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.

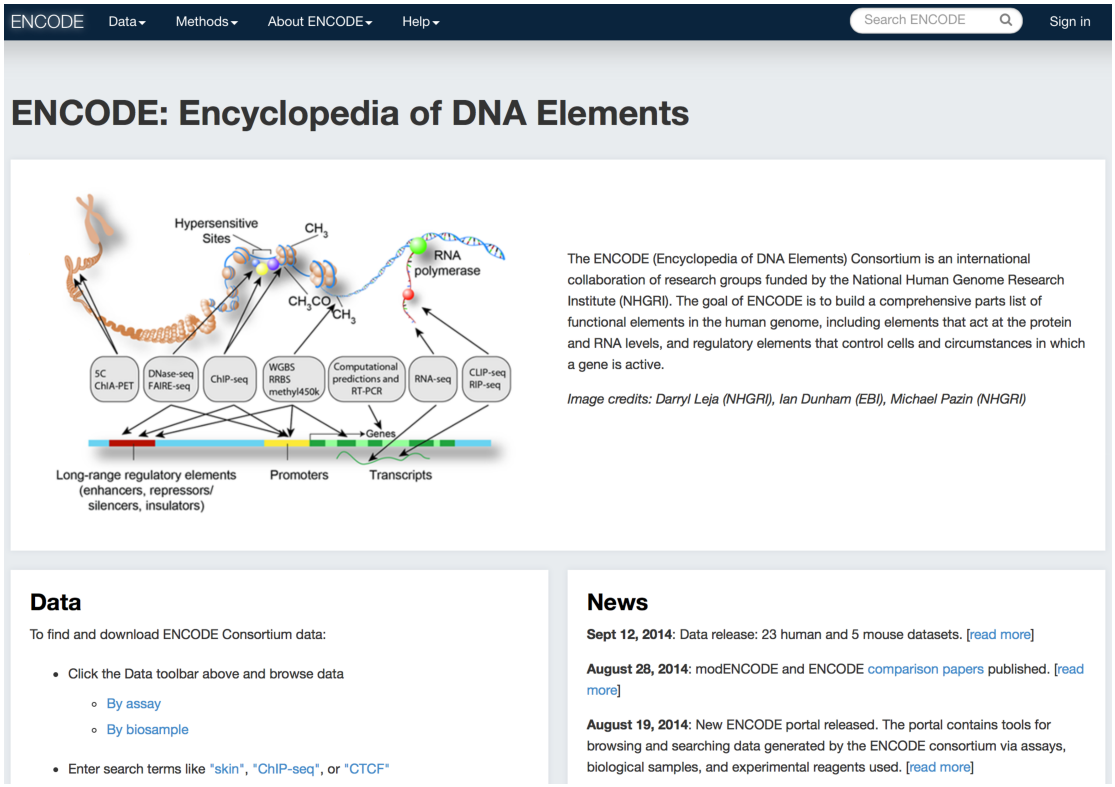
The ENCODE portal

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.

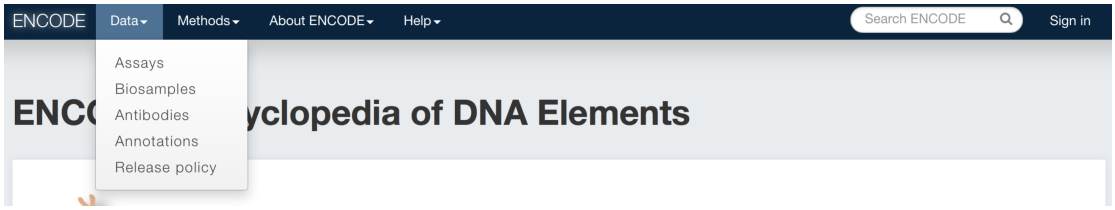
Worked example 1: the ENCODE portal

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).

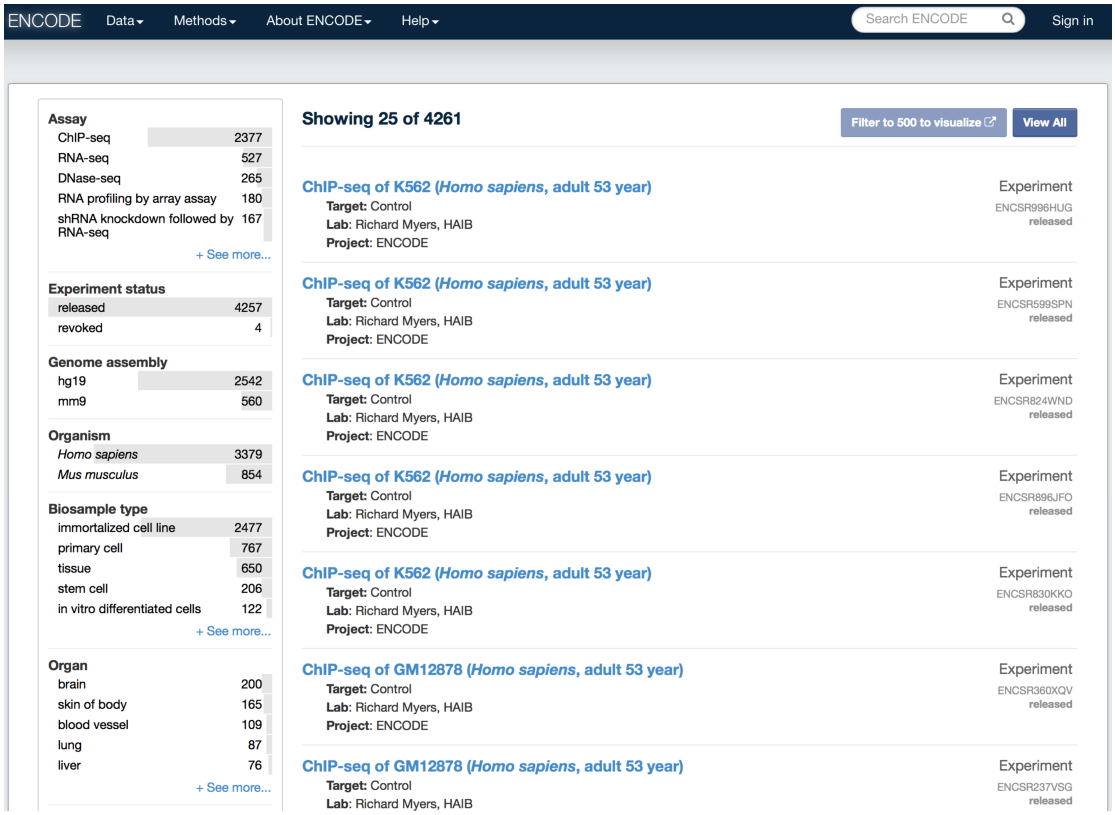


(2) Click on the “Data” drop-down menu in the toolbar.



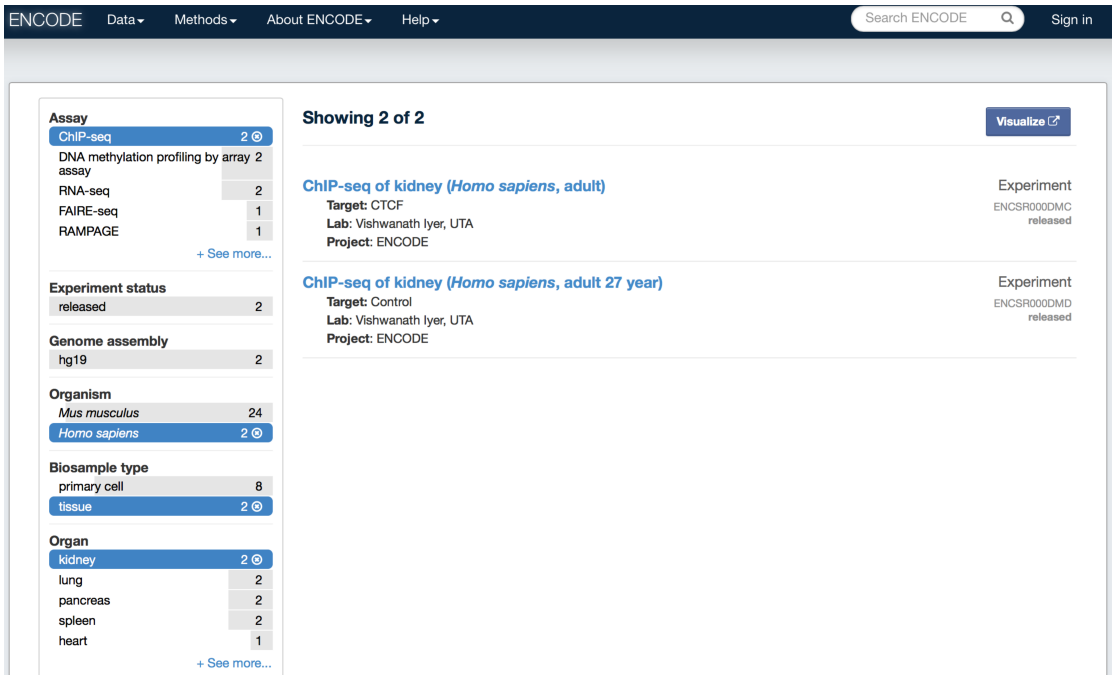
Data can be browsed via assays, biosamples, and antibodies used.

(3) Select “Assays”.



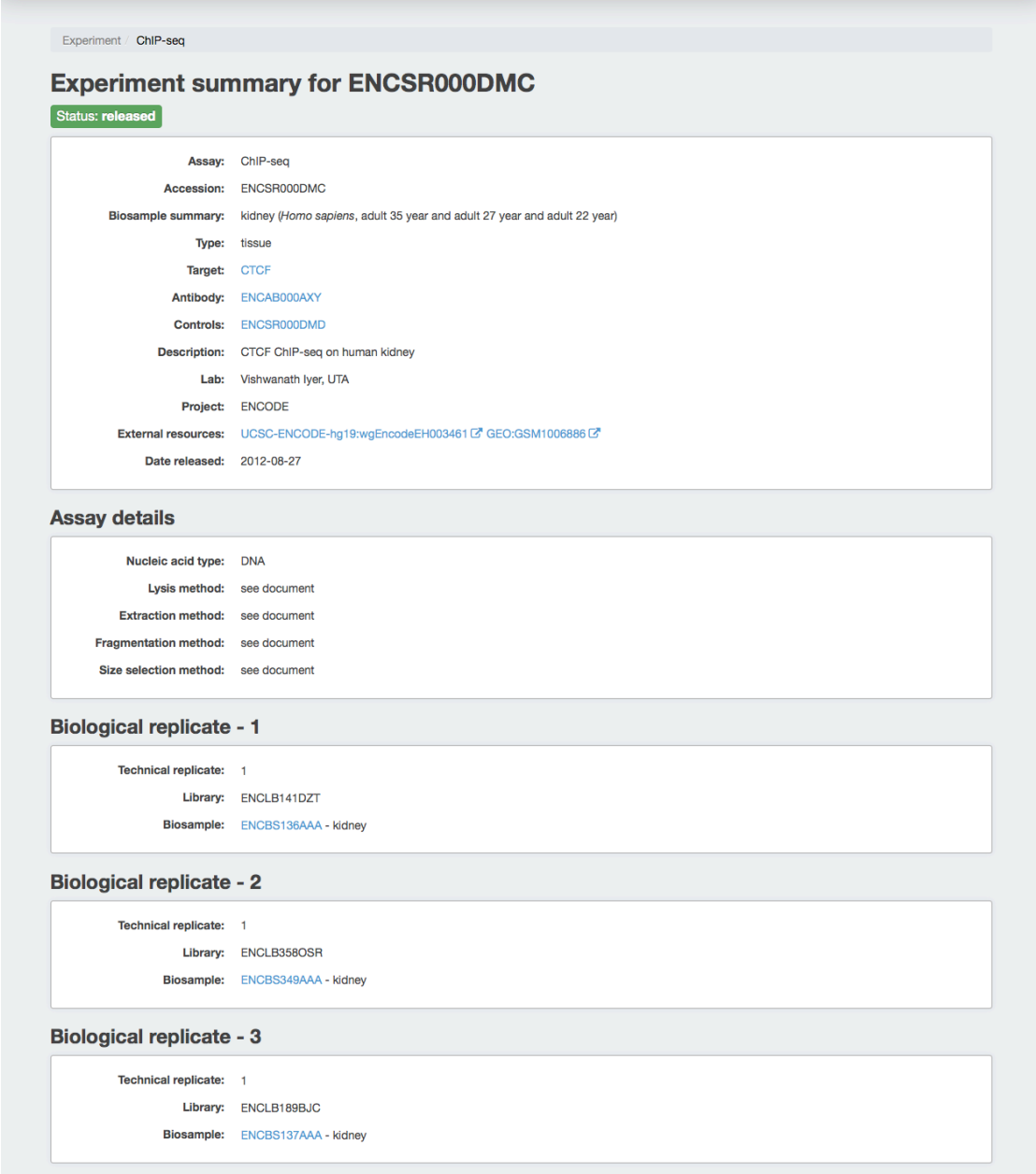
The "Assays" page lists all assays that have been used to generate ENCODE data. The results can be narrowed and filtered by selecting one or more values in a metadata category on the left hand side of the page. Multiple values from each facet can be selected at any one time.

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

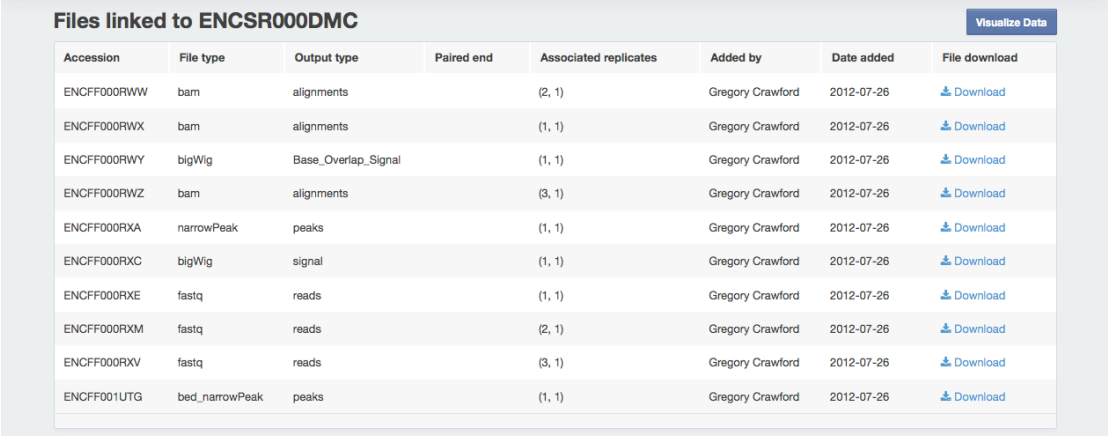


The results show that there are two datasets that match our search criteria, one containing CTCF binding data, and a control dataset. CTCF is a transcriptional repressor (http://en.wikipedia.org/wiki/CTCF).

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



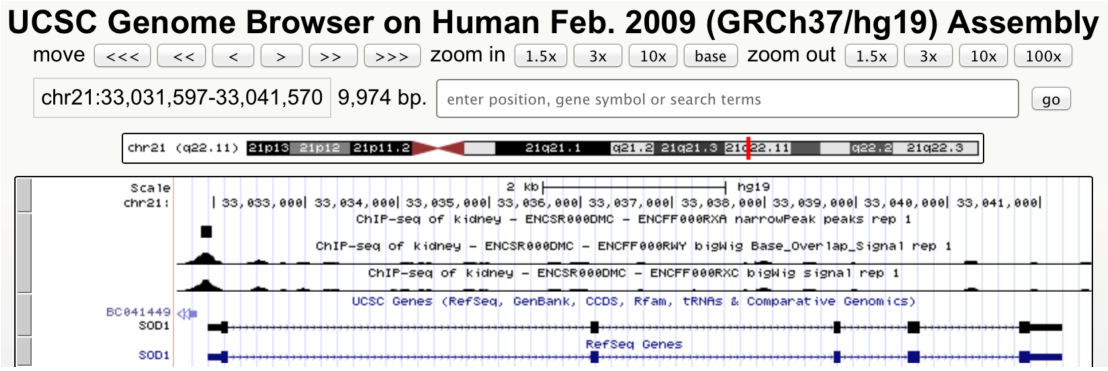
Details about the dataset are shown. At the bottom of the page data can be downloaded in various formats. Clicking on [Visualize Data] launches a UCSC Genome Browser view.



Files linked to ENCSR000DMC Visualize Data

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].

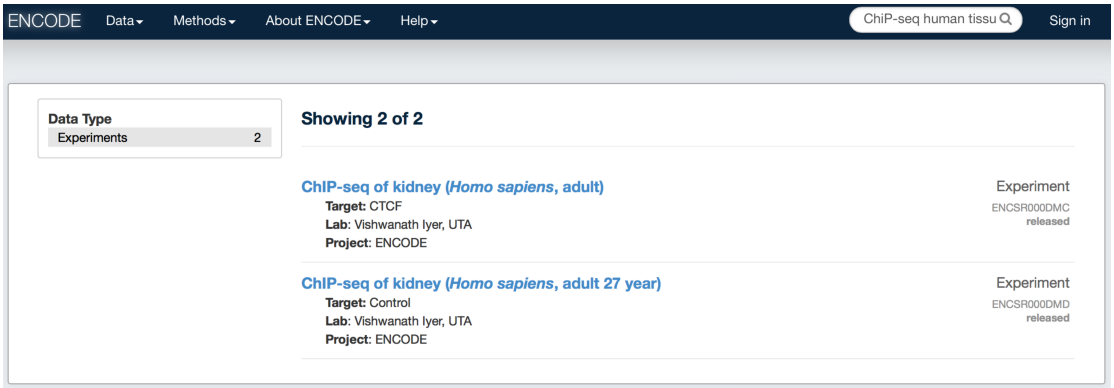


Tracks containing the kidney CTCF ChIP-seq data have been added to the browser view.

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”.



This gives the same result as we got by browsing by assay.

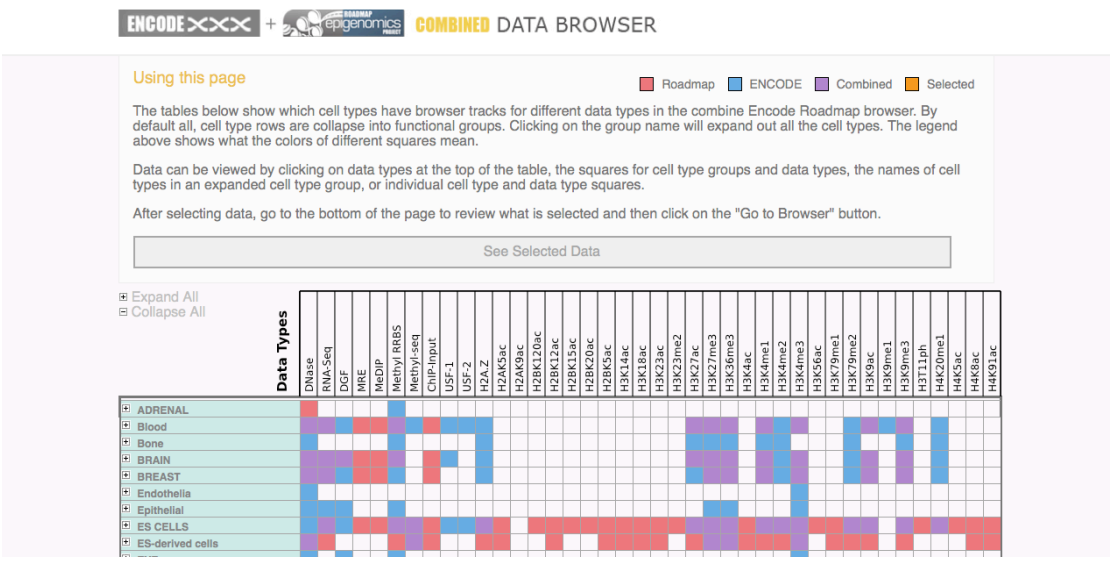
More information about how to access data via the ENCODE portal can be found at https://www.encodeproject.org/help/getting-started.

The ENCODE Roadmap Browser and the IHEC Data Portal

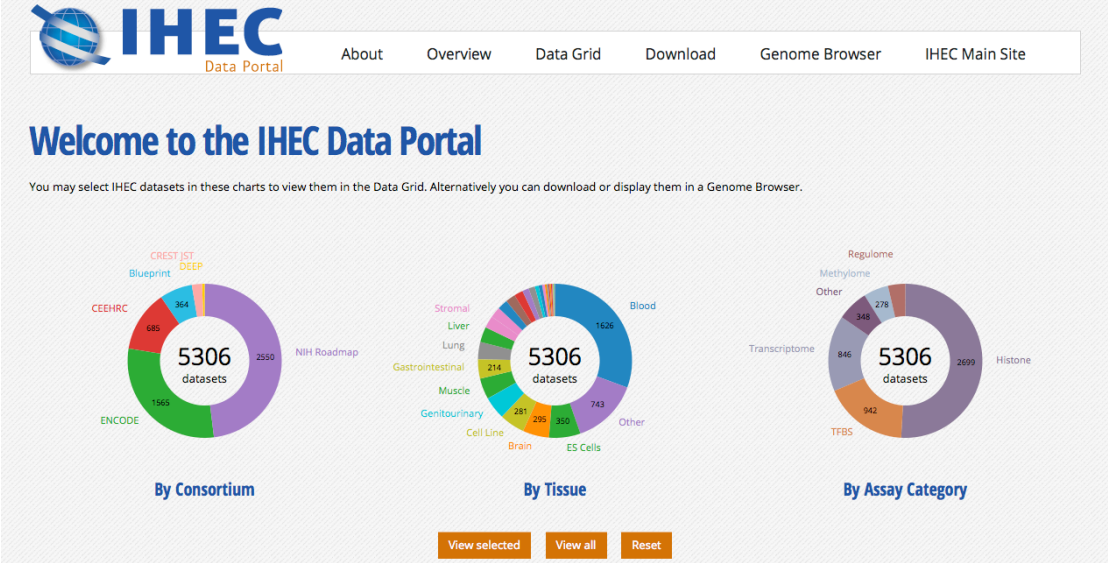
ENCODE data can also be searched along with data from other consortia.

The ENCODE Roadmap browser (http://www.encode-roadmap.org) allows searching of ENCODE data and data from the Roadmap Epigenomics project (http://www.roadmapepigenomics.org).

Data can be selected by selecting boxes from a matrix. The matrix is organised by data types (columns) and cell types (rows). After the data have been selected they subsequently can be visualised in the UCSC genome browser.



The IHEC (International Human Epigenome Consortium) Data Portal (http://epigenomesportal.ca/ihec/index.html) allows searching of ENCODE data and data from multiple other epigenomics projects.



The UCSC Genome Browser

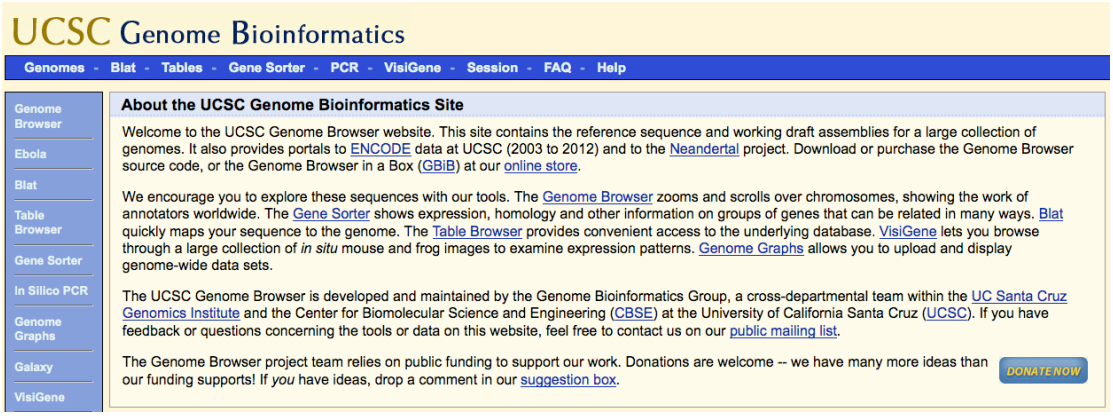
UCSC coordinated data for the ENCODE Consortium from its inception in 2003 (Pilot phase) to the end of the first 5 year phase of whole-genome data production in 2012. All data produced by ENCODE investigators and the results of ENCODE analysis projects from this period are hosted in the UCSC Genome Browser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

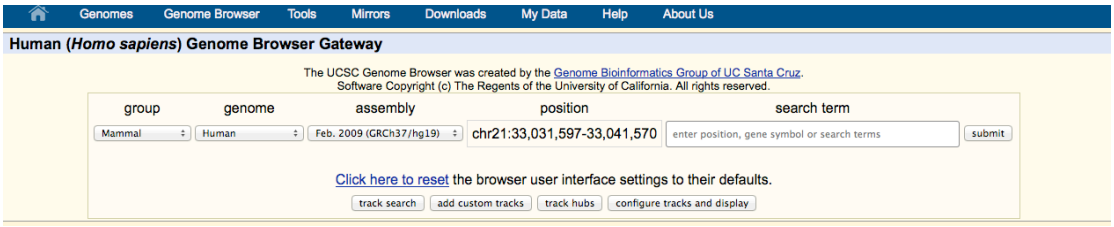
Worked example 2: the UCSC Genome Browser

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/).

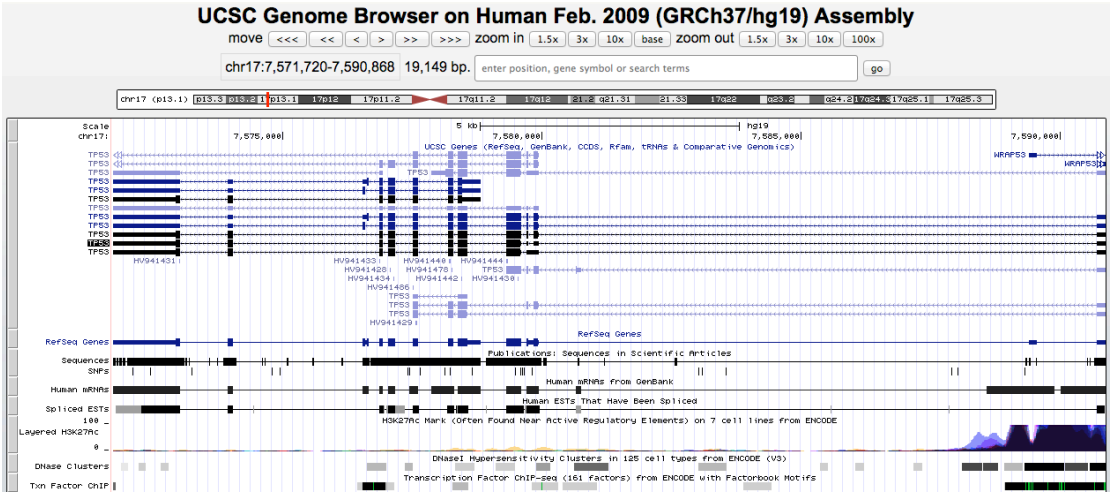


(2) From the blue navigation links on the left side of the page, click the "Genome Browser" link.



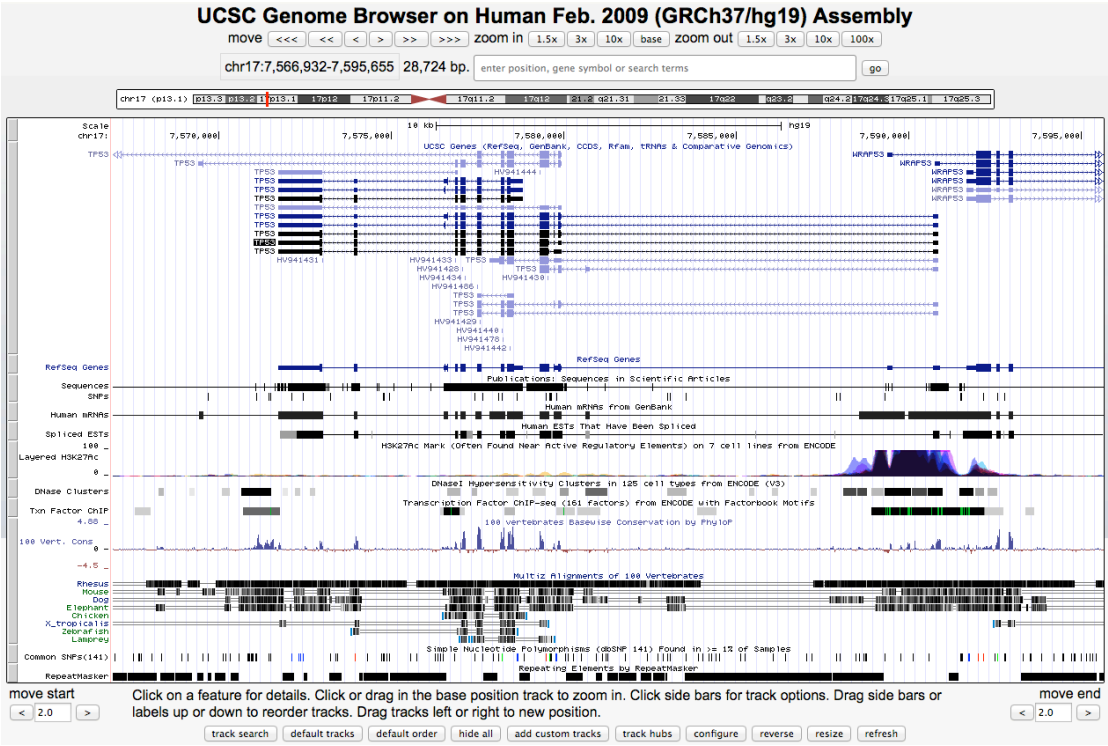
(3) On the "Human Genome Browser Gateway" interface, click the "Click here to reset the browser user interface settings to their defaults." link. This will ensure that any prior activity on the browser has been cleared out and that everyone is starting with default settings.

(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.

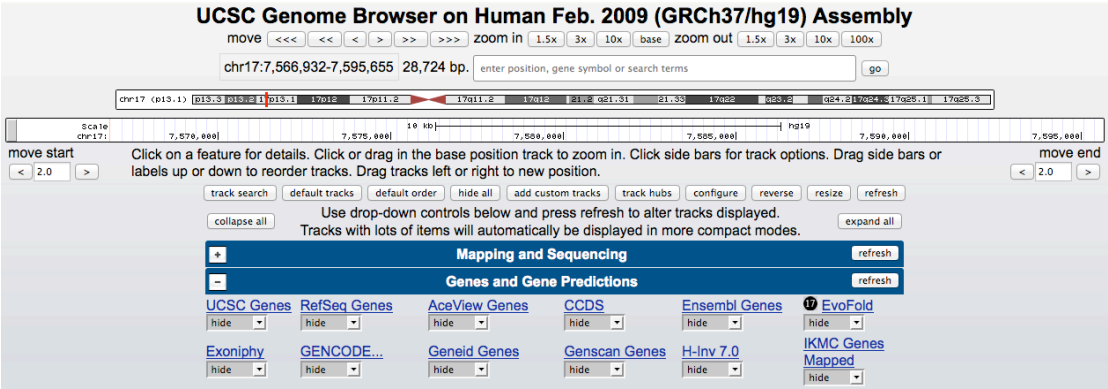


Note that we are using the GRCh37/hg19 assembly, because the ENCODE data haven't been mapped to the GRCh38/hg38 assembly.

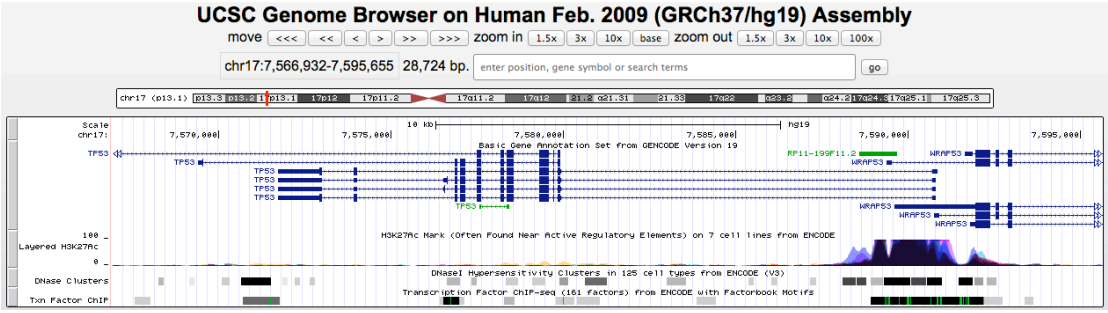
(5) In the TP53 region on the browser, examine the features briefly. Then click the "zoom out [1.5x]" button near the top. Assess the features again.



(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.)

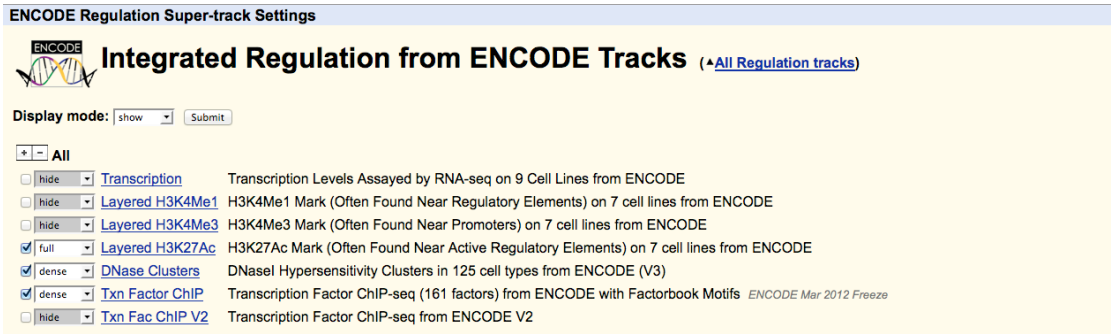


(7) Add the "GENCODE …" track (from the "Genes and Gene Predictions" group) and the "ENCODE Regulation …" track (from the "Regulation" group) by choosing "show" in the respective pull down menus and clicking a [refresh] button.

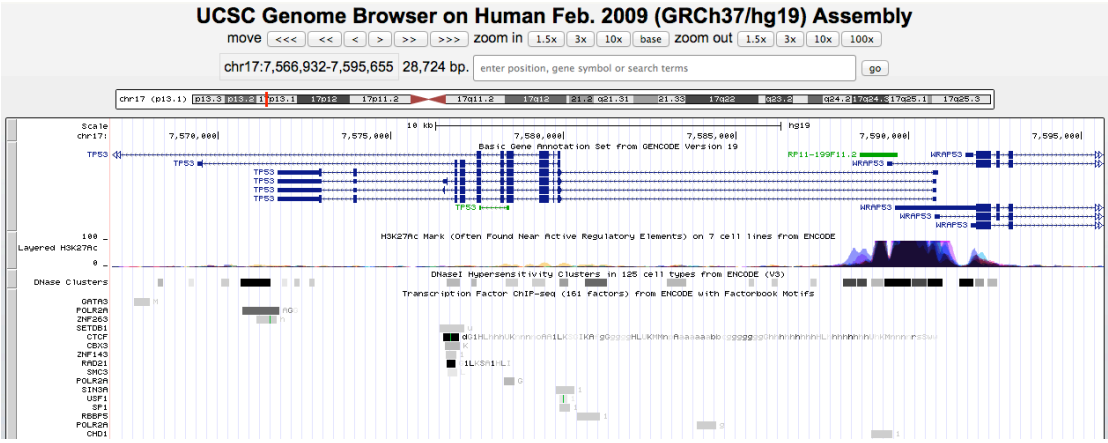


Examine the display. Note that the “Txn Factor ChIP” track shows data blocks, but not individual transcription factors. Also note that the “Layered H3K27Ac” track appears to contain multiple data sets of various colours.

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

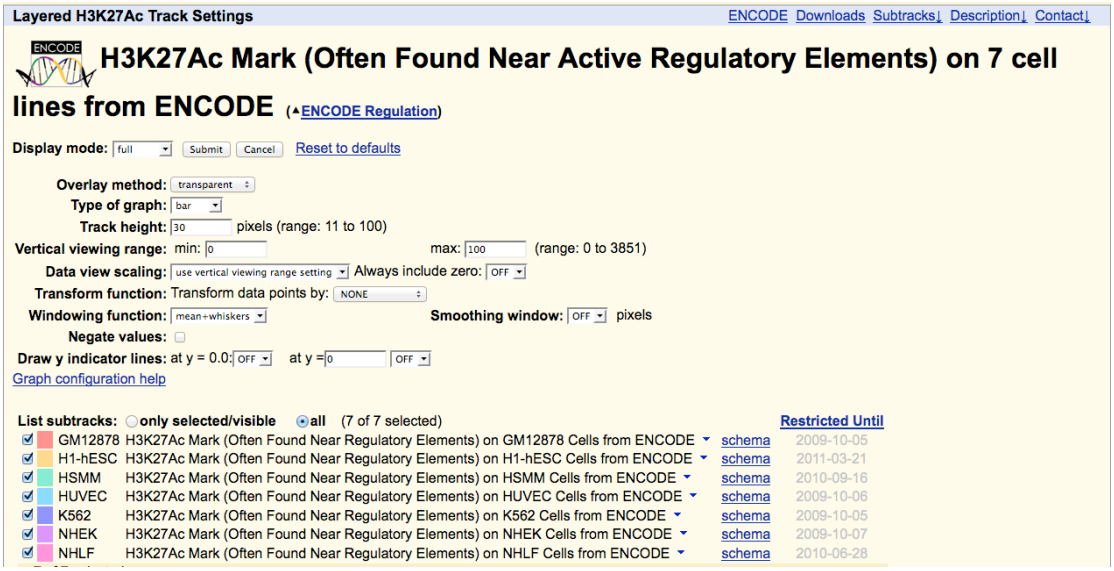


(9) By default the “Txn Factor ChIP” track is visible in “dense” mode. Change this to “full”. Click the [Submit] button.



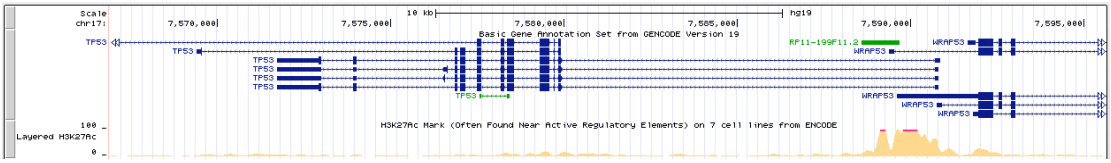
Examine the display again. Note that individual transcription factors can be identified by name using the labels on the left. Note that the letter codes near the blocks correspond to cell lines that have been used in experiments for this data. Click some of the blocks to note the cell lines and signal 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.



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.

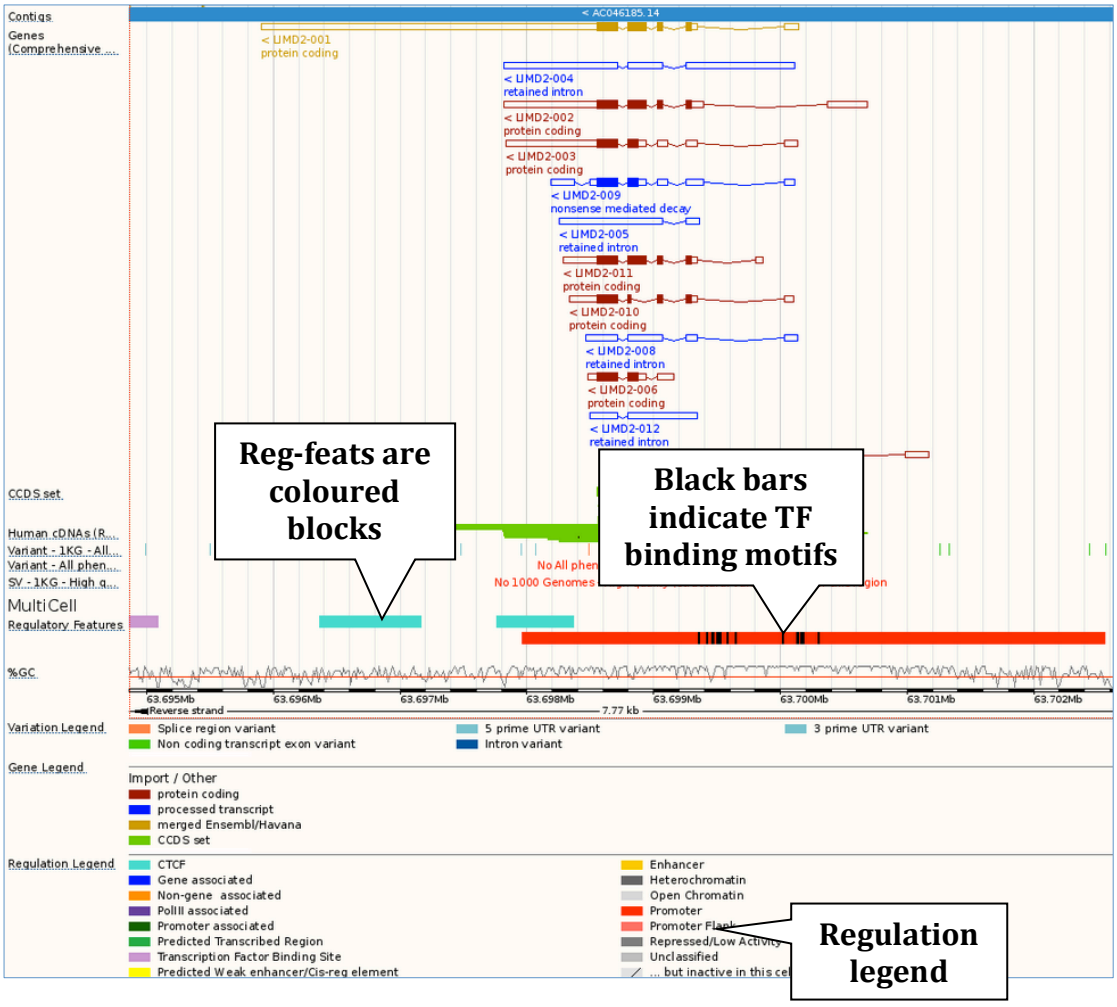
The Ensembl Genome Browser

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.

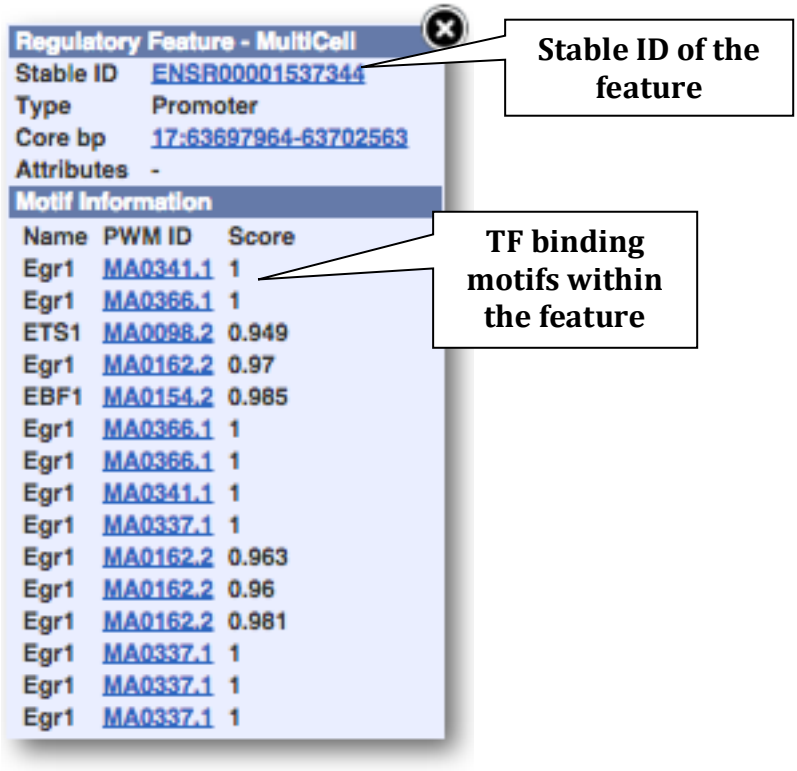
We'll start by searching for the human *LIMD2* (LIM domain containing 2) gene on the Ensembl homepage (http://www.ensembl.org) and jumping to the "Location" tab. Zoom out a little to see the gene plus some of the flanking regions.



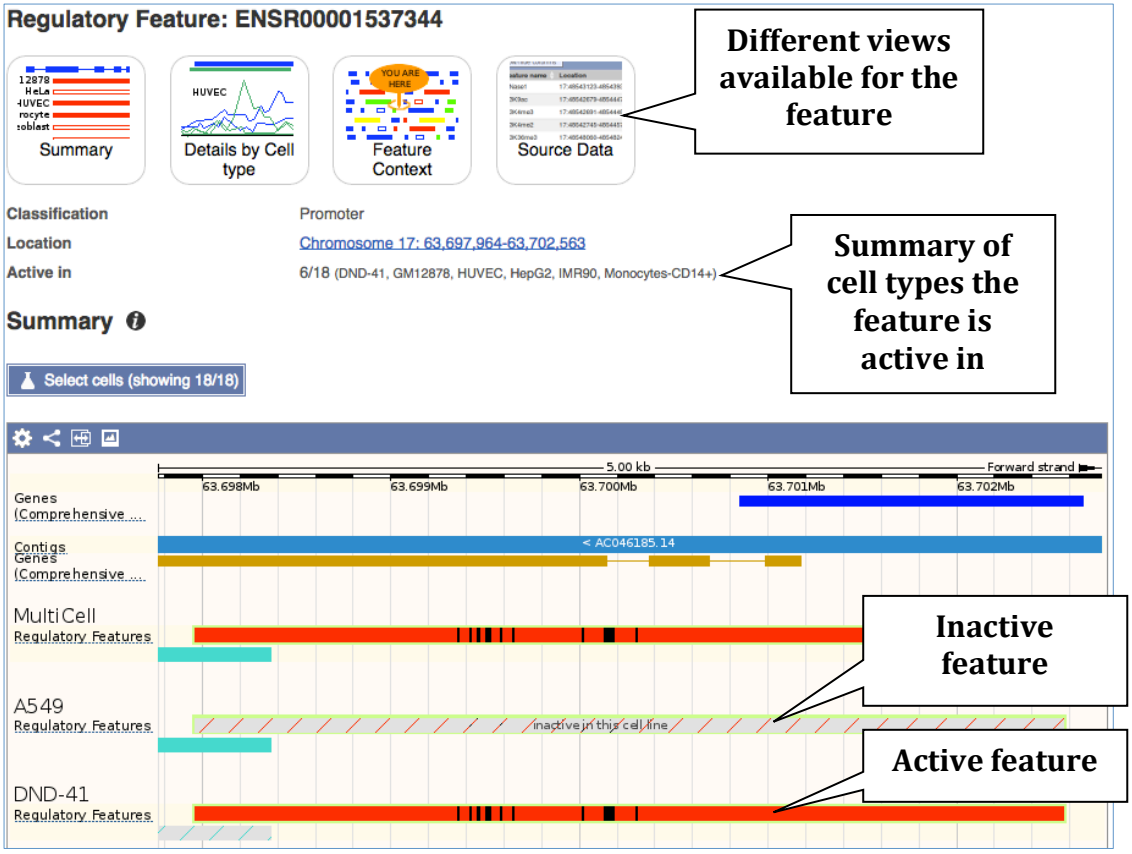
The "MultiCell Regulatory Features" are shown by default. In this region we can see a large red promoter, two turquoise CTCF binding sites and a lilac

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.

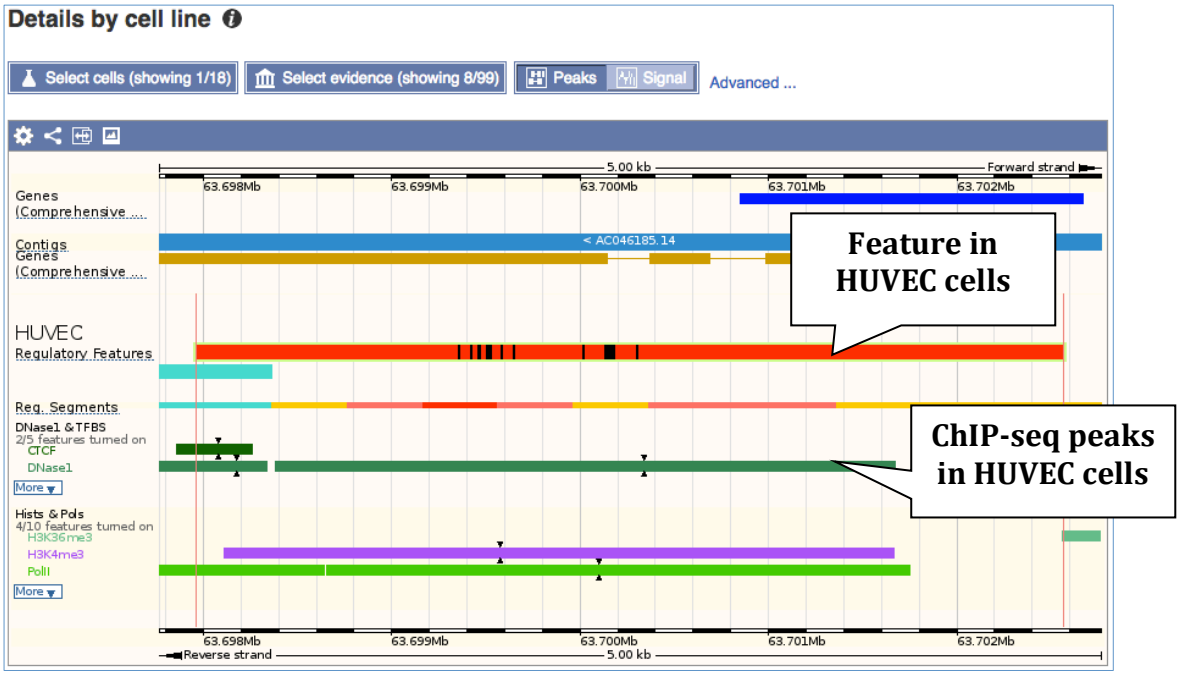
You can also click on the regulatory features to learn more. Click on the red promoter to get a pop-up.



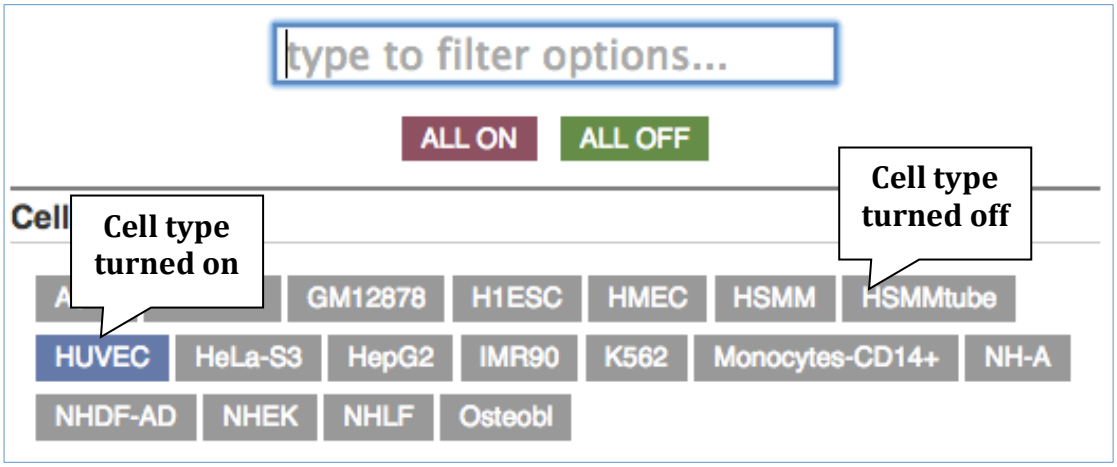
Click on the stable ID, ENSR00001537344, to jump to the "Regulation" tab.



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.



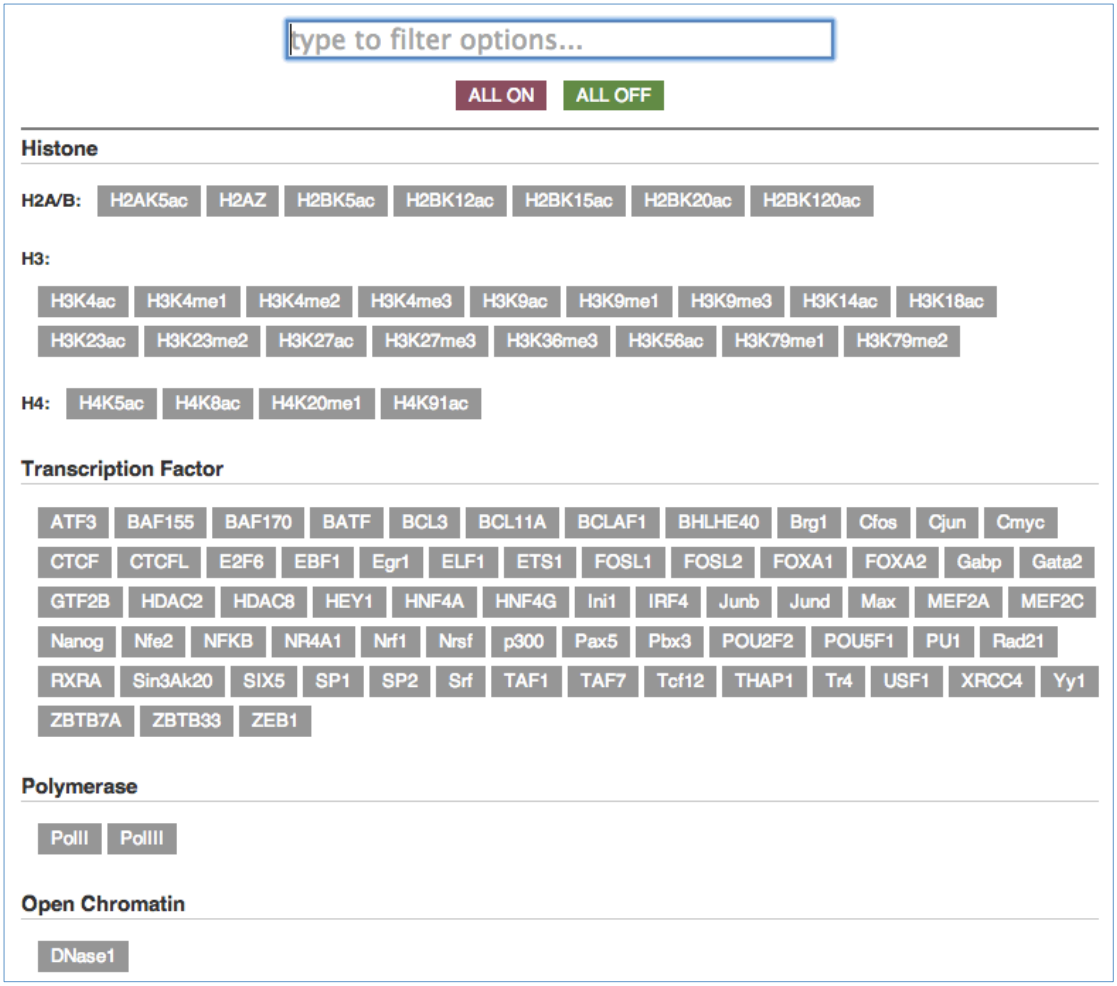
At the moment, this page is only displaying data in HUVEC cells and only for a limited amount of evidence. Click on the [Select cells] button to add more.



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.

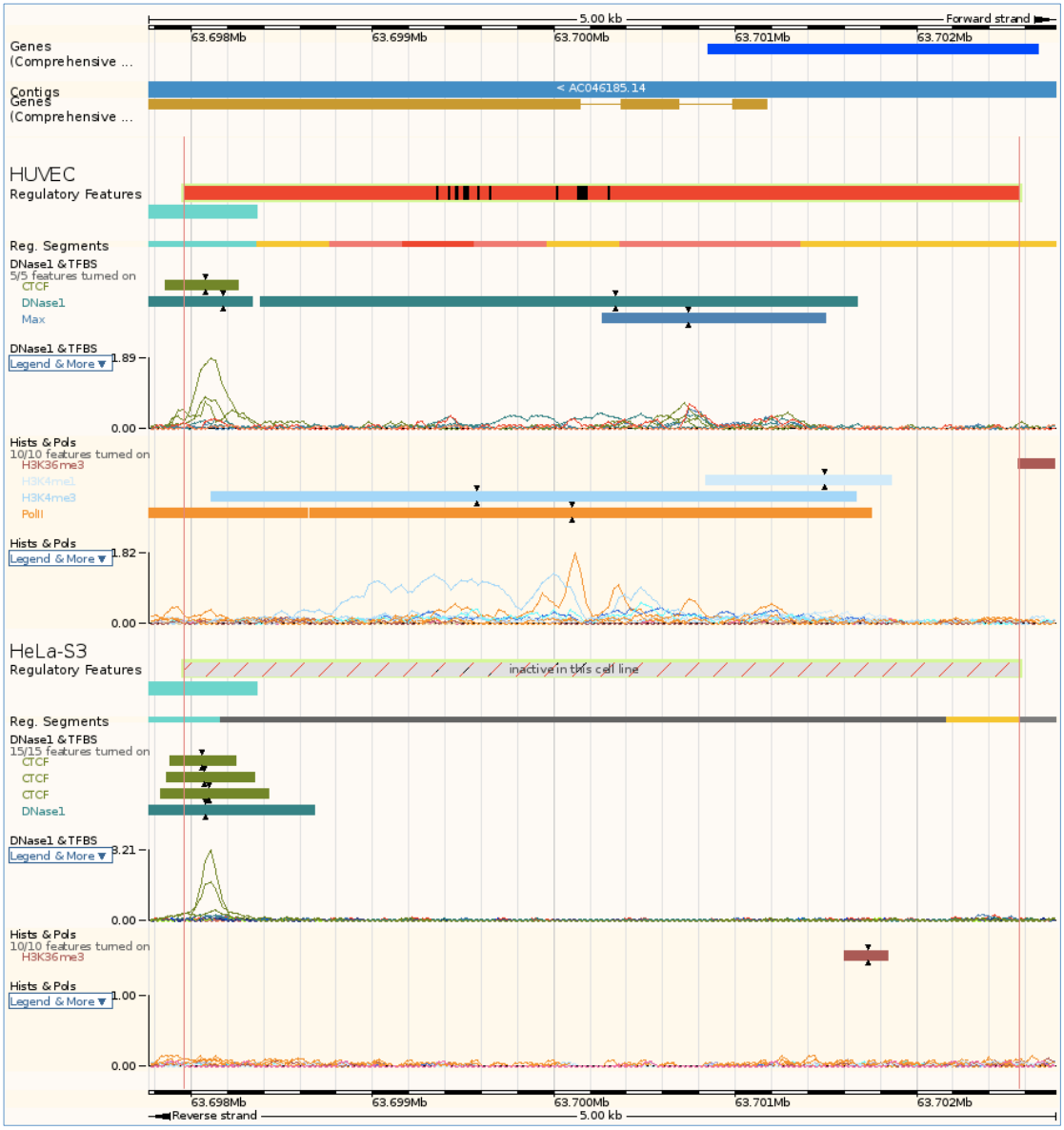
Let's add a cell type where the promoter is inactive – HeLa-S3. Now close the menu.

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



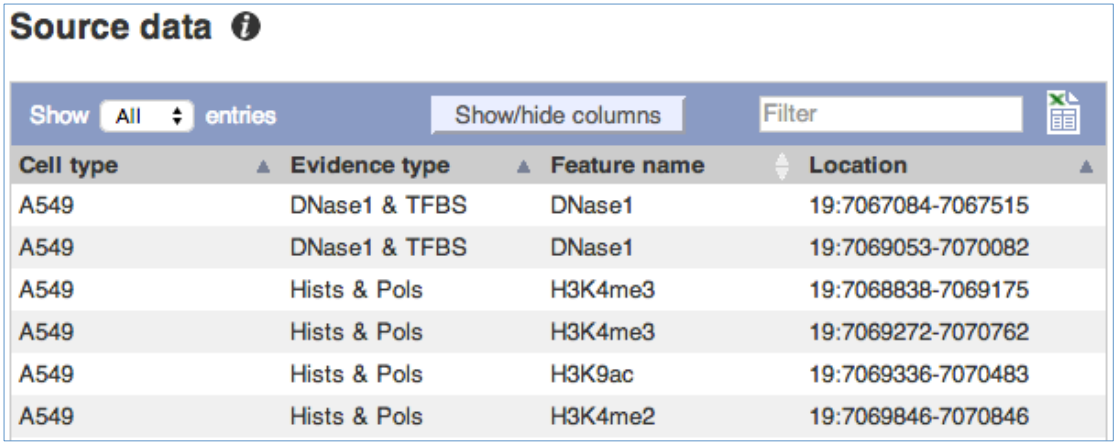
Choose [ALL ON] to get all the possible evidence, then close the menu.

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 DNaseI 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 DNaseI sensitivity in both cell types.

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



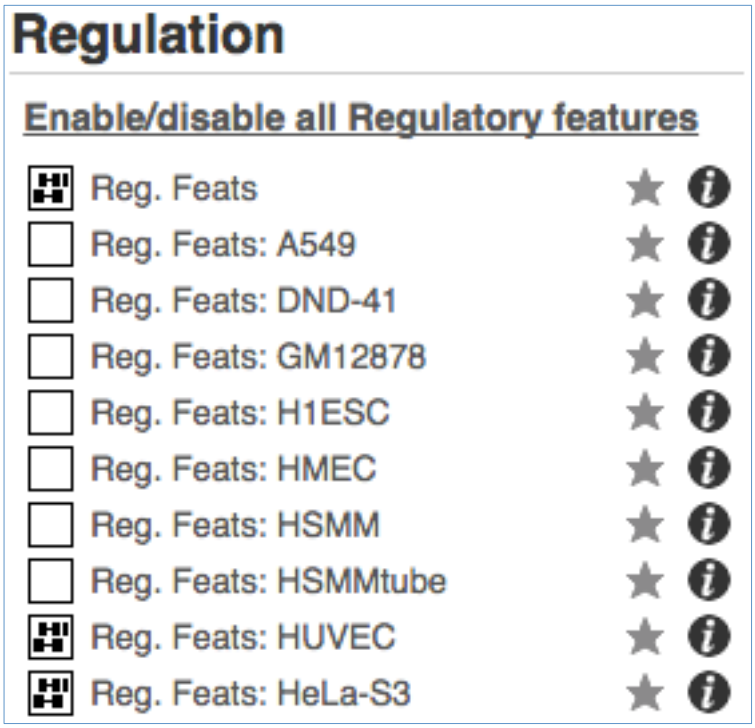
Source data

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

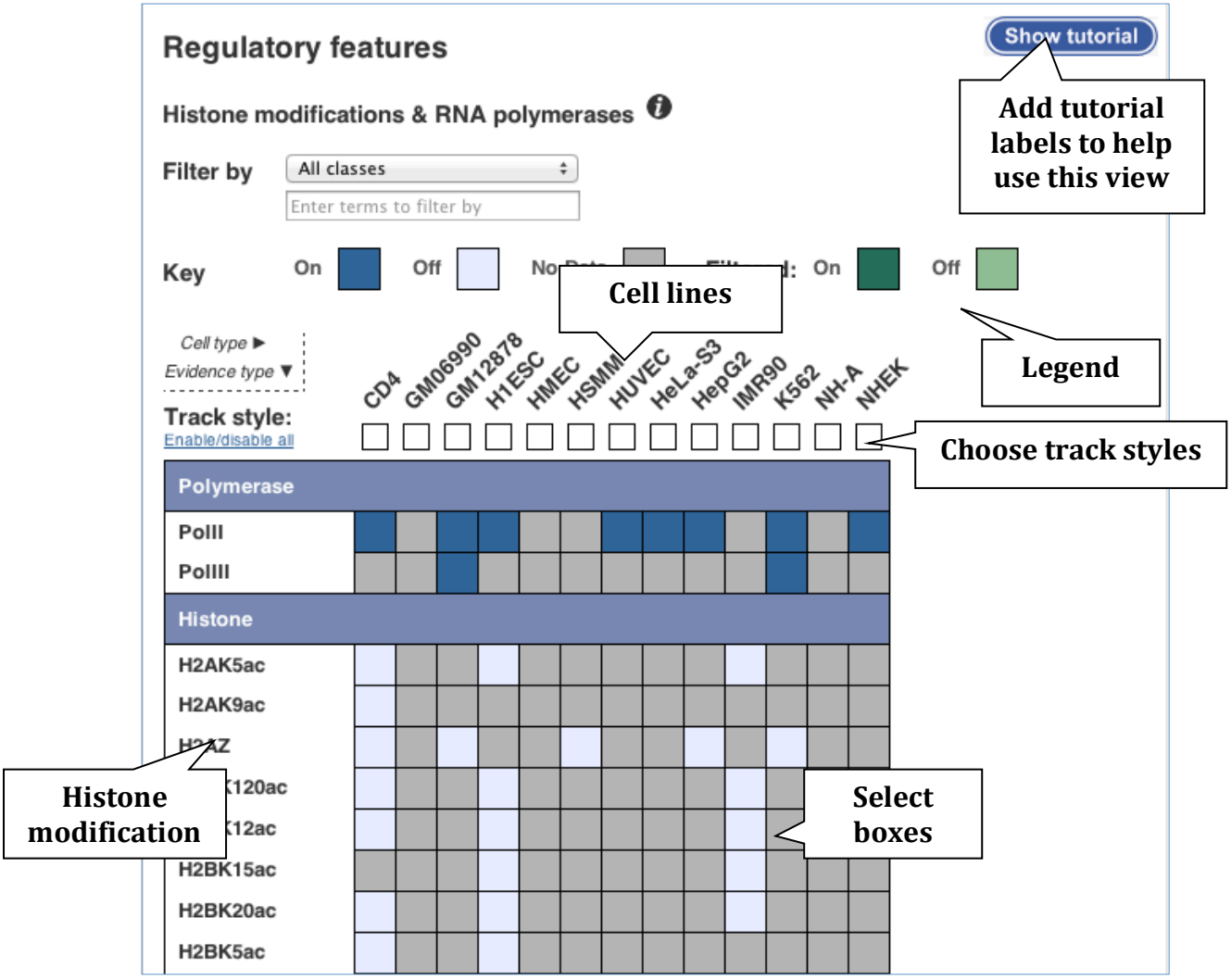
If you're interested in looking at regulatory features in detail across a region, you can do so in the "Location" tab.

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



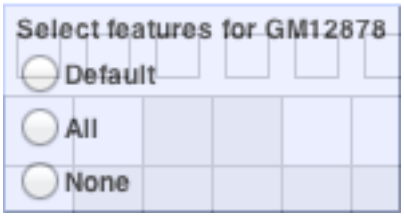
The "MultiCell Reg. Feats" are already on. Turn on the tracks for the "Reg. Feats: HUVEC" and "Reg. Feats: HeLa-S3".

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".

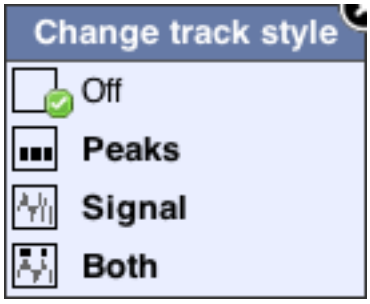


You can turn on a single track by clicking on the box in the matrix. Note that certain tracks are already selected for all cell lines by default (PolII, PolIII, H3K27me3, H3K36me3, H3K4me3, H3K9me3). However, these will appear in the "Region in detail" view only if you specify a track style for the cell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

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

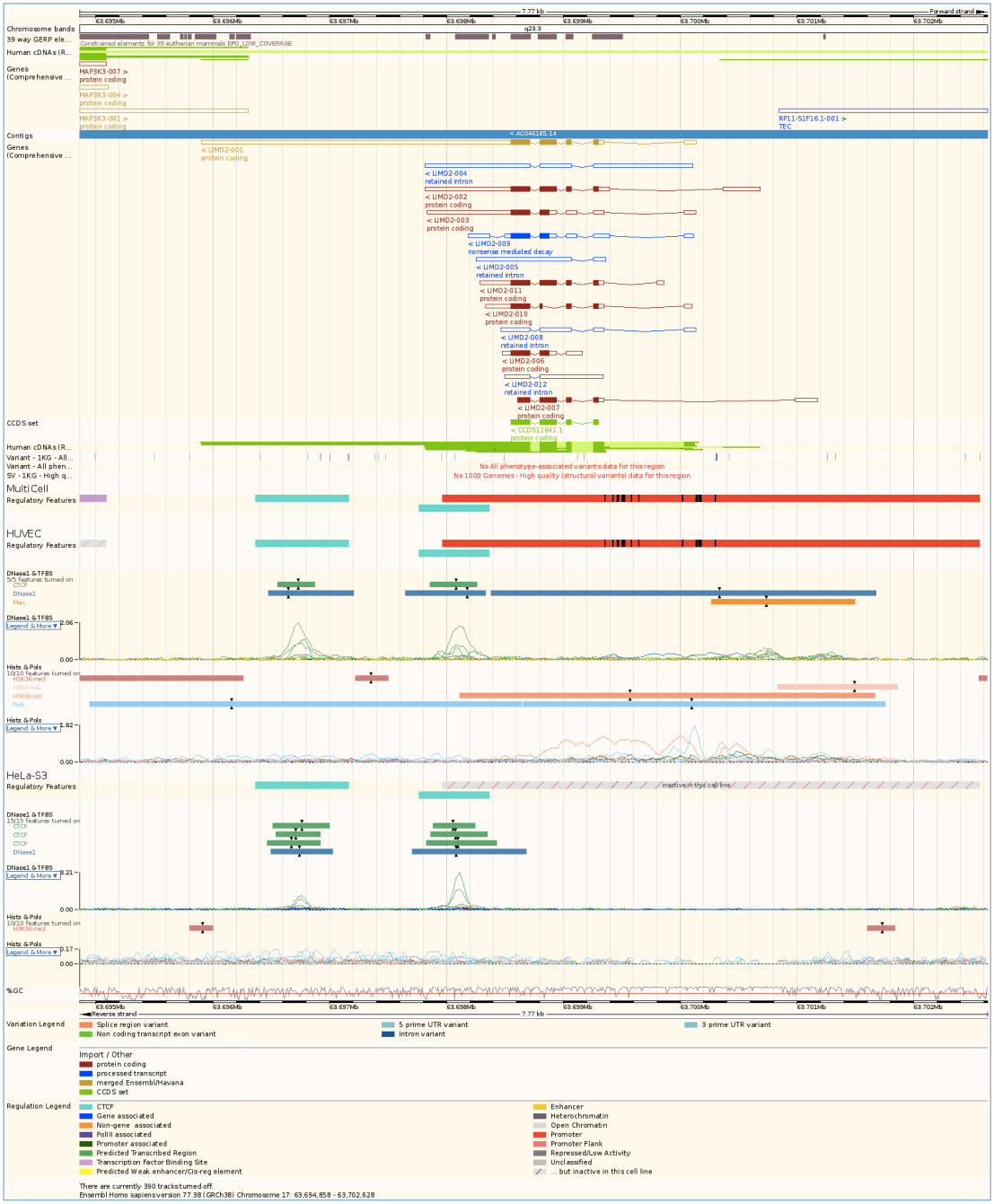


Now choose the track style for the tracks you've switched on. Click on the track style box for HUVEC and HeLa-S3 and select "Both".



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.

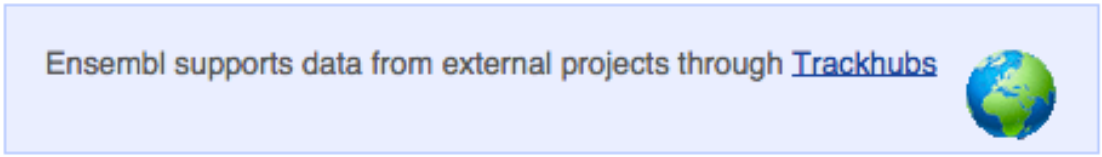


You can also get regulation data in the “Gene” tab, by clicking on “Regulation” in the left-hand menu.

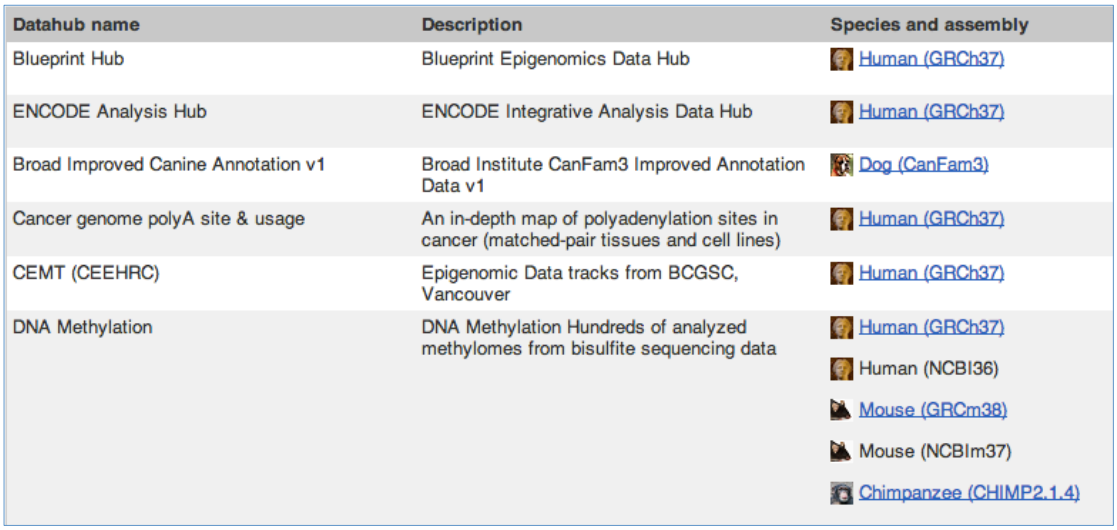
The Ensembl Regulatory Build incorporates data from sources such as ENCODE, Blueprint and the Roadmap Epigenomics project. To see the data directly from these sources, you can add so-called track hubs

Worked example 4: the Ensembl Genome Browser track hubs

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



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



Datahub 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 (NCBIm37) Chimpanzee (CHIMP2.1.4)

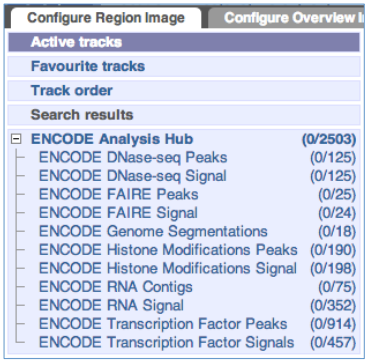
The table contains a brief description of each hub, plus the assembly that the hub is based on, as a link. Click on the link to turn on the hub. If the hub is based on a genome assembly which is not the current assembly in Ensembl, the link will also jump you to an archive with the previous assembly.

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.

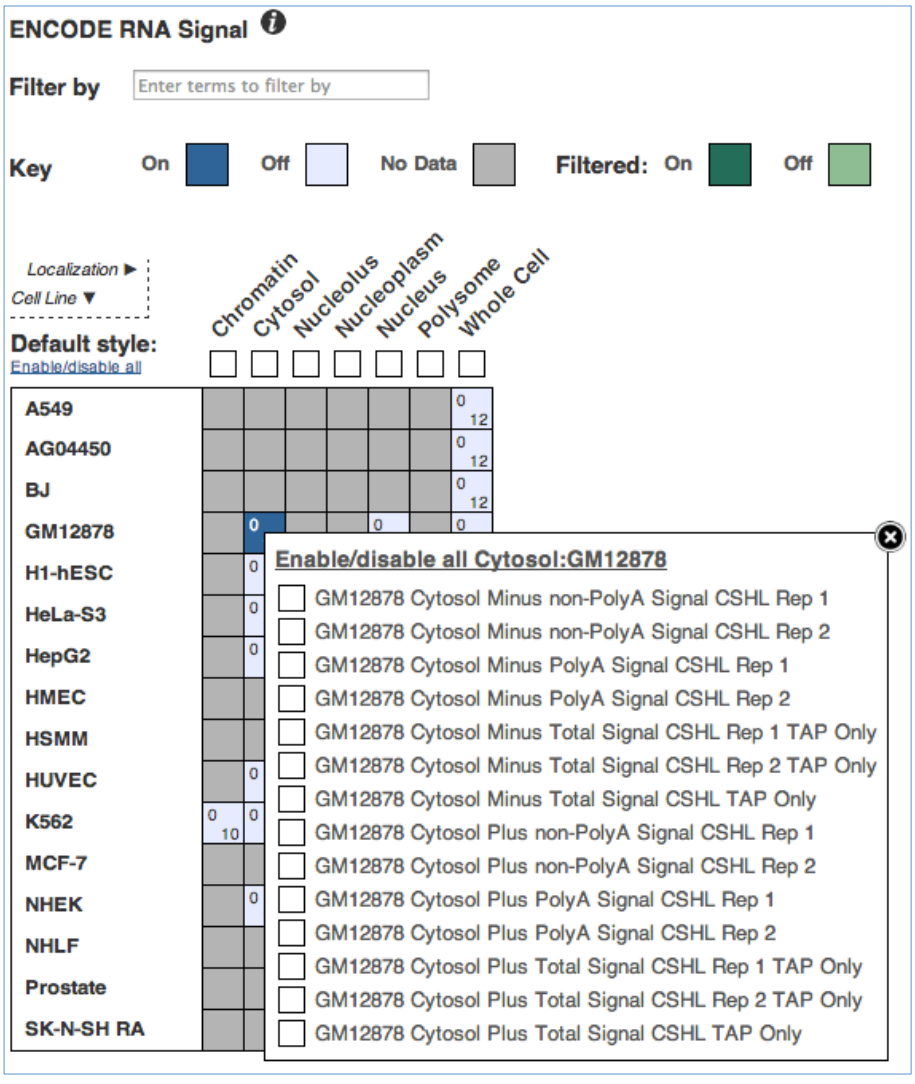
Click on the link “Human (GRCh37)” for the “ENCODE Analysis Hub”.

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.



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.



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 colocate 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*?

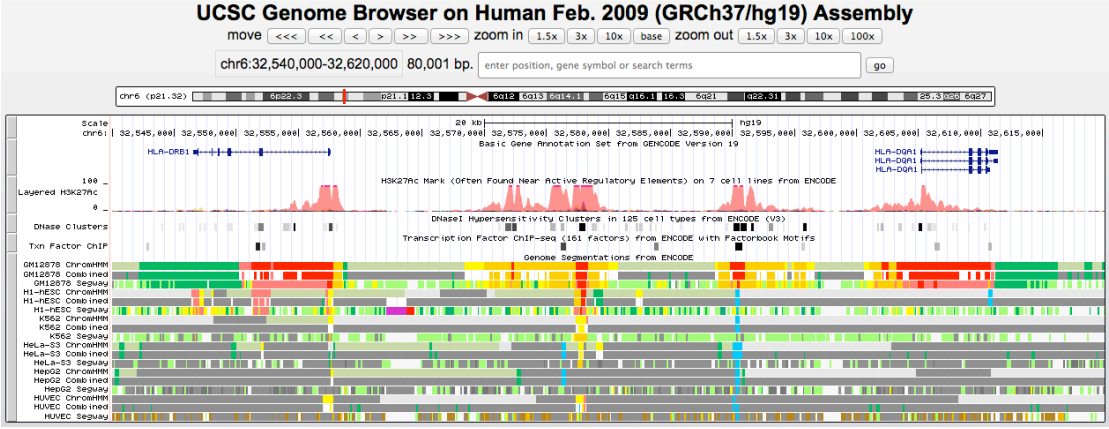
(e) Turn on the "CpG island" track. Are there any CpG islands in the *STX7* gene region?

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 H3K37Ac" track. Two of the CTCF binding sites colocate with a region of high acetylation.

(b) "CTCF enriched elements" are coloured blue in the "Genome Segments" tracks. The CTCF binding sites that colocate 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

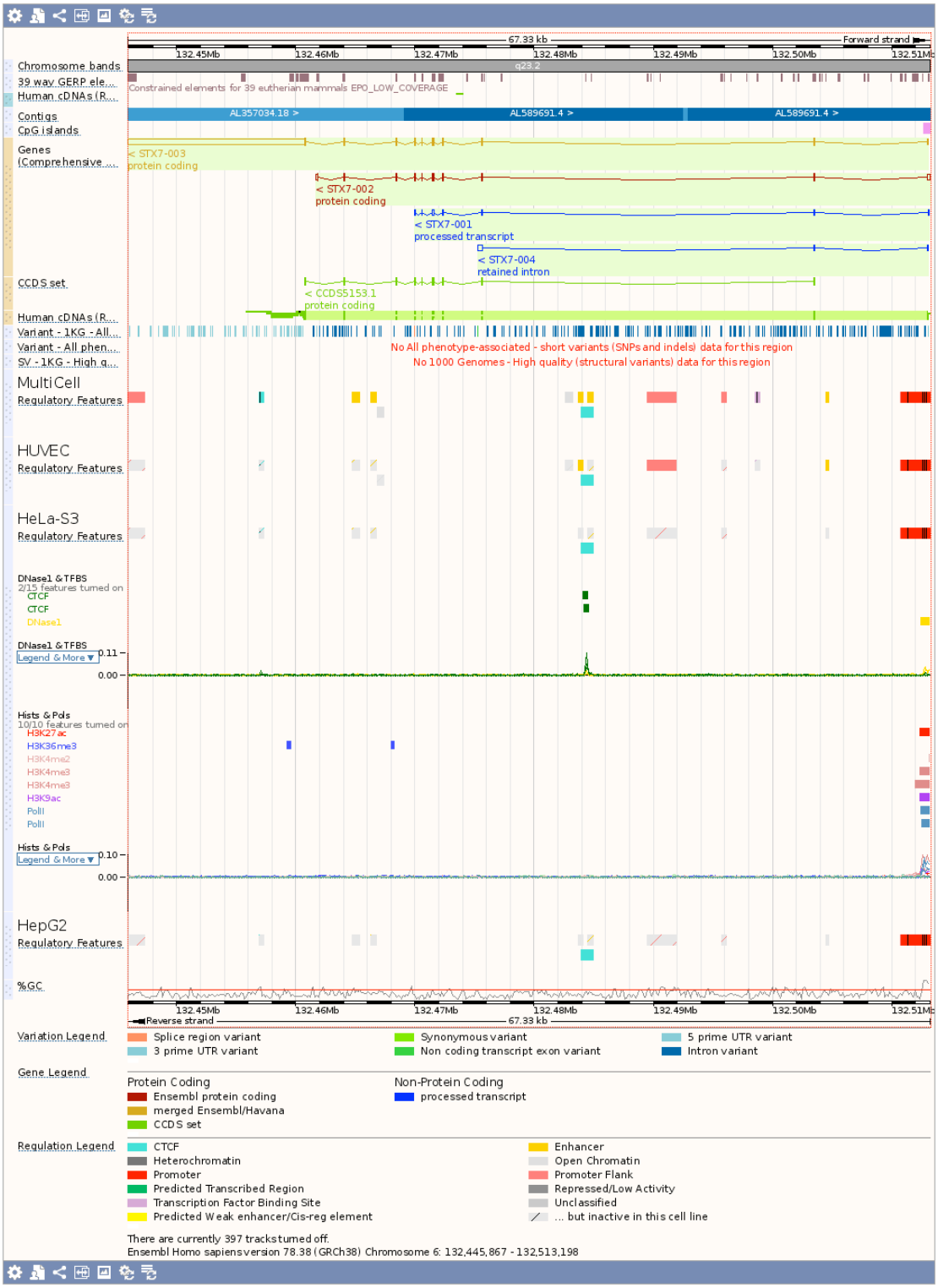
(a) Yes, there are five enhancers (coloured in dark yellow) predicted in the region of the *STX7* gene, two near the 3' end, two near the middle and one near the 5' end.

(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, HsK4me2, H3K4me3 and H3K9ac.

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


67.33 kb
Forward strand
132.45Mb
132.46Mb
132.47Mb
132.48Mb
132.49Mb
132.50Mb
132.51Mb
Chromosome bands
q23.2
39 way GERP ele...
Constrained elements for 39 eutherian mammals EPO_LOW_COVERAGE
Human cDNAs (R...
Contigs
AL357034.18 >
AL589691.4 >
AL589691.4 >
CpG islands
Genes (Comprehensive ...
< STX7-003
protein coding
< STX7-002
protein coding
< STX7-001
processed transcript
< STX7-004
retained intron
CCDS set
< CCDS5153.1
protein coding
Human cDNAs (R...
Variant - 1KG - All...
Variant - All phen...
SV - 1KG - High q...
No All phenotype-associated - short variants (SNPs and indels) data for this region
No 1000 Genomes - High quality (structural variants) data for this region
MultiCell
Regulatory Features
HUVEC
Regulatory Features
HeLa-S3
Regulatory Features
DNase1 & TFBS
2/15 features turned on
CTCF
CTCF
DNase1
DNase1 & TFBS
Legend & More
0.11
0.00
Hists & Pols
1/10 features turned on
H3K27ac
H3K36me3
H3K4me2
H3K4me3
H3K4me3
H3K9ac
PolII
PolII
Hists & Pols
Legend & More
0.10
0.00
HepG2
Regulatory Features
%GC
Reverse strand
67.33 kb
Variation Legend
Splice region variant
3 prime UTR variant
Synonymous variant
Non coding transcript exon variant
5 prime UTR variant
Intron variant
Gene Legend
Protein Coding
Ensembl protein coding
merged Ensembl/Havana
CCDS set
Non-Protein Coding
processed transcript
Regulation Legend
CTCF
Heterochromatin
Promoter
Predicted Transcribed Region
Transcription Factor Binding Site
Predicted Weak enhancer/Cis-reg element
Enhancer
Open Chromatin
Promoter Flank
Repressed/Low Activity
Unclassified
... but inactive in this cell line
There are currently 397 tracks turned off.
Ensembl Homo sapiens version 78.38 (GRCh38) Chromosome 6: 132,445,867 - 132,513,198

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/)