

Module 4: Working with ENCODE Data

Aim

Learn how to explore data from the ENCODE (Encyclopedia of DNA Elements) project using:

- the ENCODE portal
- the ENCODE Roadmap Browser and the IHEC Data Portal
- the UCSC Genome Browser
- the Ensembl Genome Browser

Introduction

The ENCODE (Encyclopedia of DNA Elements) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

ENCODE investigators employ a variety of assays and methods to identify functional elements. The discovery and annotation of gene elements is accomplished primarily by sequencing a diverse range of RNA sources, comparative genomics, integrative bioinformatic methods, and human curation. Regulatory elements are typically investigated through DNA hypersensitivity assays, assays of DNA methylation, and immunoprecipitation (IP) of proteins that interact with DNA and RNA, i.e., modified histones, transcription factors, chromatin regulators, and RNA-binding proteins, followed by sequencing.

Data from the ENCODE project can be accessed in a variety of ways.

The ENCODE portal

The primary source for data and information about the ENCODE project is the ENCODE portal at <https://www.encodeproject.org/>. The portal contains tools for browsing and searching data generated by the ENCODE consortium via assays, biological samples, and experimental reagents used.

Worked example 1: the ENCODE portal

In this worked example we will look whether there are any ENCODE data sets available containing ChIP-seq data for human kidney tissue.

(1) Go to the ENCODE website (<https://www.encodeproject.org>).

The ENCODE (Encyclopedia of DNA Elements) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

Image credits: Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

Data

To find and download ENCODE Consortium data:

- Click the Data toolbar above and browse data
 - By assay
 - By biosample
- Enter search terms like "skin", "ChIP-seq", or "CTCF"

News

Sept 12, 2014: Data release: 23 human and 5 mouse datasets. [\[read more\]](#)

August 28, 2014: modENCODE and ENCODE [comparison papers](#) published. [\[read more\]](#)

August 19, 2014: New ENCODE portal released. The portal contains tools for browsing and searching data generated by the ENCODE consortium via assays, biological samples, and experimental reagents used. [\[read more\]](#)

(2) Click on the “Data” drop-down menu in the toolbar.

Data can be browsed via assays, biosamples, and antibodies used.

(3) Select “Assays”.

The screenshot shows the ENCODE Data Browser interface. On the left, there are several facets for filtering data: Assay, Experiment status, Genome assembly, Organism, Biosample type, and Organ. The main content area shows a list of experiments, with the first five visible. The filters on the left are set to show all data.

Facet	Value	Count
Assay	ChIP-seq	2377
	RNA-seq	527
	DNase-seq	265
	RNA profiling by array assay	180
	shRNA knockdown followed by RNA-seq	167
	+ See more...	
Experiment status	released	4257
	revoked	4
Genome assembly	hg19	2542
	mm9	560
Organism	<i>Homo sapiens</i>	3379
	<i>Mus musculus</i>	854
Biosample type	immortalized cell line	2477
	primary cell	767
	tissue	650
	stem cell	206
	in vitro differentiated cells	122
Organ	brain	200
	skin of body	165
	blood vessel	109
	lung	87
	liver	76
	+ See more...	

Showing 25 of 4261

Filter to 500 to visualize View All

Experiment	Target	Lab	Project
ChIP-seq of K562 (<i>Homo sapiens</i> , adult 53 year)	Control	Richard Myers, HAIB	ENCSR996HUG released
ChIP-seq of K562 (<i>Homo sapiens</i> , adult 53 year)	Control	Richard Myers, HAIB	ENCSR996HUG released
ChIP-seq of K562 (<i>Homo sapiens</i> , adult 53 year)	Control	Richard Myers, HAIB	ENCSR996HUG released
ChIP-seq of K562 (<i>Homo sapiens</i> , adult 53 year)	Control	Richard Myers, HAIB	ENCSR996HUG released
ChIP-seq of K562 (<i>Homo sapiens</i> , adult 53 year)	Control	Richard Myers, HAIB	ENCSR996HUG released
ChIP-seq of GM12878 (<i>Homo sapiens</i> , adult 53 year)	Control	Richard Myers, HAIB	ENCSR996HUG released
ChIP-seq of GM12878 (<i>Homo sapiens</i> , adult 53 year)	Control	Richard Myers, HAIB	ENCSR996HUG released

The "Assays" page lists all assays that have been used to generate ENCODE data. The results can be narrowed and filtered by selecting one or more values in a metadata category on the left hand side of the page. Multiple values from each facet can be selected at any one time.

(4) Select "ChIP-seq", "Homo sapiens", "tissue" and "kidney".

The screenshot shows the ENCODE Data Browser interface with filters applied. The filters on the left are: Assay (ChIP-seq), Organism (Homo sapiens), Biosample type (tissue), and Organ (kidney). The main content area shows a list of experiments, with the first two visible.

Facet	Value	Count
Assay	ChIP-seq	2
	DNA methylation profiling by array	2
	RNA-seq	2
	FAIRE-seq	1
	RAMPAGE	1
Experiment status	released	2
Genome assembly	hg19	2
Organism	<i>Mus musculus</i>	24
	<i>Homo sapiens</i>	2
Biosample type	primary cell	8
	tissue	2
Organ	kidney	2
	lung	2
	pancreas	2
	spleen	2
	heart	1
	+ See more...	

Showing 2 of 2

Visualize

Experiment	Target	Lab	Project
ChIP-seq of kidney (<i>Homo sapiens</i> , adult)	CTCF	Vishwanath Iyer, UTA	ENCSR000DMC released
ChIP-seq of kidney (<i>Homo sapiens</i> , adult 27 year)	Control	Vishwanath Iyer, UTA	ENCSR000DMD released

The results show that there are two datasets that match our search criteria, one containing CTCF binding data, and a control dataset. CTCF is a transcriptional repressor (<http://en.wikipedia.org/wiki/CTCF>).

(5) Click on the link for the CTCF dataset, “ChIP-seq of kidney (*Homo sapiens*, adult)”.

Experiment / ChIP-seq

Experiment summary for ENCSR000DMC

Status: released

Assay: ChIP-seq

Accession: ENCSR000DMC

Biosample summary: kidney (*Homo sapiens*, adult 35 year and adult 27 year and adult 22 year)

Type: tissue

Target: [CTCF](#)

Antibody: [ENCAB000AXY](#)

Controls: [ENCSR000DMD](#)

Description: CTCF ChIP-seq on human kidney

Lab: Vishwanath Iyer, UTA

Project: ENCODE

External resources: [UCSC-ENCODE-hg19:wgEncodeEH003461](#) [GEO:GSM1006886](#)

Date released: 2012-08-27

Assay details

Nucleic acid type: DNA

Lysis method: see document

Extraction method: see document

Fragmentation method: see document

Size selection method: see document

Biological replicate - 1

Technical replicate: 1

Library: ENCLB141DZT

Biosample: [ENCBS136AAA](#) - kidney

Biological replicate - 2

Technical replicate: 1

Library: ENCLB358OSR

Biosample: [ENCBS349AAA](#) - kidney

Biological replicate - 3

Technical replicate: 1

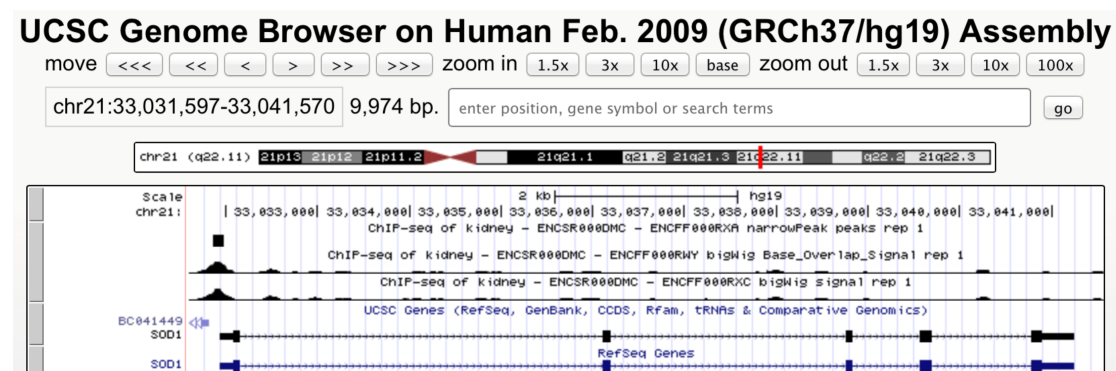
Library: ENCLB189BJC

Biosample: [ENCBS137AAA](#) - kidney

Details about the dataset are shown. At the bottom of the page data can be downloaded in various formats. Clicking on [Visualize Data] launches a UCSC Genome Browser view.

Accession	File type	Output type	Paired end	Associated replicates	Added by	Date added	File download
ENCFF000RWW	bam	alignments		(2, 1)	Gregory Crawford	2012-07-26	Download
ENCFF000RWX	bam	alignments		(1, 1)	Gregory Crawford	2012-07-26	Download
ENCFF000RWY	bigWig	Base_Overlap_Signal		(1, 1)	Gregory Crawford	2012-07-26	Download
ENCFF000RWZ	bam	alignments		(3, 1)	Gregory Crawford	2012-07-26	Download
ENCFF000RXA	narrowPeak	peaks		(1, 1)	Gregory Crawford	2012-07-26	Download
ENCFF000RXC	bigWig	signal		(1, 1)	Gregory Crawford	2012-07-26	Download
ENCFF000RXE	fastq	reads		(1, 1)	Gregory Crawford	2012-07-26	Download
ENCFF000RXM	fastq	reads		(2, 1)	Gregory Crawford	2012-07-26	Download
ENCFF000RXV	fastq	reads		(3, 1)	Gregory Crawford	2012-07-26	Download
ENCFF001UTG	bed_narrowPeak	peaks		(1, 1)	Gregory Crawford	2012-07-26	Download

(6) Click on [Visualize Data].



Tracks containing the kidney CTCF ChIP-seq data have been added to the browser view.

Data can also be searched using the “Search ENCODE” search box present in the tool bar on the ENCODE portal pages.

(8) Go back to <https://www.encodeproject.org>.

(9) Type “ChIP-seq human tissue kidney” in the “Search ENCODE” search box”.

ENCODE Data Methods About ENCODE Help

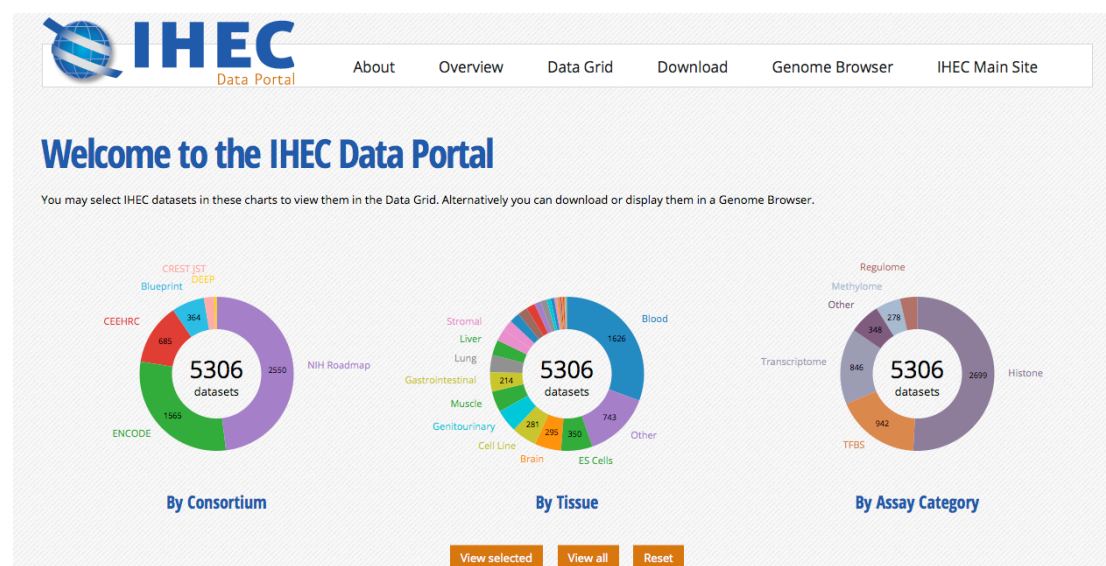
ChIP-seq human tissue Q Sign in

Data Type
Experiments 2

Showing 2 of 2

ChIP-seq of kidney (*Homo sapiens*, adult)
Target: CTCF
Lab: Vishwanath Iyer, UTA
Project: ENCODE
Experiment: ENCSR000DMC released

ChIP-seq of kidney (*Homo sapiens*, adult 27 year)
Target: Control
Lab: Vishwanath Iyer, UTA
Project: ENCODE
Experiment: ENCSR000DMC released



The UCSC Genome Browser

UCSC coordinated data for the ENCODE Consortium from its inception in 2003 (Pilot phase) to the end of the first 5 year phase of whole-genome data production in 2012. All data produced by ENCODE investigators and the results of ENCODE analysis projects from this period are hosted in the UCSC Genome Browser and database.

All the ENCODE data that are hosted as browser tracks in the UCSC Genome Browser are visually summarised in the ENCODE Experiment Matrix (<http://genome.ucsc.edu/ENCODE/dataMatrix/encodeDataMatrixHuman.html>).

Worked example 2: the UCSC Genome Browser

In this worked example we will explore the region of the *TP53* (Tumor protein p53) gene for transcription factor binding data and histone marks that are often found near active regulatory elements. We will also determine if these histone marks are indicated in human embryonic stem cells.

(1) Go to the UCSC Genome Browser homepage (<http://genome.ucsc.edu/>).

UCSC

Genome Bioinformatics

[Genomes](#) - [Blat](#) - [Tables](#) - [Gene Sorter](#) - [PCR](#) - [VisiGene](#) - [Session](#) - [FAQ](#) - [Help](#)

Genome Browser

Ebola

Blat

Table Browser

Gene Sorter

In Silico PCR

Genome Graphs

Galaxy

VisiGene

About the UCSC Genome Bioinformatics Site

Welcome to the UCSC Genome Browser website. This site contains the reference sequence and working draft assemblies for a large collection of genomes. It also provides portals to [ENCODE](#) data at UCSC (2003 to 2012) and to the [Neandertal](#) project. Download or purchase the Genome Browser source code, or the Genome Browser in a Box ([GBIB](#)) at our [online store](#).

We encourage you to explore these sequences with our tools. The [Genome Browser](#) zooms and scrolls over chromosomes, showing the work of annotators worldwide. The [Gene Sorter](#) shows expression, homology and other information on groups of genes that can be related in many ways. [Blat](#) quickly maps your sequence to the genome. The [Table Browser](#) provides convenient access to the underlying database. [VisiGene](#) lets you browse through a large collection of *in situ* mouse and frog images to examine expression patterns. [Genome Graphs](#) allows you to upload and display genome-wide data sets.

The UCSC Genome Browser is developed and maintained by the Genome Bioinformatics Group, a cross-departmental team within the [UC Santa Cruz Genomics Institute](#) and the Center for Molecular Science and Engineering ([CBSE](#)) at the University of California Santa Cruz ([UCSC](#)). If you have feedback or questions concerning the tools or data on this website, feel free to contact us on our [public mailing list](#).

The Genome Browser project team relies on public funding to support our work. Donations are welcome -- we have many more ideas than our funding supports! If you have ideas, drop a comment in our [suggestion box](#).

DONATE NOW

(2) From the blue navigation links on the left side of the page, click the “Genome Browser” link.

[Home](#)
[Genomes](#)
[Genome Browser](#)
[Tools](#)
[Mirrors](#)
[Downloads](#)
[My Data](#)
[Help](#)
[About Us](#)

Human (*Homo sapiens*) Genome Browser Gateway

The UCSC Genome Browser was created by the [Genome Bioinformatics Group of UC Santa Cruz](#).
 Software Copyright (c) The Regents of the University of California. All rights reserved.

group

genome

assembly

position

search term

Mammal

Human

Feb. 2009 (GRCh37/hg19)

chr21:33,031,597-33,041,570

enter position, gene symbol or search terms

submit

[Click here to reset](#) the browser user interface settings to their defaults.

track search

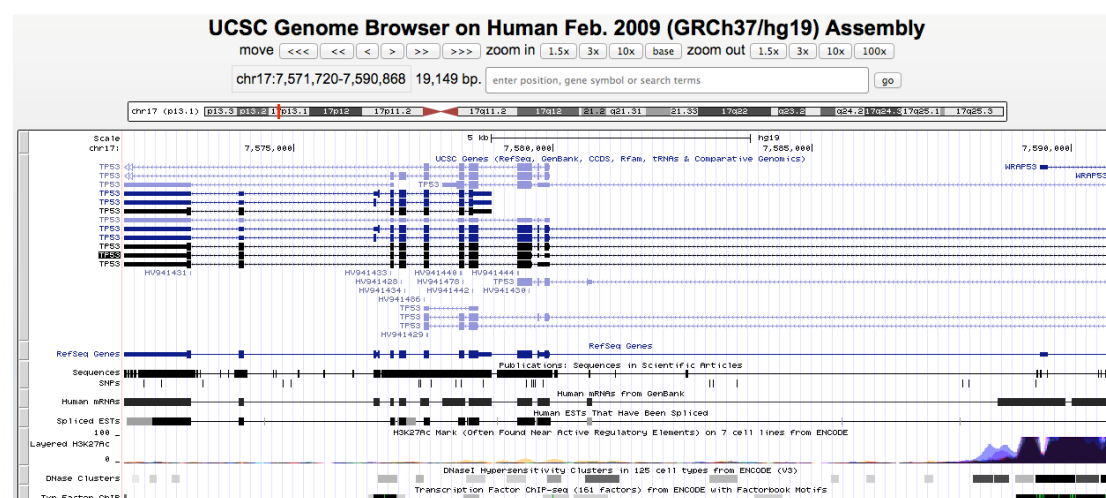
add custom tracks

track hubs

configure tracks and display

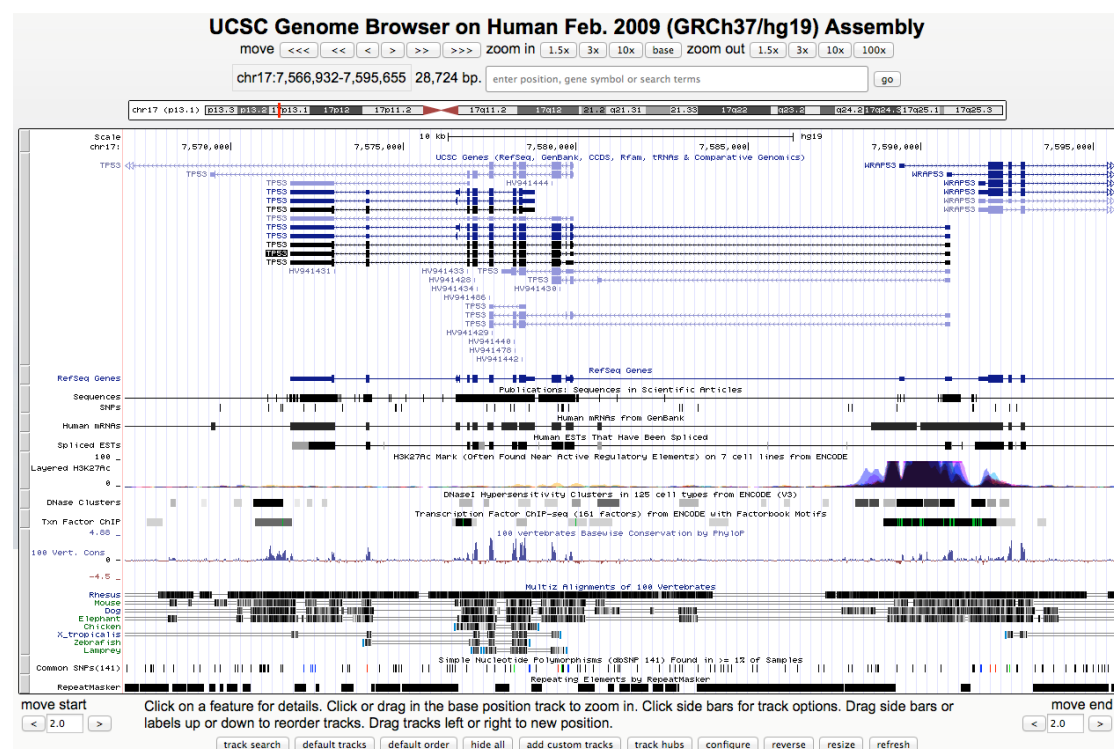
(3) On the “Human Genome Browser Gateway” interface, click the “Click here to reset the browser user interface settings to their defaults.” link. This will ensure that any prior activity on the browser has been cleared out and that everyone is starting with default settings.

(4) Choose the “Human” genome and the “Feb. 2009 (GRCh37/hg19)” assembly. Enter the text “tp53” in the “search term” box. Choose “TP53 (Homo sapiens tumor protein p53 (TP53), transcript variant 1, mRNA.)” from the resulting drop down list. Click the [submit] button.

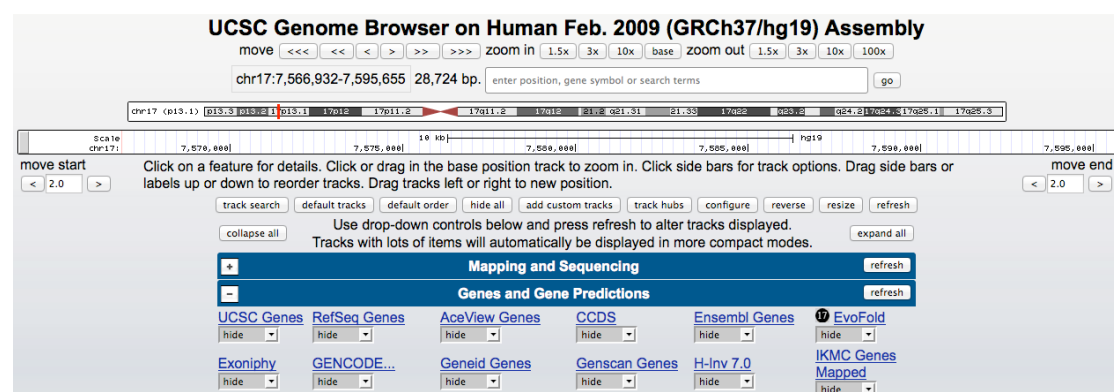


Note that we are using the GRCh37/hg19 assembly, because the ENCODE data haven't been mapped to the GRCh38/hg38 assembly.

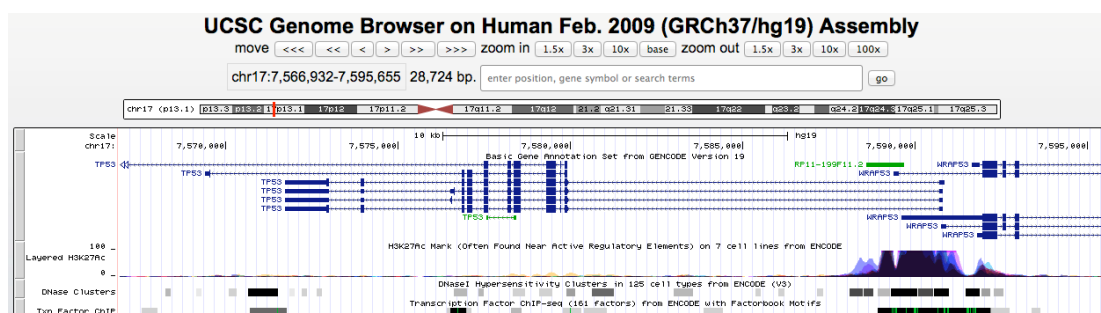
(5) In the TP53 region on the browser, examine the features briefly. Then click the “zoom out [1.5x]” button near the top. Assess the features again.



(6) Click the [hide all] button in the middle of the resulting 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.)

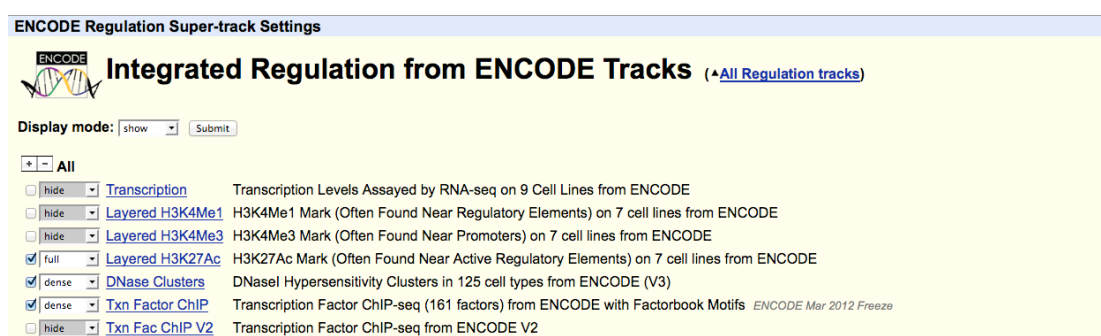


(7) Add the “GENCODE ...” track (from the “Genes and Gene Predictions” group) and the “ENCODE Regulation ...” track (from the “Regulation” group) by choosing “show” in the respective pull down menus and clicking a [refresh] button.

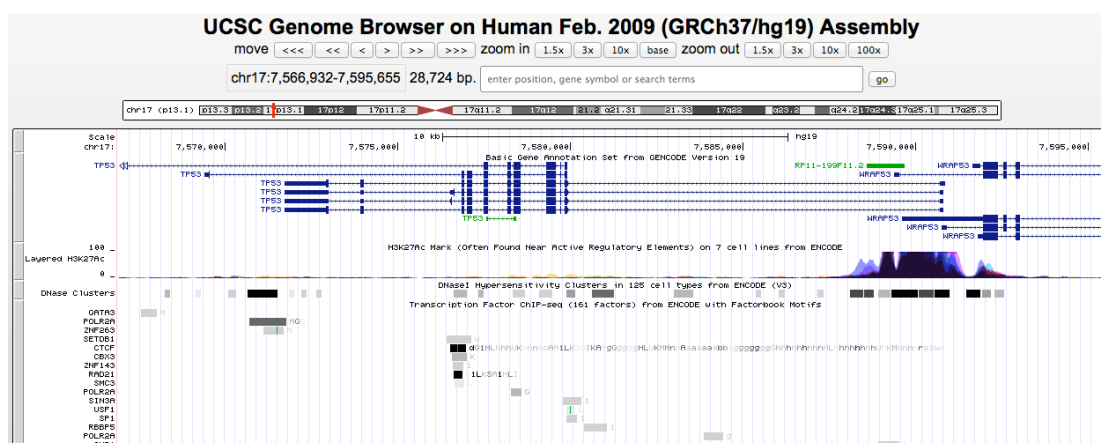


Examine the display. Note that the “Txn Factor ChIP” track shows data blocks, but not individual transcription factors. Also note that the “Layered H3K27Ac” track appears to contain multiple data sets of various colours.

(8) Click the “ENCODE Regulation ...” hyperlink (in the “Regulation” section) to look at the component tracks of this super-track.



(9) By default the “Txn Factor ChIP” track is visible in “dense” mode. Change this to “full”. Click the [Submit] button.



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.

(10) Click the grey bar to the left of the “Layered H3K27Ac” track to go to the controls for that track.

Layered H3K27Ac Track Settings [ENCODE](#) [Downloads](#) [Subtracks](#) [Description](#) [Contact](#)

H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE ([ENCODE Regulation](#))

Display mode:

Overlay method:

Type of graph:

Track height: pixels (range: 11 to 100)

Vertical viewing range: min: max: (range: 0 to 3851)

Data view scaling: Always include zero:

Transform function: Transform data points by:

Windowing function: Smoothing window: pixels

Negate values: ☐

Draw y indicator lines: at y = 0.0: at y = 0:

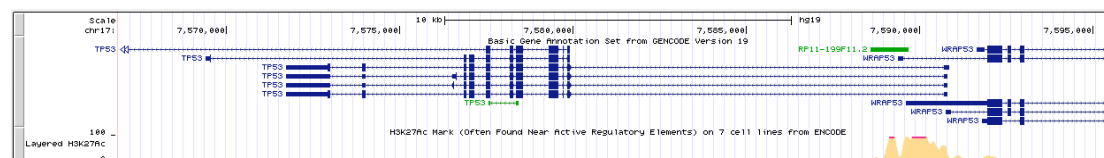
[Graph configuration help](#)

List subtracks: ☐ only selected/visible ☒ all (7 of 7 selected)

Cell Line	Track Name	Restricted Until
GM12878	H3K27Ac Mark (Often Found Near Active Regulatory Elements) on GM12878 Cells from ENCODE	2009-10-05
H1-hESC	H3K27Ac Mark (Often Found Near Active Regulatory Elements) on H1-hESC Cells from ENCODE	2011-03-21
HSMM	H3K27Ac Mark (Often Found Near Active Regulatory Elements) on HSMM Cells from ENCODE	2010-09-16
HUVEC	H3K27Ac Mark (Often Found Near Active Regulatory Elements) on HUVEC Cells from ENCODE	2009-10-06
K562	H3K27Ac Mark (Often Found Near Active Regulatory Elements) on K562 Cells from ENCODE	2009-10-05
NHEK	H3K27Ac Mark (Often Found Near Active Regulatory Elements) on NHEK Cells from ENCODE	2009-10-07
NHLF	H3K27Ac Mark (Often Found Near Active Regulatory Elements) on NHLF Cells from ENCODE	2010-06-28

On this histone mark page, note that there are various cell line data sets, which have colour codes. One of the lines is H1-hESC, which is a human embryonic stem cell line.

(11) Uncheck all cell line boxes except H1-hESC. Click the [Submit] button.



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 colour overlays.

(12) Return to the histone mark page by clicking the grey bar to the left of the “Layered H3K27Ac” track. Turn on or off various cell lines to view the data. Return to the viewer each time by clicking the [Submit] button.

The various data types in this region should help you to understand possible features of regulation of the genes in this area.

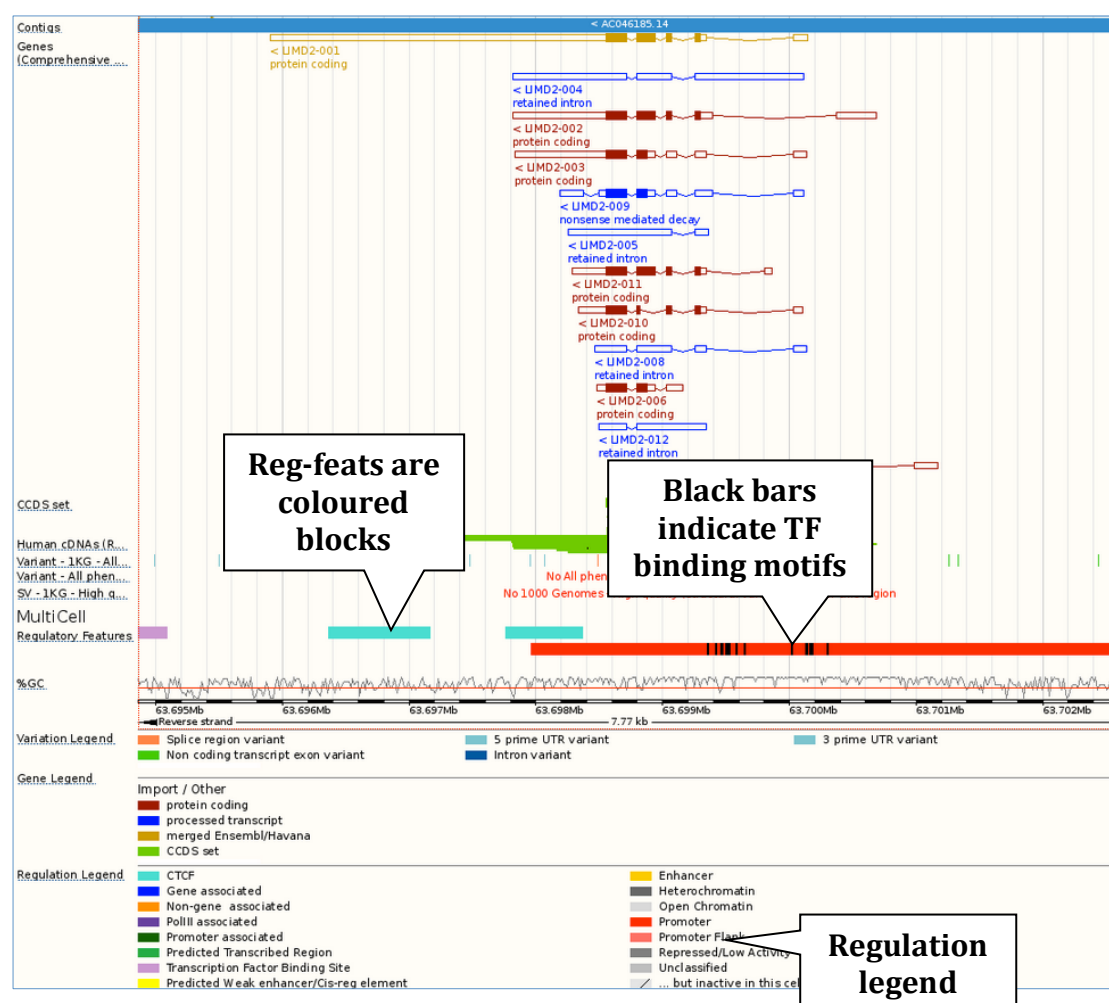
The Ensembl Genome Browser

The Ensembl project (<http://www.ensembl.org>) uses data from the ENCODE project, as well as data from other projects/publications, to predict sequences potentially involved in gene regulation. The regulatory features resulting from this Regulatory Build as well as the data on which they are based can be explored in the browser.

Worked example 3: the Ensembl Genome Browser

In this worked example we're going to have a look for regulatory features in the region of a gene and investigate their activity in different cell types.

We'll start by searching for the human *LIMD2* (LIM domain containing 2) gene on the Ensembl homepage (<http://www.ensembl.org>) and jumping to the "Location" tab. Zoom out a little to see the gene plus some of the flanking regions.



The "MultiCell Regulatory Features" are shown by default. In this region we can see a large red promoter, two turquoise CTCF binding sites and a lilac

transcription factor binding site (don't worry if you have zoomed out further or not as far and can see more/less). Refer to the legend at the bottom to see what the different colours mean.

You can also click on the regulatory features to learn more. Click on the red promoter to get a pop-up.

Regulatory Feature - MultiCell

Stable ID [ENSR00001537344](#)

Type Promoter

Core bp [17:63697964-63702563](#)

Attributes -

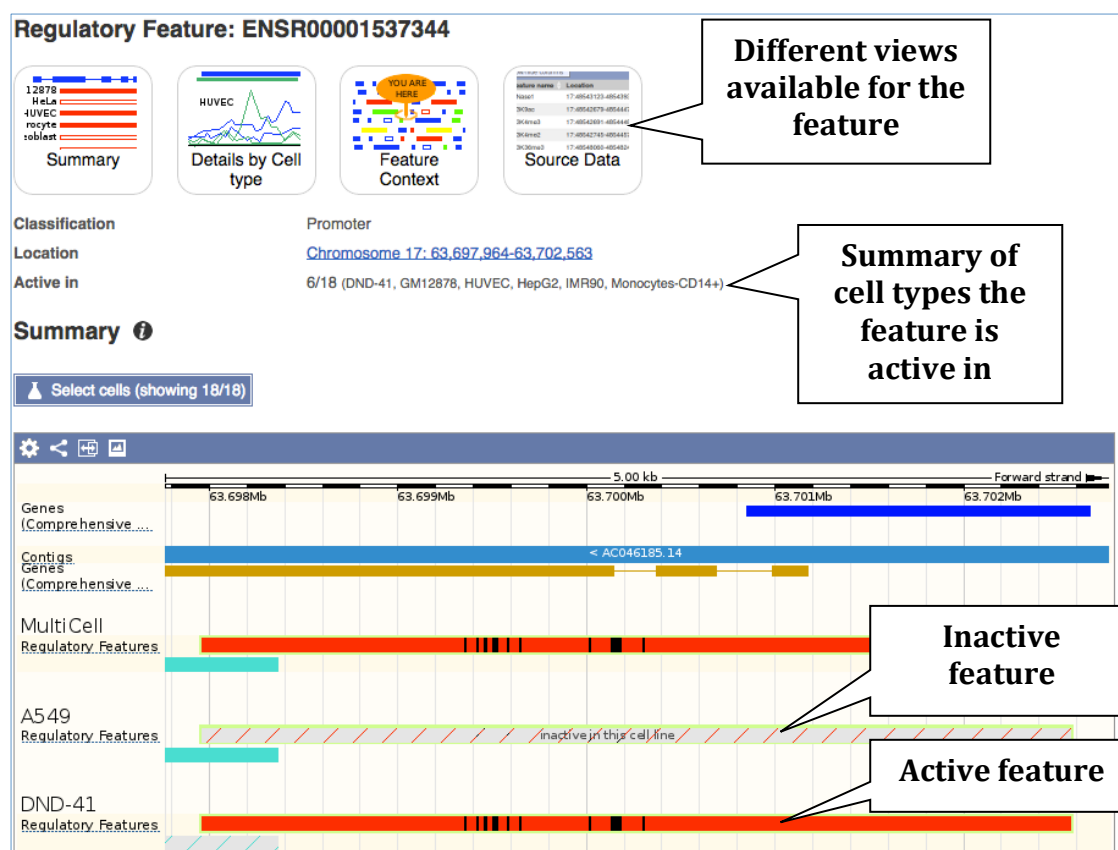
Motif Information

Name	PWM ID	Score
Egr1	MA0341.1	1
Egr1	MA0366.1	1
ETS1	MA0098.2	0.949
Egr1	MA0162.2	0.97
EBF1	MA0154.2	0.985
Egr1	MA0366.1	1
Egr1	MA0366.1	1
Egr1	MA0341.1	1
Egr1	MA0337.1	1
Egr1	MA0162.2	0.963
Egr1	MA0162.2	0.96
Egr1	MA0162.2	0.981
Egr1	MA0337.1	1
Egr1	MA0337.1	1
Egr1	MA0337.1	1

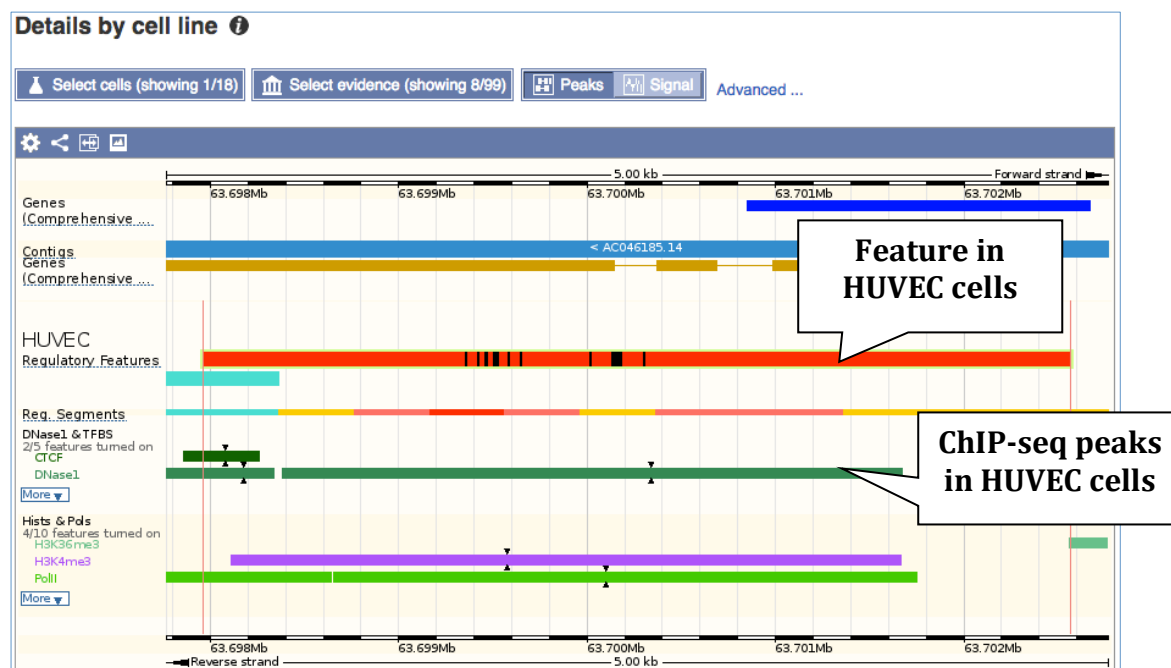
Stable ID of the feature

TF binding motifs within the feature

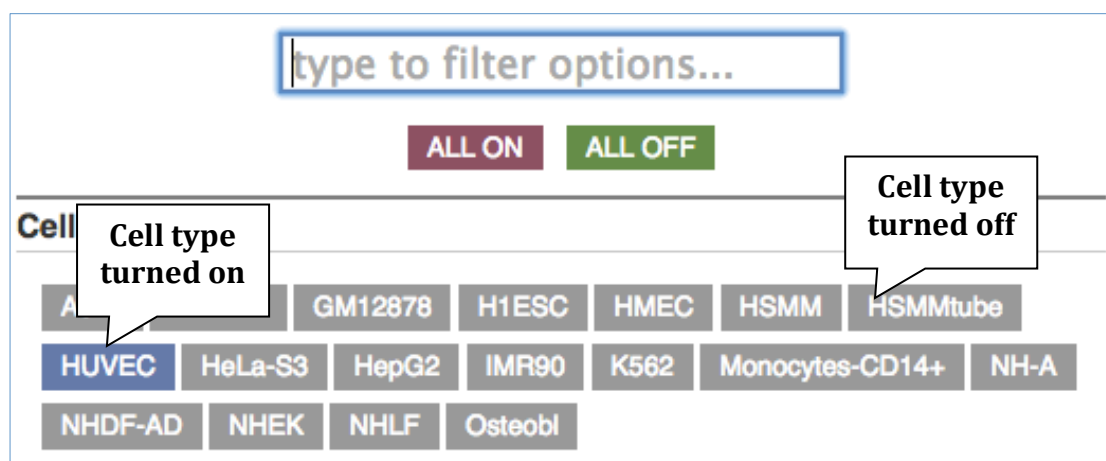
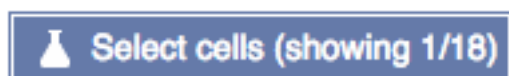
Click on the stable ID, ENSR00001537344, to jump to the “Regulation” tab.



We can see that this promoter is active in six out of the 18 cell types currently in Ensembl. We can explore more detailed data in “Details by Cell type” – click on the icon at the top.



At the moment, this page is only displaying data in HUVEC cells and only for a limited amount of evidence. Click on the [Select cells] button to add more.



We can add cell types by clicking on them. If the cell type is turned on it's blue, if it's off it's grey. You can turn them on or off by clicking on them, or turn everything on or off using the [ALL ON] and [ALL OFF] buttons at the top.

Let's add a cell type where the promoter is inactive – HeLa-S3. Now close the menu.

We can change which evidence we can see, using the [Select evidence] button.



Type to filter options...

ALL ON
ALL OFF

Histone

H2A/B: H2AK5ac H2AZ H2BK5ac H2BK12ac H2BK15ac H2BK20ac H2BK120ac

H3:

H3K4ac H3K4me1 H3K4me2 H3K4me3 H3K9ac H3K9me1 H3K9me3 H3K14ac H3K18ac
H3K23ac H3K23me2 H3K27ac H3K27me3 H3K36me3 H3K56ac H3K79me1 H3K79me2

H4: H4K5ac H4K8ac H4K20me1 H4K91ac

Transcription Factor

ATF3 BAF155 BAF170 BATF BCL3 BCL11A BCLAF1 BHLHE40 Brg1 Cfos Cjun Cmyc
CTCF CTCF1 E2F6 EBF1 Egr1 ELF1 ETS1 FOSL1 FOSL2 FOXA1 FOXA2 Gabp Gata2
GTF2B HDAC2 HDAC8 HEY1 HNF4A HNF4G Ini1 IRF4 Junb Jund Max MEF2A MEF2C
Nanog Nfe2 NFKB NR4A1 Nrf1 Nrf2 p300 Pax5 Pbx3 POU2F2 POU5F1 PU1 Rad21
RXRA Sin3Ak20 SIX5 SP1 SP2 Srf TAF1 TAF7 Tcf12 THAP1 Tr4 USF1 XRCC4 Yy1
ZBTB7A ZBTB33 ZEB1

Polymerase

PolII PolIII

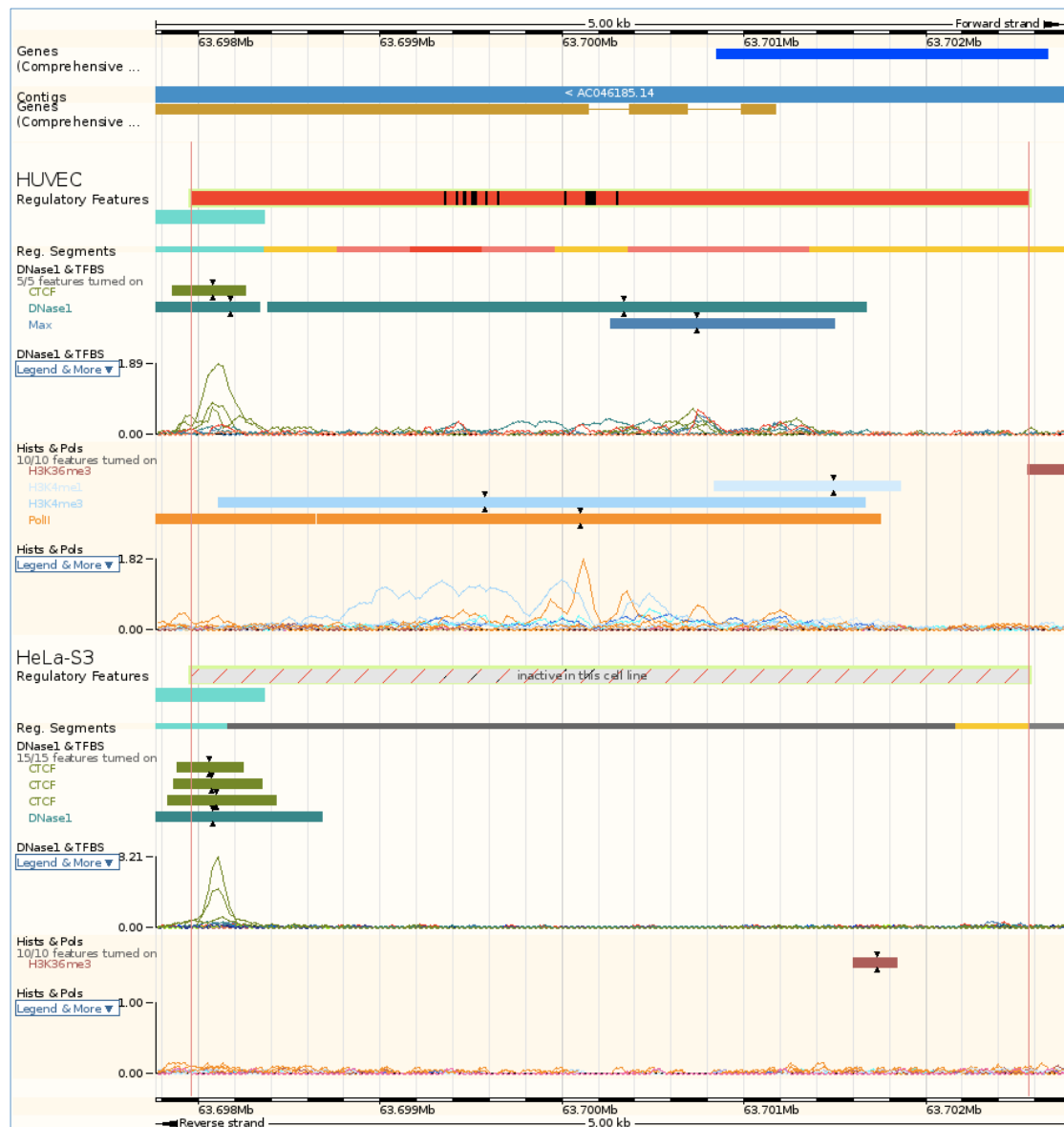
Open Chromatin

DNase1

Choose [ALL ON] to get all the possible evidence, then close the menu.

Lastly, we are currently only seeing the peaks. In order to see the signal too, select the [Signal] button.





Now we can see the active feature in HUVEC compared to the inactive feature in HeLa-S3. In HUVEC, we can see peaks of Max and PolII binding across the promoter, plus H3K4me3 and H3K4me1 modifications and DNase1 sensitivity, whereas there is no such activity in HeLa-S3. In contrast, the CTCF binding site at the left is active, and shows CTCF binding and DNase1 sensitivity in both cell types.

If you would like to see these data in table format, click on the “Source data” icon.

Source data 

Show **All**  entries Show/hide columns Filter 

Cell type	Evidence type	Feature name	Location
A549	DNase1 & TFBS	DNase1	19:7067084-7067515
A549	DNase1 & TFBS	DNase1	19:7069053-7070082
A549	Hists & Pols	H3K4me3	19:7068838-7069175
A549	Hists & Pols	H3K4me3	19:7069272-7070762
A549	Hists & Pols	H3K9ac	19:7069336-7070483
A549	Hists & Pols	H3K4me2	19:7069846-7070846

If you're interested in looking at regulatory features in detail across a region, you can do so in the "Location" tab.

Now click on [Configure this page]. Go to "Regulatory features" in the left hand menu.

Regulation

Enable/disable all Regulatory features

☒	Reg. Feats	★	
☐	Reg. Feats: A549	★	
☐	Reg. Feats: DND-41	★	
☐	Reg. Feats: GM12878	★	
☐	Reg. Feats: H1ESC	★	
☐	Reg. Feats: HMEC	★	
☐	Reg. Feats: HSMM	★	
☐	Reg. Feats: HSMMtube	★	
☒	Reg. Feats: HUVEC	★	
☒	Reg. Feats: HeLa-S3	★	

The "MultiCell Reg. Feats" are already on. Turn on the tracks for the "Reg. Feats: HUVEC" and "Reg. Feats: HeLa-S3".

We can also turn on the evidence tracks. There are two menus for this: "Open chromatin & TFBS" and "Histones & polymerases". Open the menu for "Histones & polymerases".

Regulatory features

Histone modifications & RNA polymerases ⓘ

Filter by: All classes
Enter terms to filter by

Key
On: Off: No data: Evidence type: On: Off:

Cell type ▶
Evidence type ▼

Track style:
[Enable/disable all](#)

Cell lines

CD4 GM06990 GM12878 H1ESC HMEC HSMM HUVEC HeLa-S3 HepG2 IMR90 K562 NH-A NHEK

Legend

Choose track styles

Histone modification

Select boxes

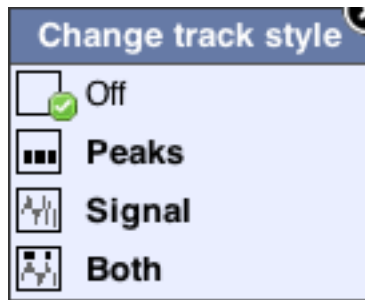
Feature	CD4	GM06990	GM12878	H1ESC	HMEC	HSMM	HUVEC	HeLa-S3	HepG2	IMR90	K562	NH-A	NHEK
Polymerase													
PolII	On	On	On	On	On	On	On	On	On	On	On	On	On
PolIII	On	On	On	On	On	On	On	On	On	On	On	On	On
Histone													
H2AK5ac	On	On	On	On	On	On	On	On	On	On	On	On	On
H2AK9ac	On	On	On	On	On	On	On	On	On	On	On	On	On
H2AK27me3	On	On	On	On	On	On	On	On	On	On	On	On	On
H2AK36me3	On	On	On	On	On	On	On	On	On	On	On	On	On
H2AK4me3	On	On	On	On	On	On	On	On	On	On	On	On	On
H2AK9me3	On	On	On	On	On	On	On	On	On	On	On	On	On
H2BK15ac	On	On	On	On	On	On	On	On	On	On	On	On	On
H2BK20ac	On	On	On	On	On	On	On	On	On	On	On	On	On
H2BK5ac	On	On	On	On	On	On	On	On	On	On	On	On	On

You can turn on a single track by clicking on the box in the matrix. Note that certain tracks are already selected for all cell lines by default (PolII, PolIII, H3K27me3, H3K36me3, H3K4me3, H3K9me3). However, these will appear in the “Region in detail” view only if you specify a track style for the cell lines.

Turn on all the tracks for HUVEC and HeLa-S3. Hover over the cell line name then select “All”.



Now choose the track style for the tracks you’ve switched on. Click on the track style box for HUVEC and HeLa-S3 and select “Both”.



There is a similar matrix for “Open chromatin & TFBS”. Use this to turn on all tracks for “HeLa-S3” and “HUVEC” in “Both” track style. Now close the configuration page.

We can now see regulatory activity across the region in both cell types.



The Ensembl Regulatory Build incorporates data from sources such as ENCODE, Blueprint and the Roadmap Epigenomics project. To see the data directly from these sources, you can add so-called track hubs

Click on “Trackhubs” on the Ensembl homepage (<http://www.ensembl.org>).

Ensembl supports data from external projects through [Trackhubs](#)



This page lists various track hubs that can be added to Ensembl.

Datashub name	Description	Species and assembly
Blueprint Hub	Blueprint Epigenomics Data Hub	Human (GRCh37)
ENCODE Analysis Hub	ENCODE Integrative Analysis Data Hub	Human (GRCh37)
Broad Improved Canine Annotation v1	Broad Institute CanFam3 Improved Annotation Data v1	Dog (CanFam3)
Cancer genome polyA site & usage	An in-depth map of polyadenylation sites in cancer (matched-pair tissues and cell lines)	Human (GRCh37)
CEMT (CEEHRC)	Epigenomic Data tracks from BCGSC, Vancouver	Human (GRCh37)
DNA Methylation	DNA Methylation Hundreds of analyzed methylomes from bisulfite sequencing data	Human (GRCh37)
		Human (NCBI36)
		Mouse (GRCm38)
		Mouse (NCBI37)
		Chimpanzee (CHIMP2.1.4)

The table contains a brief description of each hub, plus the assembly that the hub is based on, as a link. Click on the link to turn on the hub. If the hub is based on a genome assembly which is not the current assembly in Ensembl, the link will also jump you to an archive with the previous assembly.

Track hubs often contain vast amounts of data, which can slow Ensembl down, so only add them if you need them, and trash them when you are finished with them.

Click on the link “Human (GRCh37)” for the “ENCODE Analysis Hub”.

This will take you directly to the “Region in detail” view. Because this is a GRCh37 track hub, this has taken you to our dedicated GRCh37 site (<http://grch37.ensembl.org>).

Open the configuration page to see that a new category, named “ENCODE Analysis Hub”, has been added to your left hand menu.

Configure Region Image		Configure Overview
Active tracks		
Favourite tracks		
Track order		
Search results		
☒	ENCODE Analysis Hub	(0/2503)
☐	ENCODE DNase-seq Peaks	(0/125)
☐	ENCODE DNase-seq Signal	(0/125)
☐	ENCODE FAIRE Peaks	(0/25)
☐	ENCODE FAIRE Signal	(0/24)
☐	ENCODE Genome Segmentations	(0/18)
☐	ENCODE Histone Modifications Peaks	(0/190)
☐	ENCODE Histone Modifications Signal	(0/198)
☐	ENCODE RNA Contigs	(0/75)
☐	ENCODE RNA Signal	(0/352)
☐	ENCODE Transcription Factor Peaks	(0/914)
☐	ENCODE Transcription Factor Signals	(0/457)

Click on the various links under “ENCODE Analysis Hub” to find the ENCODE configuration matrices, which work in the same way as the “Open chromatin & TFBS” and “Histones & polymerases” matrices, except that some have multiple options (indicated by numbers within the boxes). If you click on these boxes, you can choose which of these options to add.

ENCODE RNA Signal

Filter by

Key

On  Off  No Data 

Filtered: On  Off 

Localization ▶

Cell Line ▼

Default style: ☐ ☐ ☐ ☐ ☐ ☐ ☐

[Enable/disable all](#)

	Chromatin	Cytosol	Nucleolus	Nucleoplasm	Nucleus	Polysome	Whole Cell
A549							0 12
AG04450							0 12
BJ							0 12
GM12878		0			0		0
H1-hESC		0					
HeLa-S3		0					
HepG2		0					
HMEC							
HSMM							
HUVEC		0					
K562		0 10	0				
MCF-7							
NHEK		0					
NHLF							
Prostate							
SK-N-SH RA							

Enable/disable all Cytosol:GM12878

☐ GM12878 Cytosol Minus non-PolyA Signal CSHL Rep 1

☐ GM12878 Cytosol Minus non-PolyA Signal CSHL Rep 2

☐ GM12878 Cytosol Minus PolyA Signal CSHL Rep 1

☐ GM12878 Cytosol Minus PolyA Signal CSHL Rep 2

☐ GM12878 Cytosol Minus Total Signal CSHL Rep 1 TAP Only

☐ GM12878 Cytosol Minus Total Signal CSHL Rep 2 TAP Only

☐ GM12878 Cytosol Minus Total Signal CSHL TAP Only

☐ GM12878 Cytosol Plus non-PolyA Signal CSHL Rep 1

☐ GM12878 Cytosol Plus non-PolyA Signal CSHL Rep 2

☐ GM12878 Cytosol Plus PolyA Signal CSHL Rep 1

☐ GM12878 Cytosol Plus PolyA Signal CSHL Rep 2

☐ GM12878 Cytosol Plus Total Signal CSHL Rep 1 TAP Only

☐ GM12878 Cytosol Plus Total Signal CSHL Rep 2 TAP Only

☐ GM12878 Cytosol Plus Total Signal CSHL TAP Only

Exercises

UCSC Genome Browser

The *HLA-DRB1* and *HLA-DQA1* genes are part of the human major histocompatibility complex class II (MHC-II) region and are located about 44 kb from each other on chromosome 6. In the paper ‘The human major histocompatibility complex class II *HLA-DRB1* and *HLA-DQA1* genes are separated by a CTCF-binding enhancer-blocking element’ (Majumder *et al.* J Biol Chem. 2006 Jul 7;281(27):18435-43) a region of high acetylation located in the intergenic sequences between *HLA-DRB1* and *HLA-DQA1* is described. This region, termed XL9, coincided with sequences that bound the insulator protein CCCTC-binding factor (CTCF). Majumder *et al.* hypothesise that the XL9 region may have evolved to separate the transcriptional units of the *HLA-DR* and *HLA-DQ* genes.

Go to the region from bp 32,540,000 to 32,620,000 on human chromosome 6 (use the GRCh37/hg19 assembly). Hide all tracks and then add the “GENCODE ...” and “ENCODE Regulation ...” tracks. Go to the configuration page for the “Txn Factor ChIP” track by clicking on the grey bar in front of it, and use the “Filter by factor” option to only show “CTCF” binding sites.

(a) Does the intergenic region between the *HLA-DRB1* and *HLA-DQA1* genes contain any CTCF binding sites? Is it a region of high acetylation? Do any of the CTCF binding sites colocalise with a region of high acetylation?

Turn on the “Genome Segments” track in “dense” mode. Go to its configuration page and turn on all available genome segmentation tracks.

(b) What colour are “CTCF enriched elements” in the “Genome Segments” tracks? Is any of the CTCF binding sites reflected in the “Genome Segments” tracks?

Ensembl Regulatory Build

(a) Go to the “Region in detail” view for the human *STX7* (Syntaxin 7) gene. Are there any predicted enhancers in this gene region? If so, where in the gene do they appear?

(b) Open the Configuration page and turn on “Regulatory features” for HUVEC, HeLa-S3, and HepG2 cell types. Are the predicted enhancers active in any of these cell types?

(c) Add DNase1 hypersensitivity data (a mark of open chromatin) for the HeLa-S3 cell type. Are there any DNase1 hypersensitive sites in the *STX7* gene in HeLa-S3 cells?

(d) Add histone modification data for the HeLa-S3 cell type. Which ones are present at the 5' end of *STX7*?

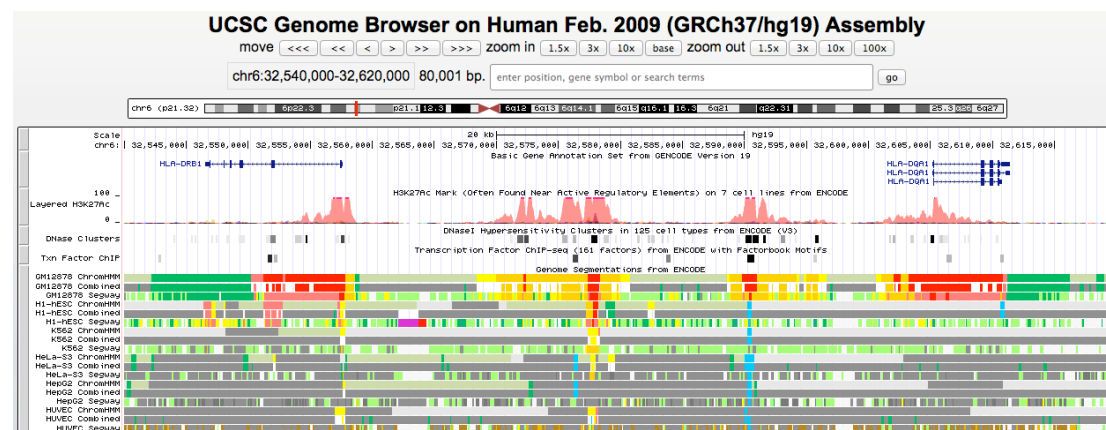
(e) Turn on the “CpG island” track. Are there any CpG islands in the *STX7* gene region?

Exercises answers

UCSC Genome Browser

(a) Yes, the intergenic region between the *HLA-DRB1* and *HLA-DQA1* genes contains four CTCF binding sites of varying length. It is also a region of high acetylation, as shown by the peaks in the “Layered H3K27Ac” track. Two of the CTCF binding sites collocate with a region of high acetylation.

(b) “CTCF enriched elements” are coloured blue in the “Genome Segments” tracks. The CTCF binding sites that collocate with a region of high acetylation are both reflected in the “Genome Segments” tracks, but only in a subset of cell types.



Ensembl Regulatory Build

(a) Yes, there are five enhancers (coloured in dark yellow) predicted in the region of the *STX7* gene, two near the 3' end, two near the middle and one near the 5' end.

(b) Two of the five predicted enhancers are active in HUVEC cells, while none of the five are active in HeLa-S3 and HepG2 cells.

(c) Yes, there's a DNase1 hypersensitive site at the 5' end of the *STX7* gene. Clicking on the coloured block shows that the source of this information is the ENCODE project.

(d) Several histone modifications are found at the 5' end of the *STX7* gene in HeLa-S3 cells, i.e. H3K27ac, H3K36me3, H3K4me2, H3K4me3 and H3K9ac.

(e) Yes, there is a CpG island at the 5' end of the *STX7* gene.



Literature:

The ENCODE Project Consortium (2012) "An integrated encyclopedia of DNA elements in the human genome". *Nature* **489** (7414): 57–74. (<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3439153/>)

The ENCODE Project Consortium (2007). "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project". *Nature* **447** (7146): 799–816. (<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2212820/>)