BGGN-213: FOUNDATIONS OF BIOINFORMATICS

Find-a-gene project assignment!

Name: Blair Chang

UCSD email: yac046@ucsd.edu

PID: A59000602

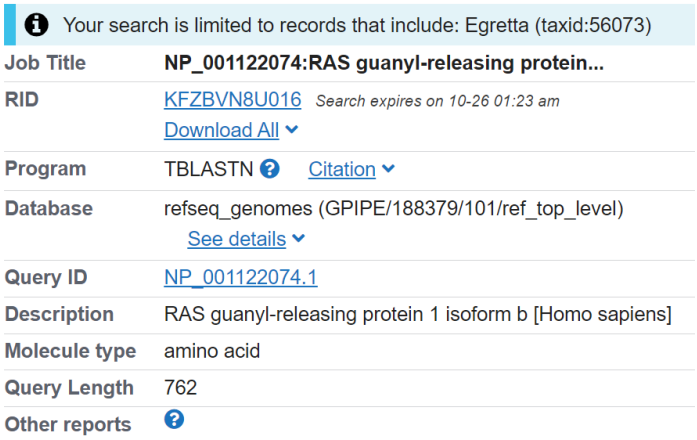
Q1. Tell me the name of a protein you are interested in. Include the species and the accession number. This can be a human protein or a protein from any other species as long as its function is known.

1. Protein name:
   RAS guanyl-releasing protein 1 (RASGRP1) isoform b
2. Species:
   Home sapiens (taxid:9606)
3. Accession number:
   NP_001122074.1
4. Function known:
   RASGRP1 is guanine nucleotide exchange factor (GEF) that activates RAS small GTPase by converting GDP to GTP. It mediates MAPK signaling in T cells and is important for T cell activation and proliferation.

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). 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.

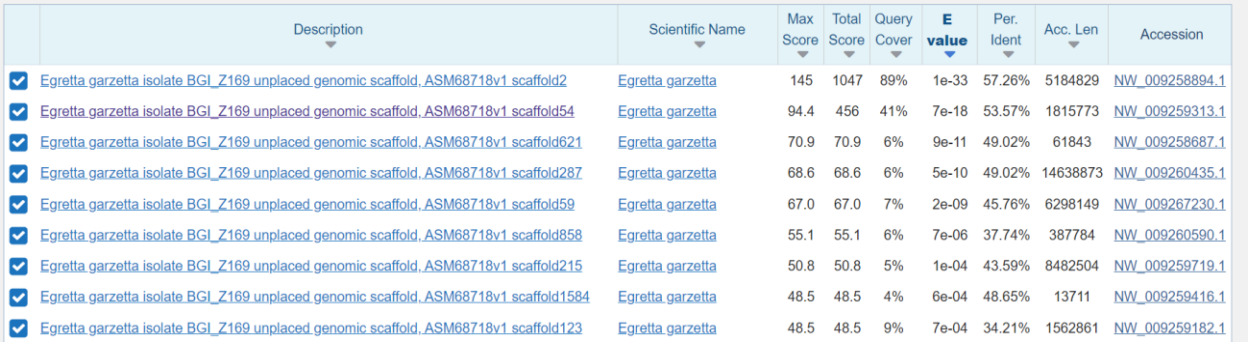
[Search Results]

1. Blast method:
   TBLASTN



Your search is limited to records that include: Egretta (taxid:56073)

Job Title	**NP_001122074:RAS guanyl-releasing protein...**
RID	KFZBVN8U016 *Search expires on 10-26 01:23 am* Download All
Program	TBLASTN Citation
Database	refseq_genomes (GPIPE/188379/101/ref_top_level) See details
Query ID	NP_001122074.1
Description	RAS guanyl-releasing protein 1 isoform b [Homo sapiens]
Molecule type	amino acid
Query Length	762
Other reports	



Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Egretta garzetta isolate BGI_Z169 unplaced genomic scaffold, ASM68718v1 scaffold2	Egretta garzetta	145	1047	89%	1e-33	57.26%	5184829	NW_009258894.1
Egretta garzetta isolate BGI_Z169 unplaced genomic scaffold, ASM68718v1 scaffold54	Egretta garzetta	94.4	456	41%	7e-18	53.57%	1815773	NW_009259313.1
Egretta garzetta isolate BGI_Z169 unplaced genomic scaffold, ASM68718v1 scaffold621	Egretta garzetta	70.9	70.9	6%	9e-11	49.02%	61843	NW_009258687.1
Egretta garzetta isolate BGI_Z169 unplaced genomic scaffold, ASM68718v1 scaffold287	Egretta garzetta	68.6	68.6	6%	5e-10	49.02%	14638873	NW_009260435.1
Egretta garzetta isolate BGI_Z169 unplaced genomic scaffold, ASM68718v1 scaffold59	Egretta garzetta	67.0	67.0	7%	2e-09	45.76%	6298149	NW_009267230.1
Egretta garzetta isolate BGI_Z169 unplaced genomic scaffold, ASM68718v1 scaffold858	Egretta garzetta	55.1	55.1	6%	7e-06	37.74%	387784	NW_009260590.1
Egretta garzetta isolate BGI_Z169 unplaced genomic scaffold, ASM68718v1 scaffold215	Egretta garzetta	50.8	50.8	5%	1e-04	43.59%	8482504	NW_009259719.1
Egretta garzetta isolate BGI_Z169 unplaced genomic scaffold, ASM68718v1 scaffold1584	Egretta garzetta	48.5	48.5	4%	6e-04	48.65%	13711	NW_009259416.1
Egretta garzetta isolate BGI_Z169 unplaced genomic scaffold, ASM68718v1 scaffold123	Egretta garzetta	48.5	48.5	9%	7e-04	34.21%	1562861	NW_009259182.1

2. Search output list (top 5 hits):

	Description	Scientific Name	Max Score	Query Cover	E-value	Per. Ident	Accession
1 Picked	Egretta garzetta isolate BGI_Z169 unplaced genomic scaffold, ASM68718v1 scaffold2	Egretta garzetta	145	89%	1e-33	57.26%	NW_009258894.1
2	Egretta garzetta isolate BGI_Z169 unplaced genomic scaffold, ASM68718v1 scaffold54	Egretta garzetta	94.4	41%	7e-18	53.57%	NW_009259313.1
3	Egretta garzetta isolate BGI_Z169 unplaced genomic scaffold, ASM68718v1 scaffold621	Egretta garzetta	70.9	6%	9e-11	49.02%	NW_009258687.1
4	Egretta garzetta isolate BGI_Z169 unplaced genomic scaffold, ASM68718v1 scaffold287	Egretta garzetta	68.6	6%	5e-10	49.02%	NW_009260435.1
5	Egretta garzetta isolate BGI_Z169 unplaced genomic scaffold, ASM68718v1 scaffold59	Egretta garzetta	67.0	7%	2e-09	45.76%	NW_009267230.1

3. Database searched:
   refseq_genomes
4. Limits applied:
   Organism limited to Egretta (taxid:56073)
5. Alignment of choice:
   Egretta garzetta isolate BGI_Z169 unplaced genomic scaffold, ASM68718v1 scaffold2

Egretta garzetta isolate BGI_Z169 unplaced genomic scaffold, ASM68718v1 scaffold2

Sequence ID: NW_009258894.1 **Length:** 5184829 **Number of Matches:** 12

Range 1: 1148188 to 1148559 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
145 bits(365)	1e-33	Compositional matrix adjust.	79/124(64%)	94/124(75%)	0/124(0%)	-2

```
Query  323      VLGEMTELLSSSRNYDNYRRAYGECTDFKIPILGVHLKDLISLYEAMPDYLEDGKVNVHK  382
                V  EMTELLSS RNYD+YRRAY EC++FKIPILGVHLKDLISLYE MPDYLED K+N++K
Sbjct  1148559  VFNEMTELLSSYRNYDSYRRAYNECSNFKIPILGVHLKDLISLYEGMPDYLEDKKINIYK  1148380

Query  383      LLALYNHISELVQLQEVAPPLEANKDlvhlltlsldlYYTEDEIYELSYAREPRNHRAPS  442
                L +LYNHI+EL+QLQE+  PLEAN DLVHLLT+SL        +  +SY  +    R P
Sbjct  1148379  LYSLYNHINELIQLQEMPLPLEANMDLVHLLTVSLLDKINMSILKIISYTSQTTK*RIPL  1148200

Query  443      VFKN  446
                +  N
Sbjct  1148199  ILHN  1148188
```

6. E-value and other alignment stats:

E-value	Score	Method	Identities	Positives	Gaps
1e-33	145 bits (365)	Compositional matrix adjust	79/124(64%)	75%	0%

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.

Egretta garzetta isolate BGI_Z169 unplaced genomic scaffold, ASM68718v1 scaffold2

Sequence ID: NW_009258894.1 **Length:** 5184829 **Number of Matches:** 12

Range 1: 1148188 to 1148559 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
145 bits(365)	1e-33	Compositional matrix adjust.	79/124(64%)	94/124(75%)	0/124(0%)	-2

```
Query  323      VLGEMTELLSSSRNYDNYRRAYGECTDFKIPILGVHLKDLISLYEAMPDYLEDGKVNVHK  382
                V  EMTELLSS RNYD+YRRAY EC++FKIPILGVHLKDLISLYE MPDYLED K+N++K
Sbjct  1148559  VFNEMTELLSSYRNYDSYRRAYNECSNFKIPILGVHLKDLISLYEGMPDYLEDKKINIYK  1148380

Query  383      LLALYNHISELVQLQEVAPPLEANKDlvhlltlsldlYYTEDEIYELSYAREPRNHRAPS  442
                L +LYNHI+EL+QLQE+  PLEAN DLVHLLT+SL        +  +SY  +    R P
Sbjct  1148379  LYSLYNHINELIQLQEMPLPLEANMDLVHLLTVSLLDKINMSILKIISYTSQTTK*RIPL  1148200

Query  443      VFKN  446
                +  N
Sbjct  1148199  ILHN  1148188
```

1. Protein sequence of choice matches Subject above:

VAQKLHQLQNFNTLMAVIGGLCHSSISRLKETSSCVPHDVIKVFNEMTELLSSYRNYDSYRRAYNECSN
FKIPILGVHLKDLISLYEGMPDYLEDKKINVYKLYSLYNHIDELIQLQEMPLPLEANMDLVHLLTLSLD
LYYTEDEIYELSYAREPRSHRAAPMTPSKPPVVADWASGVAPKPDPKTISKHVQRMVDSVFKNYDHDQD
GYISQEEFEKIAASFPFSFCVMAKDWEG

2. Name in header:
   >AGX29484.1 RAS guanyl releasing protein 1, partial [Serinus canaria]

3. Species:
   Egretta garzetta

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.

1. Blastp output list with identities and Evalue:

Job Title	**Protein Sequence**
RID	KPA7X1CE013 *Search expires on 10-28 11:05 am* Download All
Program	BLASTP Citation
Database	nr See details
Query ID	lcl|Query_91565
Description	unnamed protein product
Molecule type	amino acid
Query Length	60
Other reports	Distance tree of results Multiple alignment MSA viewer

Description	Scientific Name	E value	Per. Ident	Accession
RAS guanyl releasing protein 1 [Serinus canaria]	Serinus canaria	7e-34	98.33%	AGX29484.1
hypothetical protein DUI87_23033 [Hirundo rustica rustica]	Hirundo rustica rustica	4e-33	98.33%	RMC00424.1
RAS guanyl-releasing protein 1 [Willisornis vidua]	Willisornis vidua	6e-33	100.00%	KAJ7412799.1
GRP1 protein [Pandion haliaetus]	Pandion haliaetus	6e-33	100.00%	NXS75269.1
PREDICTED: RAS guanyl-releasing protein 1-like [Cariama cristata]	Cariama cristata	1e-32	100.00%	XP_009696774.1

2. Top alignment shown with alignment statistics:

RAS guanyl releasing protein 1, partial [Serinus canaria]

Sequence ID: AGX29484.1 Length: 235 Number of Matches: 1

Range 1: 43 to 102 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
125 bits(314)	7e-34	Compositional matrix adjust.	59/60(98%)	60/60(100%)	0/60(0%)

```
Query  1    VFNEMTELLSSYRNYDSYRRAYNECSNFKIPILGVHLKDLISLYEGMPDYLEDKKINIYK  60
           VFNEMTELLSSYRNYDSYRRAYNECSNFKIPILGVHLKDLISLYEGMPDYLEDKKIN+YK
Sbjct  43   VFNEMTELLSSYRNYDSYRRAYNECSNFKIPILGVHLKDLISLYEGMPDYLEDKKINVYK  102
```

3. Results indicates a “novel” gene found:

   My top match has 98.33% identity to a different species - Serinus canaria. Thus, this gene from NW_009258894.1 and its corresponding protein are novel.

"Hi! It's me, again!"

Q5. Generate a multiple sequence alignment with your novel protein, your original query protein, and a group of other members of this family from different species. A typical number of proteins to use in a multiple sequence alignment for this assignment purpose is a minimum of 5 and a maximum of 20 - although the exact number is up to you. Include the multiple sequence alignment in your report. Use Courier font with a size appropriate to fit page width.

Re-labeled sequences for my alignment:

>HUMAN_RASGRP1

MGTLGKAREAPRKPSHGCRAASKARLEAKPANSPFPSHPSLAHITQFRMMVSLGHLAKGASLDDLIDSCIQSFDADGNLCR
SNQLLQVMLTMHRIVISSAELLQKVITLYKDALAKNSPGLCLKICYFVRYWITEFWVMFKMDASLTDTMEEFQELVKAKGE
ELHCRLIDTTQINARDWSRKLTQRIKSNTSKKRKVSLLFDHLEPEELSEHLTYLEFKSFRRISFSDYQNYLVNSCVKENPT
MERSIALCNGISQWVQLMVLSRPTPQLRAEVFIKFIQVAQKLHQLQNFNTLMAVIGGLCHSSISRLKETSSHVPHEINKVL
GEMTELLSSSRNYDNYRRAYGECTDFKIPILGVHLKDLISLYEAMPDYLEDGKVNVHKLLALYNHISELVQLQEVAPPLEA
NKDLVHLLTLSLDLYYTEDEIYELSYAREPRNHRAPPLTPSKPPVVVDWASGVSPKPDPKTISKHVQRMVDSVFKNYDHDQ
DGYISQEEFEKIAASFPFSFCVMDKDREGLISRDEITAYFMRASSIYSKLGLGFPHNFQETTYLKPTFCDNCAGFLWGVIK
QGYRCKDCGMNCHKQCKDLVVFECKKRAKNPVAPTENNTSVGPVSNLCSLGAKDLLHAPEEGPFTFPNGEAVEHGEESKDR
TIMLMGVSSQKISLRLKRAVAHKATQTESQPWIGSEGPSGPFVLSSPRKTAQDTLYVLPSPTSPCPSPVLVRKRAFVKWEN
KDSLIKSKEELRHLRLPTYQELEQEINTLKADNDALKIQLKYAQKKIESLQLEKSNHVLAQMEQGDCS

>Novel_protein_Serinus_canaria

VAQKLHQLQNFNTLMAVIGGLCHSSISRLKETSSCVPHDVIKVFNEMTELLSSYRNYDSYRRAYNECSNFKIPILGVHLKD
LISLYEGMPDYLEDKKINVYKLYSLYNHIDELIQLQEMPLPLEANMDLVHLLTLSLDLYYTEDEIYELSYAREPRSHRAAP
MTPSKPPVVADWASGVAPKPDPKTISKHVQRMVDSVFKNYDHDQDGYISQEEFEKIAASFPFSFCVMAKDWEG

>Amazona_guildingii

VFNEMTELLSSYRNYDSYRRAYNECSNFKIPILGVHLKDLISLYEGMPDYLEDKKINIYKLYSLYNHINELIQLQEMPLPL
EANMDLVHLLTLSLDLYYTEDEIYELSYAREPRSHRAAPLTPSRPPVVADWASGVAPKPDPKTISKHVQRMVDSVFKNYDH
DQDGYISQEEFEKIAASFPFSFCVMAKDWEGLISRDEITAYFMRASSIYSKLGLGFAHNFQETTYLRPTFCDNCAGFLWGV
IKQGYRCKDCGMNCHKQCKDLVVIECKRRPKTSIPDSSPTSALASSLCPVGVKEQFHGQEEGLFTFPNGEVVEHSEDSKDR
TIMLMGSSAQKISVRLKPSVVHEGTQTDPVLLAGDVSRRQIEKKEHKMPENPYLQLAPPSPFPSPILGRKKAYVKWENKDS
SQKKKEEHYSCKPSYQELEQERNILKAHNEGLKIQLEQAHKTIESLTIHRRNHVVDNLQHRDCS

>Neopelma_chrysocephalum

MGTLGKRRENQQSAQACSTAPESALELKQTSHCPSLSNHTQVMMVPLGHLAKGATLEDLLETCIQSFDLEGNAYQNNQLLK
IILAMHQFIISSADMLQKLIDLYLNALENNSSMLCVKICYFVRYWITEFWIMFKMDSKLSTTMEEFQELVRANGEELHCRL
IDTSQINSRDWSRKLTQRVKANTSKKRKVSLLFDHLEPEELSDHLTYLEFKSFRRITFSDYQNYIVNSCVKENPTMERSIS
LCNGISQWVQLMVLSRPTPQLRAEVFIKFIHVAQKLHQLQNFNTLMAVIGGLCHSSISRLKETSSCVPHDVIKVFNEMTEL
LSSYRNYDSYRRAYNECSNFKIPILGVHLKDLISLYEGMPDYLEDKKINIYKLYSLYNHINELIQLQEMPLPLEANMDLVH
LLTLSLDLYYTEDEIYELSYAREPRSHRAAPLTPCKPPVVADWASGVAPKPDPKTISKHVQRMVDSVFKNYDHDQDGYISQ
EEFEKIAASFPFSFCVMAKDWEGLISRDEITAYFMRASSIYSKLGLGFAHNFQETTYLRPTFCDNCAGFLWGVIKQGYRCK
DCGMNCHKQCKDLVVIECKRRPKTSVADSSPTSALASSLCPVGVKEQFHGKKRSH

>Theropithecus_gelada
GAAAGTCTCAAACCAGTTATTCACCTGAGCCACCTGGATGAACTTGATGAAGACTTCTGCTCGGAGCTGCGGGGTGGGGCGGCTGAGAA
CCATCAGTTGTACCCACTGGGAGATACCGTTGCACAGAGCAATAGATCTCTCCATGGTGGGGTTTTCCTTCACACAGCTATTTACAAGG
TAATTCTGATAATCAGA

>Sagittarius_serpentarius

DLEGNAYQNNQLLKIILAMHQFIISSADMLQKLFSTYLNALENKSSALCVKICYFVRYWITEFWVMFKMDSKLSTTMEEFQ
ELVKANGEELHCHLIDTTQINSRDWSRKLTQRVKANTSKKRKVSLLFDHLEPEELSDHLTYLEFKSFRRISFSDYQNYIVN
SCVKENPTMERSIALCNGISQWVQLMVLSRPTPQLRAEVFIKFIHVAQKLHQLQNFNTLMAVIGGLCHSSISRLKETSSCV
PHDVIKVFNEMTELLSSYRNYDSYRRAYNECSNFKIPILGVHLKDLISLYEGMPNYLEDKKINIYKLYSLYNHINELIQLQ
EMPLPLEANMDLVHLLTLSLDLYYTEDEIYELSYAREPRSHRAAPLTPSRPPVVADWASGVAPKPDPKTISKHVQRMVDSV
FKNYDHDQDGYISQEEFEKIAASFPFSFCVMAKDW

>Macaca_mulatta
TACAAATCCAGCACAGTTGTCACAAAAAGTGGGCTTCAGGTAGGTGGTCTCTTGGAAGTTGTGAGGAAAGCCCAGGCCCAGCTTGGAAT
AGATTGAGCTGGCTCGCATGAAGTAGGCTGTGATCTCATCCCTGCTGATGAGGCCTTCCCTGCCAGCAAATGACCAAGGCAAGGATGTG
AGTATACG

Multiple Sequence Alignment using MUSCLE at EMBL-EBI:

CLUSTAL multiple sequence alignment by MUSCLE (3.8)

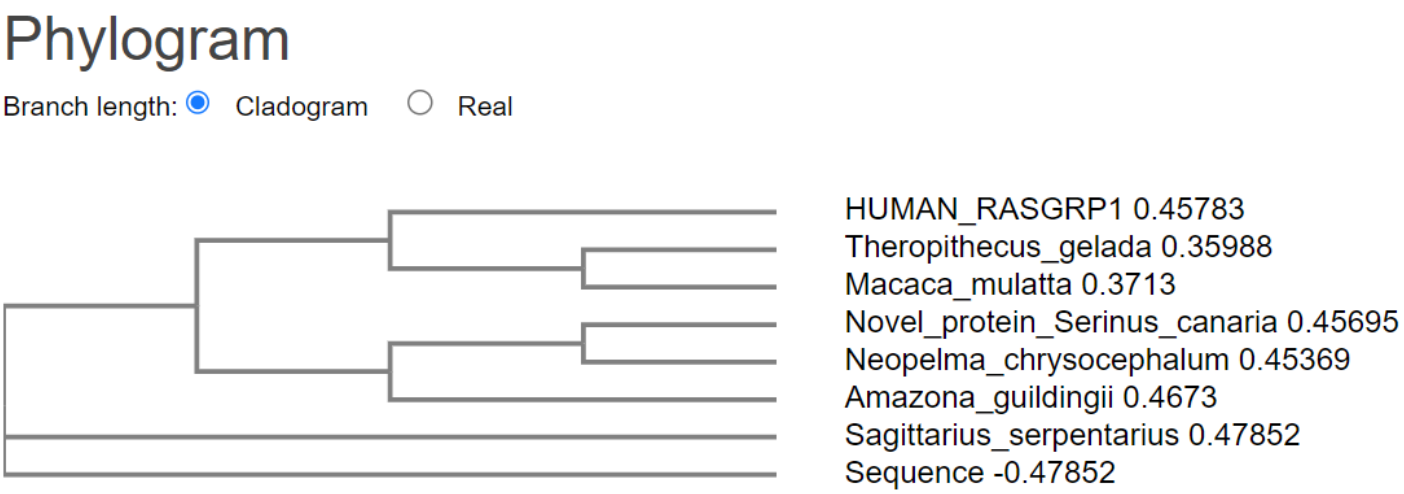
```
HUMAN_RASGRP1                  MGTLGKAREAPRKPSHGCRAASKARLEAKPANSPFPSHPSLAHITQFRMMVSLGHLAKGA
Amazona_guildingii             ------------------------------------------------------------
Sagittarius_serpentarius       ------------------------------------------------------------
Neopelma_chrysocephalum        MGTLGKRRENQQSAQACSTAPESALELKQTSHCPSLSNHTQV------MMVPLGHLAKGA
Novel_protein_Serinus_canaria  ------------------------------------------------------------
Theropithecus_gelada           ------------------------------------------------------------
Macaca_mulatta                 ------------------------------------------------------------

HUMAN_RASGRP1                  SLDDLIDSCIQSFDADGNLCRSNQLLQVMLTMHRIVISSAELLQKVITLYKDALAKNSPG
Amazona_guildingii             ------------------------------------------------------------
Sagittarius_serpentarius       ------------DLEGNAYQNNQLLKIILAMHQFIISSADMLQKLFSTYLNALENKSSA
Neopelma_chrysocephalum        TLEDLLETCIQSFDLEGNAYQNNQLLKIILAMHQFIISSADMLQKLIDLYLNALENNSSM
Novel_protein_Serinus_canaria  ------------------------------------------------------------
Theropithecus_gelada           ------------------------------------------------------GAAAG
Macaca_mulatta                 ------------------------------------------------------------

HUMAN_RASGRP1                  LCLKICYFVRYWITEFWVMFKMDASLTDTMEEFQELVKAKGEELHCRLIDTTQINARDWS
Amazona_guildingii             ------------------------------------------------------------
Sagittarius_serpentarius       LCVKICYFVRYWITEFWVMFKMDSKLSTTMEEFQELVKANGEELHCHLIDTTQINSRDWS
Neopelma_chrysocephalum        LCVKICYFVRYWITEFWIMFKMDSKLSTTMEEFQELVRANGEELHCRLIDTSQINSRDWS
Novel_protein_Serinus_canaria  ------------------------------------------------------------
Theropithecus_gelada           TC----------------------------------------------------------
Macaca_mulatta                 ------------------------------------------------------------

HUMAN_RASGRP1                  RKLTQRIKSNTSKKRKVSLLFDHLEPEELSEHLTYLEFKSFRRISFSDYQNYLVNSCVKE
Amazona_guildingii             ------------------------------------------------------------
Sagittarius_serpentarius       RKLTQRVKANTSKKRKVSLLFDHLEPEELSDHLTYLEFKSFRRISFSDYQNYIVNSCVKE
Neopelma_chrysocephalum        RKLTQRVKANTSKKRKVSLLFDHLEPEELSDHLTYLEFKSFRRITFSDYQNYIVNSCVKE
Novel_protein_Serinus_canaria  ------------------------------------------------------------
Theropithecus_gelada           ------------------------------------------------------------
Macaca_mulatta                 ------------------------------------------------------------

HUMAN_RASGRP1                  NPTMERSIALCNGISQWVQLMVLSRPTPQLRAEVFIKFIQVAQKLHQLQNFNTLMAVIGG
Amazona_guildingii             ------------------------------------------------------------
Sagittarius_serpentarius       NPTMERSIALCNGISQWVQLMVLSRPTPQLRAEVFIKFIHVAQKLHQLQNFNTLMAVIGG
Neopelma_chrysocephalum        NPTMERSISLCNGISQWVQLMVLSRPTPQLRAEVFIKFIHVAQKLHQLQNFNTLMAVIGG
Novel_protein_Serinus_canaria  ----------------------------------------VAQKLHQLQNFNTLMAVIGG
Theropithecus_gelada           ------------------------------------------------------TCAAACCAG
Macaca_mulatta                 ------------------------------------------------------TACAAA
```

```
HUMAN_RASGRP1                   LCHSSISRLKETSSHVPHEINKVLGEMTELLSSSRNYDNYRRAYGECTDFKIPILGV--H
Amazona_guildingii              ----------------------VFNEMTELLSSYRNYDSYRRAYNECSNFKIPILGV--H
Sagittarius_serpentarius        LCHSSISRLKETSSCVPHDVIKVFNEMTELLSSYRNYDSYRRAYNECSNFKIPILGV--H
Neopelma_chrysocephalum         LCHSSISRLKETSSCVPHDVIKVFNEMTELLSSYRNYDSYRRAYNECSNFKIPILGV--H
Novel_protein_Serinus_canaria   LCHSSISRLKETSSCVPHDVIKVFNEMTELLSSYRNYDSYRRAYNECSNFKIPILGV--H
Theropithecus_gelada            TTATTCACCTGAGCCAC-----CTGGATGAA---------------CTTGATGAAGACTT
Macaca_mulatta                  TCCAGCACAGTTGTCACAAAAAGTGGG-------------------CTTCAGGTAGG--T
                                                        .                    *:      *

HUMAN_RASGRP1                   LKDLISLYEAMPDYLEDGKVNVHKLLALYNHISELVQLQEVAPPLEANKDLVHLLTLSLD
Amazona_guildingii              LKDLISLYEGMPDYLEDKKINIYKLYSLYNHINELIQLQEMPLPLEANMDLVHLLTLSLD
Sagittarius_serpentarius        LKDLISLYEGMPNYLEDKKINIYKLYSLYNHINELIQLQEMPLPLEANMDLVHLLTLSLD
Neopelma_chrysocephalum         LKDLISLYEGMPDYLEDKKINIYKLYSLYNHINELIQLQEMPLPLEANMDLVHLLTLSLD
Novel_protein_Serinus_canaria   LKDLISLYEGMPDYLEDKKINVYKLYSLYNHIDELIQLQEMPLPLEANMDLVHLLTLSLD
Theropithecus_gelada            CTGCTCGGAGCTGCGGGGT-----------------------------------------
Macaca_mulatta                  GGTCTCTTGGAAGTTGTGA-----------------------------------------
                                     .   . ..

HUMAN_RASGRP1                   LYYTEDEIYELSYAREPRNHRAPPLTPSKPPVVVDWASGVSPKPDPKTISKHVQRMVDSV
Amazona_guildingii              LYYTEDEIYELSYAREPRSHRAAPLTPSRPPVVADWASGVAPKPDPKTISKHVQRMVDSV
Sagittarius_serpentarius        LYYTEDEIYELSYAREPRSHRAAPLTPSRPPVVADWASGVAPKPDPKTISKHVQRMVDSV
Neopelma_chrysocephalum         LYYTEDEIYELSYAREPRSHRAAPLTPCKPPVVADWASGVAPKPDPKTISKHVQRMVDSV
Novel_protein_Serinus_canaria   LYYTEDEIYELSYAREPRSHRAAPMTPSKPPVVADWASGVAPKPDPKTISKHVQRMVDSV
Theropithecus_gelada            -----------------GGGGCGGCTGAGAACCATCAGTTGTACCCACTGGGAGATACCG
Macaca_mulatta                  ------------------GGAAAGCCCAGGCCCAGCTTGGAATAGATTGAGCTGGCTCGC
                                                  .  .    .    .  :  ..      .  .   .

HUMAN_RASGRP1                   FKNYDHDQDGYISQEEFEKIAASFPFSFCVMDKDREGLISRDEITAYFMRASSIYSKLGL
Amazona_guildingii              FKNYDHDQDGYISQEEFEKIAASFPFSFCVMAKDWEGLISRDEITAYFMRASSIYSKLGL
Sagittarius_serpentarius        FKNYDHDQDGYISQEEFEKIAASFPFSFCVMAKDW------------------------
Neopelma_chrysocephalum         FKNYDHDQDGYISQEEFEKIAASFPFSFCVMAKDWEGLISRDEITAYFMRASSIYSKLGL
Novel_protein_Serinus_canaria   FKNYDHDQDGYISQEEFEKIAASFPFSFCVMAKDWEG------------------------
Theropithecus_gelada            TTGCACAGAGCAATAG-----ATCTCTCCATGGTGGG------------------------
Macaca_mulatta                  ATGAAGTAGGCTGT-GATCTCATCCCTGCTGATGAGG------------------------
                                 ..      *  .        *:   : *.

HUMAN_RASGRP1                   GFPHNFQETTYLKPTFCDNCAGFLWGVIKQGYRCKDCGMNCHKQCKDLVVFECKKRAKNP
Amazona_guildingii              GFAHNFQETTYLRPTFCDNCAGFLWGVIKQGYRCKDCGMNCHKQCKDLVVIECKRRPKTS
Sagittarius_serpentarius        ------------------------------------------------------------
Neopelma_chrysocephalum         GFAHNFQETTYLRPTFCDNCAGFLWGVIKQGYRCKDCGMNCHKQCKDLVVIECKRRPKTS
Novel_protein_Serinus_canaria   ------------------------------------------------------------
Theropithecus_gelada            ----------------------------------GTTTTCCTTCACACAGCTATTTACA
Macaca_mulatta                  ---------------CCTTC---------------CCTGCCAGCAAATGACCAAGGCAA

HUMAN_RASGRP1                   VAPTENNTSVGPVSNLCSLGAKDLLHAPEEGPFTFPNGEAVEHGEESKDRTIMLMGVSSQ
Amazona_guildingii              IPDSSPTSALA--SSLCPVGVKEQFHGQEEGLFTFPNGEVVEHSEDSKDRTIMLMGSSAQ
Sagittarius_serpentarius        ------------------------------------------------------------
Neopelma_chrysocephalum         VADSSPTSALA--SSLCPVGVKEQFHGKKRSH--------------------------
Novel_protein_Serinus_canaria   ------------------------------------------------------------
Theropithecus_gelada            AGGTAATTCTGATAATCAGA----------------------------------------
Macaca_mulatta                  GGATGTGAGTA------------------------------------------------

HUMAN_RASGRP1                   KISLRLKRAVAHKATQTESQPWIGSEGPSGPFVLSSPRKTAQDTLYVLPSPTSPCPSPVL
Amazona_guildingii              KISVRLKPSVVHEGTQTDPVLLAGDVSRRQ---IEKKEHKMPENPYLQLAPPSPFPSPIL
Sagittarius_serpentarius        ------------------------------------------------------------
Neopelma_chrysocephalum         ------------------------------------------------------------
Novel_protein_Serinus_canaria   ------------------------------------------------------------
Theropithecus_gelada            ------------------------------------------------------------
Macaca_mulatta                  ------------------------------------------------------------

HUMAN_RASGRP1                   VRKRAFVKWENKDSLIKSKEELRHLRLPTYQELEQEINTLKADNDALKIQLKYAQKKIES
Amazona_guildingii              GRKKAYVKWENKDSSQKKKEEHYSCK-PSYQELEQERNILKAHNEGLKIQLEQAHKTIES
Sagittarius_serpentarius        ------------------------------------------------------------
Neopelma_chrysocephalum         ------------------------------------------------------------
Novel_protein_Serinus_canaria   ------------------------------------------------------------
```

```
Theropithecus_gelada            ------------------------------------------------------------
Macaca_mulatta                  ------------------------------------------------------------


HUMAN_RASGRP1                   LQLEKSNHVLAQMEQGDCS
Amazona_guildingii              LTIHRRNHVVDNLQHRDCS
Sagittarius_serpentarius        -------------------
Neopelma_chrysocephalum         -------------------
Novel_protein_Serinus_canaria   -------------------
Theropithecus_gelada            -------------------
Macaca_mulatta                  ---------------TACG
```

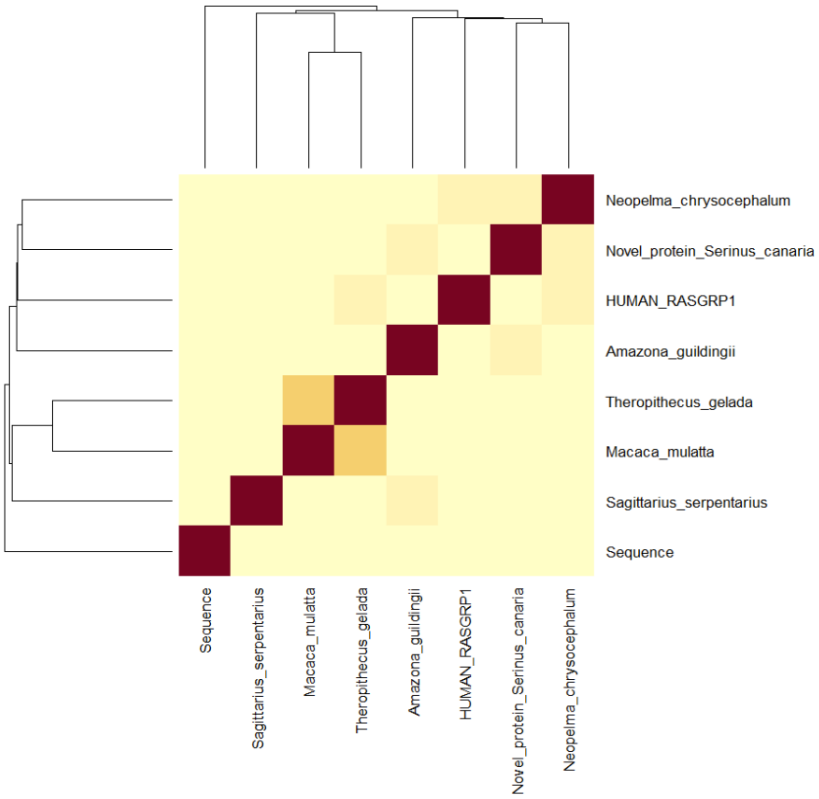
Q6. Create a phylogenetic tree, using either a parsimony or distance-based approach. Bootstrapping and tree rooting are optional. Use “simple phylogeny” online from the EBI or any respected phylogeny program (such as MEGA, PAUP, or Phylip). Paste an image of your Cladogram or tree output in your report.

Using simple phylogeny from EBI to create a phylogenetic tree (distance-based approach).



Q7. Generate a sequence identity based heatmap of your aligned sequences using R. Making heatmap using R/Bio3D

1. Converting alignment file to fasta file by MEGA.
2. Drawing heatmap on R.



Q8. Using R/Bio3D (or an online blast server if you prefer), search the main protein structure database for the most similar atomic resolution structures to your aligned sequences.

Since my consensus sequence has lots of gap positions, I used my original sequence of Novel_protein_Serinus_canaria for the search.

My top 3 unique hits are all from humans. They are RasGRP1, RasGRP4, RasGRP2. They are different proteins from RASGRP protein family.

ID	Method	Resolution	Source	E-value	Identity
4L9M_A	X-ray Diffraction	3 Å	Homo sapiens	1e-149	87.23%
6AXG_A	X-ray Diffraction	3.302 Å	Homo sapiens	1e-58	59.02%
6AXF_A	X-ray Diffraction	3.1 Å	Homo sapiens	3e-48	51.20%

Q9. Using AlphaFold notebook generate a structural model using the default parameters for your novel protein sequence.

Q10. Perform a “Target” search of ChEMBEL (https://www.ebi.ac.uk/chembl/) with your novel sequence. Are there any Target Associated Assays and ligand efficiency data reported that may be useful starting points for exploring potential inhibition of your novel protein? If there are no assays listed here simply list “non available as of [date]”.

There are 8 Target Associated Assays for my novel protein.

One is from an article published in J Med Chem (2018) 61:6261-6276: Activation of GFP-tagged RasGRP1 expressed in HEK293 cells assessed as ERK1/2 phosphorylation after 30 mins by immunoblot method. This paper proposed a ligand of Ras Guanine-Releasing Protein 3 (RasGRP3).

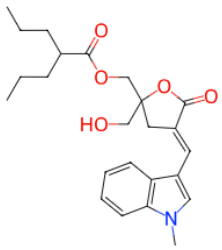
Assay ID: CHEMBL4137150
(https://www.ebi.ac.uk/chembl/assay_report_card/CHEMBL4137150/)

I think this ligand is promising since my interested protein (for Q1) is RasGRP1 in humans, which also belongs to RASGRP family. Thus, it’s likely that my novel protein has binding motif for this ligand.

Yes, there’s ligand efficiency data in my report.

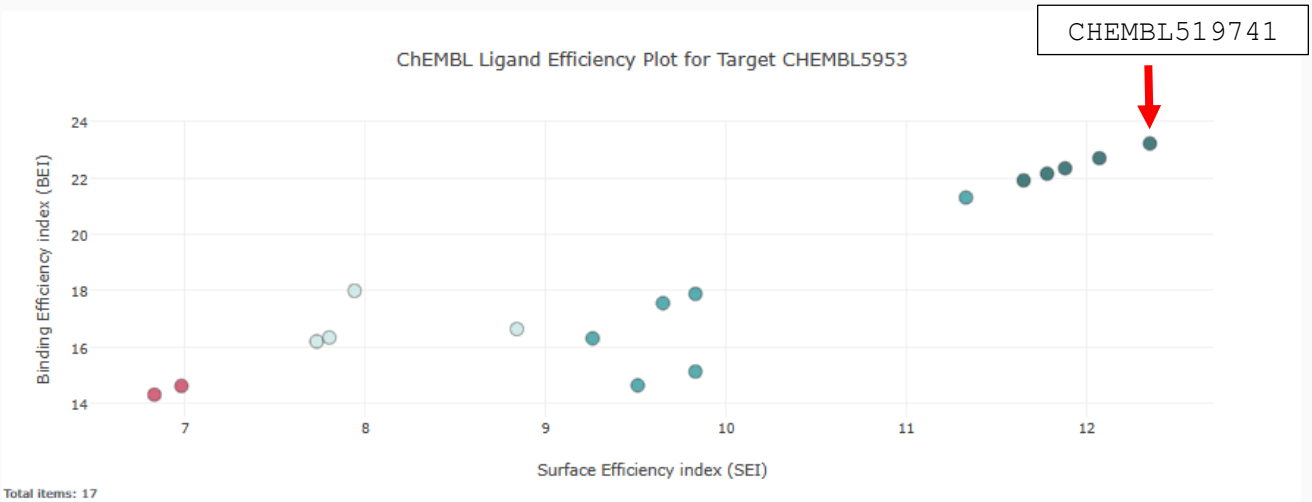
CHEMBL519741 shows the highest Binding Efficiency index (BEI).

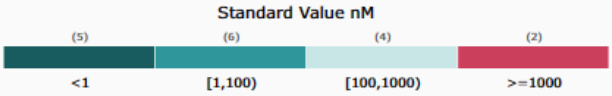
It is a small molecule that has a Molecular Weight of 413.51.



Ligand Efficiencies

See all bioactivities for target CHEMBL5953 used in this visualisation





The Ligand Efficiency chart plots Binding Efficiency Index (BEI) against Surface Efficiency Index (SEI), where:

SEI = (-log10(Standard Value*10^-9))*100/PSA
BEI = (-log10(Standard Value*10^-9))*1000/MWT