TEACHING THE GENOME GENERATION

BIOINFORMATICS EXERCISE

BEFORE YOU BEGIN

These exercises can be assigned as homework.

The Jackson Laboratory
Leading the search for tomorrow's cures

PRE-REQUISITES & GOALS

STUDENT PRE-REQUISITES

Prior to implementing this lab, students should understand:

- All previous pre-requisites
- The benefits of knowing a DNA sequence and its applications
- How DNA sequences code for proteins.
- The affect a mutation in the DNA sequence has on the protein.
- The function of a Stop codon.
- How to read a phylogenetic tree

STUDENT LEARNING GOALS

1. Explore the NCBI, OMIM and MGI database websites.
2. Locate gene files for several genes involved in human genetic diseases.
3. Compare the nomenclature of human genes and similar genes found in other species.
4. Visualize the structure of a gene and its mutations as demonstrated in an online database, (intron/exon structure).
5. Discuss variations within genes.
6. Compare genes and proteins from across species.
7. Find animal models for human disease.
8. Perform single and multiple alignments.
9. Discuss the relationship between DNA and amino acids.
10. Describe evolutionary connections between species based on amino acid sequences of homologous genes.

A searchable gene database is available through the National Center for Biotechnology Information (NCBI), in association with the National Library of Medicine (NLM) and the National Institutes of Health (NIH). Knowing this will help you remember the web address: *www.ncbi.nlm.nih.gov* Although NCBI provides a vast amount of information related to biotechnology, this exercise will focus on the database of genes.

BACKGROUND

HUMAN DISEASE	CAUSATIVE GENES
Multiple diseases	APOE, APOB
Cancer	BRCA1
Multiple hereditary exostoses	EXT1, EXT2
Polycystic kidney disease	PKD1, PKD3
Huntingtons Disease	HTT
Fragile X syndrome (mental retardation)	FMR
Neurofibromatosis	NF1
Cystic Fibrosis	CFTR (ABCC7)
Sickle Cell anemia	HBB
Marfan Syndrome	FBN1
Tay-sachs Disease	HEXA
Duchenne muscular dystrophy	DMD
TtGG genes (no disease)	ACE, CYP2C19, OXTR

Many, if not most, genetic diseases are caused by multiple genes. To the left is a list of some genetic diseases and their known causative genes. Although the genes listed are known to play a role in these human diseases, it is most likely that other genes and non-coding DNA sequences are also involved.

PROTOCOL STRUCTURE

PART 1	35 minutes
PART 2	25 minutes
PART 3	15 minutes
PART 4	35 minutes
PART 5	30 minutes
PART 6	25 minutes

PROCEDURE

PART 1 - Introduction to Databases

☐ STEP 1
Start at the NCBI homepage:
www.ncbi.nlm.nih.gov

☐ STEP 2
Use the **All Databases** drop down menu to select **Gene**.

☐ STEP 3
Type a gene symbol (from the list above, CYP2C19 in this example) in the search box and hit enter or **Search**.

NOTE: By convention, gene names are ALL CAPS for Human genes and only the first letter capitalized for other species. (e.g. CYP2C19, Human gene; Cyp2c19 mouse gene)

☐ STEP 4
Explore the results on this page. Fill in the blank:

Human gene ID: 1557
Full human gene name: cytochrome P450 family 2 subfamily C member 19
List three other species with the same gene: Pan troglodytes (chimp), Bos taurus (cattle), Xenopus tropicalis (frog)

☐ STEP 5
Click on the link for the human gene that you searched.

NOTE: This link should be in ALL CAPS

☐ STEP 6
Find the gray **Summary** bar.

Write a brief description of the gene summary:

This gene encodes a member of the family of cytochrome P450 enzymes which catalyze metabolic reactions, specifically drugs. It is located in the ER.

☐ STEP 7
Find the gray **Genomic context** bar. What is the:

Chromosomal location: 10q23.33
Exon count: 9

☐ STEP 8
Scroll down to the gray **Genomic regions, transcripts and products** bar.

NOTES:

a. Green boxes represent exons.

b. Horizontal green lines represent introns.

☐ STEP 9
Find and click on the **ATG** icon  at the top of the genome browser view.

What is the function of this button?

Hint: look in the dark horizontal bars at the top of the window

This button zooms in to the nucleotide level instead of looking at the whole gene.

NOTE: In this view, there may be red bars with

IDs like rs936726927, which indicate that a variant in that location is known to exist in humans

☐ STEP 10

Scroll down to the gray **Bibliography** bar. Click on **What's a GeneRIF** in the **GeneRIFs: Gene References into Functions** section.

How does a GeneRIF article compare to other cited articles in the Bibliography?

GeneRIF articles discover and study the function of the gene as opposed to other articles which may look at variant prevalence in populations, therapies, or technologies that utilize the gene.

☐ STEP 11

Navigate back one page and in the **Related information** panel on the right, click on the **OMIM** link.

NOTE: Online Mendelian Inheritance in Man (OMIM) is a curated database on human medical genetics and has valuable summaries of gene-disease relationships. It includes molecular genetics, clinical information and historic perspective.

☐ STEP 12

Type in the full name of the gene in the search box and click on the result that best matches your query.

☐ STEP 13

Click on **Allelic Variants Table View** in the left hand panel to see the common variation(s) in the gene you are investigating.

NOTE: For CYP2C19, four selected variants are described and can be explored further. Number 0.003 (rs4986893) is the variant genotyped in TtGG.

☐ STEP 14

Return to the NCBI homepage by navigating to: *www.ncbi.nlm.nih.gov*

☐ STEP 15

Use the **All Databases** drop down menu to navigate to **HomoloGene.**

NOTE: A homolog is a gene related to a second gene by descent from a common ancestral DNA sequence. The term, homolog, may apply to the relationship between genes separated by the event of speciation (see ortholog) or to the relationship between genes separated by the event of genetic duplication (see paralog).

☐ STEP 16

Type your gene abbreviation (from the list above, switching to ACE for the example) in the search box and hit enter or **Search.**

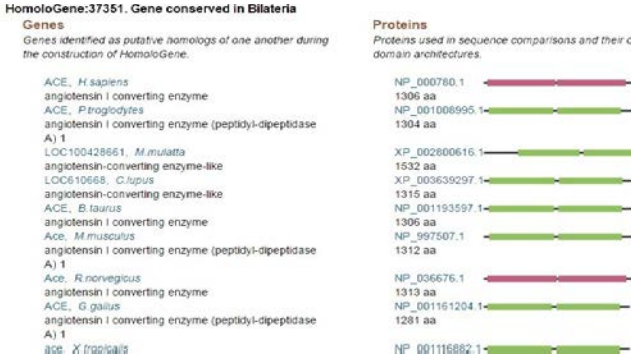
☐ STEP 17

If a **Search results** page appears, click on the link that is most relevant to your gene (for ACE this is the sixth result).

☐ STEP 18

This should take you to a page that looks similar to the image below.

NOTE: All genes in all organisms will not share the same gene nomenclature.



☐ STEP 19

Find the **Protein Alignments** heading and click **Show Multiple alignment.**

NOTE: Amino acid alignments will be shown in repeating 50 amino acid segments with the various species on separate rows.

☐ STEP 20

Scroll over the **XP** and **NP** links to make the species name appear.

☐ STEP 21

Click on a species name that you do not recognize other than **H. sapiens.**

What is the Latin and common name of the Source Organism? Hint: find SOURCE line

Canis lupus familiaris (dog)

☐ STEP 22

Navigate back one page to the **Multiple Sequence Alignment** page.

Provide an example of one amino acid where the organism you chose and humans differ (provide the amino acid number and the evolutionary change).

By looking at the stacks of amino acid single letter codes in vertical columns, look for columns where there are multiple letters. If you are following along with the ACE example, there is are two - - symbols in the human row (meaning no amino acid) and a PG in the *C. lupus* row at position #8 in the amino acid sequence. Therefore, there is an insertion of PG (or proline and glycine) at amino acid 8 in dogs.

☐ STEP 23

Navigate back one page and under the **Genes** heading click on the *M. musculus* (mouse) link.

☐ STEP 24

Find the gray **Summary** bar and the **Primary source** header and click on the **MGI** link.

☐ STEP 25

You are now in the Mouse Genome Informatics (MGI) database where you can find extensive resources regarding mouse genetics.

☐ STEP 26

Scroll down to **Mutations, alleles and phenotypes.**

How many alleles are available for study in mice? 20

☐ STEP 27

Click on the number after **All mutations/alleles.**

☐ STEP 28

Pick an allele to investigate.

Allele: Ace^{tm1Keb}

☐ STEP 29

Click on the allele symbol and find the **Mutation description** section.

Summarize the mutation description:

There should be a detailed description of how this allele was generated or found. In this example, it is a knockout allele created by insertion of a Neomycin selection cassette into intron 12, the insertion disrupts gene function.

☐ STEP 30

Find the **Phenotypes** heading.

What physiological systems are affected by altering this gene in the mouse? this gene seems to mostly affect cardiovascular systems

PART 2 - Explore the human CYP2C19 protein

☐ STEP 1

Go to the NCBI homepage.
www.ncbi.nlm.nih.gov

☐ STEP 2

Use the **All Databases** drop down menu to navigate to **Protein.**

☐ STEP 3

Search for CYP2C19 in the search box.

☐ STEP 4

Scroll down to find **cytochrome P450 2C19 precursor [Homo sapiens]** with accession number **NP_000760.1** and click on the link.

☐ STEP 5

At the top of the window, use the **GenPept** drop down menu to change the setting from **GenPept** to **FASTA.**

NOTE: The sequence is a series of letters that represent amino acids, not DNA bases. This is the amino acid sequence of the protein.

☐ STEP 6

Copy the sequence, include it's header.

☐ STEP 7

Go to the NCBI BLAST server:
blast.ncbi.nlm.nih.gov/Blast.cgi

☐ STEP 8

Click the **Protein BLAST** icon.

☐ STEP 9

Paste the FASTA sequence in to the box at the top of the window under the **Enter Query Sequence** header.

☐ STEP 10

Under the **Choose Search Set** header, find **Database** and ensure **Non-redundant protein sequences (nr)** is selected.

☐ STEP 11

Under the **Program Selection** header, ensure **blastp** is selected.

☐ STEP 12

Click on the large blue **BLAST** button on the bottom of the window.

NOTES:

1. Be patient, the interface will take a few minutes to load.

2. Under **Graphic Summary**, the top panel displays the protein family to which the query belongs, in this case, **p450 superfamily**

3. The next graphical panel demonstrates alignment quality of the hits to the query input sequence; red bars represent well-aligned sequences. The length of the bar represents the length of the alignment; the longer the line, the longer the duration of similarity.

4. Under **Descriptions** you will find a box containing the "hits" or the protein sequences in the NCBI database that align to the human CYP2C19. The first is your input sequence – Homo sapiens.

5. **Max score**, **Query cover** and **Ident** are measures of how similar each hit is to the input sequence.

☐ STEP 13

In order to see a two species alignment, click on one of the **Pan troglodytes (chimpanzee)** links.

NOTES:

1. In the alignment, the first row is your query sequence, the bottom row is the hit sequence and the middle row is the consensus between the two sequences.

2. Perfect **Identities** match is indicated by the same letter, **Positives**, or amino acids with similar chemical properties, are indicated by + and **Gaps** are indicated by —.

☐ STEP 14

Scroll back to the top of the "hit" list and find a hit from *Nomascus leucogenys*.

What is the common name of ***N. leucogenys***? Hint: click the Accession number

northern white cheeked gibbon

What is the percent identity of CYP2C19 between human and ***N. leucogneys***?

Hint: Look in the Ident column 97%

What does this value indicate about the evolutionary relationship between humans, chimps and this species?

because the similarity between gibbons and humans is lower than chimps and humans, this may indicate that chimps are more similar to humans than gibbons

☐ STEP 15

Return to the "hit" list and perform the same exercise for *Equus caballus*.

What is the common name of *E. caballus*? horse

What is the percent identity between human and *E. caballus*? 82%

What does this value indicate about the evolutionary relationship between the four species?

for similar reasons as above, horses are less similar to humans than chimps or gibbons

PART 3 - Identify an unknown sequence

☐ STEP 1

Return to the NCBI BLAST server: *blast.ncbi.nlm.nih.gov/Blast.cgi*

☐ STEP 2

Click the icon for **Nucleotide BLAST**.

☐ STEP 3

Copy and past the following *unknown sequence* into the box at the top of the screen under the **Enter Query Sequence** header.

☐ STEP 4

Ensure that **Others** and **Nucleotide collection (nr/nt)** is selected under the **Database** header and **Highly similar sequences (megablast)** is selected under **Program Selection**.

☐ STEP 5

Click on the large blue **BLAST** button on the bottom of the window.

Given the "hit" list, what gene does this sequence represent? OXTR

What species does this sequence derive from? Explain.

It is likely that this sequence derives from human as it is a 100% identity match for several Homo sapien sequences

PART 5 - Analysis of DNA sequences using Clustal Omega

☐ STEP 1

Navigate to the Clustal Omega website.
www.ebi.ac.uk/Tools/msa/clustalo

☐ STEP 2

Under **STEP 1 - Enter your input sequences**, set the drop down menu to **DNA** and copy and paste the *eleven species sequences* provided for the ACTN3 gene.

NOTE: Each sequence must begin on a new line.

☐ STEP 3

Under **STEP 2 - Set your parameters**, select **Clustal w/ character counts.**

☐ STEP 4

Click on the green **Submit** button.

☐ STEP 5

View your results in the **Alignment** tab and scroll to the bottom to see how many bases you submitted for each species.

NOTE: A perfect match between the species is indicated by *, if there is a two or more base polymorphic site, it is indicated by a space.

What is the range of sequence lengths of the ACTN3 gene you submitted for the 11 species?

2,793-2,952

☐ STEP 6

Click on the **Results Summary** tab and click on the link under **Percent Identity Matrix.**

NOTE: This table shows how the DNA sequence for each of the 11 species compares to the other 10 and itself. The identity of the rows matches the columns. Therefore, the percent identity is 100% along the diagonal.

☐ STEP 7

Using the **Homo (sapien)** row, list the 10 other species and their percent identity to humans for ACTN3 and complete the table below.

Genus Name	Common name	% Identity
Mus	mouse	86.81
Rattus	rat	87.04
Equus	horse	92.63
Balaenoptera	whale	90.40
Odobenus	walrus	91.39
Pteropus	bat	91.23
Macaca	macaque	97.18
Gorilla	gorilla	98.78
Pan	chimp	98.78
Pygmy	gorilla	99.68

☐ STEP 8

Go back to your results and click on the **Phylogentic Tree** tab, scroll down to **Phylogram** and ensure **Cladogram** is selected to visualize the relationships between the species.

NOTE: Branches (or nodes) in the tree represent common ancestors between species or groups of species. Just because two species names appear next to each other in the list, does not mean they are closely related.

☐ STEP 9

Toggle to the **Real** view of branch length.

What happens to the tree? Why do you think this is so? the tree shrinks in Real view, as this view represents evolutionary time

What do you think the numbers represent after the species name in the tree? pairwise distance to the nearest neighbor

Choose any two pairs of species and explain how they are evolutionarily linked. HINT: Keep in mind that this analysis is evaluating a single gene.

varied responses

Does anything about this tree surprise you? varied responses

☐ STEP 10

Scroll to the top of the page and click the **Share** icon at the top right to send an email to your teacher with your results.

PART 6 - Analysis of protein sequences using Clustal Omega: Solve a real world problem

☐ STEP 1

Navigate to the Clustal Omega website.
www.ebi.ac.uk/Tools/msa/clustalo

☐ STEP 2

Under **STEP 1 - Enter your input sequences**, set the drop down menu to **Protein** and copy and paste the *seven species sequences* provided.

NOTE: Each sequence must begin on a new line.

☐ STEP 3

Under **STEP 2 - Set your parameters**, select **Clustal w/ character counts**.

☐ STEP 4

Click on the green **Submit** button.

☐ STEP 5

View your results in the **Alignments** tab and scroll to the bottom to see how many amino acids you submitted for each species.

NOTE: A perfect match between the species is indicated by *, if there is a two amino acid polymorphism at that location it is indicated by a period (.), if it is a three amino acid polymorphic site it is indicated by a colon (:), and if it is a 4 or more amino acid polymorphic site it is indicated by a space.

What is the range of sequence lengths you submitted for the 7 species?

657-628

What is the minimum number of genomic DNA bases this represents?

1971-1884

Can you precisely determine the number of DNA bases? Why or why not?

No, as these species may have variable number or sizes of introns

☐ STEP 6

Click on the **Results Summary** tab and click on the link under **Percent Identity Matrix**.

Which two protein sequences share the most similarity?

NOTE: The identity of the rows matches the columns.
Human and Dog

☐ STEP 7

Go back to your results and click on the **Phylogentic Tree** tab, scroll down to **Phylogram** and ensure **Cladogram** is selected to visualize the relationships between the species.

☐ STEP 8

Scroll to the top of the page and click the **Share** icon at the top right to send an email to your teacher with your results.

Looking at the **Cladogram** for these seven species, describe what the visualization means.

Hint: Describe the groupings, the nodes and what the numbers mean (toggle to Real view as well).
Answers should be a summary of previous responses.

How many proteins were evaluated? 2

How many species were evaluated? 4

Does anything about this tree surprise you? varied responses - tree gets wonky when evaluating two genes. Genes cluster, not species.

The primary article citation for this "Real World Investigation" can be found at:
www.ncbi.nlm.nih.gov/pubmed/22203040
Tossidou, I. et al. CD2AP regulates SUMOylation of CIN85 in podocytes. Mol Cell Biol. 2012 Mar;32(6):1068-79.
The investigator was trying to determine if two genes in the same family were evolutionarily related to each other.