UCSC Genome Browser - Supplementary ENCODE Exercises
Andrew Oler (BCBB/OCICB/NIAID)
February 23, 2021

1. **Using chromatin marks, RNA-seq, and CAGE tags to identify transcription start sites (TSSs).**
    a. Load the following:
        i. GENCODE gene track ("Genes and Gene Prediction Tracks" group),
        ii. RNA-seq for MCF-7
            1. click on "ENCODE RNA-seq" in "Expression" Group.
            2. click on CSHL Long RNA-seq
            3. unselect all tracks.
            4. select all MCF-7 tracks
            5. set visibility
                a. Maximum display mode: full
                b. Plus Signal, Minus Signal: full
                c. Alignments: dense
                d. Contigs, Splice Junctions: hide
            6. click Submit
        iii. Integrated Regulation from ENCODE
            1. In "Regulation" group -- click on Integrated Regulation track
            2. select all except "Txn Fac ChIP V2"
            3. Set visibility
                a. Transcription, H3K4me1, H3K4me3, H3K27ac: full
                b. DNase Clusters: pack
                c. Txn Factor ChIP: dense
        iv. RNA Pol II ChIP-seq for MCF-7
            1. click on "ENC TF Binding" track in "Regulation" group.
            2. click on "UTA TFBS"
            3. unselect all and select Pol2 for MCF-7
            4. set visibility to full
        v. 5' cap tags
            1. click on "RIKEN CAGE" in "Expression" group
            2. unselect all and select MCF-7
            3. set visibility
                a. Maximum: full
                b. HMM, Alignments: hide
                c. Minus Signal, Plus Signal: full
            4. set max range of Signal to 50 (click on Minus Signal and Plus Signal to change)
        vi. DNase I for MCF-7
            1. click on "ENC DNase" in "Regulation" group
            2. unselect all
            3. click on "Duke DNaseI HS"
            4. unselect all

5. select MCF-7 “None”
6. keep defaults for visibility (full signal, pack peaks)

vii. ChromHMM
1. click on “track hubs” and turn on Roadmap data
2. hide all Roadmap tracks except ChromHMM.
3. edit ChromHMM tracks to show “Breast Myoepithelial Cells”

b. Click “default order”
c. Go to SMAD4 gene
d. Instead of steps a-c, you can use this link, which has all tracks configured:
   i. http://genome.ucsc.edu/cgi-bin/hgTracks?hgS_doOtherUser=submit&hgS_otherUserName=andrew.oler&hgS_otherUserSessionName=hg19_MCF7
e. Look for CAGE tags to match with all isoforms in UCSC gene track (not including the isoform overlapping the ELAC1 gene). Which is the dominant TSS?
f. Which chromatin modifications (e.g., H3K27ac, H3K4me1, H3K4me3), DNase I peaks and Pol II peaks are found at each TSS?
g. Looking at RNA-seq signal, can you see which isoforms are expressed?
h. Bonus: Why is there signal on the opposite strand upstream of the dominant TSS?
i. Which cellular fraction (e.g., cytosol, nuclear, etc.) has intronic RNA-seq signal?
j. Bonus: What is the DNaseI peak about 20kb upstream of the dominant TSS? (Take a guess)
k. Is there any evidence for expression of the non-coding isoform of SMAD4 (shared TSS with ELAC1) based on splice junction reads? Hint: expand “Long RNA-seq” alignment tracks by right-clicking on the track name in the left side and set visibility to “squish”.
l. Go to the RCAN1 gene.
   i. Zoom to the TSS region.
   ii. Which TSS is active in MCF-7? What evidence is there to support this?
   iii. What is the peak of active chromatin about 20kb upstream? (Take a guess)
m. Go to GART gene
   i. Zoom to the TSS region.
   ii. Which TSS is active in MCF-7? What evidence is there to support this?
n. Go to the DYRK1A gene.
   i. Zoom to the TSS region plus first intron. chr21:38,729,475-38,824,553
   ii. Which TSS is active?
   iii. Transcripts from which TSS are sent to the cytoplasm? Transcripts starting from which TSS’s are only found in the nucleus?
o. Bonus: Go to the MTHFR gene.
   i. Pan to the TSS region.
   ii. Which isoforms are expressed in MCF-7?

p. A few other examples with clear TSS signals if time permits: C21orf58, WRB

2. **Tissue-specific gene expression.**

Hide all MCF-7 tracks.

a. Load the following tracks:
    i. RNA-seq (mRNA).
        1. click on “RNA” in “Roadmap...” group
        2. unselect all smRNA. Make sure mRNA tracks are all selected.
        3. select visibility for Coverage: pack
        4. click Coverage and set max to 50.
b. A shortcut to loading the tracks is to use this link: http://genome.ucsc.edu/cgi-bin/hgTracks?hgS_doOtherUser=submit&hgS_otherUserName=andrew.oler&hgS_otherUserSessionName=hg19_Roadmap_mRNA
c. Go to the NEUROD1 gene.
    i. What tissue(s) is NEUROD1 expressed in?
d. Try these other genes to look at what tissues they are expressed in, as well as what isoforms are expressed:
    i. IL6
    ii. IL6R
    iii. TBX21
    iv. MYC
e. Load smRNA tracks as well
    i. In Roadmap group, like above in 3.a., but select all smRNA now and deselect all mRNA.
    ii. Set Coverage max to 10000
    iii. Set Coverage visibility to full
f. Go to the MIRLET7A1 gene.
    i. Which is the abundant miRNA and which is the miR-star sequence (i.e., complement to the functional miRNA)?
    ii. What tissue is this expressed in?

3. **Transcription Factor and Histone profiles around genes.**

Hide all Roadmap RNA-seq signal tracks.

a. Load these tracks
    i. All H1 tracks in Roadmap
        1. click on “By Sample” in “Roadmap...” group.
        2. unselect all
        3. select H1es
        4. select visibility: pack
    ii. ChromHMM for H1 (“Roadmap...” group)
b. If a.i. doesn’t work (sometimes, tracks don’t load properly from Roadmap Hub), try loading similar tracks this way
    i. click on “ENCODE Histone Modification” in “Regulation” Group.
    ii. click on “Broad Histone”
    iii. unselect all tracks.
    iv. select all H1-hESC tracks
    v. set visibility
        1. Maximum display mode: full

2. Signal: pack
3. Peaks: hide

vi. Data view scaling: “use vertical viewing range setting” and set max to 25.

c. To toggle between dense and pack, click on one of the H1 track names in the browser.
d. Go to the AGTRAP gene. chr1:11,791,124-11,829,179
    i. What histone modifications are present at the TSS?
    ii. Over the body of the gene?
    iii. At the transcription termination site (TTS)?
    iv. Upstream of the TSS?
    v. Downstream of the gene?
    vi. What histone modifications are *absent* from the TSS or over the body of the gene?
    vii. What is the Chromatin state in the ChromHMM track for each of these regions?
e. Look at the C1orf167 gene to the right of AGTRAP.
    i. What is the Chromatin state in H1 ChromHMM for the TSS of this gene?
    ii. What histone modifications are present/absent to give this signature?
f. Go to the TMPO gene. chr12:98,889,289-98,949,474
    i. What histone modifications are present at the TSS, body of the gene, etc.?
    ii. What is the promoter-looking element 10kb upstream of the TMPO TSS?