Alex Gilson
AS.410.635.81

Part 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.

(0.25 pts) How many unique exons intersect a CpG region? **542 regions found**.

(0.25 pts) How many unique exons do not intersect CpG regions? **4,296 regions found**

(0.25 pts) How many unique CpG regions intersect an exon? **200 regions found**

(0.25 pts) How many unique CpG regions do not intersect exons? **45 regions found**

(0.5 pts) 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). **File named: "AG Workflow constructed from history 'Unit 7-8 HW'"**

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

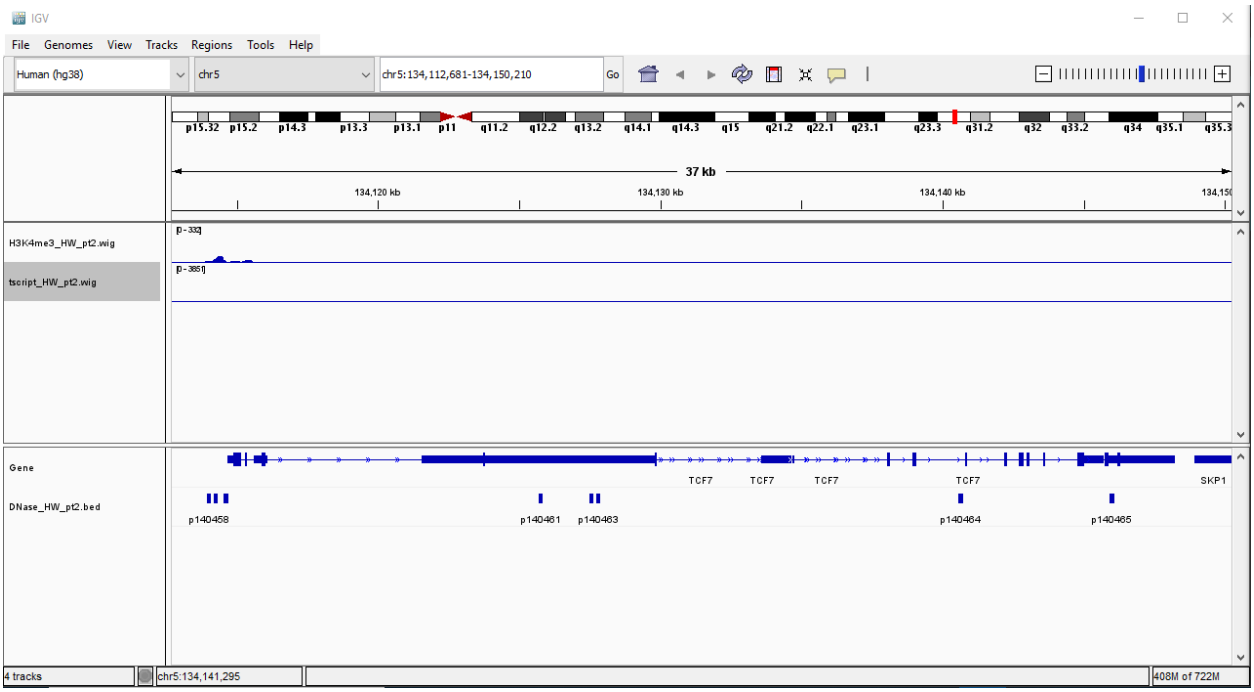
DNase HS (hypersensitivity), table HUVEC Pk (output a BED file)

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

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

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.

(0.25 pts) Submit a screenshot that includes the TCF7 gene and the 8 closest DHS regions.



(0.40 pts) 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.)

Two located upstream of TSS, one within TSS, three within exon 1, one within an intron in the middle of the gene, and the last being in exon 9.

(0.50 pts) Based on the H3K4me3 track, do you think the TCF7 gene is expressed in HUVEC cells? Why or why not?

No, TCF7 is most likely not expressed in HUVEC cells, as the H3K4me3 track is shown to have a large amount of methylation before and at the TSS of the gene, indicating it is transcriptionally inactive.

Part 3

(0.5 pts) 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. **SEE FILE “HW_pt3.tsv”**

(Optional) Duplicate this search in biomaRt using Ensembl Fungi. Submit your R code and your results (as a tab-delimited file). I should be able to run just your submitted R code and get the same results you submit.

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

(0.25 pts) 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? A**T5G13630.1, encodes magnesium chelatase involved in plastid-to-nucleus signal transduction.**

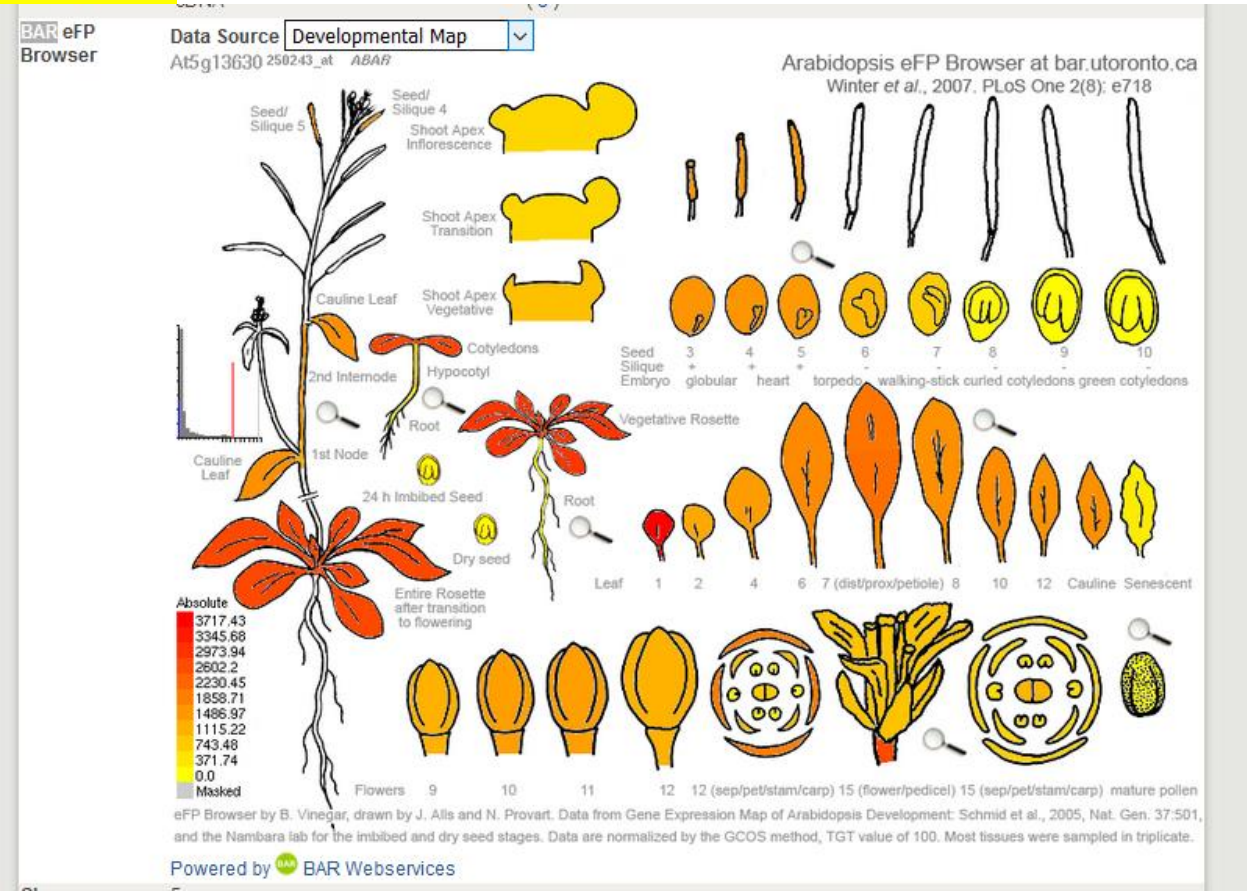
(0.10 pts) In what cellular compartment(s) does this protein function? **Chloroplast & Mitochondria**

(0.10 pts) How many protein-coding gene models exist for this protein? **2, one splice variant: AT5G13630.2**

(0.25 pts) Briefly describe how the protein-coding gene models differ. **Splice variation**

(0.25 pts) 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?

Cauline Leaf:



Part 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:

(0.25 pts) How many unique H3K4me3 regions intersect a coding exon? **10 regions**

```
[agilson2@bfx3 ~]$ 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
```

(0.25 pts) How many unique H3K4me3 regions do not intersect a coding exon? **245 regions**

```
[agilson2@bfx3 ~]$ bedtools intersect -v  -a hs_chr20_H3K4me3.bed -b hs_chr20_refseq.bed | wc -l
245
```

(0.40 pts) Submit your command line code.

[agilson2@bfx3 ~]$ bedtools intersect -u -a hs_chr20_H3K4me3.bed -b hs_chr20_refseq.bed | wc -l

[agilson2@bfx3 ~]$ bedtools intersect -v -a hs_chr20_H3K4me3.bed -b hs_chr20_refseq.bed | wc -l