Step by step summary of how to do a sequence alignment/comparison

1. Open Ensembl Genome Browser (http://uswest.ensembl.org/index.html)

2. Choose species from the drop-down search menu or "All genomes" button.

3. Search for gene of interest in the appropriate genome on ensembl. (Select the species from the drop-down search menu, and type in the name of the gene you're looking for.)

4. Select the best matching transcript from the list of results. To get the cDNA sequence, click “cDNA seq.” under the transcript you chose. In a moment, you will see all the nucleotides in the sequence, translated into amino acids for protein (those random-seeming letters under each set of three nucleotides).

5. From the bottom of the left menu, click “BLAST this sequence”

6. Under "Enter the query sequence" find the sequence; highlight and copy it.

7. In another window, open BLAST program (http://blast.ncbi.nlm.nih.gov/Blast.cgi)

8. Under Basic Blast choose “nucleotide blast”. This means you are going to compare a nucleotide sequence to a nucleotide sequence. (It is also possible to compare protein sequences to predicted DNA sequences, and vice-versa.)

9. Paste the cDNA sequence into the box "Enter query sequence / Enter accession number(s) etc." For the record, you are entering in FASTA format.

10. Under “Choose Search Set” select “Human genomic + transcript”. That means you'll be looking in the human genome for DNA sequences that are similar to the cDNA you entered from your test organism.

11. Under "Program Selection," try "Highly similar sequences" first. If you don't get a match, repeat with "Somewhat similar sequences (blastn)"

12. Click “BLAST” and wait.

Ta-da! A whole bunch of data will appear. You can ignore most of it, including the colorful graphic at the top. Here's what you should look for:

- Under "Descriptions" / "Sequences producing significant alignments", look at the first item in the list. It will probably say Homo sapiens and the name of the cDNA you searched, mRNA. This is your match, the sequence that the BLAST program pulled out of the entire ~3 billion nucleotides of the human genome. The two numbers of interest are Query Cover % and Ident %. **Query cover** is a measure of how much of the sequence you tested overlaps with human sequence. In some cases, only a tiny bit of the query sequence is found in humans. **Ident %** is the percent of nucleotides in the overlapping area that are absolutely identical between humans and the query species. Generally, both query cover and percent identity go up or down together.
- Under "Alignments", look at the first one. It shows the actual DNA sequence you entered (Query) lined up with a human gene (Sbjct). Here you can see where the percentages for cover and identity were calculated.

Print out the Descriptions and the *first* alignment for your data.

Examples of genes to test:

A) DNA polymerase: a molecular machine that copies DNA, allowing cells to reproduce. One part of this machine is called ***DNA polymerase beta*** *(or DNApol beta or polB). You can try some kind of RNA polymerase, too, if you wish.*

B) ATP synthase: this molecular machine performs the last step in cellular respiration (making ATP from food in the presence of air). Search for the ***ATP synthase gamma subunit****.*

C) Glucokinase: this protein performs the first step in glycolysis. Search for ***glucokinase 4*** *or its cousin, hexokinase.*