Using Ensembl tools for browsing ENCODE data

Aim

Learn how to search and navigate the Ensembl website with a focus on exploring ENCODE/GENCODE data and data generated by the Ensembl Regulatory Build.

Introduction

The Ensembl project (http://www.ensembl.org) provides genome resources for chordate genomes with a particular focus on human genome data as well as data for key model organisms such as mouse, rat and zebrafish. The total number of supported species is 78 as of Ensembl release 74 (December 2013). Of these, 66 species appear on the main Ensembl website and nine species are provided on the Ensembl preview site (Pre! Ensembl; http://pre.ensembl.org) with preliminary support. For all species on the main site, we provide comprehensive, evidence-based gene annotations and comparative resources including alignments and homology, orthology and paralogy relationships based on Ensembl Gene Trees. We integrate these annotations with a large number of external data sources including InterPro, UniProt and Pfam. Twenty of our most popular species also include dedicated variation resources derived from dbSNP, DGVa and other sources. The Ensembl Regulatory Build provides regulatory annotation on the human and mouse genomes and incorporates data from the ENCODE and Roadmap Epigenomics Projects.

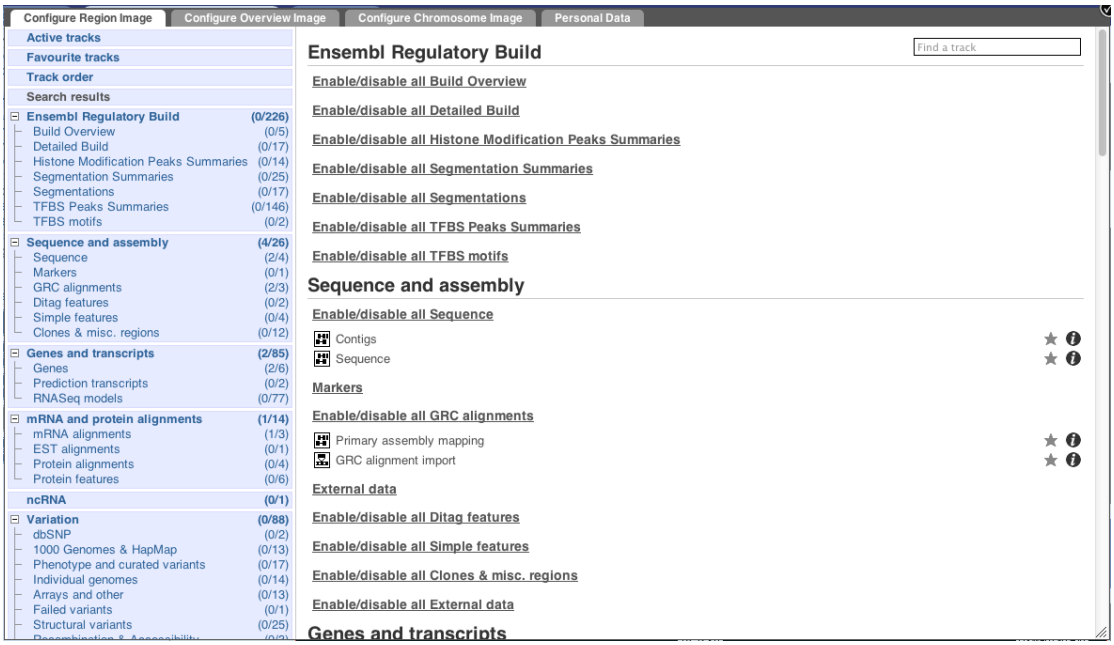
Ensembl data are accessible through an interactive website, flat files, the data retrieval tool BioMart, direct database querying, a set of Perl APIs and a REST API. We support those who use multiple web-based genome bioinformatics sites by providing links to the Vega, UCSC Genome Browser and NCBI's MapViewer on all our Location pages. We also support user data upload and visualisation using BAM, BigWig, VCF and other common data formats.

Further reading

Please see http://www.ensembl.org/info/about/publications.html.

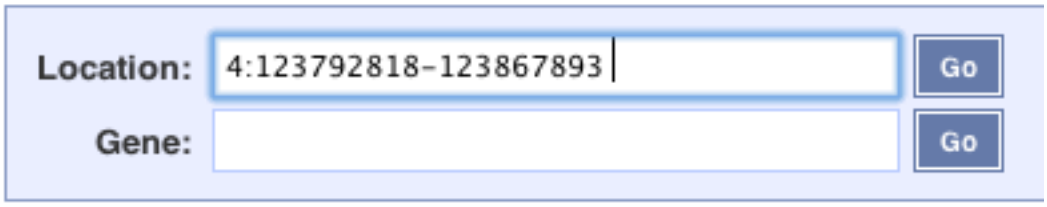
Demo: ENCODE data in the Ensembl genome browser

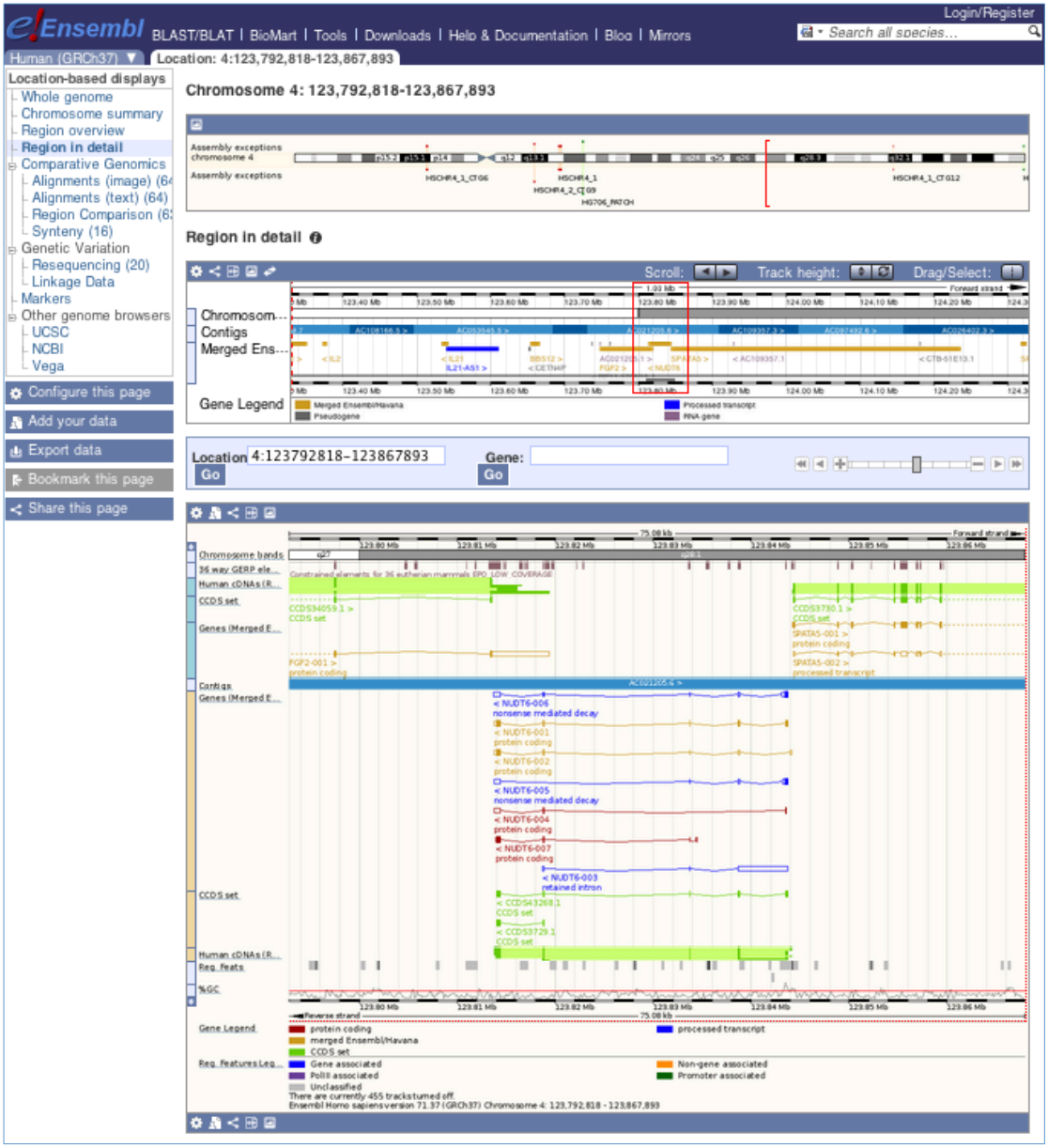
In order to demo the updated regulatory display, we need to use a track hub. To load this hub into Ensembl go to: http://www.ensembl.org/Homo_sapiens/Location/View?g=ENSG00000130544;contigviewbottom=url:http://ngs.sanger.ac.uk/production/ensembl/regulation/hub.txt;format=DATAHUB;menu=Ensembl%20Regulatory%20Build#modal_config_viewbottom-Ensembl_Regulatory_Build_BuildOverview



This will take us to an Ensembl Region in Detail page, with a Configure menu open. Let's close this menu and go to a region we're interested in.

Put 4:123792818-123867893 into the Location box and click Go.

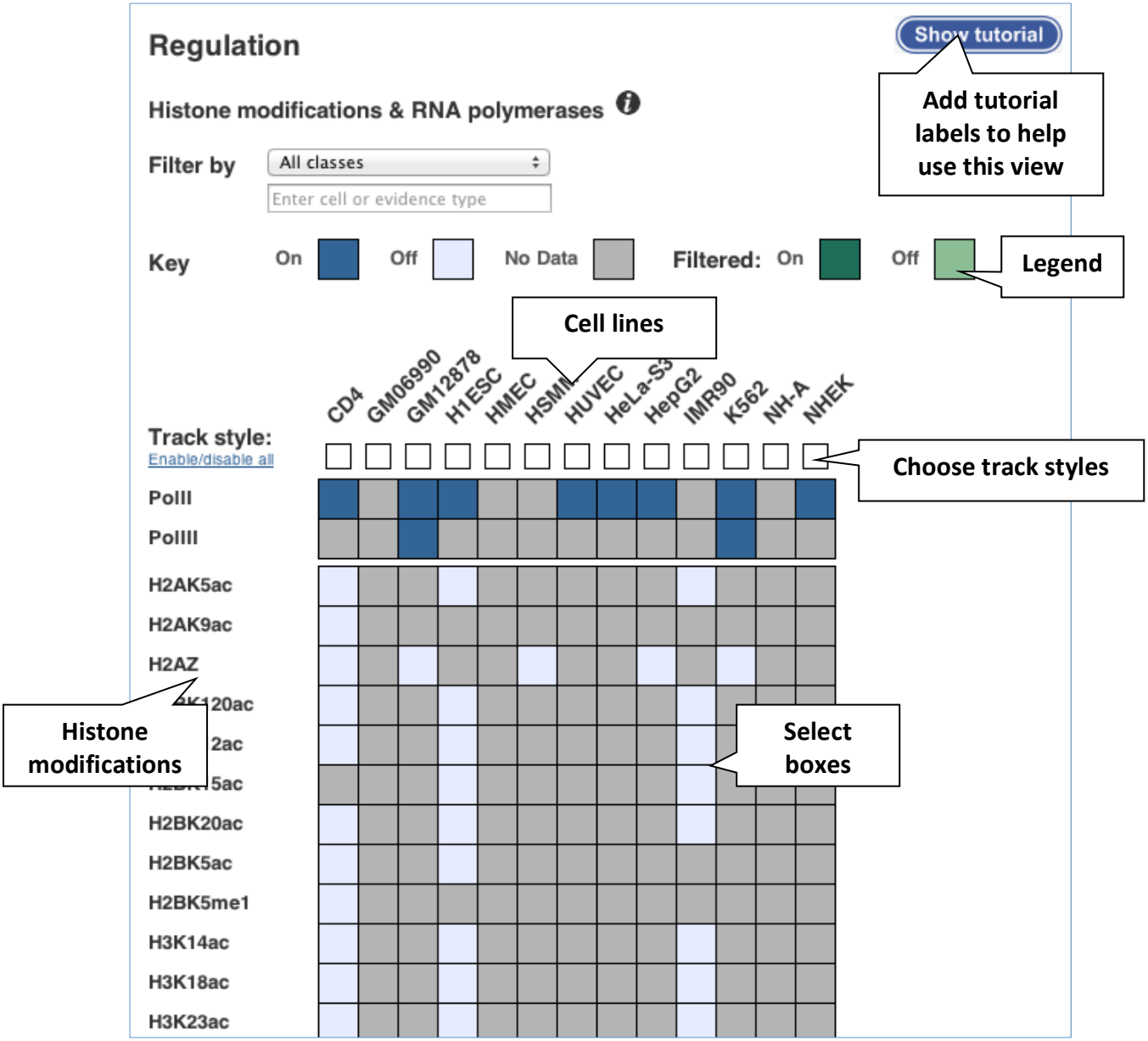




This is the page we looked at earlier with the GENCODE gene set.

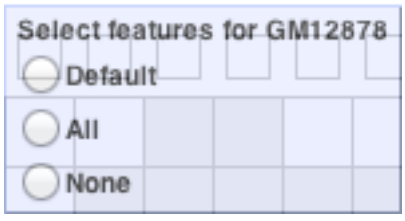
Open Configure this page to add some tracks. First, we're going to add ChIP-seq data for histone modifications and polymerase binding.

Click on Histones & polymerases under Regulation in the left-hand menu. These data are part of the existing regulatory build, so are partway down the menu.

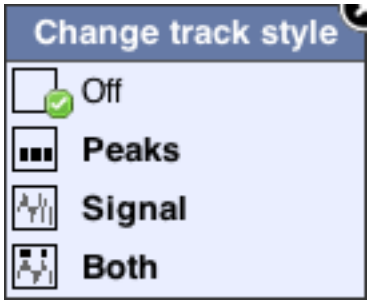


You can turn on a single track by clicking on the box in the matrix. Note that certain tracks are selected for all cell lines by default (PolII, PolIII, H3K27me3, H3K36me3, H3K4me3, H3K9me3). These will appear in the Region in detail view only if you specify a track style for the cell lines.

Turn on all the tracks for GM12878. Hover over the cell line name then select All.

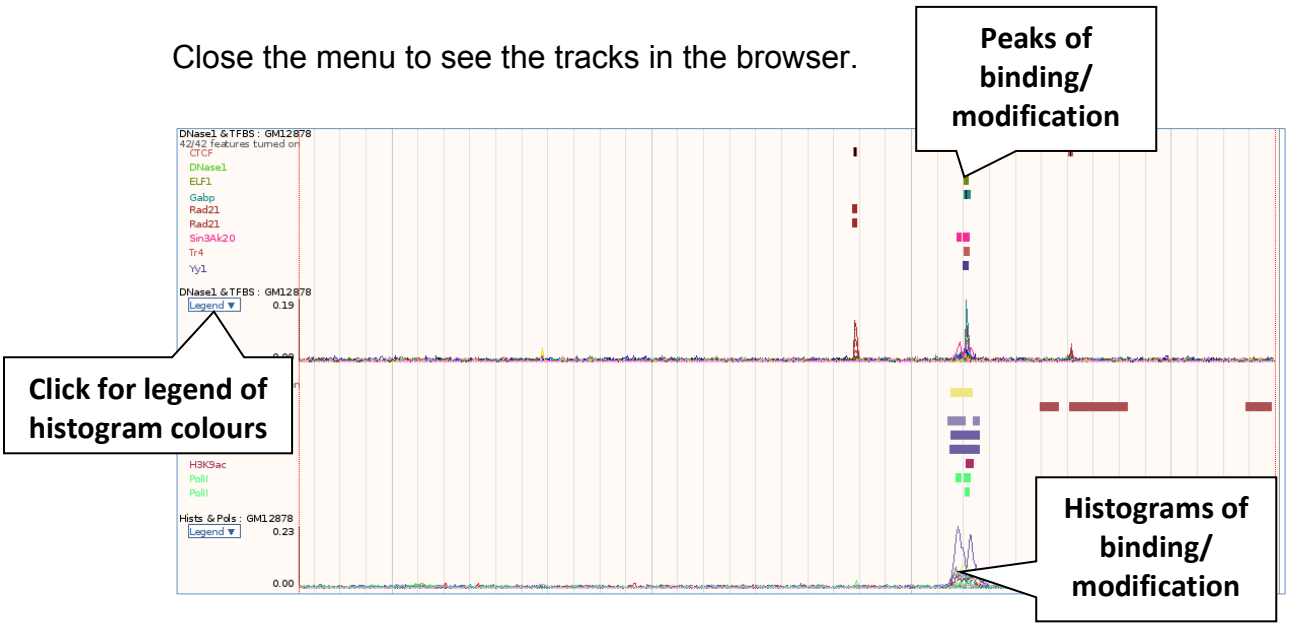


Now choose the track style for the tracks you've switched on. Click on the track style box for GM12878 and select Both.



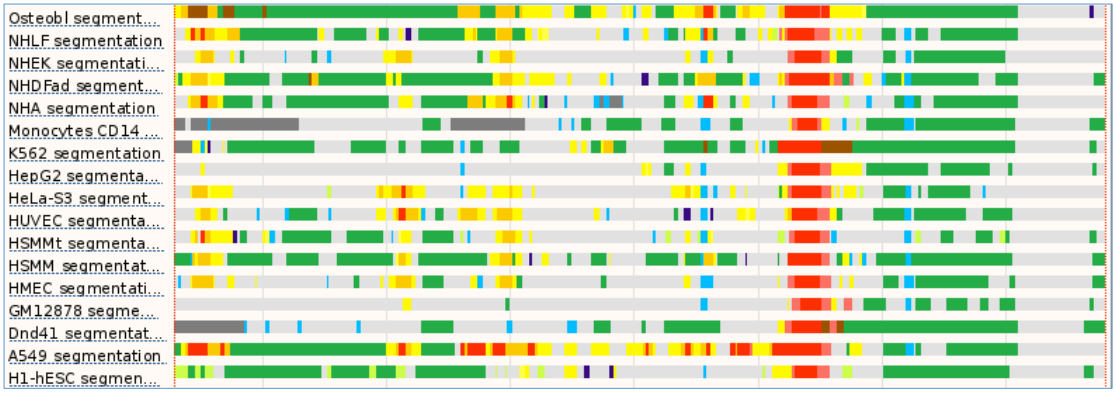
There is a similar matrix for Open chromatin &TFBS. Use this to turn on all tracks for GM12878 in Both.

Close the menu to see the tracks in the browser.



This data is processed to produce the segmentation in the different cell lines. Let's have a look at this. Open the Configure this page menu again.

The segmentation data are part of the new style of data, so are part of the track hub we added and are at the top of the menu. Click on Segmentations under Ensembl Regulatory Build. Click on Enable/disable all Segmentations then select Compact. This will turn on the segmentations for all cell lines.



We can see the coloured segments that describe different chromatin features. The colours follow the following code:

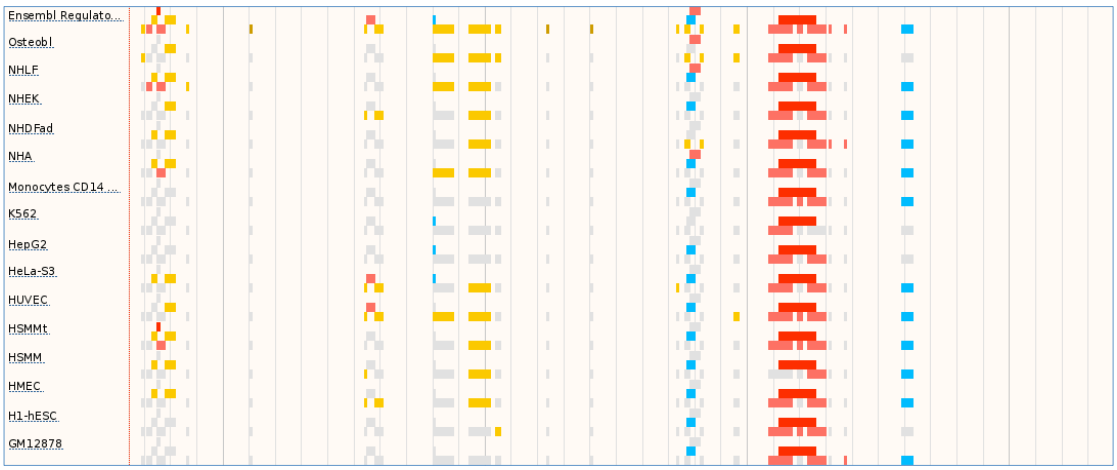
Transcribed region **Strong distal element** **Promoter**
CTCF enriched **Weak distal element** **Promoter flank**
DNase site **5' transcribed genebody** **Polycomb repressed**
Inactive region

You can click on any of the segments to get a pop-up describing them. When this data is released fully, they will have links to explore further.

Compare the histone modification and TFBS tracks to the segments. To do this, pick up and drag the tracks to be side-by-side. Can you see the correlation between them?

We can see the build that these segments produced. Configure this page then add the Ensembl Regulatory build in Normal under Build Overview. Open Detailed Build and click on Enable/disable all Detailed Build, selecting Normal.

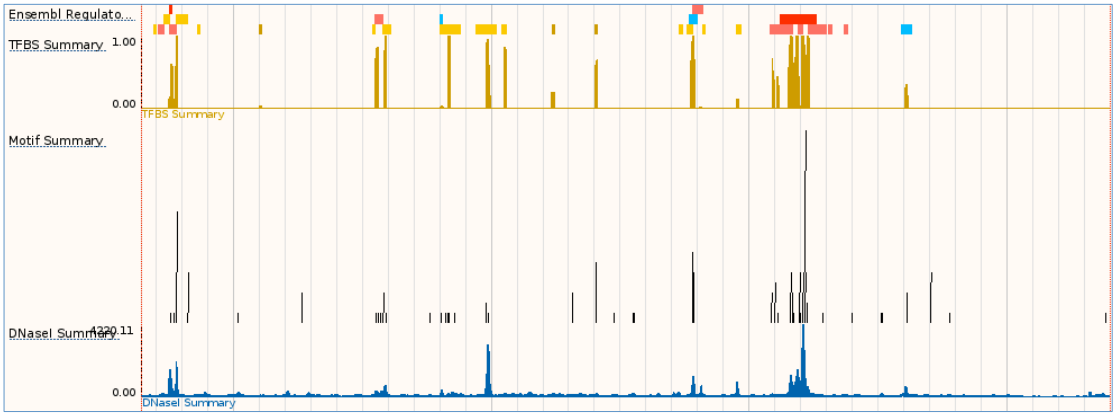
Drag the Ensembl Regulatory Build track so that it directly above the Detailed Build tracks.



The regulatory build indicates all the features known across the genome. They follow the same colour code as the segments. Under this we have the detailed build, which indicates whether that feature in active in the cell lines.

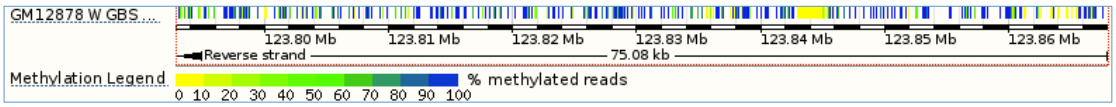
We can see that the all of the features line up perfectly between the different tracks, but are shown either coloured or in grey. The colours indicate that the activity occurs in that cell line, whilst grey indicates inactivity.

Summaries of transcription factor binding and DNase I sensitivity across all cell lines and binding motifs are also available. Click on Configure this page and find them all under Build overview. Turn on DNase I summary and TFBS summary in Wiggle, and Motif summary in Separate.



You can find out more about the individual binding motifs by clicking on them.

You can also add methylation data using Configure this page. Find it under DNA methylation and turn on GM12878 RRBS ENCODE and GM12878 WGBS ENCODE.



Now that you've got the view how you want it, you might like to show something you've found to a colleague or collaborator. Click on the Share this page button to generate a link. Email the link to someone else, so that they can see the same view as you, including all the tracks you've added. These links contain the Ensembl release number, so if a new release or even assembly comes out, your link will just take you to the archive site for the release it was made on.

To return this to the default view, go to Configure this page and select Reset configuration at the bottom of the menu.

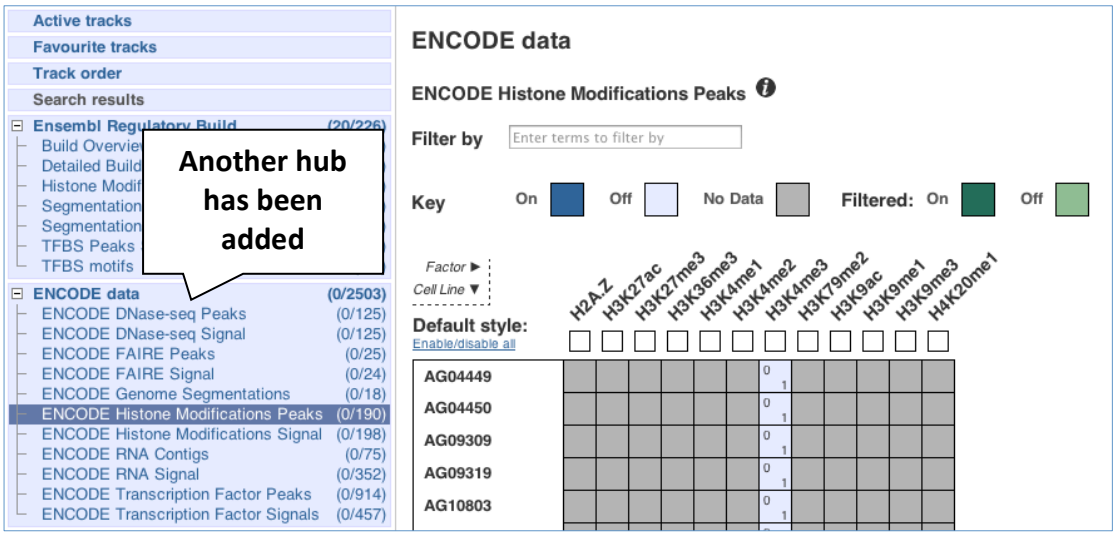
This only shows you a subset of ENCODE data. To get the full set of ENCODE data, you need to add the ENCODE track hub.

Go to the front page of Ensembl (ensembl.org) and click on the ENCODE icon.

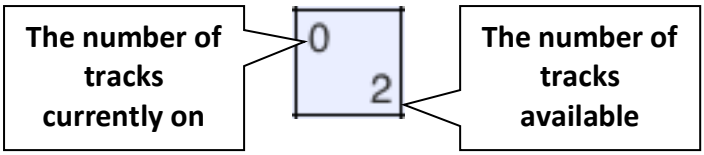
This page contains information about the ENCODE data and how it is incorporated into Ensembl.

Add the ENCODE hub by clicking on the Link to add the ENCODE track hub.

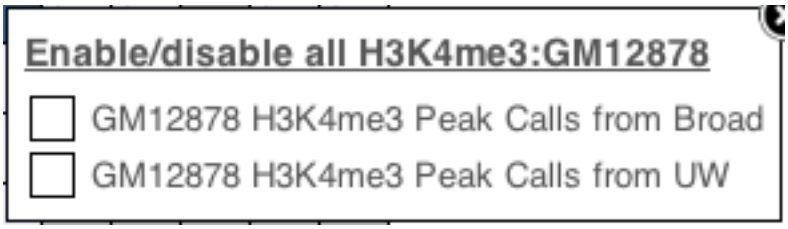
This will take you directly to the matrices for adding ENCODE data to the Region in detail view.



This gives us another hub to add tracks from. These give us matrices which work in the very same way as the tracks in Ensembl. You'll see the boxes have numbers on them.



This occurs when there is more than one dataset displaying the same data from different sources. Click on a box with a number greater than 1 to get a pop-up and select your tracks of interest.



Exercises

Exercise 1 – Gene regulation: Human STX7

(a) Find the Location tab (Region in detail page) for the *STX7* gene.

(b) Click Configure this page and on the Ensembl Regulatory Build menu in the left hand side. Turn on the Ensembl Regulatory Build and Detailed Build data for HUVEC, HeLa-S3, and HepG2 cell types. Do any of these cells show predicted enhancer activity in the *STX7* region?

(c) Use Configure this page to add supporting data indicating open chromatin for HeLa-S3 cells. Are there sites enriched for marks of open chromatin (DNase1 and FAIRE) in HeLa cells at the 5' end of *STX7*?

(d) Configure this page once again to add histone modification supporting data for the same cell type as above (e.g.HeLa-S3). Which ones are present at the 5' end of *STX7*?

(e) Find the same histone modifications in HeLa-S3 in the ENCODE track hub. Do any of them have more than one dataset? Turn on the tracks where there's more than one. Is there a difference between the same data from multiple datasets? Why would this be?

(f) Is there any data to support methylated CpG sites in this region (5' end) of *STX7* in B-cells?

(g) Create a Share link for this display. Email it to yourself then open the link.

Exercise 2 – Regulatory features in human

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.

(a) Go to the region from 32,540,000 to 32,620,000 bp on human chromosome 6

(b) Are there any CTCF enriched regions annotated in the intergenic region between the *HLA-DRB1* and *HLA-DQA1* genes?

(c) Are there CTCF binding motifs in these regions? Are there any other motifs that colocalise with the CTCF motifs?

(d) Has the CTCF binding detected at these position been observed in all cell/tissue types analysed?

(e) Have a look at the Regulatory supporting evidence - Histones & Polymerases configuration matrix. For which cell/tissue type are the most histone acetylation data sets available? In this cell/tissue type, is the region that shows CTCF binding also a region of high acetylation, as found by Majumder *et al*?

Exercise Answers

Exercise 1 – Gene regulation: Human STX7

(a) Search for human gene *STX7* from the home page. Click on Location in the search results.

(b) The predicted enhancer segments are coloured in golden yellow. Two are active in the HUVEC cell type only (out of the three cells chosen).

(c) Configure this page and click on Open chromatin &TFBS. Turn on both peaks and signal for DNase 1 and FAIRE in HeLa-S3 cells (the boxes in this configure this page window will turn blue. For more information on how to select and view the supporting data, click on Show tutorial in the pop up window). Close the menu.

There are two DNase 1 hypersenstive sites in the 5' exon of *STX7*. Click on the coloured block to find out that the DNase1 enriched sites in HeLa-S3 cells come from the ENCODE project. There is no FAIRE site known in this region.

(d) Configure this page and click on Histones & polymerases. Change the Filter by menu from All classes to Histone. Select the all the histone modifications available for HeLa cells (some of them might be on by default). Save and close the menu.

H3K4me3, H3K9ac and H3K27ac sites have been found in the 5' region of *STX7* in HeLa-S3 cells.

(e) Configure this page and click on ENCODE Histone Modifications Peaks. Type hela into the Filter by box to narrow down to HeLa-S3.

H3K4me3 has two datasets, from Broad and UW.

Select both in Normal. Close the menu.

There are peaks for both in the same location, however they differ in size. This is due to the nature of the experiments themselves. Nucleosomes are 140bp wide, so a single histone modification will cover a wide range. This gives blurry edges, meaning that different reads will give slightly different results.

(f) Click on configure this page and choose the DNA Methylation menu. Scroll down to Enable/disable all External data then turn on the first track in the list (MeDIP-chip B-cells). Save and close the menu.

The CpG sites at the 5' end of *STX7* are not highly methylated (note the yellow/green bars). Yellow, green, and blue bars represent unmethylated, intermediately methylated, and methylated regions, respectively. For more information on human DNA methylation DAS tracks, see:
www.ensembl.org/info/docs/funcgen/index.html

(g) Click Share this page in the side menu.
Select the link and copy.
Go into your email account and compose an email to yourself.
Paste the link in, then send.
Open the email and click on your link.

Exercise 25 – Regulatory features in human

(a) Go to the Ensembl homepage (http://www.ensembl.org/).
Select Search: Human and type 6:32540000-32620000 in the search box.
Click Go.

You may want to turn off all tracks that you added to the display in the previous exercises as follows:

Click Configure this page in the side menu.
Click Reset configuration.
SAVE and close.

(b) Click Configure this page in the side menu.
Click on Ensembl Regulatory Build – Build Overview.
Select Ensembl Regulatory Build – Normal.
SAVE and close.

Yes, there are two CTCF-enriched regions 32590001-32590800 and 32576201-32576600.

(c) Click Configure this page in the side menu.
Click on Ensembl Regulatory Build – Build Overview.
Select Motif summary – Separate with labels.
SAVE and close.

Yes, in each there is a CTCF binding motif, at 32540403-325890415 and 32576368-32576380. The one at 32540403-325890415 colocalises with Rad21, RFX5, ZNF143 and BCLAF1 motifs.

(d) Click Configure this page in the side menu.
Click on Regulation – Open chromatin and TFBS.
CTCF is already selected by default. Turn on all cell lines in Peaks.
You may also want to turn off the tracks for DNase I and FAIRE.
SAVE and close.

CTCF binding has been detected at the position of the 32590001-32590800 CTCF enriched region in eleven of the cell/tissue types analysed. (CD4, GM06990, GM12878, H1ESC, HMEC, HSMM, HUVEC, HeLa-S3, HepG2, NH-A, NHEK)

(e) Click Configure this page in the side menu.
Click on Regulation – Histones & polymerases.

According to the Histones & Polymerases configuration matrix the most information on histone acetylation is available for CD4 cells.

Hover over CD4 in the Histones & Polymerases configuration matrix.
Select Select features for CD4 - All.
SAVE and close.

Yes, the region that shows CTCF binding is also a region of high acetylation of histone 2A, 2B, 3 and 4 in CD4 cells.

Learn More

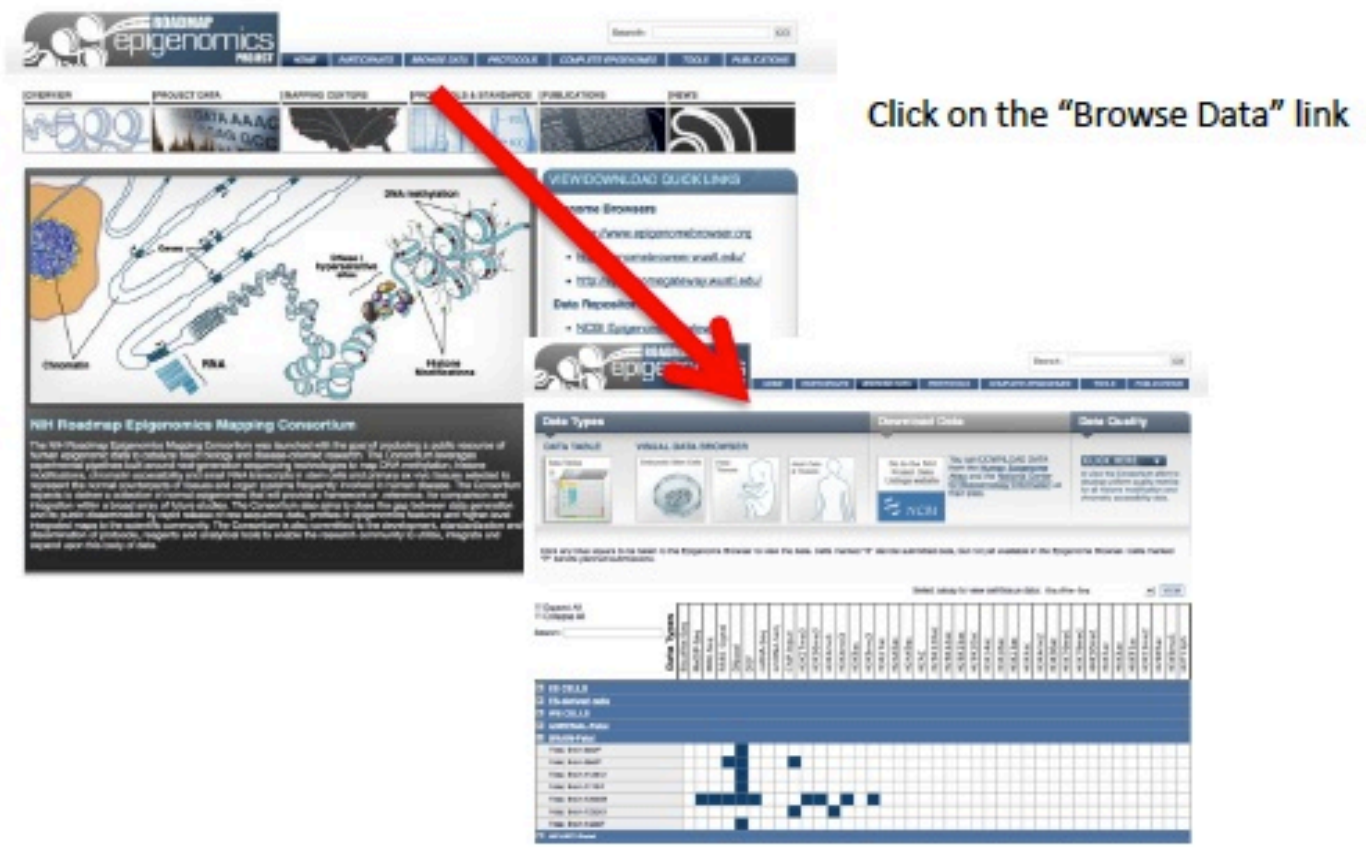
Visit our blog posts at:
http://www.ensembl.info/blog/2013/12/26/the-new-ensembl-regulatory-annotation/
http://www.ensembl.info/blog/2014/01/01/computing-ensembls-new-regulatory-annotation/

See also:
L. Ward and M. Kellis (2012) Interpreting noncoding genetic variation in complex traits and human disease. Nat. Biotech. 30:1095-1106 (http://www.nature.com/nbt/journal/v30/n11/full/nbt.2422.html)
The ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nat. 489:57-74 (http://www.nature.com/nature/journal/v489/n7414/pdf/nature11247.pdf)
B.E. Bernstein et al. (2010) The NIH Roadmap Epigenomics Mapping Consortium. Nat. Biotech. 28:1045-48 (http://www.nature.com/nbt/journal/v28/n10/abs/nbt1010-1045.html)
J. Ernst et al. (2012) ChromHMM: automating chromatin-state discovery and characterization. Nat. Meth. 9:215-16 (http://www.nature.com/nmeth/journal/v9/n3/abs/nmeth.1906.html)
P. Carninci et al. (2006) Genome-wide analysis of mammalian promoter architecture and evolution. Nat. Genet. 38:626-35 (http://www.nature.com/ng/journal/v38/n6/full/ng1789.html)
A. Visel et al. (2006) VISTA Enhancer Browser - a database of tissue-specific human enhancers. Nucl. Acids Res. 35:D88-D92 (http://nar.oxfordjournals.org/content/35/suppl_1/D88.short)
A. Mathelier et al. (2013) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucl. Acids Res. In press. (http://nar.oxfordjournals.org/content/early/2013/11/04/nar.gkt997.full)
L. Arbiza et al. (2013) Genome-wide inference of natural selection on human transcription factor binding sites. Nat. Genet. 45:723-9 (http://www.nature.com/ng/journal/v45/n7/full/ng.2658.html)
M.M. Hoffman et al. (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat. Meth. 9:473-6 (http://www.nature.com/nmeth/journal/v9/n5/full/nmeth.1937.html)
J. Harrow et al. (2012) GENCODE: The reference human genome annotation for The ENCODE Project. Genome Res. 22:1760-74 (http://genome.cshlp.org/content/22/9/1760.short)
D. Adams et al. (2012) BLUEPRINT to decode the epigenetic signature written in blood. Nat. Biotech. 30:224-6. (http://www.nature.com/nbt/journal/v30/n3/full/nbt.2153.html)

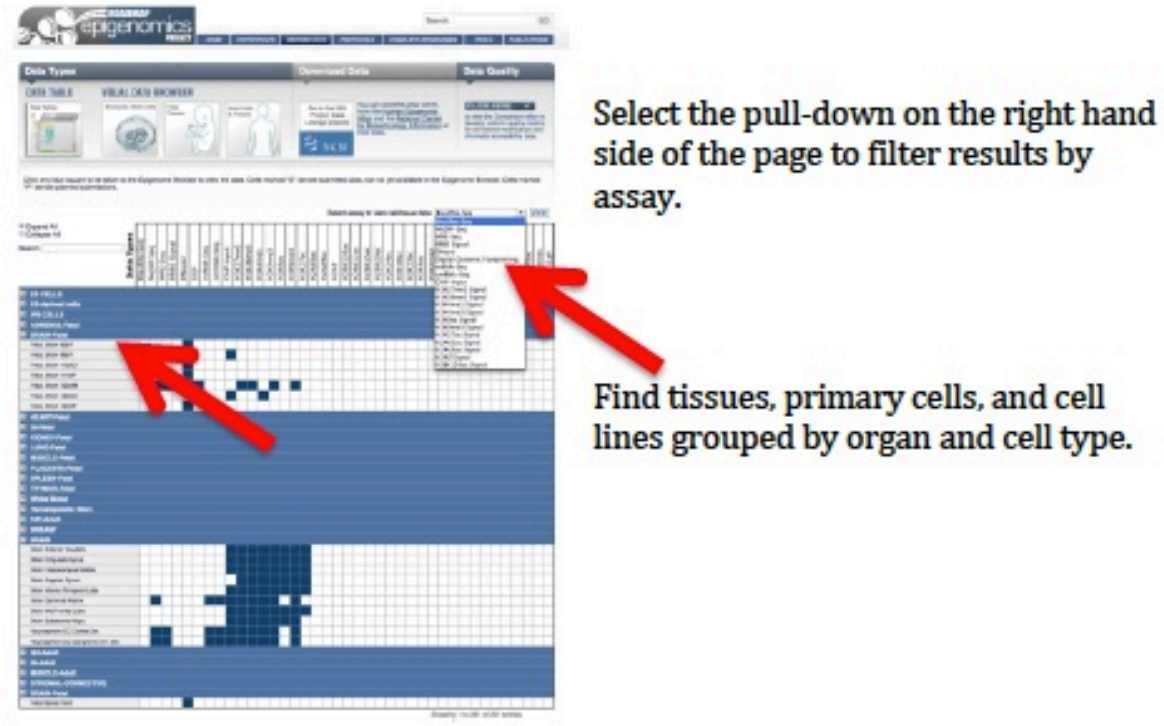
Accessing Roadmap Epigenomics Data

http://www.genome.gov/27555330

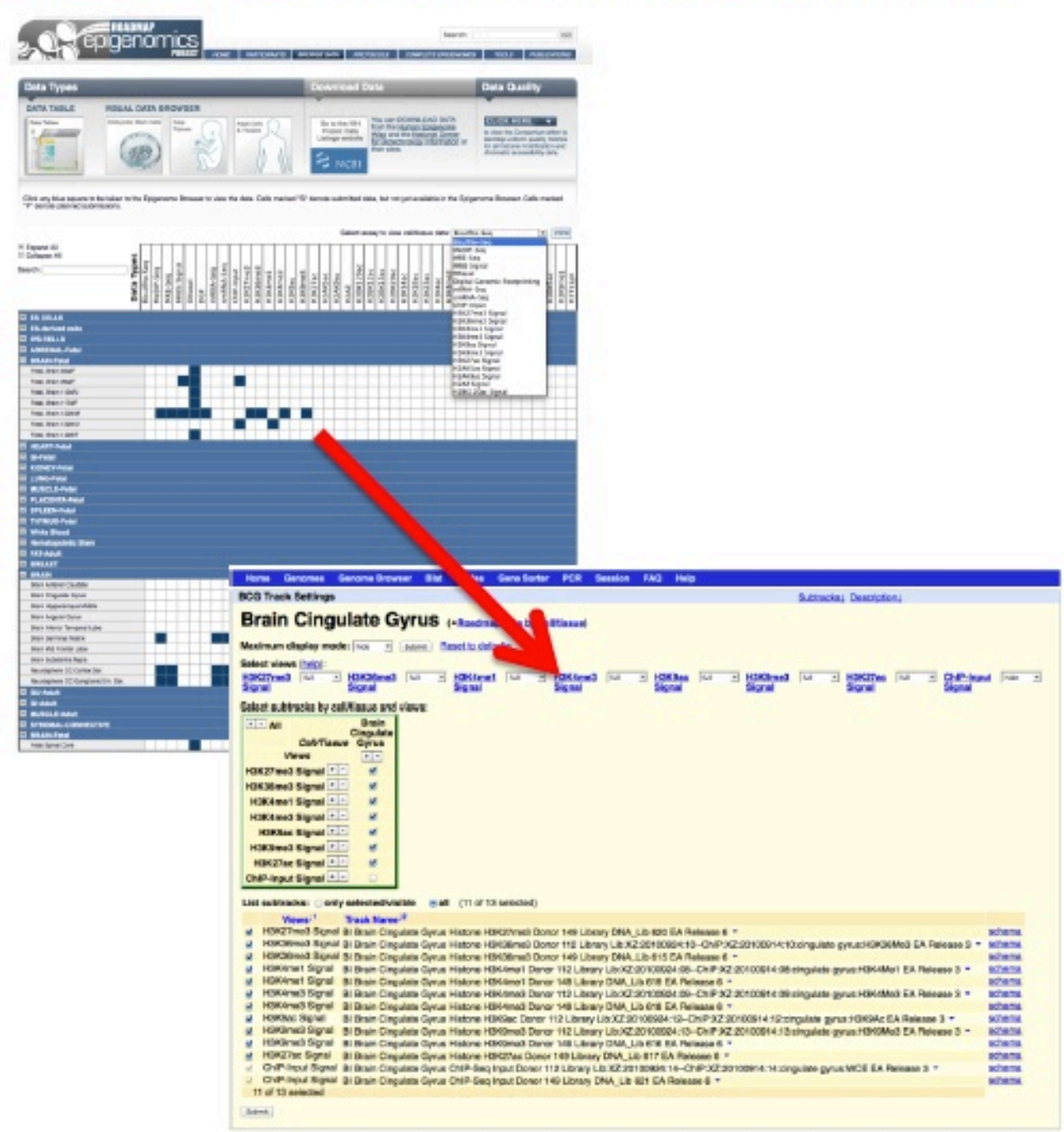
1. Go to http://roadmapepigenomics.org



2. Select tissue and assays



3. Click on a blue box to view data for that combination of cell type and assay



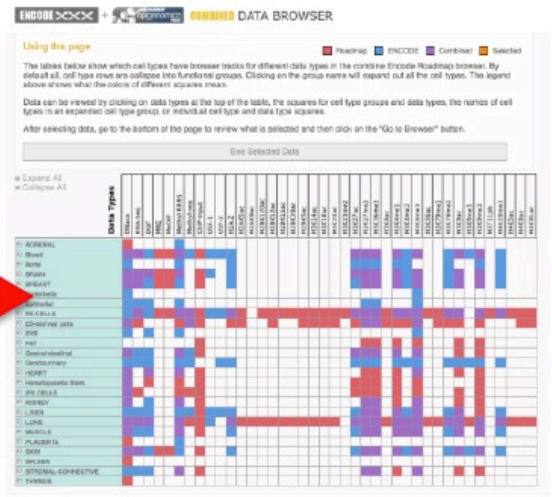
Finding Human Tissues in ENCODE and Roadmap Epigenomics Data

http://www.genome.gov/27555330

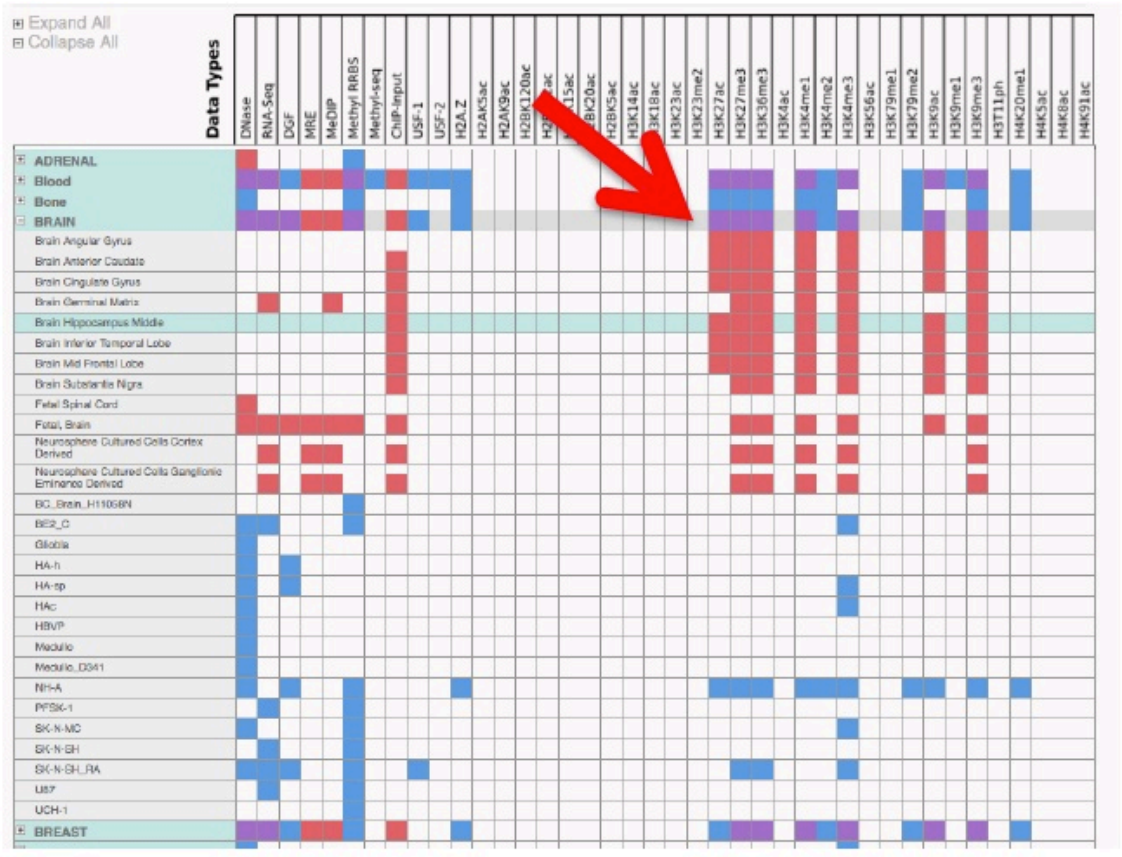
The ENCODE and Roadmap Epigenomics projects have a wide variety of data available for a wide assortment of cell types.

1) Go to http://www.encode-roadmap.org

The colored boxes indicate where there is data available from the ENCODE project or the Roadmap Epigenomics project. Click on the tissue or organ name to expand the available tissues.

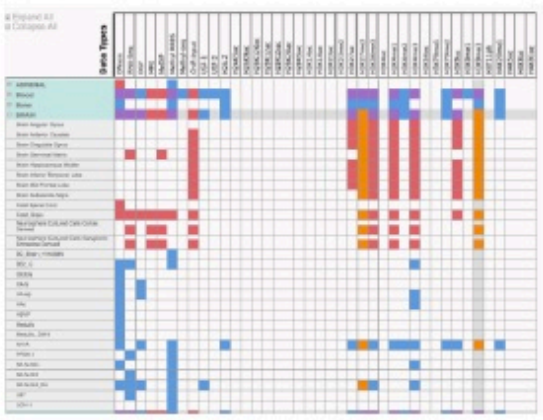


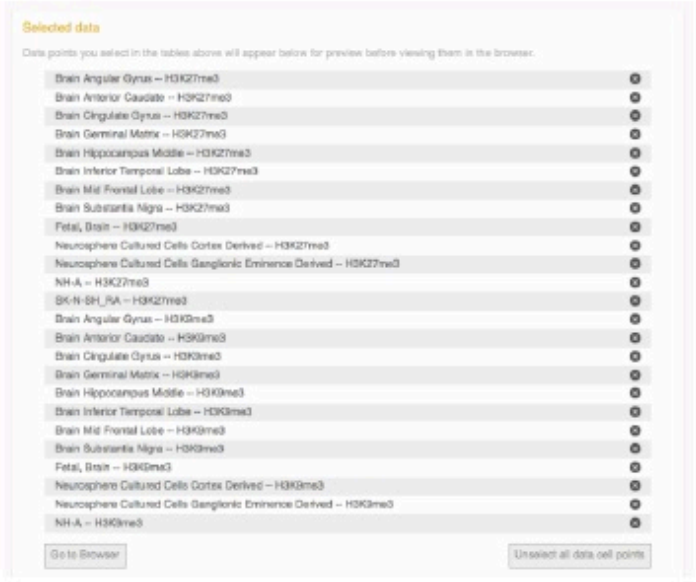
2) Select the datasets you want to see



Find the cell type (tissue, primary cell, or cell line) on the vertical axis of the matrix. Find the assay or target on the horizontal axis of the matrix. Click on the box to select the dataset you want to view.

Clicking on a box for a header, such as brain, will select all cell types listed under that heading. Selected datasets are highlighted in orange.





Selected datasets also appear as a list on the bottom of the page.

4) Click the "View in Browser" button to see track displayed. Then click the left bar to click through to the related Track Setting page.

There is a selection of ENCODE associated resources available from the NHGRI ENCODE website, which may be found here

http://www.genome.gov/27553900

A selection of these is found in the appendix section of the manual.