# Assignment #2

**Part I:**
Before starting this assignment, be sure to go through the notes on "Sequence Databases" and "Genes and Databases" and read Chapters 1 and 2 of the textbook. In this assignment you want to become familiar with searching and gathering information from a number of freely accessible biological databases.

Search for information on one of the following human genes (you will be assigned one):

NPC1 - Niemann-Pick C1 protein
Nav1.5 – Sodium ion channel
KIF1A – Kinesin family member 1A
CLOCK – Circadian regulator
SPR – Sepiapterin Reductase
PolB – DNA polymerase beta chain A
HBB – hemoglobin subunit beta
CFTR – cystic fibrosis transmembrane regulator

in each the following (at least four) databases :

1) GenBank
2) OMIM
3) Ensembl
4) Other biological sequence, protein structure, or function databases that interest you.

Your report should be well formatted with a list of the names at the top of all the people who contributed to the report. Use separate sections and subsections clearly labeled.
From each of three databases: GenBank, Ensembl, and one of your choice, assemble 3 useful and "complete" human sequences for your gene:

- ◦ genomic DNA
- ◦ matching cDNA
- ◦ matching protein – make sure it is the full protein

A total of 9 sequences. For each sequence, show

i. the GenBank or Ensemble accession numbers (or similar identification)
ii. the length of the sequence
iii. human chromosome locus
iv. the sequence itself - copy the sequence from the database site and paste it into the document.
    Use good formatting to make this easy to read:
    1. use a fixed-width font like `Courier` for the nucleotides or amino acids.
    2. single spaced lines
    3. numbers on the left side
    4. each sequence line should fit on a single line

    Note for genomic sequences: if your sequence is more than 10,000 bp long, you don't have to list the sequence in your document – just provide a link to the website with the sequence.
v. genomic DNA exon positions and alternative splice variants – a "screenshot" of a graphical view will work
vi. a reference for the earliest (journal article or other) source of the sequence

Write a 1 page summary (12 pt font, single spaced) of the function of the gene and one or more genetic diseases known to be caused by mutations in the gene – while you can copy and paste sequence and reference information, the gene function and disease information should be a summary in your own words – **do not** copy and paste.

**Part II**

Go through the Lecture on "Sequence Alignment I" (Lecture 5) and read up on BLAST.

Bioinformatics studies have revealed a surprising amount of similarity between genes in very different creatures. In some cases, these similarities may provide new avenues for understanding clinically important genes in humans. In this part of the assignment you will look for plant homologs of human genes, while reinforcing the skills needed to retrieve sequence information from publicly available databases.

Assemble the following specific information clearly marked as asked below:

1. List the accession number and mRNA sequence for your gene from Part I.
2. Underline the translation start and stop codons in this mRNA. Highlight the 5' UTR (yellow), the 3' UTR (red) and coding sequence(s) (green)
3. Retrieve the human protein sequence corresponding to your gene from NCBI, list the sequence accession number and paste the full amino acid sequence into your document. If the protein amino acid sequence does not **match the mRNA nucleotide sequence**, explain the discrepancy.
4. Take the protein sequence and use the BLASTP program on the TAIR website (http://www.arabidopsis.org/Blast/) to search for Arabidopsis proteins with sequences similar to your protein sequence. Paste the Arabidopsis protein sequence that best matches your human sequence and make sure the similarities are clearly displayed – Use a font like **`Courier`** font to make your report readable.  
Note that the best BLAST match may not give you the best match as far similar function between the human and plant sequences – look closely at the top 5 (at least) BLAST results.
5. A BLASTP "hit" with an E-value of less than $10^{-30}$ is considered "stringent" – meaning that it is not likely to be a random match with the query sequence. Note the E-value of the Blast results: how many good "hits" do you get?  
Describe the functions of the three best Arabidopsis matches and explain whether the functions of these proteins resemble those of the human sequences