LAB 1a — EXPLORING NCBI

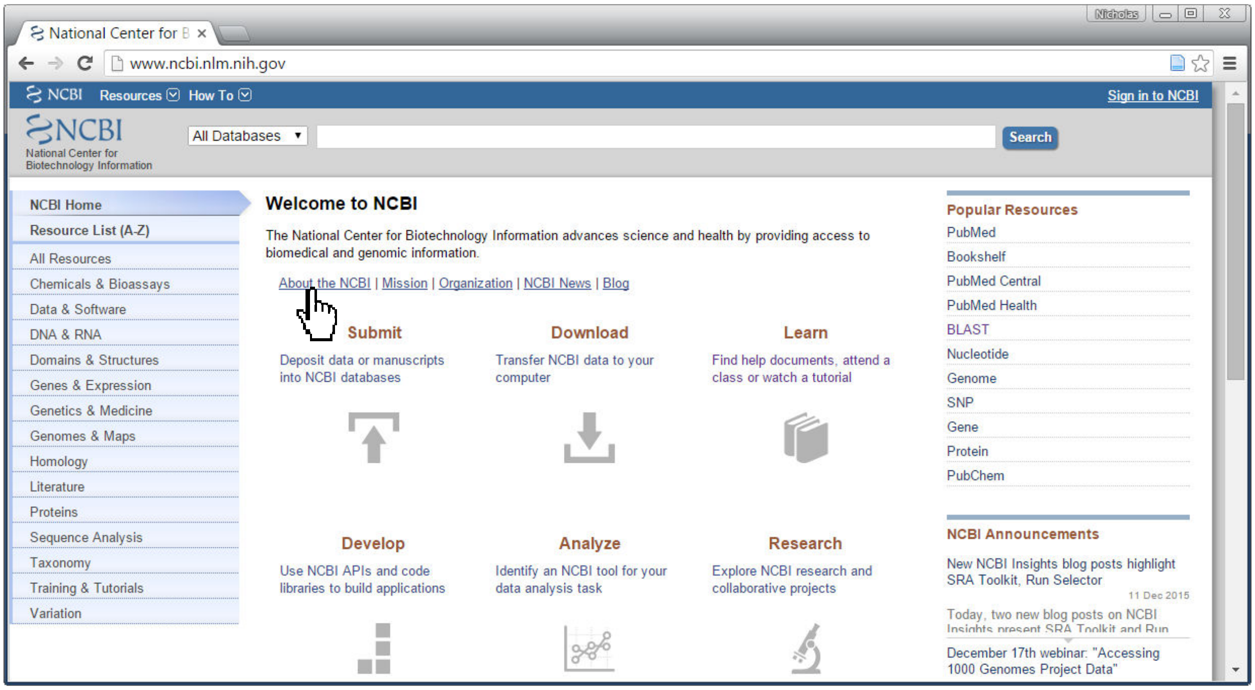
[Software needed: web access]

The National Center for Biotechnology Information (NCBI) maintained by the US National Library of Medicine and National Institutes of Health is one of the world's most important resources and repositories for biological data. This fantastic online resource provides an extensive network of databases cataloging an ever-growing wealth of genetic, medical, and biochemical information from all walks and crawls of life. Entire genomes, from viruses to humans, are compiled, organized, and cross-referenced within these networks, such that surfing the genome can be almost as easy as surfing the web.

But you have to know a) what you're looking *for*, and b) what you're looking *at* to get anything out of these databases. This is what this first lab is going to help you do. Note that Google and other search engines typically do not index database-driven websites, which is why it cannot be used for searching for information that is stored at NCBI.

The primary portal for accessing data at NCBI is called *GQuery*. But first, let's start by visiting NCBI's website and examining the interface, which undergoes constant change.

1. Open your Web browser and go to NCBI's homepage: www.ncbi.nlm.nih.gov. This page provides links to all of NCBI databases and resources. It's worth exploring here just to get a better idea of the scope of NCBI. If you click **About the NCBI** you will be taken to a page summarizing some of these resources. You can also check out the *NCBI handbook* (http://www.ncbi.nlm.nih.gov/books/NBK21101/) for more information.



The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

Deposit data or manuscripts into NCBI databases

Transfer NCBI data to your computer

Find help documents, attend a class or watch a tutorial

Use NCBI APIs and code libraries to build applications

Identify an NCBI tool for your data analysis task

Explore NCBI research and collaborative projects

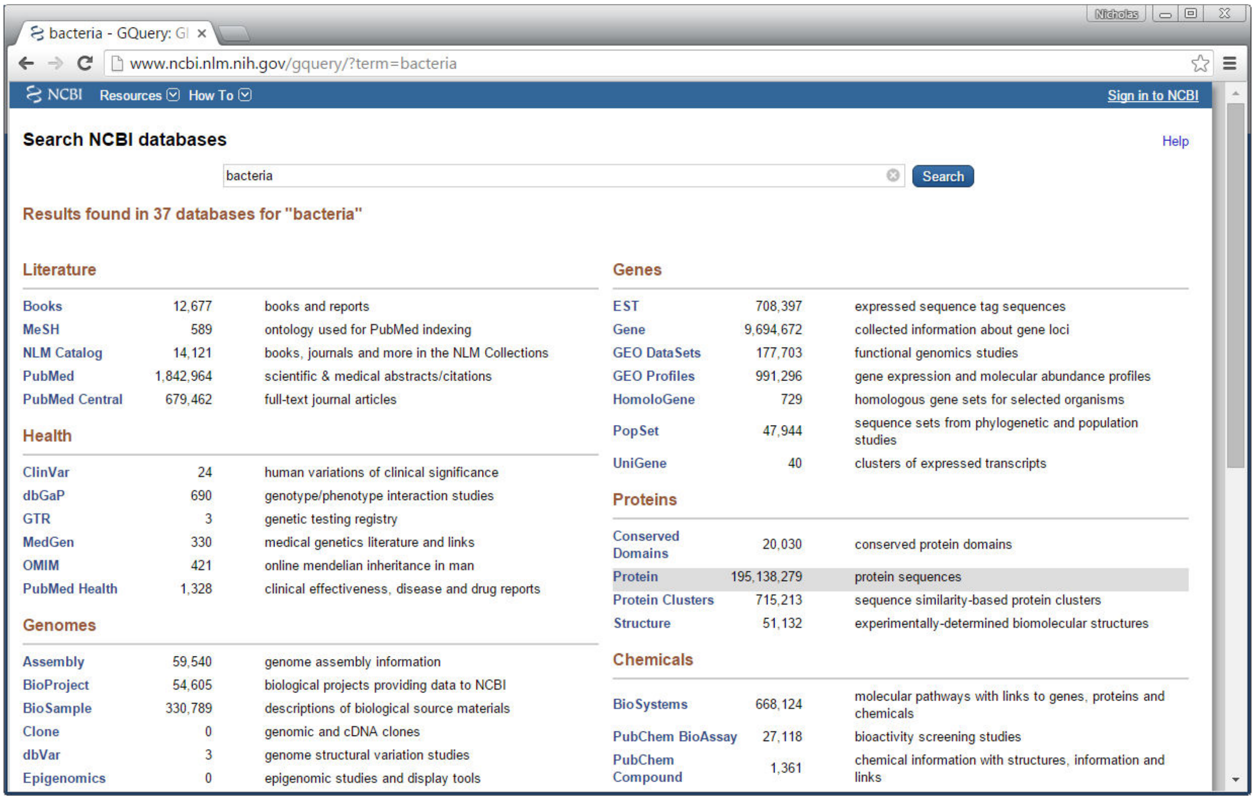
New NCBI Insights blog posts highlight SRA Toolkit, Run Selector

Today, two new blog posts on NCBI Insights present SRA Toolkit and Run

December 17th webinar: "Accessing 1000 Genomes Project Data"

Figure 1. The NCBI homepage.

2. Now let's move to the *Search NCBI Databases* (also known as *GQuery*) portal – select **All Databases** from the navigation bar at the top of the NCBI start page, by clicking "Search" on the empty field. First, scan down the assortment of databases queried through this portal. You will notice there is everything from the biomedical literature at PubMed to nucleotide databases, taxonomy databases, protein structure databases, and expression profile databases. Let's see what happens when you do an unguided search on the site. In the "Search across databases" box, type in ***bacteria***. The output is a summary page of the number of hits in each section. A search of ***bacteria*** gives millions of hits – not very helpful. We need specifics.



Search NCBI databases

Results found in 37 databases for "bacteria"

Literature

Books	12,677	books and reports
MeSH	589	ontology used for PubMed indexing
NLM Catalog	14,121	books, journals and more in the NLM Collections
PubMed	1,842,964	scientific & medical abstracts/citations
PubMed Central	679,462	full-text journal articles

Health

ClinVar	24	human variations of clinical significance
dbGaP	690	genotype/phenotype interaction studies
GTR	3	genetic testing registry
MedGen	330	medical genetics literature and links
OMIM	421	online mendelian inheritance in man
PubMed Health	1,328	clinical effectiveness, disease and drug reports

Genomes

Assembly	59,540	genome assembly information
BioProject	54,605	biological projects providing data to NCBI
BioSample	330,789	descriptions of biological source materials
Clone	0	genomic and cDNA clones
dbVar	3	genome structural variation studies
Epigenomics	0	epigenomic studies and display tools

Genes

EST	708,397	expressed sequence tag sequences
Gene	9,694,672	collected information about gene loci
GEO DataSets	177,703	functional genomics studies
GEO Profiles	991,296	gene expression and molecular abundance profiles
HomoloGene	729	homologous gene sets for selected organisms
PopSet	47,944	sequence sets from phylogenetic and population studies
UniGene	40	clusters of expressed transcripts

Proteins

Conserved Domains	20,030	conserved protein domains
Protein	195,138,279	protein sequences
Protein Clusters	715,213	sequence similarity-based protein clusters
Structure	51,132	experimentally-determined biomolecular structures

Chemicals

BioSystems	668,124	molecular pathways with links to genes, proteins and chemicals
PubChem BioAssay	27,118	bioactivity screening studies
PubChem Compound	1,361	chemical information with structures, information and links

Figure 2. The *Search NCBI Databases* portal page with ***bacteria*** used as a search word.

3. Usually when searching these databases, you have either a region of DNA or a protein (or protein function) of interest. For this lab you'll be using a gene from *Arabidopsis thaliana*, a small flowering plant that is like the fruit fly of the plant world as it has a comparatively rapid life cycle and requires little space to grow. The protein product of this gene is recorded under accession number NP_001318308, and it is an E3 ligase, involved in ubiquitination of proteins, which is a signal for their degradation.

4. Go back to the NCBI *GQuery* portal page and try a more focused search. Use the search terms found associated with the gene sequence we'll be using with the GenBank Field Qualifiers shown below (a full list of qualifiers is presented in Appendix 1). Try the four different searches presented below and look at the number records, specifically "Protein" records, found:

   - gene keywords
     e.g. ***ubiquitin-protein ligase***
   - gene keyword AND organism
     e.g. ***ubiquitin-protein ligase AND Arabidopsis thaliana***
   - gene keyword [PROT] AND organism [ORGN]
     e.g. ***ubiquitin-protein ligase [PROT] AND Arabidopsis thaliana [ORGN]***
   - accession or GI number
     e.g. ***NP_001318308***

   Lab Quiz Question 1
   *Answer lab quiz questions while doing lab!

   That narrowed things down significantly!

   Note that using parentheses can be very helpful in making sure you get exactly what you want. For example:

   - ***SMC AND (yeast [ORGN] OR Arabidopsis [ORGN])***

   is a very different search than

   ➢ ***SMC AND yeast [ORGN] OR Arabidopsis [ORGN]***

   Also, using quotation marks can also dramatically affect your search (ie: 16s rRNA vs. "16s rRNA").

   Finally, always capitalize the Boolean operators such as AND / OR / NOT. Ultimately, the most specific search items you can use are accession numbers.

Box 1. Accession Numbers, Version Numbers, and GI Numbers

An **Accession number** is a unique identifier for a particular sequence record. An accession number is assigned to a specific record and stays with that record forever. In other words, Accession numbers track a particular record and do not change even if the information in the record is changed at the author's request (e.g. if a better annotation or more complete sequence is provided). Accession numbers are usually a combination of a letter(s) and numbers, such as a single letter followed by five digits (e.g., U12345) or two letters followed by six digits (e.g., AF123456).

Version numbers follow the Accession number and indicate the revision history of that entry starting with 1 and increasing with each revision. The standard format is *Accession.Version*.

A **GI number** (GenInfo Identifier – sometimes written in lower case, "gi") was simply a series of digits that was, until recently, assigned consecutively to each sequence record processed by NCBI. The GI system of identifiers ran in parallel to the Accession.Version system; therefore, if the DNA or protein sequence changed in any way, it would receive a new GI number.

Example: When a new entry was submitted to GenBank it was assigned an accession number (say AF000001). Since this is the first version the Accession would be appended with '.1', so it would look like AF000001.1. At the same time was given a GI number (say GI:1234567). Now imagine that the researcher who originally submitted the record wanted to update the information. The updated record would keep the same Accession number, but would increase in version number (AF000001.2). The new record would have been given a completely new GI number (say GI:9876543).

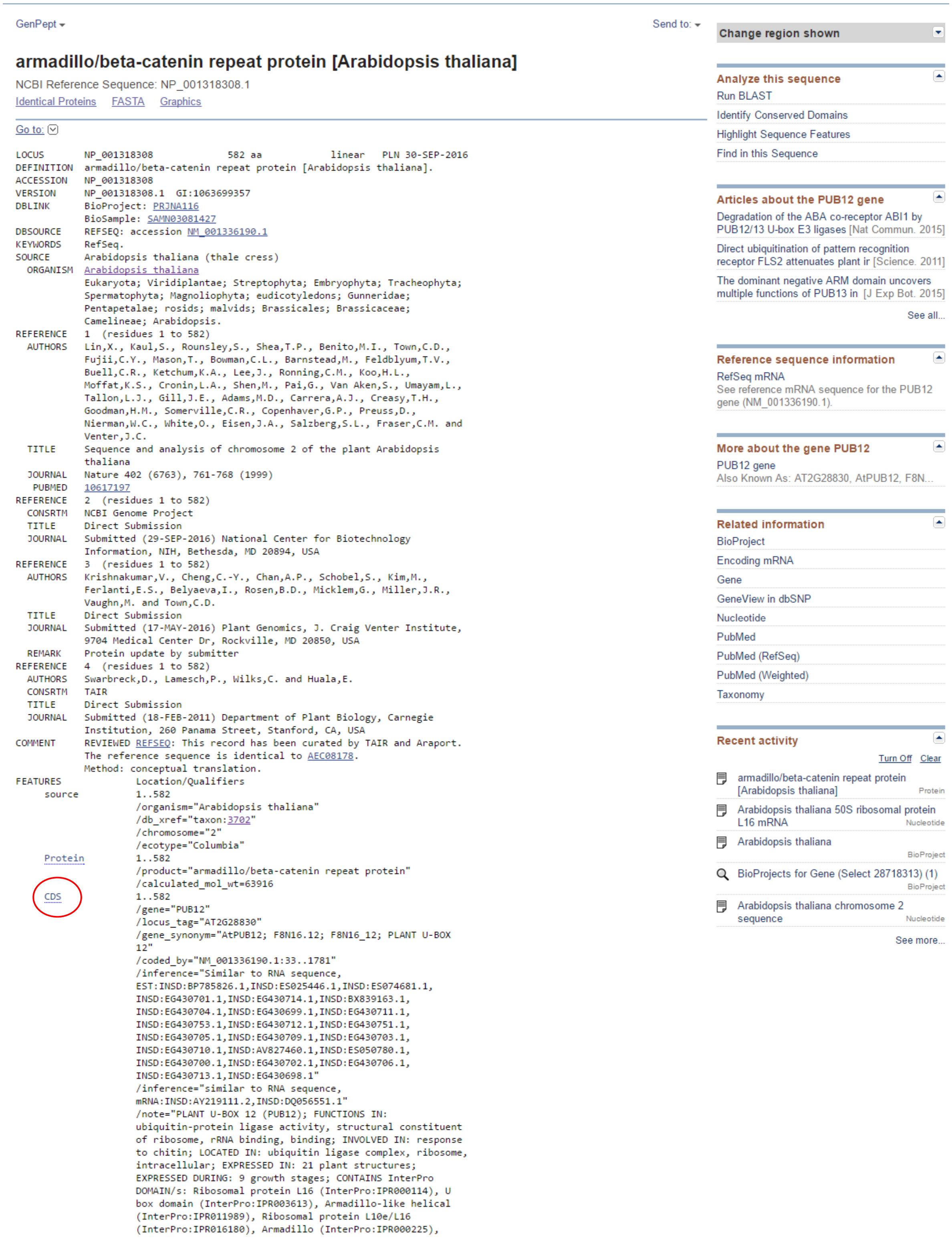
Why is this important? The *Accession number* will always give you the most up-to-date information on a record, while the *GI number* and/or *Accession.Version* will always take you to a specific record. There are times when you want the most current information, and other times when you want to point to a particular piece of information from a particular point in time (e.g. a particular record that you did an analysis with), even if more information has been subsequently added. Note that as of September 2016, NCBI started phasing out the use of GI numbers. The use of *Accession.Version* form is now recommended to access a particular record, instead of the GI number.

Box 2. NCBI Help

This is a good time to get familiar with NCBI's thorough **Help** index for future reference. With this index, you should be able to access most of the background you need for understanding how these databases work on your own (there's also an NCBI YouTube channel, if you're so inclined to acquire your information that way).

1. To the right of the search text box on the *GQuery* portal page is the **Help** icon. Click on it.
2. You are now in Entrez Help. The Entrez collection of databases is queried when you use the GQuery interface. Note the section in the right sidebar that explains everything from search options to saving sets of records.
3. Notice that under the section **Using the Advanced Search Page to Construct Complex Search Statements** some other appropriate qualifiers are given.

5. Search for our accession number of interest (e.g. NP_001318308 from above) through the *GQuery* portal page. It should give you one protein sequence hit. Click on it (it is a hyperlink) so that you get its full GenBank description.



armadillo/beta-catenin repeat protein [Arabidopsis thaliana]

NCBI Reference Sequence: NP_001318308.1

```
LOCUS       NP_001318308             582 aa            linear   PLN 30-SEP-2016
DEFINITION  armadillo/beta-catenin repeat protein [Arabidopsis thaliana].
ACCESSION   NP_001318308
VERSION     NP_001318308.1  GI:1063699357
DBLINK      BioProject: PRJNA116
            BioSample: SAMN03081427
DBSOURCE    REFSEQ: accession NM_001336190.1
KEYWORDS    RefSeq.
SOURCE      Arabidopsis thaliana (thale cress)
  ORGANISM  Arabidopsis thaliana
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; eudicotyledons; Gunneridae;
            Pentapetalae; rosids; malvids; Brassicales; Brassicaceae;
            Camelineae; Arabidopsis.
REFERENCE   1  (residues 1 to 582)
  AUTHORS   Lin,X., Kaul,S., Rounsley,S., Shea,T.P., Benito,M.I., Town,C.D.,
            Fujii,C.Y., Mason,T., Bowman,C.L., Barnstead,M., Feldblyum,T.V.,
            Buell,C.R., Ketchum,K.A., Lee,J., Ronning,C.M., Koo,H.L.,
            Moffat,K.S., Cronin,L.A., Shen,M., Pai,G., Van Aken,S., Umayam,L.,
            Tallon,L.J., Gill,J.E., Adams,M.D., Carrera,A.J., Creasy,T.H.,
            Goodman,H.M., Somerville,C.R., Copenhaver,G.P., Preuss,D.,
            Nierman,W.C., White,O., Eisen,J.A., Salzberg,S.L., Fraser,C.M. and
            Venter,J.C.
  TITLE     Sequence and analysis of chromosome 2 of the plant Arabidopsis
            thaliana
  JOURNAL   Nature 402 (6763), 761-768 (1999)
   PUBMED   10617197
REFERENCE   2  (residues 1 to 582)
  CONSRTM   NCBI Genome Project
  TITLE     Direct Submission
  JOURNAL   Submitted (29-SEP-2016) National Center for Biotechnology
            Information, NIH, Bethesda, MD 20894, USA
REFERENCE   3  (residues 1 to 582)
  AUTHORS   Krishnakumar,V., Cheng,C.-Y., Chan,A.P., Schobel,S., Kim,M.,
            Ferlanti,E.S., Belyaeva,I., Rosen,B.D., Micklem,G., Miller,J.R.,
            Vaughn,M. and Town,C.D.
  TITLE     Direct Submission
  JOURNAL   Submitted (17-MAY-2016) Plant Genomics, J. Craig Venter Institute,
            9704 Medical Center Dr, Rockville, MD 20850, USA
  REMARK    Protein update by submitter
REFERENCE   4  (residues 1 to 582)
  AUTHORS   Swarbreck,D., Lamesch,P., Wilks,C. and Huala,E.
  CONSRTM   TAIR
  TITLE     Direct Submission
  JOURNAL   Submitted (18-FEB-2011) Department of Plant Biology, Carnegie
            Institution, 260 Panama Street, Stanford, CA, USA
COMMENT     REVIEWED REFSEQ: This record has been curated by TAIR and Araport.
            The reference sequence is identical to AEC08178.
            Method: conceptual translation.
FEATURES             Location/Qualifiers
     source          1..582
                     /organism="Arabidopsis thaliana"
                     /db_xref="taxon:3702"
                     /chromosome="2"
                     /ecotype="Columbia"
     Protein         1..582
                     /product="armadillo/beta-catenin repeat protein"
                     /calculated_mol_wt=63916
     CDS             1..582
                     /gene="PUB12"
                     /locus_tag="AT2G28830"
                     /gene_synonym="AtPUB12; F8N16.12; F8N16_12; PLANT U-BOX
                     12"
                     /coded_by="NM_001336190.1:33..1781"
                     /inference="Similar to RNA sequence,
                     EST:INSD:BP785826.1,INSD:ES025446.1,INSD:ES074681.1,
                     INSD:EG430701.1,INSD:EG430714.1,INSD:BX839163.1,
                     INSD:EG430704.1,INSD:EG430699.1,INSD:EG430711.1,
                     INSD:EG430753.1,INSD:EG430712.1,INSD:EG430751.1,
                     INSD:EG430705.1,INSD:EG430709.1,INSD:EG430703.1,
                     INSD:EG430710.1,INSD:AV827460.1,INSD:ES050780.1,
                     INSD:EG430700.1,INSD:EG430702.1,INSD:EG430706.1,
                     INSD:EG430713.1,INSD:EG430698.1"
                     /inference="similar to RNA sequence,
                     mRNA:INSD:AY219111.2,INSD:DQ056551.1"
                     /note="PLANT U-BOX 12 (PUB12); FUNCTIONS IN:
                     ubiquitin-protein ligase activity, structural constituent
                     of ribosome, rRNA binding, binding; INVOLVED IN: response
                     to chitin; LOCATED IN: ubiquitin ligase complex, ribosome,
                     intracellular; EXPRESSED IN: 21 plant structures;
                     EXPRESSED DURING: 9 growth stages; CONTAINS InterPro
                     DOMAIN/s: Ribosomal protein L16 (InterPro:IPR000114), U
                     box domain (InterPro:IPR003613), Armadillo-like helical
                     (InterPro:IPR011989), Ribosomal protein L10e/L16
                     (InterPro:IPR016180), Armadillo (InterPro:IPR000225),
```

Analyze this sequence

Run BLAST
Identify Conserved Domains
Highlight Sequence Features
Find in this Sequence

Articles about the PUB12 gene

Degradation of the ABA co-receptor ABI1 by PUB12/13 U-box E3 ligases [Nat Commun. 2015]

Direct ubiquitination of pattern recognition receptor FLS2 attenuates plant ir [Science. 2011]

The dominant negative ARM domain uncovers multiple functions of PUB13 in [J Exp Bot. 2015]

See all...

Reference sequence information

RefSeq mRNA
See reference mRNA sequence for the PUB12 gene (NM_001336190.1).

More about the gene PUB12

PUB12 gene
Also Known As: AT2G28830, AtPUB12, F8N...

Related information

BioProject
Encoding mRNA
Gene
GeneView in dbSNP
Nucleotide
PubMed
PubMed (RefSeq)
PubMed (Weighted)
Taxonomy

Recent activity

armadillo/beta-catenin repeat protein [Arabidopsis thaliana] Protein
Arabidopsis thaliana 50S ribosomal protein L16 mRNA Nucleotide
Arabidopsis thaliana BioProject
BioProjects for Gene (Select 28718313) (1) BioProject
Arabidopsis thaliana chromosome 2 sequence Nucleotide

See more...

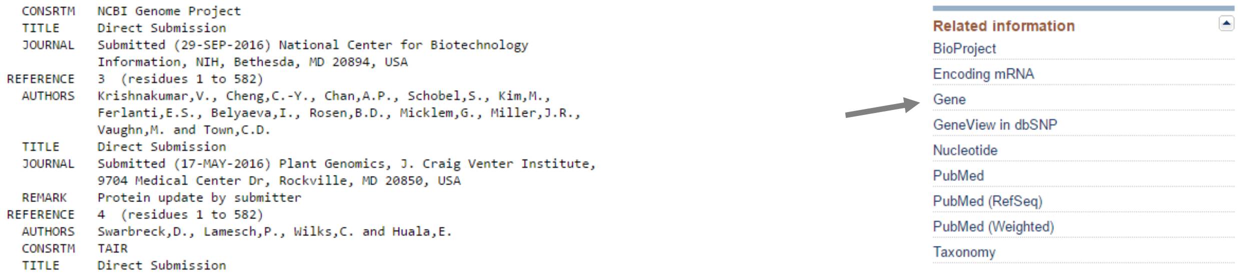
Figure 3. GenBank record for accession NP_001318308, in GenPept format.

6. Notice all the hyperlinks within the text. It looks messy, but is in fact straightforward. For example, for taxonomic information, click on the **SOURCE ORGANISM** hyperlink. Some records have links to the primary publication where this sequence was originally cited in a **PUBMED** number hyperlink (not the case in the above example, but there is a PubMed reference for the sequence). Click around on different links and see what you find.
    a. *What is the taxonomic lineage of your organism?*
    b. *Has the genome of this organism been sequenced, i.e. is there a Genome Project?*
    c. *If so, can you find the accession for the full sequence or one of the chromosomes?*

➢ **To find out much more information on the structure of the GenBank file at http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html**

7. Go back to the GenBank record and click on the **CDS** link, just above the actual sequence (circled in red in Figure 3 on the previous page).
    a. *Where did this take you or what happened when you did this?*

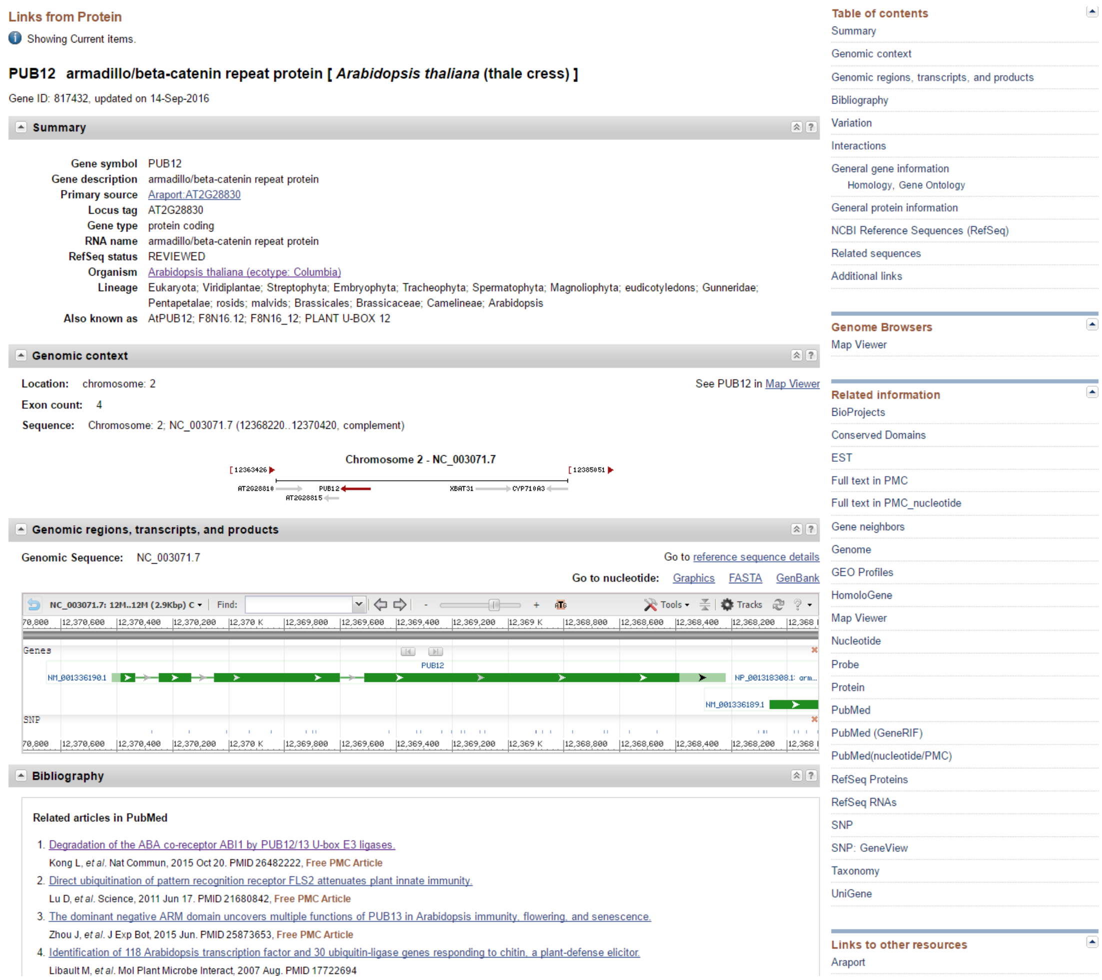
8. Go back to the GenBank record and examine the **Related Information** section on the lower right. This gives you direct links to other databases with information on this query. Find the **Gene** link.



```
  CONSRTM   NCBI Genome Project
  TITLE     Direct Submission
  JOURNAL   Submitted (29-SEP-2016) National Center for Biotechnology
            Information, NIH, Bethesda, MD 20894, USA
REFERENCE   3  (residues 1 to 582)
  AUTHORS   Krishnakumar,V., Cheng,C.-Y., Chan,A.P., Schobel,S., Kim,M.,
            Ferlanti,E.S., Belyaeva,I., Rosen,B.D., Micklem,G., Miller,J.R.,
            Vaughn,M. and Town,C.D.
  TITLE     Direct Submission
  JOURNAL   Submitted (17-MAY-2016) Plant Genomics, J. Craig Venter Institute,
            9704 Medical Center Dr, Rockville, MD 20850, USA
  REMARK    Protein update by submitter
REFERENCE   4  (residues 1 to 582)
  AUTHORS   Swarbreck,D., Lamesch,P., Wilks,C. and Huala,E.
  CONSRTM   TAIR
  TITLE     Direct Submission
```

Figure 4. The **Related Information** menu for NP_001318308, to the right of the record. The arrow is pointing to the “Gene” link.

9. Select **Gene** from the **Related Information** menu. This is a great starter resource at NCBI. Scroll through the different sections. Use them to answer the following questions.
    a. *Where is your gene’s location in the genome? (Tip: hover with your cursor over the green bars in the “Genomic regions, transcripts, and products” section; the green bars represent the gene in the sequence viewer)*
    b. *How many exons do you see in this gene? Tip: how many green boxes are there?*
    c. *What are the names of the genes surrounding it (i.e. what is its “Genomic context”)?*
    d. *Does it have any conserved domains? What are they called? (Tip: use the “Related Information” link to Conserved Domains on the right of the **Gene** page)*
    e. *After exploring conserved domains go back to the **Gene** page. What biological process (Gene Ontology terms) is this gene involved with (scroll down!)?*



Links from Protein

Showing Current items.

PUB12 armadillo/beta-catenin repeat protein [*Arabidopsis thaliana* (thale cress)]

Gene ID: 817432, updated on 14-Sep-2016

Summary

Gene symbol	PUB12
Gene description	armadillo/beta-catenin repeat protein
Primary source	Araport:AT2G28830
Locus tag	AT2G28830
Gene type	protein coding
RNA name	armadillo/beta-catenin repeat protein
RefSeq status	REVIEWED
Organism	Arabidopsis thaliana (ecotype: Columbia)
Lineage	Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons; Gunneridae; Pentapetalae; rosids; malvids; Brassicales; Brassicaceae; Camelineae; Arabidopsis
Also known as	AtPUB12; F8N16.12; F8N16_12; PLANT U-BOX 12

Genomic context

Location: chromosome: 2

See PUB12 in Map Viewer

Exon count: 4

Sequence: Chromosome: 2; NC_003071.7 (12368220..12370420, complement)

Genomic regions, transcripts, and products

Genomic Sequence: NC_003071.7

Go to reference sequence details

Go to nucleotide: Graphics FASTA GenBank

Bibliography

Related articles in PubMed

1. Degradation of the ABA co-receptor ABI1 by PUB12/13 U-box E3 ligases.
   Kong L, *et al.* Nat Commun, 2015 Oct 20. PMID 26482222, Free PMC Article
2. Direct ubiquitination of pattern recognition receptor FLS2 attenuates plant innate immunity.
   Lu D, *et al.* Science, 2011 Jun 17. PMID 21680842, Free PMC Article
3. The dominant negative ARM domain uncovers multiple functions of PUB13 in Arabidopsis immunity, flowering, and senescence.
   Zhou J, *et al.* J Exp Bot, 2015 Jun. PMID 25873653, Free PMC Article
4. Identification of 118 Arabidopsis transcription factor and 30 ubiquitin-ligase genes responding to chitin, a plant-defense elicitor.
   Libault M, *et al.* Mol Plant Microbe Interact, 2007 Aug. PMID 17722694

Table of contents

- Summary
- Genomic context
- Genomic regions, transcripts, and products
- Bibliography
- Variation
- Interactions
- General gene information
  - Homology, Gene Ontology
- General protein information
- NCBI Reference Sequences (RefSeq)
- Related sequences
- Additional links

Genome Browsers

- Map Viewer

Related information

- BioProjects
- Conserved Domains
- EST
- Full text in PMC
- Full text in PMC_nucleotide
- Gene neighbors
- Genome
- GEO Profiles
- HomoloGene
- Map Viewer
- Nucleotide
- Probe
- Protein
- PubMed
- PubMed (GeneRIF)
- PubMed(nucleotide/PMC)
- RefSeq Proteins
- RefSeq RNAs
- SNP
- SNP: GeneView
- Taxonomy
- UniGene

Links to other resources

- Araport

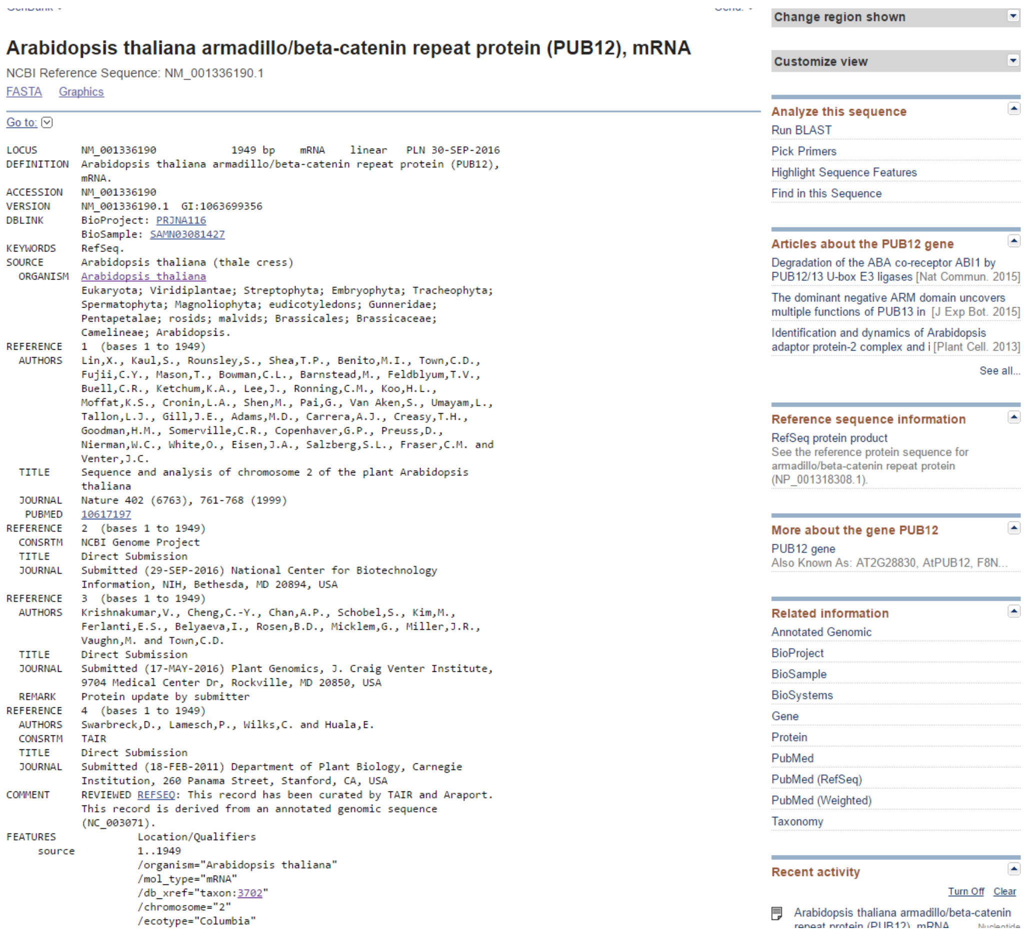
Figure 5. GenBank **Gene** page for At2g28830 (also known as PUB12), the gene that encodes NP_001318308.

10. On the Gene page, there are also **Additional links** to examine a gene's structure, function and phylogenetic relationships further. The navigation sidebar on the right has an "Additional links" hyperlink which will take you to the bottom of the page, where they're found for most genes. Click [+] Gene LinkOut to see them.

    a. Click on Additional Links. What kind of information is in this section?

➢ Click around and explore the variety of ways that data for PUB12 are interconnected and displayed (don't worry, you can't break anything). Using the **Related Information** links can you find any publications associated with this gene? What about gene expression data? The next page shows the related "RefSeq" record for the corresponding mRNA (NCBI's RefSeq aims to provide canonical "reference" sequences – genomic, mRNA, CDS, protein etc. – for many model organisms).

b. Why is the length of the mRNA different from the value you can calculate from the start and stop positions in Question 9a?



Arabidopsis thaliana armadillo/beta-catenin repeat protein (PUB12), mRNA

NCBI Reference Sequence: NM_001336190.1

```
LOCUS       NM_001336190            1949 bp    mRNA    linear   PLN 30-SEP-2016
DEFINITION  Arabidopsis thaliana armadillo/beta-catenin repeat protein (PUB12),
            mRNA.
ACCESSION   NM_001336190
VERSION     NM_001336190.1  GI:1063699356
DBLINK      BioProject: PRJNA116
            BioSample: SAMN03081427
KEYWORDS    RefSeq.
SOURCE      Arabidopsis thaliana (thale cress)
  ORGANISM  Arabidopsis thaliana
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; eudicotyledons; Gunneridae;
            Pentapetalae; rosids; malvids; Brassicales; Brassicaceae;
            Camelineae; Arabidopsis.
REFERENCE   1  (bases 1 to 1949)
  AUTHORS   Lin,X., Kaul,S., Rounsley,S., Shea,T.P., Benito,M.I., Town,C.D.,
            Fujii,C.Y., Mason,T., Bowman,C.L., Barnstead,M., Feldblyum,T.V.,
            Buell,C.R., Ketchum,K.A., Lee,J., Ronning,C.M., Koo,H.L.,
            Moffat,K.S., Cronin,L.A., Shen,M., Pai,G., Van Aken,S., Umayam,L.,
            Tallon,L.J., Gill,J.E., Adams,M.D., Carrera,A.J., Creasy,T.H.,
            Goodman,H.M., Somerville,C.R., Copenhaver,G.P., Preuss,D.,
            Nierman,W.C., White,O., Eisen,J.A., Salzberg,S.L., Fraser,C.M. and
            Venter,J.C.
  TITLE     Sequence and analysis of chromosome 2 of the plant Arabidopsis
            thaliana
  JOURNAL   Nature 402 (6763), 761-768 (1999)
   PUBMED   10617197
REFERENCE   2  (bases 1 to 1949)
  CONSRTM   NCBI Genome Project
  TITLE     Direct Submission
  JOURNAL   Submitted (29-SEP-2016) National Center for Biotechnology
            Information, NIH, Bethesda, MD 20894, USA
REFERENCE   3  (bases 1 to 1949)
  AUTHORS   Krishnakumar,V., Cheng,C.-Y., Chan,A.P., Schobel,S., Kim,M.,
            Ferlanti,E.S., Belyaeva,I., Rosen,B.D., Micklem,G., Miller,J.R.,
            Vaughn,M. and Town,C.D.
  TITLE     Direct Submission
  JOURNAL   Submitted (17-MAY-2016) Plant Genomics, J. Craig Venter Institute,
            9704 Medical Center Dr, Rockville, MD 20850, USA
  REMARK    Protein update by submitter
REFERENCE   4  (bases 1 to 1949)
  AUTHORS   Swarbreck,D., Lamesch,P., Wilks,C. and Huala,E.
  CONSRTM   TAIR
  TITLE     Direct Submission
  JOURNAL   Submitted (18-FEB-2011) Department of Plant Biology, Carnegie
            Institution, 260 Panama Street, Stanford, CA, USA
COMMENT     REVIEWED REFSEQ: This record has been curated by TAIR and Araport.
            This record is derived from an annotated genomic sequence
            (NC_003071).
FEATURES             Location/Qualifiers
     source          1..1949
                     /organism="Arabidopsis thaliana"
                     /mol_type="mRNA"
                     /db_xref="taxon:3702"
                     /chromosome="2"
                     /ecotype="Columbia"
```

Analyze this sequence

- Run BLAST
- Pick Primers
- Highlight Sequence Features
- Find in this Sequence

Articles about the PUB12 gene

- Degradation of the ABA co-receptor ABI1 by PUB12/13 U-box E3 ligases [Nat Commun. 2015]
- The dominant negative ARM domain uncovers multiple functions of PUB13 in [J Exp Bot. 2015]
- Identification and dynamics of Arabidopsis adaptor protein-2 complex and i [Plant Cell. 2013]

See all...

Reference sequence information

RefSeq protein product
See the reference protein sequence for armadillo/beta-catenin repeat protein (NP_001318308.1).

More about the gene PUB12

PUB12 gene
Also Known As: AT2G28830, AtPUB12, F8N...

Related information

- Annotated Genomic
- BioProject
- BioSample
- BioSystems
- Gene
- Protein
- PubMed
- PubMed (RefSeq)
- PubMed (Weighted)
- Taxonomy

Recent activity

Arabidopsis thaliana armadillo/beta-catenin repeat protein (PUB12), mRNA

Figure 6. RefSeq RNA linked from **Gene** page for At2g28830.

Box 3. Helpful Hints for NCBI searches

On most NCBI search pages (except, oddly, GQuery) click on "Save Search" below the search box. Register for an account and save your search. You can also combine previous searches using the **History** tab and the search numbers listed within it, as well as save your searches by registering for a *My NCBI* account, so you don't have to keep redoing the same searches in the future.

Lab 1b — Basic BLAST (*blastn*)

One of the most important bioinformatic strategies used for the functional annotation of genes and genomes is to predict the function of uncharacterized genes or proteins based on their similarity to sequences with better functional annotations. BLAST is perhaps the single most important tool for finding database sequences that are similar to a query sequence of interest.

Box 4. BLAST and Homology

The Basic Local Alignment and Search Tool (BLAST) is a very powerful approach to identifying database sequences that share local similarity to a query sequence (see below for definitions). There is a very important chain of assumptions used in biological research that is generally followed when using BLAST:

- Homologous genes share sequence similarity
  - Orthologous genes have the highest similarity among multiple species
    - Orthologous genes most likely have similar functions
      - Consequently, sequences that are most similar between multiple species share similar functions

Note, it is very important to understand that these are only assumptions, and there are many reasons and instances where these assumptions prove to be false. Nevertheless, they are a reasonable starting place.

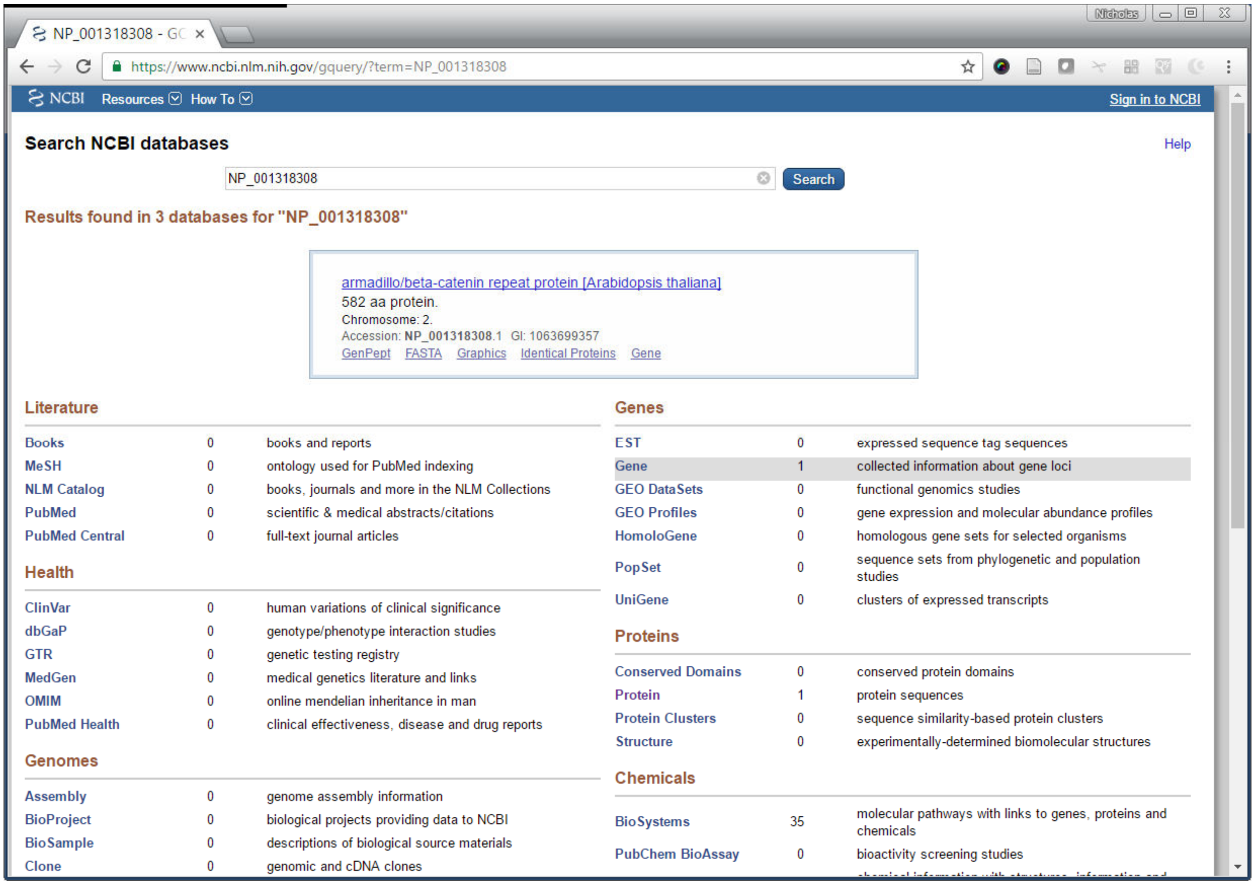
Definitions:

- **Similar sequences** – sequences that share a significant number of residues (nucleotides or amino acids). Sequences can be similar due to homology or simply by chance. The higher the similarity between sequences, the more likely they are to be homologous.
- **Homologous sequences** – sequences that are related through common ancestry. Homology is qualitative – two sequences either are, or are not related through common ancestry. Homologous sequences can vary greatly in their level of *similarity* – from 100% to 0%.
- **Orthologous sequences** – sequences that are related through a past speciation event. Orthologous sequences are assumed to share common functions.
- **Paralogous sequences** – sequences that are related through a past gene duplication event. Genes often diverge in function after duplicating; therefore, paralogous sequences are not assumed to share a common function.
- **Query sequence** – your sequence; the sequence you are interested in finding more about.
- **High Scoring Segment Pair** (HSP) – 'hits' to the database. A subsequence match between your query sequence and a database sequence returned by BLAST.
- **Local alignment** – a sequence alignment that extends only across part of the sequence.
- **Global alignment** – a sequence alignment that extends across the entire sequence (from end to end).

1. First, we need a query sequence for the search. Let's start with our given gene again, but this time we'll the nucleotide sequence corresponding to the protein sequence, not the protein sequence. First try finding the gene's DNA sequence using GQuery again.

   - On the *Search NCBI Databases (GQuery)* Portal (All Databases) page, search for your given protein sequence again using the Accession number. Using the protein from the first part of this lab, we would search for ***NP_001318308***.

- The first page that comes up is the summary page. Once you're on this page you can move to the database of interest. In this case you probably don't have hits in too many databases since you had a very specific search.



Search NCBI databases

NP_001318308

Results found in 3 databases for "NP_001318308"

armadillo/beta-catenin repeat protein [Arabidopsis thaliana]
582 aa protein.
Chromosome: 2.
Accession: **NP_001318308.1** GI: 1063699357
GenPept FASTA Graphics Identical Proteins Gene

Literature

Books	0	books and reports
MeSH	0	ontology used for PubMed indexing
NLM Catalog	0	books, journals and more in the NLM Collections
PubMed	0	scientific & medical abstracts/citations
PubMed Central	0	full-text journal articles

Health

ClinVar	0	human variations of clinical significance
dbGaP	0	genotype/phenotype interaction studies
GTR	0	genetic testing registry
MedGen	0	medical genetics literature and links
OMIM	0	online mendelian inheritance in man
PubMed Health	0	clinical effectiveness, disease and drug reports

Genomes

Assembly	0	genome assembly information
BioProject	0	biological projects providing data to NCBI
BioSample	0	descriptions of biological source materials
Clone	0	genomic and cDNA clones

Genes

EST	0	expressed sequence tag sequences
Gene	1	collected information about gene loci
GEO DataSets	0	functional genomics studies
GEO Profiles	0	gene expression and molecular abundance profiles
HomoloGene	0	homologous gene sets for selected organisms
PopSet	0	sequence sets from phylogenetic and population studies
UniGene	0	clusters of expressed transcripts

Proteins

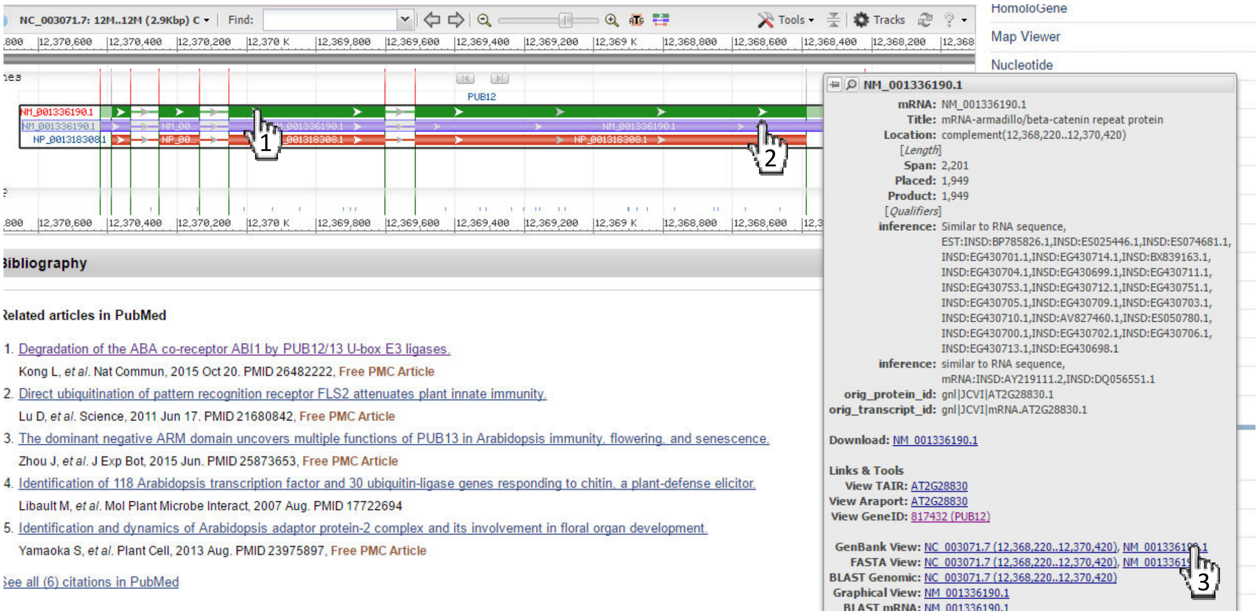
Conserved Domains	0	conserved protein domains
Protein	1	protein sequences
Protein Clusters	0	sequence similarity-based protein clusters
Structure	0	experimentally-determined biomolecular structures

Chemicals

BioSystems	35	molecular pathways with links to genes, proteins and chemicals
PubChem BioAssay	0	bioactivity screening studies

Figure 7. *GQuery* portal queried for NP_001318308 (partial view), with Gene results highlighted (numbers of results may differ slightly depending on when you're accessing NCBI).

- Try clicking the **Gene** link. Does the Gene page give you the gene sequence alone? What do you get instead? Note the context specific link menus that pop up when you hover over the graphic of the gene with your mouse pointer. You can click on the green boxes denoting the exons of the gene to get links to various sequences and analyses associated with the gene. Note that the green track is a composite of the mRNA and CDS tracks – click on either the NM_ or NP_ number to see the deconvolution of the green track (Figure 8).



ibliography

elated articles in PubMed

1. Degradation of the ABA co-receptor ABI1 by PUB12/13 U-box E3 ligases.
   Kong L, *et al*. Nat Commun, 2015 Oct 20. PMID 26482222, Free PMC Article
2. Direct ubiquitination of pattern recognition receptor FLS2 attenuates plant innate immunity.
   Lu D, *et al*. Science, 2011 Jun 17. PMID 21680842, Free PMC Article
3. The dominant negative ARM domain uncovers multiple functions of PUB13 in Arabidopsis immunity, flowering, and senescence.
   Zhou J, *et al*. J Exp Bot, 2015 Jun. PMID 25873653, Free PMC Article
4. Identification of 118 Arabidopsis transcription factor and 30 ubiquitin-ligase genes responding to chitin, a plant-defense elicitor.
   Libault M, *et al*. Mol Plant Microbe Interact, 2007 Aug. PMID 17722694
5. Identification and dynamics of Arabidopsis adaptor protein-2 complex and its involvement in floral organ development.
   Yamaoka S, *et al*. Plant Cell, 2013 Aug. PMID 23975897, Free PMC Article

ee all (6) citations in PubMed

NM_001336190.1

mRNA: NM_001336190.1
Title: mRNA-armadillo/beta-catenin repeat protein
Location: complement(12,368,220..12,370,420)
[*Length*]
Span: 2,201
Placed: 1,949
Product: 1,949
[*Qualifiers*]
inference: Similar to RNA sequence, EST:INSD:BP785826.1,INSD:ES025446.1,INSD:ES074681.1, INSD:EG430701.1,INSD:EG430714.1,INSD:BX839163.1, INSD:EG430704.1,INSD:EG430699.1,INSD:EG430711.1, INSD:EG430753.1,INSD:EG430712.1,INSD:EG430751.1, INSD:EG430705.1,INSD:EG430709.1,INSD:EG430703.1, INSD:EG430710.1,INSD:AV827460.1,INSD:ES050780.1, INSD:EG430700.1,INSD:EG430702.1,INSD:EG430706.1, INSD:EG430713.1,INSD:EG430698.1
inference: similar to RNA sequence, mRNA:INSD:AY219111.2,INSD:DQ056551.1
orig_protein_id: gnl|JCVI|AT2G28830.1
orig_transcript_id: gnl|JCVI|mRNA.AT2G28830.1

Download: NM_001336190.1

Links & Tools
View TAIR: AT2G28830
View Araport: AT2G28830
View GeneID: 817432 (PUB12)

GenBank View: NC_003071.7 (12,368,220..12,370,420), NM_00133619[illegible].1
FASTA View: NC_003071.7 (12,368,220..12,370,420), NM_0013361[illegible]
BLAST Genomic: NC_003071.7 (12,368,220..12,370,420)
Graphical View: NM_001336190.1
BLAST mRNA: NM_001336190.1

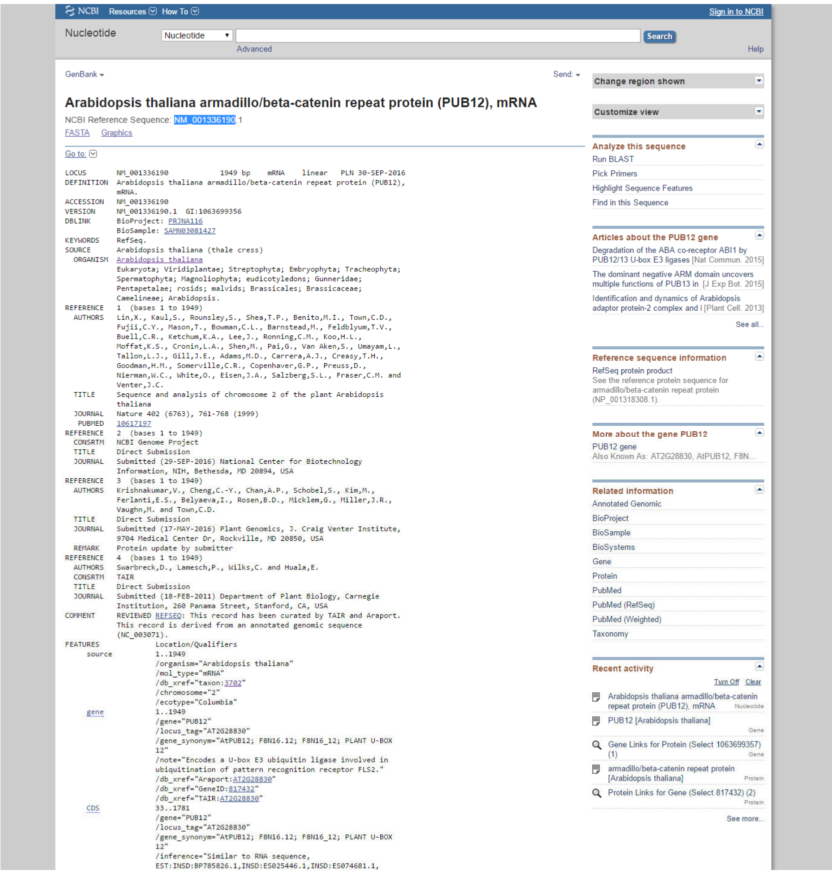
Figure 8. Part of the Gene page for NP_001318308, showing pop-up to sequence links. 1. Click the green bars to make mRNA and protein tracks appear; 2. hover over the mRNA track to see info panel; 3. click on NM_001336190 link to see GenBank record.

- Click on the mRNA link (***NM_001336190*** – the "M" in the accession number denotes mRNA – you may notice that this record is identical to the "RefSeq" record you accessed in a different way in Step 10 of the first part of the lab) and select **GenBank View** (you may need to scroll to the right to access this link; see Figure 8). This takes you to the mRNA that encodes the protein you have been looking at. Notice the feature list in the record. One Feature in the GenBank record is **gene**, and corresponds to base position 1 – 1949 on this record. Another features is the coding sequence (**CDS**), which corresponds to base position 33 – 1781.

   a. *Given your biology background knowledge, why do you think these are different?*

- On the pop-up on the Gene page click on the Nucleotide Link [***NC_003071.7 (12368220..12370420)***], and select **GenBank View**. This takes you to the genomic region that encodes the mRNA you were just looking at. Notice how the **gene** feature corresponds to positions 1 – 2201, while the **mRNA** feature corresponds to positions 1 to 1296, 1383 to 1832, 1916 to 2032, and 2116 to 2201, and the **CDS** feature corresponds to positions 169 to 1296, 1383 to 1832, 1916 to 2032, and 2116 to 2169.

   b. *Again, why are these different? Tip: recall the Central Dogma of Molecular Biology.*



Arabidopsis thaliana armadillo/beta-catenin repeat protein (PUB12), mRNA

NCBI Reference Sequence: NM_001336190.1

```
LOCUS       NM_001336190            1949 bp    mRNA    linear   PLN 30-SEP-2016
DEFINITION  Arabidopsis thaliana armadillo/beta-catenin repeat protein (PUB12),
            mRNA.
ACCESSION   NM_001336190
VERSION     NM_001336190.1  GI:1063699356
DBLINK      BioProject: PRJNA116
            BioSample: SAMN03081427
KEYWORDS    RefSeq.
SOURCE      Arabidopsis thaliana (thale cress)
  ORGANISM  Arabidopsis thaliana
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; eudicotyledons; Gunneridae;
            Pentapetalae; rosids; malvids; Brassicales; Brassicaceae;
            Camelineae; Arabidopsis.
REFERENCE   1  (bases 1 to 1949)
  AUTHORS   Lin,X., Kaul,S., Rounsley,S., Shea,T.P., Benito,M.I., Town,C.D.,
            Fujii,C.Y., Mason,T., Bowman,C.L., Barnstead,M., Feldblyum,T.V.,
            Buell,C.R., Ketchum,K.A., Lee,J., Ronning,C.M., Koo,H.L.,
            Moffat,K.S., Cronin,L.A., Shen,M., Pai,G., Van Aken,S., Umayam,L.,
            Tallon,L.J., Gill,J.E., Adams,M.D., Carrera,A.J., Creasy,T.H.,
            Goodman,H.M., Somerville,C.R., Copenhaver,G.P., Preuss,D.,
            Nierman,W.C., White,O., Eisen,J.A., Salzberg,S.L., Fraser,C.M. and
            Venter,J.C.
  TITLE     Sequence and analysis of chromosome 2 of the plant Arabidopsis
            thaliana
  JOURNAL   Nature 402 (6763), 761-768 (1999)
   PUBMED   10617197
REFERENCE   2  (bases 1 to 1949)
  CONSRTM   NCBI Genome Project
  TITLE     Direct Submission
  JOURNAL   Submitted (29-SEP-2016) National Center for Biotechnology
            Information, NIH, Bethesda, MD 20894, USA
REFERENCE   3  (bases 1 to 1949)
  AUTHORS   Krishnakumar,V., Cheng,C.-Y., Chan,A.P., Schobel,S., Kim,M.,
            Ferlanti,E.S., Belyaeva,I., Rosen,B.D., Micklem,G., Miller,J.R.,
            Vaughn,M. and Town,C.D.
  TITLE     Direct Submission
  JOURNAL   Submitted (17-MAY-2016) Plant Genomics, J. Craig Venter Institute,
            9704 Medical Center Dr, Rockville, MD 20850, USA
  REMARK    Protein update by submitter
REFERENCE   4  (bases 1 to 1949)
  AUTHORS   Swarbreck,D., Lamesch,P., Wilks,C. and Huala,E.
  CONSRTM   TAIR
  TITLE     Direct Submission
  JOURNAL   Submitted (18-FEB-2011) Department of Plant Biology, Carnegie
            Institution, 260 Panama Street, Stanford, CA, USA
COMMENT     REVIEWED REFSEQ: This record has been curated by TAIR and Araport.
            This record is derived from an annotated genomic sequence
            (NC_003071).
FEATURES             Location/Qualifiers
     source          1..1949
                     /organism="Arabidopsis thaliana"
                     /mol_type="mRNA"
                     /db_xref="taxon:3702"
                     /chromosome="2"
                     /ecotype="Columbia"
     gene            1..1949
                     /gene="PUB12"
                     /locus_tag="AT2G28830"
                     /gene_synonym="AtPUB12; F8N16.12; F8N16_12; PLANT U-BOX
                     12"
                     /note="Encodes a U-box E3 ubiquitin ligase involved in
                     ubiquitination of pattern recognition receptor FLS2."
                     /db_xref="Araport:AT2G28830"
                     /db_xref="GeneID:817432"
                     /db_xref="TAIR:AT2G28830"
     CDS             33..1781
                     /gene="PUB12"
                     /locus_tag="AT2G28830"
                     /gene_synonym="AtPUB12; F8N16.12; F8N16_12; PLANT U-BOX
                     12"
                     /inference="Similar to RNA sequence,
                     EST:INSD:BP785826.1,INSD:ES025446.1,INSD:ES074681.1,
```

Analyze this sequence
- Run BLAST
- Pick Primers
- Highlight Sequence Features
- Find in this Sequence

Articles about the PUB12 gene
- Degradation of the ABA co-receptor ABI1 by PUB12/13 U-box E3 ligases [Nat Commun. 2015]
- The dominant negative ARM domain uncovers multiple functions of PUB13 in [J Exp Bot. 2015]
- Identification and dynamics of Arabidopsis adaptor protein-2 complex and i [Plant Cell. 2013]

See all...

Reference sequence information

RefSeq protein product

See the reference protein sequence for armadillo/beta-catenin repeat protein (NP_001318308.1).

More about the gene PUB12

PUB12 gene

Also Known As: AT2G28830, AtPUB12, F8N...

Related information
- Annotated Genomic
- BioProject
- BioSample
- BioSystems
- Gene
- Protein
- PubMed
- PubMed (RefSeq)
- PubMed (Weighted)
- Taxonomy

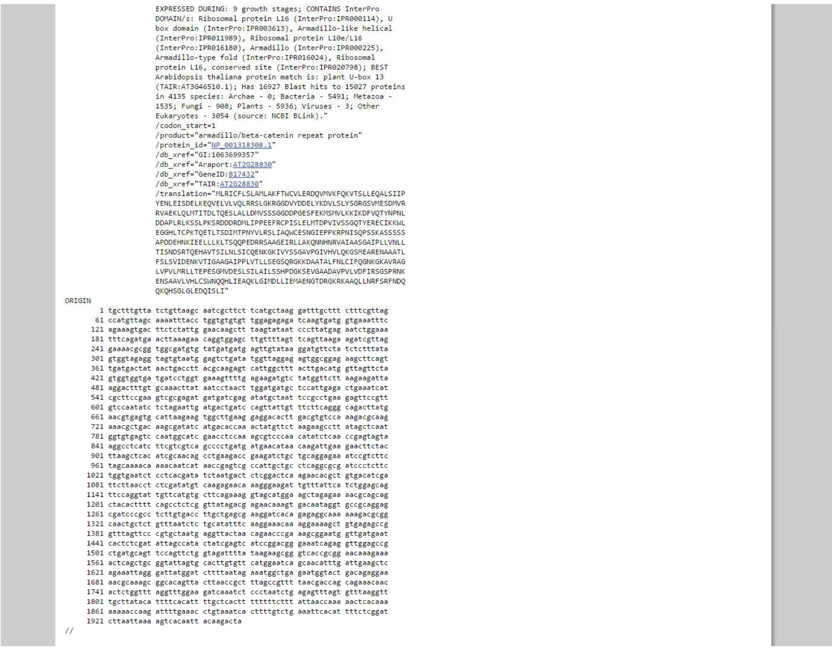
Recent activity

Turn Off Clear

- Arabidopsis thaliana armadillo/beta-catenin repeat protein (PUB12), mRNA Nucleotide
- PUB12 [Arabidopsis thaliana] Gene
- Gene Links for Protein (Select 1063699357) (1) Gene
- armadillo/beta-catenin repeat protein [Arabidopsis thaliana] Protein
- Protein Links for Gene (Select 817432) (2) Protein

See more...

…

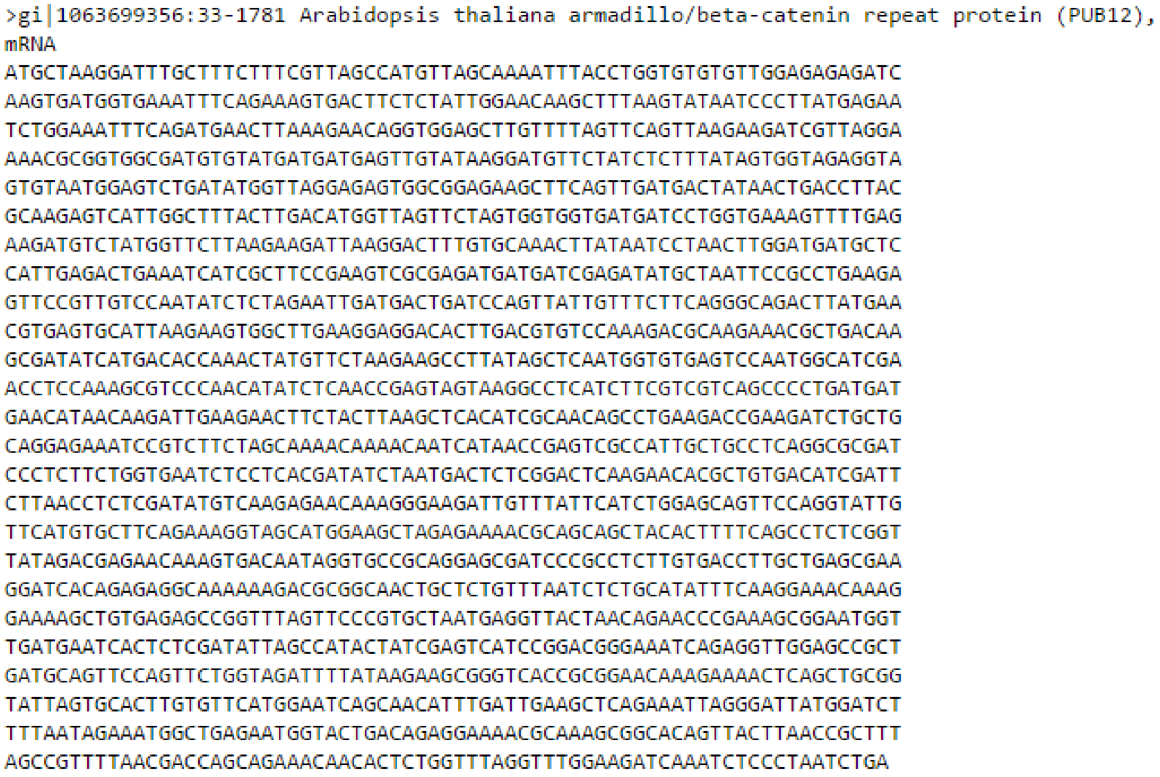


```
                     EXPRESSED DURING: 9 growth stages; CONTAINS InterPro
                     DOMAIN/s: Ribosomal protein L16 (InterPro:IPR000114), U
                     box domain (InterPro:IPR003613), Armadillo-like helical
                     (InterPro:IPR011989), Ribosomal protein L10e/L16
                     (InterPro:IPR016180), Armadillo (InterPro:IPR000225),
                     Armadillo-type fold (InterPro:IPR016024), Ribosomal
                     protein L16, conserved site (InterPro:IPR020798); BEST
                     Arabidopsis thaliana protein match is: plant U-box 13
                     (TAIR:AT3G46510.1); Has 16927 Blast hits to 15027 proteins
                     in 4135 species: Archae - 0; Bacteria - 5491; Metazoa -
                     1535; Fungi - 908; Plants - 5936; Viruses - 3; Other
                     Eukaryotes - 3054 (source: NCBI BLink)."
                     /codon_start=1
                     /product="armadillo/beta-catenin repeat protein"
                     /protein_id="NP_001318308.1"
                     /db_xref="GI:1063699357"
                     /db_xref="Araport:AT2G28830"
                     /db_xref="GeneID:817432"
                     /db_xref="TAIR:AT2G28830"
                     /translation="MLRICFLSLAMLAKFTWCVLERDQVMVKFQKVTSLLEQALSIIP
                     YENLEISDELKEQVELVLVQLRRSLGKRGGDVYDDELYKDVLSLYSGRGSVMESDMVR
                     RVAEKLQLMTITDLTQESLALLDMVSSSGGDDPGESFEKMSMVLKKIKDFVQTYNPNL
                     DDAPLRLKSSLPKSRDDDRDMLIPPEEFRCPISLELMTDPVIVSSGQTYERECIKKWL
                     EGGHLTCPKTQETLTSDIMTPNYVLRSLIAQWCESNGIEPPKRPNISQPSSKASSSSS
                     APDDEHNKIEELLLKLTSQQPEDRRSAAGEIRLLAKQNNHNRVAIAASGAIPLLVNLL
                     TISNDSRTQEHAVTSILNLSICQENKGKIVYSSGAVPGIVHVLQKGSMEARENAAATL
                     FSLSVIDENKVTIGAAGAIPPLVTLLSEGSQRGKKDAATALFNLCIFQGNKGKAVRAG
                     LVPVLMRLLTEPESGMVDESLSILAILSSHPDGKSEVGAADAVPVLVDFIRSGSPRNK
                     ENSAAVLVHLCSWNQQHLIEAQKLGIMDLLIEMAENGTDRGKRKAAQLLNRFSRFNDQ
                     QKQHSGLGLEDQISLI"
ORIGIN      
        1 tgctttgtta tctgttaagc aatcgcttct tcatgctaag gatttgcttt ctttcgttag
       61 ccatgttagc aaaatttacc tggtgtgtgt tggagagaga tcaagtgatg gtgaaatttc
      121 agaaagtgac ttctctattg gaacaagctt taagtataat cccttatgag aatctggaaa
      181 tttcagatga acttaaagaa caggtggagc ttgttttagt tcagttaaga agatcgttag
      241 gaaaacgcgg tggcgatgtg tatgatgatg agttgtataa ggatgttcta tctctttata
      301 gtggtagagg tagtgtaatg gagtctgata tggttaggag agtggcggag aagcttcagt
      361 tgatgactat aactgacctt acgcaagagt cattggcttt acttgacatg gttagttcta
      421 gtggtggtga tgatcctggt gaaagttttg agaagatgtc tatggttctt aagaagatta
      481 aggactttgt gcaaacttat aatcctaact tggatgatgc tccattgaga ctgaaatcat
      541 cgcttccgaa gtcgcgagat gatgatcgag atatgctaat tccgcctgaa gagttccgtt
      601 gtccaatatc tctagaattg atgactgatc cagttattgt ttcttcaggg cagacttatg
      661 aacgtgagtg cattaagaag tggcttgaag gaggacactt gacgtgtcca aagacgcaag
      721 aaacgctgac aagcgatatc atgacaccaa actatgttct aagaagcctt atagctcaat
      781 ggtgtgagtc caatggcatc gaacctccaa agcgtcccaa catatctcaa ccgagtagta
      841 aggcctcatc ttcgtcgtca gcccctgatg atgaacataa caagattgaa gaacttctac
      901 ttaagctcac atcgcaacag cctgaagacc gaagatctgc tgcaggagaa atccgtcttc
      961 tagcaaaaca aaacaatcat aaccgagtcg ccattgctgc ctcaggcgcg atccctcttc
     1021 tggtgaatct cctcacgata tctaatgact ctcggactca agaacacgct gtgacatcga
     1081 ttcttaacct ctcgatatgt caagagaaca aagggaagat tgtttattca tctggagcag
     1141 ttccaggtat tgttcatgtg cttcagaaag gtagcatgga agctagagaa aacgcagcag
     1201 ctacactttt cagcctctcg gttatagacg agaacaaagt gacaataggt gccgcaggag
     1261 cgatcccgcc tcttgtgacc ttgctgagcg aaggatcaca gagaggcaaa aaagacgcgg
     1321 caactgctct gtttaatctc tgcatatttc aaggaaacaa aggaaaagct gtgagagccg
     1381 gtttagttcc cgtgctaatg aggttactaa cagaacccga aagcggaatg gttgatgaat
     1441 cactctcgat attagccata ctatcgagtc atccggacgg gaaatcagag gttggagccg
     1501 ctgatgcagt tccagttctg gtagatttta taagaagcgg gtcaccgcgg aacaaagaaa
     1561 actcagctgc ggtattagtg cacttgtgtt catggaatca gcaacatttg attgaagctc
     1621 agaaattagg gattatggat cttttaatag aaatggctga gaatggtact gacagaggaa
     1681 aacgcaaagc ggcacagtta cttaaccgct ttagccgttt taacgaccag cagaaacaac
     1741 actctggttt aggtttggaa gatcaaatct ccctaatctg agagtttagt gtttaaggtt
     1801 tgcttataca ttttcacatt ttgctcactt ttttttcttt attaaccaaa aactcacaaa
     1861 aaaaaccaag attttgaaac ctgtaaatca cttttgtctg aaattcacat tttctcggat
     1921 cttaattaaa agtcacaatt acaagacta
//
```

…

Figure 9. GenBank record for NM_001336190 mRNA.

- Let's return the mRNA record we were previously working with (NM_001336190). Click on the **CDS** link. Now you are looking at the information for the coding sequence, as opposed to the whole gene or protein (highlighted in brown).
- Using the “Display: FASTA” option in the grey bar at the bottom of the page generate a FASTA-formatted version of the CDS.
- Now you have the sequence in the most basic and easily managed format – **FASTA** format. FASTA format is simply a header line that starts with a ‘>’ followed by text describing the sequence, and then the actual sequence beginning on the next line. The sequence can be either DNA or protein, and may be continuous (scrolling off the page), or cut into more manageable lengths typically ranging between 60-80 residues.



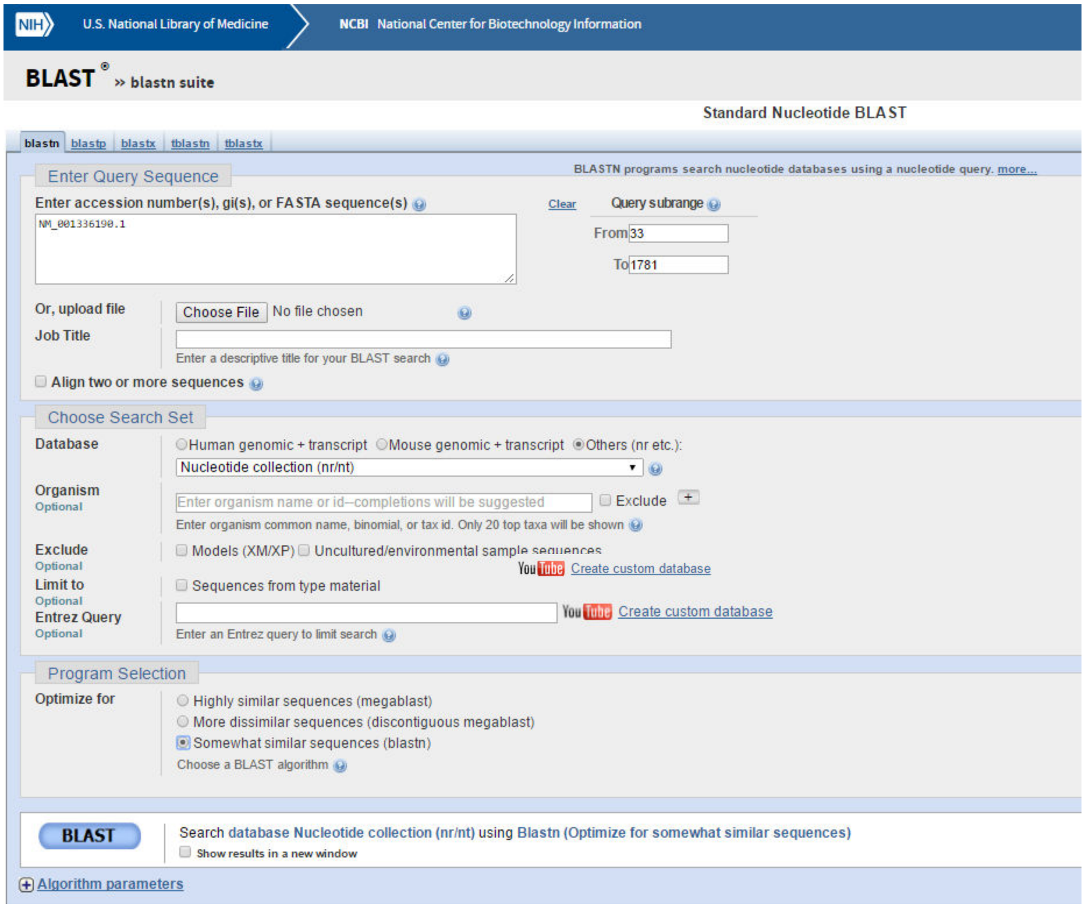
```
>gi|1063699356:33-1781 Arabidopsis thaliana armadillo/beta-catenin repeat protein (PUB12),
mRNA
ATGCTAAGGATTTGCTTTCTTTCGTTAGCCATGTTAGCAAAATTTACCTGGTGTGTGTTGGAGAGAGATC
AAGTGATGGTGAAATTTCAGAAAGTGACTTCTCTATTGGAACAAGCTTTAAGTATAATCCCTTATGAGAA
TCTGGAAATTTCAGATGAACTTAAAGAACAGGTGGAGCTTGTTTTAGTTCAGTTAAGAAGATCGTTAGGA
AAACGCGGTGGCGATGTGTATGATGATGAGTTGTATAAGGATGTTCTATCTCTTTATAGTGGTAGAGGTA
GTGTAATGGAGTCTGATATGGTTAGGAGAGTGGCGGAGAAGCTTCAGTTGATGACTATAACTGACCTTAC
GCAAGAGTCATTGGCTTTACTTGACATGGTTAGTTCTAGTGGTGGTGATGATCCTGGTGAAAGTTTTGAG
AAGATGTCTATGGTTCTTAAGAAGATTAAGGACTTTGTGCAAACTTATAATCCTAACTTGGATGATGCTC
CATTGAGACTGAAATCATCGCTTCCGAAGTCGCGAGATGATGATCGAGATATGCTAATTCCGCCTGAAGA
GTTCCGTTGTCCAATATCTCTAGAATTGATGACTGATCCAGTTATTGTTTCTTCAGGGCAGACTTATGAA
CGTGAGTGCATTAAGAAGTGGCTTGAAGGAGGACACTTGACGTGTCCAAAGACGCAAGAAACGCTGACAA
GCGATATCATGACACCAAACTATGTTCTAAGAAGCCTTATAGCTCAATGGTGTGAGTCCAATGGCATCGA
ACCTCCAAAGCGTCCCAACATATCTCAACCGAGTAGTAAGGCCTCATCTTCGTCGTCAGCCCCTGATGAT
GAACATAACAAGATTGAAGAACTTCTACTTAAGCTCACATCGCAACAGCCTGAAGACCGAAGATCTGCTG
CAGGAGAAATCCGTCTTCTAGCAAAACAAAACAATCATAACCGAGTCGCCATTGCTGCCTCAGGCGCGAT
CCCTCTTCTGGTGAATCTCCTCACGATATCTAATGACTCTCGGACTCAAGAACACGCTGTGACATCGATT
CTTAACCTCTCGATATGTCAAGAGAACAAAGGGAAGATTGTTTATTCATCTGGAGCAGTTCCAGGTATTG
TTCATGTGCTTCAGAAAGGTAGCATGGAAGCTAGAGAAAACGCAGCAGCTACACTTTTCAGCCTCTCGGT
TATAGACGAGAACAAAGTGACAATAGGTGCCGCAGGAGCGATCCCGCCTCTTGTGACCTTGCTGAGCGAA
GGATCACAGAGAGGCAAAAAAGACGCGGCAACTGCTCTGTTTAATCTCTGCATATTTCAAGGAAACAAAG
GAAAAGCTGTGAGAGCCGGTTTAGTTCCCGTGCTAATGAGGTTACTAACAGAACCCGAAAGCGGAATGGT
TGATGAATCACTCTCGATATTAGCCATACTATCGAGTCATCCGGACGGGAAATCAGAGGTTGGAGCCGCT
GATGCAGTTCCAGTTCTGGTAGATTTTATAAGAAGCGGGTCACCGCGGAACAAAGAAAACTCAGCTGCGG
TATTAGTGCACTTGTGTTCATGGAATCAGCAACATTTGATTGAAGCTCAGAAATTAGGGATTATGGATCT
TTTAATAGAAATGGCTGAGAATGGTACTGACAGAGGAAAACGCAAAGCGGCACAGTTACTTAACCGCTTT
AGCCGTTTTAACGACCAGCAGAAACAACACTCTGGTTTAGGTTTGGAAGATCAAATCTCCCTAATCTGA
```

Figure 10. Sequence in FASTA text format.

2. Let’s do some BLASTing! Use the “Run BLAST” link in the “Analyze This Sequence” part of the webpage. [Or open a new tab or window in your browser and go back to the NCBI home page (www.ncbi.nlm.nih.gov), then select **BLAST** from the Resources dropdown along the top, under the DNA&RNA subsection].

   There are lots of options here. We will discuss some of these next lab, but right now let’s work with the simplest. Since our sequence is a nucleotide sequence, we want to do a ***nucleotide blast***.

   - On the BLAST page, note that under the **Enter Query Sequence** section, the NCBI system has automatically entered the **accession number** (but you can also enter **a GI number,** or **FASTA sequence**) and **subrange** (we'll be searching with just the coding sequence part of the mRNA sequence). You could also copy-and-paste the FASTA formatted CDS sequence you found as in Figure 10 into the query box *without* defining a subrange – you should be clear on the difference between an mRNA sequence and coding sequence at this point...



NIH U.S. National Library of Medicine
NCBI National Center for Biotechnology Information
BLAST ® » blastn suite

Standard Nucleotide BLAST

blastn blastp blastx tblastn tblastx

Enter Query Sequence

BLASTN programs search nucleotide databases using a nucleotide query. more...

Enter accession number(s), gi(s), or FASTA sequence(s): NM_001336190.1

Clear

Query subrange

From 33

To 1781

Or, upload file: Choose File No file chosen

Job Title:

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: ○Human genomic + transcript ○Mouse genomic + transcript ●Others (nr etc.):

Nucleotide collection (nr/nt)

Organism Optional: Enter organism name or id--completions will be suggested □ Exclude +

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

Exclude Optional: □ Models (XM/XP) □ Uncultured/environmental sample sequences

YouTube Create custom database

Limit to Optional: □ Sequences from type material

Entrez Query Optional:

YouTube Create custom database

Enter an Entrez query to limit search

Program Selection

Optimize for:

○ Highly similar sequences (megablast)

○ More dissimilar sequences (discontiguous megablast)

● Somewhat similar sequences (blastn)

Choose a BLAST algorithm

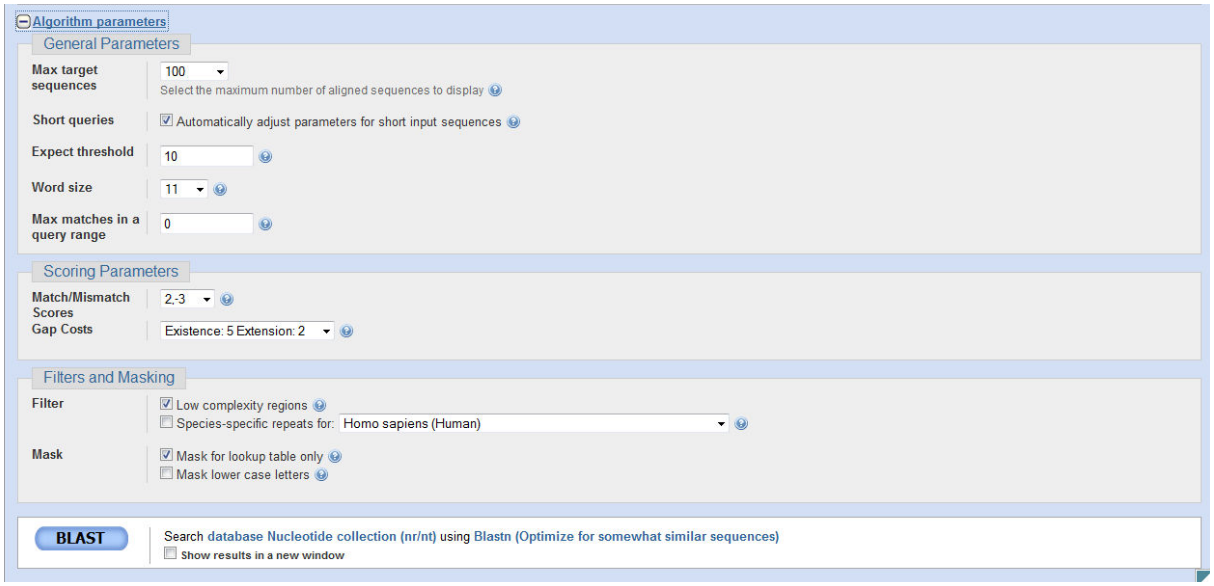
Search database Nucleotide collection (nr/nt) using Blastn (Optimize for somewhat similar sequences)

□ Show results in a new window

Figure 11. The blastn query page, with optimization for "Somewhat similar sequences (blastn)" selected.

- Scan the sections of the page. You have quite a bit of control over how the algorithm runs (particularly if you click [+] **Algorithm parameters** near the bottom.
- We want to query the full NCBI database; the NCBI linking system has automatically changed the default **Database** (which is Human) to ***Other*** and ***Nucleotide collection (nr/nt)*** because our sequence is non-human. The nr database is the non-redundant collection of sequences in GenBank.
- Change the **Program Selected / Optimized for** to ***Somewhat similar sequences (blastn)***.
- Note all the small question mark icons around the page. Click any one of these to find out more about the associated parameter. For example, by clicking the question mark in the **Program Selection** section you get a very brief summary of the different methods. By clicking **more** you jump to a new page with full documentation for the algorithms.

a. When would you want to use megaBLAST? What about discontinuous megaBLAST? (if you have time, try each to see how your results differ)



⊟ Algorithm parameters

General Parameters

Max target sequences: 100 ▾
Select the maximum number of aligned sequences to display

Short queries: ☑ Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: 11 ▾

Max matches in a query range: 0

Scoring Parameters

Match/Mismatch Scores: 2,-3 ▾

Gap Costs: Existence: 5 Extension: 2 ▾

Filters and Masking

Filter: ☑ Low complexity regions
☐ Species-specific repeats for: Homo sapiens (Human) ▾

Mask: ☑ Mask for lookup table only
☐ Mask lower case letters

Search database Nucleotide collection (nr/nt) using Blastn (Optimize for somewhat similar sequences)
☐ Show results in a new window

Figure 12. Algorithm parameters for blastn.

- Open the Algorithm Parameters near the bottom.
  - *b. What is the* ***Expect threshold****?*
  - *c. What would happen if you decreased it? Increased it?*
  - *d. What would be the effect of increasing the* ***Word size****?*
  - *e. Why is there a* ***Low complexity regions*** *filter? Should we keep it on?*

Lab Quiz Question 2

- Make sure you have your query sequence entered in the input box, and check the box next to **Show results in a new** window near the **BLAST** button. Now (finally) click the **BLAST** button.
- While BLAST is running or after the search is complete you can choose to adjust the format of the search results by clicking on the **Format options** link. We won’t do this right now, as the defaults usually work fine.

Box 5. How Good is My Hit?

The quality of a BLAST HSP is quantified in a number of different ways. It is important that you understand the differences between these metrics and use the appropriate one.

- Identity – the extent to which two sequences are invariant. A very poor measure since it doesn’t take into account the subtleties of sequence relationships (e.g. a small region of a highly conserved domain within two sequences that are otherwise very poorly conserved).
- Bit score – the alignment score (S). A very precise measure that is normalized over the particular score system employed. Suffers from the disadvantage of being dependent on the length of the query.

- E value – the expect value. A value that is based on the number of different alignments with scores at least as good as that observed, which are expected to occur simply by chance. The lower the E value, the more significant the score. This is by far the best metric to use since results of different searches in the same database can be readily compared. Note that E value is dependent on the size of the database (n) and the length of the query sequence (m). The same sequence searched on different databases containing identical hit sequences would result in different E values being reported.

$$E = mn2^{-S}$$

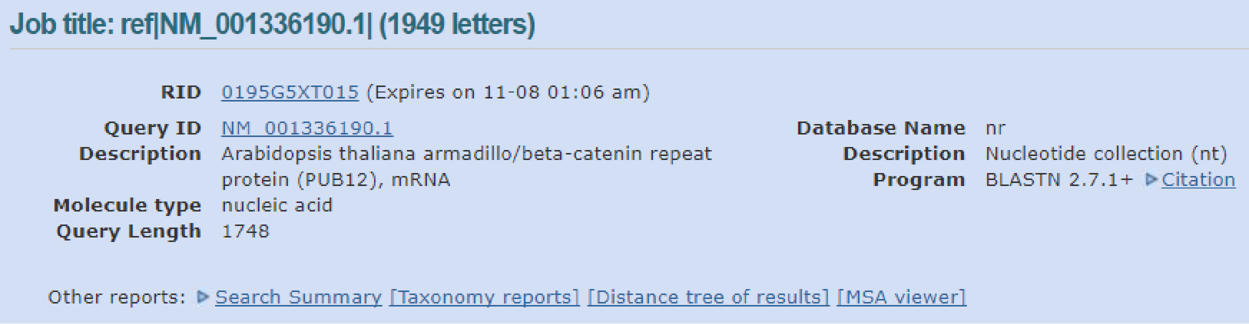
We'll go into greater detail about this calculation in next week's class.

3. The Results page is broken up into sections.

- At the very top is the job summary, which simply shows details about your query and the database searched. You can find more details about your search by clicking **Search Summary.**

a. How many sequences are in the nr database?

b. What sequences are not included in the nr database? (Trick question: this information is actually available by clicking on the question mark beside the Database option on the input page!)



Job title: ref|NM_001336190.1| (1949 letters)

RID 0195G5XT015 (Expires on 11-08 01:06 am)

Query ID	NM_001336190.1	Database Name	nr
Description	Arabidopsis thaliana armadillo/beta-catenin repeat protein (PUB12), mRNA	Description	Nucleotide collection (nt)
Molecule type	nucleic acid	Program	BLASTN 2.7.1+ ▷Citation
Query Length	1748		

Other reports: ▷Search Summary [Taxonomy reports] [Distance tree of results] [MSA viewer]

Figure 13. Blastn output Search Summary.

- Next is the **Graphic Summary**. Scroll your mouse over the coloured bars.

c. What do the coloured bars mean?

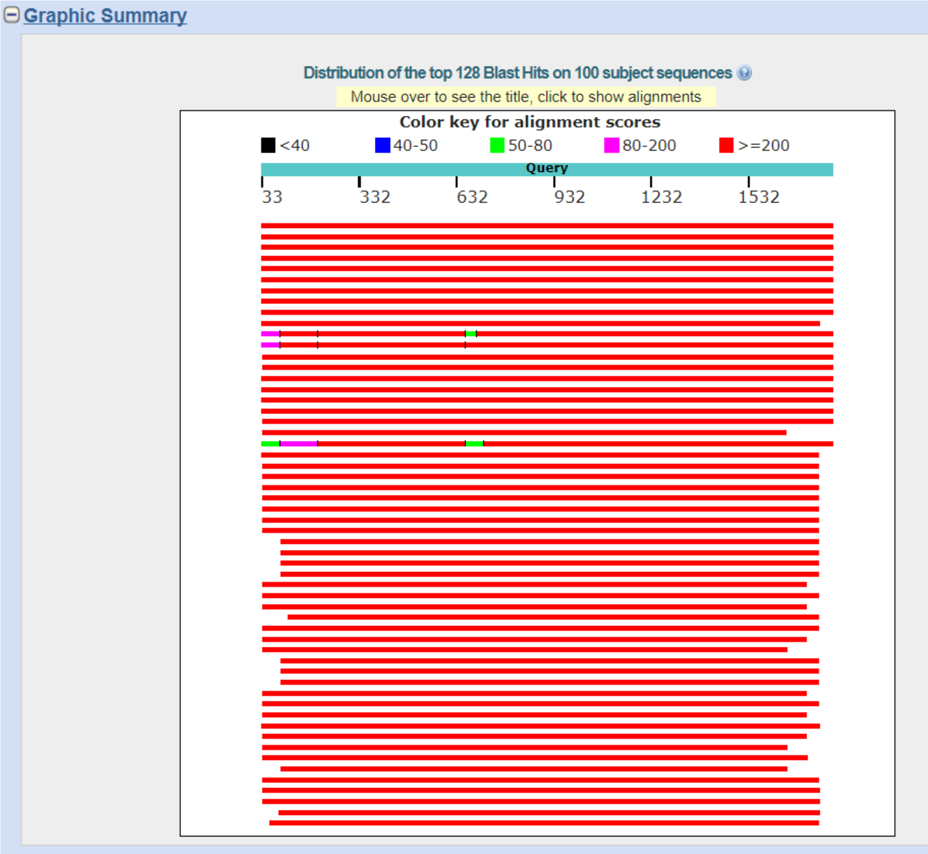
d. How does the colour code work?

e. What information is displayed in the box near the top of the graphic summary?

f. What do you notice about the significance values as you move down the graphical summary?

g. What is the genus and species of the top (best) hit?

h. What happens if you click on one of the entries?



Graphic Summary

Distribution of the top 128 Blast Hits on 100 subject sequences

Mouse over to see the title, click to show alignments

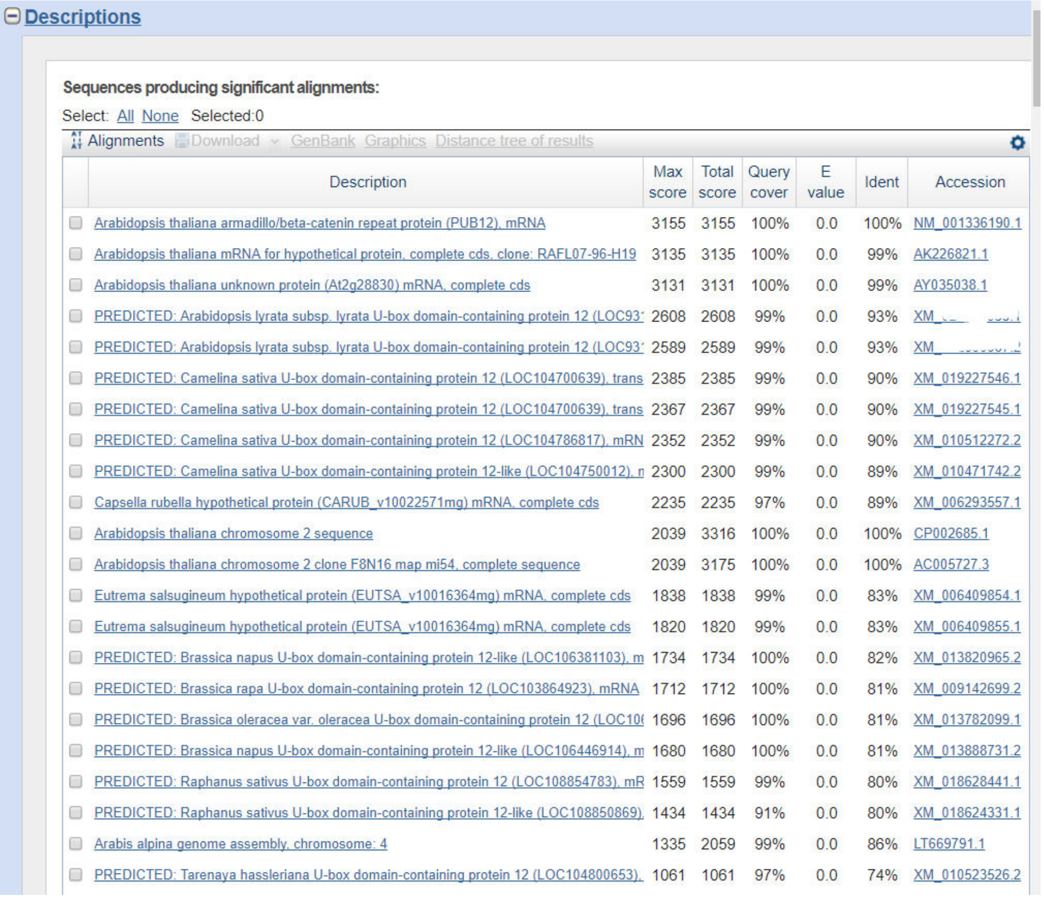
Figure 14. Blastn output Graphic Summary.

- The **Descriptions** section is next, listing:
  - **Description** [hyperlinked to corresponding **Alignment(s)** in Alignments section]
  - **Max Score** – the alignment bit score
  - **Total Score** – another alignment bit score which may differ from the **Max Score** if your query matched a single database entry in multiple regions.
  - **Query Coverage** – what percent of the query had similarity to the database hit.
  - **E-value** – probably the best measure of hit quality. Smaller numbers mean better hits, with 0.0 being the best value possible.
  - **Identity** – the highest identity found between query and HSP.
  - **Accession** – linked to the indicated sequence at NCBI

i. *How many sequence matches are listed for this query sequence? How are they ordered? (you can sort these segments in other ways, like by identity, score, and query start position.)*

j. *What happens if you click the **Accession** hotlink?*

k. *What happens if you click the **Alignments** hotlink?*



⊖ Descriptions

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

	Description	Max score	Total score	Query cover	E value	Ident	Accession
□	Arabidopsis thaliana armadillo/beta-catenin repeat protein (PUB12), mRNA	3155	3155	100%	0.0	100%	NM_001336190.1
□	Arabidopsis thaliana mRNA for hypothetical protein, complete cds, clone: RAFL07-96-H19	3135	3135	100%	0.0	99%	AK226821.1
□	Arabidopsis thaliana unknown protein (At2g28830) mRNA, complete cds	3131	3131	100%	0.0	99%	AY035038.1
□	PREDICTED: Arabidopsis lyrata subsp. lyrata U-box domain-containing protein 12 (LOC93	2608	2608	99%	0.0	93%	XM_[illegible]
□	PREDICTED: Arabidopsis lyrata subsp. lyrata U-box domain-containing protein 12 (LOC93	2589	2589	99%	0.0	93%	XM_[illegible]
□	PREDICTED: Camelina sativa U-box domain-containing protein 12 (LOC104700639), trans	2385	2385	99%	0.0	90%	XM_019227546.1
□	PREDICTED: Camelina sativa U-box domain-containing protein 12 (LOC104700639), trans	2367	2367	99%	0.0	90%	XM_019227545.1
□	PREDICTED: Camelina sativa U-box domain-containing protein 12 (LOC104786817), mRN	2352	2352	99%	0.0	90%	XM_010512272.2
□	PREDICTED: Camelina sativa U-box domain-containing protein 12-like (LOC104750012), n	2300	2300	99%	0.0	89%	XM_010471742.2
□	Capsella rubella hypothetical protein (CARUB_v10022571mg) mRNA, complete cds	2235	2235	97%	0.0	89%	XM_006293557.1
□	Arabidopsis thaliana chromosome 2 sequence	2039	3316	100%	0.0	100%	CP002685.1
□	Arabidopsis thaliana chromosome 2 clone F8N16 map mi54, complete sequence	2039	3175	100%	0.0	100%	AC005727.3
□	Eutrema salsugineum hypothetical protein (EUTSA_v10016364mg) mRNA, complete cds	1838	1838	99%	0.0	83%	XM_006409854.1
□	Eutrema salsugineum hypothetical protein (EUTSA_v10016364mg) mRNA, complete cds	1820	1820	99%	0.0	83%	XM_006409855.1
□	PREDICTED: Brassica napus U-box domain-containing protein 12-like (LOC106381103), m	1734	1734	100%	0.0	82%	XM_013820965.2
□	PREDICTED: Brassica rapa U-box domain-containing protein 12 (LOC103864923), mRNA	1712	1712	100%	0.0	81%	XM_009142699.2
□	PREDICTED: Brassica oleracea var. oleracea U-box domain-containing protein 12 (LOC10	1696	1696	100%	0.0	81%	XM_013782099.1
□	PREDICTED: Brassica napus U-box domain-containing protein 12-like (LOC106446914), m	1680	1680	100%	0.0	81%	XM_013888731.2
□	PREDICTED: Raphanus sativus U-box domain-containing protein 12 (LOC108854783), mR	1559	1559	99%	0.0	80%	XM_018628441.1
□	PREDICTED: Raphanus sativus U-box domain-containing protein 12-like (LOC108850869),	1434	1434	91%	0.0	80%	XM_018624331.1
□	Arabis alpina genome assembly, chromosome: 4	1335	2059	99%	0.0	86%	LT669791.1
□	PREDICTED: Tarenaya hassleriana U-box domain-containing protein 12 (LOC104800653),	1061	1061	97%	0.0	74%	XM_010523526.2

Figure 15. Blastn output descriptions

Lab Quiz Question 3

- Finally we get down to the actual HSP **Alignments**.
    - Compare the information presented for the first HSP alignment to the first entry in the graphical summary and HSP summary.
    - As you scroll down the alignments, you will see the alignment quality drop – that is, the e-value increases.

        l. *What do the vertical bars (|)represent between the **Query** and the **Sbjct** (database sequence)?*

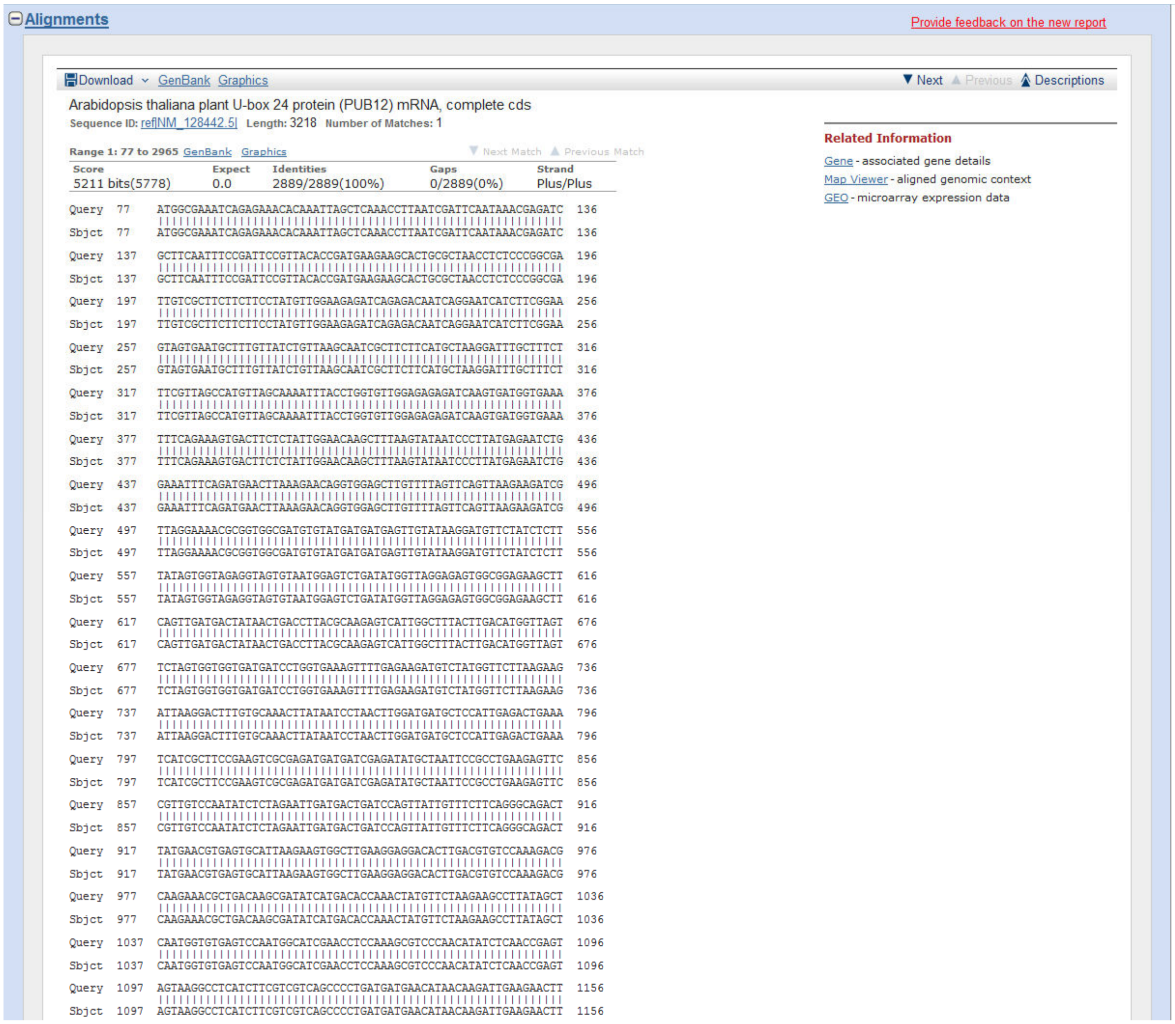
        m. *What does **Strand=Plus/Plus**, **Strand=Plus/Minus** mean? Hint: are genes always in the same direction on a piece of chromosomal DNA?*
- Go back to the top of the page and click **Formatting options**. Change the **Alignment View** to **Query-anchored with dots for identities**. Click **Reformat** and score down to the HSP alignment section.

    n. *Describe the difference between this format and the previous format. Can you imagine cases where the different formats might be most useful?*

    o. *Play with these format options to get a feel for what they mean.*

- Return the formatting to the original **Pairwise** format. Go back to the graphical summary. If there are any low-scoring segments (i.e.: green or blue-coded blocks), click on one.

n. *What is its E-value?*
o. *Does it have a high percent identity? If so, why would BLAST give it such a poor E-value?*
p. *Do you think these hits are homologous? Why or why not?*



Alignments

Provide feedback on the new report

Download GenBank Graphics ▼ Next ▲ Previous ▲ Descriptions

Arabidopsis thaliana plant U-box 24 protein (PUB12) mRNA, complete cds

Sequence ID: ref|NM_128442.5| Length: 3218 Number of Matches: 1

Range 1: 77 to 2965 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
5211 bits(5778)	0.0	2889/2889(100%)	0/2889(0%)	Plus/Plus

```
Query  77    ATGGCGAAATCAGAGAAACACAAATTAGCTCAAACCTTAATCGATTCAATAAACGAGATC  136
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  77    ATGGCGAAATCAGAGAAACACAAATTAGCTCAAACCTTAATCGATTCAATAAACGAGATC  136

Query  137   GCTTCAATTTCCGATTCCGTTACACCGATGAAGAAGCACTGCGCTAACCTCTCCCGGCGA  196
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  137   GCTTCAATTTCCGATTCCGTTACACCGATGAAGAAGCACTGCGCTAACCTCTCCCGGCGA  196

Query  197   TTGTCGCTTCTTCTTCCTATGTTGGAAGAGATCAGAGACAATCAGGAATCATCTTCGGAA  256
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  197   TTGTCGCTTCTTCTTCCTATGTTGGAAGAGATCAGAGACAATCAGGAATCATCTTCGGAA  256

Query  257   GTAGTGAATGCTTTGTTATCTGTTAAGCAATCGCTTCTTCATGCTAAGGATTTGCTTTCT  316
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  257   GTAGTGAATGCTTTGTTATCTGTTAAGCAATCGCTTCTTCATGCTAAGGATTTGCTTTCT  316

Query  317   TTCGTTAGCCATGTTAGCAAAATTTACCTGGTGTTGGAGAGAGATCAAGTGATGGTGAAA  376
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  317   TTCGTTAGCCATGTTAGCAAAATTTACCTGGTGTTGGAGAGAGATCAAGTGATGGTGAAA  376

Query  377   TTTCAGAAAGTGACTTCTCTATTGGAACAAGCTTTAAGTATAATCCCTTATGAGAATCTG  436
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  377   TTTCAGAAAGTGACTTCTCTATTGGAACAAGCTTTAAGTATAATCCCTTATGAGAATCTG  436

Query  437   GAAATTTCAGATGAACTTAAAGAACAGGTGGAGCTTGTTTTAGTTCAGTTAAGAAGATCG  496
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  437   GAAATTTCAGATGAACTTAAAGAACAGGTGGAGCTTGTTTTAGTTCAGTTAAGAAGATCG  496

Query  497   TTAGGAAAACGCGGTGGCGATGTGTATGATGATGAGTTGTATAAGGATGTTCTATCTCTT  556
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  497   TTAGGAAAACGCGGTGGCGATGTGTATGATGATGAGTTGTATAAGGATGTTCTATCTCTT  556

Query  557   TATAGTGGTAGAGGTAGTGTAATGGAGTCTGATATGGTTAGGAGAGTGGCGGAGAAGCTT  616
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  557   TATAGTGGTAGAGGTAGTGTAATGGAGTCTGATATGGTTAGGAGAGTGGCGGAGAAGCTT  616

Query  617   CAGTTGATGACTATAACTGACCTTACGCAAGAGTCATTGGCTTTACTTGACATGGTTAGT  676
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  617   CAGTTGATGACTATAACTGACCTTACGCAAGAGTCATTGGCTTTACTTGACATGGTTAGT  676

Query  677   TCTAGTGGTGGTGATGATCCTGGTGAAAGTTTTGAGAAGATGTCTATGGTTCTTAAGAAG  736
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  677   TCTAGTGGTGGTGATGATCCTGGTGAAAGTTTTGAGAAGATGTCTATGGTTCTTAAGAAG  736

Query  737   ATTAAGGACTTTGTGCAAACTTATAATCCTAACTTGGATGATGCTCCATTGAGACTGAAA  796
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  737   ATTAAGGACTTTGTGCAAACTTATAATCCTAACTTGGATGATGCTCCATTGAGACTGAAA  796

Query  797   TCATCGCTTCCGAAGTCGCGAGATGATGATCGAGATATGCTAATTCCGCCTGAAGAGTTC  856
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  797   TCATCGCTTCCGAAGTCGCGAGATGATGATCGAGATATGCTAATTCCGCCTGAAGAGTTC  856

Query  857   CGTTGTCCAATATCTCTAGAATTGATGACTGATCCAGTTATTGTTTCTTCAGGGCAGACT  916
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  857   CGTTGTCCAATATCTCTAGAATTGATGACTGATCCAGTTATTGTTTCTTCAGGGCAGACT  916

Query  917   TATGAACGTGAGTGCATTAAGAAGTGGCTTGAAGGAGGACACTTGACGTGTCCAAAGACG  976
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  917   TATGAACGTGAGTGCATTAAGAAGTGGCTTGAAGGAGGACACTTGACGTGTCCAAAGACG  976

Query  977   CAAGAAACGCTGACAAGCGATATCATGACACCAAACTATGTTCTAAGAAGCCTTATAGCT  1036
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  977   CAAGAAACGCTGACAAGCGATATCATGACACCAAACTATGTTCTAAGAAGCCTTATAGCT  1036

Query  1037  CAATGGTGTGAGTCCAATGGCATCGAACCTCCAAAGCGTCCCAACATATCTCAACCGAGT  1096
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1037  CAATGGTGTGAGTCCAATGGCATCGAACCTCCAAAGCGTCCCAACATATCTCAACCGAGT  1096

Query  1097  AGTAAGGCCTCATCTTCGTCGTCAGCCCCTGATGATGAACATAACAAGATTGAAGAACTT  1156
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1097  AGTAAGGCCTCATCTTCGTCGTCAGCCCCTGATGATGAACATAACAAGATTGAAGAACTT  1156
```

Related Information

Gene - associated gene details
Map Viewer - aligned genomic context
GEO - microarray expression data

Figure 16. Blastn output alignments.

End of Lab!

Lab 1 Objectives

By the end of Lab 1 (comprising the lab including its boxes, and the lecture), you should:

- know how to search for records at NCBI, both using search terms or identifiers (first part of lab) and GQuery, or using a nucleotide sequence and BLAST;
- know the difference between a GenBank accession number, a version number, and a GI number;
- understand the difference between the nucleotide sequence database part of GenBank and the protein sequence part of it;
- know the parts of a GenBank record and be able to switch between sequence formats (e.g. to FASTA format);
- be familiar with the interconnectedness of various NCBI databases and be able to call up linked records with ease;
- be able to use nucleotide BLAST (Blastn) to search GenBank, and be able to interpret the output – what does the E-value tell you etc.?;
- understand the meaning of homologous, orthologous, and paralogous sequences;
- be able to use the Help function to address any question you may have with regards to the NCBI interface (if you have any questions on background material, check in with the forums for this course on Coursera!).

Do not hesitate to post any questions you might have to the Forum section of the Coursera website for this course if you do not understand any of the above after reading the relevant material.

Further Reading

Chapter 2 "Information Organization and Sequence Databases" in *Concepts in Bioinformatics and Genomics* by Jamil Momand and Alison McCurdy, Oxford University Press, 2017. pp 21-37.

SF Altschul , TL Madden , AA Schaffer , J Zhang , Z Zhang , W Miller , and DJ Lipman (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids Res. 25: 3389-3402.

NM Luscombe, D Greenbaum, M Gerstein (2001) What is bioinformatics? An introduction and overview. Yearkbook of Medical Informatics 2001:83.

CA Kerfeld, KM Scott (2011) Using BLAST to Teach ''E-value-tionary'' Concepts. PLoS Biol 9(2): e1001014. http://dx.doi.org/10.1371/journal.pbio.1001014.

Appendix 1: GenBank Field Qualifiers

From http://www.ncbi.nlm.nih.gov/books/NBK3837/#EntrezHelp.Entrez_Searching_Options

Accession [ACCN]
Contains the unique accession number of the sequence or record, assigned to the nucleotide, protein, structure, genome record, or PopSet by a sequence database builder. The Structure database accession index contains the PDB IDs but not the MMDB IDs.

All Fields [ALL]
Contains all terms from all searchable database fields in the database.

Author Name [AUTH]
Contains all authors from all references in the database records. The format is last name space first initial(s), without punctuation (e.g., marley jf).

EC/RN Number [ECNO]
Number assigned by the Enzyme Commission or Chemical Abstract Service (CAS) to designate a particular enzyme or chemical, respectively.

Feature Key [FKEY]
Contains the biological features assigned or annotated to the nucleotide sequences and defined in the DDBJ/EMBL/GenBank Feature Table (http://www.ncbi.nlm.nih.gov/projects/collab/FT/index.html). Not available for the Protein or Structure databases.

Filter [FILT]
Contains predetermined or filtered subsets of the various databases. These subsets or filters are created by grouping records that are commonly linked to other GQuery databases or within the same database. For example, the PopSet database Filter index includes PopSet all, PopSet medline, PopSet nucleotide, and PopSet protein. The PopSet medline filter includes all PopSet records with links to PubMed; the PopSet nucleotide filter includes all PopSet records with links to the nucleotide database; and, the PopSet protein filter includes all PopSet records with links to the protein database. The PopSet all filter includes all PopSet records.

Gene Name [GENE]
Contains the standard and common names of genes found in the database records. This field is not available in Structure database.

Issue [ISS]
Contains the issue number of the journal in which the data were published.

Journal Name [JOUR]
Contains the name of the journal in which the data were published. Journal names are indexed in the database in abbreviated form (e.g., J Biol Chem). Journals are also indexed by their by ISSNs. Browse the index if you do not know the ISSN or are not sure how a particular journal name is abbreviated.

Keyword [KYWD]
Contains special index terms from the controlled vocabularies associated with the GenBank, EMBL, DDBJ, SWISS-Prot, PIR, PRF, or PDB databases. Browse the Keyword indexes of the individual databases to become familiar with these vocabularies. A Keyword index is not available in the Structure database.

Modification Date [MDAT]
Contains the date that the most recent modification to that record is indexed in GQuery, in the format YYYY/MM/DD (e.g., 1999/08/05). A year alone, (e.g., 1999) will retrieve all records modified for that year; a year and month (e.g., 1999/03) retrieves all records modified for that month that are indexed in GQuery.

Molecular Weight [MOLWT]
Molecular weight of a protein, in Daltons (Da), calculated by the method described in the Searching by Molecular

Weight section of the GQuery help document. Note that molecular weight must be entered as a fixed 6 digit field, filled with leading zeros (not letter O), e.g., 002002 [MOLWT]

Organism [ORGN]
Contains the scientific and common names for the organisms associated with protein and nucleotide sequences.

Page Number [PAGE]
Contains the number of the first journal page of the article in which the data were published.

Primary Accession [PACC]
Contains the primary accession number of the sequence or record, assigned to the nucleotide, protein, structure, genome record, or PopSet by a sequence database builder. A Primary Accession index is not available in the Structure database.

Properties [PROP]
Contains properties of the nucleotide or protein sequence. For example, the Nucleotide database's Properties index includes molecule types, publication status, molecule locations, and GenBank divisions. A Properties index is not available in the Structure database.

Protein Name [PROT]
Contains the standard names of proteins found in database records. Common names may not be indexed in this field so it is best to also consider All Fields or Text Words. A Protein Name index is not available in the Structure database.

Publication Date [PDAT]
Contains the date that records are released into GQuery, in the format YYYY/MM/DD (e.g., 1999/08/05). It is the date the entry first appeared in GenBank explicitly indexed in GQuery. A year alone, (e.g., 1999) will retrieve all records for that year; a year and month (e.g., 1999/03) will retrieve all records released into GenBank for that month.

SeqID String [SQID]
Contains the special string identifier, similar to a FASTA identifier, for a given sequence. A SeqID String index is not available in the Structure database.

Sequence Length [SLEN]
Contains the total length of the sequence. Sequence Length indexes are not available in the Structure or PopSet databases.

Substance Name [SUBS]
Contains the names of any chemicals associated with this record from the CAS registry and the MEDLINE Name of Substance field. Substance Name indexes are not available in the Genome or PopSet databases.

Text Word [WORD]
Contains all of the "free text" associated with a record.

Title Word [TITL]
Includes only those words found in the definition line of a record. The definition line summarizes the biology of the sequence and is carefully constructed by database staff. A standard definition line will include the organism, product name, gene symbol, molecule type and whether it is a partial or complete cds. Title Word indexes are not available in the Structure or PopSet databases.

Uid [UID]
Contains the Medline unique identifier for records that contain published references that are linked to PubMed. The Uid index is not browsable.

Volume [VOL]
Contains the volume number of the journal in which the data were published.