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 at http://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 (2007-2012).

In Phase III, beginning in 2012, the DCC is managed at Stanford University and ENCODE production data continues to be routed to UCSC for validation, quality review, database storage, visualization, and dissemination to other public databases. At this time more than 2700 distinct ENCODE experiments have been processed by the DCC and made publicly available.

Other organisms represented at the UCSC Genome Browser site include 11 non-human primates, 34 other mammals including marsupials and a monotreme, 18 non-mammalian vertebrates, 23 invertebrates and yeast. 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. For human and mouse assemblies, we also offer a locally generated UCSC Genes track based upon RefSeq, GenBank and CCDS 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

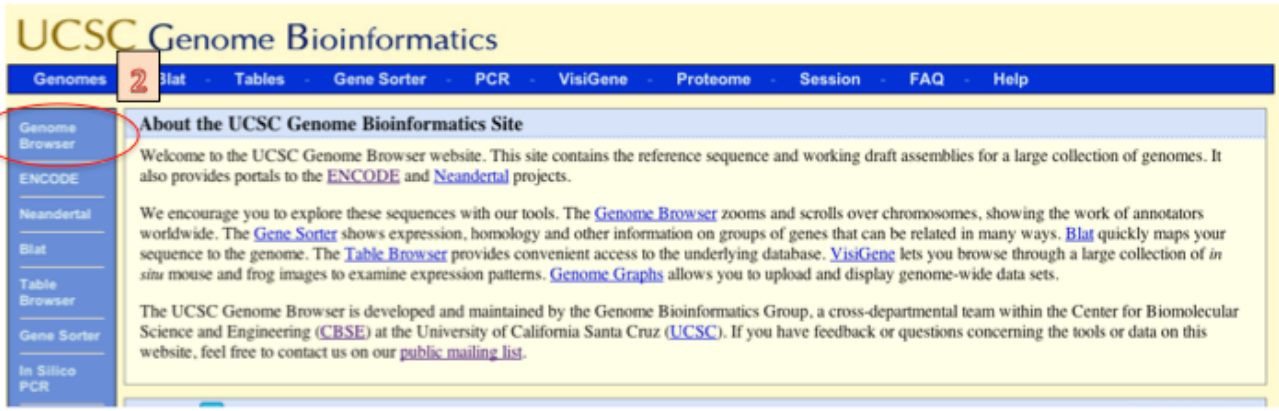
assemblies. We also support user data upload and visualization, and offer a data-hub mechanism allowing visualization of user data hosted remotely.

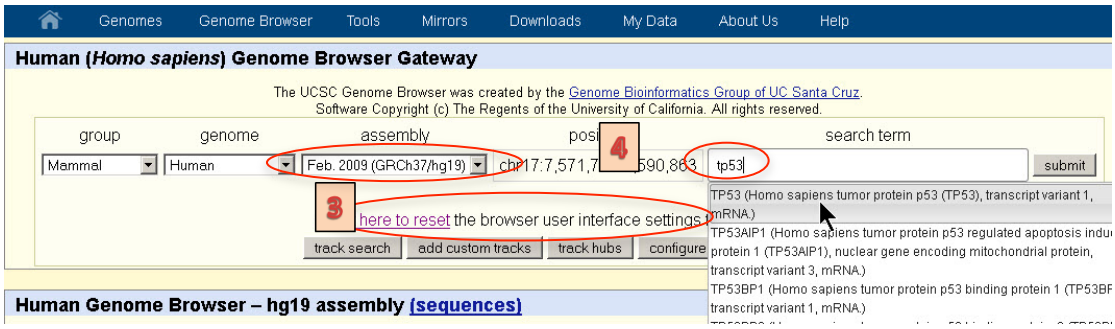
Step-by-step screenshots for the Exercises.

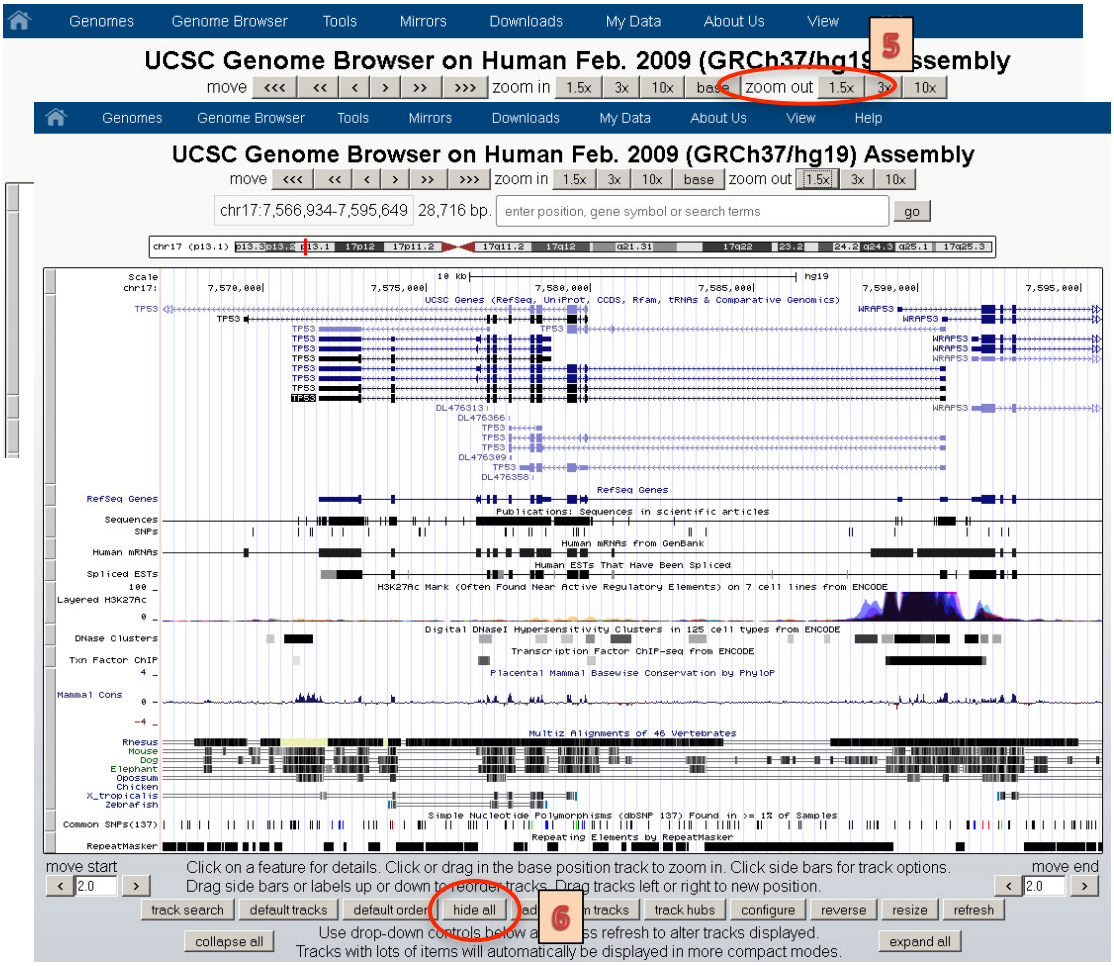
Worked Example 1:

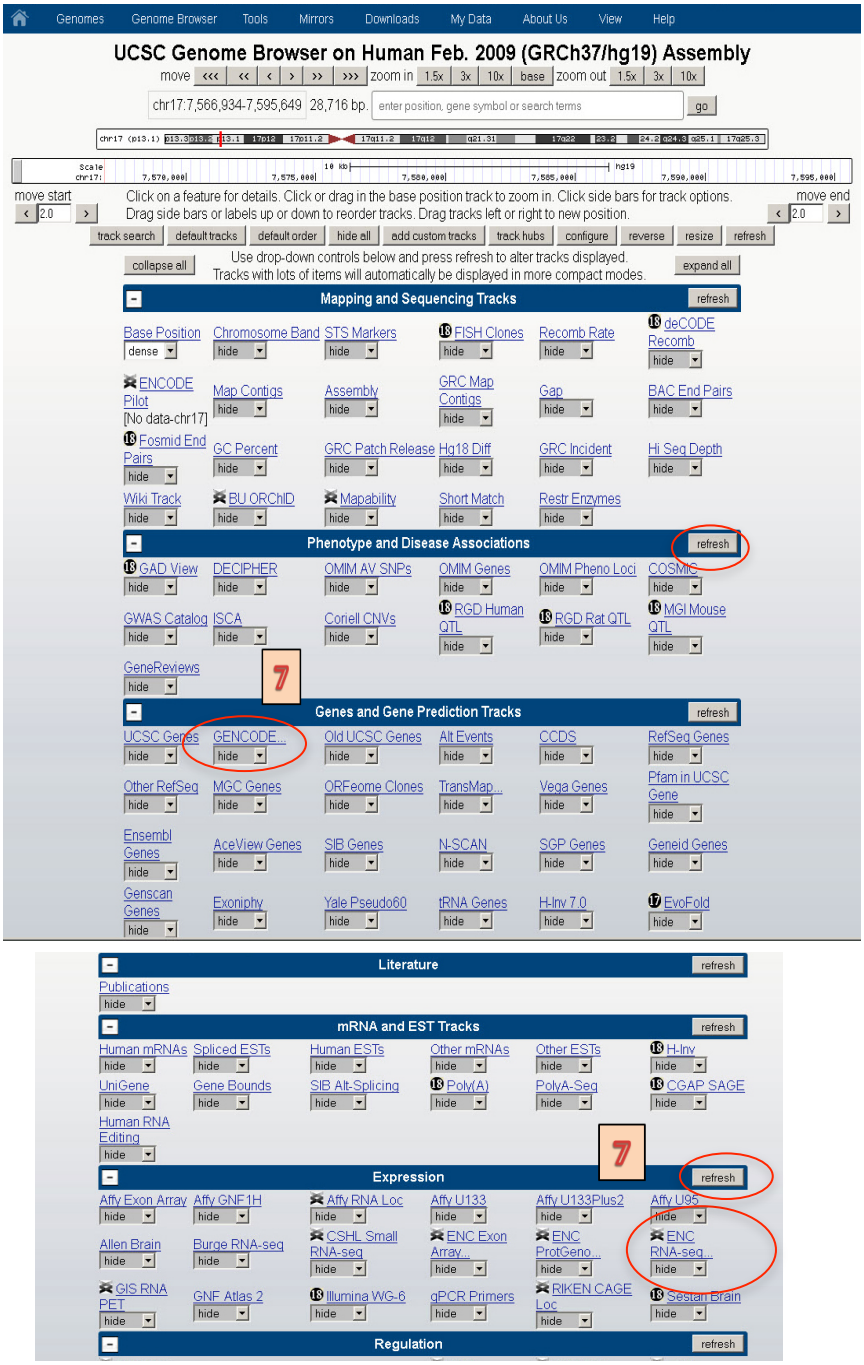
Examining RNA expression in the vicinity of the TP53 gene

Browse to genome.ucsc.edu.



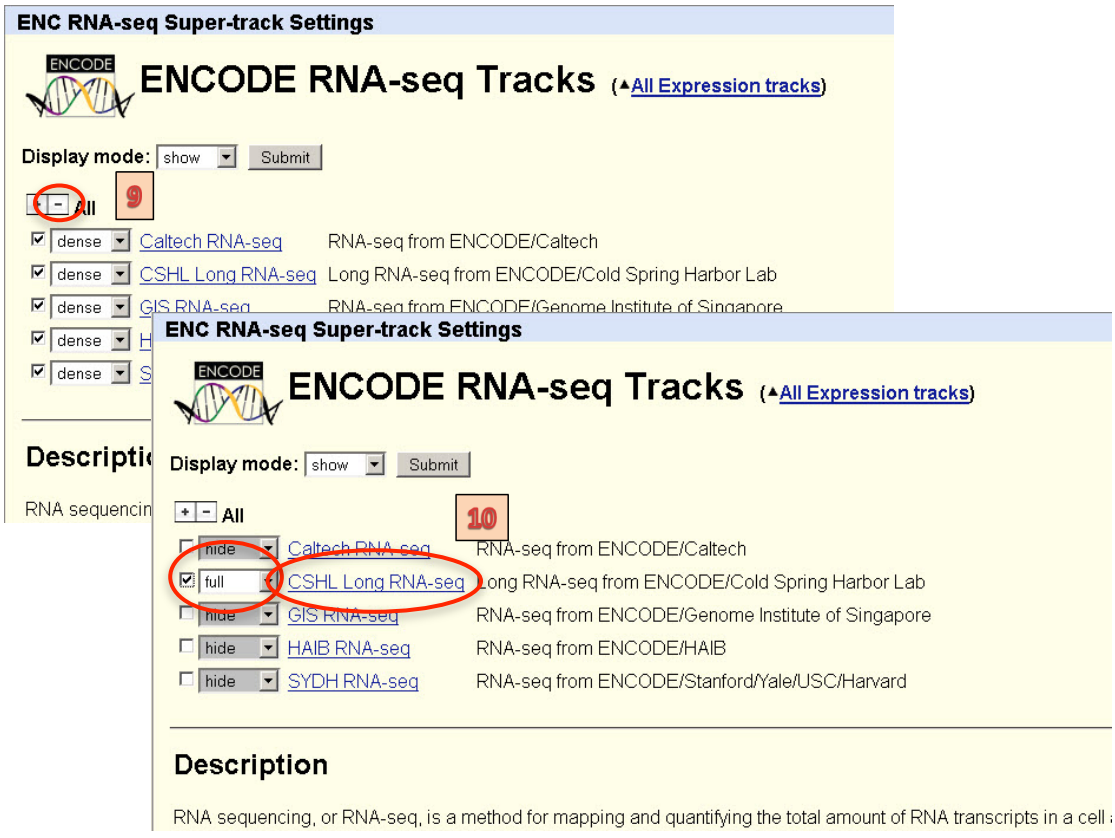


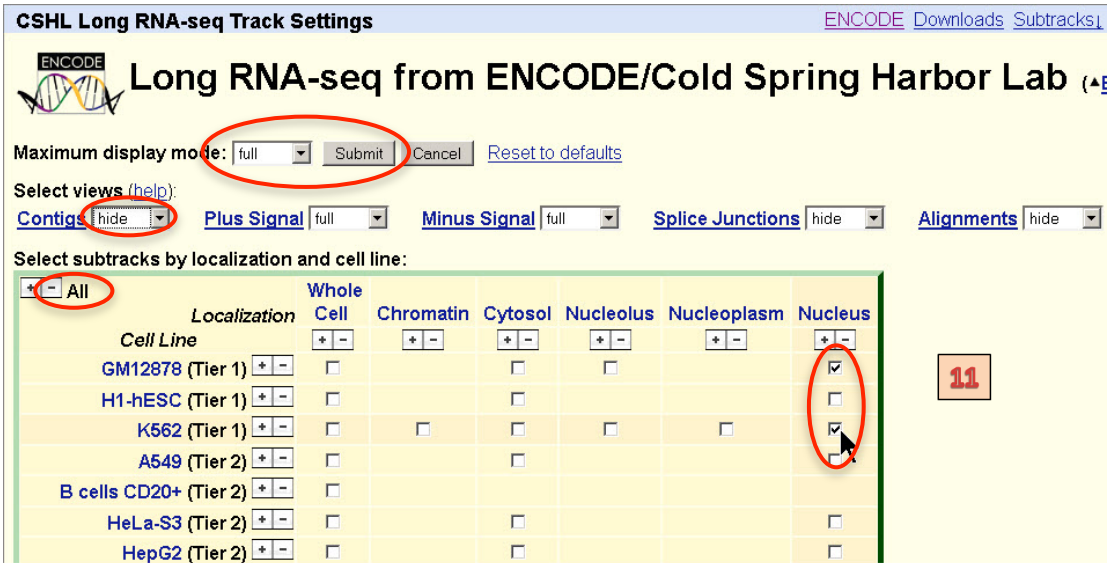

Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
About Us
View
Help
5
UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly
move
zoom in
zoom out
base
chr17:7,566,934-7,595,649 28,716 bp.
enter position, gene symbol or search terms
go
RefSeq Genes
Sequences
SNPs
Human mRNAs
Spliced ESTs
Layered H3K27Ac
DNase Clusters
Txn Factor ChIP
Mammal Cons
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish
Common SNPs(137)
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 tracks left or right to new position.
track search
default tracks
default order
hide all
track hubs
configure
reverse
resize
refresh
6
collapse all
expand all
Tracks with lots of items will automatically be displayed in more compact modes.

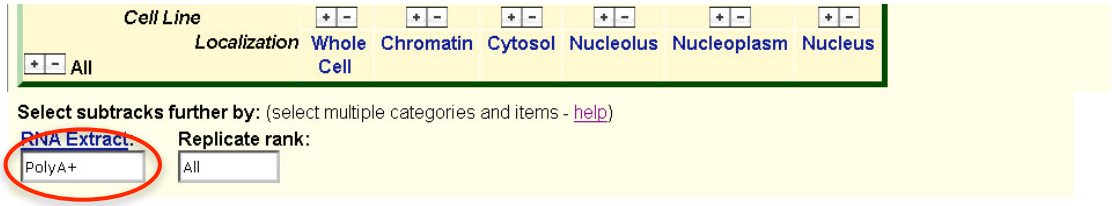

Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
About Us
View
Help
UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly
move
zoom in
zoom out
chr17:7,566,934-7,595,649 28,716 bp.
enter position, gene symbol or search terms
go
Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options.
Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.
move start
move end
track search
default tracks
default order
hide all
add custom tracks
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
Mapping and Sequencing Tracks
Base Position
Chromosome Band
STS Markers
FISH Clones
Recomb Rate
deCODE Recomb
ENCODE Pilot
[No data-chr17]
Map Contigs
Assembly
GRC Map Contigs
Gap
BAC End Pairs
Fosmid End Pairs
GC Percent
GRC Patch Release
Hg18 Diff
GRC Incident
Hi Seq Depth
Wiki Track
BU ORChID
Mapability
Short Match
Restr Enzymes
Phenotype and Disease Associations
GAD View
DECIPHER
OMIM AV SNPs
OMIM Genes
OMIM Pheno Loci
COSMIC
GWAS Catalog
ISCA
Coriell CNVs
RGD Human QTL
RGD Rat QTL
MGI Mouse QTL
GeneReviews
7
Genes and Gene Prediction Tracks
UCSC Genes
GENCODE...
Old UCSC Genes
Alt Events
CCDS
RefSeq Genes
Other RefSeq
MGC Genes
ORFeome Clones
TransMap...
Vega Genes
Pfam in UCSC Gene
Ensembl Genes
AceView Genes
SIB Genes
N-SCAN
SGP Genes
Geneid Genes
Genscan Genes
Exoniphy
Yale Pseudo60
tRNA Genes
H-Inv 7.0
EvoFold

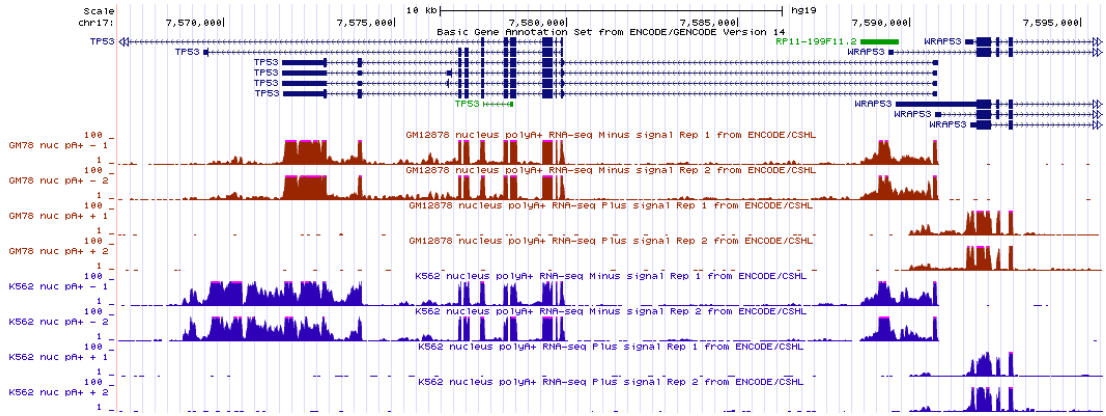
Literature
Publications
mRNA and EST Tracks
Human mRNAs
Spliced ESTs
Human ESTs
Other mRNAs
Other ESTs
H-Inv
UniGene
Gene Bounds
SIB Alt-Splicing
Poly(A)
PolyA-Seq
CGAP SAGE
Human RNA Editing
7
Expression
Affy Exon Array
Affy GNF1H
Affy RNA Loc
Affy U133
Affy U133Plus2
Affy U95
Allen Brain
Burge RNA-seq
CSHL Small RNA-seq
ENC Exon Array...
ENC ProtGeno...
ENC RNA-seq...
GIS RNA PET
GNF Atlas 2
Illumina WG-6
qPCR Primers
RIKEN CAGE Loc
Sestan Brain
Regulation


Basic Gene Annotation Set from ENCODE/GENCODE Version 14
RNA-seq from ENCODE/Caltech
Long RNA-seq from ENCODE/Cold Spring Harbor Lab
RNA-seq from ENCODE/Genome Institute of Singapore
RNA-seq from ENCODE/HAIB
Expression
refresh
Affy Exon Array
Affy GNF1H
Affy RNA Loc
Affy U133
Affy U133Plus2
Affy U95
Allen Brain
Burge RNA-seq
CSHL Small RNA-seq
ENC Exon Array...
ENC ProtGeno...
ENC RNA-seq...
GIS RNA PET
GNF Atlas 2
Illumina WG-6
qPCR Primers
RIKEN CAGE Loc
Sestan Brain
8


ENC RNA-seq Super-track Settings
ENCODE RNA-seq Tracks
(▲All Expression tracks)
Display mode: show Submit
All
9
dense Caltech RNA-seq RNA-seq from ENCODE/Caltech
dense CSHL Long RNA-seq Long RNA-seq from ENCODE/Cold Spring Harbor Lab
dense GIS RNA-seq RNA-seq from ENCODE/Genome Institute of Singapore
Description
RNA sequencing
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 GIS RNA-seq RNA-seq from ENCODE/Genome Institute of Singapore
hide HAIB RNA-seq RNA-seq from ENCODE/HAIB
hide SYDH RNA-seq RNA-seq from ENCODE/Stanford/Yale/USC/Harvard
Description
RNA sequencing, or RNA-seq, is a method for mapping and quantifying the total amount of RNA transcripts in a cell

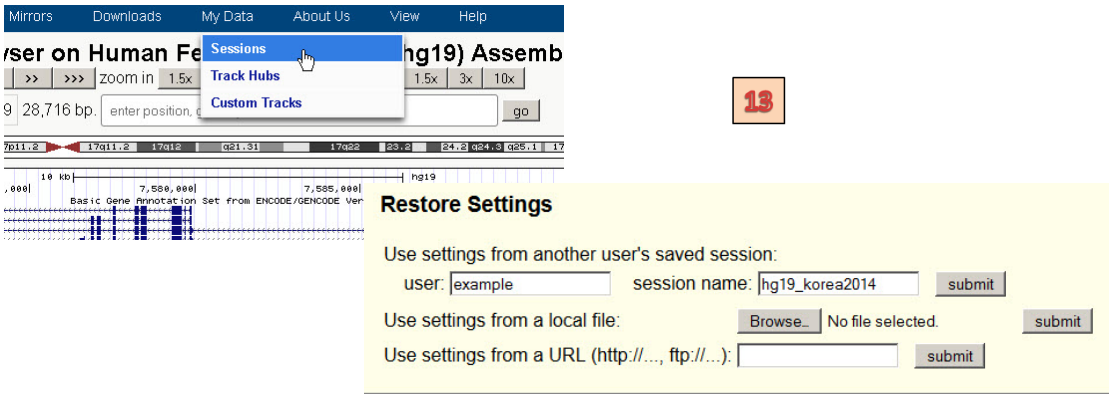

CSHL Long RNA-seq Track Settings
ENCODE Downloads Subtracks
Long RNA-seq from ENCODE/Cold Spring Harbor Lab
Maximum display mode: full Submit Cancel Reset to defaults
Select views (help):
Contigs hide Plus Signal full Minus Signal full Splice Junctions hide Alignments hide
Select subtracks by localization and cell line:
All
Localization
Cell Line
Whole Cell
Chromatin
Cytosol
Nucleolus
Nucleoplasm
Nucleus
GM12878 (Tier 1)
H1-hESC (Tier 1)
K562 (Tier 1)
A549 (Tier 2)
B cells CD20+ (Tier 2)
HeLa-S3 (Tier 2)
HepG2 (Tier 2)
11


Cell Line
Localization
All
Whole Cell
Chromatin
Cytosol
Nucleolus
Nucleoplasm
Nucleus
Select subtracks further by: (select multiple categories and items - help)
RNA Extract:
PolyA+
Replicate rank:
All

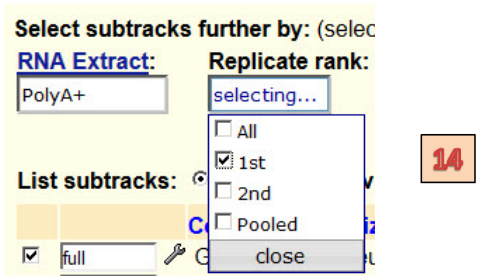


12

This view can be obtained directly using the session tool.
Under “My Data, Sessions” … “Restore Settings” :
user: example
session name: hg19_korea2014



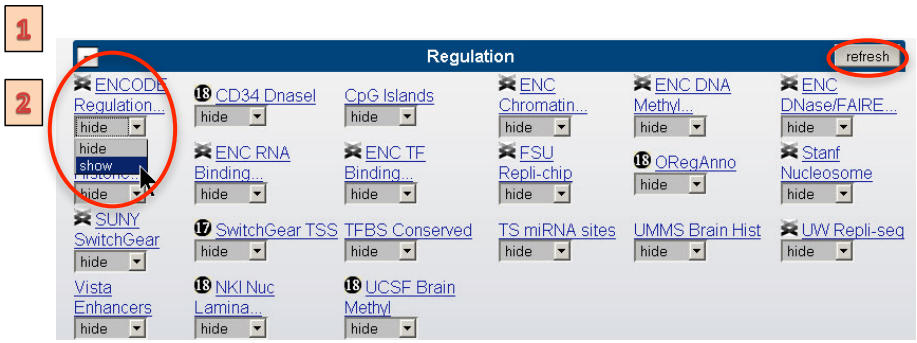
13



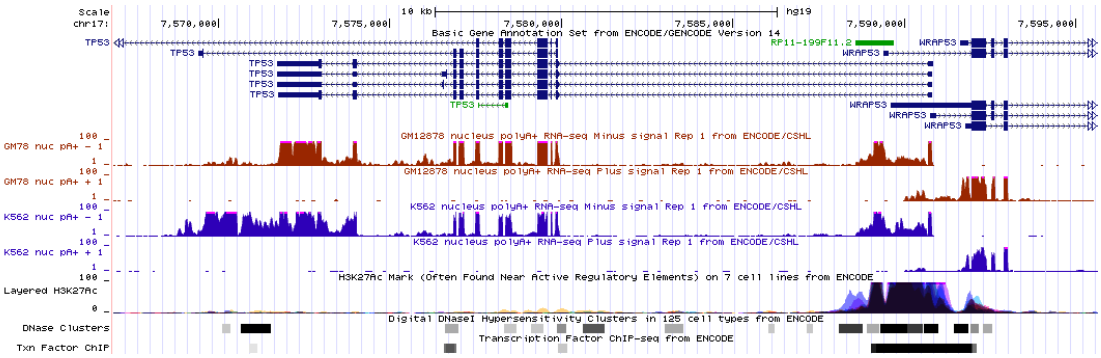
14

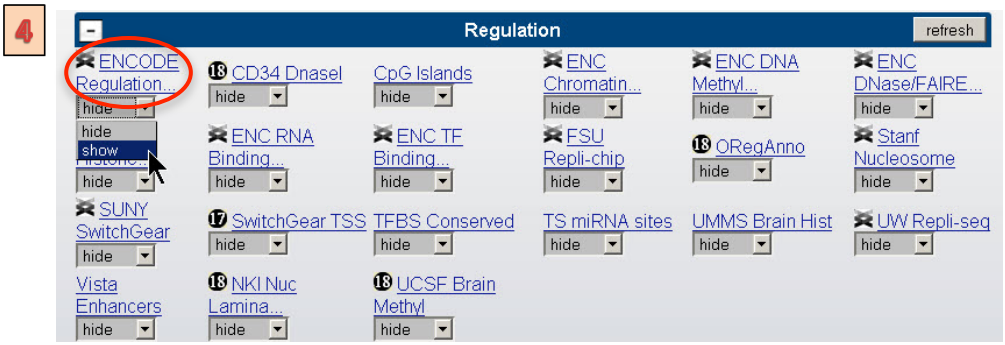
Worked Example 2:

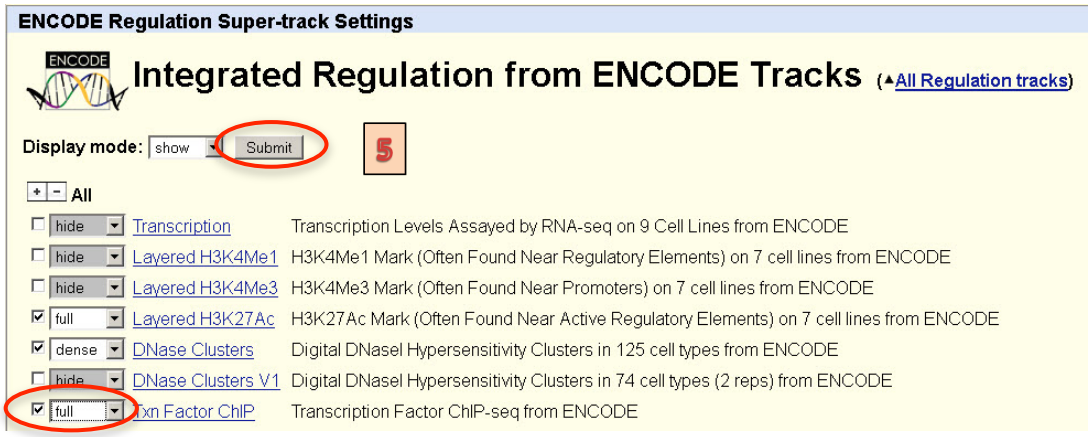
Exploring TFBS and Histone Marks in the TP53 region

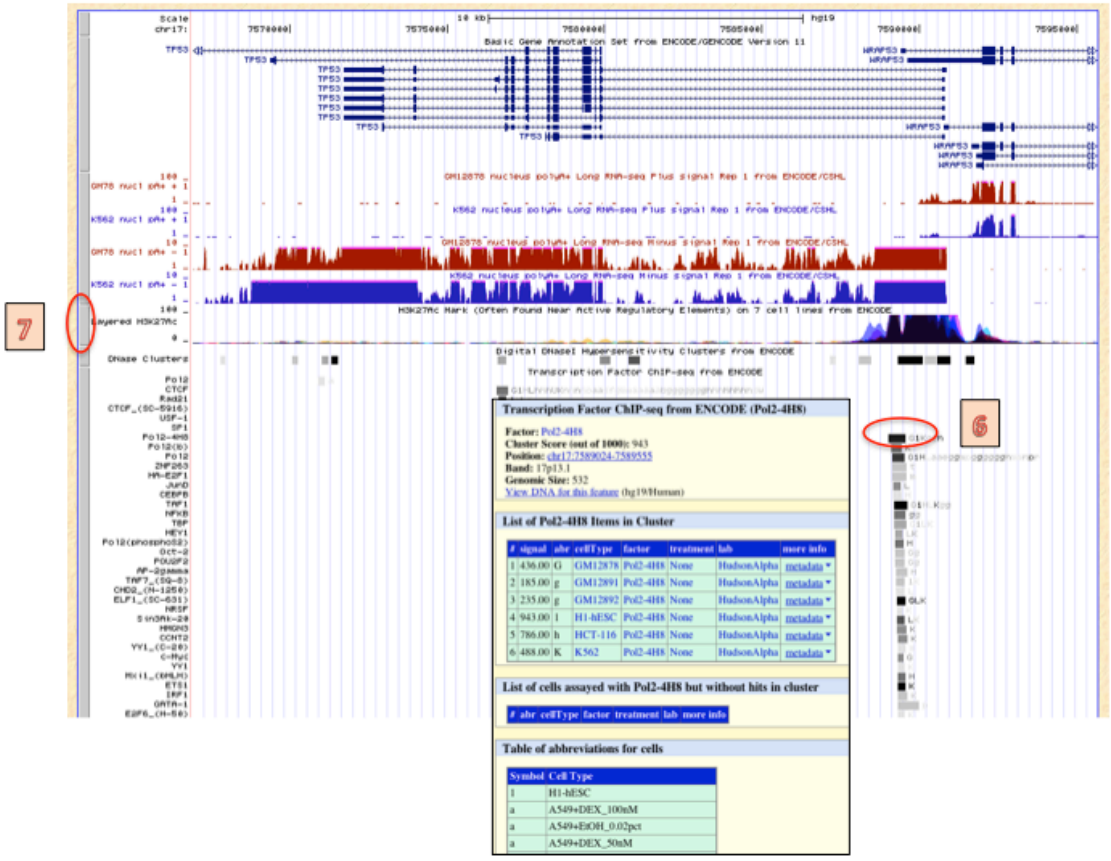

1
2
Regulation
refresh
ENCODE Regulation...
hide
show
CD34 Dnasel
CpG Islands
ENC Chromatin...
ENC DNA Methyl...
ENC DNase/FAIRE...
ENC RNA Binding...
ENC TF Binding...
FSU Repli-chip
ORegAnno
Stanf Nucleosome
SUNY SwitchGear
SwitchGear TSS
TFBS Conserved
TS miRNA sites
UMMS Brain Hist
UW Repli-seq
Vista Enhancers
NKI Nuc Lamina...
UCSF Brain Methyl

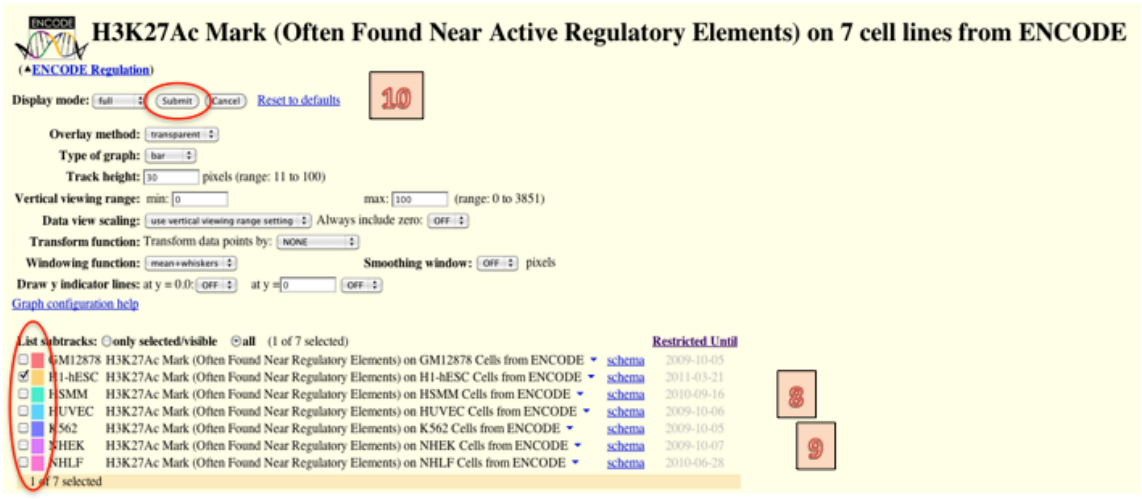
3

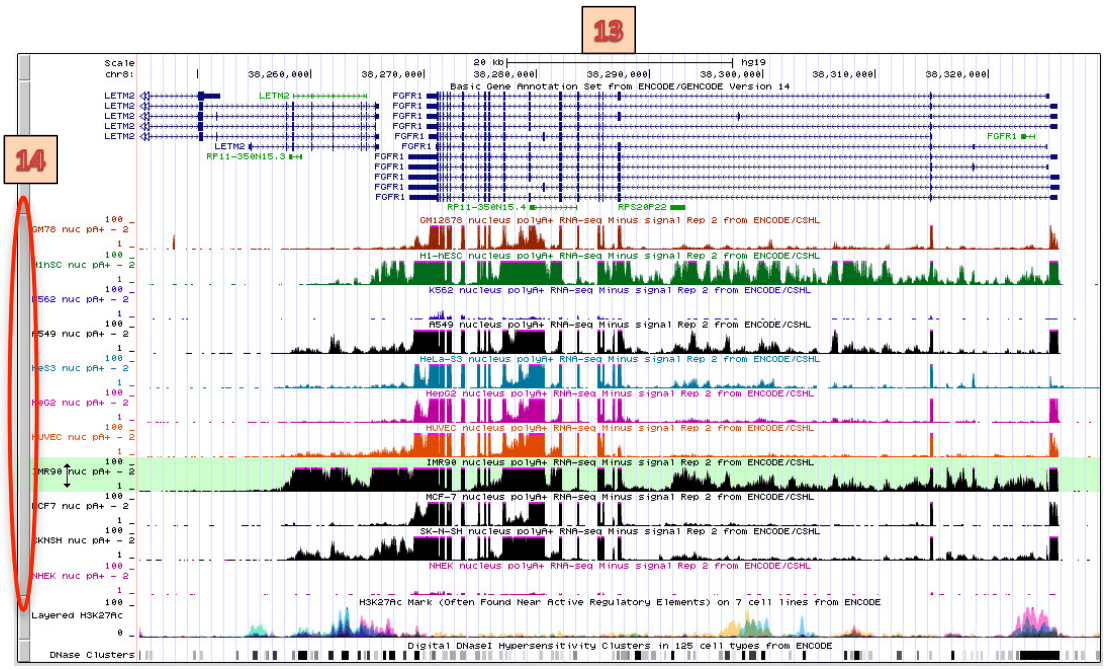

Basic Gene Annotation Set from ENCODE/GENCODE Version 14
GM12878 nucleus polyA+ RNA-seq Minus signal Rep 1 from ENCODE/CSHL
GM12878 nucleus polyA+ RNA-seq Plus signal Rep 1 from ENCODE/CSHL
K562 nucleus polyA+ RNA-seq Minus signal Rep 1 from ENCODE/CSHL
K562 nucleus polyA+ RNA-seq Plus signal Rep 1 from ENCODE/CSHL
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Digital DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE
Transcription Factor ChIP-seq from ENCODE

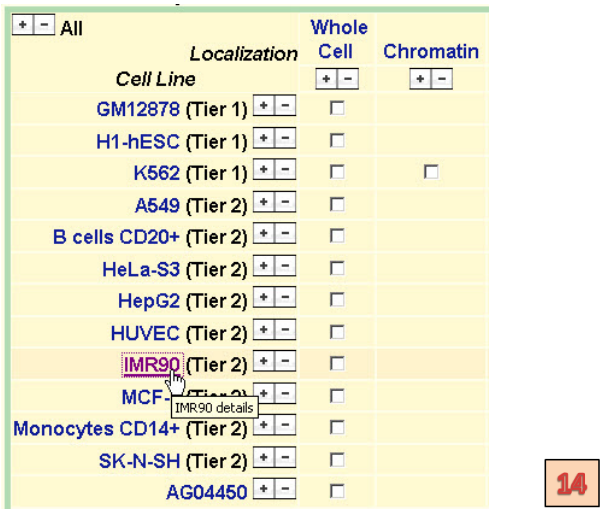

4
Regulation
refresh
ENCODE Regulation...
hide
show
CD34 Dnasel
CpG Islands
ENC Chromatin...
ENC DNA Methyl...
ENC DNase/FAIRE...
ENC RNA Binding...
ENC TF Binding...
FSU Repli-chip
ORegAnno
Stanf Nucleosome
SUNY SwitchGear
SwitchGear TSS
TFBS Conserved
TS miRNA sites
UMMS Brain Hist
UW Repli-seq
Vista Enhancers
NKI Nuc Lamina...
UCSF Brain Methyl


ENCODE Regulation Super-track Settings
Integrated Regulation from ENCODE Tracks
All Regulation tracks
Display mode: show Submit
5
All
hide Transcription Transcription Levels Assayed by RNA-seq on 9 Cell Lines from ENCODE
hide Layered H3K4Me1 H3K4Me1 Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
hide Layered H3K4Me3 H3K4Me3 Mark (Often Found Near Promoters) on 7 cell lines from ENCODE
full Layered H3K27Ac H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
dense DNase Clusters Digital DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE
hide DNase Clusters V1 Digital DNaseI Hypersensitivity Clusters in 74 cell types (2 reps) from ENCODE
full Txn Factor ChIP Transcription Factor ChIP-seq from ENCODE


7
6
Transcription Factor ChIP-seq from ENCODE (Pol2-4H8)
Factor: Pol2-4H8
Cluster Score (out of 1000): 943
Position: chr17:7589024-7589555
Band: 17p13.1
Genomic Size: 532
View DNA for this feature (hg19/Human)
List of Pol2-4H8 Items in Cluster
signal abr cellType factor treatment lab more info
1 436.00 G GM12878 Pol2-4H8 None HudsonAlpha metadata
2 185.00 g GM12891 Pol2-4H8 None HudsonAlpha metadata
3 235.00 g GM12892 Pol2-4H8 None HudsonAlpha metadata
4 943.00 1 H1-hESC Pol2-4H8 None HudsonAlpha metadata
5 786.00 h HCT-116 Pol2-4H8 None HudsonAlpha metadata
6 488.00 K K562 Pol2-4H8 None HudsonAlpha metadata
List of cells assayed with Pol2-4H8 but without hits in cluster
abr cellType factor treatment lab more info
Table of abbreviations for cells
Symbol Cell Type
1 H1-hESC
a A549+DEX_100nM
a A549+EtOH_0.02pct
a A549+DEX_50nM


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
12





14

Cell, tissue or DNA sample. Cell line or tissue used as the source of experimental material.

cell	Tier	Description	Lineage	Tissue	Karyotype	Sex	Documents	Vendor ID	Term ID	*Label*
IMR90	2	fetal lung fibroblasts, newly promoted to tier 2: not in 2011 analysis	endoderm	lung	normal	F	Stam	ATCC CCL-186	BTO:0001229	IMR90