National
Human Genome
Research Institute

How to Display and Download ENCODE Data

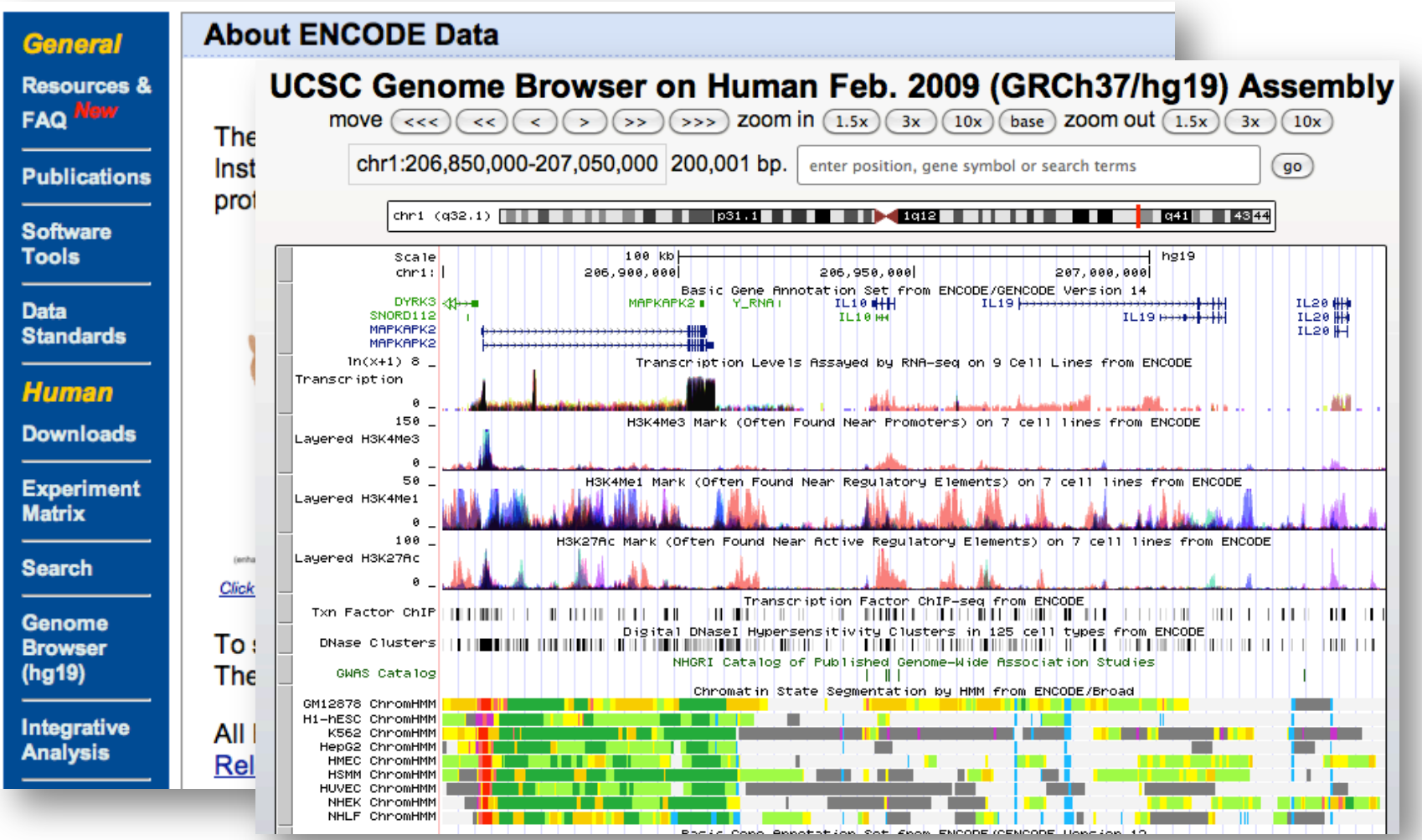
Mike Pazin

NHGRI, NIH

Goals of ENCODE

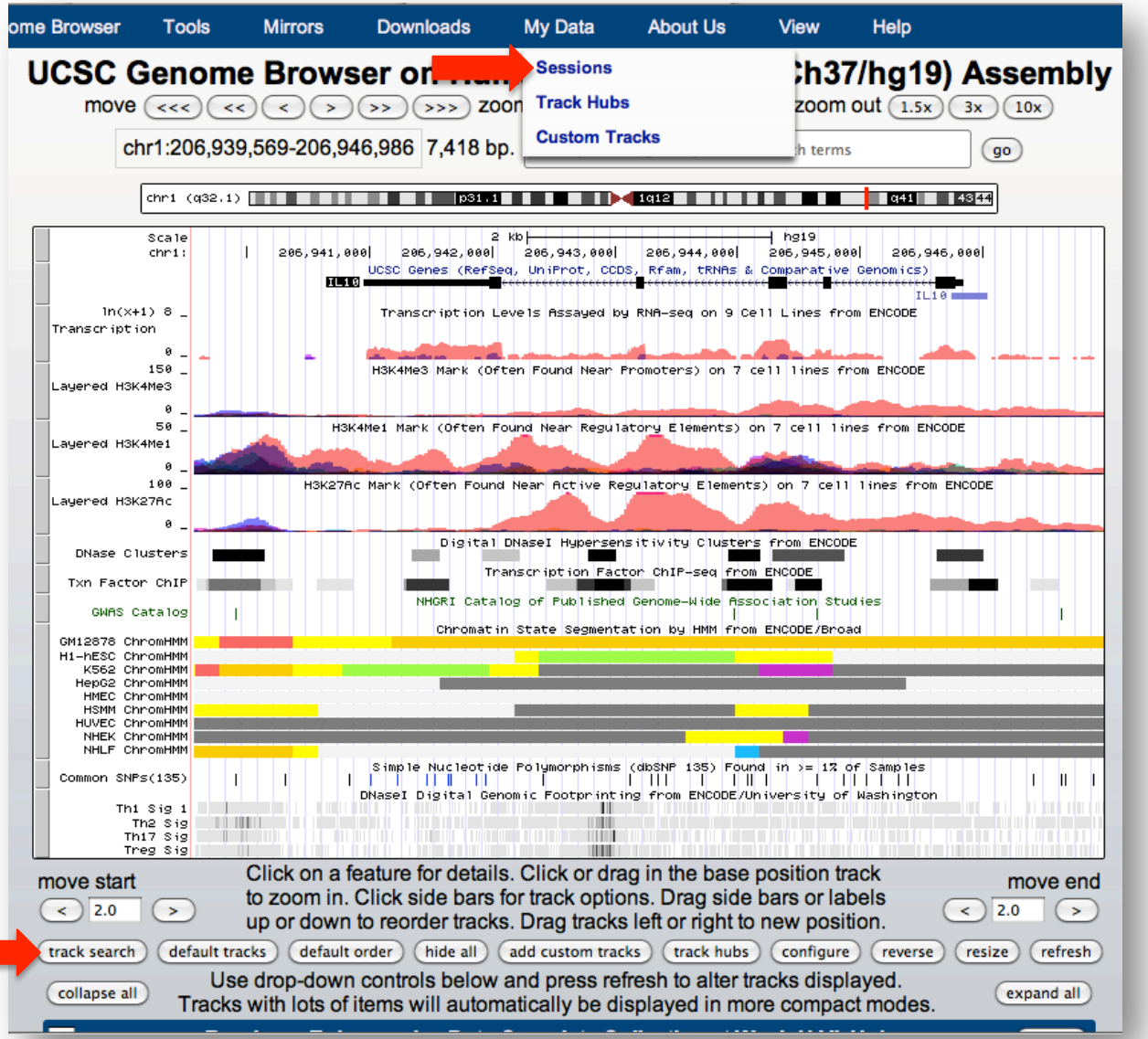
- Catalog all functional elements in the genome
- Freely available resource for all biologists
- Human as well as other species
- Project components:
  - Data generation
  - Data analysis
  - Data repository

Display of ENCODE Data



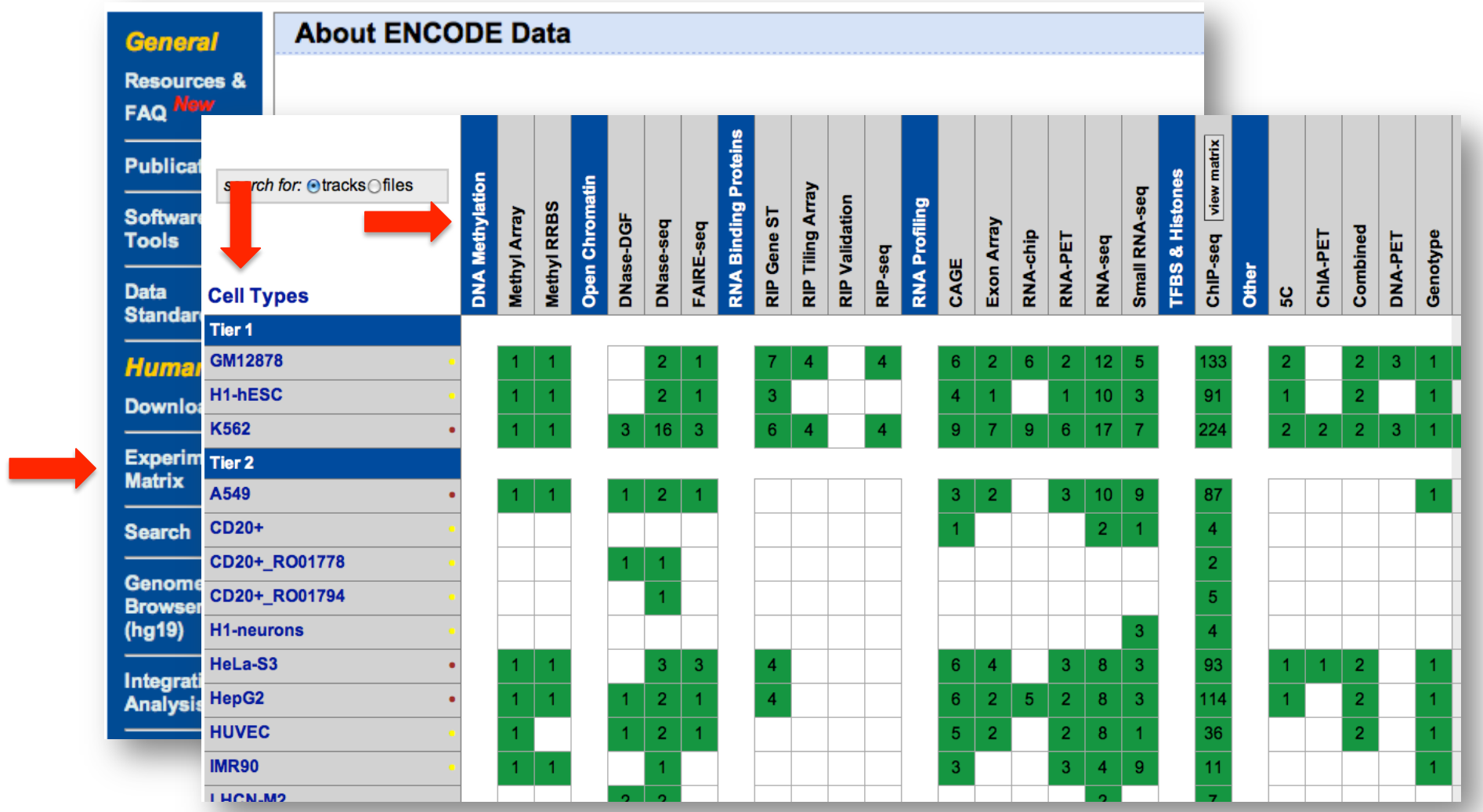
http://encodeproject.org

Configuring The Display



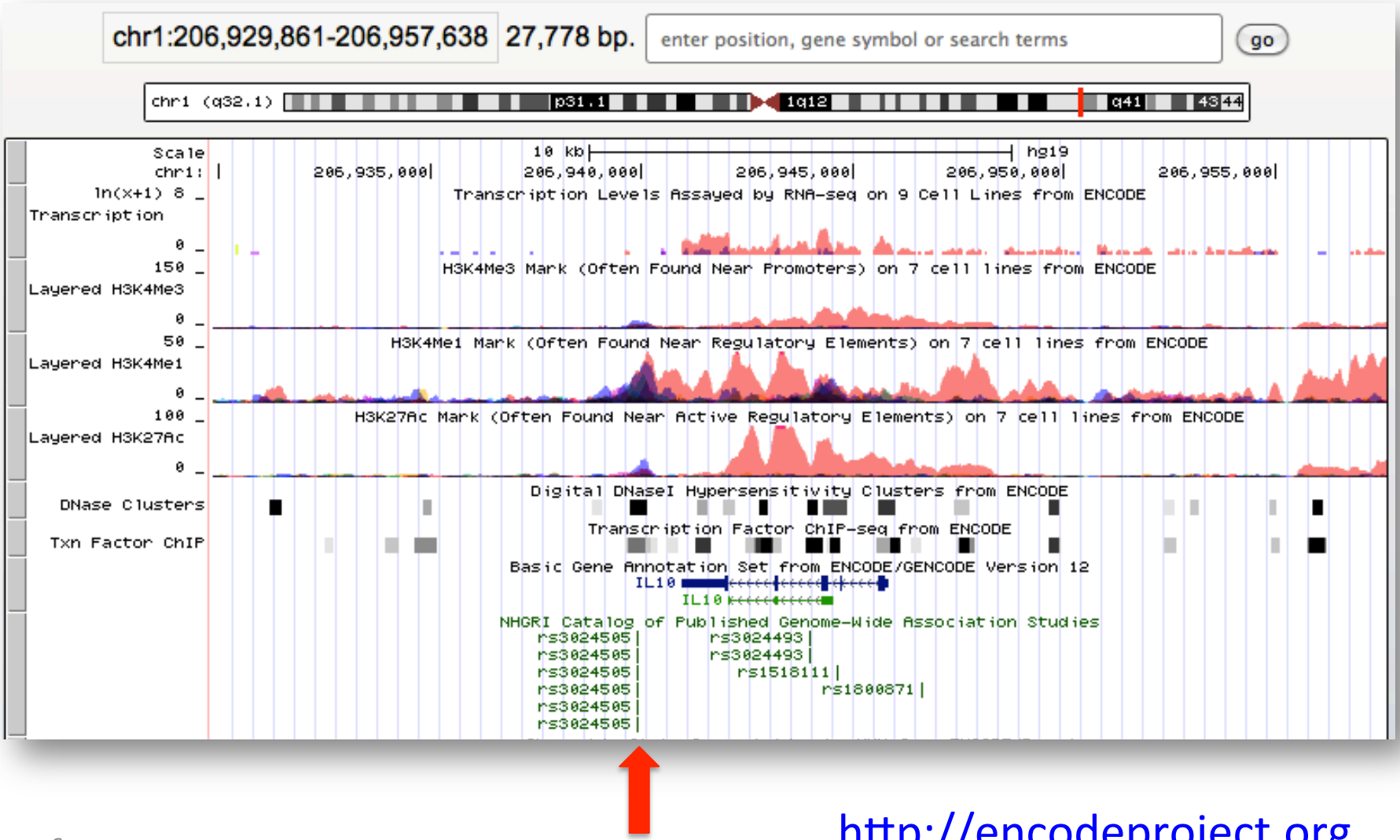
Handout

ENCODE Experiment Matrix



http://encodeproject.org

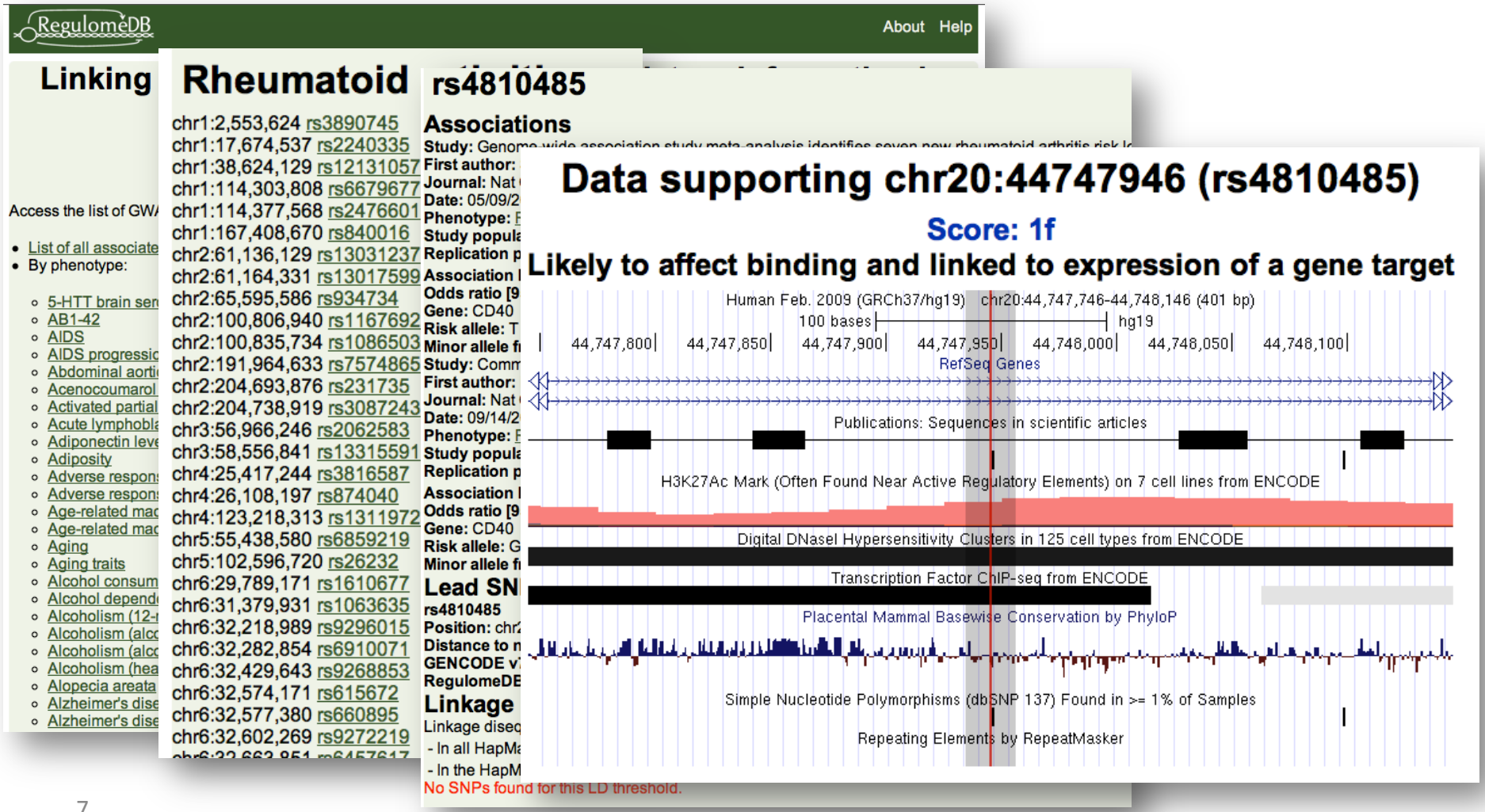
ENCODE Browser
Viewing Locus of Interest



http://encodeproject.org

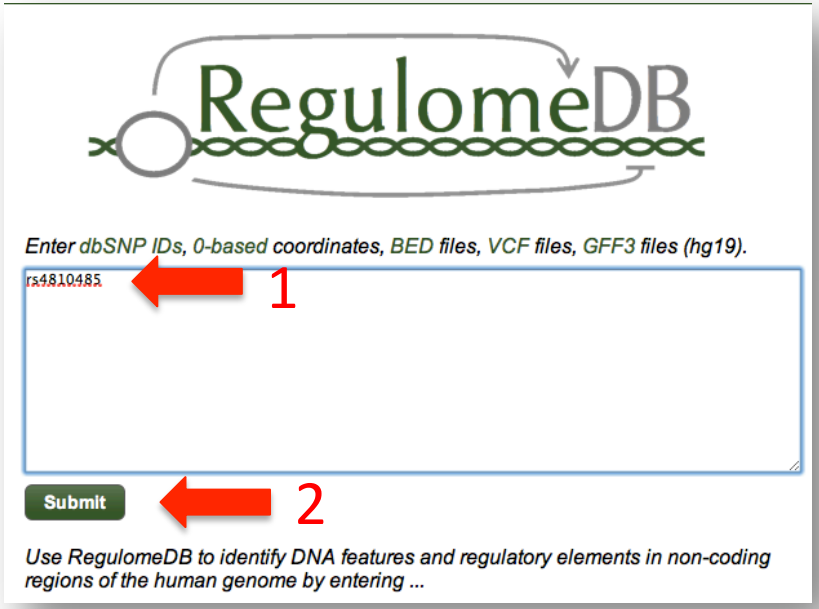
RegulomeDB Disease Database

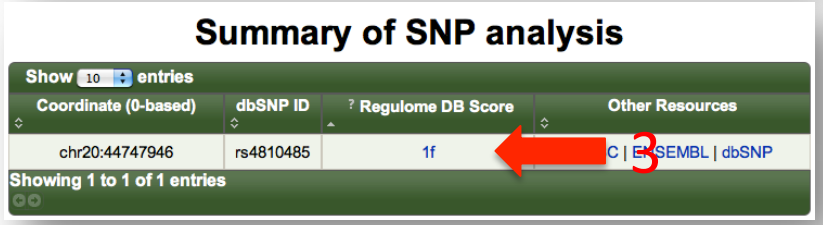
http://regulome.stanford.edu/GWAS; Schaub…Snyder, Genome Research 22-1748,2012

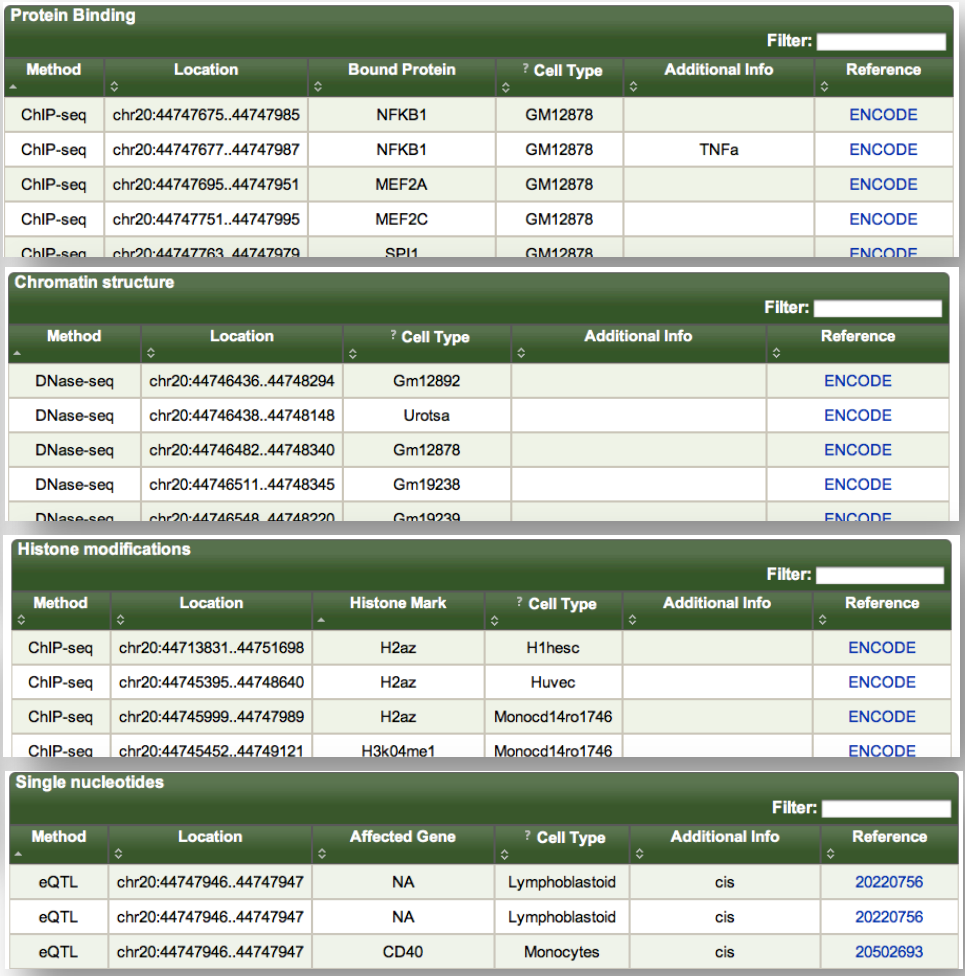


RegulomeDB Annotation of SNPs

http://regulome.stanford.edu/ ; Boyle…Snyder, Genome Research 22-1790,2012





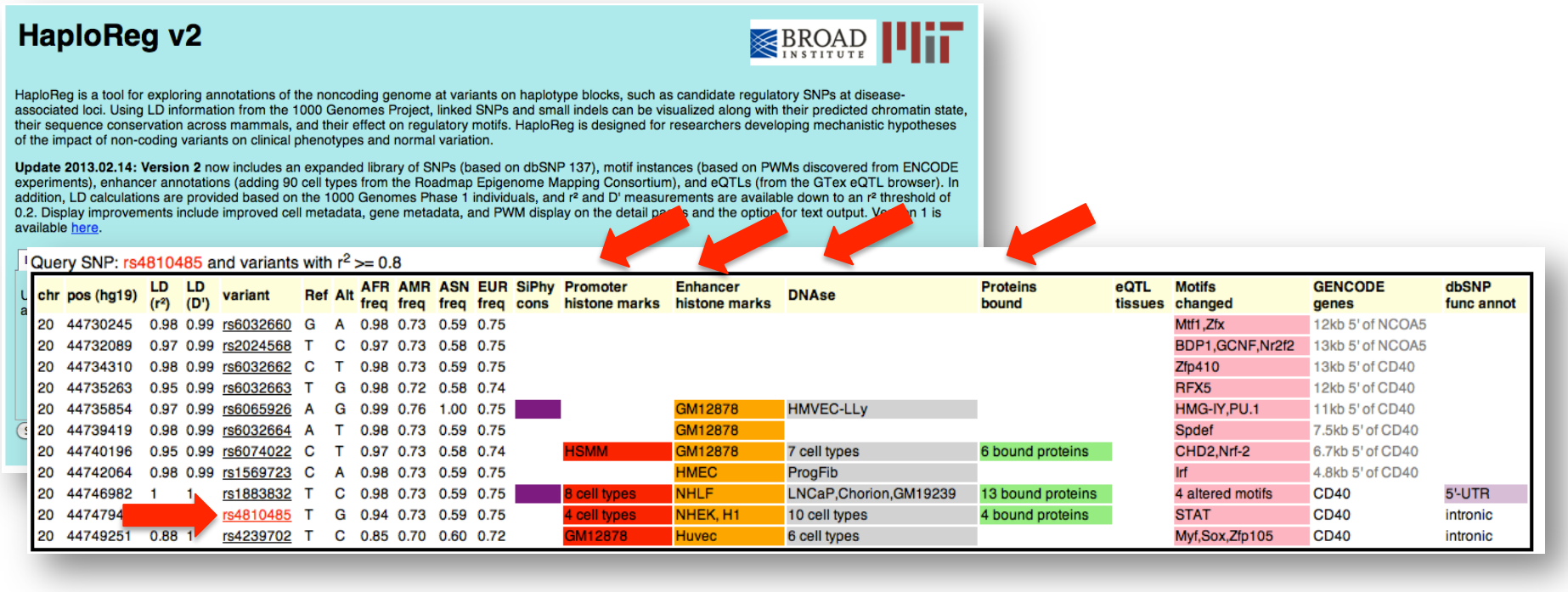


http://www.genome.gov/2755193

HaploReg v2 Annotation of SNPs

www.broadinstitute.org/mammals/haploreg/

Ward and Kellis, Nucleic Acids Research 40-D930, 2011



HaploReg v2

HaploReg is a tool for exploring annotations of the noncoding genome at variants on haplotype blocks, such as candidate regulatory SNPs at disease-associated loci. Using LD information from the 1000 Genomes Project, linked SNPs and small indels can be visualized along with their predicted chromatin state, their sequence conservation across mammals, and their effect on regulatory motifs. HaploReg is designed for researchers developing mechanistic hypotheses of the impact of non-coding variants on clinical phenotypes and normal variation.

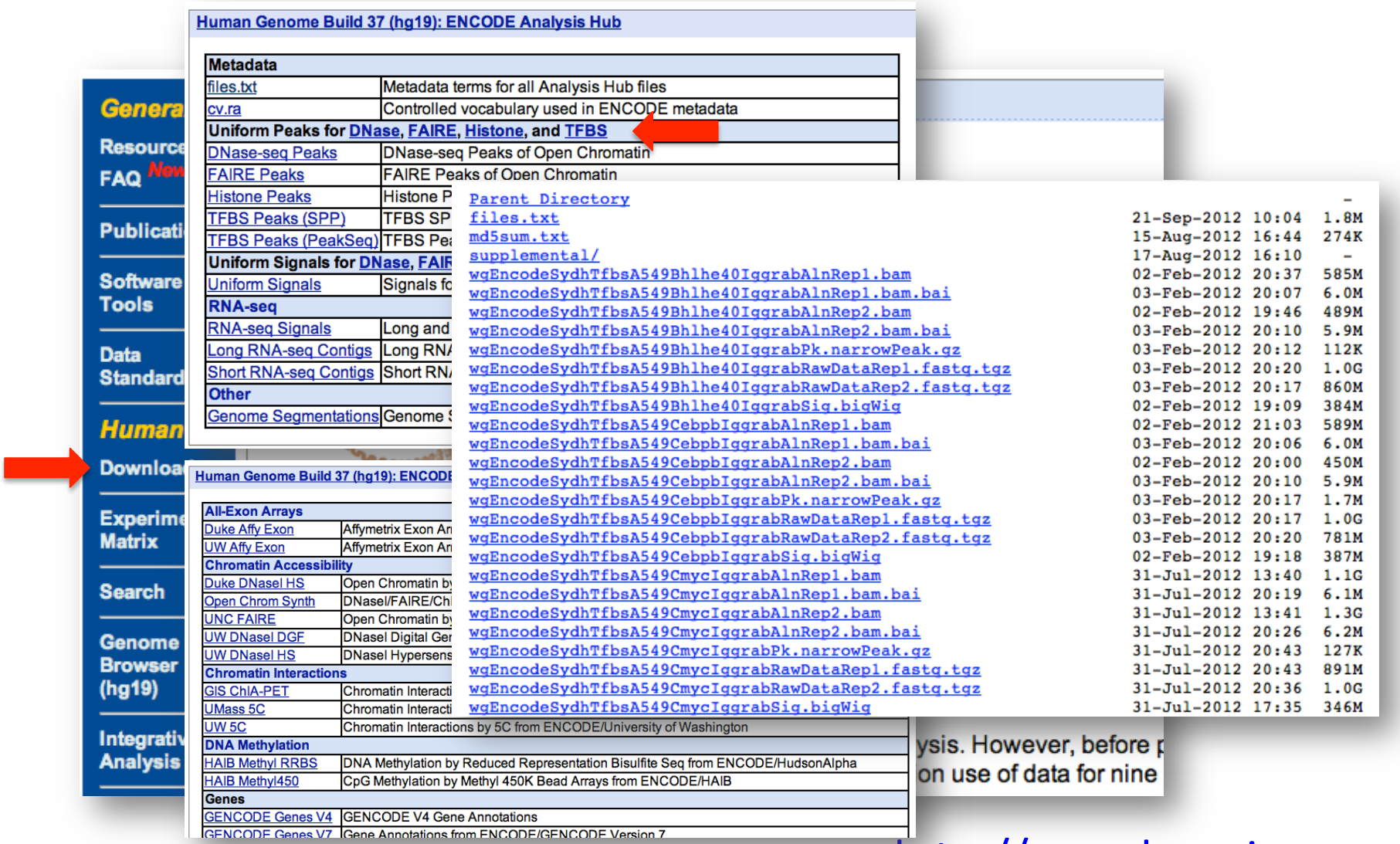
Update 2013.02.14: Version 2 now includes an expanded library of SNPs (based on dbSNP 137), motif instances (based on PWMs discovered from ENCODE experiments), enhancer annotations (adding 90 cell types from the Roadmap Epigenome Mapping Consortium), and eQTLs (from the GTex eQTL browser). In addition, LD calculations are provided based on the 1000 Genomes Phase 1 individuals, and r² and D' measurements are available down to an r² threshold of 0.2. Display improvements include improved cell metadata, gene metadata, and PWM display on the detail pa[illegible]s and the option for text output. Ve[illegible]n 1 is available here.

Query SNP: rs4810485 and variants with r^2 >= 0.8

chr	pos (hg19)	LD (r^2)	LD (D')	variant	Ref	Alt	AFR freq	AMR freq	ASN freq	EUR freq	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse	Proteins bound	eQTL tissues	Motifs changed	GENCODE genes	dbSNP func annot
20	44730245	0.98	0.99	rs6032660	G	A	0.98	0.73	0.59	0.75							Mtf1,Zfx	12kb 5' of NCOA5	
20	44732089	0.97	0.99	rs2024568	T	C	0.97	0.73	0.58	0.75							BDP1,GCNF,Nr2f2	13kb 5' of NCOA5	
20	44734310	0.98	0.99	rs6032662	C	T	0.98	0.73	0.59	0.75							Zfp410	13kb 5' of CD40	
20	44735263	0.95	0.99	rs6032663	T	G	0.98	0.72	0.58	0.74							RFX5	12kb 5' of CD40	
20	44735854	0.97	0.99	rs6065926	A	G	0.99	0.76	1.00	0.75			GM12878	HMVEC-LLy			HMG-IY,PU.1	11kb 5' of CD40	
20	44739419	0.98	0.99	rs6032664	A	T	0.98	0.73	0.59	0.75			GM12878				Spdef	7.5kb 5' of CD40	
20	44740196	0.95	0.99	rs6074022	C	T	0.97	0.73	0.58	0.74		HSMM	GM12878	7 cell types	6 bound proteins		CHD2,Nrf-2	6.7kb 5' of CD40	
20	44742064	0.98	0.99	rs1569723	C	A	0.98	0.73	0.59	0.75			HMEC	ProgFib			Irf	4.8kb 5' of CD40	
20	44746982	1	1	rs1883832	T	C	0.98	0.73	0.59	0.75		8 cell types	NHLF	LNCaP,Chorion,GM19239	13 bound proteins		4 altered motifs	CD40	5'-UTR
20	4474794			rs4810485	T	G	0.94	0.73	0.59	0.75		4 cell types	NHEK, H1	10 cell types	4 bound proteins		STAT	CD40	intronic
20	44749251	0.88	1	rs4239702	T	C	0.85	0.70	0.60	0.72		GM12878	Huvec	6 cell types			Myf,Sox,Zfp105	CD40	intronic

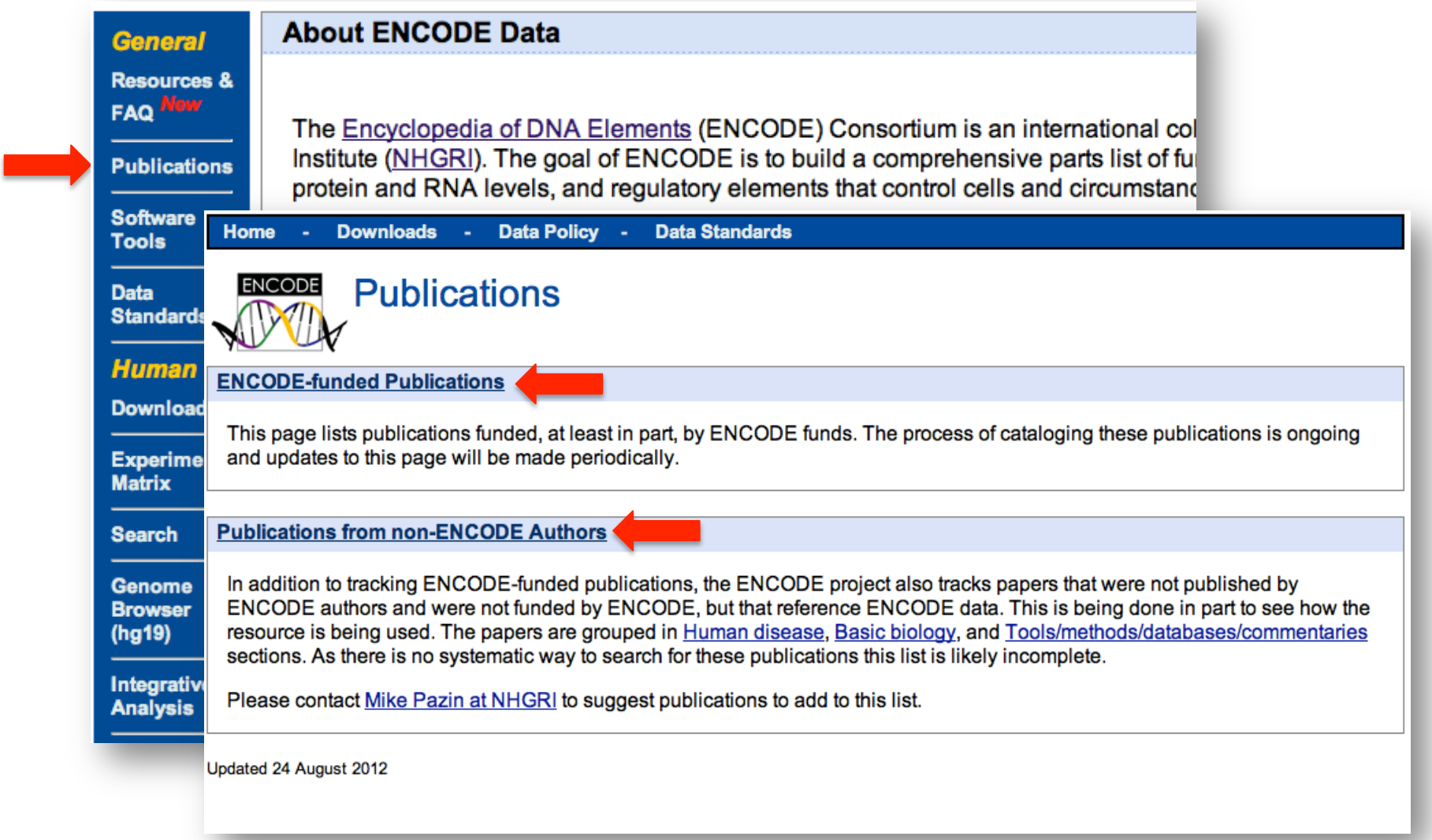
http://www.genome.gov/2755193

Downloading ENCODE Data



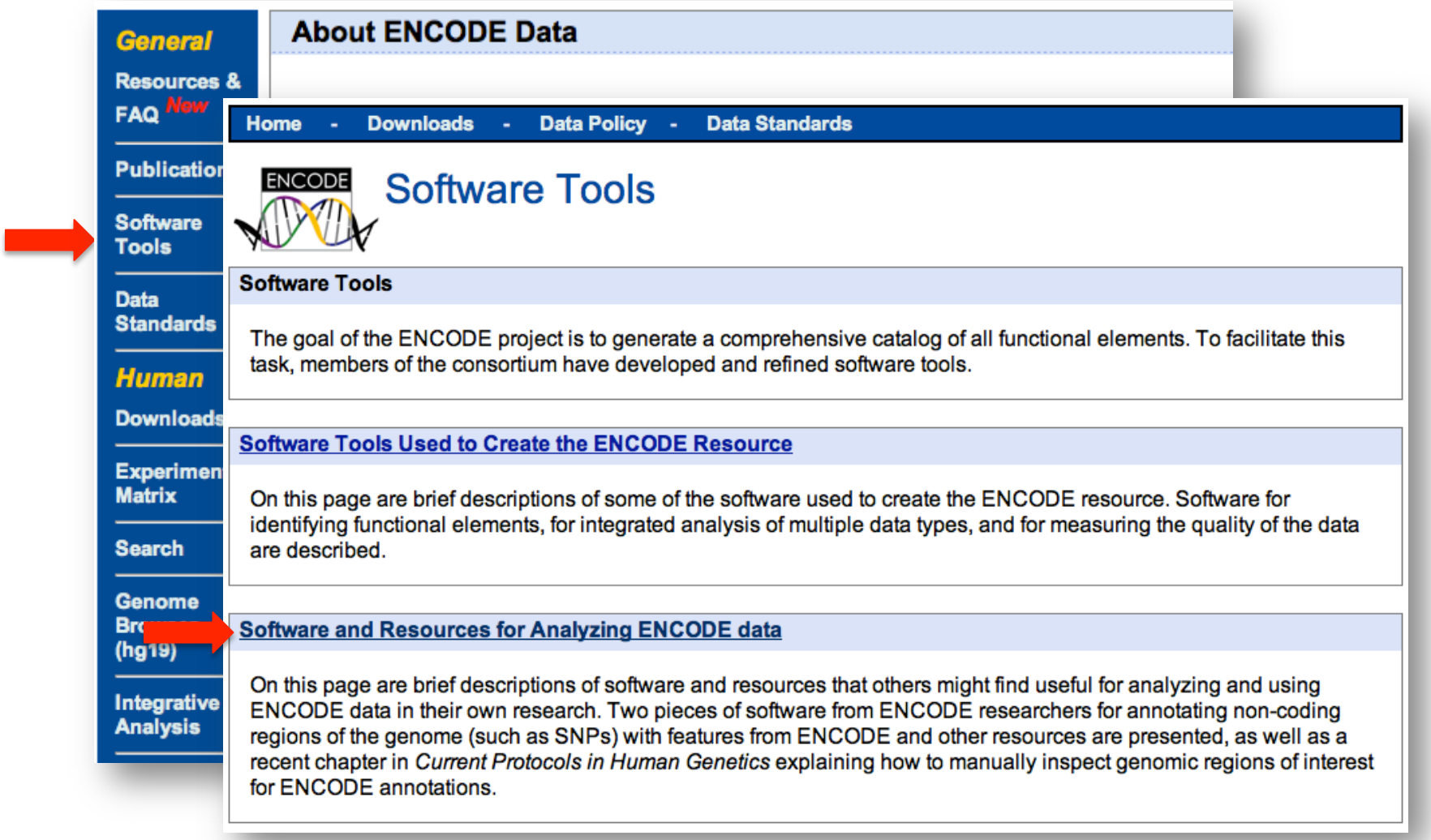
http://encodeproject.org

Publications



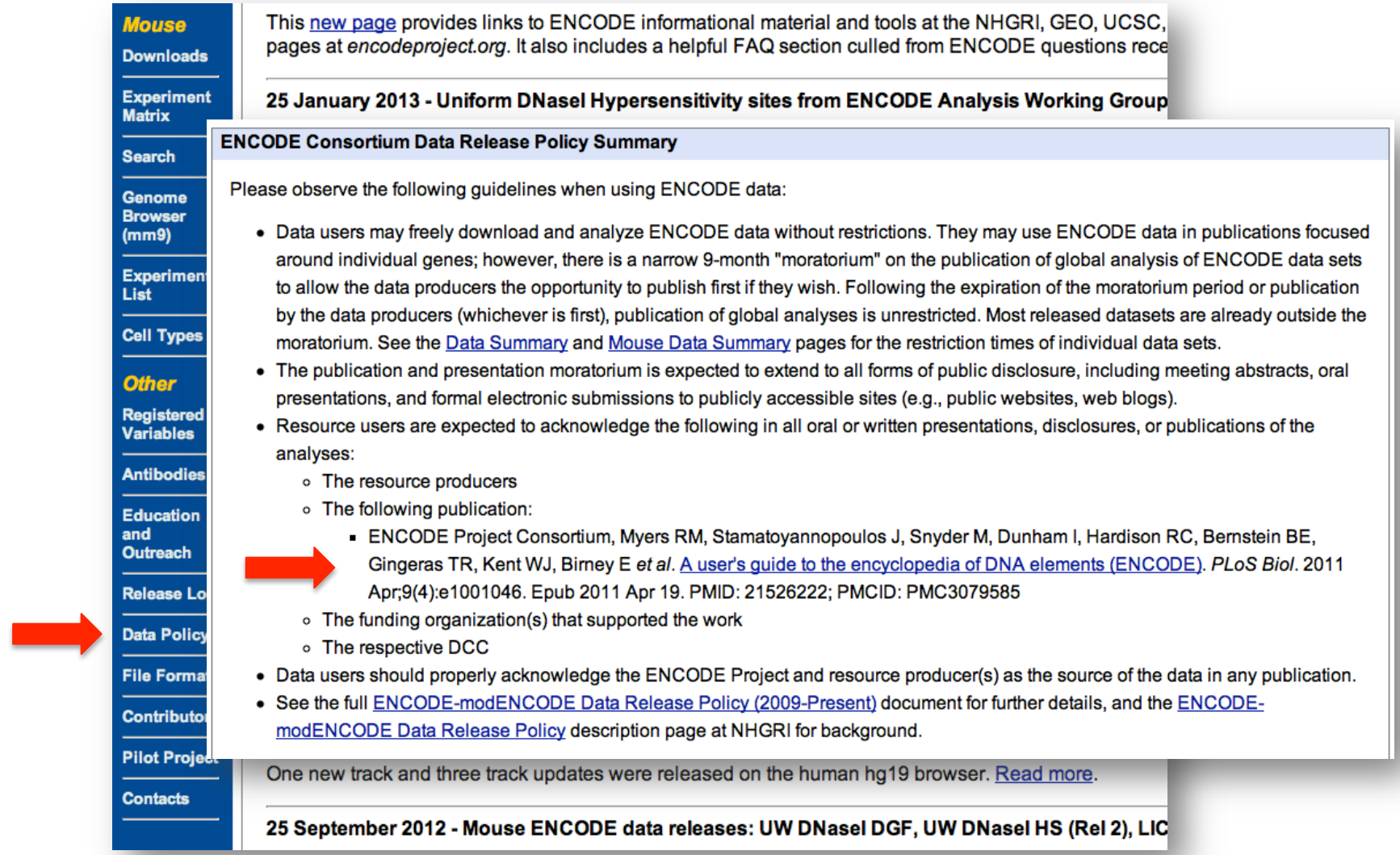
http://encodeproject.org

ENCODE Software Tools

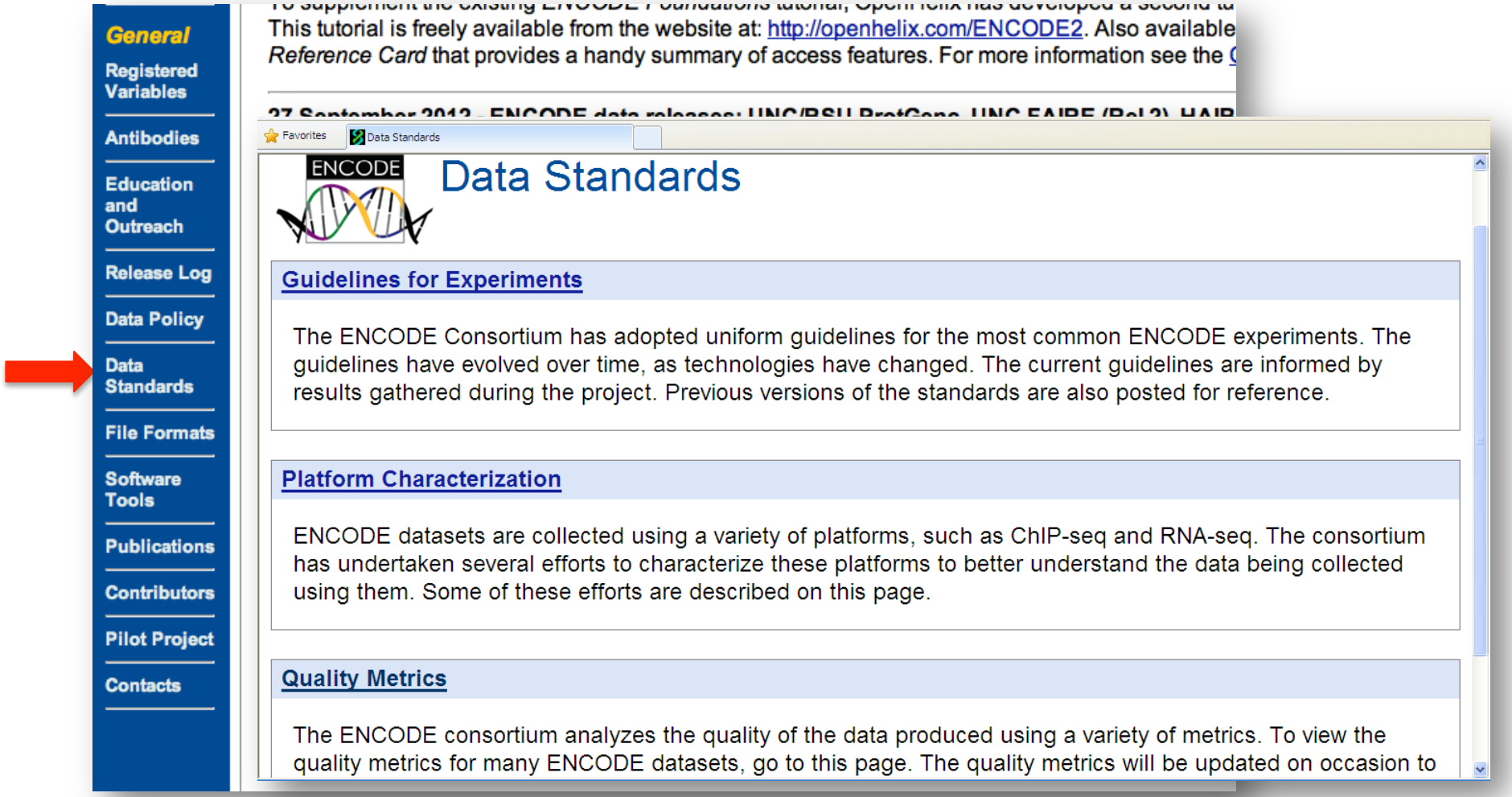


http://encodeproject.org

ENCODE Citation



ENCODE Data Standards



http://encodeproject.org

NIH Roadmap Epigenomic Mapping Consortium

A **public community resource of epigenomic data** in primary human cells and tissues.

Range of cells/tissues covered:

Currently 125 cell/tissue types represented including….
iPS and ES cells, some differentiated forms
Fetal tissues (heart, brain, kidney, lung, others)
Adult primary cells and tissues (hematopoietic, brain regions, breast cell types, liver, kidney, colon, muscle, adipocytes, others)

Some samples will have:

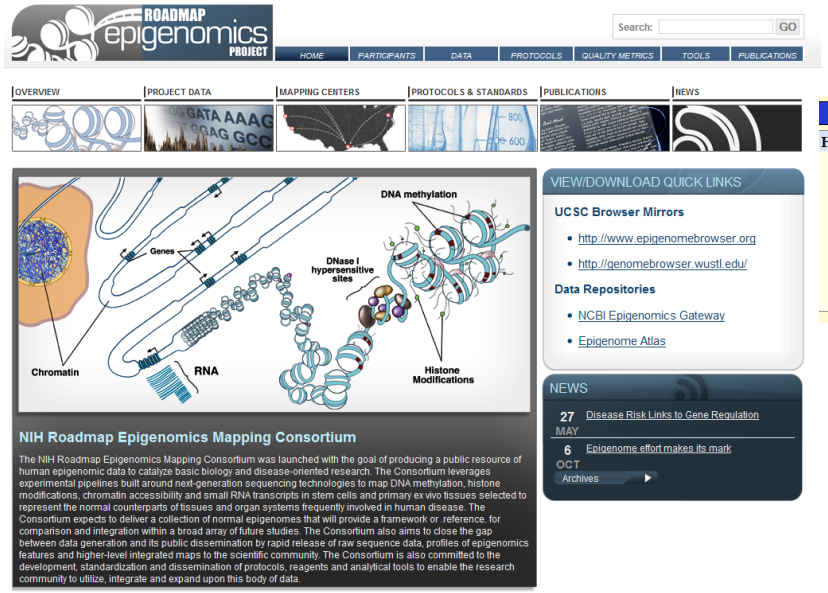
Expanded panel of histone modifications (currently 20 additional)

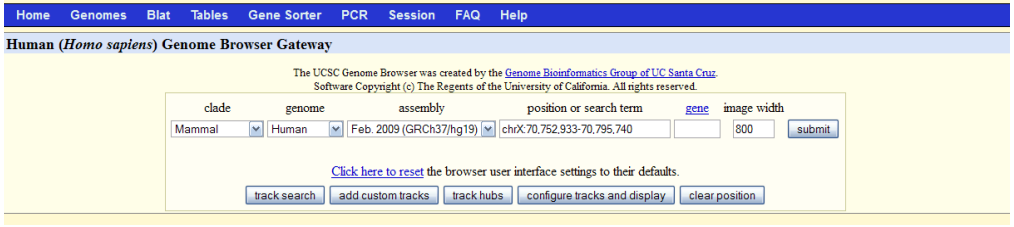
Most samples will have:

DNA methylation data (RRBS, MRE-seq, MeDIP-seq, whole genome bisulfite seq)
ChIP-seq data (currently H3K27me3, H3K36me3, H3K4me1, H3K4me3, H3K9me3)
DNase I hypersensitivity data
Gene expression data (arrays or RNA-seq)

Can download:

.wig, .bed, some .bam, some SRA, working on peak calls





View Roadmap data at http://genome.ucsc.edu via TRACK HUB – click 'track hubs' to load Roadmap data and summary data tracks on UCSC.

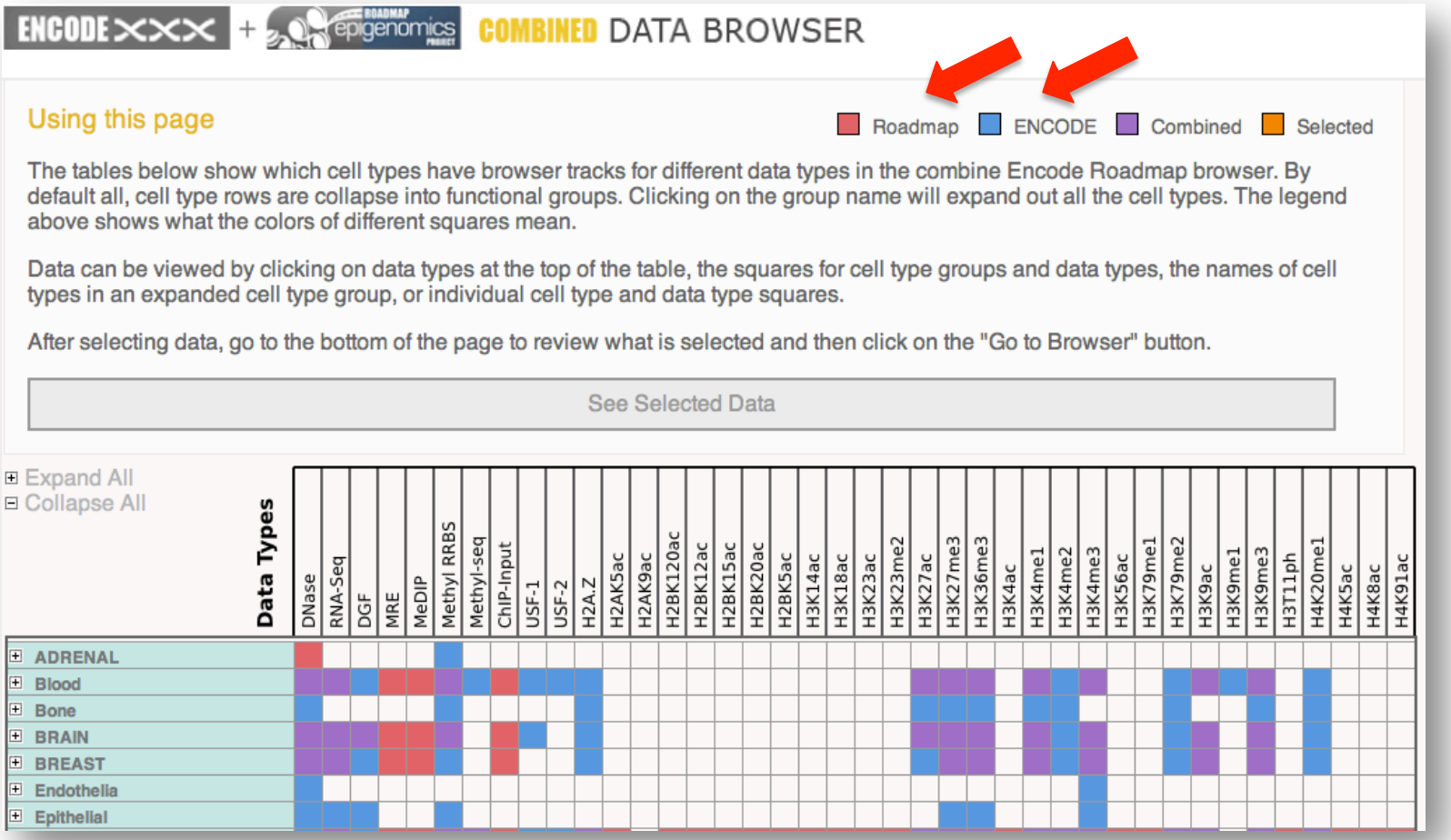
Where is the data?

Find data, protocols, and analysis/viewing tools from the Roadmap Epigenomic Mapping Consortium at these sites:

http://roadmapepigenomics.org
http://ncbi.nlm.nih.gov/epigenomics
http://ncbi.nlm.nih.gov/geo/roadmap/epigenomics
http://epigenomeatlas.org
http://vizhub.wustl.edu
http://genome.ucsc.edu (via track hub)

http://roadmapepigenomics.org – Find and view data, protocols, links to other sites associated with the program.

ENCODE Browser for ENCODE and Roadmap Epigenomics data



http://encode-roadmap.org

How Can ENCODE Data Be Used?

- A standard problem: many genetic findings for human disease map to non-protein coding regions of the human genome
  - What is the functional variant?
  - What is the target gene?
  - What is the target cell type?
  - What is the function of the variant?
- Standard Browser view for loci of interest
- HaploReg and RegulomeDB searches for loci of interest
- Search a cell type across all data types for loci of interest
- Search a data type across all cell types for loci of interest

ENCODE Production

Brad Bernstein (Manolis Kellis, Tony Kouzarides, Eric Lander, John Rinn)

Tom Gingeras (Roderic Guigo, Carrie Davis, Alexandre Reymond, David Spector, Greg Hannon, Michael Brent, Stylianos Antonarakis, Yijun Ruan, Yoshihide Hayashizaki)

Rick Myers (Barbara Wold, Ross Hardison, Flo Pauli Behn, Ali Mortazavi, Tim Reddy, Greg Cooper, Devin Absher)

Mike Snyder (Peggy Farnham, Sherman Weissman, Kevin White, Kevin Struhl, Mark Gerstein,)

John Stamatoyannopoulos (Evan Eichler, George Stamatoyannopoulos, Job Dekker, Maynard Olson, Michael Dorschner, Patrick Navas, Phil Green)

Brent Graveley (Gene Yeo, Chris Burge, Xiang-Dong Fu)

Bing Ren (Joe Ecker, Len Pennacchio, Axel Visel, Wei Wang)

Mike Cherry (Jim Kent, Eurie Hong, David Haussler, Kate Rosenbloom)

Zhiping Weng (Mark Gerstein, Manolis Kellis, Roderic Guigo, Shirley Liu, Bill Noble, Rafael Irizarry)

Tim Hubbard (Alexandre Reymond, Alfonso Valencia, David Haussler, Ewan Birney, Jim Kent, Manolis Kellis, Mark Gerstein, Michael Brent, Roderic Guigo)

Ewan Birney (Jim Kent, Mark Gerstein, Bill Noble, Peter Bickel, Ross Hardison, Zhiping Weng)

Greg Crawford (Ewan Birney, Jason Lieb, Terry Furey, Vishy Iyer)

NHGRI: Elise Feingold, Peter Good, Sherry Zhou, Yekaterina Vaydylevich, Laura Dillon, Rebecca Lowdon, Leslie Adams, Caroline Kelly, Shaila Chhibba)

Additional ENCODE Participants: Elliott Marguiles, Eric Green, Job Dekker, Laura Elnitski, Len Pennachio, Dave Gilbert, Jochen Wittbrodt

Questions? pazinm@mail.nih.gov