National
Human Genome
Research Institute

How to Display and Download ENCODE Data

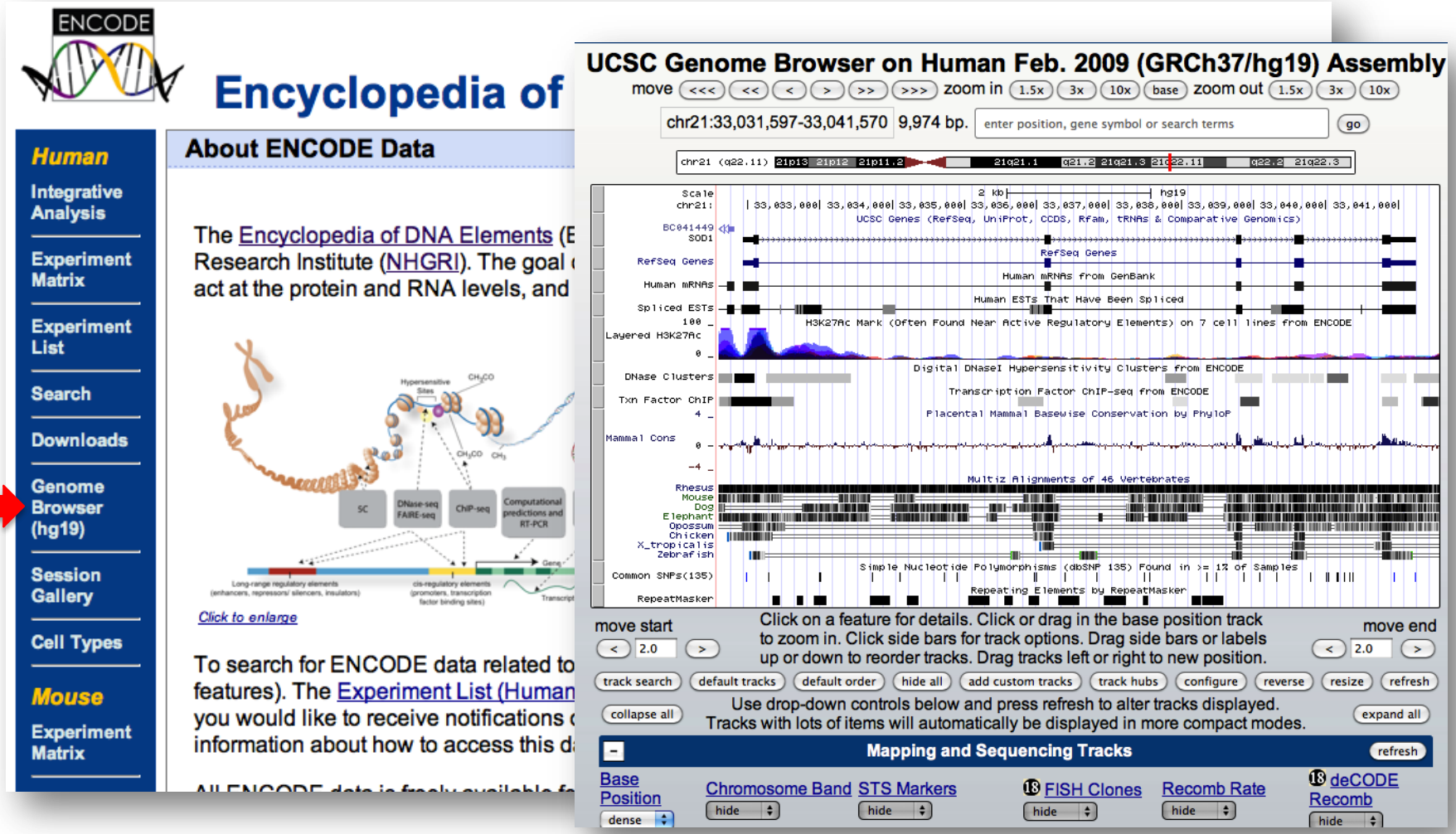
Mike Pazin

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

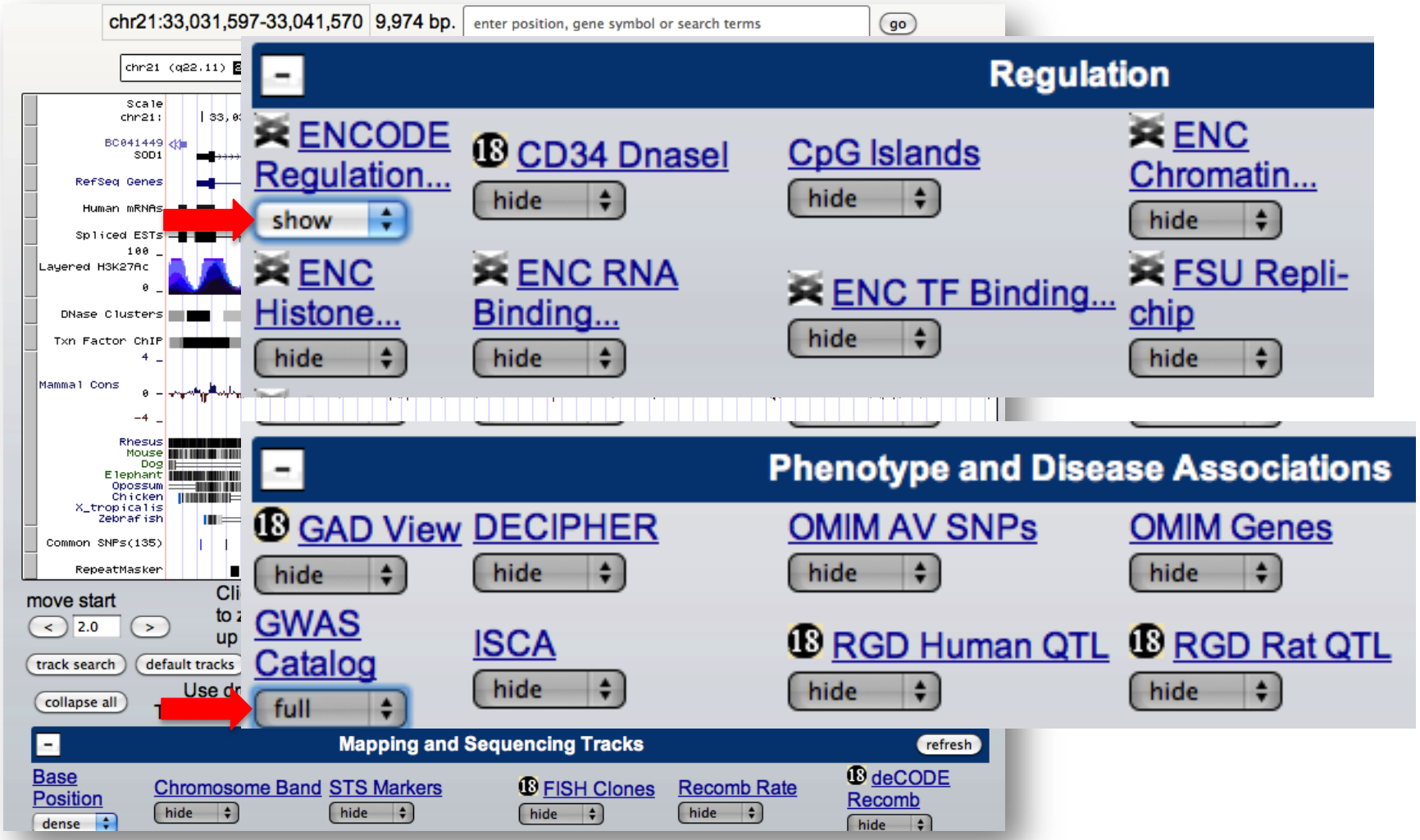
How Can ENCODE Data Be Displayed?



Handout

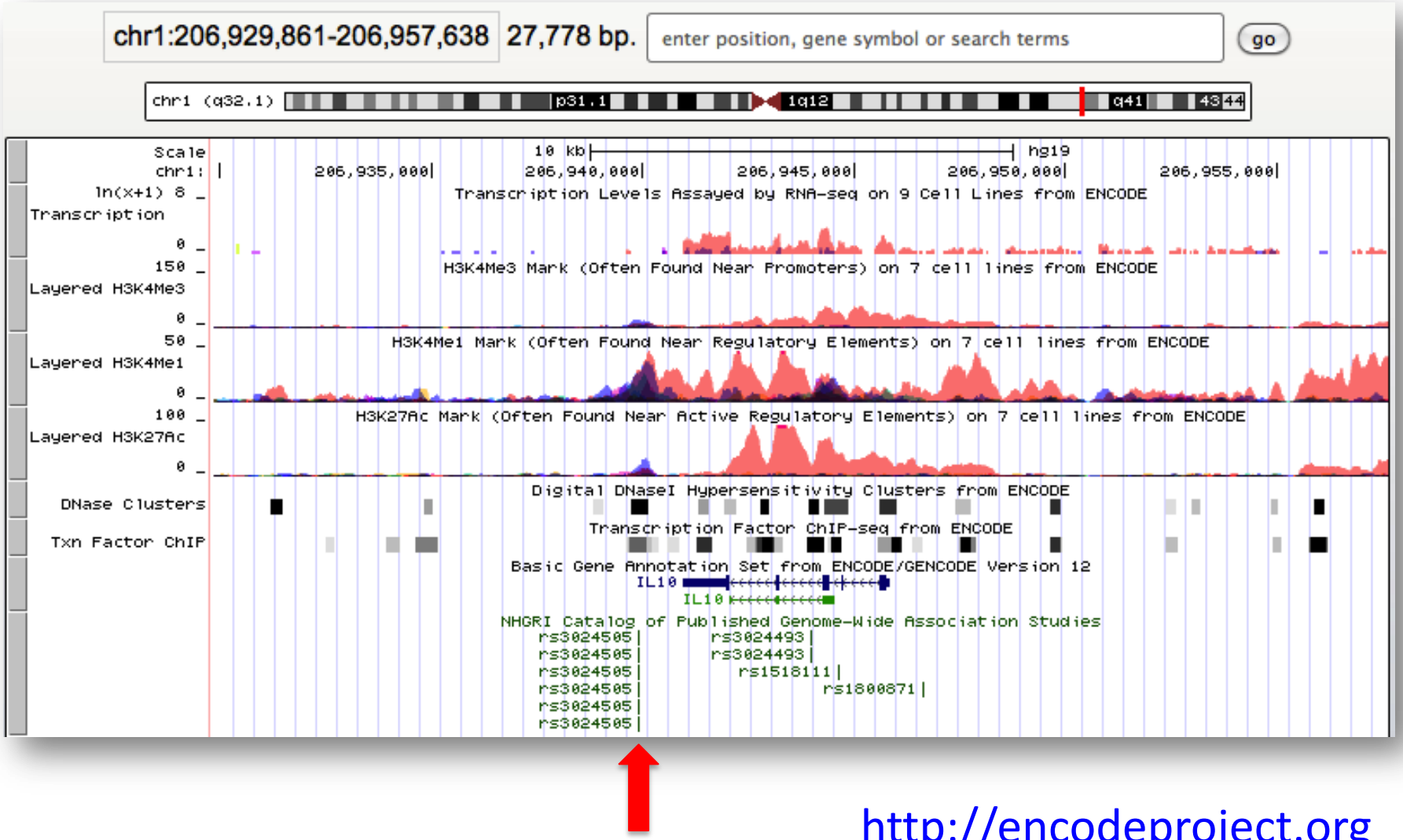
http://encodeproject.org

How Can The Display Be Configured?



http://encodeproject.org

ENCODE Browser, Locus of Interest

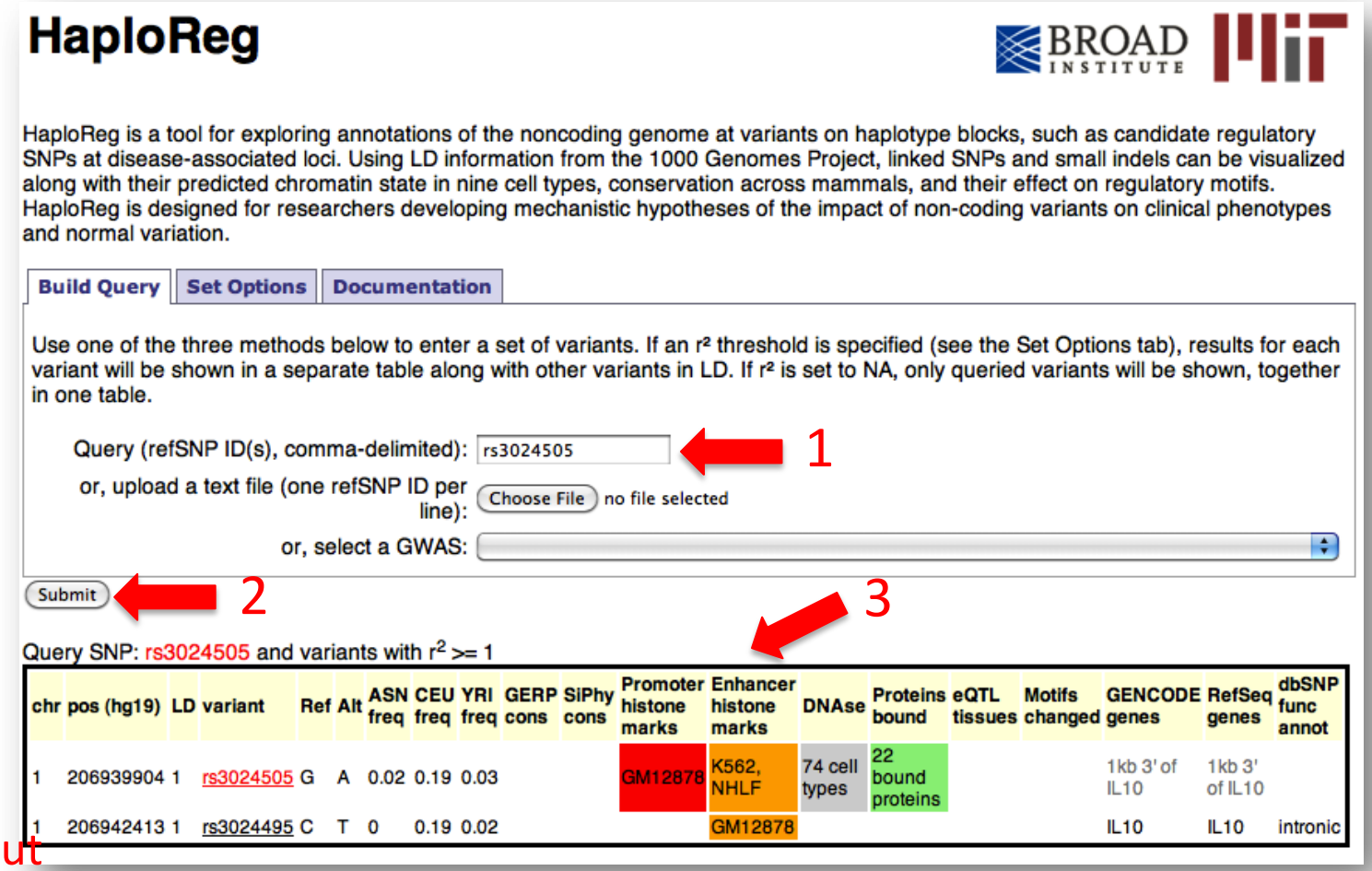


http://encodeproject.org

ENCODE Data From HaploReg

http://www.broadinstitute.org/mammals/haploreg/haploreg.php

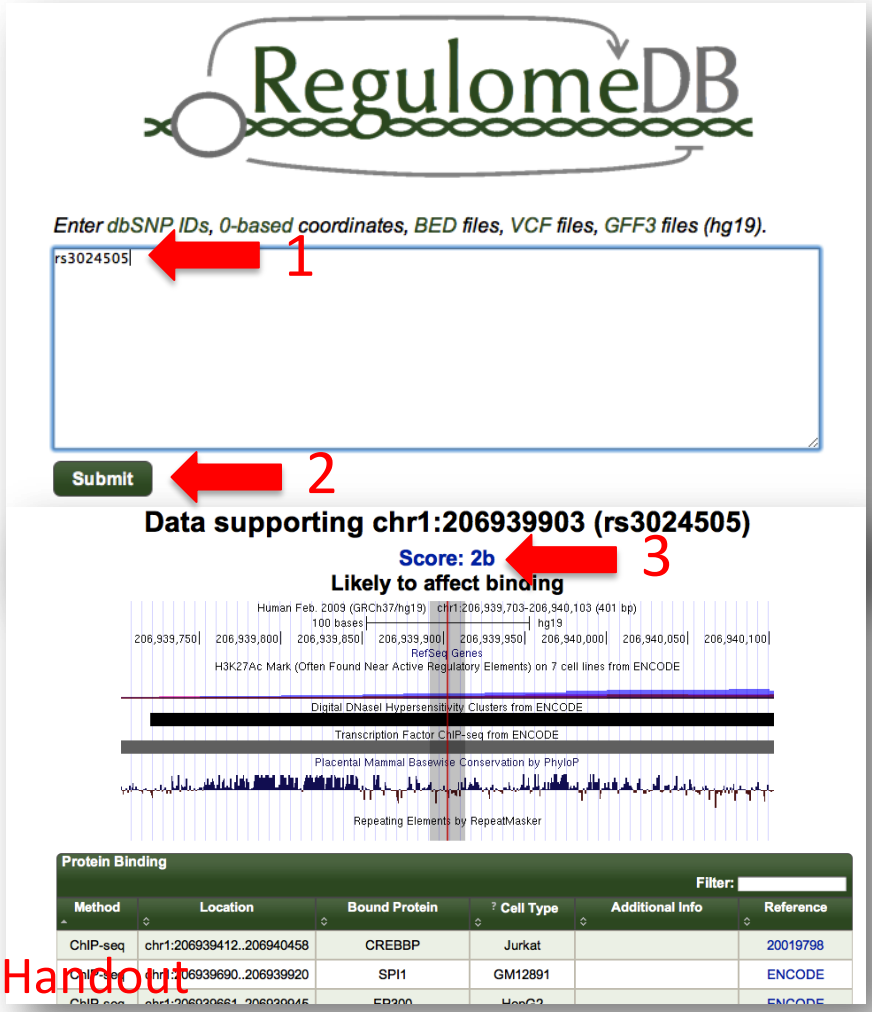
Ward and Kellis, Nucleic Acids Research 40-D930, 2011



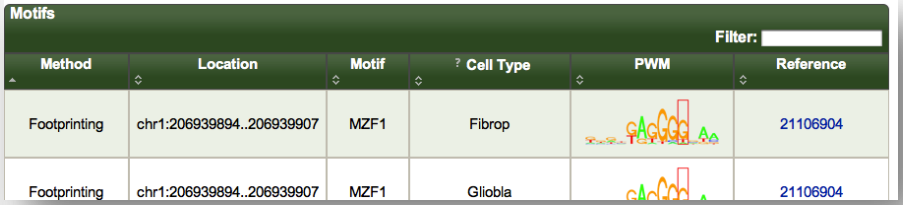
Handout

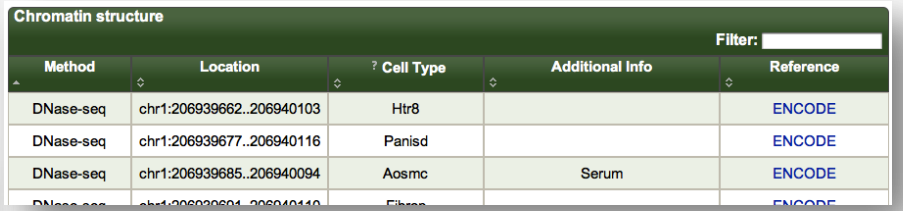
ENCODE Data From RegulomeDB

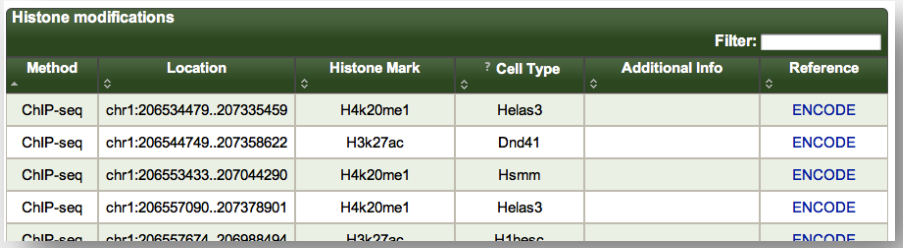
http://regulome.stanford.edu/ ; Boyle…Snyder, Genome Research 22-1790,2012



Handout



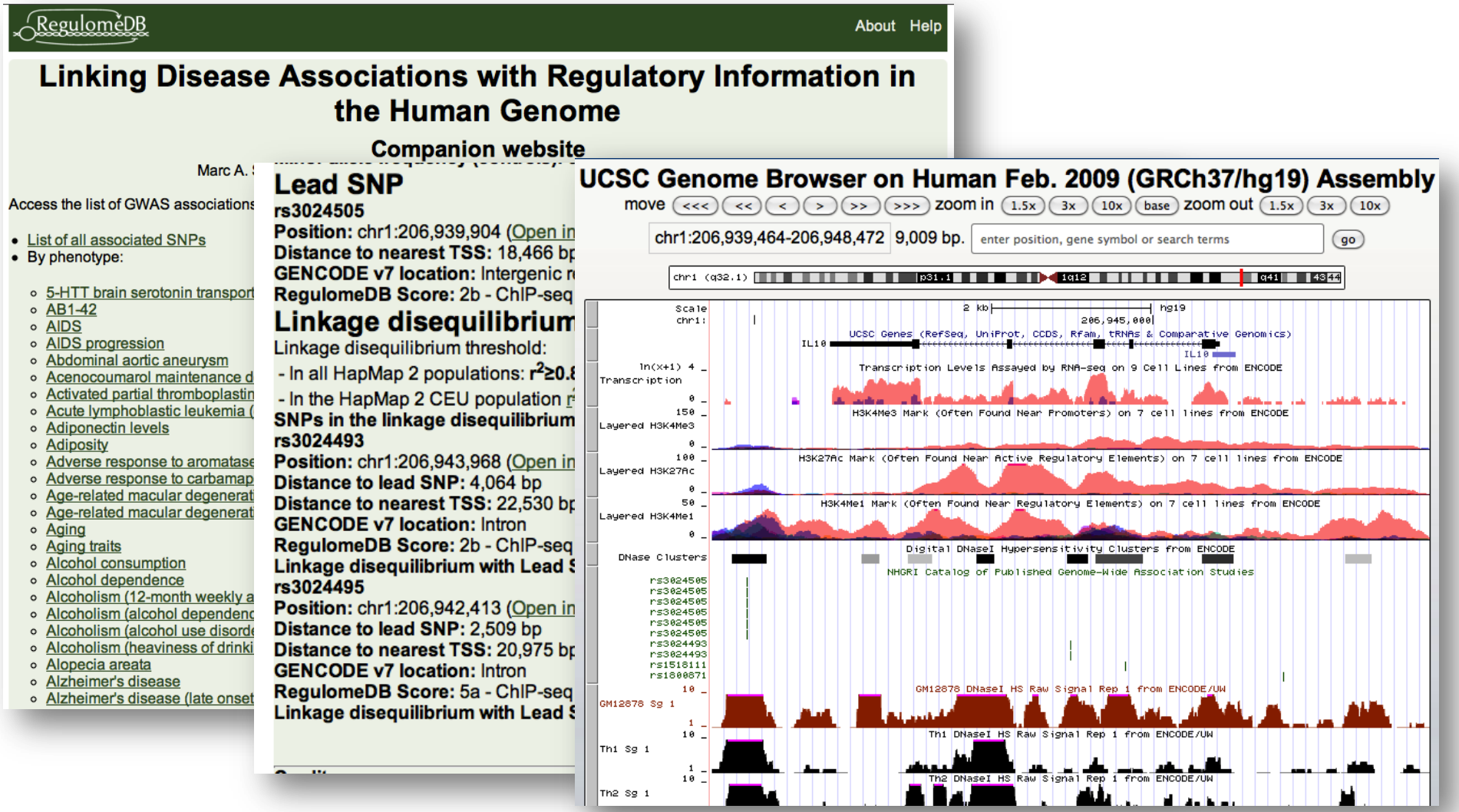




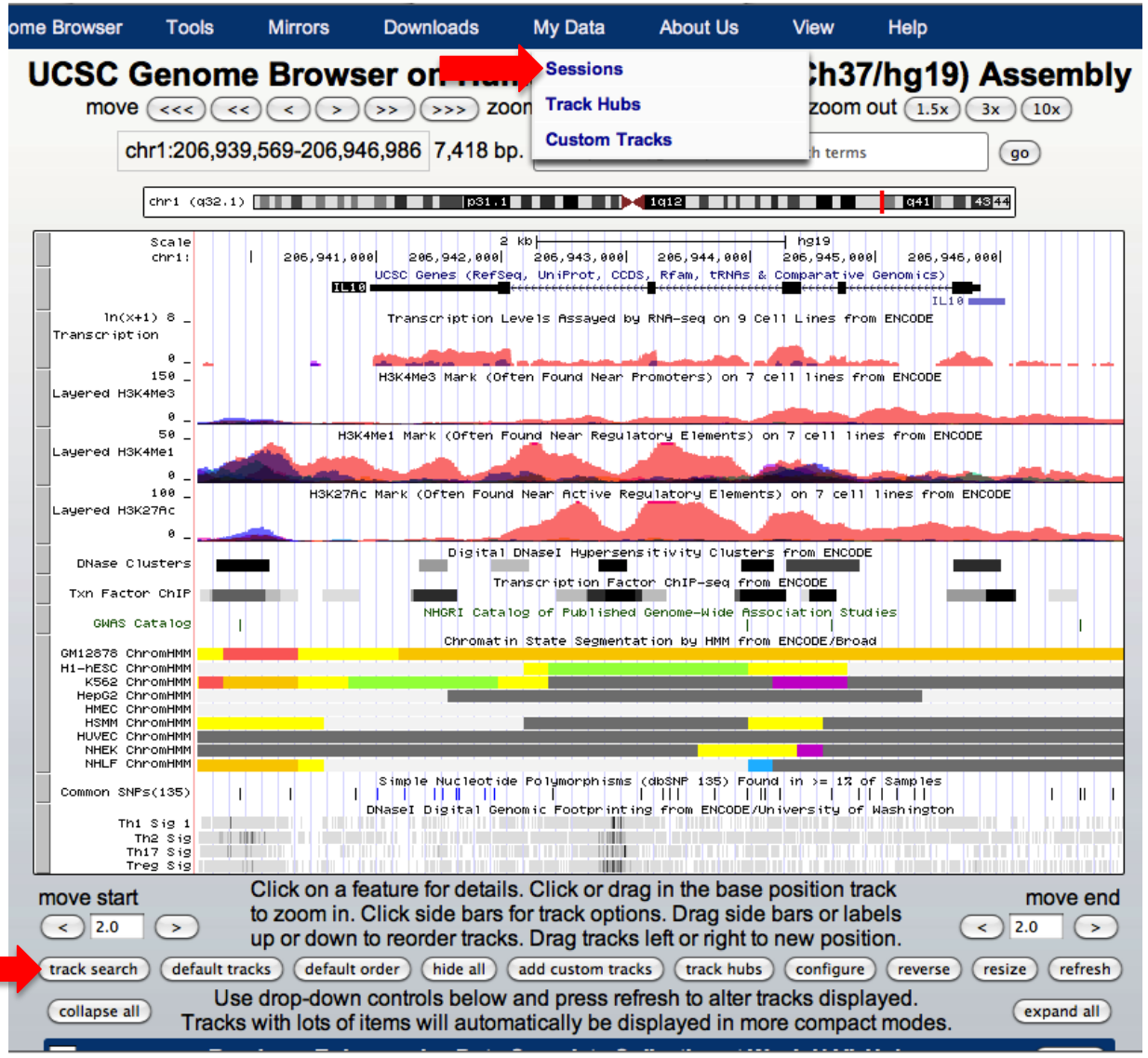
RegulomeDB Disease Database

National Human Genome Research Institute

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

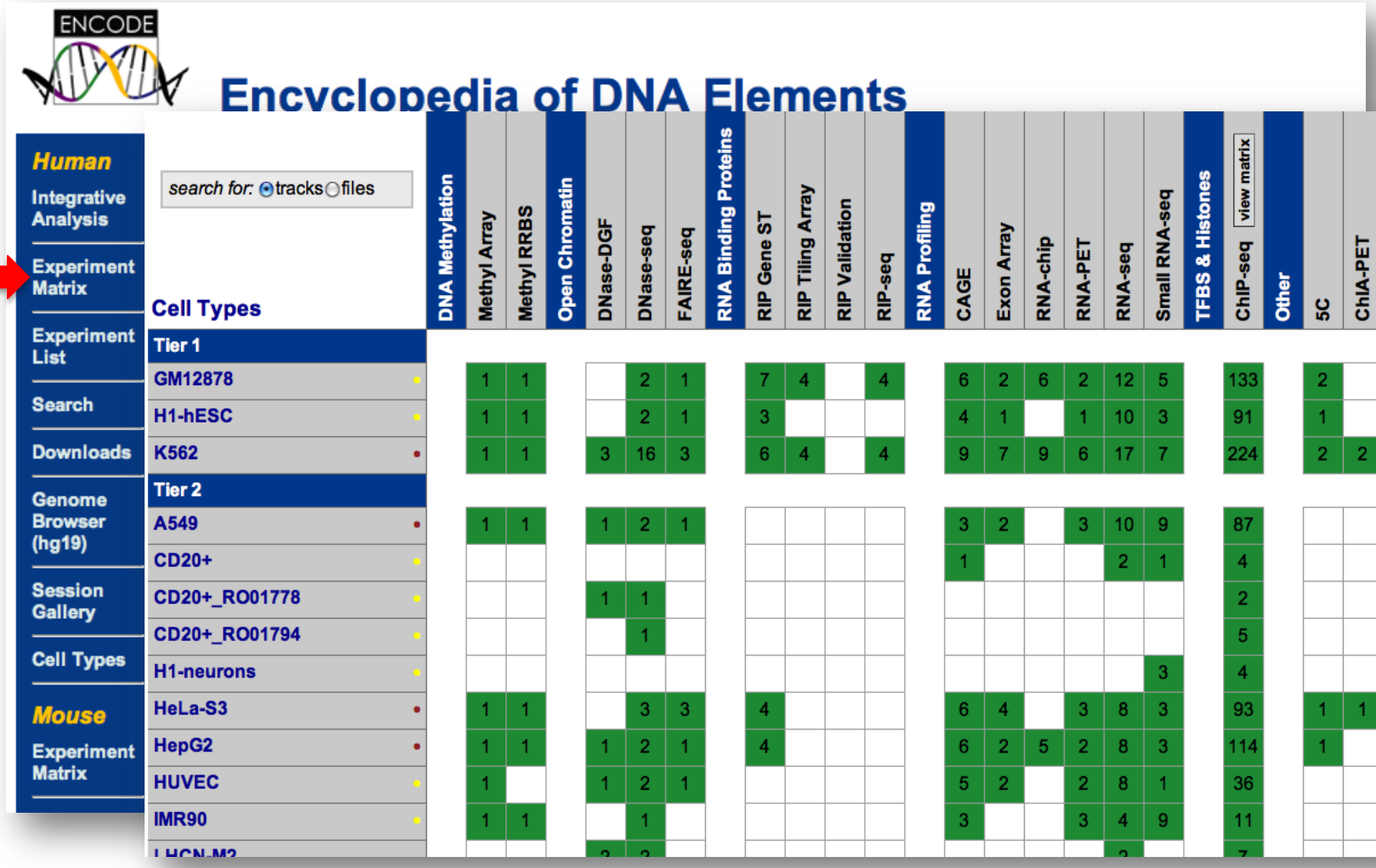


How Can The Display Be Configured?



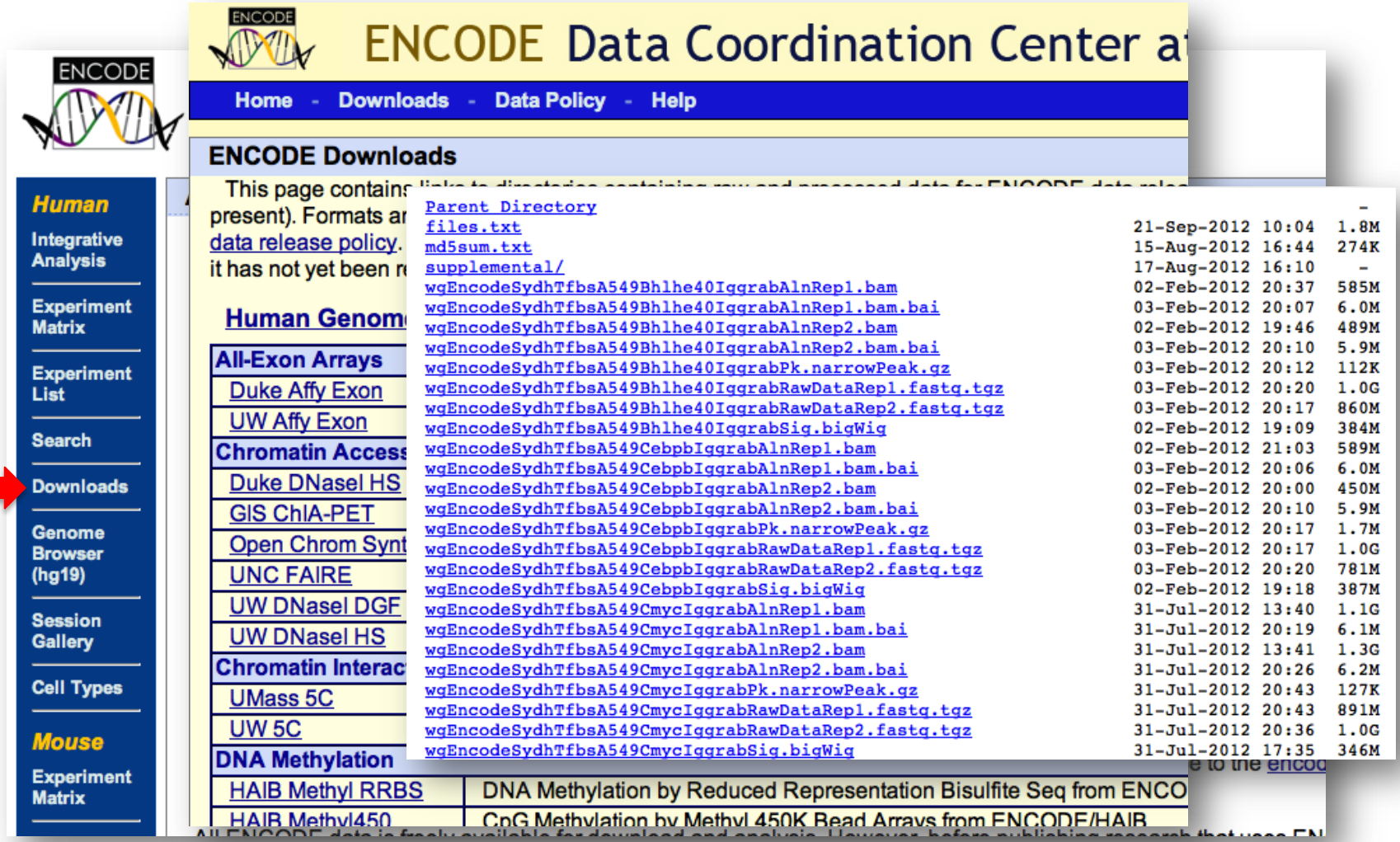
Handout

ENCODE Experiment Matrix



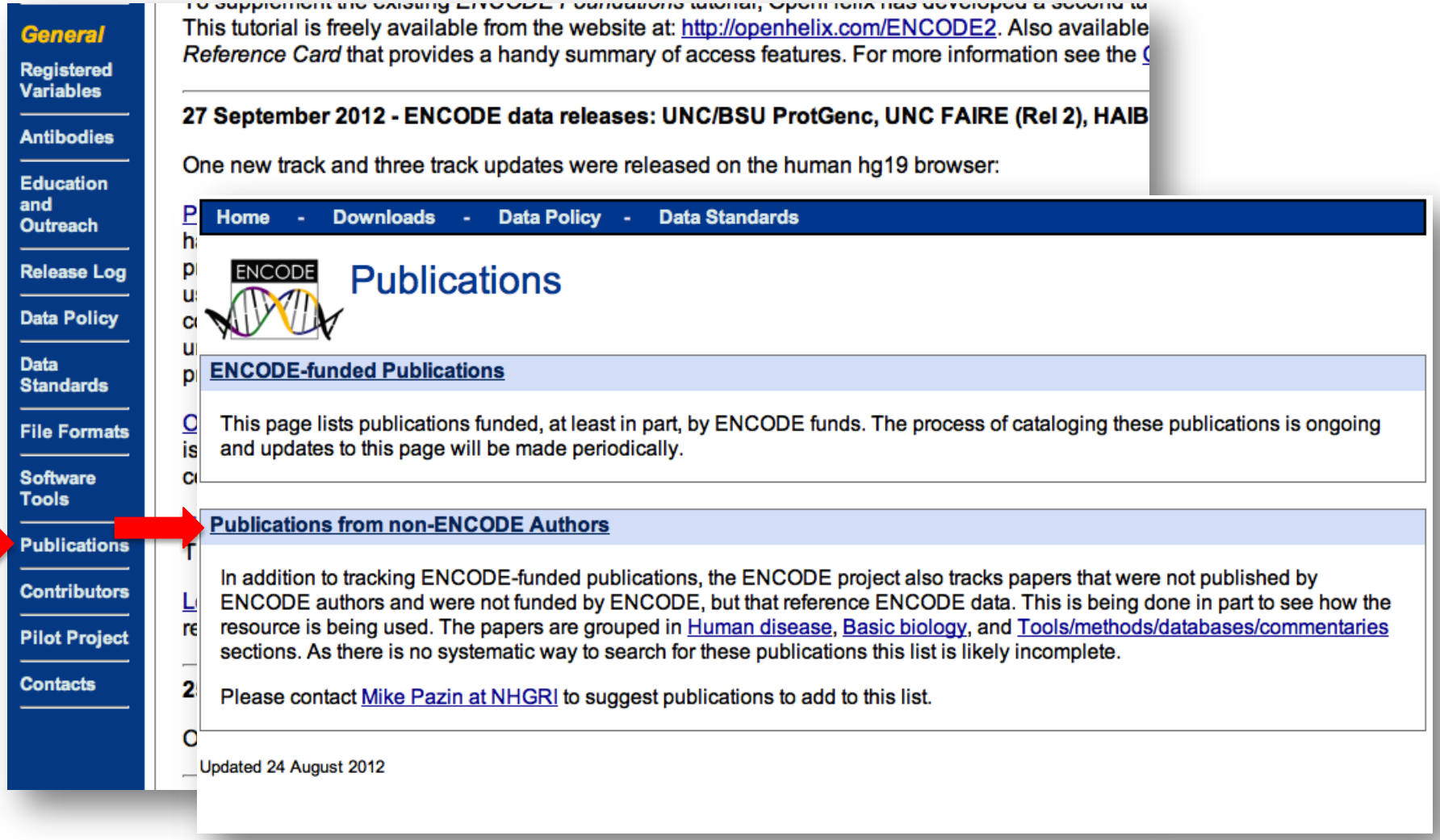
http://encodeproject.org

How Can ENCODE Data Be Downloaded?



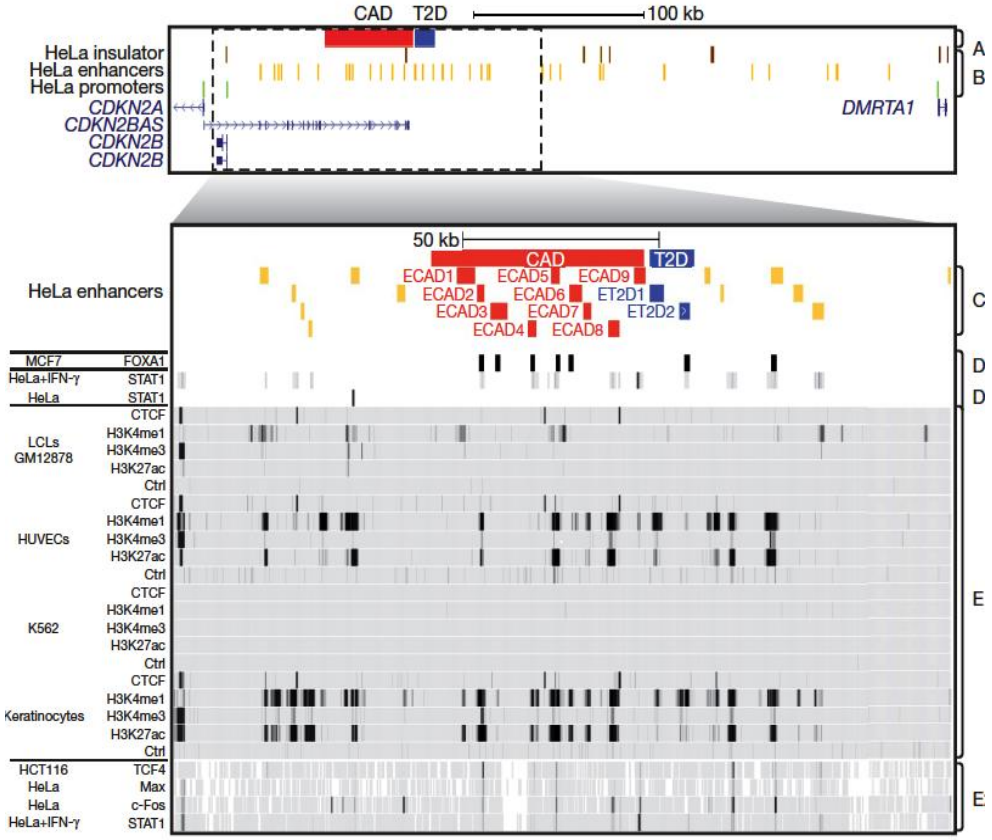
http://encodeproject.org

Publications



http://encodeproject.org

Over 170 Publications Using ENCODE Data



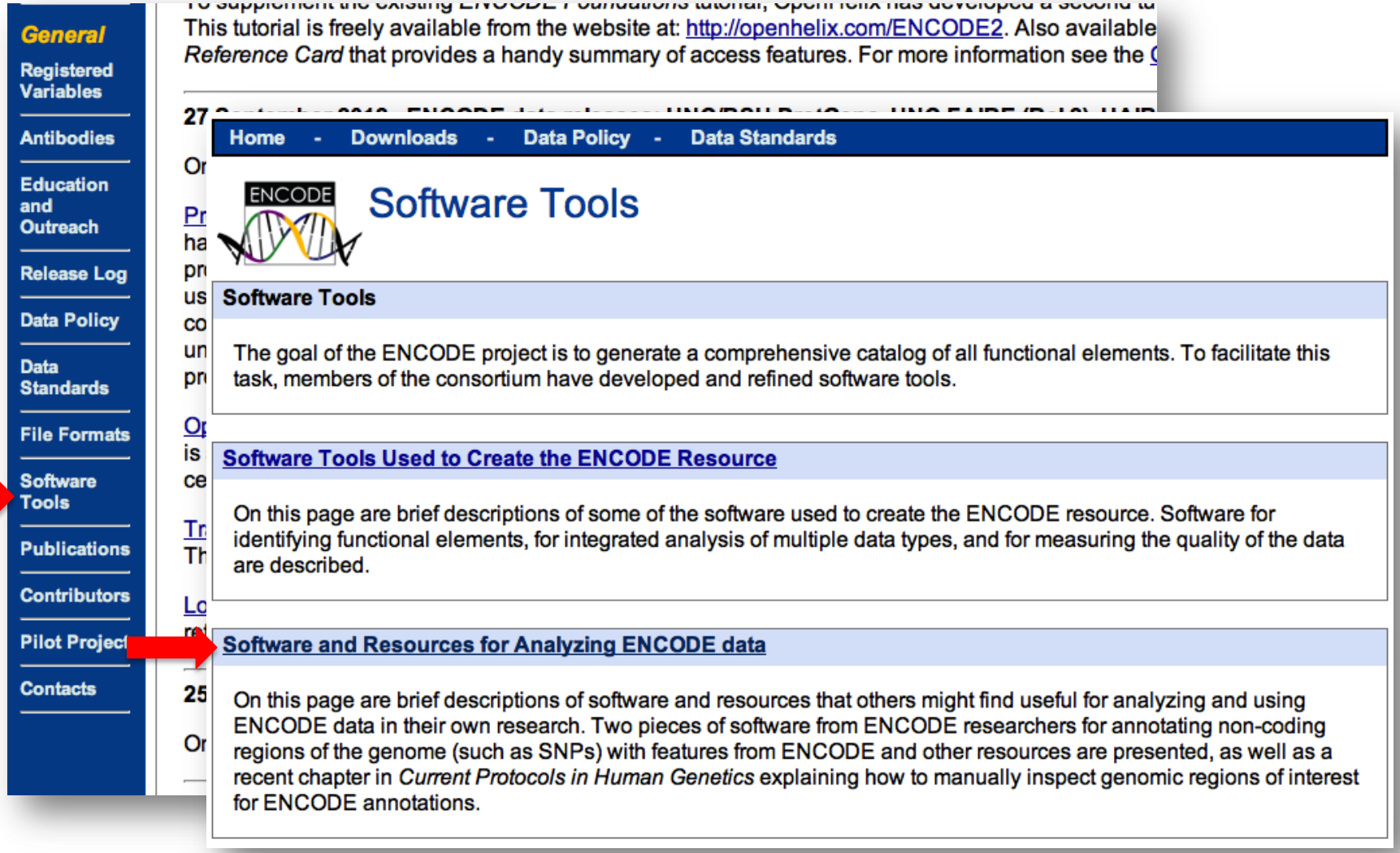
Frazer lab, Nature 470-264,2011

Table 1 PBC risk loci at genome-wide significance

Chr.	SNP[a]	RA[b]	RAF[c]	P[d]	OR	:e)	Nearby gene(s)[f]	Functional annotation[g]
1p31	rs72678531	G	0.17	2.47×10^{-38}	1.61 (	(31,842)	*Il12RB2*	OC
1q31	rs2488393	A	0.21	4.29×10^{-12}	1.28 (	:1 (438,818)	*DENND1B*	
2q32	rs3024921	A	0.06	2.59×10^{-18}	1.62 (	.7 (0)	*STAT1, STAT4*	
2q32	Second signal rs7574865	A	0.22	1.38×10^{-13}	1.31 (	'9 (29,292)	*STAT1, STAT4*	
3q13	rs2293370	G	0.8	6.84×10^{-16}	1.39 (	)8 (136,058)	*TMEM39A, POGLUT1, TIMMDC1, CD80*	NS
3q25	rs2366643	A	0.57	3.92×10^{-22}	1.35 (	'0 (16,805)	*IL12A*	OC
3q25	Second signal rs62270414	G	0.15	5.74×10^{-17}	1.41 (	'6 (52,623)	*IL12A*	OC
3q25	Third signal rs668998	G	0.43	4.73×10^{-9}	1.26 (	5 (5,550)	*IL12A*	OC
3q25	Fourth signal rs80014155	A	0.004	2.64×10^{-11}	3.44 (	7 (68,660)	*IL12A*	
4q24	rs7665090	G	0.52	8.48×10^{-14}	1.26 (	1 (0)	*MANBA, NFKB1*	NS, eQTL
5p13	rs6871748	A	0.72	2.26×10^{-13}	1.3 (1	(35,833)	*IL7R, CAPSL, SPEF2, UGT3A1*	NS
7q32	rs35188261	A	0.17	6.52×10^{-22}	1.52 (	.0 (126,258)	*IRF5, TNP03*	OC
7q32	Second signal rs3807307	G	0.47	4.12×10^{-9}	1.22 (	.6 (6,713)	*IRF5, TNP03*	OC, eQTL
11q23	rs80065107	A	0.79	7.20×10^{-16}	1.39 (	2 (133,223)	*DDX6*	OC
12p13	rs1800693	G	0.4	1.18×10^{-14}	1.27 (	,802)	*TNFRSF1A, LTBR, SCNN1A*	OC
12p13	Second signal rs11064157	A	0.25	1.69×10^{-9}	1.23 (		*TNFRSF1A, LTBR, SCNN1A*	OC
12q24	**rs11065979**	A	0.44	2.87×10^{-9}	1.2 (1	7 (726,106)	*ATXN2, BRAP, SH2B3*	NS
14q24	rs911263	A	0.71	9.95×10^{-11}	1.26 (	(0)	*RAD51B*	
16p13	rs1646019	G	0.71	6.72×10^{-15}	1.31 (	(18,452)	*SOCS1, CLEC16A, PRM1, PRM2*	OC, eQTL
16p13	Second signal rs12708715	C	0.68	2.19×10^{-13}	1.29 (	(118,128)	*SOCS1,CLEC16A, PRM1, PRM2*	OC, eQTL
16p13	Third signal rs80073729	A	0.004	2.69×10^{-8}	2.96 (	(0)	*SOCS1, CLEC16A, PRM1, PRM2*	OC
16q24	rs11117433	G	0.77	1.41×10^{-9}	1.26 (	(0)	*IRF8*	OC
17q12	rs8067378	G	0.52	6.05×10^{-14}	1.26 (	(177,700)	*ORMDL3, ZPBP2, GSDMB, IKZF3*	NS, OC, eQTL
17q21	**rs17564829**	G	0.24	2.15×10^{-9}	1.25 (	(1,164,644)	*CRHR1, MAPT*	NS, OC, eQTL
19p12	**rs34536443**	G	0.95	1.23×10^{-12}	1.91 (	(0)	*TYK2*	NS
22q13	rs2267407	A	0.23	1.29×10^{-13}	1.29 (	(9,600)	*SYNGR1, PDGFB, RPL3*	OC, eQTL

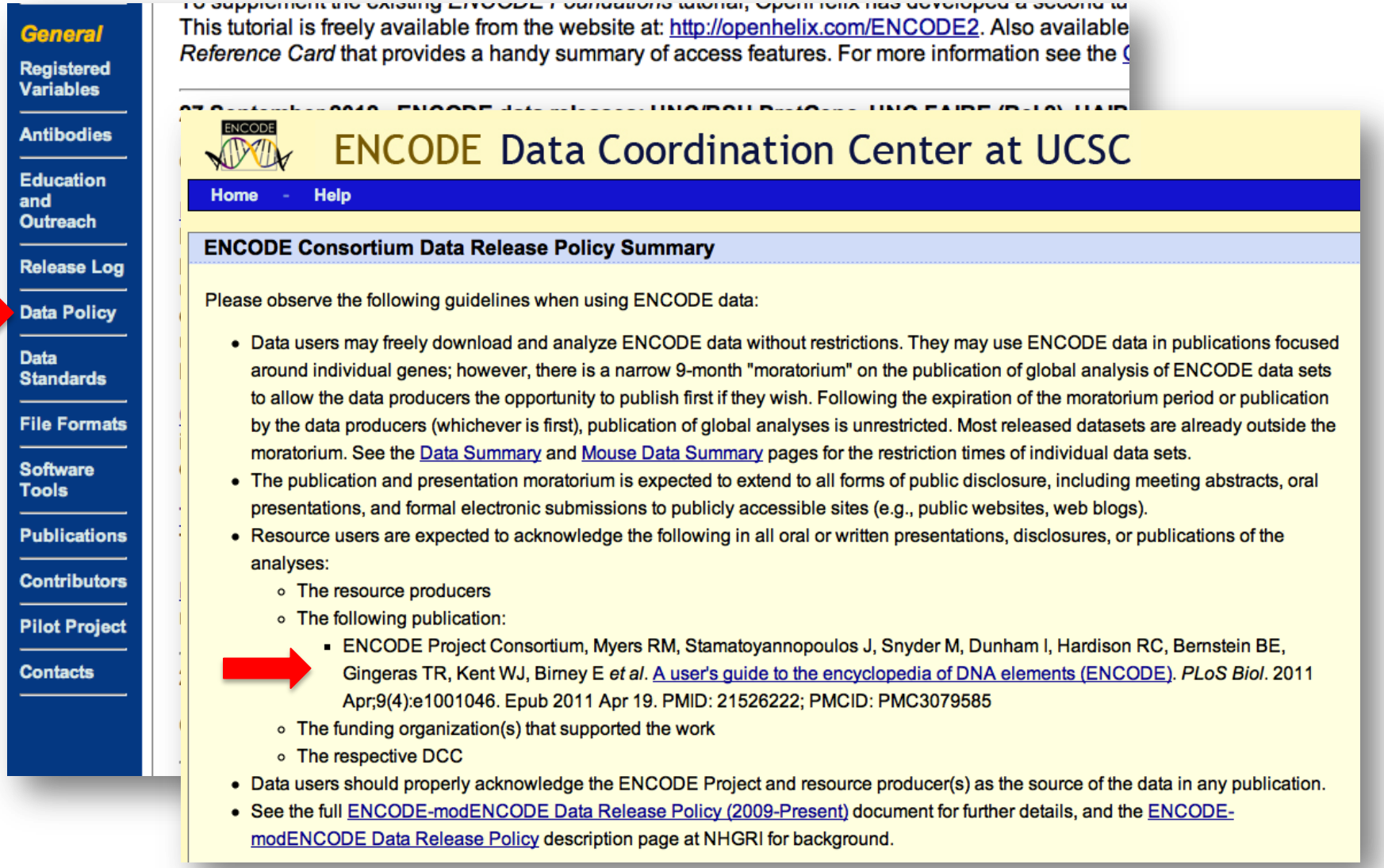
Anderson lab, Nature Genetics 44-1137,2012

ENCODE Software Tools



http://encodeproject.org

ENCODE Citation



General
Registered Variables
Antibodies
Education and Outreach
Release Log
Data Policy
Data Standards
File Formats
Software Tools
Publications
Contributors
Pilot Project
Contacts

This tutorial is freely available from the website at: http://openhelix.com/ENCODE2. Also available
Reference Card that provides a handy summary of access features. For more information see the

ENCODE Data Coordination Center at UCSC

Home - Help

ENCODE Consortium Data Release Policy Summary

Please observe the following guidelines when using ENCODE data:

- Data users may freely download and analyze ENCODE data without restrictions. They may use ENCODE data in publications focused around individual genes; however, there is a narrow 9-month "moratorium" on the publication of global analysis of ENCODE data sets to allow the data producers the opportunity to publish first if they wish. Following the expiration of the moratorium period or publication by the data producers (whichever is first), publication of global analyses is unrestricted. Most released datasets are already outside the moratorium. See the Data Summary and Mouse Data Summary pages for the restriction times of individual data sets.
- The publication and presentation moratorium is expected to extend to all forms of public disclosure, including meeting abstracts, oral presentations, and formal electronic submissions to publicly accessible sites (e.g., public websites, web blogs).
- Resource users are expected to acknowledge the following in all oral or written presentations, disclosures, or publications of the analyses:
  - The resource producers
  - The following publication:
    - ENCODE Project Consortium, Myers RM, Stamatoyannopoulos J, Snyder M, Dunham I, Hardison RC, Bernstein BE, Gingeras TR, Kent WJ, Birney E *et al*. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol*. 2011 Apr;9(4):e1001046. Epub 2011 Apr 19. PMID: 21526222; PMCID: PMC3079585
  - The funding organization(s) that supported the work
  - The respective DCC
- Data users should properly acknowledge the ENCODE Project and resource producer(s) as the source of the data in any publication.
- See the full ENCODE-modENCODE Data Release Policy (2009-Present) document for further details, and the ENCODE-modENCODE Data Release Policy description page at NHGRI for background.

ENCODE Consortium

Brad Bernstein (Eric Lander, Manolis Kellis/Luke Ward, Tony Kouzarides)

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

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

Jim Kent (David Haussler, Kate Rosenbloom)

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

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

Rick Myers (Barbara Wold)

Scott Tenenbaum (Luiz Penalva)

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

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

Zhiping Weng (Nathan Trinklein, Rick Myers)

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

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

.. and many senior scientists, postdocs, students, technicians, computer scientists, statisticians and administrators in these groups

Questions? NHGRI booth (# 825) noon- 1 PM; pazinm@mail.nih.gov

National
Human Genome
Research Institute

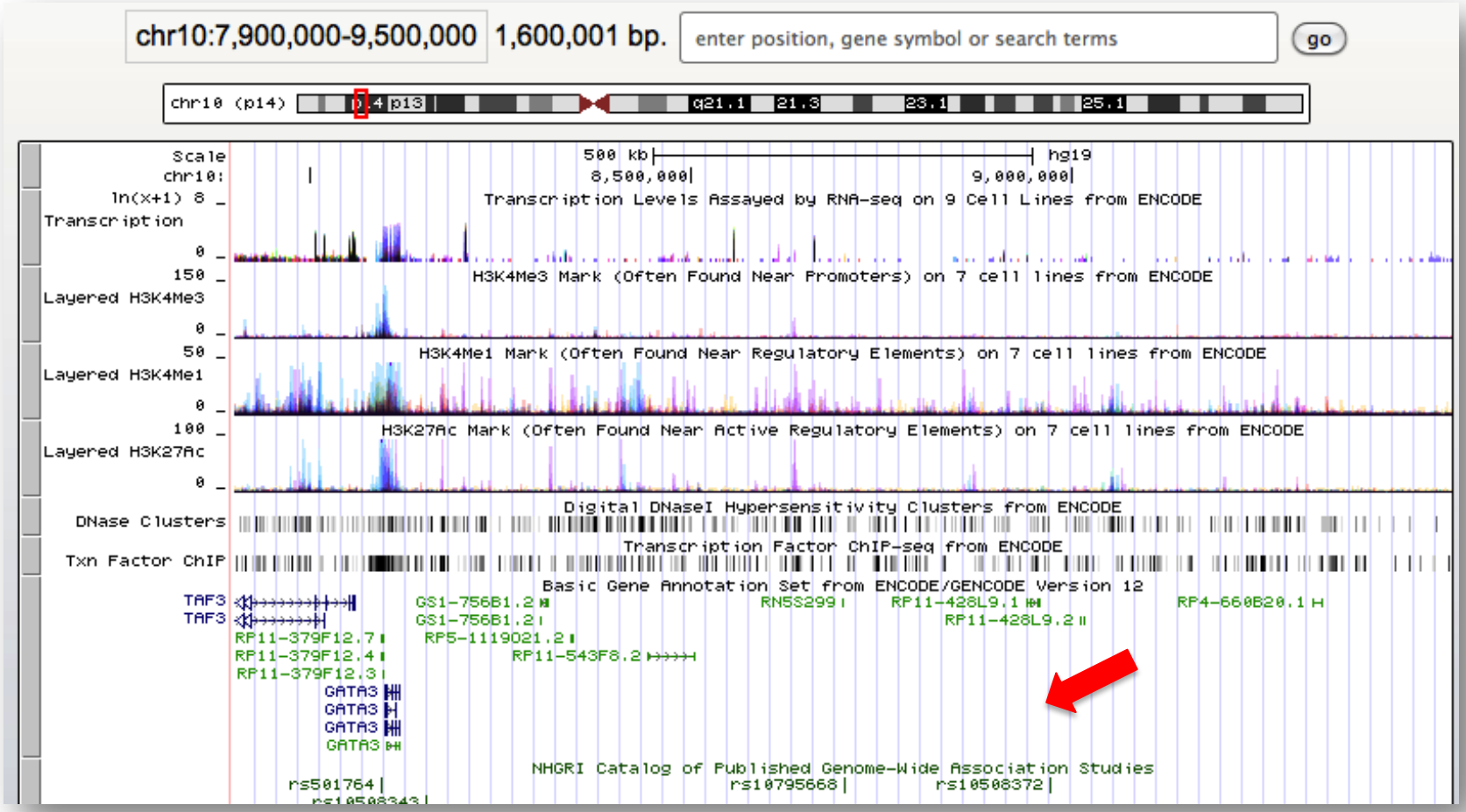
What Do The Data Mean?

- Some standard interpretations-
  - RNA
  - Histone modifications
  - DNase
  - Transcription Factor ChIP

How Can ENCODE Data Be Used?

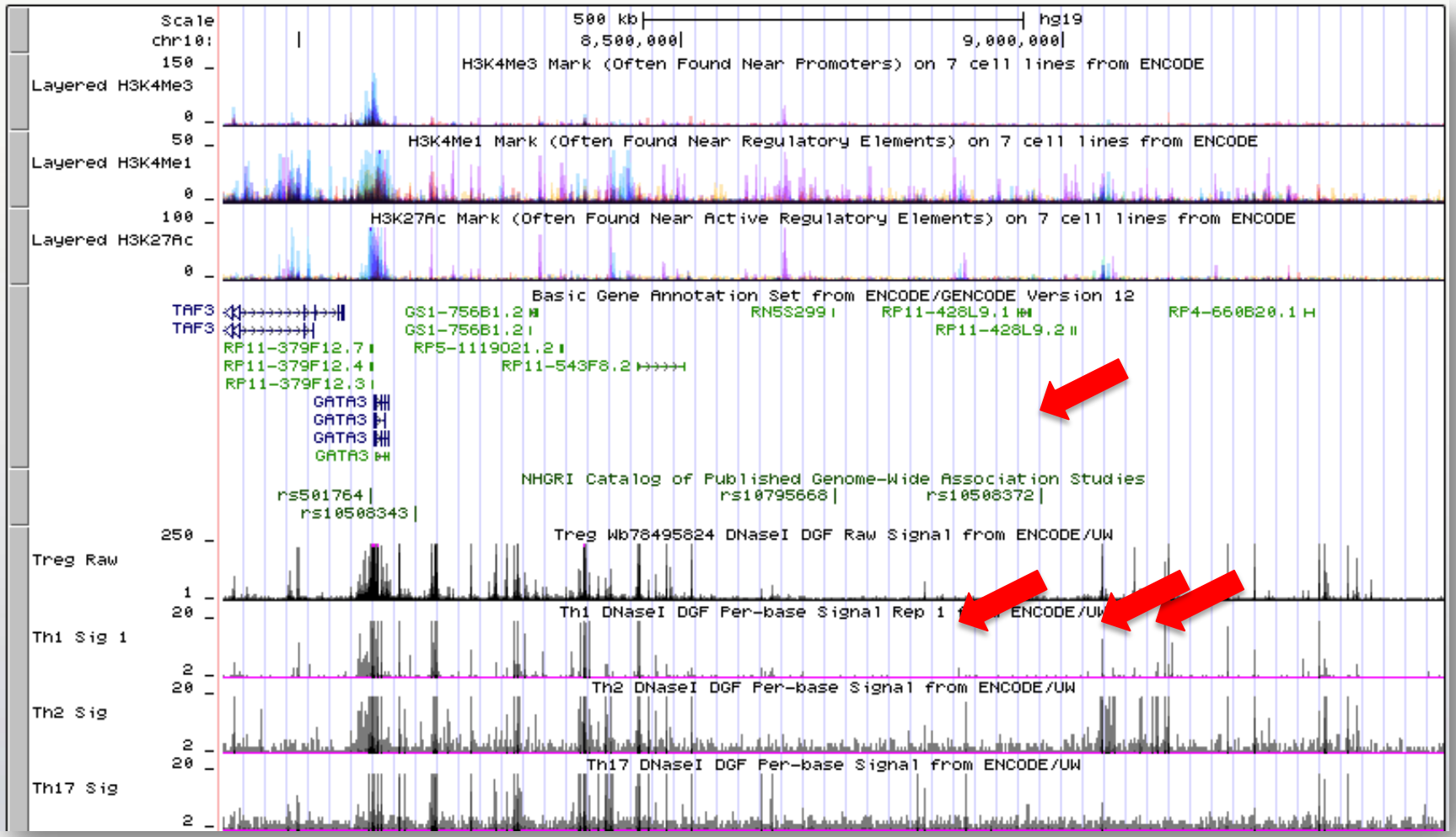
- A standard problem: many genetic findings for human disease map to non-protein coding regions of the human genome
  - What is the functional variant?
  - What is the target gene?
  - What is the target cell type?
  - What is the function of the variant?
- Standard Browser view for loci of interest
- 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 Browser, Locus of Interest



http://encodeproject.org

ENCODE Browser, Cells of Interest



http://encodeproject.org

NIH Roadmap Epigenomic Mapping Consortium

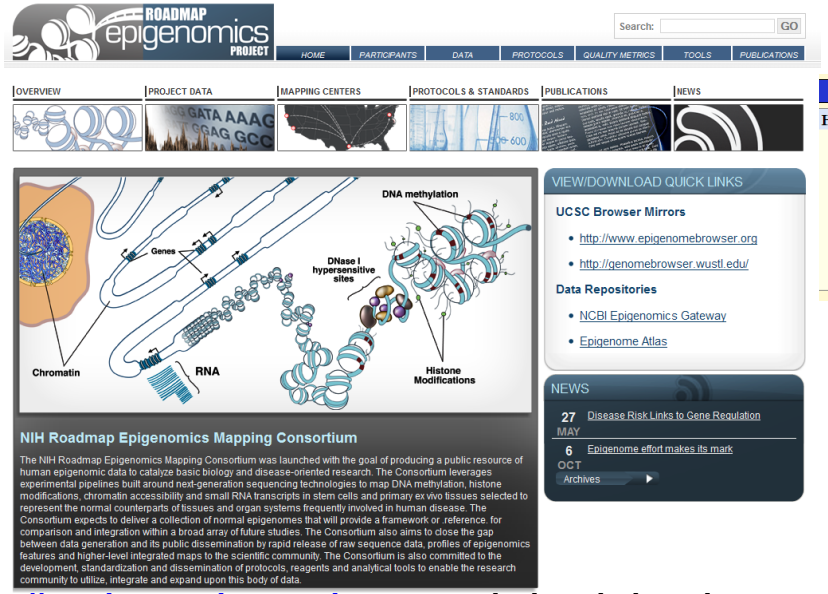
A **public community resource of epigenomic data** in primary human cells and tissues.

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)



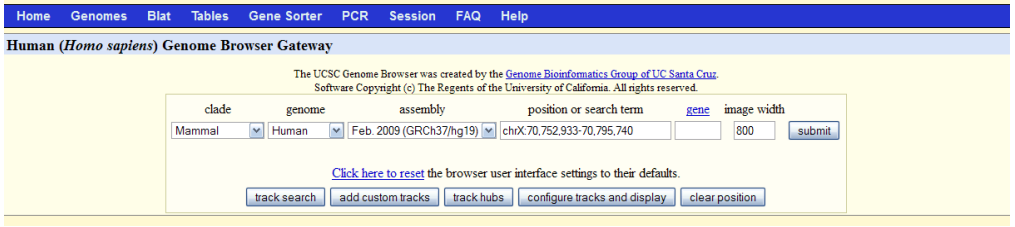
http://roadmapepigenomics.org – Find and view data, protocols, links to other sites associated with the program.

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



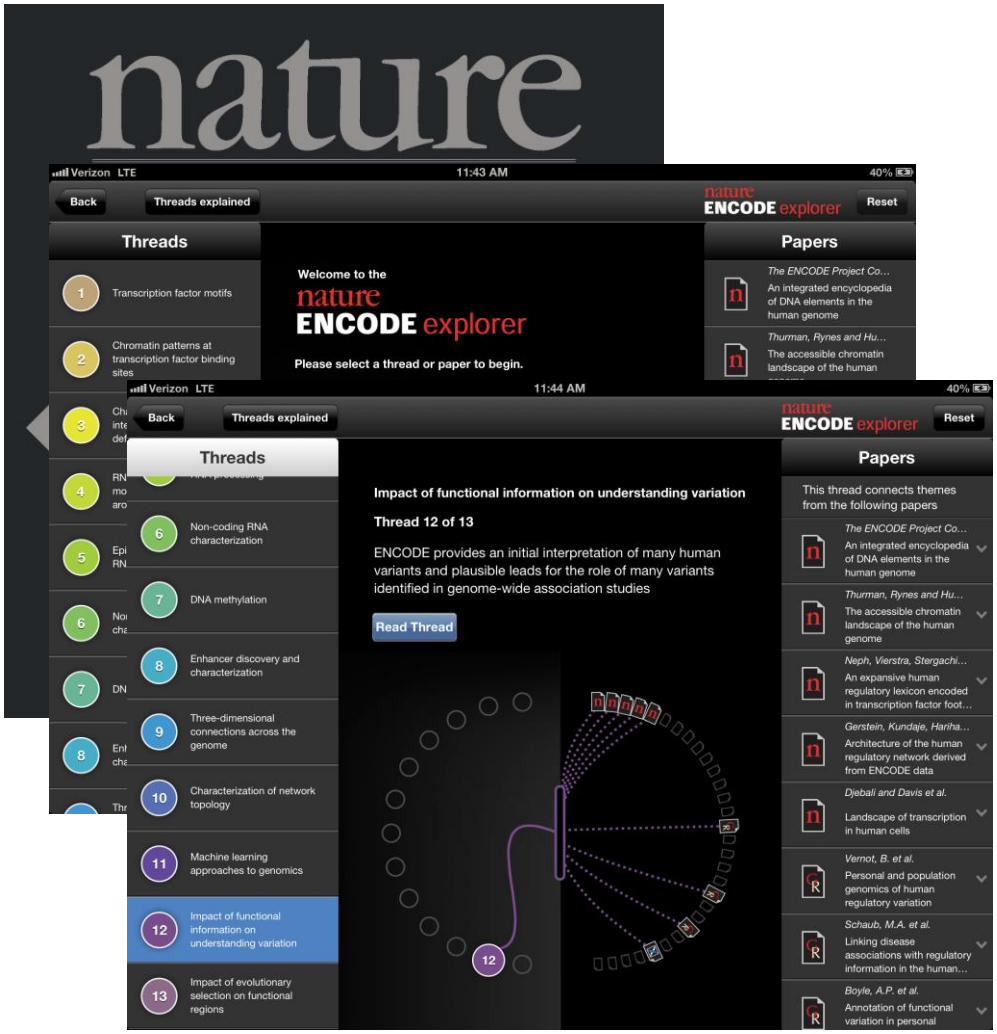
View Roadmap data at http://genome.ucsc.edu via TRACK HUB – click 'track hubs' to load Roadmap data and summary data tracks on UCSC.

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)

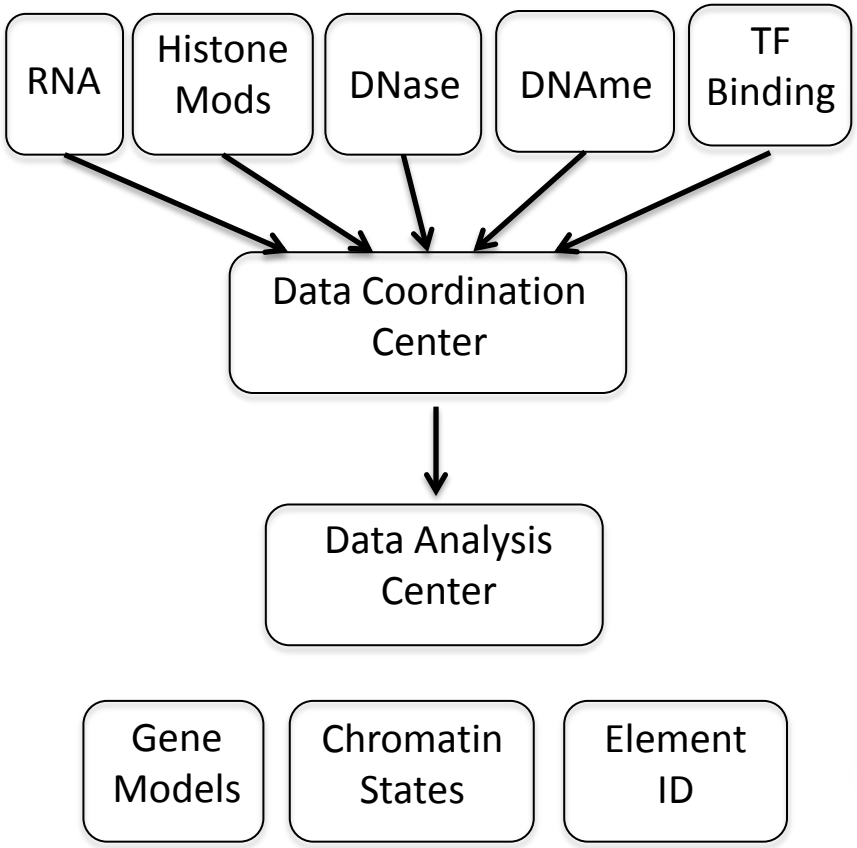
Recent ENCODE Publications

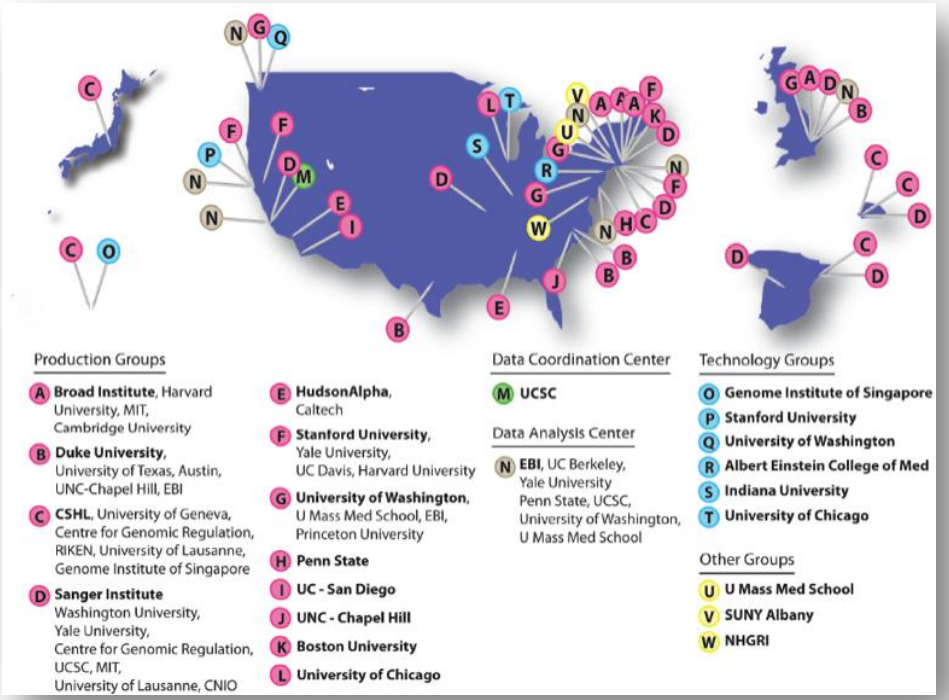


- More than 30 papers in Nature, Genome Research, Genome Biology, Science, Cell
- Publishing innovations
- ENCODE increased our understanding of non-coding DNA, and human disease

From www.nature.com/encode

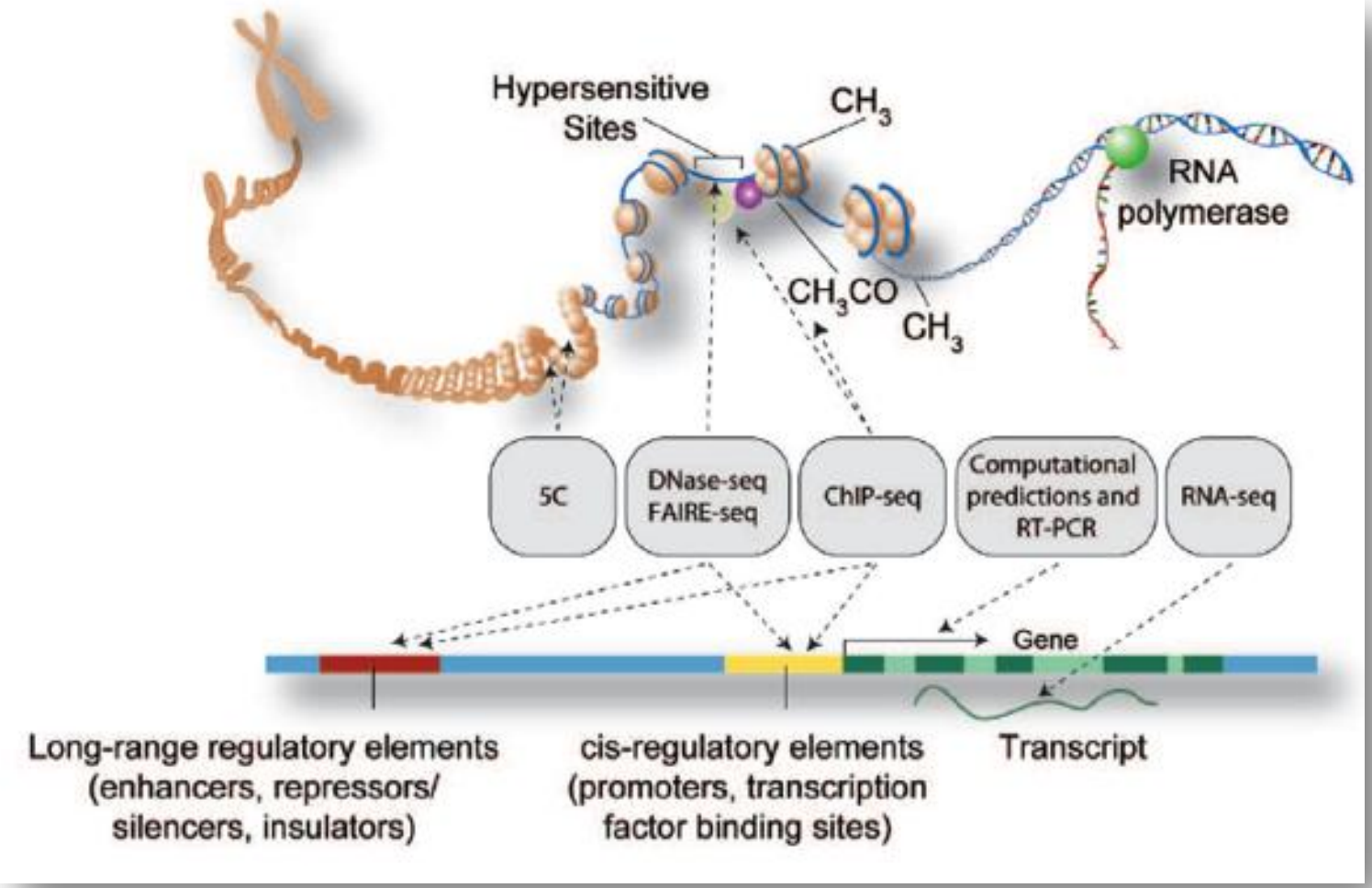
ENCODE Structure





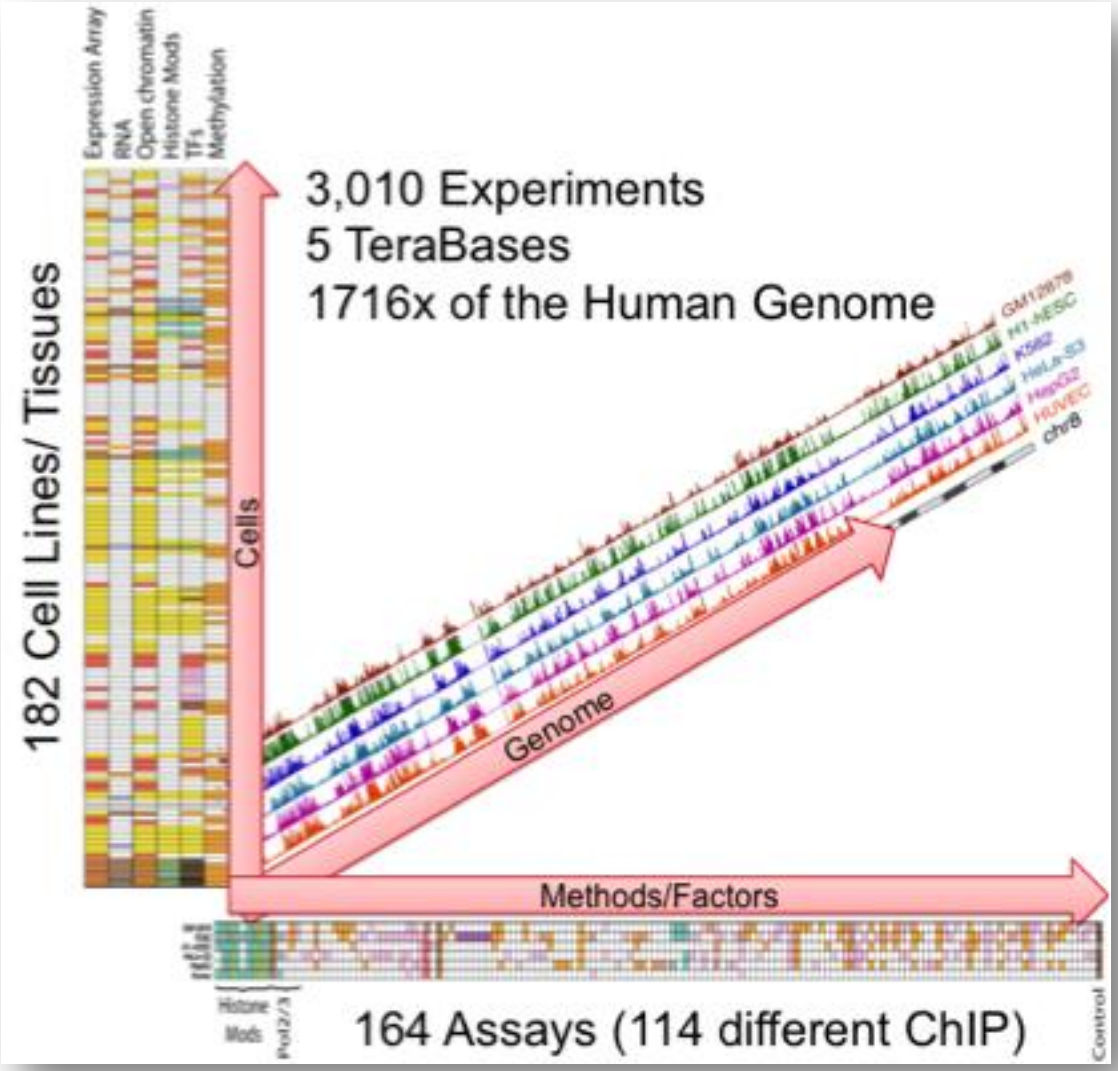
Modified from PLoS Biol 9-e1001046,2011

What Data Are Available?



Modified from PLoS Biol 9-e1001046,2011

ENCODE Dimensions



From Ewan Birney

Raw Genomic Coverage of Elements

Biochemical Mark	Genomic Coverage
Any ENCODE mark	80 %
Any ENCODE RNA	62 %
Any Histone modification	56 %
DHS or TF ChIP	19.4 %
Open chromatin	15.2 %
TF ChIP	8.1 %
DHS footprint	5.7 %
Purifying selection	~3-8 %
Exons (GENCODE)	2.94 %
Protein-coding regions (GENCODE)	1.22 %

Nature 489-57,2012