National Human Genome Research Institute

Using ENCODE Data To Interpret Disease-associated Genetic Variation

Mike Pazin

National Human Genome Research Institute, NIH

ENCODE Users Meeting

June 8, 2016

Welcome

National Human Genome Research Institute

- Objectives
  - We want to tell the community about the ENCODE resource
  - We want to hear community experiences and suggestions

Elise Feingold

Dan Gilchrist

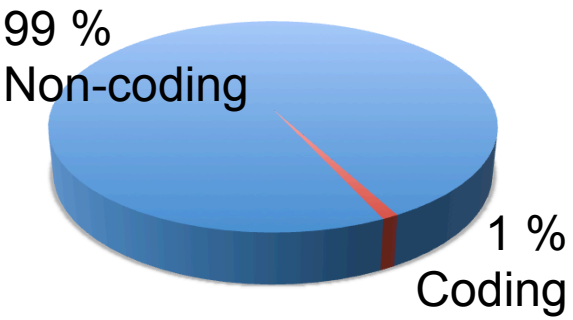
Overview

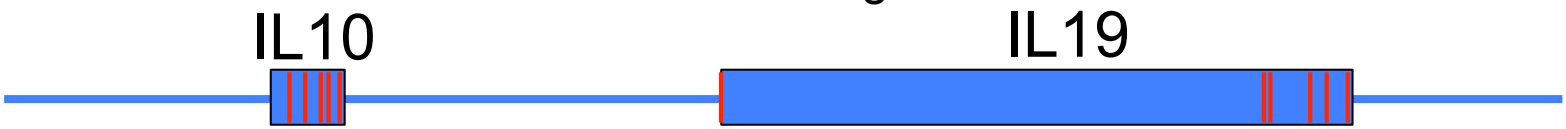
- The ENCODE Resource
- Use of ENCODE to illuminate the role of genetic variation in human disease
- Accessing ENCODE materials

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 some candidate functional elements (but not when or where they act)
- Regulatory regions aren't always in the same order as gene targets

➢ Need unbiased experimental investigation





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.

PMID: 22955828, PMID: 25439723, PMID: 23128226
PMID: 17477822, PMID: 25679767

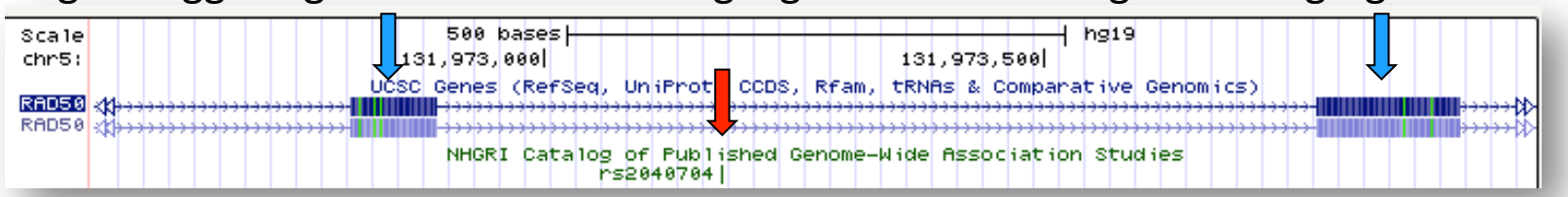
National Human Genome Research Institute

1,500 Letters Of Our 3 Billion Letter Genome

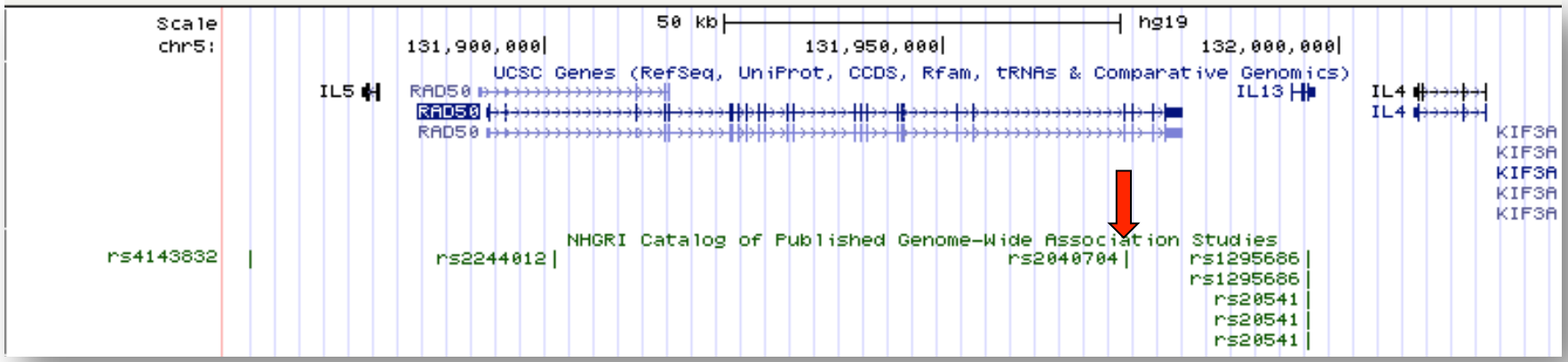
agccaagcagcaaagttttgctgctgtttatttttgtagctcttactatattctacttttaccattgaaaatattgaggaagttatt
tatatttctatttttatatattatatattttatgtattttaatattactattacacataattatttttatatatatgaagtaccaatg
acttccttttccagagcaataatgaaatttcacagtatgaaaatggaagaaatcaataaaattatacgtgacctgtggcgaa
gtacctatcgtggacaaggtgagtaccatggtgtatcacaaatgctctttccaaagccctctccgcagctcttccccttatga
cctctcatcatgccagcattacctccctggacccctttctaagcatgtctttgagattttctaagaattcttatcttggcaacatc
ttgtagcaagaaaatgtaaagttttctgttccagagcctaacaggacttacatatttgactgcagtaggcattatatttagctg
atgacataataggttctgtcatagtgtagatagggataagccaaaatgcaataagaaaaaccatccagaggaaactctttt
ttttttctttttctttttttttttttccagatggagtctcgcacttctctgtcacccgggctggagcgcagtggtgcaatcttggctca
ctgcaacctccacctcctgggttcaggtgattctcccacctcagcctcccgagtagtagctggaattacaggtgcgcgctccc
acacctggctaatttttgtattcttagtagagatggggtttcaccatgttggccaggctggtctcaaactcctgccctcaggtg
atctgcccaccttggcctcccagtgttgggtttacaggcgtgagccaccgcgcctggcctggaggaaactcttaacagggaa
actaagaaagagttgaggctgaggaactggggcatctgggttgcttctggccagaccaccaggctcttgaatcctcccagc
cagagaaagagtttccacaccagccattgttttcctctggtaatgtcagcctcatctgttgttcctaggcttacttgatatgtttg
taaatgacaaaaggctacagagcataggttcctctaaaatattcttcttcctgtgtcagatattgaatacatagaaatacggt
ctgatgccgatgaaaatgtatcagcttctgataaaaggcggaattataactaccgagtggtgatgctgaagggagacacag
ccttggatatgcgaggacgatgcagtgctggacaaaaggcaggtatctcaaaagcctggggagccaactcacccaagtaa
ctgaaagagagaaacaaacatcagtgcagtggaagcacccaaggctacacctgaatggtgggaagctctttgctgctata
taaaatgaatcaggctcagctactattattacactctcctgaagctaaccaacatttcctgcaacattatgtagactt

Maps And Annotation Help Us To Understand The Sequence

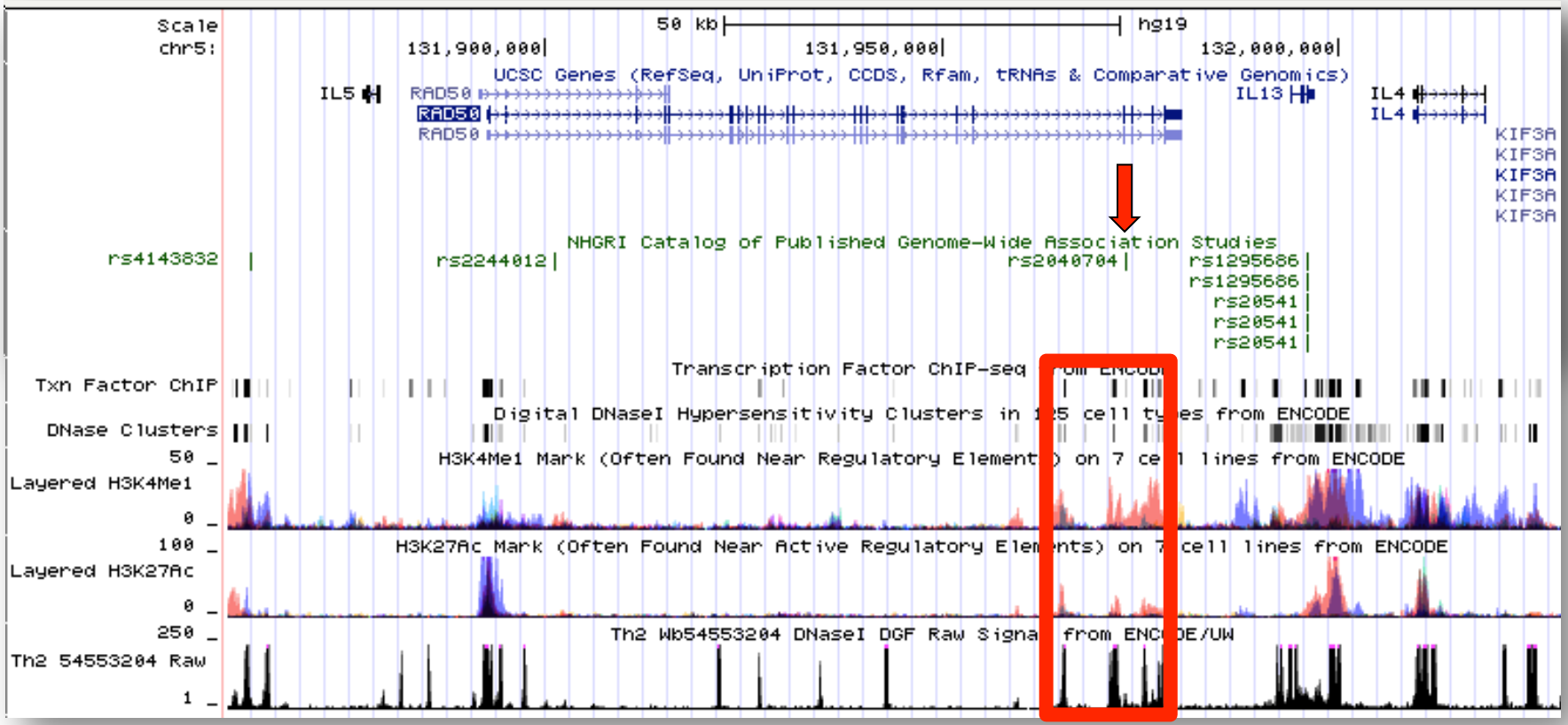
agccaagcagcaaagttttgctgctgtttatttttgtagctcttactatattctacttttaccattgaaaatattgaggaagttatt
tatatttctatttttatatattatatattttatgtattttaatattactattacacataattatttttatatatatgaagtaccaatg
acttccttttccagagcaataatgaaatttcacagtatgaaaatggaagaaatcaataaaattatacgtgacctgtggcgaa
gtacctatcgtggacaaggtgagtaccatggtgtatcacaaatgctctttccaaagccctctccgcagctcttccccttatga
cctctcatcatgccagcattacctccctggacccctttctaagcatgtctttgagattttctaagaattcttatcttggcaacatc
ttgtagcaagaaaatgtaaagttttctgttccagagcctaacaggacttacatatttgactgcagtaggcattatatttagctg
atgacataataggttctgtcatagtgtagatagggataagccaaaatgcaataagaaaaaccatccagaggaaactctttt
ttttttctttttctttttttttttttccagatggagtctcgcacttctctgtcacccgggctggagcgcagtggtgcaatcttggctca
ctgcaacctccacctcctgggttcaggtgattctcccacctcagcctcccgagtagtagctggaattacaggtgcgcgctccc
acacctggctaatttttgtattcttagtagagatggggtttcaccatgttggccaggctggtctcaaactcctgccctcaggtg
atctgcccaccttggcctcccagtgttgggtttacaggcgtgagccaccgcgcctggcctggaggaaactcttaacagggaa
actaagaaagagttgaggctgaggaactggggcatctgggttgcttctggccagaccaccaggctcttgaatcctcccagc
cagagaaagagtttccacaccagccattgttttcctctggtaatgtcagcctcatctgttgttcctaggcttacttgatatgtttg
taaatgacaaaaggctacagagcataggttcctctaaaatattcttcttcctgtgtcagatattgaatacatagaaatacggt
ctgatgccgatgaaaatgtatcagcttctgataaaaggcggaattataactaccgagtggtgatgctgaagggagacacag
ccttggatatgcgaggacgatgcagtgctggacaaaaggcaggtatctcaaaagcctggggagccaactcacccaagtaa
ctgaaagagagaaacaaacatcagtgcagtggaagcacccaaggctacacctgaatggtgggaagctctttgctgctata
taaaatgaatcaggctcagctactattattacactctcctgaagctaaccaacatttcctgcaacattatgtagactt



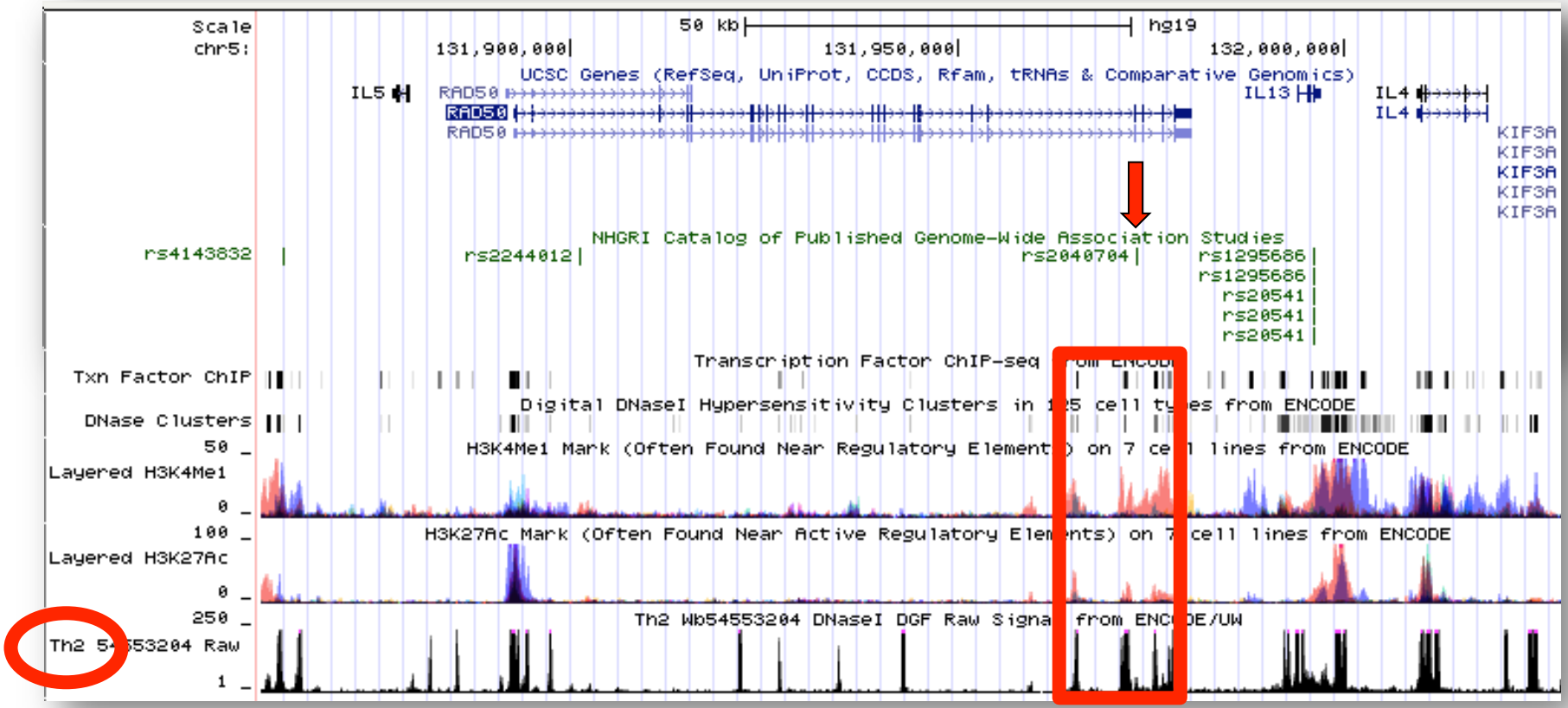
Richer Maps Provide More Information



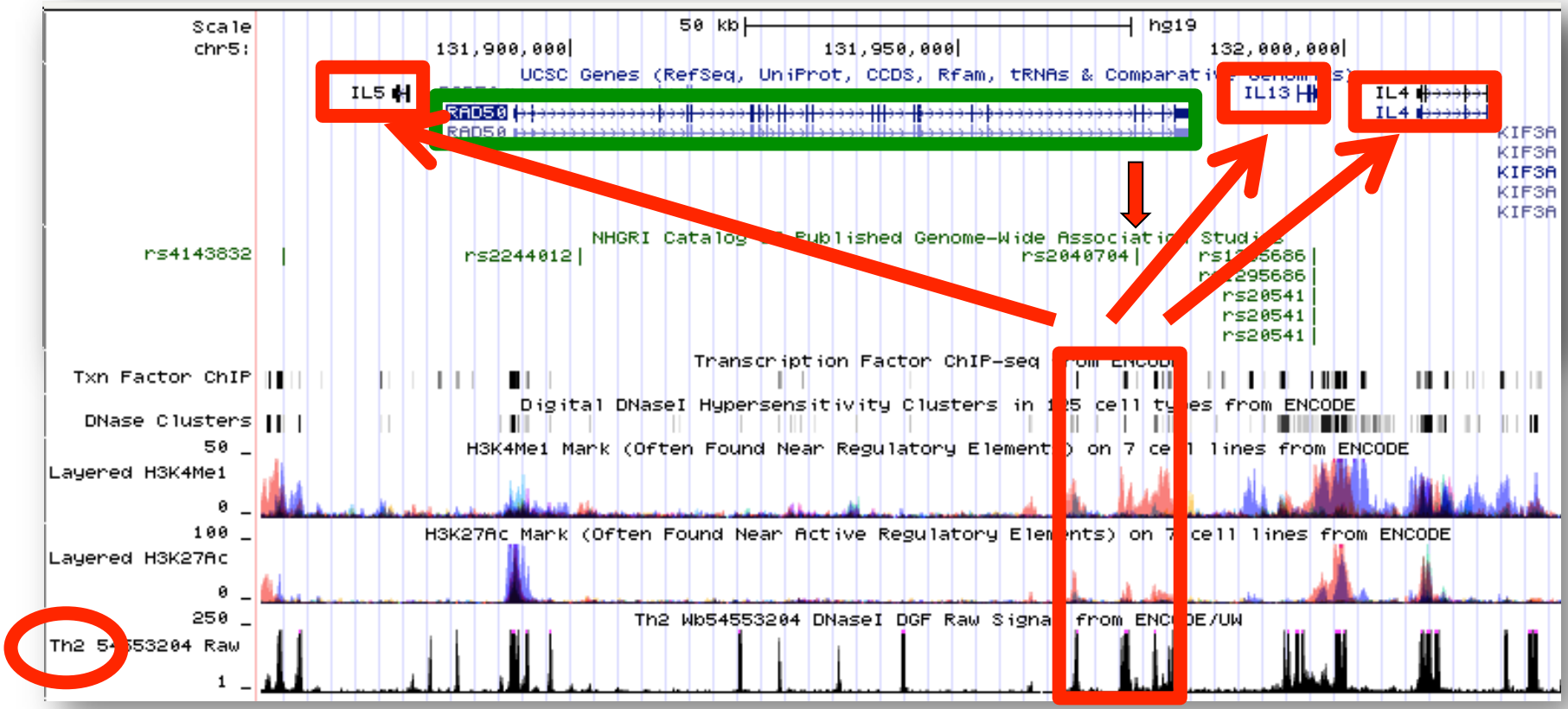
Richer Maps Provide More Information



Richer Maps Provide More Information



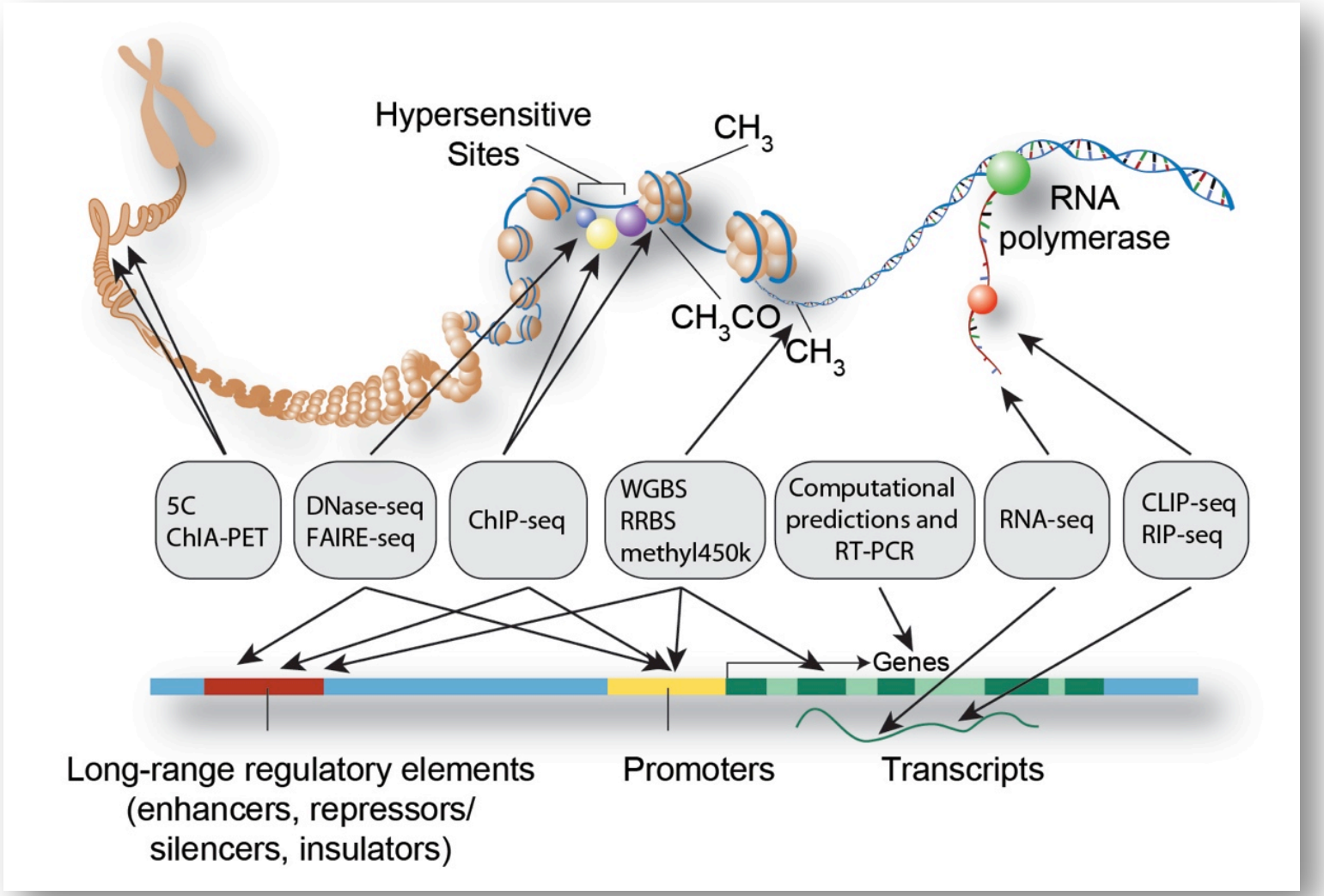
Richer Maps Provide More Information



ENCODE: Encyclopedia Of DNA Elements

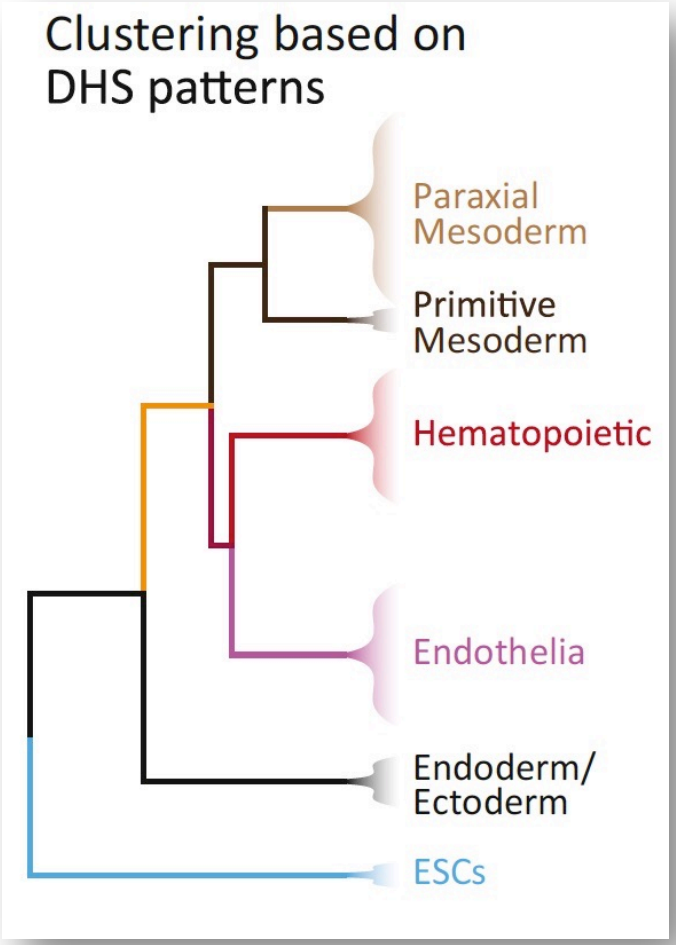
- Identify all candidate functional elements in the genome
- Make resource freely available to community
  - genetic basis of disease
  - gene regulation

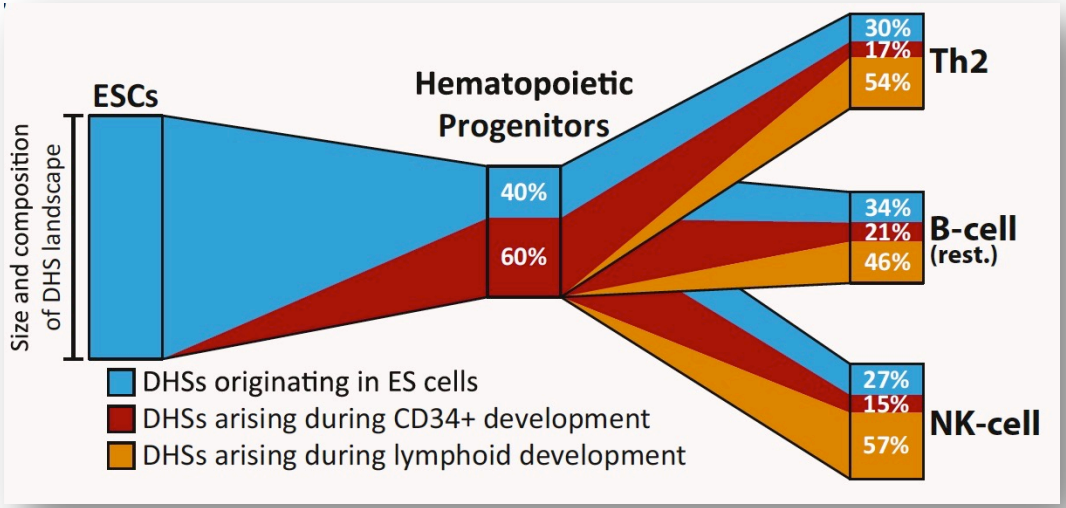
ENCODE Data Types



Modified from PLoS Biol 9:e1001046, 2011
Science 306:636, 2004

ENCODE Data Are Cell-Type Specific





Stamatoyannopoulos, Cell 154:888, 2013

ENCODE Accomplishments

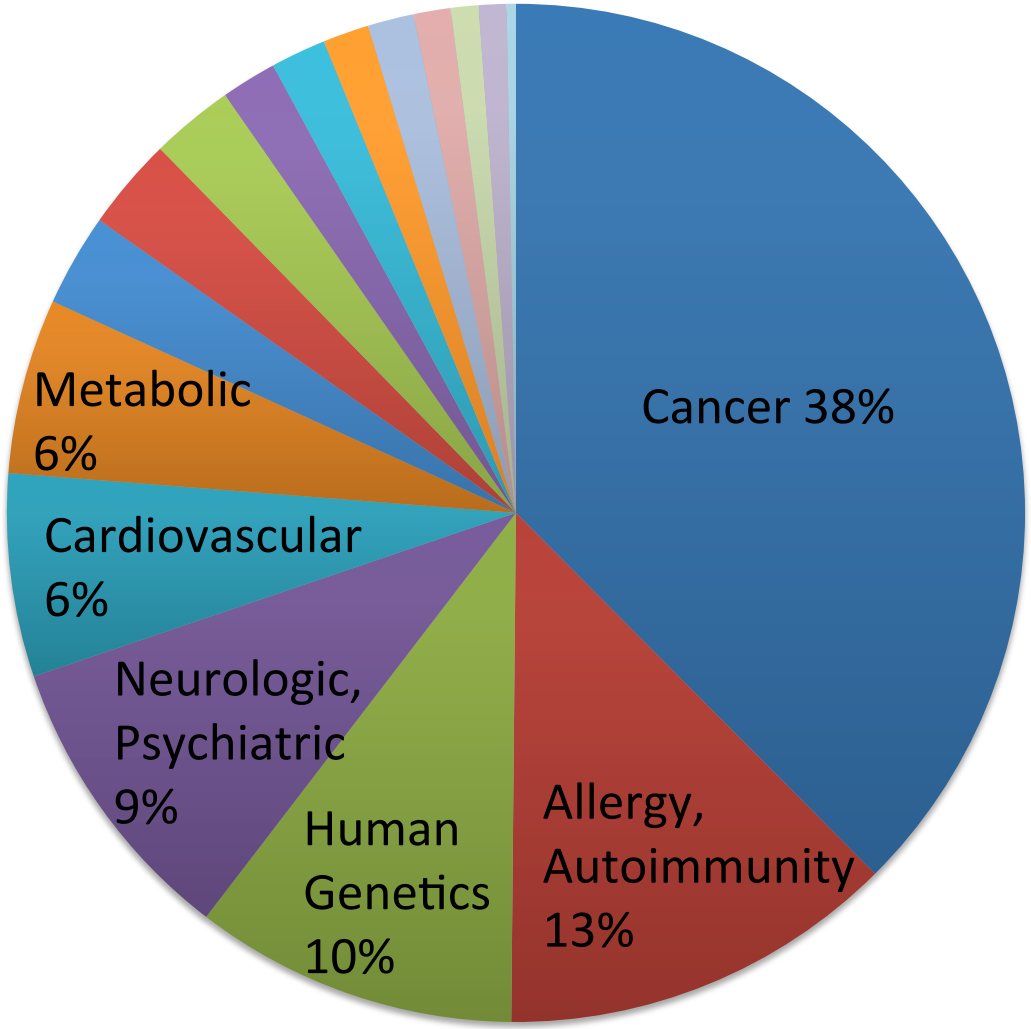
- Sharing 1000s of datasets
  - No embargo
  - Unrestricted access
  - High quality
  - Uniformly processed
- Sharing software
- Data interoperability

Publications Using ENCODE Data

Hundreds of Consortium publications

~1500 community publications using ENCODE data:

- ~675 Human Disease
- ~600 Basic Biology
- ~225 Methods/Software Development



Summary- ENCODE Resource

- Freely shared catalog of genomic data and 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

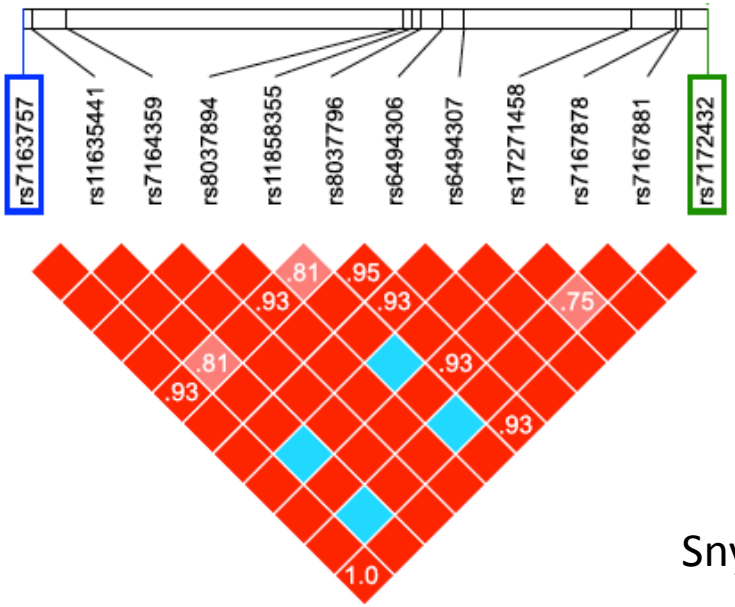
Standard ENCODE Use Cases: Hypothesis Generation

Major use: Hypothesis generation and refinement

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

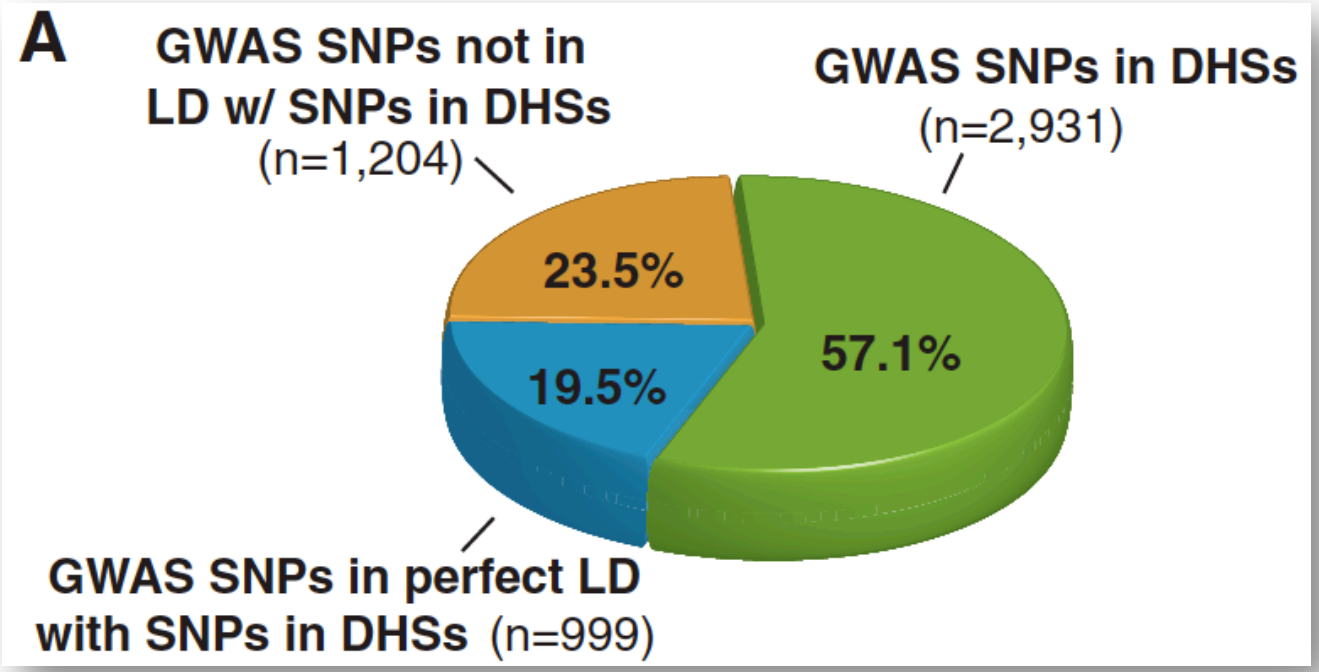
Prediction of Causal Variants

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



Snyder, Genome Research 22:1748, 2012

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



Stamatoyannopoulos, Science 337:1190, 2012

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 r² 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 r² is set to NA, only queried variants will be shown, together in one table.

Query (comma-delimited list of rsIDs OR a single region as chrN:start-end):

1

or, upload a text file (one refSNP ID per line): Choose File no file selected

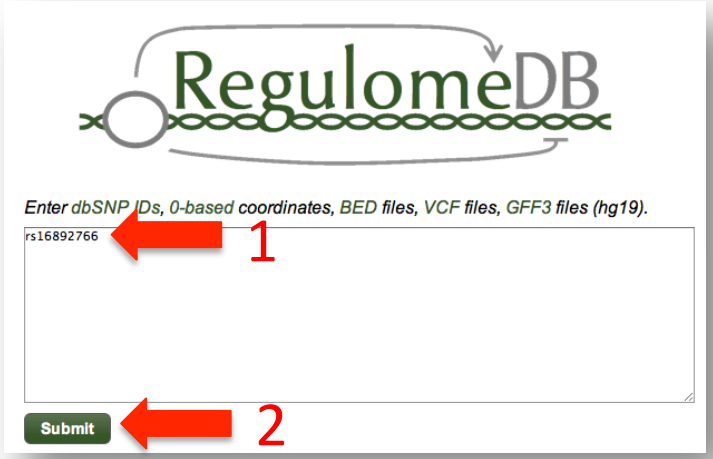
Query SNP: rs16892766 and variants with r^2 >= 0.8

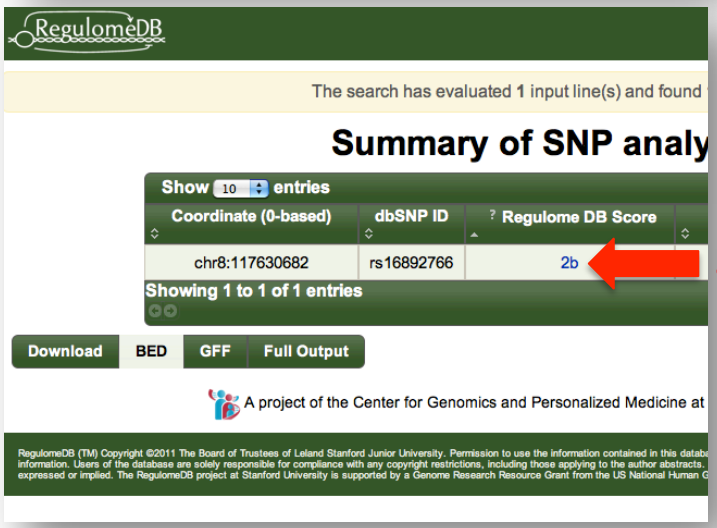
chr	pos (hg38)	LD (r^2)	LD (D')	variant	Ref	Alt	AFR freq	AMR freq	ASN freq	EUR freq	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse	Proteins bound	Motifs changed	NHGRI/EBI GWAS hits	GRASP QTL hits	Selected eQTL hits	GENCODE genes	dbSNP func annot
8	116618444	1	1	rs16892766	A	C	0.12	0.08	0.00	0.09		FAT	STRM, LNG, GI	7 tissues	FOXA1,GR	Rhox11	2 hits	1 hit	3 hits	24kb 3' of EIF3H	
8	116618773	0.97	1	rs200235517	CG	C	0.47	0.10	0.04	0.09			7 tissues			Irf				23kb 3' of EIF3H	
8	116618774	0.97	1	rs58147231	GA	G	0.47	0.10	0.04	0.09			7 tissues			Irf			1 hit	23kb 3' of EIF3H	
8	116623363	0.89	0.97	rs16888589	A	G	0.12	0.08	0.00	0.09		ESDR	8 tissues			Ik-1,STAT			3 hits	19kb 3' of EIF3H	

www.broadinstitute.org/mammals/haploreg/

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

ENCODE Data From RegulomeDB





Protein Binding

Filter:

Method	Location	Bound Protein	Cell Type	Additional Info	Reference
ChIP-seq	chr8:117630539..117630739	FOXA1	ECC-1	DMSO_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
DNase-seq	chr8:117630500..117630710	Nha		ENCODE
DNase-seq	chr8:117630500..117630770	Hah		ENCODE
DNase-seq	chr8:117630510..117630704	Aosmc	Serum	ENCODE
DNase-seq	chr8:117630520..117630790	Hvmf		ENCODE
DNase-seq	chr8:117630625..117631002	Htr8		ENCODE
FAIRE	chr8:117630519..117630761	Medullo		ENCODE

Histone modifications

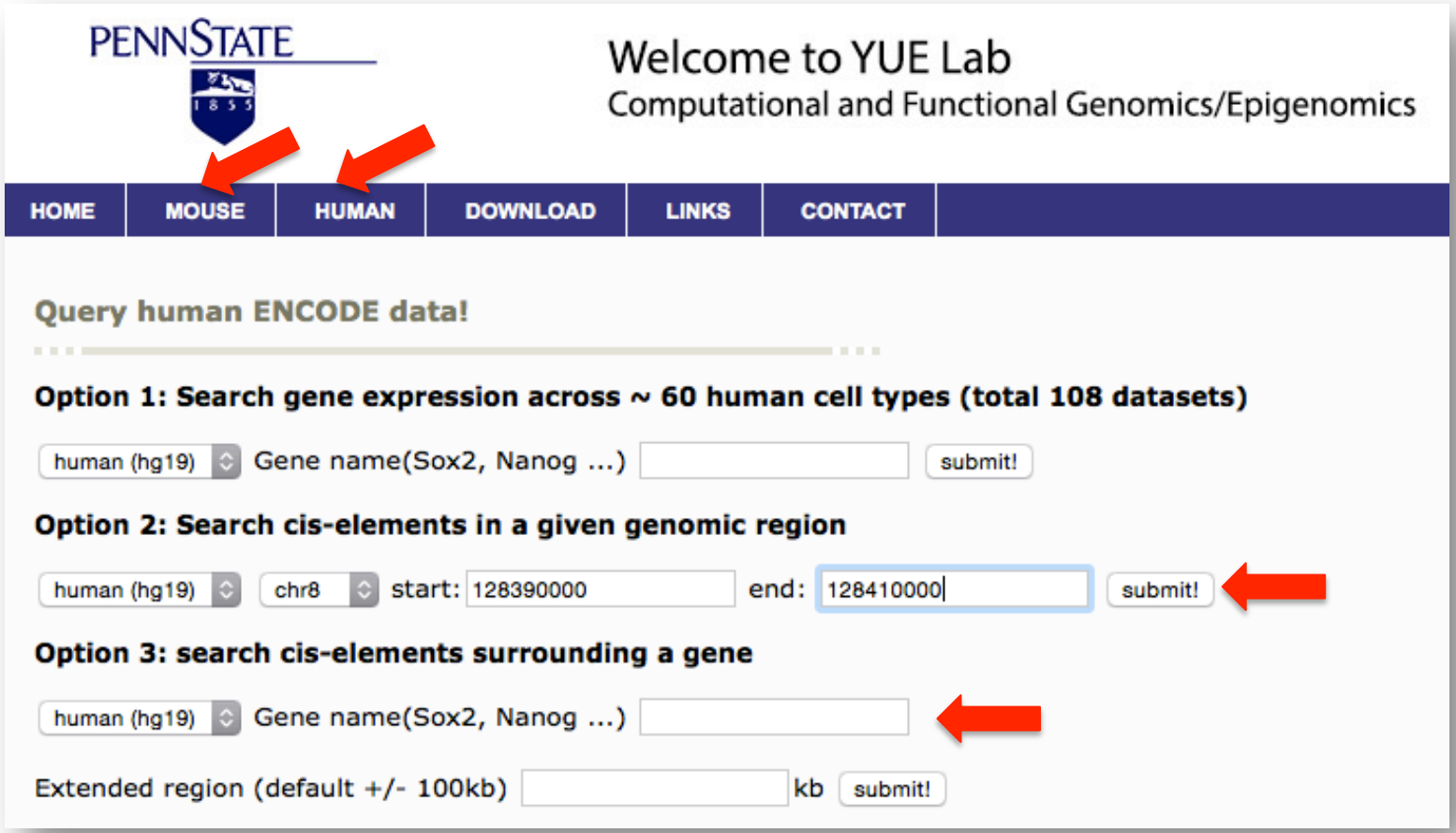
Filter:

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
ChIP-seq	chr8:117409399..118413945	H4k20me1	Hmec		ENCODE
ChIP-seq	chr8:117555446..118475451	H4k20me1	Nhlf		ENCODE
ChIP-seq	chr8:117384499..117650386	H2az	Dnd41		ENCODE

http://regulomedb.org/

Cherry, Snyder, Genome Research 22-1790,2012

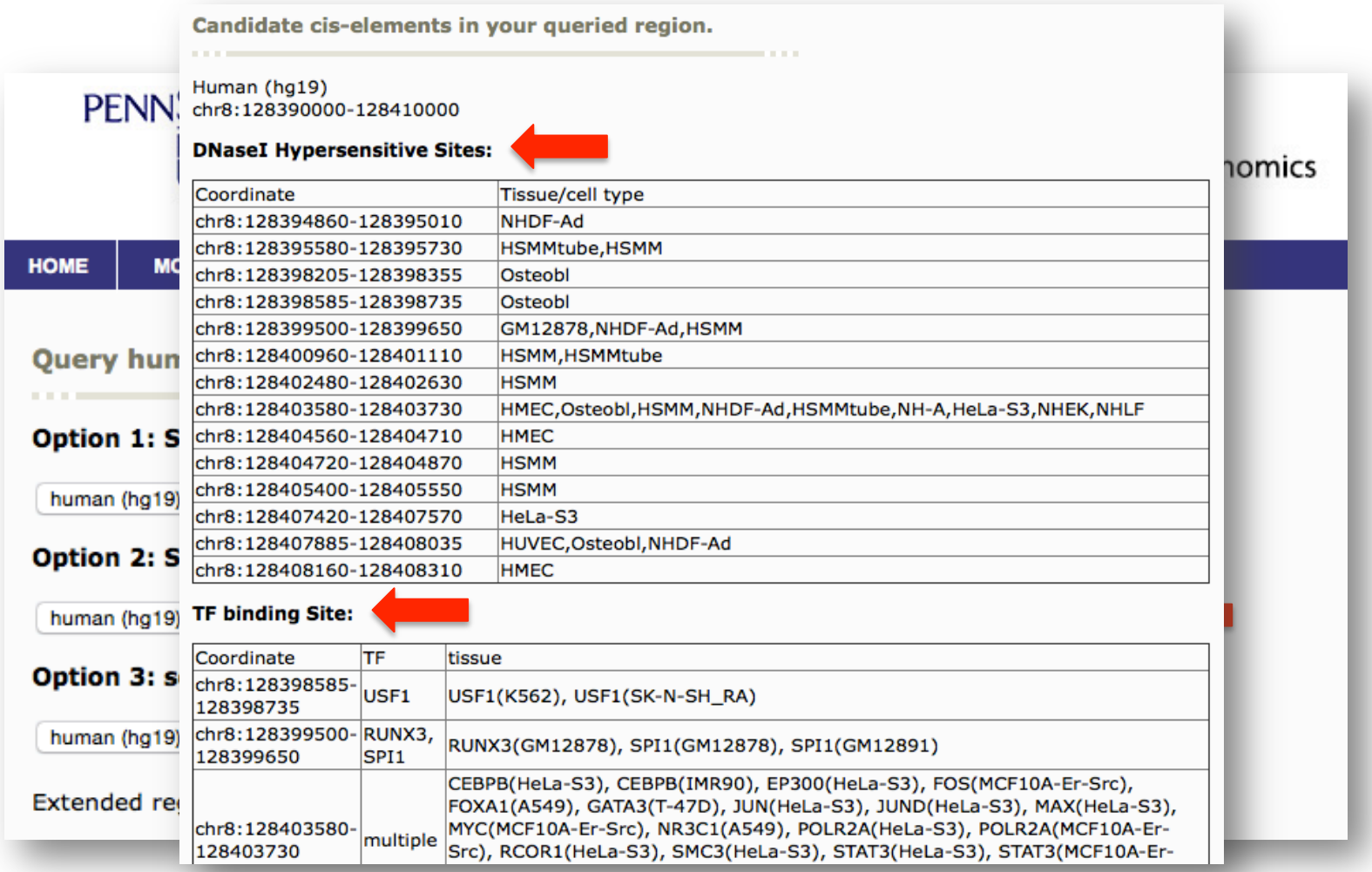
ENCODE cis-element Browser



Workshop Session 3

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

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:

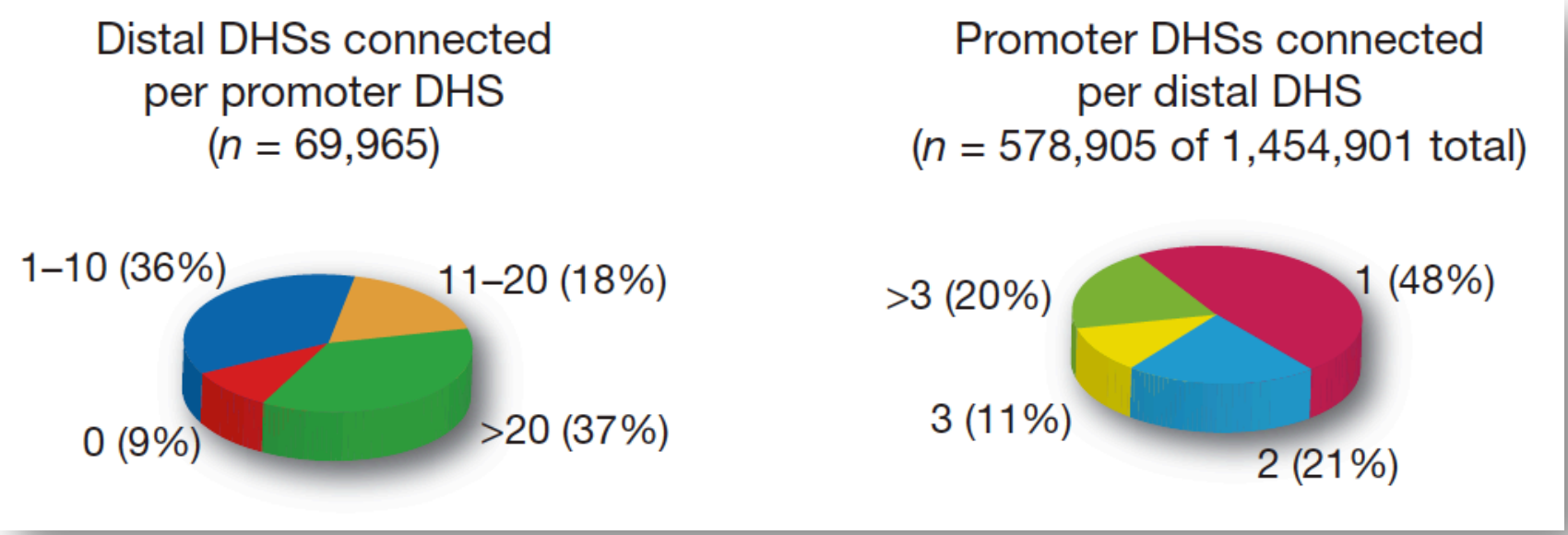
Coordinate	TF	tissue
chr8:128398585-128398735	USF1	USF1(K562), USF1(SK-N-SH_RA)
chr8:128399500-128399650	RUNX3, SPI1	RUNX3(GM12878), SPI1(GM12878), SPI1(GM12891)
chr8:128403580-128403730	multiple	CEBPB(HeLa-S3), CEBPB(IMR90), EP300(HeLa-S3), FOS(MCF10A-Er-Src), FOXA1(A549), GATA3(T-47D), JUN(HeLa-S3), 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-

Workshop Session 3

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

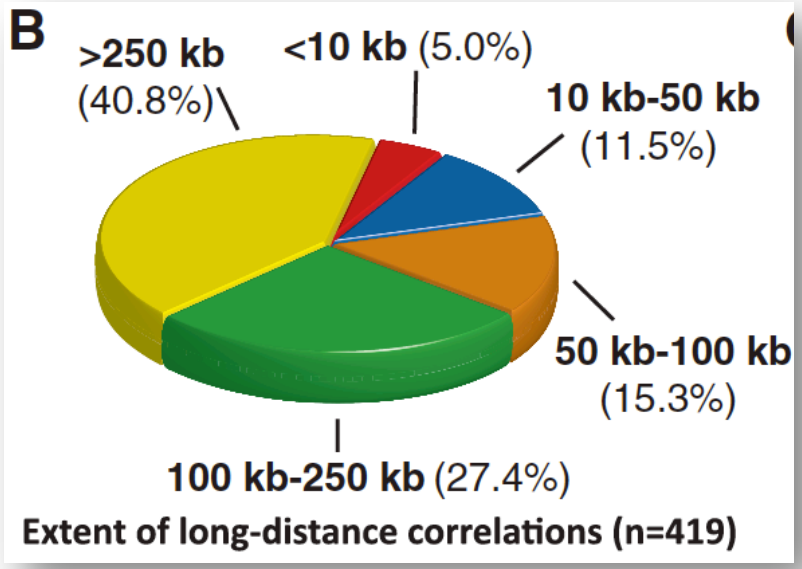
Prediction of Target Genes

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



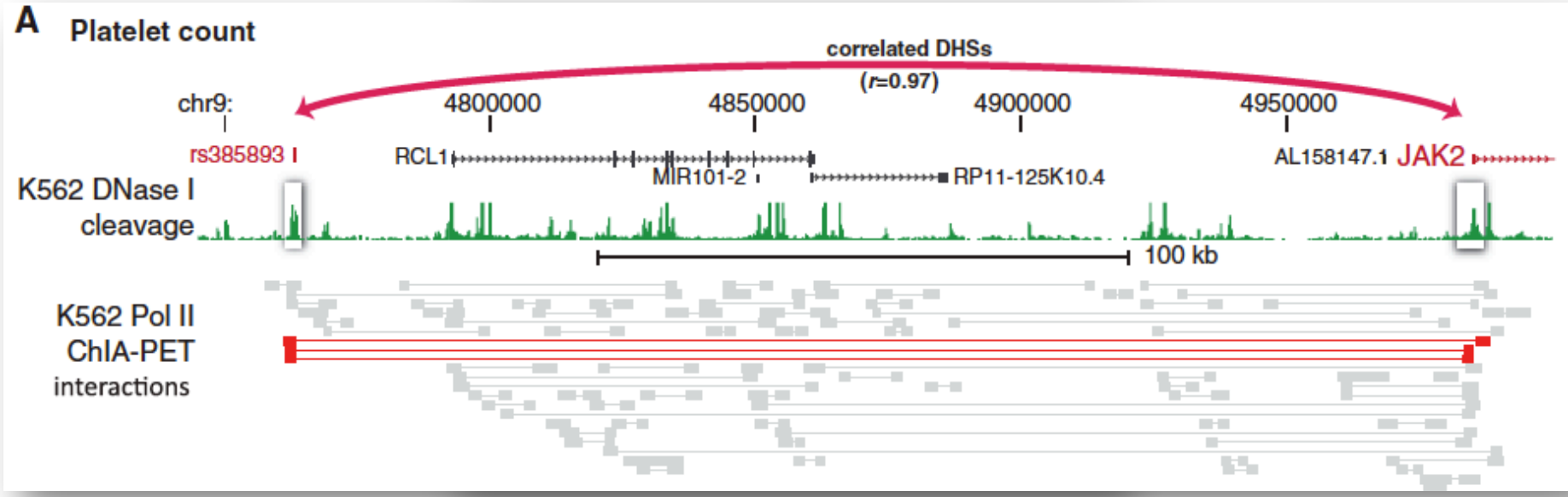
Stamatoyannopoulos, Crawford, Nature 489:75, 2012

Many GWAS Associations Are Predicted To Be Linked To Distal Genes



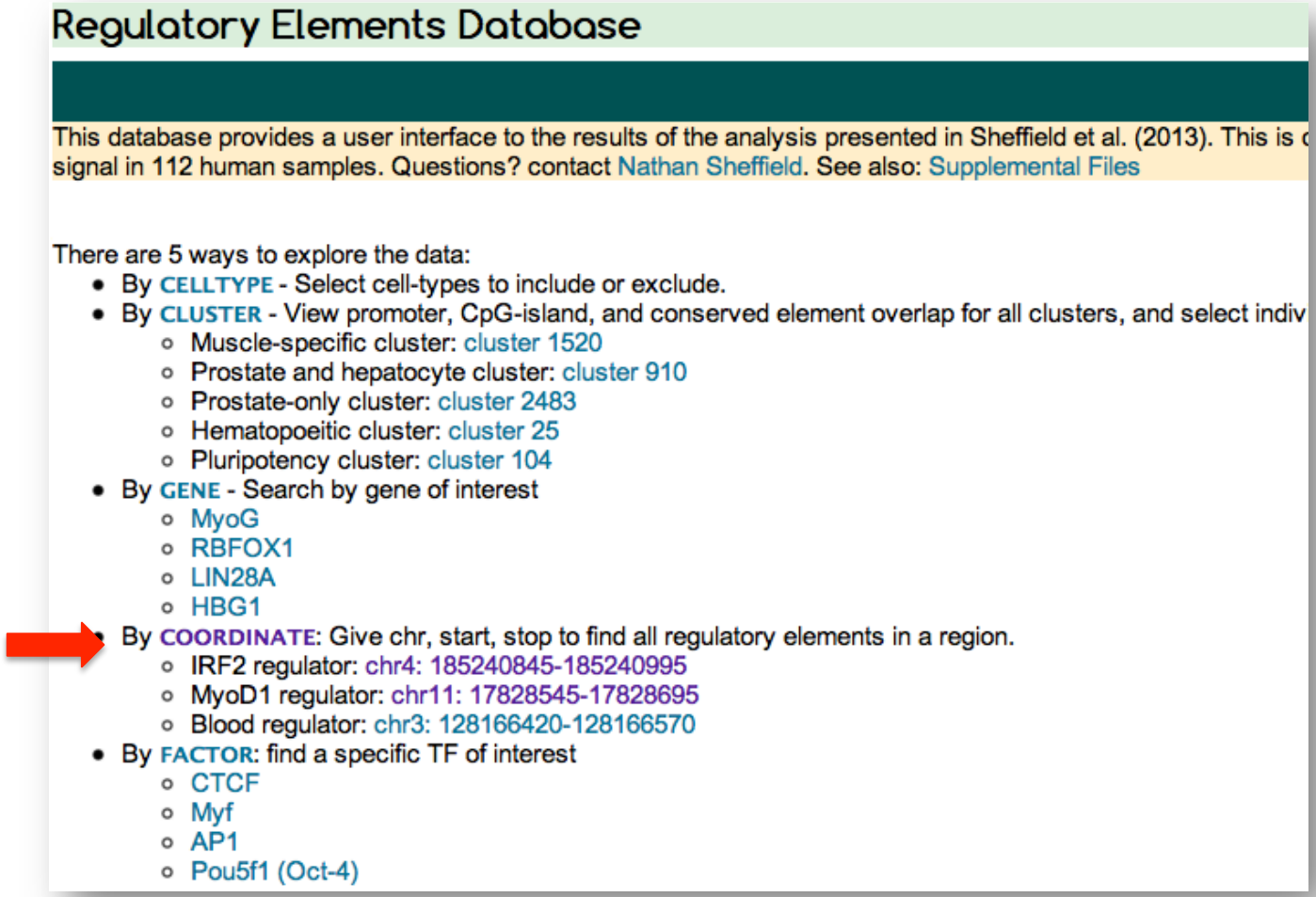
Stamatoyannopoulos, Science 337:1190, 2012

Many GWAS Associations Are Predicted To Be Linked To Distal Genes



Stamatoyannopoulos, Science 337:1190, 2012

Prediction of Linkage Between Regulatory Elements and Genes



Regulatory Elements Database

This database provides a user interface to the results of the analysis presented in Sheffield et al. (2013). This is c
signal in 112 human samples. Questions? contact Nathan Sheffield. See also: Supplemental Files

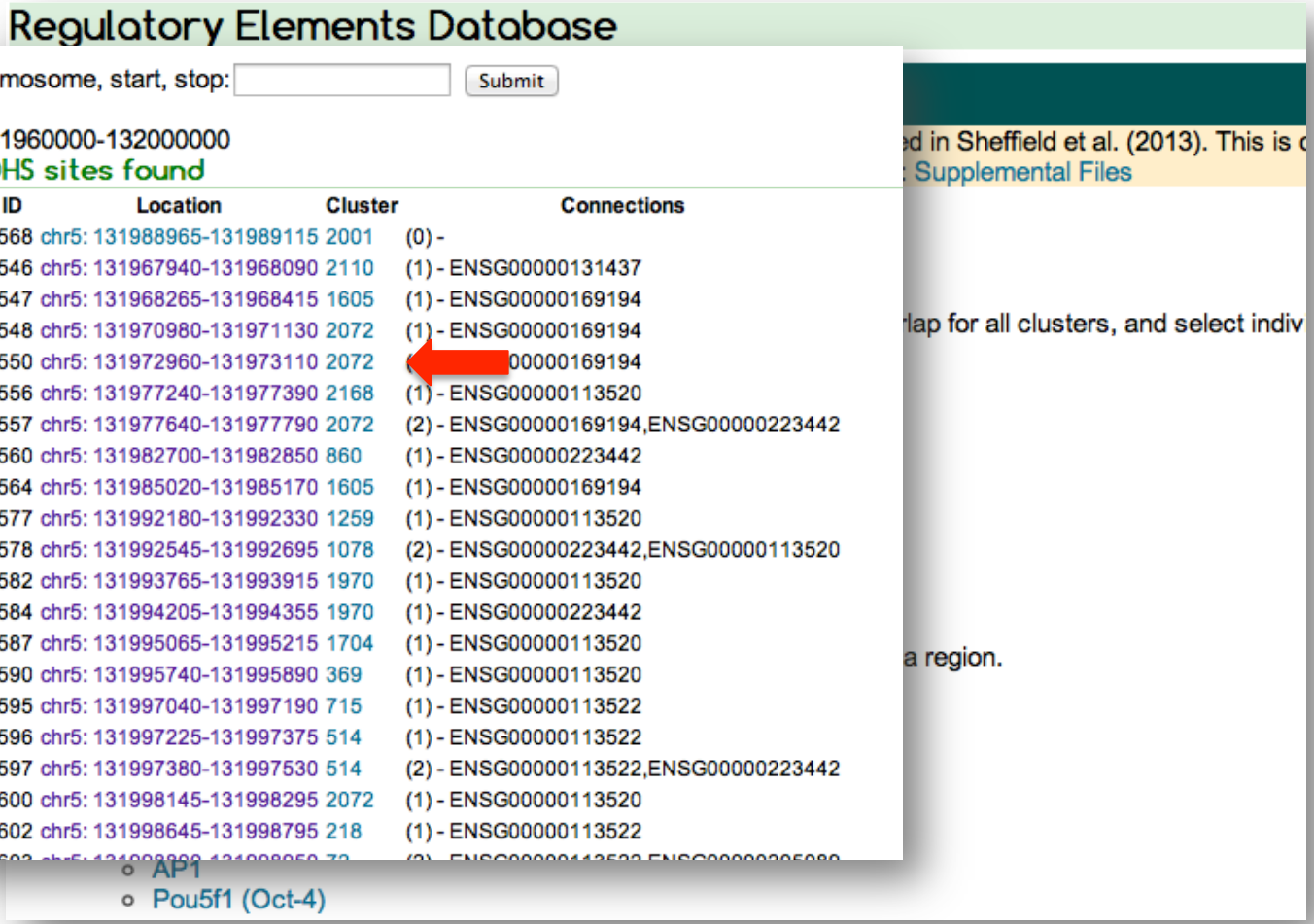
There are 5 ways to explore the data:

- By CELLTYPE - Select cell-types to include or exclude.
- By CLUSTER - View promoter, CpG-island, and conserved element overlap for all clusters, and select indiv
  - Muscle-specific cluster: cluster 1520
  - Prostate and hepatocyte cluster: cluster 910
  - Prostate-only cluster: cluster 2483
  - Hematopoeitic cluster: cluster 25
  - Pluripotency cluster: cluster 104
- By GENE - Search by gene of interest
  - MyoG
  - RBFOX1
  - LIN28A
  - HBG1
- By COORDINATE: Give chr, start, stop to find all regulatory elements in a region.
  - IRF2 regulator: chr4: 185240845-185240995
  - MyoD1 regulator: chr11: 17828545-17828695
  - Blood regulator: chr3: 128166420-128166570
- By FACTOR: find a specific TF of interest
  - CTCF
  - Myf
  - AP1
  - Pou5f1 (Oct-4)

http://dnase.genome.duke.edu

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

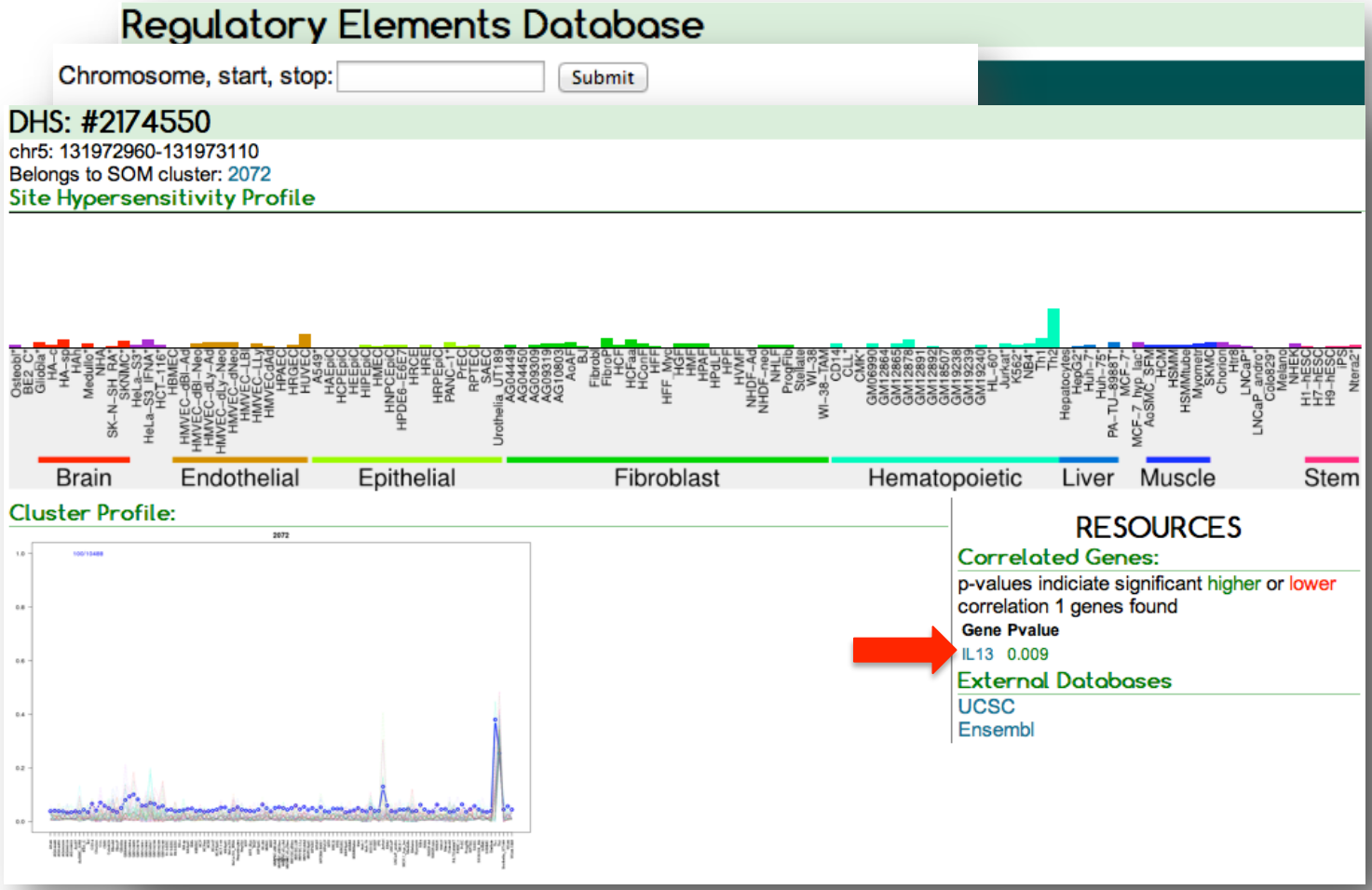
Prediction of Linkage Between Regulatory Elements and Genes



http://dnase.genome.duke.edu

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

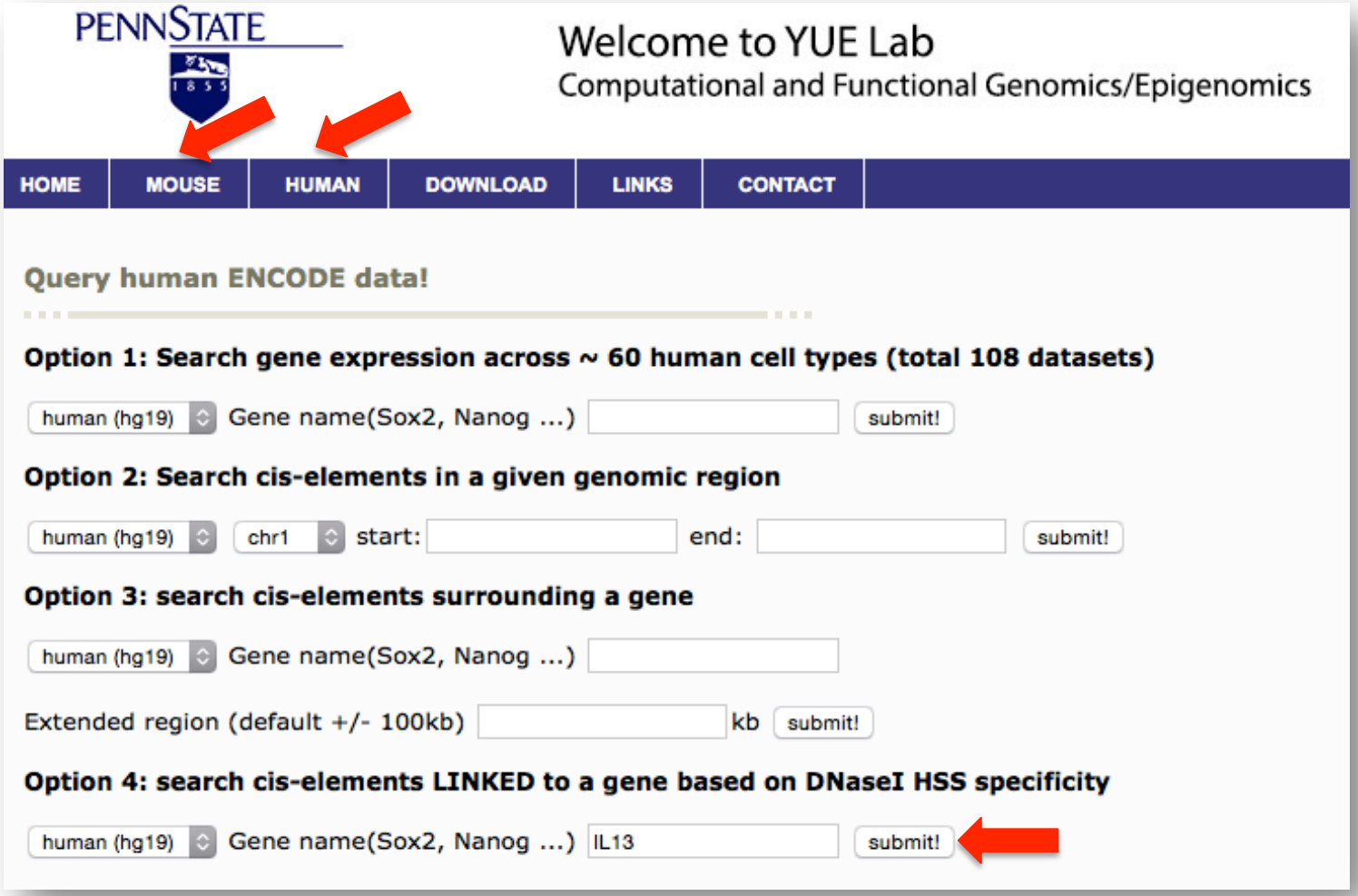
Prediction of Linkage Between Regulatory Elements and Genes



http://dnase.genome.duke.edu

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

ENCODE cis-element Browser



Workshop Session 3

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

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

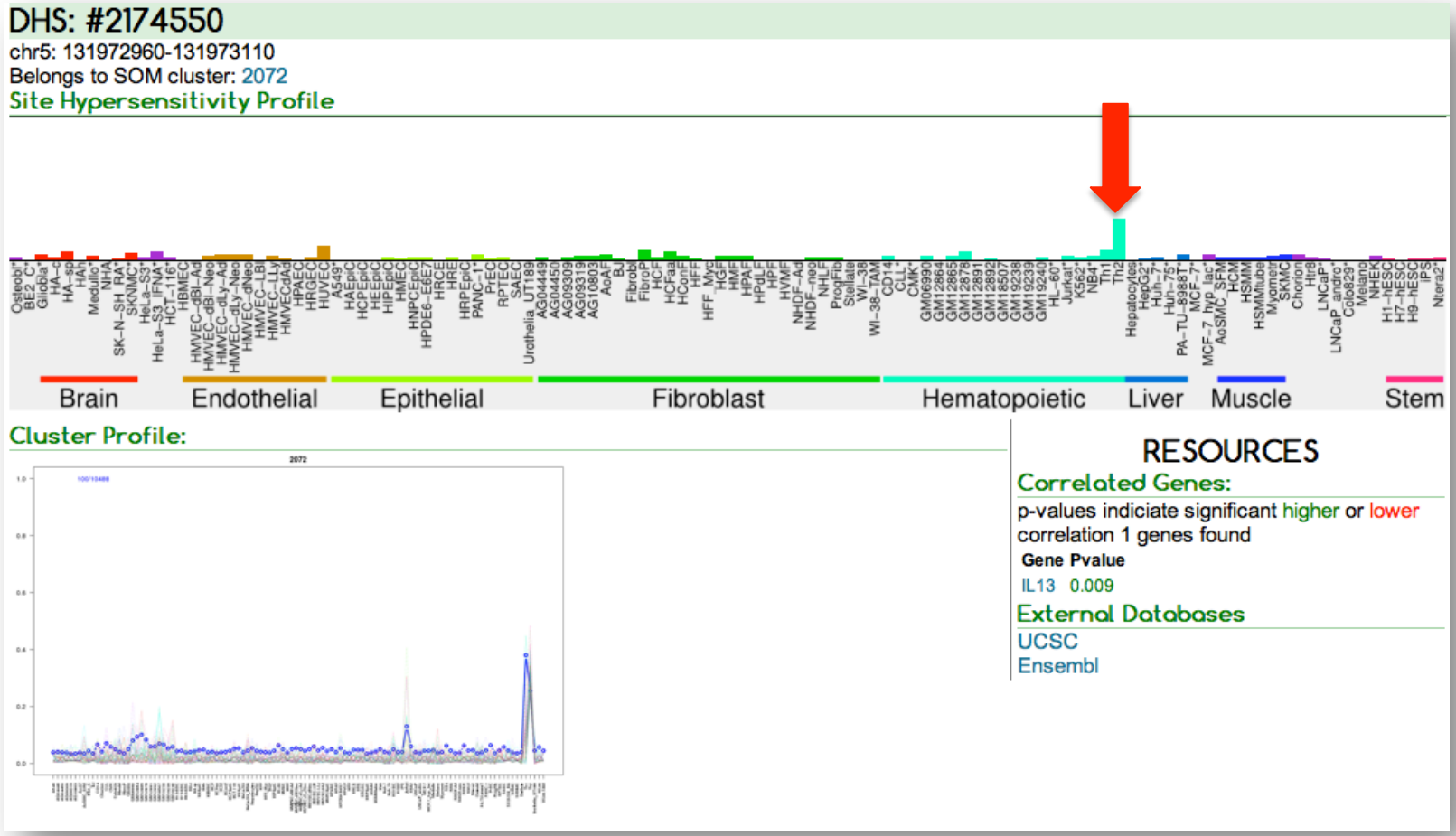
Workshop Session 3

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

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

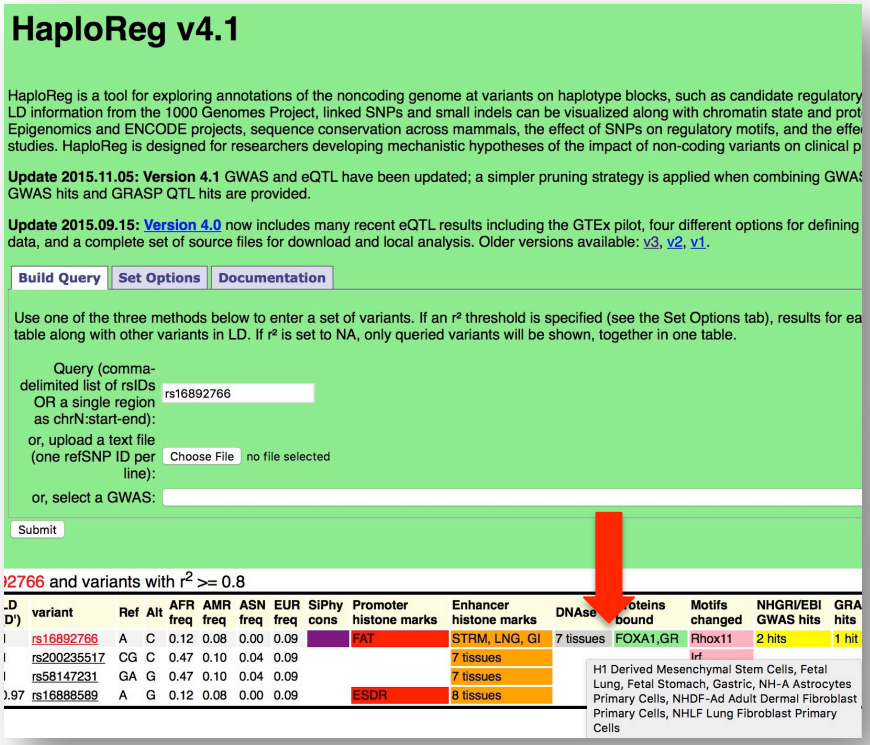
Prediction of Linkage Between Regulatory Elements and Cell Type

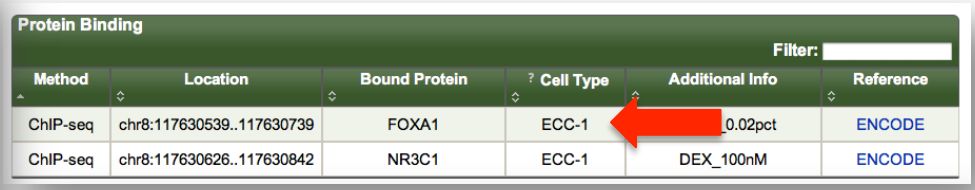


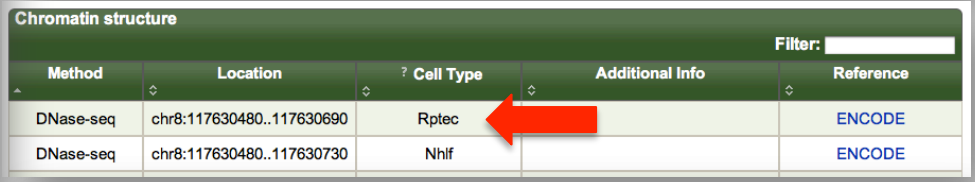
http://dnase.genome.duke.edu

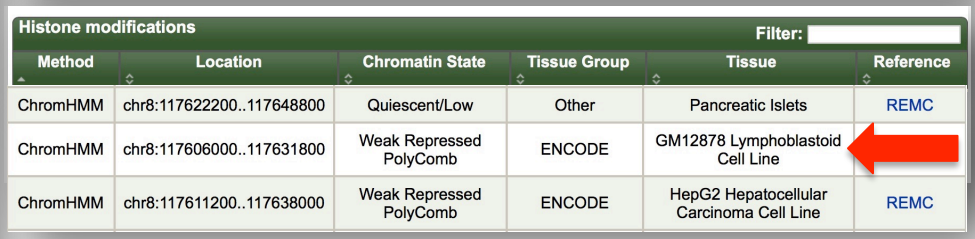
Furey, Crawford, Stamatoyannopoulos, Genome Res. 23:777, 2013

Prediction of Linkage Between Regulatory Elements and Cell Type









Workshop Session 3

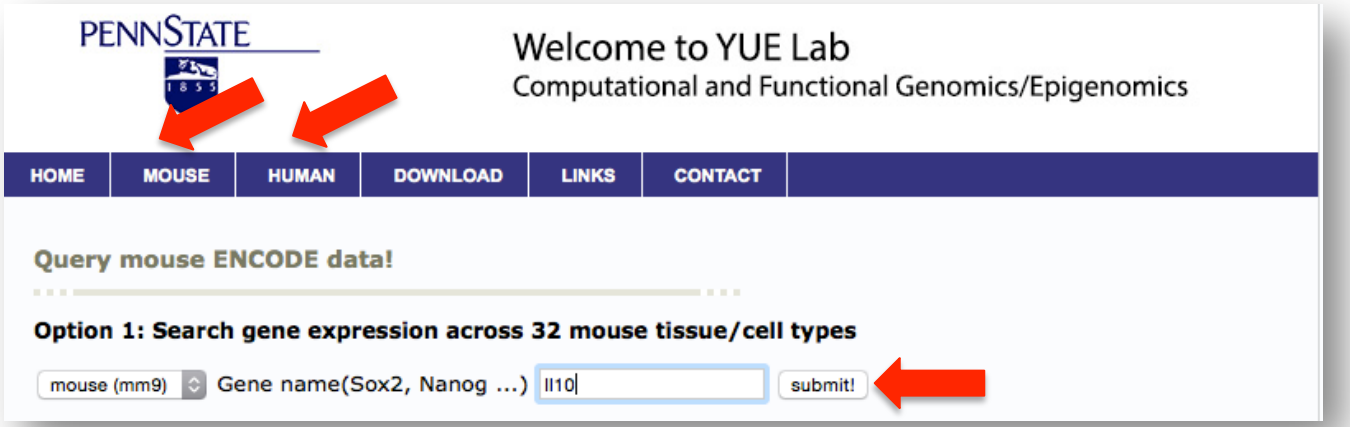
www.broadinstitute.org/mammals/haploreg/

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

http://regulomedb.org/

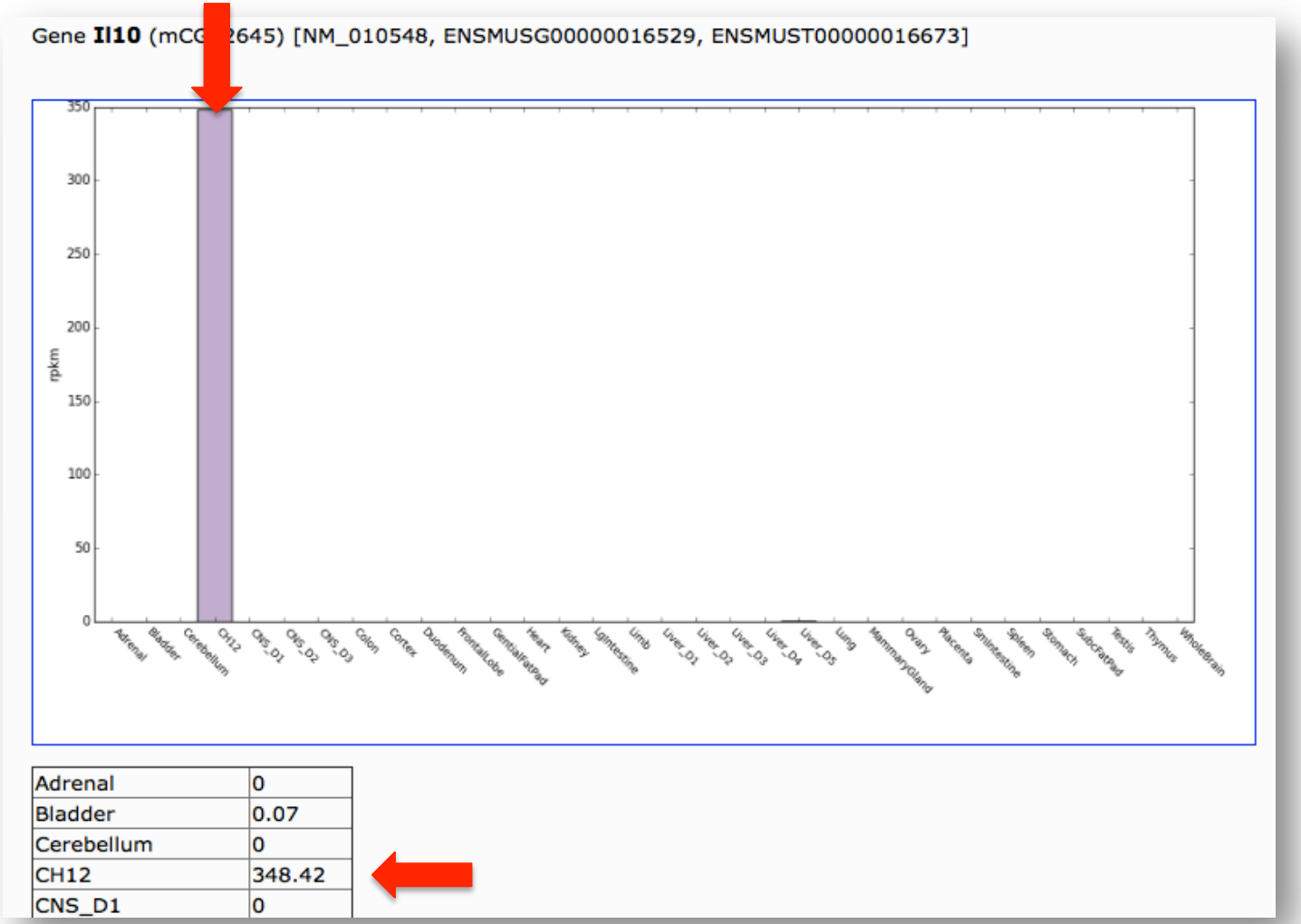
Cherry, Snyder, Genome Research 22-1790,2012

ENCODE cis-element Browser



Workshop Session 3
https://www.encodeproject.org/data/annotations/

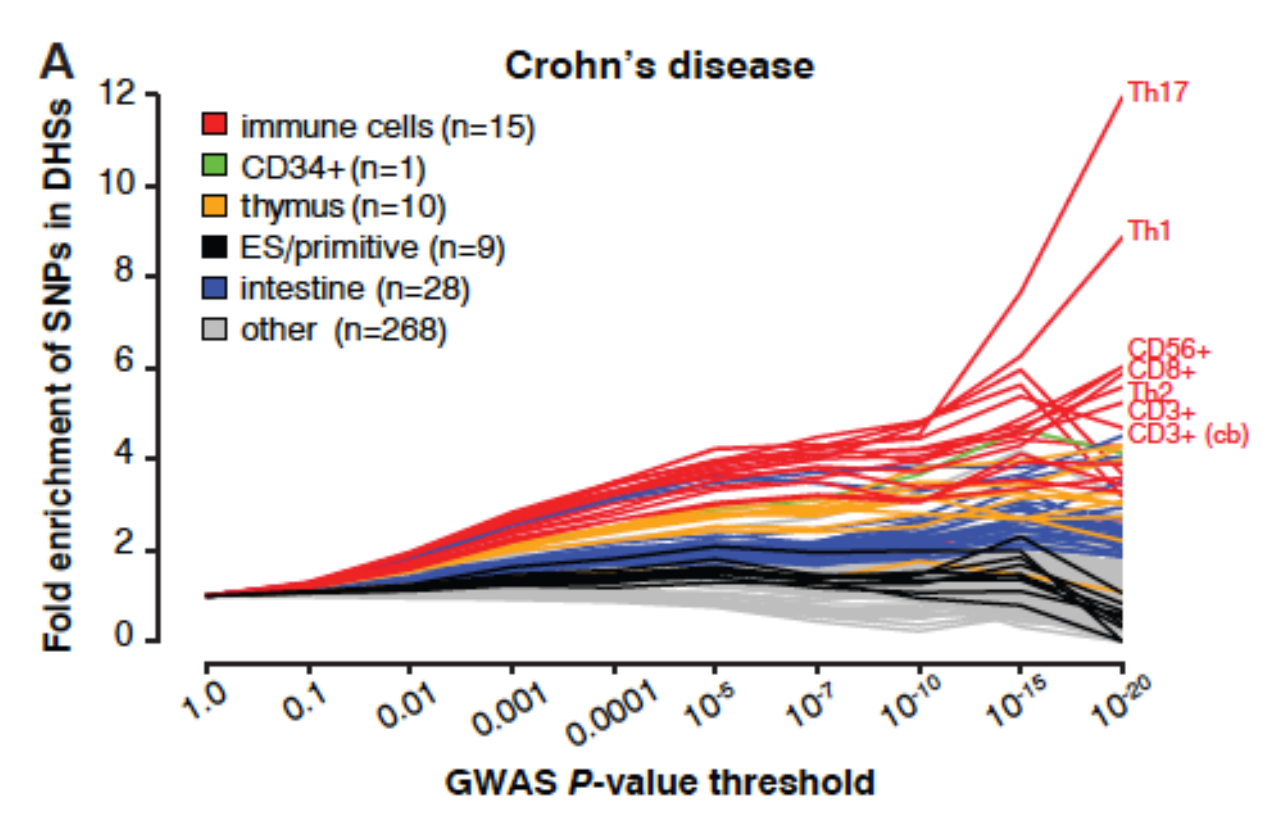
ENCODE cis-element Browser



Adrenal	0
Bladder	0.07
Cerebellum	0
CH12	348.42
CNS_D1	0

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

ENCODE And Epigenomics Data Can Be Used To Predict Cell Types



https://github.com/mauranolab/GWAS_plots

Stamatoyannopoulos, Science 337:1190, 2012

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
- Prediction of upstream regulators

➢ Genetic v. epigenetic

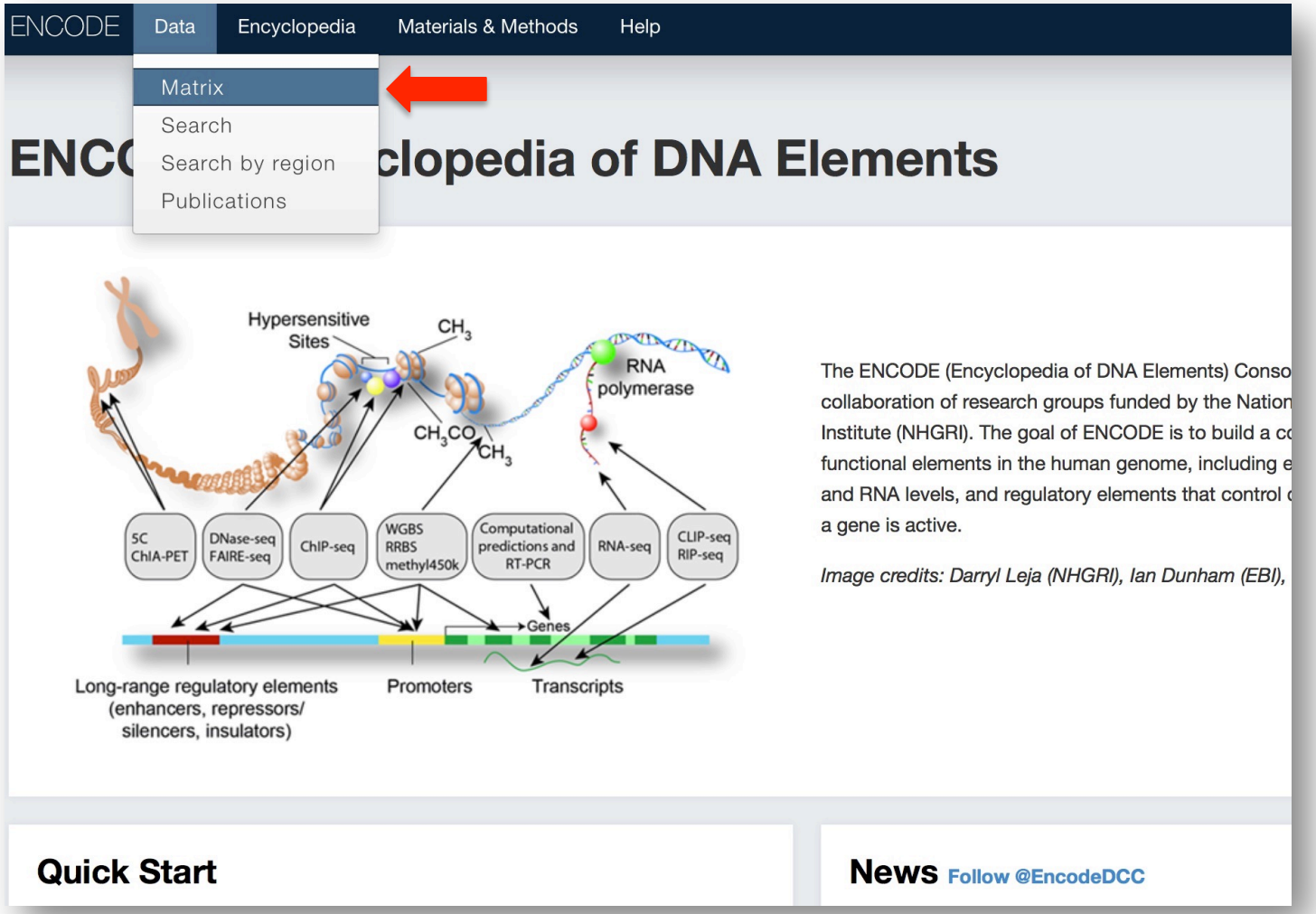
➢ Germline v. somatic

Overview

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

Workshop Session 2, 1, 5

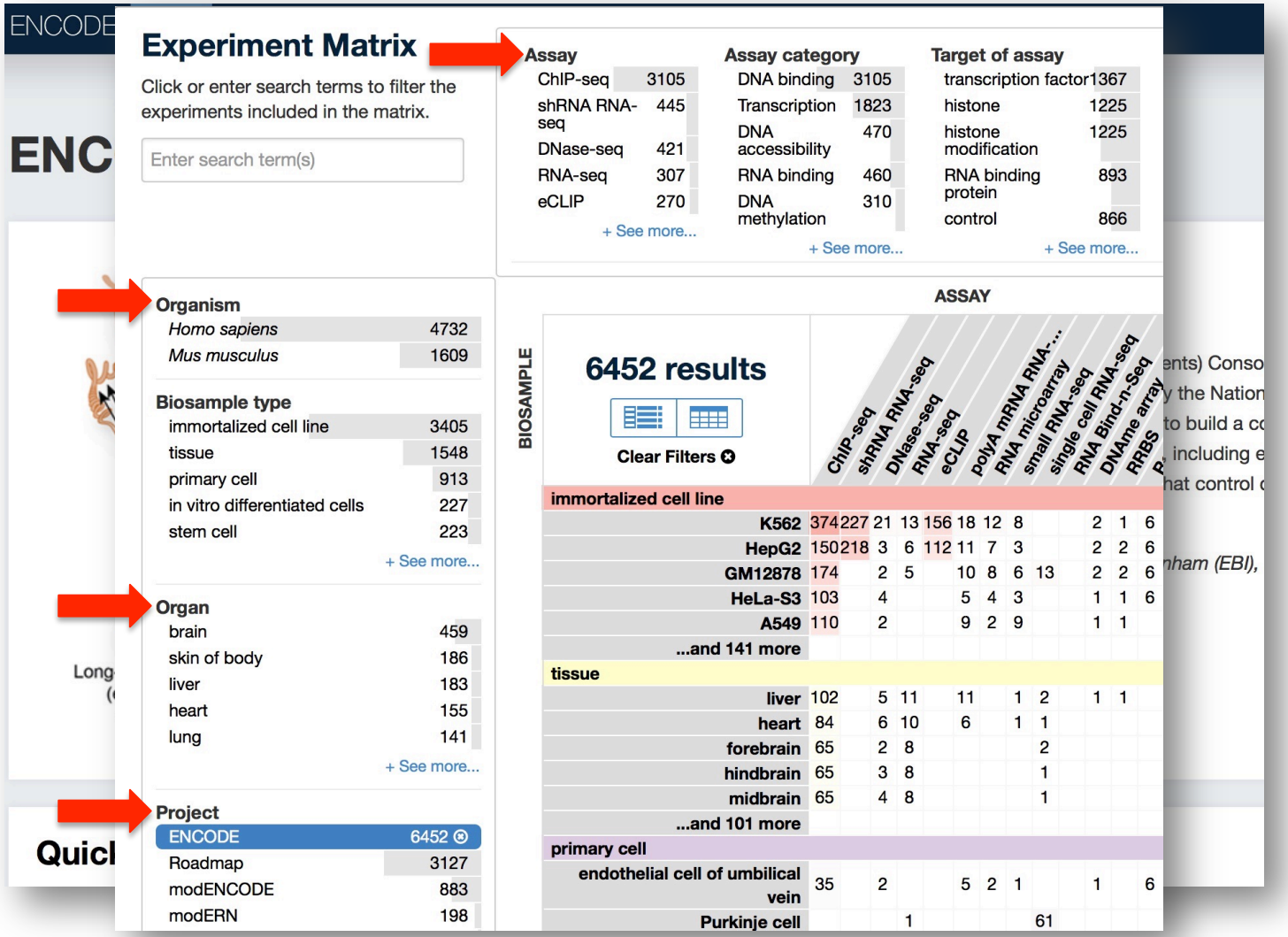
ENCODE Data



Workshop Session 2

https://www.encodeproject.org

ENCODE Data



Workshop Session 2

https://www.encodeproject.org

ENCODE Data

Data Use, Software, and Analysis Release Policies

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.

Data Use Policy for External Users

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:

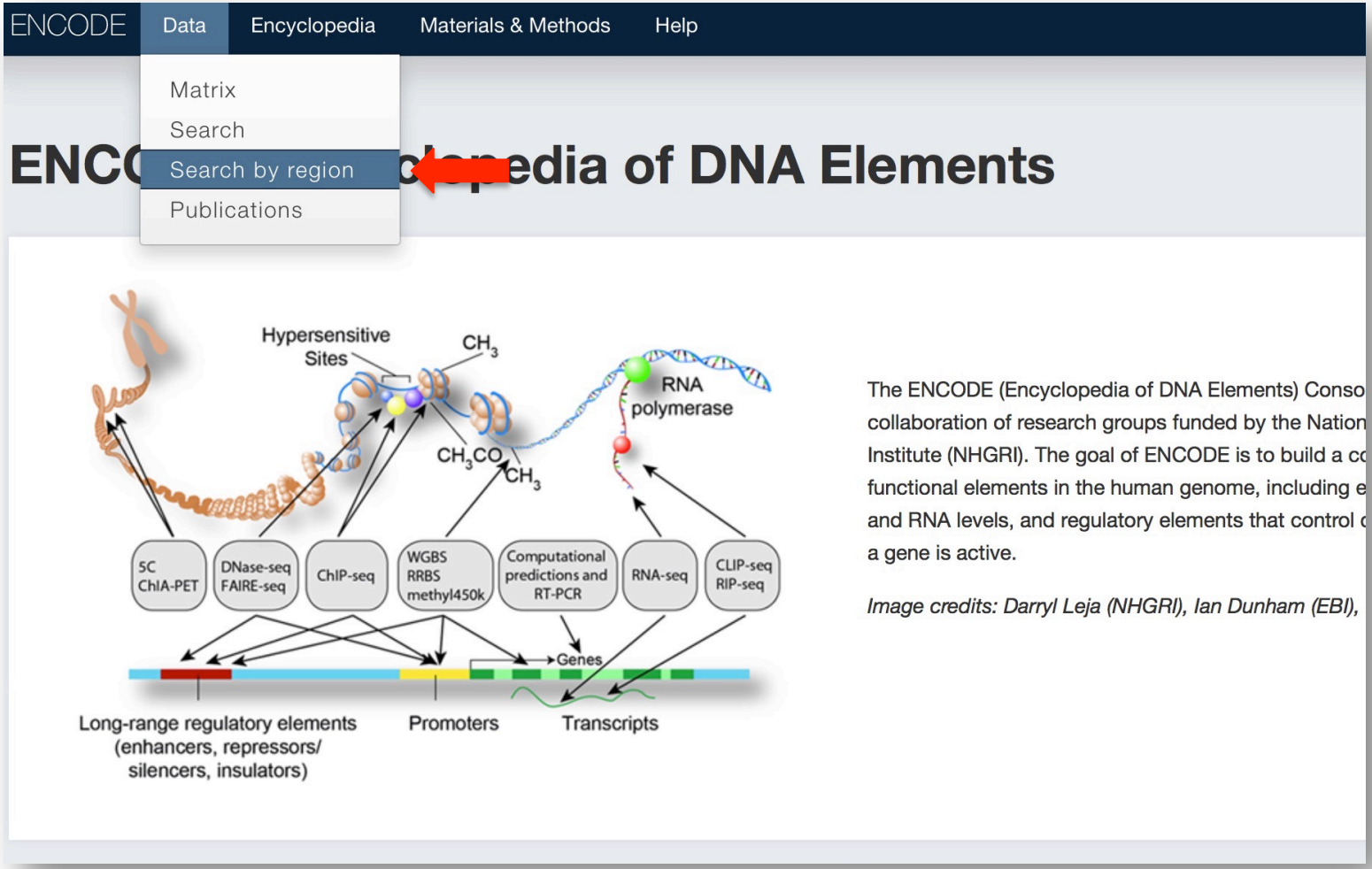
1. Cite the Consortium's most recent integrative publication (PMID: 22955616; PMC: PMC3439153);
2. Reference the ENCODE Data Coordination Center (DCC) or GEO accession numbers of the datasets (DCC accession: ENCSR037HRJ; GEO accession: GSE30567);
3. And acknowledge the ENCODE Consortium and the ENCODE production laboratory(s) generating the particular dataset(s)

modERN	198		Purkinje cell	1	61

Workshop Session 2

https://www.encodeproject.org

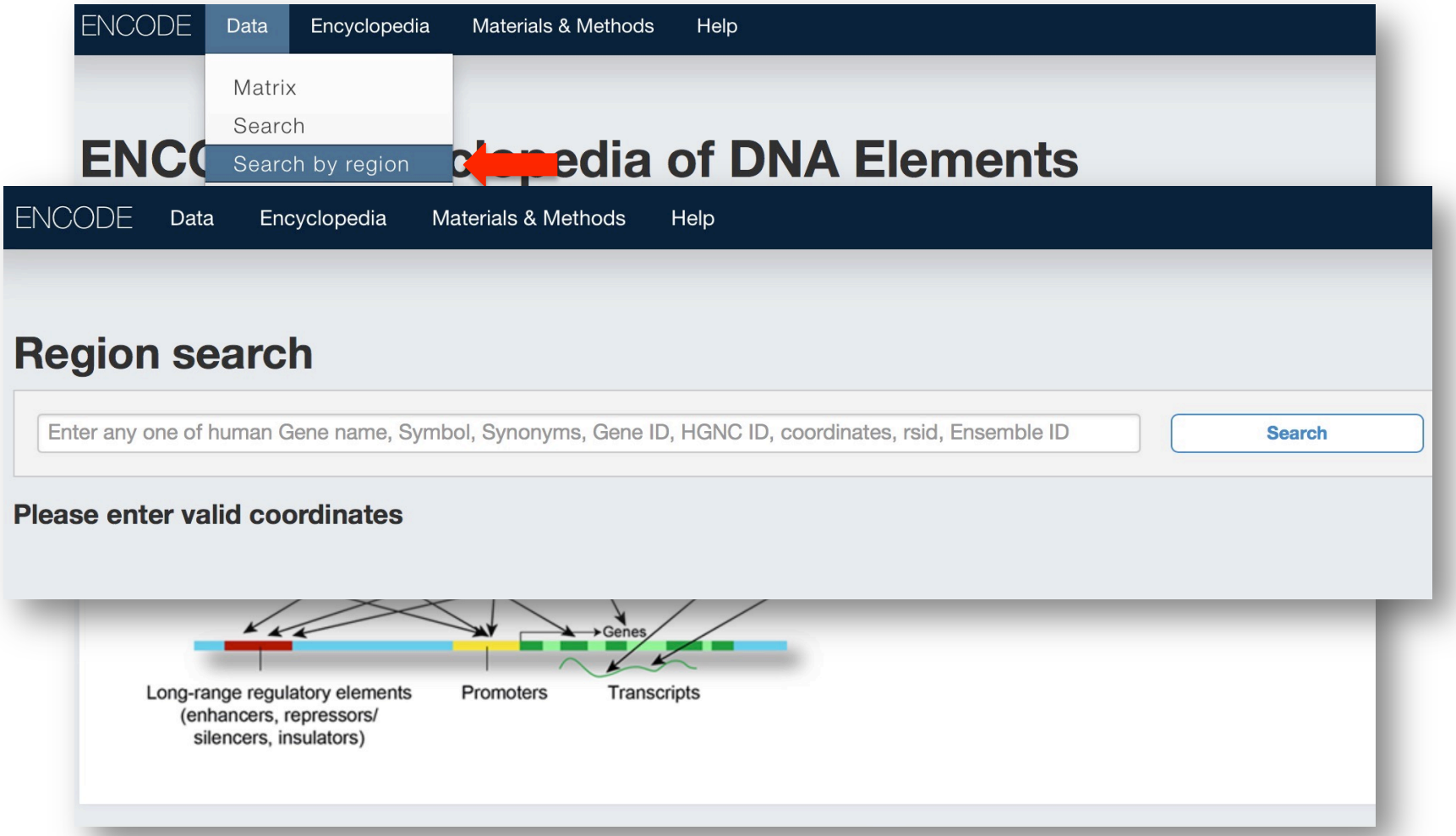
ENCODE Data



Workshop Session 2

https://www.encodeproject.org

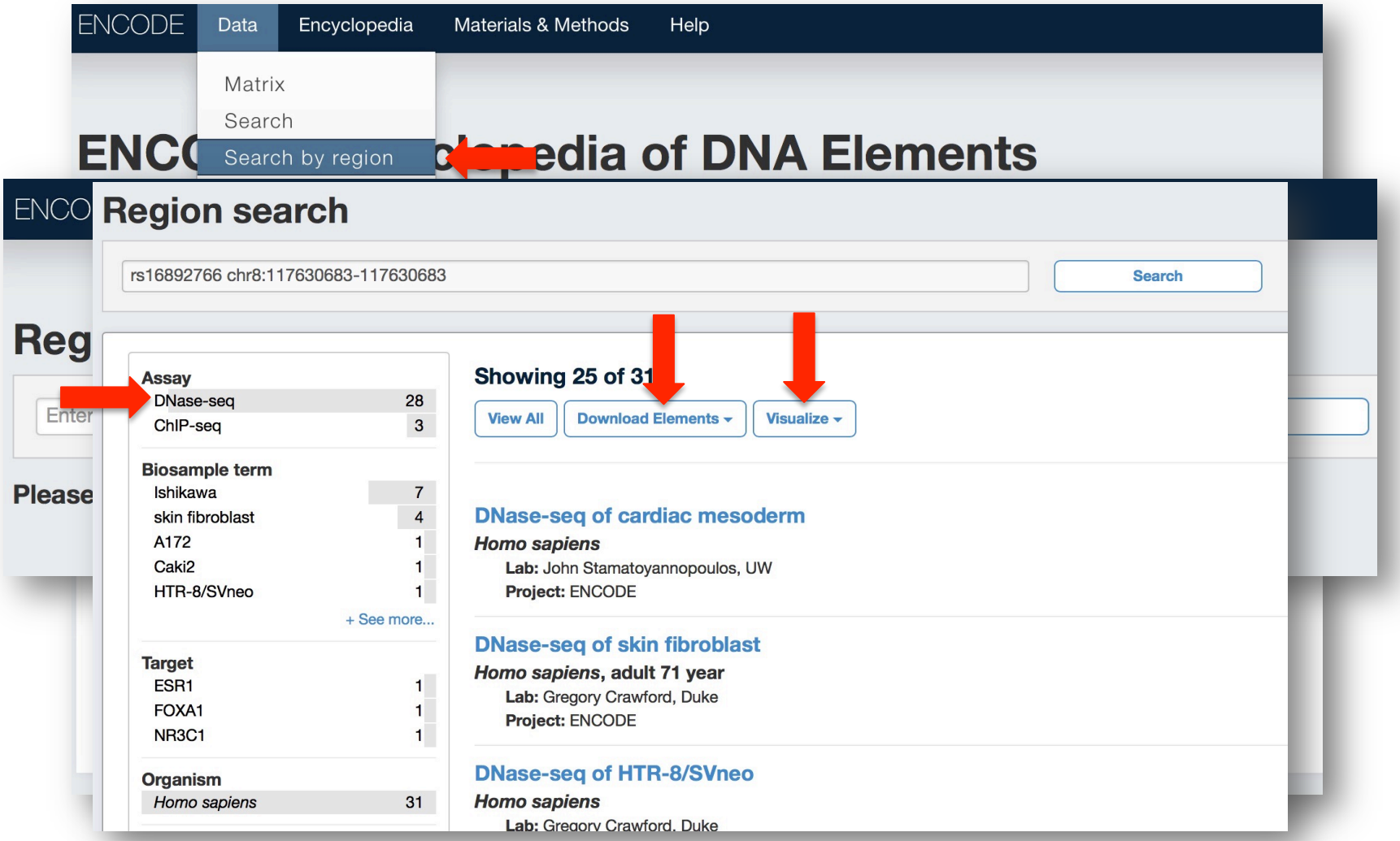
ENCODE Data



Workshop Session 2

https://www.encodeproject.org

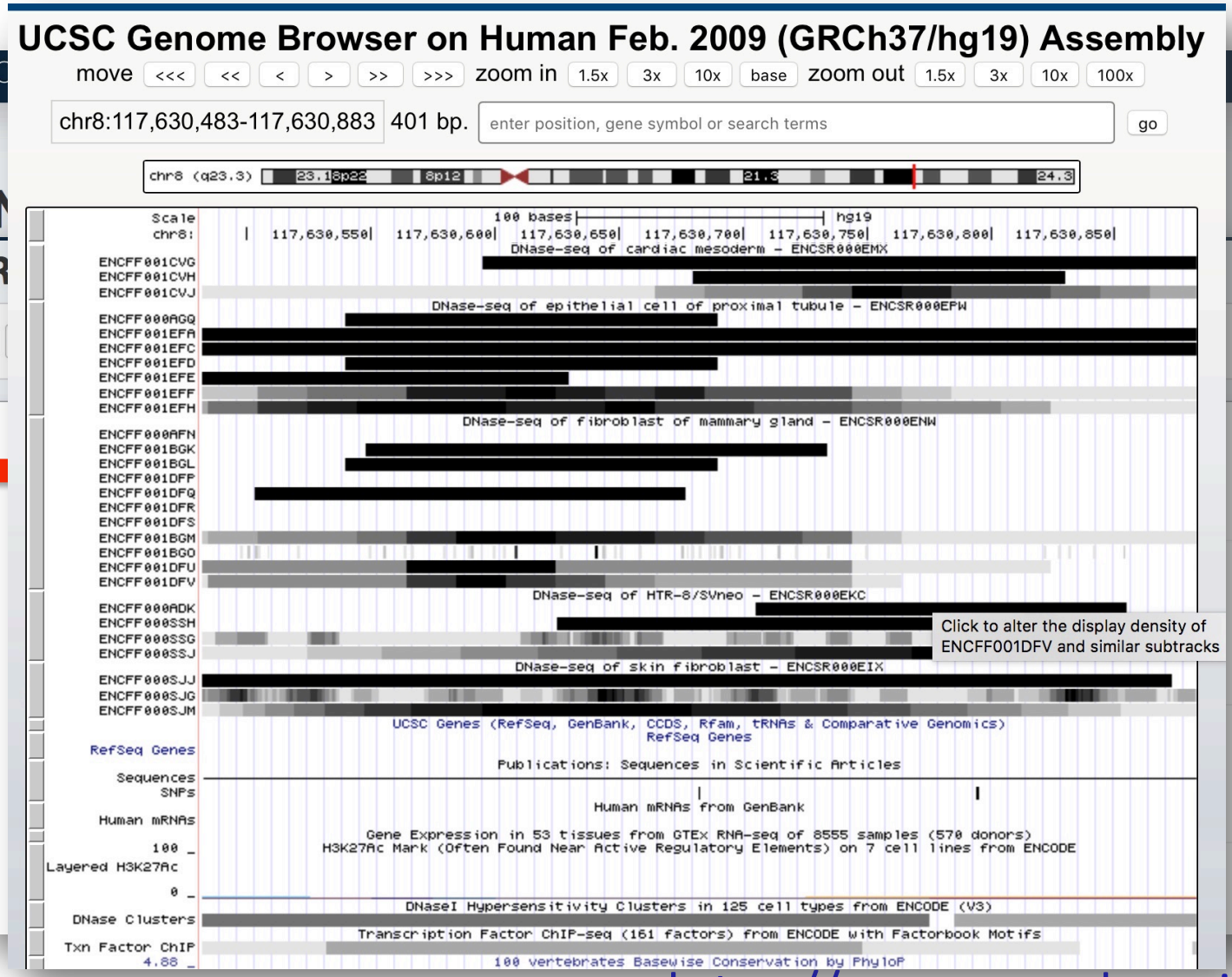
ENCODE Data



Workshop Session 2

https://www.encodeproject.org

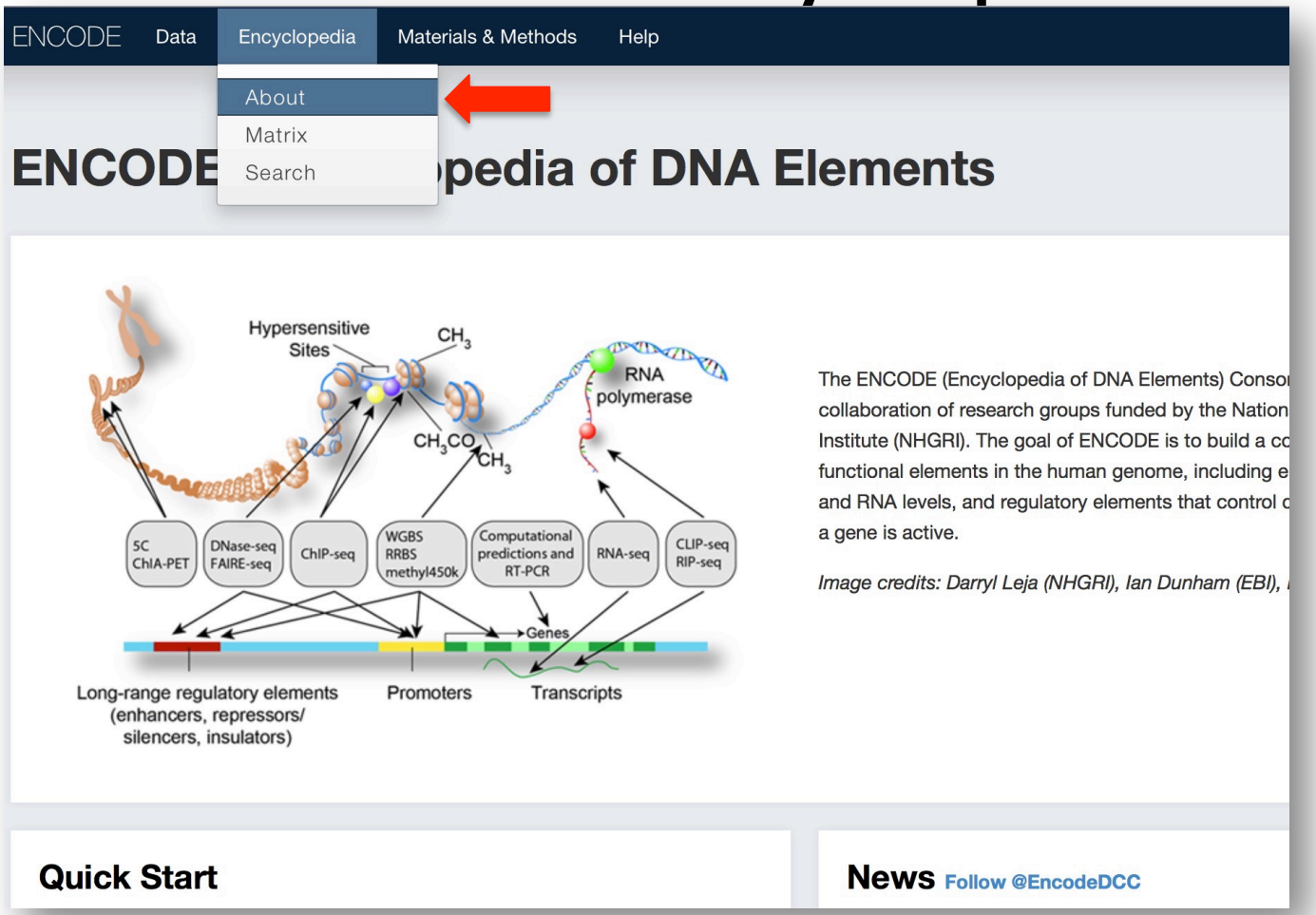
ENCODE Data



Workshop Session 2

https://www.encodeproject.org

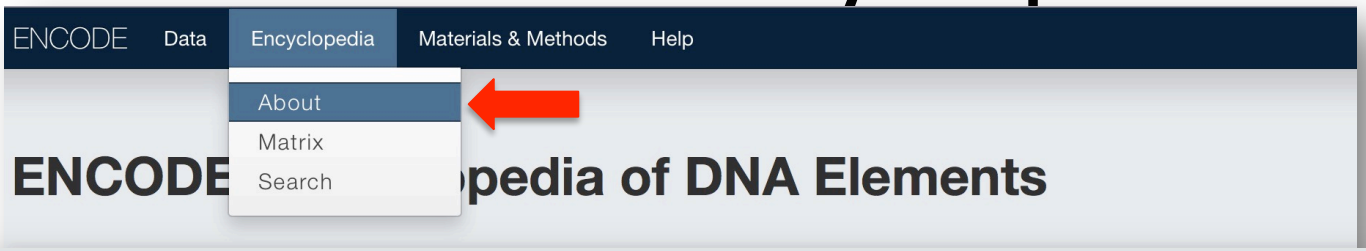
ENCODE Encyclopedia



Workshop Session 1, 5

https://www.encodeproject.org

ENCODE Encyclopedia

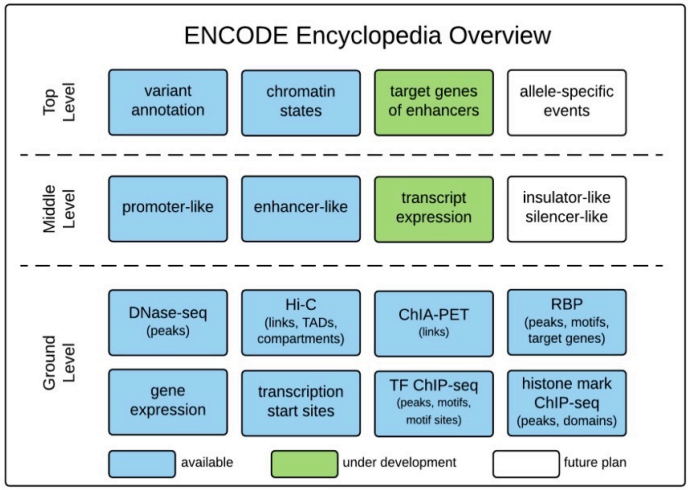


ENCODE Encyclopedia: Genomic annotations

Introduction

The ENCODE Consortium not only produces data, but also analyzes the data in an integrative fashion. The ENCODE Encyclopedia organizes the most salient analysis products into annotations, and provides tools to search and visualize them. The Encyclopedia has three levels of annotations:

- Ground level annotations are typically derived directly from the experimental data.
- Middle level annotations integrate multiple types of experimental data and multiple ground level annotations.
- Top level annotations integrate a broad range of experimental data and ground and middle level annotations.



Workshop Session 1, 5

https://www.encodeproject.org

ENCODE Encyclopedia

ENCODE Data Encyclopedia Materials & Methods Help

About

Ground Level Annotations

Gene expression (RNA-seq)

The expression levels of genes annotated by GENCODE 19 in over 100 human cell types and 70 mouse cell types.

[Long RNA-seq Data | Query | Download | Method]

BRCA1 Gene Expression

Transcription factor binding (TF ChIP-seq)

Peaks (enriched genomic regions) of TFs computed from ~900 human and mouse ChIP-seq experiments.

[Raw Data | Peaks]

Visualize sequence motifs and other information [Factorbook]

CTCF Motif from Factorbook

Histone mark enrichment (ChIP-seq)

Peaks of a variety of histone marks computed from ~600 ChIP-seq experiments.

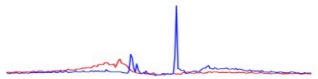
[Raw Data | Peaks]

H3K27ac from e11.5 Neural Tube

Open chromatin (DNase-seq)

DNase I hypersensitive sites (also known as DNase-seq peaks) computed from ~300 human and mouse experiments.

[Raw Data | Peaks]

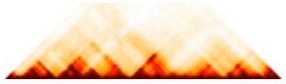


CTCF DHS Profile

Topologically associating domains (TADs) and compartments (Hi-C)

TADs and A and B compartments computed from 12 human cell lines.

[Raw Data | Visualize]



K562 Interaction Matrix

Workshop Session 1, 5

https://www.encodeproject.org

ENCODE Encyclopedia

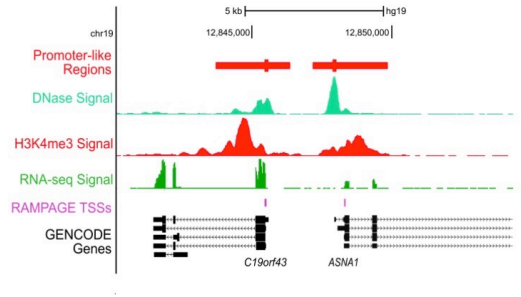
ENCODE Data Encyclopedia Materials & Methods Help

Middle Level Annotations

Promoter-like regions

DNase hypersensitivity and histone modification H3K4me3 are well-known indicators of promoter function. We have developed an unsupervised method that combines DNase and H3K4me3 signals in the same cell type to predict promoter-like regions. When used to predict ranked gene expression from RNA-seq data, our method shows higher accuracy than DNase and H3K4me3 individually. We have applied this method to 107 human cell types and 14 mouse cell types with both DNase and H3K4me3 data generated by the ENCODE and Roadmap Epigenomic consortia. For cell and tissues types with only H3K4me3 data, we centered predictions on H3K4me3 peaks and ranked them by H3K4me3 signals. You can query these promoter-like regions by genomic locations, nearby genes, or SNPs, and visualize them in the UCSC and WashU genome browsers.

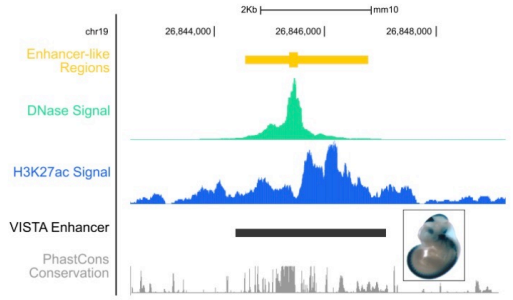
[Visualize | | Method]



Enhancer-like regions

DNase hypersensitivity and histone modification H3K27ac are well-known indicators of enhancer function. We have developed an unsupervised method that combines DNase and H3K27ac signals in the same cell type to predict enhancer-like regions. When tested on mouse transgenic assays, our method shows higher accuracy than DNase and H3K27ac individually. We have applied this method to 47 human cell types and 14 mouse cell types with both DNase and H3K27ac data generated by the ENCODE and Roadmap Epigenomic consortia. For cell and tissues types with only H3K27ac or DNase data, we rank the peaks using the available data and make predictions of enhancer-like regions. You can query these enhancers by genomic locations, nearby genes, or SNPs, and visualize them in the UCSC and WashU genome browsers.

[Visualize | Method]



Enhancer-like genomic regions were tested on VISTA experimentally-validated enhancer elements: [VISTA]

Workshop Session 1, 5

https://www.encodeproject.org

ENCODE Encyclopedia

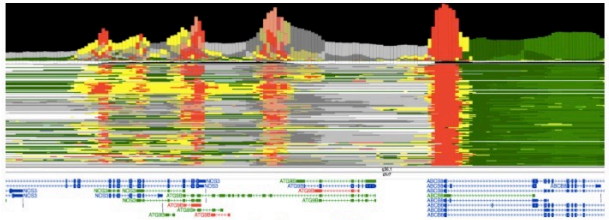
ENCODE Data Encyclopedia Materials & Methods Help

Top Level Annotations

Chromatin states

Semi-automated genomic annotation methods such as ChromHMM and Segway take as input a panel of epigenomic data (including histone mark ChIP-seq and DNase-seq) in a particular cell type and use machine learning methods to simultaneously partition the genome into segments and assign chromatin states to these segments; the states are assigned such that two segments with the same state exhibit similar epigenomic patterns. The procedure is "semi-automated" because states are then manually compared with known biological information in order to designate each state as an enhancer-like, promoter-like, gene body, etc.

[Search]

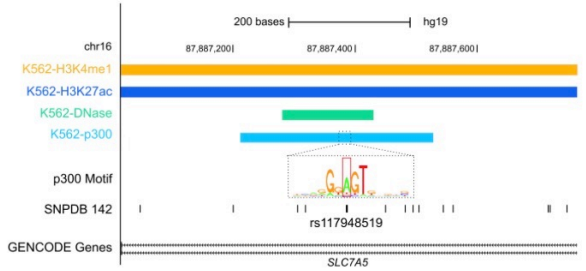


Epilogos

Variant Annotation

Over the past decade, Genome Wide Association Studies (GWAS) have provided insights into how genetic variations contribute to human diseases. However, over 80% of the variants reported by GWAS are in noncoding regions of the genome and the mechanism of how they contribute to disease onset is unknown. By integrating data from the ENCODE project and other public sources, RegulomeDB and HaploReg are two resources developed by ENCODE labs to aid the research community in annotating GWAS variants. FunSeq2 is another ENCODE resource for annotating both germline and somatic variants, particularly in the noncoding regions of cancer genomes.

[RegulomeDB | HaploReg | FunSeq2]



TADs

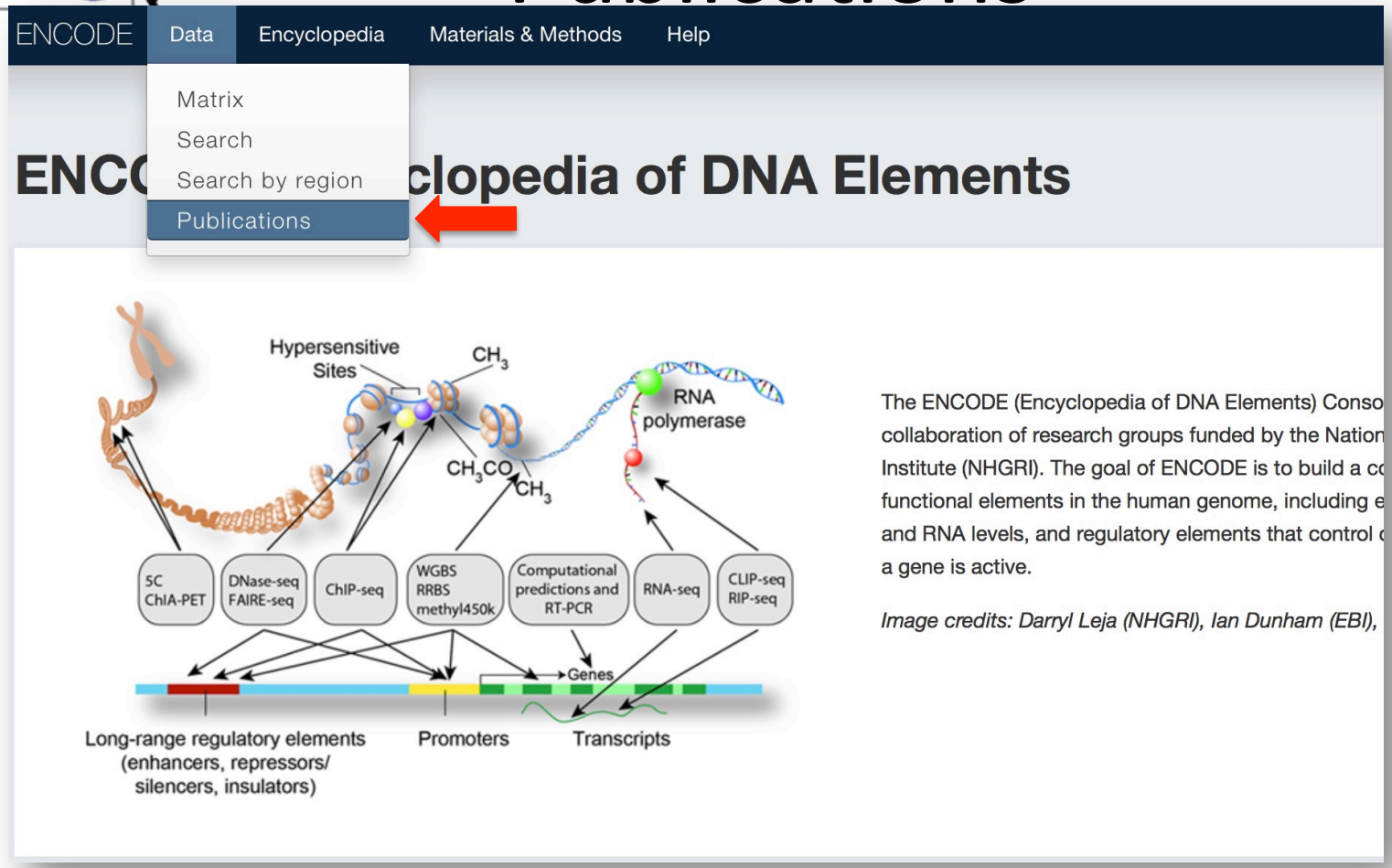
[Raw

Enhancer-like genomic regions were tested on VISTA experimentally-validated enhancer elements: [VISTA]

Workshop Session 1, 5

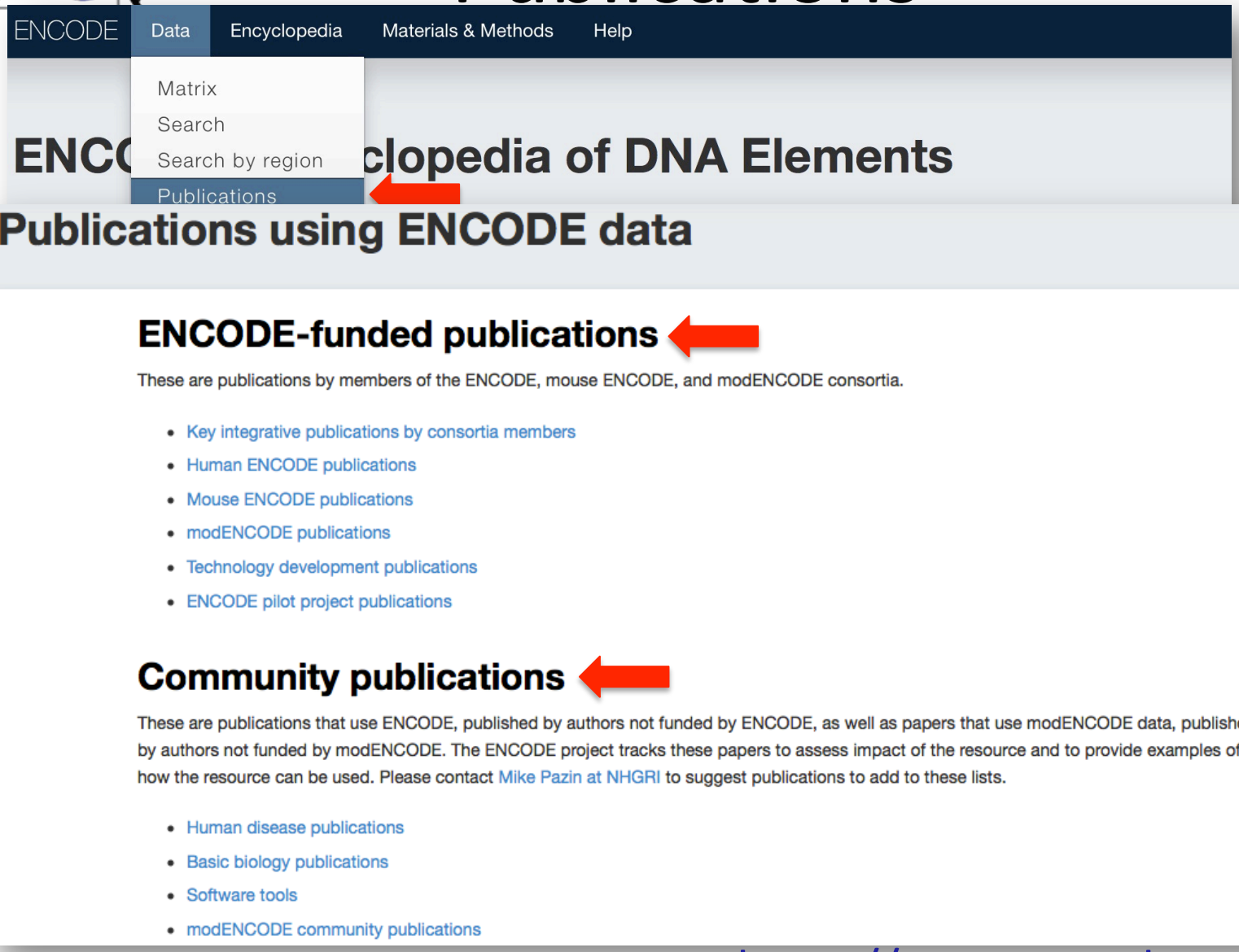
https://www.encodeproject.org

Publications



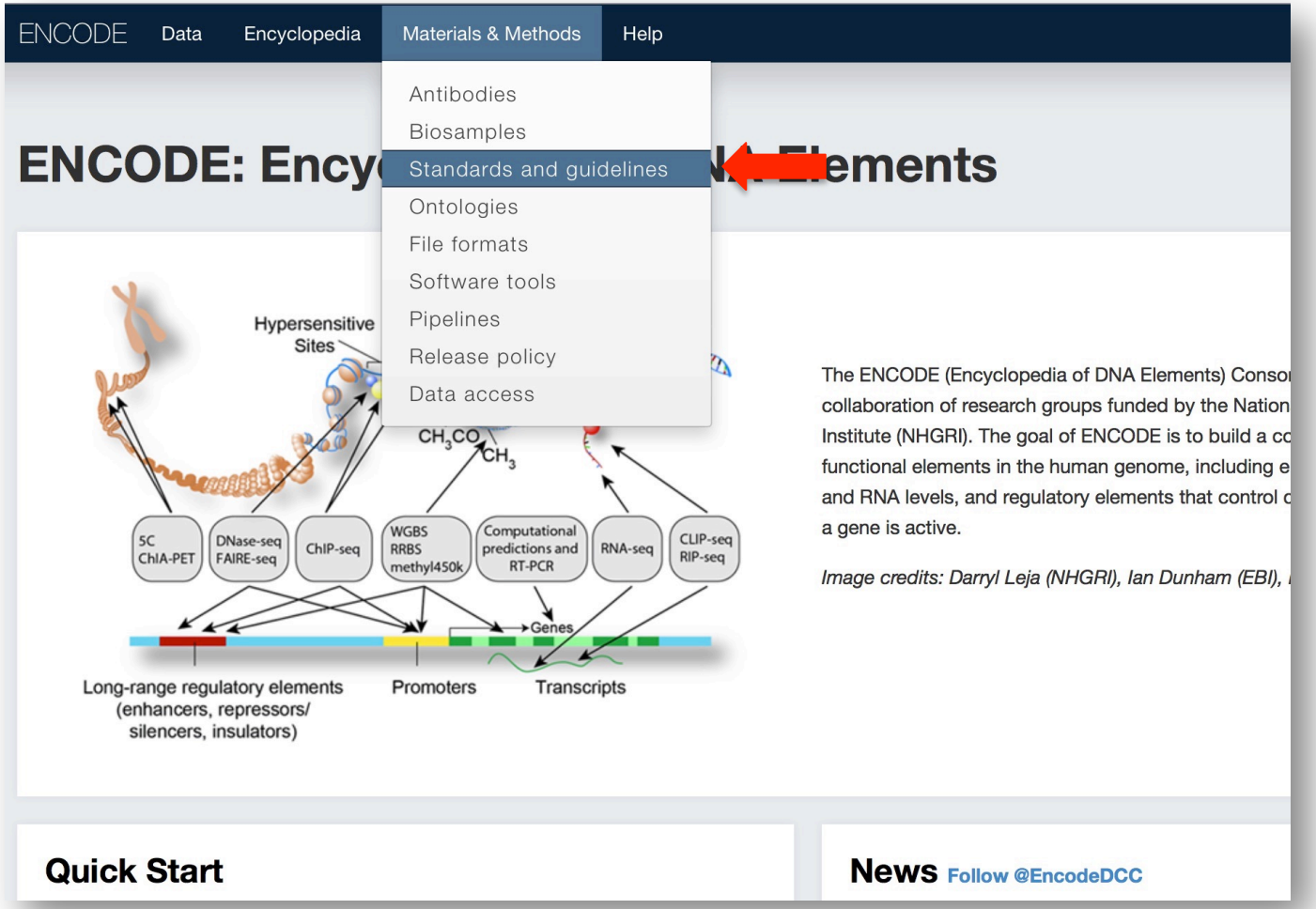
https://www.encodeproject.org

Publications



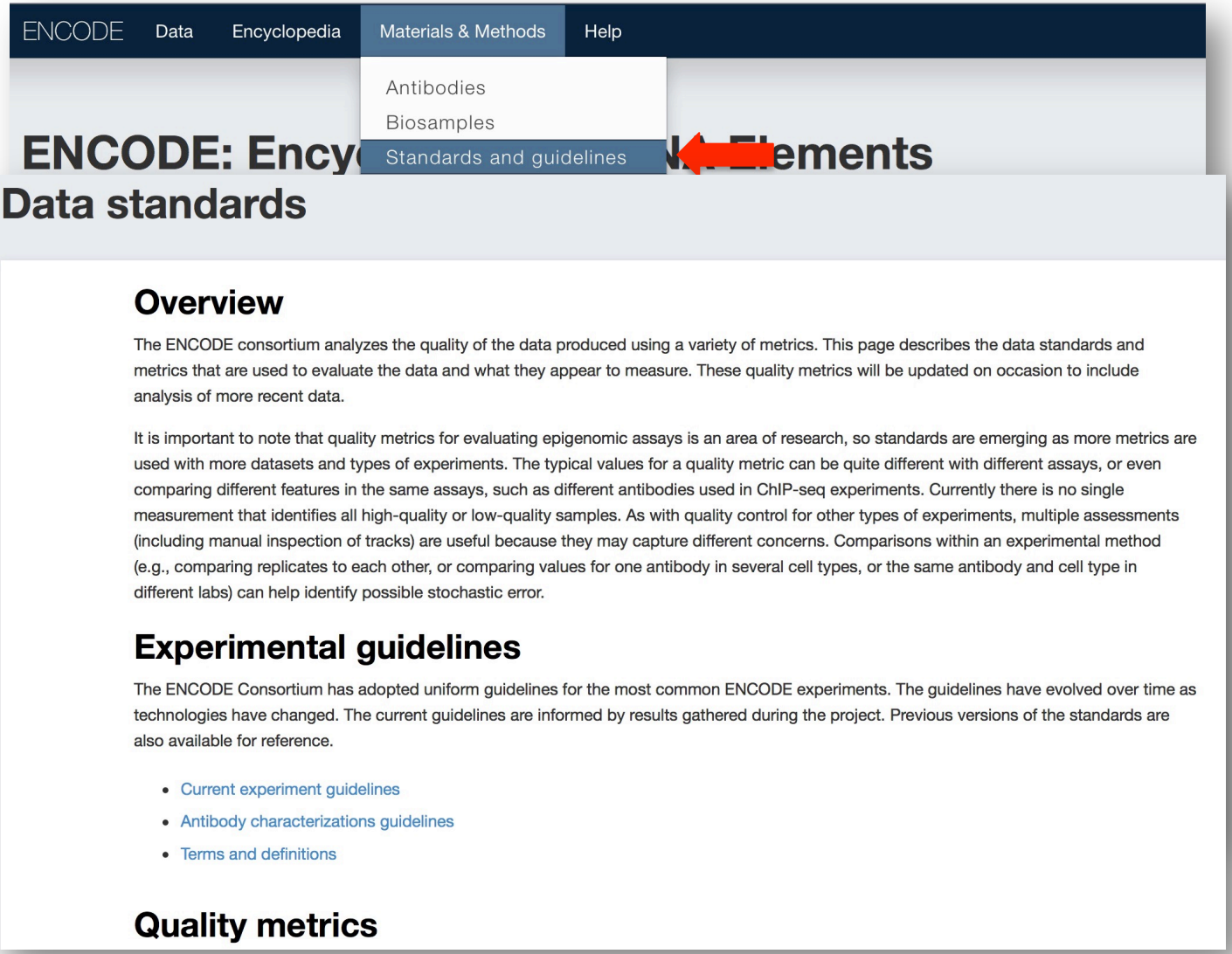
https://www.encodeproject.org

ENCODE Data Standards



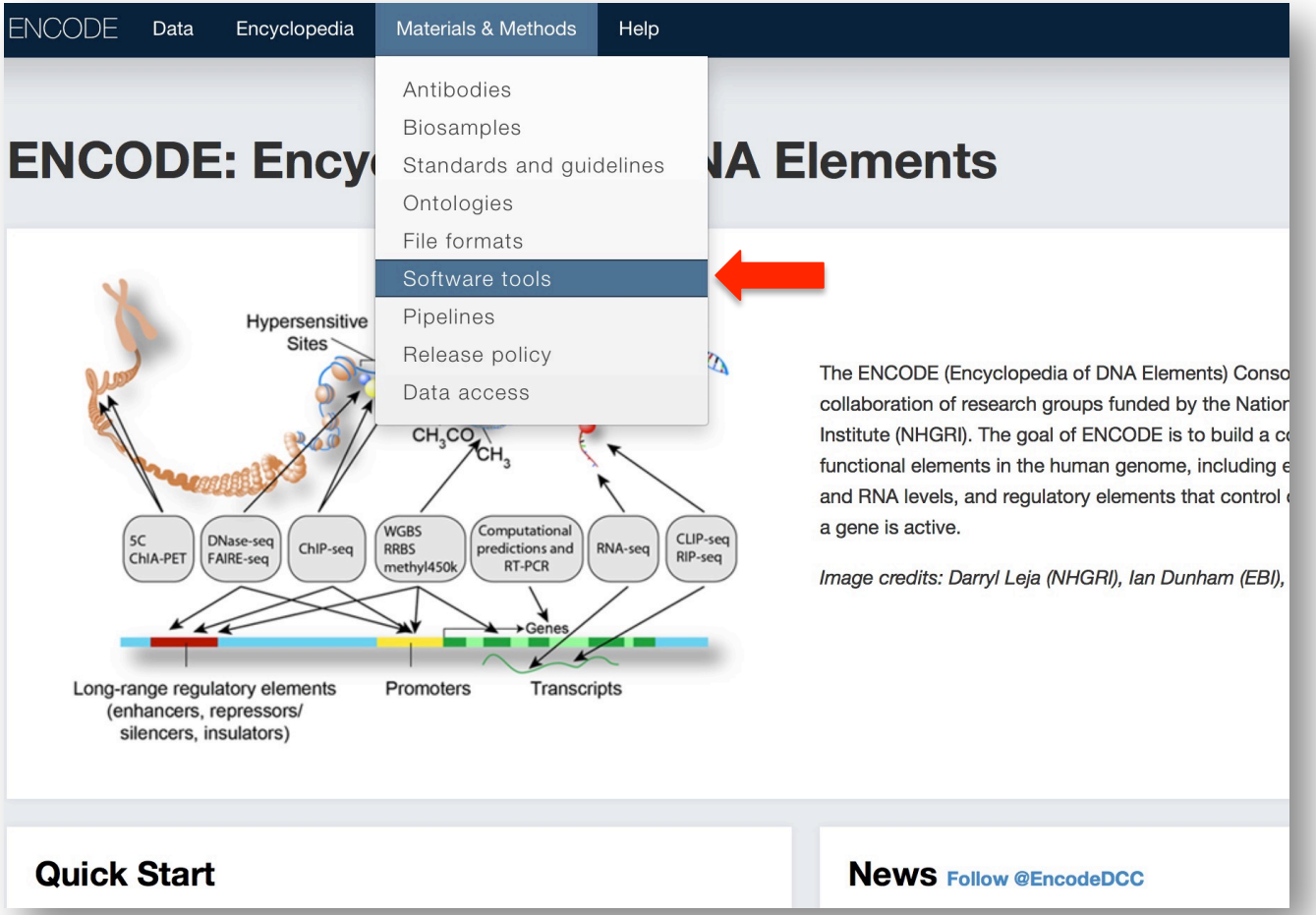
https://www.encodeproject.org

ENCODE Data Standards



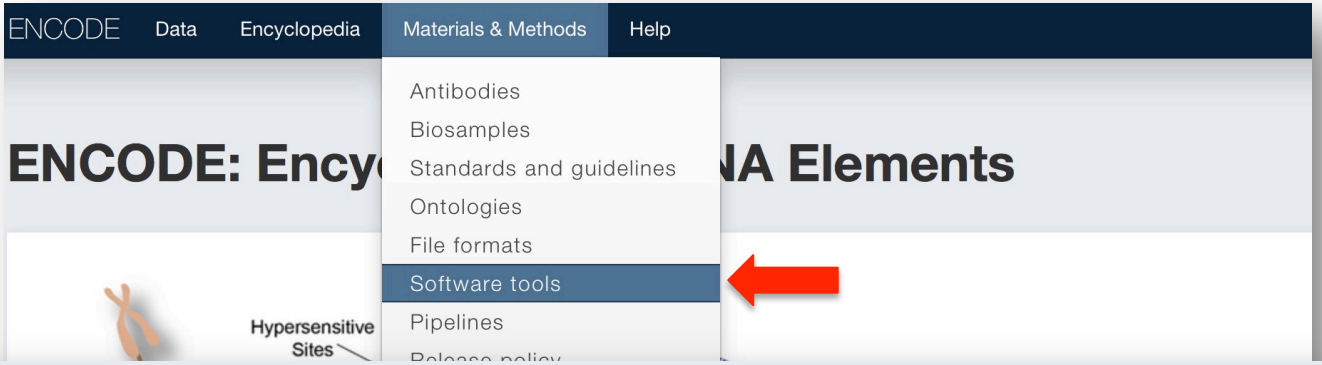
https://www.encodeproject.org

ENCODE Software Tools



https://www.encodeproject.org

ENCODE Software Tools



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. View all software used or developed by the ENCODE Consortium or select one of the following:

- 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.

https://www.encodeproject.org

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 8 consortia

IHEC International Human Epigenome Consortium

ROADMAP epigenomics PROJECT

ENCODE

BLUEPRINT epigenome

CEEHRC

CREST IHEC International Human Epigenome Consortium, Team Japan

DEEP

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/

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