Obesity Induced Changes in Gene Expression

This project will focus on analyzing simulated RNAseq data in order to evaluate differential gene expression between control and treatment samples. Your goal is to combine your knowledge of Unix, Python/R, and bioinformatic tools *muscle* and *hmmer* to "replicate" a recent study of gene expression in kidney tissues of normal (control) and obese mice (Kuhns & Pluznick 2017). RNAseq involves high-throughput sequencing of cDNA libraries constructed from all expressed RNA in the target tissue. Usually, RNAseq sequences are filtered and aligned to a transcriptome or genome assembly with established bioinformatic pipelines, including tools such as *Bowtie* or *Tophat*. Expression levels of each transcript in each tissue sample are then quantified/converted to counts using statistical models implemented in programs such as *HTSeq* and differential expression assessed with various programs/packages (e.g. *edgeR* or *DESeq*). These tools build on smaller workflows and algorithms that we will replicate in this project. You will approximate an established pipeline by building protein models of 6 target transcripts, searching simulated transcriptome data for "hits," and counting those hits as a proxy for gene expression.

Begin by inspecting your files, you should have five `.fasta` files to work with. Recall that a `.fasta` file is commonly used to store a number of biological sequences (RNA, DNA, or protein). The convention for these files is for each sequence to be delimited by a line that starts with `>` and then contains a sequence identifier of some sort. Following this line, the actual sequence information follows on one or more lines. The next sequence starts with another `>` and sequence identifier, followed by the actual sequence information, and so on. The files for your project are:

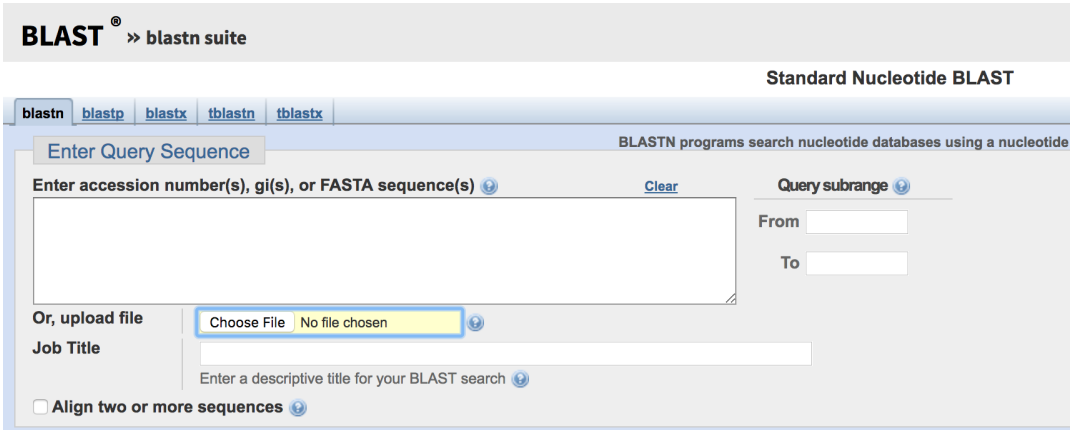
1.*uniquetranscripts.fasta*: A fasta file including the sequences of 6 transcripts that showed differential expression between kidney tissues of normal (control) and obese mice in Kuhns & Pluznick 2017.

2&3.*control1.fasta* & *control2.fasta*: Transcript sequences produced in a "new *RNAseq* experiment" with kidney tissue from two normal (control) male mice. These sequences have been translated into amino acid sequences.

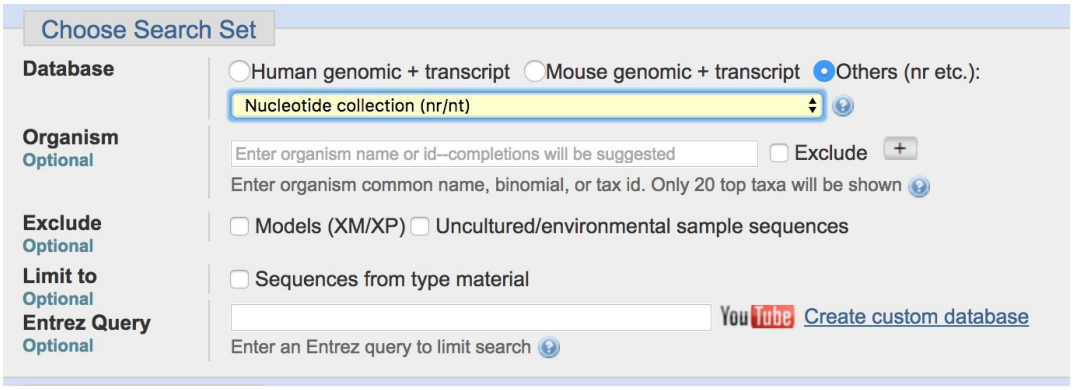
4&5.*obese1.fasta* & *obese2.fasta*: Transcript sequences produced in a "new *RNAseq* experiment" with kidney tissue from two diet-induced obese male mice. These sequences have been translated into amino acid sequences.

Use BLAST to identify the genes encoded by the 6 differentially expressed transcripts listed in *uniquetranscripts.fasta*.

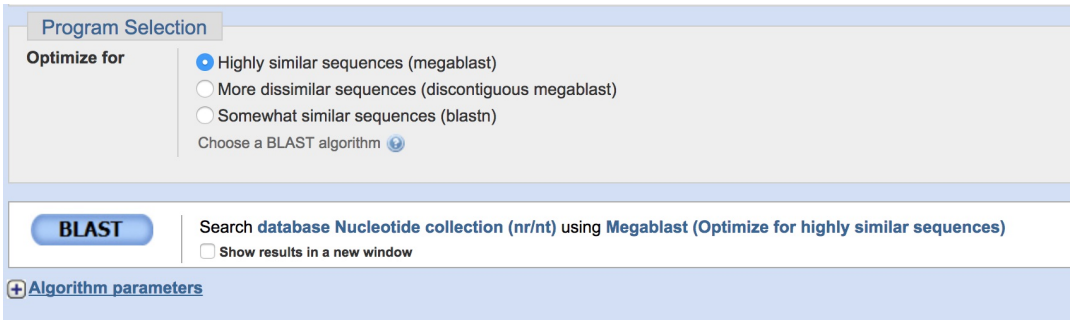
Go to https://blast.ncbi.nlm.nih.gov/Blast.cgi and select *nucleotide BLAST*, because your data are transcriptome (cDNA) sequences. Upload *uniquetranscripts.fasta* via the *choose file* button shown below.



For this assignment, search the *Nucleotide collection (nr/nt)* Database.

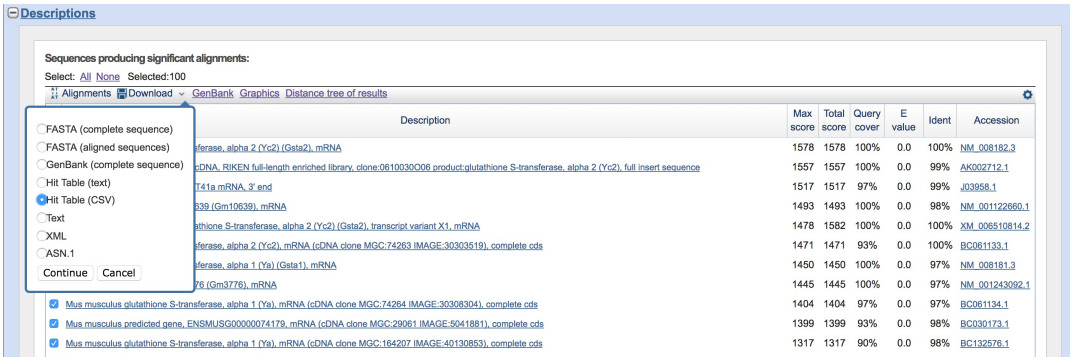


Also, use the default optimization for *Highly similar sequences (megablast)*. And click *BLAST* to initiate the search.



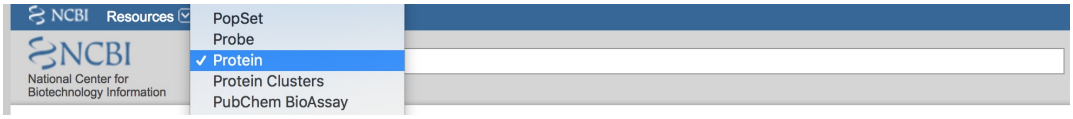
When the searches are complete, your browser will redirect to the results page. The drop down menu next to *Results for:* will allow you to toggle between results for the 6 transcripts in your fasta file.

Scroll down to the *Descriptions* heading for your table of hits, which you should **save to your repo**. The best hits, listed first, will have extremely low E-values and high (~100%) Identity. By selecting *All* (top left of the table) the *Download* button becomes available and you can choose the download format from the dropdown menu. **Using Unix commands or R/Python, make a single table that includes the top hit for each transcript.**

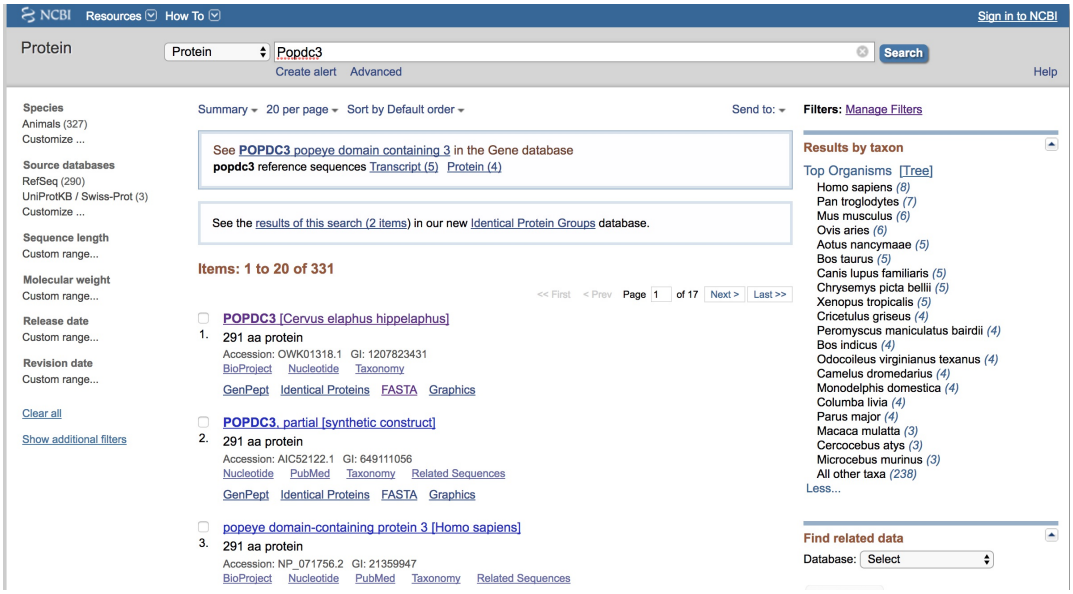


Search the NCBI protein database for amino acid sequences corresponding to these 6 transcripts.

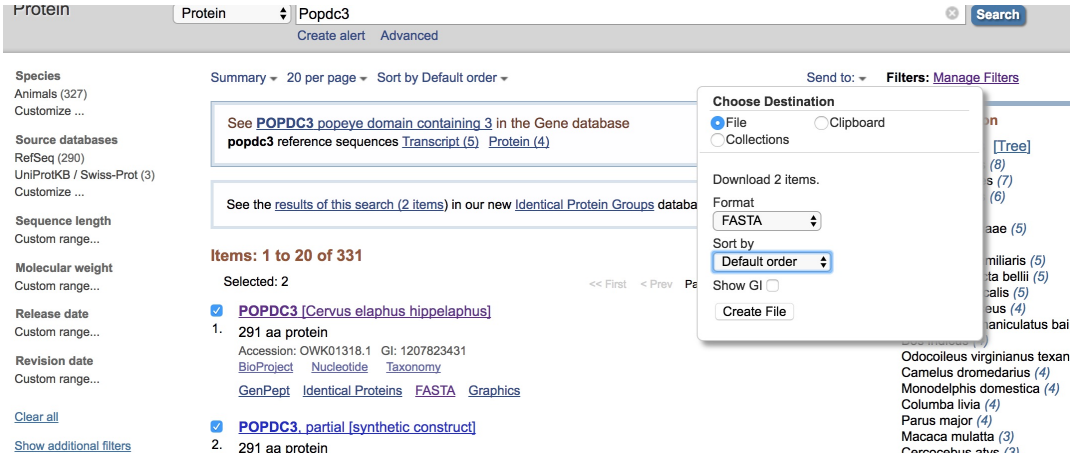
Point your browser to the NCBI homepage https://www.ncbi.nlm.nih.gov/ and choose the *Protein* database.



Based on the name of the protein listed in the top hit for each transcript in the previous step, search the NCBI protein data base for sequences. By selecting *Animals* under *Species* at the top left of the page, the option to filter by taxon should appear at the top right.



Depending on how conserved a protein is, the protein sequence can vary substantially between distantly related organisms (view taxa as a *TREE* to display relationships). This could change the efficacy of the HMM model you will build below. Choose 10 protein sequences from mice (*Mus musculus*) and other closely related organisms (if available) by checking the boxes to the left of the listed matches. Download your selected sequences by clicking on the arrow next to *Send to:*, in the upper right corner, selecting *File* and choosing *FASTA* from the dropdown menu. **You should save one fasta file of protein sequences per identified transcript (6 total).**



Build a Hidden Markov Model for each of the 6 transcript proteins and search the 4 translated "RNAseq files."

For each of the 6 transcripts make a muscle alignment from your downloaded protein sequences. Using these alignments, construct 6 HMM protein models using hmmbuild. Finally, search all 4 translated "RNAseq files" for each of the 6 HMM protein models using hmmsearch. **Use a bash script to loop over the 6 transcript files and 4 "RNAseq files," executing these commands.**

Graph the "expression levels" of each protein in each of the "RNAseq files."

Based on the counts of hmm hits for each transcript (our measure of RNA expression) in each "RNAseq file," make a **graphical comparison of expression levels** across the 2 normal and 2 obese mice. **Qualitatively compare these results to those reported in Kuhns & Pluznick 2017.**

Short Answer: Extending your work

1. For 2 of your 6 transcripts, return to the original *BLAST* search and change the *Optimize for* option. Quantitatively, how do *discontinuous megablast* and *blastn* change your table of *BLAST* hits? This may be easier to explore, if you also restrict the *Database* option to either *Human* or *Mouse*. (You do not need to repeat any other steps here.)

2. For 2 of your 6 transcripts, return to the NCBI protein search and explore the effects of phylogenetic relatedness of your amino acid sequences on the performance of your HMM model. For example, what would happen if you built your HMM protein model using more distantly related mammals (e.g. primates)? Would you still get the same quality hits if your HMM protein model was based on non-mammalian sequences? Pick one of the "RNAseq files" to search in order to test your hypotheses. Compare e-values among HMMs built from differing taxa.