Exercises for the ENCODE data in the UCSC Genome Browser

1) Using RNA-seq data, examine expression of RNA in the vicinity of the TP53 gene. From the CSHL Long RNA-seq track, determine which strand is transcribed into Poly-A+ RNA and then found in the nuclear fraction of K562 and GM12878 cells.

Skills: Use RNA-seq data to evaluate RNA presence in a region; become aware of the cellular fraction data that is available.

2) In the region we are exploring, let's add transcription factor binding data and histone marks that are often found near active regulatory elements. Let's also determine if these histone marks are indicated in human embryonic stem cells.

Skills: Explore TFBS data; examine features associated with histone modifications; visualize cell type specific data.

3) Use the Table Browser to locate NFKB transcription factor binding signals that are greater than 500 on chromosome 21. Let's intersect that with RNA-seq data indicating presence of RNA in epidermal keratinocyte cells (NHEK cells).

Skills: Table Browser to query ENCODE data; use filters and intersections to generate a complex customized query of the data.

For additional guidance and ways to interact with the ENCODE data, access this open access publication in PLoS Biology: http://bit.ly/plosENC

Citation: The ENCODE Project Consortium (2011) A User's Guide to the Encyclopedia of DNA Elements (ENCODE). PLoS Biol 9(4): e1001046. doi:10.1371/journal.pbio.1001046

UCSC ENCODE 2 Exercises, version 2.
Correspond to the data available in October 2012.

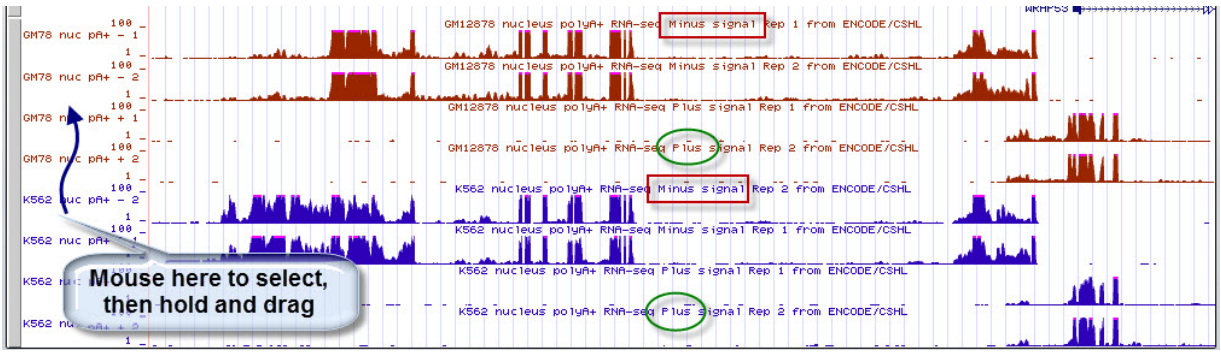
Seattle
12600 SE 38th Street, Suite 230
Bellevue, WA 98006
(425) 401-1400

Boston
65 Main Street
Somerville, MA 02145
(617) 627-9398

San Francisco
193 Haight Street
San Francisco, CA 94012
(415) 252-1519

1) Using RNA-seq data, examine expression of RNA in the vicinity of the TP53 gene. From the CSHL Long RNA-seq track, determine which strand is transcribed into Poly-A+ RNA and then found in the nuclear fraction of K562 and GM12878 cells.

Step	Action	✓
1	Go to the UCSC Genome Browser homepage, **genome.ucsc.edu**	
2	From the blue navigation links on the left side of the page, **click the link for Genome Browser.**	
3	From the Gateway interface**, click the link that says "Click here to reset the browser user interface settings to their defaults."** This will ensure that any prior activity on the Browser has been cleared out and that everyone is starting with default settings.	
4	Choose the Human **February 2009** assembly. Enter the text **tp53** in the gene box. Choose the TP53 item in the list. Click **submit**.	
5	In the TP53 region on the browser, examine the features briefly. Then **click the "zoom out" 1.5x button** near the top. Assess the features again.	
6	**Click the "hide all" button** in the middle of the resulting Genome viewer page. *(We want to reduce what's in the display to reduce the burden on the servers, and to focus on our features of interest.)*	
7	**Add back 2 tracks** to the viewer: • **GENCODE Genes in "show"** visibility (from the Genes and Gene Predictions group) • **ENC RNA-seq…in "show"** (from the Expression group) **Click a refresh button** to add these tracks back to the viewer. It may take a while for this to load, as there is a lot of data here.	
8	RNA seq data from multiple labs, cell lines, and experiment types are shown. Let's focus on the Long RNA-seq data. You can see there is signal across this region indicating RNA transcription in this region of the genome in this mode. But we'd like to distinguish which RNA-seq data corresponds to which genes in this region. **Return to the ENC RNA-seq… hyperlink and click it to access individual tracks from this super-track.**	
9	Note all the RNA-seq… component tracks are in "dense" visibility at this time. **Turn all of them to "hide" except for the CSHL Long RNA-seq menu.**	
10	**Click the CSHL Long RNA-seq hyperlink. Examine the options you can set to explore this track's data.** *(continued on next page)*	

11	**At the top, set the Maximum display mode to "full".** **Make these changes** on the Long RNA-seq settings: ***Set Contig view to "hide".** ***Select the GM12878 cell line "nucleus" localization checkbox.** ***Unselect all other localization checkboxes except K562 "nucleus".** **Below the cell line table, in the RNA Extract box choose only Poly A+:** ***Poly-A+ should be checked. Unselect any others.** **Click "Submit" when these changes have been made.**	
12	Back on the viewer, examine the data. **Use the select/drag feature of the left label area to move the plus and minus data sets together.** 	
13	**Note the data which derives from the Plus strand and which from the Minus strand.** It appears that the WRAP53 RNA derives from the plus strand, and the TP53 RNA from the minus strand. This will help you to orient when looking for transcription factor binding sites or other genomic features.	

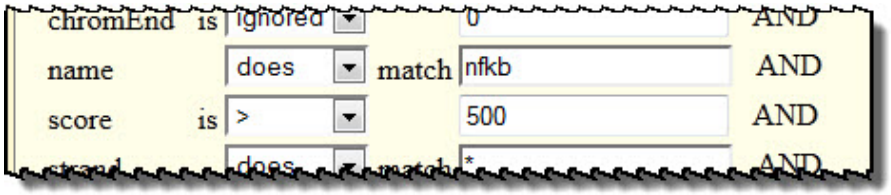
This exercise was inspired by the Figure 3 illustration in the ENCODE User Guide paper. See that figure legend and the accompanying text for more assessments of the data and the features in this region: ***http://bit.ly/plosENC***

2) In the region we are exploring, let's add transcription factor binding data and histone marks that are often found near active regulatory elements. Let's also determine if these histone marks are indicated in human embryonic stem cells.

Step	Action	✓
1	On the browser view that we established in exercise 1, **scroll down to the Regulation Group.**	
2	**Locate the ENCODE Regulation**… track. **Choose "show"** in the pulldown menu. **Click a "refresh" button.**	
3	**Examine the display.** New data appears in the viewer beneath the RNA-seq data. *Note that the Transcription Factor ChIP-seq from ENCODE track shows data blocks, but not individual transcription factors. *Note that the H3K27Ac histone mark track appears to have multiple data sets of various colors.	
4	**Return to the ENCODE Regulation menu area. Click the hyperlink to look at the component tracks of this super-track.**	
5	By default Txn Factor ChIP is visible in "dense" mode. **Set that menu to "full". Click the "Submit" button.**	
6	Examine the display again. Note that individual transcription factors can be identified by name using the labels on the left. Note that the letter codes near the blocks correspond to cell lines that have been used in experiments for this data. **Click some of the blocks** to note the cell lines and signal levels observed in them. Return to the viewer for the next steps.	
7	**Return to the ENCODE Regulation menu area. Click the hyperlink to look at the component tracks of this super-track again.**	
8	On the Integrated Regulation page, **click the hyperlink for Layered H3K27Ac to go to the controls for that track.**	
9	On this histone mark page, note that there are various cell line data sets, which have color codes. One of the lines is **H1-hESC**, which is a human embryonic stem cell line.	
10	**Uncheck all cell line boxes except H1-hESC.**	
11	**Click the "Submit" button at the top** to return to the genome viewer.	
12	Note that we can now see that **there is signal associated with this histone mark in stem cells** in this region. This was difficult to examine before because of the other color overlays.	
13	**Return to the histone mark page** as we did in steps 7-8, or by clicking the gray bar to the left of the browser track. **Turn on or off various cell lines** to view the data. **Return to the viewer** each time by clicking "Submit".	
14	The various data types in this region should help you to understand possible features of regulation of the genes in this area.	

This exercise was inspired by the Figure 5 illustration in the ENCODE User Guide paper. See that figure legend and the accompanying text for more assessments of the data and the features in this region: **http://bit.ly/plosENC**

3) Use the Table Browser to locate NFKB transcription factor binding signals that are greater than 500 on chromosome 21. Let's intersect that with RNA-seq data indicating presence of RNA in epidermal keratinocyte cells (NHEK cells).

Step	Action	✓
1	From the genome browser, **click the navigation bar option Tools menu item called Table Browser.**	
2	**At the Table Browser, begin to establish the query with these choices:** ***Mammal, human, February 2009** ***Regulation group, Txn Factor ChIP track** ***table wgEncodeRegTfbsClusteredV2** ***region: position chr21. Click the lookup button to load the chr21 range.**	
3	Next we'll set a filter. **Click the "create" button.** We want the factor NFKB, and signals to be over 500. ***in the name area chose "does" match nfkb *[remove the asterisk]*** ***in the score choose "is >" and type 500 in the text box** **Your filter should look like this:**  **Click submit.**	
4	**Click the summary/statistics button to assess the results at this point.** This will provide a sense of how many results return with these settings. If there were too many or too few, you might want to adjust the filters accordingly. **Return to the table browser interface** with the back button.	
5	Let's take a look at the output at this point. In the "output format" area **select "all fields from selected table".**	
6	**Click "get output"** to see the results in table form. Note that the Name field has our choice, and all the scores are over 500.	
	Let's intersect this data with some other data. Let's require that this NFKB data also overlap with RNA-seq evidence in a particular cell type of our choice.	
7	**Use the back button** to go back to the Table Browser interface. It should still have all of your previous choices and settings.	
8	**Find the "intersection" option, and click the "create" button.** *(continued on the next page)*	

9	Make these choices in the "Intersect with" interface: ***In the group menu, select Expression**. ***Track choice: select CSHL Long RNA-seq.** *(This is only because we are already familiar with this track and have some of it visible in the browser. You could choose any of the data sets later.)* ***Table selection: choose nuclear NHEK polyA+ first data set**. Looks like: **NHEK nuc pA+ + 1 wgEncodeCshlLongRnaSeqNhekNucleusPapPlusRawSigRep3** *(This is the CSHL Long data set Plus track)*	
11	Ensure that the Intersect **radio button is set to the first choice** for "any overlap".	
12	**Click submit.**	
13	This time let's **choose "output format" as "hyperlinks to Genome Browser".** This will allow us to quickly inspect some of the results visually.	
14	**Click the "get output" button**.	
15	**Click on some of the links to explore** the region of the browser that meets these criteria. Zoom out for larger scope. Do you see the NHEK data in the current view? If not, **go to the ENC RNA seq… super-track** and **access the CSHL Long RNA-seq track details** as we did before. **Select NHEK nuclear data** to add it to the viewer. **Submit.** Show or hide various expression tracks, transcription factor tracks, or any other features you are interested in. You may need to turn on or off tracks in the browser because they were not on when we were using it before.	

This was designed to emphasize that you might choose things in the table browser that are not yet visible in the graphical browser. Remember to adjust the settings back in the viewer to see the items you need.