EDVOTEK®

THE BIOTECHNOLOGY EDUCATION COMPANY®

Revised and Updated

Edvo-Kit # 339

Edvo-Kit #339

# Sequencing the Human Genome

## Experiment Objective:

In this experiment, students will read DNA sequences obtained from automated DNA sequencing techniques. The data will be analyzed using publicly available databases to identify genes and gene products. The impact of Genomics will be discussed in the context of today's society.

See page 3 for storage instructions.

1.800.EDVOTEK • Fax 202.370.1501 • www.edvotek.com • info@edvotek.com

# Table of Contents



## Experiment Components

This experiment contains a total of twelve sections of automated DNA sequence printouts. Students can use any Human Genome sequence database to perform the activities in this lab. For purposes of simplification we have chosen to illustrate the database offered by the NCBI.

## Requirements

- Computer with Internet access

# Background Information

The haploid human genome comprises approximately three billion base pairs of DNA that are organized into 23 chromosomes. The order of these nucleotides creates genes, which are discrete units of genetic information that contain the instructions to build and maintain an organism. DNA sequencing is the process of determining the precise order of these nucleotides. In 2001, the first draft sequences of the human genome were published, one by an international coalition known as the Human Genome Project and one by a company called Celera. Although these initial studies took multiple years and required a lot of people, advances in sequencing technology have made acquiring full genome data simple and fast. In 2015 a full sequence could be acquired in a day for around one thousand dollars. Given that much of this sequence data is freely available online, the new challenge in genetics involves finding creative and efficient ways to analyze and manage the vast amounts of data being generated. This has resulted in the growing field of bioinformatics – a discipline that blends computer science, biology, and information technology.

Information from the human genome can help us better understand our physiology and the biological basis of inherited diseases. For example, DNA sequencing of individual patients, known as personalized medicine, is changing the role of genetics in medicine. Personalized medicine uses an individual's genetic profile to guide decisions regarding the prevention, diagnosis, and treatment of disease. Table 1 highlights five promising areas of personalized medicine. Although these tests and services provide amazing therapeutic possibilities, they also generate patient information that is unprecedented in its detail and permanence. This raises new legal and logistical questions about protecting doctor-patient confidentially. For example, do parents have the right to order genetic tests for their minor children, or can insurance companies increase rates or deny service due to a potential genetic issue? The Genetic Information Nondiscrimination Act of 2008 addresses some of these concerns by prohibiting the use of genetic information in employment and health insurance decisions. In depth discussion is needed to balance improvements to human health with the ethical consequences.

Besides a role in health care, human genome sequencing has, and will continue to have, a large impact on our understanding of human history, evolution, and general biology. In the field of phylogeography, scientists examine current and ancient geographic patterns of molecular genetic variation to learn more about human's global expansion and adaptation. Studies at the genomic scale have shown extensive interbreeding between separated populations, far more than was previously estimated based on individual loci studies. Another promising area is genomic comparison between humans and other organisms that help connect the biology of model organisms to human physiology. Comparative genomic studies are also helping to decipher the roles of protein coding genes, noncoding RNAs, and regulatory sequences in evolution. Similarly by studying the structure and activity of the human genome itself we can ask questions about the function of DNA at the levels of genes, RNA transcripts, and

| Promising Area of Application | Goal |
|---|---|
| Molecular characterization of rare disease | Provide a definitive diagnosis and suggest new treatment options |
| Individualized cancer treatments | Predict a patients response to a specific targeted therapy |
| Pharmacogenomic | Optimize drug therapy to ensure maximum efficacy with minimal adverse effects |
| Population screening for disease risk | Increase individuals health knowledge and encourage proactive health changes |
| Preconception and prenatal screening | Inform parents about the risk of disease in offspring |

**Table 1:** Promising Areas in Personalized Medicine

protein products. However, for genomic data to provide insight into these questions, scientists must address the computational challenges around data analysis, integration, and visualization.

## DNA SEQUENCING AND DATABASE SEARCHES

The first step in a sequencing project is obtaining the raw data – the precise order of the four nucleotides: A, G, C, and T. There are several approaches to generating sequence information and new methods are emerging each year. Two popular methods are chain termination sequencing and sequencing by synthesis. Chain termination sequencing, often called Sanger sequencing, allows researchers to generate long DNA reads of a target sequence, also known as the template. The DNA template is combined with a DNA primer, the DNA polymerase I (DNA Pol I) enzyme, and a mixture of two types of free nucleotides, deoxynucleotides (dNTPs) and dideoxynucleotides (ddNTPs) (Figure 1A). During the sequencing reaction, DNA Pol I uses the DNA template and adds dNTPs to the primer to form a complementary strand of DNA. Occasionally, the DNA Pol I will instead add a ddNTP to the DNA strand, terminating the reaction (Figure 1B). This end-termination is due to the lack of a 3′ hydroxyl group on the ddNTPs (Figure 1A) making it impossible for the polymerase to add another nucleotide to the end of the strand. Therefore, the reaction result is a series of DNA fragments of differing sizes that can be separated by capillary electrophoresis (Figures 1B, 1C). Importantly, each ddNTP is labeled with a different fluorescent marker, allowing the fluorescence of a particular DNA strand to be "read" by a laser (Figure 1C). The four different ddNTP fluorescence colors are then automatically detected and the fluorescence intensity translated into a data "peak" that represents the order of nucleotides in the template DNA (Figure 1D). Sanger sequencing was introduced in 1977 and was the major method used to create the first genome sequence. It is still used today because of its ability to generate long sequence reads (500-800 bp).

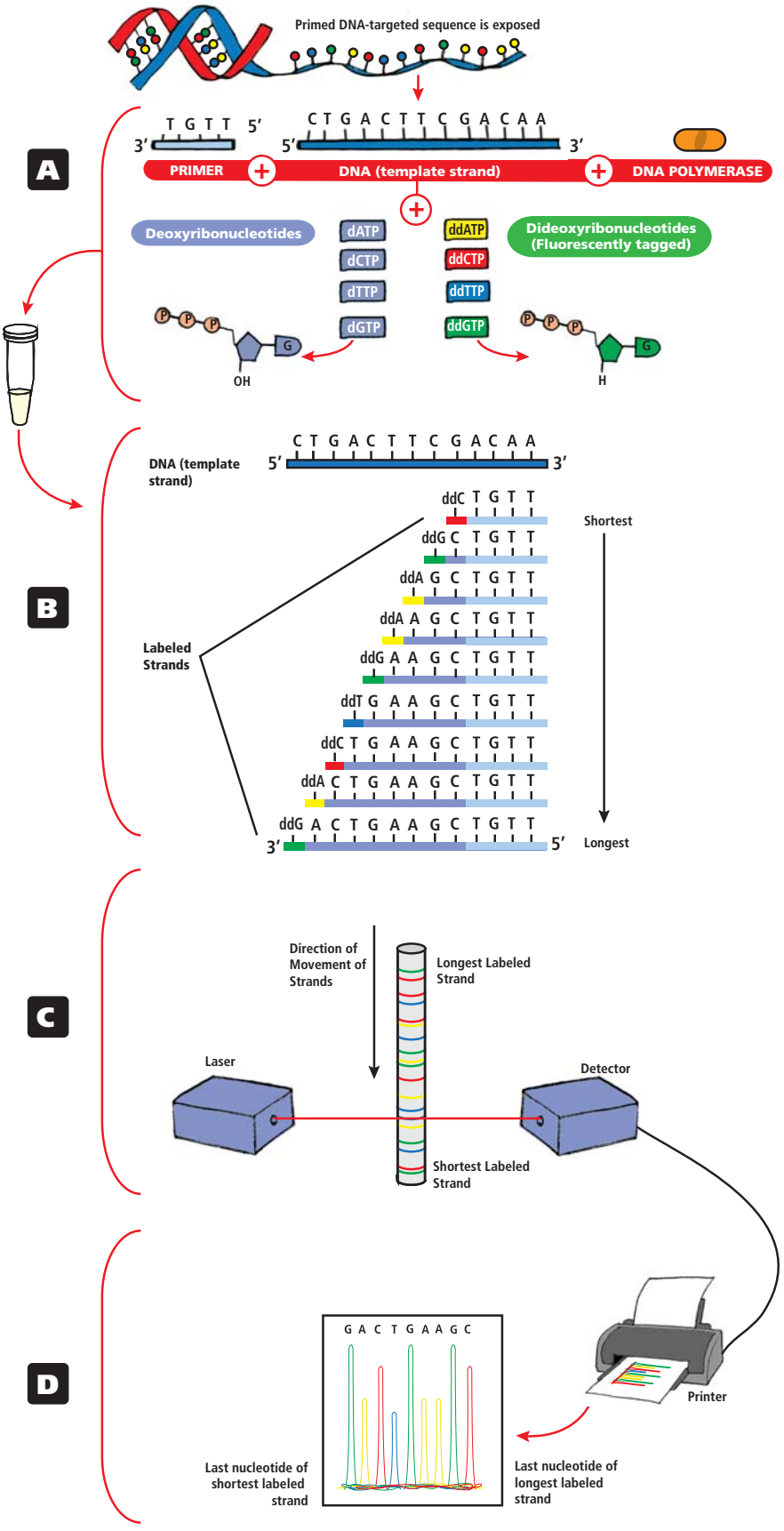
In sequencing by synthesis, a single strand of DNA that complements the strand of interest is built nucleotide by nucleotide with each added nucleotide releasing a specific signal. Before sequencing, the DNA sample must be prepared and immobilized. This involves randomly fragmenting the DNA of interest, anchoring the DNA fragment to a solid surface, and eliminating one of the strands. These steps produce detectable and distinct areas of short, identical DNA molecules. These DNA molecules are then allowed to rebuild themselves into double stranded DNA. As each new nucleotide is added to the complimentary strand a signal unique to that nucleotide is released. Theses signals are recorded on a computer and translated into sequence information. The advantage of this method is that many different DNA strands can be examined in a single run. This means that sequencing by synthesis has a much higher throughput and a lower cost than Sanger sequencing. One disadvantage of this technique is that it produces shorter reads (50-150 bp), meaning that many reads must performed to sequence a single gene.

## INTERPRETING DNA SEQUENCE INFORMATION

After obtaining DNA sequencing data, molecular biologists will often search public databases for similar sequences. Performing such a search can reveal research already performed on the sequenced gene, including the three-dimensional structure of the gene product, diseases associated with the sequence, and in which tissues the gene is active. In cases where the region has not been specifically studied, finding similar sequences can provide clues to the sequence's function and its evolutionary relationship to other human genes.

One of the largest and most influential databases is known as GenBank. This free, open source database contains over a trillion nucleotide bases of publicly available sequence data. Each entry in GenBank contains a sequence and an accession number as well as supporting bibliographic and biological annotations such as author references and taxonomic data. The NCBI (National Center for Biotechnology Information) oversees and maintains the database as a whole but each entry is submitted directly by individual laboratories. Direct submission has allowed the database to keep pace with the rapid growth in sequence data production. However, it also means that heterogeneity in entry quality exists, especially in the certainty of each nucleotide's identity and in the extent of attached annotation. These can vary depending on the goals of the study, the physical properties of the DNA region(s), and the chosen sequencing method. To address this, GenBank classifies the sequence information based on the sequencing strategy used to obtain the data.

**Figure 1: Sanger DNA Sequencing.**
(A) Setting up the sequencing reaction.
(B) Incorporation of the ddNTPs create different size DNA fragments.
(C) The labeled mixture is sequenced using capillary gel electrophoresis. A laser detects the dluorescent label on each of the ddNTPs.
(D) The information is analyzed using a computer.



Associated with this database are several useful bioinformatics tools including the Basic Local Alignment Search Tool, or BLAST. The BLAST tool finds regions of local similarity between a user's DNA sequence and sequences in the GenBank database. Such similarity suggests homology, the existence of shared ancestry between the genes. In BLAST terminology the user's input sequence is known as the **query sequence**, sequences in the database are known as **target sequences**, and sequences with similarities to the input sequence are **hits**. The user can draw inference about the putative molecular function of the query sequence by looking at the hits. BLAST is also used to identify unknown species, locate known protein domains, and find potential chromosome locations.

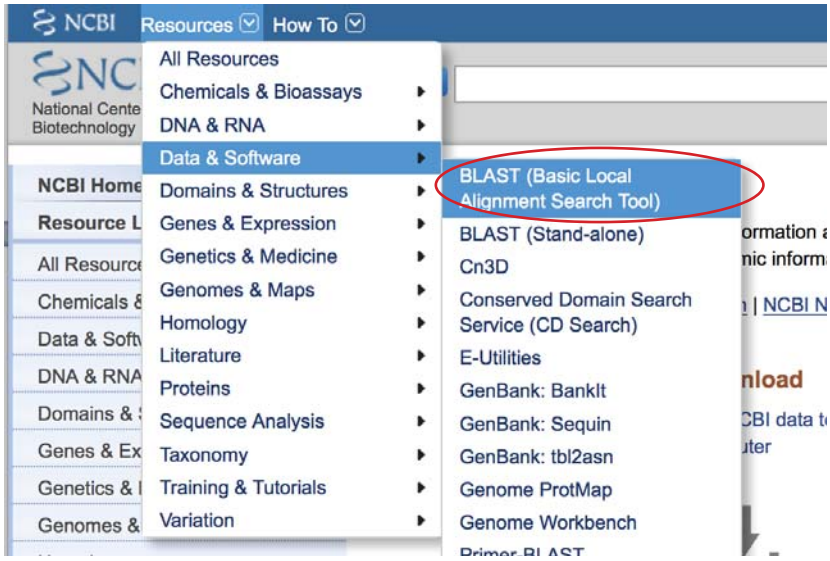
BLAST takes a heuristic approach to the problem of searching through such a mammoth database of target sequences. This means that it takes shortcuts in order to find sequence matches in a reasonable time frame. These shortcuts are based on the assumption that biologically similar sequences will contain short stretches of very high scoring matches. BLAST attempts to find these high scoring segment pairs by removing low complexity regions, dividing the sequence into much shorter seeds, and then scanning the database for matches. Once it has generated a list of high scoring matches BLAST extends the seeds to see if they are contained in longer high scoring alignments. By searching the GenBank database this way BLAST can return results very quickly although it sacrifices some accuracy and precision.

BLAST is popular not only because of its speed but also because it computes the statistical significance of the solutions. In addition to the accession number, description, and genome link BLAST provides a score, bit score, and e value. The score, S, is a raw measure of the quality of alignment between the query and the hit. User chosen variables that incorporate molecular and biochemical concepts heavily influence this value. The bit score is the raw score adjusted for the size of the database and the sequence length. The e value translates to the probability due to chance that there is another alignment with a similarity score greater than the given S score. Scores, bits scores, and e values are a good first indicator of similarity between sequences, however the alignment itself should also be examined to ensure accuracy.

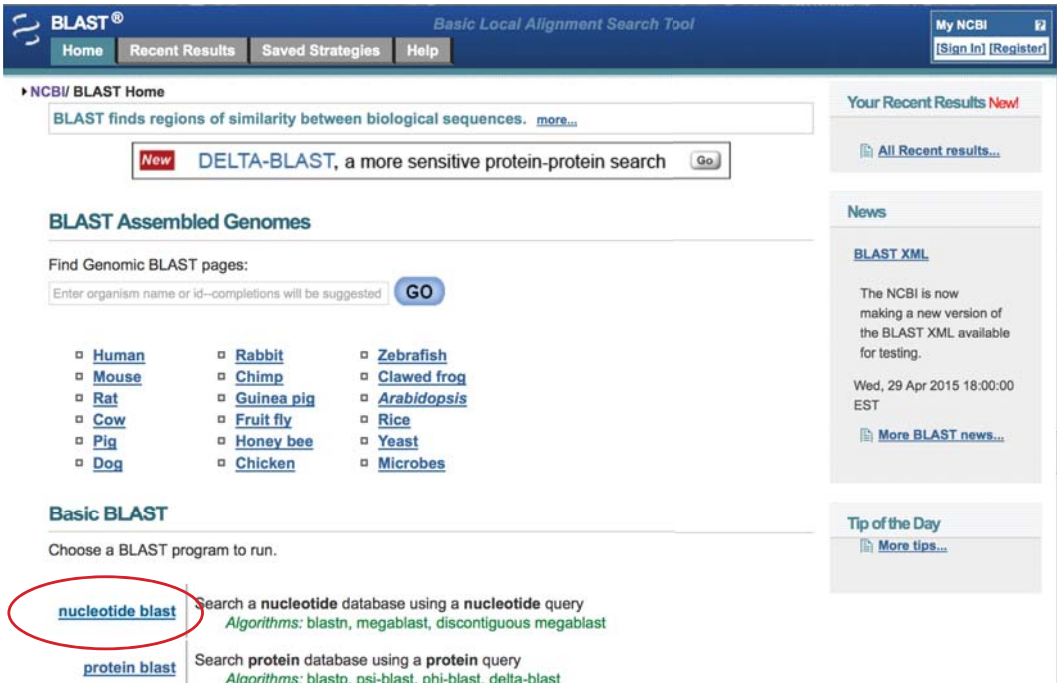
This exercise introduces students to genomics and bioinformatics. In order to gain experience in database searching, students will use the free service offered by the National Center for Biotechnology (NCBI). At present, GenBank comprises several databases including the GenBank and EMBL nucleotide sequences, the non-redundant GenBank CDS (protein sequences) translations, and the EST (expressed sequence tags) database. For this experiment, we recommend using a database offered by the NCBI. These exercises will involve using BLASTN, whereby a nucleotide sequence will be compared to other sequences in the nucleotide database. For each of the three sequences, students should identify a potential human disease and discuss related bioethical issues.

# Experiment Procedures

1. Type: **www.ncbi.nlm.nih.gov** to log on to the NCBI web page.
2. On the top left of the screen click on the drop down menu "Resources".
3. Click on "Data and Software", then click on "BLAST" (Basic Local Alignment Search Tool.)



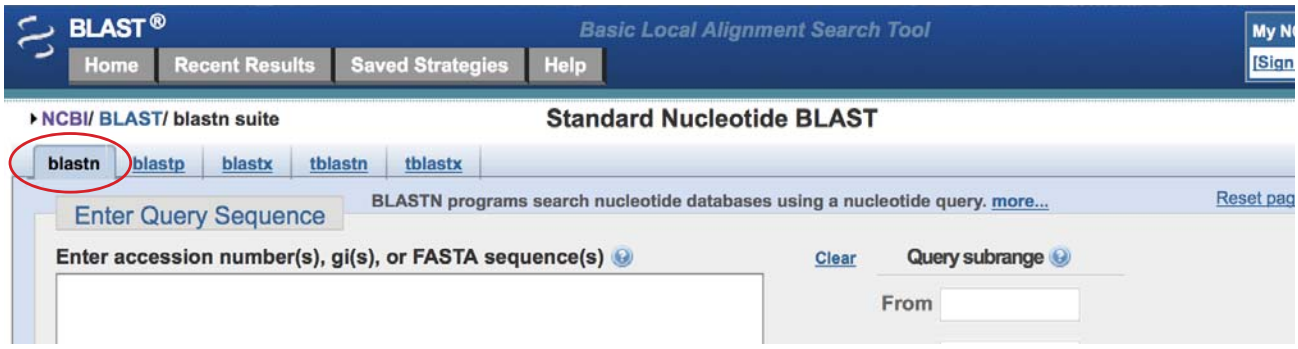

4. On the new BLAST Home screen select "nucleotide blast" which is the first opinion under the Basic BLAST list.



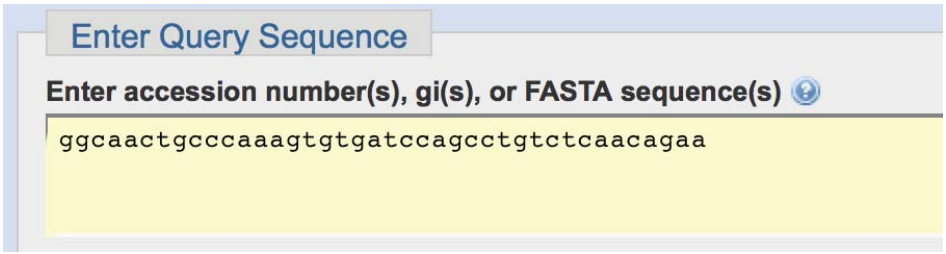

*continued...*

## Experiment Procedures, continued

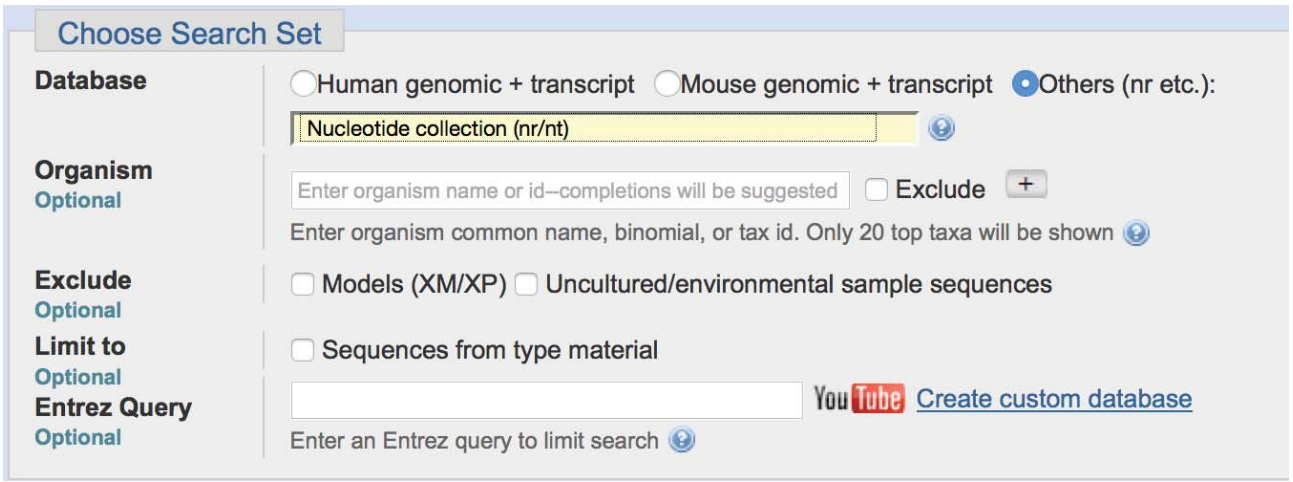
5. On the new screen make sure the tab selected is "blastn".




6. Enter the nucleotide sequence into the large box in the "Enter Query Sequence" section; be careful to type the following sequence exactly: ggcaactgcccaaagtgtgatccagcctgtctcaacagaa

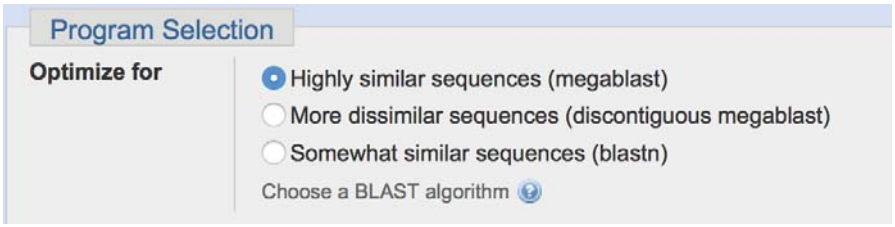



7. Under "Choose Search Set" make sure that "Others (nr etc)" is selected and that "Nucleotide collection (nr/nt)" is highlighted in the dropdown menu. The remaining entries should be left blank.



## Experiment Procedures, continued

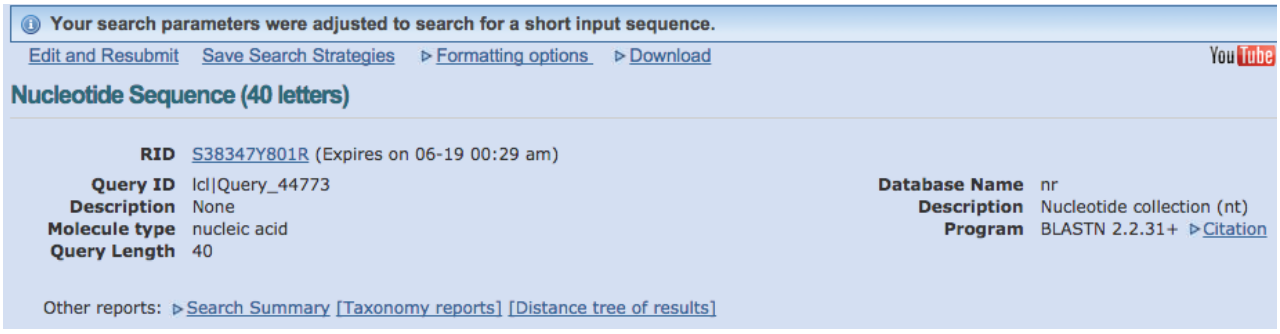
8. Under "Program Selection" select "High similar sequence (megablast)"




9. Click on the blue "BLAST" query box.


10. Once the "BLAST" query box has been clicked you will be assigned an ID#. Record this number so you can check your results at a later time.

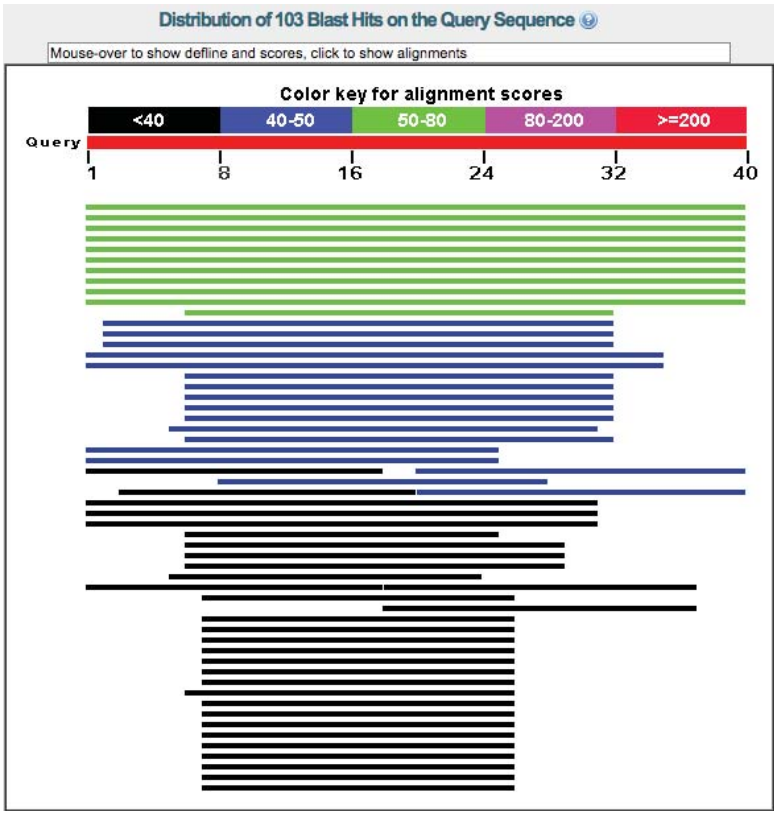
11. Examine the BLASTN search report. The report includes:
    a. **Search summary report** shows an overview of the BLASTN search parameters.



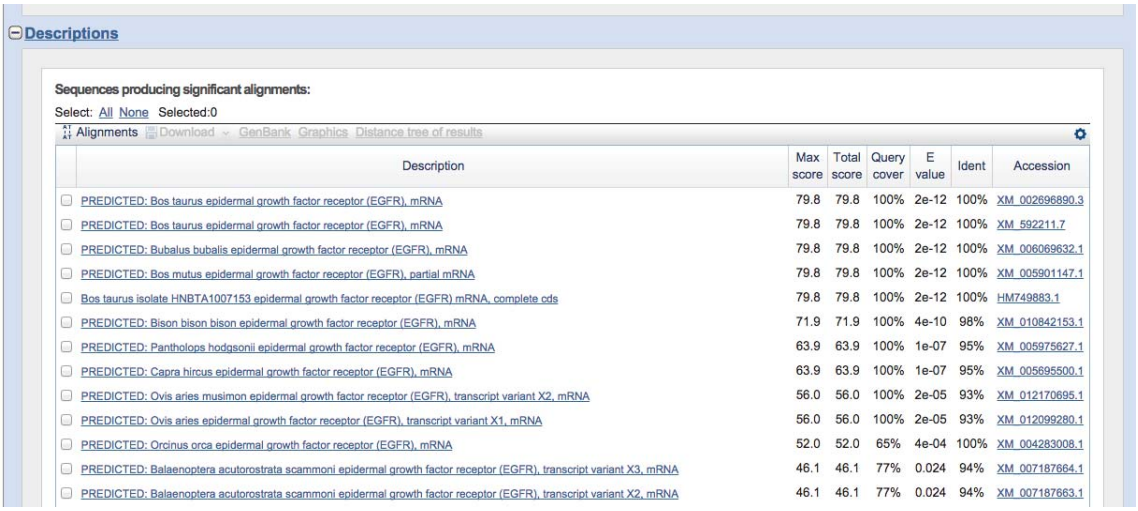

*continued...*

## Experiment Procedures, continued

b. **Graphic Summary section** that shows the alignment of database matches to the query sequence. The color of the boxes corresponds to the score of the alignment with red representing the highest alignment scores.



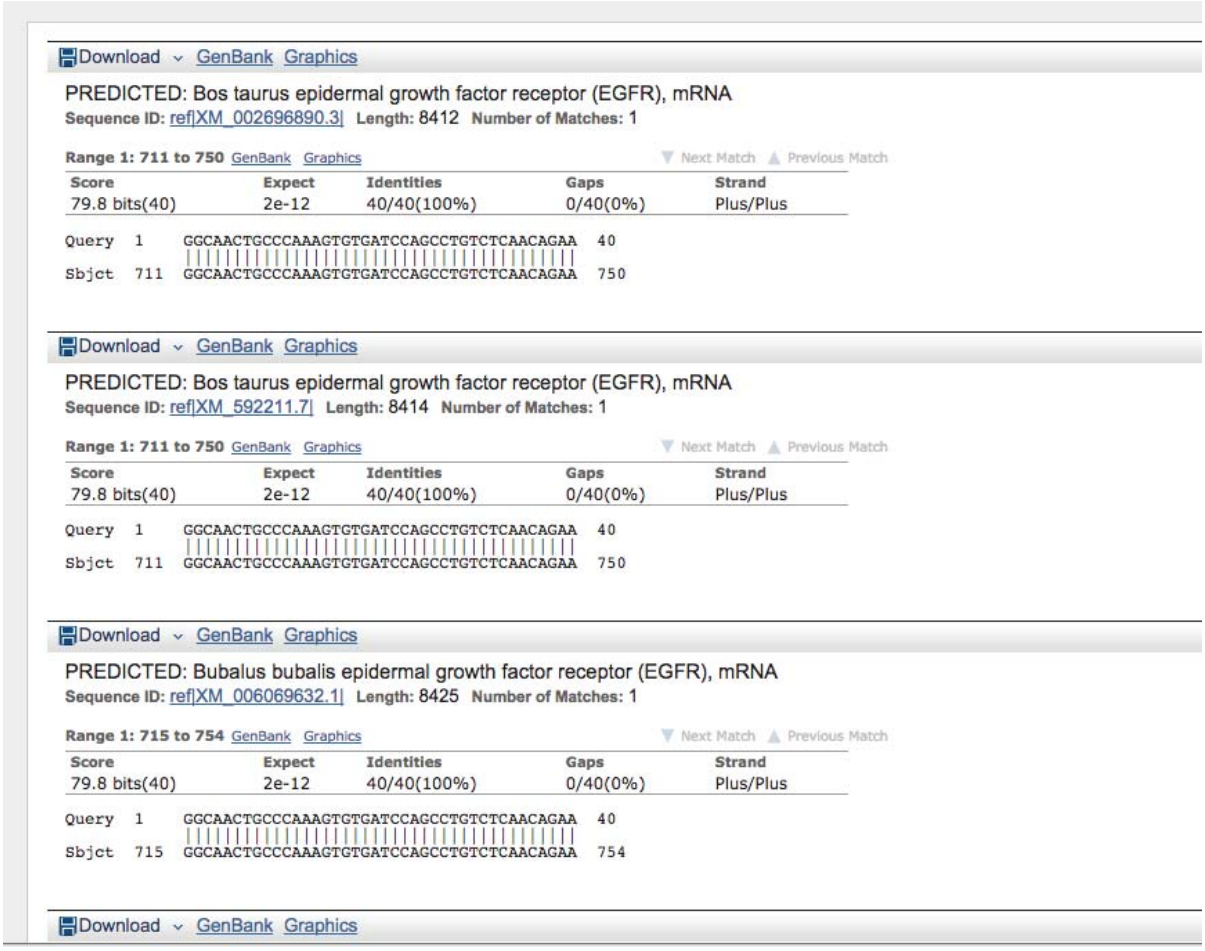

c. **Description section** that shows all the sequences in the database with significant sequence homology to our sequence. By default the results are sorted according to the E-value but you can click on the column header to sort the results by different categories. Notice that there can be several different entries with identical high scores.




| Description | Max score | Total score | Query cover | E value | Ident | Accession |
|---|---|---|---|---|---|---|
| PREDICTED: Bos taurus epidermal growth factor receptor (EGFR), mRNA | 79.8 | 79.8 | 100% | 2e-12 | 100% | XM_002696890.3 |
| PREDICTED: Bos taurus epidermal growth factor receptor (EGFR), mRNA | 79.8 | 79.8 | 100% | 2e-12 | 100% | XM_592211.7 |
| PREDICTED: Bubalus bubalis epidermal growth factor receptor (EGFR), mRNA | 79.8 | 79.8 | 100% | 2e-12 | 100% | XM_006069632.1 |
| PREDICTED: Bos mutus epidermal growth factor receptor (EGFR), partial mRNA | 79.8 | 79.8 | 100% | 2e-12 | 100% | XM_005901147.1 |
| Bos taurus isolate HNBTA1007153 epidermal growth factor receptor (EGFR) mRNA, complete cds | 79.8 | 79.8 | 100% | 2e-12 | 100% | HM749883.1 |
| PREDICTED: Bison bison bison epidermal growth factor receptor (EGFR), mRNA | 71.9 | 71.9 | 100% | 4e-10 | 98% | XM_010842153.1 |
| PREDICTED: Pantholops hodgsonii epidermal growth factor receptor (EGFR), mRNA | 63.9 | 63.9 | 100% | 1e-07 | 95% | XM_005975627.1 |
| PREDICTED: Capra hircus epidermal growth factor receptor (EGFR), mRNA | 63.9 | 63.9 | 100% | 1e-07 | 95% | XM_005695500.1 |
| PREDICTED: Ovis aries musimon epidermal growth factor receptor (EGFR), transcript variant X2, mRNA | 56.0 | 56.0 | 100% | 2e-05 | 93% | XM_012170695.1 |
| PREDICTED: Ovis aries epidermal growth factor receptor (EGFR), transcript variant X1, mRNA | 56.0 | 56.0 | 100% | 2e-05 | 93% | XM_012099280.1 |
| PREDICTED: Orcinus orca epidermal growth factor receptor (EGFR), mRNA | 52.0 | 52.0 | 65% | 4e-04 | 100% | XM_004283008.1 |
| PREDICTED: Balaenoptera acutorostrata scammoni epidermal growth factor receptor (EGFR), transcript variant X3, mRNA | 46.1 | 46.1 | 77% | 0.024 | 94% | XM_007187664.1 |
| PREDICTED: Balaenoptera acutorostrata scammoni epidermal growth factor receptor (EGFR), transcript variant X2, mRNA | 46.1 | 46.1 | 77% | 0.024 | 94% | XM_007187663.1 |

## Experiment Procedures, continued

d. **Alignment section** that shows alignment blocks for each BLAST hit. Each alignment block begins with a summary that includes the Max score and expected value, sequence identity, the number of gaps in the alignment, and the orientation of the query sequence relative to the subject sequence.




12. Select a sequence to focus on in-depth. You can do this by clicking on a colored bar in the graphic section, clicking on the sequence name in the description section, or scrolling down to the alignment section. Then click on the sequence ID. This brings up additional information about the subject sequence, including the gene name, the genus and species of origin, and articles written about the gene. After performing this search, the top hit should be *Bos taurus* epidermal growth factor receptor (EGFR), mRNA. Sequence ID: ref|XM_002696890.3|. If top hit does not match, try re-entering the sequence. Be sure to double check the search parameters before BLASTN searching.

## Experiment Procedures, continued

### EXERCISE 1

Now that you have familiarity with the entry and submission process of BLAST, read the DNA sequence analysis from the automated gel run sequence printout (any lane) and find the gene that this sequence fingerprint identifies.

To do this:

(1) Identify the nucleotide sequence (100-200 nucleotides) from the DNA Sequence readout.
(2) Type at least 70 bases in the query box of the BLAST program at the NCBI website. The bases can be from any region of the sequence, but they should be contiguous.
(3) Examine the BLASTN search report, identify a likely gene, and examine the gene ID for detailed information.

Once the gene has been identified, answer the following questions:

(a) What is the name of this gene?
(b) Compared to the GenBank entry, what strand have you read?
(c) Can you find a paper that has been written about this gene? Write down the name of one of the contributing authors.
(d) Identify a disease caused by mutations in this gene. What would be the implications if a doctor tested for this disease? What if an employer or an insurance company tested for this disease?

**Remember the following:**

- The automated sequence differentiates the bases as follows: A is green, C is blue, G is black, and T is red.
- DNA is double stranded and contains a top (5′→3′) and bottom (3′→5′) strand (sometimes this corresponds to the coding and noncoding strands.) A DNA sequence is always entered in the 5′→3′ direction.
- Sometimes it is difficult to read a nucleotide peak. This is particularly true at the beginning and end of a sequence read where peaks may overlap. Such ambiguous places are often labeled with a N rather than one of the four nucleotides. It is often best to skip sections with lots of Ns
- Because researchers call the same gene different names, several possibilities may exist for each sequence.
- By clicking on the GenBank accession number you can access additional information such as the protein/ amino acid sequence, descriptions of the sequence/ gene, and the contributing scientists names.
- Some high-scoring sequence matches may be predicted genes that do not have any research papers associated with them. Students may have to check several matches before they will find a sequence with an associated reference.

### EXERCISE 2

Exchange your automated data sequence printout with another group and submit the sequence to BLAST analysis. Write down the gene. Select a sequence associated with a published paper, record the title and first author of the paper.

# Study Questions

1. What is bioinformatics? How have advances in sequencing technology affected this field?
2. Name two sequencing methods and describe the trade off between the production rate and the length of the sequences produced.
3. What assumption does BLAST make? What are the advantages and disadvantages of making this assumption?

## DISCUSSION QUESTIONS

These questions are complex with no correct or incorrect answers. They are provided to stimulate discussions about the application of genomics in the 21st Century.

1. Should we do prenatal tests for diseases that are currently incurable?
2. Should we test our children for adult onset disease?
3. What role should the government play in establishing guidelines for human genetic tests?

# Instructor's Guide

## DNA SEQUENCES

### Sequence 1

LANE 1:
GNNNNNTGGNNNNNNNATANTTGCGGCCGCGGTTTTNTTTTTTTTNTTNNNCNNGGAGCACAANCNAATGNANTGTGTTGTTGTGGCGARGGC-
GARGGCGCCGTTHHTAAAACTGTCTCCTGATATCCTACACAACAAACAAATTTCAT
LANE 2:
CGGAGTATGTACCGACTGTTTTTGACAACTATGCAGTCACAGTTATGATTGGTGGAGAACCATATACTCTTGGACTTTTTGATACTGCAGGGCAAGTTATGA-
CAGATTACGACCGCTCACTTATCCACAAACAG
LANE 3:
ATGTATTTCTAGTCTGTTTTTCAGTGGTCTCTCCATCTTCATTTGAAAACGTGAAAGAAAAGTGGTGCCTGAGATAACTCACCACTGTCCAAGACNCCTTTCTT-
GCTTGTTGGGACTCAAATTGATCTCAACGAGATGACCC
LANE 4:
CTCTACTATTGAGAAACTTGCCAAGAACAAACAGAAGCCTATCACTCCANAGACTGGGTGAAAAGCTGGCCCGTGACCTGAANGCNGTCAAAGTATGTG-
GAGTGTTCTGCACTTACACAGCAGANGTCTGAAAAATGTGTTNATGAAGC

### Sequence 2

LANE 1:
TGCNNNNNTGGNTNNGGNNNNNATTGNNTCNCTNTACCATGCNNGNGCACAANGTTTTTTTTTTTTTTTTTTGGGCAAAGCGTACAAAGGTTC-
CAAGGGACAGGACCAAGAACGAGGGGCTGAGACATTTACAACAGCAGGCATT

**NOTE: For best results, we recommending eliminating the first 73 nucleotides (in red) when analyzing this lane.*

LANE 2:
TTTCTCTTCCTCTTCTTCACGGGAGGCGGGCANAGGACTGCTCGGATCGCTTCGTCAAACACTGTCTTGAGGCCTCNCTGTGTGAGCGCCGAGCACTCCAG-
GTATTTTACAGCACCAATCTCCTTANCCATGGCTANAC
LANE 3:
CCCTGCGGATAGGTGATGGGAGTCAGCTTCTTCBCCTTCAGTTTCNCNATCGTGTCTTTATCATCCCTAAGATCAAGTTTAGTTCCCACTANGATGATGGGAGT-
GTTGGGACAGTGGTGCCGCACCTCAGGATACCACTT
LANE 4:
TGCACGGACATTTTCAAATGATGCAGGACTCACAAGGGAAAAGCAAATTAAGAACACATCTGTTTGCGGATAGGATAGGGGGCGTATTCTGTCATA-
ATCTTCTTGTCCAGCTGTATCCCAAAAAGCCCAGATTCACCGGTTT

### Sequence 3

LANE 1:
TGNNNNNNTGNNNNNNNGNNANAACGAAGTGCAGACTCAAAAGTGCCATCTCCCTCCCGACCATTGGAGGATCCCAAGCTCTATGTTGCCCTTATTGT-
CACCAGTGACATTTAATTCCAAACAGGAGTCCTTCGGGCCAGCAA
LANE 2:
GCTGCCCAGGCTTAGCTGCGAGCCCGTCATGGAGGAAAAAAGCTCAGGAGAAAACCAGTCTGTTGGAGAATGGGACAGTCCACCAGGGAGACACCTC-
GTGGGGCTCCAGCGGTTCTGCATCTCAGTCCAGCCAAGGCAGA
LANE 3:
GACAGCCACTCCTCCAGCCTGTCCGAACAGTACCCCGACTGGGCCAGCCGAGGACATGTTTGACCATCCCACCCCATGCGAGCTCATCAAGGGGAAAGAC-
TAAGTCAGAGGAGTCCCTCTCTGACCTTACAGGTTCCCTCCCTCCTCTCC
LANE 4:
CCTGCAAGCTTGATCTTGGGCCCTCACTTTTGGATGANGTGCTGAATGTTATGGATAAAAATAAGTAACTCGAGCATGCATCTAGAAGGGCCTATTCTATA-
ATGTCACCTAAATGCTAAACCTCGCTGATCAGCCTCGACTGTGCCNT

**NOTE: For best results, we recommending eliminating the last 70 nucleotides (in red) when analyzing this lane.*

# Answers to Exercises

Signal transduction is an important process by which cell receptors act as signal amplifiers. The cell surface membrane contains complex molecular sensors that receive, amplify, and react to extracellular signals involved in cell growth and differentiation. The loss of a cell's ability to react to these signals can result in cell transformation and the onset of various cancers.

The set of three sequences correspond to nucleotides for distinct genes which are functionally related and play important roles in signal transduction. The identity of these genes (and their protein products) can be determined using the NCBI website as described previously in this experiment. Please note that details about Gene ID, strand, references, and author information will be different depending upon which hit a student focuses on. However, the gene itself should match the information presented below. After researching potential diseases related to these genes, students should discuss the related bioethical issues.

### Sequence 1: Cell division cycle 42 (GTP-binding protein) or *CDC42*

*CDC42* is a key intracellular regulatory molecule involved in cell shape. In mammalian cells CDC42 protein regulates the actin cytoskeleton (the scaffolding of a cell) to produce hair like structures called filopodia, which are involved in cell sensing and cell movement. Mutations in CDC42 and its associated proteins are correlated to certain types of cancer.

### Sequence 2: Ras-related C3 botulinum toxin substrate 1 (rho family, small GTP binding protein Rac1) or *RAC1*

The *RAC1* gene codes for a regulatory protein that alters the actin cytoskeleton to produce membrane ruffles, which are important for cell motility. RAC1 is important in a number of diseases, including cancer. Many tumors will utilize RAC1 signaling to gain increased motility and invasiveness, one of the early stages required for cancer to spread throughout the body.

### Sequence 3: CDC42 effector protein (Rho GTPase binding) 3 or *CDC42EP3*

*CDC42* (sequence 1) regulates the formation of F-actin-containing structures through its interaction with several downstream effector proteins. CDC42EP3, one of these effector proteins, is involved in binding the protein CDC42, thereby changing the shape of a cell. A cell senses the external environment by setting up signaling networks that tell the cell how to respond to external signals. Defects in these signaling molecules are involved in many diseases, including cancer.

**Please refer to the kit insert for the Answers to Study Questions**