Seong Tae Gwon
A12364788
sgwon@ucsd.edu

Find a Gene Project

[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.

Name: Sonic hedgehog protein isoform 1 preproprotein (SHH)

Accession: NP_000184.1

Species: Homo Sapiens

Known Function: chemical signal to control embryonic development

[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).

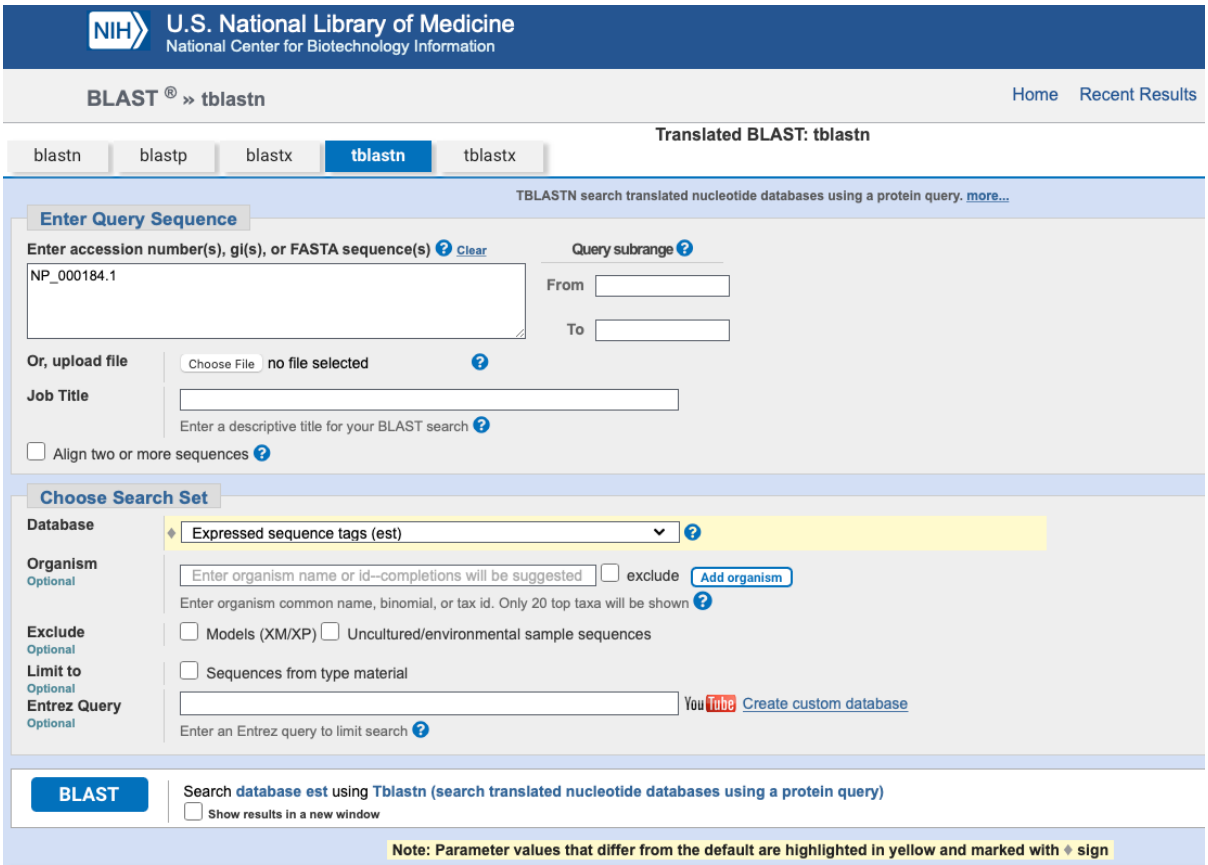
Method: NIH TBLASTN (2.12.0) search against nematode ESTs

Database: Expressed Sequence Tags (est)

Organism: All species

Seong Tae Gwon
A12364788
sgwon@ucsd.edu

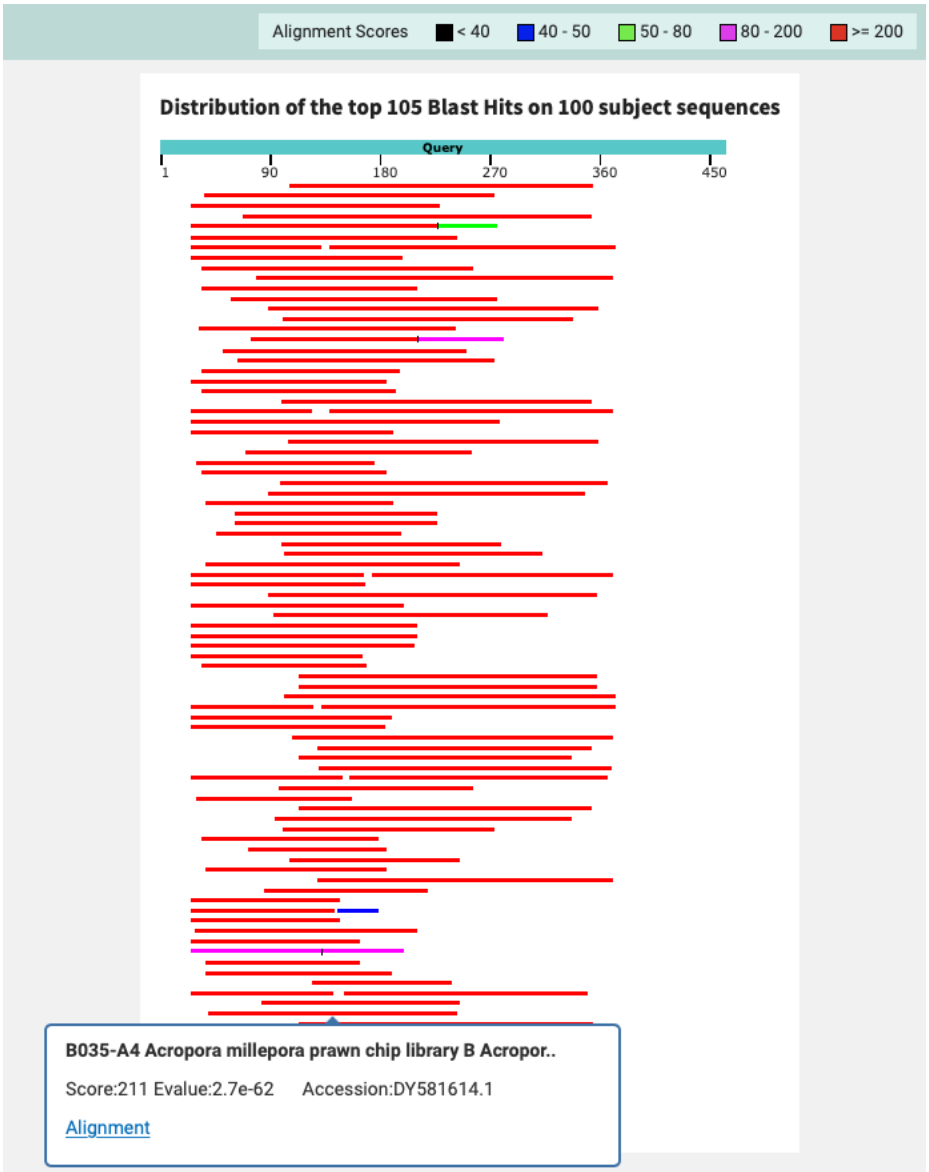
Also include the output of that BLAST search in your document. If appropriate, change the font to `Courier size 10` so that the results are displayed neatly.



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.

Chosen match: Accession DY581614.1 - a 859 base pair clone from *Acropora millepora* (a species of branching stony coral). See below for alignment details.

Seong Tae Gwon
A12364788
sgwon@ucsd.edu



☑ B035-A4 Acropora millepora prawn chip library B Acropora millepora cDNA clone B035-A4, mRNA sequence	Acropora millepora	211	211	44%	3e-62	53.14%	859	DY581614.1

B035-A4 Acropora millepora prawn chip library B Acropora millepora cDNA clone B035-A4, mRNA sequence

Sequence ID: DY581614.1 Length: 859 Number of Matches: 1

Range 1: 223 to 843 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
211 bits(538)	3e-62	Compositional matrix adjust.	110/207(53%)	136/207(65%)	3/207(1%)	+1

```
Query  40   TPLAYKQFIPNVAEKTLGASGRYEGKISRNSERFKELTPNYNPDIIFKDEENTGADRLMT  99
            +PL    Q +P+++E + GASG   +GKI+RNS   F++L P YN   IIFKDEE TGADRLM+
Sbjct  223  SPLMLYQCVPDLSENSQGASGPAKGKITRNSPEFEKLEPCYNTAIIFKDEEGTGADRLMS  402

Query  100  QRCKDKLNALAISVMNQWPGVKLRVTEGWDEDGHHSEESLHYEGRAVDITTSDRDRS--K  157
            +RCK+KL   LA   V NQWP +KL VTE WDE G HS+ SLHYEGRAVD+   SD   +S   K
Sbjct  403  KRCKEKLIELASLVKNQWPSLKLVVTEAWDEQGQHSKNSLHYEGRAVDLRLSDTYKSNPK  582

Query  158  YGMLARLAVEAGFDWVYYESKAHIHCSVKAENSV-AAKSGGCFPGSATVHLEQGGTKLVK  216
             +L RLAV AGFD+V YESK HIH SV+ ++ V    K   GCF    +TV LE G     V
Sbjct  583  LALLGRLAVNAGFDYVLYESKTHIHASVREDSYVDKTKRTGCFSSESTVRLENGAVLRVD  762

Query  217  DLSPGDRVLAADDQGRLLYSDFLTFLD  243
             L    DRV       G + YS+ + F D
Sbjct  763  HLKISDRVQVMMQDGTIGYSEVIMFAD  843
```

Seong Tae Gwon
A12364788
sgwon@ucsd.edu

Alignment details:

B035-A4 Acropora millepora prawn chip library B Acropora millepora cDNA clone B035-A4, mRNA sequence
Sequence ID: DY581614.1Length: 859Number of Matches: 1
Range 1: 223 to 843GenBankGraphicsNext MatchPrevious Match

Score	Expect	Method	Identities	Positives	Gaps
211 bits(538)	3e-62	Compositional matrix adjust.	110/207(53%)	136/207(65%)	3/207(1%)

```
Query  40   TPLAYKQFIPNVAEKTLGASGRYEGKISRNSERFKELTPNYNPDIIFKDEENTGADRLMT  99
            +PL   Q +P+++E + GASG  +GKI+RNS  F++L P YN  IIFKDEE TGADRLM+
Sbjct  223  SPLMLYQCVPDLSENSQGASGPAKGKITRNSPEFEKLEPCYNTAIIFKDEEGTGADRLMS  402

Query  100  QRCKDKLNALAISVMNQWPGVKLRVTEGWDEDGHHSEESLHYEGRAVDITTSDRDRS--K  157
            +RCK+KL  LA  V NQWP +KL VTE WDE G HS+ SLHYEGRAVD+  SD  +S  K
Sbjct  403  KRCKEKLIELASLVKNQWPSLKLVVTEAWDEQGQHSKNSLHYEGRAVDLRLSDTYKSNPK  582

Query  158  YGMLARLAVEAGFDWVYYESKAHIHCSVKAENSV-AAKSGGCFPGSATVHLEQGGTKLVK  216
              +L RLAV AGFD+V YESK HIH SV+ ++ V   K  GCF   +TV LE G    V
Sbjct  583  LALLGRLAVNAGFDYVLYESKTHIHASVREDSYVDKTKRTGCFSSESTVRLENGAVLRVD  762

Query  217  DLSPGDRVLAADDQGRLLYSDFLTFLD  243
             L   DRV      G + YS+ + F D
Sbjct  763  HLKISDRVQVMMQDGTIGYSEVIMFAD  843
```

[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.

Chosen sequence:

```
>A. millepora protein (sequence taken from BLAST result)
SPLMLYQCVPDLSENSQGASGPAKGKITRNSPEFEKLEPCYNTAIIFKDEEGTGADRLMSKRCK
EKLIELASLVKNQWPSLKLVVTEAWDEQGQHSKNSLHYEGRAVDLRLSDTYKSNPKLALLGRLA
```

Seong Tae Gwon
A12364788
sgwon@ucsd.edu

VNAGFDYVLYESKTHIHASVREDSYVDKTKRTGCFSSESTVRLENGAVLRVDHLKISDRVQVMM QDGTIGYSEVIMFAD

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.

Name: *A. millepora* SHH

Species: *Acropora millepora*

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Cnidaria; Anthozoa; Hexacorallia; Scleractinia; Astrocoeniina; Acroporidae; Acropora

[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.

Details:

A BLASTP search against NR database (see setup in first screen-shot below) yielded a top hit result is to a protein from *Acropora millepora* (a species of branching stony coral).

See additional screen shots below for top hits and selected alignment details:

Seong Tae Gwon
A12364788
sgwon@ucsd.edu

U.S. National Library of Medicine
National Center for Biotechnology Information

BLAST ® » blastp suite

Standard Protein BLAST

blastn | blastp | blastx | tblastn | tblastx

BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

Query subrange

From

To

```
>A. millepora protein (sequence taken from BLAST result)
SPLMLYQCVPDLSENSQGASGPAKGKITRNSPEFEKLEPCYNTAIIFKDEEGT
GADRLMSKRCKEKLIEL
ASLVKNQWPSLKLVVTEAWDEQGQHSKNSLHYEGRAVDLRLSDTYKSNPKL
ALLGRLAVNAGFDYVLYES
KTHIHASVREDSYVDKTKRTGCFSSESTVRLENGAVLRVDHLKISDRVQVMM
QDGTIGYSEVIMFAD
```

Or, upload file: Choose File no file selected

Job Title: A. millepora protein (sequence taken from...

Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism (Optional): Enter organism name or id--completions will be suggested exclude Add organism

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude (Optional): Models (XM/XP) Non-redundant RefSeq proteins (WP) Uncultured/environmental sample sequences

Program Selection

Algorithm:
- Quick BLASTP (Accelerated protein-protein BLAST)
- blastp (protein-protein BLAST)
- PSI-BLAST (Position-Specific Iterated BLAST)
- PHI-BLAST (Pattern Hit Initiated BLAST)
- DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

BLAST

Search database nr using Blastp (protein-protein BLAST)

Show results in a new window

Sequences producing significant alignments

Download New Select columns Show 100

select all 100 sequences selected

GenPept Graphics Distance tree of results Multiple alignment New MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
tiggy-winkle hedgehog protein-like [Acropora millepora]	Acropora millepora	431	431	100%	3e-149	99.03%	404	XP_029200285.2
tiggy-winkle hedgehog protein-like [Acropora millepora]	Acropora millepora	430	430	100%	3e-149	99.03%	404	XP_044179040.1
PREDICTED: tiggy-winkle hedgehog protein-like [Acropora digitifera]	Acropora digitifera	430	430	100%	6e-149	99.03%	404	XP_015756004.1
indian hedgehog protein-like [Stylophora pistillata]	Stylophora pistillata	314	314	99%	2e-103	74.27%	402	XP_022799046.1
indian hedgehog protein-like [Orbicella faveolata]	Orbicella faveolata	313	313	99%	3e-103	74.76%	399	XP_020632016.1
indian hedgehog protein-like isoform X1 [Pocillopora damicornis]	Pocillopora damicornis	308	308	99%	3e-101	72.68%	403	XP_027044648.1
indian hedgehog protein-like isoform X2 [Pocillopora damicornis]	Pocillopora damicornis	306	306	98%	5e-101	72.55%	357	XP_027044649.1
hypothetical protein pdam_00015707 [Pocillopora damicornis]	Pocillopora damicornis	308	308	99%	1e-100	72.68%	430	RMX45416.1
hedgehog [Artemia franciscana]	Artemia franciscana	233	233	100%	1e-71	54.11%	421	AAP38182.1
sonic hedgehog protein A [Cryptotermes secundus]	Cryptotermes secundus	232	232	100%	2e-71	56.25%	409	XP_023707185.1

Seong Tae Gwon
A12364788
sgwon@ucsd.edu

Download GenPept Graphics ▼ Next ▲ Previous ◀Descriptions

tiggy-winkle hedgehog protein-like [Acropora millepora]

Sequence ID: XP_029200285.2 Length: 404 Number of Matches: 1

Range 1: 40 to 246 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
431 bits(1107)	3e-149	Compositional matrix adjust.	205/207(99%)	206/207(99%)	0/207(0%)

```
Query  1    SPLMLYQCVPDLSENSQGASGPAKGKITRNSPEFEKLEPCYNTAIIFKDEEGTGADRLMS  60
            SPLM YQCVPDLSENSQGASGPAKGKITRNSPEFEKLEPCYNTAIIFKDEEGTGADRLMS
Sbjct  40   SPLMQYQCVPDLSENSQGASGPAKGKITRNSPEFEKLEPCYNTAIIFKDEEGTGADRLMS  99

Query  61   KRCKEKLIELASLVKNQWPSLKLVVTEAWDEQGQHSKNSLHYEGRAVDLRLSDTYKSNPK  120
            KRCKEKLIELASLVKNQWPSLKLVVTEAWDEQGQHSKNSLHYEGRAVDLRLSDTYKSNPK
Sbjct  100  KRCKEKLIELASLVKNQWPSLKLVVTEAWDEQGQHSKNSLHYEGRAVDLRLSDTYKSNPK  159

Query  121  LALLGRLAVNAGFDYVLYESKTHIHASVREDSYVDKTKRTGCFSSESTVRLENGAVLRVD  180
            LALLGRLAVNAGFDYVLYESKTHIHASVREDSYVDKTK+TGCFSSESTVRLENGAVLRVD
Sbjct  160  LALLGRLAVNAGFDYVLYESKTHIHASVREDSYVDKTKKTGCFSSESTVRLENGAVLRVD  219

Query  181  HLKISDRVQVMMQDGTIGYSEVIMFAD  207
            HLKISDRVQVMMQDGTIGYSEVIMFAD
Sbjct  220  HLKISDRVQVMMQDGTIGYSEVIMFAD  246
```

Download GenPept Graphics ▼ Next ▲ Previous ◀Descriptions

tiggy-winkle hedgehog protein-like [Acropora millepora]

Sequence ID: XP_044179040.1 Length: 404 Number of Matches: 1

Range 1: 40 to 246 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
430 bits(1106)	3e-149	Compositional matrix adjust.	205/207(99%)	206/207(99%)	0/207(0%)

```
Query  1    SPLMLYQCVPDLSENSQGASGPAKGKITRNSPEFEKLEPCYNTAIIFKDEEGTGADRLMS  60
            SPLM YQCVPDLSENSQGASGPAKGKITRNSPEFEKLEPCYNTAIIFKDEEGTGADRLMS
Sbjct  40   SPLMQYQCVPDLSENSQGASGPAKGKITRNSPEFEKLEPCYNTAIIFKDEEGTGADRLMS  99

Query  61   KRCKEKLIELASLVKNQWPSLKLVVTEAWDEQGQHSKNSLHYEGRAVDLRLSDTYKSNPK  120
            KRCKEKLIELASLVKNQWPSLKLVVTEAWDEQGQHSKNSLHYEGRAVDLRLSDTYKSNPK
Sbjct  100  KRCKEKLIELASLVKNQWPSLKLVVTEAWDEQGQHSKNSLHYEGRAVDLRLSDTYKSNPK  159

Query  121  LALLGRLAVNAGFDYVLYESKTHIHASVREDSYVDKTKRTGCFSSESTVRLENGAVLRVD  180
            LALLGRLAVNAGFDYVLYES+THIHASVREDSYVDKTKRTGCFSSESTVRLENGAVLRVD
Sbjct  160  LALLGRLAVNAGFDYVLYESETHIHASVREDSYVDKTKRTGCFSSESTVRLENGAVLRVD  219

Query  181  HLKISDRVQVMMQDGTIGYSEVIMFAD  207
            HLKISDRVQVMMQDGTIGYSEVIMFAD
Sbjct  220  HLKISDRVQVMMQDGTIGYSEVIMFAD  246
```