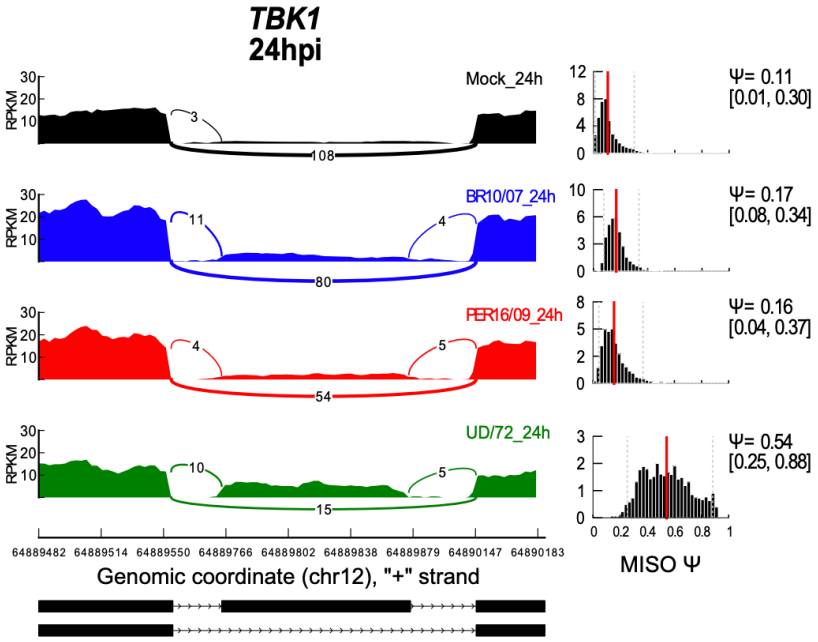
UCSC Genome Browser exercises
Andrew Oler (BCBB/OCICB/NIAID)
February 23, 2021

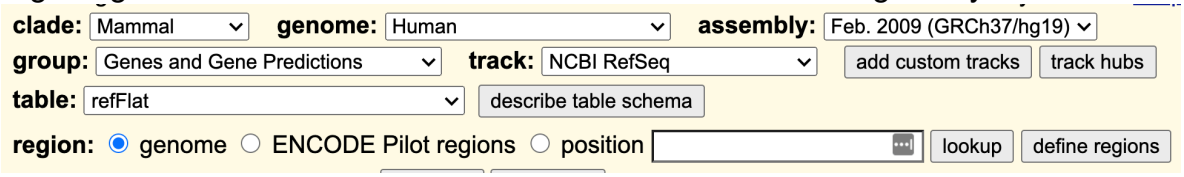
1. **Quick orientation.** Go to the UCSC Genome Browser: genome.ucsc.edu. Make sure "hg19" version is selected. Click on "Genome Browser" at the top. Click on the "default tracks" button below the browser window. Click on the "resize" button. Search for WARS gene in search bar at the top. Select WARS from the drop-down list and click "go". After it loads, zoom out 3x. What can you learn about the gene?
    a. How many transcript isoforms are there for this gene in the UCSC Gene track?
    b. What tissues is the gene expressed in, according to GTEx?
    c. What is the function of the gene? (Click on the top transcript in the "UCSC Genes" track and read the description)
    d. What do the conserved peaks correspond to?
    e. What strand is the gene found on (forward or reverse)? Is the transcription start site on the right side of the page or the left?
    f. How many transcription start sites?
    g. Zoom in on the transcription start region by highlighting a region in the coordinates track and clicking "Zoom In".
    h. What does the H3K27ac profile look like (i.e., high, low) with respect to the transcription start? (configure track to see better). How does H3K27ac correlate to DNAseI Hypersensitivity Clusters?
    i. What repeat sequence is near the transcription start?
    j. What transcription factor binding sites are present in the transcription start site region?
2. **Save PDF or Post-script file of current view**
    a. Still in the region of the transcription start site of WARS gene, click "View -> PDF/PS"
    b. Click "Download the current browser graphic in PDF" or "Download the current browser graphic in EPS". These can be used to make figures for publications.
3. **Download sequence data**
    a. Go to WARS gene.
    b. Where does protein translation start for each transcript?
    c. Zoom in to the common translation start site of WARS.
    d. Click View -> DNA
    e. Select "extended case/color options"
    f. Select "Reverse complement", default "Lower", and check "Toggle case" for NCBI RefSeq; click "submit"
    g. The first exon should be uppercase and you should be able to identify the ATG start codon about 70 bases into the exon.
4. **Liftover: Convert coordinates from hg19 to hg38**
    a. Click on the coordinates
    b. Copy coordinates from the box on the right
    c. Click on Tools -> LiftOver
    d. Select hg19 for "Original" and hg38 for "New Assembly"
    e. Paste coordinates in the box and click Submit
    f. Click "View Conversions" to download the result. This should be opened in a text editor.
5. **BLAT search**
    a. For this exercise we'll look at *TBK1* novel exon inclusion induced by infection of flu virus. (See PMID 29976658.) In H3N2-Udorn/72-infected BEAS-2B (lung epithelial) cells, we see highest levels of exon inclusion compare to mock or other infections. Inclusion of this novel exon results in a frameshift, predicted to interfere with TBK1

protein expression.



b. Open the file "BEAS2B_Udorn_24h_TBK1_RNAseq_reads.txt" file in a text editor and copy all. These are RNA-seq reads that map to this intronic region.
c. In the Genome Browser, select Tools -> BLAT
d. Paste the sequences and click "Submit"
e. All of the reads should have a match to chr12. Click "browser" for the first match.
f. Where are the boundaries of the novel exon? Does the novel exon contain canonical donor and acceptor sites? (e.g., GT-AG)

6. **Loading ENCODE data tracks**
   a. While still looking at *TBK1* novel exon region, find "ENCODE RNA-seq" in the Expression category. Click on the link.
   b. Click on the "-" to deselect all tracks, then select "CSHL Long RNA-seq" and set to "full".
   c. Click on CSHL Long RNA-seq
   d. Set Maximum display mode to "full"; set Contigs, Splice Junctions and Alignments to "hide"; set Plus Signal and Minus Signal to "full". Select GM12878 and K562 "Whole cell" column. Click submit.
   e. Does the RNA-seq track line up with the novel *TBK1* exon?
   f. What approximate fraction of the transcripts include the novel *TBK1* exon in GM12878? (~8-10%)
7. **Table Browser – download gene table**
   a. Click on Tools -> Table Browser
   b. Select Human genome, hg19 assembly
   c. Under group select "Genes and Gene Predictions"; track "NCBI RefSeq"; table "refFlat"; region "genome". We use "refFlat" because that includes the gene symbol.



   d. For output format, some tools ask for GTF format (e.g., HISAT2, DESeq2) so you can use that. BED file is useful for other applications and we'll use that for now. Select "BED – browser extensible data".

e. Enter a name for the output file “hg19_refFlat_exons_2021-02-23.bed.gz”, select “gzip compressed” under file type returned, and click submit.
f. On the next screen, under “Create one BED record per,” select “Exons”. This will make one row per exon. Click “get BED” and it will download the file. The file should be about 5MB and will take a few seconds to download.

8. **Table Browser – intersection and create custom track**
   a. Go to the Table Browser and select refFlat table as before.
   b. Click on intersection “create”.
   c. Select group “COVID-19”; track “Rare Harmful Vars”; table “covidMuts”
   d. Select “All refFlat records that have **any** overlap with Rare Harmful Vars” and click submit.
   e. Select output format “custom track” and leave output file blank. Click “get output”
   f. On the next screen, name this “refFlat_int_covidMuts” and write something similar in the description.
   g. Select “Coding Exons” and click “get custom track in genome browser”. Notice that the track has been added but you might not see any records.
      i. If you don’t see any records, you can go to Table Browser and click “manage custom tracks”. Click on the hyperlink in the “Pos” column. Should be “chr4”. This will take you to the first record of the custom track.
      ii. Or type “TBK1” in the search bar as there should be exons there in the custom track
   h. Turn on the COVID-19 Rare Harmful Vars track. It’s possible that you don’t see any variants are showing up for the exons in view. Zoom out a little (try 10x, and an additional 3x if needed). Now you should be able to see which exons are overlapping the variants. Note that our query was for *genes* overlapping COVID-19 variants, not *exons*, which is why there are some exons not overlapping COVID-19 variants. Next, we’ll make another custom track containing *only* the exons that overlap COVID-19 variants.
   i. Go to the Table Browser. Your custom track might already be selected. If not, select group “Custom Tracks” and it should populate the track and table. The intersection will stay the same. For output format keep it custom track. Click “get output”
   j. On the next screen name the new custom track “refFlat_coding_exact_overlap_covidMuts” and click “get custom track in genome browser”. The new custom track only contains exons that overlap the variants. Next, we’ll download these exons to a file.
   k. Go to the Table Browser. Select the custom track with “exact_overlap” in the name. Clear the intersection. For output format, select “BED” and click “get output”. Select “Whole Gene” and click “get BED”. The results will show on the screen. You can copy and paste this into a text editor and save it for your own analysis. For example, you could screen your own cohort of COVID-19 patients for rare variants in these coding exons.