

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 (nor does it handle sequence searching well, especially in the case of protein sequences).

The primary portal for accessing data at NCBI is called *Search NCBI*. 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.

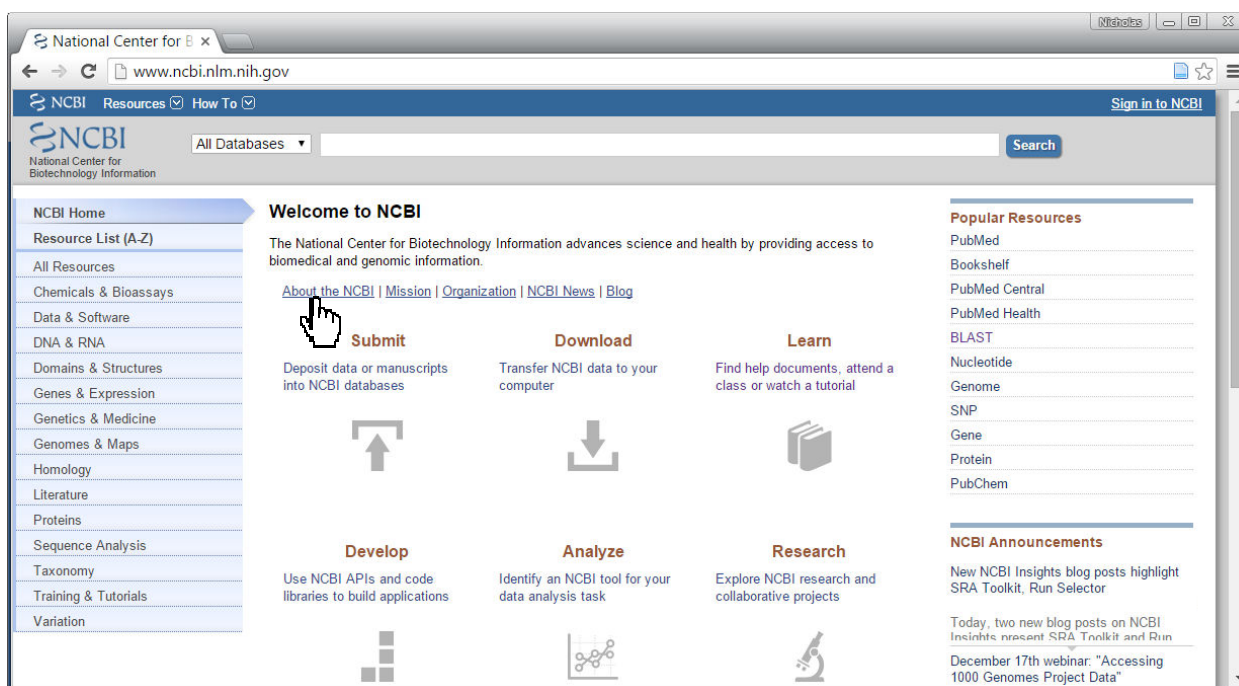


Figure 1. The NCBI homepage.

2. Now let's move to the *Search NCBI* (formerly known as *GQuery* or *Entrez*) portal – select **All Databases** from the navigation bar at the top of the NCBI start page and click “Search” beside 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 NCBI" 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.

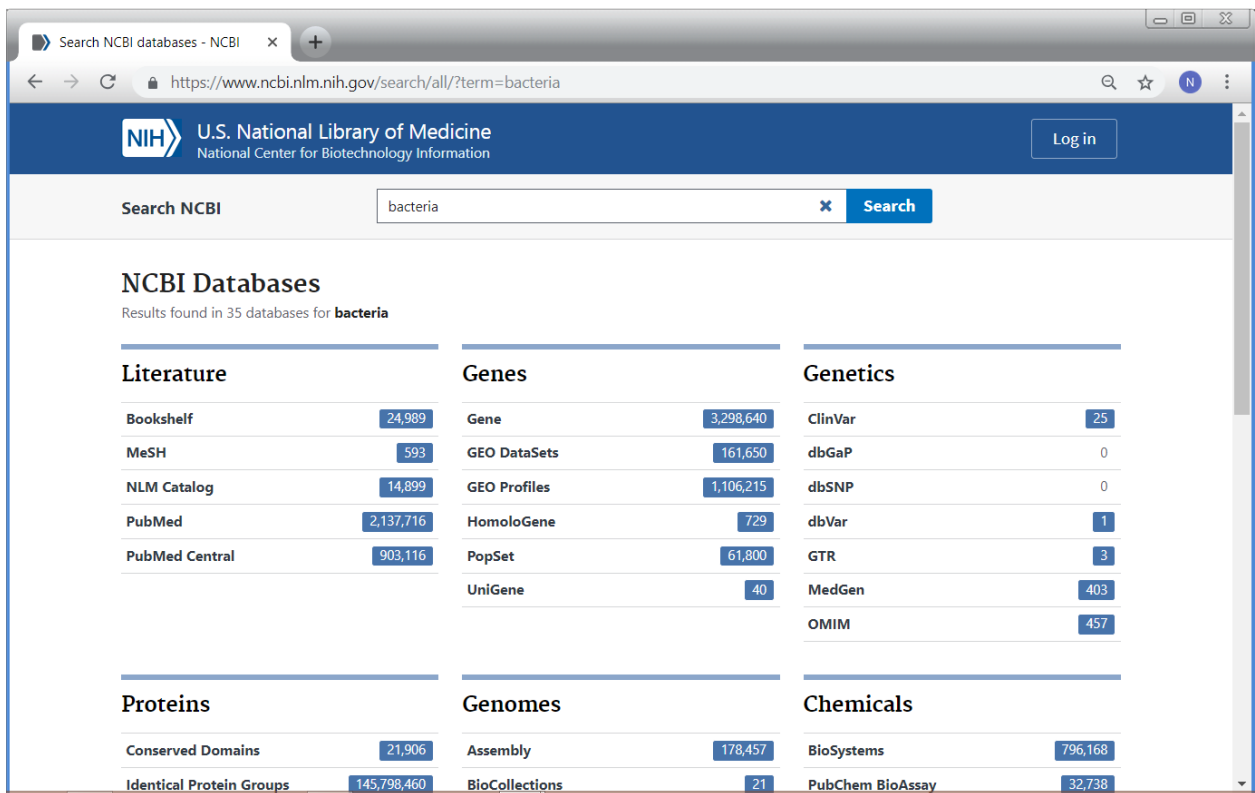


Figure 2. The *Search NCBI* portal page with *bacteria* used as a search word.

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 *Search NCBI* 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*

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.

Lab Quiz Question 1

*Answer lab quiz
questions while doing lab!

Box 1. Accession Numbers and 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 *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 for accessing a particular record, instead of the GI number. GI numbers are not to be confused with Entrez Gene IDs, which are another referencing system that NCBI uses entirely!

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'd like to check that out too).

1. At the bottom left of the NCBI homepage find the "NCBI Help Manual" link. Click on it. Then access the "Entrez Help" section.
2. You are now in Entrez Help. The Entrez collection of databases is queried when you use the *Search NCBI* interface. Note the contents that explain everything from search options to saving sets of records.
3. Notice that under the section **Entrez Searching Options** some other appropriate qualifiers are given, as illustrated on the previous page.

5. Search for our accession number of interest (e.g. NP_001318308 from above) through the *Search NCBI* portal page. It should give you **1** protein sequence hit. Click on it (it is a hyperlink) so that you get its full GenBank description (you can also click on the "[armadillo/beta-catenin repeat protein \[Arabidopsis thaliana\]](#)" link at the top of the page as the NCBI system recognizes that you've entered a protein identifier and hence provides some summary information for that above the numerical overview of results).

GenPept ▾ Send to: ▾ **Change region shown** ▾

armadillo/beta-catenin repeat protein [Arabidopsis thaliana]

NCBI Reference Sequence: NP_001318308.1
[Identical Proteins](#) [FASTA](#) [Graphics](#)

[Go to: \(v\)](#)

LOCUS	NP_001318308	582 aa	linear	PLN 14-FEB-2019
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; Pentapetales; rosids; malvids; Brassicales; Brassicaceae; Camelineae; Arabidopsis. 1 (residues 1 to 582)			
REFERENCE	1 (residues 1 to 582) 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. Sequence and analysis of chromosome 2 of the plant Arabidopsis thaliana NATURE 402 (6763), 761-768 (1999) 10617197			
REFERENCE	2 (residues 1 to 582) NCBI Genome Project Direct Submission Submitted (20-MAR-2017) National Center for Biotechnology Information, NIH, Bethesda, MD 20894, USA 3 (residues 1 to 582) 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. Direct Submission Submitted (17-MAY-2016) Plant Genomics, J. Craig Venter Institute, 9704 Medical Center Dr, Rockville, MD 20850, USA Protein update by submitter 4 (residues 1 to 582) Swarbreck,D., Lamesch,P., Wilks,C. and Huala,E. TAIR Direct Submission Submitted (18-FEB-2011) Department of Plant Biology, Carnegie Institution, 260 Panama Street, Stanford, CA, USA			
REMARK	REVIEWED REFSEQ : This record has been curated by TAIR and Araport. The reference sequence is identical to AEC08178 . Method: conceptual translation.			
FEATURES	<p>source</p> <p>1..582 /organism="Arabidopsis thaliana" /db_xref="taxon:3702" /chromosome="2" /ecotype="Columbia"</p> <p>1..582 /product="armadillo/beta-catenin repeat protein" /calculated_mol_wt=63916</p> <p>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),</p>			

[Protein](#)

CDS

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

Turn Off Clear

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.
- What is the taxonomic lineage of your organism?*
 - Has the genome of this organism been sequenced, i.e. is there a Genome Project?*
 - 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).
- 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.

CONSRTH NCBI Genome Project
 TITLE Direct Submission
 JOURNAL Submitted (20-MAR-2017) 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.
 CONSRTH TAIR
 TITLE Direct Submission

Related information

- ...
- Conserved Domains (Full)
- Domain Relatives
- Encoding mRNA
- Gene
- GeneView in dbSNP
- Genome
- PubMed (RefSeq)
- PubMed (Weighted)

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.
- 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)*
 - How many exons do you see in this gene? Tip: how many green boxes are there?*
 - What are the names of the genes surrounding it (i.e. what is its “Genomic context”)?*
 - 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)*
 - 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 18-Apr-2019

Summary

Gene symbol: PUB12
Gene description: armadillo/beta-catenin repeat protein
Primary source: [Aradip: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; Pentapetales; rosids; malvids; Brassicales; Brassicaceae; Camelinae; Arabidopsis
Also known as: AtPUB12, F8N16.12, F8N16.12, PLANT U-BOX 12
Summary: Encodes a U-box E3 ubiquitin ligase involved in ubiquitination of pattern recognition receptor FLS2.

Genomic context

Location: chromosome: 2
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. [Regulation of Arabidopsis brassinosteroid receptor BRI1 endocytosis and degradation by plant U-box PUB12/PUB13-mediated ubiquitination.](#)
Zhou J, et al. Proc Natl Acad Sci U S A. 2018 Feb 20. PMID 29432171. Free PMC Article
2. [Arabidopsis ubiquitin ligase PUB12 interacts with and negatively regulates Chitin Elicitor Receptor Kinase 1 \(CERK1\).](#)
Yamaguchi K, et al. PLoS One. 2017. PMID 29182677. Free PMC Article
3. [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
4. [Direct ubiquitination of pattern recognition receptor FLS2 attenuates plant innate immunity.](#)
Lu D, et al. Science. 2011 Jun 17. PMID 21680842. Free PMC Article

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

[Aradip](#)

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 RNA" record for the corresponding encoding 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
[FASTA](#) [Graphics](#)

[Go to:](#)

LOCUS NM_001336190 1949 bp mRNA linear PLN 14-FEB-2019
DEFINITION Arabidopsis thaliana armadillo/beta-catenin repeat protein (PUB12), mRNA.
ACCESSION NM_001336190 GI:1063699356
VERSION NM_001336190.1
DBLINK BioProject: [PRJNA116](#)
 BioSample: [SAMN03081427](#)
KEYWORDS RefSeq.
SOURCE Arabidopsis thaliana (thale cress)
ORGANISM Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons; Gunneridae; Pentapetales; 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.H., Koo, H.L., Moffatt, K.S., Cronin, L.A., Shen, M., Pal, O., 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 [10617187](#)
REFERENCE 2 (bases 1 to 1949)
CONSTRM NCBI Genome Project
TITLE Direct Submission
JOURNAL Submitted (14-FEB-2019) 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., Hicklem, 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 Saarbreck, D., Lamesch, P., Wilks, C. and Huala, E.
CONSTRM 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
 source
 1..1949
 /organism="Arabidopsis thaliana"
 /mol_type="mRNA"
 /db_xref="taxon:3702"
 /chromosome="2"
 /ecotype="Columbia"

Change region shown

Customize view

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 (DI IR17) mRNA 30,000,000

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

Box 3. Helpful Hints for NCBI searches

On most NCBI search pages (except, oddly, *Search NCBI*) click on “Save Search” or “Create Alert” 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; Altschul *et al.*, 1997) 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 use the nucleotide sequence corresponding to the protein sequence, not the protein sequence. First try finding the gene’s DNA sequence using Search NCBI tool again.

- On the *Search NCBI (GQuery)* Portal page, search “All Databases” 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 - NCBI

https://www.ncbi.nlm.nih.gov/search/all/?term=NP_001318308

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

Search NCBI NP_001318308 Search

PROTEIN SEQUENCE

armadillo/beta-catenin repeat protein [Arabidopsis thaliana]

thale cress

582 aa protein

NP_001318308.1

FASTA Identical proteins Gene

Download

NCBI Databases

Results found in 4 databases for NP_001318308

Literature	Genes	Genetics
Bookshelf 0	Gene 1	ClinVar 0
MeSH 0	GEO DataSets 0	dbGaP 0
NLM Catalog 0	GEO Profiles 0	dbSNP 0
PubMed 0	HomoloGene 0	dbVar 0
PubMed Central 0	PopSet 0	GTR 0
	UniGene 0	MedGen 0
		OMIM 0

Proteins	Genomes	Chemicals
Conserved Domains 0	Assembly 0	BioSystems 49
Identical Protein Groups 1	BioCollections 0	PubChem BioAssay 0
Protein 1	BioProject 0	PubChem Compound 0

Figure 7. Search NCBI 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).

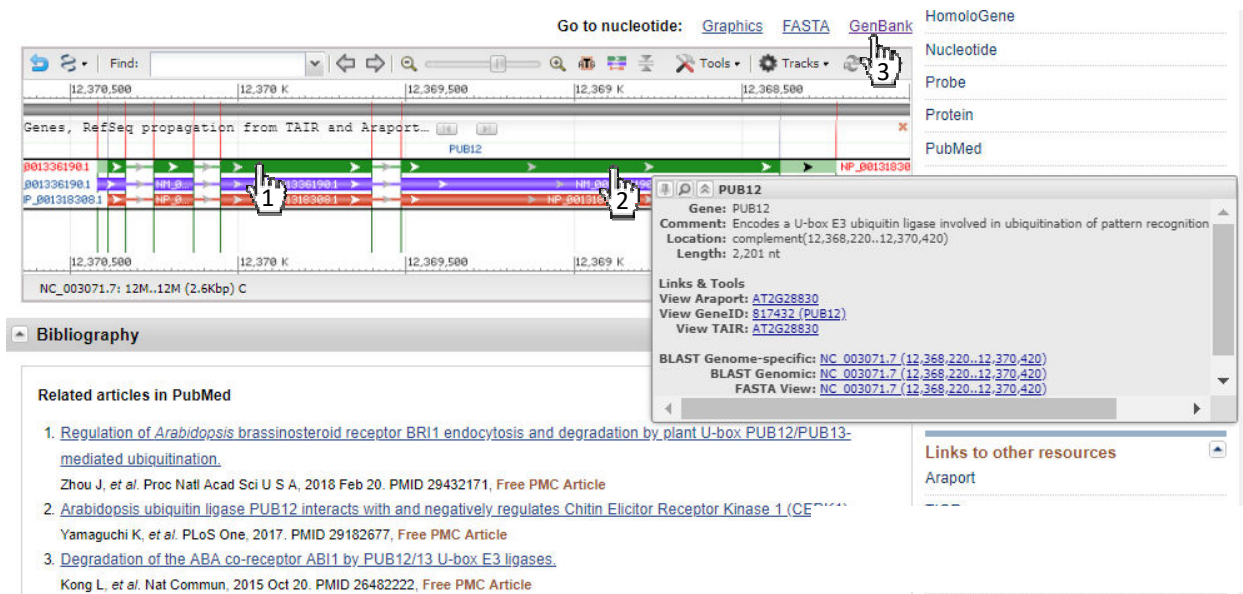


Figure 8. Part of the Gene page for NP_001318308, showing pop-up to sequence links. 1. Click the green bars to make mRNA and protein tracks appear; 2. hover over the mRNA track to see info panel; 3. Click “Genbank” link to see Genbank record for the genomic region for this gene.

- Click on the RefSeq RNAs link in the “Related information” panel on the right. This takes you to the mRNA that encodes the protein you have been looking at (we are accessing the same record you accessed in Step 10 of the first part of the lab). 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?
- Above the Sequence Viewer panel, click on the “Go to nucleotide: Genbank” link (see Step 3. in Figure 8 above). You will be taken 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 1–86, 170–286, 370–819, and 906–2201 and the **CDS** feature corresponds to nucleotide positions 33–86, 170–286, 370–819, and 906–2033. You may have remarked that the sequence from the chromosome has been reverse complemented.
 - b. Again, why are these different? Tip: recall the Central Dogma of Molecular Biology!



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.

```
>NM_001336190.1:33-1781 Arabidopsis thaliana armadillo/beta-catenin repeat protein
(PUB12), mRNA
ATGCTAAGGATTGCTTCTTTCTGTTAGCCATGTTAGCAAAATTTACCTGGTGTGTGTTGGAGAGAGATC
AAGTGATGGTGAAATTCAGAAAGTGACTTCTCTATTGGAACAAGCTTTAAGTATAATCCCTTATGAGAA
TCTGGAATTTTCAGATGAACCTAAAGAACAGGTGGAGCTTGTGTTAGTTACAGTTAAGAAGATCGTTAGGA
AAACGCGGTGGCGATGTGTATGATGATGAGTTGTATAAGGATGTTCTATCTTTATAGTGGTAGAGGTA
GTGTAATGGAGTCTGATATGGTTAGGAGAGTGGCGGAGAAGCTTCAGTTGATGACTATAACTGACCTTAC
GCAAGAGTCATTGGCTTTACTTGACATGGTTAGTTCTAGTGGTGGTATGATCCTGGTGAAAGTTTGAG
AAGATGTCTATGGTTCTTAAGAAGATTAAGGACTTTGTGCAAACCTATAATCCCTAAGTTGGATGATGCTC
CATTGAGACTGAAATCATCGCTTCCGAAGTCGCGAGATGATGATCGAGATATGCTAATCCGCCTGAAGA
GTTCCGTTGTCCAATATCTCTAGAATTGATGACTGATCCAGTTATTGTTTCTTCAGGGCAGACTTATGAA
CGTGAGTGCATTAAGAAGTGGCTTGAAGGAGGACACTTGACGTGTCCAAAGACGCAAGAAACGCTGACAA
GCGATATCATGACACCAAACTATGTTCTAAGAAGCCTTATAGCTCAATGGTGTGAGTCCAATGGCATCGA
ACCTCCAAAGCGTCCCAACATATCTCAACCGAGTAGTAAGGCCTCATCTTCGTGTCAGCCCTGATGAT
GAACATAACAAGATTGAAGAAGTTCTACTTAAGCTCACATCGCAACAGCCTGAAGACCGAAGATCTGCTG
CAGGAGAAATCCGTCTCTAGCAAAACAAACAATCATAACCGAGTCGCCATTGCTGCCTCAGGCGCGAT
CCCTCTCTGTTGGAATCTCTCACGATATCTAATGACTCTCGGACTCAAGAACACGCTGTGACATCGATT
CTTAACCTCTCGATATGTCAAGAGAACAAGGAAGATTGTTTATTCATCTGGAGCAGTTCCAGGTATTG
TTCATGTGCTTCAGAAAGGTAGCATGGAAGCTAGAGAAAACGCAGCAGTACACTTTTCAGCCTCTCGGT
TATAGACGAGAACAAGTGACAATAGGTGCCGAGGAGCGATCCCGCCTCTTGTGACCTTGCTGAGCGAA
GGATCACAGAGAGGCAAAAAAGACGCGGCAACTGCTCTGTTAATCTCTGCATATTTCAAGGAAACAAAG
GAAAAGCTGTGAGAGCCGTTTAGTTCCCGTCTAATGAGGTTACTAACAGAACCCGAAAGCGGAATGGT
TGATGAATCACTCTCGATATTAGCCATACTCATCGAGTCATCCGACGGGAAATCAGAGGTTGGAGCCGCT
GATGCAAGTCCAGTTCTGGTAGATTTTATAAGAAGCGGGTCACCGCGGAACAAAGAAAACCTCAGCTGCGG
TATTAGTGCATTGTGTTTCATGGAATCAGCAACATTTGATTGAAGCTCAGAAATTAGGGATTATGGATCT
TTTAATAGAAATGGCTGAGAATGGTACTGACAGAGGAAAACGCAAGCGGCACAGTTACTTAACCGCTTT
AGCCGTTTAAACGACCAGCAGAAACAACACTCTGGTTAGGTTTGAAGATCAAATCTCCCTAATCTGA
```

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...

The screenshot displays the NCBI BLAST Standard Nucleotide BLAST interface. At the top, the NIH and NCBI logos are present. The main heading is 'BLAST >> blastn suite'. Below this, the 'Standard Nucleotide BLAST' title is shown. The 'Enter Query Sequence' section contains a text box with 'NM_001336190.1' and a 'Query subrange' section with 'From: 33' and 'To: 1781'. There are links for 'Clear' and 'more...'. Below this is an 'Or, upload file' section with a 'Choose File' button and a 'Job Title' field. The 'Choose Search Set' section includes a 'Database' dropdown set to 'Nucleotide collection (nr/nt)', an 'Organism' field, and checkboxes for 'Exclude' and 'Limit to'. The 'Program Selection' section has radio buttons for 'Highly similar sequences (megablast)', 'More dissimilar sequences (discontiguous megablast)', and 'Somewhat similar sequences (blastn)', with the latter selected. At the bottom, a 'BLAST' button is next to the text 'Search database Nucleotide collection (nr/nt) using Blastn (Optimize for somewhat similar sequences)'. A link for 'Algorithm parameters' is at the very bottom.

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)

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?
- 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.

Lab Quiz
Question 2

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	NM_001336190:Arabidopsis thaliana armadillo/beta-catenin...		
RID	RZNNHFAV015	Search expires on 09-18 02:36 am	Download All ▼
Program	BLASTN ?	Citation ▼	
Database	nr	See details ▼	
Query ID	NM_001336190.1		
Description	Arabidopsis thaliana armadillo/beta-catenin repeat protein (...		
Molecule type	nucleic acid		
Query Length	1748		
Other reports	Distance tree of results	MSA viewer	?

Descriptions	Graphic Summary	Alignments	Taxonomy
--------------	------------------------	------------	----------

Figure 13. Blastn output Search Summary.

- Explore the **Graphic Summary** tab. 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 when you hover on an entry?
 - 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?

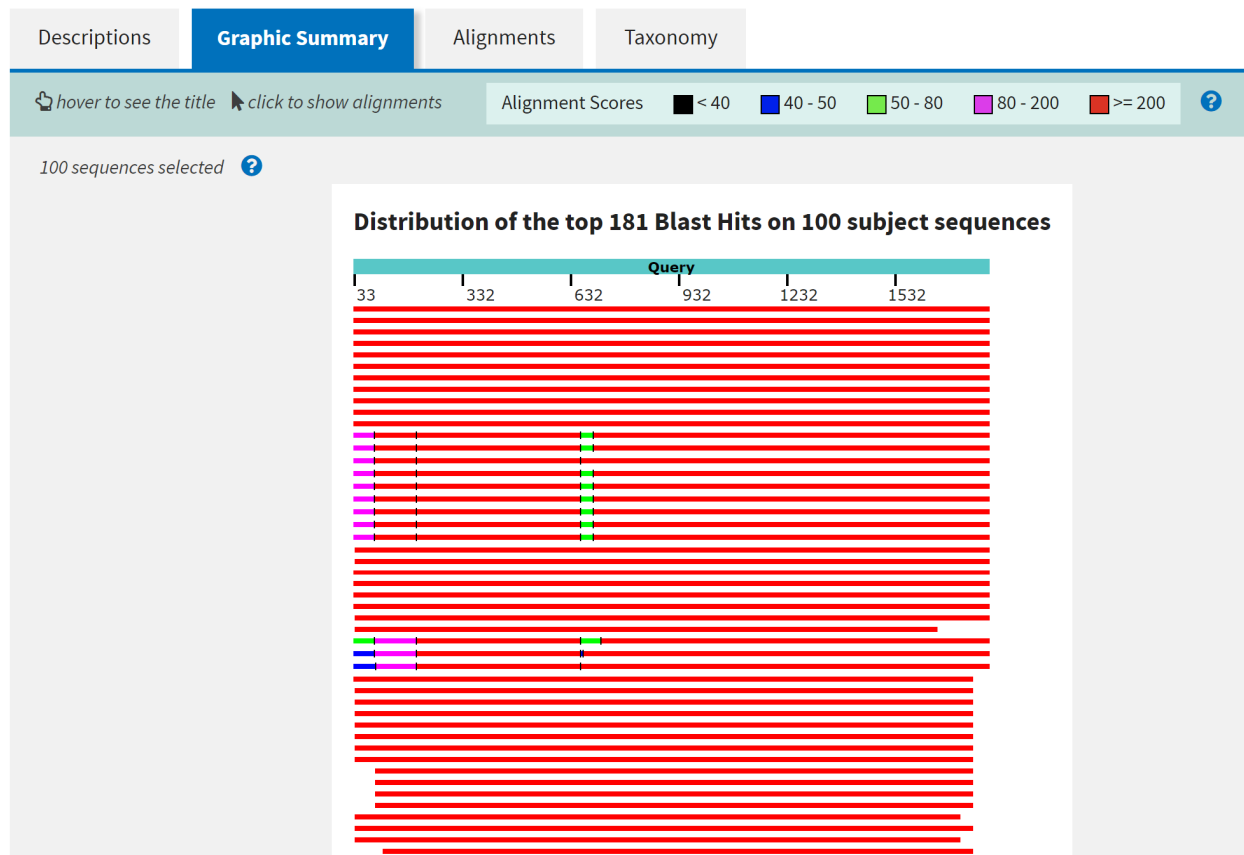


Figure 14. Blastn output Graphic Summary.

- The **Descriptions** tab lists the following:
 - **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?*

<div> <div>Descriptions</div> <div>Graphic Summary</div> <div>Alignments</div> <div>Taxonomy</div> </div>							
Sequences producing significant alignments				Download	Manage Columns	Show	100
<input checked="" type="checkbox"/> select all 100 sequences selected				GenBank	Graphics	Distance tree of results	
	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
<input checked="" type="checkbox"/>	Arabidopsis thaliana armadillo/beta-catenin repeat protein (PUB12), mRNA	3155	3155	100%	0.0	100.00%	NM_001336190.1
<input checked="" type="checkbox"/>	Arabidopsis thaliana mRNA for hypothetical protein, complete cds, clone: RAFL07-96-H19	3136	3136	100%	0.0	99.77%	AK226821.1
<input checked="" type="checkbox"/>	Arabidopsis thaliana unknown protein (At2g28830) mRNA, complete cds	3131	3131	100%	0.0	99.71%	AY035038.1
<input checked="" type="checkbox"/>	PREDICTED: Arabidopsis lyrata subsp. lyrata U-box domain-containing protein 12 (LOC9:2608)	2608	2608	99%	0.0	93.14%	XM_021028953.1
<input checked="" type="checkbox"/>	PREDICTED: Arabidopsis lyrata subsp. lyrata U-box domain-containing protein 12 (LOC9:2590)	2590	2590	99%	0.0	92.91%	XM_002880967.2
<input checked="" type="checkbox"/>	PREDICTED: Camelina sativa U-box domain-containing protein 12 (LOC104700639), tran	2386	2386	99%	0.0	90.33%	XM_019227546.1
<input checked="" type="checkbox"/>	PREDICTED: Camelina sativa U-box domain-containing protein 12 (LOC104700639), tran	2367	2367	99%	0.0	90.10%	XM_019227545.1
<input checked="" type="checkbox"/>	PREDICTED: Camelina sativa U-box domain-containing protein 12 (LOC104786817), mR	2353	2353	99%	0.0	89.93%	XM_010512272.2
<input checked="" type="checkbox"/>	PREDICTED: Camelina sativa U-box domain-containing protein 12-like (LOC104750012),	2301	2301	99%	0.0	89.07%	XM_010471742.2
<input checked="" type="checkbox"/>	PREDICTED: Capsella rubella U-box domain-containing protein 12 (LOC17887788), trans	2299	2299	99%	0.0	89.21%	XM_023784454.1
<input checked="" type="checkbox"/>	PREDICTED: Capsella rubella U-box domain-containing protein 12 (LOC17887788), trans	2280	2280	99%	0.0	88.98%	XM_023784453.1
<input checked="" type="checkbox"/>	Arabidopsis thaliana genome assembly, chromosome: 2	2039	3318	100%	0.0	100.00%	LR699746.2
<input checked="" type="checkbox"/>	Arabidopsis thaliana chromosome 2 sequence	2039	3318	100%	0.0	100.00%	CP002685.1
<input checked="" type="checkbox"/>	Arabidopsis thaliana chromosome 2 clone F8N16 map mi54, complete sequence	2039	3175	100%	0.0	100.00%	AC005727.3
<input checked="" type="checkbox"/>	Arabidopsis thaliana genome assembly, chromosome: 2	2025	3269	100%	0.0	99.73%	LR699761.1
<input checked="" type="checkbox"/>	Arabidopsis thaliana genome assembly, chromosome: 2	1912	3169	100%	0.0	97.52%	LR699766.1
<input checked="" type="checkbox"/>	Arabidopsis thaliana genome assembly, chromosome: 2	1912	3174	100%	0.0	97.52%	LR699756.1
<input checked="" type="checkbox"/>	Arabidopsis thaliana genome assembly, chromosome: 2	1912	3178	100%	0.0	97.52%	LR699751.1
<input checked="" type="checkbox"/>	Arabidopsis thaliana genome assembly, chromosome: 2	1908	3156	100%	0.0	97.43%	LR699771.1
<input checked="" type="checkbox"/>	Arabidopsis thaliana genome assembly, chromosome: 2	1908	3156	100%	0.0	97.43%	LR215053.1
<input checked="" type="checkbox"/>	PREDICTED: Eutrema salsudineum U-box domain-containing protein 12 (LOC18027496),	1839	1839	99%	0.0	83.37%	XM_006409854.2

Figure 15. Blastn output descriptions

Lab Quiz
Question 3

- Finally we can explore the actual HSP **Alignments** in the Alignments tab.
 - 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.
 - What do the vertical bars (|) represent between the **Query** and the **Subject** (database sequence)?
 - 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 scroll down to the HSP alignment section.
 - Describe the difference between this format and the previous format. Can you imagine cases where the different formats might be most useful?
 - 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?

Download

GenBank

Graphics

Next

Previous

Descriptions

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

Sequence ID: [NM_001336190.1](#) Length: 1949 Number of Matches: 1

Range 1: 33 to 1781

[GenBank](#)

[Graphics](#)

Next Match

Previous Match

Score	Expect	Identities	Gaps	Strand
3155 bits(3498)	0.0	1749/1749(100%)	0/1749(0%)	Plus/Plus
Query 33	ATGCTAAGGATTGCTTTCTTTTCGTTAGCCATGTTAGCAAAATTTACCTGGTGTGTGTG	92		
Sbjct 33	ATGCTAAGGATTGCTTTCTTTTCGTTAGCCATGTTAGCAAAATTTACCTGGTGTGTGTG	92		
Query 93	GAGAGAGATCAAGTGATGGTGAAATTCAGAAAGTGACTTCTCTATTGGAAACAAGCTTTA	152		
Sbjct 93	GAGAGAGATCAAGTGATGGTGAAATTCAGAAAGTGACTTCTCTATTGGAAACAAGCTTTA	152		
Query 153	AGTATAATCCCTTATGAGAATCTGGAAATTCAGATGAACCTAAAGAACAGGTGGAGCTT	212		
Sbjct 153	AGTATAATCCCTTATGAGAATCTGGAAATTCAGATGAACCTAAAGAACAGGTGGAGCTT	212		
Query 213	GTTTTAGTTTCAGTTAAGAAGATCGTTAGGAAAACGCGGTGGCGATGTGTATGATGATGAG	272		
Sbjct 213	GTTTTAGTTTCAGTTAAGAAGATCGTTAGGAAAACGCGGTGGCGATGTGTATGATGATGAG	272		
Query 273	TTGTATAAGGATGTTCTATCTCTTTATAGTGGTAGAGGTAGTGTAAAGGAGCTGATATG	332		
Sbjct 273	TTGTATAAGGATGTTCTATCTCTTTATAGTGGTAGAGGTAGTGTAAAGGAGCTGATATG	332		
Query 333	GTTAGGAGAGTGGCGGAGAACTTCAGTTGATGACTATAACTGACCTTACGCAAGAGTCA	392		
Sbjct 333	GTTAGGAGAGTGGCGGAGAACTTCAGTTGATGACTATAACTGACCTTACGCAAGAGTCA	392		
Query 393	TTGGCTTACTTGACATGGTTAGTTCTAGTGGTGGTGATGATCCTGGTGAAAGTTTGGAG	452		
Sbjct 393	TTGGCTTACTTGACATGGTTAGTTCTAGTGGTGGTGATGATCCTGGTGAAAGTTTGGAG	452		
Query 453	AAGATGCTCTATGGTTCTTAAGAAGATTAAAGGACTTTGTGCAAACTTATAATCCTAACTTG	512		
Sbjct 453	AAGATGCTCTATGGTTCTTAAGAAGATTAAAGGACTTTGTGCAAACTTATAATCCTAACTTG	512		
Query 513	GATGATGCTCATTGAGACTGAAATCATCGCTCCGAAGTCGCGAGATGATGATCGAGAT	572		
Sbjct 513	GATGATGCTCATTGAGACTGAAATCATCGCTCCGAAGTCGCGAGATGATGATCGAGAT	572		
Query 573	ATGCTAATTCGCTGAAGAGTTCCGTTGTCCAATATCTCTAGAATTGATGACTGATCCA	632		
Sbjct 573	ATGCTAATTCGCTGAAGAGTTCCGTTGTCCAATATCTCTAGAATTGATGACTGATCCA	632		
Query 633	GTTATTGTTTCTTCAGGGCAGACTTATGAACGTGAGTGCAATTAAGAAGTGGCTTGAAGGA	692		
Sbjct 633	GTTATTGTTTCTTCAGGGCAGACTTATGAACGTGAGTGCAATTAAGAAGTGGCTTGAAGGA	692		
Query 693	GGACACTTGACGTGTCGCAAGACGCAAGAAACGCTGACAAGCGATATCATGACACCAAAAC	752		
Sbjct 693	GGACACTTGACGTGTCGCAAGACGCAAGAAACGCTGACAAGCGATATCATGACACCAAAAC	752		
Query 753	TATGTTCTAAGAAGCCTTATAGCTCAATGGTGTGAGTCCAATGGCATCGAACCTCCAAAG	812		
Sbjct 753	TATGTTCTAAGAAGCCTTATAGCTCAATGGTGTGAGTCCAATGGCATCGAACCTCCAAAG	812		
Query 813	CGTCCCAACATATCTCAACCGAGTAGTAAGGCCATCTTCTGTCGTGAGCCCTGATGAT	872		
Sbjct 813	CGTCCCAACATATCTCAACCGAGTAGTAAGGCCATCTTCTGTCGTGAGCCCTGATGAT	872		
Query 873	GAACATAACAAGATTGAAGAACTTCTACTTAAGCTCACATCGCAACAGCCTGAAGACCGA	932		
Sbjct 873	GAACATAACAAGATTGAAGAACTTCTACTTAAGCTCACATCGCAACAGCCTGAAGACCGA	932		
Query 933	AGATCTGCTGCAGGAGAAATCCGTTCTTAGCAAAAACAAATCATAACCGAGTCGCC	992		
Sbjct 933	AGATCTGCTGCAGGAGAAATCCGTTCTTAGCAAAAACAAATCATAACCGAGTCGCC	992		
Query 993	ATTGCTGCCTCAGGCGCGATCCCTCTTCTGGTGAATCTCCTCACGATATCTAATGACTCT	1052		
Sbjct 993	ATTGCTGCCTCAGGCGCGATCCCTCTTCTGGTGAATCTCCTCACGATATCTAATGACTCT	1052		

Related Information

[Gene](#) - associated gene details

[New Genome Data Viewer](#) - aligned genomic context

Related Information

[Gene](#) - associated gene details
[New Genome Data Viewer](#) - aligned genomic context

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 Search NCBI / 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. Yearbook 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.