Variant Annotation using ENCODE Data: An Introduction to RegulomeDB and HaploReg

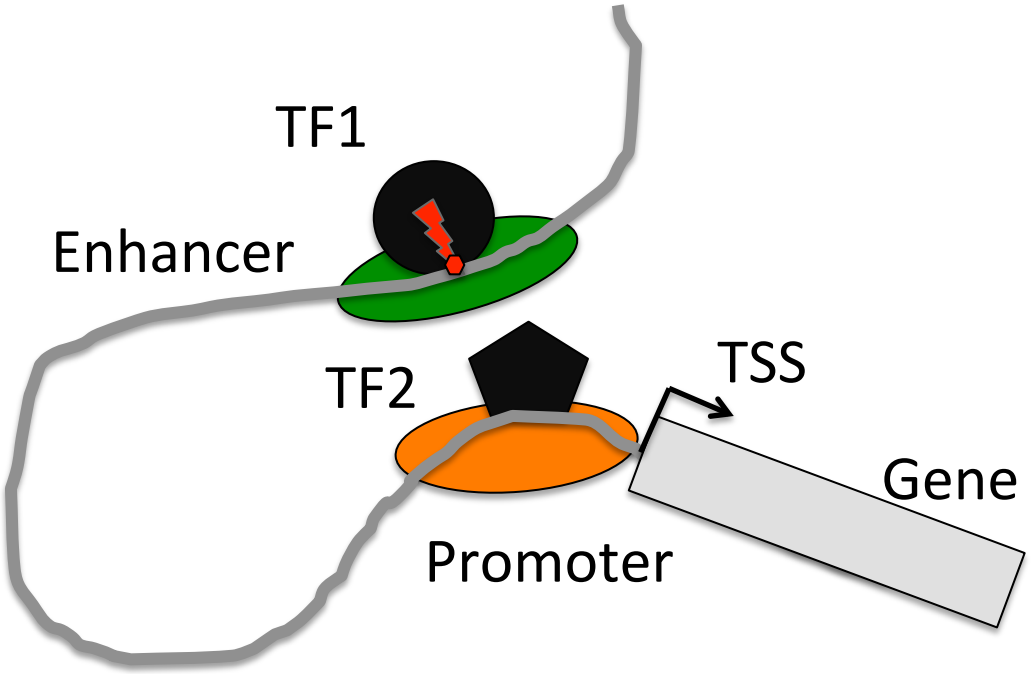
Collin Melton

Snyder Lab

Stanford University

ENCODE Data Provides Insight into the Function of Non-Coding Variants

- GWAS variants
- WGS study variants
  - Cancer
  - Inherited Disease



Variant Annotation Tools

http://regulomedb.org/

HaploReg v4.1

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*

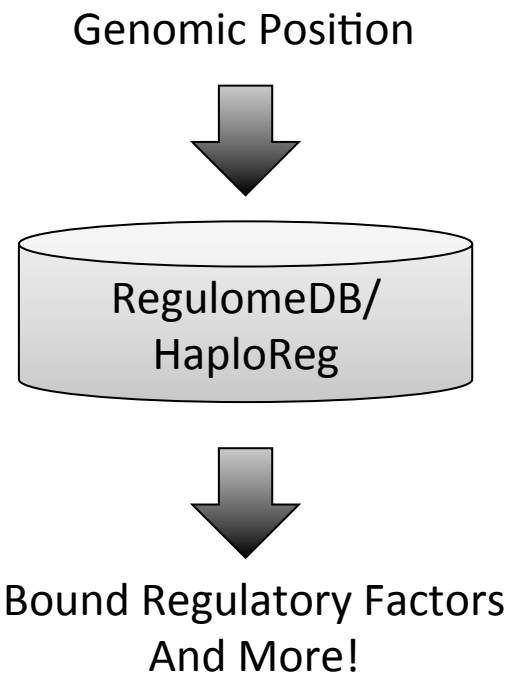
Nucleic Acids Research

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

These Tools Provide a Simple Web Interface for Retrieving Site-Specific Regulatory Data



RegulomeDB Classification Strategy

Score	eQTL	TF Binding	Matched TF Motif	Any TF Motif	Matched DNase Footprint	Any DNase Footprint	DNase peak
1a	x	x	x		x		x
1b	x	x		x		x	x
1c	x	x	x				x
1d	x	x		x			x
1e	x	x	x				
1f	x	x					x
2a		x	x		x		x
2b		x		x		x	x
2c		x	x				x
3a		x		x			x
3b		x	x				
4		x					x
5		x					x
6							

RegulomeDB Data Sources

Data Type	Types	Features
Transcription Factor ChIP-Seq (ENCODE)	707 conditions / cell lines	13,208,383
Transcription Factor ChIP-Seq (non-ENCODE)	32 conditions / cell lines	397,534
Transcription Factor ChIP-exo	1 condition	35,161
Chromatin States (REMC)	127 Cell Lines	55,605,005
DNase I Hypersensitive Sites	204 conditions / cell lines	42,738,084
FAIRE Sites	25 conditions / cell lines	4,816,196
DNase I Footprints	50 cell lines / 3 Methods	128,850,659
Predicted Binding (PWMs)	1,896 motifs (TRANSFAC, Jaspar, UniPROBE, Jolma et al.)	339,409,461
eQTLs	142,945 SNPs	142,945
dsQTLs	6,069 SNPs	6,069
Differentially Methylated Regions	1 Cell type	992
Manual Annotations	6 Genomic Regions	282
VISTA Enhancers	1,448 Enhancers	1,325
Validated SNPs affecting binding	855 SNPs	855

Example Usage of RegulomeDB

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 We have also added Chromatin States from the Roadmap Epigenome Consortium (unpublished) as well as updates to DNase footprinting, PWMs, and DNA Methylation.

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

```
# zero-based example - this is a comment and will be ignored
# Single nucleotides can be submitted
11    5248049  5248050
11:5248050-5248050
X:146993388..146993389
chrX:55041618-55041619
# Coordinate ranges can be submitted
3     128210000     128212040
11    5246900  5247000
19    12995238      12998702
11:5248050-5248050
14:100705102-100705102
```

Submit

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

The search has evaluated **16** input line(s) and found **37** SNP(s).

Summary of SNP analysis

Show 10 entries

Coordinate (0-based)	dbSNP ID	? Regulome DB Score	Other Resources
chrX:146993388	n/a	2a	UCSC | ENSEMBL
chrX:146993388	n/a	2a	UCSC | ENSEMBL
chr19:12995421	rs16978757	2b	UCSC | ENSEMBL | dbSNP
chr19:12995421	rs16978757	2b	UCSC | ENSEMBL | dbSNP
chr19:12996739	rs2072597	2b	UCSC | ENSEMBL | dbSNP
chr19:12996739	rs2072597	2b	UCSC | ENSEMBL | dbSNP
chr19:12998101	rs79334031	2b	UCSC | ENSEMBL | dbSNP
chr19:12998101	rs79334031	2b	UCSC | ENSEMBL | dbSNP
chr3:128210062	rs4496503	2b	UCSC | ENSEMBL | dbSNP
chr3:128210062	rs4496503	2b	UCSC | ENSEMBL | dbSNP

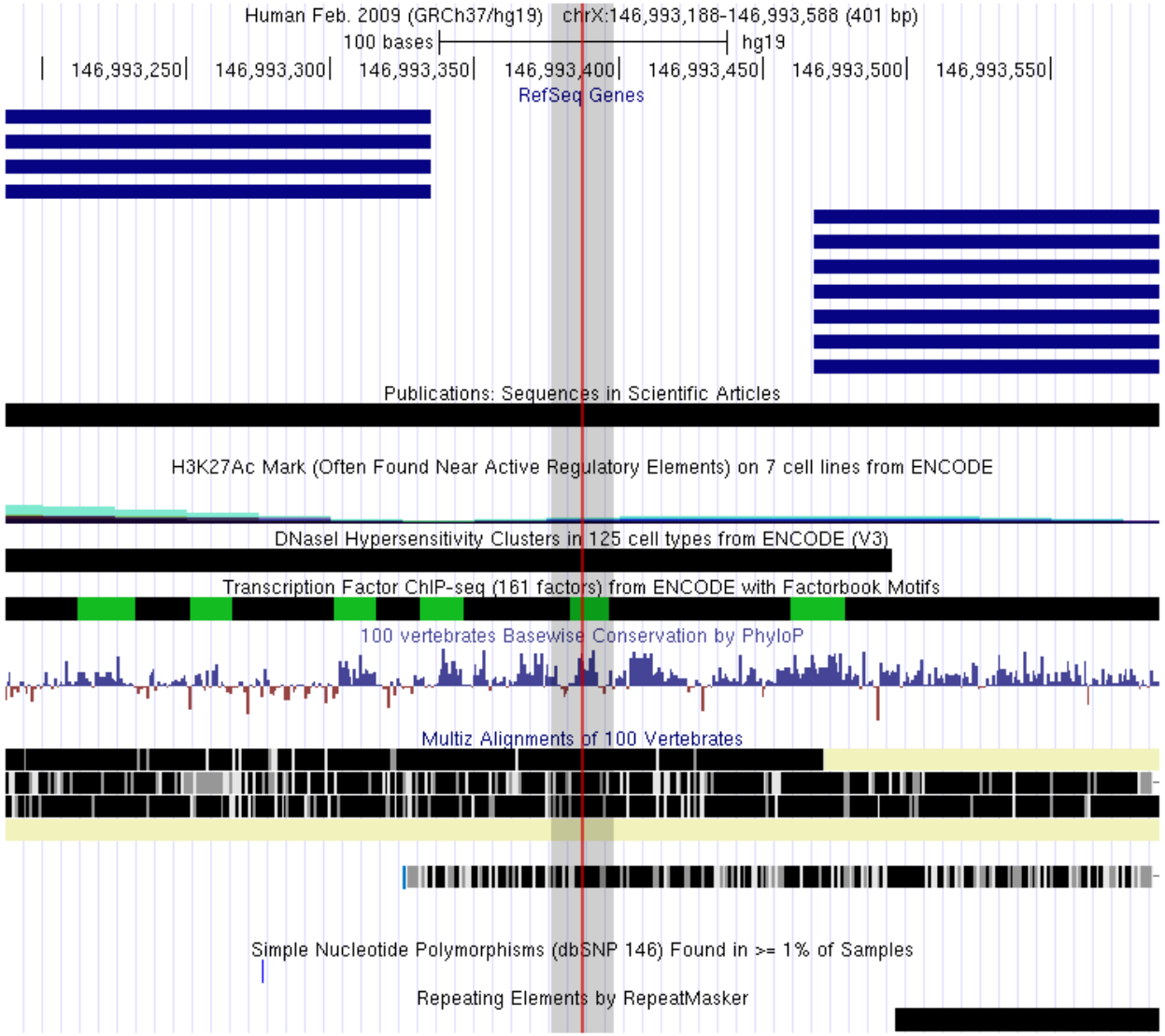
Showing 1 to 10 of 37 entries

Download | BED | GFF | Full Output

Data supporting chrX:146993388 (n/a)

Score: 2a

Likely to affect binding



Protein Binding Filter:

Protein Binding

Filter:

Method	Location	Bound Protein	? Cell Type	Additional Info	Reference
ChIP	chrX:146993319..146993471	TFAP2A	Hela		15930016
ChIP-seq	chrX:146992920..146993464	POLR2A	HeLa-S3		ENCODE
ChIP-seq	chrX:146992934..146993478	E2F1	MCF-7		ENCODE
ChIP-seq	chrX:146992934..146993478	E2F1	MCF-7		ENCODE
ChIP-seq	chrX:146992937..146993427	POLR2A	GM12878		ENCODE
ChIP-seq	chrX:146992958..146993468	E2F6	HeLa-S3		ENCODE
ChIP-seq	chrX:146992966..146993422	E2F4	HeLa-S3		ENCODE
ChIP-seq	chrX:146992970..146993500	POLR2A	GM19193		ENCODE
ChIP-seq	chrX:146992975..146993463	POLR2A	GM12878		ENCODE
ChIP-seq	chrX:146992976..146993500	POLR2A	GM18505		ENCODE
ChIP-seq	chrX:146992984..146993464	POLR2A	GM12892		ENCODE
ChIP-seq	chrX:146992993..146993457	POLR2A	HCT-116		ENCODE
ChIP-seq	chrX:146993002..146993438	POLR2A	GM18526		ENCODE
ChIP-seq	chrX:146993024..146993408	MYC	K562	ifng30	ENCODE
ChIP-seq	chrX:146993025..146993402	E2F6	K562		ENCODE
ChIP-seq	chrX:146993385..146993589	POLR2A	HepG2		ENCODE
ChIP-seq	chrX:146993006..146993430	POLR2A	K562	ifng30	ENCODE
ChIP-seq	chrX:146993004..146993440	POLR2A	K562		ENCODE
ChIP-seq	chrX:146993019..146993469	POLR2A	NB4		ENCODE
ChIP-seq	chrX:146993007..146993477	POLR2A	PANC-1		ENCODE
ChIP-seq	chrX:146992963..146993479	POLR2A	Raji		ENCODE
ChIP-seq	chrX:146993268..146993592	STAT2	K562	ifna6h	ENCODE
ChIP-seq	chrX:146993275..146993585	TAF1	H1-hESC		ENCODE

Motifs

Filter:

Method	Location	Motif	Cell Type	PWM	Reference
DMS footprinting	chrX:146993383..146993394		Fibroblast		9328468
DMS footprinting	chrX:146993386..146993392		Fibroblast		9199556
Footprinting	chrX:146993357..146993399		SkMC		24071585
Footprinting	chrX:146993357..146993401		K562		24071585
Footprinting	chrX:146993382..146993396	Stra13	Helas3		21106904
Footprinting	chrX:146993382..146993396	USF	Chorion		21106904
Footprinting	chrX:146993382..146993396	USF	Chorion		21106904
Footprinting	chrX:146993382..146993396	Stra13	AosmcSerumfree		21106904
Footprinting	chrX:146993382..146993396	USF	Helas3		21106904
Footprinting	chrX:146993382..146993396	USF	Helas3		21106904
Footprinting	chrX:146993382..146993396	USF	Helas3		21106904

Chromatin structure

Filter:

Method	Location	Cell Type	Additional Info	Reference
DNase-seq	chrX:146992606..146995122	Psoasmuscleoc		ENCODE
DNase-seq	chrX:146992761..146995023	Myometr		ENCODE
DNase-seq	chrX:146992448..146993661	Hepatocytes		ENCODE
DNase-seq	chrX:146992595..146994377	Cerebellumoc		ENCODE
DNase-seq	chrX:146992601..146994458	Gcbcell		ENCODE
DNase-seq	chrX:146992675..146994460	Progfib		ENCODE
DNase-seq	chrX:146992687..146994410	Adultcd4th0		ENCODE
DNase-seq	chrX:146992687..146994614	Heartoc		ENCODE
DNase-seq	chrX:146992697..146994421	Huvec		ENCODE
DNase-seq	chrX:146992727..146994513	Frontalcortexoc		ENCODE
DNase-seq	chrX:146992765..146994349	Adultcd4th1		ENCODE
DNase-seq	chrX:146992767..146994456	Gm20000		ENCODE
DNase-seq	chrX:146992775..146994531	H1hesc		ENCODE
DNase-seq	chrX:146992780..146994452	8988t		ENCODE
DNase-seq	chrX:146992781..146994360	Fibropag20443		ENCODE
DNase-seq	chrX:146992781..146994360	Fibrop		ENCODE
DNase-seq	chrX:146992781..146994491	Gliobla		ENCODE
DNase-seq	chrX:146992775..146994726	Hsmm		ENCODE
DNase-seq	chrX:146992782..146994427	Naivebcell		ENCODE
DNase-seq	chrX:146992788..146994776	A549		ENCODE

Histone modifications Filter:

Method	Location	Chromatin State	Tissue Group	Tissue	Reference
ChromHMM	chrX:146668200..146994400	Quiescent/Low	Heart	Fetal Heart	REMC
ChromHMM	chrX:146930600..146993800	Quiescent/Low	Other	Placenta Amnion	REMC
ChromHMM	chrX:146808200..146993400	Quiescent/Low	Digestive	Small Intestine	REMC
ChromHMM	chrX:146989600..146993800	Weak Repressed PolyComb	ESC	ES-WA7 Cell Line	REMC
ChromHMM	chrX:146962600..146993600	Quiescent/Low	Digestive	Sigmoid Colon	REMC
ChromHMM	chrX:146992400..146994600	Active TSS	Digestive	Rectal Mucosa Donor 31	REMC
ChromHMM	chrX:146992400..146995000	Active TSS	Blood & T-cell	Primary T CD8+ memory cells from peripheral blood	REMC
ChromHMM	chrX:146992400..146995600	Active TSS	Sm. Muscle	Duodenum Smooth Muscle	REMC
ChromHMM	chrX:146992400..146996600	Active TSS	ES-deriv	hESC Derived CD184+ Endoderm Cultured Cells	REMC
ChromHMM	chrX:146992600..146994800	Active TSS	Mesench	Mesenchymal Stem Cell Derived Adipocyte Cultured Cells	REMC
ChromHMM	chrX:146992600..146994800	Active TSS	Adipose	Adipose Nuclei	REMC
ChromHMM	chrX:146992600..146994800	Active TSS	Brain	Brain Inferior Temporal Lobe	REMC
ChromHMM	chrX:146992600..146995000	Active TSS	Mesench	Bone Marrow Derived Cultured Mesenchymal Stem Cells	REMC

Related data Filter:

Method	Location	? Cell Type	Annotation	Reference
Manual	chrX:146992974..146993653		Intergenic Region	7825604

Programmatic Access

https://github.com/aboyle/RegulomeDB-Tools