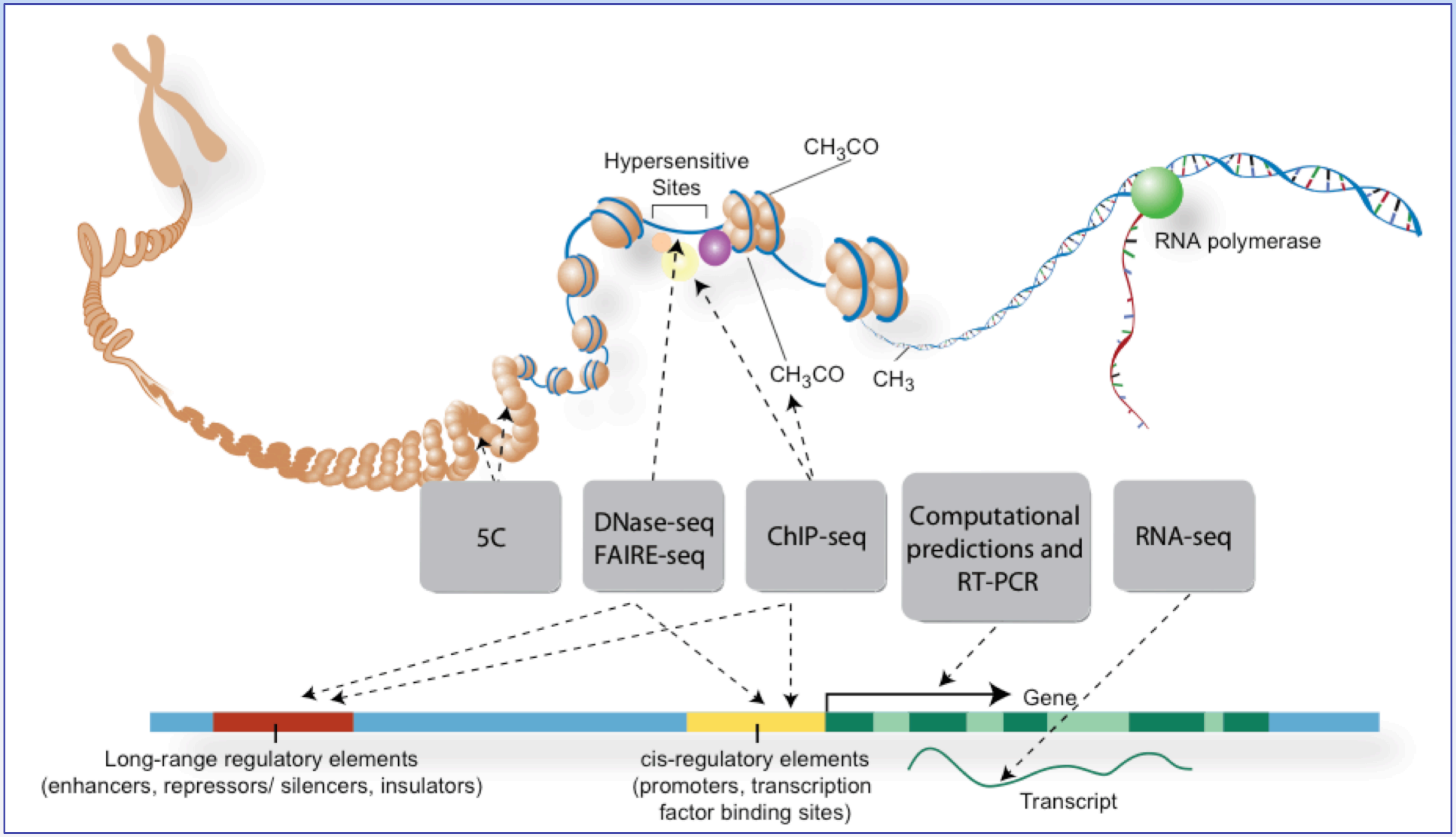
Introduction to the ENCODE DCC


Hypersensitive Sites
CH_3CO
RNA polymerase
CH_3CO
CH_3
5C
DNase-seq FAIRE-seq
ChIP-seq
Computational predictions and RT-PCR
RNA-seq
Gene
Long-range regulatory elements (enhancers, repressors/ silencers, insulators)
cis-regulatory elements (promoters, transcription factor binding sites)
Transcript

Jim Kent and Kate Rosenbloom – University of California Santa Cruz

ENCODE Project

- Not to be confused with ENCODE pilot project that just covered 1% of human genome. Current ENCODE is full genome on human and mouse.
- 32 biology labs organized into 19 grants, plus an Analysis Working Group and a Data Coordination Center (DCC)
- I'm the principal investigator of the DCC
- ENCODE's overall goal is to identify and characterize all functional elements of the genome.
- ENCODE DCC's job is to make data accessible and clear, to put it in UCSC Genome Browser, and to help other databases at NCBI, EBI, and elsewhere import ENCODE data as well.

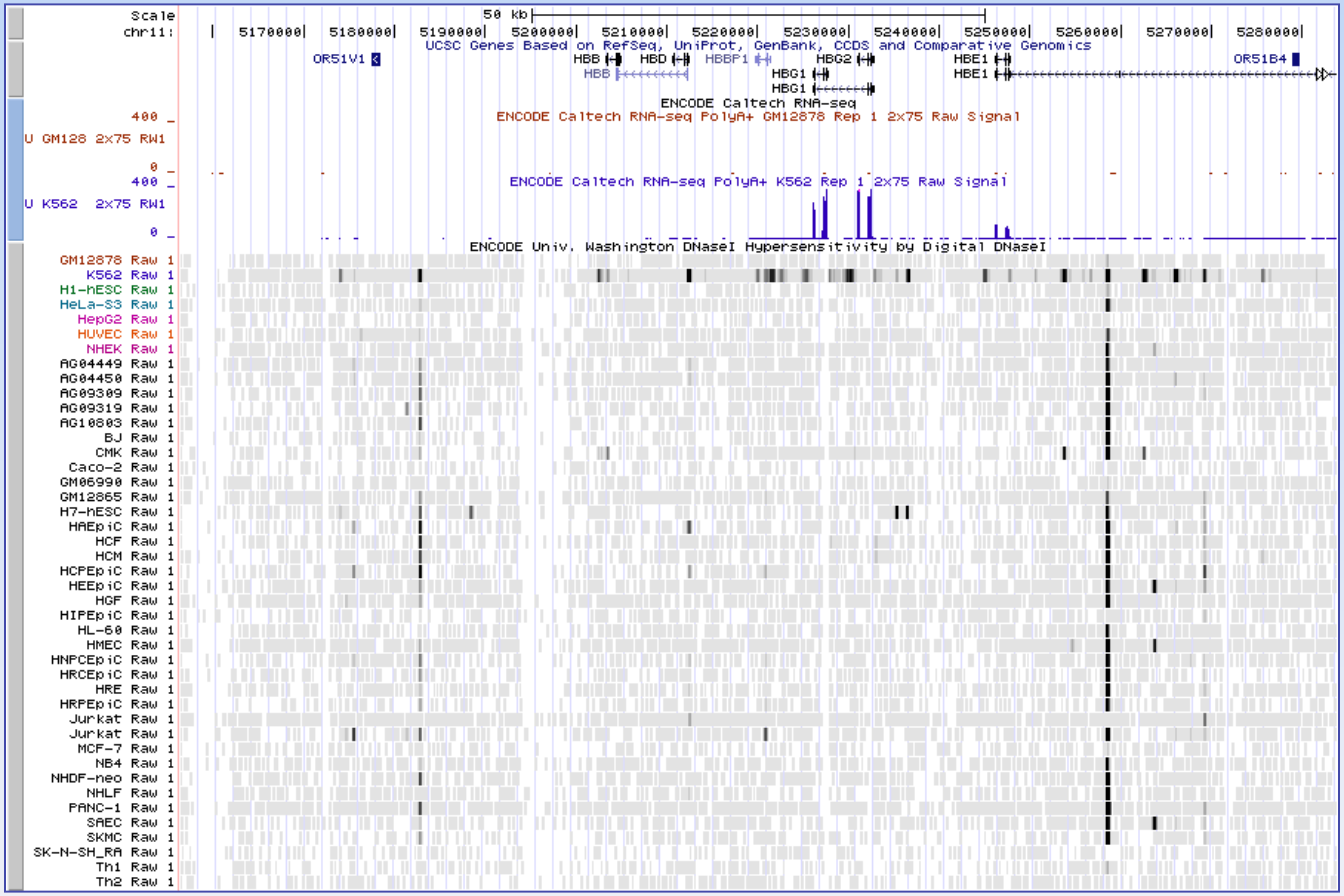
ENCODE assays on regulation of transcription

- Opening/closing chromatin
  - DNase hypersensitivity
  - Chromatin immunoprecipitation & sequencing (ChIP-seq) of histone marks
- Binding expressive/inhibitory transcription factors.
  - ChIP-seq of various transcription factors
- RNA transcription (or not)
  - mRNA sequencing of ENCODE cell lines
  - Exotic RNA sequencing – short/long polyA+/- localized to nucleus, cytoplasm, polysome, nucleoplasm, nuclear matrix, mitochondria, etc.

ENCODE DNase Hypersensitivity

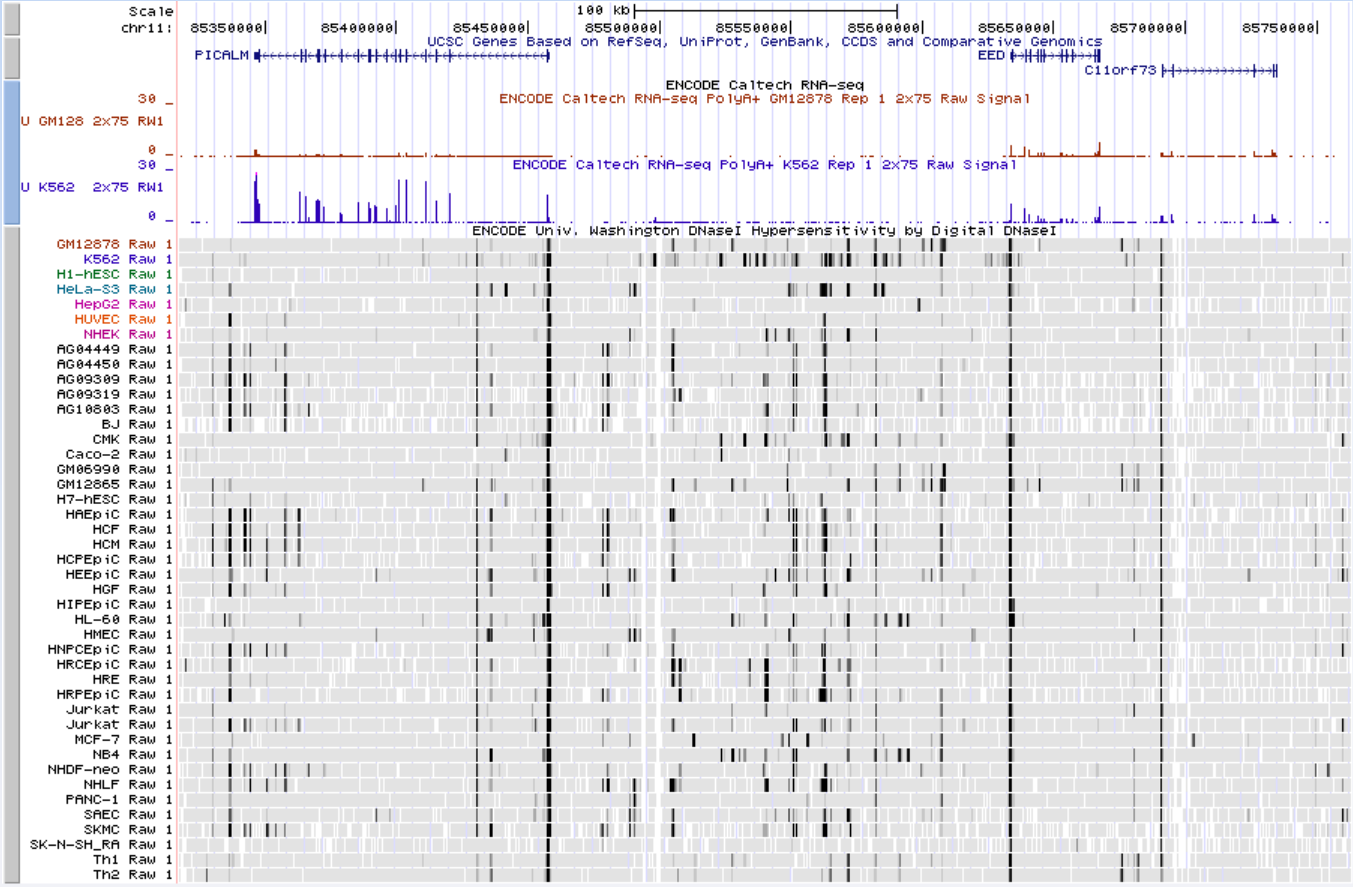
- Several genome-wide high throughput methods being used in ENCODE. All involve DNA-seq
- Data currently available for 388 cell lines and tissues
- Main artifacts to watch for:
  - DNA present in cell in multiple copies:
    - Mitochondria, centromeric repeats, other repeats
    - Generally such regions ignored except in "raw" data.
  - Sequencing biases (highly g/c rich regions etc.)
  - In general artifacts easier to work around than those associated with DNA-chip based assays.

UW DNaseI at Hemoglobin Beta



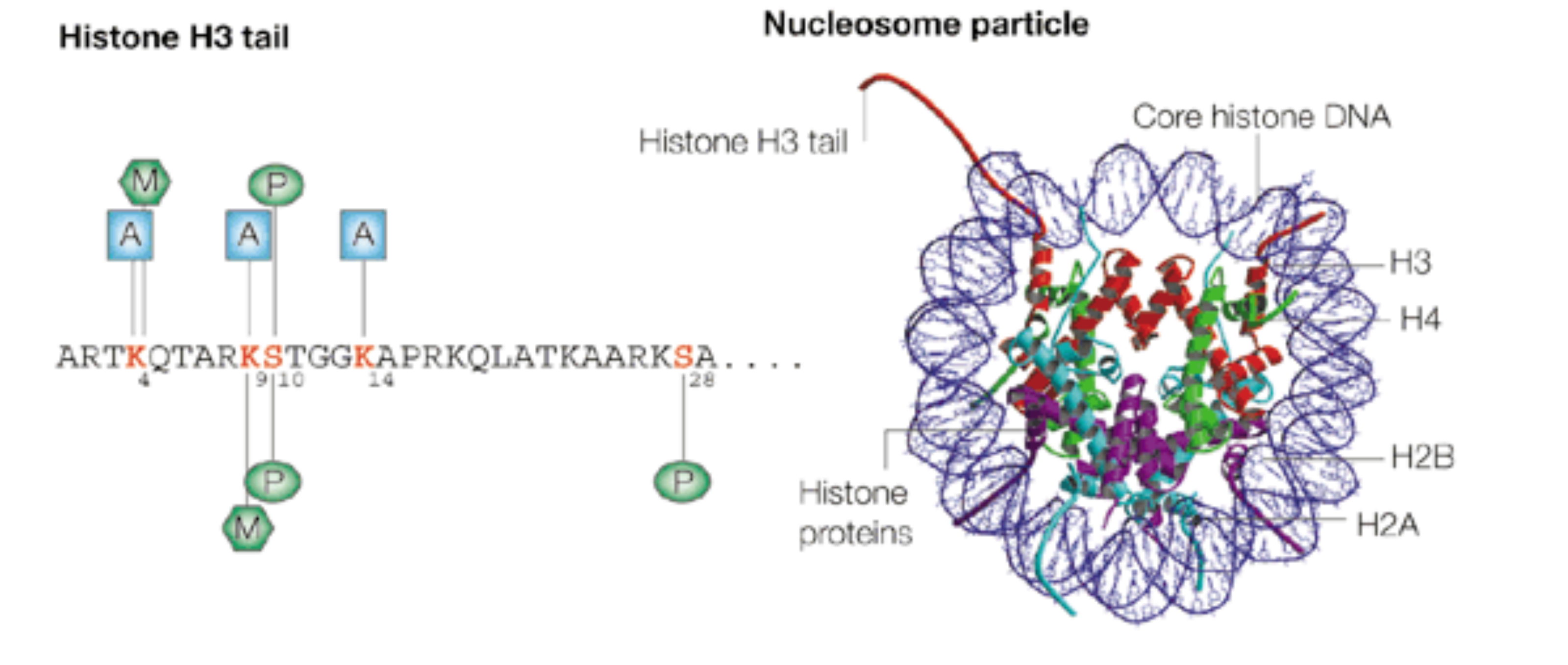
Top track shows genes in the Hemoglobin beta (HBB) locus.
Next track shows RNA levels in GM12878 and K562 cell lines.
The last track is density plots of DNAse hypersensitivity in many cell lines.
K562, a cell line similar to a red blood cell precursor, shows much RNA and DNAase activity.

A more typical locus - PICALM



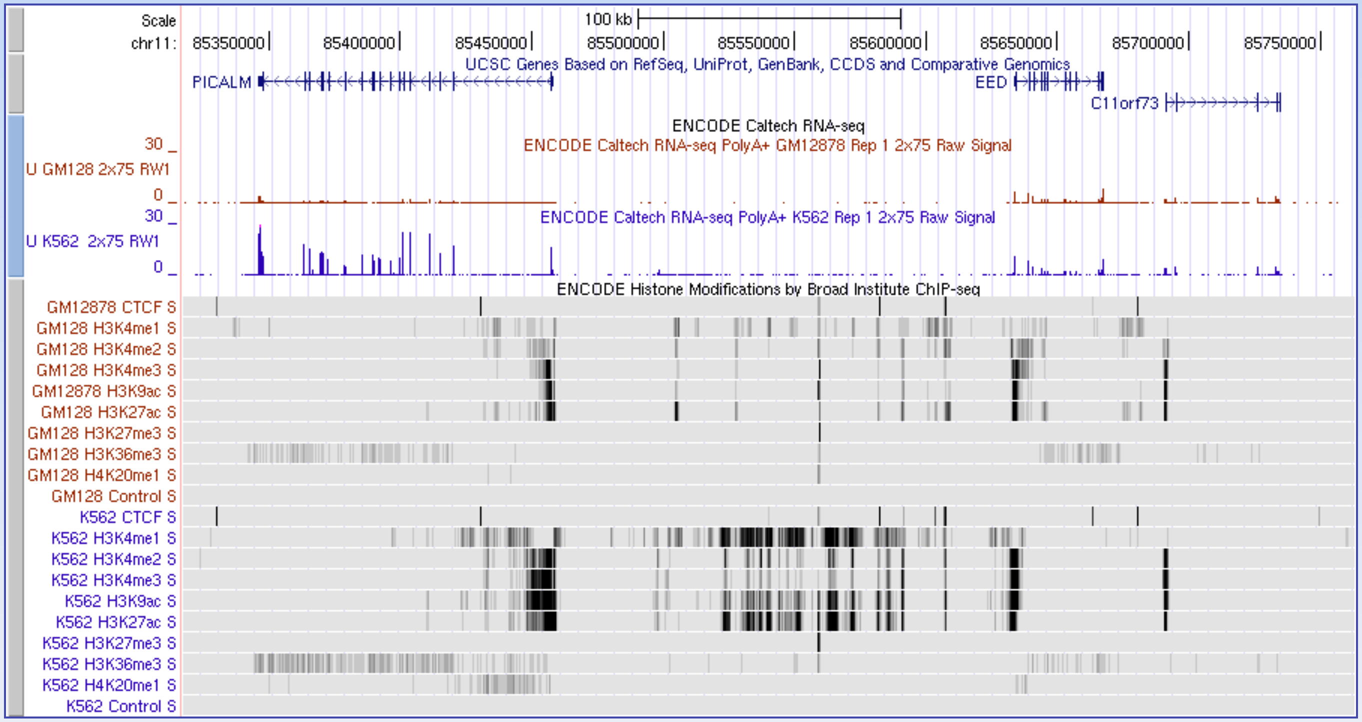
DNase patterns typically are less specific to a single cell type as seen here

Histone Mark and related ChIP-SEQ



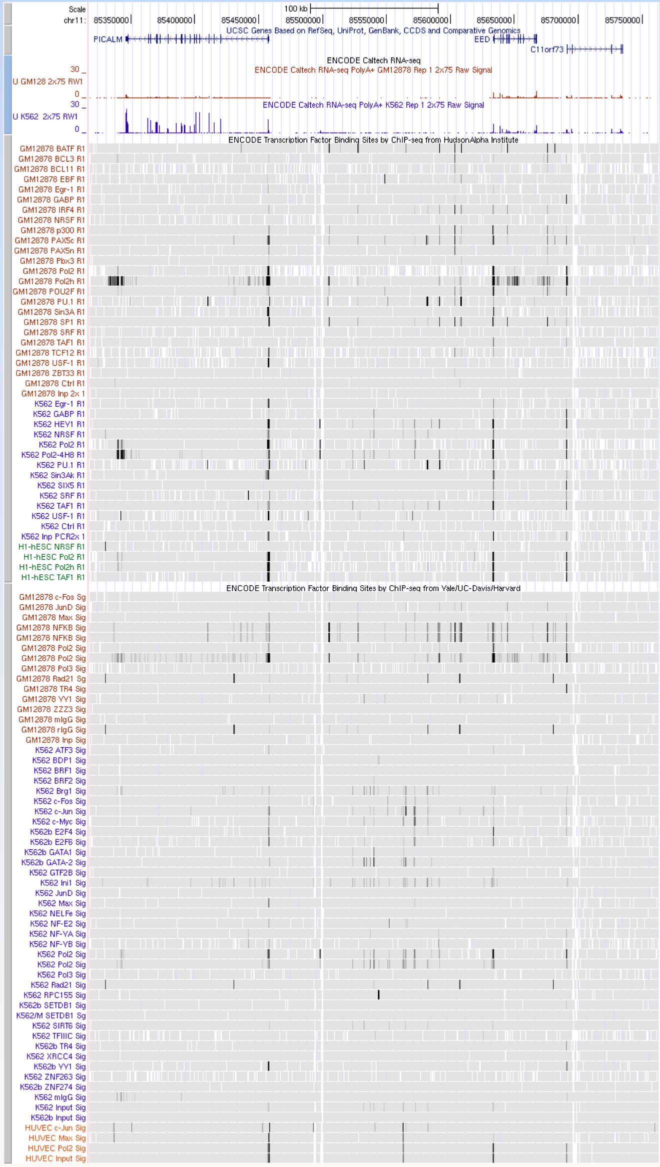
- Various histone marks give a broad picture of promoters, enhancers, repressed regions, transcribed regions
- ENCODE data sets currently include 12 histone marks + CTCF (insulator mark) in 67 cell lines. ~12 cell lines have near complete histone mark coverage

Histone marks on 2 cell lines



Histone mark data at the same locus in two cell lines, GM12878 (red) and K562 (blue). Different marks are associated with promoters, transcribed regions, silencers, enhancers, etc. Most marks are darker in K562, which is more actively transcribing this region.

Transcription Factor ChIP-Seq



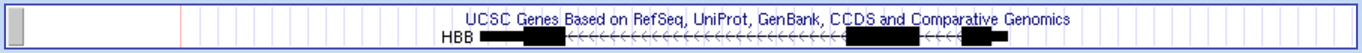
ENCODE has data on 160 factors – most in several cell lines where they are expressed. More coming.

Making data fit on a single screen

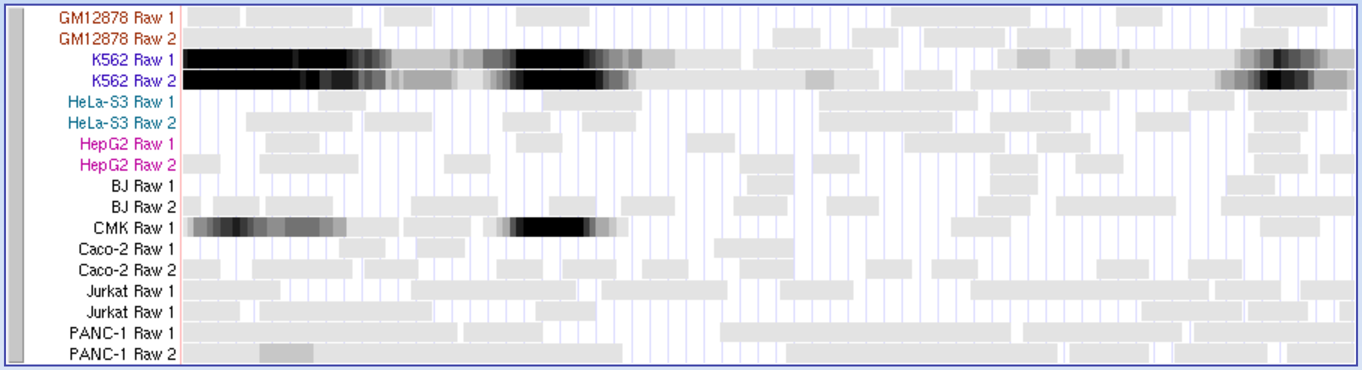
- All of the ENCODE data is excellent, but there is so much of it, it can be hard to know if you’ve seen everything relevant.
- Problem most acute in transcription factor ChIP-SEQ, but really a problem everywhere.
- Lately UCSC has developed several ways of visually summarizing the data.

Integrating DNase across cell lines

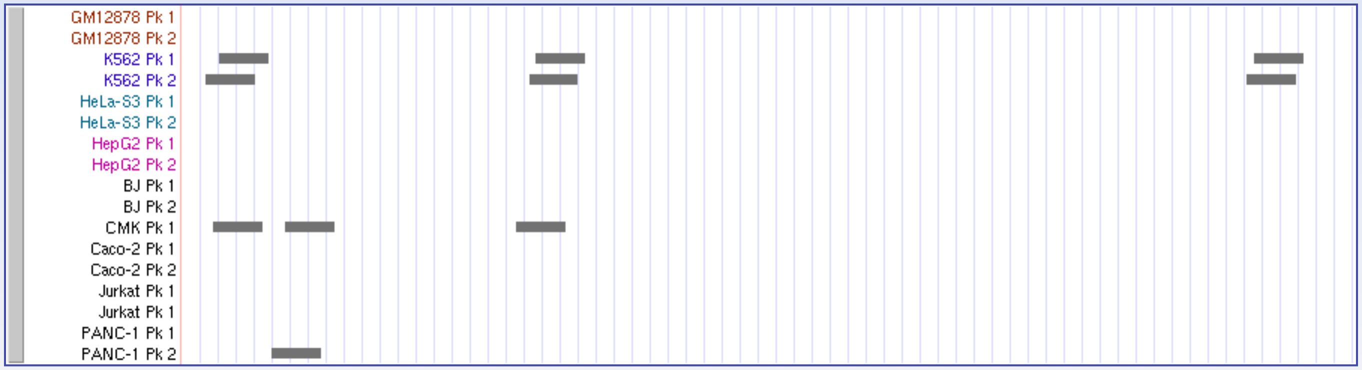
HBB Gene


UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
HBB

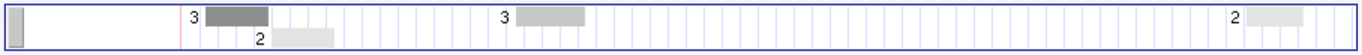
DNAseI signal



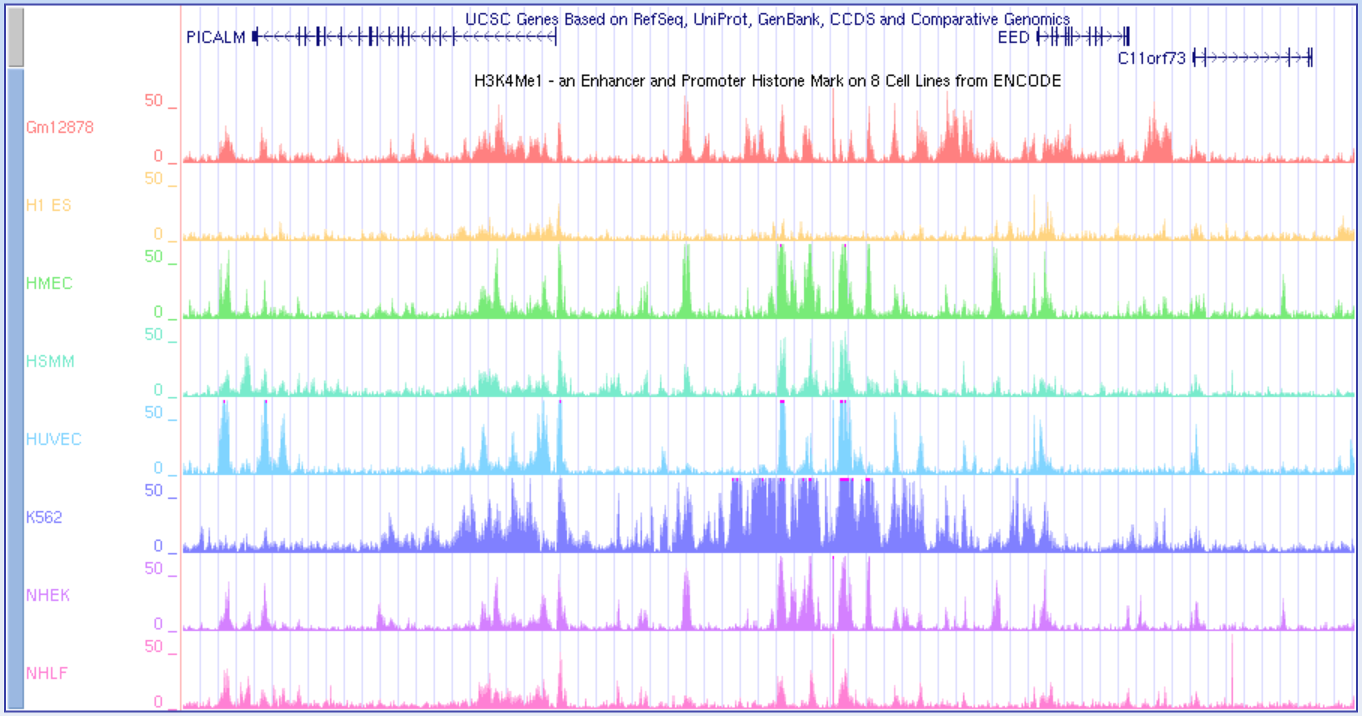
peaks

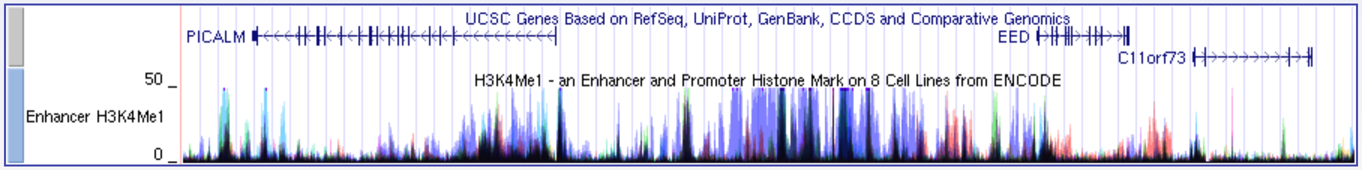


clustered peaks

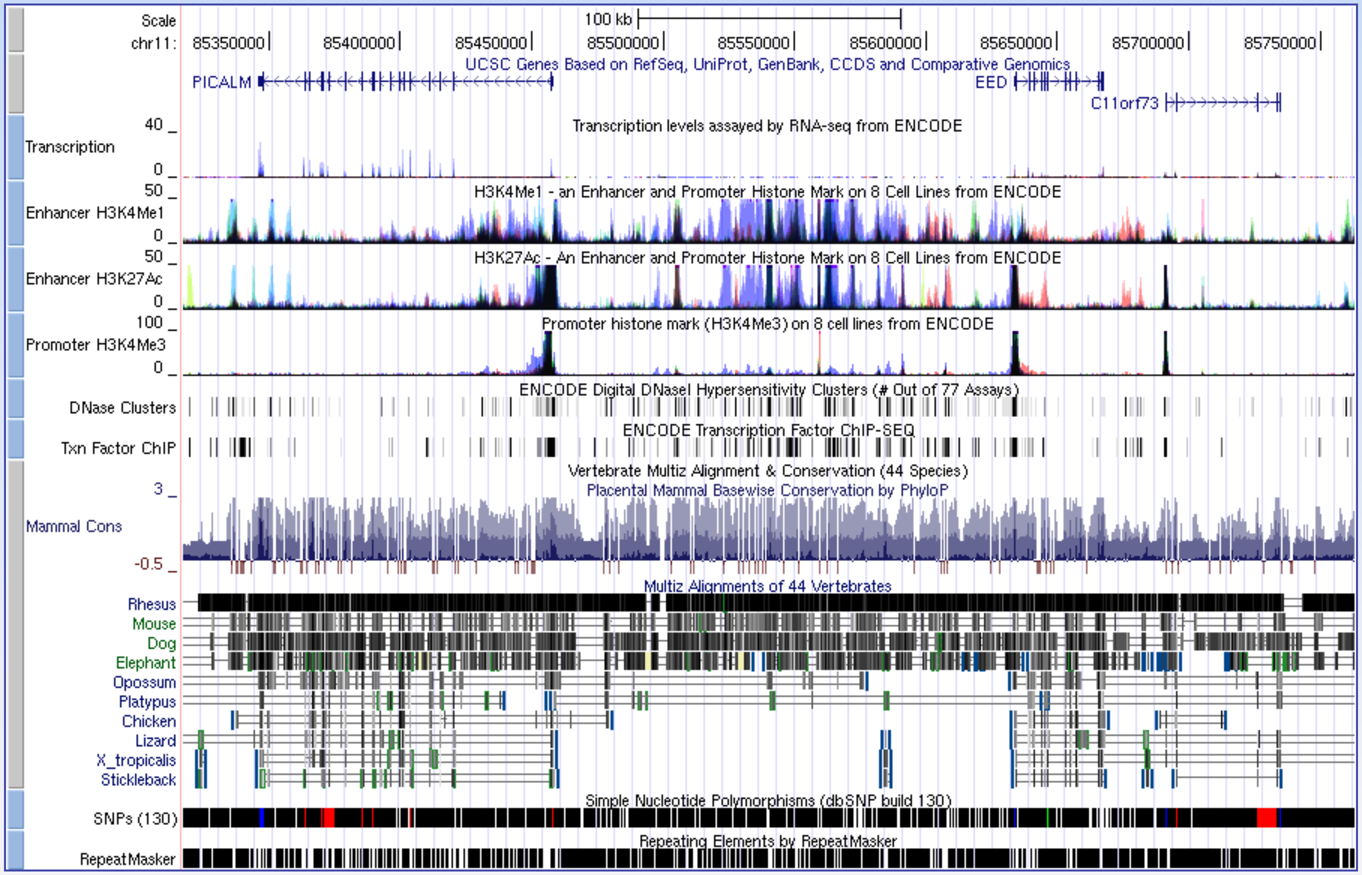


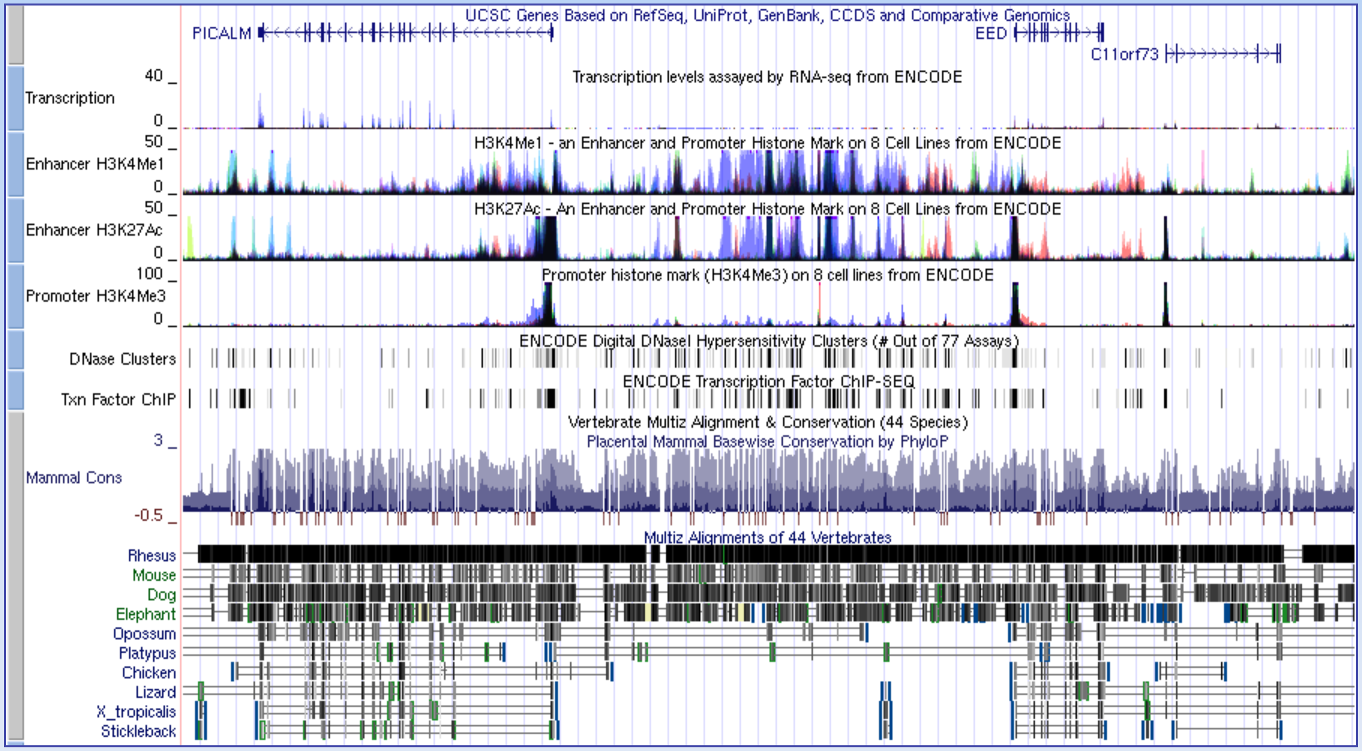
Rainbow overlay for histone marks


UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
PICALM
EED
C11orf73
H3K4Me1 - an Enhancer and Promoter Histone Mark on 8 Cell Lines from ENCODE
Gm12878
H1 ES
HMEC
HSMM
HUVEC
K562
NHEK
NHLF
50
0


UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
PICALM
EED
C11orf73
H3K4Me1 - an Enhancer and Promoter Histone Mark on 8 Cell Lines from ENCODE
Enhancer H3K4Me1
50
0

Integrated regulatory tracks in context with other genomics information at UCSC


Scale
100 kb
chr11:
85350000
85400000
85450000
85500000
85550000
85600000
85650000
85700000
85750000
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
PICALM
EED
C11orf73
40
Transcription
0
Transcription levels assayed by RNA-seq from ENCODE
50
Enhancer H3K4Me1
0
H3K4Me1 - an Enhancer and Promoter Histone Mark on 8 Cell Lines from ENCODE
50
Enhancer H3K27Ac
0
H3K27Ac - An Enhancer and Promoter Histone Mark on 8 Cell Lines from ENCODE
100
Promoter H3K4Me3
0
Promoter histone mark (H3K4Me3) on 8 cell lines from ENCODE
DNase Clusters
ENCODE Digital DNaseI Hypersensitivity Clusters (# Out of 77 Assays)
Txn Factor ChIP
ENCODE Transcription Factor ChIP-SEQ
Vertebrate Multiz Alignment & Conservation (44 Species)
Placental Mammal Basewise Conservation by PhyloP
3
Mammal Cons
-0.5
Multiz Alignments of 44 Vertebrates
Rhesus
Mouse
Dog
Elephant
Opossum
Platypus
Chicken
Lizard
X_tropicalis
Stickleback
Simple Nucleotide Polymorphisms (dbSNP build 130)
SNPs (130)
Repeating Elements by RepeatMasker
RepeatMasker

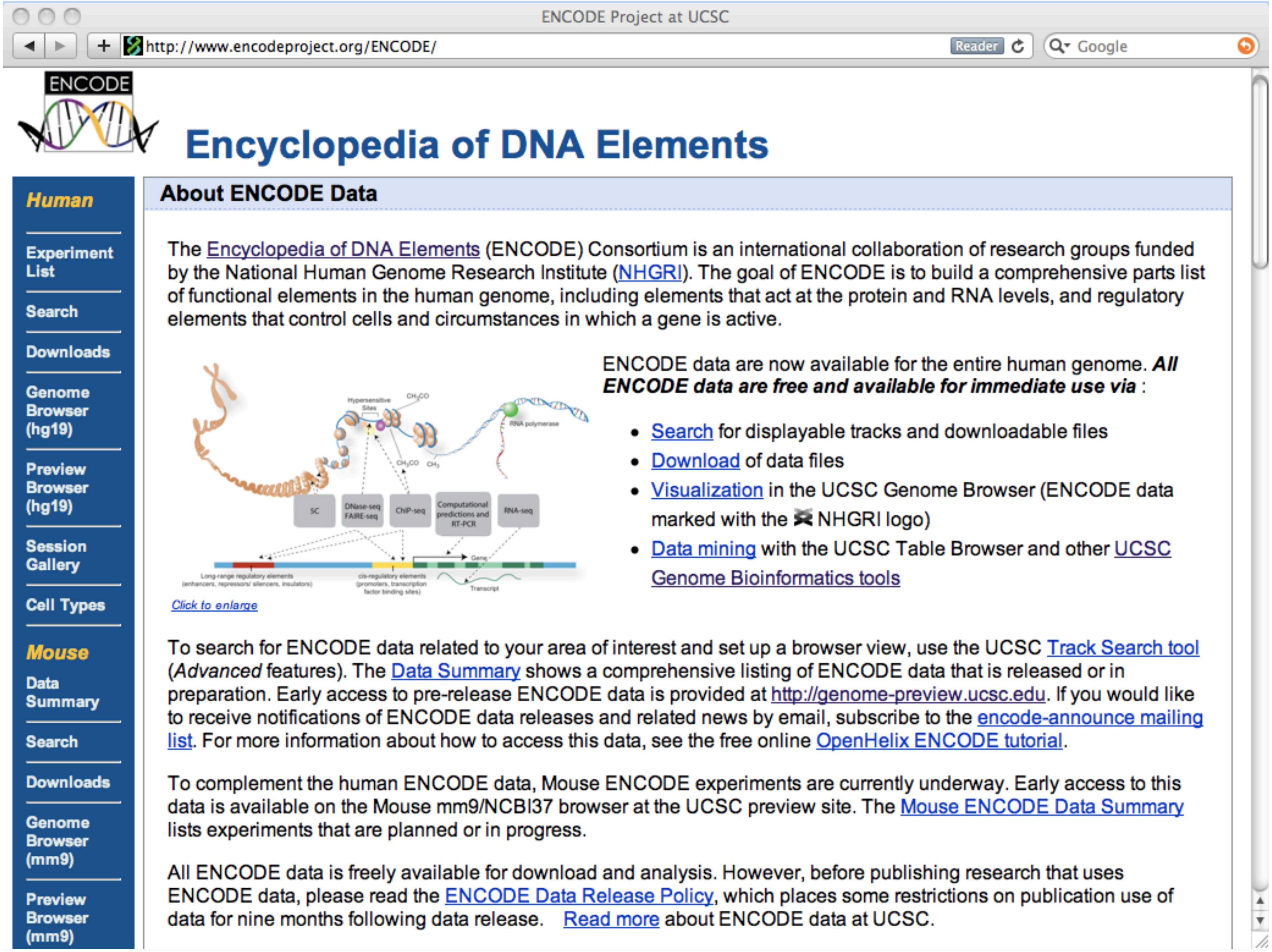

UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
PICALM
EED
C11orf73
40
Transcription
0
Transcription levels assayed by RNA-seq from ENCODE
50
Enhancer H3K4Me1
0
H3K4Me1 - an Enhancer and Promoter Histone Mark on 8 Cell Lines from ENCODE
50
Enhancer H3K27Ac
0
H3K27Ac - An Enhancer and Promoter Histone Mark on 8 Cell Lines from ENCODE
100
Promoter H3K4Me3
0
Promoter histone mark (H3K4Me3) on 8 cell lines from ENCODE
DNase Clusters
ENCODE Digital DNaseI Hypersensitivity Clusters (# Out of 77 Assays)
Txn Factor ChIP
ENCODE Transcription Factor ChIP-SEQ
Vertebrate Multiz Alignment & Conservation (44 Species)
Placental Mammal Basewise Conservation by PhyloP
3
Mammal Cons
-0.5
Multiz Alignments of 44 Vertebrates
Rhesus
Mouse
Dog
Elephant
Opossum
Platypus
Chicken
Lizard
X_tropicalis
Stickleback

- ENCODE regulatory data:
  – Histone marks –characterization of large regions into promoter/ enhancer/repressed
  – DNAse hypersensitivity - defines smaller regions as regulatory
  – Transcription factor chromatin immunoprecipitation – what regulatory factors bind in a smaller region.
  – Chromatin conformation capture – just starting to ramp up.
- Available at http://genome.ucsc.edu

Accessing ENCODE Data at DCC

- http://www.encodeproject.org
  - ENCODE portal. Describes project overall, project news, tables and spreadsheets for all experiments
- http://genome.ucsc.edu
  - ENCODE data integrated into UCSC Genome Browser on hg19 and mm9 assemblies
- http://genome-preview.ucsc.edu
  - Includes not-yet-reviewed data

Much of the data also is at NCBI (GEO) and Ensembl.



Encyclopedia of DNA Elements

About ENCODE Data

The Encyclopedia of DNA Elements (ENCODE) 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.

Click to enlarge

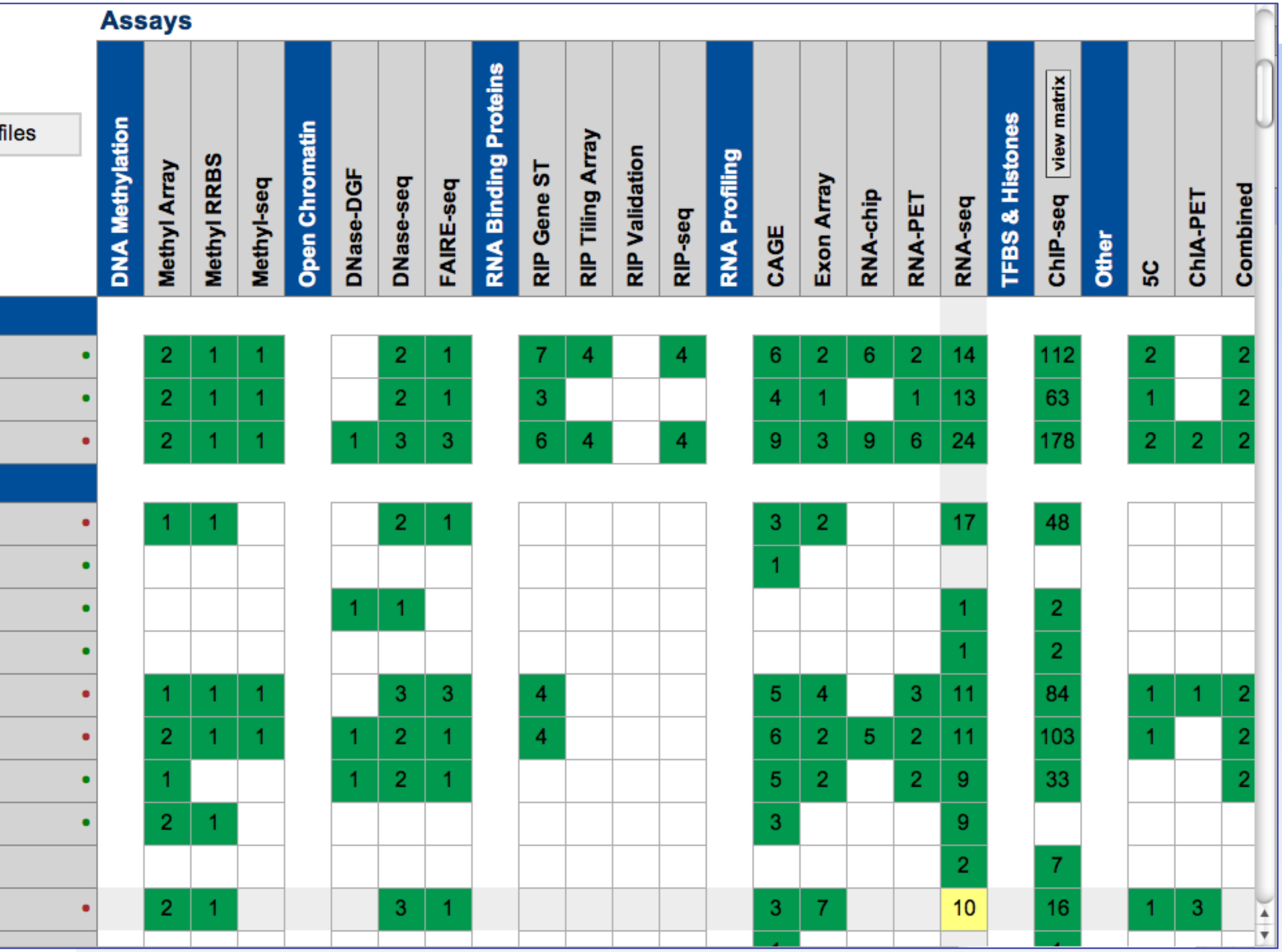
ENCODE data are now available for the entire human genome. ***All ENCODE data are free and available for immediate use via*** :

- Search for displayable tracks and downloadable files
- Download of data files
- Visualization in the UCSC Genome Browser (ENCODE data marked with the NHGRI logo)
- Data mining with the UCSC Table Browser and other UCSC Genome Bioinformatics tools

To search for ENCODE data related to your area of interest and set up a browser view, use the UCSC Track Search tool (*Advanced* features). The Data Summary shows a comprehensive listing of ENCODE data that is released or in preparation. Early access to pre-release ENCODE data is provided at http://genome-preview.ucsc.edu. If you would like to receive notifications of ENCODE data releases and related news by email, subscribe to the encode-announce mailing list. For more information about how to access this data, see the free online OpenHelix ENCODE tutorial.

To complement the human ENCODE data, Mouse ENCODE experiments are currently underway. Early access to this data is available on the Mouse mm9/NCBI37 browser at the UCSC preview site. The Mouse ENCODE Data Summary lists experiments that are planned or in progress.

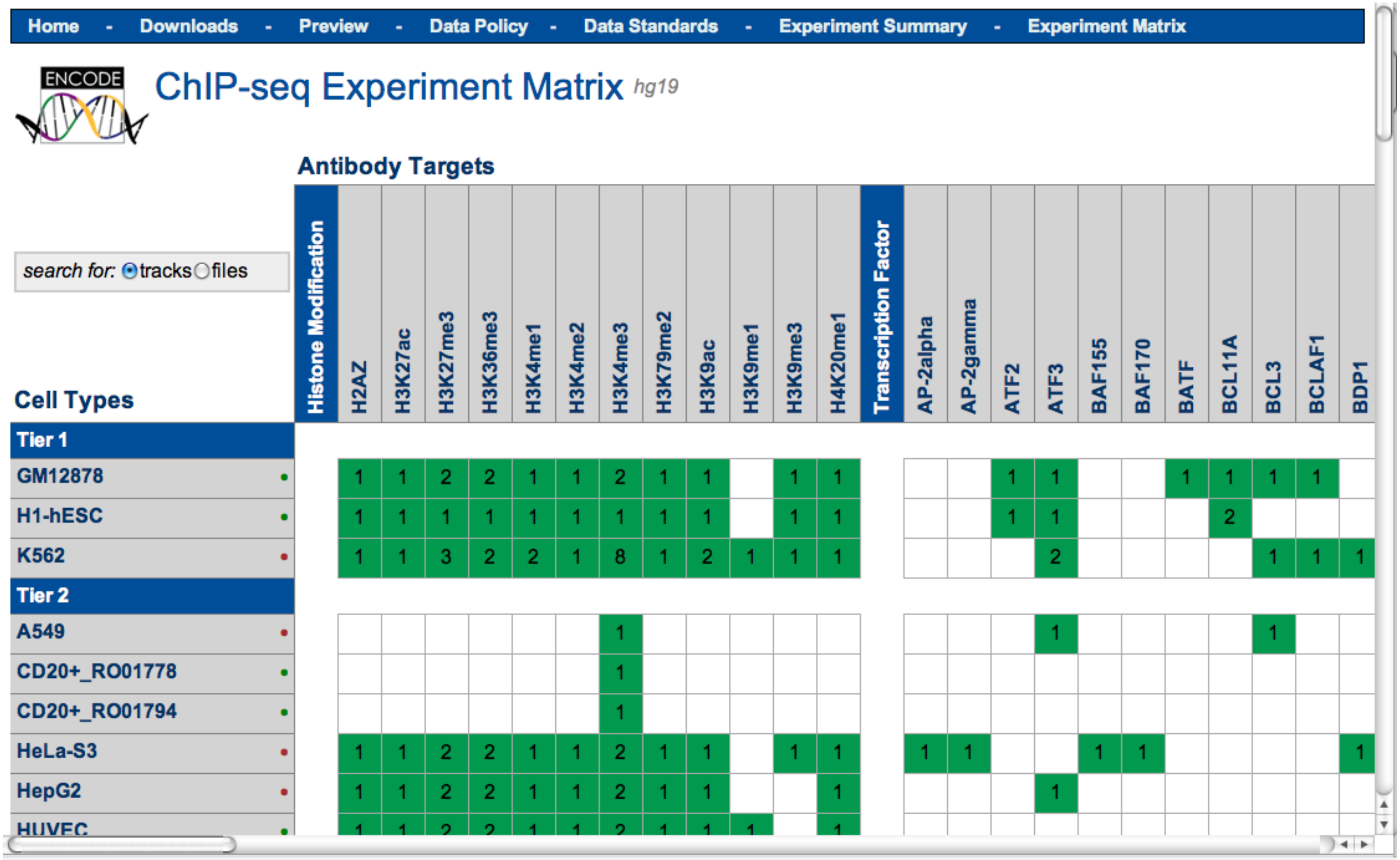
All ENCODE data is freely available for download and analysis. However, before publishing research that uses ENCODE data, please read the ENCODE Data Release Policy, which places some restrictions on publication use of data for nine months following data release. Read more about ENCODE data at UCSC.



Assays

	DNA Methylation	Methyl Array	Methyl RRBS	Methyl-seq	Open Chromatin	DNase-DGF	DNase-seq	FAIRE-seq	RNA Binding Proteins	RIP Gene ST	RIP Tiling Array	RIP Validation	RIP-seq	RNA Profiling	CAGE	Exon Array	RNA-chip	RNA-PET	RNA-seq	TFBS & Histones	ChIP-seq view matrix	Other	5C	ChIA-PET	Combined
🟢		2	1	1			2	1		7	4		4		6	2	6	2	14		112		2		2
🟢		2	1	1			2	1		3					4	1		1	13		63		1		2
🔴		2	1	1		1	3	3		6	4		4		9	3	9	6	24		178		2	2	2
🔴		1	1				2	1							3	2			17		48				
🟢															1										
🟢						1	1												1		2				
🟢																			1		2				
🔴		1	1	1			3	3		4					5	4		3	11		84		1	1	2
🔴		2	1	1		1	2	1		4					6	2	5	2	11		103		1		2
🟢		1				1	2	1							5	2		2	9		33				2
🟢		2	1												3				9						
																			2		7				
🔴		2	1				3	1							3	7			10		16		1	3	

Experiment matrix link off of ENCODE Portal, provides overview of number of experiments of various types on various cells. Clicking on a cell brings up list of individual tracks or files. It's a big matrix, note size of thumb on scrollbar.



Home - Downloads - Preview - Data Policy - Data Standards - Experiment Summary - Experiment Matrix
ENCODE

ChIP-seq Experiment Matrix *hg19*

search for: ◉tracks ○files

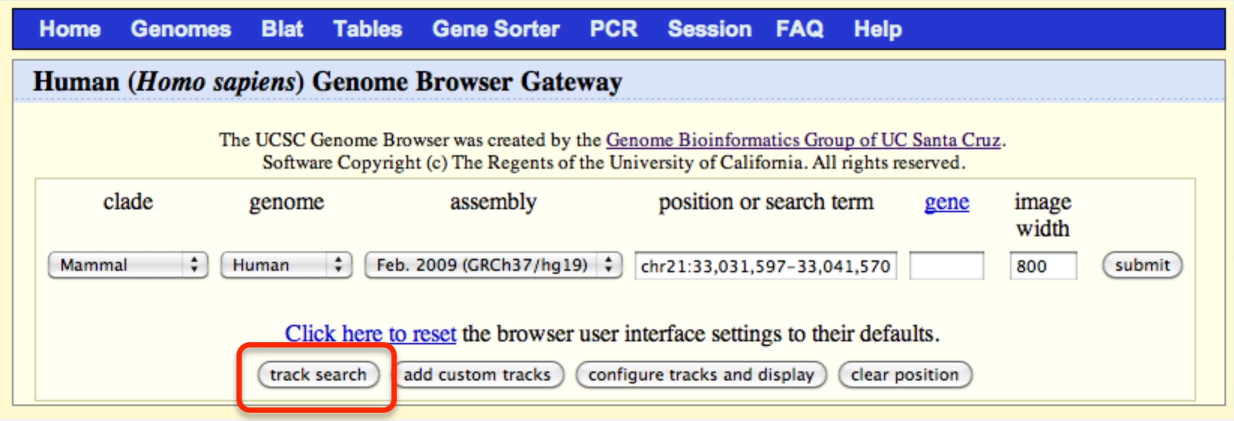
Antibody Targets

Cell Types		Histone Modification	H2AZ	H3K27ac	H3K27me3	H3K36me3	H3K4me1	H3K4me2	H3K4me3	H3K79me2	H3K9ac	H3K9me1	H3K9me3	H4K20me1	Transcription Factor	AP-2alpha	AP-2gamma	ATF2	ATF3	BAF155	BAF170	BATF	BCL11A	BCL3	BCLAF1	BDP1
Tier 1																										
GM12878	●		1	1	2	2	1	1	2	1	1		1	1				1	1			1	1	1	1	
H1-hESC	●		1	1	1	1	1	1	1	1	1		1	1				1	1				2			
K562	●		1	1	3	2	2	1	8	1	2	1	1	1					2					1	1	1
Tier 2																										
A549	●								1										1					1		
CD20+_RO01778	●								1																	
CD20+_RO01794	●								1																	
HeLa-S3	●		1	1	2	2	1	1	2	1	1		1	1		1	1			1	1					1
HepG2	●		1	1	2	2	1	1	2	1	1			1					1							
HUVEC	●		1	1	2	2	1	1	2	1	1	1		1												

ChIP-seq experiments have their own submatrix. This is an even bigger matrix. Note size of both horizontal and vertical scroller thumbs.

Track Search

- Can do a free-form (Google-style) search or search metadata field-by-field



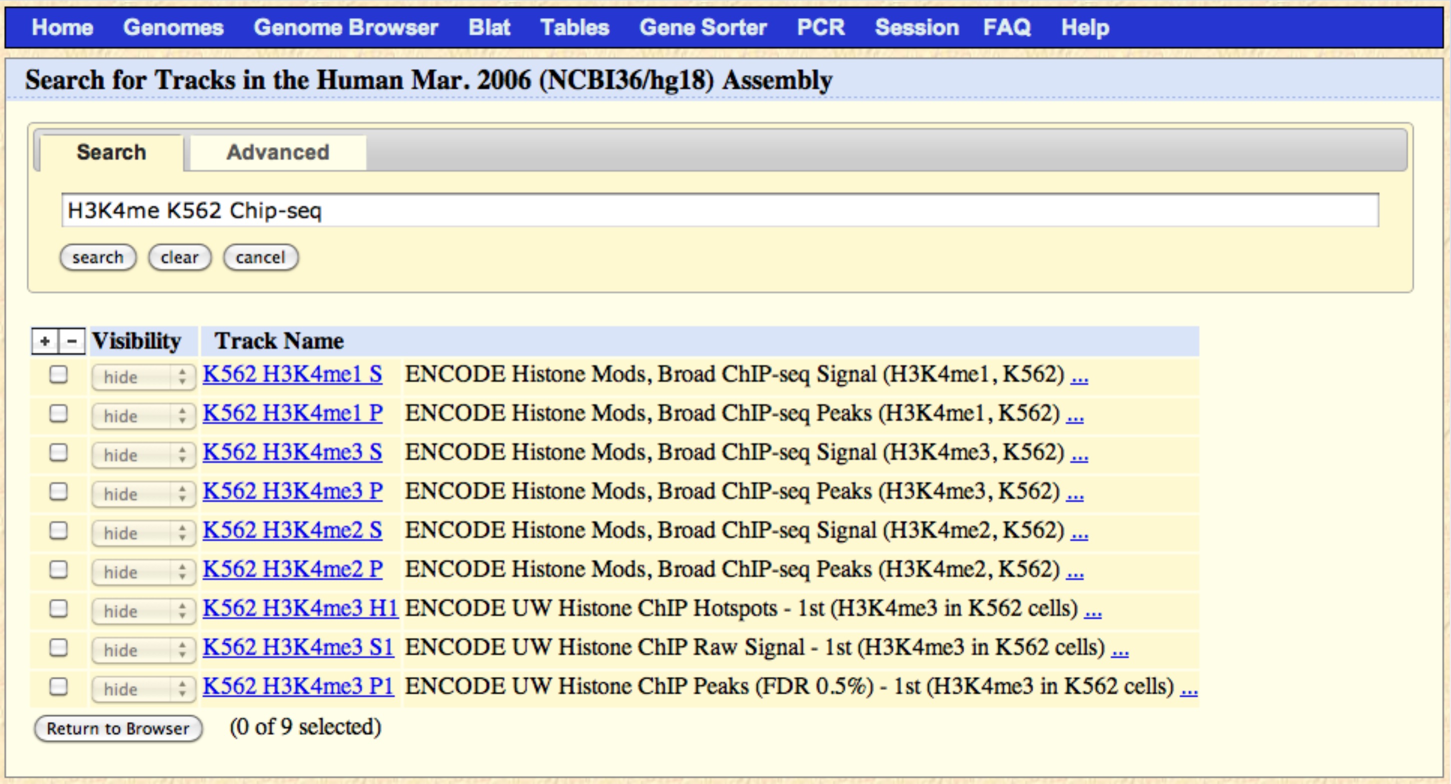
Human (*Homo sapiens*) Genome Browser Gateway

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.

clade	genome	assembly	position or search term	gene	image width	
Mammal	Human	Feb. 2009 (GRCh37/hg19)	chr21:33,031,597-33,041,570		800	submit

Click here to reset the browser user interface settings to their defaults.

Free text search



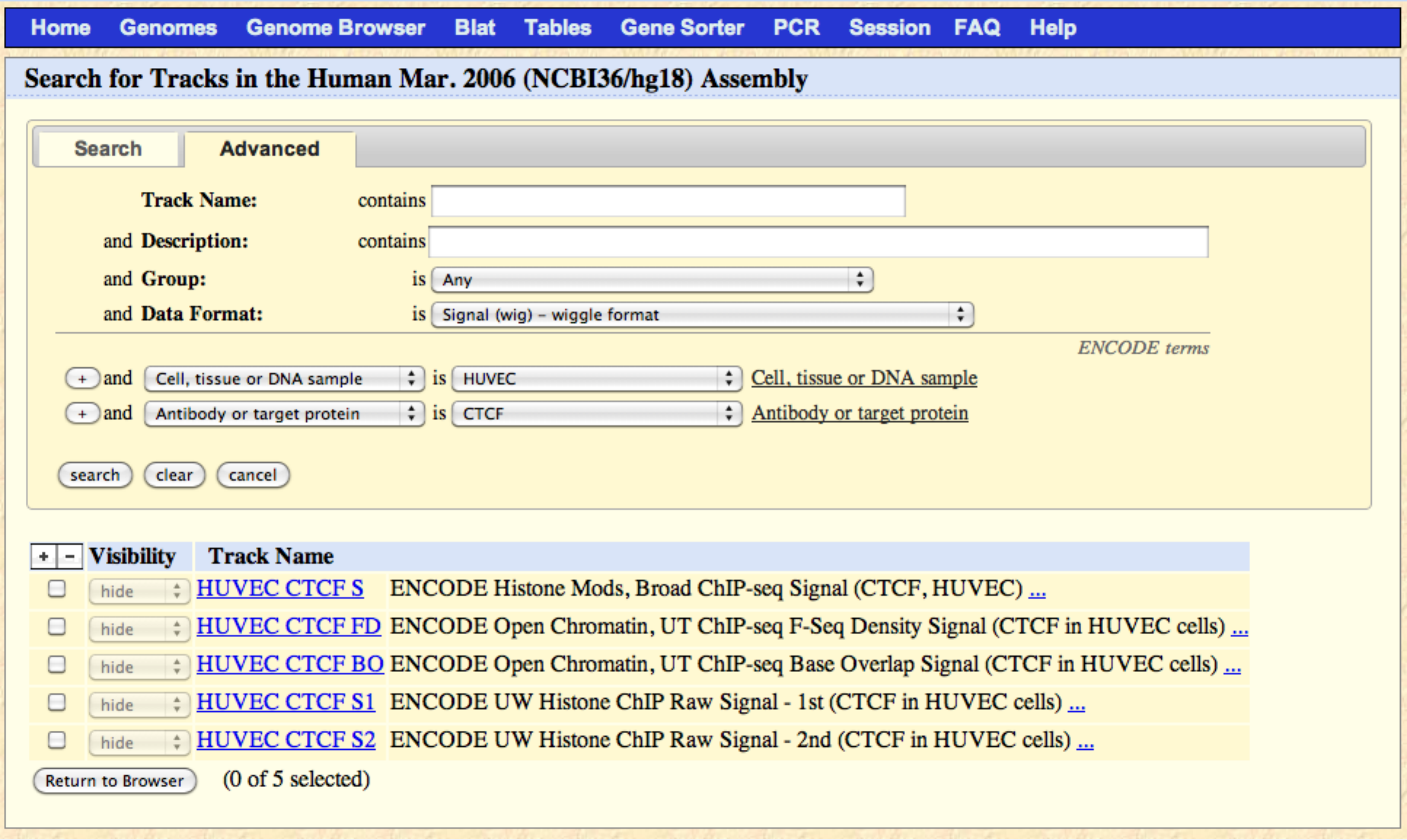
Search for Tracks in the Human Mar. 2006 (NCBI36/hg18) Assembly

H3K4me H3K4me K562 Chip-seq

+ −	Visibility	Track Name	
□	hide	K562 H3K4me1 S	ENCODE Histone Mods, Broad ChIP-seq Signal (H3K4me1, K562) ...
□	hide	K562 H3K4me1 P	ENCODE Histone Mods, Broad ChIP-seq Peaks (H3K4me1, K562) ...
□	hide	K562 H3K4me3 S	ENCODE Histone Mods, Broad ChIP-seq Signal (H3K4me3, K562) ...
□	hide	K562 H3K4me3 P	ENCODE Histone Mods, Broad ChIP-seq Peaks (H3K4me3, K562) ...
□	hide	K562 H3K4me2 S	ENCODE Histone Mods, Broad ChIP-seq Signal (H3K4me2, K562) ...
□	hide	K562 H3K4me2 P	ENCODE Histone Mods, Broad ChIP-seq Peaks (H3K4me2, K562) ...
□	hide	K562 H3K4me3 H1	ENCODE UW Histone ChIP Hotspots - 1st (H3K4me3 in K562 cells) ...
□	hide	K562 H3K4me3 S1	ENCODE UW Histone ChIP Raw Signal - 1st (H3K4me3 in K562 cells) ...
□	hide	K562 H3K4me3 P1	ENCODE UW Histone ChIP Peaks (FDR 0.5%) - 1st (H3K4me3 in K562 cells) ...

(0 of 9 selected)

Advanced field by field search



Search for Tracks in the Human Mar. 2006 (NCBI36/hg18) Assembly

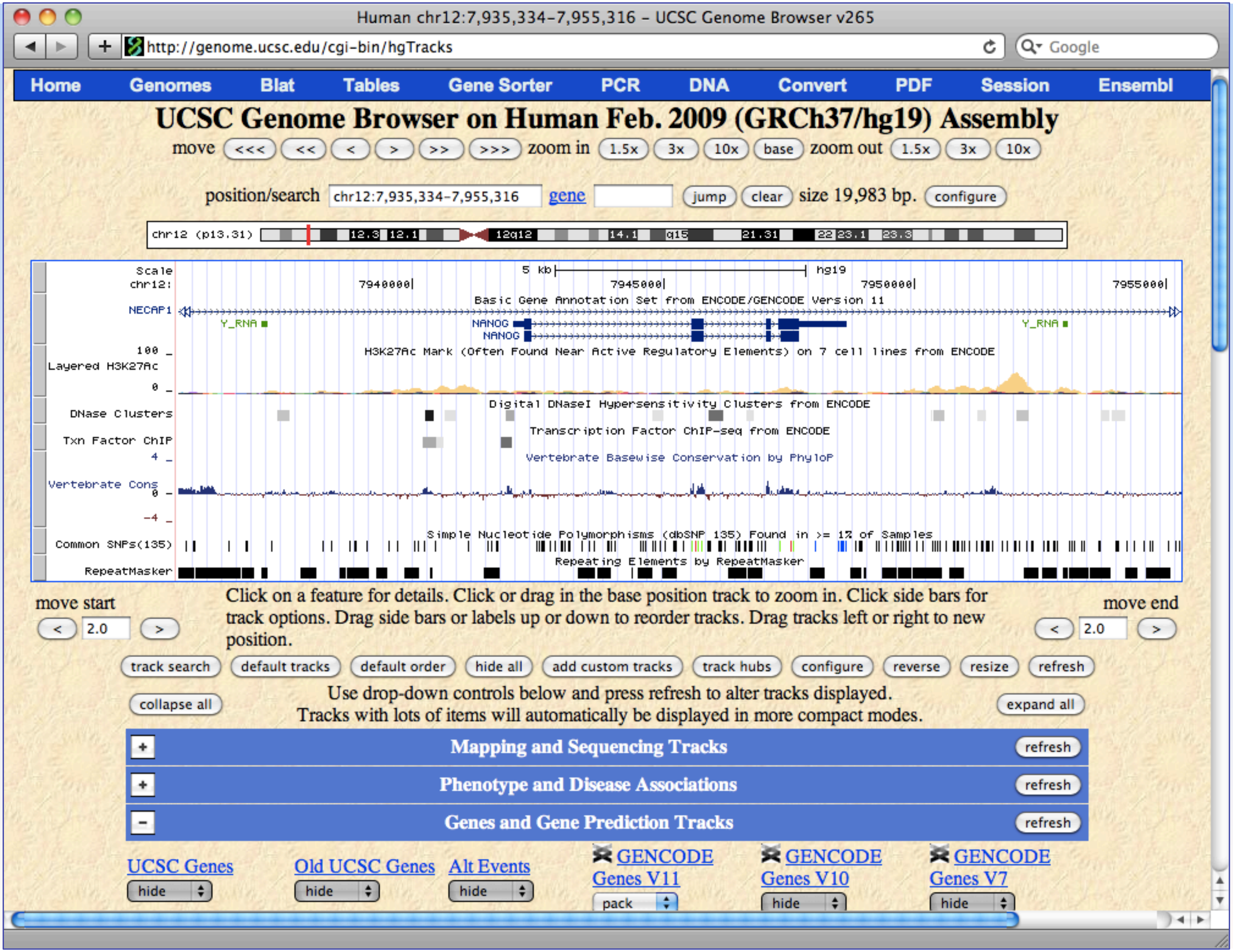
Track Name:	contains	
and **Description:**	contains	
and **Group:**	is	Any
and **Data Format:**	is	Signal (wig) - wiggle format

ENCODE terms

+ and	Cell, tissue or DNA sample	is HUVEC	Cell, tissue or DNA sample
+ and	Antibody or target protein	is CTCF	Antibody or target protein

+ −	Visibility	Track Name	
□	hide	HUVEC CTCF S	ENCODE Histone Mods, Broad ChIP-seq Signal (CTCF, HUVEC) ...
□	hide	HUVEC CTCF FD	ENCODE Open Chromatin, UT ChIP-seq F-Seq Density Signal (CTCF in HUVEC cells) ...
□	hide	HUVEC CTCF BO	ENCODE Open Chromatin, UT ChIP-seq Base Overlap Signal (CTCF in HUVEC cells) ...
□	hide	HUVEC CTCF S1	ENCODE UW Histone ChIP Raw Signal - 1st (CTCF in HUVEC cells) ...
□	hide	HUVEC CTCF S2	ENCODE UW Histone ChIP Raw Signal - 2nd (CTCF in HUVEC cells) ...

(0 of 5 selected)



UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

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.

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.