Find-A-Gene – BGGN 213 – Spring 2019 A Geffre – A53267765

Please find below my submission for Q1-Q4, for the Find-A-Gene Project:

Questions:

[Q1]

Tell me the name of a protein you are interested in. Include the species and the accession number. This can be a human protein or a protein from any other species as long as it's function is known. If you do not have a favorite protein, select human RBP4 or KIF11. Do not use beta globin as this is in the worked example report that I provide you with online.

Protein of Interest: PGRP-S2; peptidoglycan recognition protein S2 (Apis mellifera)

Accession Number: Gene ID: 412484

Gene: NM_001163716.1 Protein: NP 001157188.1

Related to anti-viral immune response in honey bees (Nazzi F, Brown SP, Annoscia D, et al. Synergistic parasite-pathogen interactions mediated by host immunity can drive the collapse of honeybee colonies. *PLoS Pathog.* 2012;8(6):e1002735. doi:10.1371/journal.ppat.1002735).

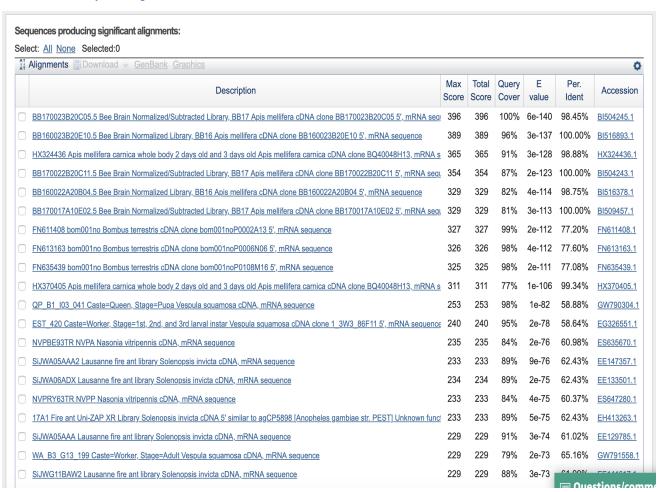
[Q2]

Perform a BLAST search against a DNA database, such as a database consisting of genomic DNA or ESTs. The BLAST server can be at NCBI or elsewhere. Include details of the BLAST method used, database searched and any limits applied (e.g. Organism).

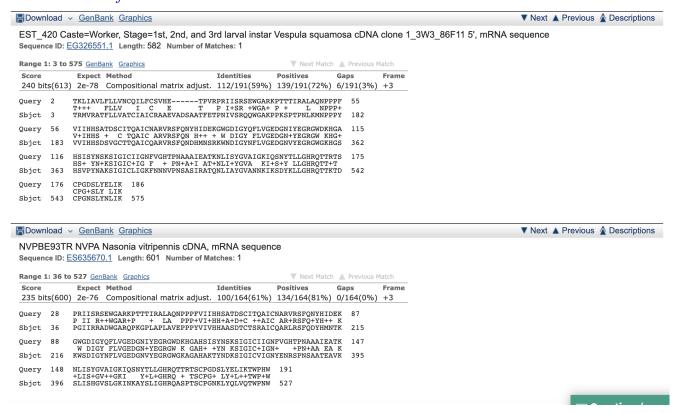
tBLASTn search ID: DA9TFDNZ014, searching *est* (Database of GenBank+EMBL+DDBJ sequences from EST Division) database.

Also include the output of that BLAST search in your document. If appropriate, change the font to Courier size 10 so that the results are displayed neatly. You can also screen capture a BLAST output (e.g. alt print screen on a PC or on a MAC press — -shift-4. The pointer becomes a bulls eye. Select the area you wish to capture and release. The image is saved as a file called Screen Shot [].png in your Desktop directory). It is not necessary to print out all of the blast results if there are many pages.

An overview of the top hits:



Some select hits of interest:



On the BLAST results, clearly indicate a match that represents a protein sequence, encoded from some DNA sequence, that is homologous to your query protein. I need to be able to inspect the pairwise alignment you have selected, including the E value and score. It should be labeled a "genomic clone" or "mRNA sequence", etc. - but include no functional annotation.

I am interested in <u>FN611408.1</u> (Bombus terrestris); <u>GW790304.1</u>, (Vespula squamosa); <u>EE147357.1</u> (Solenopsis invicta), <u>ES647280.1</u> (Nasonia vitripens) etc.

All of these sequences appear to be hitherto unannotated mRNAs, all of which belong to Hymenopterans. *Vespula squamosa* is likely the most under-annotated, as it is not a model organism or agriculturally-relevant species.

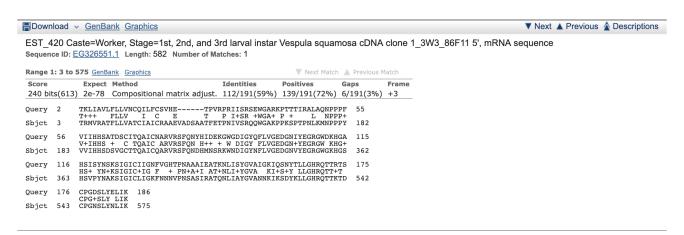
In general, [Q2] is the most difficult for students because it requires you to have a "feel" for how to interpret BLAST results. You need to distinguish between a perfect match to your query (i.e. a sequence that is not "novel"), a near match (something that might be "novel", depending on the results of [Q4]), and a non-homologous result. If you are having trouble finding a novel gene try restricting your search to an organism that is poorly annotated.

Gather information about this "novel" protein. At a minimum, show me the protein sequence of the "novel" protein as displayed in your BLAST results from [Q2] as FASTA format (you can copy and paste the aligned sequence subject lines from your BLAST result page if necessary) or translate your novel DNA sequence using a tool called EMBOSS Transeq at the EBI. Don't forget to translate all six reading frames; the ORF (open reading frame) is likely to be the longest sequence without a stop codon. It may not start with a methionine if you don't have the complete coding region. Make sure the sequence you provide includes a header/subject line and is in traditional FASTA format. Here, tell me the name of the novel protein, and the species from which it derives. It is very unlikely (but still definitely possible) that you will find a novel gene from an organism such as S. cerevisiae, human or mouse, because those genomes have already been thoroughly annotated. It is more likely that you will discover a new gene in a genome that is currently being sequenced, such as bacteria or plants or protozoa.

Below follows the FASTA sequence for the cDNA clone from V. squamosa

>GW790304.1 QP_B1_I03_041 Caste=Queen, Stage=Pupa Vespula squamosa cDNA, mRNA sequence

BLAST gives this translated protein sequence:



I put this sequence through EMBOSS Transeq (EBI) for the following translations:

>GW790304.1_1 QP_B1_I03_041 Caste=Queen, Stage=Pupa Vespula squamosa cDNA, mRNA sequence REWYEORFYLLORALOFAVORKLOIRLOPLKHLISFODNNGELNRRKALRLILK*IHLLT

LLFIIRIQLVVPLKQFVRLELEVFRMII*TRENGMTSATTFWSVKMVMFTKVVAGVNMAP
IQSRTMPRVSVFALLVNLTITYRIQRVFEQHKI**LTE*LTTKSNPIISFLAIDKPLKLI
VLEILYTI*LKHGLTGPTRHKKYNRYAIKLIYQV*RLIVLDSTIIGN*RSSDR*LLLTFL
*LTMDTFKFDFRANIFT*CFVKEE

>GW790304.1_2 QP_B1_I03_041 Caste=Queen, Stage=Pupa Vespula squamosa cDNA, mRNA sequence

```
ENGTSNVSTCCNVHCNLPCSGSCRFGCNL*NT*YRFKTTMGS*TAEKPYA*S*NESTSLR
CYSSFGFSWLYHSSNLSG*S*KFSE*SYELEKME*HRLQLFGR*RW*CLRRSWLG*TWLP
FSPVQCQEYRYLPYW*I*Q*RTEFSEYSSNTKFDSLRSS*QQNQIRL*ASWPSTNH*N*L
SWKFSIQFD*NMASLDRHAIRNIIVMRLN*STKYNVLLFWTRRSLEINDHRIDNFC*LFF
N*RWIPSNSIFVRTFLLDVS*RKX
>GW790304.1_3 QP_B1_I03_041 Caste=Queen, Stage=Pupa Vespula
```

>GW790304.1_3 QP_B1_I03_041 Caste=Queen, Stage=Pupa Vespula squamosa cDNA, mRNA sequence

RMVRATFLLVATCIAICRAAEVADSAATFETPNIVSRQQWGAKPPKSPTPNLKMNPPPYV VIHHSDSVGCTTQAICQARVRSFQNDHMNSRKWNDIGYNFLVGEDGNVYEGRGWGKHGSH SVPYNAKSIGICLIGKFNNNVPNSASIRATQNLIAYGVANNKIKSDYKLLGHRQTTKTDC PGNSLYNLIKTWPHWTDTP*EI*SLCD*INLPSITSYCFGLDDHWKLTIIGSITFVDFSL INDGYLQIRFSCEHFYLMFRKGRX

(This appears to be the best option to go with; longest ORF.) >GW790304.1_4 QP_B1_I03_041 Caste=Queen, Stage=Pupa Vespula squamosa cDNA, mRNA sequence

LFLYETSSKNVRTKIEFEGIHR*LKKSQQKLSIR*SLISNDRRVQNNKTLYLVD*FNRIT
IIFLMACRSSEAMF*SNCIENFQDNQF*WFVDGQEAYNRI*FCC*LLRKLSNFVLLEYSL
NSVRYC*IYQ*GKYRYSWHCTGLNGSHVYPSHDLRKHYHLHRPKSCSRCHSIFSSSYDHS
ENF*L*PDKLLEWYNQLNPNDE*QRKEVDSF*D*A*GFSAV*LPIVVLKRY*VFQRLQPN
LQLPLHGKLQCTLQQVETLLVPFS

>GW790304.1_5 QP_B1_I03_041 Caste=Queen, Stage=Pupa Vespula squamosa cDNA, mRNA sequence

LPLRNIK*KCSHENRI*RYPSLIKEKSTKVIDPMIVNFQ*SSSPKQ*DVILGRLI*SHND
YISYGVSVQ*GHVLIKLYREFPGQSVLVVCRWPRSL*SDLILLLATP*AIKFCVARILAE
FGTLLLNLPIRQIPILLALYGTEWEPCLPQPRPS*TLPSSPTKKL*PMSFHFLEFI*SF*
KLLTLA*QIA*VVQPTESE**ITT*GGGFILRLGVGLFGGLAPHCCLETILGVSKVAAES
ATSAARQIAMHVATSRNVARTILX

>GW790304.1_6 QP_B1_I03_041 Caste=Queen, Stage=Pupa Vespula squamosa cDNA, mRNA sequence

SSTKHQVKMFARKSNLKVSIVN*RKVNKSYRSDDR*FPMIVESKTIRRYTW*INLIA*R LYFLWRVGPVRPCFNQIV*RISRTISFSGLSMAKKLIIGFDFVVSYSVSYQILCCSNTR*IRYVIVKFTNKANTDTLGIVRD*MGAMFTPATTFVNITIFTDQKVVADVIPFSRVHMIIL KTSNSSLTNCLSGTTN*IRMMNNNVRRWIHFKIRRRAFRRFSSPLLS*NDIRCFKGCSRICNFRCTANCNARCNK*KRCSYHSR

[Q4]

Prove that this gene, and its corresponding protein, are novel. For the purposes of this project, "novel" is defined as follows. Take the protein sequence (your answer to [Q3]), and use it as a query in a blastp search of the nr database at NCBI.

- If there is a match with 100% amino acid identity to a protein in the database, from the same species, then your protein is NOT novel (even if the match is to a protein with a name such as "unknown"). Someone has already found and annotated this sequence, and assigned it an accession number.
- If the top match reported has less than 100% identity, then it is likely that your protein is novel, and you have succeeded.
- If there is a match with 100% identity, but to a different species than the one you started with, then you have likely succeeded in finding a novel gene.
- If there are no database matches to the original query from [Q1], this indicates that you have partially succeeded: yes, you may have found a new gene, but noit is not actually homologous to the original query. You should probably start over.

The BLASTp results from the selected sequence above follow below:

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
PREDICTED: peptidoglycan-recognition protein SC2-like [Polistes dominula]	293	293	68%	3e-97	74.03%	XP_015177188.1
PREDICTED: peptidoglycan recognition protein 3-like [Polistes canadensis]	276	451	64%	9e-88	74.10%	XP_014602399.1
peptidoglycan-recognition protein SC2 isoform X1 [Ooceraea biroi]	264	264	73%	5e-86	62.44%	XP_011344962.1
PREDICTED: peptidoglycan-recognition protein SC2-like [Trachymyrmex cornetzi]	264	264	67%	5e-86	65.54%	XP_018369940.1
PREDICTED: peptidoglycan-recognition protein SC2-like [Atta cephalotes]	262	262	66%	2e-85	65.34%	XP_012059667.1
Peptidoglycan-recognition protein SC2 [Acromyrmex echinatior]	263	263	64%	2e-85	66.67%	EGI64536.1
PREDICTED: peptidoglycan-recognition protein SC2-like [Atta colombica]	260	260	66%	9e-85	64.77%	XP_018046500.1
peptidoglycan-recognition protein SB1-like [Pogonomyrmex barbatus]	260	260	73%	2e-84	59.28%	XP_011631523.1
peptidoglycan-recognition protein SC2 isoform X2 [Ooceraea biroi]	258	258	63%	5e-84	68.86%	XP_011344963.1
peptidoglycan-recognition protein precursor [Myrmica ruginodis]	257	257	64%	2e-83	65.09%	ACT66860.1
peptidoglycan-recognition protein precursor [Myrmica lobicornis]	256	256	67%	3e-83	63.84%	ACT66861.1
Peptidoglycan recognition protein 3 [Trachymyrmex cornetzi]	262	412	64%	4e-83	67.25%	KYN14761.1
peptidoglycan-recognition protein precursor [Myrmica brevispinosa]	255	255	64%	6e-83	65.09%	ACT66868.1
peptidoglycan-recognition protein precursor [Myrmica alaskensis]	255	255	63%	7e-83	64.88%	ACT66869.1
peptidoglycan-recognition protein precursor [Myrmica sulcinodis]	254	254	65%	1e-82	64.16%	ACT66863.1
peptidoglycan-recognition protein precursor [Myrmica fracticornis]	254	254	64%	2e-82	64.50%	ACT66867.1
PREDICTED: peptidoglycan-recognition protein LF-like [Habropoda laboriosa]	263	447	63%	2e-82	70.91%	XP_017789839.1
peptidoglycan-recognition protein precursor [Myrmica sabuleti]	254	254	65%	2e-82	63.22%	ACT66865.1
peptidoglycan-recognition protein SC2 [Bombus terrestris]	254	254	73%	4e-82	61.14%	XP_012170795.1

The top hits are not exact matches, but appear to be related to the function from the original A.

mellifera gene, as well as other closely related species, suggesting there is homology among these sequences, and potentially similar function in this V . $squamosa$ sequence.							