

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. The **Science Primer** provides a nice introduction to some of the tools and methods used. You can also check out the *NCBI handbook* (<http://www.ncbi.nlm.nih.gov/books/NBK21101/>) for more information.

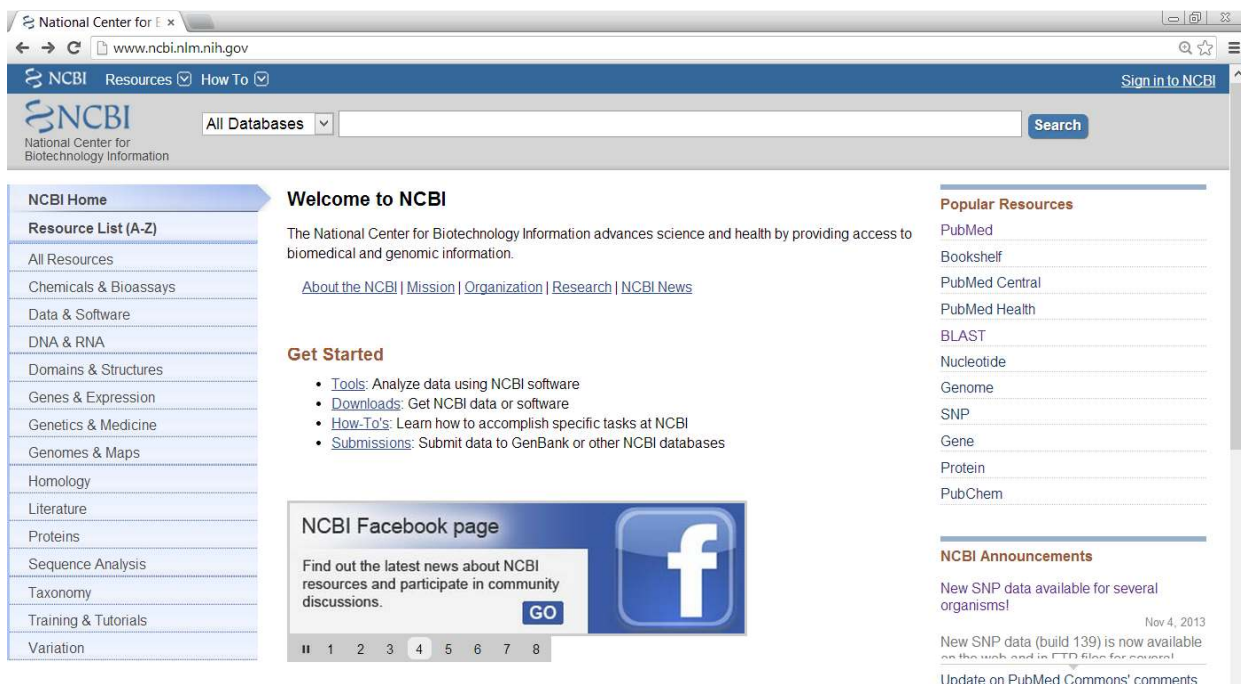


Figure 1. The NCBI homepage.

- Now let's move to the *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 *GQuery*. 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. A search of *bacteria* gives thousands of hits – not very helpful. We need specifics.

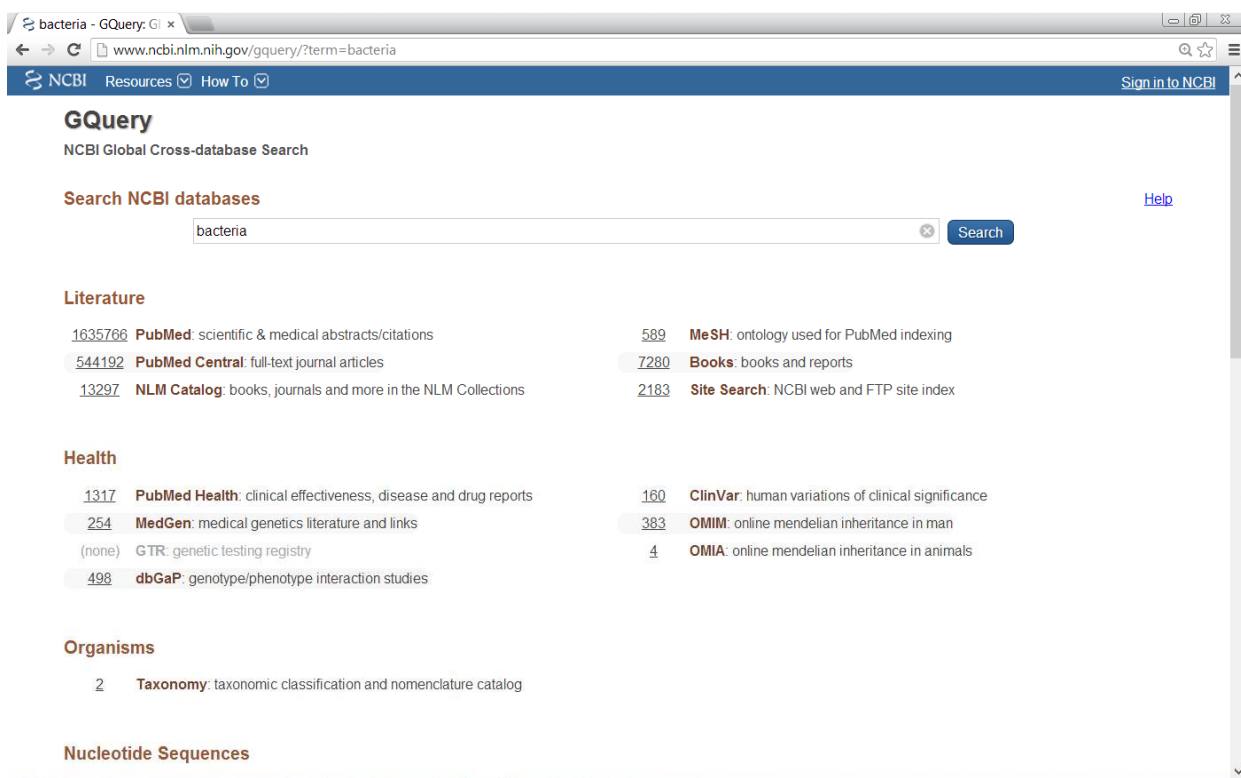


Figure 2. The *GQuery* portal page with *bacteria* used as a search word.

- 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_565676, and it is a structural component of the ribosome.

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:

- gene keywords
e.g. *structural constituent of ribosome*
- gene keyword AND organism
e.g. *structural constituent of ribosome AND Arabidopsis thaliana*
- gene keyword [PROT] AND organism [ORGN]
e.g. *structural constituent of ribosome [PROT] AND Arabidopsis thaliana [ORGN]*
- accession or gi number
e.g. *NP_565676*

Lab Quiz
Question 1

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 gi or 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") is simply a series of digits that are assigned consecutively to each sequence record processed by NCBI. The GI system of identifiers runs parallel to the accession.version system; therefore, if the DNA or protein sequence changes in any way, it will receive a new GI number,

Example: When a new entry is submitted to GenBank it will be assigned an accession number (say AF000001). Since this is the first version the Accession will be appended with ‘.1’, so it will look like AF000001.1. At the same time it will be given a GI number (say GI:1234567). Now imagine that the researcher who originally submitted the record wants to update the information. The updated record will keep the same Accession number, but increase in version number (AF000001.2), which the new record will be 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 will always take you back 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.

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 your given accession number through the *GQuery* portal page (e.g. NP_565676 from above). It should give you one protein sequence hit. Click on it and the following link so that you get its full GenBank description.

Protein Protein Limits Advanced

[Display Settings:](#) ☒ GenPept [Send to:](#) ☒

plant U-box 24 protein [Arabidopsis thaliana]

NCBI Reference Sequence: NP_565676.4
[FASTA](#) [Graphics](#)

[Go to:](#) ☒

LOCUS NP_565676 962 aa linear PLN 28-MAY-2011
 DEFINITION plant U-box 24 protein [Arabidopsis thaliana].
 ACCESSION NP_565676
 VERSION NP_565676.4 GI:240254552
 DBSOURCE REFSEQ: accession [NM_128442.5](#)
 KEYWORDS .
 SOURCE Arabidopsis thaliana (thale cress)
 ORGANISM [Arabidopsis thaliana](#)
 Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
 Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons;
 rosids; malvids; Brassicales; Brassicaceae; Camelineae;
 Arabidopsis.
 REFERENCE 1 (residues 1 to 962)
 AUTHORS Lin,X., Kaul,S., Rounsley,S., Shea,T.F., Benito,M.I., Town,C.D.,
 Fujii,C.Y., Mason,T., Bowman,C.L., Barnstead,M., Feldblum,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 962)
 CONSRM Arabidopsis TAIR10 Release
 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. The
 reference sequence is identical to [AEC08178](#).
 On Jun 19, 2009 this sequence version replaced gi:[238622842](#).
 Method: conceptual translation.
 FEATURES
 Location/Qualifiers
 source
 1..962
 /organism="Arabidopsis thaliana"
 /db_xref="taxon:3702"
 /chromosome="2"
 /ecotype="Columbia"
[Protein](#)
 1..962
 /product="plant U-box 24 protein"
 /calculated_mol_wt=106787
[Region](#)
 262..298
 /region_name="RING"
 /note="RING-finger (Really Interesting New Gene) domain, a
 ...
 CDS
 1..962
 /gene="FUB12"
 /locus_tag="AT2G28830"
 /gene_synonym="AtFUB12; F8N16.12; F8N16_12; PLANT U-BOX
 12; FUB12"
 /coded_by="NM_128442.5:77..2965"
 /inference="Similar to DNA
 sequence:INSD:AY219111.2,INSD:DQ056551.1"
 /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"
 /note="PLANT U-BOX 12 (FUB12); 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),
 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 1692 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)."
 /db_xref="GeneID:817432"
 /db_xref="TAIR:AT2G28830"
 ORIGIN
 1 maksekhhkila qtlidseinei asidsdevtpm kkhcanlerr lalllpmlae irdnqessse
 61 vvnallsvkg silhakdlle fshvskiyl vlerdqvmvk fgkvtslleg alsiiipenl
 121 eisdelkeqv elivqlrrs lgrgddvdyd delykdvlsl ysgrgsvmes dmrvrrvaeik
 181 qlmtitdltg esallldmvs seggdpgges fekmsvmlkk ikdfvqtnp nlddaplrlk
 241 selpkdrdd rdmllppeef rcpisilelmt dpvivssggt yerecikkvl egghltcpkt
 301 qetltsdimt pnyvlrslia qwoceangiep pkprnpsqps skassssssap ddehknkieel
 361 llkltsgqpe drsaaageir llakqnnhnr vaiaasgaip llvnltism dertqehavt
 421 silnlsicqe nkgkivysg avpgivhvlq kysmearena aatlflslavi denkvitgaa

Figure 3. GenBank record for accession NP_565676, 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.

```

/product= plant U-BOX 23 protein
/calculated_mol_wt=106787
262..298
/region_name="RING"
/note="RING-finger (Really Interesting New Gene) domain, a
specialized type of Zn-finger of 40 to 60 residues that
binds two atoms of zinc; defined by the 'cross-brace'
motif C-X2-C-X(9-39)-C-X(1-3)-
H-X(2-3)-(N/C/H)-X2-C-X(4-48)C-X2-C; probably involved
in...; cd00162"
/db_xref="CDD:29102"
order(262,265,279,282,285,297)
/site_type="other"
/note="cross-brace motif"
/db_xref="CDD:29102"
Region
<353..426
/region_name="ARM"
/note="Armadillo/beta-catenin-like repeats. An
approximately 40 amino acid long tandemly repeated
sequence motif first identified in the Drosophila segment
polarity gene armadillo; these repeats were also found in
the mammalian armadillo homolog beta-catenin; cl02500"
/db_xref="CDD:207616"
order(370,374,377,381,417,420,423..424)
/site_type="other"
/note="protein binding surface [polypeptide binding]"
/db_xref="CDD:28904"
Site
392..509
/region_name="ARM"
/note="Armadillo/beta-catenin-like repeats. An
approximately 40 amino acid long tandemly repeated
sequence motif first identified in the Drosophila segment
polarity gene armadillo; these repeats were also found in
the mammalian armadillo homolog beta-catenin; cd00020"
/db_xref="CDD:28904"
order(416,420,424,455,459,462,466,500,503,506..507)
/site_type="other"
/note="protein binding surface [polypeptide binding]"
/db_xref="CDD:28904"
Region
480..592
/region_name="ARM"
/note="Armadillo/beta-catenin-like repeats. An
approximately 40 amino acid long tandemly repeated
sequence motif first identified in the Drosophila segment
polarity gene armadillo; these repeats were also found in
the mammalian armadillo homolog beta-catenin; cd00020"
/db_xref="CDD:28904"

```

LinkOut to external resources

Subcellular localisation in Arabidopsis thaliana
[Bio-Array Resource]

Expression profiles in Arabidopsis thaliana
[Bio-Array Resource]

The Arabidopsis Information Resource
[The Arabidopsis Information R...]

Related information

[BLink](#)

[Related Sequences](#)

[Identical Proteins](#)

[BioProject](#)

[BioSystems](#)

[CDD Search Results](#)

[Conserved Domains \(Concise\)](#)

[Conserved Domains \(Full\)](#)

[Encoding mRNA](#)

[Gene](#)

[Genome](#)

[HomoloGene](#)

[Map Viewer](#)

[Nucleotide](#)

[Protein Clusters](#)

[PubMed](#)

[PubMed \(RefSeq\)](#)

[PubMed \(Weighted\)](#)

[Related Structures \(List\)](#)

[Related Structures \(Summary\)](#)

[Taxonomy](#)

[UniGene](#)

Figure 4. The **Links** (“Related Information”) menu

9. Select **Gene** from the **Links** 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 position in the genome (tip: mouse-over the green bar, which represents the gene in the sequence viewer)?
 - b. What are the names of the genes surrounding it (genomic context)?
 - c. Does it have any conserved domains (scroll down to the Genome Annotation section)? What are they called?
 - d. What biological process (Gene Ontology terms) is this gene involved with (again, scroll down!)?

The screenshot displays the NCBI Gene database entry for PUB12 (AT2G28830) in Arabidopsis thaliana. The page is organized into several sections:

- Summary:** Provides basic information including the gene symbol (PUB12), description (plant U-box 24 protein), primary source (TAIR:AT2G28830), locus tag (AT2G28830), gene type (protein coding), RNA name (plant U-box 24 protein), RefSeq status (REVIEWED), organism (Arabidopsis thaliana), lineage, and also known as (AtPUB12; F8N16.12; F8N16_12; PLANT U-BOX 12; PUB12).
- Genomic context:** Shows the gene's location on Chromosome 2 (NC_003071.7) and a diagram of the genomic region with surrounding genes (RLS3, RT262651, PUB12, XBRT31, CYP71A3).
- Genomic regions, transcripts, and products:** Displays the genomic sequence (NC_003071) and a detailed view of the gene structure with exons and introns. The sequence viewer shows the gene's position on the chromosome and provides links to reference sequence details, graphics, FASTA, and GenBank.
- Related information:** A sidebar on the right lists various resources and links, including BioProjects, Conserved Domains, EST, Full text in PMC, Genome, GEO Profiles, HomoloGene, Map Viewer, Nucleotide, Probe, Protein, Protein Clusters, PubMed, PubMed (GeneRIF), RefSeq Proteins, RefSeq RNAs, Taxonomy, and UniGene.
- Links to other resources:** A section at the bottom right lists external databases such as KEGG, MIPS, TAIR, and TIGR.

Figure 5. GenBank Gene page for AT2G28830.

10. On the Gene page, there are also other links (see the sidebar on the right) to examine a gene's structure, function and phylogenetic relationships further.
 - Click on **Additional Links**. a. What kind of information does this section tell you?
 - Return to the Gene page and click on **Map Viewer** from the **Related Information** menu.
 - Use the selector on the left hand side of the screen to zoom in and out. Scroll along the genome to see the order of genes. Use the gene's locus tag (found on the Gene page) to

find your gene this way using the Search function. You may have to zoom in pretty far. You can also scroll along the genome by clicking the small up and down arrows at the top and bottom of the **Genes_seq** genome graphic.

- Click on the small black box pointed to by the gene's locus tag.
 - b. *How many exons do you see in this gene? Tip: this can also be determined from the Gene page's sequence viewer entry...how many green bars are there?*
- Go back to the Map Viewer.
- Click around and explore the variety of ways that data is interconnected and displayed (don't worry, you can't break anything).

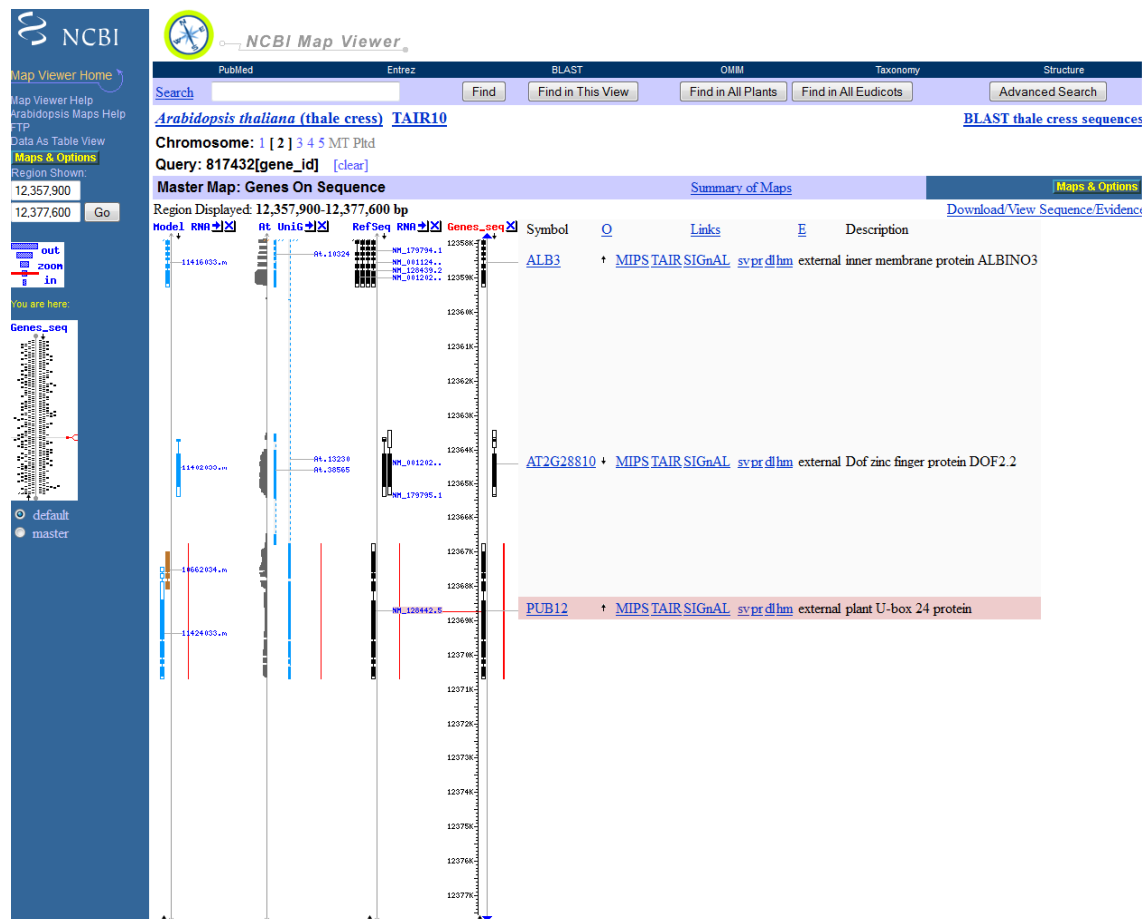


Figure 6. NCBI Map Viewer for part of *Arabidopsis thaliana* chromosome 2.

Box 3. Helpful Hints for GQuery searches

Go back to NCBI GQuery, search for your gene again using your given accession number. Click on “Save Search” beside 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 *myNCBI* 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 of interest.

Box 4. BLAST and Homology

The Basic Local Alignment and Search Tool (BLAST) is a very power 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 its corresponding nucleotide sequence, not its protein sequence. First try finding the gene's DNA sequence using GQuery again.
 - On the *GQuery* Portal (All Databases) page, search for your given protein sequence again using the Accession or GI number (or alternatively, go back to the search you saved in your NCBI account). Using the protein from the first part of this lab, we would search for **NP_565676**.
 - 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.

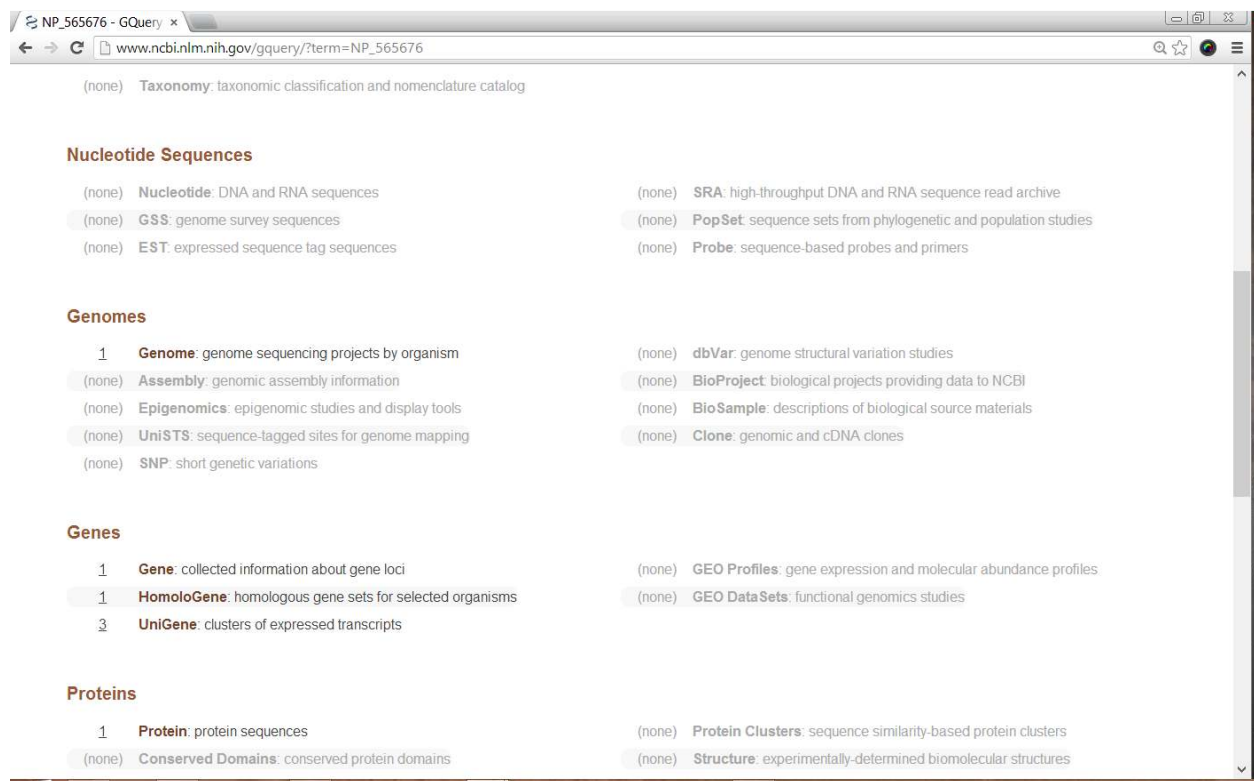



Figure 7. *GQuery* portal queried for NP_565676 (partial view).

- 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  icon in the pop up menu 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)

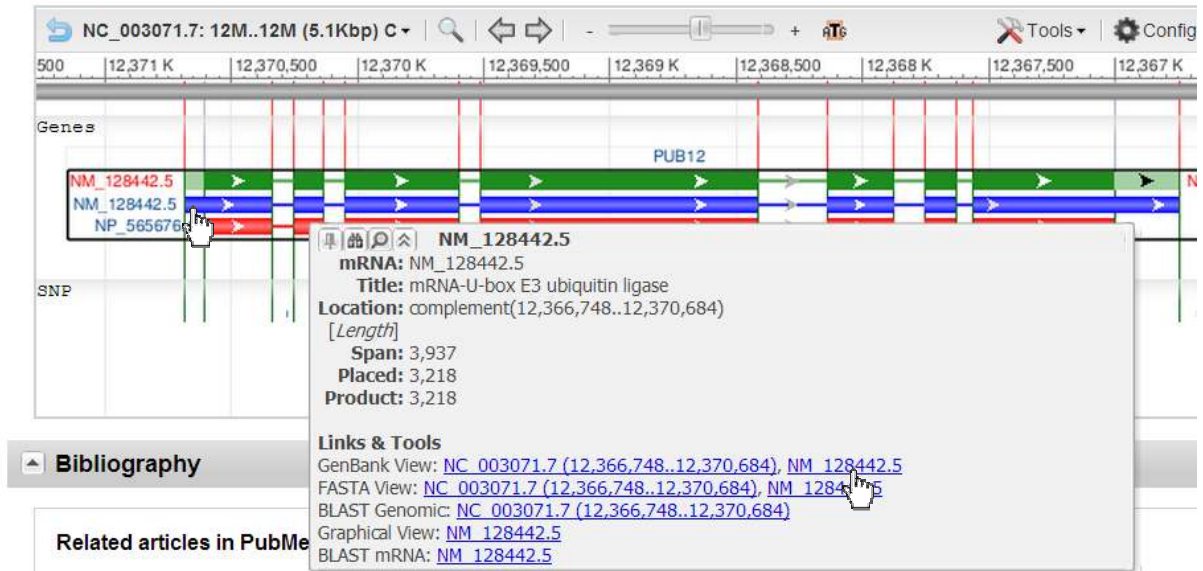


Figure 8. Part of the Gene page for NP_565676, showing pop-up to sequence links.

- Click on the mRNA link (**NM_128442.5** – the “M” in the accession number denotes mRNA) and select **GenBank View** (you may need to scroll to the right to access this link). This takes you to the mRNA that encodes the protein you have been looking at. Notice the feature list in the record. One feature is **gene**, and corresponds to base position 1 – 3218 on this record. Another features is the coding sequence (**CDS**), which corresponds to base position 77 – 2965.
 - 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**), 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 – 3937, while the **mRNA** feature corresponds to positions 1-814, 882-1007, 1129-1394, 1670-2768, 2855-3304, 3388-3504, and 3592-3937, while the **CDS** feature corresponds to positions 254-814, 882-1007, 1129-1394, 1670-2768, 2855-3304, 3388-3504, and 3592-3861.
 - Again, why are these different? Tip: recall the Central Dogma of Molecular Biology.

Figure 9. GenBank record for NM_128442 mRNA.

- Let's return the mRNA record we were previously working with (NM_128442). 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|240254551:77-2965 Arabidopsis thaliana plant U-box 24 protein (PUB12) mRNA,
complete cds
ATGGCGAAATCAGAGAAACACAAATTAGCTCAACCTTAATCGATTCAATAACGAGATCGCTTCAATTT
CCGATTCCGTTACACCGATGAAGAAGCACTGCGCTAACCTTCCCGCGGATTGTCGCTTCTTCTCTAT
GTTGGAAGAGATCAGAGACAATCAGGAATCATCTCGGAAGTAGTGAATGCTTTGTTATCTGTTAAGCAA
TCGCTTCTTCATGCTAAGGATTTGCTTTCTTTCGTTAGCCATGTTAGCAAAATTTACCTGGTGTGGAGA
GAGATCAAGTATGTTGAAATTTAGAAAGTGAATTTCTCTATTGGAACAGCTTTAAGTATAATCCCTTA
TGGAATCTGGAAATTTAGATGAATTAAGAACAGGTGGAGCTTGTTTAGTTTCAGTTAAGAGATCGG
TTAGGAACACGCGTGGCGATGTTATGATGATGAGTTGTTAAGGATGTTCTATCTTTATATGTTGTA
GAGGTAGTGTAAATGGAGTCTGATATGTTAGGAGAGTGGCGGAGAAGCTTCAGTTGATGACTATACTGA
CCTTACGCAAGAGTCATTGGCTTTACTTGACATGTTAGTTCTAGTGGTGGTGGTGGTGGTGGTGGTGGT
TTTGAGAAGATGCTATGTTCTTAAGAAGATTAAAGGACTTTGTGCAAACTTATAATCCTAACTTGGATG
ATGCTCCATTGAGACTGAATCATCGCTTCCGAACTCGCGAGATGATGATCGAGATATGCTAATTCGCGC
TGAAGAGTTCGTTGTTCCAAATATCTCTAGAATTGATGACTGATCCAGTTATTGTTTCTTCAGGGCAGACT
TATGAACGTGAGTGATTAAGAAGTGGCTTGAAGAGGACACTTGACGTGTCCAAAGACGCAAGAAACGC
TGACAAGCGATATCATGACCAAACTATGTTCTAAGAACCTTATAGCTCAATGGTGTGATCCAAATGG
CATCGAAGCTCCAAAGCGTCCCAACATATCTCAACGAGTAGTAAGGCTCATCTTCGTGCTCAGGCCCT
GATGATGAACATAACAGATTGAAGAATTTCTCTTAAGCTCACATCGCAACAGCTGAAGACCGAAGAT
CTGCTGAGGAGAAATCCGCTCTCTAGCAAAACAAACAATCATAACCGAGTCGCCATTGCTGCTCAGG
CGCGATCCCTCTTCTGTTGAATCTCTCACGATATCTAATGACTCTCGGACTCAAGAACACGCTGTGACA
TCGATTCTTAACCTCTCGATATGTCAGAGAACAAAGGGAAGATTGTTTATCATCTGGAGCAGTCCAG
GTATTGTTCAATGCTTTCAGAAAGGTAGCATGGAAGCTAGAGAAACGACAGCTACACTTTTCAGCCT
CTCGGTTATAGACGAGAACAAAGTGAACAATAGTGGCGGAGGAGCGATCCGCTCTTGTGACCTTGTG
AGCGAAGGATCAGAGAGGCAAAAAGACGCGGCAACTGCTCTGTTTAACTCTGATATTTCAAGGAA
ACAAAGGAAAGCTGTGAGAGCGGTTTAGTTCCCGTGCTAATGAGGTTACTAACAGAACCCGAAAGCGG
AATGGTTGATGAATCACTCTCGATATTAGCCATACTATCGAGTCATCCGAGCGGAAATCAGAGTTGGA
GCCGCTGATGAGTTCAGTCTCTGGTAGATTTTAAAGAAGCGGGTCACCGCGGAACAAAGAACTCAG
CTCGGTTATTAGTGCACCTTGTGTTCTAGGAATCAGCAACATTTGATTGAAGCTCAGAAATTAGGATTAT
GGATCTTTTAAAGAAATGGCTGAGAATGTTACTGACAGGAAACGCAAGCGGCACAGTTACTTAAAC
CGCTTTAGCCGTTTAAACGACCAAGCAGAAACAACTCTGTTTAAAGGCAAAATAAATGTACAAAAAA
ATGATTTAATCTTCAGCTTGAGCAGCAAGTTCTATAGATCTGAGAGCAGCAAGATGAGAGATTCATGTT
TAGCCGCGCTGTTGAACATCAACGTCAGATAGCCGAGGATTTCTAGCTTGTACCATCTCTATCTCCC
ACTGCTGTTCTGCTATGTTCTGTTCTTTCCGAAATTAAGTCTGTTGATTCCACTTCTCGATTCCCT
TTTTACTCCAGACTTCATCAATCCCAAGAGACTCTTGAAGAGTCCCTTAACTTAGAAGGCTTAAC
ATGTAACCAAGCGAAGAGAGATGATCTCTTTCCACAGATTAAACACACGCTTCTCAACACCACT
GGTTCCTGCTTCGACAGGTTCTTGGAACTTGGCAGTTCAGATGCACAACTTCTCCGCGAGAGTGAATC
GTGTGAGAGAGGTCCACGAGACTTCCAAACAGAAAGAAACAAACAAAAAAGTTCCGCTCAACGA
AAAGAAACAAAAAAGAGTTCCATCAGCGATATCCCAAGAAAGAAAGTTCCAGAAACAT
CATCAGAGGAAGATTAATAAGAGGATCTCTCAGGGGTATATTGTAGTAGATATGCTCTTCAACAC
TTGAACAGCTTGGATCACTCTAGACAAATAGAAAGCAGGACGACGAGCAATGACACGAAATATAGGACG
TGTTTAACTGTTGAGTTTATATTTGACAGAAACAGTTACAGTAAGACCTCCTGAAACGCGTATG
GGTCGTGGGAAAGGAGCTCCAGCGTTTTGGGTAGCTGTGGTTAAACAGGTAAATCAATTTAGAAATGG
GTGGTGTTCGAAAAAGTAGTAGAAGCTATTCTATAGCGCATCAAAGTTGCTGCAAAACCAA
ATTCATCATTTCTAAATAA
```

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. 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**). You could also copy-and-paste the FASTA formatted mRNA sequence you found in the previous step into the query box.

The screenshot displays the NCBI BLAST Standard Nucleotide BLAST interface. The top navigation bar includes links for Home, Recent Results, Saved Strategies, and Help. The main heading is 'Standard Nucleotide BLAST'. The 'Enter Query Sequence' section contains a text box with 'NM_128442.5' and a 'Browse...' button. The 'Choose Search Set' section has radio buttons for 'Human genomic + transcript', 'Mouse genomic + transcript', and 'Others (nr etc.)', with 'Others (nr etc.)' selected. Below this, there are fields for 'Organism' and 'Exclude' options. The 'Program Selection' section has radio buttons for 'Highly similar sequences (megablast)', 'More dissimilar sequences (discontiguous megablast)', and 'Somewhat similar sequences (blastn)', with 'Somewhat similar sequences (blastn)' selected. At the bottom, there is a 'BLAST' button and a checkbox for 'Show results in a new window'.

Figure 11. The blastn query page.

- 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 and 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

BLAST 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.
 - What is the **Expect threshold**?*
 - What would happen if you decreased it? Increased it?*
 - What would be the effect of increasing the **Word size**?*
 - 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 probability 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?*

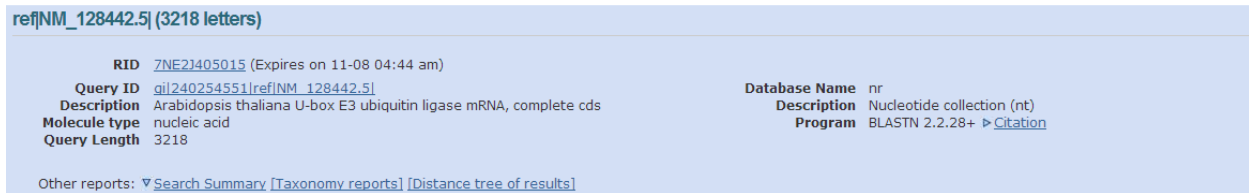


Figure 13. Blastn output search summary.

- 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?*

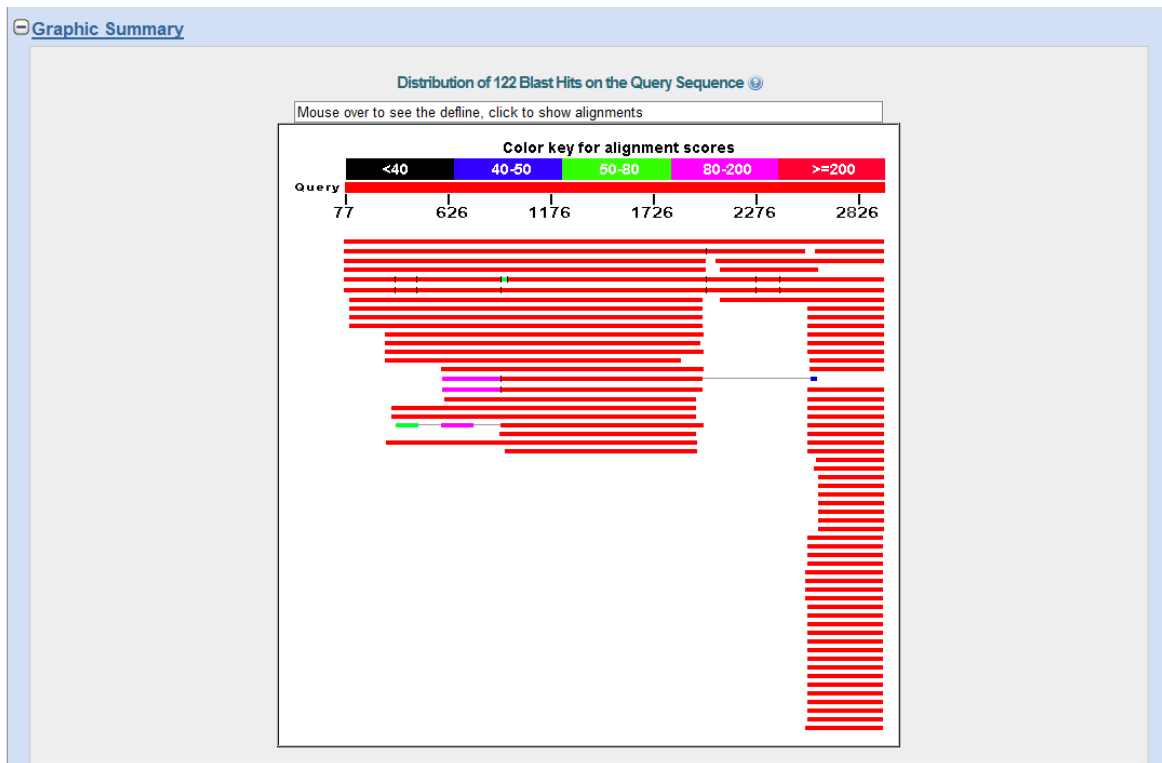


Figure 14. Blastn output graphic summary.

- The **Descriptions** section is next, listing:
 - **Description**
 - **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 [Provide feedback on the new report](#)

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	Max ident	Accession
Arabidopsis thaliana plant U-box 24 protein (PUB12) mRNA, complete cds	5211	5211	89%	0.0	100%	NM_128442.5
Arabidopsis thaliana mRNA for hypothetical protein, complete cds, clone: RAF107-96-H19	3492	4453	76%	0.0	100%	AK226821.1
Arabidopsis thaliana unknown protein (At2g28830) mRNA, complete cds	3487	3487	60%	0.0	99%	AY035038.1
Arabidopsis lyrata subsp. lyrata armadillo/beta-catenin repeat family protein, mRNA	2852	2852	60%	0.0	93%	XM_002880967.1
Arabidopsis thaliana chromosome 2, complete sequence	1986	5439	89%	0.0	100%	CP002685.1
Arabidopsis thaliana chromosome 2 clone F8N16 map mi54, complete sequence	1986	5253	89%	0.0	100%	AC005727.3
Arabidopsis thaliana hypothetical protein (At2g28820/F8N16.11) mRNA, complete cds	1629	1629	28%	0.0	100%	AY219111.2
Arabidopsis thaliana hypothetical protein (At2g28820) mRNA, complete cds	939	939	16%	0.0	99%	DQ056551.1
Arabidopsis lyrata subsp. lyrata predicted protein, mRNA	930	930	27%	0.0	83%	XM_002880966.1
Arabidopsis thaliana U-box domain-containing protein 13 (PUB13) mRNA, complete cds	728	728	58%	0.0	69%	NM_114518.3
Arabidopsis thaliana arm repeat containing protein homolog (At3g46510) mRNA, complete cds	728	728	58%	0.0	69%	AY128813.1
Arabidopsis thaliana arm repeat containing protein homolog (F12A12.30) mRNA, complete cds	728	728	58%	0.0	69%	AY042791.1
Arabidopsis lyrata subsp. lyrata armadillo/beta-catenin repeat family protein, mRNA	713	713	58%	0.0	69%	XM_002877444.1
Ricinus communis Spotted leaf protein, putative, mRNA	673	673	52%	0.0	69%	XM_002534059.1
Populus trichocarpa predicted protein, mRNA	654	654	52%	0.0	69%	XM_002298327.1
PREDICTED: Vitis vinifera U-box domain-containing protein 13-like (LOC100258708), mRNA	625	625	52%	7e-175	68%	XM_002283956.2
Populus trichocarpa predicted protein, mRNA	623	623	49%	3e-174	69%	XM_002313404.1
Vitis vinifera clone SS04FA1YB06	558	558	43%	9e-155	69%	FQ394758.1
Arabidopsis thaliana chromosome 3, complete sequence	547	908	44%	2e-151	85%	CP002686.1
Arabidopsis thaliana DNA chromosome 3, BAC clone F12A12	547	639	43%	2e-151	72%	AL133314.1
PREDICTED: Glycine max U-box domain-containing protein 13-like (LOC100812160), mRNA	529	529	41%	4e-146	69%	XM_003537948.1
PREDICTED: Glycine max U-box domain-containing protein 13-like (LOC100815575), mRNA	502	502	50%	6e-138	67%	XM_003539631.1
Glycine max cDNA, clone: GMFL01-30-O19	502	502	50%	6e-138	67%	AK245454.1
Vitis vinifera contig VV78X076016.5, whole genome shotgun sequence	486	647	43%	5e-133	72%	AM428113.2
Glycine max strain Williams 82 clone GMI_WB0059N17, complete sequence	479	479	32%	7e-131	70%	AC235295.1
Solanum lycopersicum cDNA, clone: LEFL1014DE07, HTC in leaf	471	471	51%	1e-128	67%	AK320949.1
Barbarea verna chloroplast DNA, complete sequence	446	446	12%	4e-121	84%	AP009370.1
Pachycladon ensyii chloroplast, complete genome	441	441	12%	2e-119	84%	JX205495.1
Nasturtium officinale chloroplast DNA, complete sequence	437	437	12%	2e-118	84%	AP009376.1
Crucihimalaya wallichii chloroplast DNA, complete sequence	437	437	12%	2e-118	84%	AP009372.1
Capsella bursa-pastoris chloroplast DNA, complete sequence	437	437	12%	2e-118	84%	AP009371.1

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

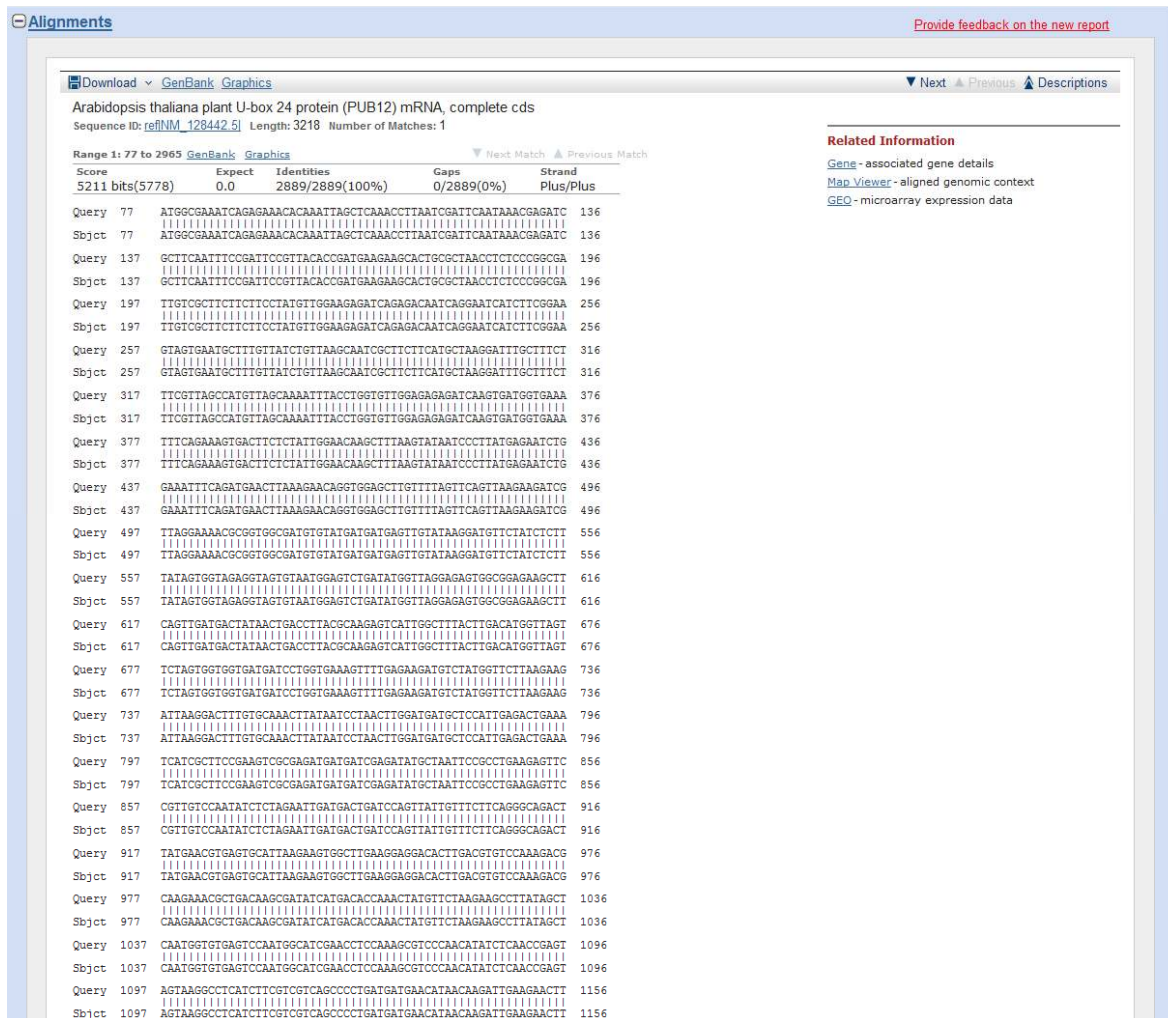


Figure 16. Blastn output alignments.

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

Section I “Introduction and Biological Databases” in *Essential Bioinformatics* by Jin Xiong, Cambridge University Press, 2006. pp 3-27.

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.