



National Human Genome
Research Institute

Using ENCODE Data to Interpret Disease-associated Genetic Variation

Dan Gilchrist

National Human Genome Research Institute, NIH

Society of Toxicology Satellite Meeting

March 17, 2016





ENCODE 2016: Research Applications and Users Meeting

June 8-10, 2016 at Stanford University

- Hands-on workshops on how to navigate, analyze, and integrate ENCODE and mouseENCODE data into your research
- Leading-edge research applications from distinguished invited speakers from across the research community
- Tutorials on newly-available informatics pipelines that greatly facilitate working with ENCODE data
- Short talks selected from abstracts
- For details and registration see:

<https://www.encode2016.org>

Overview

- The ENCODE Resource
- Use Cases of ENCODE to illuminate the role of genetic variation in human disease
- Accessing ENCODE materials

Overview

- The ENCODE Resource
- Use Cases of ENCODE to illuminate the role of genetic variation in human disease
- Accessing ENCODE materials

ENCODE:

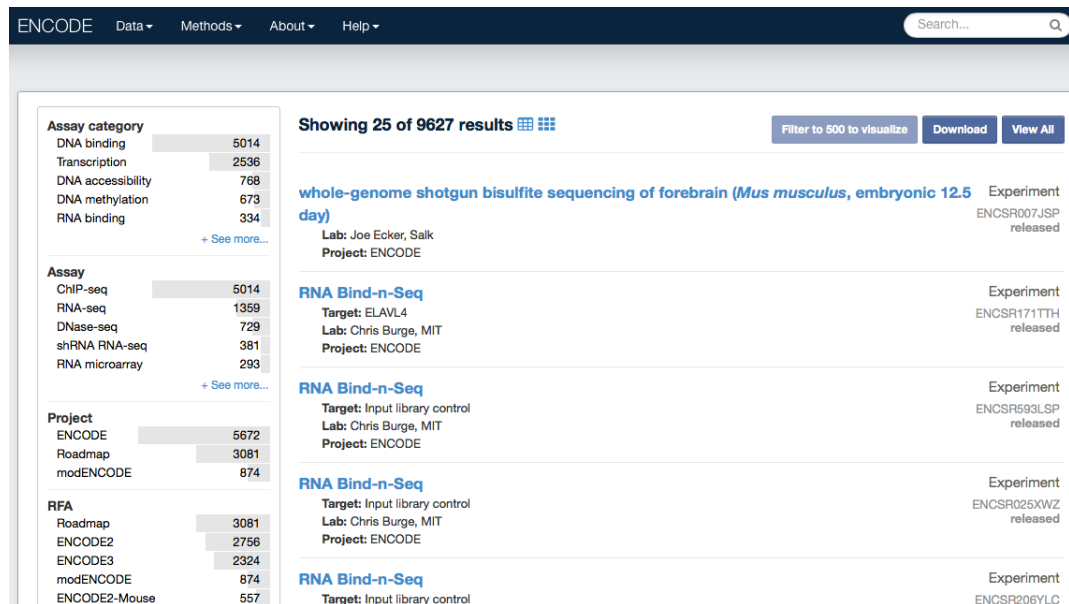
Encyclopedia Of DNA Elements

- Identify all candidate functional elements in the genome
 - Promoters, enhancers, transcribed regions
- Make resource freely available to community for use in studies of:
 - genetic basis of disease
 - gene regulation
 - whatever you want

Research Parasites Welcome!

ENCODE: Encyclopedia Of DNA Elements

- Thousands of data sets
- Hundreds of different bio-samples
- Data sets related to exposure



ENCODE:

Encyclopedia Of DNA Elements

ENCODE 'Phase 4' RFAs –

Applications Due March 21st

- Expanding the Encyclopedia of DNA Elements in Human and Mouse
- Characterizing ENCODE Functional Elements
- Computational Analysis of ENCODE Data
- Data Coordination Center
- Data Analysis Center

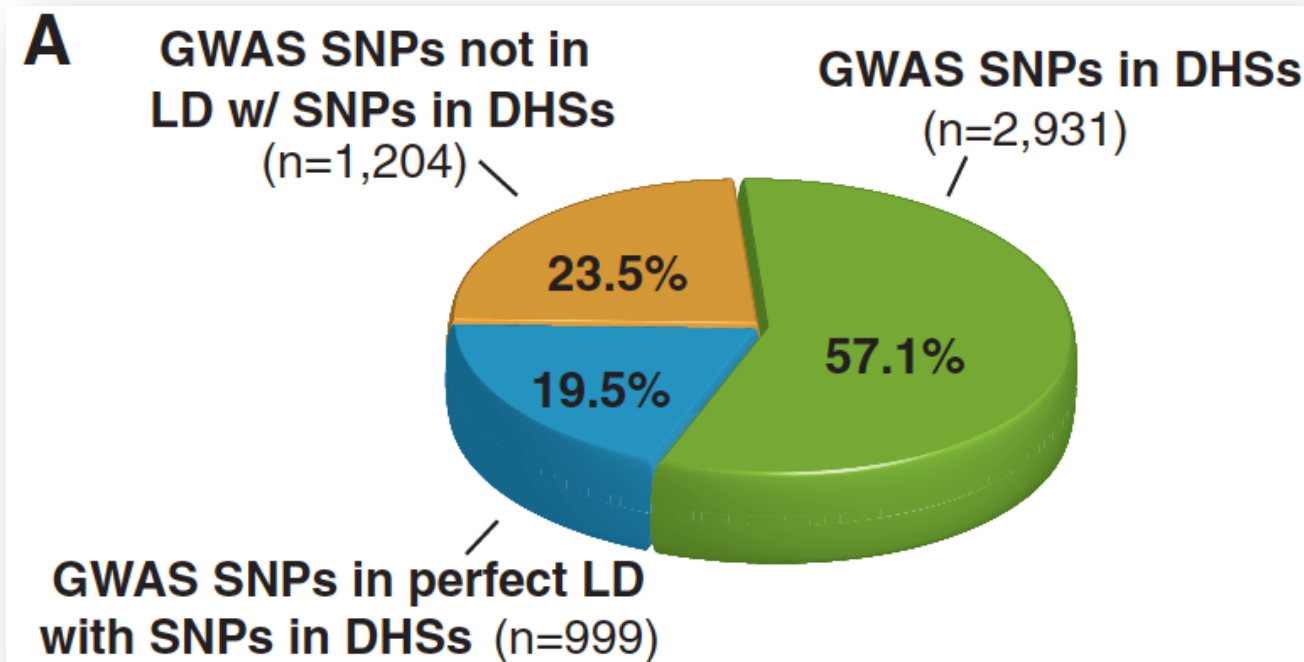
Understanding Non-coding DNA Is Central To NHGRI Goals

Non-coding DNA is important for disease and gene regulation

- Vast majority of common disease associations and heritability lie outside of protein-coding regions
- Non-coding DNA variants are known to cause human diseases and alter human traits (FXS, ALS)

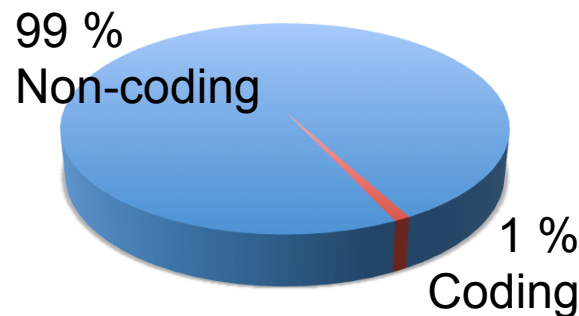
Functional information is needed to interpret the role of genetic variation in human disease, and to apply genomics in the clinic.

Many GWAS Associations Lie In Regions Annotated By ENCODE And Epigenomics Data



Reading The Human Genome Is Difficult

- Genetic code very powerful for 1% of the human genome
 - No correspondingly powerful regulatory code
 - Sequence conservation can identify candidate functional elements (but not when or where they act)
- Need unbiased experimental investigation

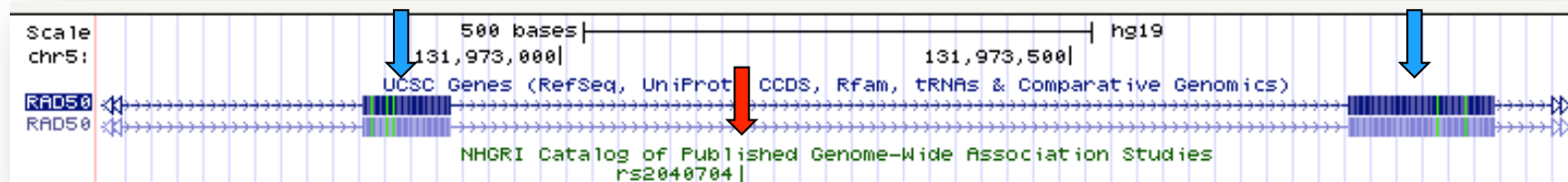


1,500 Letters Of Our 3 Billion Letter Genome

agccaagcagcaaagtttgctgctgtttatgtttagctcttactatattctacttttaccattgaaaatattgaggaagtatt
tatatttctatttttatatattatattttatgtattttaataattactattacacataattatttttatatatatgaagtaccaatg
acttccttttccagagcaataatgaaatttcacagtatgaaaatggaagaaatcaataaaattatacgtgacctgtggcgaa
gtacctatcgtggacaaggtgagtaccatgggtgtatcacaatgctctttccaaagccctctccgcagctcttccccttatga
cctctcatcatgccagcattacctcctggaccccttctaagcatgtctttgagattttctaagaattcttatcttggcaacatc
ttgtagcaagaaaatgtaaagttttctgttccagagcctaacaggacttacatatttgactgcagtaggcattatatttagctg
atgacataataggttctgtcatagtgtagatagggataagccaaaatgcaataagaaaaaccatccagaggaaactctttt
tttttcttttcttttttttttccagatggagtctcgcacttctctgtcaccgggctggagcgagtggtgcaatcttggctca
ctgcaacctccacctcctgggttcaggtgattctccacctcagcctcccgagtagtagctggaattacaggtgcgcgctccc
acacctggctaatttttgtattcttagtagagatgggggttcacatgttggccaggctggtctcaaactcctgccctcaggtg
atctgcccaccttggcctcccagtggttgggttacaggcgtgagccaccgcgctggcctggaggaaactcttaacagggaa
actaagaaagagttgaggctgaggaactggggcatctgggttgcttctggccagaccaccaggctcttgaatcctccagc
cagagaaagagttccacaccagccattgtttcctctggtaatgtcagcctcatctgttgttctaggcttacttgatatgtttg
taaatagacaaaaggctacagagcataggttcctctaaaatattcttcttctgtgtcagatattgaatacatagaaatacgg
ctgatgccgatgaaaatgtatcagcttctgataaaaggcgggaattataactaccgagtgggtgatgctgaaggagacacag
ccttggatatgagaggacgatgcagtgtggacaaaaggcaggtatctcaaaagcctggggagccaactcacccaagtaa
ctgaaagagagaaacaaacatcagtgcagtggaagcacccaaggctacacctgaatggtgggaagctctttgctgtata
taaatgaatcaggctcagctactattattacactctcctgaagctaaccaacatttctgcaacattatgtagactt

Maps And Annotation Help Us To Understand The Sequence

agccaagcagcaaagtttgcctgctgtttatTTTTgtagctcttactatattctacttttaccattgaaaatattgaggaagtatt
 tatatttctatttttatatatattatattttatgtattttaataattactattacacataattatttttatatatatgaagtaccaatg
 acttccttttccagagcaataatgaaatttcacagtatgaaaatggaagaaatcaataaaattatacgtgacctgtggcgaa
 gtacctatcgtggacaaggtgagtaccatggtgtatcacaaatgctctttccaaagccctctccgcagctcttccccttatga
 cctctcatcatgccagcattacctccctggaccccttctaagcatgtctttgagattttctaagaattcttatcttggcaacatc
 tttagcaagaaaatgtaaagtttctgttccagagcctaacaggacttacatatttgactgcagtaggcattatatttagctg
 atgacataataggttctgtcatagtgtagatagggataagccaaatgcaataagaaaaaccatccagaggaaactctttt
 tttttctttttctttttttttttccagatggagtctcgcacttctctgtcaccgggctggagcgcagtggtgcaatcttggctca
 ctgcaacctccacctcctgggttcaggtgatttcccacctcagcctcccagtagtagctggaattacaggtgcgcgctccc
 acacctggctaatttttgtattcttagtagagatgggggtttcacatgttggccaggctggtctcaaactcctgccctcaggtg
 atctgccaccttggcctcccagtggttgggttacaggcgtgagccaccgcgctggcctggaggaaactcttaacagggaa
 actaagaaagagttgaggctgaggaaactggggcatctgggttgcttctggccagaccaccagggtcttgaatcctcccagc
 cagagaaagagtttccacaccagccattgtttcctctggtaatgtcagcctcatctgttgttcttaggcttacttgatatgtttg
 taaatgacaaaaggctacagagcataggttcctctaaaatattcttcttctgtgtcagatattgaatacatagaaatcgggt
 ctgatgccgatgaaaatgtatcagcttctgataaaaggcgggaattataactaccgagtgggtgatgctgaaggagacacag
 cttggatatgagaggacgatgcagtgtggacaaaaggcaggtatctcaaagcctggggagccaactcacccaagtaa
 ctgaaagagagaaacaaacatcagtgagtggaagcacccaaggctacacctgaatggtgggaagctctttgctgctata
 taaaatgaatcaggctcagctactattattacactctcctgaagctaaccaacatttcttgcaacattatgtagactt

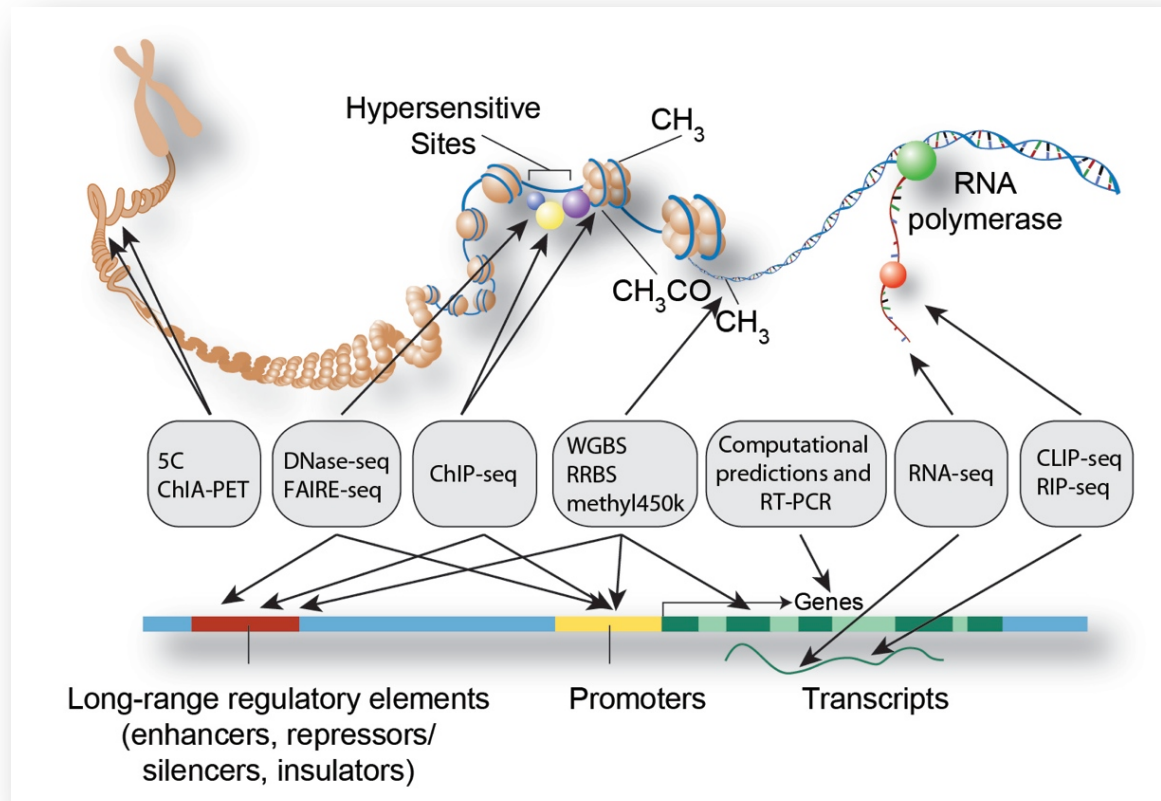


Richer Maps Provide More Information



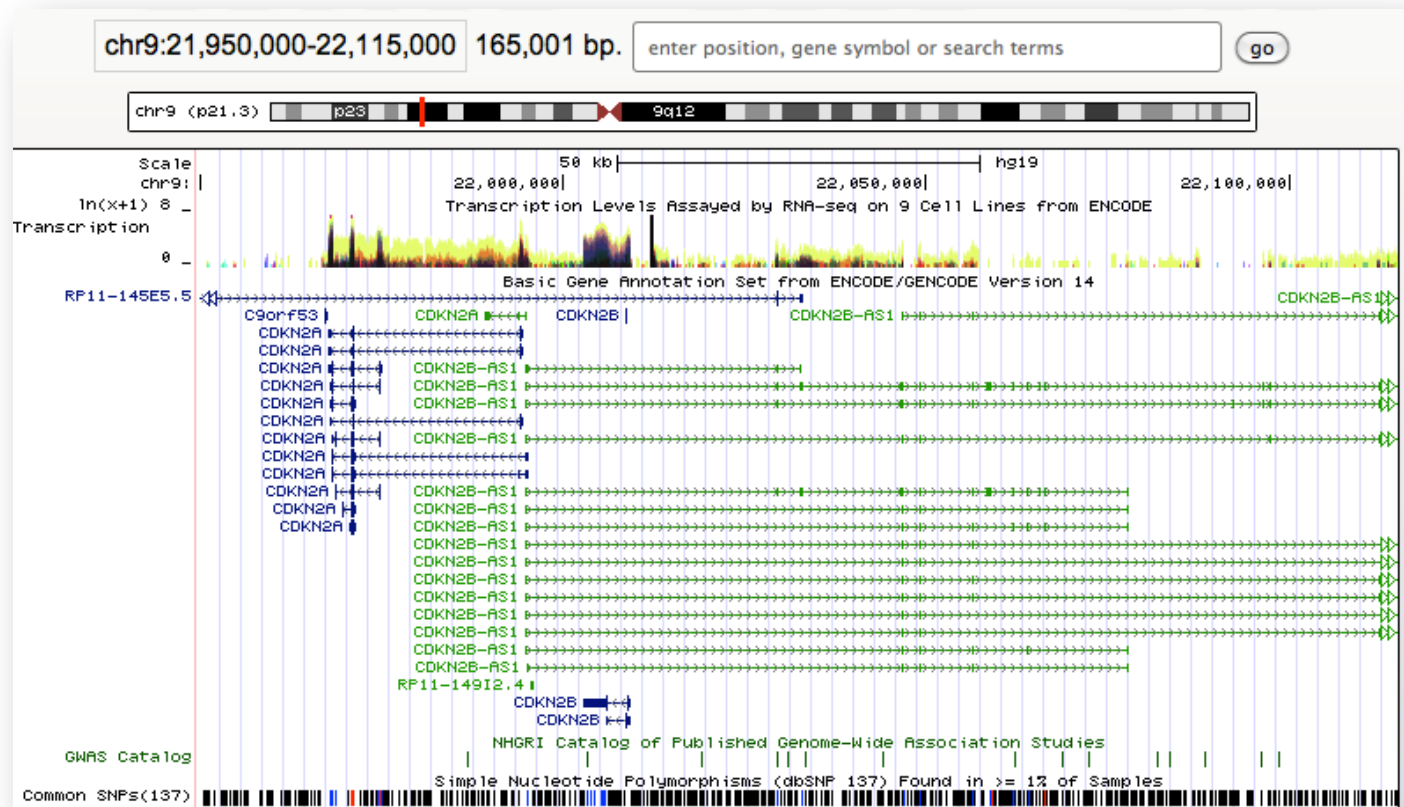


ENCODE Data: Built on Decades of Research Into Gene Regulation



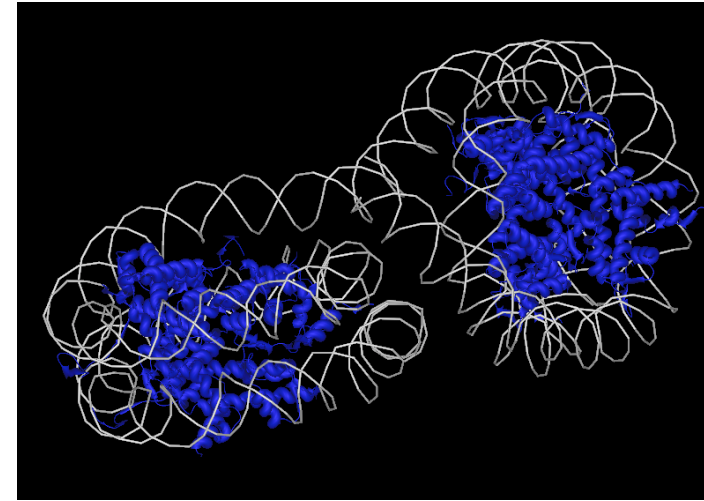
RNA Data

- RNA amount (mRNA and short RNA)
- Cell Specificity
- Transcript modeling (GENCODE)

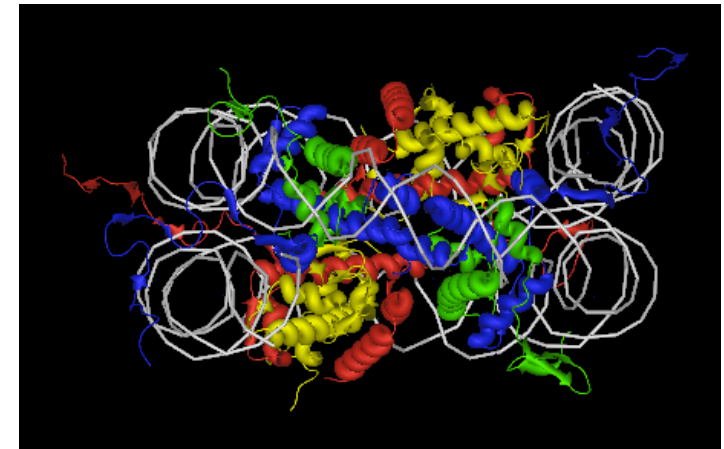


ENCODE Chromatin Structure Data

- DNase
- Histone modifications
- DNA methylation
 - Enhancers
 - Promoters
 - Transcription factor footprints
 - Transcribed regions
 - Active and repressed regions
 - Cell specificity



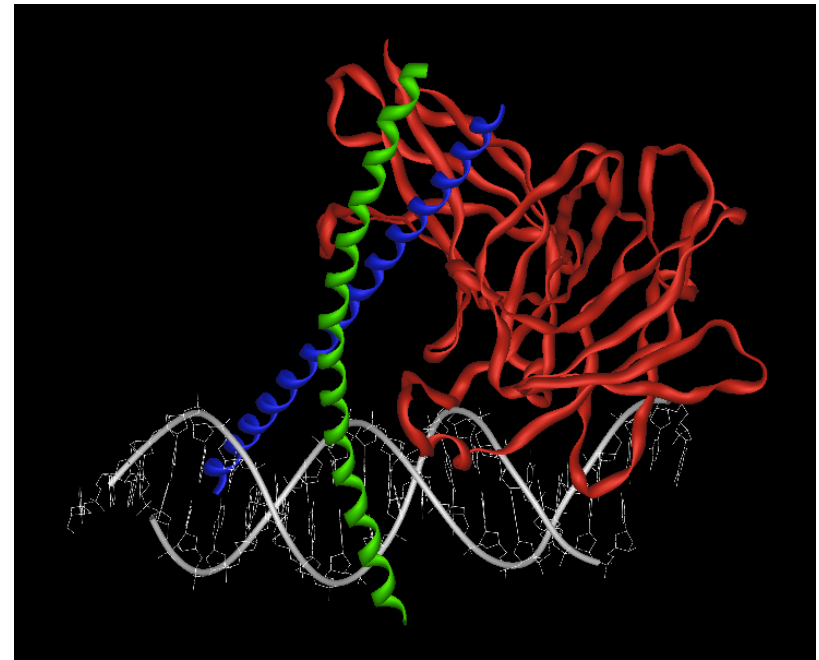
Richmond, PDB 1ZBB



Richmond, PDB 1KX5

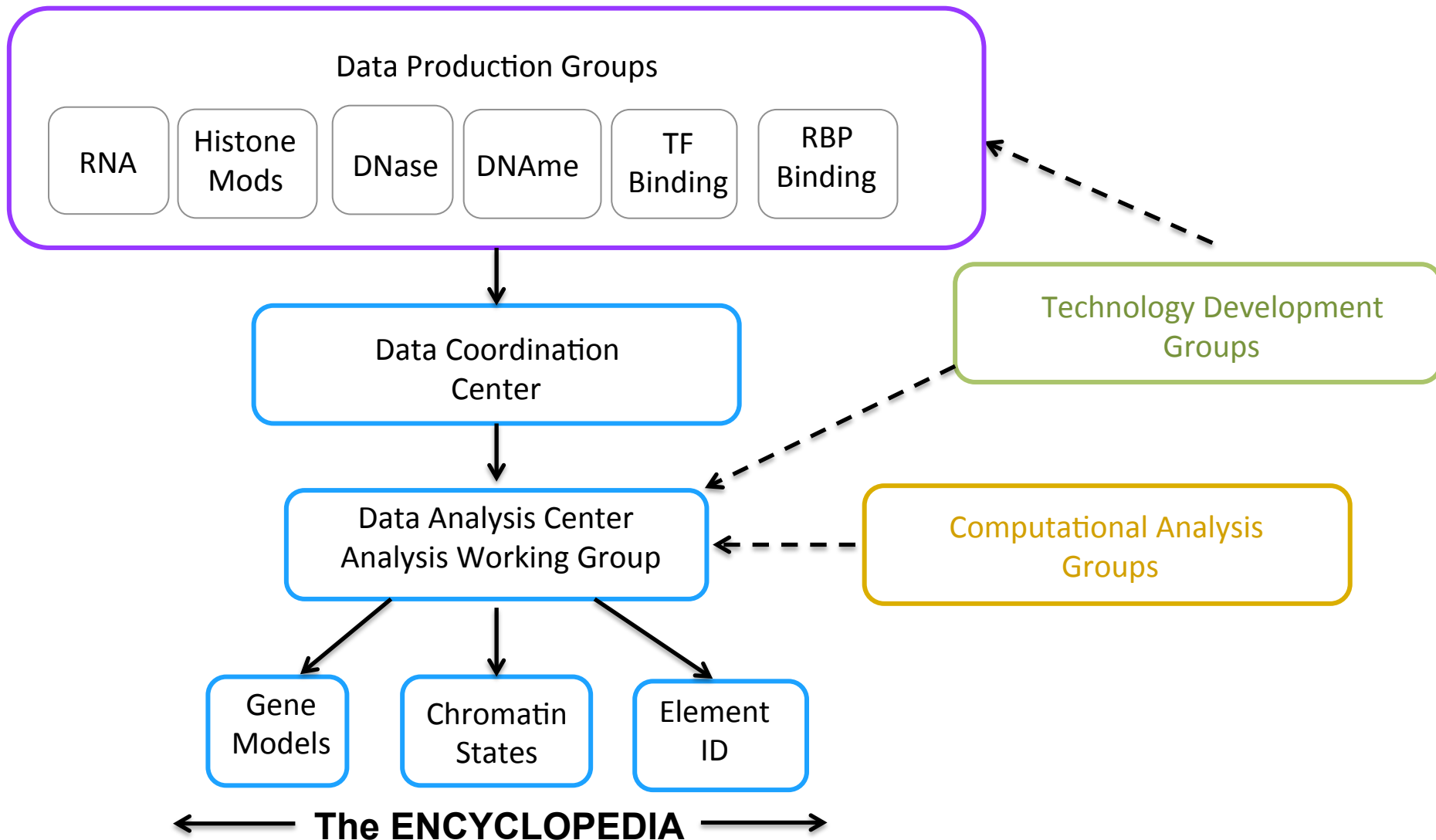
ENCODE Nucleic Acid Binding Data

- DNA binding proteins (Transcription factors)
 - Activators
 - Repressors
 - Remodelers
 - RNA Polymerases
 - Cell specificity
- RNA binding proteins
 - RNA Splicing
 - Translation
 - RNA Stability
 - RNA Localization
 - Cell specificity



Harrison, PDB_1A02

ENCODE Consortium Structure





Summary- ENCODE Resource

- Freely shared catalog of candidate genomic functional elements
- ENCODE is built upon established techniques and interpretations developed for the study of gene regulation
- ENCODE maps can be used to make predictions about genome function

Overview

- The ENCODE Resource
- Use of ENCODE to illuminate the role of genetic variation in human disease
- Accessing ENCODE materials

Using ENCODE Data: Access for Scientists with Varying Expertise

- Many tools for deriving biological insight and hypotheses from ENCODE data
- Many do not require informatics experience
- Many do not require deep functional genomics background
- Encouraging broad use of ENCODE resources

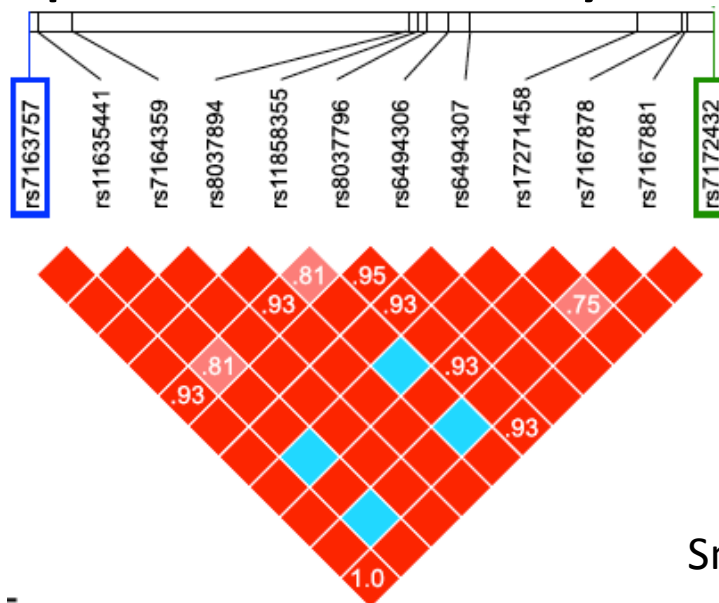
Standard ENCODE Use Cases:

Hypothesis Generation

- Prediction of causal variants/regulatory elements
- Prediction of target genes
- Prediction of target cell types
- Applied to germline genetic variation, but many other applications (e.g., differentially methylated regions)

Prediction of Causal Variants

- Multiple variants may be in linkage disequilibrium
- The causal variant may not have been tested during data collection
- Multiple variants may be causal



ENCODE/Epigenomics Data From HaploReg

HaploReg v4.1



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 chromatin state and protein binding annotation from the Roadmap Epigenomics and ENCODE projects, sequence conservation across mammals, the effect of SNPs on regulatory motifs, and the effect of SNPs on expression from eQTL studies. HaploReg is designed for researchers developing mechanistic hypotheses of the impact of non-coding variants on clinical phenotypes and normal variation.

Update 2015.11.05: Version 4.1 GWAS and eQTL have been updated; a simpler pruning strategy is applied when combining GWAS; and links out to other NHGRI/EBI GWAS hits and GRASP QTL hits are provided.

Update 2015.09.15: Version 4.0 now includes many recent eQTL results including the GTEx pilot, four different options for defining enhancers using Roadmap Epigenomics data, and a complete set of source files for download and local analysis. Older versions available: [v3](#), [v2](#), [v1](#).

[Build Query](#) [Set Options](#) [Documentation](#)

Use one of the three methods below to enter a set of variants. If an r^2 threshold is specified (see the Set Options tab), results for each variant will be shown in a separate table along with other variants in LD. If r^2 is set to NA, only queried variants will be shown, together in one table.

Query (comma-delimited list of

Query SNP: **rs6575353** and variants with $r^2 \geq 0.8$

chr	pos (hg38)	LD (r²)	LD (D')	variant	Ref	Alt	AFR freq	AMR freq	ASN freq	EUR freq	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse	Proteins bound	Motifs changed	NHGRI/EBI GWAS hits	GRASP QTL hits	Selected eQTL hits	GENCODE genes	dbSNP func annot
14	93736606	0.83	-0.96	rs77582682	G	A	0.03	0.06	0.00	0.08				9 tissues	CTCF,RAD21	4 altered motifs				PRIMA1	intronic
14	93738999	1	1	rs6575353	G	A	0.71	0.85	0.51	0.92						Sox,TCF4	1 hit	1 hit		PRIMA1	intronic
14	93739616	0.95	1	rs12896080	G	A	0.61	0.84	0.51	0.91						4 altered motifs				PRIMA1	intronic
14	93739927	0.83	1	rs4905084	A	G	0.60	0.83	0.51	0.90						4 altered motifs				PRIMA1	intronic
14	93742184	0.95	1	rs4900194	A	G	0.61	0.84	0.51	0.91			4 tissues	BRN		Gfi1,KAP1,Smad3		1 hit		PRIMA1	intronic
14	93743570	0.97	1	rs11160137	A	G	0.69	0.85	0.52	0.91			5 tissues	ESC,HRT		4 altered motifs				PRIMA1	intronic
14	93743836	0.85	0.93	rs12587586	T	G	0.59	0.84	0.52	0.91			6 tissues			Roaz				PRIMA1	intronic

www.broadinstitute.org/mammals/haploreg/

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

ENCODE Data From RegulomeDB



RegulomeDB has been updated to Version 1.1. This includes bringing our database up-to-date with current ENCODE releases: Xie et al. (2013) and Boyle et al. (2014). We have also added Chromatin States from the Roadmap Epigenome Consortium (unpublished) as well as updates to DNase footprinting, PWMs, and DNA Methylation.

Enter dbSNP IDs, 0-based coordinates, BED files, VCF files, GFF3 files (hg19).

rs4900194

Submit

Use RegulomeDB to identify DNA features and regulatory elements in non-coding regions of the human genome by entering ...

dbSNP IDs

Single nucleotides

A chromosomal region

Enter dbSNP ID(s) (example) or upload a list of dbSNP IDs to identify DNA features and regulatory elements that contain the coordinate of the SNP(s).



A project of the Center for Genomics and Personalized Medicine at Stanford University

Summary of SNP

Chromatin structure Filter:

Method	Location	Cell Type	Additional Info	Reference
DNase-seq	chr14:94207796..94208834	Cerebellumoc		ENCODE
DNase-seq	chr14:94208390..94208598	Paniclets		ENCODE
DNase-seq	chr14:94208414..94209060	T47d		ENCODE
DNase-seq	chr14:94208516..94209130	Hepatocytes		ENCODE

Histone modifications Filter:

Method	Location	Chromatin State	Tissue Group	Tissue	Reference
ChromHMM	chr14:94201400..94210600	Weak Repressed PolyComb	ESC	ES-UCSF4 Cell Line	REMC
ChromHMM	chr14:94201400..94211200	Quiescent/Low	ESC	H1 Cell Line	REMC
ChromHMM	chr14:94201400..94211400	Quiescent/Low	ESC	HUES6 Cell Line	REMC
ChromHMM	chr14:94201400..94211600	Quiescent/Low	ESC	ES-I3 Cell Line	REMC

Show 10 entries

Coordinate (0-based)	dbSNP ID	Regulome DB Score	Other Resources
chr14:94208529	rs4900194	5	UCSC ENSEMBL dbSNP

Showing 1 to 1 of 1 entries

<http://regulomedb.org/>

RegulomeDB

Linkin

Access the list

- List of all as
- By phenoty

C colorectal cancer

chr1:222,045,446 rs6
chr1:222,164,948 rs6
chr3:169,492,101 rs1
chr6:160,840,252 rs7
chr8:117,630,683 rs16892766
chr8:128,407,443 rs1
chr8:128,413,305 rs6
chr8:128,424,792 rs7
chr10:8,701,219 rs10
chr11:111,171,709 rs1
chr12:51,155,663 rs1
chr14:54,410,919 rs4
chr15:32,994,756 rs4
chr16:68,820,946 rs9
chr18:46,453,463 rs16892766
chr19:33,532,300 rs1
chr20:6,404,281 rs96
chr20:60,921,044 rs4

rs16892766

Association Study: A genomewide association study of type 2 diabetes mellitus [PMID:1837290]
First author: Teasdale
Journal: Nat Genet
Date: 03/30/2008
Phenotype: C
Study population: European
Replication population: European
Association P-value: 1.1e-10
Odds ratio [95% CI]: 1.12 [1.05, 1.19]
Gene: TRPS1 -
Risk allele: A
Minor allele frequency: 0.0001

Lead SNP

rs16892766

Position: chr8:117,630,683
Distance to nearest gene: 117,630,683 bp from TRPS1
Gencode v7 ID: TRPS1
RegulomeDB Score: 2b
Linkage disequilibrium: r²=1.0 with all HapMap SNPs in LD

Data supporting chr8:117630682 (rs16892766)

Score: 2b

Likely to affect binding

Human Feb. 2009 (GRCh37/hg19) chr8:117,630,482-117,630,882 (401 bp)
100 bases | hg19
| 117,630,550 | 117,630,600 | 117,630,650 | 117,630,700 | 117,630,750 | 117,630,800 | 117,630,850 |
RefSeq Genes
Publications: Sequences in scientific articles
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Digital DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE
Transcription Factor ChIP-seq from ENCODE
Placental Mammal Basewise Conservation by PhyloP
Simple Nucleotide Polymorphisms (dbSNP 137) Found in >= 1% of Samples
Repeating Elements by RepeatMasker

- In all HapMap 2 populations: $r^2 \geq 0.8$ $r^2 \geq 0.9$ $r^2 = 1.0$
- In the HapMap 2 CEU population $r^2 \geq 0.8$ $r^2 \geq 0.9$ $r^2 = 1.0$
No SNPs found for this LD threshold.

Cherry, Snyder, Genome Research 22-1790, 2012



ENCODE cis-element Browser

Candidate cis-elements in your queried region.

Human (hg19)
chr8:128390000-128410000

DNaseI Hypersensitive Sites: 

Coordinate	Tissue/cell type
chr8:128394860-128395010	NHDF-Ad
chr8:128395580-128395730	HSMMtube,HSMM
chr8:128398205-128398355	Osteobl
chr8:128398585-128398735	Osteobl
chr8:128399500-128399650	GM12878,NHDF-Ad,HSMM
chr8:128400960-128401110	HSMM,HSMMtube
chr8:128402480-128402630	HSMM
chr8:128403580-128403730	HMEC,Osteobl,HSMM,NHDF-Ad,HSMMtube,NH-A,HeLa-S3,NHEK,NHLF
chr8:128404560-128404710	HMEC
chr8:128404720-128404870	HSMM
chr8:128405400-128405550	HSMM
chr8:128407420-128407570	HeLa-S3
chr8:128407885-128408035	HUVEC,Osteobl,NHDF-Ad
chr8:128408160-128408310	HMEC

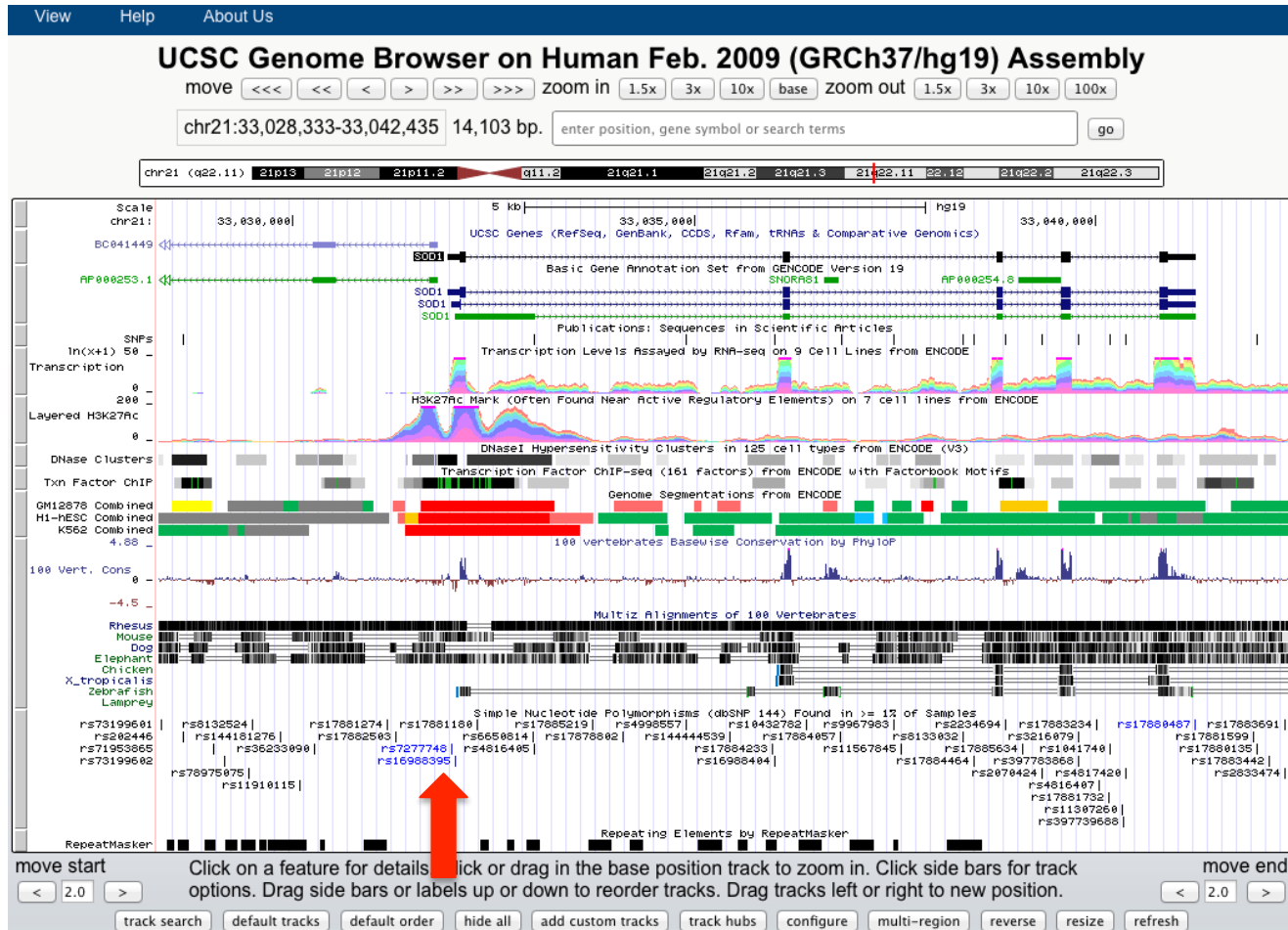
TF binding Site: 

Coordinate	TF	tissue
chr8:128398585-128398735	USF1	USF1(K562), USF1(SK-N-SH_RA)
chr8:128399500-128399650	RUNX3, SPI1	RUNX3(GM12878), SPI1(GM12878), SPI1(GM12891)
chr8:128403580-128403730	multiple	CEBPB(HeLa-S3), CEBPB(IMR90), EP300(HeLa-S3), FOS(MCF10A-Er-Src), FOXA1(A549), GATA3(T-47D), JUN(HeLa-S3), JUND(HeLa-S3), MAX(HeLa-S3), MYC(MCF10A-Er-Src), NR3C1(A549), POLR2A(HeLa-S3), POLR2A(MCF10A-Er-Src), RCOR1(HeLa-S3), SMC3(HeLa-S3), STAT3(HeLa-S3), STAT3(MCF10A-Er-Src), TAF1(HeLa-S3), TBP(HeLa-S3), TCF7L2(HeLa-S3), TFAP2A(HeLa-S3), TFAP2C(HeLa-S3)



ENCODE Browser

Viewing Locus Of Interest

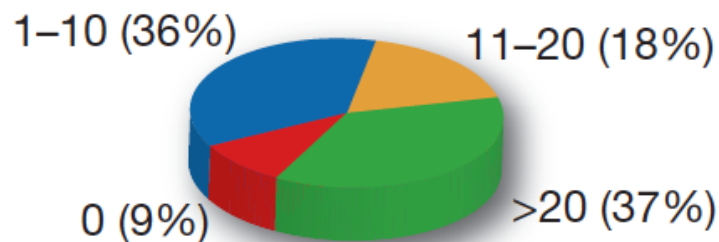


<https://genome.ucsc.edu/encode/>

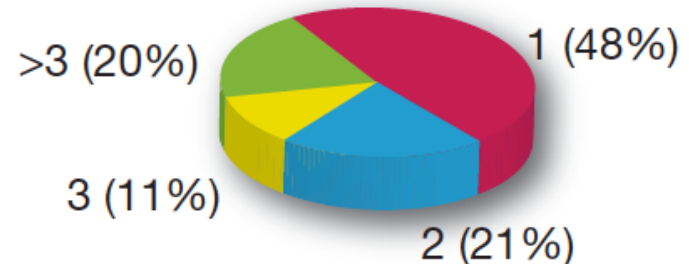
Prediction of Target Genes

- Regulatory regions can operate on multiple, distal genes
- The target gene could be a non-coding RNA

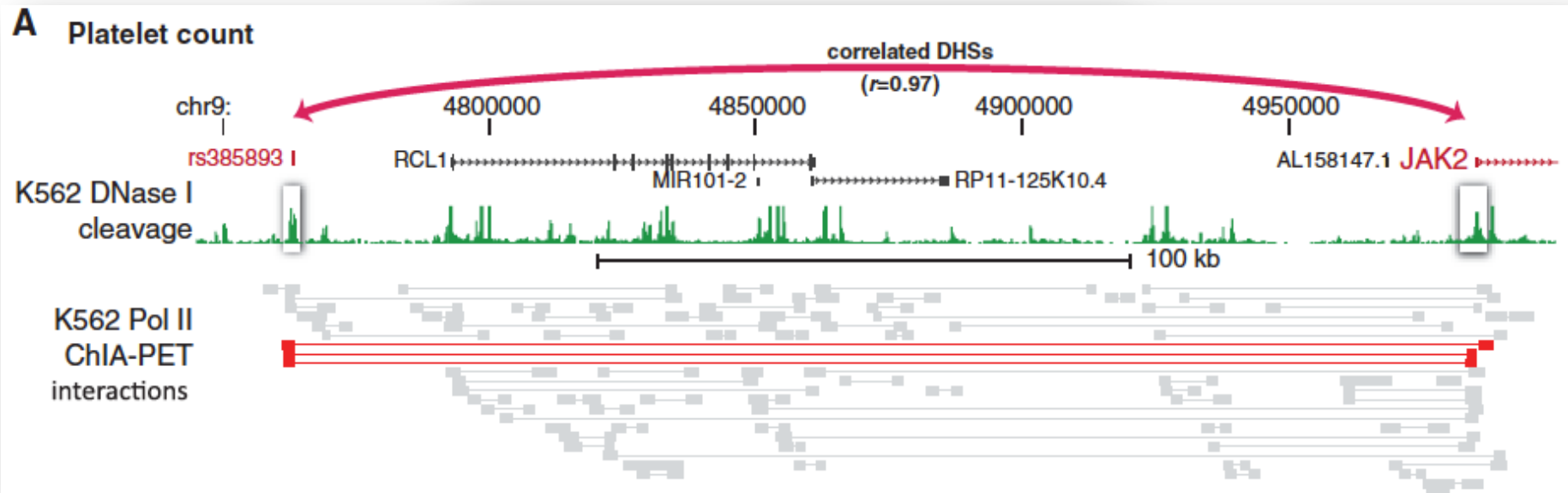
Distal DHSs connected
per promoter DHS
($n = 69,965$)



Promoter DHSs connected
per distal DHS
($n = 578,905$ of 1,454,901 total)



Many GWAS Associations Lie In Regions Linked To Distal Genes

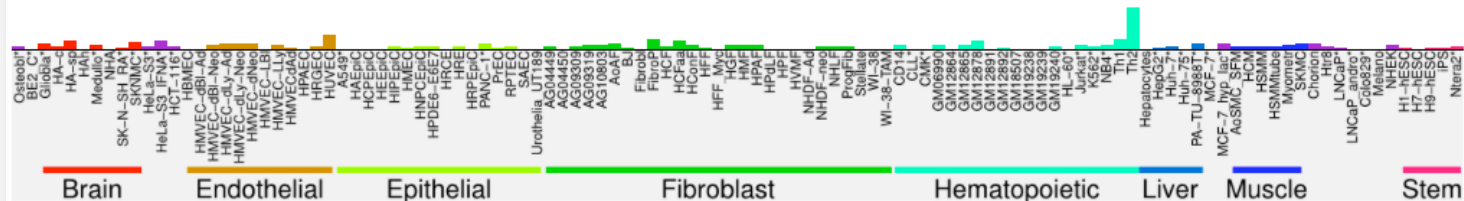


Prediction of Linkage Between Regulatory Elements and Genes

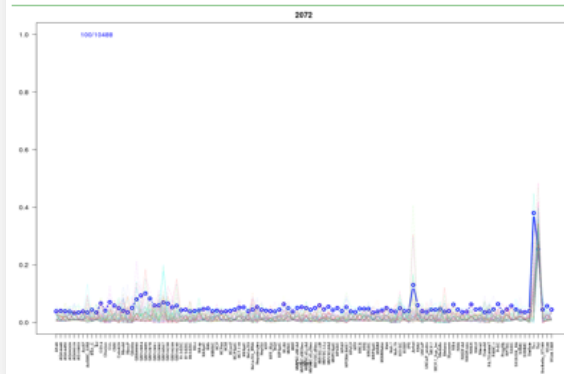
Regulatory Elements Database

Chromosome, start, stop: **DHS: #2174550**

chr5: 131972960-131973110

Belongs to SOM cluster: [2072](#)[Site Hypersensitivity Profile](#)

Cluster Profile:



RESOURCES

Correlated Genes:

p-values indicate significant **higher** or **lower** correlation 1 genes found

Gene Pvalue

IL13 0.009

External Databases

[UCSC](#)

[Ensembl](#)

<http://dnase.genome.duke.edu>

Furey, Crawford, Stamatoyannopoulos, Genome Res. 23:777, 2013



ENCODE cis-element Browser

Cis-elements linked to your queried gene.

Human (hg19)

Gene **IL13** [NM_002188, ENSG00000169194, ENST00000304506]

Cis-element lined by DNaseI Hypersensitive Sites Linkage:

Proximal DHS (TSS)	start	end	Gene	Distal DHS	start	end	correlation
chr5	131992140	131992290	IL13	chr5	131512800	131512950	0.743283
chr5	131992140	131992290	IL13	chr5	131558440	131558590	0.761866
chr5	131992140	131992290	IL13	chr5	131571820	131571970	0.782866
chr5	131992140	131992290	IL13	chr5	131720440	131720590	0.766176
chr5	131992140	131992290	IL13	chr5	131732540	131732690	0.739405
chr5	131992140	131992290	IL13	chr5	131745200	131745350	0.765629
chr5	131992140	131992290	IL13	chr5	131747860	131748010	0.749684
chr5	131993580	131993730	IL13	chr5	131917860	131918010	0.797702
chr5	131993580	131993730	IL13	chr5	131921920	131922070	0.800141
chr5	131993580	131993730	IL13	chr5	131970980	131971130	0.772113
chr5	131993580	131993730	IL13	chr5	131971640	131971790	0.763557
chr5	131993580	131993730	IL13	chr5	131972960	131973110	0.797839
chr5	131993580	131993730	IL13	chr5	131977060	131977210	0.848905
chr5	131993580	131993730	IL13	chr5	131990420	131990570	0.855445
chr5	131993580	131993730	IL13	chr5	131993880	131994030	0.769354
chr5	131993580	131993730	IL13	chr5	132011780	132011930	0.756074
chr5	131993580	131993730	IL13	chr5	132077740	132077890	0.820222
chr5	131993580	131993730	IL13	chr5	132083520	132083670	0.770558
chr5	131993580	131993730	IL13	chr5	132164240	132164390	0.837496
chr5	131993580	131993730	IL13	chr5	132200200	132200350	0.752104

Prediction of Target Cell Types

- Some diseases are known to affect multiple cell types
- The defect may not be intrinsic to the cell type with obvious pathology
- The disease etiology may not be completely known

Prediction of Linkage Between Regulatory Elements and Cell Type

BROAD INSTITUTE **MIT**

the noncoding genome at variants on haplotype blocks, such as candidate regulatory SNPs at disease-1000 Genomes Project, linked SNPs and small indels can be visualized along with their predicted cross mammals, and their effect on regulatory motifs. HaploReg is designed for researchers developing coding variants on clinical phenotypes and normal variation.

in beta.

an expanded library of SNPs (based on dbSNP 137), motif instances (based on PWMs discovered from (adding 90 cell types from the Roadmap Epigenome Mapping Consortium), and eQTLs (from the GTEx provided based on the 1000 Genomes Phase 1 individuals, and r^2 and D' measurements are available. Comments include improved cell metadata, gene metadata, and PWM display on the detail pages and the e.

on

set of variants. If an r^2 threshold is specified (see the Set Options tab), results for each variant will be variants in LD. If r^2 is set to NA, only queried variants will be shown, together in one table.

a single
art-end):

er line): no file selected

GWAS:

SIPhy cons	Promoter histone marks	Enhancer histone marks	DNAse	Proteins bound	eQTL tissues	Motifs changed	GENCODE genes	dist	fu
			4 cell types	FOXA1,GR		Rhox11	24kb 3' of EIF3H		
				Th1,AoSMC,NH-A,RPTEC			33kb 3' of EIF3H		
							23kb 3' of EIF3H		
							19kb 3' of EIF3H		

Protein Binding					
Filter:					
Method	Location	Bound Protein	? Cell Type	Additional Info	Reference
ChIP-seq	chr8:117630539..117630739	FOXA1	ECC-1	D_0.02pct	ENCODE
ChIP-seq	chr8:117630626..117630842	NR3C1	ECC-1	DEX_100nM	ENCODE

Chromatin structure					Filter:
Method	Location	Cell Type	Additional Info	Reference	
DNase-seq	chr8:117630480..117630690	Rptec		ENCODE	
DNase-seq	chr8:117630480..117630730	Nhlf		ENCODE	

Histone modifications					
Filter:					
Method	Location	Histone Mark	? Cell Type	Additional Info	Reference
ChIP-seq	chr8:110578383..117647033	H3k09me3	Dnd41		ENCODE
ChIP-seq	chr8:116009496..120997897	H2az	Hepg2		ENCODE

www.broadinstitute.org/mammals/haploreg/

<http://regulomedb.org/>

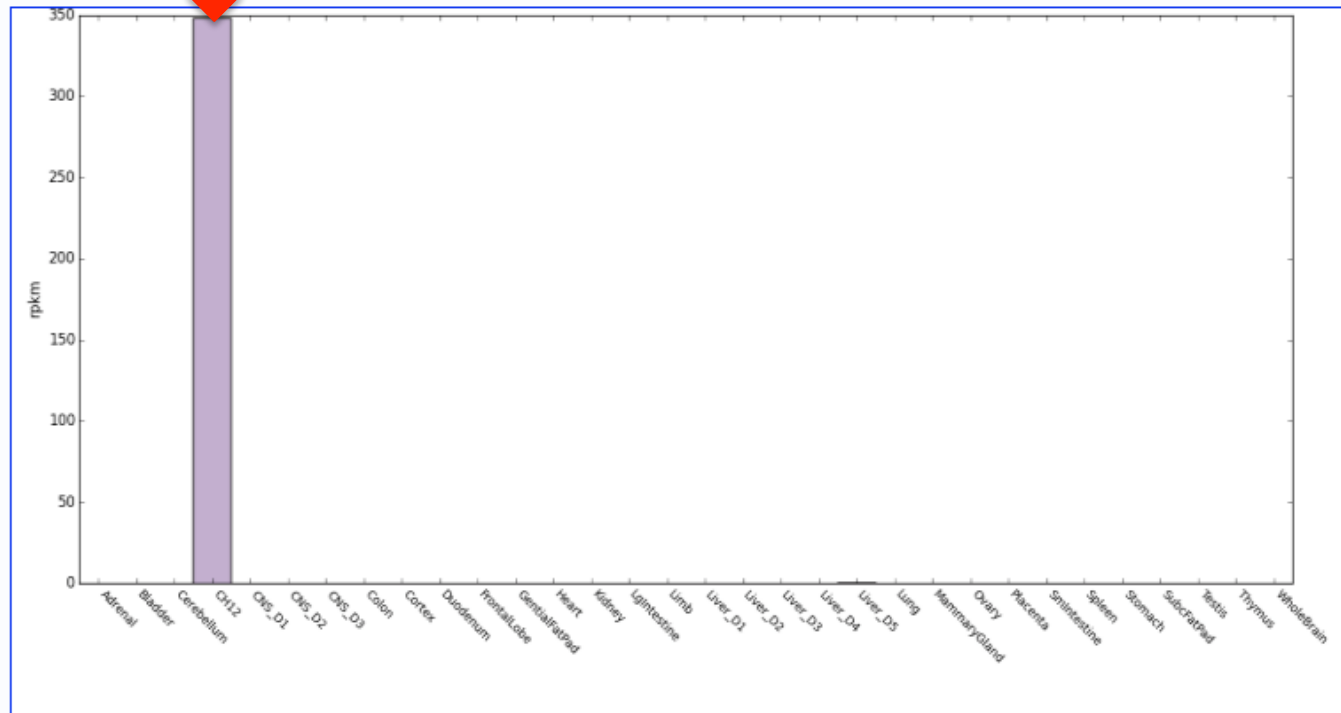
Ward and Kellis, Nucleic Acids Research 40-D930, 2011

Cherry, Snyder, Genome Research 22-1790, 2012



ENCODE cis-element Browser

Gene **II10** (mCC2645) [NM_010548, ENSMUSG00000016529, ENSMUST00000016673]



Adrenal	0
Bladder	0.07
Cerebellum	0
CH12	348.42
CNS_D1	0
CNS_D2	0.01

<https://www.encodeproject.org/data/annotations/>

Summary- ENCODE Use Cases

Major use: Hypothesis generation and refinement

- Prediction of causal variants/regulatory elements
- Prediction of target genes
- Prediction of target cell types

Overview

- The ENCODE Resource
- Use of ENCODE by the research community
- **Accessing ENCODE materials**

ENCODE Data Standards

ENCODE

Data ▾

Methods ▾

About ENCODE ▾

Help ▾

Search ENCODE

Data standards

Overview

The ENCODE consortium analyzes the quality of the data produced using a variety of metrics. This page describes the data standards and metrics that are used to evaluate the data and what they appear to measure. These quality metrics will be updated on occasion to include analysis of more recent data.

It is important to note that quality metrics for evaluating epigenomic assays is an area of research, so standards are emerging as more metrics are used with more datasets and types of experiments. The typical values for a quality metric can be quite different with different assays, or even comparing different features in the same assays, such as different antibodies used in ChIP-seq experiments. Currently there is no single measurement that identifies all high-quality or low-quality samples. As with quality control for other types of experiments, multiple assessments (including manual inspection of tracks) are useful because they may capture different concerns. Comparisons within an experimental method (e.g., comparing replicates to each other, or comparing values for one antibody in several cell types, or the same antibody and cell type in different labs) can help identify possible stochastic error.

Experimental guidelines

The ENCODE Consortium has adopted uniform guidelines for the most common ENCODE experiments. The guidelines have evolved over time as technologies have changed. The current guidelines are informed by results gathered during the project. Previous versions of the standards are also available for reference.

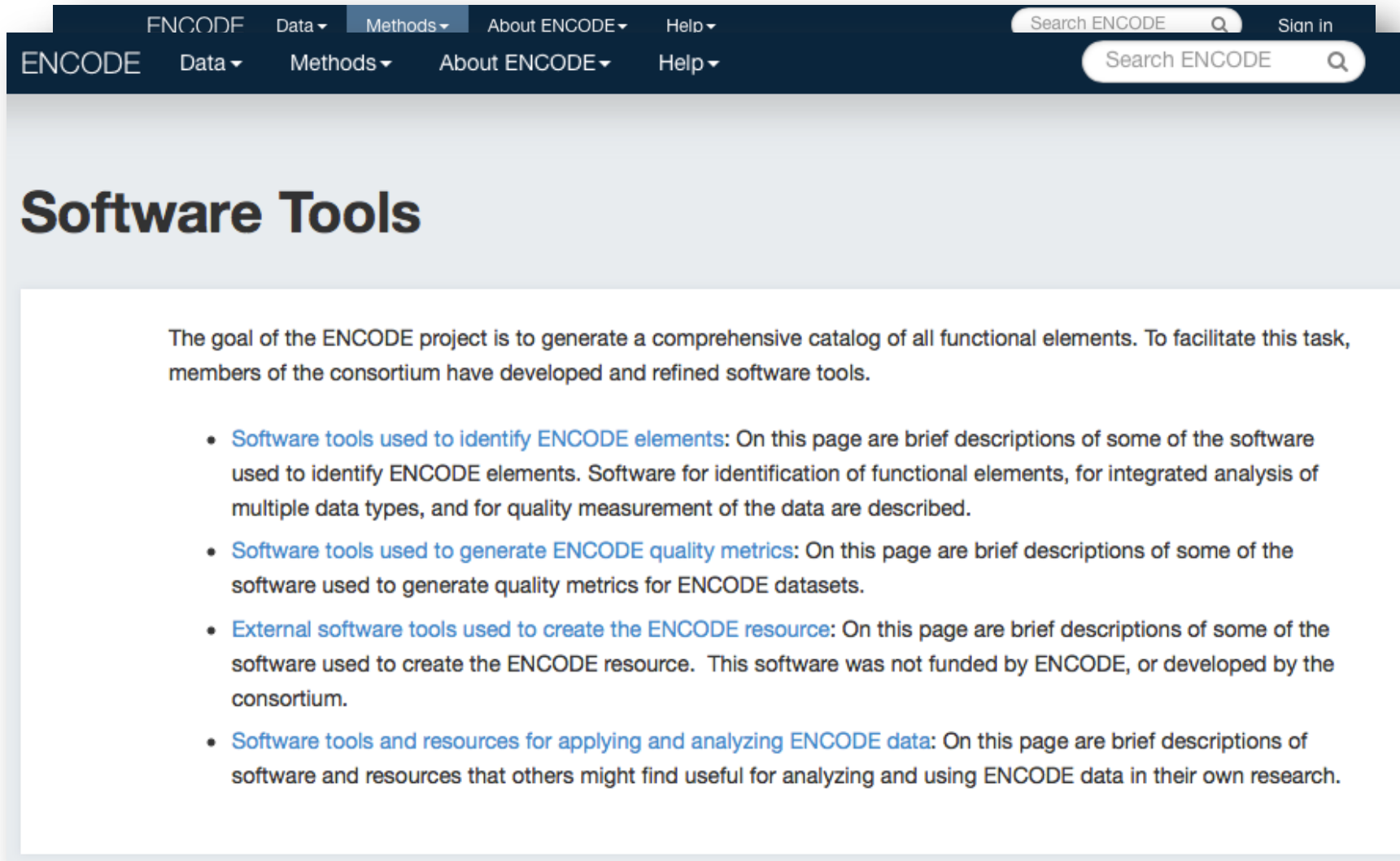
- [Current experiment guidelines](#)
- [Antibody characterizations guidelines](#)

Quality metrics

The ENCODE consortium analyzes the quality of the data produced using a variety of metrics. Those generated for datasets published as part of

<https://www.encodeproject.org>

ENCODE Software Tools

A screenshot of the ENCODE project's 'Software Tools' page. The page has a dark blue header with navigation links: 'ENCODE', 'Data', 'Methods', 'About ENCODE', and 'Help'. There are also search bars and a 'Sign in' link. The main content area is white with a light blue header. The title 'Software Tools' is in a large, bold, dark blue font. Below the title, there is a paragraph explaining the goal of the ENCODE project. This is followed by a bulleted list of four categories of software tools, each with a brief description. The list items are: 'Software tools used to identify ENCODE elements', 'Software tools used to generate ENCODE quality metrics', 'External software tools used to create the ENCODE resource', and 'Software tools and resources for applying and analyzing ENCODE data'.

ENCODE Data Methods About ENCODE Help Search ENCODE Sign in

Software Tools

The goal of the ENCODE project is to generate a comprehensive catalog of all functional elements. To facilitate this task, members of the consortium have developed and refined software tools.

- **Software tools used to identify ENCODE elements:** On this page are brief descriptions of some of the software used to identify ENCODE elements. Software for identification of functional elements, for integrated analysis of multiple data types, and for quality measurement of the data are described.
- **Software tools used to generate ENCODE quality metrics:** On this page are brief descriptions of some of the software used to generate quality metrics for ENCODE datasets.
- **External software tools used to create the ENCODE resource:** On this page are brief descriptions of some of the software used to create the ENCODE resource. This software was not funded by ENCODE, or developed by the consortium.
- **Software tools and resources for applying and analyzing ENCODE data:** On this page are brief descriptions of software and resources that others might find useful for analyzing and using ENCODE data in their own research.

Downloading and Visualizing

ENCODE

Data ▾

Methods ▾

About ENCODE ▾

Help ▾

Search ENCODE



Data Use Policy for External Users

The goal of the Encyclopedia of DNA Elements (ENCODE) Project is to build a comprehensive catalog of candidate functional elements in the genome. The catalog includes genes (protein-coding and non-protein coding), transcribed regions, and regulatory elements, as well as information about the tissues, cell types and conditions where they are found to be active. The current phase of ENCODE (2012-2016) greatly expands the number of cell types, data types and assays and includes the study of both the human and mouse genomes.

Like the Human Genome Project, the ENCODE Project seeks rapid data dissemination and use by the entire scientific community. Accordingly, to encourage the widest possible use of the datasets, all data produced will be available for unrestricted use immediately upon release to public databases, eliminating the nine-month moratorium previously used by ENCODE.

 **External data users may freely download, analyze and publish results based on any ENCODE data without restrictions as soon as they are released.** This applies to all datasets, regardless of type or size, and includes no grace period for ENCODE data producers, either as individual members or as part of the Consortium. Researchers using unpublished ENCODE data are encouraged to contact the data producers to discuss possible coordinated publications; however, this is optional. The Consortium will continue to publish the results of its own analysis efforts in independent publications.

 We request that researchers who use ENCODE datasets (published or unpublished) in publications and talks cite the ENCODE Consortium in all of the following ways:

skin of body

125

Lab: John Stamatoyannopoulos, UW

Experiment
ENCSR958IUL
released

Experiment
ENCSR810JIM
released

Experiment
ENCSR000CXE
released

Experiment
ENCSR331IKA
released

Experiment
ENCSR109NSK
released

Experiment
ENCSR345QON

<https://www.encodeproject.org>

Downloading and Visualizing

The screenshot displays the ENCODE Project website interface. The top navigation bar includes links for ENCODE, Data, Methods, About, and Help. A search bar is located in the top right corner. The main content area is titled "Experiment Matrix" and features a table with columns for Assay, Assay category, Target of assay, Date released, and Available data. A sidebar on the left provides filters for Assay category, Assay, Project, and RFA. The main table shows 17 results for DNase-seq experiments, including details such as the experiment name, lab, project, and release status.

Experiment Matrix

Click or enter experiment ID

Enter search

Assay category

- DNA binding
- Transcription
- DNA accessibility
- DNA methylation
- RNA binding

Assay

- ChIP-seq
- RNA-seq
- DNase-seq
- shRNA RNA-seq
- RNA microarray

Project

- ENCODE
- Roadmap
- modENCODE

RFA

- Roadmap
- ENCODE2
- ENCODE3
- modENCODE
- ENCODE2-Mouse

Organism

- Homo sapiens
- Mus musculus
- Drosophila
- Caenorhabditis
- Drosophila

Biosample

- immortal
- tissue
- primary
- stem cell
- whole organism

Assay category

- DNA accessibility 17

Assay

- ChIP-seq 82
- RNA-seq 19
- DNase-seq 17
- RNA microarray 10
- WGBS 7

Project

- Roadmap 14
- ENCODE 3

RFA

- Roadmap 14
- ENCODE2 1
- ENCODE2-Mouse 1
- ENCODE3 1

Experiment status

- released 17

Genome assembly (visualization)

- hg19 2
- mm9 1

Organism

- Homo sapiens 16
- Mus musculus 1

Showing 17 of 17 results

Visualize **Download**

DNase-seq of heart (*Homo sapiens*, fetal 96 day)

Lab: John Stamatoyannopoulos, UW
Project: Roadmap

Experiment
ENCSR539OPF released

DNase-seq of heart (*Homo sapiens*, child 3 year)

Lab: John Stamatoyannopoulos, UW
Project: Roadmap

Experiment
ENCSR305UJX released

DNase-seq of heart (*Homo sapiens*, fetal 96 day)

Lab: John Stamatoyannopoulos, UW
Project: Roadmap

Experiment
ENCSR911LT released

DNase-seq of heart (*Homo sapiens*, fetal 101 day)

Lab: John Stamatoyannopoulos, UW
Project: Roadmap

Experiment
ENCSR366EGE released

DNase-seq of heart (*Homo sapiens*, fetal 117 day)

Lab: John Stamatoyannopoulos, UW
Project: Roadmap

Experiment
ENCSR174JMM released

DNase-seq of heart (*Homo sapiens*, fetal 96 day)

Lab: John Stamatoyannopoulos, UW
Project: Roadmap

Experiment
ENCSR536NGW released

DNase-seq of heart (*Homo sapiens*, fetal 103 day)

Lab: John Stamatoyannopoulos, UW
Project: Roadmap

Experiment
ENCSR536NGW released

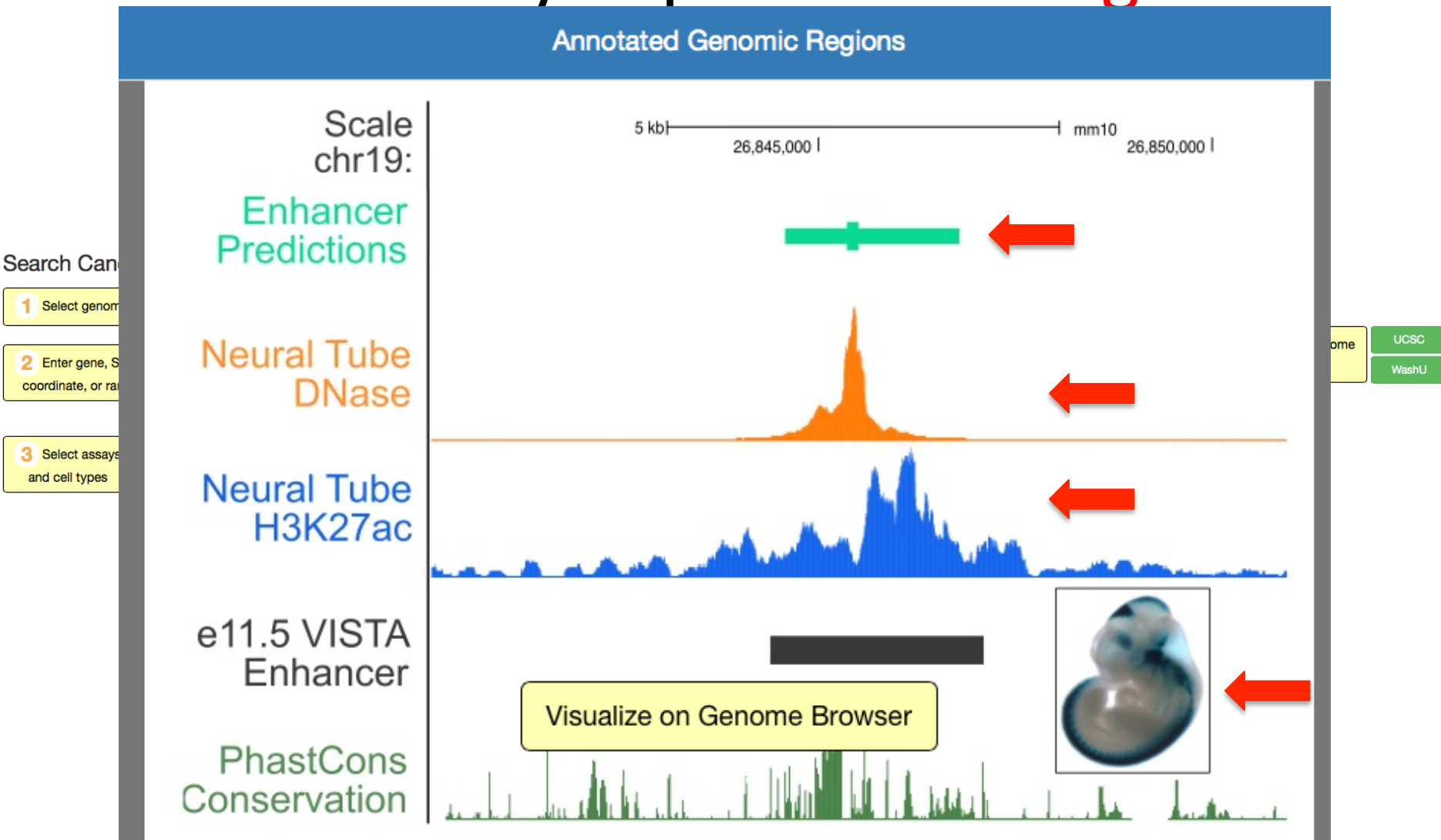
ENCODE Encyclopedia Prototype

ENCOD

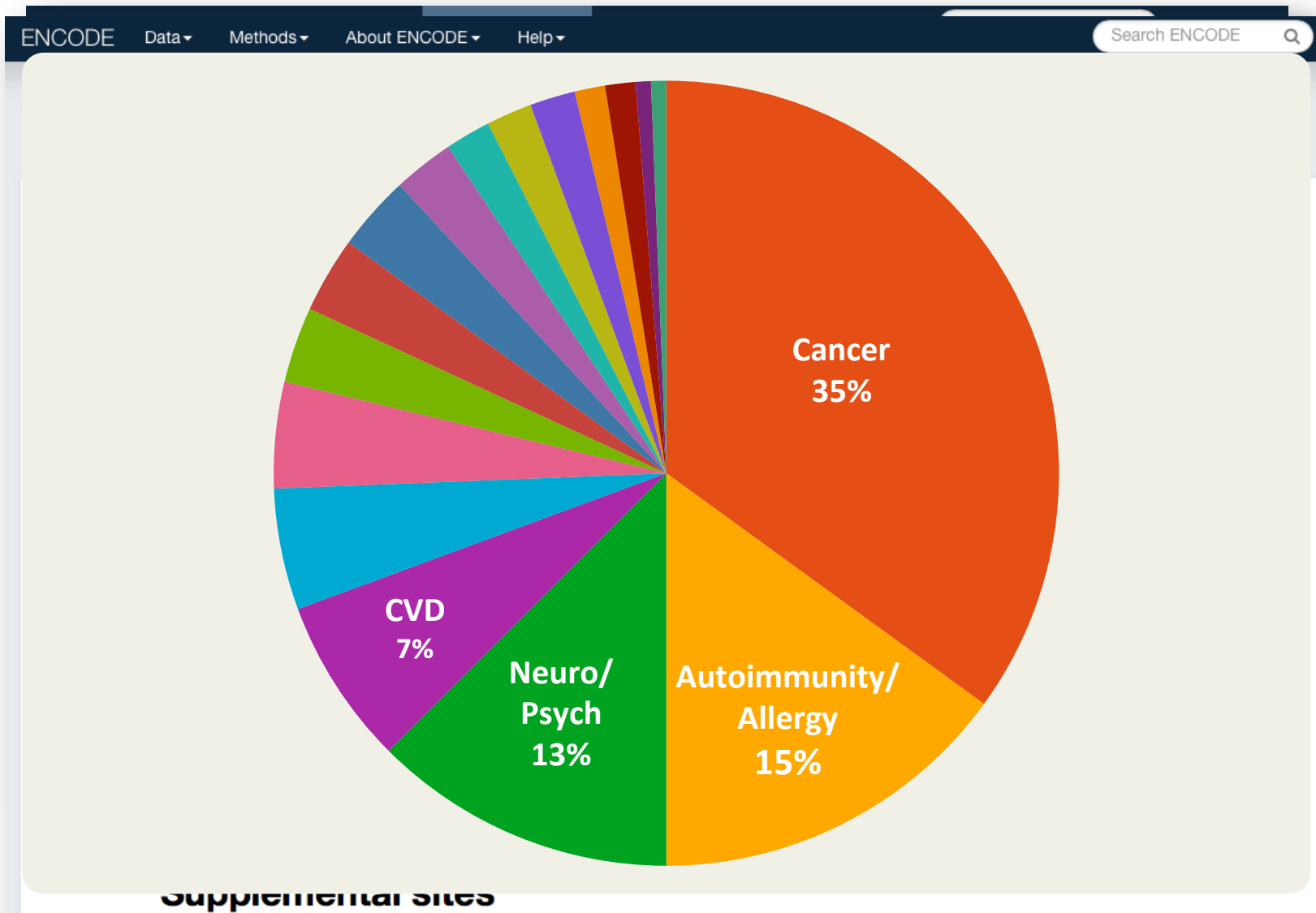
Genomic annotations

- Distal DNase peaks [[Download](#)]
 - Proximal DNase peaks [[Download](#)]
 - Distal H3K27ac annotations (cell type specific) [[Download](#)]
 - Distal H3K4me1 annotations (cell type specific) [[Download](#)]
 - Distal H3K4me3 annotations (cell type specific) [[Download](#)]
 - Distal H3K9ac annotations (cell type specific) [[Download](#)]
 - Proximal H3K27ac annotations (cell type specific) [[Download](#)]
 - Proximal H3K4me1 annotations (cell type specific) [[Download](#)]
 - Proximal H3K4me3 annotations (cell type specific) [[Download](#)]
 - Proximal H3K9ac annotations (cell type specific) [[Download](#)]
 - Distal TF binding sites [[Download](#)]
 - Proximal TF binding sites [[Download](#)]
- Gene expression matrix over ~60 cell types with genes annotated by GENCODE 19 [[Download data](#) | [Download methods](#)]
 - Transcription start site (TSS) lists [[View README](#)]
 - GENCODE v19 TSS [[Download](#)]
 - GENCODE v19 TSS stratified by strict Fantom5 CAGE clusters [[Download](#)]
 - GENCODE v19 TSS stratified by robust Fantom5 CAGE clusters [[Download](#)]
 - GENCODE v19 TSS stratified by permissive Fantom5 CAGE clusters [[Download](#)]

ENCODE Encyclopedia: Coming Soon!



Publications





Goals Of ENCODE

- Catalog all functional elements in the genome
- Develop freely available resource for research community

ENCODE data are being used in the study of human disease and basic biology

ENCODE Accomplishments

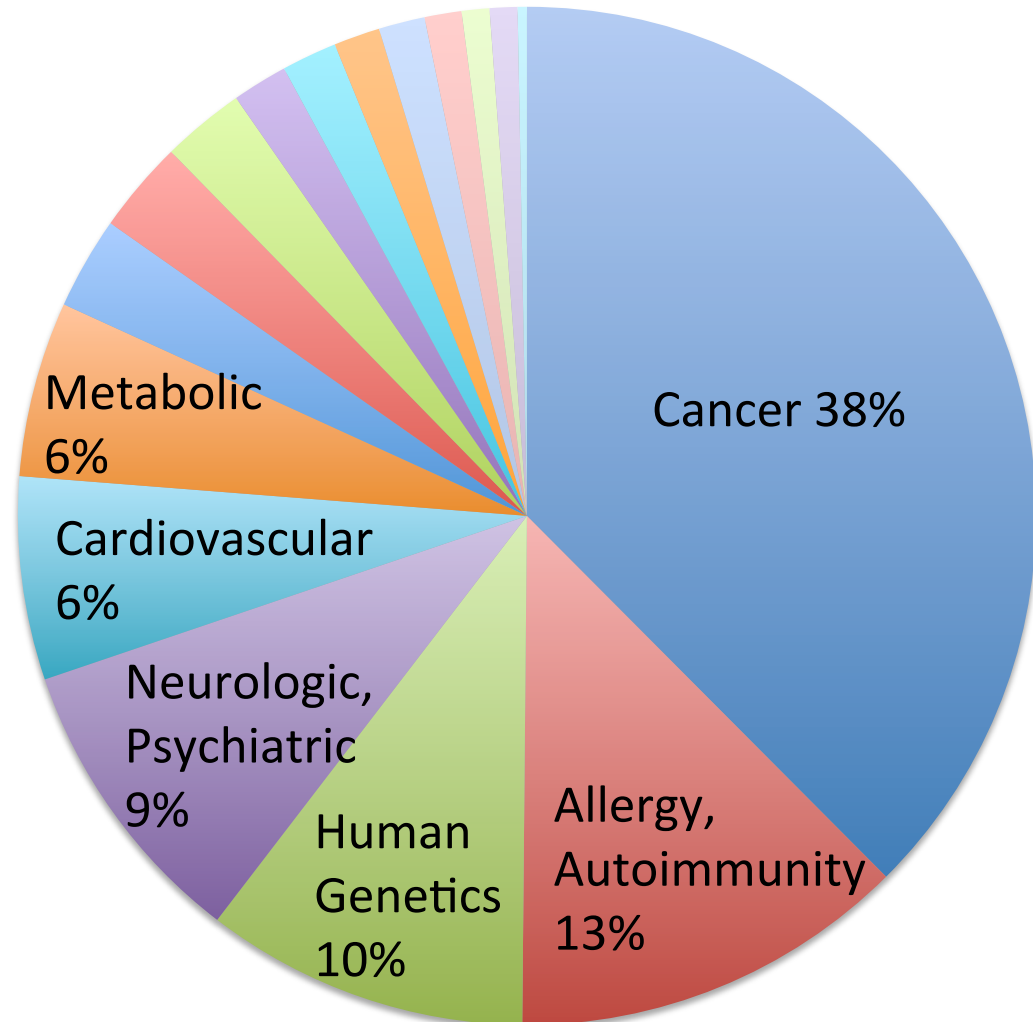
- Sharing 1000s of datasets
 - No embargo
 - High quality
 - Uniformly processed
- Data interoperability
- Sharing software

Publications Using ENCODE Data

Hundreds of Consortium publications

~1500 community publications using ENCODE data:

~340 Human Disease
~500 Basic Biology
~170 Methods/Software Development



ENCODE Consortium



49

Elise Feingold



Peter Good



Mike Pazin



Summary- Accessing ENCODE Resources

- ENCODE portal <https://www.encodeproject.org>
 - Display/download ENCODE and Roadmap Epigenomics data
 - Data Standards
 - Software tools
 - Publications
 - Encyclopedia prototype
- ENCODE Analysis Tools
 - RegulomeDB <http://regulomedb.org/>
 - HaploReg <http://www.broadinstitute.org/mammals/haploreg/>
 - Regulatory Elements Database <http://dnase.genome.duke.edu>
 - RegulomeDB GWAS Database <http://www.regulomedb.org/GWAS/>
- ENCODE Tutorials
 - <http://www.genome.gov/27553900>
 - <https://www.encodeproject.org/tutorials/>
<http://www.ncbi.nlm.nih.gov/pubmed/25762420>
- ENCODE mailing list :
 - <https://mailman.stanford.edu/mailman/listinfo/encode-announce>
- IHEC resources
 - IHEC Home Page <http://ihec-epigenomes.org>
 - IHEC Data Portal <http://epigenomesportal.ca/ihec/>

International Human Epigenome Consortium (IHEC)

- Data Portal: <http://epigenomesportal.ca/ihec/>
- Goal: Coordinate production of 1000 human epigenome maps for cellular states relevant to health and disease <http://ihec-epigenomes.org>
- Can view by consortium, by assay, by cell type
- Data from 7 consortia



A Toxicology User's Guide to the Roadmap Epigenomics and ENCODE Data Resources

Thursday, March 17, 1:00 PM–6:00 PM

Room 205

Hosted by: *Ivan Rusyn, Texas A&M University, College Station, TX; and Lisa Chadwick, NIEHS, Research Triangle Park, NC.*

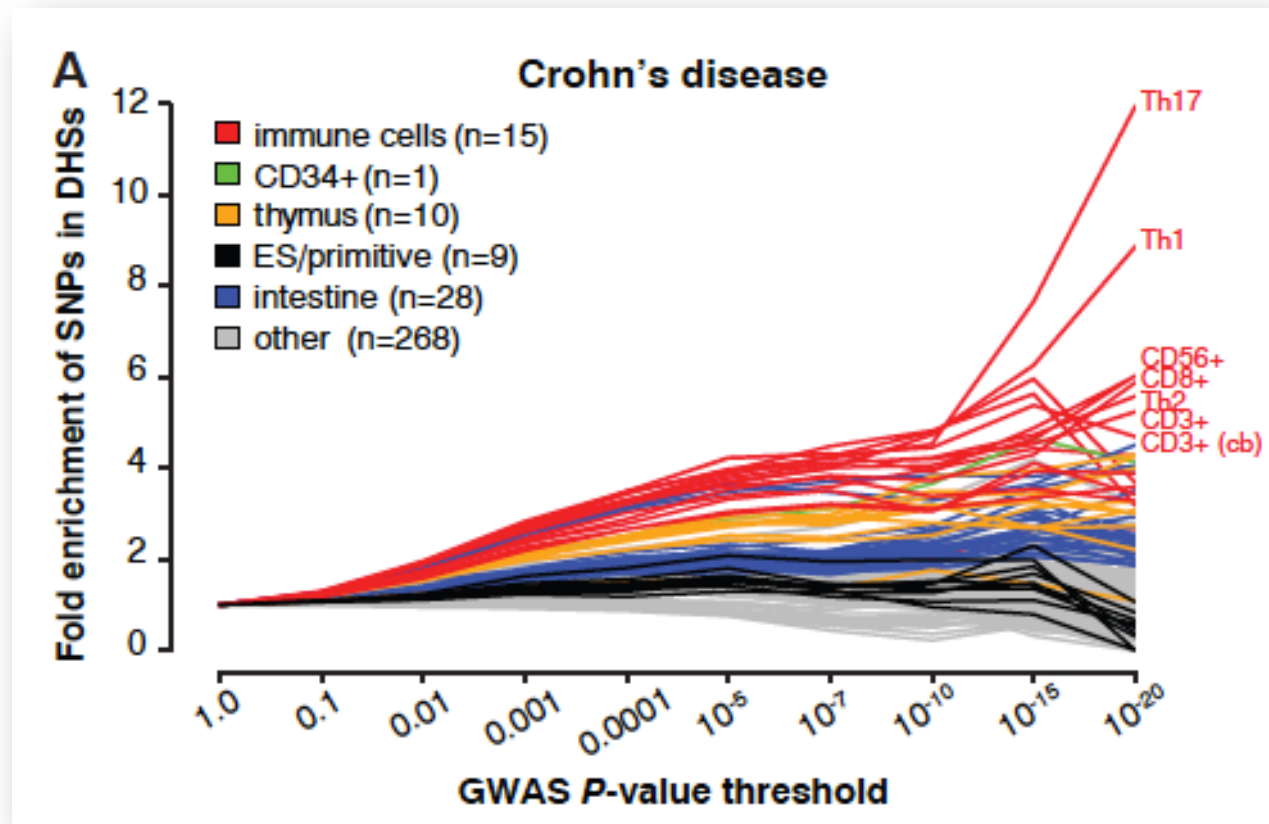
Purpose of the Meeting: Improvements in DNA sequencing technologies have resulted in an exponential increase in the amount of genomic and epigenomic data available. Some of these data have been generated as part of large-scale, focused mapping efforts aimed at understanding how genes are regulated, such as the NIH Roadmap Epigenomics Program, and ENCODE (Encyclopedia of DNA Elements). Efforts such as these can be extremely valuable for hypothesis generation and data mining, but can only be useful if one knows what is available and how to use it. This SOT satellite meeting will provide toxicology researchers with an overview of these two NIH-funded programs, introduce attendees to the informatics tools that have been developed to help navigate these large datasets, and walk through several use cases. The meeting will be of broad interest to researchers interested in learning more about how environmental exposure might impact gene regulation.

Registration: Open registration. No fee to register and attend.

Lectures followed by Q&A and a poster session.

For more information on this Satellite Meeting, contact [Lisa Chadwick](#).

ENCODE And Epigenomics Data Can Be Used To Predict Cell Types



Prediction of Linkage Between Regulatory Elements and Genes



Data from Table S7, Stamatoyannopoulos, Crawford, Nature 489:75, 2012