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-througput sequencing of cDNA libraries constructed from all expressed RNA in the target tissue. 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*). In this project, you will approximate this process 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 six to work with:

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.

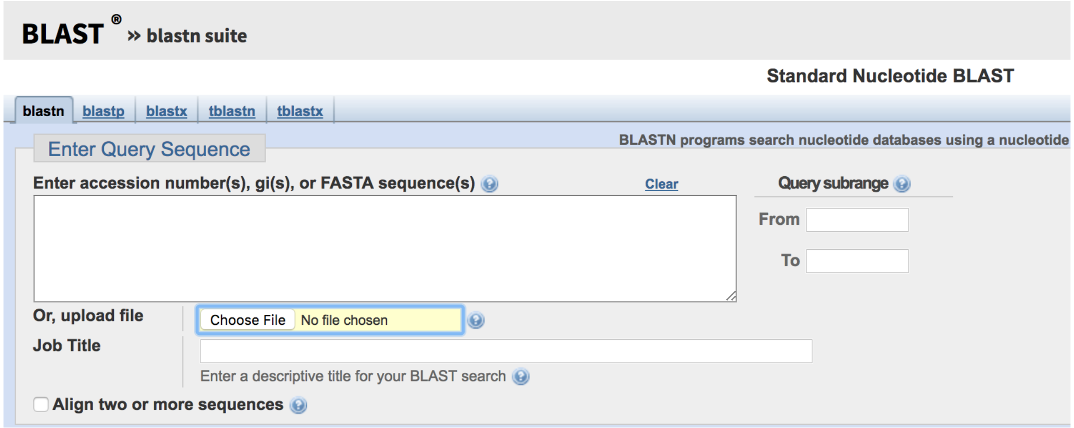
2.*codonmap.txt*: A tab delimited text file with amino acid abbreviation in the first column and associated three base pair codons in the second.

3&4.*control1.fasta* & *control2.fasta*: Transcript sequences produced in a “new *RNAseq* experiment” with kidney tissue from two normal (control) male mice.

5&6.*obese1.fasta* & *obese2.fasta*: Transcript sequences produced in a “new *RNAseq* experiment” with kidney tissue from two diet-induced obese male mice.

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.

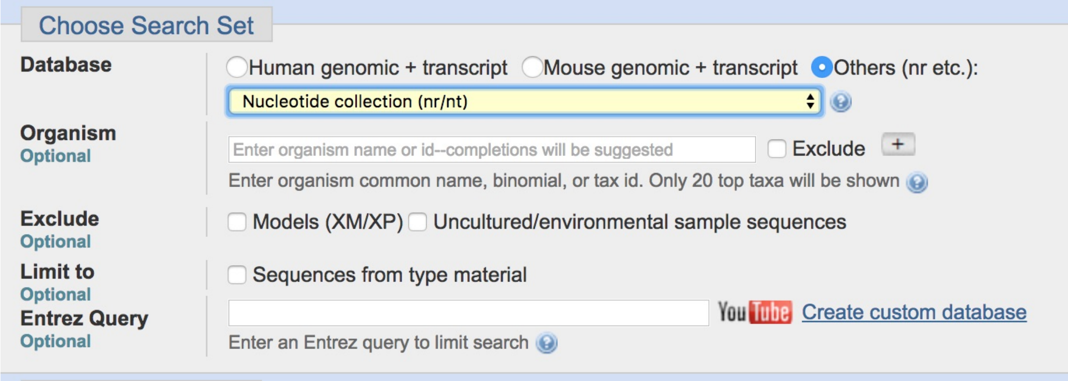


BLASTN programs search nucleotide databases using a nucleotide

No file chosen

Enter a descriptive title for your BLAST search

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



Choose Search Set

Database: ○ Human genomic + transcript ○ Mouse genomic + transcript ● Others (nr etc.):

Nucleotide collection (nr/nt)

Organism Optional: Enter organism name or id--completions will be suggested □ Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

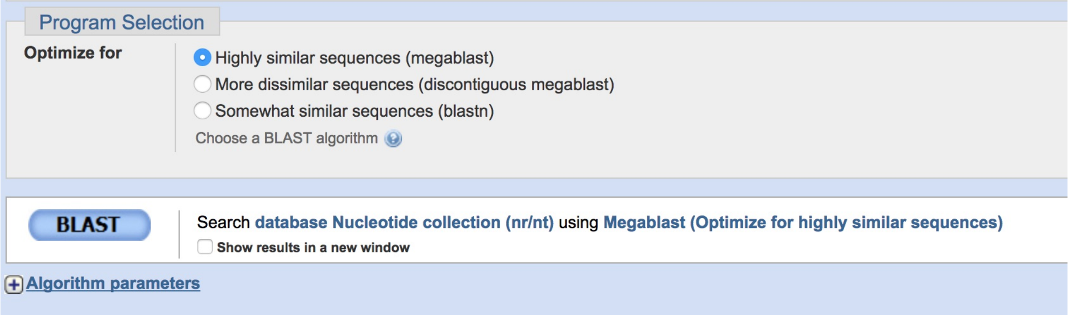
Exclude Optional: □ Models (XM/XP) □ Uncultured/environmental sample sequences

Limit to Optional: □ Sequences from type material

Entrez Query Optional: YouTube Create custom database

Enter an Entrez query to limit search

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



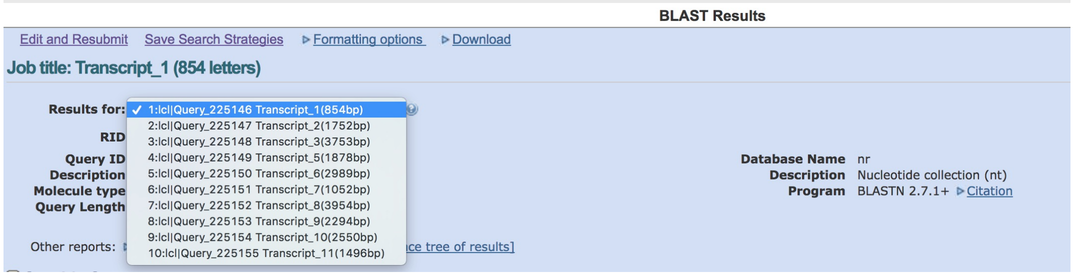
Program Selection

Optimize for: ● Highly similar sequences (megablast)
○ More dissimilar sequences (discontiguous megablast)
○ Somewhat similar sequences (blastn)
Choose a BLAST algorithm

BLAST — Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)
□ Show results in a new window

⊕ Algorithm parameters

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.



BLAST Results

Edit and Resubmit Save Search Strategies ▷ Formatting options ▷ Download

Job title: Transcript_1 (854 letters)

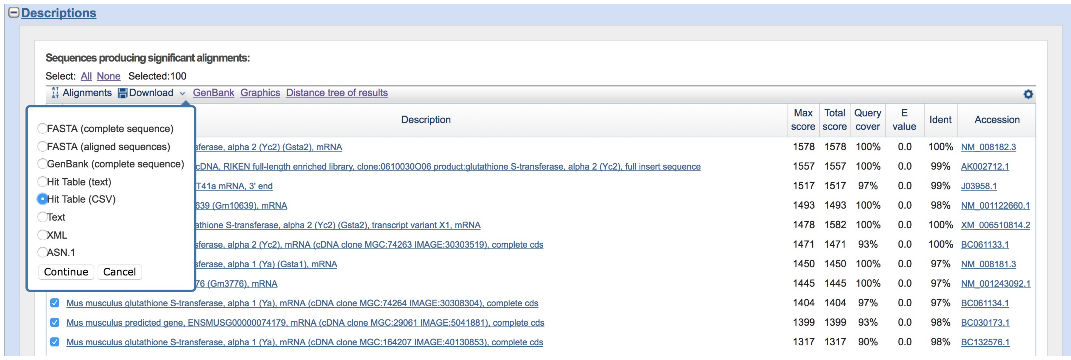
Results for:
✓ 1:lcl|Query_225146 Transcript_1(854bp)
2:lcl|Query_225147 Transcript_2(1752bp)
3:lcl|Query_225148 Transcript_3(3753bp)
4:lcl|Query_225149 Transcript_5(1878bp)
5:lcl|Query_225150 Transcript_6(2989bp)
6:lcl|Query_225151 Transcript_7(1052bp)
7:lcl|Query_225152 Transcript_8(3954bp)
8:lcl|Query_225153 Transcript_9(2294bp)
9:lcl|Query_225154 Transcript_10(2550bp)
10:lcl|Query_225155 Transcript_11(1496bp)

RID
Query ID
Description
Molecule type
Query Length

Database Name nr
Description Nucleotide collection (nt)
Program BLASTN 2.7.1+ ▷ Citation

Other reports: ce tree of results]

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, make a single table that includes the top hit for each transcript.**



⊟ Descriptions

Sequences producing significant alignments:

Select: All None Selected:100

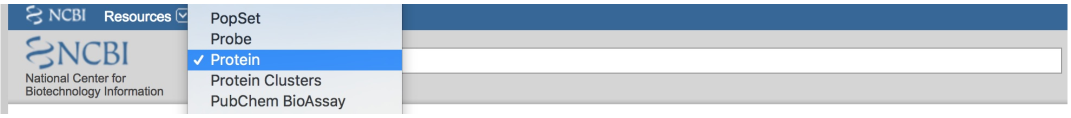
Alignments Download ⌄ GenBank Graphics Distance tree of results

○ FASTA (complete sequence)
○ FASTA (aligned sequences)
○ GenBank (complete sequence)
○ Hit Table (text)
● Hit Table (CSV)
○ Text
○ XML
○ ASN.1
Continue Cancel

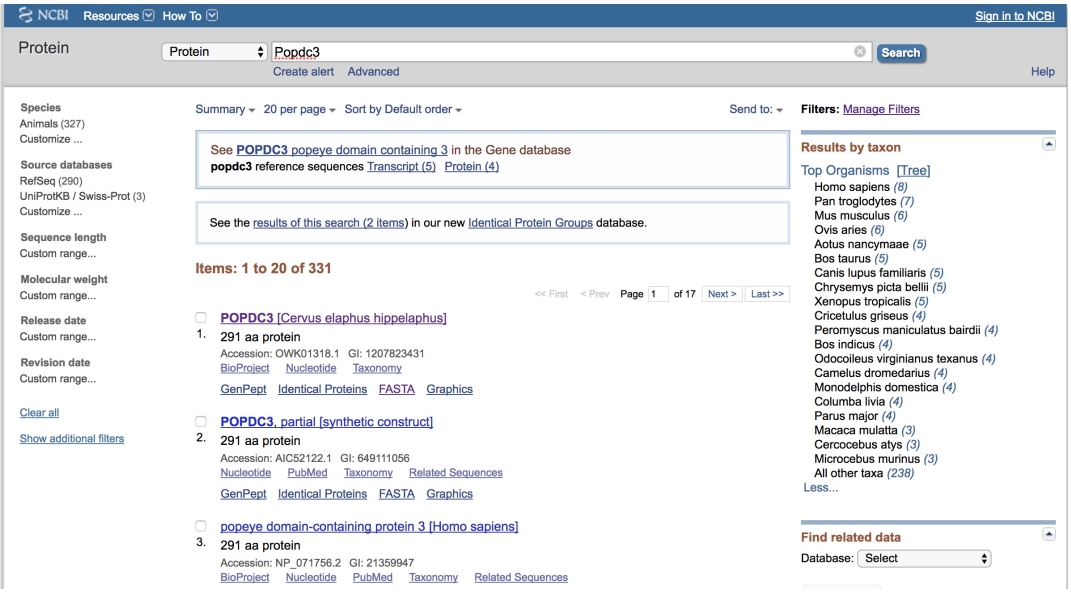
Description	Max score	Total score	Query cover	E value	Ident	Accession
sferase, alpha 2 (Yc2) (Gsta2), mRNA	1578	1578	100%	0.0	100%	NM_008182.3
cDNA, RIKEN full-length enriched library, clone:0610030O06 product:glutathione S-transferase, alpha 2 (Yc2), full insert sequence	1557	1557	100%	0.0	99%	AK002712.1
T41a mRNA, 3' end	1517	1517	97%	0.0	99%	J03958.1
639 (Gm10639), mRNA	1493	1493	100%	0.0	98%	NM_001122660.1
athione S-transferase, alpha 2 (Yc2) (Gsta2), transcript variant X1, mRNA	1478	1582	100%	0.0	100%	XM_006510814.2
sferase, alpha 2 (Yc2), mRNA (cDNA clone MGC:74263 IMAGE:30303519), complete cds	1471	1471	93%	0.0	100%	BC061133.1
sferase, alpha 1 (Ya) (Gsta1), mRNA	1450	1450	100%	0.0	97%	NM_008181.3
76 (Gm3776), mRNA	1445	1445	100%	0.0	97%	NM_001243092.1
☑ Mus musculus glutathione S-transferase, alpha 1 (Ya), mRNA (cDNA clone MGC:74264 IMAGE:30308304), complete cds	1404	1404	97%	0.0	97%	BC061134.1
☑ Mus musculus predicted gene, ENSMUSG00000074179, mRNA (cDNA clone MGC:29061 IMAGE:5041881), complete cds	1399	1399	93%	0.0	98%	BC030173.1
☑ Mus musculus glutathione S-transferase, alpha 1 (Ya), mRNA (cDNA clone MGC:164207 IMAGE:40130853), complete cds	1317	1317	90%	0.0	98%	BC132576.1

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.



Protein: Popdc3

Species
Animals (327)
Customize ...

Source databases
RefSeq (290)
UniProtKB / Swiss-Prot (3)
Customize ...

Sequence length
Custom range...

Molecular weight
Custom range...

Release date
Custom range...

Revision date
Custom range...

See POPDC3 popeye domain containing 3 in the Gene database
popdc3 reference sequences Transcript (5) Protein (4)

See the results of this search (2 items) in our new Identical Protein Groups database.

Items: 1 to 20 of 331

Page 1 of 17

1. □ POPDC3 [Cervus elaphus hippelaphus]
291 aa protein
Accession: OWK01318.1 GI: 1207823431
BioProject Nucleotide Taxonomy
GenPept Identical Proteins FASTA Graphics

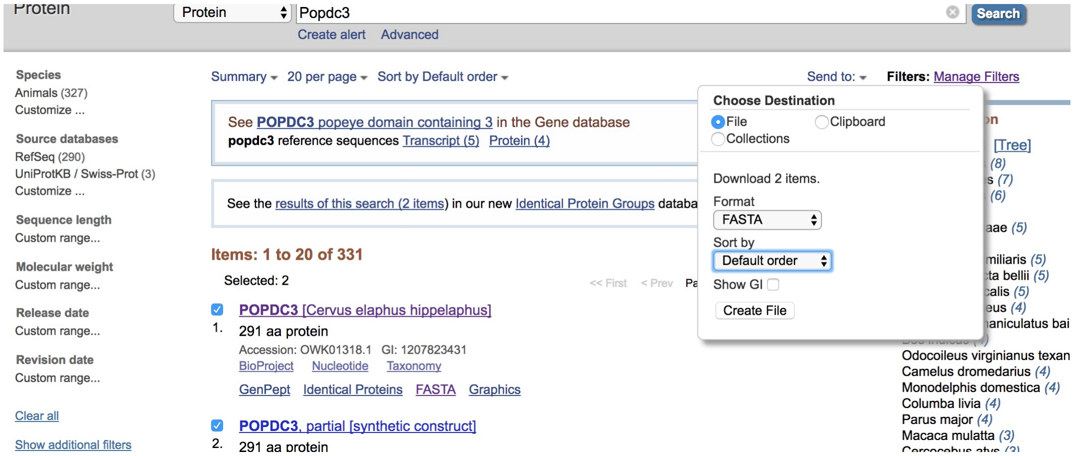
2. □ POPDC3, partial [synthetic construct]
291 aa protein
Accession: AIC52122.1 GI: 649111056
Nucleotide PubMed Taxonomy Related Sequences
GenPept Identical Proteins FASTA Graphics

3. □ popeye domain-containing protein 3 [Homo sapiens]
291 aa protein
Accession: NP_071756.2 GI: 21359947
BioProject Nucleotide PubMed Taxonomy Related Sequences

Results by taxon
Top Organisms [Tree]
Homo sapiens (8)
Pan troglodytes (7)
Mus musculus (6)
Ovis aries (6)
Aotus nancymaae (5)
Bos taurus (5)
Canis lupus familiaris (5)
Chrysemys picta bellii (5)
Xenopus tropicalis (5)
Cricetulus griseus (4)
Peromyscus maniculatus bairdii (4)
Bos indicus (4)
Odocoileus virginianus texanus (4)
Camelus dromedarius (4)
Monodelphis domestica (4)
Columba livia (4)
Parus major (4)
Macaca mulatta (3)
Cercocebus atys (3)
Microcebus murinus (3)
All other taxa (238)

Database: Select

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 10-20 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).**



Protein: Popdc3

Species
Animals (327)
Customize ...

Source databases
RefSeq (290)
UniProtKB / Swiss-Prot (3)
Customize ...

Sequence length
Custom range...

Molecular weight
Custom range...

Release date
Custom range...

Revision date
Custom range...

See POPDC3 popeye domain containing 3 in the Gene database
popdc3 reference sequences Transcript (5) Protein (4)

See the results of this search (2 items) in our new Identical Protein Groups databa

Items: 1 to 20 of 331

Selected: 2

1. ☑ POPDC3 [Cervus elaphus hippelaphus]
291 aa protein
Accession: OWK01318.1 GI: 1207823431
BioProject Nucleotide Taxonomy
GenPept Identical Proteins FASTA Graphics

2. ☑ POPDC3, partial [synthetic construct]
291 aa protein

Choose Destination
● File ○ Clipboard
○ Collections

Download 2 items.
Format: FASTA
Sort by: Default order
Show GI □

Odocoileus virginianus texan
Camelus dromedarius (4)
Monodelphis domestica (4)
Columba livia (4)
Parus major (4)
Macaca mulatta (3)

Translate the 4 provided files of "RNAseq data."

Write a python or R script to translate nucleotide sequences into amino acid sequences. Your script should read the codon map (*codonmap.txt*) and the nucleotide fasta file you want to translate and write a corresponding fasta file containing the translated amino acid sequences. (If you want to try out something new, look up how to pass arguments to a python or R script.) I won't give too many suggestions here, because we want you to decompose the problem as a group. But, here are a few hints for finding the first open reading frame in each sequence. All of the correct start codons are the first in frame, so you could count nucleotides by 3's from the beginning of the sequence or you could use a regular expression. However, beware of greedy regex!! You'll want to stop at the first stop codon in frame (not the last!). **Apply this script to the 4 files of "RNAseq data" (*control1.fasta*, *control2.fasta*, *obese1.fasta* & *obese2.fasta*).**

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.**

Further Exploration

1. For a few (2-3) of your 6 transcripts, return to the original *BLAST* search and change the *Optimize for* option. Qualitatively, 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 a few (2-3) 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.