Jaimee Beckett
Graded Homework 7-8

1. Attached are two bed files: hs_5q31_exons.bed and hs_5q31_cpg.bed. The first file contains exons in chromosomal region 5q31. The second file contains CpG regions in the same chromosomal region. Use Galaxy to answer the following.
   a. How many unique exons intersect a CpG region?
      542

   b. How many unique exons do not intersect CpG regions?
      4,296

   c. How many unique CpG regions intersect an exon?
      200

   d. How many unique CpG regions do not intersect exons?
      45

   e. Submit a Galaxy workflow showing all the steps used to answer parts a-d (HINT: In history panel, click on the cog > "Extract Workflow" > select only the steps that contribute to answering the HW problems > Name workflow something meaningful > "Create workflow". Then click "Workflow" at the top of Galaxy, click on your workflow name > "Share or Download". Submit the .ga file)
      See Galaxy-Workflow-Question_1-_CPG_and_Exon_Overlap_Workflow.ga

2. Use the UCSC Table Browser to retrieve three ENCODE tracks from region chr5:134000000-134250000 of hg38

   DNase HS (hypersensitivity), table HUVEC Pk (output a BED file)

   Layered H3K4me3, HUVEC Cells (output a WIG (data points) file)

   Any other ENCODE track of your choice (output appropriate file type)
   - **Transcription, HUVEC (wgEncodeRegTxnCaltechRNASeqHuvecR2x75l1200SigPooled)**

   Using IGV, load all three tracks to hg38. Set any WIG formatted files to "Autoscale". Rename the tracks so I know which they are. Change up the colors if you feel so inclined.
   a. Submit a screenshot that includes the TCF7 gene and the 8 closest DHS regions.
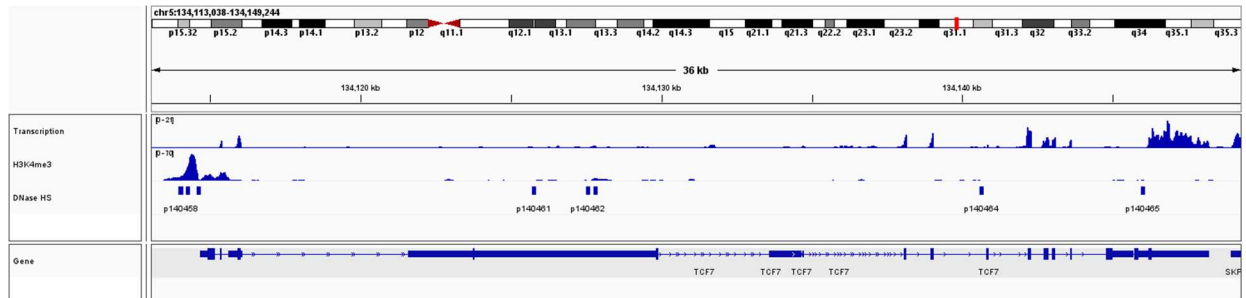
b. For the eight DHS regions that are closest to the TCF7 gene, describe what genomic features they are closest to or overlap (e.g. Located in intron. Located upstream of TSS.)
p140458, located upstream of TSS
p140459, located upstream of TSS
p140460, located upstream of and overlaps TSS
p140461, located in intron
p140462, located in intron
p140463, located in intron
p140464, located in intron
p140465, located in intron

c. Based on the H3K4me3 track, do you think the TCF7 gene is expressed in HUVEC cells? Why or why not?

I think the TCF7 gene is expressed in HUVEC cells because there is a large H3K4me3 signal right before the TSS. H3K4me3 is usually associated with gene activation.

3. Go to YeastMine and click "Templates" in the top menu. Find the template for retrieving genes annotated to a specific GO (gene ontology) ID. Find all genes associated with GO:0006623 (protein targeting to vacuole). Export the results as a tab-separated file and submit the file.

There are 123 results, see attached file: yeastMine_GOTerm_Genes.tsv

4. Using the attached Arabidopsis thaliana protein sequence (tair_plant.fasta):

a. Go to the TAIR website and, using the BLASTP tool, search for the closest match in the TAIR10 database. Click on the best-matching link. What is the ID and description of this protein?

AT5G13630.1 CHLH, Encodes magnesium chelatase involved in plastid-to-nucleus signal transduction

b. In what cellular compartment(s) does this protein function?

Chloroplast, chloroplast inner membrane, mitochondrion, plastid, and magnesium chelatase complex

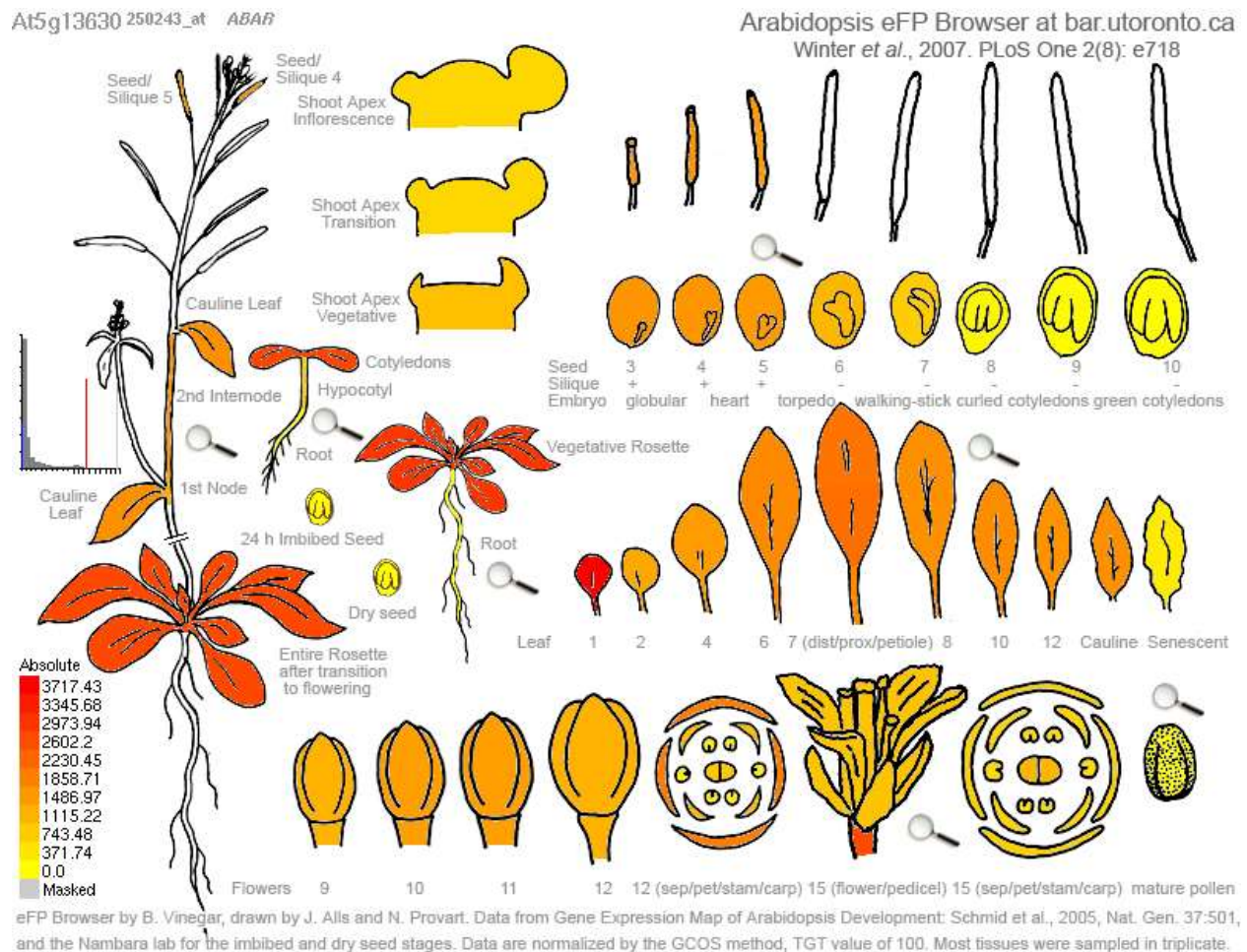c. How many protein-coding gene models exist for this protein?

2: AT5G13630.1 and AT5G13630.2

d. Briefly describe how the protein-coding gene models differ.

The to models are splice variants of each other.

e. Scroll down to BAR eFP browser and check "Developmental Map". And, describe in which of the tissue types shown in the graph is this protein expressed the lowest?

The protein is expressed the lowest in Silique – seeds 8, 9, and 10, in Senescent leaves, 24 hour imbibed seeds, dry seeds, shoot apex inflorescence, shoot apex transition, and mature pollen

Arabidopsis eFP Browser at bar.utoronto.ca
Winter et al., 2007. PLoS One 2(8): e718

eFP Browser by B. Vinegar, drawn by J. Alls and N. Provart. Data from Gene Expression Map of Arabidopsis Development: Schmid et al., 2005, Nat. Gen. 37:501, and the Nambara lab for the imbibed and dry seed stages. Data are normalized by the GCOS method, TGT value of 100. Most tissues were sampled in triplicate.

5. Load the attached hs_chr20* files to the BFX server. The hs_chr20_H3K4me3 BED file represents active H3K4me3 states on chr 20. The hs_chr20_refseq BED file represents coding exons on chr 20. Use bedtools intersect with the H3K4me3 file as the A file and the refseq file as the B file to answer the following:

a. How many unique H3K4me3 regions intersect a coding exon?

   10

b. How many unique H3K4me3 regions do not intersect a coding exon?

   245

c. Submit your command line code.

```
jbecket5@bfx3:~/hs_chr20
[jbecket5@bfx3 hs_chr20]$ bedtools intersect -u -a hs_chr20_H3K4me3.bed -b hs_chr20_refseq.bed
chr20    95834    96109    chr20.28
chr20    96309    96534    chr20.29
chr20    144884   146934   chr20.96
chr20    157484   157709   chr20.114
chr20    158709   158934   chr20.117
chr20    187584   188034   chr20.156
chr20    189259   189684   chr20.159
chr20    226859   227534   chr20.214
chr20    257534   257959   chr20.252
chr20    257984   260084   chr20.253
[jbecket5@bfx3 hs_chr20]$ bedtools intersect -u -a hs_chr20_H3K4me3.bed -b hs_chr20_refseq.bed | wc -l
10
[jbecket5@bfx3 hs_chr20]$ bedtools intersect -v -a hs_chr20_H3K4me3.bed -b hs_chr20_refseq.bed | wc -l
245
[jbecket5@bfx3 hs_chr20]$ _
```