Julie Garcia
Exam #1
February 10, 2018

Part 1 - 8 points

I used Glimmer to predict the ORF and extract sequences from the *Halanaerobium* species sequence file halan.fasta. The species *Halanaerobium pravalens* was used as a reference organism with the file hprev_genome.fasta.

1. Below is a screenshot of the .predict file created by running the glimmer commands. I also attached the file 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
halan.predict (END)
```

2. The glimmer code to produce the output above (halan.predict):

```
long-orfs hprev_genome.fasta hprev.longorfs
extract -t hprev_genome.fasta hprev.longorfs > hprev.train
build-icm -r hprev.icm < hprev.train
glimmer3 -o50 -g110 -t30 halan.fasta hprev.icm halan
```

And the code to extract the sequences:

```
extract halan.fasta halan.predict > halan.glimmer
```

3. The DNA sequence of the first ORF in FASTA format (I also attached halan.glimmer with all of the ORF sequences):

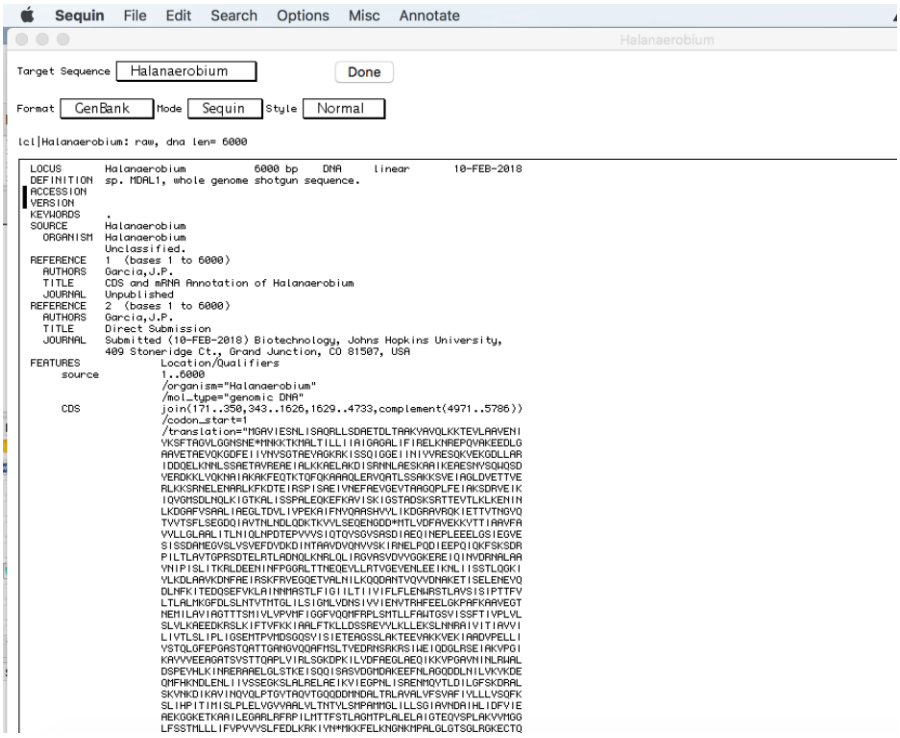
>orf00001 171 350 len=180

ATGGGGGCAGTAATTGAAAGTAATTTAATTTCGGCTCAGAGATTGTTAAGTGATGCAGAA

ACAGATTTAACTGCTGCAAAATATGCCGTGCAGTTAAAAAAGACAGAAGTTTTGGCTGCA

GTAGAAAATATATATAAGAGCTTTACTGCAGGAGTATTAGGAGGTAATAGTAATGAATAA

4. Below is a screenshot of only the CDS annotation in Sequin (I also attached the generated GeneBank file with both CDS and mRNA annotated):



FGENESB did not contain *Halanaerobium prevalens* to use as a reference organism, so I looked up *Halanaerobium* in the NCBI Taxonomy Browser to find the closest similar organism to use in FGENESB as a comparison. I decided to use *Halobacterium* as my comparison organism. I also tried using Bacteria Generic and got the same result. Below is the result of the FGENESB prediction, which I also attached in a text file.

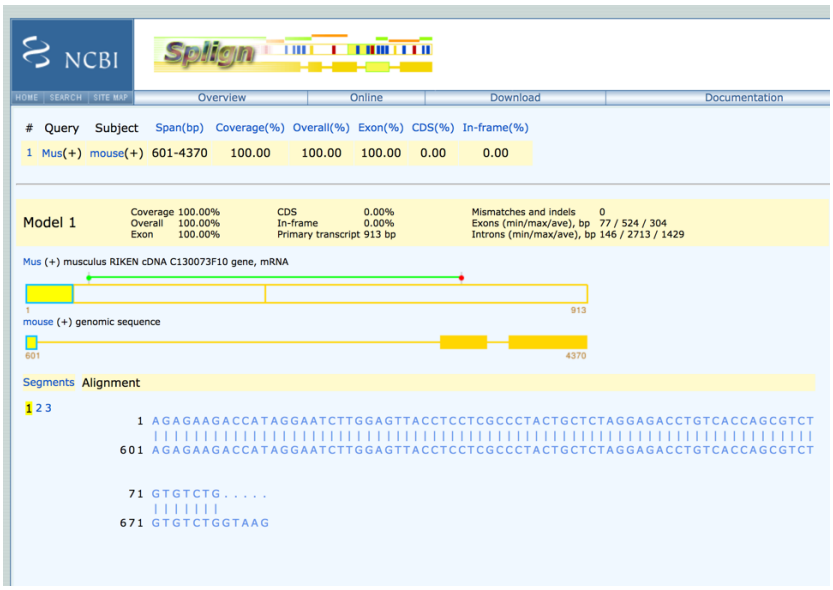
5. Below is a screenshot of the final Sequin file that includes CDS and mRNA annotations.

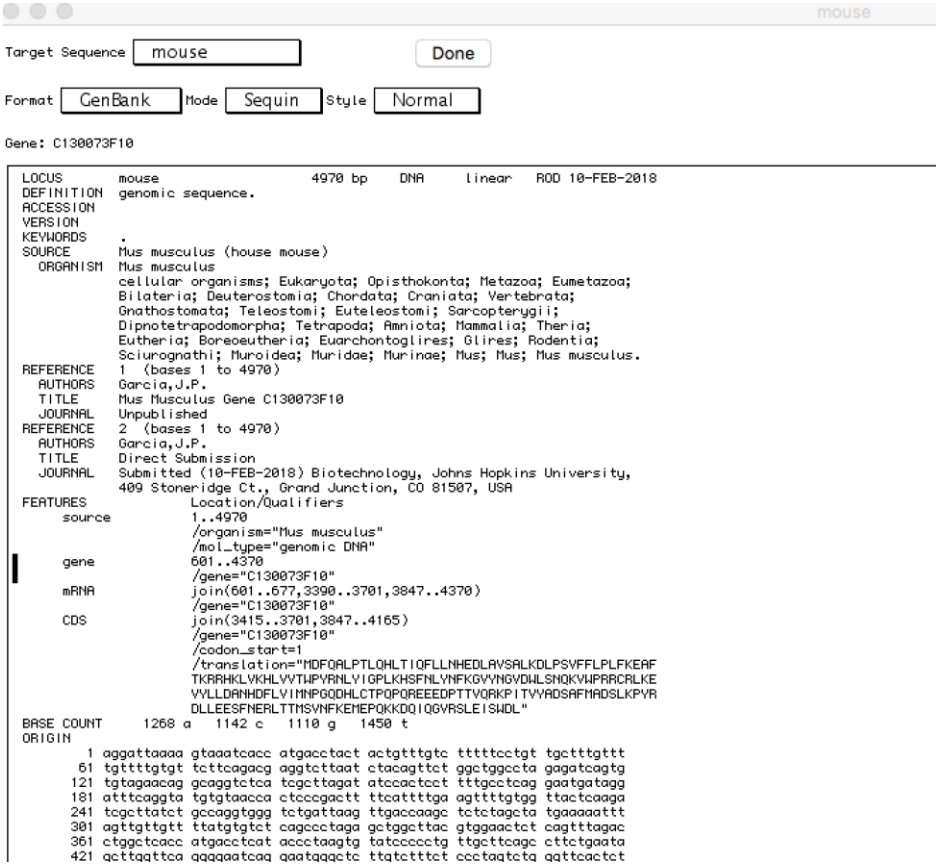
```
            409 Stoneridge Ct., Grand Junction, CO 81507, USA
FEATURES             Location/Qualifiers
     source          1..6000
                     /organism="Halanaerobium"
                     /mol_type="genomic DNA"
     mRNA            join(<3..350,343..1626,1629..4733)
     CDS             join(171..350,343..1626,1629..4733,complement(4971..5786))
                     /codon_start=1
                     /translation="MGAVIESNLISAQRLLSDAETDLTAAKVAVQLKKTEVLAAVENI
                     VKSFTAGVLGGNSNE*MNKKTKMALTILLIIAIGAGALIFIRELKNREPQVAKEEDLG
                     AAVETAEVQKGDFEIIYNYSGTAEYAGKRKISSQIGGEIINIYVRESQKVEKGDLLAR
                     IDDQELKNNLSSAETAVREAEIALKKAELAKDISRNNLAESKAAIKEAESNYSQWQSD
                     VERDKKLYQKNAIAKAKFEQTKTQFQKAAAQLERVQATLSSAKKSVEIAGLDVETTVE
                     RLKKSRNELENARLKFKDTEIRSPISAEIVNEFAEVGEVTAAGQPLFEIAKSDRVEIK
                     IQVGMSDLNQLKIGTKALISSPALEQKEFKAVISKIGSTADSKSRTTEVTLKLKENIN
                     LKDGAFVSAALIAEGLTDVLIVPEKAIFNYQAASHVYLIKDGRAVRQKIETTVTNGYQ
                     TVVTSFLSEGDQIAVTNLNDLQDKTKVYLSEQENGDD*MTLVDFAVEKKVTTIAAVFA
                     VVLLGLAALITLNIQLNPDTEPVVVSIQTQVSGVSASDIAEQINEPLEEELGSIEGVE
                     SISSDAMEGVSLVSVEFDYDKDINTAAVDVQNVVSKIRNELPQDIEEPQIQKFSKSDR
                     PILTLAVTGPRSDTELRTLADNQLKNRLQLIRGVASVDVYGGKEREIQINVDRNALAA
                     YNIPISLITKRLDEENINFPGGRLTTNEQEYLLRTVGEYENLEEIKNLIISSTLQGKI
                     YLKDLAAVKDNFAEIRSKFRVEGQETVALNILKQQDANTVQVVDNAKETISELENEYQ
                     DLNFKITEDQSEFVKLAINNMASTLFIGIILTIIVIFLFLENWRSTLAVSISIPTTFV
                     LTLALMKGFDLSLNTVTMTGLILSIGMLVDNSIVVIENVTRHFEELGKPAFKAAVEGT
                     NEMILAVIAGTTTSMIVLVPVMFIGGFVQQMFRPLSMTLLFAWTGSVISSFTIVPLVL
                     SLVLKAEEDKRSLKIFTVFKKIAALFTKLLDSSREVYLKLLEKSLNNRAIVITIAVVI
                     LIVTLSLIPLIGSEMTPVMDSGQSVISIETEAGSSLAKTEEVAKKVEKIAADVPELLI
                     YSTQLGFEPGASTQATTGANGVQQAFMSLTYEDRNSRKRSIWEIQDGLRSEIAKVPGI
                     KAVVVEEAGATSVSTTQAPLVIRLSGKDPKILYDFAEGLAEQIKKVPGAVNINLRWAL
                     DSPEYHLKINRERAAELGLSTKEISQQISASVDGMDAKEEFNLAGQDDLNILVKYKDE
                     QMFHKNDLENLIIVSSEGKSLALRELAEIKVIEGPNLISRENMQYTLDILGFSKDRAL
                     SKVNKDIKAVINQYQLPTGYTAQVTGQQDDMNDALTRLAVALVFSVAFIYLLLVSQFK
                     SLIHPITIMISLPLELVGVVAALVLTNTYLSMPAMMGLILLSGIAVNDAIHLIDFVIE
                     AEKGGKETKAAILEGARLRFRPILMTTFSTLAGMTPLALELAIGTEQVSPLAKVVMGG
                     LFSSTMLLLIFVPVVYSLFEDLKRKIYN*MKKFELKNGNKMPALGLGTSGLRGKECTQ
                     VVKEALELGVRQVDTADMYGNHRAIAEALNESDVRREDLFITSKIQSEDLEVRQLKKT
                     ASRLLDELDLKVFDLLLIHWPSPEVPVEESLKAMKELKEAGKAKNIGVSNFTIPLLKK
                     ALAAYPDLITVNQVEFHPTLYQKELLDFAFKNDIILTAYAPLAQGEVFENSVLKSLGE
                     KYDKSPAQLALRWLVEKNIAVIPKASSKAHLKNNLEIFDWDFPIDAARENELLDQNNR
                     LIDPGYPNFD"
     mRNA            complement(4971..5786)
BASE COUNT     2058 a    910 c   1213 g   1819 t
ORIGIN
```

6. In the screenshots above you can see that FGENESB and Glimmer predicted almost the same coding regions, except for one spot where FGENESB predict the first CDS region to start at 3 and Glimmer predicted it to start at 171.

Part 2 - 4 points

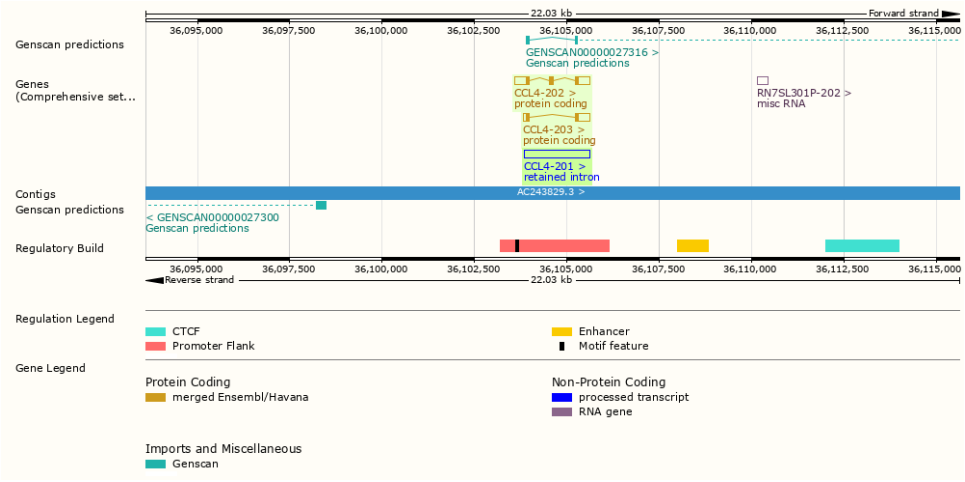
1 & 2. I ran Splign to align the mouse cDNA sequence to the genomic sequence provided. The results of Splign are shown here. One CDS was found with 3 exons. The annotated sequence in GeneBank is also shown below and the GeneBank file is attached.

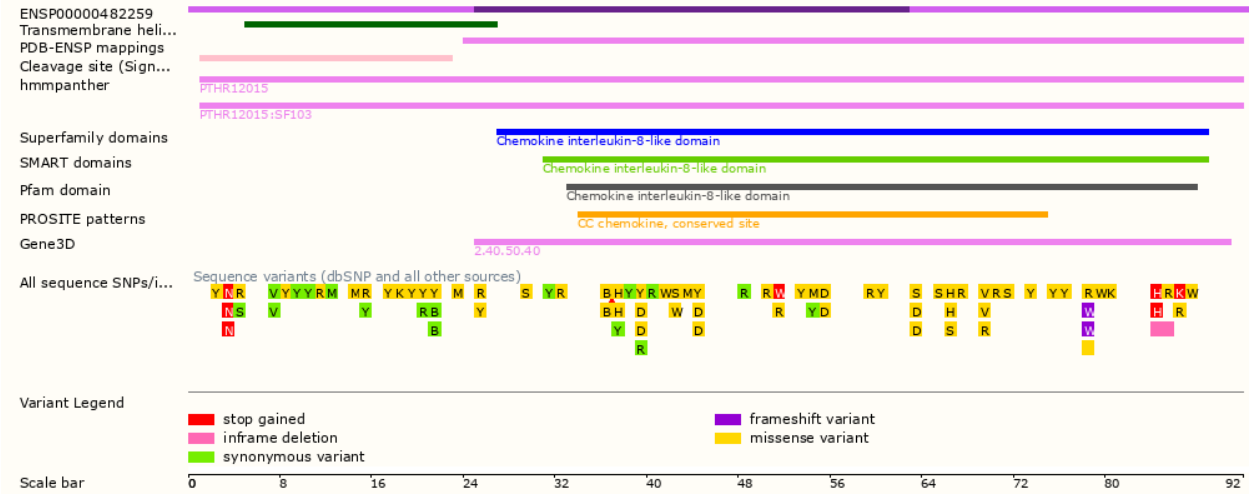





```
LOCUS       mouse                   4970 bp    DNA     linear   ROD 10-FEB-2018
DEFINITION  genomic sequence.
ACCESSION
VERSION
KEYWORDS    .
SOURCE      Mus musculus (house mouse)
  ORGANISM  Mus musculus
            cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa;
            Bilateria; Deuterostomia; Chordata; Craniata; Vertebrata;
            Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii;
            Dipnotetrapodomorpha; Tetrapoda; Amniota; Mammalia; Theria;
            Eutheria; Boreoeutheria; Euarchontoglires; Glires; Rodentia;
            Sciurognathi; Muroidea; Muridae; Murinae; Mus; Mus; Mus musculus.
REFERENCE   1  (bases 1 to 4970)
  AUTHORS   Garcia,J.P.
  TITLE     Mus Musculus Gene C130073F10
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 4970)
  AUTHORS   Garcia,J.P.
  TITLE     Direct Submission
  JOURNAL   Submitted (10-FEB-2018) Biotechnology, Johns Hopkins University,
            409 Stoneridge Ct., Grand Junction, CO 81507, USA
FEATURES             Location/Qualifiers
     source          1..4970
                     /organism="Mus musculus"
                     /mol_type="genomic DNA"
     gene            601..4370
                     /gene="C130073F10"
     mRNA            join(601..677,3390..3701,3847..4370)
                     /gene="C130073F10"
     CDS             join(3415..3701,3847..4165)
                     /gene="C130073F10"
                     /codon_start=1
                     /translation="MDFQALPTLQHLTIQFLLNHEDLAVSALKDLPSVFFLPLFKEAF
                     TKRRHKLVKHLVVTWPVRNLVIGPLKHSFNLVNFKGVVNGVDWLSNQKVWPRRCRLKE
                     VVLLDANHDFLVIMNPGQDHLCTPQPQREEEDPTTVQRKPITVVRADSAFMADSLKPVR
                     DLLEESFNERLTTMSVNFKEMEPQKKDQIQGVRSLEISWDL"
BASE COUNT     1268 a   1142 c   1110 g   1450 t
ORIGIN
        1 aggattaaaa gtaaatcacc atgacctact actgtttgtc tttttcctgt tgctttgttt
       61 tgttttgtgt tcttcagacg aggtcttaat ctacagttct ggctggccta gagatcagtg
      121 tgtagaacag gcaggtctca tcgcttagat atccactcct tttgcctcag gaatgatagg
      181 atttcaggta tgtgtaacca ctcccgactt ttcattttga agttttgtgg ttactcaaga
      241 tcgcttatct gccaggtggg tctgattaag ttgaccaagc tctctagcta tgaaaaattt
      301 agttgttgtt ttatgtgtct cagccctaga gctggcttac gtggaactct cagtttagac
      361 ctggctcacc atgacctcat accctaagtg tatccccctg ttgcttcagc cttctgaata
      421 gcttgattca gaggaatcag gaatgggctc ttgtctttct ccctagtctg gattcactct
```

Part 3 - 12 points

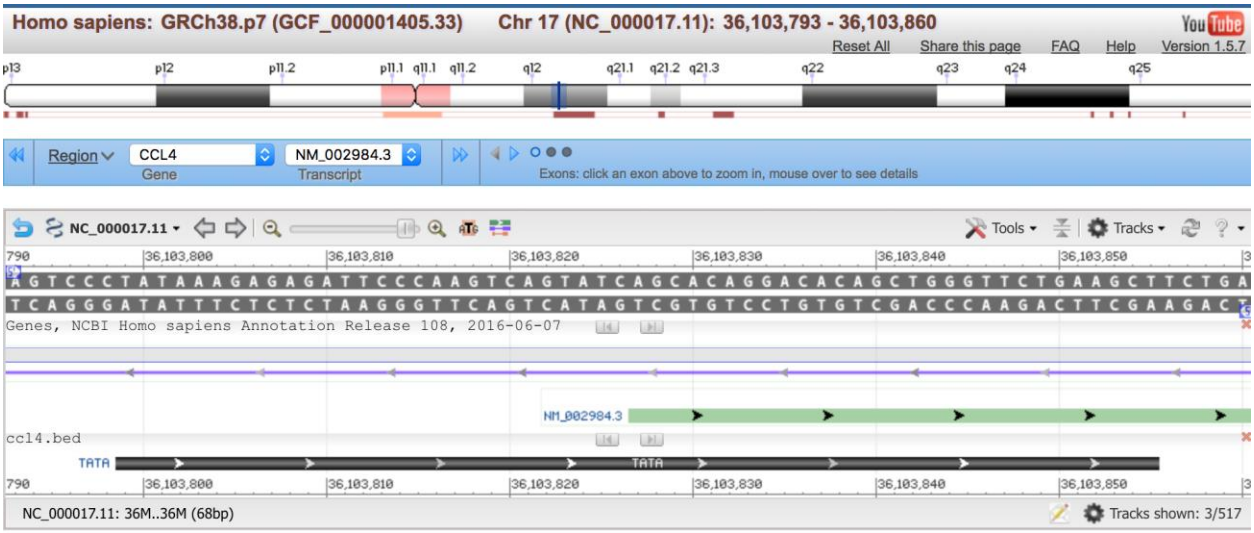
1. In the human CCL4, the transcript described on NCBI gene contains all three exons, the first including the 5' UTR, the second being only a coding sequence, and the third containing the 3' UTR. This primary transcript encodes a protein 92aa long, while the second transcript on Ensembl skips the middle exon, only includes the first and third exons and encodes a shorter protein of 52aa. The third transcript is a retained intron transcript and does not encode a protein.
2. 

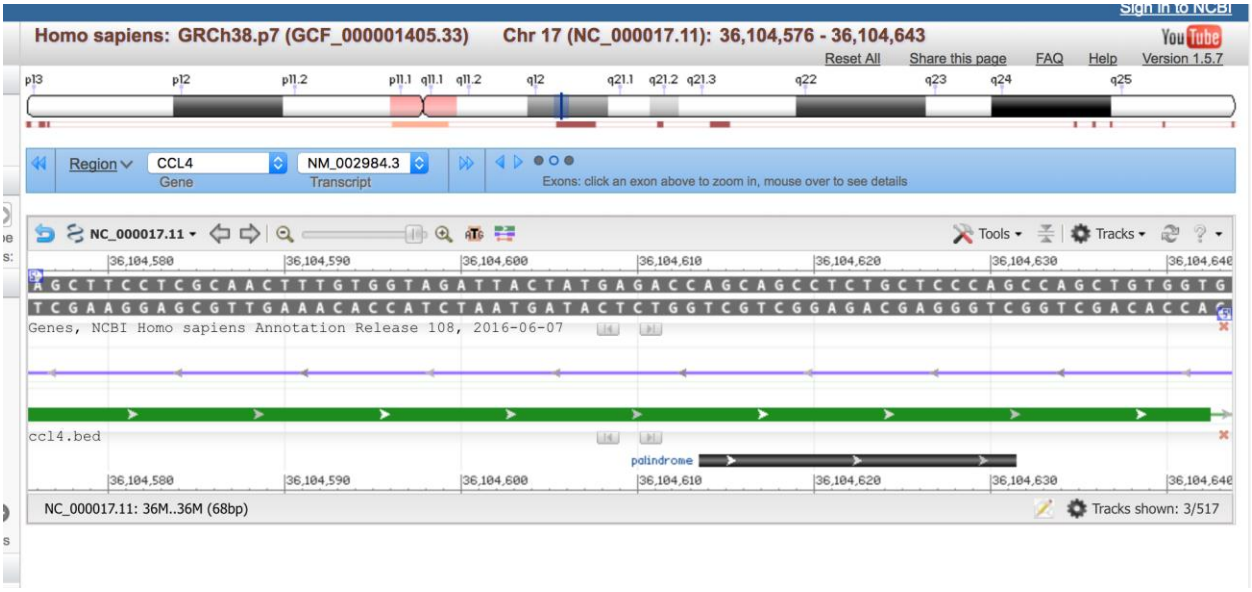
3. The gene encodes a protein that contains a Chemokine interleukin-8-like domain at position 33-88aa as can be seen in the image from Ensembl below. It is found in the primary transcript of CCL4-202, but not in either of the alternative transcripts CCL4-201 and CCL4-203.


4. I created a BED file containing the TATA box and the DNA location of the palindromic sequence. Below is a screenshot and I attached the file to my submission. Below are two screenshots of the BED file loaded into Variation Viewer, one with the TATA Box location and one with the palindrome location.

ccl4

```
chr17   36103798        36103855        TATA    0       +
chr17   36104613        36104631        palindrome      0       +
```





Part 4 - 10 points

Search OMIM.org for "huntington's disease". The first five entires 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". Use biomaRt to retrieve two tables with the following attributes, limiting to the five MIM values you found:

```
huntingtons <- c("143100", "613004", "604802",  "606483", "603218")
huntingtons_genes <- getBM(attributes=c("entrezgene", "hgnc_symbol", "ensembl_gene_id"),
                filters="mim_morbid_accession", values=huntingtons, mart=ensembl)
huntingtons_genes
```

```
  entrezgene hgnc_symbol ensembl_gene_id
1       9096       TBX18 ENSG00000112837
2       3064         HTT ENSG00000197386
3       5621        PRNP ENSG00000171867
```

```
huntingtons_transcripts <- getBM(attributes=c("hgnc_symbol", "ensembl_gene_id",
"ensembl_transcript_id"),
                filters="mim_morbid_accession", values=huntingtons, mart=ensembl)
huntingtons_transcripts
```

```
  hgnc_symbol ensembl_gene_id ensembl_transcript_id
1       TBX18 ENSG00000112837       ENST00000330469
2       TBX18 ENSG00000112837       ENST00000606784
3       TBX18 ENSG00000112837       ENST00000369663
4       TBX18 ENSG00000112837       ENST00000607343
5       TBX18 ENSG00000112837       ENST00000606521
6       TBX18 ENSG00000112837       ENST00000606325
7       TBX18 ENSG00000112837       ENST00000606621
8         HTT ENSG00000197386       ENST00000355072
9         HTT ENSG00000197386       ENST00000506137
10        HTT ENSG00000197386       ENST00000512909
11        HTT ENSG00000197386       ENST00000510626
12        HTT ENSG00000197386       ENST00000509618
13        HTT ENSG00000197386       ENST00000513639
14        HTT ENSG00000197386       ENST00000513326
15        HTT ENSG00000197386       ENST00000509043
16        HTT ENSG00000197386       ENST00000502820
17        HTT ENSG00000197386       ENST00000509751
18        HTT ENSG00000197386       ENST00000512068
19        HTT ENSG00000197386       ENST00000513806
20        HTT ENSG00000197386       ENST00000508321
21       PRNP ENSG00000171867       ENST00000379440
22       PRNP ENSG00000171867       ENST00000430350
23       PRNP ENSG00000171867       ENST00000424424
24       PRNP ENSG00000171867       ENST00000457586
```

Because we included the Ensembl Transcript Id in the second query, the table lists each gene id and each transcript for each gene. For example, gene ENSG00000112837 has 7 transcripts with the transcript ids shown in the third column.