Alexander Gilson

AS.410.635.81

EXAM 1

Part 1 - 6 points

There are two sequences attached for this part of the exam. The first file is halan.fasta. It is part of a contig from a strain of *Halanaerobium*. The genome is incomplete, and was submitted by a group at the University of Pittsburgh on September 6, 2016. Some CDS regions have been annotated by NCBI Prokaryotic Pipeline Analysis. No mRNAs have been annotated.
The second sequence is hprev_genome.fasta, which is the complete genome of *Halanaerobium praevalens*. This is a related species, isolated from the Great Salt Lake sediment in Utah. Its genome should be suitable for use as a training set in Glimmer.
On the BFX server, run Glimmer. Train it with the *H. praevalens* genome but run Glimmer on the halan.fasta sequence.
Then run FGENESB to compare with the Glimmer result. Submit the following:

1. (1 point) The .predict file with the ORF locations from Glimmer on the BFX server.

```
[agilson2@bfx3 ~]$ less halan.predict
```

```
>Halanaerobium sp. MDAL1, whole genome shotgun sequence
orf00001      171      350  +3    11.68
orf00003      343     1626  +1     8.96
orf00004     1629     4733  +3     6.58
orf00005     5786     4971  -3     8.13
```

2. (1 point) The code you ran to produce the output for number 1 (Glimmer code).

```
[agilson2@bfx3 ~]$ long-orfs -n -t 1.15 hprev_genome.fasta hprev.longorfs
Starting at Wed Mar  3 10:10:38 2021

Sequence file = hprev_genome.fasta
Excluded regions file = none
Circular genome = true
Initial minimum gene length = 90 bp
Determine optimal min gene length to maximize number of genes
Maximum overlap bases = 30
Start codons = atg,gtg,ttg
Stop codons = taa,tag,tga
Sequence length = 2309262
Final minimum gene length = 280
Number of genes = 1811
Total bases = 1822515
[agilson2@bfx3 ~]$
```

```
[agilson2@bfx3 ~]$ extract -t hprev_genome.fasta hprev.longorfs > hprev.train
[agilson2@bfx3 ~]$ build-icm -r hprev.icm < hprev.train
[agilson2@bfx3 ~]$ glimmer3 -o50 -g110 -t30 halan.fasta hprev.icm halan
```

```
Sequence file = halan.fasta
Number of sequences = 1
ICM model file = hprev.icm
Excluded regions file = none
List of orfs file = none
Input is NOT separate orfs
Independent (non-coding) scores are used
Circular genome = true
Truncated orfs = false
Minimum gene length = 110 bp
Maximum overlap bases = 50
Threshold score = 30
Use first start codon = false
Start codons = atg,gtg,ttg
Start probs = 0.600,0.300,0.100
Stop codons = taa,tag,tga
GC percentage = 35.4%
Ignore score on orfs longer than 503
Analyzing Sequence #1
Start Find_Orfs
Start Score_Orfs
Start Process_Events
Start Trace_Back
```

```
[agilson2@bfx3 ~]$ extract -t halan.fasta halan.predict > halan.glimmer
ERROR:  Skipped following coord line
>Halanaerobium sp. MDAL1, whole genome shotgun sequence
```

3. (1 point) The DNA sequence of the first ORF in FASTA format.

```
[agilson2@bfx3 ~]$ less halan.glimmer
```

```
>orf00001  171 350  len=177
ATGGGGGCAGTAATTGAAAGTAATTTAATTTCGGCTCAGAGATTGTTAAGTGATGCAGAA
ACAGATTTAACTGCTGCAAAATATGCCGTGCAGTTAAAAAAGACAGAAGTTTTGGCTGCA
GTAGAAAATATATATAAGAGCTTTACTGCAGGAGTATTAGGAGGTAATAGTAATGAA
```

4. (1 point) Report every predicted CDS of the halan.fasta file, based entirely on the Glimmer result.

CDS 1: 171..350 (orf1)

CDS 2: 343..1626 (orf3)

CDS 3: 1629..4733 (orf4)

CDS 4: 5786..4971 on the reverse strand (orf5)

5. (1 point) Report all possible mRNA molecules based on the FGENESB prediction.

Using “Bacterial generic” as the basis for this prediction

```
 Prediction of potential genes in microbial genomes
 Time:   Tue Jan  1 00:00:00 2005
 Seq name: Halanaerobium sp. MDAL1, whole genome shotgun sequence
 Length of sequence - 6000 bp
 Number of predicted genes - 4
 Number of transcription units - 2, operons - 1
    N      Tu/Op   Conserved  S             Start         End    Score
                   pairs(N/Pv)
    1     1 Op  1     .       +    CDS          3 -       350    298
    2     1 Op  2     .       +    CDS        343 -      1626   1063
    3     1 Op  3     .       +    CDS       1629 -      4733   1901
    4     2 Tu  1     .       -    CDS       4971 -      5786    654
Predicted protein(s):
>GENE     1         3  -       350    298    115 aa, chain +
GIKESEINLDNLESEIRIELSSLLRELELTKLNLETAAKNLKRAKLEYQSTKNRYQMGAV
IESNLISAQRLLSDAETDLTAAKYAVQLKKTEVLAAVENIYKSFTAGVLGGNSNE
>GENE     2       343  -      1626   1063    427 aa, chain +
MNKKTKMALTILLIIAIGAGALIFIRELKNREPQVAKEEDLGAAVETAEVQKGDFEIIYN
YSGTAEYAGKRKISSQIGGEIINIYVRESQKVEKGDLLARIDDQELKNNLSSAETAVREA
EIALKKAELAKDISRNNLAESKAAIKEAESNYSQWQSDYERDKKLYQKNAIAKAKFEQTK
TQFQKAAAQLERVQATLSSAKKSVEIAGLDVETTVERLKKSRNELENARLKFKDTEIRSP
ISAEIVNEFAEVGEVTAAGQPLFEIAKSDRVEIKIQVGMSDLNQLKIGTKALISSPALEQ
KEFKAVISKIGSTADSKSRTTEVTLKLKENINLKDGAFVSAALIAEGLTDVLIVPEKAIF
NYQAASHVYLIKDGRAVRQKIETTVTNGYQTVVTSFLSEGDQIAVTNLNDLQDKTKVYLS
EQENGDD
>GENE     3      1629  -      4733   1901   1034 aa, chain +
MTLVDFAVEKKYTTIAAVFAVVLLGLAALITLNIQLNPDTEPVVVSIQTQYSGVSASDIA
EQINEPLEEELGSIEGVESISSDAMEGVSLVSVEFDYDKDINTAAVDVQNVVSKIRNELP
QDIEEPQIQKFSKSDRPILTLAVTGPRSDTELRTLADNQLKNRLQLIRGVASVDVYGGKE
REIQINVDRNALAAYNIPISLITKRLDEENINFPGGRLTTNEQEYLLRTVGEYENLEEIK
NLIISSTLQGKIYLKDLAAVKDNFAEIRSKFRVEGQETVALNILKQQDANTVQVVDNAKE
TISELENEYQDLNFKITEDQSEFVKLAINNMASTLFIGIILTIIVIFLFLENWRSTLAVS
ISIPTTFVLTLALMKGFDLSLNTVTMTGLILSIGMLVDNSIVVIENVTRHFEELGKPAFK
AAVEGTNEMILAVIAGTTTSMIVLVPVMFIGGFVQQMFRPLSMTLLFAWTGSVISSFTIV
PLVLSLVLKAEEDKRSLKIFTVFKKIAALFTKLLDSSREYYLKLLEKSLNNRAIVITIAV
VILIVTLSLIPLIGSEMTPVMDSGQSYISIETEAGSSLAKTEEVAKKVEKIAADVPELLI
YSTQLGFEPGASTQATTGANGVQQAFMSLTYEDRNSRKRSIWEIQDGLRSEIAKVPGIKA
YVVEEAGATSVSTTQAPLVIRLSGKDPKILYDFAEGLAEQIKKVPGAVNINLRWALDSPE
YHLKINRERAAELGLSTKEISQQISASVDGMDAKEEFNLAGQDDLNILVKYKDEQMFHKN
DLENLIIVSSEGKSLALRELAEIKVIEGPNLISRENMQYTLDILGFSKDRALSKVNKDIK
AVINQYQLPTGYTAQVTGQQDDMNDALTRLAVALVFSVAFIYLLLVSQFKSLIHPITIMI
SLPLELVGVVAALVLTNTYLSMPAMMGLILLSGIAVNDAIHLIDFVIEAEKGGKETKAAI
LEGARLRFRPILMTTFSTLAGMTPLALELAIGTEQYSPLAKVVMGGLFSSTMLLLIFVPV
VYSLFEDLKRKIYN
>GENE     4      4971  -      5786    654    271 aa, chain -
MKKFELKNGNKMPALGLGTSGLRGKECTQVVKEALELGYRQVDTADMYGNHRAIAEALNE
SDVRREDLFITSKIQSEDLEYRQLKKTASRLLDELDLKYFDLLLIHWPSPEVPVEESLKA
MKELKEAGKAKNIGVSNFTIPLLKKALAAYPDLITVNQVEFHPTLYQKELLDFAFKNDII
LTAYAPLAQGEVFENSVLKSLGEKYDKSPAQLALRWLVEKNIAVIPKASSKAHLKNNLEI
FDWDFPIDAAREMELLDQNNRLIDPGYPNFD
```

The run of the halan.fasta file on FGENESB shows 4 total predicted mRNA locations in the genome. These occur at 3..350, 343..1626, 1629..4733, & 5786..4971.

6. (1 point) Point out every location where FGENESB and Glimmer differ in CDS prediction.

There are two noticeable differences that can be seen from the Glimmer predictions and those produced by FGENESB.

First off, the location of the beginning of orf1 in Glimmer, is 171, whereas the location in FGENESB is 3, but both end on location 350. Due to this, the first transcript predicted by FGENESB is 298nt long vs. the 177nt length of the Glimmer prediction.

The list for FGENESB for the last predicted CDS lists the locations as 4971..5786, and the Glimmer prediction shows 5786..4971, so a discrepancy can be seen in how the two prediction programs handle those sequences that can be found on the - strand.

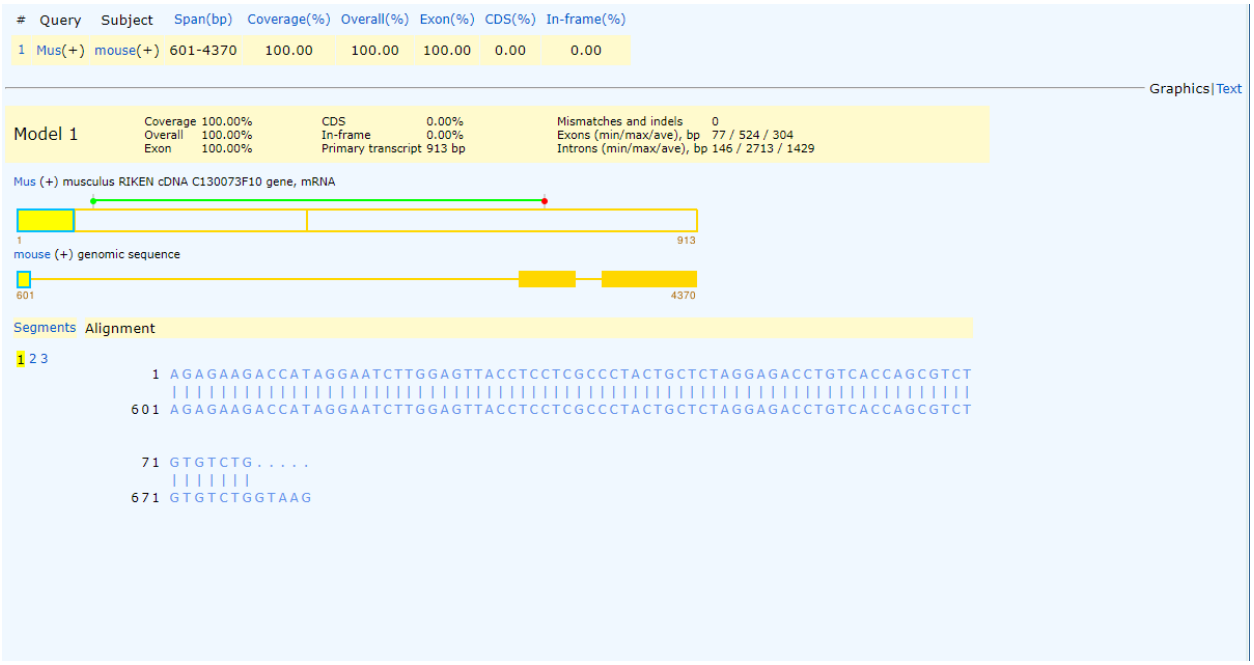
Part 2 - 4 points

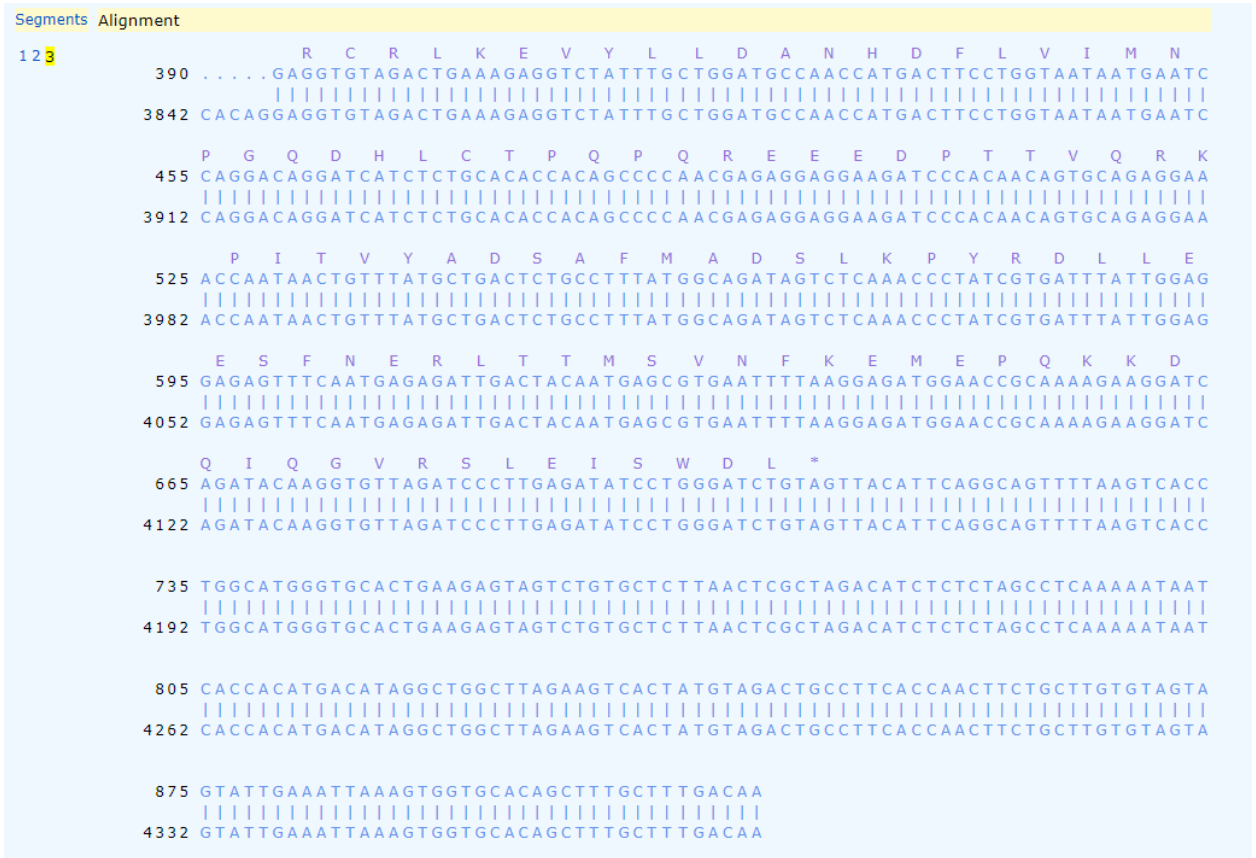
Use the two attached sequences: mouse_genomic.txt and mouse_cdna.txt from the organism *Mus musculus*. The cDNA is an alternately spliced transcript that was verified by NCBI on September 13, 2016 based on RNA-seq data. Run Splign to get the mRNA coordinates and the cDNA coordinates of the genomic sequence:

1. (4 points) Report the coordinates of the mRNA & the CDS locations based on the provided genomic sequence.

mRNA coordinates: **104..709 (**length: 605)

CDS coordinates: **3416..4166** (length: 750)





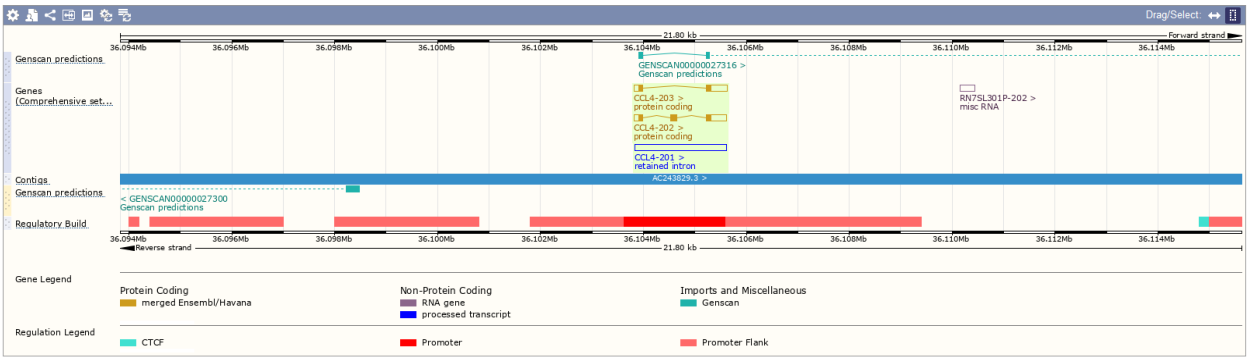
Part 3 - 6 points

The human CCL4 gene has three exons. In 1999, the gene was called the Act-2 cytokine gene. In a paper in the *Journal of Biological Chemistry* from that year, Monica Napolitano *et al.* reported some gene features. There is a TATA box upstream from the transcription start site. The transcription start site is on

chromosome 17 at position 36,103,827. There is also a palindromic sequence in exon 2. The authors reported an unknown significance for the palindrome.
Submit a report that includes the following:
1. (1 points) NCBI reports one transcript with three exons on its gene page for human CCL4. Ensembl lists three transcripts for the same gene. Explain in a sentence or two how the other two transcripts differ from the three-exon transcript. Include protein lengths in your description.

- NM_002984.4(ENST00000615863.2) is the three exon transcript of the CCL4 gene, with a transcript length of 660nt and a protein length of 92aa, with two other transcripts listed by ensembl. The first(ENST00000621626.1) is listed as protein coding with a protein length of 52aa being much shorter than the first transcript, and the other(ENST00000613947.1) is listed as not protein coding with a much longer length of base pairs at 1774 due to having a retained intron from alternative splicing in the CCL4 gene.

2. (1 points) Include an image of the Gene Summary image with the Genscan prediction track turned on. The CCL4-201 transcript should be visible.



3. (1 points) In Ensembl, find and report the amino acid location range for the Chemokine interleukin-8-like domain according to the Pfam database for CCL4-202 transcript.

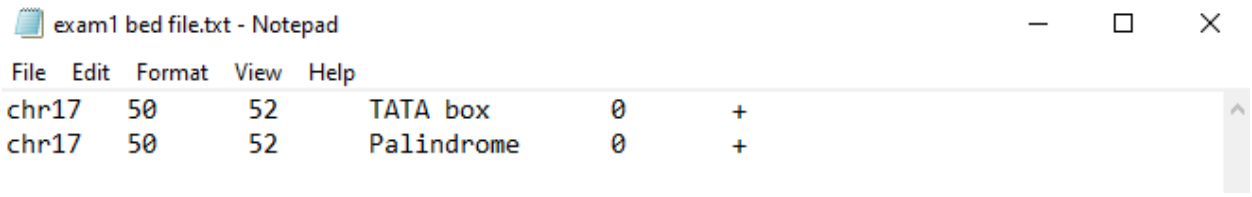
Pfam	33	86	IL8	PF00048	IPR001811 [Display all genes with this domain]

- The Pfam database showed the AA range of the IL8 domain to be 33..86.

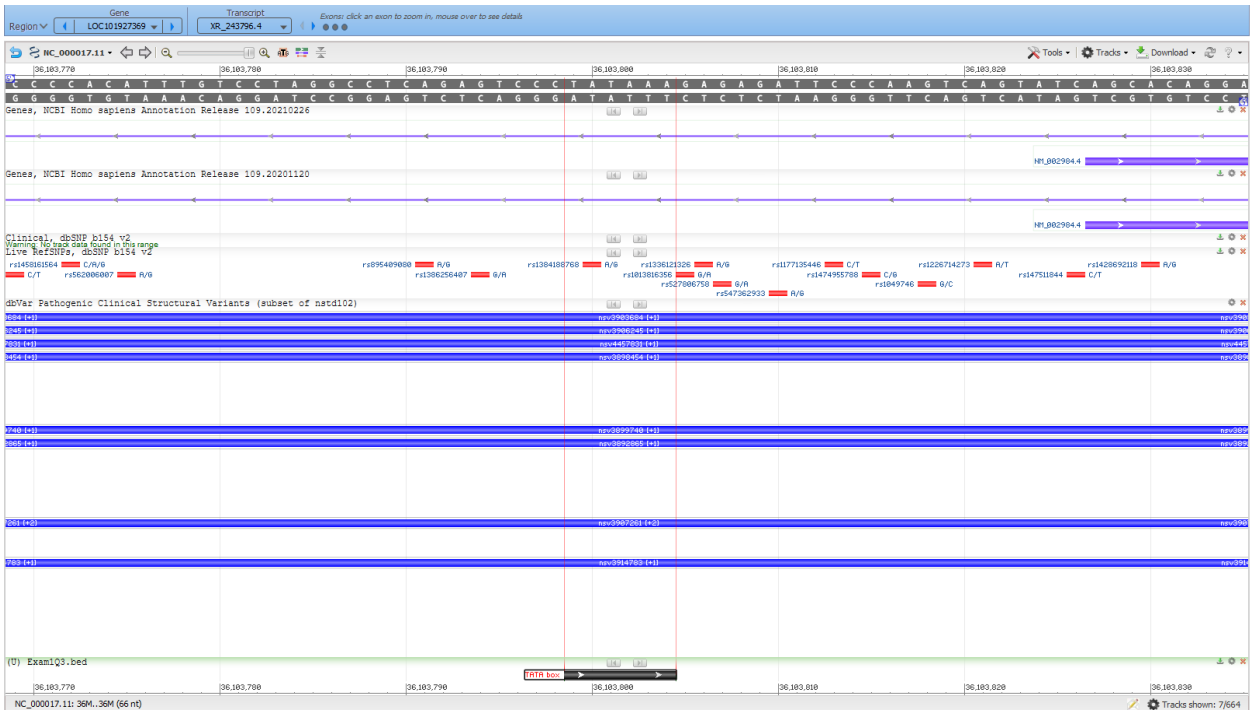
4. (1 points) Create and submit a two-line 0-indexed BED file that includes the following locations:

a. The TATA box, which begins 28 bases upstream from the transcription start site and ends 23 bases upstream from that same site. **36103827-28 = 36103799 36103827-23 = 36103804**

b. The DNA location of the palindromic sequence, which runs from amino acid 40 through amino acid 45 in the protein. **36104569(8 with zero-base)..36104586**



5. (1 points) Display the BED file at the NCBI Variation Viewer for the CCL4 gene. Submit a screenshot that shows both the TATA box (labeled) and the transcription start site. It should be zoomed in far enough that I can see the nucleotides in the sequence.



6. (1 points) Submit another screenshot that shows the palindromic sequence (labeled) in exon 2. Also zoomed in far enough for me to see the nucleotides in the sequence.

Part 4 - 8 points

Search OMIM.org for "huntington's disease". The first five entries all have this or a similar phrase in the title. Record the five identifiers (six-digit numbers) of those five records. The corresponding biomaRt filter name for these identifiers is "mim_morbid_accession". Use biomaRt to retrieve two tables with the following attributes, limiting to the five MIM values you found:


```
> library(bioMart)
> exam = useMart('ensembl')
> exam = useDataset('hsapiens_gene_ensembl',mart=ensembl)
```

First table (2 points)= Entrez Gene ID, HGNC symbol, Ensembl Gene ID

```
> exam1 = exam
> exam1 = getBM(attributes=c('entrezgene_id', 'hgnc_symbol', 'ensembl_gene_id'),
filters='mim_morbid_accession', values = c('603218', '604802','143100', '606438', '607136'), mart =
ensembl)
> exam1
```

	entrezgene_id	hgnc_symbol	ensembl_gene_id
1	3064	HTT	ENSG00000197386
2	5621	PRNP	ENSG00000171867
3	57338	JPH3	ENSG00000154118
4	6908	TBP	ENSG00000112592

****an entry was not found from my search for 604802 Huntingon Disease-Like 3; HDL3. This may be partially due to the fact that that entry on OMIM lacks identifiers beyond it's OMIM ID.****

Second table (2 points)= HGNC symbol, Ensembl Gene ID, Ensembl Transcript ID

```
> exam2 = exam
> exam2 = getBM(attributes = c('hgnc_symbol', 'ensembl_gene_id', 'ensembl_transcript_id'),
filters='mim_morbid_accession', values = c('603218', '604802','143100', '606438', '607136'), mart =
ensembl)
> exam2
```

	hgnc_symbol	ensembl_gene_id	ensembl_transcript_id
1	HTT	ENSG00000197386	ENST00000680239
2	HTT	ENSG00000197386	ENST00000680956
3	HTT	ENSG00000197386	ENST00000680360
4	HTT	ENSG00000197386	ENST00000681528
5	HTT	ENSG00000197386	ENST00000647962
6	HTT	ENSG00000197386	ENST00000649900
7	HTT	ENSG00000197386	ENST00000680291
8	HTT	ENSG00000197386	ENST00000355072
9	HTT	ENSG00000197386	ENST00000648150
10	HTT	ENSG00000197386	ENST00000506137
11	HTT	ENSG00000197386	ENST00000512909
12	HTT	ENSG00000197386	ENST00000510626
13	HTT	ENSG00000197386	ENST00000649131
14	HTT	ENSG00000197386	ENST00000509618
15	HTT	ENSG00000197386	ENST00000650588
16	HTT	ENSG00000197386	ENST00000650595
17	HTT	ENSG00000197386	ENST00000513639
18	HTT	ENSG00000197386	ENST00000513326
19	HTT	ENSG00000197386	ENST00000509043
20	HTT	ENSG00000197386	ENST00000509751
21	HTT	ENSG00000197386	ENST00000512068
22	HTT	ENSG00000197386	ENST00000513806
23	HTT	ENSG00000197386	ENST00000508321
24	PRNP	ENSG00000171867	ENST00000430350
25	PRNP	ENSG00000171867	ENST00000379440
26	PRNP	ENSG00000171867	ENST00000424424
27	PRNP	ENSG00000171867	ENST00000457586
28	JPH3	ENSG00000154118	ENST00000537256
29	JPH3	ENSG00000154118	ENST00000301008
30	JPH3	ENSG00000154118	ENST00000284262
31	JPH3	ENSG00000154118	ENST00000563609
32	TBP	ENSG00000112592	ENST00000421512
33	TBP	ENSG00000112592	ENST00000230354
34	TBP	ENSG00000112592	ENST00000423353
35	TBP	ENSG00000112592	ENST00000636632
36	TBP	ENSG00000112592	ENST00000446829
37	TBP	ENSG00000112592	ENST00000392092
38	TBP	ENSG00000112592	ENST00000540980
39	TBP	ENSG00000112592	ENST00000616883

Submit both tables, either as a spreadsheet or as a screenshot (2 points for each table). Also submit your R code (1 point for each table). For the final two points, explain why there are different numbers of lines in the two tables.

- The second table contains more lines because it lists every transcript that is associated with the ensembl gene id in addition to listing the 4 genes that were found in my search. HTT is shown to have the most transcripts of the genes listed, with PRNP and JPH3 having the least in the second table. This method works well for finding transcript IDs quickly from a simple OMIM search, while the first table works to give more information about the OMIM entries themselves in its findings.