**Gen462 LAB - NCBI Sequence Databases and Tools**

**1. Introduction**

The [**National Center for Biotechnology Information (NCBI)**](http://www.ncbi.nlm.nih.gov/) is a great resource and starting point for all things related to bioinformatics.  For our first stop, take a look at the [**site map for NCBI**](http://www.ncbi.nlm.nih.gov/guide/sitemap/). Remember that we cannot cover every aspect of this database. So, please explore this site as much as possible by yourself. Another nice resource you may wish to return to is the NCBI [**Education**](http://www.ncbi.nlm.nih.gov/Education/) page, which has a number of how-tos available as well as mini-courses on a variety of topics related to bioinformatics. The main search portal at NCBI is the [**Entrez Query page**](http://www.ncbi.nlm.nih.gov/sites/gquery).  Entrez allows you to search all of the databases at NCBI at the same time.

**Getting started:**

**Let’s say that you’ve heard about the disease “dyskeratosis congenita” and you’re interested in knowing more about it. Wikipedia or Google can give you basic information about this disease, but the results are not well organized to pursue further research. For research purposes, the best place to search for new things is NCBI.**

**On** the [**Entrez Query page**](http://www.ncbi.nlm.nih.gov/sites/gquery)**, enter “dyskeratosis congenita” (no quotes necessary) in the search box to search for that particular disease in all databases. Click on OMIM to go to the OMIM results page. OMIM (**[**Online Mendelian Inheritance in Man**](http://www.ncbi.nlm.nih.gov/omim/)**)** contains information about inherited traits, especially diseases. **Go back to the Entrez search results and click on “Gene” to go to the Gene results page. Browse through the text and answer the question below.**

1. **Dyskeratosis congenita can be caused by mutations in a wide variety of genes. Identify three such genes. (3 pt)**

**Click on Gene results to learn more about these genes and answer questions 2 – 4. (Note the list of links on the left-hand side.)**

1. **What is the full name of the TINF2 gene? (1 pt)**
2. **On which chromosome, and where on that chromosome, is the TERC gene located? (1 pt)**
3. What are the RefSeq accession numbers for the mRNA and protein sequences for isoform 1 of the TERT gene? (2 pt)

**Clicking on the RefSeq accession number for a protein takes you to its entry in the NCBI’s Protein database. Open the Gene entry for TERT, then click its accession number, NP\_937983.2, to learn more about this protein. The default display setting for this entry is GenPept format. As you can see, a lot of information is available in this format, especially in the features section. Experiment with other formats, such as FASTA (the most widely used sequence format), by selecting them from the upper left dialog. (Note:** [READSEQ](http://www.ebi.ac.uk/cgi-bin/readseq.cgi) is a very useful utility for converting between sequence formats, which is often necessary when using bioinformatics tools which only accept a certain format).

1. **What is the GI number for this protein? (This is a unique identifier for every record in GenBank in NCBI.) (1 pt)**
2. **Paste the protein sequence for NP\_937983.2 in FASTA format below. (1 pt)**

**The right-hand side of the Protein record page has links to a wide variety of sources about the protein, both internal and external to NCBI. In particular, the Related Information header contains cross-references to other NCBI databases, similar to the links under the Links header you clicked from the OMIM or Gene results page.**

1. **Name any 4 links that are available for this record. (1 pt)**
2. **Identify the conserved domains (Concise) in this protein. (1 pt)**
3. **Identify (by its 4-letter PDB ID) at least 1 related structure for this protein. (1 pt)**

**This only scratches the surface of the data available in NCBI databases, but the steps we’ve went through are a good start to finding relevant peer-reviewed literature to learn more from.**

**2. Introduction to Gene Identification, and Gene Annotation**

**Goal:** Introduce you to a few of the genetic resources that are available to evolutionary biologists.

By the end of this lab, you should know how to:

* find molecular resources for your species of interest, including microsatellite markers and protein coding genes, and the corresponding publications.
* given a sequence of interest, identify what gene it is and homologous sequences from other species.
* Learn how to find genes involved in particular biological processes, and vice versa.

Every manuscript considered for publication that contains genetic data must submit their data to one of the following databases, making it freely accessible to the scientific community and the public: European Molecular Biology Lab – EMBL; Data DNA Bank of Japan – DDBI; National Center for Biotechnology Information – NCBI, also referred to as GenBank.

These databases are linked. If we search one, we are searching all of them. Therefore, we will focus on NCBI from the NIH (National Institute of Health) and NLM (National Library of Medicine).

**3. Find Resources for YOUR Species.**

*Back on the main page NCBI:* [***http://www.ncbi.nlm.nih.gov/***](http://www.ncbi.nlm.nih.gov/)

Click on the heading ***All databases.*** This takes you to the *Entrez* search engine that will search all of the databases.

**Type in the Genus and species name** for an organism that you are interested in (or just the Genus or Family). Click ***Go***

Investigate the results. Pubmed lists publications. Click on the **Taxonomy** linkand see the lineage for your species. If you click on the big blue link of the species name a table on the right comes up with the all the Entrez records.

Click on **Nucleotides.** This lists all the core nucleotide sequences from your species that have been submitted to GenBank. If populated, click on Genome, etc to see more complex nucleotide-based data resources. Notice in the search window there is now a taxon ID – this is specific for your species. You can use this and other key words to search for specific types of sequences in your species. For example, typing “txid35005[Organism:exp] microsatellite” will find all the microsatellite sequences for the organism txid35005 (Thamnophis elegans – garter snake). If you wanted to search for the mitochondrial gene ND4 you could use “txid35005[Organism:exp] NADH dehydrogenase subunit 4”, or use the abbreviated form “txid35005[Organism:exp] ND4”, or just type “Thamnophis elegans ND4” searching All databases. As a new user, I would recommend searching within the Taxonomy browser, especially if you are working on a more obscure organism.

**Click on an entry** to see the “flat file” of information about that submitted sequence. This gives you all the information about that sequence including who submitted it and the associated journal publication. Explore the information on this page. Below all the information about the author and origin of the sequence, you can see the details of the sequence itself. This includes the name of the genes and which part of the sequence is the coding region (CDS) that becomes the mRNA, intron and exon boundaries and so on.

Notice the links towards the top. Click on **FASTA.** This gives you the sequence in the most commonly used format for sequences to be used in other programs for (example) sequence alignments and phylogenetic analyses.

**4. Learn how to identify the source and identity of a gene using BLAST.**

From the NCBI website: “The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.”

On the main NCBI page, click on **BLAST**

Note the different BLAST programs available depending on what your query sequence (or the sequence you want to search) is and what database you want to use. We will try the nucleotide blast (blastn) and the translated protein blast (tblastx).

Click **nucleotide blast**. Notice all the (?) these are help files for more information.

**Paste this sequence** into the search window

>contig04613

GTGGATATTGGTCCTCATCGTaaGGAATCGAGTAGGACTCATCTGGATAAATGAGGAGGA

GAAGaGAAAGCAGACCTtGGTTGATTGCAGGCTGAGGGGTGGAAGGGTGTTTATCCTGAC

ATTCAAGGAAGTATCGGCACTCTCGTACgAGcAAATCACGGTCAACCTGAAATCACTATC

TCTTGAGATCCAGCGTGGGGgTTGATCTGGCCAAAAGGAACTCACTTCATTTTCATAAAC

CATCTTCTCACCAACAACCTTCAGTGTGGTCCCGCAATCGTTCAGCGGCACACGGAAGTG

GACTCGGTCTCTCGAAGCCGACCTGGCGATGGGTCCACAGCTTGGGTCTCTGAGCCTGAG

GGTATCCAGGTTCAAGTCCGGTTtCGTACTGCTGCTCGCTATCTCAAAaTCCaTGTGTCC

GTCTTAaGTGCAAATGGTAACTATTGGTGCATTTTGTACACAAGGACATTCTGGGTAGCT

CACCATGGTGGCTATTTCCCcACGATAGTCAAAaTTCAATGTCAAGGGTGCGATATACAA

CTGAAAACCGGa

Under Choose Search Set and then Database, **choose Other Database; nucleotide collection (nr/nt.)** – this will search all nucleotide sequences in the database.

Next, choose your BLAST algorithm under Program selection. Click on the (?) under the choices to read more about the algorithm options (megablast, discontiguous blast, blastn) and when someone may want to use each one. This time let’s choose **blastn**, so click on that one and then click **BLAST** at the bottom of the page.

When the results appear, you’ll see a Graphic Summary of the results, Descriptions of the results, and the specific alignments of your searched sequence to those it matched (or “hit”) in the database at the bottom of the page.

The Graphic Summary shows how well other sequences in the database hit the one you uploaded into the program (your “query”) with the different colored bars (Red is best, Black worst). The position and length of the bars show where the match occurred on your query sequence according to the representation of your query sequence above the graph (red bar labeled “Query” and with tick marks denoting the length of the query sequence). Hover your mouse over the top 5 bars and write down what they say. This information is for the sequence in the database that hit your query.

Below the graph is a list of hits with the following information:

*Accession number*- sequence ID number in the database. This link will take you to the flat-file (try it, you can use the back arrow on your browser to come back to this results page).

*Query Coverage* – how much of your searched sequence matched the listed database sequence

*Score (bits)* – statistical value, the higher the better.

*E-value* – tells you the probability that the hit you got was purely due to chance as opposed to due to an actual similarity between the two sequences. The closer the value to 0, the better (so 1e-100 is better than 1e-5 because 1e-100 is closest to 0).

*Max Ident*- how much of the part of your query sequence that did match a sequence in the database matched exactly.

Below this are the actual alignments for each of the hits.

In this area, you will find that many of the hits align to full reference genomes now (they used to align to genes). To find specific information for the genes of interest, scroll down to the alignments, and next to the range description above the alignment visual, click the Graphics link. You will see your query sequence along the bottom and any genes that align to that sequence above. Hold your cursor on the gene to get a description of the gene name, the title of the gene, and other information.

**10. Write down the best 5 hits for your query sequence and the information for each one (including e-value). From these hits, write what gene your query sequence is. Click on one of the hits to go to the detailed flat file to see more details about the sequence.**

Next, try the same sequence with **tblastx.** Go back to the main BLAST webpage, and click Nucleotide BLAST. In the top left are tabs; click tblastx. Enter the sequence and run BLAST. **This might take a while**. Make sure that upon getting the results, you click “Total Score” in the heading for the results to sort results by Total Score.

**11. Write down the top 5 hits and their information with tblastx. Are they the same results as what you got for blastn? What about the e-values; why are they different? Can you guess what species this sequence may have come from? Why? Choose which algorithm you think gives the best hit when searching for the identity of an unknown sequence through BLAST (blastn or tblastx).**

**5. Learn about Gene Ontology (GO) annotation.**

**Read about Gene Ontology: http://geneontology.org/**

To get a feel for the organization of this database, go to the home menu of this page and search the GO term: **sperm-egg recognition GO:0035036**

A biological processes page should come up describing this GO term. GO terms are arranged in hierarchical tree of biological processes. If you click the “Ancestors and Children” tab, you can see the “Ancestor” processes for which “sperm-egg recognition” is listed as a “child” term, all of which are somehow associated with reproduction.

Now click on the **[link] gene and gene products.** This lists all the genes so far identified to be involved in this Biological Process. Set species to Homo sapiens and apply this filter.

In this list, find the gene abbreviation that we found in the BLAST (problem 2: “ZP2”). Click on the ZP2 entry to view all of the functional annotations for ZP2. This lists all of the direct annotation in the database with corresponding reference links.

Next, go to the protein information database by clicking on the **Q05996** link on the “Database” line of the “Gene product information” entry near the top of the page. Here we see different types of associations (ontologies) for this protein: Biological Processes, Cellular components, and also Molecular Functions. Next to each are Evidence codes that describe how they came up with these associations and direct access to the reference.

**Click on some of these evidence codes**

**6. Mini Case Study to practice using your new Bioinformatic Tools**

**“Biological Processes in the Community Ecology of a Sloth”**



Suppose you are interested in understanding the ecological community of a tree sloth, and how the *biological processes* within the community change when the tree sloth is munching on leaves in the hot sun at the top of the tree verse hanging out sleeping on a shady limb. You perform an experiment taking the following samples in each environment (hot & sunny versus cool & shady): a hair sample that would contain algae, and a sample from the gut which would contain the microbes. From each of these samples you isolate all the mRNA that is being produced (all the genes that are being expressed). You accidently mix the RNA from each sample together – but still kept the two treatments separate. Since you can’t go back to Costa Rica to get more samples, you decide to proceed and depend on your new bioinformatic skills to sort out the results. So your samples are now…

*Cool & Shady Sample*: pool of mRNA from hair (with algae), gut (with microbes), and liver expressed at cool temperature

*Hot & Sunny Sample:* pool of mRNA from hair (with algae), gut (with microbes), and liver expressed at warm temperature

You send these two samples to the sequencing facility to sequence all the RNA using a method called RNA-seq. After all the analyses are done, you are given a list of sequences that are differentially expressed between your two treatments. These sequences represent genes that were up-regulated in hot treatment relative to cool treatment – this means that these genes were “turned-on” or “turned-up” to produce more RNA copies in the hot conditions; or were “turned-off” or “turned-down” to produce less RNA in the shady condition.

**At the bottom of this assignment are a few those differentially expressed sequences. Use these sequences to address the following questions.**

**12) Which sequences come from the sloth, the algae growing on the sloth, and the microbes in the gut of the sloth? To figure this out, take each sequence and BLAST it. Because we are working on a community of organisms that are likely NOT represented in this genetic database we are looking for evolutionarily similar hits. As sequences evolve slower at the protein level then nucleotide level (i.e. many nucleotide mutations are “sense” mutations that don’t change the protein) it would be best to start with *blastn,* and then perhaps try *tblastx* that will translate our sequence into the 6 possible protein sequences and compare it to the translated version of the nucleotide database .**

**What are the top hits?** Hint – if you hit whole genomes in your blast search, you can scroll down to the alignments to see the specific region or gene in the genome that matches your query.

**Fill out the table. Tip: To find the gene when you are given the full genome result, click the name of the result, then click Graphics under the results. Click the colored bar to get information on the gene your sequence aligns to, and find the gene name in the title label.**

|  |
| --- |
| **Sequence Species/organism Gene e-value** |
| **1** |
| **2** |
| **3** |
| **4** |

**13) Which sequences do you think belonged to the… (1 point)**

Sloth:

Gut Microbes:

Algae:

**14) What are the biological processes that each of these genes are involved in?**

**To investigate use GO Ontology. Type in the gene name (select the gene) and hit Search, then go to Genes and Gene Products. Click the gene name and look at the GO class (direct) column to see gene functions.**

You can also search using Uniprot if you don’t find an easy explanation for the biological process.

|  |
| --- |
| **Sequence, Gene Biological process** |
| **1** |
| **2** |
| **3** |
| **4** |

**15. Write a few sentences interpreting what is happening in this little community at the gene expression level inferred by the biological processes.**

**Sequence 1**

TTATTCAATATCAATATGGTTTTCGTCGTCATCCGTCTCATCCATCTTAGGAAGAGTAATTGTCAAAACA

CCATTCTTATATTGAGCCTTGATATCCTTCTTATTTACATCCGGCAAACGATATTGCCGACTCATTTGGC

CATACCGACGCTCAGAATGCAAAACATTGCCATCCTTGTCACCATCATCATCAAAAGTGTCCCGATGACC

AGTAACTGTCAAAATATCATTAGCATAATTAATGTGAATATCCTTCTTATCAAATCCAGGCATGTCAATC

TTTACGGTATAATTCTTGTCATCTTCTGTAACATCTGTCTTCATGTGATCAGCCACCGGAGTTAGGTTAG

AGAAGAAATCATCGTTCATCCAGTTACGCATGTTCATCAAACTATCAAATAAATTATTACGGTTTTGTAA

TTCATTAGCCAT

**Sequence 2**

ATGGAGCTGCTAAAGCTGAACCGGAGCGCGCAGGGGTCCGGAGCCGGCCCGGGGGCTTCCCTGTGCCGCG

CGGGGGGCGCCCTCCTCAACAGCAGCGGTGCGGGCAATCTCAGCTGCCAGCCGCCTCGCCTCCGCGGAGC

CGGGACACGAGNNANATTGGAGCTGGCCATTAGGGTCACCCTTTATGGAGTGATCTTTCTGATGAGTGTT

GGAGGAAATGTGCTCATCATCGTGGTCCTGGGACTGAGTCGCCGGCTCAGGACTGTCACCAACGCCTTCC

TGCTCTCACTGGCTGTCAGCGACCTCCTGCTGGCTGTGGCTTGCATGGCCTTCACCCTCCTGCCCAATCT

CATGGGCACGTTCATGTTTGGCACAGTCGTCTGTAAGGCAGTTTCCTACCTCATGGGGGTGTCTGTGAGT

GTGTCCACACTAAGGCTTGTGGCCATCGCCCTGGAGCGATACAGCGCTATCTGCCGGCCGCTACAAGCAC

GCGTGTGGCAGACGTGTTCCCATGCGGCTCGTGTGATCATCGCCACTTGGATGCTCTCTGGACTGCTCAT

GGTGCCCTACCCGGAGTACACCGCCGTACAGCCCGCAGGAGGGGCCCGGGCGCTGCAGTGCGTGCATCGT

TGGCCCAGTGCGCGGGTCCGCCAAACCTGGTCGGTACTGCTGCTCCTGCTTTTGTTCTTCGTCCCAGGCG

TGGTTATGGCTGTGCCCTACGGGCTCATCTCCCGCGAGCTCTACTTAGGGCTTCGCTTCGACGAGGACAG

CGACAGCGAAAGCCTAGTCCGAAGCCAAGGAGGGCTGCGGGGTGGGGCGGGACCAGGTCCTGCCCCCCCC

AATGGGAGTTGCCCGCCGGAGGGCGGGCTGGCTGGCGAGGACGGCGACGGCTGCTACGTGCAGCTTCCGC

GCTCGCGTCAGACTCTGGAGCTGTCCGCGCTGACCGCGCCCACTCCTGGGCCCGGAGGTGGCCCCCGGCC

CTACCAGGCCAAGGTGTTGGCCAAGAAGCGCGTGGTGCGGATGCTGCTGGTGATCGTCGTGCTTTTTTTC

CTGTGTTGGTTGCTACTGTATAGTGCCAACACGTGGCGTGCCTTCGACAGCTCTGGTGCACACCGCGCAC

TTTCAGGAGCGCCTATCTCTTTCATCCACTTGCTGAGCTACGCCTCAGCCTGCGTCAACCCCCTGGTCTA

**Sequence 3** GAGGCTGCACAAACAACTAAACTCGCCTTTAGGACAAGTCTTACTATCTCCTTACCGATTCGCTCGGGGA

CCGCCGATACCCTCGCTTGTTGCGTCAAGATGCTTCGACTACTAGGTCTGGCGCTGCGCGCTCATGGCCT

GGCCGTGAGCGGCGCCCGGATGTTGCAAAGCCGACTCATGCTTCAGTCCTCCGCGGCGTCACTGTCCACC

AGCGCATCAAACCAGCAAGAGTCGAGTCCTCCCGCAGAGGGGACAGAAACGGTATCCAACACCTACATCG

ACAAGGACGGAAAGGAGCACACAGTAGCAGCGCCTATCGCGAAGAACCTGCTGGAGATAGCGCATGAGAA

CGAAATCGACTTGGAAGTCGCTTGCGAGGGCTCCCTCGCTTGCTCCACGTGTCACCTCATCTTCGAGGAC

GAGGCCACGTACAAGAAGCTGCCTGAGCCGCACGAGGATTAGCTGGACATGCTGGACCTGGCCTTTGGGC

TGACCGACACCTCGCGGCTGGGCTGCCAGGTGTTGGCCATCAAGGATCTGGAGGGCGTGCGGGTGCGCAT

CCCCTCCGCCTCCCGCAACTTCTACGTGGACGGCCACAAGCCCAAGCCCCACT

**Sequence 4**

CGAAATTATCGAGCAGCGACCAAGCGAAATAGCCCCGGACGTCCACCCATTGGGCGATGGCCTCGTGCAG

CGCGCGCAGATGGCTGCGGAATAGGCGATGCGGCGCGTGTCGTCGATCCTCCATTCTCTCCCTCAGGA

TCGGCGAAGGCCGCGCCGATTTCGGTGATGTAGAGAGGGATGTCCCCGTACCGGGCCTTGACCCACGCCA

AGACGTCCTTCAGCCCTTCCGGATAGACCTCCCAGCCCATCTCGGTATGCGCGACGCCGCGGTGGGGCAC

GGCGGTTACCTCGAGCGGGCCGCCCGACGGATCGTGTCGGACTACGGCGCTGGTGTAGTAGATGATGCCC

AGGTAATCGATGGGCTCCCGGATCACGCGCAGGTCTTCGCTCTCGAAGCTCGGCCAGTGGATAGCGGAAA

TTTCCGCCAGCTCGTCGGCATAGGCACCGTGCAAGACCGGATCGAGATACTGGCGATTCATCTAGGCGTG

GGCGCGCTCGGCGGCGGCACGGTCGTCCCGGCTGTCGGTCAGGGCATATTTGGGCTCGAGATGACCACC

AGTCCGATCTGCCCCCGGTCGTCGGCGCGAAACGCCTGTACGGCCAAGGCGTGGGCCCGCAGAAGGTTGT

GCGTGACCCAGGGCGCATGCTTCAGGGACCTGTGACCCGGAGGATGTACGCCCGAGACATAACCCGCGTC

CATGATGACCCAGGGCTCCTTCAGCGTCGCCCACAAATCGATCTCGTTGCCCAGCGCTCTGATGACGGTG

TGCGCATAAT