The ENCODE Encyclopedia

&

Variant Annotation Using RegulomeDB and HaploReg

Jill E. Moore
Weng Lab
University of Massachusetts Medical School
October 10, 2015

Where's the Encyclopedia?

- ENCODE: **Enc**yclopedia **O**f **D**NA **E**lements

- So far ENCODE data producers have generated thousands of experiments in humans
  - 200+ DNase-seq
  - 800+ Transcription Factor (TF) ChIP-seq
  - 300+ Histone Mark ChIP-seq
  - RNA-seq, RNA-binding, DNAme

- How do we:
  - Integrate different experiments and assays?
  - Find functional annotations
  - Build and visualize the encyclopedia?

Genomic Annotations

- Gene expression
- Transcription start sites (TSS)
- Uniformly processed peaks from DNase-seq, histone mark ChIP-seq, and TF ChIP-seq
- 3D chromatin contacts from Hi-C and ChIA-PET
- Candidate enhancers and promoters
- Semi-automated genome annotations (ChromHMM and Segway)
- Target genes of regulatory elements

Genomic Annotations

- Gene expression
- Transcription start sites (TSS)
- Uniformly processed peaks from DNase-seq, histone mark ChIP-seq, and TF ChIP-seq
- 3D chromatin contacts from Hi-C and ChIA-PET
- Candidate enhancers and promoters
- Semi-automated genome annotations (ChromHMM and Segway)
- Target genes of regulatory elements

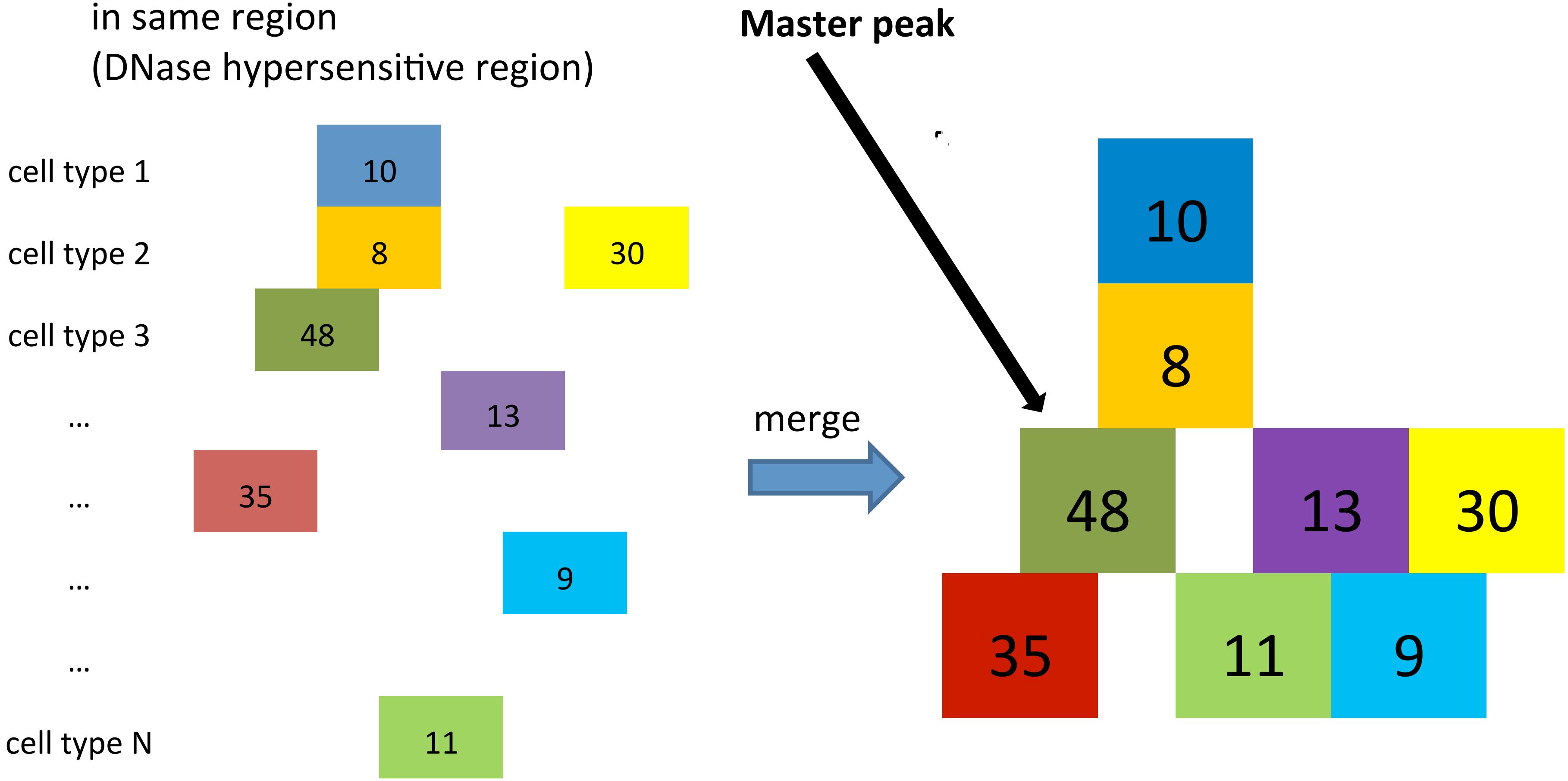
Step 1: Define DNase Master Peaks (MPs)

Master peaks:

- Are a set of unique, non-overlapping peaks
- Are a “representative” peak in a region of overlapping peaks
- Span all datasets
- Collectively cover ~20% of the genome
- Incorporates ENCODE and Roadmap DNase data

Step 1: Define DNase Master Peaks (MPs)

Peaks present across cell types
in same region
(DNase hypersensitive region)



Master peak file created by Stam lab (UW)

Step 2: Separate DNase MPs by Genetic Context

DNase master peaks are separated into:

- TSS-proximal = within a 2kb window centered on any GENCODE V19 transcription start site (TSS)

- TSS-distal = all other peaks

Step 3: Annotate DNase MPs

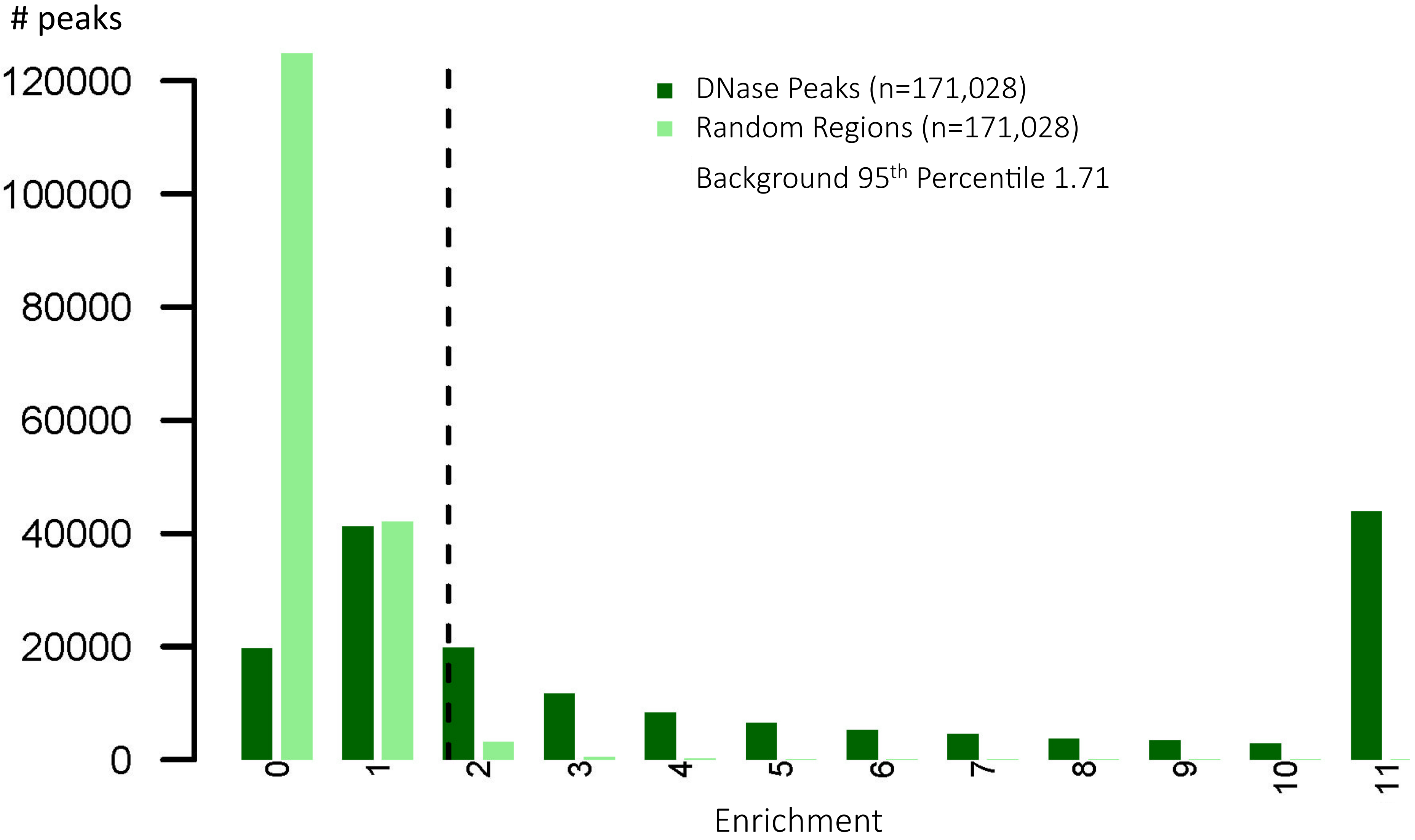
- Intersect with TF ChIP-seq peaks from all cell types

- Enrichment in histone mark signal:

  - For each master peak, we calculated histone signal in a 1000 bp window centered on the peak

  - We converted signal percentile using a background distribution calculated from randomly chosen 1000-bp genomic regions (excluding DNase peaks and ENCODE blacklist regions)

Enrichment of TSS-distal DNase MPs in GM12878 with H3K27ac Signal from GM12878



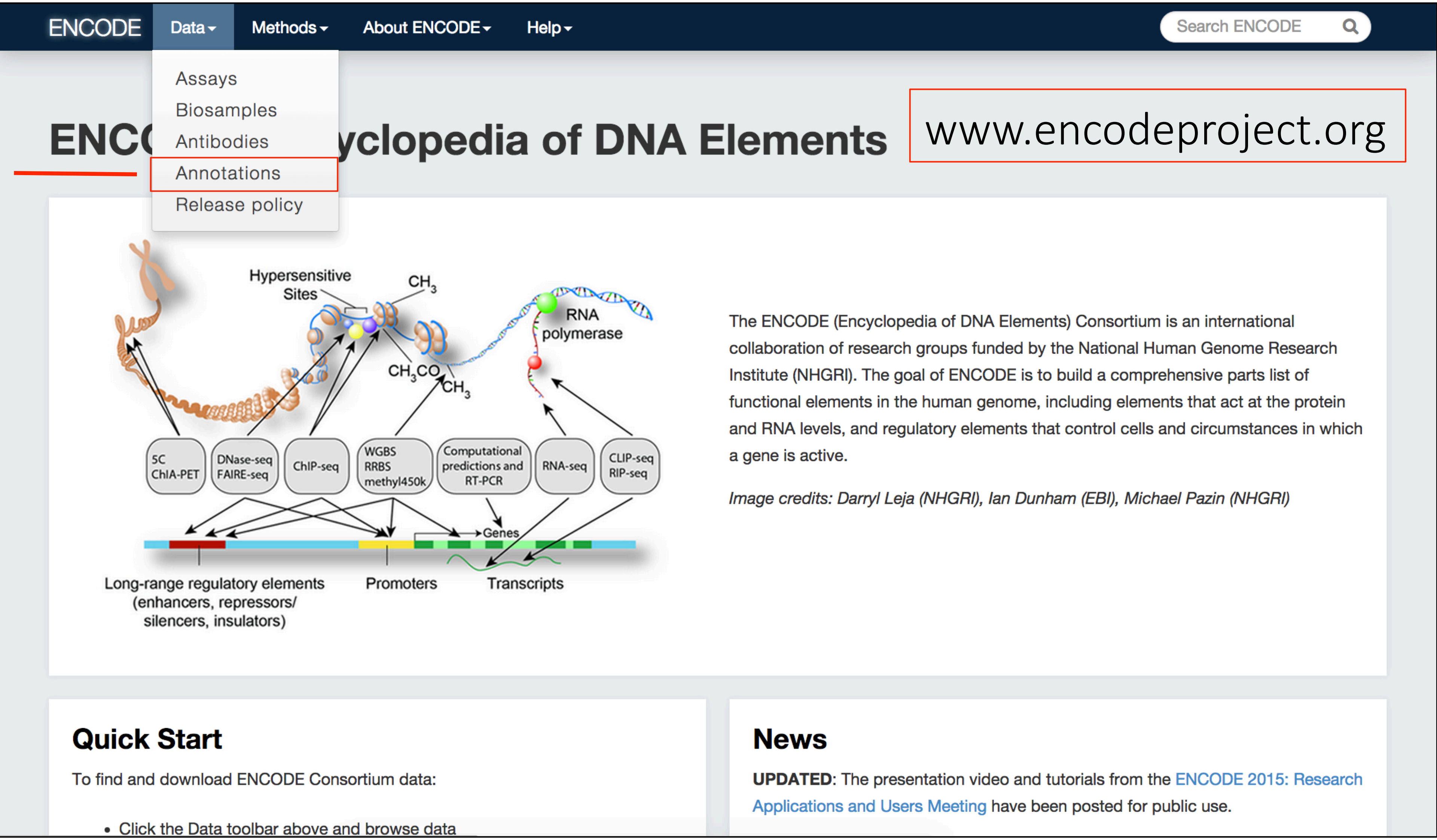
Selection of Histone Marks

- H3K4me3 - enriched at actively transcribed promoters
- H3K9ac - enriched at promoters and enhancers
- H3K27ac   enriched at active enhancers
- H3K4me1 - enriched at enhancers (both active and poised)

Current Annotations

- Proximal Regulatory Elements = proximal DNase MPs
- Distal Regulatory Elements = distal DNase MPs
- Proximal TF Binding = proximal DNase MPs + TF peaks
- Distal TF Binding = distal DNase MPs + TF peaks
- Candidate Promoters = proximal DNase MPs + enrichment in histone mark
- Candidate Enhancers = distal DNase MPs + enrichment in histone mark

How can I access these annotations?



www.encodeproject.org

ENCODE: Encyclopedia of DNA Elements

The ENCODE (Encyclopedia of DNA Elements) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

Image credits: Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

Quick Start

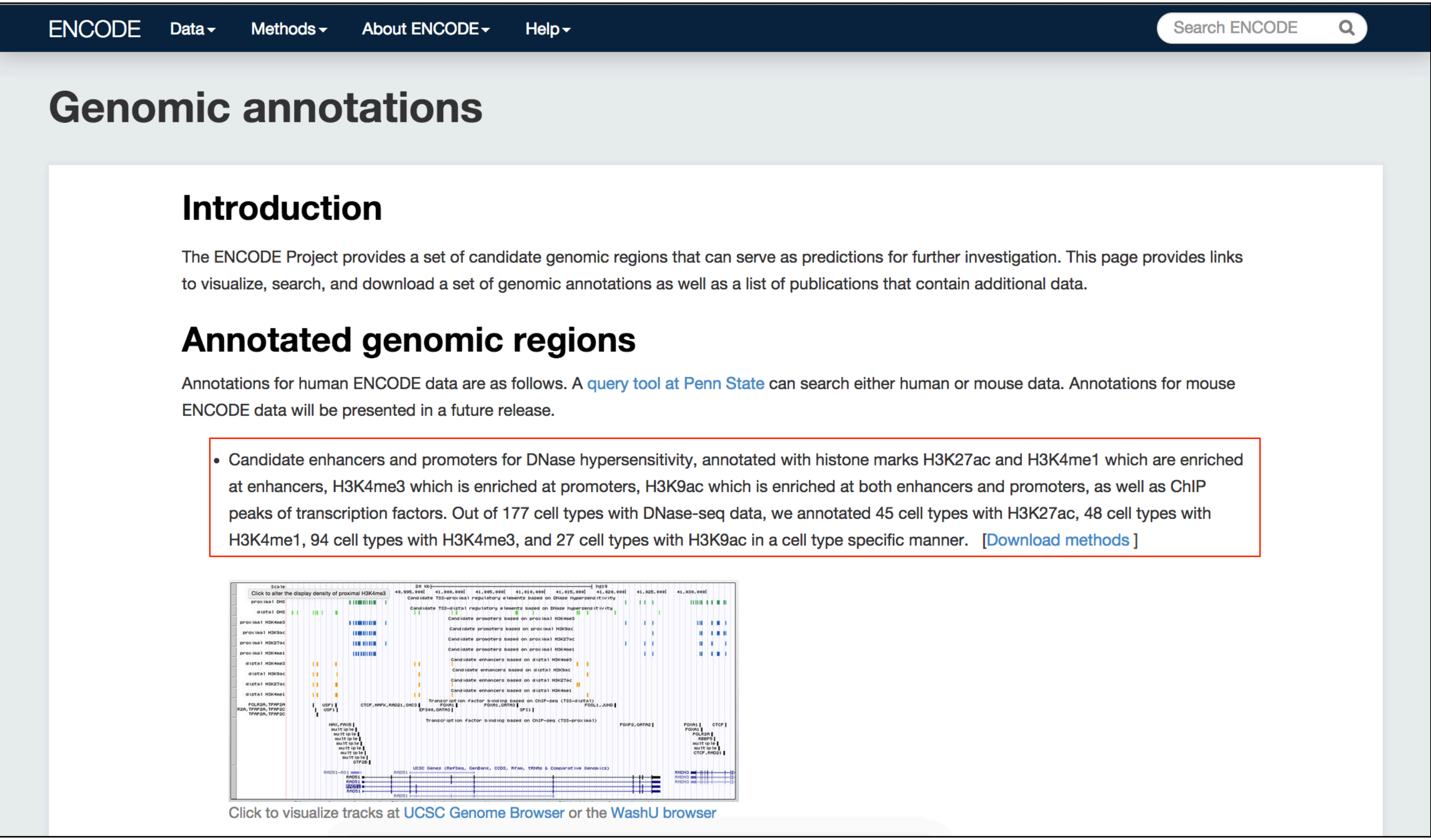
To find and download ENCODE Consortium data:

- Click the Data toolbar above and browse data

News

UPDATED: The presentation video and tutorials from the ENCODE 2015: Research Applications and Users Meeting have been posted for public use.

How can I access these annotations?



Genomic annotations

Introduction

The ENCODE Project provides a set of candidate genomic regions that can serve as predictions for further investigation. This page provides links to visualize, search, and download a set of genomic annotations as well as a list of publications that contain additional data.

Annotated genomic regions

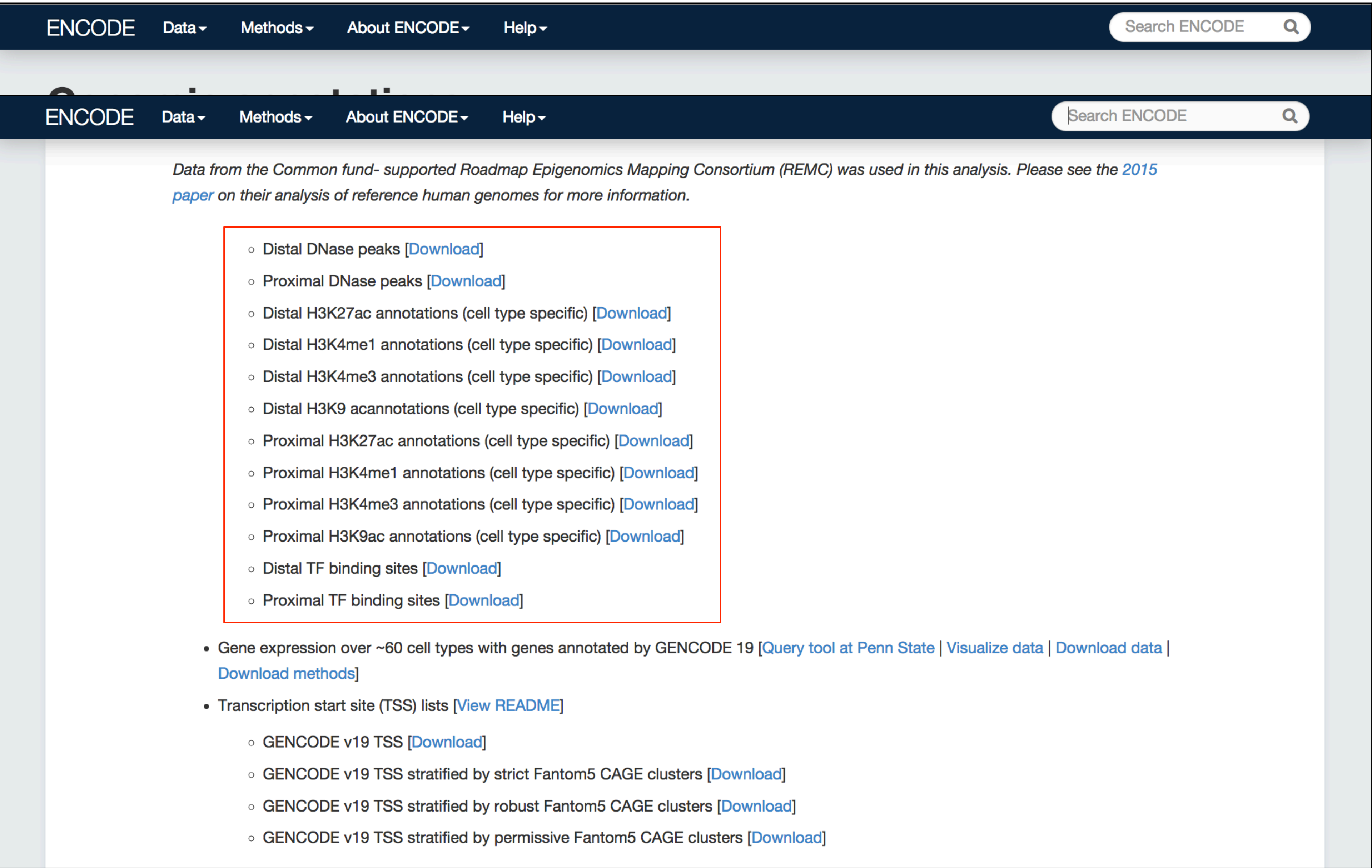
Annotations for human ENCODE data are as follows. A query tool at Penn State can search either human or mouse data. Annotations for mouse ENCODE data will be presented in a future release.

- Candidate enhancers and promoters for DNase hypersensitivity, annotated with histone marks H3K27ac and H3K4me1 which are enriched at enhancers, H3K4me3 which is enriched at promoters, H3K9ac which is enriched at both enhancers and promoters, as well as ChIP peaks of transcription factors. Out of 177 cell types with DNase-seq data, we annotated 45 cell types with H3K27ac, 48 cell types with H3K4me1, 94 cell types with H3K4me3, and 27 cell types with H3K9ac in a cell type specific manner. [Download methods]

Click to alter the display density of proximal H3K4me3
proximal DHS
distal DHS
proximal H3K4me3
proximal H3K9ac
proximal H3K27ac
proximal H3K4me1
distal H3K4me3
distal H3K9ac
distal H3K27ac
distal H3K4me1
Candidate TSS-proximal regulatory elements based on DNase hypersensitivity
Candidate TSS-distal regulatory elements based on DNase hypersensitivity
Candidate promoters based on proximal H3K4me3
Candidate promoters based on proximal H3K9ac
Candidate promoters based on proximal H3K27ac
Candidate promoters based on proximal H3K4me1
Candidate enhancers based on distal H3K4me3
Candidate enhancers based on distal H3K9ac
Candidate enhancers based on distal H3K27ac
Candidate enhancers based on distal H3K4me1
Transcription factor binding based on ChIP-seq (TSS-distal)
Transcription factor binding based on ChIP-seq (TSS-proximal)
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
RAD51

Click to visualize tracks at UCSC Genome Browser or the WashU browser

How can I access these annotations?

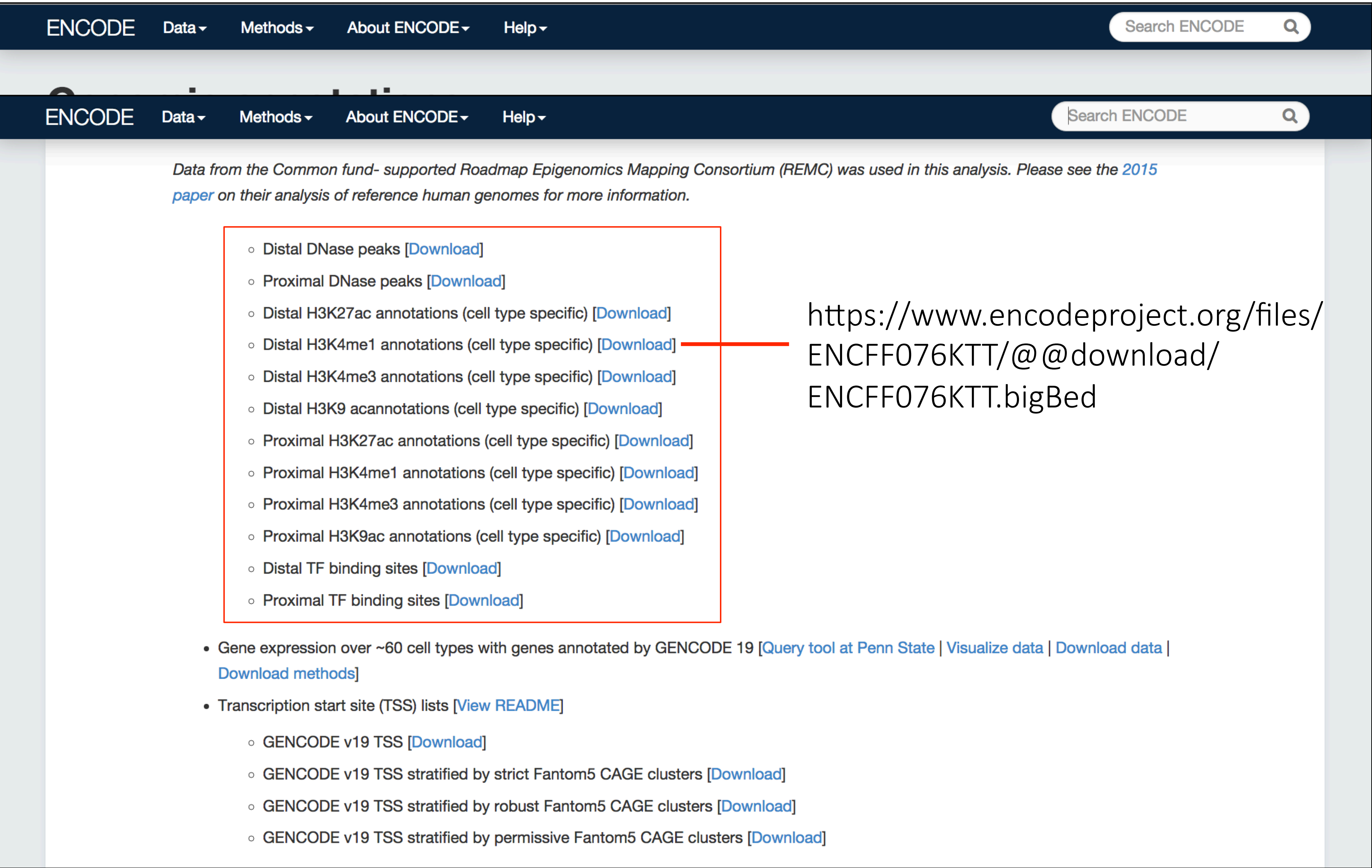


Data from the Common fund- supported Roadmap Epigenomics Mapping Consortium (REMC) was used in this analysis. Please see the 2015 paper on their analysis of reference human genomes for more information.

- 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 H3K9 acannotations (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 over ~60 cell types with genes annotated by GENCODE 19 [Query tool at Penn State | Visualize data | 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]

How can I access these annotations?



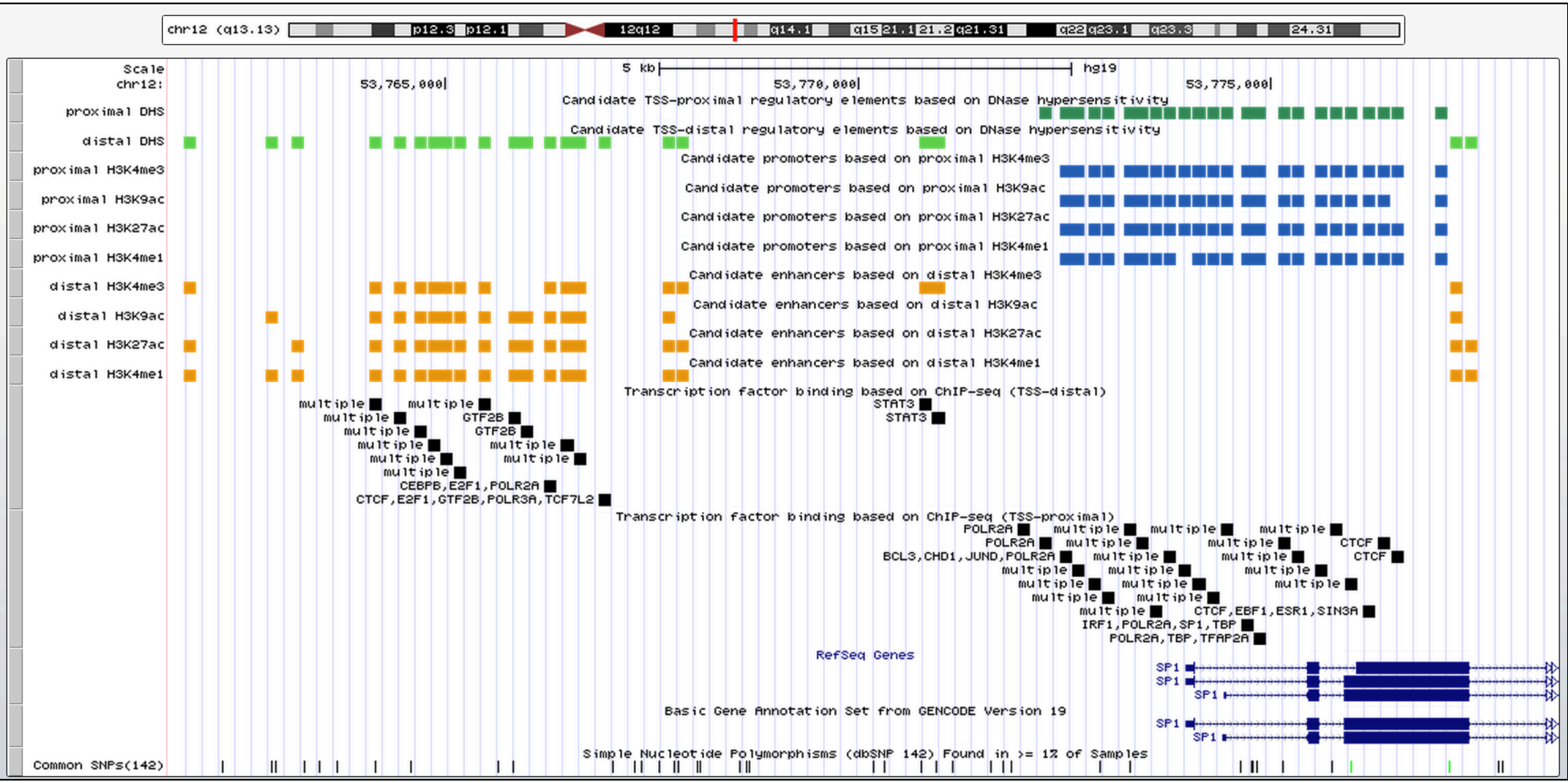
Data from the Common fund- supported Roadmap Epigenomics Mapping Consortium (REMC) was used in this analysis. Please see the 2015 paper on their analysis of reference human genomes for more information.

- 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 H3K9 acannotations (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]

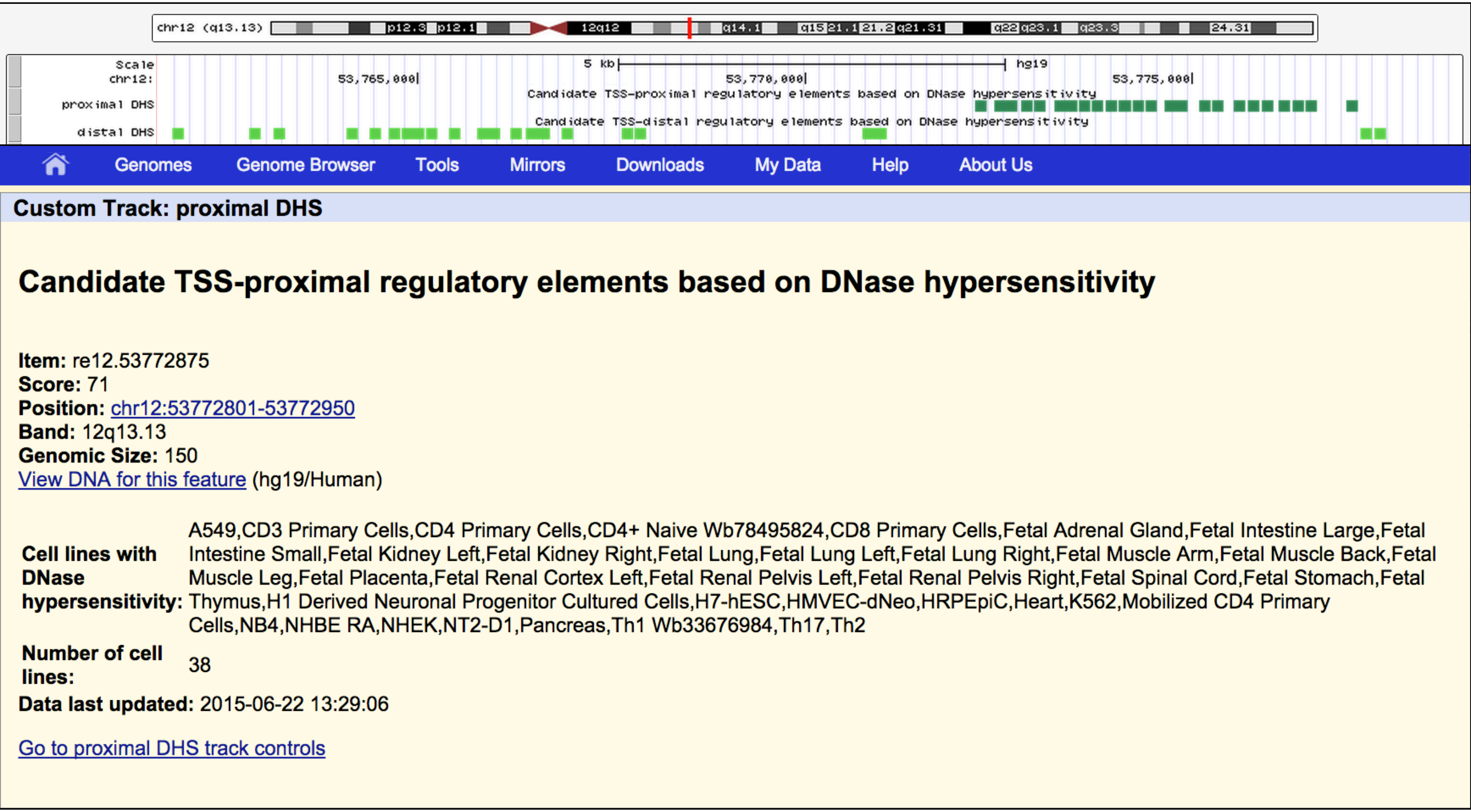
- Gene expression over ~60 cell types with genes annotated by GENCODE 19 [Query tool at Penn State | Visualize data | Download data | Download methods]
- Transcription start site (TSS) lists [View README]
  - GENCODE v19 TSS [Download]
  - GENCODE v19 TSS stratified by strict Fantom5 CAGE clusters [Download]
  - GENCODE v19 TSS stratified by robust Fantom5 CAGE clusters [Download]
  - GENCODE v19 TSS stratified by permissive Fantom5 CAGE clusters [Download]

https://www.encodeproject.org/files/
ENCFF076KTT/@@download/
ENCFF076KTT.bigBed

UCSC Genome Browser


chr12 (q13.13)
p12.3 p12.1 12q12 q14.1 q15 21.1 21.2 q21.31 q22 q23.1 q23.3 24.31
Scale
chr12:
5 kb
hg19
53,765,000
53,770,000
53,775,000
Candidate TSS-proximal regulatory elements based on DNase hypersensitivity
proximal DHS
Candidate TSS-distal regulatory elements based on DNase hypersensitivity
distal DHS
Candidate promoters based on proximal H3K4me3
proximal H3K4me3
Candidate promoters based on proximal H3K9ac
proximal H3K9ac
Candidate promoters based on proximal H3K27ac
proximal H3K27ac
Candidate promoters based on proximal H3K4me1
proximal H3K4me1
Candidate enhancers based on distal H3K4me3
distal H3K4me3
Candidate enhancers based on distal H3K9ac
distal H3K9ac
Candidate enhancers based on distal H3K27ac
distal H3K27ac
Candidate enhancers based on distal H3K4me1
distal H3K4me1
Transcription factor binding based on ChIP-seq (TSS-distal)
multiple
multiple
multiple
GTF2B
multiple
GTF2B
multiple
multiple
multiple
multiple
multiple
CEBPB,E2F1,POLR2A
CTCF,E2F1,GTF2B,POLR3A,TCF7L2
STAT3
STAT3
Transcription factor binding based on ChIP-seq (TSS-proximal)
POLR2A
multiple
multiple
multiple
POLR2A
multiple
multiple
CTCF
BCL3,CHD1,JUND,POLR2A
multiple
multiple
CTCF
multiple
multiple
multiple
multiple
multiple
multiple
multiple
multiple
multiple
CTCF,EBF1,ESR1,SIN3A
IRF1,POLR2A,SP1,TBP
POLR2A,TBP,TFAP2A
RefSeq Genes
SP1
SP1
SP1
Basic Gene Annotation Set from GENCODE Version 19
SP1
SP1
Simple Nucleotide Polymorphisms (dbSNP 142) Found in >= 1% of Samples
Common SNPs(142)

UCSC Genome Browser



Custom Track: proximal DHS

Candidate TSS-proximal regulatory elements based on DNase hypersensitivity

Item: re12.53772875
Score: 71
Position: chr12:53772801-53772950
Band: 12q13.13
Genomic Size: 150
View DNA for this feature (hg19/Human)

Cell lines with DNase hypersensitivity:	A549,CD3 Primary Cells,CD4 Primary Cells,CD4+ Naive Wb78495824,CD8 Primary Cells,Fetal Adrenal Gland,Fetal Intestine Large,Fetal Intestine Small,Fetal Kidney Left,Fetal Kidney Right,Fetal Lung,Fetal Lung Left,Fetal Lung Right,Fetal Muscle Arm,Fetal Muscle Back,Fetal Muscle Leg,Fetal Placenta,Fetal Renal Cortex Left,Fetal Renal Pelvis Left,Fetal Renal Pelvis Right,Fetal Spinal Cord,Fetal Stomach,Fetal Thymus,H1 Derived Neuronal Progenitor Cultured Cells,H7-hESC,HMVEC-dNeo,HRPEpiC,Heart,K562,Mobilized CD4 Primary Cells,NB4,NHBE RA,NHEK,NT2-D1,Pancreas,Th1 Wb33676984,Th17,Th2
Number of cell lines:	38

Data last updated: 2015-06-22 13:29:06

Go to proximal DHS track controls

UCSC Genome Browser

Other useful tracks:

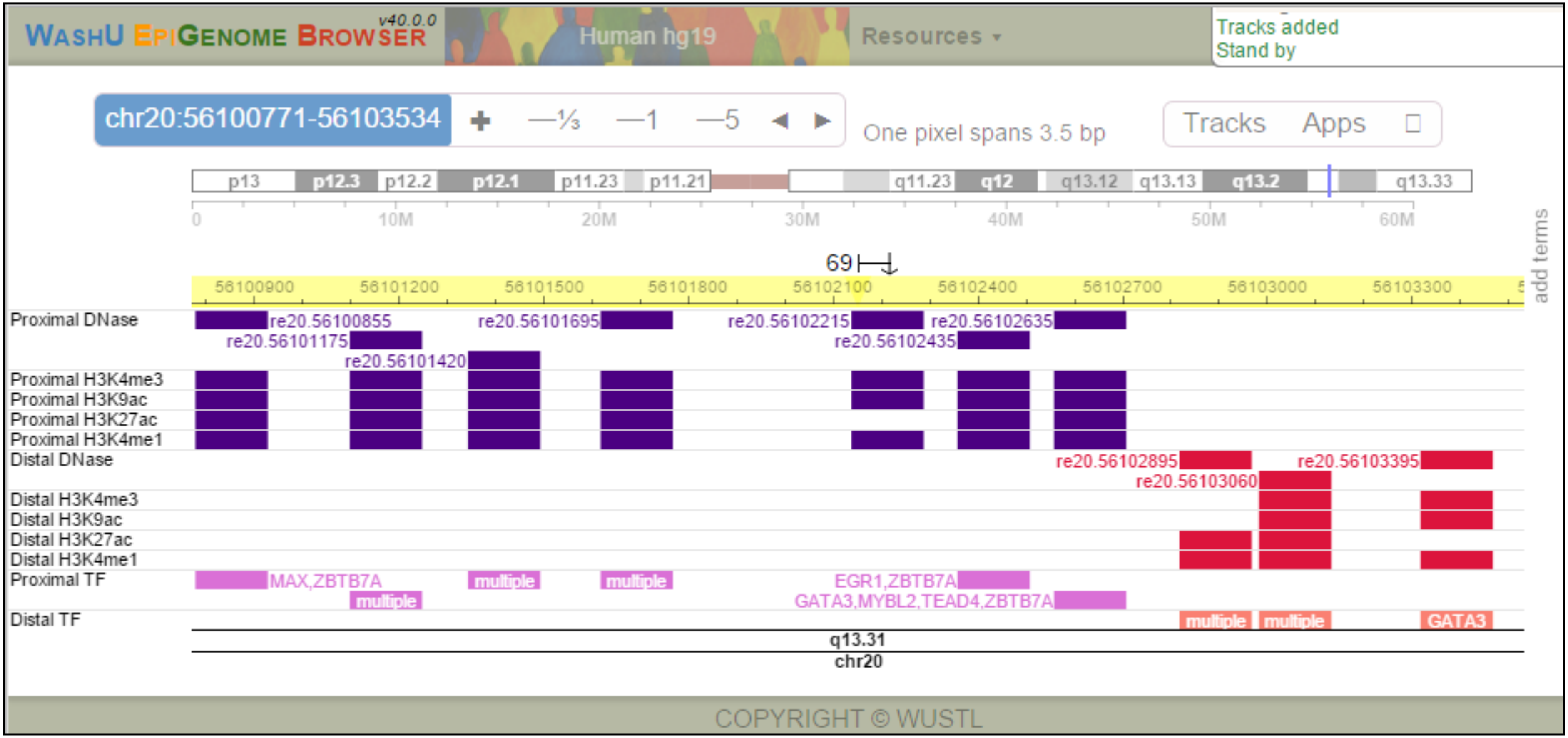
- UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)

- GENCODE Gene Annotation Tracks

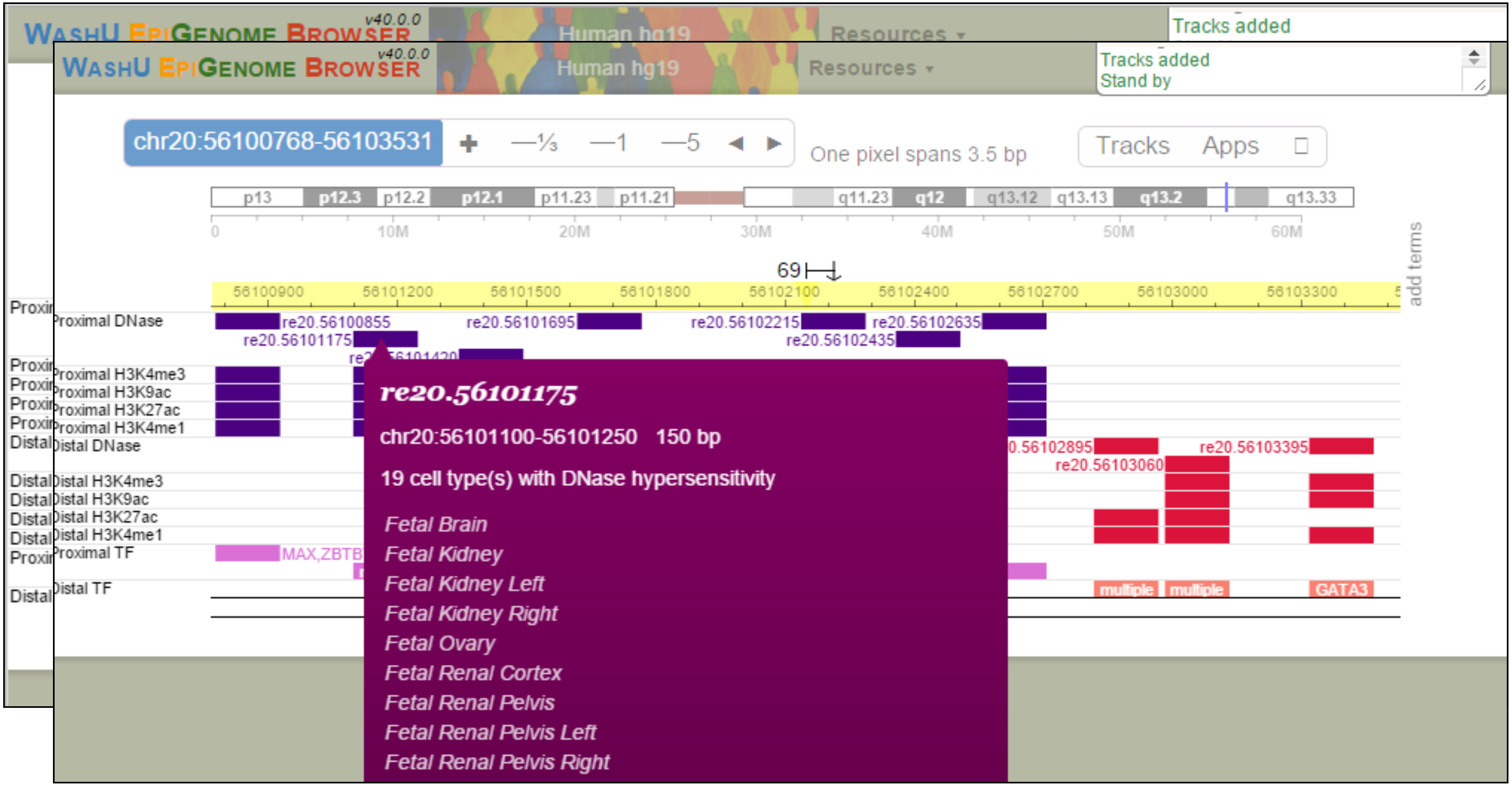
- Integrated Regulation from ENCODE Tracks

- Genome Segmentations from ENCODE (ChromHMM, Segway)

WashU Epigenome Browser



WashU Epigenome Browser



Tracks added
Stand by

re20.56101175

chr20:56101100-56101250 150 bp

19 cell type(s) with DNase hypersensitivity

Fetal Brain
Fetal Kidney
Fetal Kidney Left
Fetal Kidney Right
Fetal Ovary
Fetal Renal Cortex
Fetal Renal Pelvis
Fetal Renal Pelvis Left
Fetal Renal Pelvis Right

Genome Browser Links

- UCSC Custom tracks
  http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg19&hgt.customText=http://zlab-trackhub.umassmed.edu/encyclopedia/ucsc_trackhub.txt

- UCSC Track Hub
  http://zlab-trackhub.umassmed.edu/encyclopedia/v1/hub.txt

  http://zlab-trackhub.umassmed.edu/encyclopedia/v1/genome.txt

  http://zlab-trackhub.umassmed.edu/encyclopedia/v1/hg19/trackDb.txt

- WashU Hammock tracks*
  http://epigenomegateway.wustl.edu/browser/?genome=hg19&datahub=http://zlab-trackhub.umassmed.edu/encyclopedia/washu_trackhub.txt

*http://wiki.wubrowse.org/Hammock

Future Directions

- Open-source codebase

- Generate mouse annotations

- Add more data!
  - Refine use of TF data
  RNA-seq
  3D contacts (ChIA-PET and Hi-C)
  ChromHMM and Segway
  Target gene prediction

Variant Annotation Using RegulomeDB and HaploReg

Motivation

- The majority of variants reported by GWAS are in noncoding regions of the genome

- The variant reported by the GWAS (lead/tagged variant) may not be causal but is in high linkage disequilibrium with the casual variant

- Using data from ENCODE, we can annotate noncoding regions of the genome and predict the function of disease associated noncoding variants

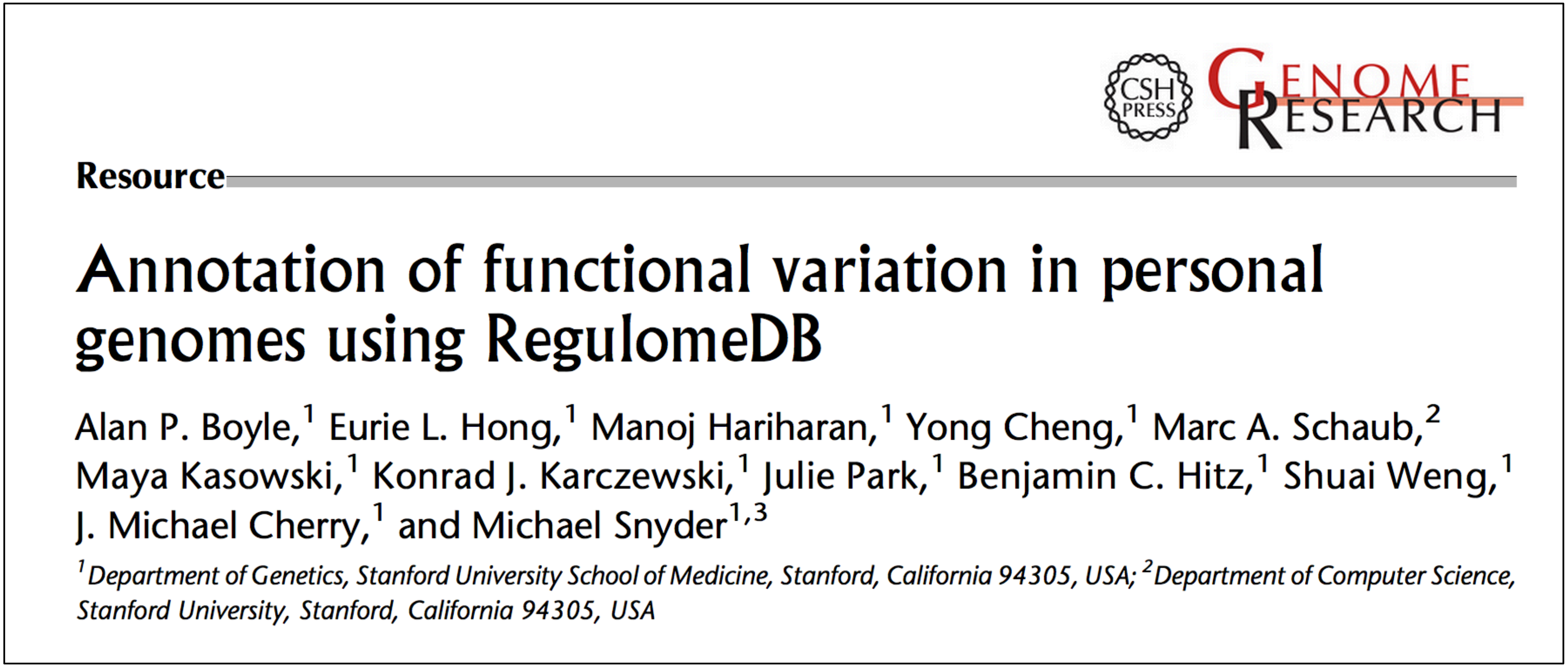
Variant Annotation Tools

RegulomeDB

http://www.regulomedb.org/

HaploReg v4

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



CSH PRESS GENOME RESEARCH

Resource

Annotation of functional variation in personal genomes using RegulomeDB

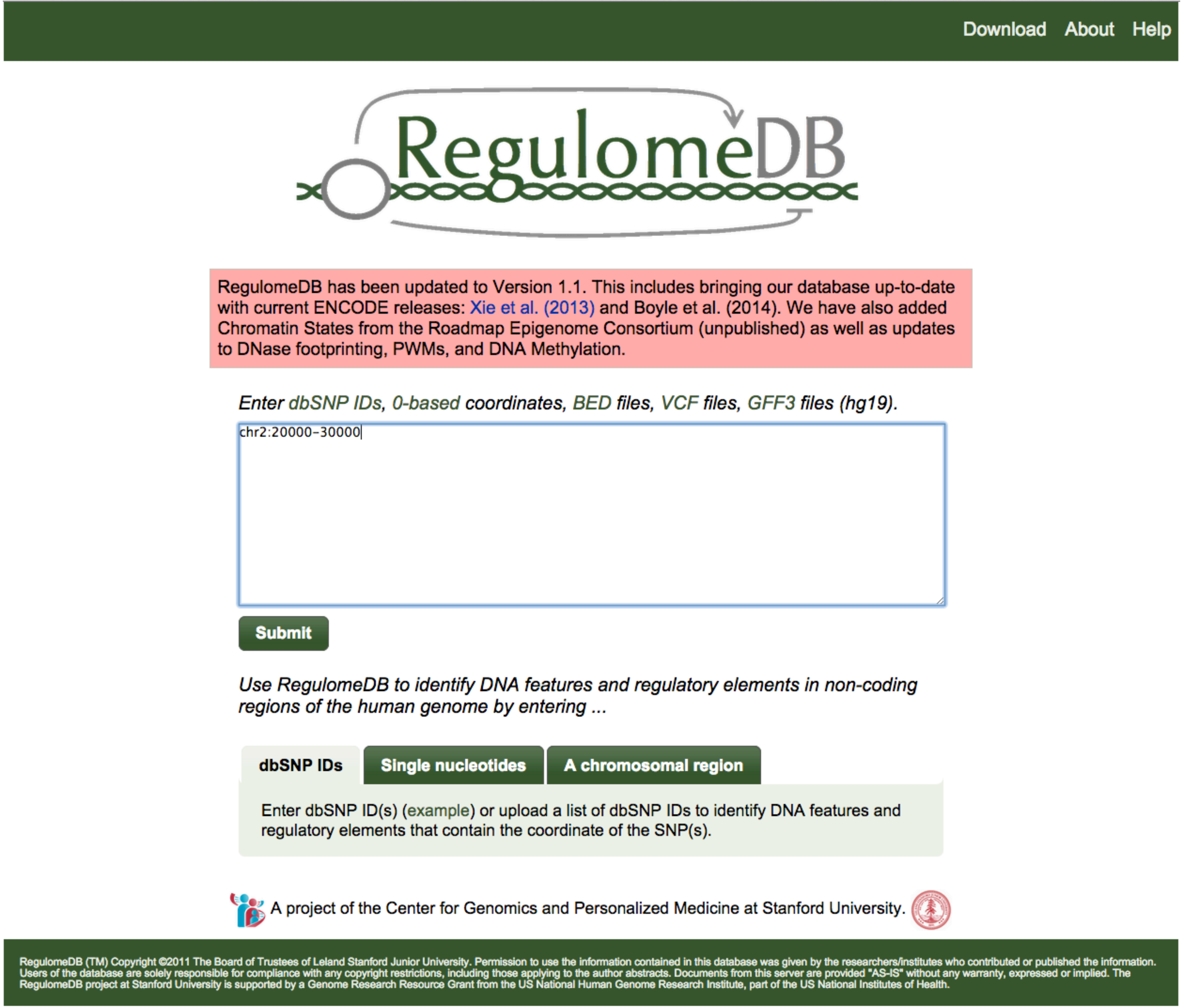
Alan P. Boyle,[1] Eurie L. Hong,[1] Manoj Hariharan,[1] Yong Cheng,[1] Marc A. Schaub,[2] Maya Kasowski,[1] Konrad J. Karczewski,[1] Julie Park,[1] Benjamin C. Hitz,[1] Shuai Weng,[1] J. Michael Cherry,[1] and Michael Snyder[1,3]

[1] Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA; [2] Department of Computer Science, Stanford University, Stanford, California 94305, USA

Table 2. RegulomeDB variant classification scheme

Category scheme	
Category	Description
	Likely to affect binding and linked to expression of a gene target
1a	eQTL + TF binding + matched TF motif + matched DNase footprint + DNase peak
1b	eQTL + TF binding + any motif + DNase footprint + DNase peak
1c	eQTL + TF binding + matched TF motif + DNase peak
1d	eQTL + TF binding + any motif + DNase peak
1e	eQTL + TF binding + matched TF motif
1f	eQTL + TF binding/DNase peak
	Likely to affect binding
2a	TF binding + matched TF motif + matched DNase footprint + DNase peak
2b	TF binding + any motif + DNase footprint + DNase peak
2c	TF binding + matched TF motif + DNase peak
	Less likely to affect binding
3a	TF binding + any motif + DNase peak
3b	TF binding + matched TF motif
	Minimal binding evidence
4	TF binding + DNase peak
5	TF binding or DNase peak
6	Motif hit

Lower scores indicate increasing evidence for a variant to be located in a functional region. Category 1 variants have equivalents in other categories with the additional requirement of eQTL information.



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.

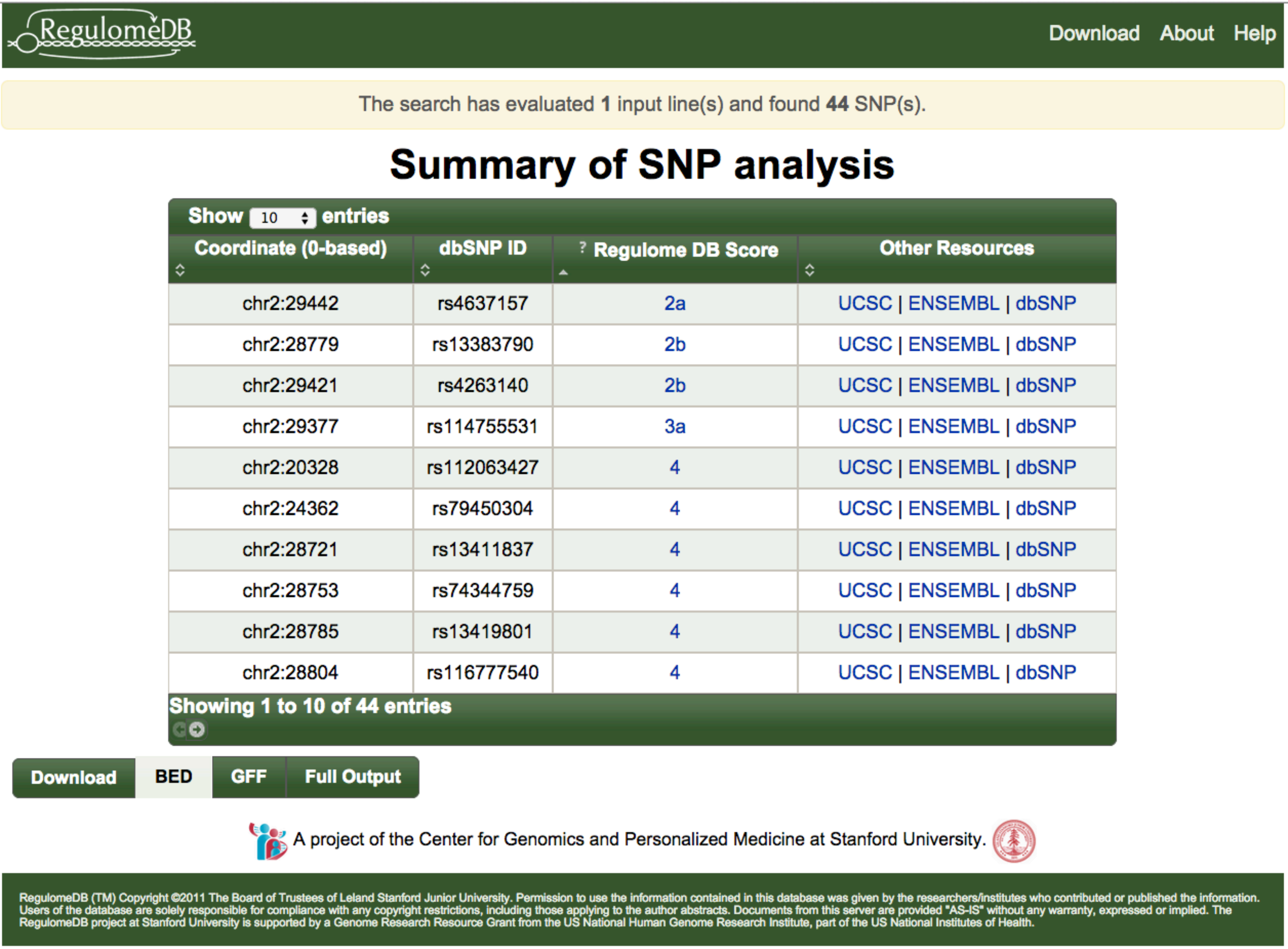
Enter dbSNP IDs, 0-based coordinates, BED files, VCF files, GFF3 files (hg19).

```
chr2:20000-30000
```

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

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.



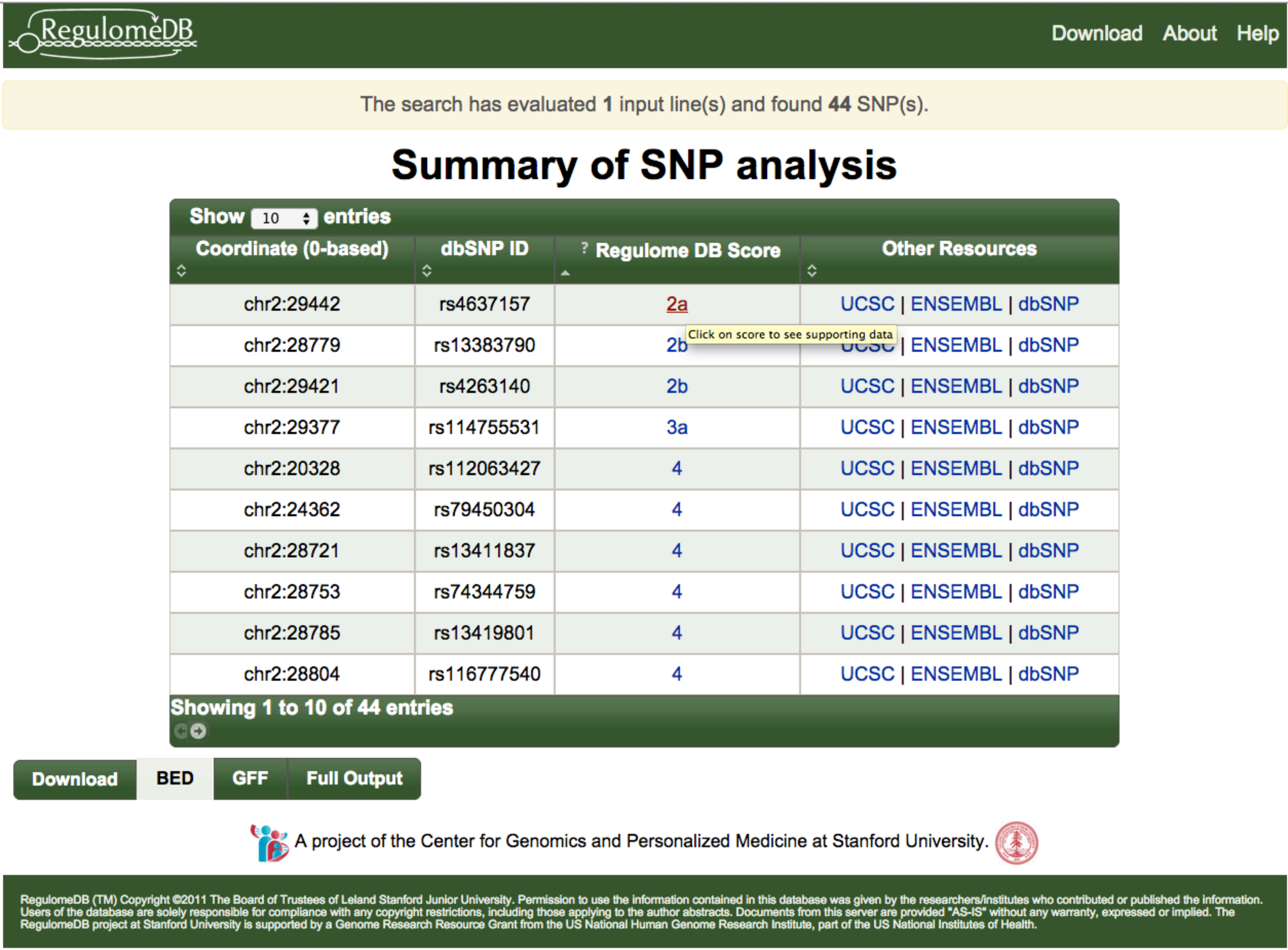
The search has evaluated **1** input line(s) and found **44** SNP(s).

Summary of SNP analysis

Coordinate (0-based)	dbSNP ID	? Regulome DB Score	Other Resources
chr2:29442	rs4637157	2a	UCSC | ENSEMBL | dbSNP
chr2:28779	rs13383790	2b	UCSC | ENSEMBL | dbSNP
chr2:29421	rs4263140	2b	UCSC | ENSEMBL | dbSNP
chr2:29377	rs114755531	3a	UCSC | ENSEMBL | dbSNP
chr2:20328	rs112063427	4	UCSC | ENSEMBL | dbSNP
chr2:24362	rs79450304	4	UCSC | ENSEMBL | dbSNP
chr2:28721	rs13411837	4	UCSC | ENSEMBL | dbSNP
chr2:28753	rs74344759	4	UCSC | ENSEMBL | dbSNP
chr2:28785	rs13419801	4	UCSC | ENSEMBL | dbSNP
chr2:28804	rs116777540	4	UCSC | ENSEMBL | dbSNP

Showing 1 to 10 of 44 entries

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



The search has evaluated **1** input line(s) and found **44** SNP(s).

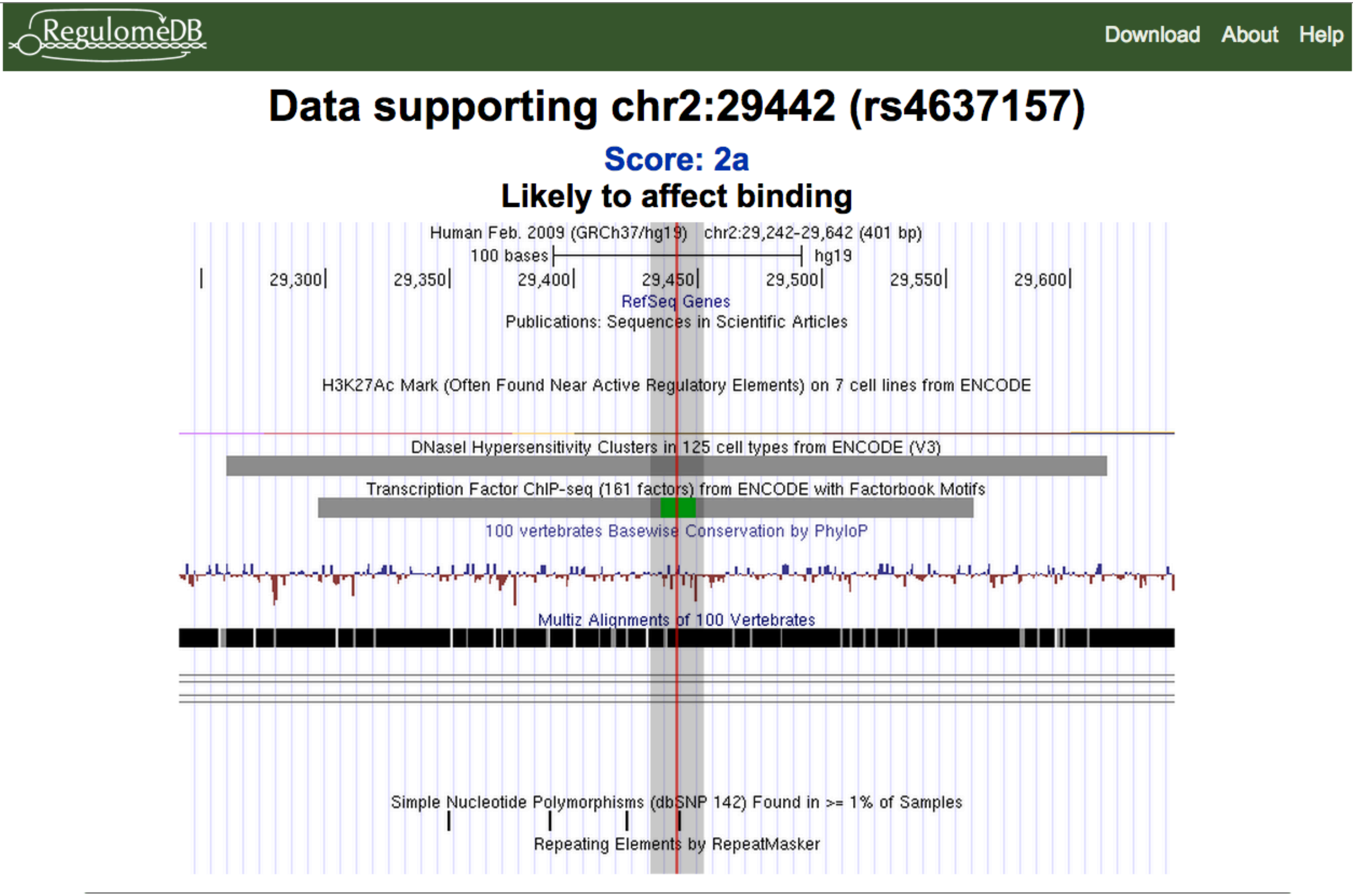
Summary of SNP analysis

Coordinate (0-based)	dbSNP ID	? Regulome DB Score	Other Resources
chr2:29442	rs4637157	2a	UCSC | ENSEMBL | dbSNP
chr2:28779	rs13383790	2b	UCSC | ENSEMBL | dbSNP
chr2:29421	rs4263140	2b	UCSC | ENSEMBL | dbSNP
chr2:29377	rs114755531	3a	UCSC | ENSEMBL | dbSNP
chr2:20328	rs112063427	4	UCSC | ENSEMBL | dbSNP
chr2:24362	rs79450304	4	UCSC | ENSEMBL | dbSNP
chr2:28721	rs13411837	4	UCSC | ENSEMBL | dbSNP
chr2:28753	rs74344759	4	UCSC | ENSEMBL | dbSNP
chr2:28785	rs13419801	4	UCSC | ENSEMBL | dbSNP
chr2:28804	rs116777540	4	UCSC | ENSEMBL | dbSNP

Click on score to see supporting data

Showing 1 to 10 of 44 entries

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

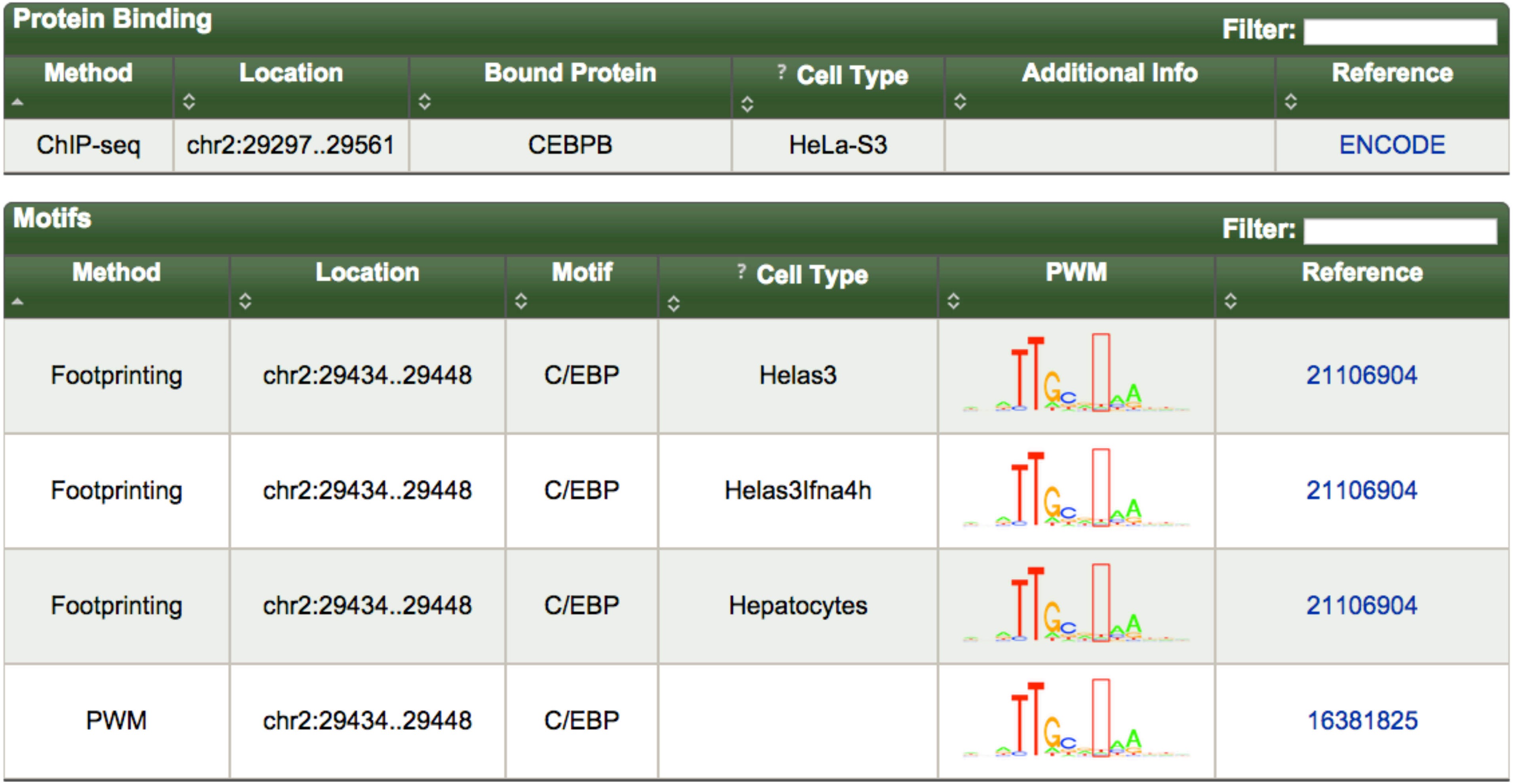

RegulomeDB
Download
About
Help

Data supporting chr2:29442 (rs4637157)

Score: 2a

Likely to affect binding

Human Feb. 2009 (GRCh37/hg19) chr2:29,242-29,642 (401 bp)
100 bases
hg19
29,300
29,350
29,400
29,450
29,500
29,550
29,600
RefSeq Genes
Publications: Sequences in Scientific Articles
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
100 vertebrates Basewise Conservation by PhyloP
Multiz Alignments of 100 Vertebrates
Simple Nucleotide Polymorphisms (dbSNP 142) Found in >= 1% of Samples
Repeating Elements by RepeatMasker



Protein Binding

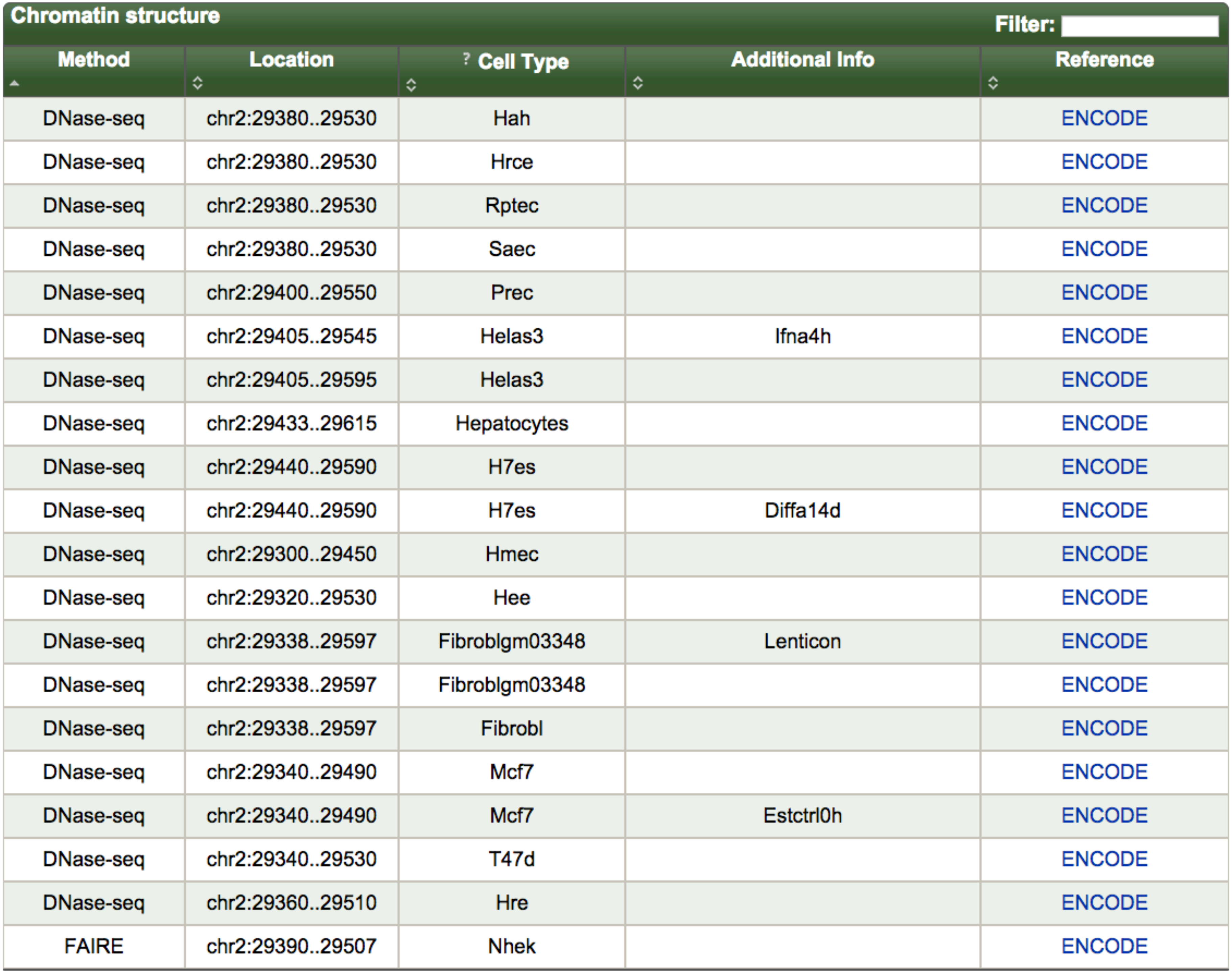
Filter:

Method	Location	Bound Protein	? Cell Type	Additional Info	Reference
ChIP-seq	chr2:29297..29561	CEBPB	HeLa-S3		ENCODE

Motifs

Filter:

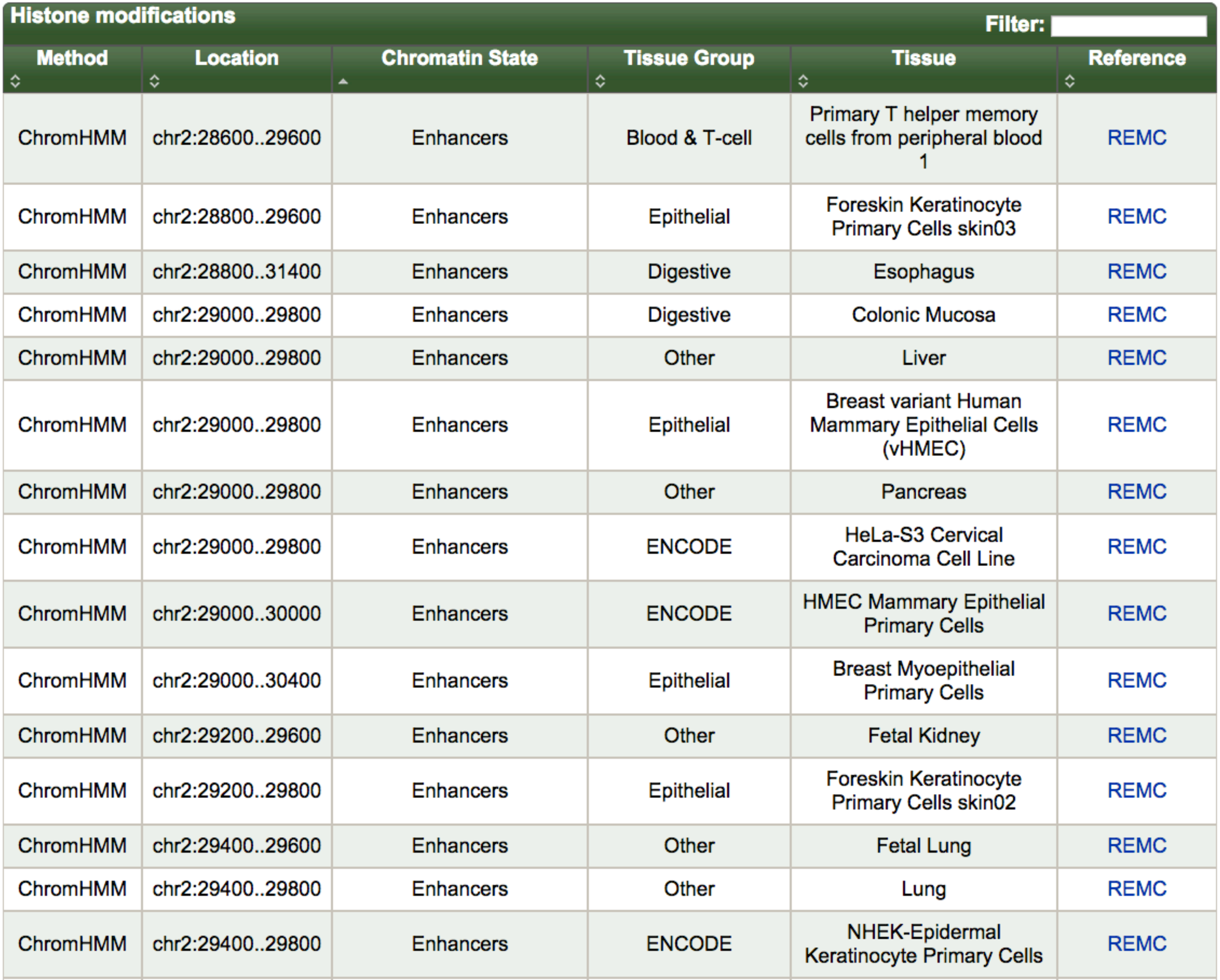
Method	Location	Motif	? Cell Type	PWM	Reference
Footprinting	chr2:29434..29448	C/EBP	Helas3		21106904
Footprinting	chr2:29434..29448	C/EBP	Helas3Ifna4h		21106904
Footprinting	chr2:29434..29448	C/EBP	Hepatocytes		21106904
PWM	chr2:29434..29448	C/EBP			16381825



Chromatin structure

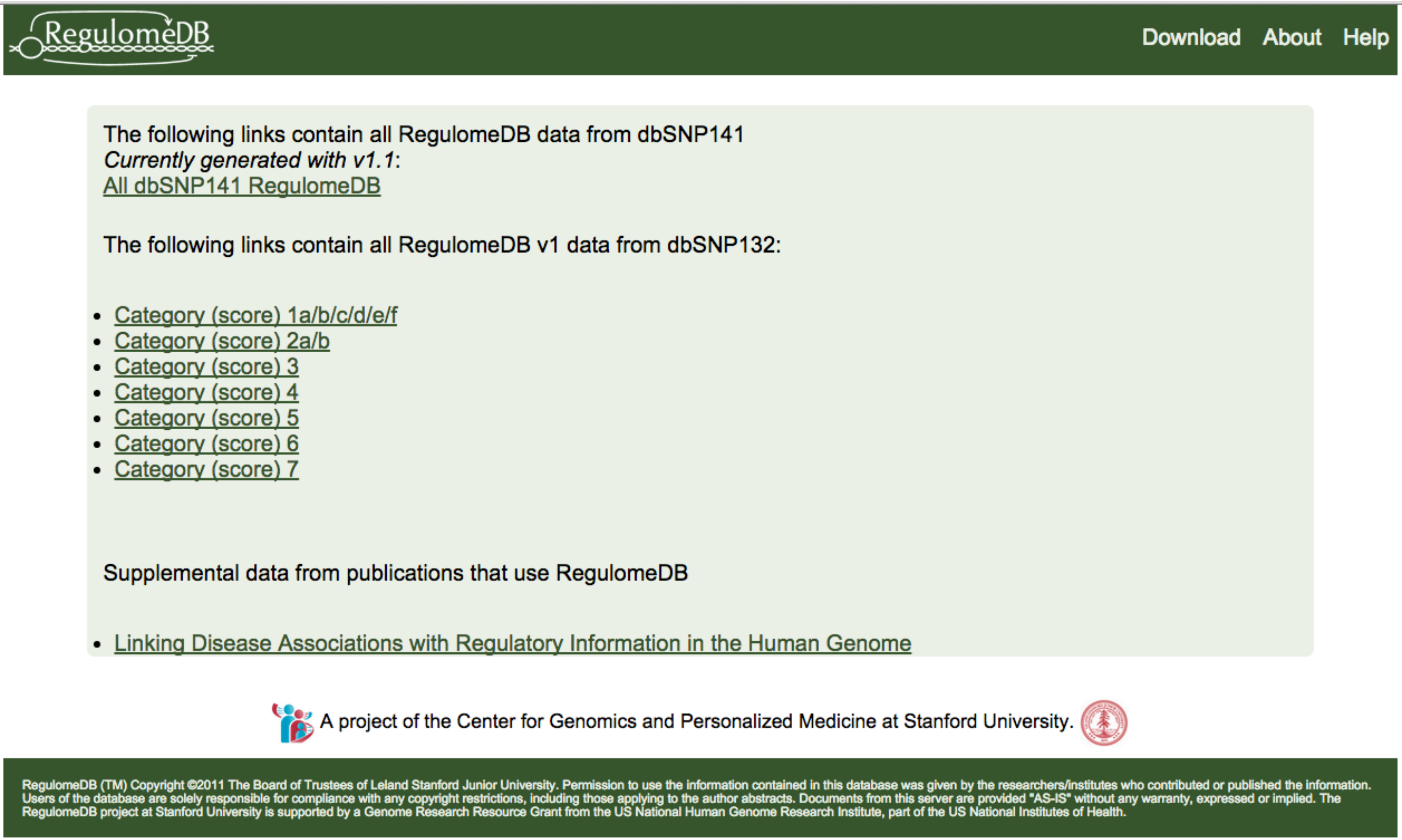
Filter:

Method	Location	? Cell Type	Additional Info	Reference
DNase-seq	chr2:29380..29530	Hah		ENCODE
DNase-seq	chr2:29380..29530	Hrce		ENCODE
DNase-seq	chr2:29380..29530	Rptec		ENCODE
DNase-seq	chr2:29380..29530	Saec		ENCODE
DNase-seq	chr2:29400..29550	Prec		ENCODE
DNase-seq	chr2:29405..29545	Helas3	Ifna4h	ENCODE
DNase-seq	chr2:29405..29595	Helas3		ENCODE
DNase-seq	chr2:29433..29615	Hepatocytes		ENCODE
DNase-seq	chr2:29440..29590	H7es		ENCODE
DNase-seq	chr2:29440..29590	H7es	Diffa14d	ENCODE
DNase-seq	chr2:29300..29450	Hmec		ENCODE
DNase-seq	chr2:29320..29530	Hee		ENCODE
DNase-seq	chr2:29338..29597	Fibroblgm03348	Lenticon	ENCODE
DNase-seq	chr2:29338..29597	Fibroblgm03348		ENCODE
DNase-seq	chr2:29338..29597	Fibrobl		ENCODE
DNase-seq	chr2:29340..29490	Mcf7		ENCODE
DNase-seq	chr2:29340..29490	Mcf7	Estctrl0h	ENCODE
DNase-seq	chr2:29340..29530	T47d		ENCODE
DNase-seq	chr2:29360..29510	Hre		ENCODE
FAIRE	chr2:29390..29507	Nhek		ENCODE



Histone modifications

Method	Location	Chromatin State	Tissue Group	Tissue	Reference
ChromHMM	chr2:28600..29600	Enhancers	Blood & T-cell	Primary T helper memory cells from peripheral blood 1	REMC
ChromHMM	chr2:28800..29600	Enhancers	Epithelial	Foreskin Keratinocyte Primary Cells skin03	REMC
ChromHMM	chr2:28800..31400	Enhancers	Digestive	Esophagus	REMC
ChromHMM	chr2:29000..29800	Enhancers	Digestive	Colonic Mucosa	REMC
ChromHMM	chr2:29000..29800	Enhancers	Other	Liver	REMC
ChromHMM	chr2:29000..29800	Enhancers	Epithelial	Breast variant Human Mammary Epithelial Cells (vHMEC)	REMC
ChromHMM	chr2:29000..29800	Enhancers	Other	Pancreas	REMC
ChromHMM	chr2:29000..29800	Enhancers	ENCODE	HeLa-S3 Cervical Carcinoma Cell Line	REMC
ChromHMM	chr2:29000..30000	Enhancers	ENCODE	HMEC Mammary Epithelial Primary Cells	REMC
ChromHMM	chr2:29000..30400	Enhancers	Epithelial	Breast Myoepithelial Primary Cells	REMC
ChromHMM	chr2:29200..29600	Enhancers	Other	Fetal Kidney	REMC
ChromHMM	chr2:29200..29800	Enhancers	Epithelial	Foreskin Keratinocyte Primary Cells skin02	REMC
ChromHMM	chr2:29400..29600	Enhancers	Other	Fetal Lung	REMC
ChromHMM	chr2:29400..29800	Enhancers	Other	Lung	REMC
ChromHMM	chr2:29400..29800	Enhancers	ENCODE	NHEK-Epidermal Keratinocyte Primary Cells	REMC



The following links contain all RegulomeDB data from dbSNP141
Currently generated with v1.1:
All dbSNP141 RegulomeDB

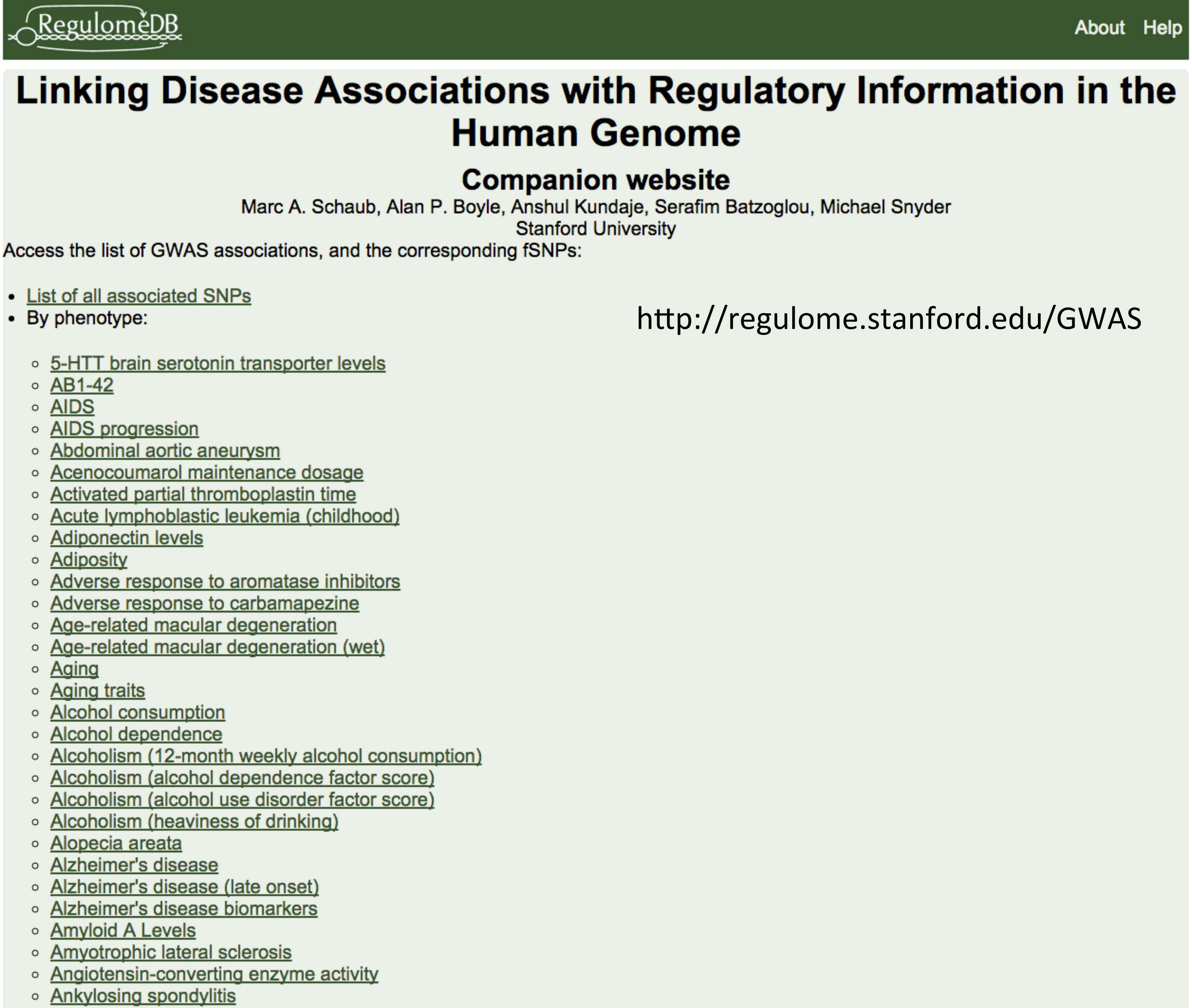
The following links contain all RegulomeDB v1 data from dbSNP132:

- Category (score) 1a/b/c/d/e/f
- Category (score) 2a/b
- Category (score) 3
- Category (score) 4
- Category (score) 5
- Category (score) 6
- Category (score) 7

Supplemental data from publications that use RegulomeDB

- Linking Disease Associations with Regulatory Information in the Human Genome

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



Linking Disease Associations with Regulatory Information in the Human Genome

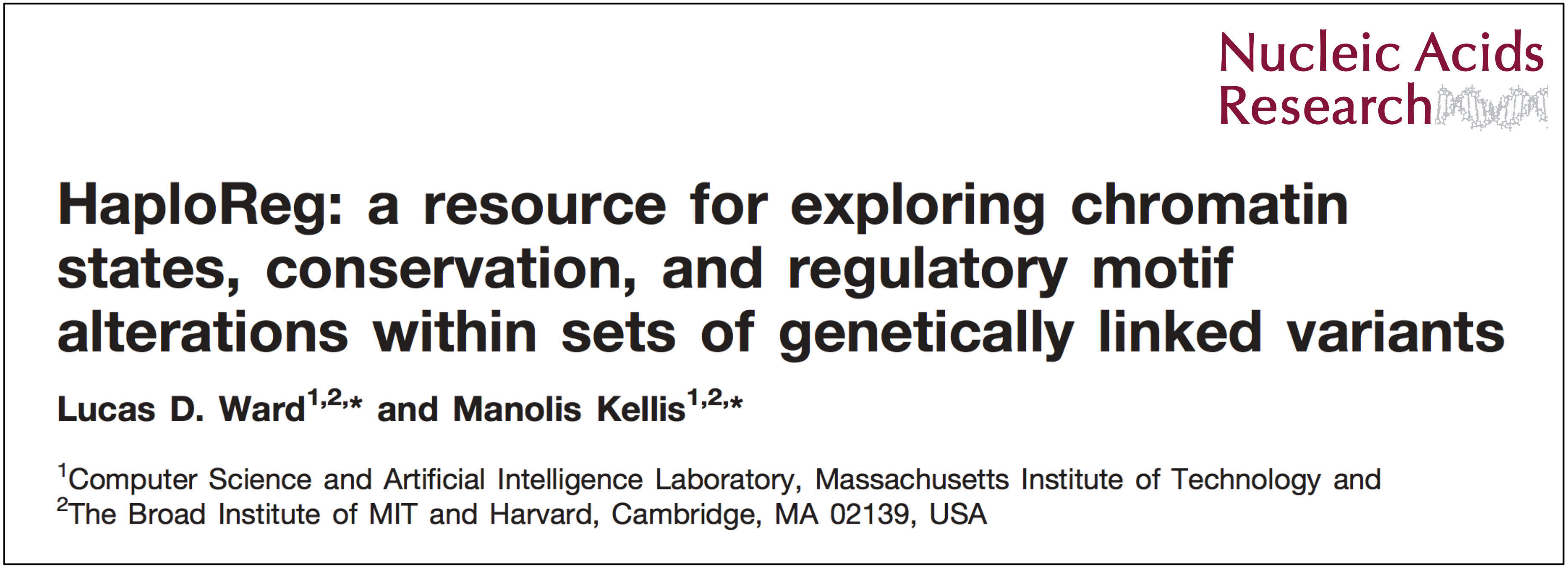
Companion website

Marc A. Schaub, Alan P. Boyle, Anshul Kundaje, Serafim Batzoglou, Michael Snyder
Stanford University

Access the list of GWAS associations, and the corresponding fSNPs:

- List of all associated SNPs
- By phenotype:
  - 5-HTT brain serotonin transporter levels
  - AB1-42
  - AIDS
  - AIDS progression
  - Abdominal aortic aneurysm
  - Acenocoumarol maintenance dosage
  - Activated partial thromboplastin time
  - Acute lymphoblastic leukemia (childhood)
  - Adiponectin levels
  - Adiposity
  - Adverse response to aromatase inhibitors
  - Adverse response to carbamapezine
  - Age-related macular degeneration
  - Age-related macular degeneration (wet)
  - Aging
  - Aging traits
  - Alcohol consumption
  - Alcohol dependence
  - Alcoholism (12-month weekly alcohol consumption)
  - Alcoholism (alcohol dependence factor score)
  - Alcoholism (alcohol use disorder factor score)
  - Alcoholism (heaviness of drinking)
  - Alopecia areata
  - Alzheimer's disease
  - Alzheimer's disease (late onset)
  - Alzheimer's disease biomarkers
  - Amyloid A Levels
  - Amyotrophic lateral sclerosis
  - Angiotensin-converting enzyme activity
  - Ankylosing spondylitis

http://regulome.stanford.edu/GWAS

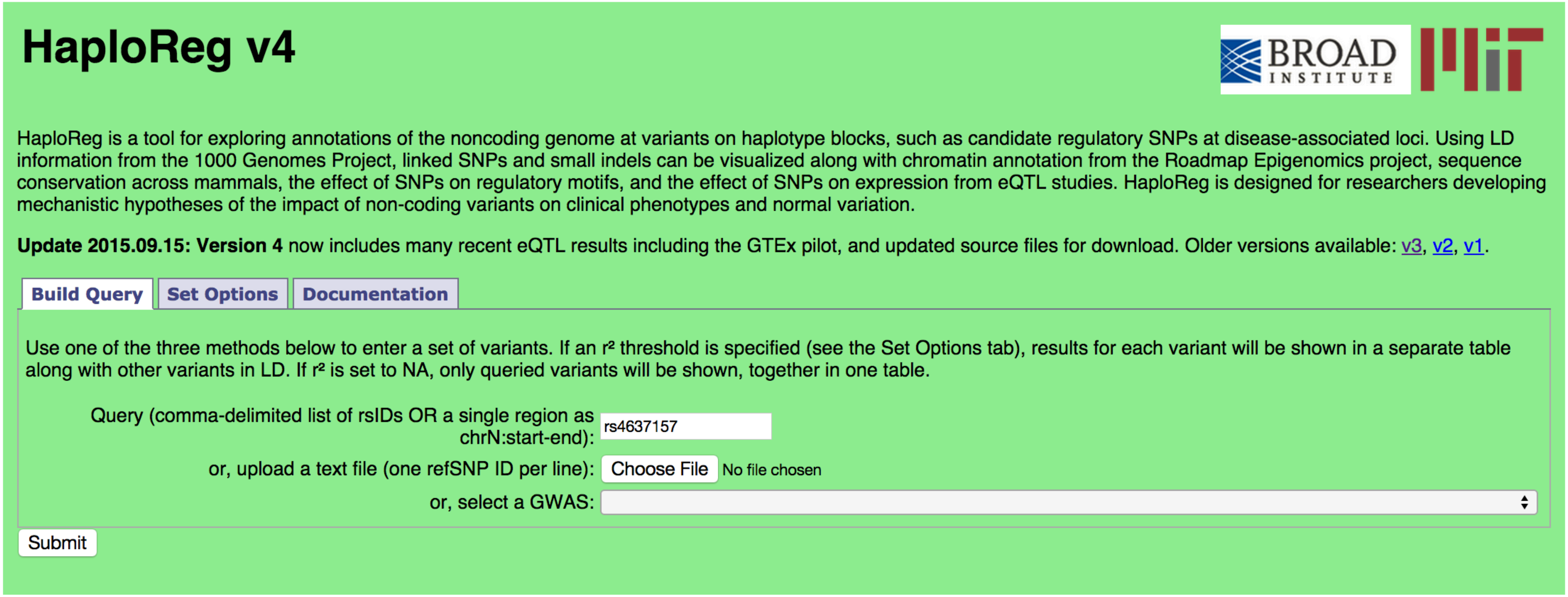


Nucleic Acids Research

HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants

Lucas D. Ward[1,2,*] **and Manolis Kellis**[1,2,*]

[1]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology and
[2]The Broad Institute of MIT and Harvard, Cambridge, MA 02139, USA



HaploReg v4

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 annotation from the Roadmap Epigenomics project, 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.

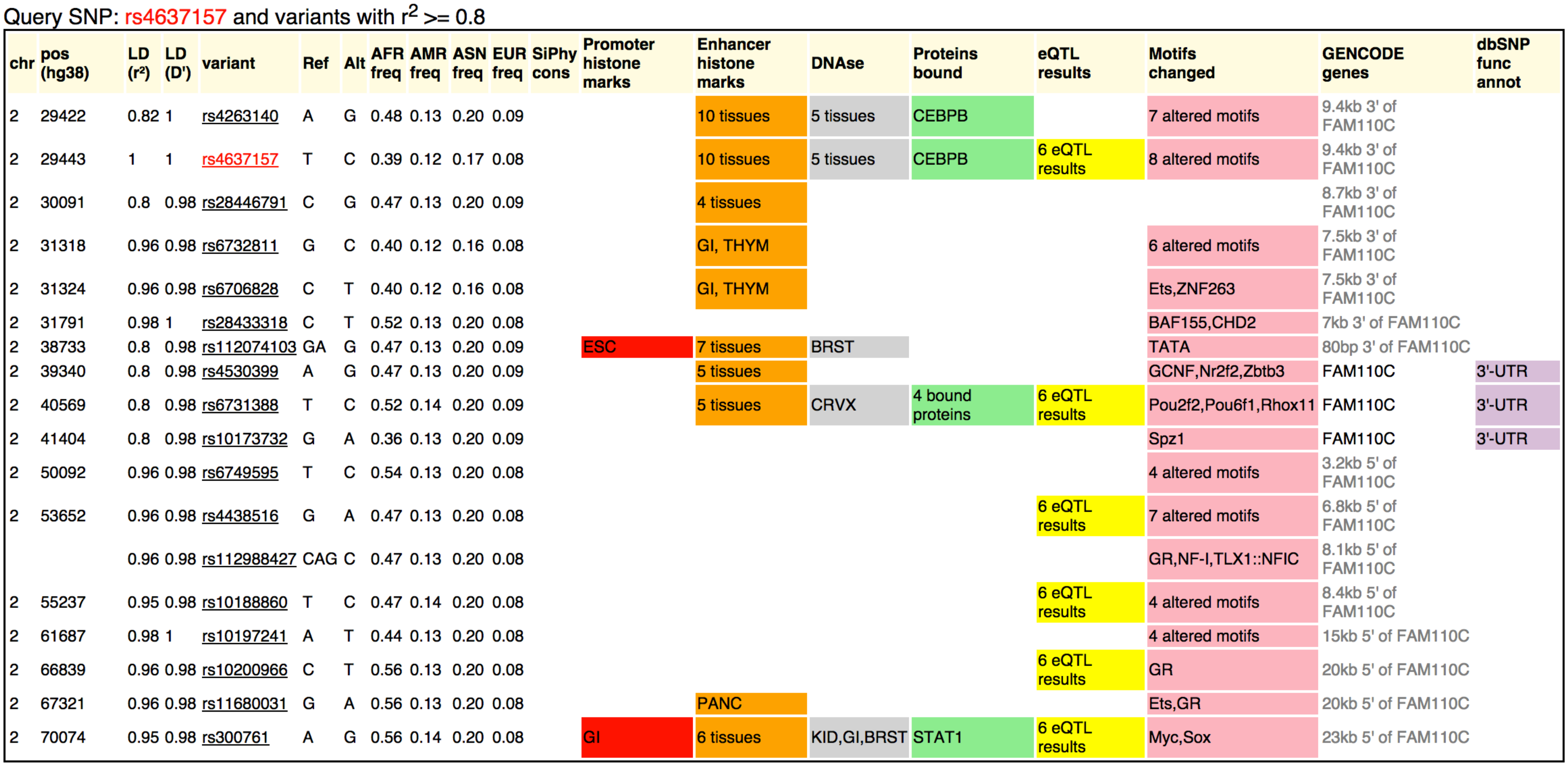
Update 2015.09.15: Version 4 now includes many recent eQTL results including the GTEx pilot, and updated source files for download. Older versions available: v3, v2, v1.

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): rs4637157

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

or, select a GWAS:



Query SNP: rs4637157 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	eQTL results	Motifs changed	GENCODE genes	dbSNP func annot
2	29422	0.82	1	rs4263140	A	G	0.48	0.13	0.20	0.09			10 tissues	5 tissues	CEBPB		7 altered motifs	9.4kb 3' of FAM110C	
2	29443	1	1	rs4637157	T	C	0.39	0.12	0.17	0.08			10 tissues	5 tissues	CEBPB	6 eQTL results	8 altered motifs	9.4kb 3' of FAM110C	
2	30091	0.8	0.98	rs28446791	C	G	0.47	0.13	0.20	0.09			4 tissues					8.7kb 3' of FAM110C	
2	31318	0.96	0.98	rs6732811	G	C	0.40	0.12	0.16	0.08			GI, THYM				6 altered motifs	7.5kb 3' of FAM110C	
2	31324	0.96	0.98	rs6706828	C	T	0.40	0.12	0.16	0.08			GI, THYM				Ets,ZNF263	7.5kb 3' of FAM110C	
2	31791	0.98	1	rs28433318	C	T	0.52	0.13	0.20	0.08							BAF155,CHD2	7kb 3' of FAM110C	
2	38733	0.8	0.98	rs112074103	GA	G	0.47	0.13	0.20	0.09		ESC	7 tissues	BRST			TATA	80bp 3' of FAM110C	
2	39340	0.8	0.98	rs4530399	A	G	0.47	0.13	0.20	0.09			5 tissues				GCNF,Nr2f2,Zbtb3	FAM110C	3'-UTR
2	40569	0.8	0.98	rs6731388	T	C	0.52	0.14	0.20	0.09			5 tissues	CRVX	4 bound proteins	6 eQTL results	Pou2f2,Pou6f1,Rhox11	FAM110C	3'-UTR
2	41404	0.8	0.98	rs10173732	G	A	0.36	0.13	0.20	0.09							Spz1	FAM110C	3'-UTR
2	50092	0.96	0.98	rs6749595	T	C	0.54	0.13	0.20	0.08							4 altered motifs	3.2kb 5' of FAM110C	
2	53652	0.96	0.98	rs4438516	G	A	0.47	0.13	0.20	0.08						6 eQTL results	7 altered motifs	6.8kb 5' of FAM110C	
		0.96	0.98	rs112988427	CAG	C	0.47	0.13	0.20	0.08							GR,NF-I,TLX1::NFIC	8.1kb 5' of FAM110C	
2	55237	0.95	0.98	rs10188860	T	C	0.47	0.14	0.20	0.08						6 eQTL results	4 altered motifs	8.4kb 5' of FAM110C	
2	61687	0.98	1	rs10197241	A	T	0.44	0.13	0.20	0.08							4 altered motifs	15kb 5' of FAM110C	
2	66839	0.96	0.98	rs10200966	C	T	0.56	0.13	0.20	0.08						6 eQTL results	GR	20kb 5' of FAM110C	
2	67321	0.96	0.98	rs11680031	G	A	0.56	0.13	0.20	0.08			PANC				Ets,GR	20kb 5' of FAM110C	
2	70074	0.95	0.98	rs300761	A	G	0.56	0.14	0.20	0.08		GI	6 tissues	KID,GI,BRST	STAT1	6 eQTL results	Myc,Sox	23kb 5' of FAM110C	

Detail view for rs4637157

[Link to dbSNP entry]

[Link to Ensembl Variation entry]

Sequence facts

chr	pos (hg19)	chr	pos (hg38)	Reference	Alternate	1000 Genomes Phase 1 Frequencies				Sequence constraint		dbSNP functional annotation
						AFR	AMR	ASN	EUR	by GERP	by SiPhy	
chr2	29443	chr2	29443	T	C	0.39	0.12	0.17	0.08	No	No	none

Closest annotated gene					
Source	Distance	Direction	ID/Link	Common name	Description
GENCODE	3'	9370	ENSG00000184731.5	FAM110C	family with sequence similarity 110, member C [Source:HGNC Symbol;Acc:33340]
RefSeq	3'	9369	NM_001077710	FAM110C	family with sequence similarity 110, member C [Source:HGNC Symbol;Acc:33340]

Regulatory chromatin states from DNAse and histone ChIP-Seq (Roadmap Epigenomics Consortium, 2015)

(Black = missing data)

Epigenome ID (EID)	Group	Mnemonic	Description	Chromatin states (Core 15-state model)	Chromatin states (25-state model using 12 imputed marks)	H3K4me1	H3K4me3	H3K27ac	H3K9ac	DNase
E017	IMR90	LNG.IMR90	IMR90 fetal lung fibroblasts Cell Line							
E002	ESC	ESC.WA7	ES-WA7 Cells					■		■
E008	ESC	ESC.H9	H9 Cells							
E001	ESC	ESC.I3	ES-I3 Cells					■		■
E015	ESC	ESC.HUES6	HUES6 Cells							■
E014	ESC	ESC.HUES48	HUES48 Cells							■
E016	ESC	ESC.HUES64	HUES64 Cells							■
E003	ESC	ESC.H1	H1 Cells							
E024	ESC	ESC.4STAR	ES-UCSF4 Cells					■	■	■
E020	iPSC	IPSC.20B	iPS-20b Cells							■
E019	iPSC	IPSC.18	iPS-18 Cells							■
E018	iPSC	IPSC.15b	iPS-15b Cells					■		■
E021	iPSC	IPSC.DF.6.9	iPS DF 6.9 Cells			H3K4me1_Enh			■	
E022	iPSC	IPSC.DF.19.11	iPS DF 19.11 Cells						■	
E007	ES-deriv	ESDR.H1.NEUR.PROG	H1 Derived Neuronal Progenitor Cultured Cells				H3K4me3_Pro			
E115	ENCODE2012	BLD.DND41.CNCR	Dnd41 TCell Leukemia Cell Line	7_Enh	19_DNase	H3K4me1_Enh		H3K27ac_Enh		■
E116	ENCODE2012	BLD.GM12878	GM12878 Lymphoblastoid Cells							
E117	ENCODE2012	CRVX.HELAS3.CNCR	HeLa-S3 Cervical Carcinoma Cell Line	7_Enh		H3K4me1_Enh				DNase
E118	ENCODE2012	LIV.HEPG2.CNCR	HepG2 Hepatocellular Carcinoma Cell Line							
E119	ENCODE2012	BRST.HMEC	HMEC Mammary Epithelial Primary Cells	7_Enh	16_EnhW1	H3K4me1_Enh				
E120	ENCODE2012	MUS.HSMM	HSMM Skeletal Muscle Myoblasts Cells							
E121	ENCODE2012	MUS.HSMMT	HSMM cell derived Skeletal Muscle Myotubes Cells							
E122	ENCODE2012	VAS.HUVEC	HUVEC Umbilical Vein Endothelial Primary Cells							
E123	ENCODE2012	BLD.K562.CNCR	K562 Leukemia Cells							
E124	ENCODE2012	BLD.CD14.MONO	Monocytes-CD14+ RO01746 Primary Cells							
E125	ENCODE2012	BRN.NHA	NH-A Astrocytes Primary Cells							
E126	ENCODE2012	SKIN.NHDFAD	NHDF-Ad Adult Dermal Fibroblast Primary Cells							
E127	ENCODE2012	SKIN.NHEK	NHEK-Epidermal Keratinocyte Primary Cells	7_Enh	16_EnhW1	H3K4me1_Enh				
E128	ENCODE2012	LNG.NHLF	NHLF Lung Fibroblast Primary Cells							
E129	ENCODE2012	BONE.OSTEO	Osteoblast Primary Cells						■	■

Proteins bound in ChIP-Seq experiments (ENCODE Project Consortium, 2011)

Cell ID	Protein
HeLa-S3	CEBPB

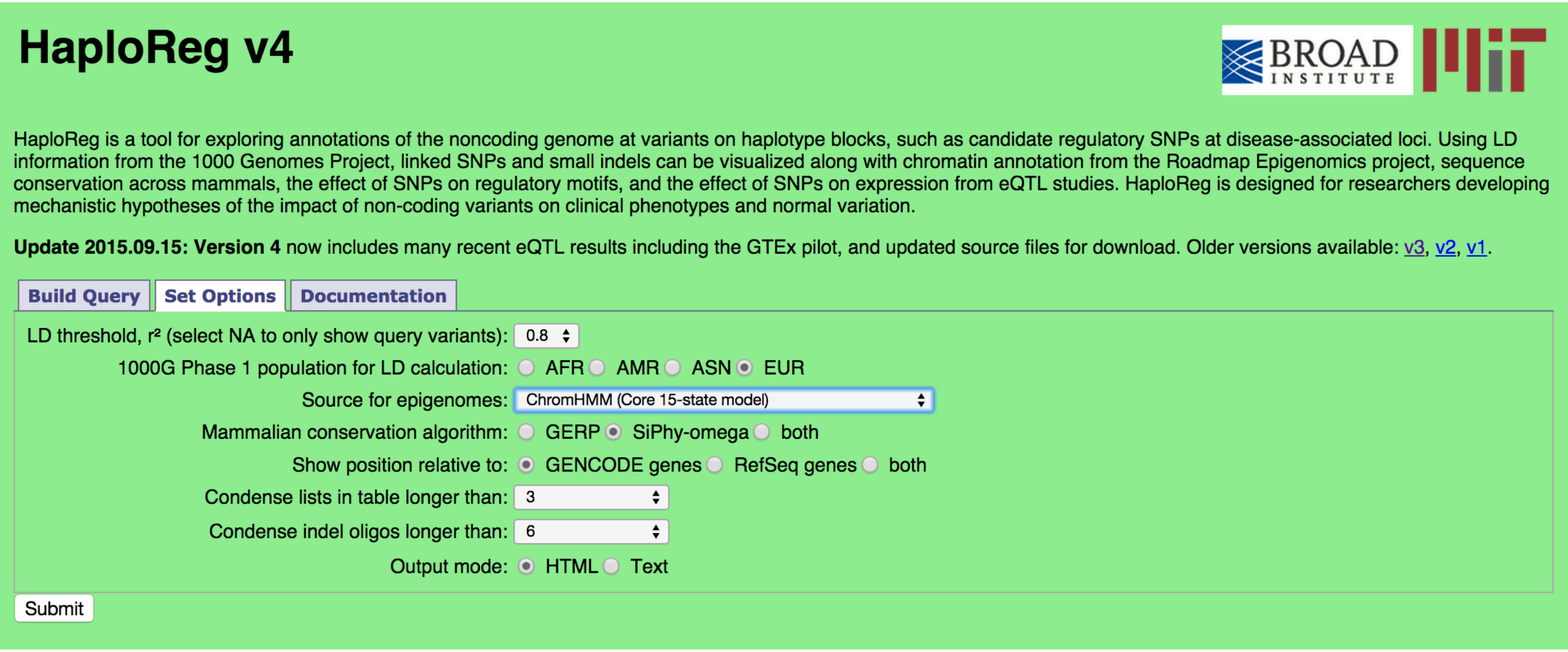
eQTL studies showing correlation of SNP with cis expression

Study ID	Paper Title	PMID	Tissue	Correlated gene
Zou2012	Brain expression genome-wide association study (eGWAS) identifies human disease-associated variants	22685416	Cerebellum	ATG4B
Zou2012	Brain expression genome-wide association study (eGWAS) identifies human disease-associated variants	22685416	Cerebellum	FAM110C
Zou2012	Brain expression genome-wide association study (eGWAS) identifies human disease-associated variants	22685416	Cerebellum	THAP4
Zou2012	Brain expression genome-wide association study (eGWAS) identifies human disease-associated variants	22685416	Temporal_Cortex	ATG4B
Zou2012	Brain expression genome-wide association study (eGWAS) identifies human disease-associated variants	22685416	Temporal_Cortex	FAM110C
Zou2012	Brain expression genome-wide association study (eGWAS) identifies human disease-associated variants	22685416	Temporal_Cortex	THAP4

Regulatory motifs altered

Position Weight Matrix ID
(Library from Kheradpour and Kellis, 2013)
	Strand	**Ref**	**Alt**	**Match on:**
				`Ref: CACACAAGATGGCTTAGGGCCAGGTTGCA`**`T`**`AATGTCCTTTTTCCTTCAGGAATGTGTGG`
				`Alt: CACACAAGATGGCTTAGGGCCAGGTTGCA`**`C`**`AATGTCCTTTTTCCTTCAGGAATGTGTGG`

AP-1_disc8	-	-31.6	-40.6	TMAYTTSCTT
CEBPA_2	-	10.4	11.3	WKDYRCAAY
CEBPB_disc1	-	12.4	14.8	RTTGYRCAAY
CEBPB_known1	+	11	11.4	NTTDCHHMABHH
CEBPB_known3	+	11.7	10.6	DNRTTGCDHMRDDN
CEBPB_known5	+	11.4	12.1	DKVTTRCDHMAYHN
GR_known3	+	6.1	6.3	KKYAYMRDVWGTYCTK
HLF	+	12.9	12.4	RTTACRYMAT
Hsf_disc1	+	13.5	12.3	VTTRYRYAAS
Myc_disc5	+	11.4	7.8	TTRCATCAKS
p300_disc2	+	12.4	11.4	NRTTKCAHMABHHHH



HaploReg v4

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 annotation from the Roadmap Epigenomics project, 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 2015.09.15: Version 4 now includes many recent eQTL results including the GTEx pilot, and updated source files for download. Older versions available: v3, v2, v1.

LD threshold, r^2 (select NA to only show query variants): 0.8

1000G Phase 1 population for LD calculation: ○ AFR ○ AMR ○ ASN ● EUR

Source for epigenomes: ChromHMM (Core 15-state model)

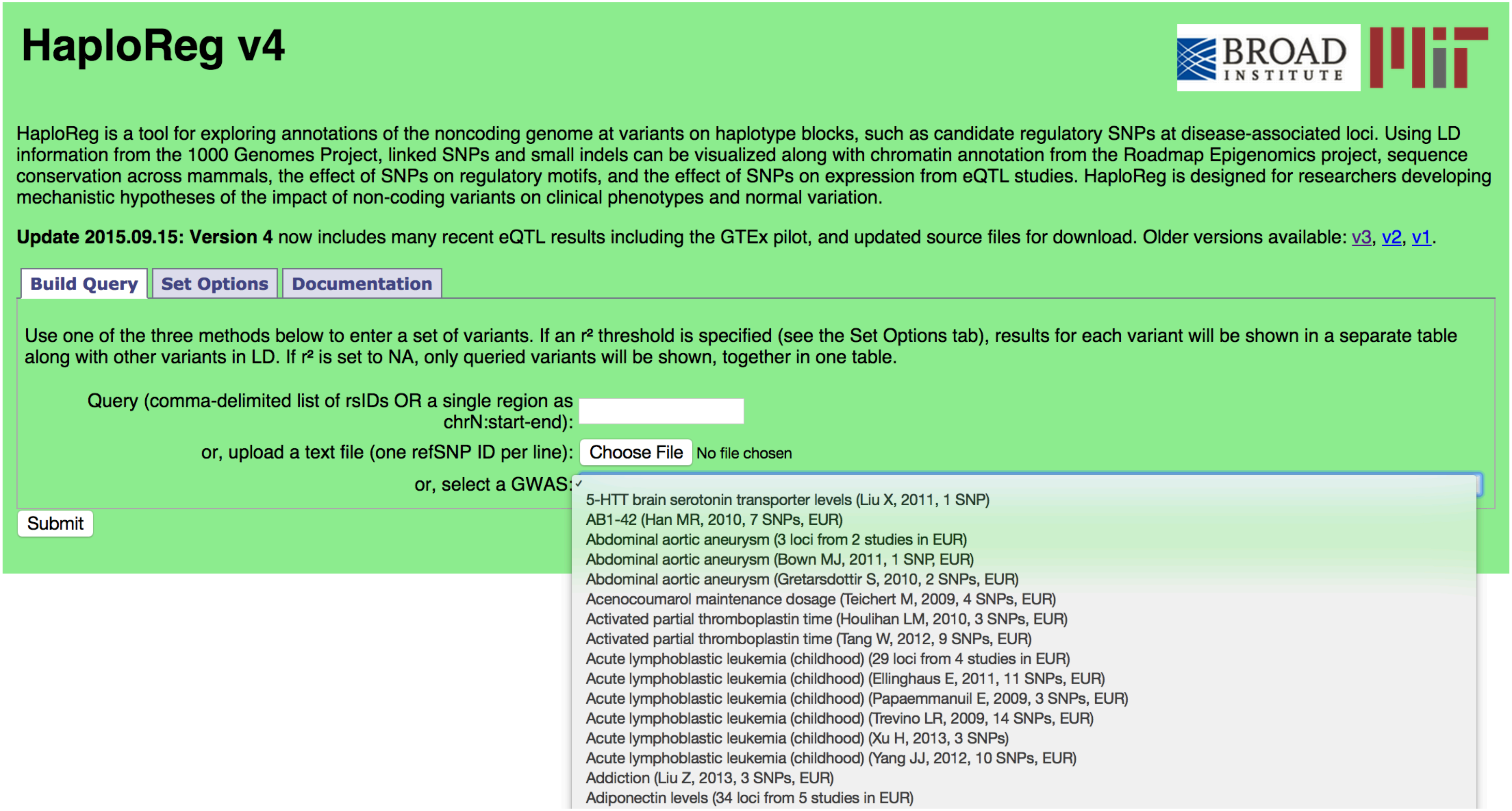
Mammalian conservation algorithm: ○ GERP ● SiPhy-omega ○ both

Show position relative to: ● GENCODE genes ○ RefSeq genes ○ both

Condense lists in table longer than: 3

Condense indel oligos longer than: 6

Output mode: ● HTML ○ Text



HaploReg v4

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 annotation from the Roadmap Epigenomics project, 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 2015.09.15: Version 4 now includes many recent eQTL results including the GTEx pilot, and updated source files for download. Older versions available: v3, v2, v1.

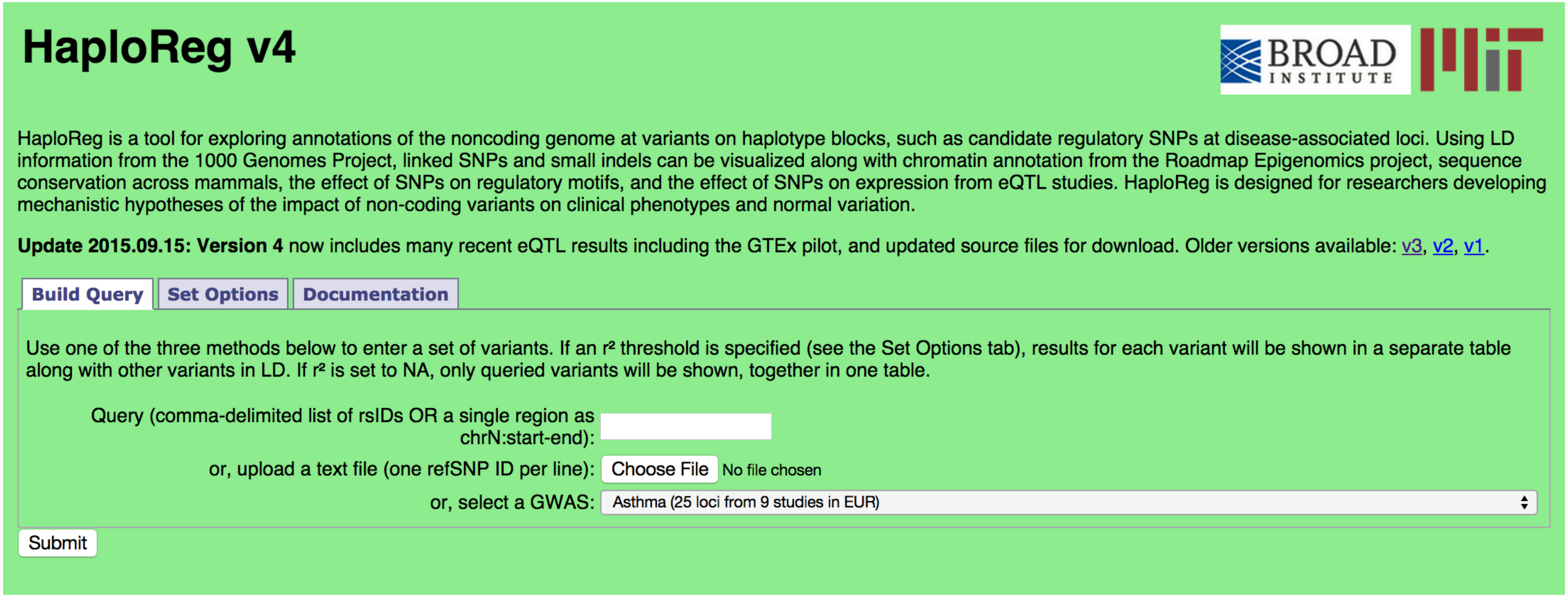
Use one of the three methods below to enter a set of variants. If an r² threshold is specified (see the Set Options tab), results for each variant will be shown in a separate table along with other variants in LD. If r² is set to NA, only queried variants will be shown, together in one table.

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

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

or, select a GWAS: ✓

5-HTT brain serotonin transporter levels (Liu X, 2011, 1 SNP)
AB1-42 (Han MR, 2010, 7 SNPs, EUR)
Abdominal aortic aneurysm (3 loci from 2 studies in EUR)
Abdominal aortic aneurysm (Bown MJ, 2011, 1 SNP, EUR)
Abdominal aortic aneurysm (Gretarsdottir S, 2010, 2 SNPs, EUR)
Acenocoumarol maintenance dosage (Teichert M, 2009, 4 SNPs, EUR)
Activated partial thromboplastin time (Houlihan LM, 2010, 3 SNPs, EUR)
Activated partial thromboplastin time (Tang W, 2012, 9 SNPs, EUR)
Acute lymphoblastic leukemia (childhood) (29 loci from 4 studies in EUR)
Acute lymphoblastic leukemia (childhood) (Ellinghaus E, 2011, 11 SNPs, EUR)
Acute lymphoblastic leukemia (childhood) (Papaemmanuil E, 2009, 3 SNPs, EUR)
Acute lymphoblastic leukemia (childhood) (Trevino LR, 2009, 14 SNPs, EUR)
Acute lymphoblastic leukemia (childhood) (Xu H, 2013, 3 SNPs)
Acute lymphoblastic leukemia (childhood) (Yang JJ, 2012, 10 SNPs, EUR)
Addiction (Liu Z, 2013, 3 SNPs, EUR)
Adiponectin levels (34 loci from 5 studies in EUR)
Submit



HaploReg v4

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 annotation from the Roadmap Epigenomics project, 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.

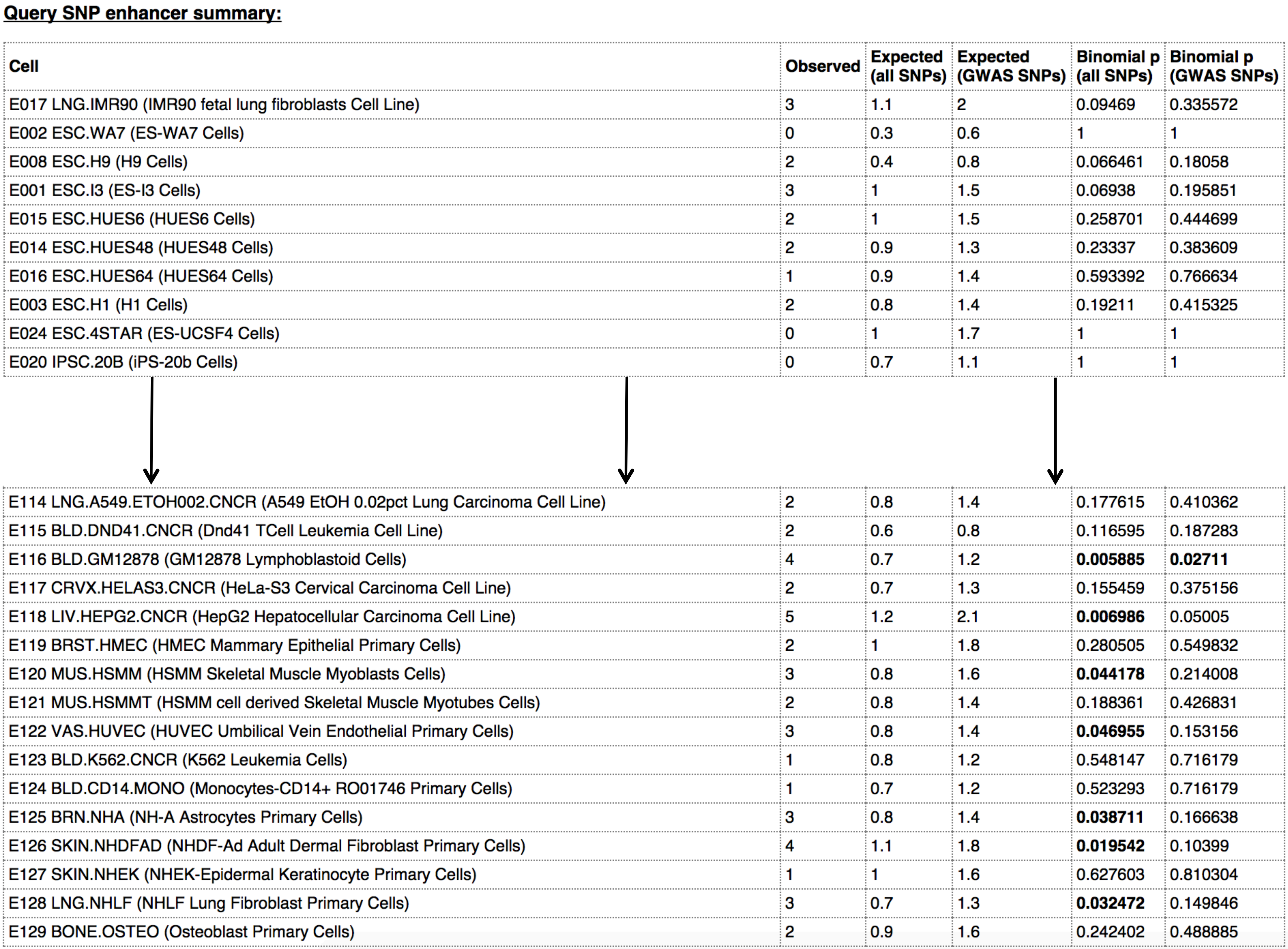
Update 2015.09.15: Version 4 now includes many recent eQTL results including the GTEx pilot, and updated source files for download. Older versions available: v3, v2, v1.

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):

or, upload a text file (one refSNP ID per line):

or, select a GWAS: Asthma (25 loci from 9 studies in EUR)



Query SNP enhancer summary:

Cell	Observed	Expected (all SNPs)	Expected (GWAS SNPs)	Binomial p (all SNPs)	Binomial p (GWAS SNPs)
E017 LNG.IMR90 (IMR90 fetal lung fibroblasts Cell Line)	3	1.1	2	0.09469	0.335572
E002 ESC.WA7 (ES-WA7 Cells)	0	0.3	0.6	1	1
E008 ESC.H9 (H9 Cells)	2	0.4	0.8	0.066461	0.18058
E001 ESC.I3 (ES-I3 Cells)	3	1	1.5	0.06938	0.195851
E015 ESC.HUES6 (HUES6 Cells)	2	1	1.5	0.258701	0.444699
E014 ESC.HUES48 (HUES48 Cells)	2	0.9	1.3	0.23337	0.383609
E016 ESC.HUES64 (HUES64 Cells)	1	0.9	1.4	0.593392	0.766634
E003 ESC.H1 (H1 Cells)	2	0.8	1.4	0.19211	0.415325
E024 ESC.4STAR (ES-UCSF4 Cells)	0	1	1.7	1	1
E020 IPSC.20B (iPS-20b Cells)	0	0.7	1.1	1	1
E114 LNG.A549.ETOH002.CNCR (A549 EtOH 0.02pct Lung Carcinoma Cell Line)	2	0.8	1.4	0.177615	0.410362
E115 BLD.DND41.CNCR (Dnd41 TCell Leukemia Cell Line)	2	0.6	0.8	0.116595	0.187283
E116 BLD.GM12878 (GM12878 Lymphoblastoid Cells)	4	0.7	1.2	**0.005885**	**0.02711**
E117 CRVX.HELAS3.CNCR (HeLa-S3 Cervical Carcinoma Cell Line)	2	0.7	1.3	0.155459	0.375156
E118 LIV.HEPG2.CNCR (HepG2 Hepatocellular Carcinoma Cell Line)	5	1.2	2.1	**0.006986**	0.05005
E119 BRST.HMEC (HMEC Mammary Epithelial Primary Cells)	2	1	1.8	0.280505	0.549832
E120 MUS.HSMM (HSMM Skeletal Muscle Myoblasts Cells)	3	0.8	1.6	**0.044178**	0.214008
E121 MUS.HSMMT (HSMM cell derived Skeletal Muscle Myotubes Cells)	2	0.8	1.4	0.188361	0.426831
E122 VAS.HUVEC (HUVEC Umbilical Vein Endothelial Primary Cells)	3	0.8	1.4	**0.046955**	0.153156
E123 BLD.K562.CNCR (K562 Leukemia Cells)	1	0.8	1.2	0.548147	0.716179
E124 BLD.CD14.MONO (Monocytes-CD14+ RO01746 Primary Cells)	1	0.7	1.2	0.523293	0.716179
E125 BRN.NHA (NH-A Astrocytes Primary Cells)	3	0.8	1.4	**0.038711**	0.166638
E126 SKIN.NHDFAD (NHDF-Ad Adult Dermal Fibroblast Primary Cells)	4	1.1	1.8	**0.019542**	0.10399
E127 SKIN.NHEK (NHEK-Epidermal Keratinocyte Primary Cells)	1	1	1.6	0.627603	0.810304
E128 LNG.NHLF (NHLF Lung Fibroblast Primary Cells)	3	0.7	1.3	**0.032472**	0.149846
E129 BONE.OSTEO (Osteoblast Primary Cells)	2	0.9	1.6	0.242402	0.488885

Acknowledgements

Weng Lab

Zhiping Weng
Michael Purcaro
Sowmya Iyer
Jie Wang
Arjan van der Velde

Stam Lab

John Stamatoyannopoulos
Bob Thurman
Richard Sandstrom

ENCODE Consortium

Brad Bernstein
Ross Hardison
Mark Gerstein
Data Production Groups

ROADMAP epigenomics PROJECT

ENCODE