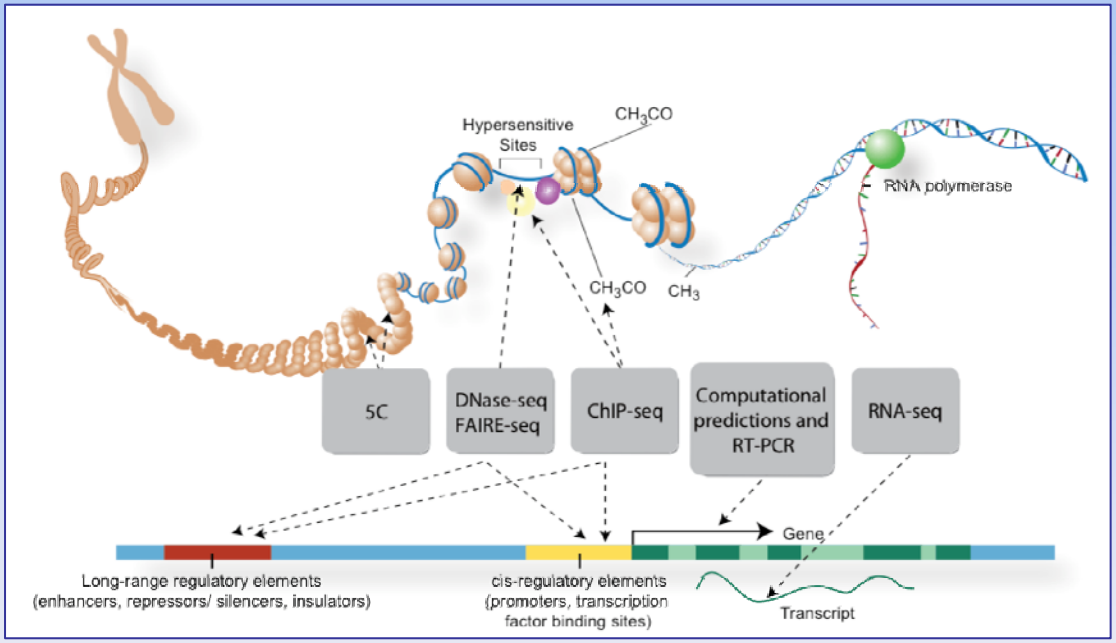
ENCODE Project

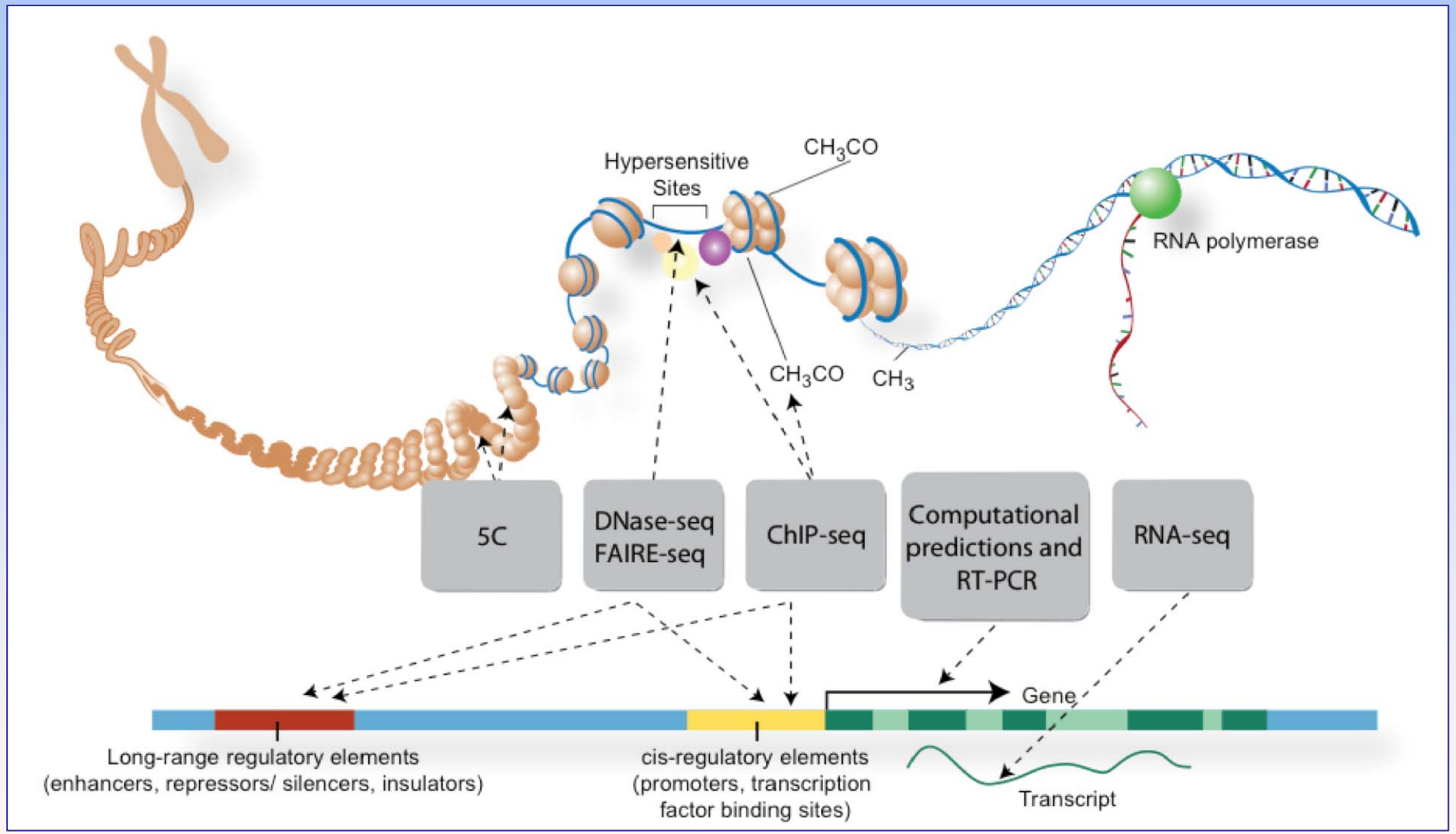


Robert Kuhn
University of California Santa Cruz

Workshop: Working with ENCODE Data
Korean Genome Organization
February 5-7, 2014
Yong Pyong, Korea

http:// genome.ucsc.edu

http:// encodeproject.org


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

ENCODE Project

- ENCODE pilot project covered only 1% of human genome. Phase II ENCODE is full-genome on human and mouse.

  Phase III is starting now. DCC at Stanford, UCSC hosts data.
- 32 biology labs organized into 19 grants + Analysis Working Group and Data Coordination Center (DCC)
- Goal: 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.

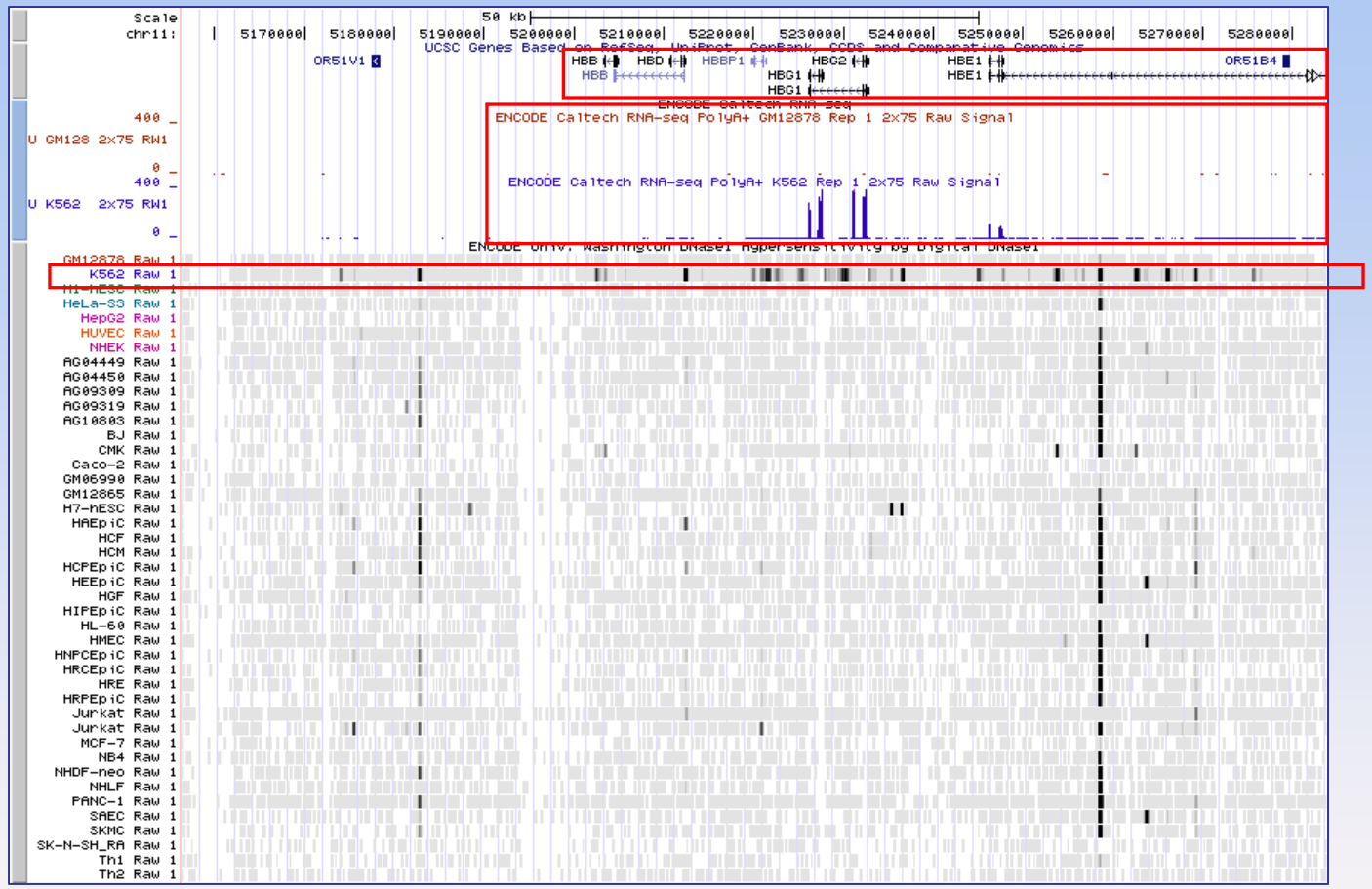
ENCODE assays on regulation of transcription

- Opening/closing chromatin
  - DNase hypersensitivity, FAIRE-seq
  - 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
  - RNA seq fractionation: short/long polyA+/- localized to nucleus, cytoplasm, polysome, nucleoplasm, nuclear matrix, mitochondria, etc.

ENCODE DNase Hypersensitivity

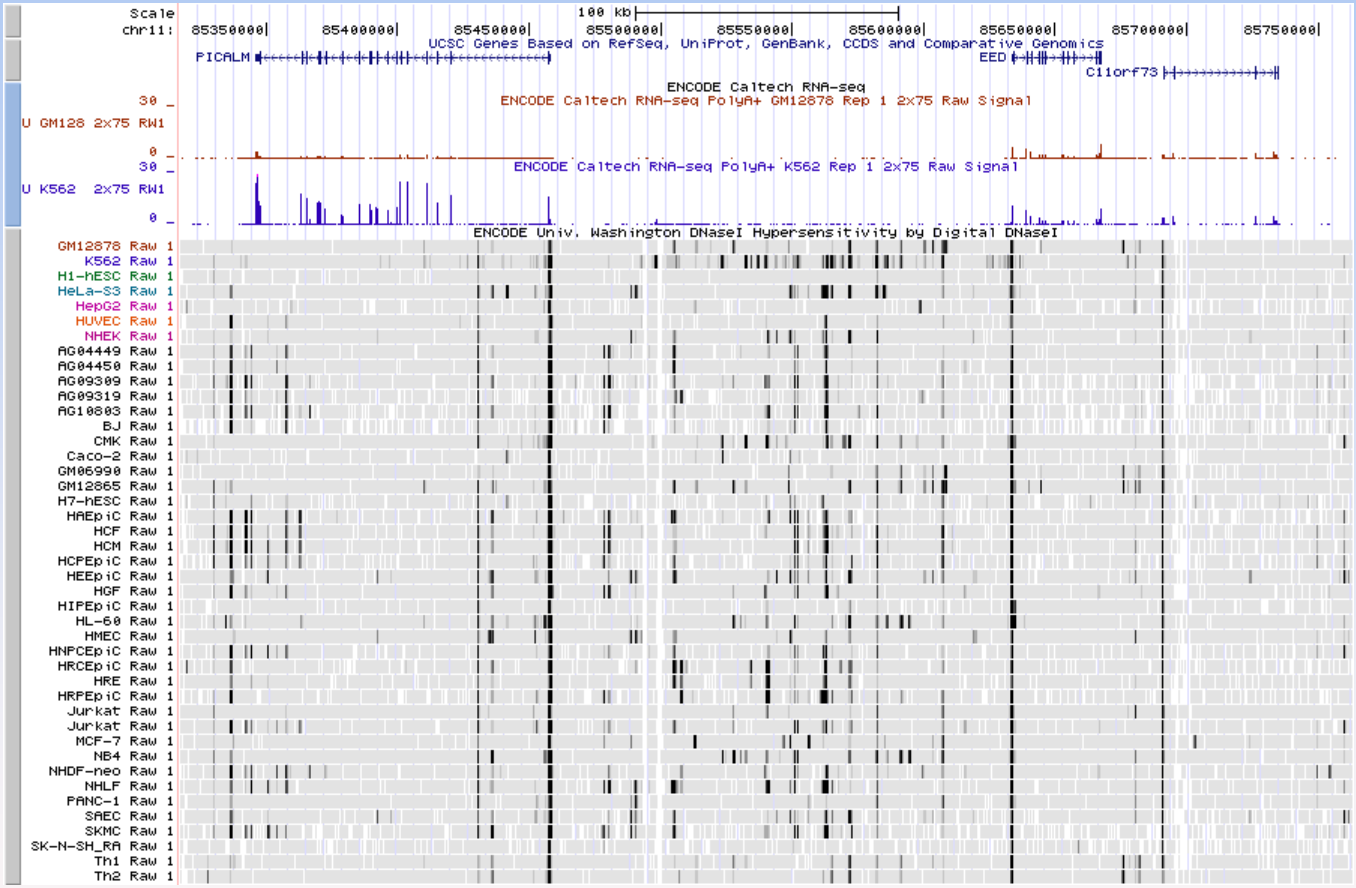
- Several genome-wide high-throughput methods used in ENCODE. All involve DNA-seq
- Data currently available for 388 cell lines and tissues
- Main alignment artifacts to watch for:
  - DNA present in cell in multiple copies:
    - Mitochondria, centromeric repeats, other repeats
    - Generally such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

UW DNaseI at Hemoglobin Beta



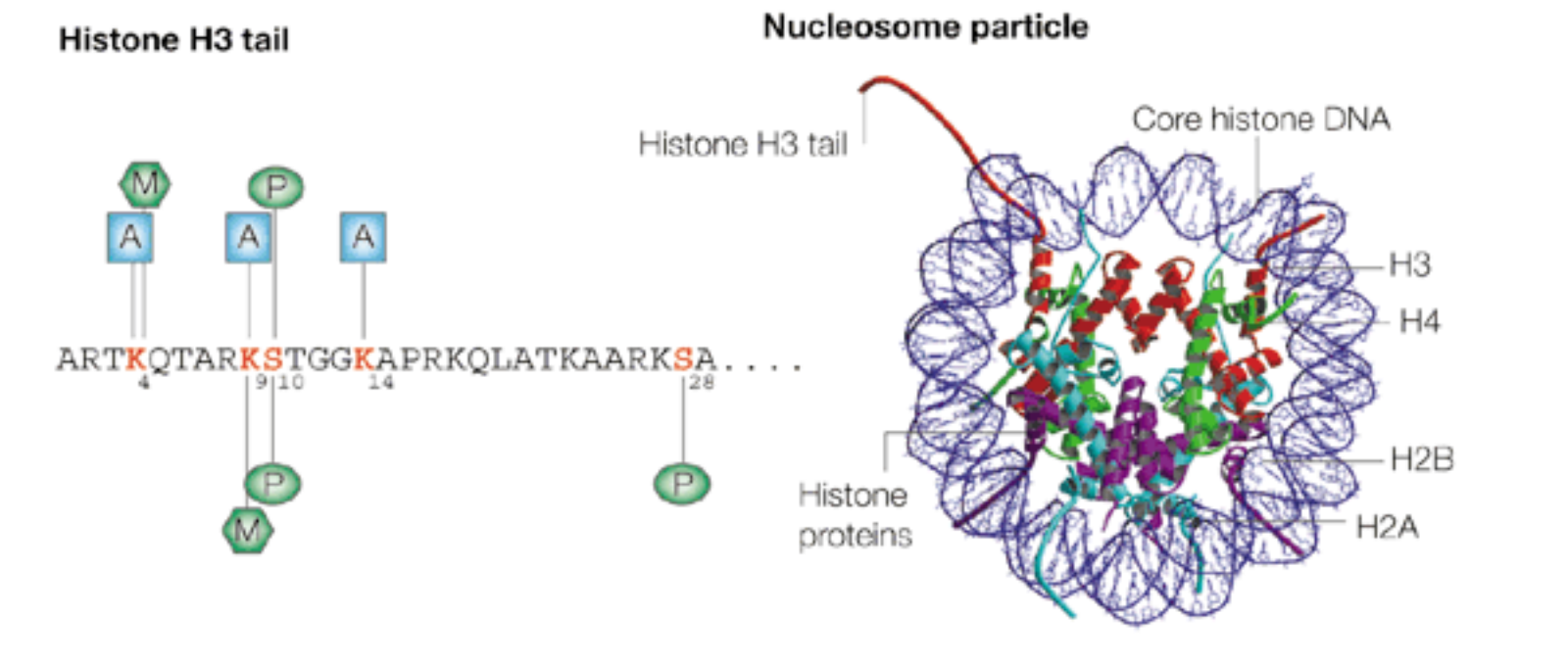
Top track shows genes in the Hemoglobin beta (HBB) locus.
Next track shows RNA levels in GM12878 and K562 cell lines.
The last track is density plots of DNAse hypersensitivity in many cell lines.
K562, a cell line similar to a red blood cell precursor, shows much RNA and DNAase activity.

A more typical locus - PICALM

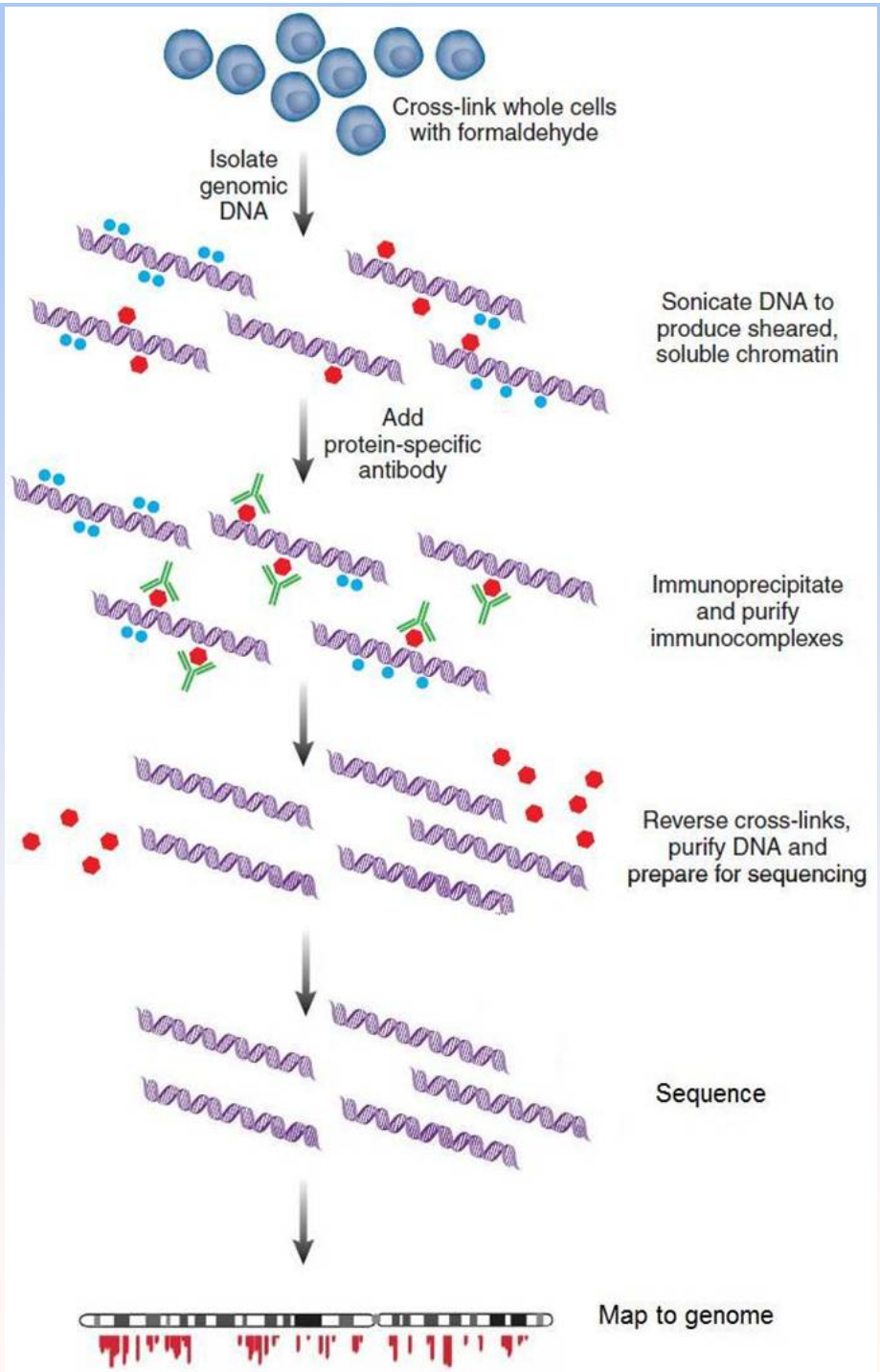


DNase patterns typically are less specific to a single cell type as seen here

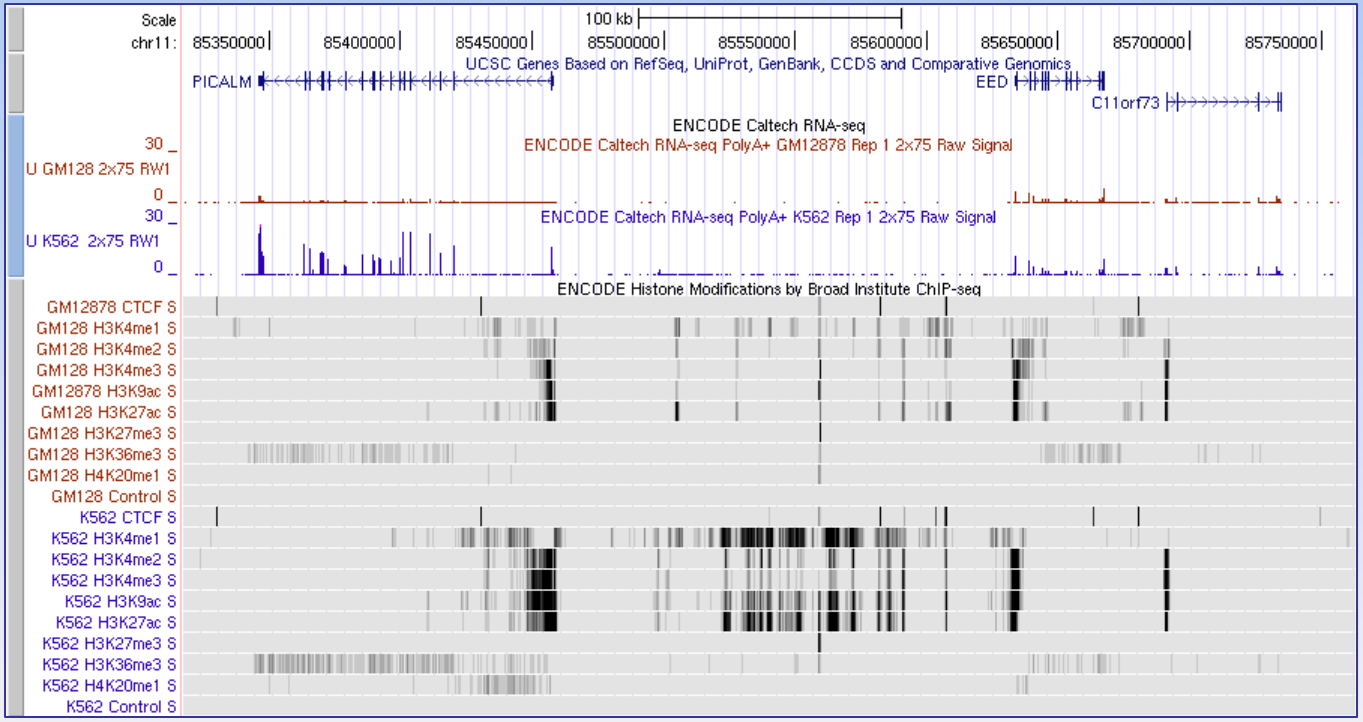
Histone Mark and related ChIP-seq



- 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

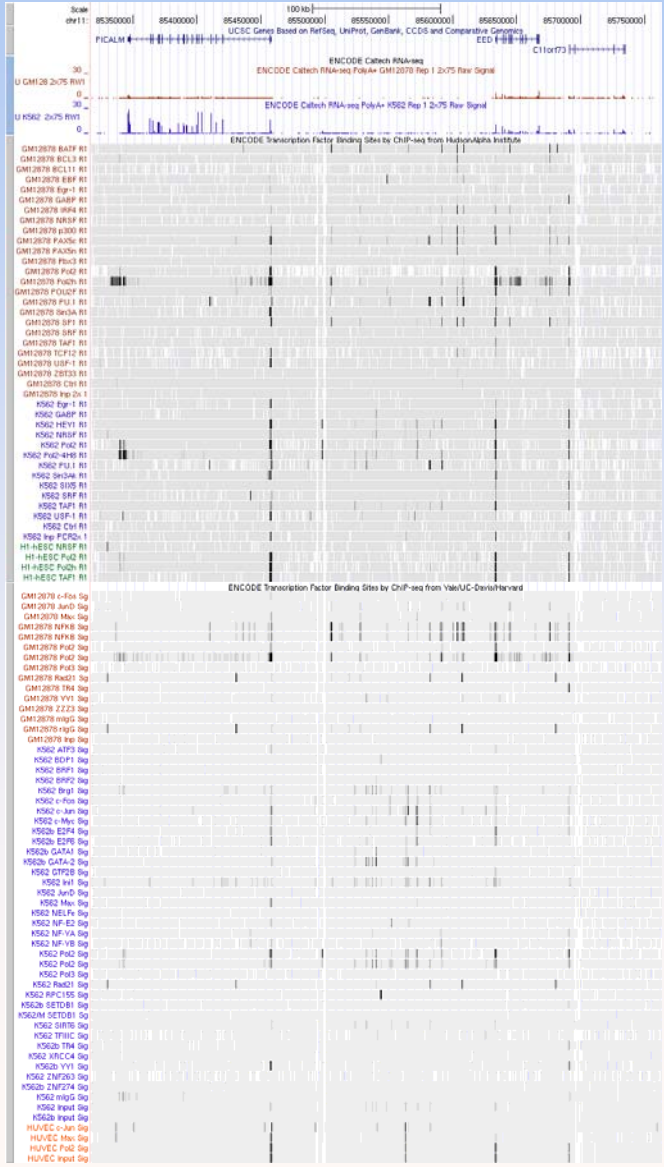

Cross-link whole cells with formaldehyde
Isolate genomic DNA
Sonicate DNA to produce sheared, soluble chromatin
Add protein-specific antibody
Immunoprecipitate and purify immunocomplexes
Reverse cross-links, purify DNA and prepare for sequencing
Sequence
Map to genome

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 – in many cell lines where they are expressed.

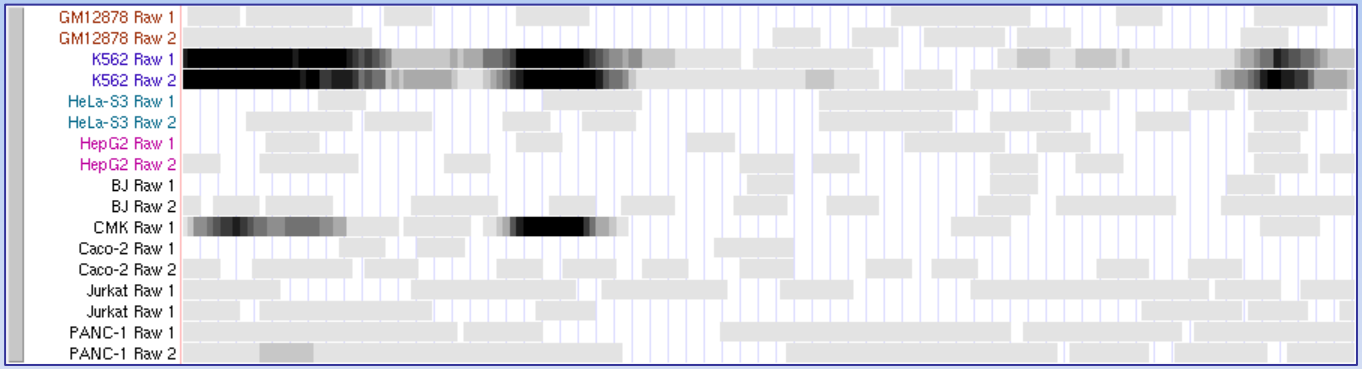
Making data fit on a single screen

- All of the ENCODE data are 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.
- UCSC has developed several ways of visually summarizing the data.

Integrating DNase across cell lines

HBB Gene

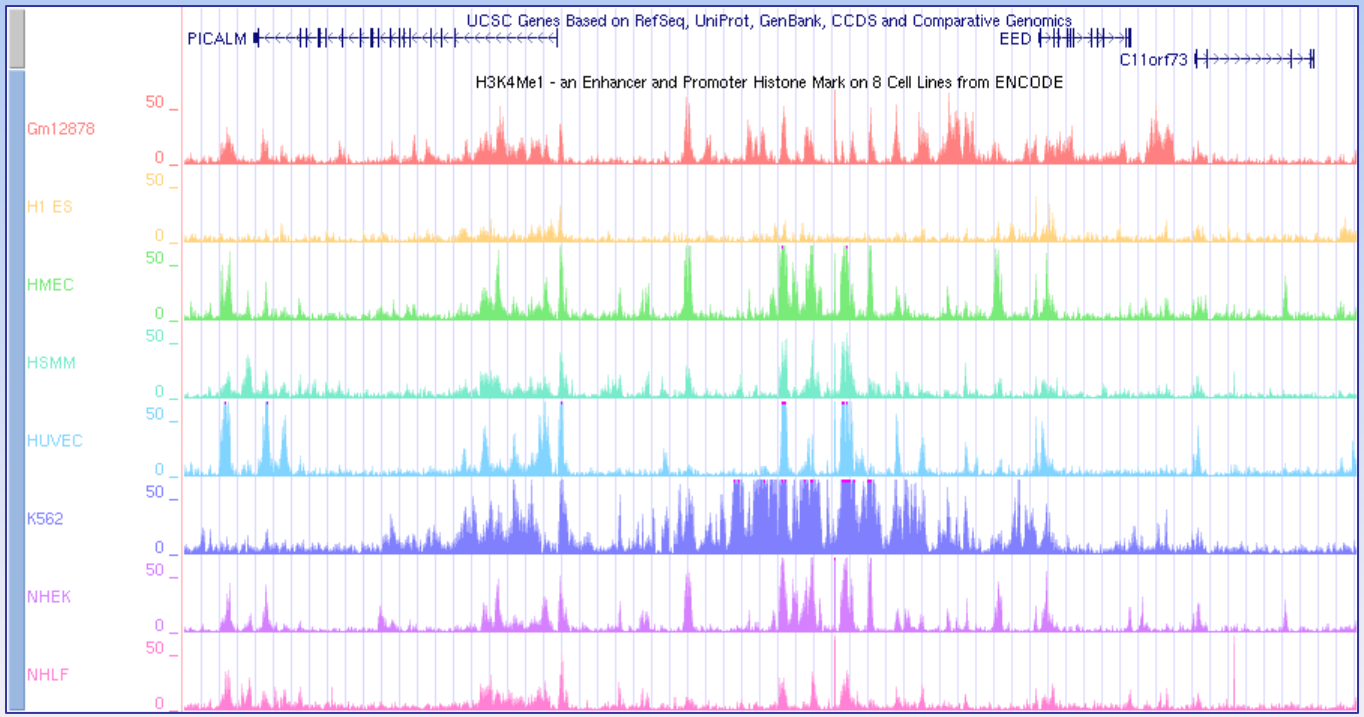
DNAseI signal



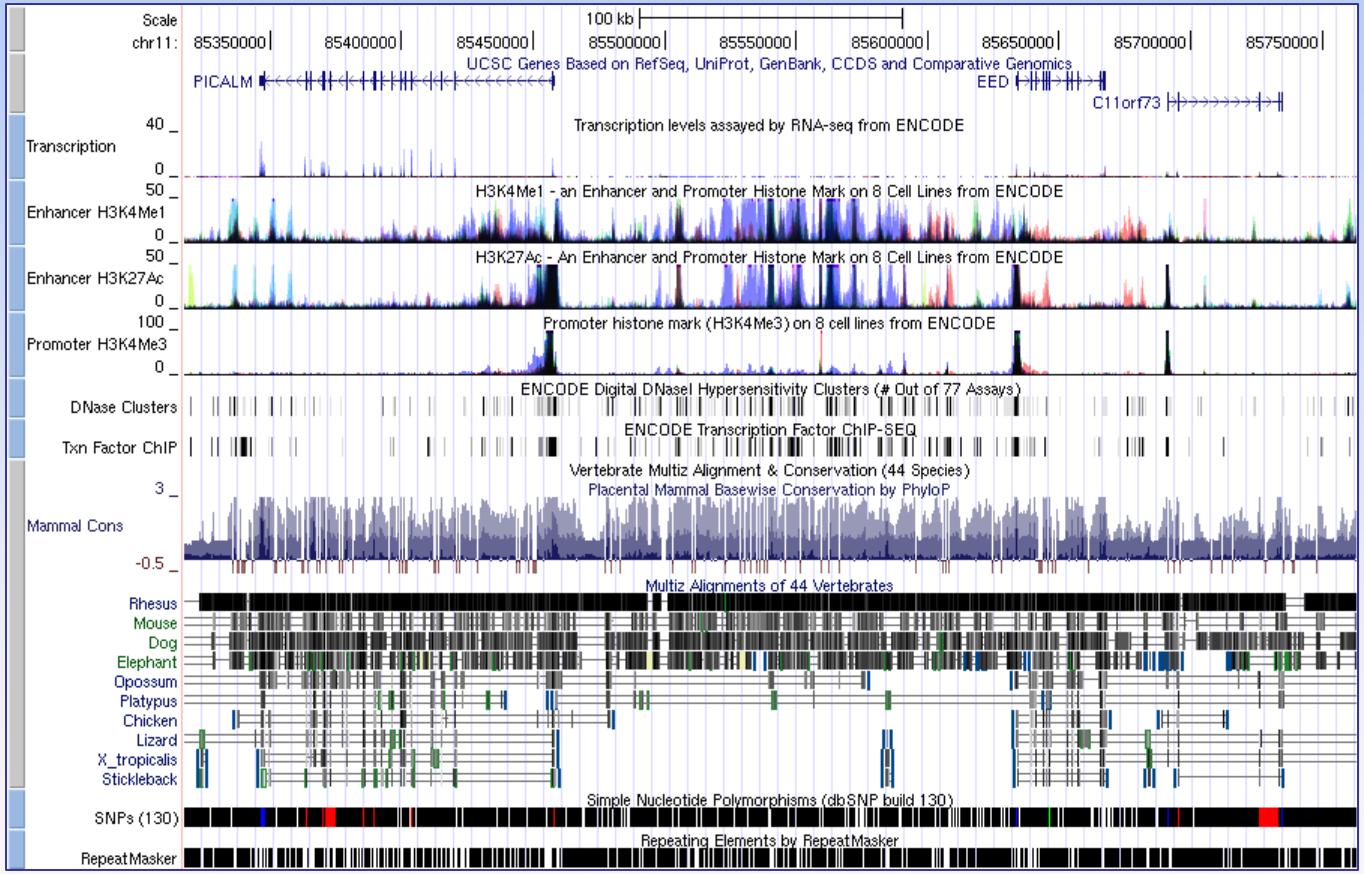
peaks

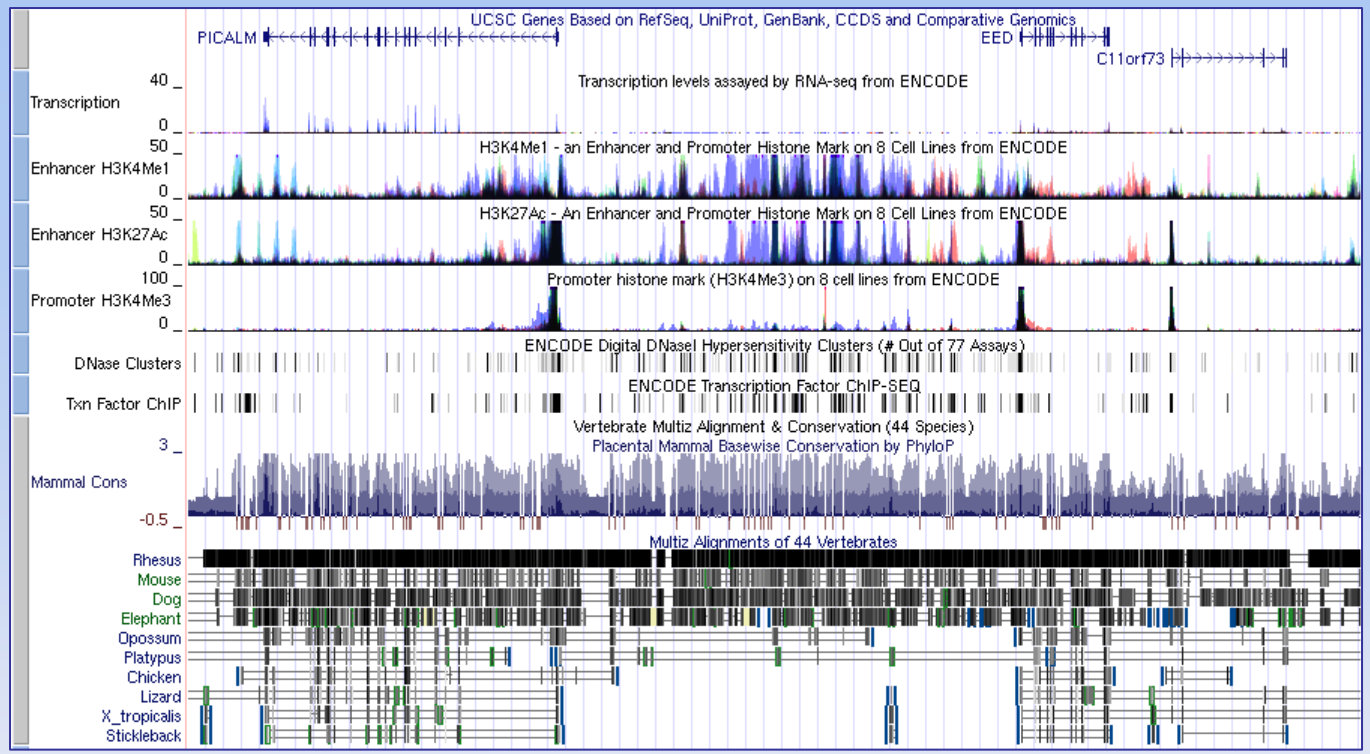
clustered peaks

Rainbow overlay for histone marks



Integrated regulatory tracks in context with other genomics information at UCSC





- 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

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

ENCODE Project at UCSC

http://www.encodeproject.org/ENCODE/

Encyclopedia of DNA Elements

Human

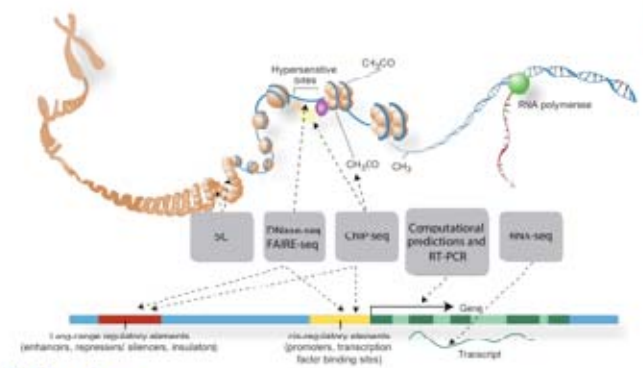
- Experiment List
- Search
- Downloads
- Genome Browser (hg19)
- Preview Browser (hg19)
- Session Gallery
- Cell Types

Mouse

- Data Summary
- Search
- Downloads
- Genome Browser (mm9)
- Preview Browser (mm9)

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.

DNA Methylation				Open Chromatin				RNA Binding Proteins					RNA Profiling						TFBS & Histones		Other			
Methyl Array	Methyl RRBS	Methyl-seq		DNase-DGF	DNase-seq	FAIRE-seq		RIP Gene ST	RIP Tiling Array	RIP Validation	RIP-seq		CAGE	Exon Array	RNA-chip	RNA-PET	RNA-seq		ChIP-seq (view matrix)		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 on scrollbar.

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

ENCODE

ChIP-seq Experiment Matrix *hg19*

search for: tracks files

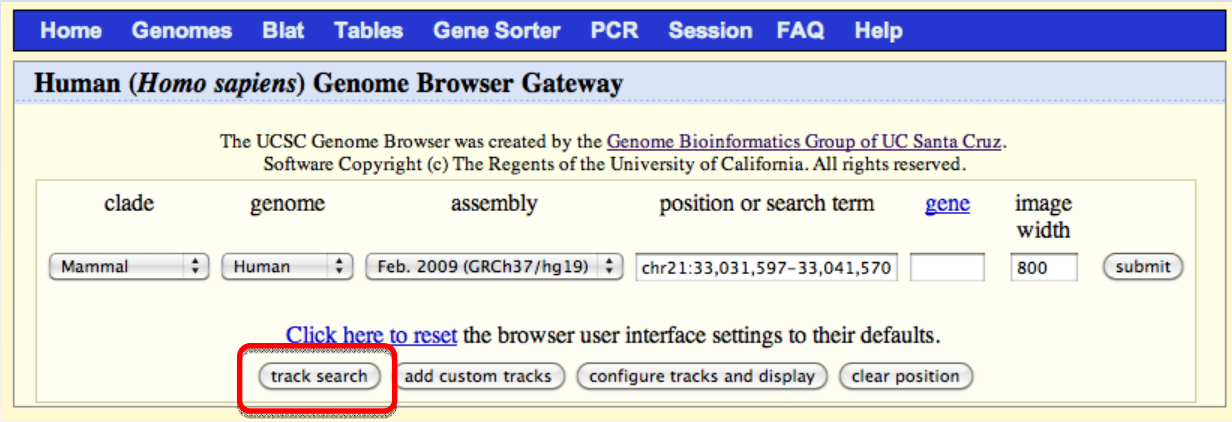
Antibody Targets

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												

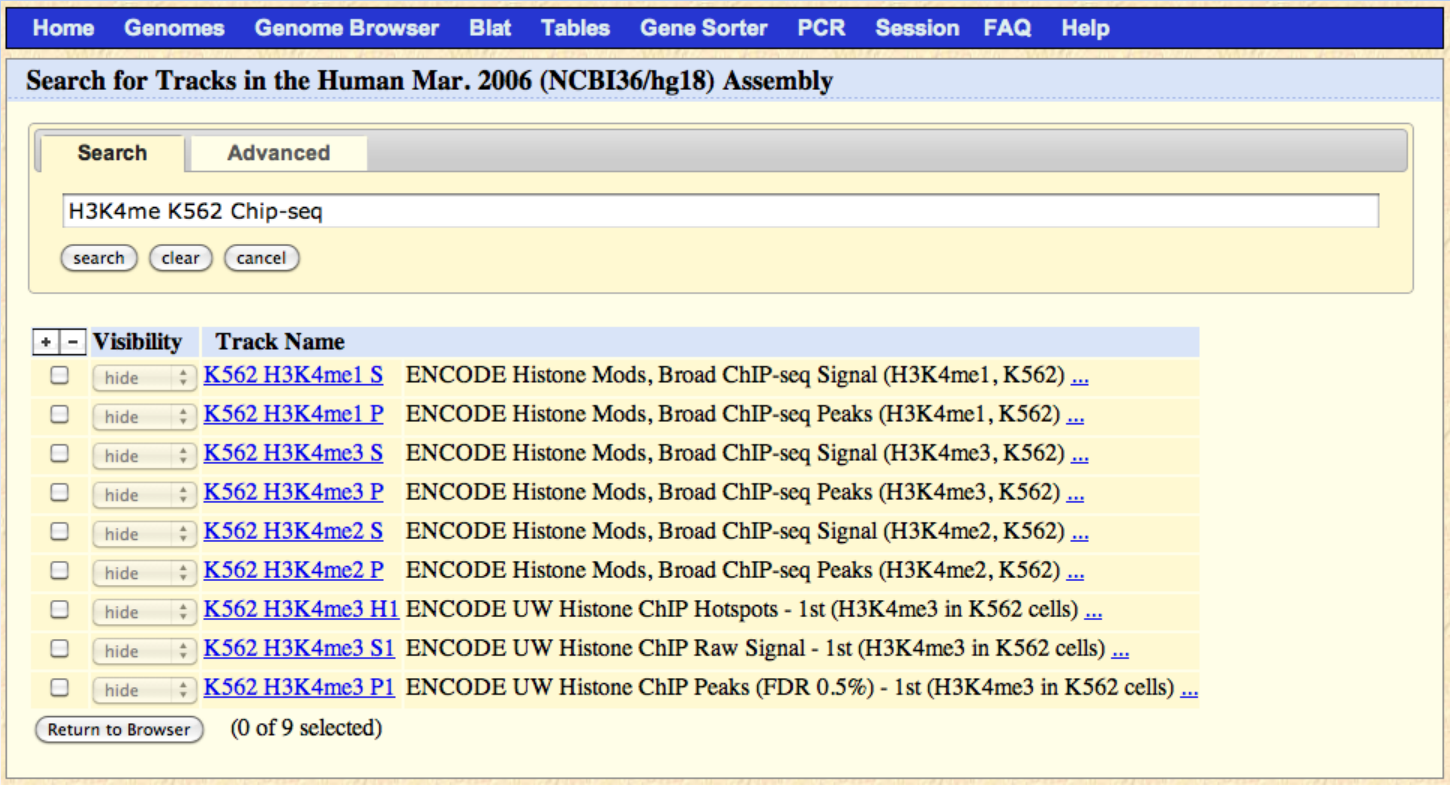
ChIP-seq experiments have their own submatrix. This is an even bigger matrix. Note size on both horizontal and vertical scrollbars.

Track Search

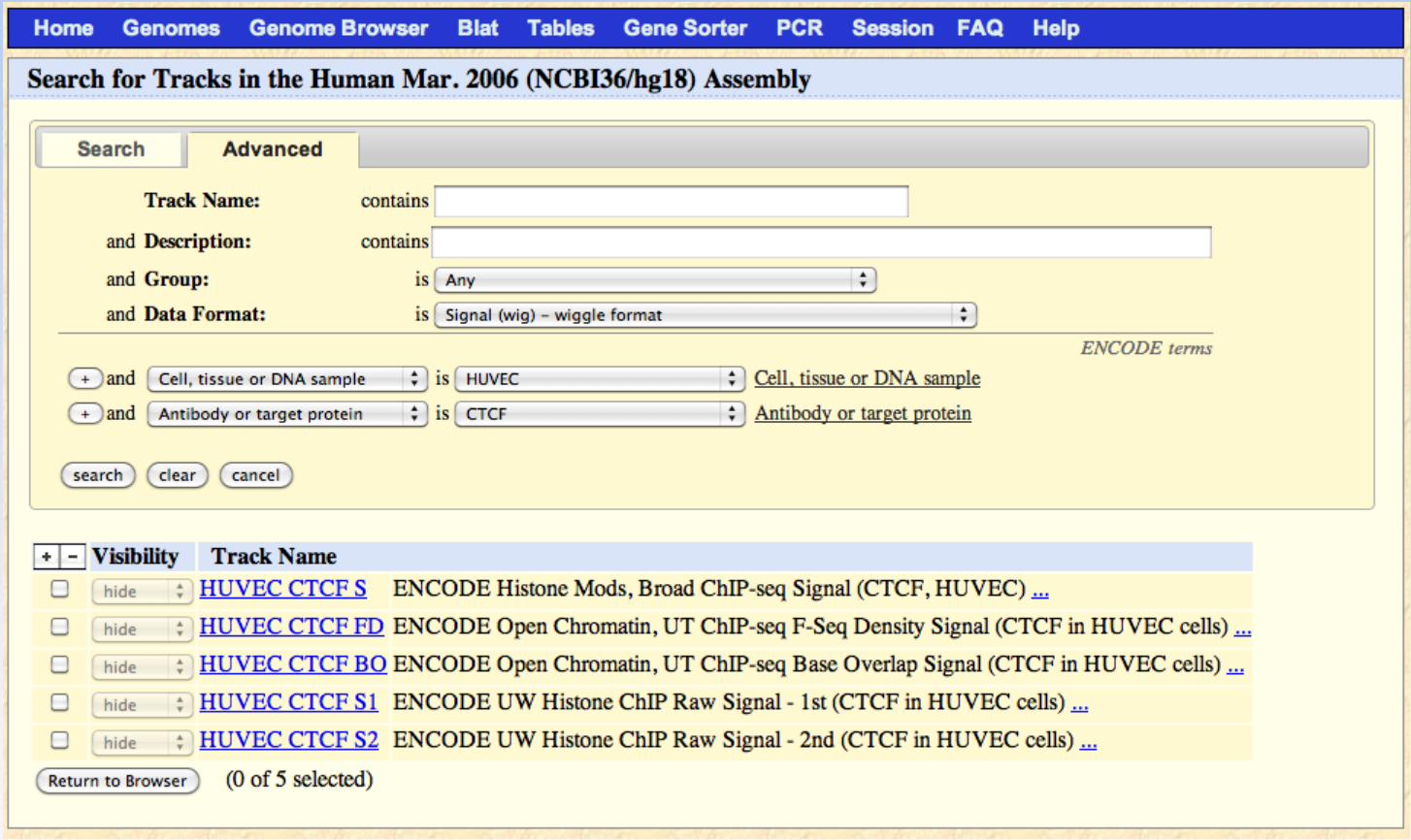
- Can do a free-form (Google-style) search or search metadata field-by-field



Free-text search



Advanced field-by-field search



Human chr12:7,935,334-7,955,316 - UCSC Genome Browser v265

http://genome.ucsc.edu/cgi-bin/hgTracks

Home | Genomes | Blat | Tables | Gene Sorter | PCR | DNA | Convert | PDF | Session | Ensembl

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

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search chr12:7,935,334-7,955,316 gene jump clear size 19,983 bp. configure

chr12 (p13.31) 12.3 12.1 12q12 14.1 q15 21.31 22 23.1 23.3

Scale chr12: 5 kb hg19 7940000 7945000 7950000 7955000

Basic Gene Annotation Set from ENCODE/GENCODE Version 11

NECAP1 Y_RNA NANOG NANOG Y_RNA

Layered H3K27Ac — H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE

DNase Clusters — Digital DNaseI Hypersensitivity Clusters from ENCODE

Txn Factor ChIP — Transcription Factor ChIP-seq from ENCODE

Vertebrate Cons — Vertebrate Basewise Conservation by PhyloP

Common SNPs(135) — Simple Nucleotide Polymorphisms (dbSNP 135) Found in >= 1% of Samples

RepeatMasker — Repeating Elements by RepeatMasker

move start < 2.0 >

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

move end < 2.0 >

track search | default tracks | default order | hide all | add custom tracks | track hubs | configure | reverse | resize | refresh

collapse all

Use drop-down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.

expand all

+ Mapping and Sequencing Tracks refresh

+ Phenotype and Disease Associations refresh

− Genes and Gene Prediction Tracks refresh

UCSC Genes hide | Old UCSC Genes hide | Alt Events hide | GENCODE Genes V11 pack | GENCODE Genes V10 hide | GENCODE Genes V7 hide

Acknowledgements

Jim Kent
Kate Rosenbloom
Tim Dreszer
Katrina Learned
Brian Lee

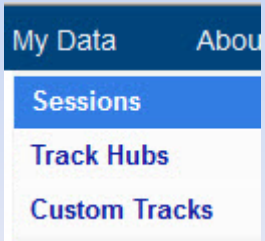
Browser Team

Hands-on Exercises

p 31 – questions, steps (Module 3)
p 22 – solutions (screen grab) (Module2)

http:// genome.ucsc.edu

My data, Sessions:



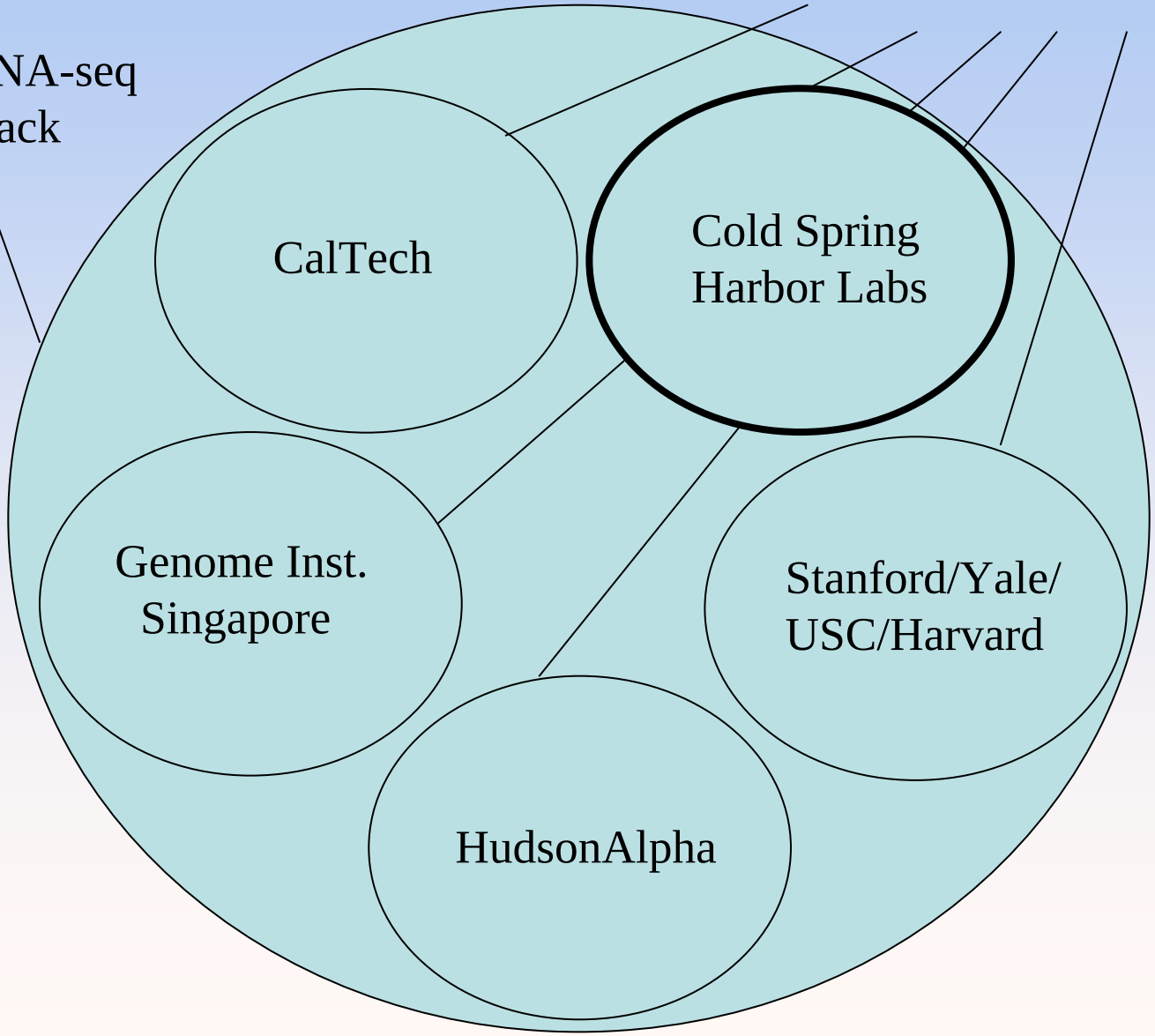
user: korea2014

pwd: ******

ENCODE RNA-seq track

5 composite tracks,
each with many subtracks

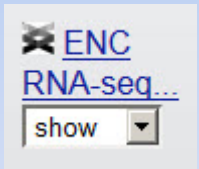
ENC RNA-seq
super track

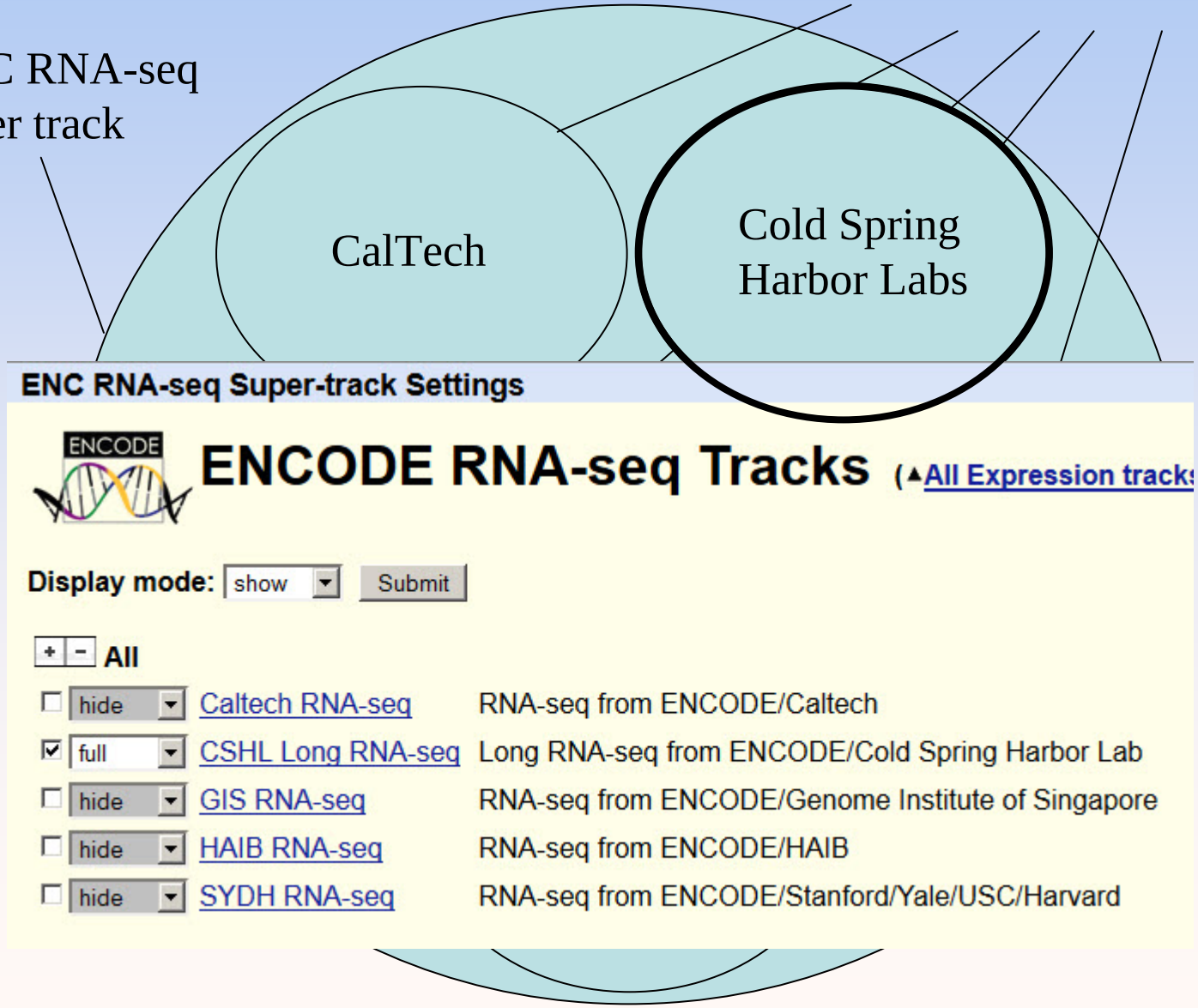


ENCODE RNA-seq track

5 composite tracks,
each with many subtracks

ENC RNA-seq
super track





ENCODE RNA-seq track

Cold Spring Harbor Labs

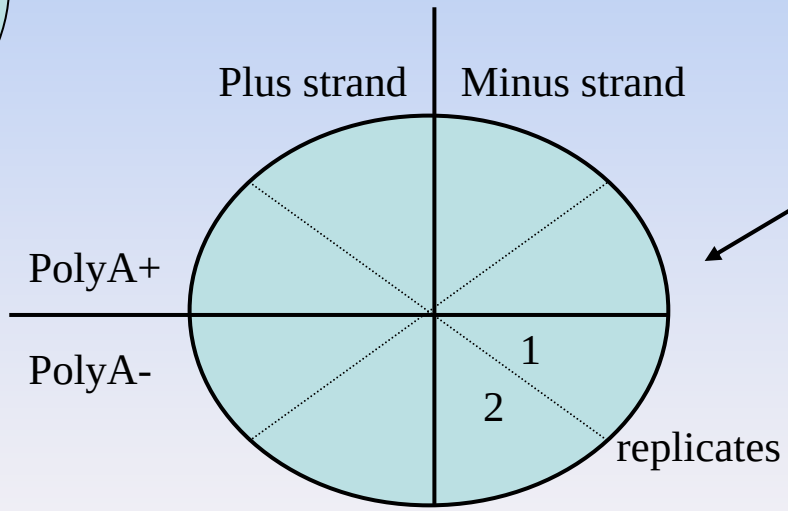
Multi-dimensional data

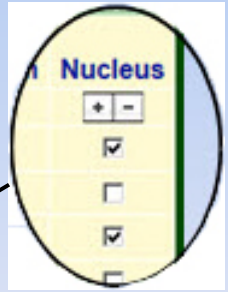
Localization / Cell Line	Whole Cell	Chromatin	Cytosol	Nucleolus	Nucleoplasm	Nucleus
GM12878 (Tier 1)	☐		☐	☐		☑
H1-hESC (Tier 1)	☐		☐			☐
K562 (Tier 1)	☐	☐	☐	☐	☐	☑
A549 (Tier 2)	☐		☐			☐
B cells CD20+ (Tier 2)	☐					
HeLa-S3 (Tier 2)	☐		☐			☐
HepG2 (Tier 2)	☐		☐			☐
HUVEC (Tier 2)	☐		☐			☐
IMR90 (Tier 2)	☐		☐			☐
MCF-7 (Tier 2)	☐		☐			☐
Monocytes CD14+ (Tier 2)	☐					
SK-N-SH (Tier 2)	☐		☐			☐
AG04450	☐					
BJ	☐					
CD34+ Mobilized	☐					
HAoAF	☐					
HAoEC	☐					

ENCODE RNA-seq track

Cold Spring Harbor Labs

Multi-dimensional data





Cell Line[1]	Localization[2]	RNA Extract[3]	Views[4]	Replicate
GM12878	Nucleus	PolyA-	Minus Signal	1st
GM12878	Nucleus	PolyA-	Minus Signal	2nd
GM12878	Nucleus	PolyA-	Plus Signal	1st
GM12878	Nucleus	PolyA-	Plus Signal	2nd
GM12878	Nucleus	PolyA+	Minus Signal	1st
GM12878	Nucleus	PolyA+	Minus Signal	2nd
GM12878	Nucleus	PolyA+	Plus Signal	1st
GM12878	Nucleus	PolyA+	Plus Signal	2nd