Using UCSC Tools for Browsing and Data Mining ENCODE Data

Aims

- Learn to locate and display ENCODE data in the UCSC Genome Browser
- Learn to retrieve ENCODE data from the UCSC Genome Browser database using the Table Browser data retrieval tool

Introduction

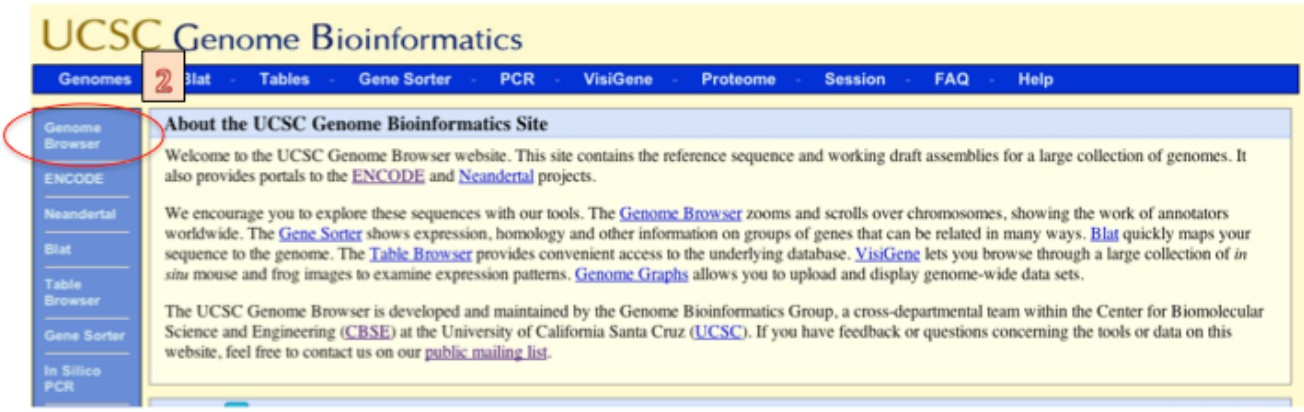
The University of California Santa Cruz (UCSC) Genome Browser athttp://genome.ucsc.edu is a web-based set of tools providing access to a database of genome sequence and annotations for visualization, comparison and analysis by the scientific, medical and academic communities. The primary mission of the site is to provide timely and convenient open access to high-quality human genome sequence and annotations in a framework that enables easy exploration from genome-wide down to the base level. Annotation datasets, or 'tracks', on the human genome cover conservation and evolutionary comparisons, gene models, regulation, expression, epigenetics and tissue differentiation, variation, phenotype and disease associations. A substantial contributor to our mission has been participation in the ENCODE project as the designated data repository in the ENCODE Pilot (2003-2007) and as the Data Coordination Center (DCC) in the ENCODE whole-genome data production phase (2007-2011). All production ENCODE data is routed to UCSC for validation, quality review, database storage, visualization, and dissemination to other public databases. At this time over 2700 distinct ENCODE experiments have been processed by the DCC and made publicly available.

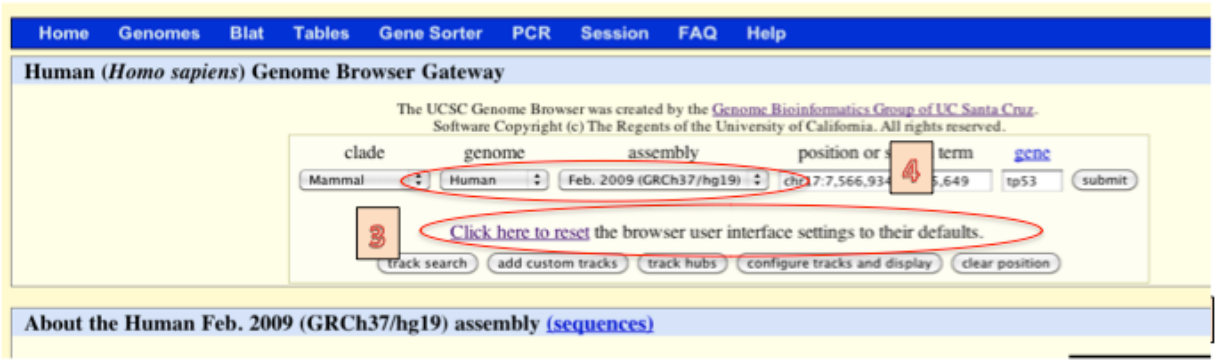
Other organisms represented at the site include 4 non-human primates, 14 other mammals including a marsupial and a monotreme, 10 non-mammalian vertebrates and 24 non-vertebrates. The Genome Browser hosts mapping and sequence annotation tracks that describe assembly, gap and GC content for all organisms in the browser database. Additionally, for most organisms we show alignments from RefSeq genes, mRNAs and ESTs from GenBank, and other gene or gene prediction tracks such as Ensembl Genes (6). For human and mouse assemblies, we also offer a locally generated UCSC Genes track based upon RefSeq, GenBank, CCDS and UniProt data. About half of the genomes hosted at UCSC include a multiple sequence alignment track and pairwise genomic alignments between assemblies to further comparative and evolutionary investigations. Expression, regulation, variation and phenotype tracks are available for many of the assemblies. We also support user data upload and visualization, and have recently introduced a data hub mechanism allowing visualization of user data hosted remotely.

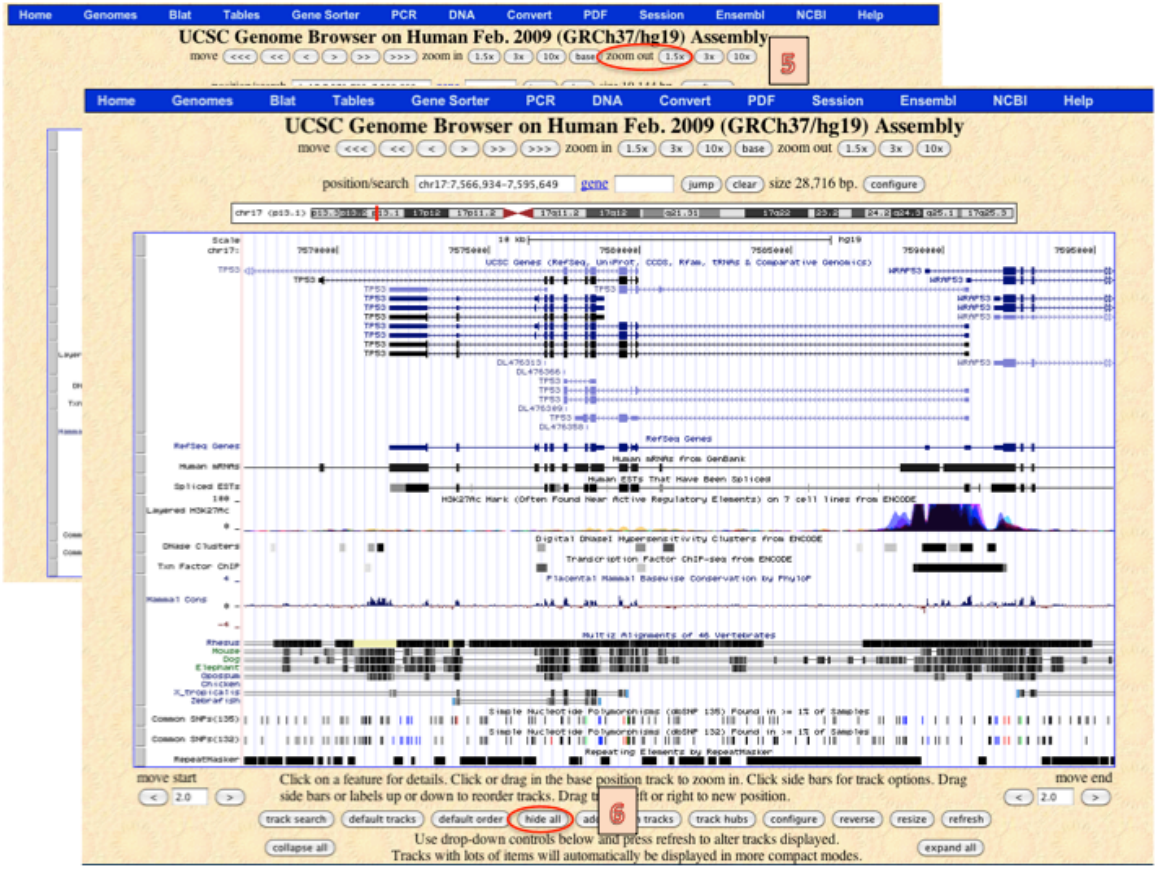
Worked Example 1: You will need to follow this through with the OpenHelix exercises at the end of this section.

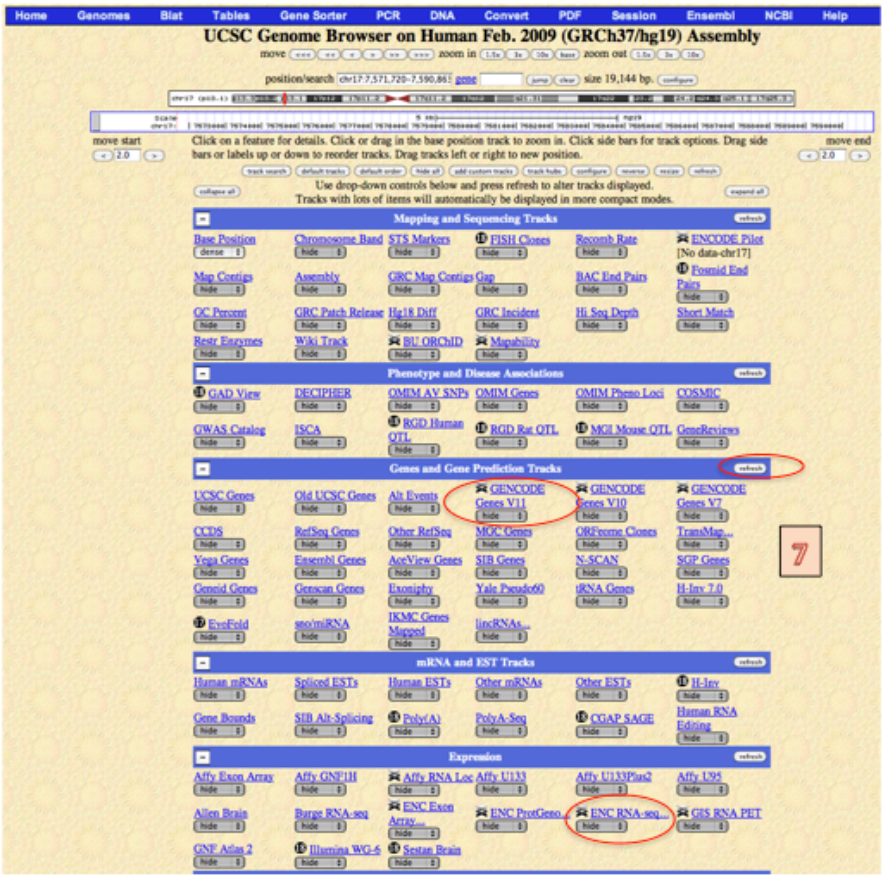
Examining RNA expression in the vicinity of the TP53 gene

1 Browse to genome.ucsc.edu.

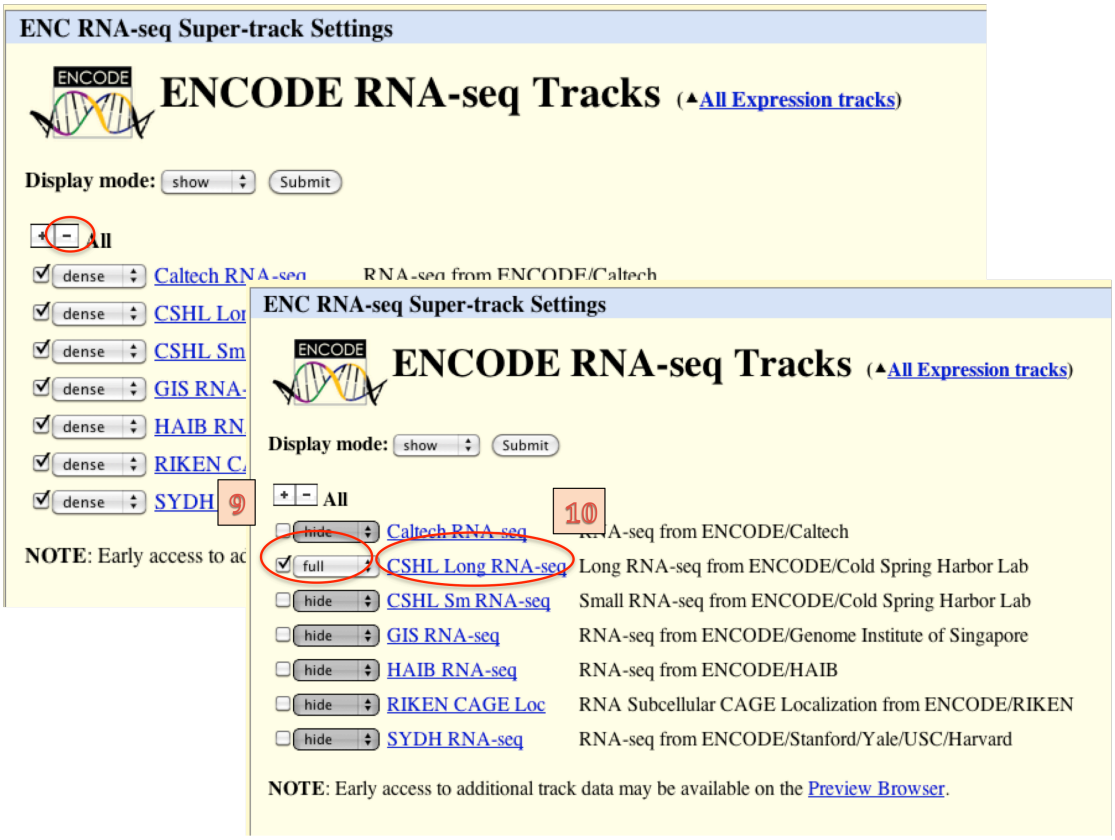


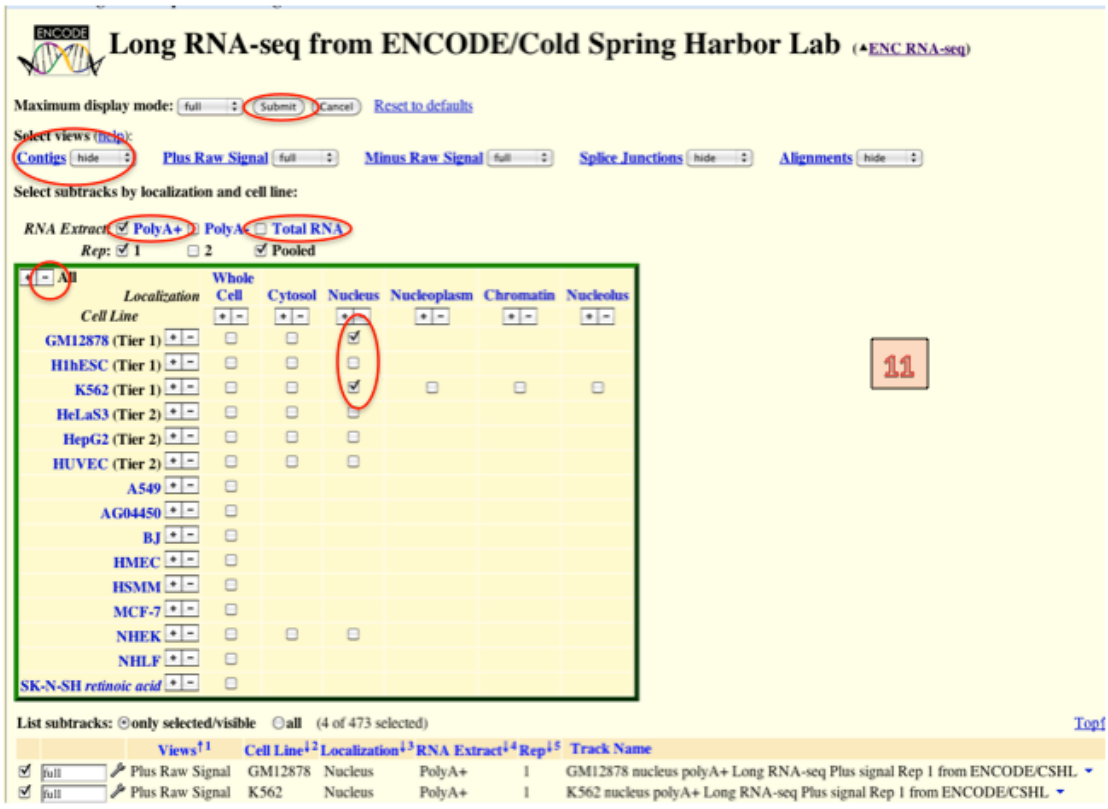


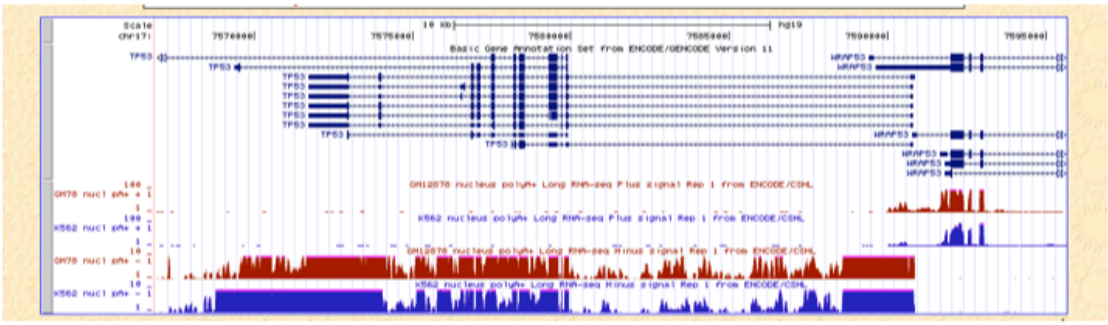

Home Genomes Blat Tables Gene Sorter PCR DNA Convert PDF Session Ensembl NCBI Help
UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly
move zoom in zoom out
5
position/search chr17:7,566,934-7,595,649 gene jump clear size 28,716 bp. configure
RefSeq Genes
Human mRNAs from GenBank
Human ESTs That Have Been Spliced
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Digital DNaseI Hypersensitivity Clusters from ENCODE
Transcription Factor ChIP-seq from ENCODE
Placental Mammal Basewise Conservation by PhyloP
Multiz Alignments of 46 Vertebrates
Repeating Elements by RepeatMasker
move start
move end
Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag
side bars or labels up or down to reorder tracks. Drag t ft or right to new position.
6
track search default tracks default order hide all track hubs configure reverse resize refresh
collapse all
Use drop-down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.
expand all


Home Genomes Blat Tables Gene Sorter PCR DNA Convert PDF Session Ensembl NCBI Help
UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly
Mapping and Sequencing Tracks
Phenotype and Disease Associations
Genes and Gene Prediction Tracks
GENCODE Genes V11
GENCODE Genes V10
GENCODE Genes V7
mRNA and EST Tracks
Expression
Affy Exon Array
Affy GNF1H
Affy RNA Loc
Affy U133
Affy U133Plus2
Affy U95
Allen Brain
Burge RNA-seq
ENC Exon Array...
ENC ProtGeno...
ENC RNA-seq...
GIS RNA PET
GNF Atlas 2
Illumina WG-6
Sestan Brain
7


Expression
refresh
Affy Exon Array
Affy GNF1H
Affy RNA Loc
Affy U133
Affy U133Plus2
Affy U95
Allen Brain
Burge RNA-seq
ENC Exon Array...
ENC ProtGeno...
ENC RNA-seq...
GIS RNA PET
GNF Atlas 2
Illumina WG-6
Sestan Brain
8


ENC RNA-seq Super-track Settings
ENCODE RNA-seq Tracks (▲All Expression tracks)
Display mode: show Submit
All
dense Caltech RNA-seq RNA-seq from ENCODE/Caltech
dense CSHL Lor
dense CSHL Sm
dense GIS RNA-
dense HAIB RN
dense RIKEN C
dense SYDH
9
NOTE: Early access to ac
ENC RNA-seq Super-track Settings
ENCODE RNA-seq Tracks (▲All Expression tracks)
Display mode: show Submit
All
10
hide Caltech RNA-seq RNA-seq from ENCODE/Caltech
full CSHL Long RNA-seq Long RNA-seq from ENCODE/Cold Spring Harbor Lab
hide CSHL Sm RNA-seq Small RNA-seq from ENCODE/Cold Spring Harbor Lab
hide GIS RNA-seq RNA-seq from ENCODE/Genome Institute of Singapore
hide HAIB RNA-seq RNA-seq from ENCODE/HAIB
hide RIKEN CAGE Loc RNA Subcellular CAGE Localization from ENCODE/RIKEN
hide SYDH RNA-seq RNA-seq from ENCODE/Stanford/Yale/USC/Harvard
NOTE: Early access to additional track data may be available on the Preview Browser.


Long RNA-seq from ENCODE/Cold Spring Harbor Lab (▲ENC RNA-seq)
Maximum display mode: full Submit Cancel Reset to defaults
Select views (hide):
Contigs hide Plus Raw Signal full Minus Raw Signal full Splice Junctions hide Alignments hide
Select subtracks by localization and cell line:
RNA Extract: PolyA+ PolyA- Total RNA
Rep: 1 2 Pooled
All
Localization Whole Cell Cytosol Nucleus Nucleoplasm Chromatin Nucleolus
Cell Line
GM12878 (Tier 1)
H1hESC (Tier 1)
K562 (Tier 1)
HeLaS3 (Tier 2)
HepG2 (Tier 2)
HUVEC (Tier 2)
A549
AG04450
BJ
HMEC
HSMM
MCF-7
NHEK
NHLF
SK-N-SH retinoic acid
11
List subtracks: only selected/visible all (4 of 473 selected) Top
Views Cell Line Localization RNA Extract Rep Track Name
full Plus Raw Signal GM12878 Nucleus PolyA+ 1 GM12878 nucleus polyA+ Long RNA-seq Plus signal Rep 1 from ENCODE/CSHL
full Plus Raw Signal K562 Nucleus PolyA+ 1 K562 nucleus polyA+ Long RNA-seq Plus signal Rep 1 from ENCODE/CSHL



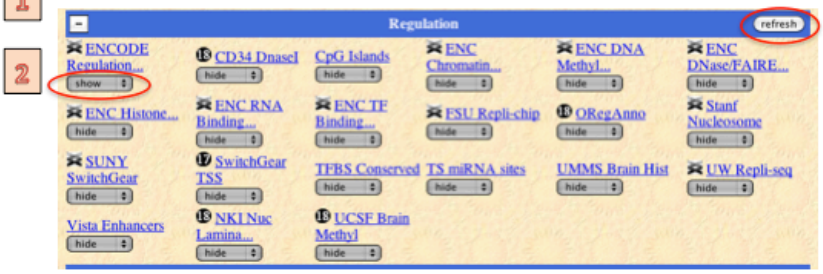
12

Worked Example 2:

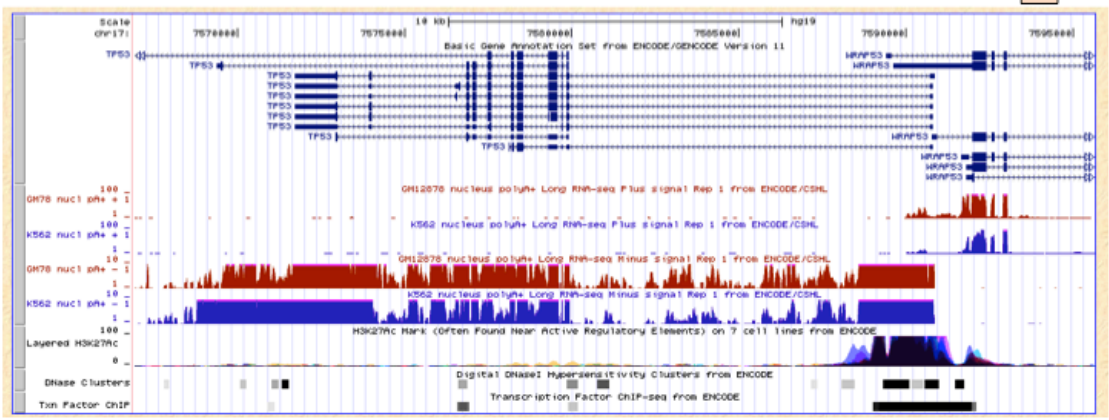
Exploring TFBS and Histone Marks in the TP53 region

1

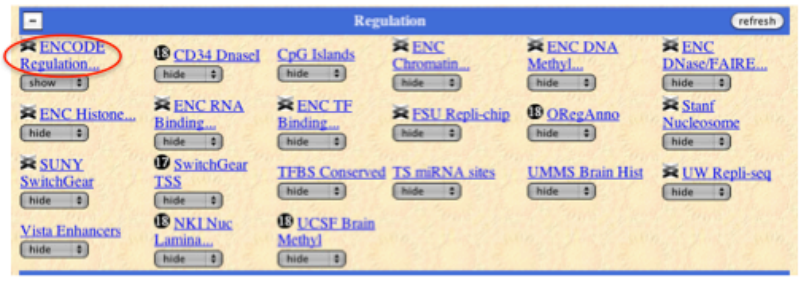
2

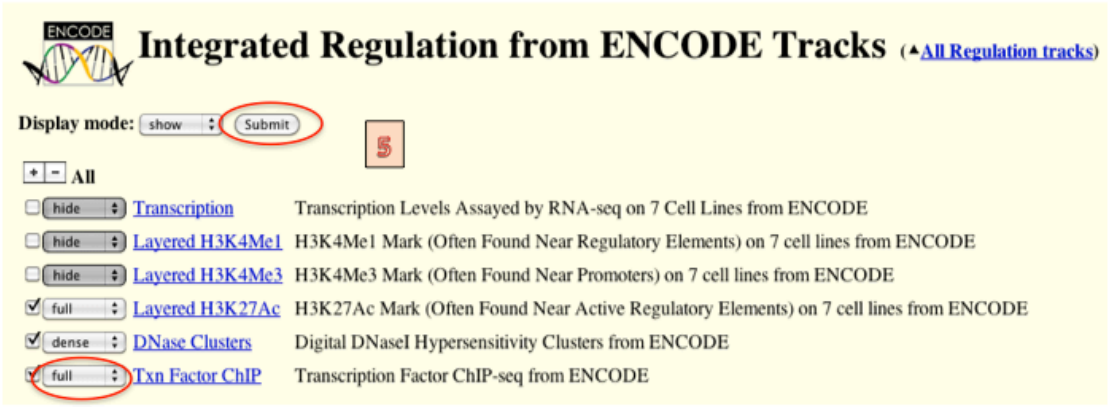


3

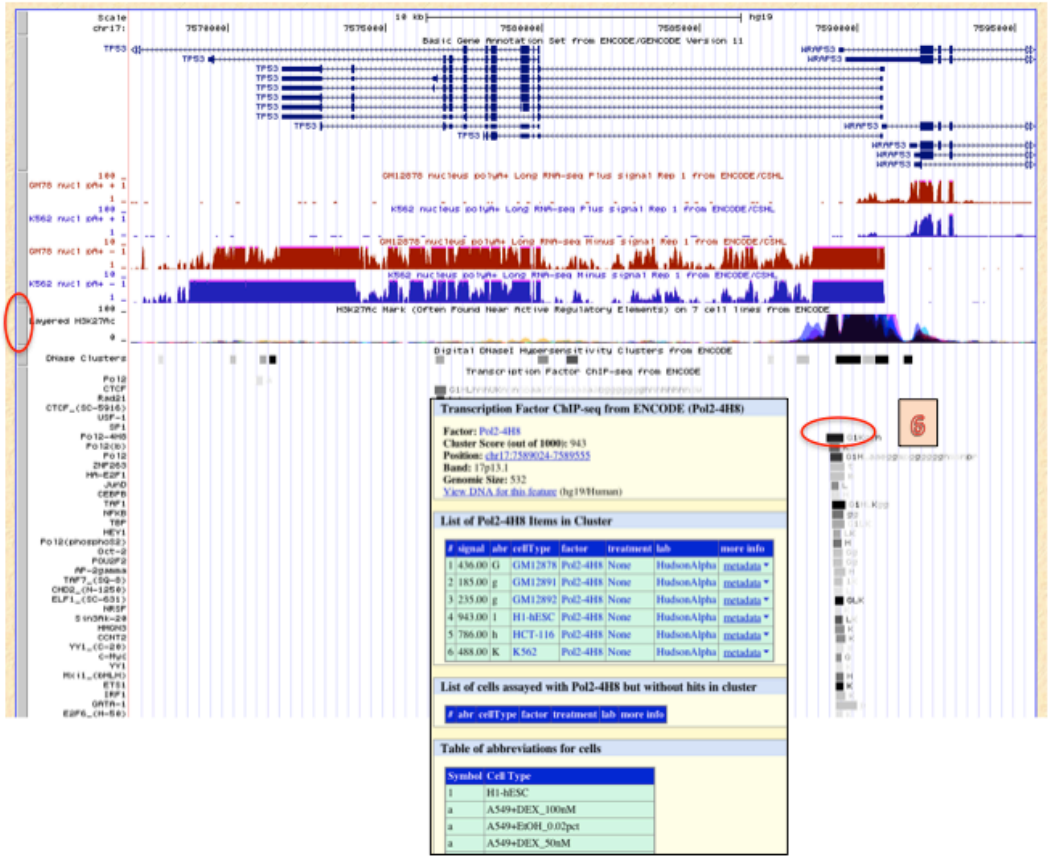


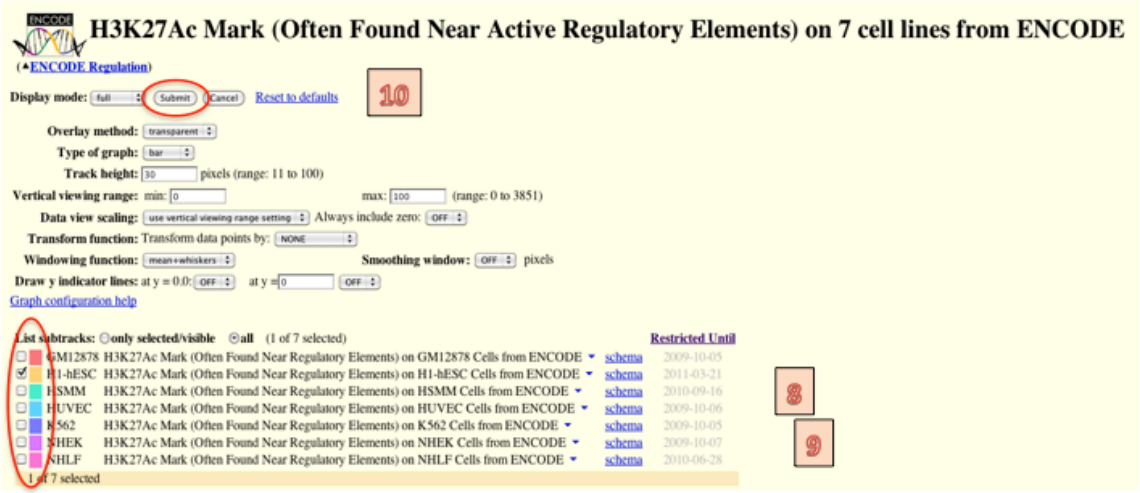
4

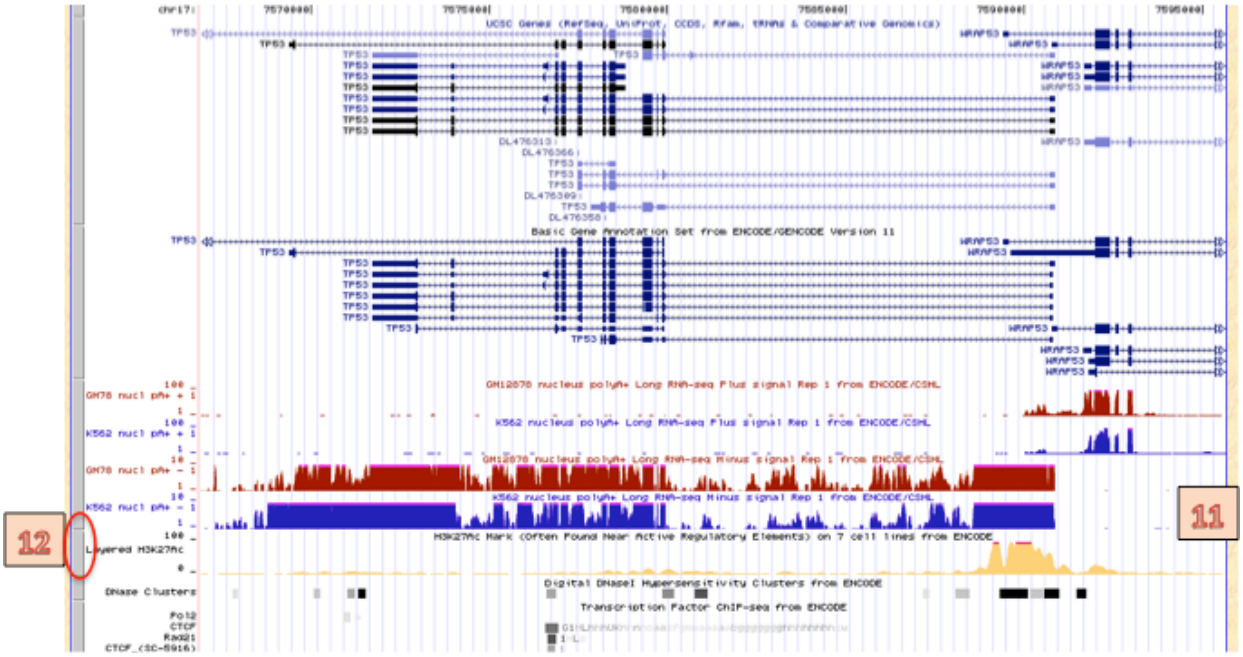

Regulation
refresh
ENCODE Regulation...
CD34 DnaseI
CpG Islands
ENC Chromatin...
ENC DNA Methyl...
ENC DNaseI/FAIRE...
ENC Histone...
ENC RNA Binding...
ENC TF Binding...
FSU Repli-chip
ORegAnno
Stanf Nucleosome
SUNY SwitchGear
SwitchGear TSS
TFBS Conserved
TS miRNA sites
UMMS Brain Hist
UW Repli-seq
Vista Enhancers
NKI Nuc Lamina...
UCSF Brain Methyl


Integrated Regulation from ENCODE Tracks
All Regulation tracks
Display mode:
show
Submit
5
All
Transcription
Transcription Levels Assayed by RNA-seq on 7 Cell Lines from ENCODE
Layered H3K4Me1
H3K4Me1 Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K4Me3
H3K4Me3 Mark (Often Found Near Promoters) on 7 cell lines from ENCODE
Layered H3K27Ac
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
DNase Clusters
Digital DNaseI Hypersensitivity Clusters from ENCODE
Txn Factor ChIP
Transcription Factor ChIP-seq from ENCODE

7

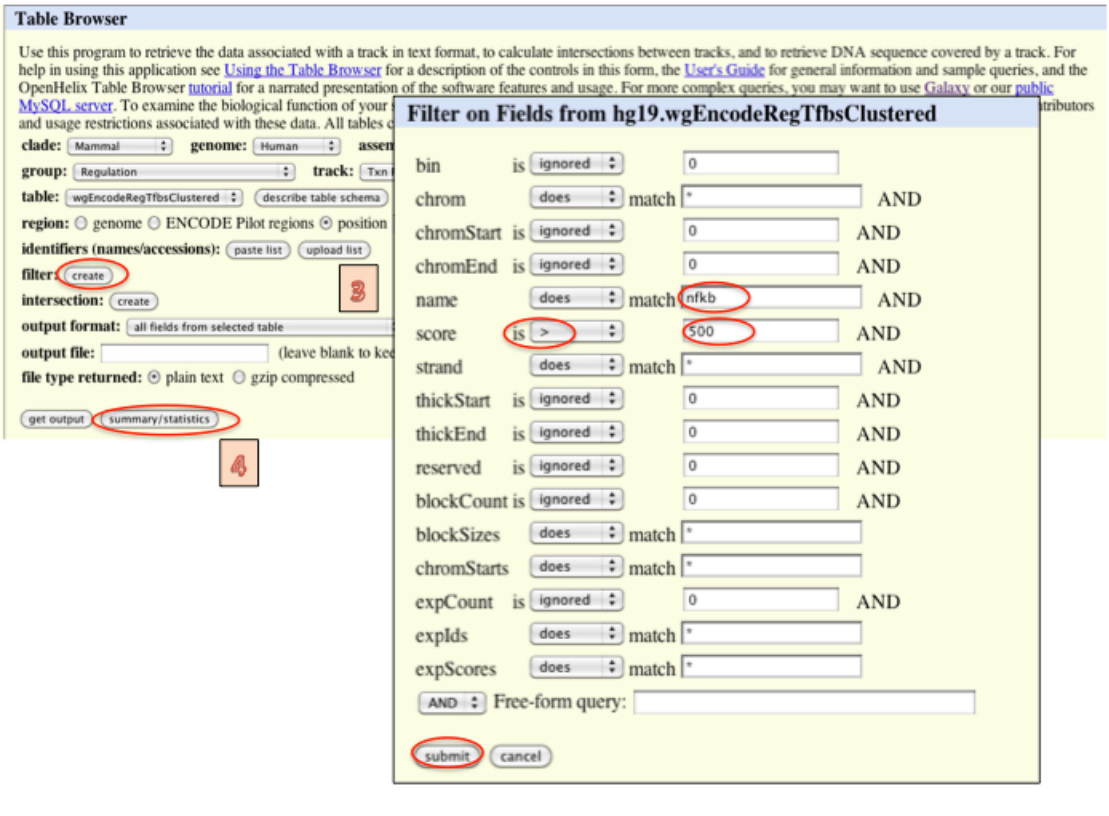

Transcription Factor ChIP-seq from ENCODE (Pol2-4H8)
List of Pol2-4H8 Items in Cluster
List of cells assayed with Pol2-4H8 but without hits in cluster
Table of abbreviations for cells
6

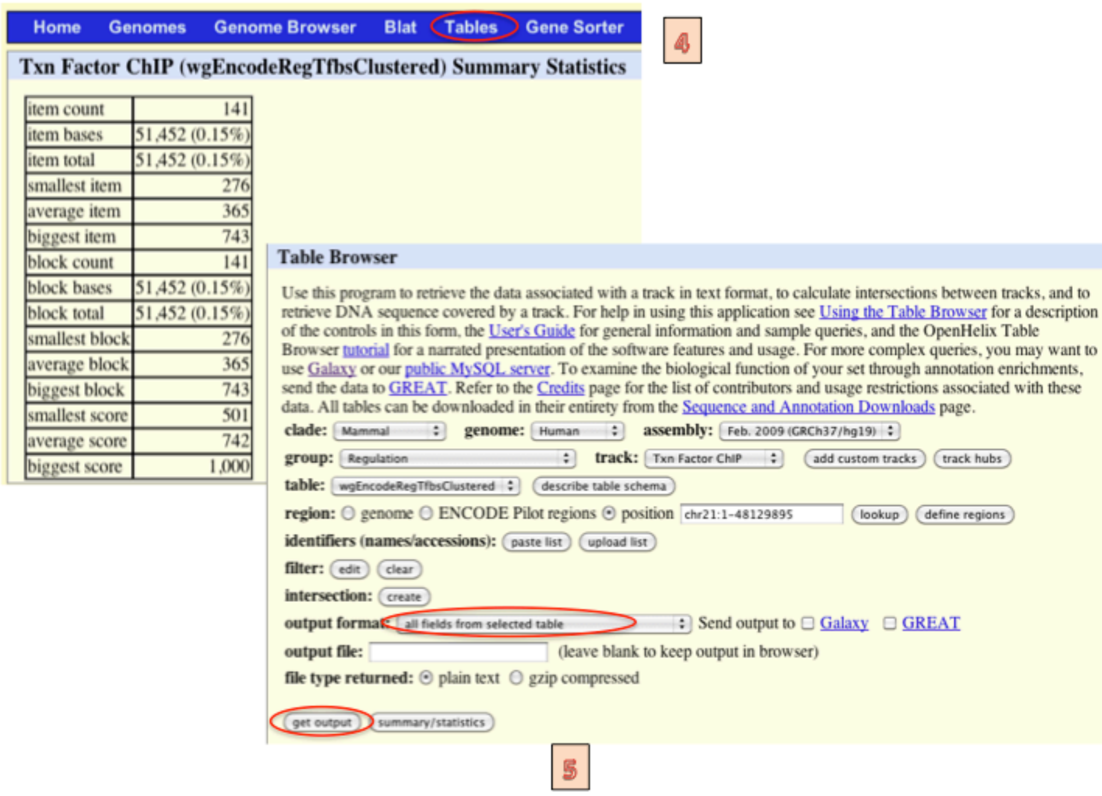

H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
(▲ENCODE Regulation)
Display mode: full Submit Cancel Reset to defaults
10
Overlay method: transparent
Type of graph: bar
Track height: 30 pixels (range: 11 to 100)
Vertical viewing range: min: 0 max: 100 (range: 0 to 3851)
Data view scaling: use vertical viewing range setting Always include zero: OFF
Transform function: Transform data points by: NONE
Windowing function: mean+whiskers Smoothing window: OFF pixels
Draw y indicator lines: at y = 0.0: OFF at y = 0 OFF
Graph configuration help
List subtracks: only selected/visible all (1 of 7 selected)
Restricted Until
GM12878 H3K27Ac Mark (Often Found Near Regulatory Elements) on GM12878 Cells from ENCODE schema 2009-10-05
H1-hESC H3K27Ac Mark (Often Found Near Regulatory Elements) on H1-hESC Cells from ENCODE schema 2011-03-21
HSMM H3K27Ac Mark (Often Found Near Regulatory Elements) on HSMM Cells from ENCODE schema 2010-09-16
HUVEC H3K27Ac Mark (Often Found Near Regulatory Elements) on HUVEC Cells from ENCODE schema 2009-10-06
K562 H3K27Ac Mark (Often Found Near Regulatory Elements) on K562 Cells from ENCODE schema 2009-10-05
NHEK H3K27Ac Mark (Often Found Near Regulatory Elements) on NHEK Cells from ENCODE schema 2009-10-07
NHLF H3K27Ac Mark (Often Found Near Regulatory Elements) on NHLF Cells from ENCODE schema 2010-06-28
1 of 7 selected
8
9

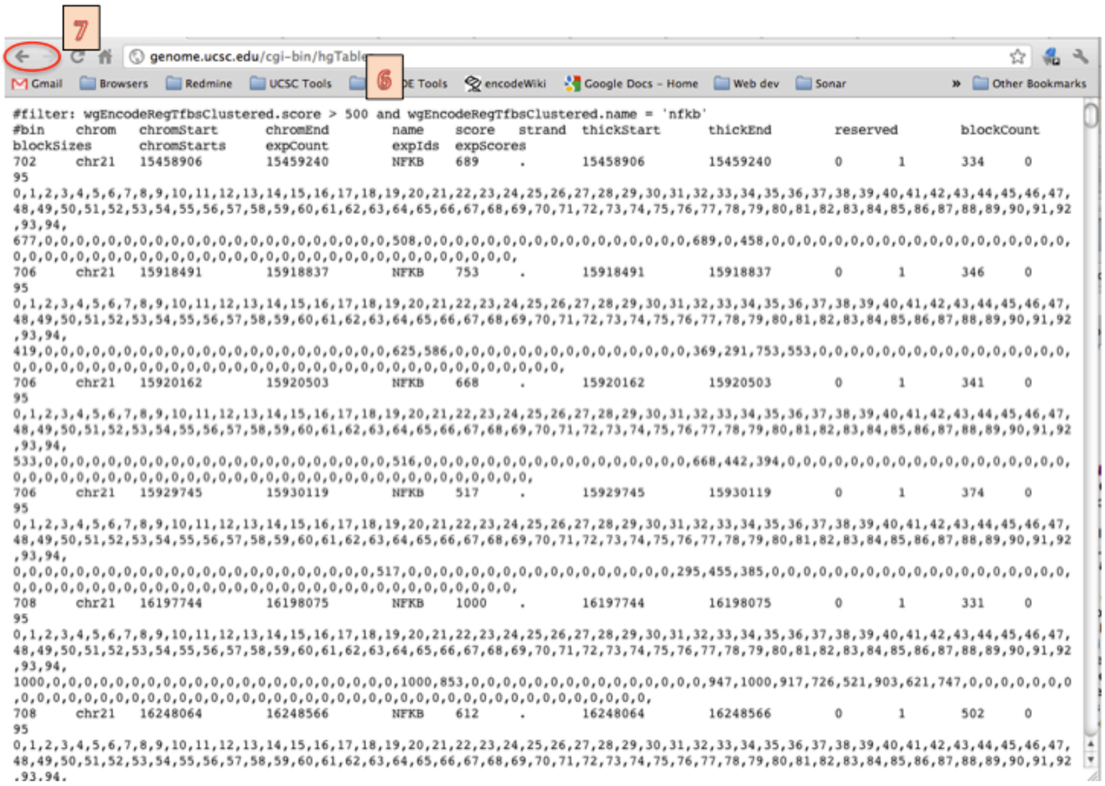

UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
Basic Gene Annotation Set from ENCODE/GENCODE Version 11
GM12878 nucleus polyA+ Long RNA-seq Plus signal Rep 1 from ENCODE/CSHL
K562 nucleus polyA+ Long RNA-seq Plus signal Rep 1 from ENCODE/CSHL
GM12878 nucleus polyA+ Long RNA-seq Minus signal Rep 1 from ENCODE/CSHL
K562 nucleus polyA+ Long RNA-seq Minus signal Rep 1 from ENCODE/CSHL
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Digital DNaseI Hypersensitivity Clusters from ENCODE
Transcription Factor ChIP-seq from ENCODE
DNase Clusters
12
11

Worked Example 3:

Intersect NFKB binding sites with RNA-seq using the Table Browser.




Home Genomes Genome Browser Blat Tables Gene Sorter
4
Txn Factor ChIP (wgEncodeRegTfbsClustered) Summary Statistics
item count 141
item bases 51,452 (0.15%)
item total 51,452 (0.15%)
smallest item 276
average item 365
biggest item 743
block count 141
block bases 51,452 (0.15%)
block total 51,452 (0.15%)
smallest block 276
average block 365
biggest block 743
smallest score 501
average score 742
biggest score 1,000
Table Browser
Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.
clade: Mammal genome: Human assembly: Feb. 2009 (GRCh37/hg19)
group: Regulation track: Txn Factor ChIP add custom tracks track hubs
table: wgEncodeRegTfbsClustered describe table schema
region: genome ENCODE Pilot regions position chr21:1-48129895 lookup define regions
identifiers (names/accessions): paste list upload list
filter: edit clear
intersection: create
output format: all fields from selected table Send output to Galaxy GREAT
output file: (leave blank to keep output in browser)
file type returned: plain text gzip compressed
get output summary/statistics
5


7
genome.ucsc.edu/cgi-bin/hgTable
6
Gmail Browsers Redmine UCSC Tools E Tools encodeWiki Google Docs - Home Web dev Sonar Other Bookmarks
#filter: wgEncodeRegTfbsClustered.score > 500 and wgEncodeRegTfbsClustered.name = 'nfkb'
#bin chrom chromStart chromEnd name score strand thickStart thickEnd reserved blockCount blockSizes chromStarts expCount expIds expScores
702 chr21 15458906 15459240 NFKB 689 . 15458906 15459240 0 1 334 0 95
706 chr21 15918491 15918837 NFKB 753 . 15918491 15918837 0 1 346 0 95
706 chr21 15920162 15920503 NFKB 668 . 15920162 15920503 0 1 341 0 95
706 chr21 15929745 15930119 NFKB 517 . 15929745 15930119 0 1 374 0 95
708 chr21 16197744 16198075 NFKB 1000 . 16197744 16198075 0 1 331 0 95
708 chr21 16248064 16248566 NFKB 612 . 16248064 16248566 0 1 502 0 95

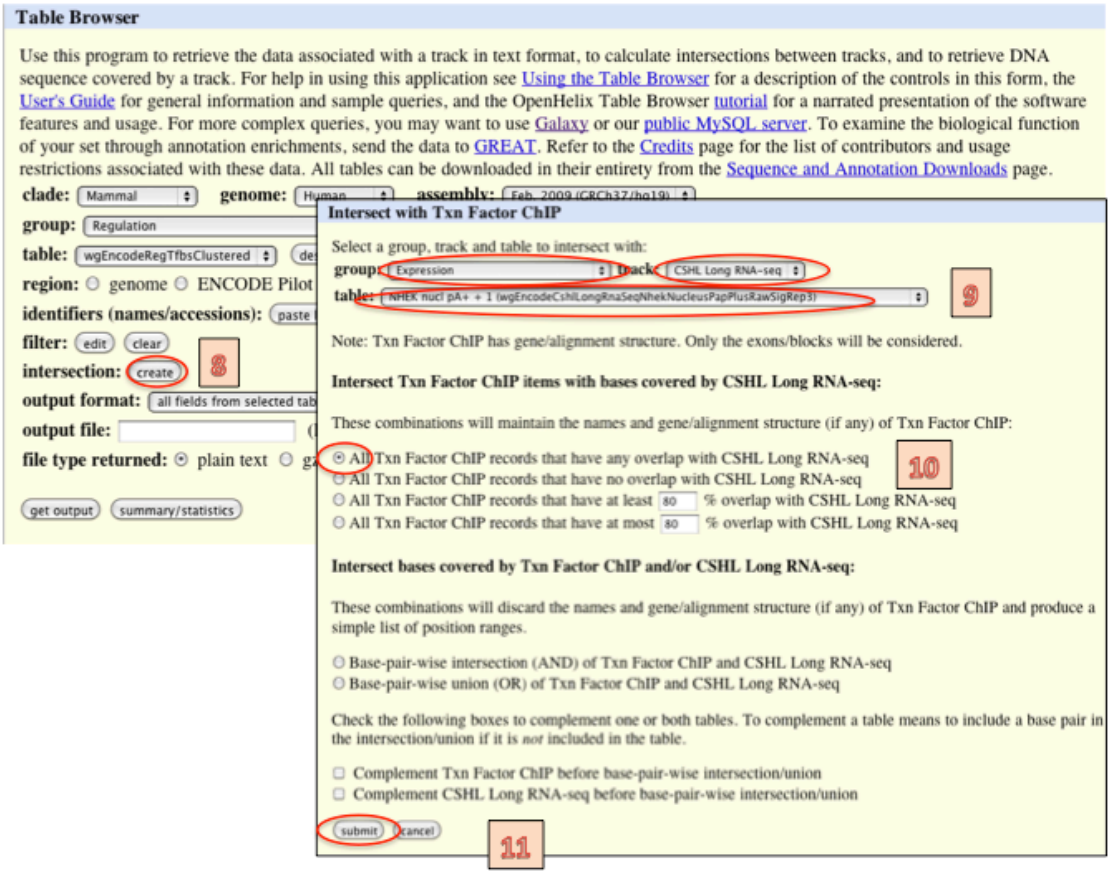


Table Browser
Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.
clade: Mammal genome: Human assembly: Feb. 2009 (GRCh37/hg19)
group: Regulation
table: wgEncodeRegTfbsClustered
region: genome ENCODE Pilot
identifiers (names/accessions): paste
filter: edit clear
intersection: create
8
output format: all fields from selected tab
output file:
file type returned: plain text
get output summary/statistics
Intersect with Txn Factor ChIP
Select a group, track and table to intersect with:
group: Expression track: CSHL Long RNA-seq
table: NHEK nucl pA+ + 1 (wgEncodeCshlLongRnaSeqNhekNucleusPapPlusRawSigRep3)
9
Note: Txn Factor ChIP has gene/alignment structure. Only the exons/blocks will be considered.
Intersect Txn Factor ChIP items with bases covered by CSHL Long RNA-seq:
These combinations will maintain the names and gene/alignment structure (if any) of Txn Factor ChIP:
All Txn Factor ChIP records that have any overlap with CSHL Long RNA-seq
All Txn Factor ChIP records that have no overlap with CSHL Long RNA-seq
All Txn Factor ChIP records that have at least 80 % overlap with CSHL Long RNA-seq
All Txn Factor ChIP records that have at most 80 % overlap with CSHL Long RNA-seq
10
Intersect bases covered by Txn Factor ChIP and/or CSHL Long RNA-seq:
These combinations will discard the names and gene/alignment structure (if any) of Txn Factor ChIP and produce a simple list of position ranges.
Base-pair-wise intersection (AND) of Txn Factor ChIP and CSHL Long RNA-seq
Base-pair-wise union (OR) of Txn Factor ChIP and CSHL Long RNA-seq
Check the following boxes to complement one or both tables. To complement a table means to include a base pair in the intersection/union if it is not included in the table.
Complement Txn Factor ChIP before base-pair-wise intersection/union
Complement CSHL Long RNA-seq before base-pair-wise intersection/union
submit cancel
11

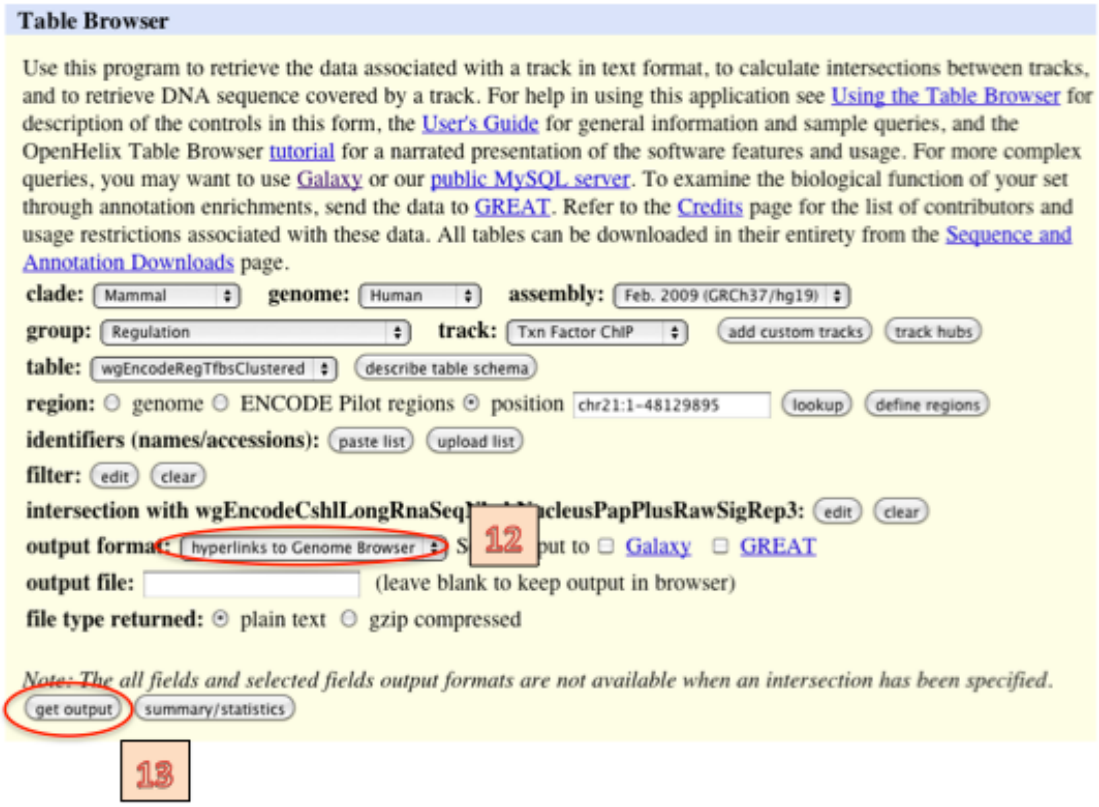
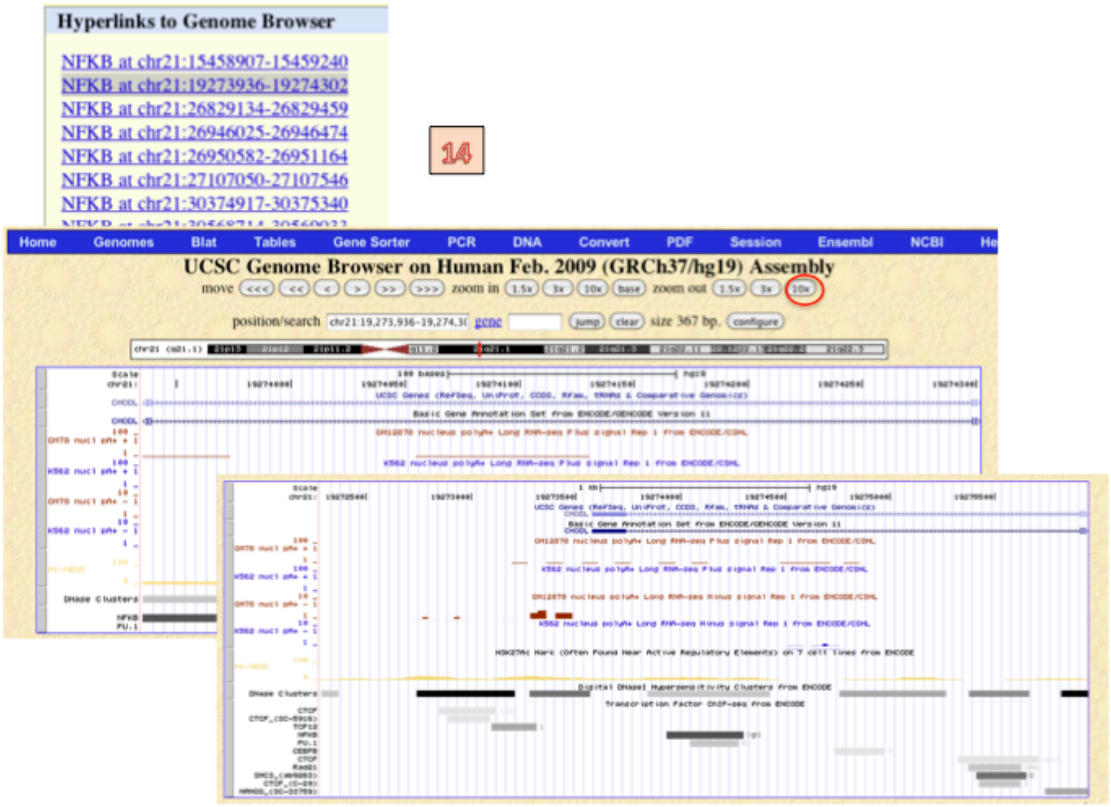


Table Browser
Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.
clade: Mammal genome: Human assembly: Feb. 2009 (GRCh37/hg19)
group: Regulation track: Txn Factor ChIP add custom tracks track hubs
table: wgEncodeRegTfbsClustered describe table schema
region: genome ENCODE Pilot regions position chr21:1-48129895 lookup define regions
identifiers (names/accessions): paste list upload list
filter: edit clear
intersection with wgEncodeCshlLongRnaSeq
cleusPapPlusRawSigRep3: edit clear
output format: hyperlinks to Genome Browser
12
put to Galaxy GREAT
output file: (leave blank to keep output in browser)
file type returned: plain text gzip compressed
Note: The all fields and selected fields output formats are not available when an intersection has been specified.
get output summary/statistics
13


Hyperlinks to Genome Browser
NFKB at chr21:15458907-15459240
NFKB at chr21:19273936-19274302
NFKB at chr21:26829134-26829459
NFKB at chr21:26946025-26946474
NFKB at chr21:26950582-26951164
NFKB at chr21:27107050-27107546
NFKB at chr21:30374917-30375340
14
Home Genomes Blat Tables Gene Sorter PCR DNA Convert PDF Session Ensembl NCBI
UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly
move zoom in zoom out
position/search gene jump clear size 367 bp. configure

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 NKFB 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.

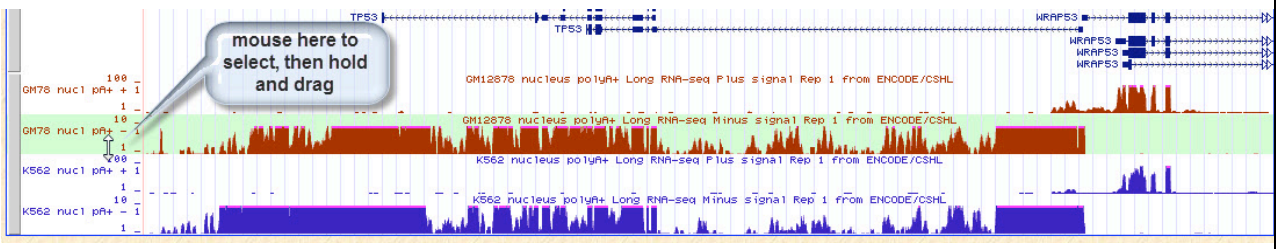
For additional guidance and ways to interact with the ENCODE data, access this open access publication in PLoS Biology: http://bit.ly/plosENC

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 1.
Correspond to the data available in March 2012.

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 V11 in "pack"** 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 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)*	

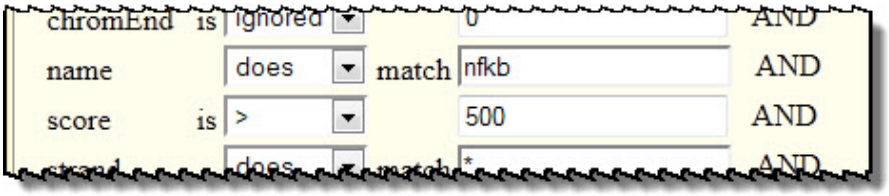
11	**At the top, set the Maximum display mode to "full".** **Make these changes** on the Long RNA-seq settings: ***Set Contig view to "hide".** ***Poly-A+ should be checked. Uncheck "Total RNA" in the extract row. Leave the other settings in that area unchanged.** ***Select the GM12878 cell line "nucleus" localization checkbox.** ***Unselect all other localization checkboxes except K562 "nucleus".** **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 GM12878 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	**Click the grey control button to the left of the Layered H3K27Ac** to go to the controls for that track**.**	
8	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.	
9	**Uncheck all cell line boxes except H1-hESC.**	
10	**Click the "Submit" button at the top** to return to the genome viewer.	
11	**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.	
12	**Return to the histone mark page** 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".	
13	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 NKFB 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 called Tables.**	
2	**At the Table Browser, begin to establish the query with these choices:** ***Mammal, human, February 2009** ***Regulation group, Txn Factor ChIP track** ***table wgEncodeRegTfbsClustered** ***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 by clicking the Tables link.**	
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**. This looks like: **NHEK nucl pA+ + 1** *(This is the CSHL Long data set Plus track)*	
10	Ensure that the Intersect **radio button is set to the first choice** for "any overlap".	
11	**Click submit.**	
12	This time let's **choose "output format" as "hyperlinks to Genome Browser".** This will allow us to quickly inspect some of the results visually.	
13	**Click the "get output" button**.	
14	**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.	

Tasks:

1.
Search for the SLC25A29 gene in UCSC and view the GENCODE geneset.
How many alternative splice variants are there and what are their biotypes?
What regulatory information can you find by investigating the ENCODE data tracks not only for this gene but for the adjacent locus?

2.
Search for RP4-550H1.6.1.
What biotype is this and what can you observe about this locus?

3.
Search for RPN2 and then zoom put so that you can also see C20orf132.
How many alternative variants are there for each of these two loci and what are their biotypes? What can you observe about these two loci?

Answers:

1.
There are 21 alternative splice variants for this locus.
8 protein coding, 3 NMD, 7 processes transcript and 2 retained intron.

Observations: Histone marks H3K4Me3 shows the classic dip at the tss.
There is TF binding, transcriptional evidence and exonic conservation back to zebrafish. Contrast this with the adjacent locus SLC25A47. This just has 2 coding transcripts and a pretty much identical conservation pattern, but appears to be transcriptionally silent in ENCODE cell-types. You can see the difference in histone modifications (very low levels detected), open chromatin (weak signals), TF binding (lower levels) and general transcription (background).

2.
This is a single-exon lncRNA locus that has a biotype of lincRNA (protein coding LOC in UCSC genes).
There is no alternative splicing. The conservation is quite good for a lncRNA, right back to Tenrec. Looking at the Open Chromatin marks there is a string signal from DNAseq and FAIRE, and the histone marks H3K4Me1, Me3 and K27Ac are all strong. Looking at the TF binding there is a huge pileup of factors with a very tight distribution and transcription looks high across the full-length of the transcript. If you look at polyA+ CAGE TSS there are strong signals on the minus strand, with the highest on H1-hESC and K562 and some expression (but not much) in GM12878. The polyA site and signal annotation shows very tight correspondence with the transcription level.

3.
C20orf132 has 10 alternative variants: 9 protein-coding and 1 processed transcript.
RPN2 has 11 alternative variants: 9 protein-coding, 1 NMD and 1 processed transcript.

They have similar levels of conservation, with C20orf132 conserved doen to platypus and RPN2 to zebrafish. The two loci have a head to head arrangement. The histone modifications have detected high levels here and the Open Chromatin marks give a strong signal from DNAse1 and FAIRE. The TF binding has a big pileup of factors and a very tight distibution.
But, if you look at the transcription you can see that RPN2 has much higher transcription levels than C20orf132. This is shown by looking at the CAGE TSS data, as its much higher for the positive strand (RPN2) than the negative strand (C20orf132) and is true in all three cell types. There is also proteogenomic data for RPN2 e.g. peptides TGQEVVFVAEPDNK and FPEEEAPSTVLSQNLFTPK are found in both mitochondrial and nuclear fraction.