Annotation Worksheet: NCBI and BLAST Due: _____

These are questions related NCBI and BLAST as well as a couple other things. The tutorial video I made should be of help. Once it has been completed, I will upload a video lecture explaining all of these topics to my YouTube channel and will provide you with a link. Feel free to type in your answers or print it out and fill it in by hand. Even if you don't have all of this memorized, this can be a good reference sheet to remind you of details should you forget later on down the line. Think of it like a study guide you would fill out for an exam in a class. A few sentences for each answer should be sufficient but this will vary. Use your own discretion but air on the side of being detailed. Feel free to include figures you find online or drawings you make if it is helpful in conveying an idea

1. Name 2 websites that are good sources for background information about a gene. What are some important things to take note of from these websites?

2. Explain this available evidence table, particularly the gene model and evidence supporting annotation columns.

Gene	Annotator	3.0 Identifier	Is 3.0 Gene model: Complete, needs more work, or you need help. Provide comments if you are not sure.	Gene Model			Evidence Supporting Annotation			
				Complete	Partial	de novo	MCOT	Iso-Seq	RNA-Seq	Ortholog
protein phosphatase PP2A 55 kDa regulatory subunit	Yasmin	Dcitr09g05720.1.1	Model complete in IRSC Training ID: Dcitr09g05720.1.1 Location: DC3.0sc09:14402937..14416984 (14.05 Kb)	X			X	X	X	X

3. Explain this copy number table. Why are copy numbers important to look at?

Gene name	Drosophila citri	Acyrthosiphon pisum	Bemisia tabaci	Cimex lectularius	Halyomorpha halys	Drosophila melanogaster	Aedes aegypti
tim	1	1*	1	1	1	1	1
per	1	1*	1	1	1	1	1
vg	1	0	1	1	1	1	1
Hr3	1	0	1	1	2	1	1
cyc	1	1*	1	1	1	1	1
clk	1	1*	0	1	0	1	0
Lirp	1	0	1		1	0	1
Yp1	1	1	1	1	2	0	2
Yp2	1	0	0	0	0	0	0
EcR	1	0	1	1	1	1	1
vri	1	1*	0	0	0	1	0
Pdp1	1	1*	1	1	0	1	0
dco	1	1*	1	1	0	0	0
cwo	1	0	0	0	1	1	1
CkIIα	1	1*	1	0	2	1	1
CkIIβ	1	1*	1	1	1	1	1
timeout	1	0	1	1	1	1	1

4. Compare and contrast ortholog, paralog and homolog.

5. What is NCBI? What are some important tools this website provides?

6. Explain some of the databases in NCBI. What is the INSDC and what sort of databases does it contain? What is RefSeq and where do its sequences come from?

7. What is BLAST? Why do we BLAST? What does this tell us?

8. What are the valid inputs for blast?

9. What is a FASTA sequence. How is it retrieved?

10. What is an accession number?

11. How do RefSeq accession numbers differ from other accession numbers? What is the difference between a RefSeq accession number Starting with and N and one starting with an X?

12. Explain the differences between these 3 accession numbers.

ATJ04204.1

XP_039488069.1

NP_001260113.1

13. When you blast, under what circumstance would you blast to insecta vs Hemiptera?

14. What does it mean if a gene is highly conserved? What would a graphic summary of a highly conserved gene look like in blast? What type of gene would likely be highly conserved?

15. What is Score? (either max or total. Doesn't matter) Percent Identity, query coverage and e-value?

16. Situation: We are working on annotating the GAPDH gene in apollo. We have 2 versions of the gene in apollo that we are considering. We expect to only find 1 copy of the gene in the genome. We blast the first version of the model. It looks like this:

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
RecName: Full=Glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic; Short=GPD-C; Short=GPDH-C...	Drosophila mela...	743	743	100%	0.0	100.00%	363	P13706.3
glycerol-3-phosphate dehydrogenase [Drosophila melanogaster]	Drosophila mela...	741	741	99%	0.0	100.00%	362	CAA43536.1
Glycerol-3-phosphate dehydrogenase 1, isoform F [Drosophila melanogaster]	Drosophila mela...	737	737	99%	0.0	100.00%	360	NP_001260112.1
glycerol-3-phosphate dehydrogenase (NAD+) [Drosophila melanogaster]	Drosophila mela...	736	736	100%	0.0	99.45%	363	CAA56497.1
glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X1 [Drosophila ananassae]	Drosophila anan...	735	735	99%	0.0	99.72%	360	XP_001962710.1
GM14480p [Drosophila melanogaster]	Drosophila mela...	734	734	99%	0.0	99.72%	360	AAL13721.1
PREDICTED: glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X1 [Drosophila takahashii]	Drosophila taka...	734	734	99%	0.0	99.44%	360	XP_017006339.1
PREDICTED: glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X1 [Drosophila biarmipes]	Drosophila biar...	734	734	99%	0.0	99.44%	360	XP_016946221.1
glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X1 [Drosophila erecta]	Drosophila erecta	733	733	99%	0.0	99.44%	360	XP_001968861.1
glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X1 [Drosophila suzukii]	Drosophila suzukii	733	733	99%	0.0	99.17%	360	XP_016938655.1
glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X1 [Drosophila kikkawai]	Drosophila kikka...	732	732	99%	0.0	99.17%	360	XP_017022143.1
PREDICTED: glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X1 [Drosophila rhopaloa]	Drosophila rhop...	731	731	99%	0.0	99.17%	360	XP_016975728.1
glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X1 [Drosophila subpulchrella]	Drosophila subp...	731	731	99%	0.0	98.89%	360	XP_037731121.1
glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X1 [Drosophila persimilis]	Drosophila persi...	731	731	99%	0.0	98.89%	360	XP_002014256.1
glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X1 [Drosophila obscura]	Drosophila obsc...	731	731	99%	0.0	99.17%	360	XP_022228818.1
glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X1 [Drosophila guanche]	Drosophila guan...	730	730	99%	0.0	98.89%	360	XP_034135187.1
PREDICTED: glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X1 [Drosophila ficusphila]	Drosophila ficus...	730	730	99%	0.0	98.89%	360	XP_017041280.1
glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X1 [Drosophila subobscura]	Drosophila subo...	728	728	99%	0.0	98.61%	360	XP_034670096.1
glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X1 [Drosophila willistoni]	Drosophila willist...	728	728	99%	0.0	98.89%	360	XP_002065366.1
glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic [Drosophila serrata]	Drosophila serrata	728	728	99%	0.0	98.61%	360	XP_020806966.1
glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X1 [Drosophila innubila]	Drosophila innu...	728	728	99%	0.0	98.33%	360	XP_034474831.1
glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X1 [Drosophila hydei]	Drosophila hydei	726	726	99%	0.0	98.33%	360	XP_023174999.1
glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X1 [Drosophila grimshawi]	Drosophila grim...	726	726	99%	0.0	97.78%	360	XP_001988153.1
glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X1 [Scaptodrosophila lebanonensis]	Scaptodrosophil...	726	726	99%	0.0	98.33%	360	XP_030377464.1
glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X1 [Drosophila mojavensis]	Drosophila moja...	725	725	99%	0.0	98.06%	360	XP_002002704.1
G-3-P dehydrogenase [Drosophila ezoana]	Drosophila ezoana	724	724	99%	0.0	98.06%	360	BAA20578.1
Glycerol-3-phosphate dehydrogenase [Drosophila montana]	Drosophila mont...	723	723	99%	0.0	98.06%	360	BAA34404.1
glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X1 [Drosophila virilis]	Drosophila virilis	723	723	99%	0.0	98.06%	360	XP_002051368.1
RecName: Full=Glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic; Short=GPD-C; Short=GPDH-C...	Drosophila kane...	722	722	99%	0.0	98.06%	360	O97463.3
G-3-P dehydrogenase [Drosophila novamexicana]	Drosophila nova...	722	722	99%	0.0	97.78%	360	BAA57829.1
G-3-P dehydrogenase [Drosophila americana americana]	Drosophila amer...	721	721	99%	0.0	97.78%	360	BAA20574.1
glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X1 [Drosophila albomicans]	Drosophila albo...	720	720	99%	0.0	97.78%	361	XP_034099380.1
Glycerol-3-phosphate dehydrogenase 1, isoform E [Drosophila melanogaster]	Drosophila mela...	716	716	96%	0.0	100.00%	350	NP_001260111.1
Glycerol-3-phosphate dehydrogenase 1, isoform G [Drosophila melanogaster]	Drosophila mela...	716	716	96%	0.0	100.00%	353	NP_001260113.1
glycerol-3-phosphate dehydrogenase [Drosophila melanogaster]	Drosophila mela...	715	715	96%	0.0	99.71%	400	CAA47892.1

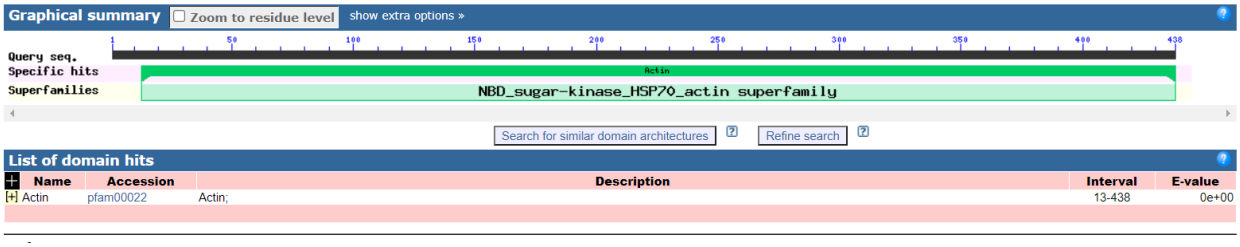
We blast the second version of the model and we get this:

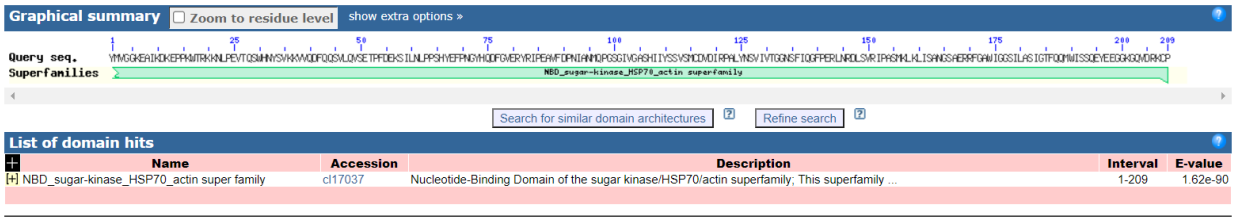
Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
sn-glycerol-3-phosphate dehydrogenase [Drosophila melanogaster]	Drosophila mela...	284	284	100%	1e-95	100.00%	322	AAA28592.1
sn-glycerol-3-phosphate dehydrogenase [Drosophila melanogaster]	Drosophila mela...	284	284	100%	1e-95	100.00%	329	AAA28591.1
glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X2 [Drosophila erecta]	Drosophila erecta	283	283	100%	6e-95	100.00%	353	XP_015015356.1
glyceraldehyde-3-phosphate dehydrogenase [Drosophila bipectinata]	Drosophila bipec...	282	282	100%	6e-95	99.29%	319	ATJ04204.1
LOW QUALITY PROTEIN: glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic [Drosophila santomea]	Drosophila sant...	283	283	100%	7e-95	100.00%	353	XP_039488069.1
Glycerol-3-phosphate dehydrogenase 1, isoform G [Drosophila melanogaster]	Drosophila mela...	283	283	100%	7e-95	100.00%	353	NP_001260113.1
glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X1 [Drosophila erecta]	Drosophila erecta	283	283	100%	7e-95	100.00%	360	XP_001968861.1
GM14480p [Drosophila melanogaster]	Drosophila mela...	283	283	100%	7e-95	100.00%	360	AAL13721.1
Glycerol-3-phosphate dehydrogenase 1, isoform E [Drosophila melanogaster]	Drosophila mela...	283	283	100%	8e-95	100.00%	350	NP_001260111.1
Glycerol-3-phosphate dehydrogenase 1, isoform F [Drosophila melanogaster]	Drosophila mela...	283	283	100%	9e-95	100.00%	360	NP_001260112.1
glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X2 [Drosophila guanche]	Drosophila guan...	283	283	100%	1e-94	99.29%	353	XP_034135189.1
glycerol-3-phosphate dehydrogenase (NAD+) [Drosophila melanogaster]	Drosophila mela...	283	283	100%	1e-94	100.00%	363	CAA56497.1
RecName: Full=Glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic; Short=GPD-C; Short=GPDH-C...	Drosophila mela...	283	283	100%	1e-94	100.00%	363	P13706.3
glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X2 [Drosophila obscura]	Drosophila obsc...	283	283	100%	1e-94	99.29%	353	XP_022228819.1
glycerol-3-phosphate dehydrogenase [Drosophila melanogaster]	Drosophila mela...	284	284	100%	1e-94	100.00%	400	CAA47892.1
glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X1 [Drosophila guanche]	Drosophila guan...	283	283	100%	1e-94	99.29%	360	XP_034135187.1
glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X1 [Drosophila obscura]	Drosophila obsc...	282	282	100%	2e-94	99.29%	360	XP_022228818.1
glycerol 3 phosphate dehydrogenase [Drosophila miranda]	Drosophila mira...	274	274	97%	2e-94	98.54%	137	AAX13132.1
GPDH [Drosophila pseudoobscura]	Drosophila pseu...	281	281	100%	2e-94	98.57%	350	AAB02947.1
PREDICTED: glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X2 [Drosophila ficusphila]	Drosophila ficus...	282	282	100%	3e-94	99.29%	353	XP_017041281.1
glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X2 [Drosophila miranda]	Drosophila mira...	282	282	100%	3e-94	98.57%	353	XP_017154656.1
PREDICTED: glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X1 [Drosophila ficusphila]	Drosophila ficus...	282	282	100%	3e-94	99.29%	360	XP_017041280.1
glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X2 [Drosophila ananassae]	Drosophila anan...	281	281	100%	3e-94	99.29%	353	XP_014762112.1
glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X3 [Scaptodrosophila lebanonensis]	Scaptodrosophil...	281	281	100%	4e-94	99.29%	353	XP_030377466.1
glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X1 [Drosophila persimilis]	Drosophila persi...	281	281	100%	4e-94	98.57%	360	XP_002014256.1
glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X2 [Scaptodrosophila lebanonensis]	Scaptodrosophil...	281	281	100%	4e-94	99.29%	353	XP_030377465.1
glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X1 [Drosophila ananassae]	Drosophila anan...	281	281	100%	4e-94	99.29%	360	XP_001962710.1
glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X2 [Drosophila subobscura]	Drosophila subo...	281	281	100%	5e-94	98.57%	353	XP_034670098.1
glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X1 [Scaptodrosophila lebanonensis]	Scaptodrosophil...	281	281	100%	5e-94	99.29%	360	XP_030377464.1
glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X2 [Drosophila albomicans]	Drosophila albo...	281	281	100%	5e-94	98.57%	353	XP_034099381.1
glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic [Drosophila busckii]	Drosophila busckii	280	280	100%	6e-94	97.86%	329	XP_033150726.1
glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X1 [Drosophila subobscura]	Drosophila subo...	281	281	100%	7e-94	98.57%	360	XP_034670096.1
glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X2 [Drosophila kikkawai]	Drosophila kikka...	280	280	100%	8e-94	98.57%	353	XP_017022151.1
glycerol-3-phosphate dehydrogenase [Drosophila melanogaster]	Drosophila mela...	281	281	99%	8e-94	100.00%	362	CAA43536.1
glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic isoform X1 [Drosophila kikkawai]	Drosophila kikka...	281	281	100%	9e-94	98.57%	360	XP_017022143.1

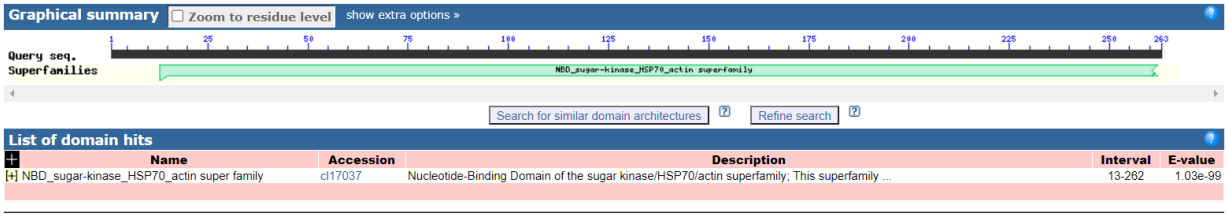
Explain in the context similarity to orthologs why we would choose one model over the other.

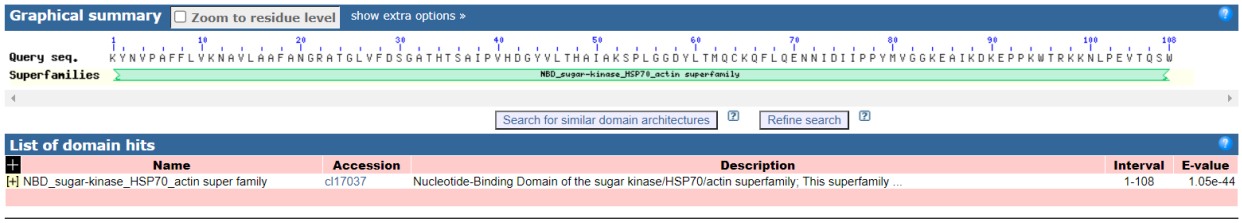
17. What is a domain?

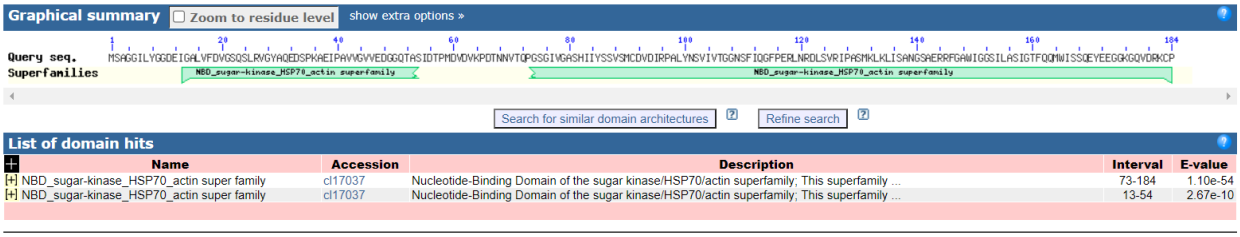
18. Say if the following domains are complete or incomplete. Explain why you think so and where you think there is some of the domain missing.





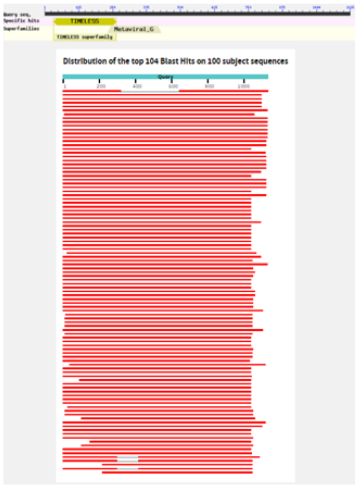




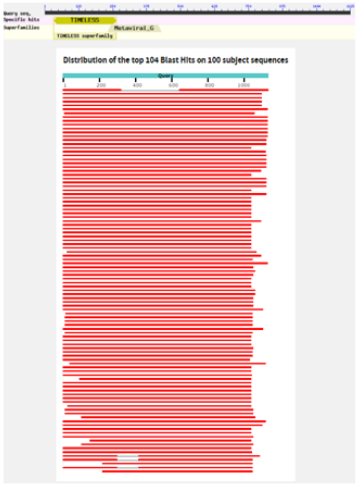


19. If we examine the following 3 orthologs and the domains look like this:

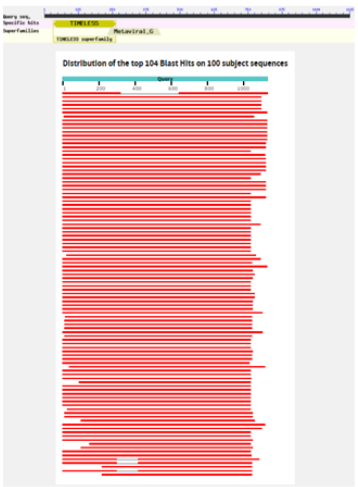
Organism 1



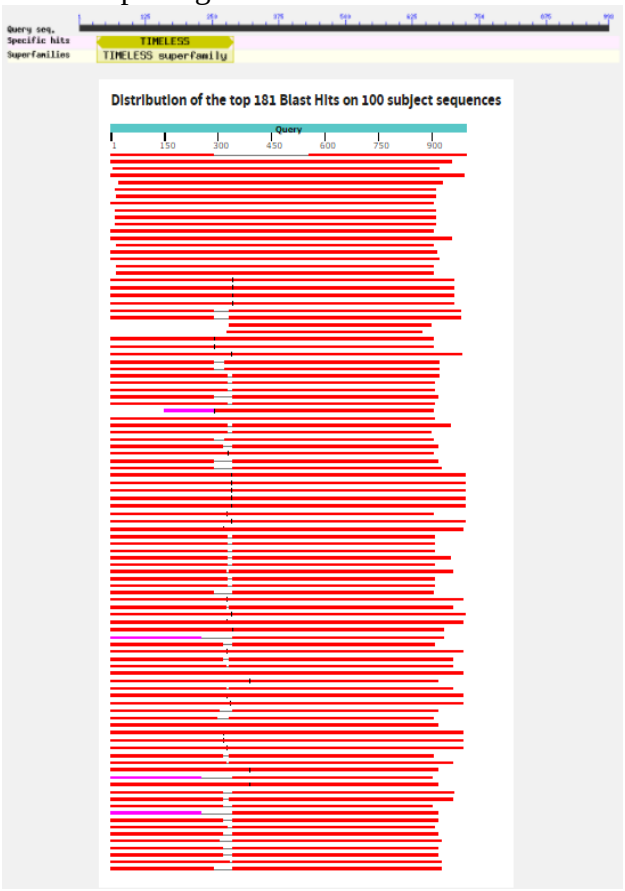
Organism 2



Organism 3



But our apollo gene model domain looks like this:



What might be occurring?

20. What is a peptide pairwise blast? Why would we perform one? Give a hypothetical example of a situation where performing a peptide pairwise blast would be beneficial.

21. What is a cDNA pairwise blast? Why might we perform a cDNA pairwise blast over a peptide pairwise blast? Give a specific example.