same scores and alignments?



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.

4) The human fragile histidine triad protein (FHIT) (SWISS-PROT: P49789) has been shown to be structurally homologous to galactose-1-phosphate uridylyltransferase. However this relationship is not apparent in an ordinary BLAST search. Perform a protein-protein blast search against the swissprot database with P49789 and search your results for galactose-1-phosphate uridylyltransferases.

>sp|**P49789**|FHIT\_HUMAN Bis(5'-adenosyl)-triphosphatase OS=Homo sapiens OX=9606 GN=FHIT PE=1 SV=3

MSFRFGQHLIKPSVVFLKTELSFALVNRKPVVPGHVLVCPLRPVERFHDLRPDEVADLFQ

TTQRVGTVVEKHFHGTSLTFSMQDGPEAGQTVKHVHVHVLPRKAGDFHRNDSIYEELQKH

DKEDFPASWRSEEEMAAEAAALRVYFQ

galactose-1-phosphate uridylyltransferases.

>NC\_000009.12:34646589-34650598 Homo sapiens chromosome 9, GRCh38.p12 Primary Assembly

ACGTGCTGGCGGCTGGCCAATCATCGGGGGCGGCGCGGGGAGGGGTGGTGTGGACGGAGAAAGTGAAAGG

TGAGGCACGGCCCTGCAGATTTTCCAGCGGATCCCCCGGTGGCCTCATGTCGCGCAGTGGAACCGATCCT

CAGCAACGCCAGCAGGCGTCAGAGGCGGACGCCGCAGCAGCAACCTTCCGGGCAAACGGTAACTGCACCG

CGGCAGGGACTCGCTGGGGCGCGGAGCCGAGCCCTCCCCTTCCTTAGGAAGCTTTCGTCCCCTCCGAAGG

TTGGAACGCTCATCCCGAGCCAGACCGACAAGGCGTACAGTCTGCAGGCCTGTACGAGCAGCAGGCCAAT

TGGCGCTGGGAAAGTCCAATCCTGGGCCTCTAGCTCCTGAGCGGGACAGGGCCGAGAGGGCGCTCCCGAG

CTTGGGCCTGCTGGTGGGTGAGACCCAGGAGAGAGGGAGCTAGAGAGCTCTGAGGACTGATCTTGACTGT

CTGCCCCCAGACCATCAGCATATCCGCTACAACCCGCTGCAGGATGAGTGGGTGCTGGTGTCAGCTCACC

GCATGAAGCGGCCCTGGCAGGGTCAAGTGGAGCCCCAGCTTCTGAAGACAGTGCCCCGCCATGACCCTCT

CAACCCTCTGTGTCCTGGGGCCATCCGAGCCAACGGAGAGGTAAGCCTGTAGAGCCCTGCATCTGCAGGC

TGGGCCACGGGGAGTAGTTCCCTCTTAGAACTGTCCTCCACCCACAGGGATAGTGAACCTCCTTCTGGGT

CATATCCCACCAAGCTTTTTGGTCCCCTAGGGTGGGCCTTCCCTACTCCCTTGTAGCCTGTCCAGTCTTT

GAAGCCCACCAGGTAACTGGTGGTATGGGGCAGTGAGTGCTTCTAGCCTATCCTTGTCGGTAGGTGAATC

CCCAGTACGATAGCACCTTCCTGTTTGACAACGACTTCCCAGCTCTGCAGCCTGATGCCCCCAGTCCAGG

TAACCTGGCTCCAACTGCTGCTGGGGAGGAGGGTGGCTAGACCTCTTGAGGGACTTCTGCTGCAGAGAGT

GATACTCCTTTACCTCAGGACCCAGTGATCATCCCCTTTTCCAAGCAAAGTCTGCTCGAGGAGTCTGGTA

ACTATGGATTTCCCCTCTTACAACTTTCAAACCAGAGTTGGAGACTCAGCATTGGGGTTCGGCCCTGCCC

GTAGCACAGCCAAGCCCTACCTCTCGGTTATCTTTTCTCCCGTCACCACCCAGTAAGGTCATGTGCTTCC

ACCCCTGGTCGGATGTAACGCTGCCACTCATGTCGGTCCCTGAGATCCGGGCTGTTGTTGATGCATGGGC

CTCAGTCACAGAGGAGCTGGGTGCCCAGTACCCTTGGGTGCAGGTTTGTGAGGTCGCCCCTTCCCCTGGA

TGGGCAGGGAGGGGGTGATGAAGCTTTGGTTCTGGGGAGTAACATTTCTGTTTCCACAGGGTGTGGTCAG

GAGGGAGTTGACTTGGTGTCTTTTGGCTAACAGAGCTCCGTATCCCTATCTGATAGATCTTTGAAAACAA

AGGTGCCATGATGGGCTGTTCTAACCCCCACCCCCACTGCCAGGTAAGGGTGTCAGGGGCTCCAGTGGGT

TTCTTGGCTGAGTCTGAGCCAGCACTGTGGACATGGGAACAGGATTAATGGATGGGACAGAGGAAATATG

CCAATGATGTGGAGGCTTGGAGGTAAAGGACCTGCCTGTTCTTCTCTGCTTTTGCCCCTTGACAGGTATG

GGCCAGCAGTTTCCTGCCAGATATTGCCCAGCGTGAGGAGCGATCTCAGCAGGCCTATAAGAGTCAGCAT

GGAGAGCCCCTGCTAATGGAGTACAGCCGCCAGGAGCTACTCAGGAAGGTGGGAGAGAGCCAAGCCCTGT

GTCCCCAAGGAGTCCCTAACTTTCTTATCCCATGAGAGAGGTGTGTAAAGGAGAAAGCTAGAGGTGAACT

AGTAGAGAGAGACTTGCTAGGAGGCCTTAGCAATAATCCAGTAATCTAAAGGAAAGATGATGGTGACTTA

GACTCGGGTGGTTAGTGGTAGAGGTGGTGAGAAGACATCAGATCCTGGGCACATTCTTTTCTTCTGCTTC

CCTTGCCTATTTGCTGACCACACTCCGGCTCCTATGTCACCTTGATGACTTCCTATCCATTCTGTCTTCC

TAGGAACGTCTGGTCCTAACCAGTGAGCACTGGTTAGTACTGGTCCCCTTCTGGGCAACATGGCCCTACC

AGACACTGCTGCTGCCCCGTCGGCATGTGCGGCGGCTACCTGAGCTGACCCCTGCTGAGCGTGATGGTCA

GTCTCCCAAGTAGGATCCTGGGGCTAGGCACTGGATGGAGGTTGCTCCCAGTAGGGTCAGCATCTGGACC

CCAGGCTGAGAGTCAGGCTCTGATTCCAGATCTAGCCTCCATCATGAAGAAGCTCTTGACCAAGTATGAC

AACCTCTTTGAGACGTCCTTTCCCTACTCCATGGGCTGGCATGGTGAGGCTTTTCAAGTACCTATATTTA

GCCCCAACACCATTTCTGGGCTCCTGGGGCTCAGCCTAGTGAACTGCAACCTCAAAGGAGCAAGCCTTGA

AACAGTTGCTGGGGGAAGTGGCCAGAGTAGAGATGCTGGGACTGAGGGTGGAGCAGCAAACTTGGTGAAA

CTACATCTCCAATGTGCTTTCTAATCTCCTGCCAGCTCTTCTCAAGCAGGGGATCCTGGGAGATGTAGTT

TTCAGATACCTGGTTGGGTTTGGGAGTAGGTGCTAACCTGGATAACTGTAAAAGGGCTCTCTCTCCCCAC

TGTCTCTCTTCTTTCTGTCAGGGGCTCCCACAGGATCAGAGGCTGGGGCCAACTGGAACCATTGGCAGCT

GCACGCTCATTACTACCCTCCGCTCCTGCGCTCTGCCACTGTCCGGAAATTCATGGTTGGCTACGAAATG

CTTGCTCAGGCTCAGAGGGACCTCACCCCTGAGCAGGTCAGGACTCAGAACAGTCTGGCGTCTCCAGACT

CTCACATGCAGTATGTGCAGGCACCTGATACTTCTGTTGCCCTTGTGCTCCAGTCATTGCACAAGGCAGA

AAACAGCTCTGGCAGGAAGGGACTGCCAAAGTTAGGAGCCCTAGGGCCTGGAAGGAGAGTATGGTCCTCA

GATCCCCCTTCTCTCCTGCTTCCTCCAGGGAACCCAACAGTCATGACCCTGATAGTTTCCCATAACAACC

TGGGCATTCCTTGGGACTCAGGAGCTGCTAAACTCTTTCATCCCCTGGTGGCTTCAGCAGTCCTTATCAC

CAGCCTCACAATCCCACAGGCCCACCCCCAGTGGGCCTGTGGCATTCATATTTCATATTCATATTTCAAA

CCACAATATCCAGCAAAATGTCTCCTGAGCACCCAGAACTCCATACCATCGGCCGGGTGTGGTGGCTCAT

GCCTTAATCCCAGCACTTTGGGAGGTCAAGATGGGAGGATTGCTTGAGCCCAGAAGTTCGAGACTAGCCT

GGGAAACATAGGAAGCCCTCGTCTCTACAAAAAAAATTTAAAAAGTTAGCCAGGTATGGTGGCATATACT

TGTGGTCCCAGATACTTGGGAGGCTGAGATAGGATCACTTGGGCCCAGGAGTTTGAGGCTGCAGTGAGCC

ATCATCATGGCATCATTGCATTCCAGCCTGGGCAACAGAGCAAGACCTCGTCTCAAAAAAAAAAAAAAAA

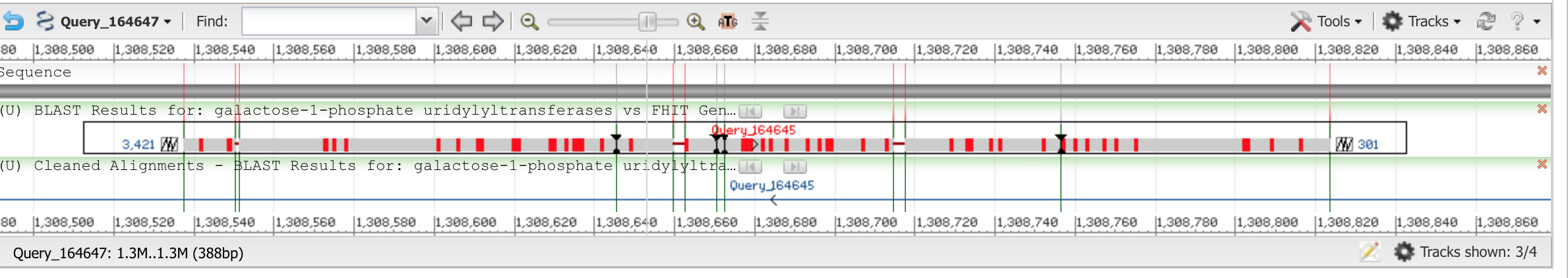
TGAAGTCCATGCCACCATTCTTGGCAGCCCAGCCCTTATCCTCCTTAATTGCTCCCTGTCCCTTTTCCAG

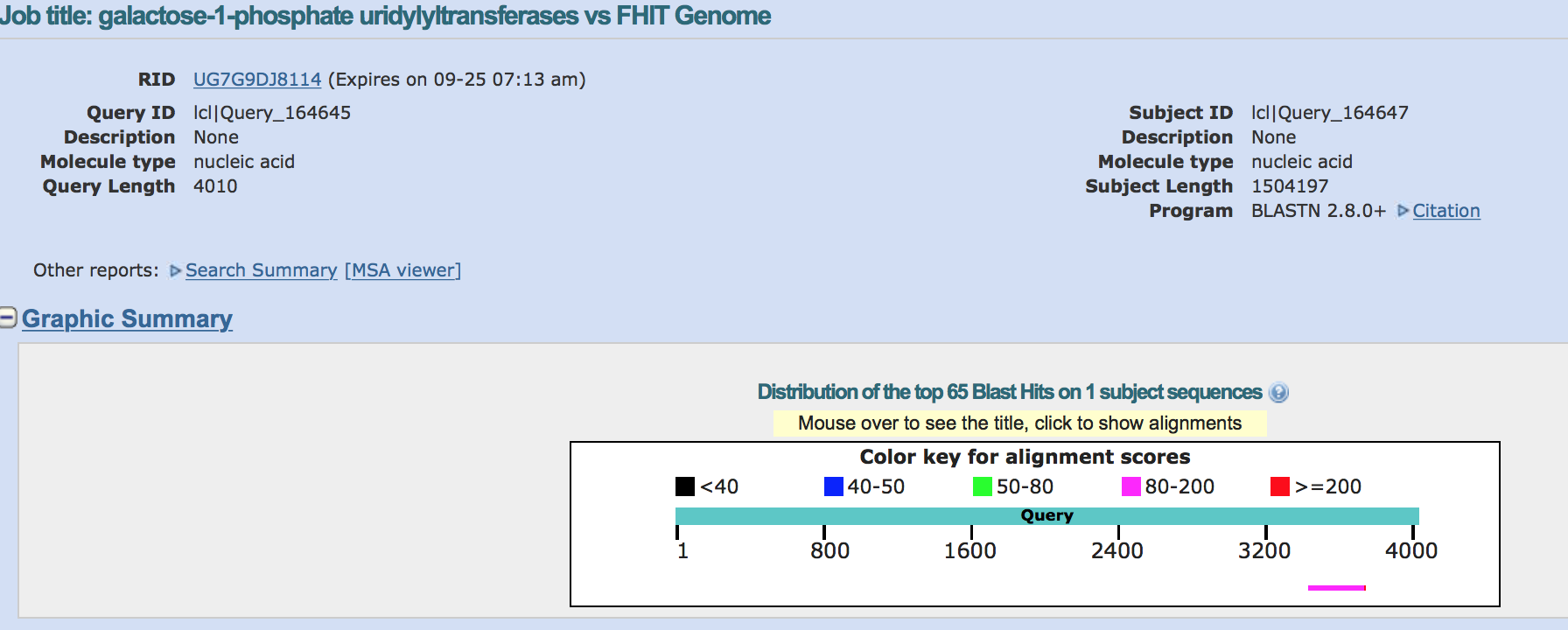
GCTGCAGAGAGACTAAGGGCACTTCCTGAGGTTCATTACCACCTGGGGCAGAAGGACAGGGAGACAGCAA

CCATCGCCTGACCACGCCGACCACAGGGCCTTGAATCCTTTTTTGTTTTCAACAGTCTTGCTGAATTAAG

CAGAAAGGGCCTTGAATCCTGGCCTGGAATTTGGGCAGATATAGCATTAATAAAACTGTGCATCTCAAAC

TTTTATCACATACTCTAATA



?????Not sure how to interpret NCBI viewer, compared to BLAST’s results. 

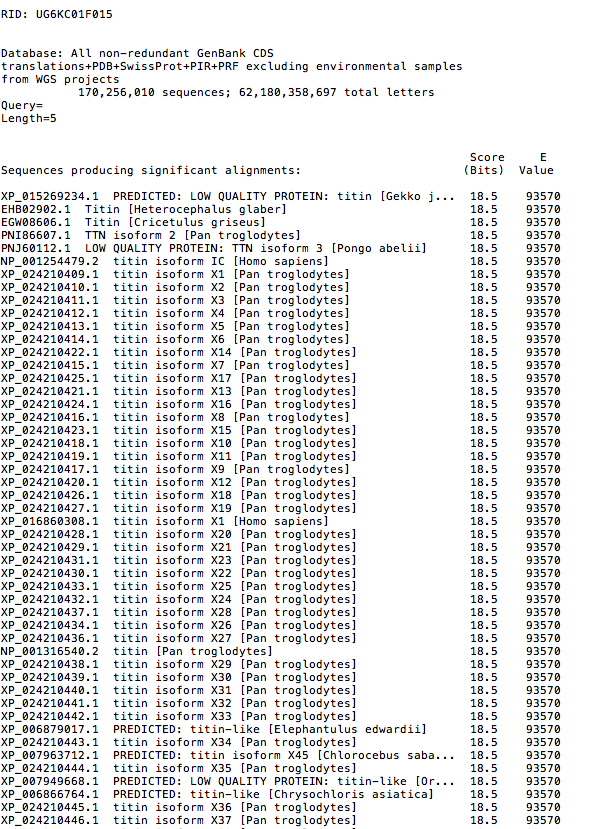
5) Find the unannotated genomic scaffold for Drosophila melanogaster, AE003584, using Entrez nucleotides. Display protein links to see the predicted proteins for this scaffold. Use the CDD search to identify conserved domains present in the tenth predicted protein (gi: 7295996) and suggest a potential function for this hypothetical protein.

6) As GenBank grows so does the number of chance occurrences of amino acid motifs that spell out words or people's names in single letter amino acid codes. One such name motif is ELVIS. Find the number of occurrences of ELVIS in the protein 'nr' database. To get any hits at all, you will have to adjust several of the advanced BLAST parameters including the Expect value, Word size and Score Matrix. Adjust some of these in the "Other advanced options" box. Options are entered command line style. For example, typing:

-e 10000

-A good rule of thumb is that the query length must be at least x2 the Word size

-Sometimes a short query sometimes does not produce results because it contains low-complexity sequence. If your query has regions of low-complexity sequence, then large portions of your query may be filtered out, essentially making your query shorter than you might have expected. Removing the filter will help in these cases.



sets the Expect value cut-off to 10000. See the BLAST "Frequently Asked Questions" linked on the left side bar of the BLAST page on "How do I perform a similarity search with a short peptide/nucleotide sequence?" for more information. We now have a page with presets optimized to find short nearly exact matches. You can cheat and run the search on this page to see the correct parameters to use.