Julie Garcia
March 1, 2018
Genomics Spring 2018
Unit 8 Graded Homework

Part 1

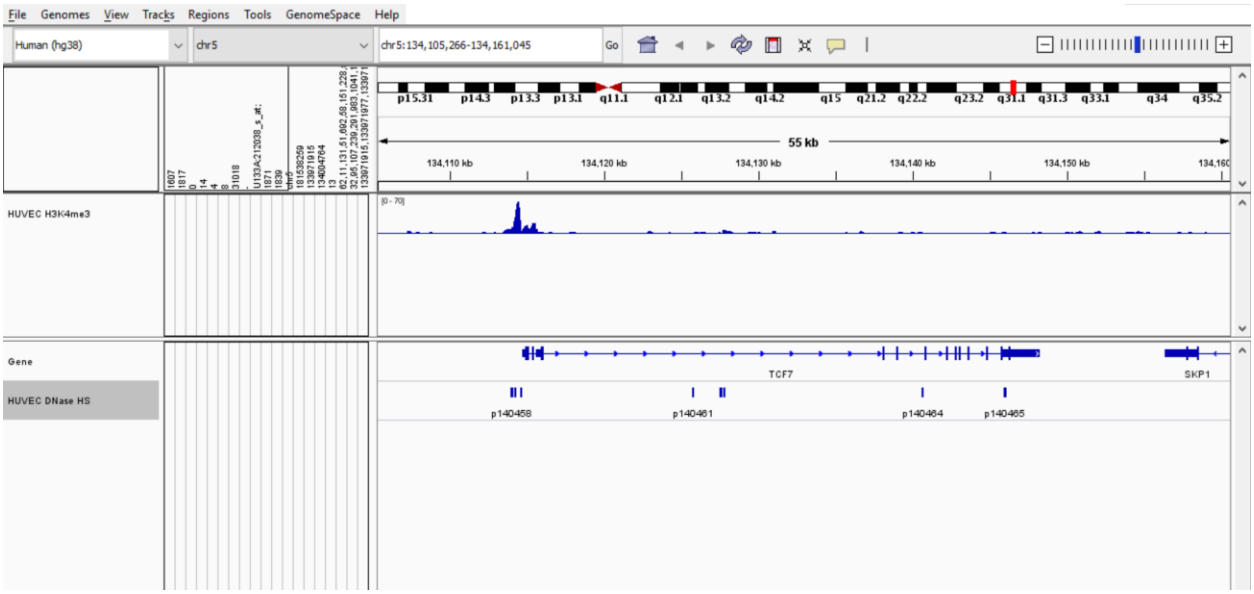
I used Galaxy to upload the two bed files from the region 5q31, one with exons, one with CpG area. Using the BEDtools > Intersect Intervals tool and choosing the exon file as the A dataset, and using the -u option to only report overlapping exons from A once, I found 542 unique exons that intersect with a CpG region. To find unique exons that DO NOT intersect with a CpG region, I ran the same command, however, I used the -v option to only report those in dataset A (exons) that do not intersect. I found 4296 unique exons that DO NOT intersect with CpG sites.

Next, in order to find unique CpG regions that do and do not intersect, I ran the same commands above, but selected the CpG dataset as the A dataset for both. There were 200 unique CpG regions intersecting exons and 45 unique CpG regions that DO NOT intersect with any exons.

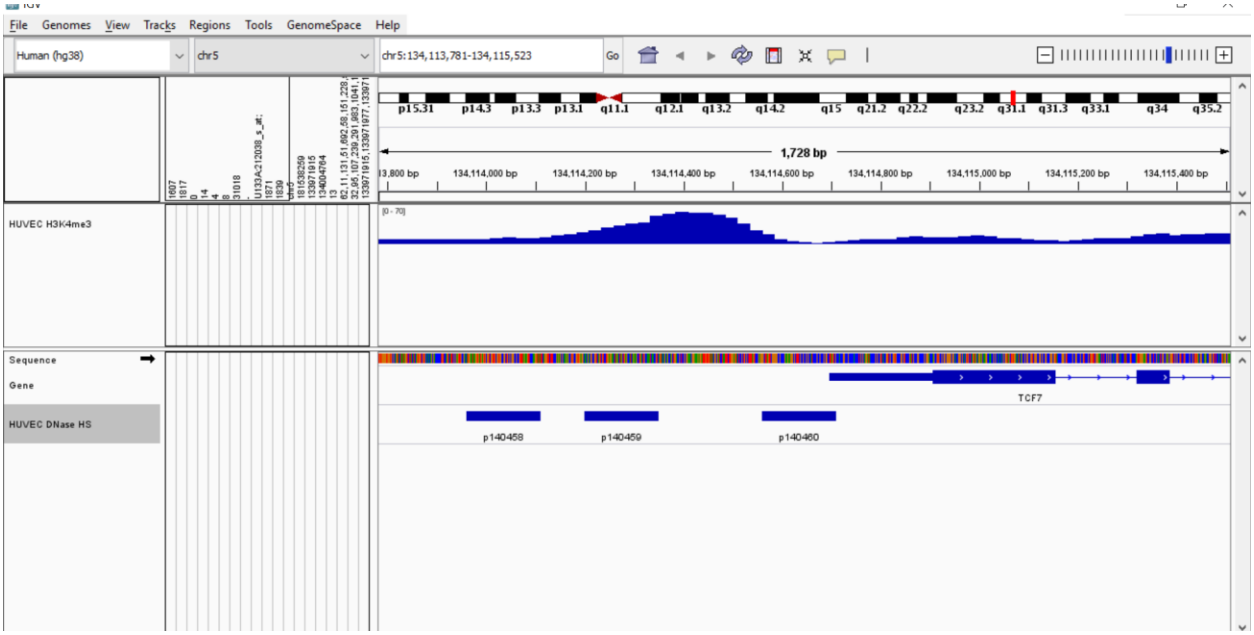
I attached the workflow file named Galaxy-Workflow-HS_5q31_overlapping_exons_and_CpGs.ga.

Part 2

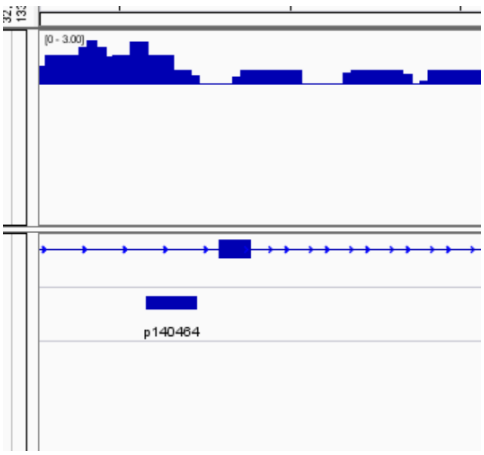
I used UCSC Table browser to output a BED file of DNase HS, Layered H3K4me3 and Gene Expression data (Affy U133) as a table for the region chr5:134000000-134250000 of hg38. I loaded all three tracks into IGV, and set the WIG files to “Autoscale”. I zoomed to the region chr5:134000000-134250000, and found the gene TCF7 shown below with the 8 closes DNase HS regions.



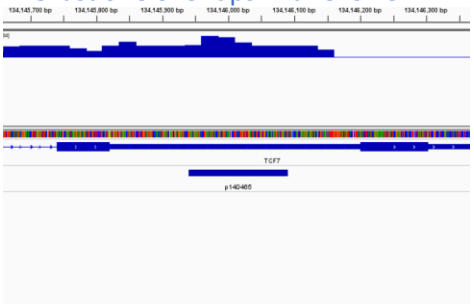
The first three DHS regions are near the promoter of the TCF7 gene, just upstream of the first exon. The third one overlaps with the first exon.



The next three overlap with a large intron downstream of exon 1 and exon 2. The next one is just upstream of exon 5 of the TCF7 gene.



The last one overlaps with exon 9.



Based on the H3K4me3 track, I do believe that the TCF7 gene is expressed in the HUVEC cells, since trimethylation of histone 3, lysine 4 generally leads to active transcription. The DNase activity near the gene (promoter region and exons) also shows that this area is associated with open chromatin, which allows for transcription.

Part 3

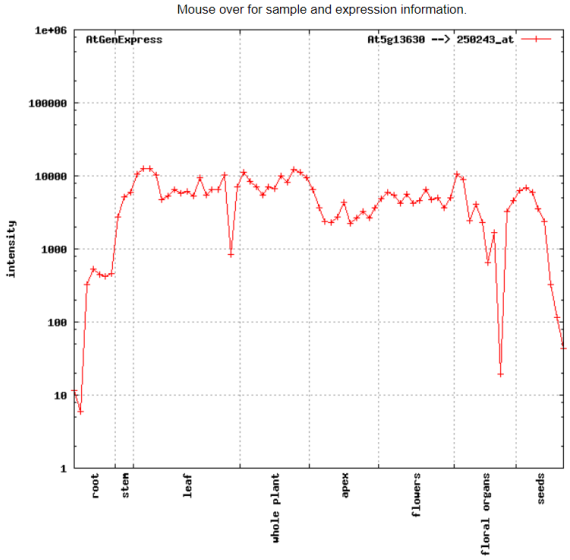
I used YeastMine to find all genes related to GO:0006623 by using the GI ID -> Genes template. Partial results are shown below, there were 71 gene records. Total results are in the attached file GOID_0006623 _Gene.tsv.

Extra credit

For 0.5 pts of extra credit, 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

I used BLASTP on the Tair website to find matches for the Arabidopsis thaliana protein sequence given. The best match was AT5G13630.1, a protein coding gene ABAR expressed in various plants, that is a part of the biochemical processes of photosynthesis and chlorophyll biosynthesis. It is a subunit of a magnesium chelatase complex and it can be found in chloroplast, chloroplast inner membrane, chloroplast stroma, and mitochondrion. Some of its functions are ATP binding, magnesium chelatase activity, and protein binding. There are two protein coding gene models for this gene. The two transcripts look almost exactly the same, except the first exon in the second transcript is mostly UTR with only a small portion protein-coding exon, while the first transcript is more than half protein-coding. From the ATGenExpress visualization, it appears that in root tissue, the expression of this gene is the lowest.



Part 5

Using bedtools on the bfx server command line. I used the following code to find the unique H3K4me3 regions that interesect with coding exons and found 10.

```
bedtools intersect -u -a  hs_chr20_H3K4me3.bed -b hs_chr20_refseq.bed
```

To find the unique H3K4me3 regions that DO NOT interset a coding exon I used the following code, counting the lines because there were so many. The number was 245.

```
bedtools intersect -v -a  hs_chr20_H3K4me3.bed -b hs_chr20_refseq.bed | wc -l
```