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 29, 2015

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

Functional Genomics Is Central To NHGRI Goals

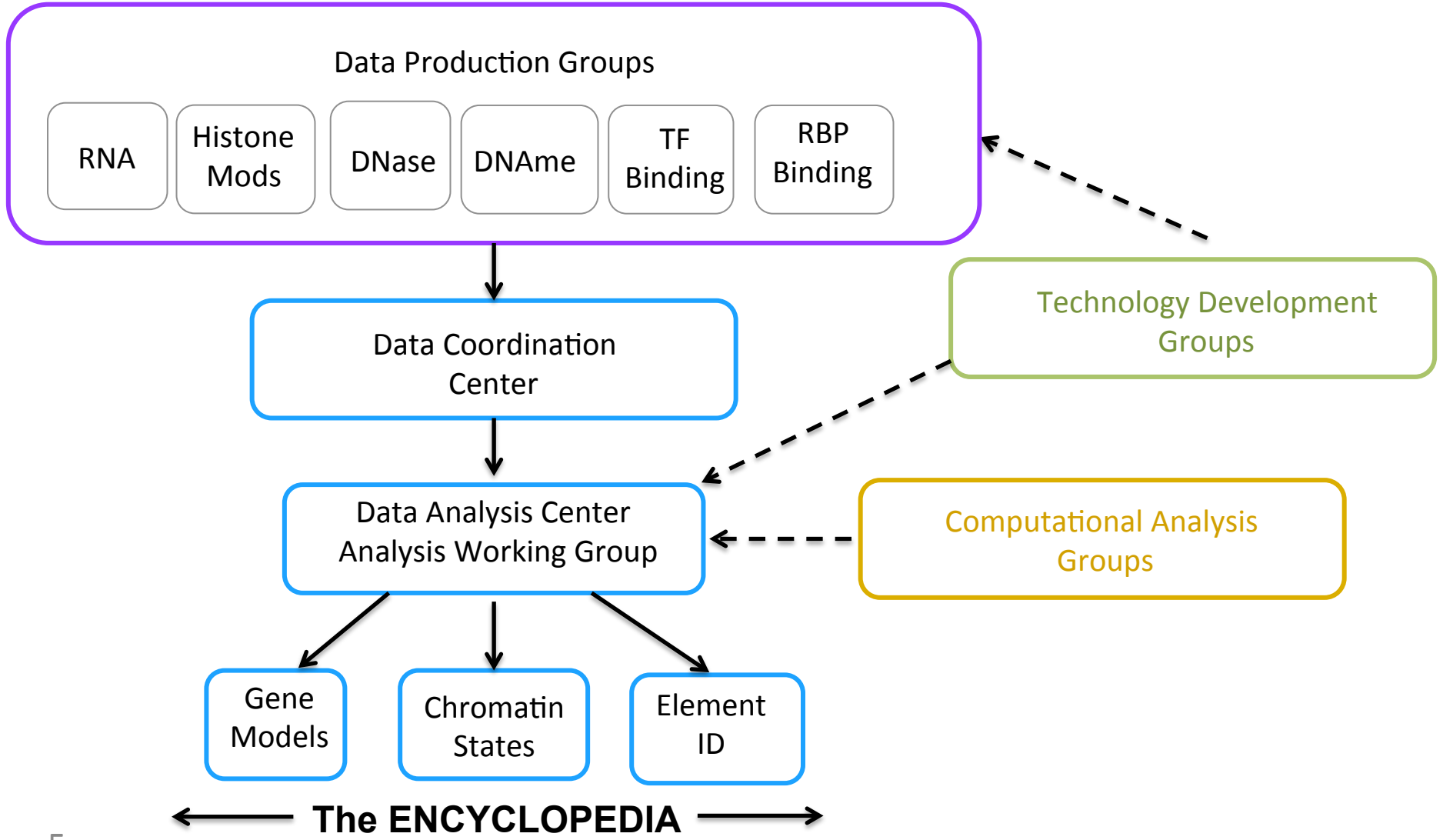
Non-coding DNA is important for disease and gene regulation

- 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.

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

ENCODE Consortium Structure



ENCODE Accomplishments

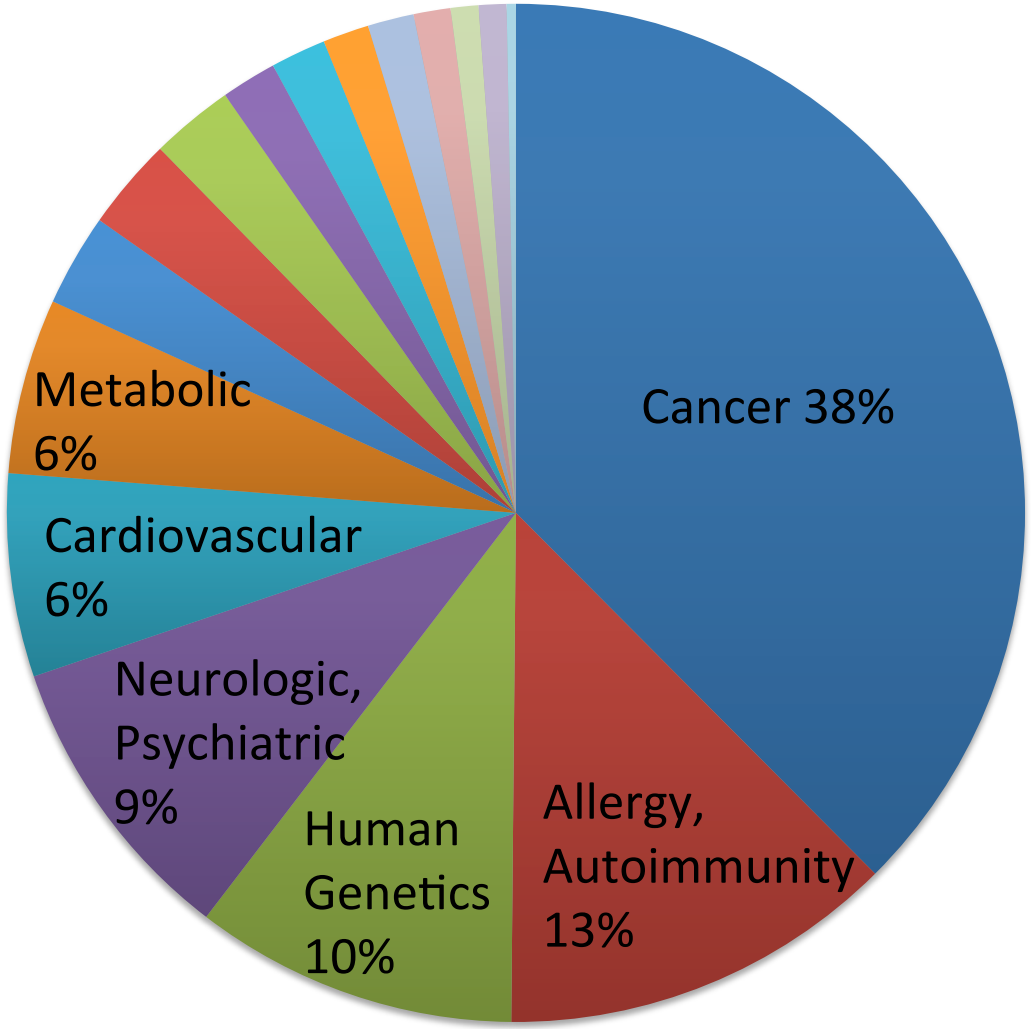
- Sharing 1000s of datasets
  - No embargo
  - High quality
  - Uniformly processed
- Sharing software
- Data interoperability
- Informed consent for unrestricted-access sharing of genomic data

Publications Using ENCODE Data

Hundreds of Consortium publications

~1000 community publications using ENCODE data:

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



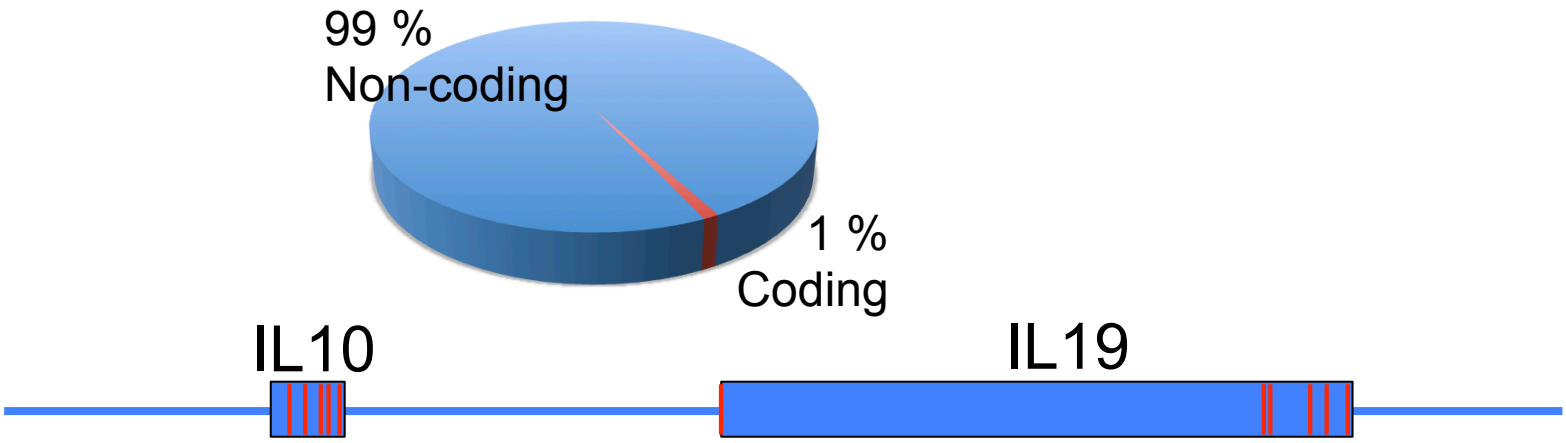
ENCODE: Encyclopedia Of DNA Elements

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

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)
- Regulatory regions aren't always in the same order as gene targets

➢ Need unbiased experimental investigation

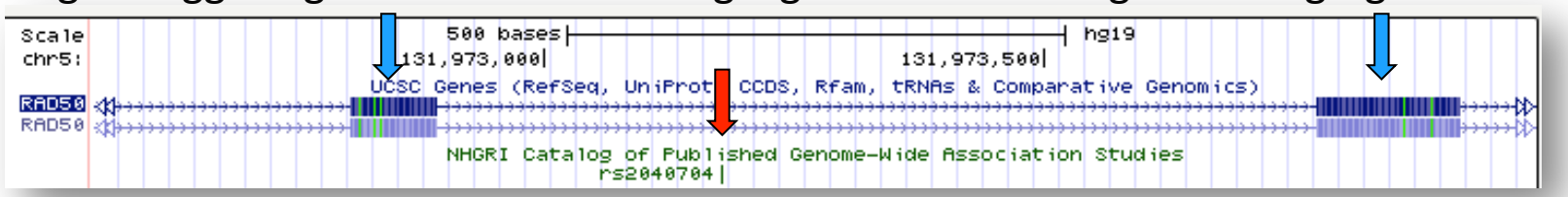


1,500 Letters Of Our 3 Billion Letter Genome

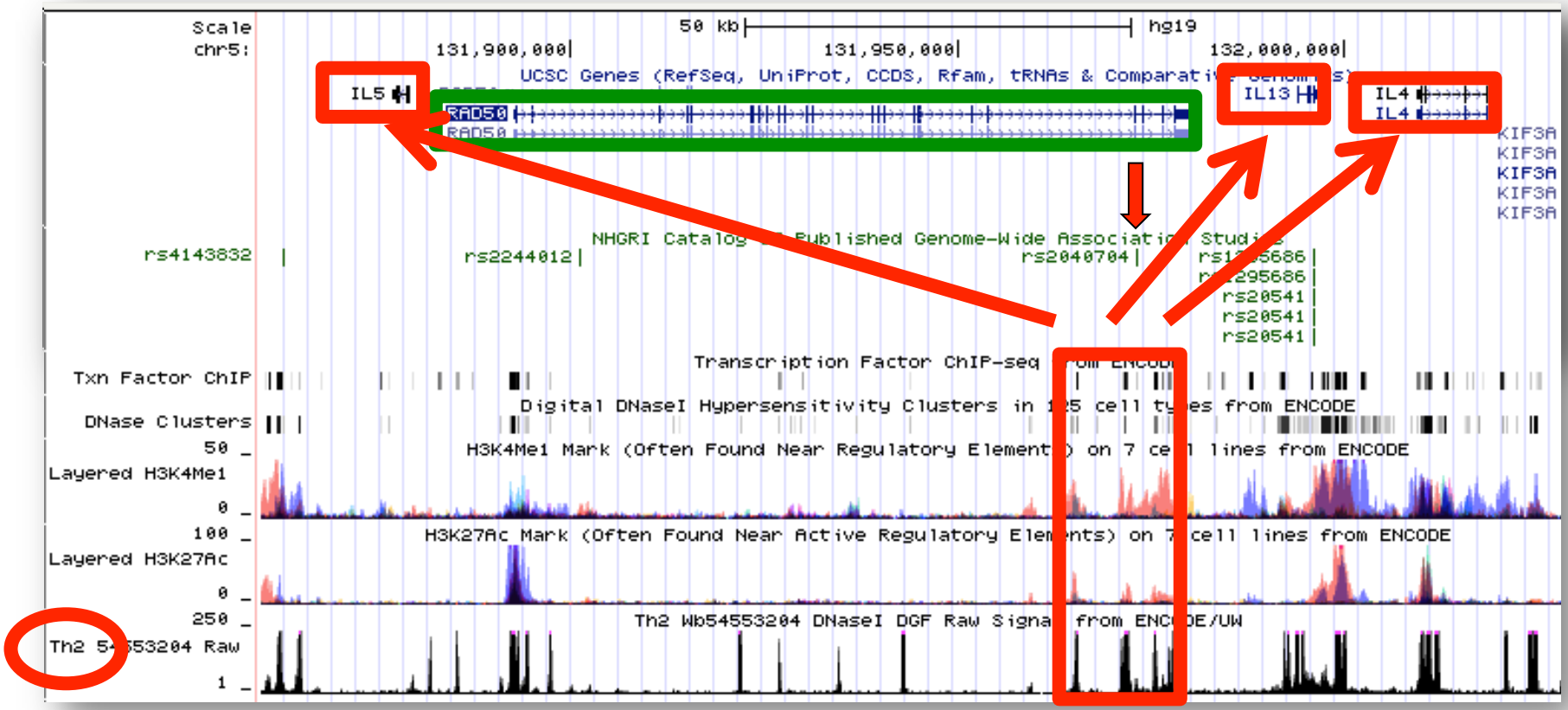
agccaagcagcaaagttttgctgctgtttatttttgtagctcttactatattctacttttaccattgaaaatattgaggaagttatt
tatatttctatttttatatattatatattttatgtattttaatattactattacacataattatttttatatatatgaagtaccaatg
acttccttttccagagcaataatgaaatttcacagtatgaaaatggaagaaatcaataaaattatacgtgacctgtggcgaa
gtacctatcgtggacaaggtgagtaccatggtgtatcacaaatgctctttccaaagccctctccgcagctcttccccttatga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

Maps And Annotation Help Us To Understand The Sequence

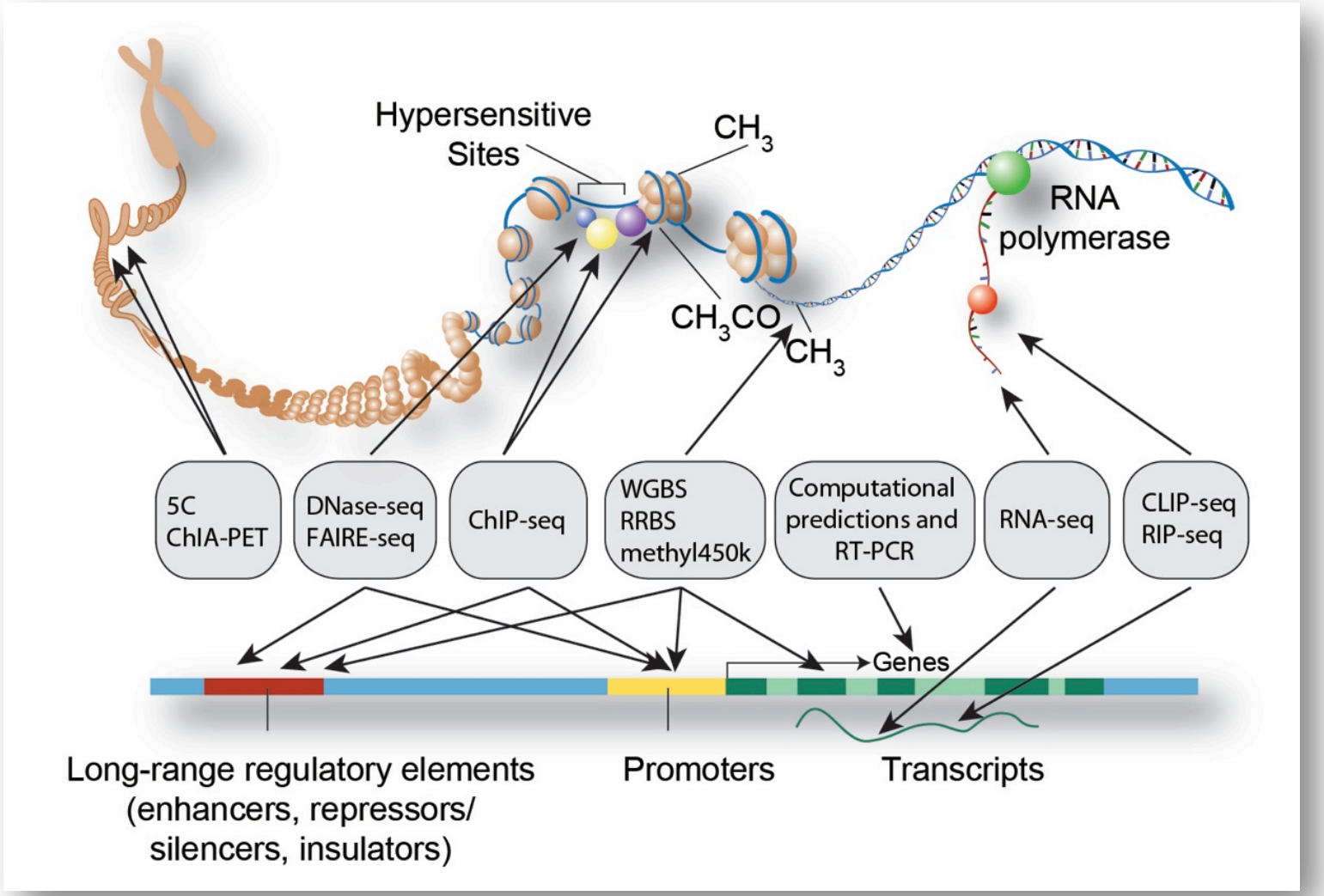
agccaagcagcaaagttttgctgctgtttatttttgtagctcttactatattctacttttaccattgaaaatattgaggaagttatt
tatatttctatttttatatattatatattttatgtattttaatattactattacacataattatttttatatatatgaagtaccaatg
acttccttttccagagcaataatgaaatttcacagtatgaaaatggaagaaatcaataaaattatacgtgacctgtggcgaa
gtacctatcgtggacaaggtgagtaccatggtgtatcacaaatgctctttccaaagccctctccgcagctcttccccttatga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



Richer Maps Provide More Information



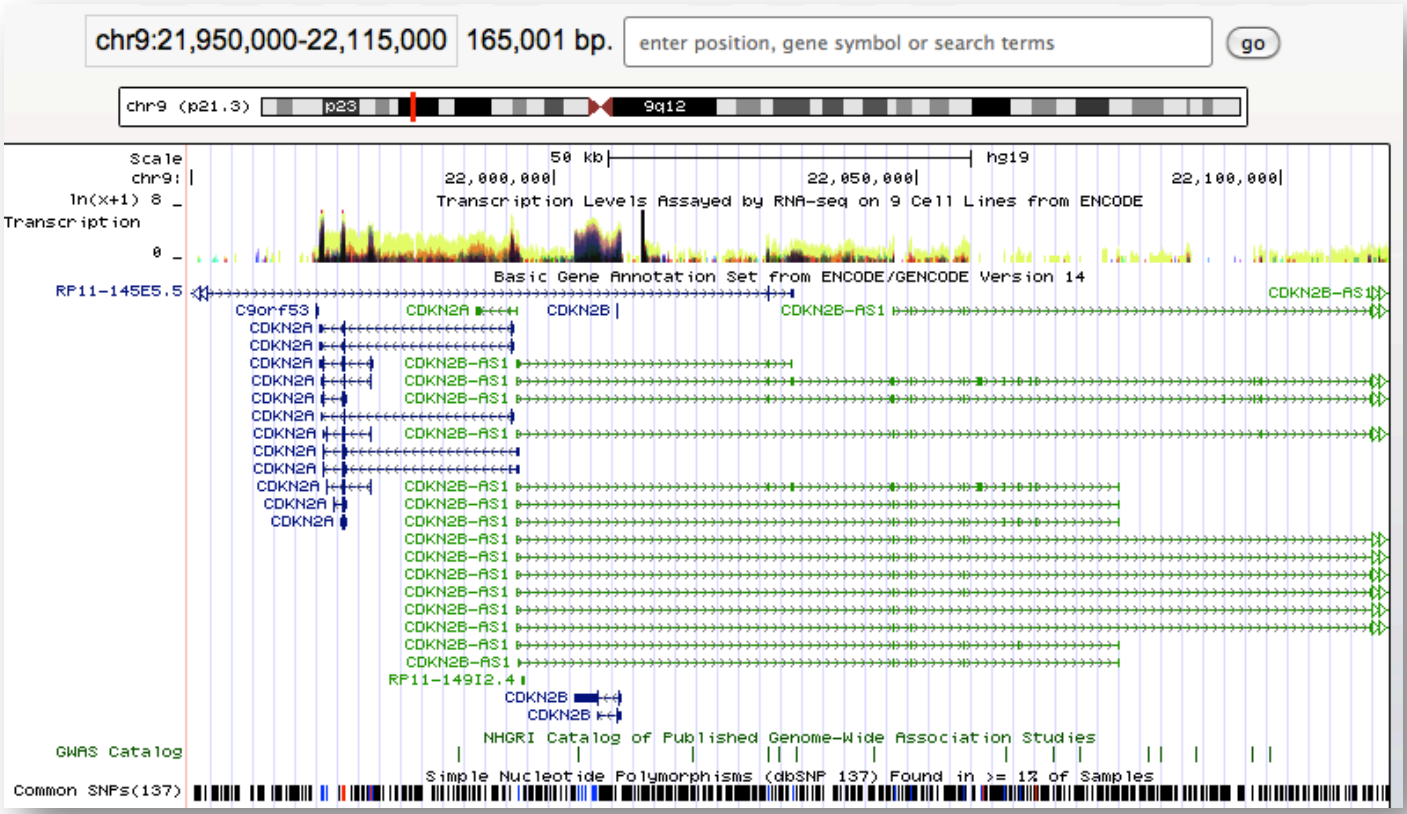
ENCODE Data Types



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

RNA Data

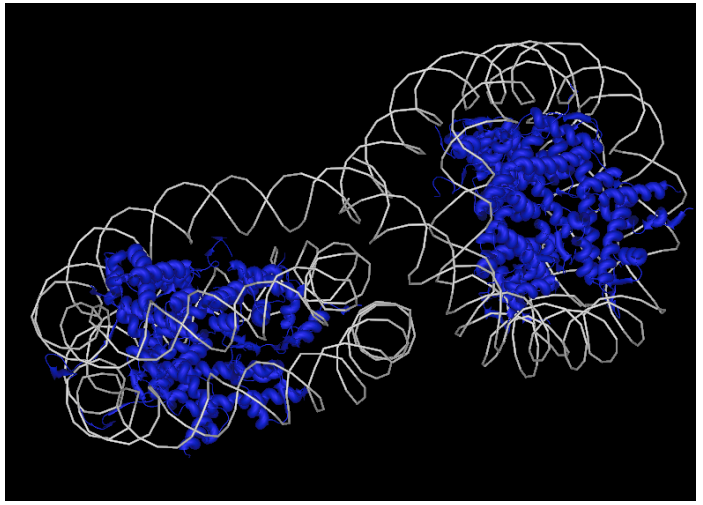
- RNA amount (mRNA and short RNA)
- Cell Specificity
- Transcript modeling (GENCODE)



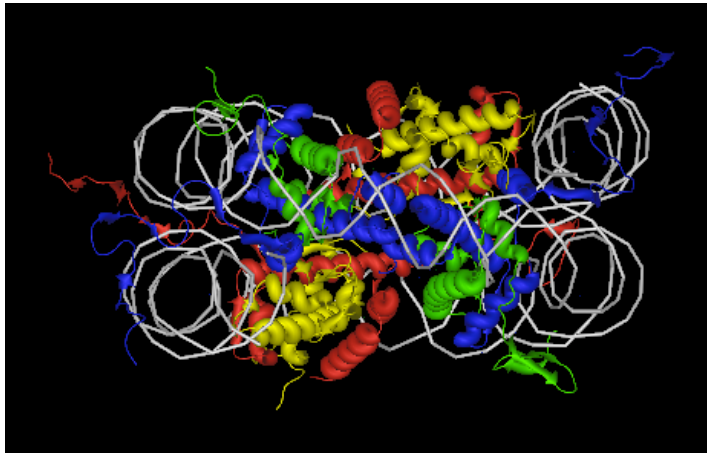
https://encodeproject.org

ENCODE Chromatin Structure Data

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



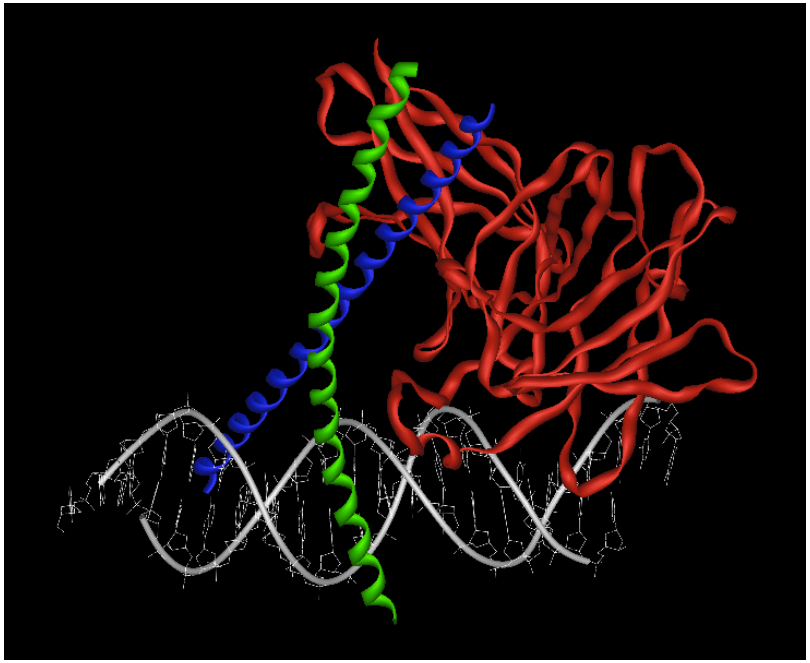
Richmond, PDB 1ZBB



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

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

Standard ENCODE Use Cases: Hypothesis Generation

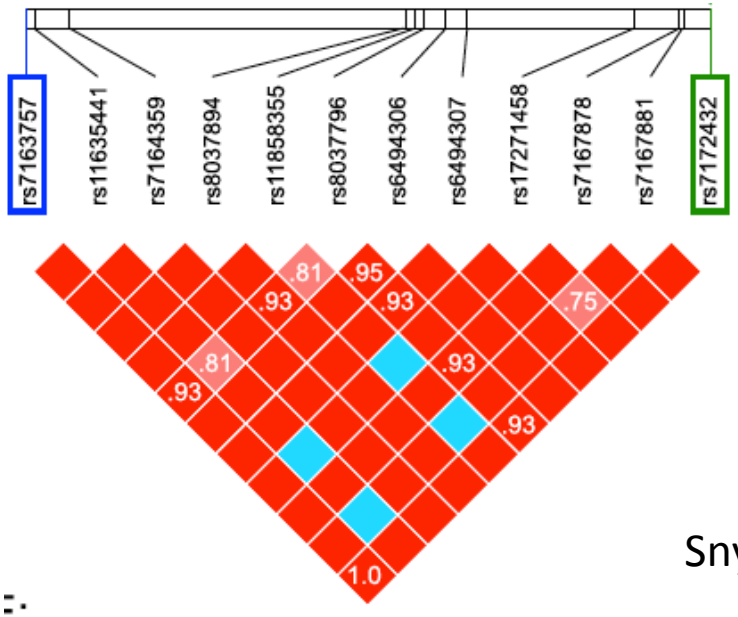
- Prediction of causal variants/regulatory elements
- Prediction of target genes
- Prediction of target cell types
- Prediction of mechanism for phenotype changes

➢ Genetic v. epigenetic

➢ Germline v. somatic

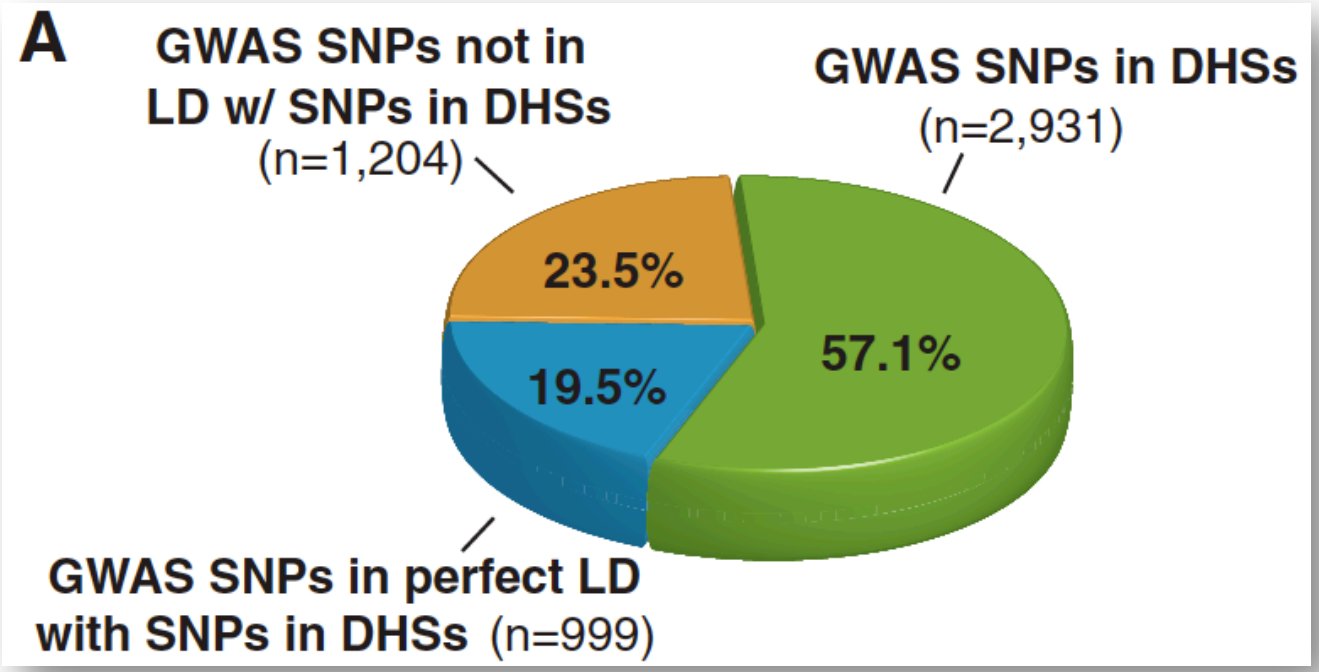
Prediction of Causal Variants

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



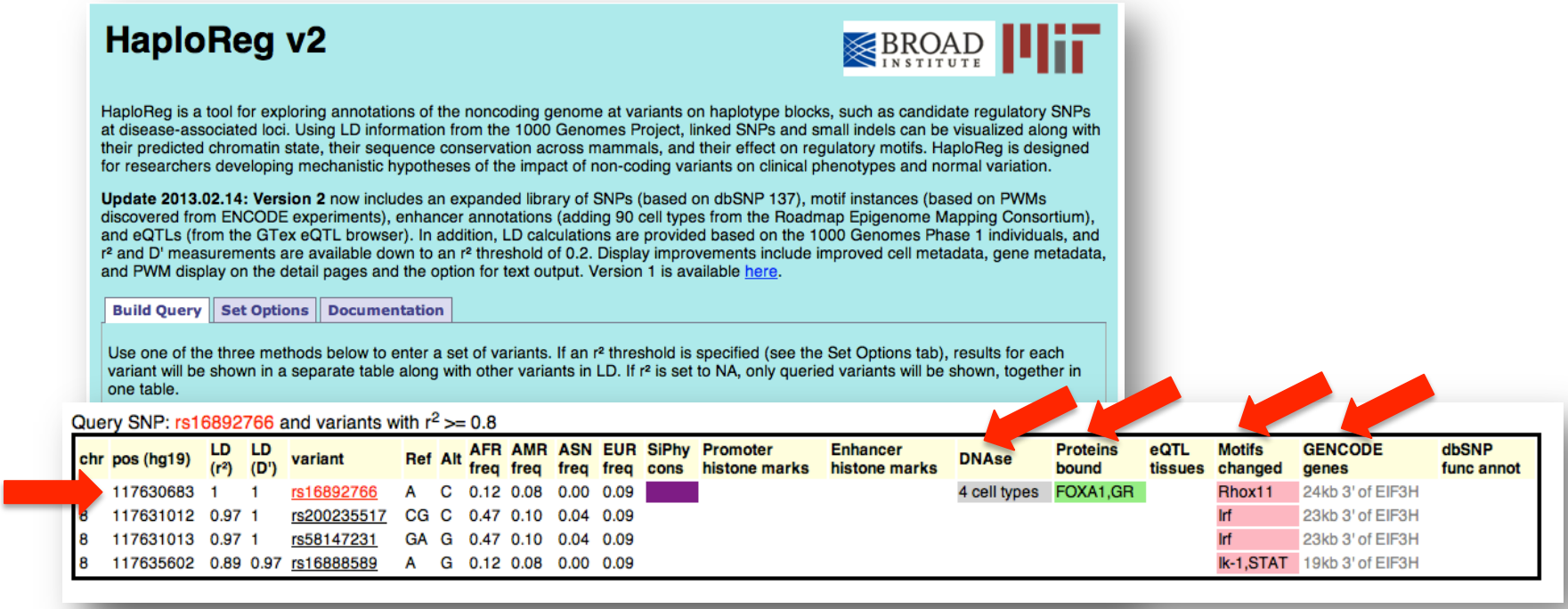
Snyder, Genome Research 22:1748, 2012

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



Stamatoyannopoulos, Science 337:1190, 2012

ENCODE/Epigenomics Data From HaploReg



HaploReg v2

BROAD INSTITUTE

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

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

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.

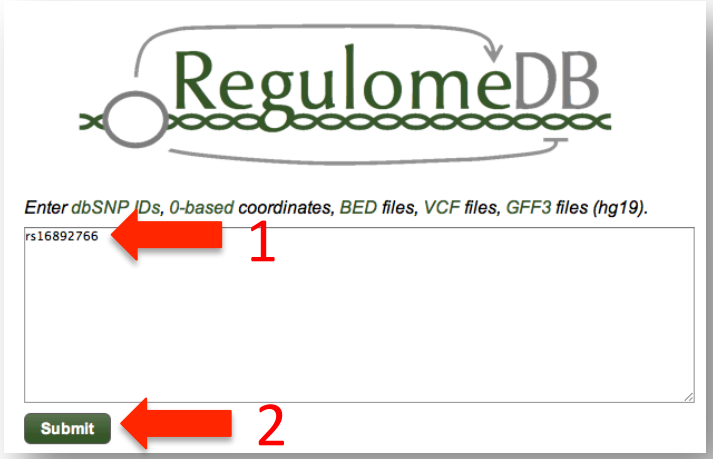
Query SNP: rs16892766 and variants with r^2 >= 0.8

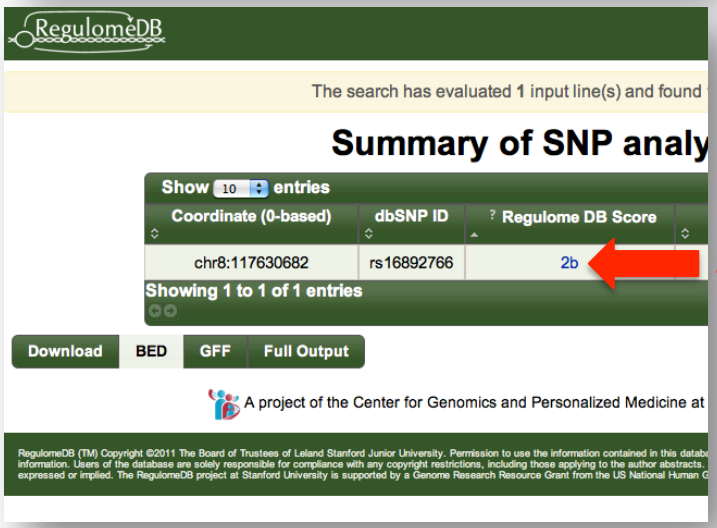
chr	pos (hg19)	LD (r^2)	LD (D')	variant	Ref	Alt	AFR freq	AMR freq	ASN freq	EUR freq	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse	Proteins bound	eQTL tissues	Motifs changed	GENCODE genes	dbSNP func annot
	117630683	1	1	rs16892766	A	C	0.12	0.08	0.00	0.09				4 cell types	FOXA1,GR		Rhox11	24kb 3' of EIF3H	
8	117631012	0.97	1	rs200235517	CG	C	0.47	0.10	0.04	0.09							Irf	23kb 3' of EIF3H	
8	117631013	0.97	1	rs58147231	GA	G	0.47	0.10	0.04	0.09							Irf	23kb 3' of EIF3H	
8	117635602	0.89	0.97	rs16888589	A	G	0.12	0.08	0.00	0.09							Ik-1,STAT	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

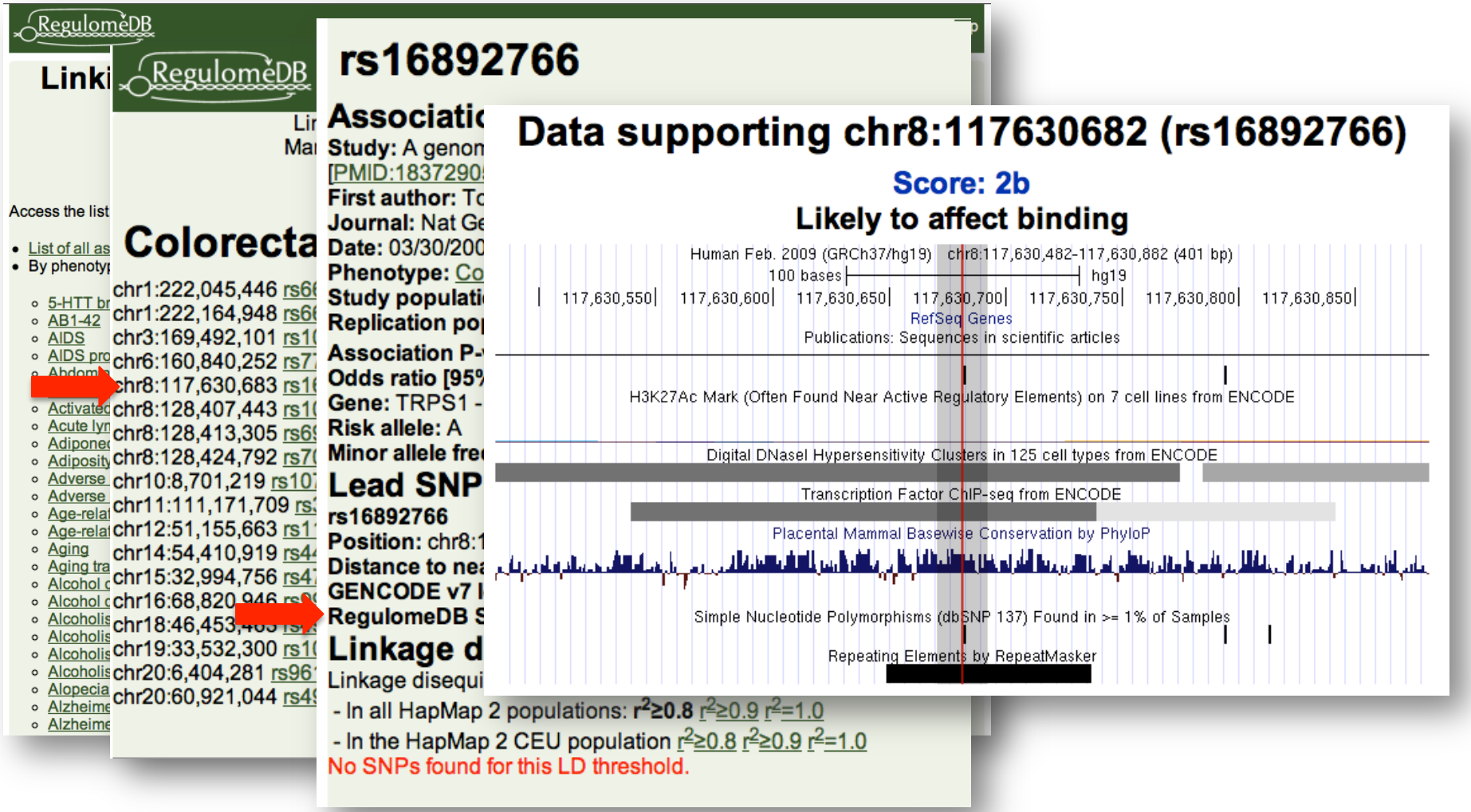
Filter:

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

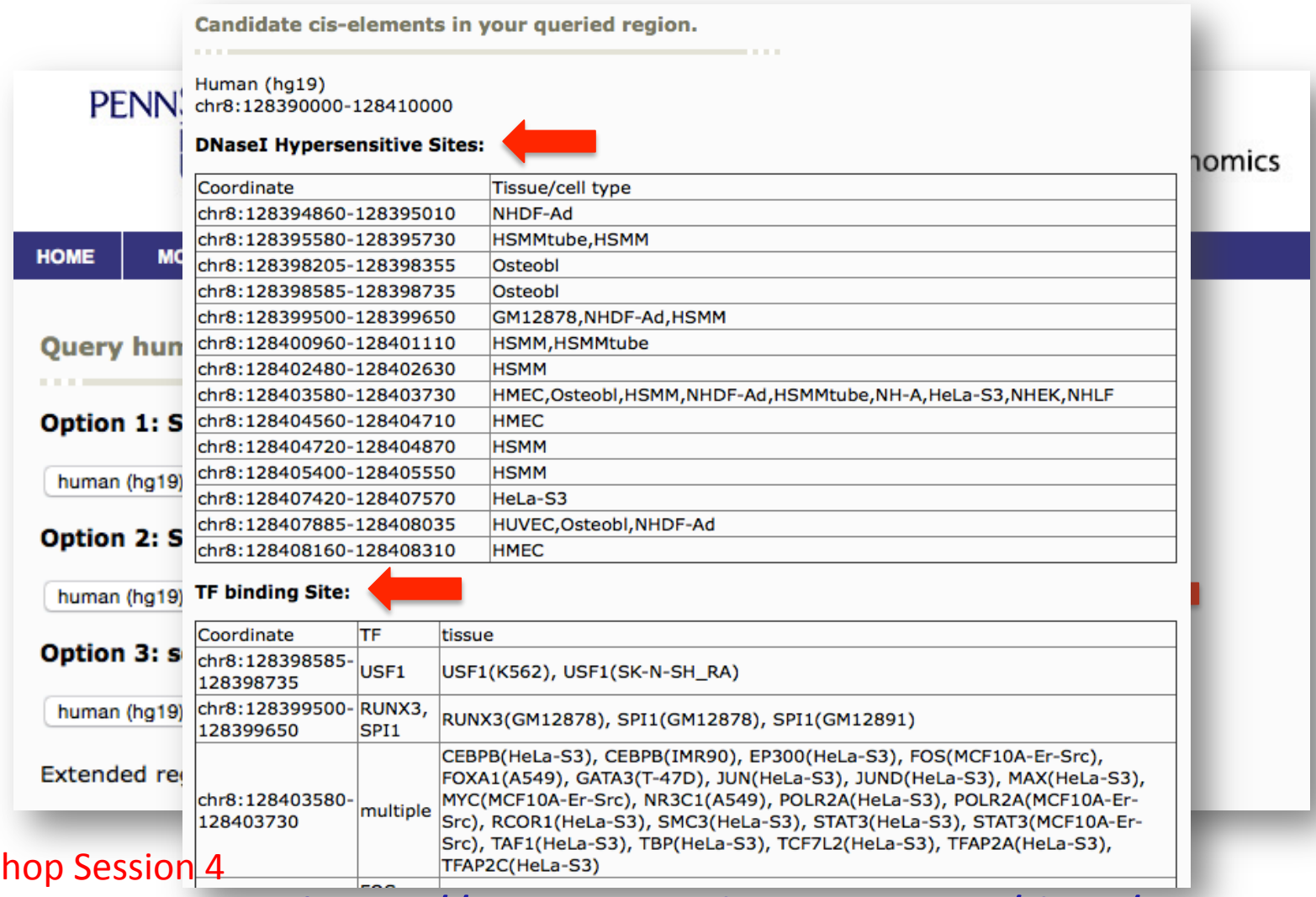
RegulomeDB GWAS Database



http://www.regulomedb.org/GWAS/

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)

Workshop Session 4

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

ENCODE Browser
Viewing Locus Of Interest

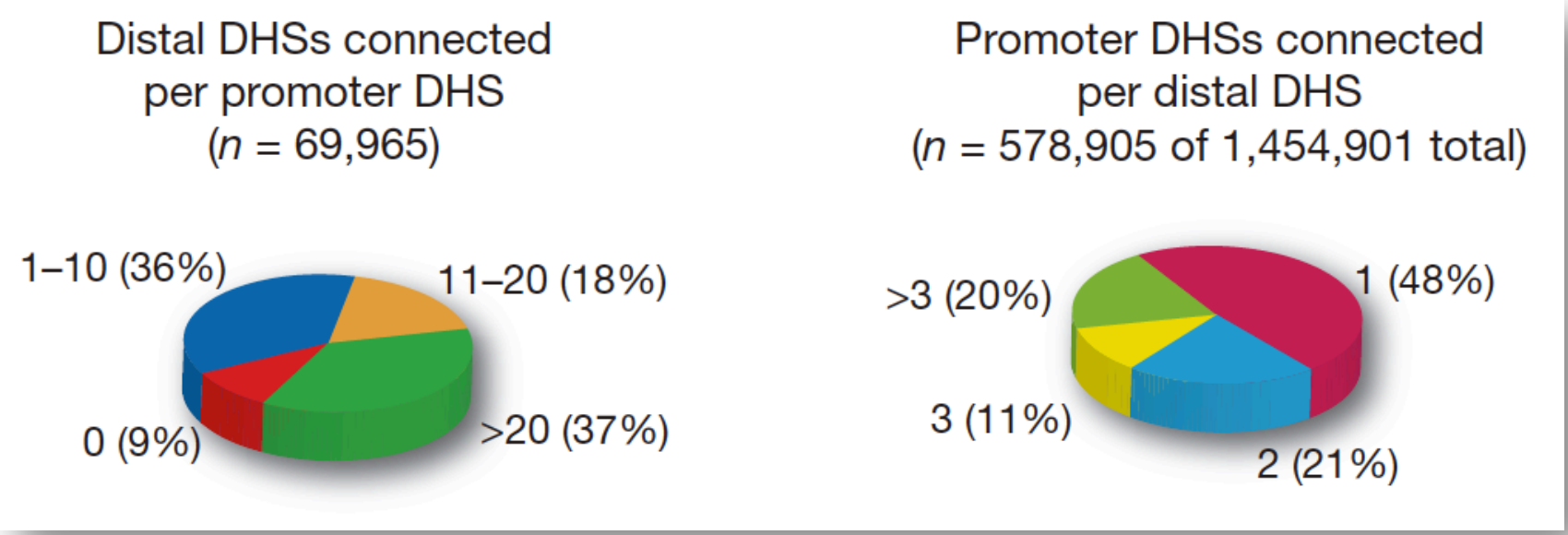
Workshop Session 1

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

"Sessions" handout at http://www.genome.gov/2755193

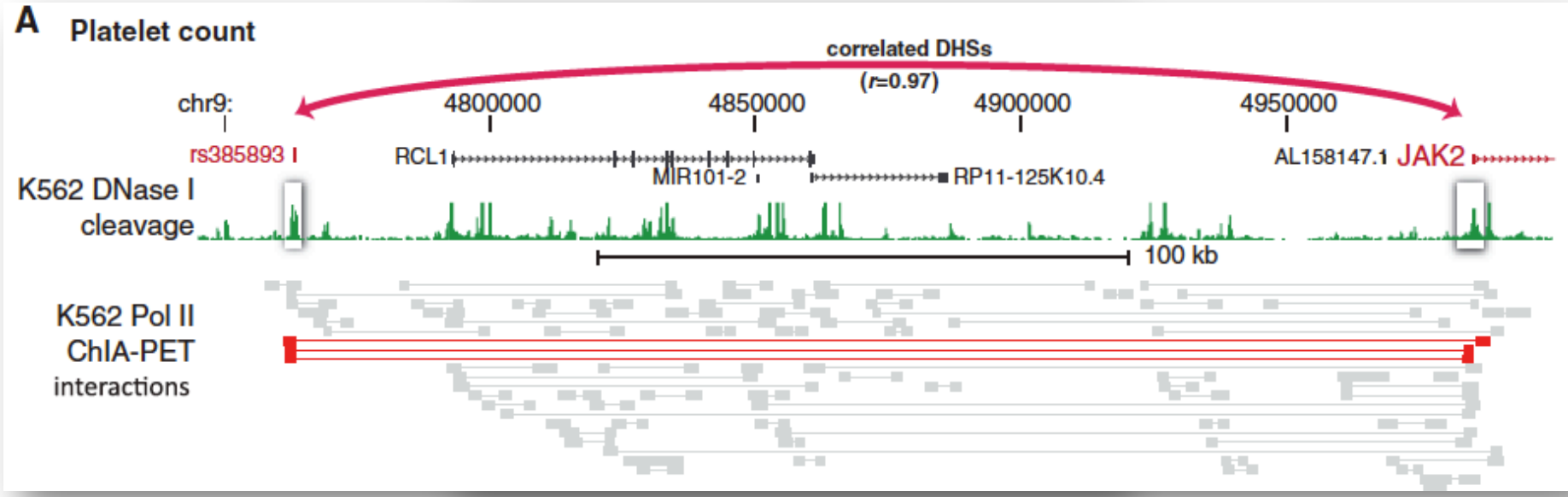
Prediction of Target Genes

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



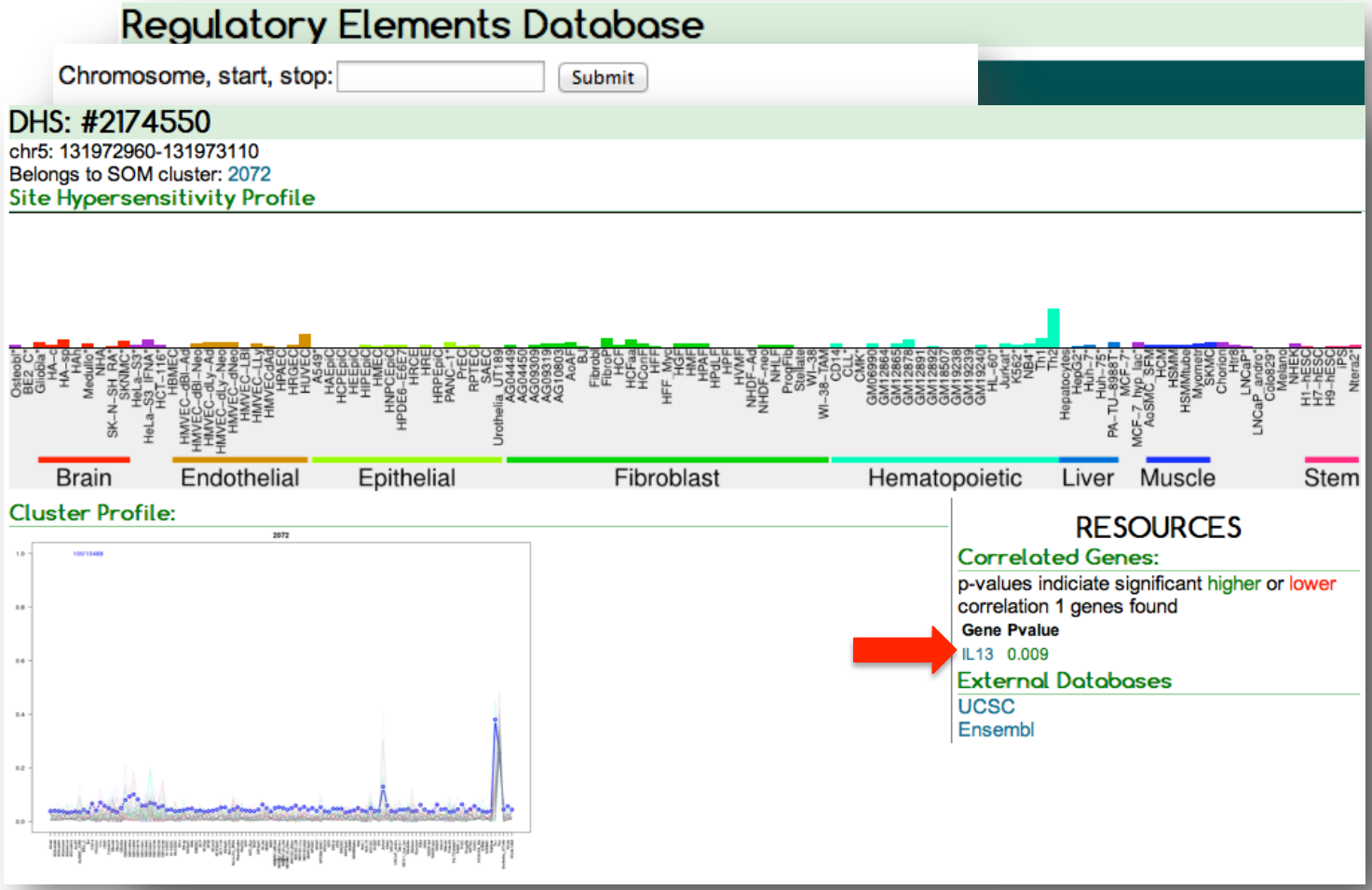
Stamatoyannopoulos, Crawford, Nature 489:75, 2012

Many GWAS Associations Lie In Regions Linked To Distal Genes



Stamatoyannopoulos, Science 337:1190, 2012

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

Cis-elements linked to your queried gene.

Human (hg19)
Gene **IL13** [NM_[illegible]2188, 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 4

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

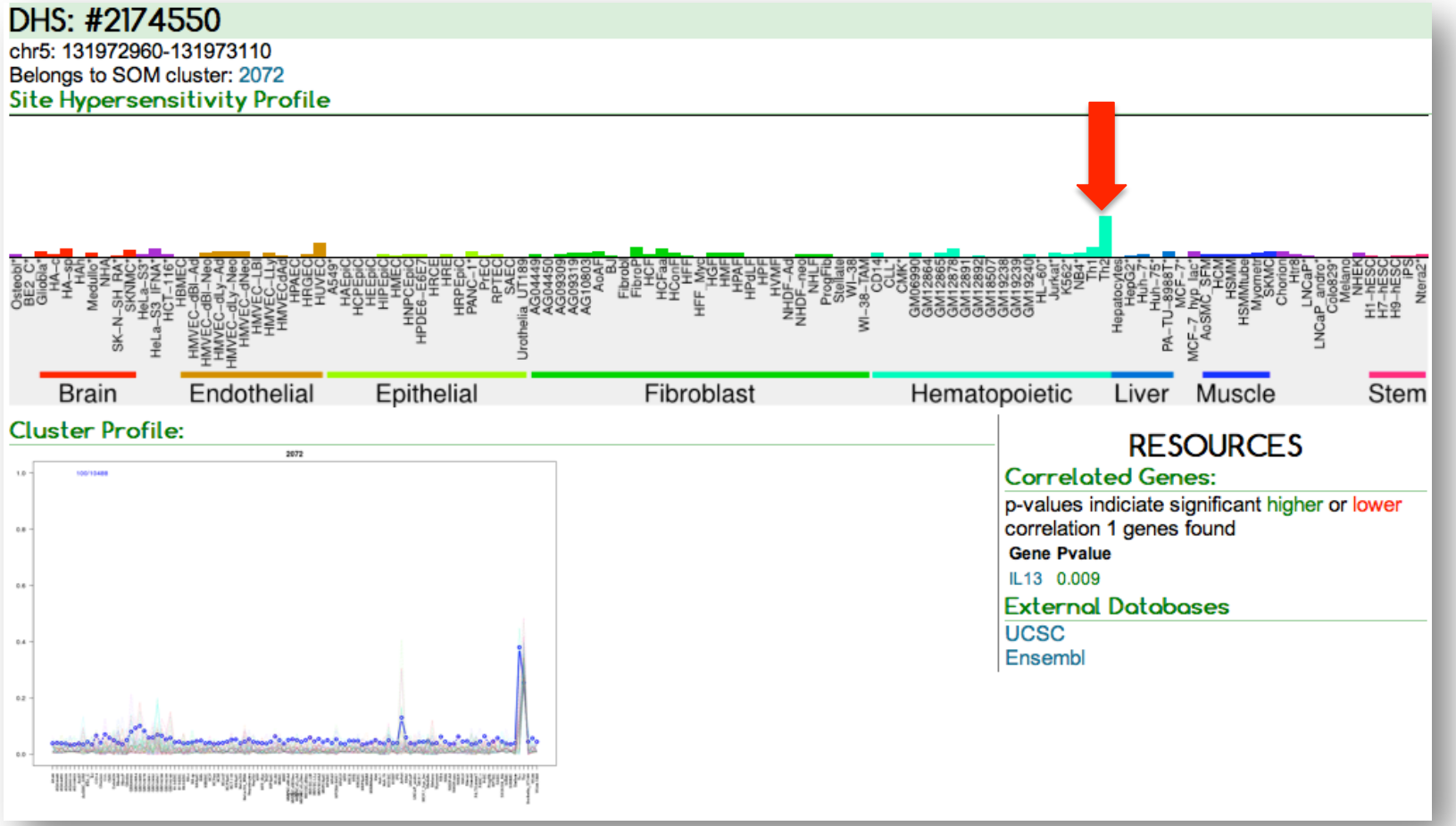
Prediction of Linkage Between Regulatory Elements and Genes

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

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

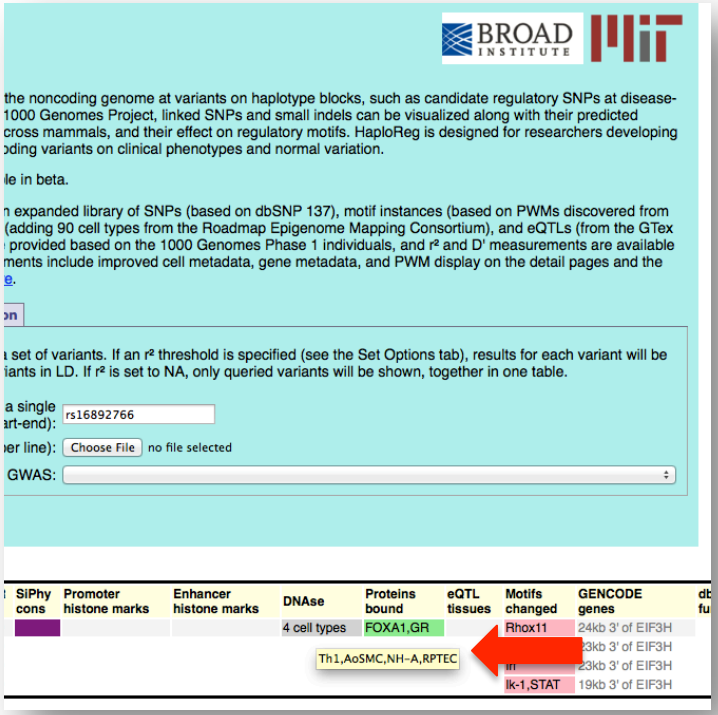
Prediction of Linkage Between Regulatory Elements and Cell Type

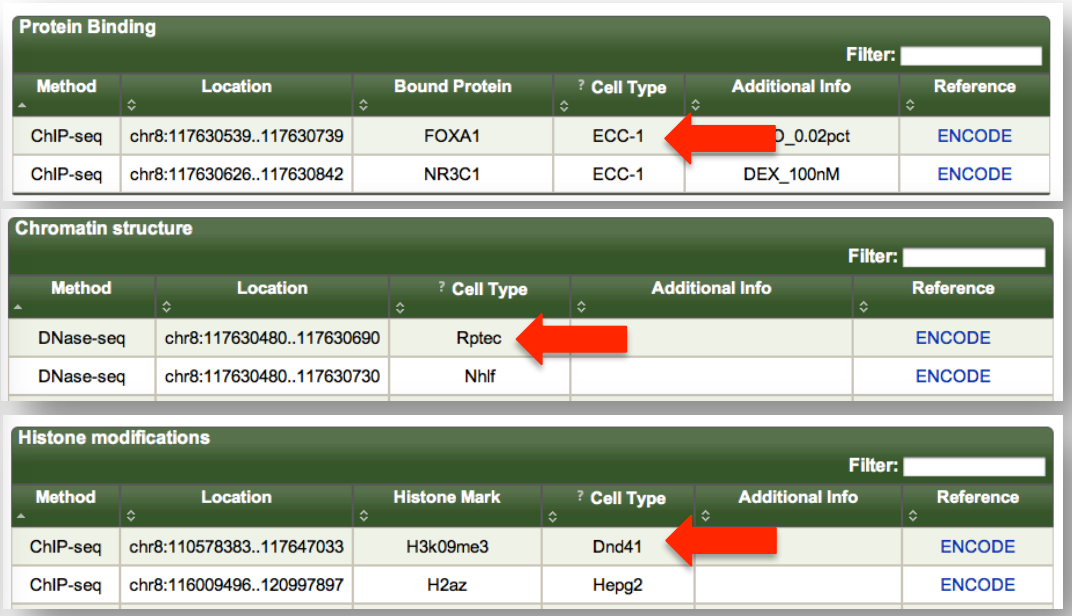


http://dnase.genome.duke.edu

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

Prediction of Linkage Between Regulatory Elements and Cell Type





Workshop Session 2

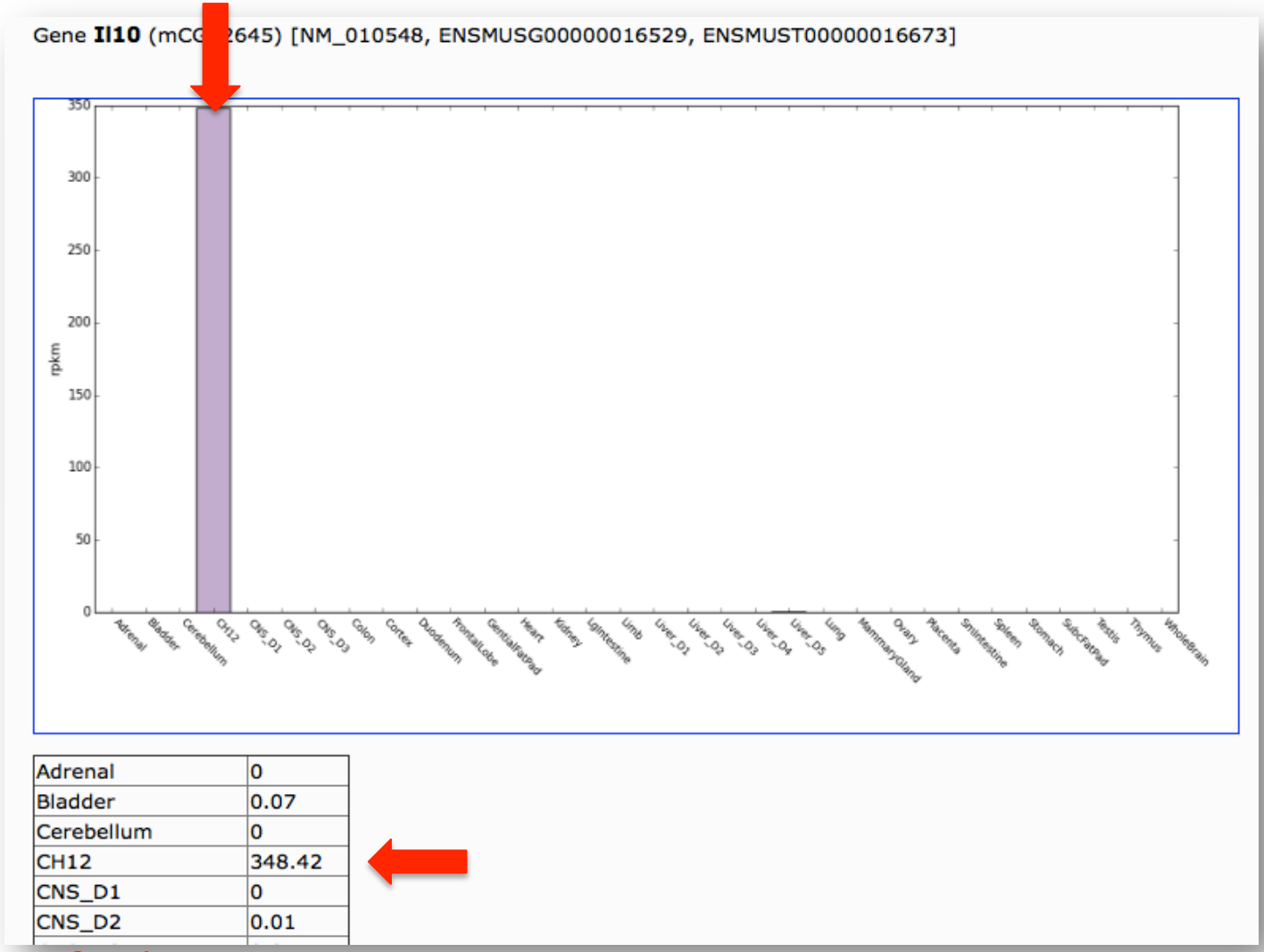
www.broadinstitute.org/mammals/haploreg/

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

http://regulomedb.org/

Cherry, Snyder, Genome Research 22-1790,2012

ENCODE cis-element Browser

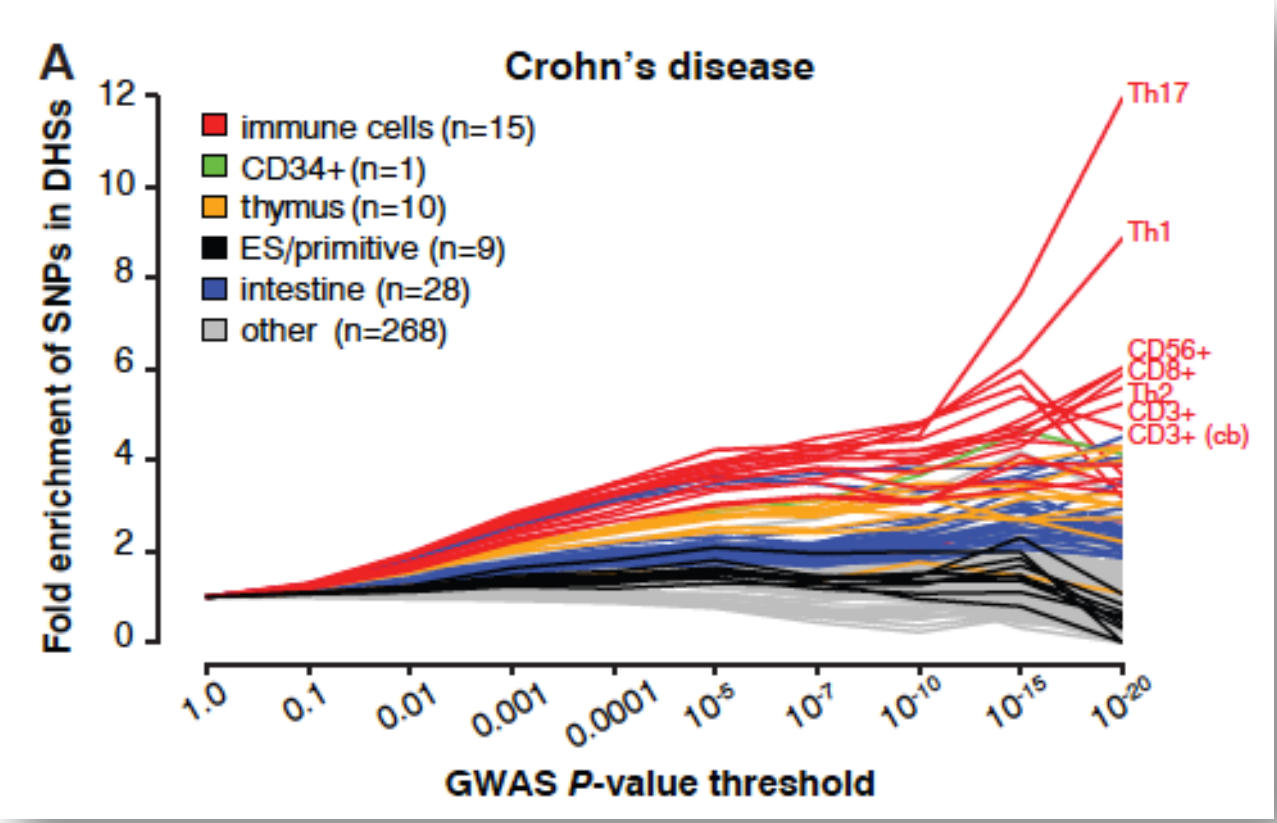


Adrenal	0
Bladder	0.07
Cerebellum	0
CH12	348.42
CNS_D1	0
CNS_D2	0.01

Workshop Session 4

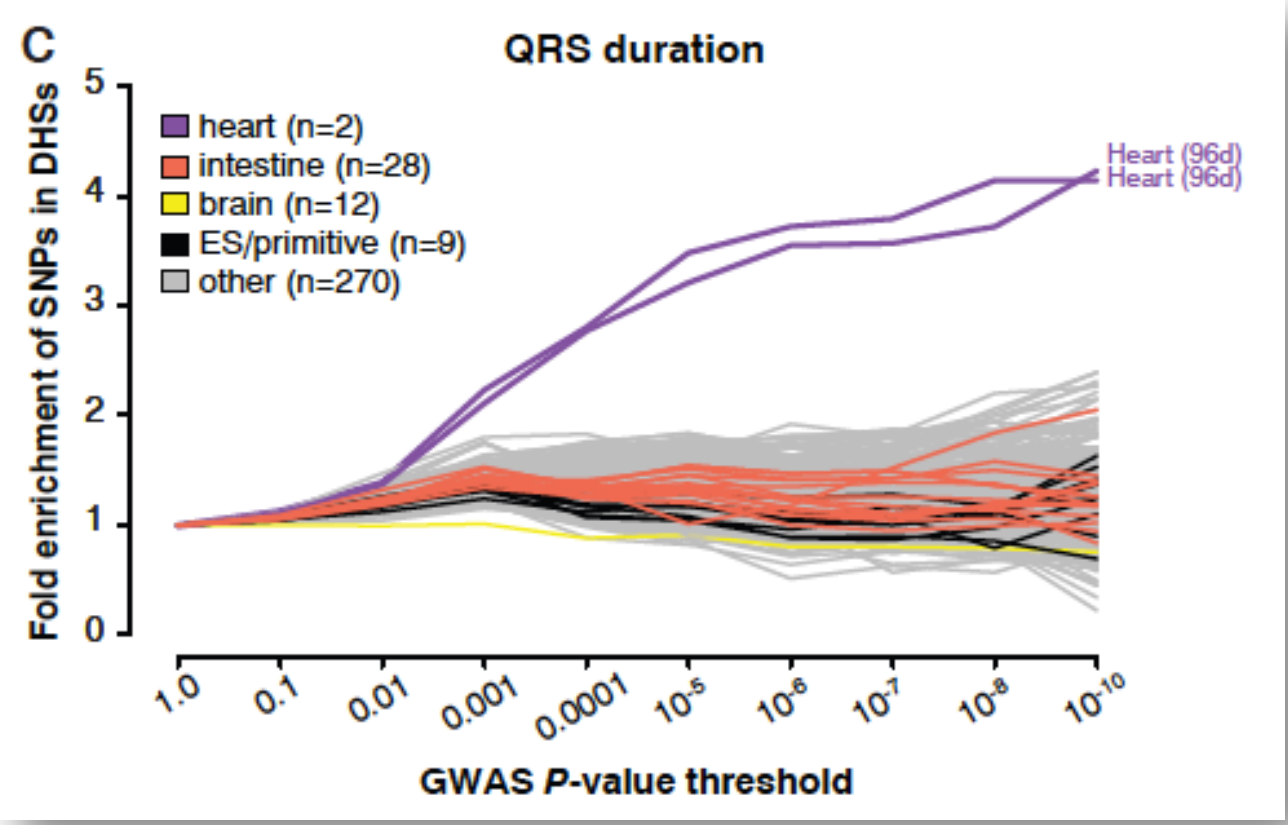
https://www.encodeproject.org/data/annotations/

ENCODE And Epigenomics Data Can Be Used To Predict Cell Types



Stamatoyannopoulos, Science 337:1190, 2012

ENCODE And Epigenomics Data Can Be Used To Identify Variants



Stamatoyannopoulos, Science 337:1190, 2012

Summary- ENCODE Use Cases

Major use: Hypothesis generation and refinement

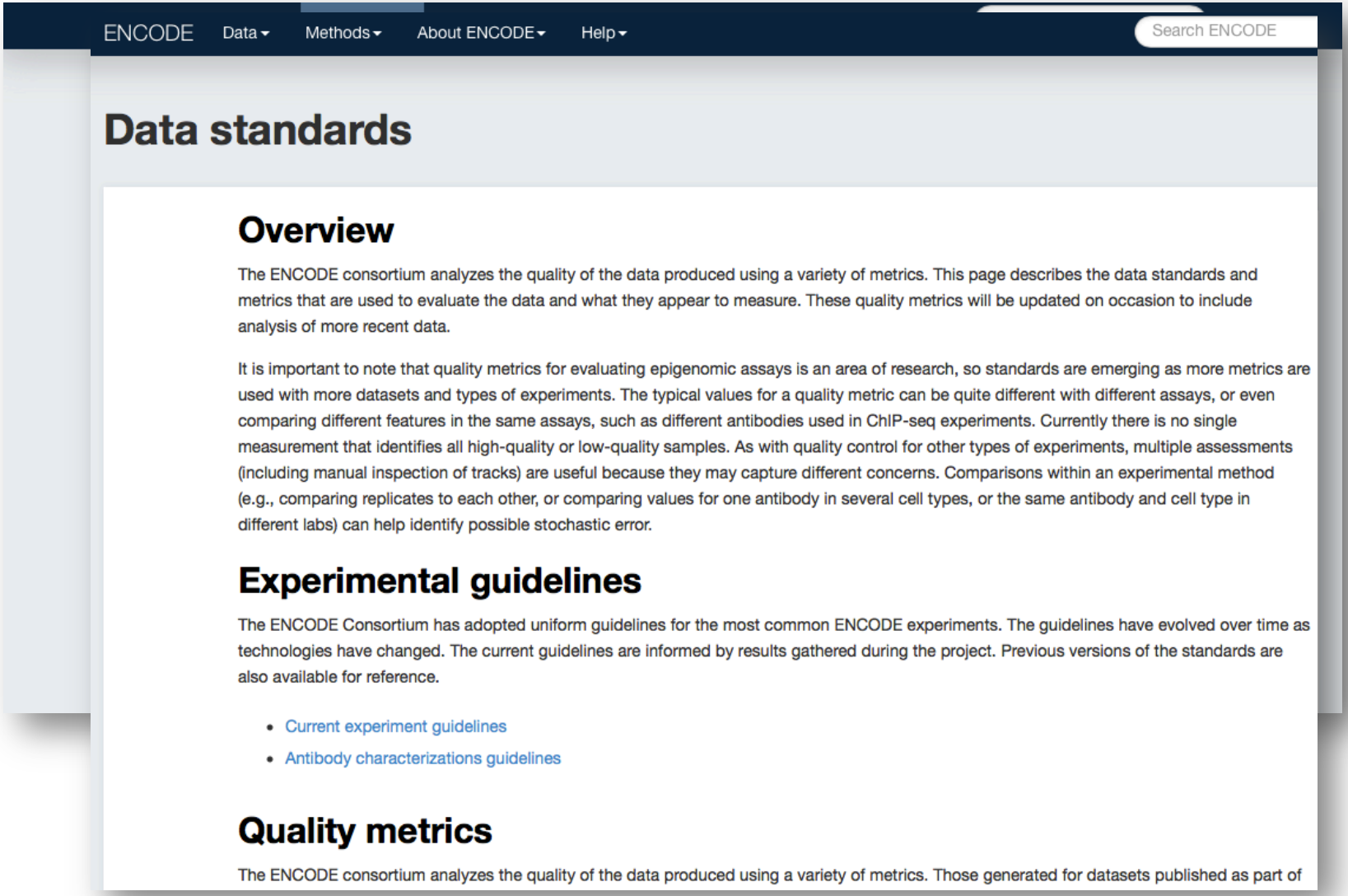
- Prediction of causal variants/regulatory elements
- Prediction of target genes
- Prediction of target cell types
- Prediction of mechanism for phenotype changes

Overview

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

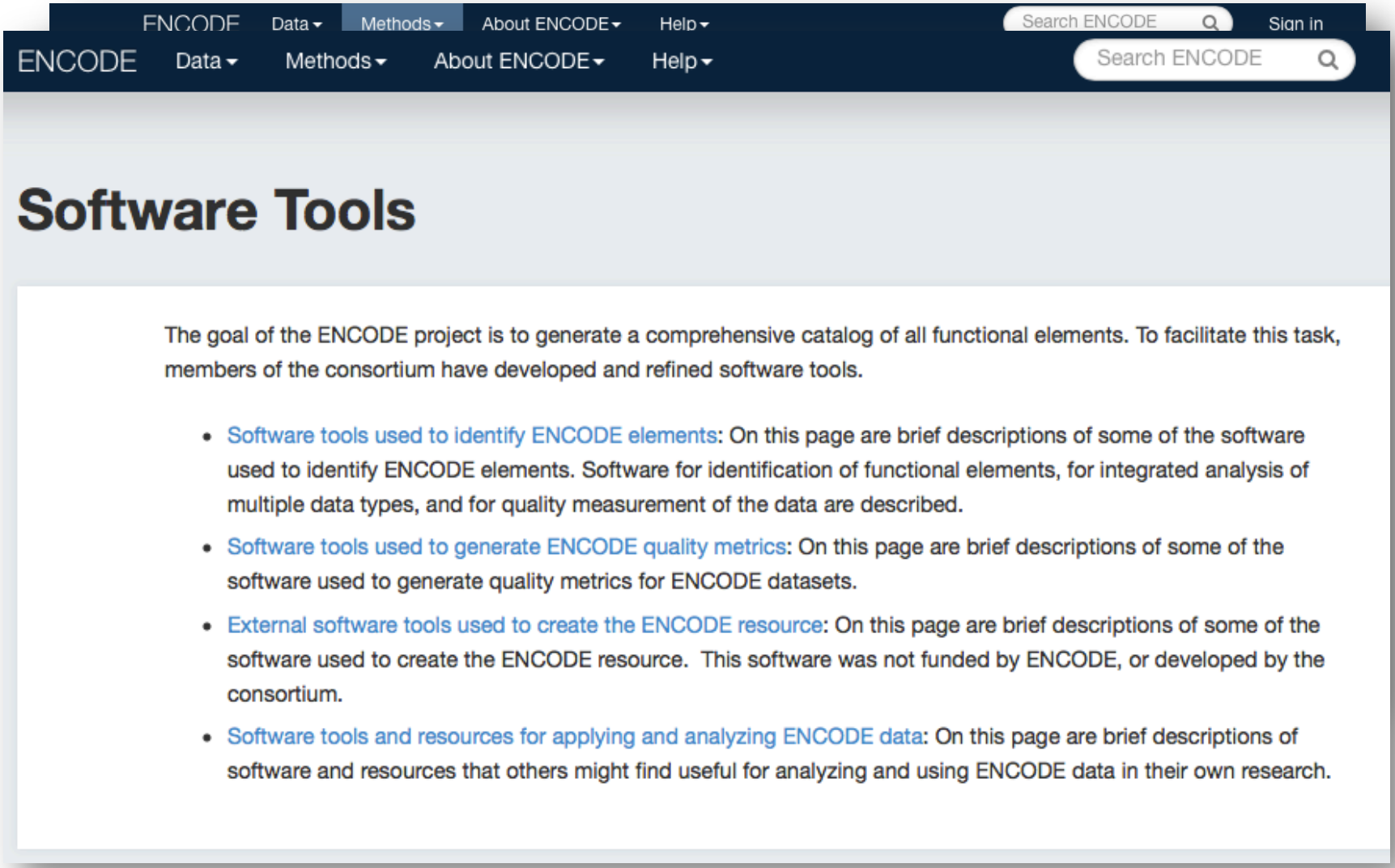
Workshop Session 1

ENCODE Data Standards



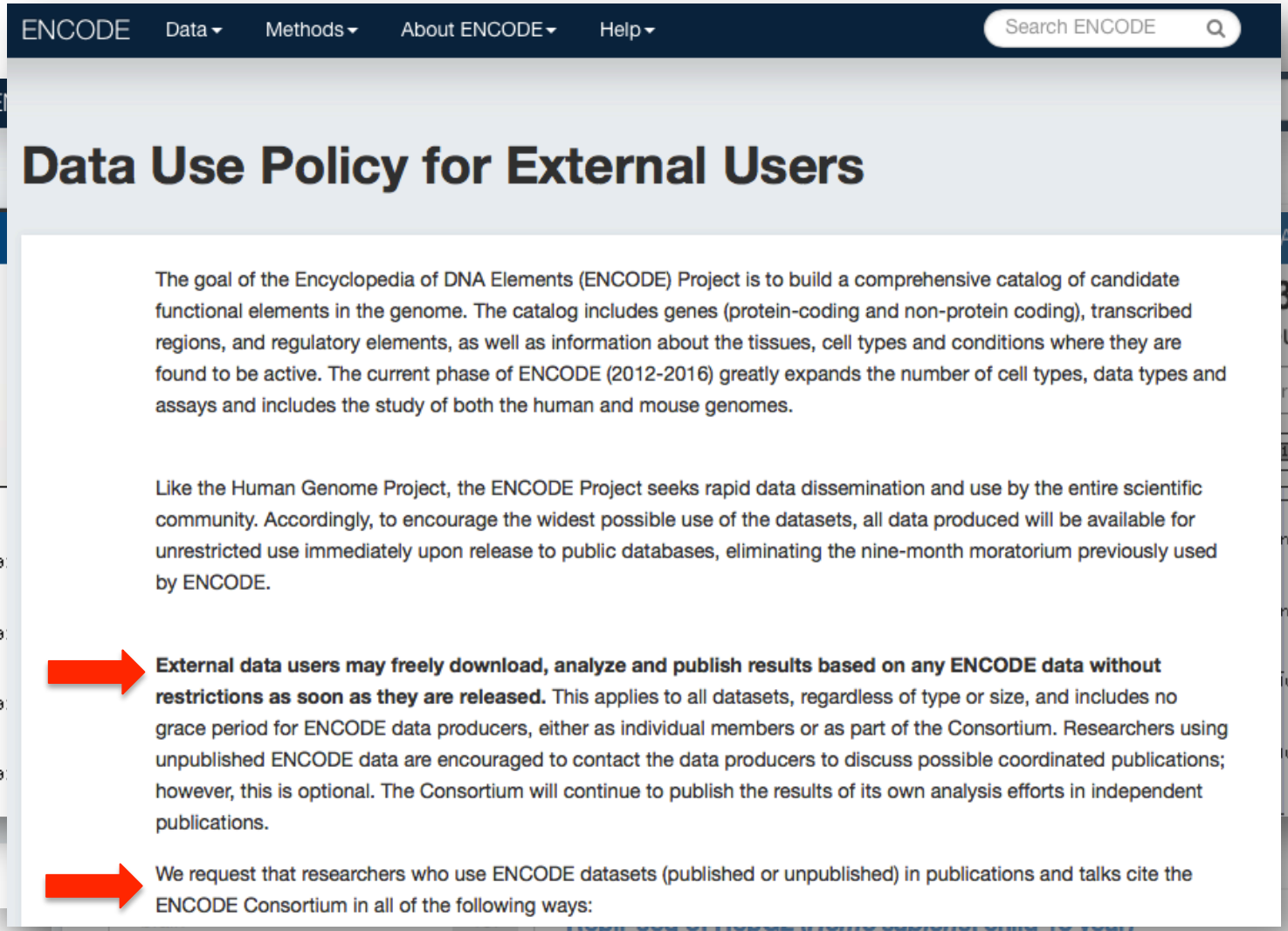
https://www.encodeproject.org

ENCODE Software Tools



https://www.encodeproject.org

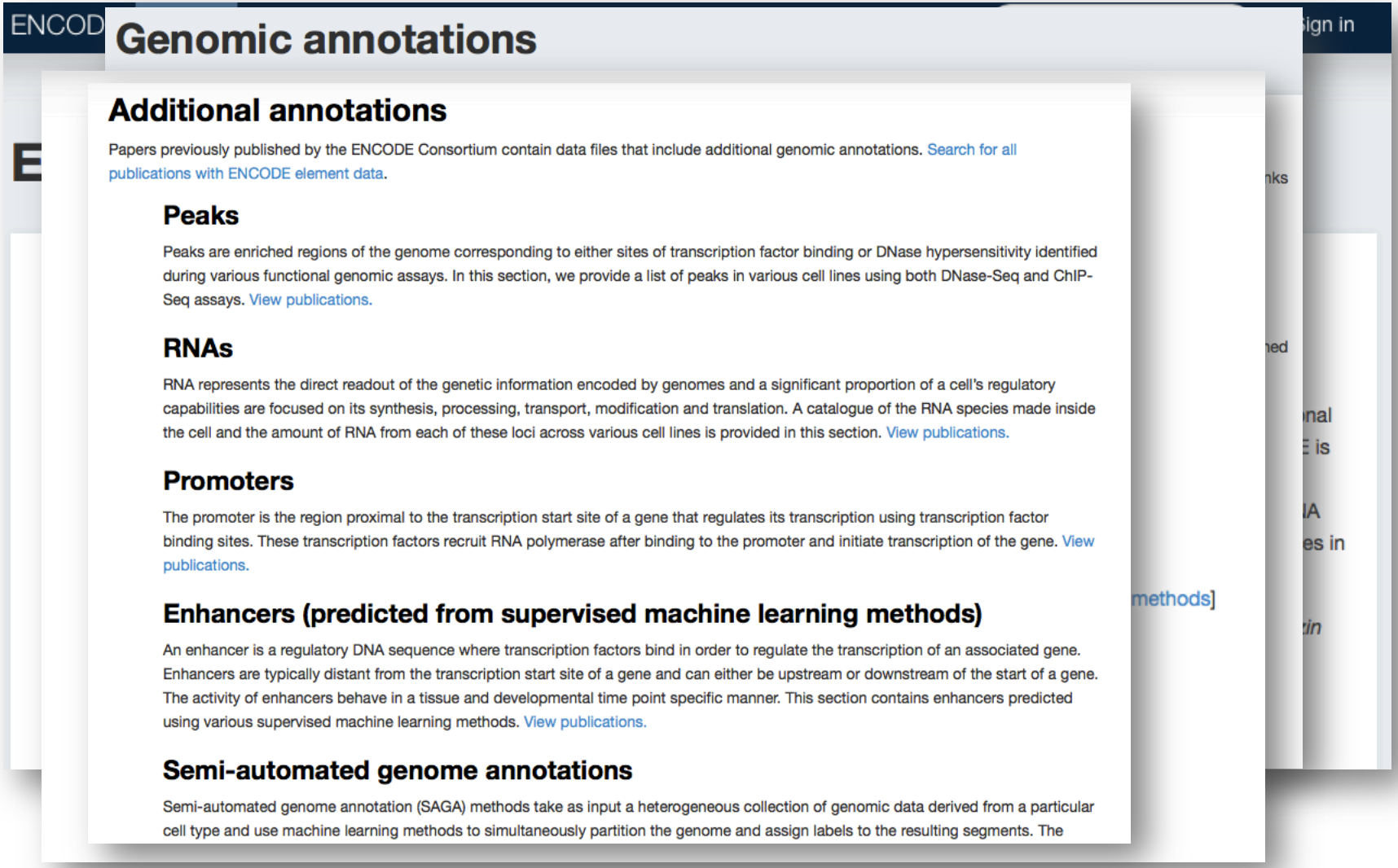
Downloading and Visualizing



Workshop Session 1

https://www.encodeproject.org

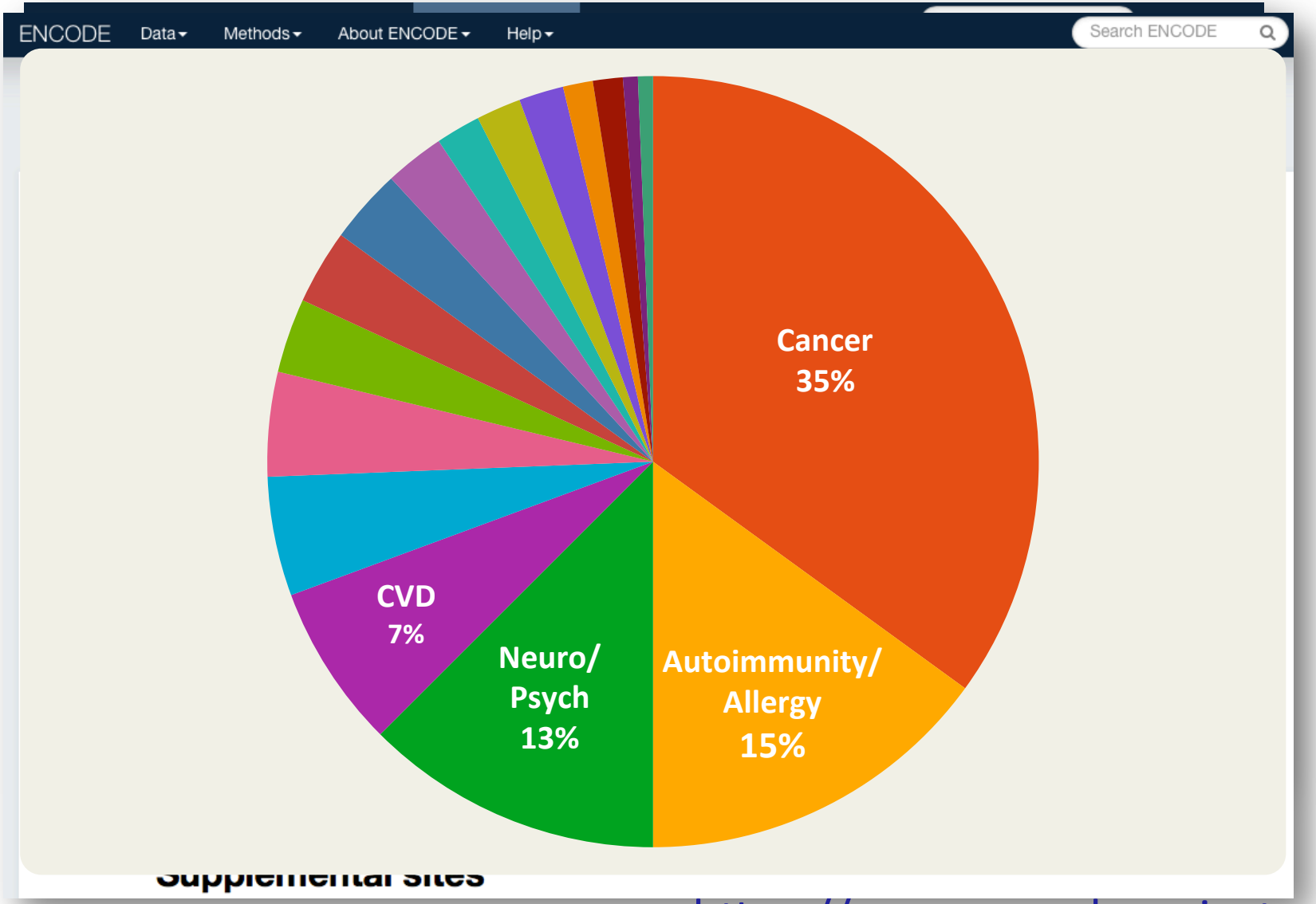
ENCODE Encyclopedia Prototype



Workshop Session 4

https://www.encodeproject.org

Publications



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 7 consortia

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

ENCODE Consortium

Elise Feingold

Peter Good

Dan Gilchrist


ENCODE