

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:

Sequences producing significant alignments:							
Select: All None Selected:0							
 Alignments	 Download	GenBank	Graphics				
	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
<input type="checkbox"/>	BB170023B20C05.5 Bee Brain Normalized/Subtracted Library, BB17 Apis mellifera cDNA clone BB170023B20C05 5', mRNA seq	396	396	100%	6e-140	98.45%	BI504245.1
<input type="checkbox"/>	BB160023B20E10.5 Bee Brain Normalized Library, BB16 Apis mellifera cDNA clone BB160023B20E10 5', mRNA sequence	389	389	96%	3e-137	100.00%	BI516893.1
<input type="checkbox"/>	HX324436 Apis mellifera carnica whole body 2 days old and 3 days old Apis mellifera carnica cDNA clone BQ40048H13, mRNA s	365	365	91%	3e-128	98.88%	HX324436.1
<input type="checkbox"/>	BB170022B20C11.5 Bee Brain Normalized/Subtracted Library, BB17 Apis mellifera cDNA clone BB170022B20C11 5', mRNA seq	354	354	87%	2e-123	100.00%	BI504243.1
<input type="checkbox"/>	BB160022A20B04.5 Bee Brain Normalized Library, BB16 Apis mellifera cDNA clone BB160022A20B04 5', mRNA sequence	329	329	82%	4e-114	98.75%	BI516378.1
<input type="checkbox"/>	BB170017A10E02.5 Bee Brain Normalized/Subtracted Library, BB17 Apis mellifera cDNA clone BB170017A10E02 5', mRNA seq	329	329	81%	3e-113	100.00%	BI509457.1
<input type="checkbox"/>	FN611408 bom001no Bombus terrestris cDNA clone bom001noP0002A13 5', mRNA sequence	327	327	99%	2e-112	77.20%	FN611408.1
<input type="checkbox"/>	FN613163 bom001no Bombus terrestris cDNA clone bom001noP0006N06 5', mRNA sequence	326	326	98%	4e-112	77.60%	FN613163.1
<input type="checkbox"/>	FN635439 bom001no Bombus terrestris cDNA clone bom001noP0108M16 5', mRNA sequence	325	325	98%	2e-111	77.08%	FN635439.1
<input type="checkbox"/>	HX370405 Apis mellifera carnica whole body 2 days old and 3 days old Apis mellifera carnica cDNA clone BQ40048H13, mRNA s	311	311	77%	1e-106	99.34%	HX370405.1
<input type="checkbox"/>	QP_B1_I03_041 Caste=Queen, Stage=Pupa Vesputa squamosa cDNA, mRNA sequence	253	253	98%	1e-82	58.88%	GW790304.1
<input type="checkbox"/>	EST_420 Caste=Worker, Stage=1st, 2nd, and 3rd larval instar Vesputa squamosa cDNA clone 1_3W3_86F11 5', mRNA sequence	240	240	95%	2e-78	58.64%	EG326551.1
<input type="checkbox"/>	NVPBE93TR NVPA Nasonia vitripennis cDNA, mRNA sequence	235	235	84%	2e-76	60.98%	ES635670.1
<input type="checkbox"/>	SiJWA05AAA2 Lausanne fire ant library Solenopsis invicta cDNA, mRNA sequence	233	233	89%	9e-76	62.43%	EE147357.1
<input type="checkbox"/>	SiJWA06ADX Lausanne fire ant library Solenopsis invicta cDNA, mRNA sequence	234	234	89%	2e-75	62.43%	EE133501.1
<input type="checkbox"/>	NVPRY63TR NVPP Nasonia vitripennis cDNA, mRNA sequence	233	233	84%	4e-75	60.37%	ES647280.1
<input type="checkbox"/>	17A1 Fire ant Uni-ZAP XR Library Solenopsis invicta cDNA 5' similar to agCP5898 [Anopheles gambiae str. PEST] Unknown func	233	233	89%	5e-75	62.43%	EH413263.1
<input type="checkbox"/>	SiJWA05AAA Lausanne fire ant library Solenopsis invicta cDNA, mRNA sequence	229	229	91%	3e-74	61.02%	EE129785.1
<input type="checkbox"/>	WA_B3_G13_199 Caste=Worker, Stage=Adult Vesputa squamosa cDNA, mRNA sequence	229	229	79%	2e-73	65.16%	GW791558.1
<input type="checkbox"/>	SiJWG11BAW2 Lausanne fire ant library Solenopsis invicta cDNA, mRNA sequence	229	229	88%	3e-73	64.00%	EE141617.1

Some select hits of interest:

Download

GenBank

Graphics

NextPreviousDescriptions

EST_420 Caste=Worker, Stage=1st, 2nd, and 3rd larval instar Vespula squamosa cDNA clone 1_3W3_86F11 5', mRNA sequence

Sequence ID: EG326551.1Length: 582Number of Matches: 1

Range 1: 3 to 575GenBankGraphics

Next MatchPrevious Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
240 bits(613)	2e-78	Compositional matrix adjust.	112/191(59%)	139/191(72%)	6/191(3%)	+3
Query 2	TKLIAVLFLVNCQILFCSVHE-----TPVRPRIISRSEWGARKPTTTIRALAQNPPPF	55				
Sbjct 3	T+++ FLLV I C E T P I+SR +WGA+ P + L NPPP+ TRMVRATFLVATCIAICRAAEVADSAATFETPNIVSRQQWAKFPKSPPTNLKMNPPFY	182				
Query 56	VIIHHSATDSCITQAICNARVRSFQNYHIDEKGWDIGYQFLVGEDGNIYEGRGWDKHGA	115				
Sbjct 183	V+IHHS + C TQAIC ARVRSFQ N H++ + W DIGY FLVGEDGN+YEGRGW KHG+ VIIHSDSVGCTTQAICARVRSFQNDHMSRKWNDIGYNFLVGEDGNVYEGRGWKGHGS	362				
Query 116	HSISYNSKSGICIGIIGNFVGHTPNAAAIEATKNLISYGVAGKIQSNYTLGHRQTTRTS	175				
Sbjct 363	HS+ YN+KSGICIG F + PN+A+I AT+NLI+YGVA KI+S+Y LLGHRQTT+T HSVFPYNAKSGICIGLIGFNNNVFNSASIRATQNLIAYGVANNKIKSDYKLLGHRQTTKTD	542				
Query 176	CPGDSLYELIK 186					575
Sbjct 543	CPG+SLY LIK CPGNSLYNLIK					

DownloadGenBankGraphics

NextPreviousDescriptions

NVPBE93TR NVPA Nasonia vitripennis cDNA, mRNA sequence

Sequence ID: ES635670.1Length: 601Number of Matches: 1

Range 1: 36 to 527GenBankGraphics

Next MatchPrevious Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
235 bits(600)	2e-76	Compositional matrix adjust.	100/164(61%)	134/164(81%)	0/164(0%)	+3
Query 28	PRIISRSEWGARKPTTTIRALAQNPPPFVIIHHSATDSCITQAICNARVRSFQNYHIDEK	87				
Sbjct 36	P II R++WGAR+P + LA PPP+VI+HH+A+D+C ++AIC AR+RSFQ+YH++ K PGIIRRADWGARKPKGPLAPLAVEPPYVIVHHAASDTCTSAICQARLSRQDYHMMTK	215				
Query 88	GWGDIGYQFLVGEDGNIYEGRGWDKHGAHSISYNSKSGICIGIIGNFVGHTPNAAAIEATK	147				
Sbjct 216	W DIGY FLVGEDGN+YEGRGW K GAH+ +YN KSGIC+IGN+ +PN+AA EA K KWSDIGYNFLVGEDGNVYEGRGWKGAGAHAKTYNDKSGICVIGNYENRSPNSAATEAVK	395				
Query 148	NLISYGVAGKIQSNYTLGHRQTTRTSCPGDSLYELIKTWPHW 191					527
Sbjct 396	+LIS+GV++GKI Y+L+GHRQ + TSCPG+ LY+L++TWP+W SLISHGVSLGKINKAYSLIGHRQASPTSCPGNKLYQLVQTWPNW					

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 [FN611408.1](#) (*Bombus terrestris*); [GW790304.1](#), (*Vespula squamosa*); [EE147357.1](#) (*Solenopsis invicta*), [ES647280.1](#) (*Nasonia vitripennis*) 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.

[Q3]

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
CGAGAATGGTACGAGCAACGTTTCTACTTGTGTCGAACGTGCATTGCAATTTGCCGTGCAGCGGAAGTTGCAGATTCCGGCT
GCAACCTTTGAAACACCTAATATCGTTTCAAGACAACAATGGGGAGCTAAACCGCCGAAAAGCCCTACGCCCTAATCTTAA
AATGAATCCACCTCCTTACGTTGTTATTTCATCATTCGGATTTCAGTTGGTTGTACCACTCAAGCAATTTGTCAGGCTAGAG
TTAGAAGTTTTCAGAATGATCATATGAACTCGAGAAAAATGGAATGACATCGGCTACAACTTTTTGGTCGGTGAAGATGGT
AATGTTTACGAAGGTCGTGGCTGGGGTAAACATGGCTCCCATTCAGTCCCGTACAATGCCAAGAGTATCGGTATTTGCCT
TATTGGTAAATTTAACAATAACGTACCGAATTCAGCGAGTATTCGAGCAACACAAAAATTTGATAGCTTACGGAGTAGCTA
ACAACAAAAATCAAATCCGATTATAAGCTTCTTGGCCATCGACAAAACCACTAAAACTGATTGTCCTGGAAATTTCTCTATAC
AATTTGATTAACACATGGCCTCACTGGACCGACACGCCATAAGAAATATAATCGTTATGCGATTAAATTAATCTACCAAG
TATAACGTCTTATTGTTTGGACTCGACGATCATTTGGAATTAACGATCATCGGATCGATAACTTTTGTGACTTTTCTT
TAATTAACGATGGATACCTTCAAATTCGATTTTCGTGCGAACATTTTACTTTGATGTTTCGTAAAGGAAGAG
```

BLAST gives this translated protein sequence:

Download ▾ GenBank Graphics ▾ Next ▴ Previous ▴ Descriptions

EST_420 Caste=Worker, Stage=1st, 2nd, and 3rd larval instar Vespula squamosa cDNA clone 1_3W3_86F11 5', mRNA sequence
Sequence ID: [EG326551.1](#) Length: 582 Number of Matches: 1

Range 1: 3 to 575 [GenBank](#) [Graphics](#) ▾ Next Match ▴ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
240 bits(613)	2e-78	Compositional matrix adjust.	112/191(59%)	139/191(72%)	6/191(3%)	+3
Query 2	TKLIAVLFLVNCQLLFCVHE-----TPVVRPRIISRSEWGARKPTTIRALAQNPPPF	55				
Sbjct 3	T+++ FLVV I C E T P I+SR +WGA+ P + L NPPP+ TRMVRATFLLVATCIAICRAAEVADSAAFTETPNIVSRQWGAKPKSPNLMNPPPY	182				
Query 56	VIIHHSATDSCITQAICNARVRSFQNYHIDEKGWDIGYQFLVGEDGNIYEGRGWDKHGA	115				
Sbjct 183	V+IHHS + C TQAIC ARVRSFQN H++ + W DIGY FLVGEDGN+YEGRGW KHG+ VVIHHSDSVGCTTQAICQARVRSFQNDHMNSRKWNDIGYNFLVGEDGNVYEGRGWKGHS	362				
Query 116	HSISYNSKSGIGICIIGNFVGHTPNAAAEATKNLISYGVAIGKIQSNYTLGLHRQTTRTS	175				
Sbjct 363	HS+ YN+KSGIG+IG F + PN+A+I AT+NLI+YGVA KI+S+Y LLGHRQTT+T HSVPYNKSGIGICLIGFNNNVPSASIRATQNLIAYGVANNIKSDYKLLGHRQTTKTD	542				
Query 176	CPGDSLYELIK 186					
Sbjct 543	CPG+SLY LIK CPGNSLYNLIK 575					

```
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
REWYEQRFYLLQRALQFAVQRKLQIRLQPLKHLISFQDNNGELNRRKALRLILK*IHLLT
LLFIIRIQLVVLKQFVRLELEVFRMII*TRENGMTSATTFWSVKMVMFTKVVAGVNMAP
IQSRTMPRVSVFALLVNLITITYRIQRVFEQHKI**LTE*LTTKSNPIISFLAIDKPLKLI
VLEILYTI*LKHGLTGPTRHKKYNRYAIKLIYQV*RLIVLDSTIIGN*RSSDR*LLLTFL
*LTMDFKFDFRANIFT*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*RTEFSEYSSNTKFDSLRS*QQNQIRL*ASWPSTNH*N*L
SWKFISIQFD*NMASLDRHAIRNIIVMRLN*STKYNVLLFWTRRSLEINDHRIDNFC*LFF
N*RWIPSNSIFVRTFLLDVS*RKX

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

RMVRATFLLVATCIAICRAAEVADSAATFETPNIVSRQQWGAKPPKSPTPNLKMNPPPYV
VIHHSDSVGCTTQAICQARVRSFQNDHMNSRKWN DIGYNFLVGEDGNVYEGRGWGKHGSH
SVPYNAKSIGICLIGKFNNNVPNSASIRATQNLIAYGVANNKIKSDYKLLGHRQTTKTDC
PGNSLYNLIKTWPHWTDTP*EI*SLCD*INLPSITSYCFGLDDHWKLTIGSITFVDFSL
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*FNRI
IIFLMACRSSEAMF*SNCIENFQDNQF*WFVDGQEAYNRI*FCC*LLRKLSNFVLLLEYS
NSVRYC*IYQ*GKYRYSWHCTGLNGSHVYPSHDLRKHYHLHRPKSCSRCHSIFSSSYDHS
ENF*L*PDKLLEWYNQLNPND*QRKEVDSF*D*A*GFSAV*LPIVVVKRY*VFQRLQPN
LQLPLHGKQLQCTLQQVETLLVPFS

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

LPLRNIK*KCSHENRI*RYPSLIKEKSTKVIDPMIVNFQ*SSSPKQ*DVLGRLI*SHND
YISYGVSVQ*GHVLIKLYREFPGQSVLVVCRWPRSL*SDLILLLATP*AIKFCVARILAE
FGTLLLNLPPIRQIPILLALYGTEWEPCLPQPRPS*TLPSSTPKKL*PMSFHFLEFI*SF*
KLLTLA*QIA*VVQPTSE**ITT*GGGFILRLGVGLFGGLAPHCCLETLGVSKVAAES
ATSAARQIAMHVATSRNVARTILX

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

SSFTKHQVKMFARKSNLKVSI*RVNKSYSRDD*FPMIVESKTIRRYTW*INLIA*R
LYFLWRVGPVPCFNQIV*RISRTISFSGLSMAKKLIIGFDFVVSYSVSYQILCCSNTR*
IRYVIVKFTNKANTDTLGIVRD*MGAMFTPATTFVNITIFTDQKVADVIPFSRVHMIIL
KTSNSSLTNCLSGTTN*IRMMNNVRRWIHFKIRRRAFRRFSSPLLS*NDIRCFKGCSRI
CNFRCTANCNARCNK*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 no it 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
<input type="checkbox"/>	PREDICTED: peptidoglycan-recognition protein SC2-like [Polistes dominula]	293	293	68%	3e-97	74.03%	XP_015177188.1
<input type="checkbox"/>	PREDICTED: peptidoglycan recognition protein 3-like [Polistes canadensis]	276	451	64%	9e-88	74.10%	XP_014602399.1
<input type="checkbox"/>	peptidoglycan-recognition protein SC2 isoform X1 [Ooceraea biroi]	264	264	73%	5e-86	62.44%	XP_011344962.1
<input type="checkbox"/>	PREDICTED: peptidoglycan-recognition protein SC2-like [Trachymyrmex cornetzi]	264	264	67%	5e-86	65.54%	XP_018369940.1
<input type="checkbox"/>	PREDICTED: peptidoglycan-recognition protein SC2-like [Atta cephalotes]	262	262	66%	2e-85	65.34%	XP_012059667.1
<input type="checkbox"/>	Peptidoglycan-recognition protein SC2 [Acromyrmex echinator]	263	263	64%	2e-85	66.67%	EGI64536.1
<input type="checkbox"/>	PREDICTED: peptidoglycan-recognition protein SC2-like [Atta colombica]	260	260	66%	9e-85	64.77%	XP_018046500.1
<input type="checkbox"/>	peptidoglycan-recognition protein SB1-like [Pogonomyrmex barbatus]	260	260	73%	2e-84	59.28%	XP_011631523.1
<input type="checkbox"/>	peptidoglycan-recognition protein SC2 isoform X2 [Ooceraea biroi]	258	258	63%	5e-84	68.86%	XP_011344963.1
<input type="checkbox"/>	peptidoglycan-recognition protein precursor [Myrmica ruginodis]	257	257	64%	2e-83	65.09%	ACT66860.1
<input type="checkbox"/>	peptidoglycan-recognition protein precursor [Myrmica lobicornis]	256	256	67%	3e-83	63.84%	ACT66861.1
<input type="checkbox"/>	Peptidoglycan recognition protein 3 [Trachymyrmex cornetzi]	262	412	64%	4e-83	67.25%	KYN14761.1
<input type="checkbox"/>	peptidoglycan-recognition protein precursor [Myrmica brevispinosa]	255	255	64%	6e-83	65.09%	ACT66868.1
<input type="checkbox"/>	peptidoglycan-recognition protein precursor [Myrmica alaskensis]	255	255	63%	7e-83	64.88%	ACT66869.1
<input type="checkbox"/>	peptidoglycan-recognition protein precursor [Myrmica sulcinodis]	254	254	65%	1e-82	64.16%	ACT66863.1
<input type="checkbox"/>	peptidoglycan-recognition protein precursor [Myrmica fracticornis]	254	254	64%	2e-82	64.50%	ACT66867.1
<input type="checkbox"/>	PREDICTED: peptidoglycan-recognition protein LF-like [Habropoda laboriosa]	263	447	63%	2e-82	70.91%	XP_017789839.1
<input type="checkbox"/>	peptidoglycan-recognition protein precursor [Myrmica sabuleti]	254	254	65%	2e-82	63.22%	ACT66865.1
<input type="checkbox"/>	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.