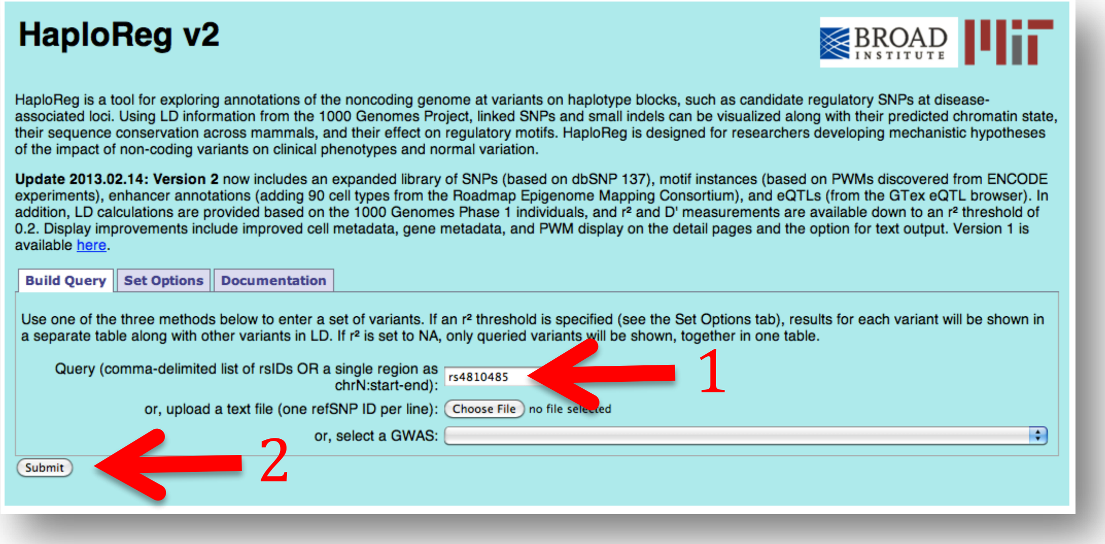
Using HaploReg and RegulomeDB to mine ENCODE data:

(Updated 17 May 2013, Mike Pazin)

HaploReg and RegulomeDB are ENCODE-funded tools described in recent publications that retrieve ENCODE annotations at SNPs of interest, as well as annotations from work by other researchers and projects.

HaploReg v2:

Go to the HaploReg site, and enter the name of the SNP of interest (Arrow 1). (Using the "Set Options" tab, the user can configure values such as the LD threshold and the population used from 1000 Genomes data used to calculate LD.)
Click on the submit button (Arrow 2)



HaploReg v2

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 r² and D' measurements are available down to an r² 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.

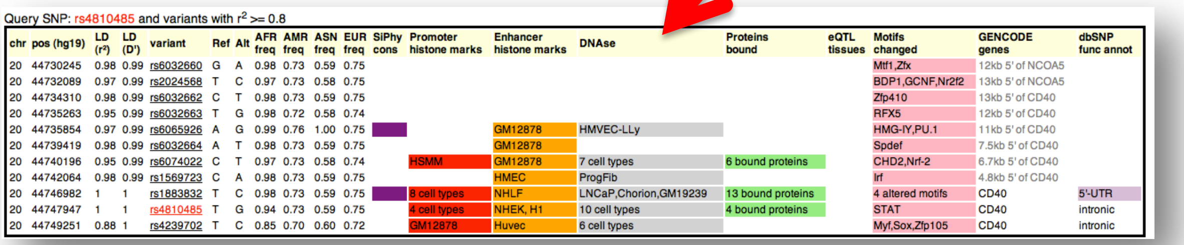
Use one of the three methods below to enter a set of variants. If an r² threshold is specified (see the Set Options tab), results for each variant will be shown in a separate table along with other variants in LD. If r² is set to NA, only queried variants will be shown, together in one table.

Query (comma-delimited list of rsIDs OR a single region as chrN:start-end): rs4810485

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

or, select a GWAS:

HaploReg retrieves the ENCODE and Roadmap Epigenomics annotations for the selected SNP, as well as other SNPs in LD (arrow 3).

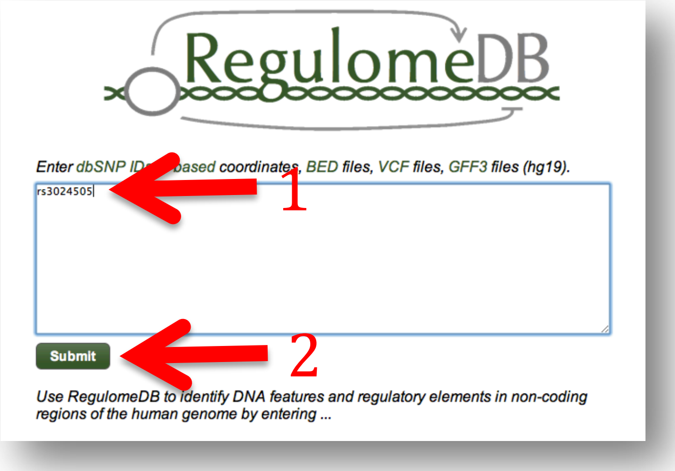


Query SNP: rs4810485 and variants with r^2 >= 0.8

chr	pos (hg19)	LD (r²)	LD (D')	variant	Ref	Alt	AFR freq	AMR freq	ASN freq	EUR freq	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse	Proteins bound	eQTL tissues	Motifs changed	GENCODE genes	dbSNP func annot
20	44730245	0.98	0.99	rs6032660	G	A	0.98	0.73	0.59	0.75							Mtf1,Zfx	12kb 5' of NCOA5	
20	44732089	0.97	0.99	rs2024568	T	C	0.97	0.73	0.58	0.75							BDP1,GCNF,Nr2f2	13kb 5' of NCOA5	
20	44734310	0.98	0.99	rs6032662	C	T	0.98	0.73	0.59	0.75							Zfp410	13kb 5' of CD40	
20	44735263	0.95	0.99	rs6032663	T	G	0.98	0.72	0.58	0.74							RFX5	12kb 5' of CD40	
20	44735854	0.97	0.99	rs6065926	A	G	0.99	0.76	1.00	0.75	🟪		GM12878	HMVEC-LLy			HMG-IY,PU.1	11kb 5' of CD40	
20	44739419	0.98	0.99	rs6032664	A	T	0.98	0.73	0.59	0.75			GM12878				Spdef	7.5kb 5' of CD40	
20	44740196	0.95	0.99	rs6074022	C	T	0.97	0.73	0.58	0.74		HSMM	GM12878	7 cell types	6 bound proteins		CHD2,Nrf-2	6.7kb 5' of CD40	
20	44742064	0.98	0.99	rs1569723	C	A	0.98	0.73	0.59	0.75			HMEC	ProgFib			Irf	4.8kb 5' of CD40	
20	44746982	1	1	rs1883832	T	C	0.98	0.73	0.59	0.75	🟪	8 cell types	NHLF	LNCaP,Chorion,GM19239	13 bound proteins		4 altered motifs	CD40	5'-UTR
20	44747947	1	1	rs4810485	T	G	0.94	0.73	0.59	0.75		4 cell types	NHEK, H1	10 cell types	4 bound proteins		STAT	CD40	intronic
20	44749251	0.88	1	rs4239702	T	C	0.85	0.70	0.60	0.72		GM12878	Huvec	6 cell types			Myf,Sox,Zfp105	CD40	intronic

RegulomeDB:

Go to the RegulomeDB site and enter the name of the SNP of interest (Arrow 1).



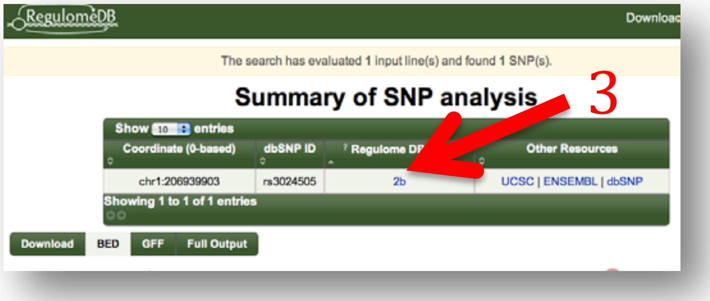
Enter dbSNP IDs, 0-based coordinates, BED files, VCF files, GFF3 files (hg19).

rs3024505

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

Click on the submit button (Arrow 2).

RegulomeDB calculates a score for the regulatory potential of this region.



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

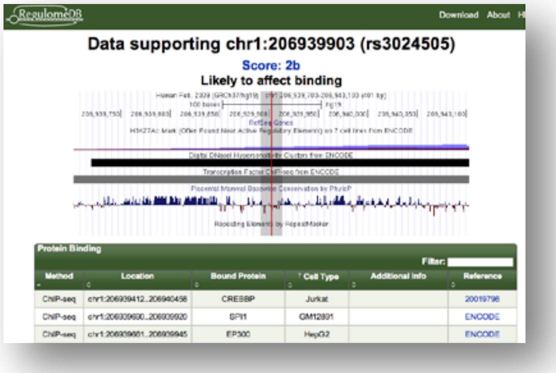
Summary of SNP analysis

Show 10 entries

Coordinate (0-based)	dbSNP ID	Regulome DB	Other Resources
chr1:206939903	rs3024505	2b	UCSC | ENSEMBL | dbSNP

Showing 1 to 1 of 1 entries

Clicking on the score (arrow 3) retrieves the ENCODE (and other) annotation for the region, including transcription factor binding, chromatin structure (DNase, FAIRE, and histone modifications), transcription factor motifs and eQTL.



RegulomeDB
Download
About

Data supporting chr1:206939903 (rs3024505)

Score: 2b

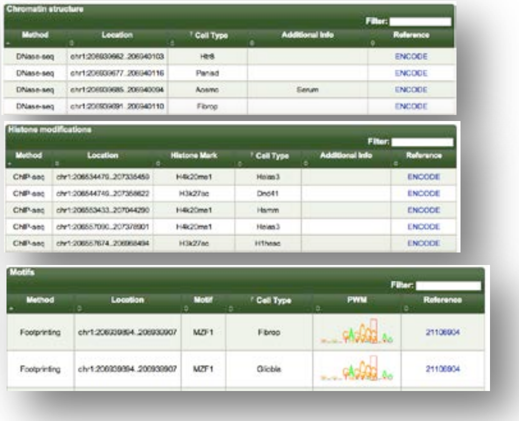
Likely to affect binding

Human Feb. 2009 (GRCh37/hg19) chr1:206,939,703-206,940,103 (401 bp)
100 bases
hg19
RefSeq Genes
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Digital DNaseI Hypersensitivity Clusters from ENCODE
Transcription Factor ChIP-seq from ENCODE
Placental Mammal Basewise Conservation by PhyloP
Repeating Elements by RepeatMasker

Protein Binding

Filter:

Method	Location	Bound Protein	Cell Type	Additional Info	Reference
ChIP-seq	chr1:206939412..206940458	CREBBP	Jurkat		20019798
ChIP-seq	chr1:206939630..206939920	SPI1	GM12891		ENCODE
ChIP-seq	chr1:206939861..206939945	EP300	HepG2		ENCODE



Chromatin structure

Filter:

Method	Location	Cell Type	Additional Info	Reference
DNase-seq	chr1:206939662..206940103	Hrt8		ENCODE
DNase-seq	chr1:206939677..206940116	Panisd		ENCODE
DNase-seq	chr1:206939685..206940094	Aosmc	Serum	ENCODE
DNase-seq	chr1:206939691..206940110	Fibrop		ENCODE

Histone modifications

Filter:

Method	Location	Histone Mark	Cell Type	Additional Info	Reference
ChIP-seq	chr1:206534470..207336459	H4k20me1	Helas3		ENCODE
ChIP-seq	chr1:206544749..207358822	H3k27ac	Dnd41		ENCODE
ChIP-seq	chr1:206553433..207044290	H4k20me1	Hsmm		ENCODE
ChIP-seq	chr1:206557090..207378901	H4k20me1	Helas3		ENCODE
ChIP-seq	chr1:206557674..206968494	H3k27ac	H1hesc		ENCODE

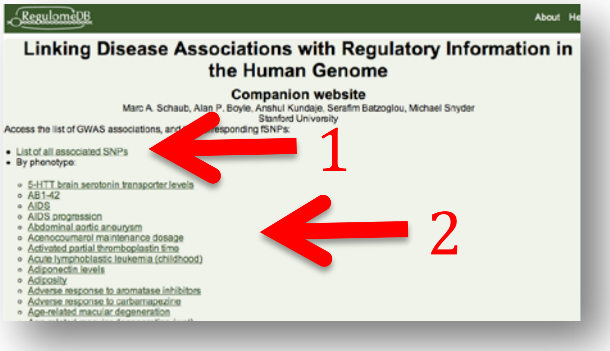
Motifs

Filter:

Method	Location	Motif	Cell Type	PWM	Reference
Footprinting	chr1:206939894..206939907	MZF1	Fibrop		21106904
Footprinting	chr1:206939894..206939907	MZF1	Gliobla		21106904

RegulomeDB Disease Association Database, a database of predicted functional SNPs, organized by disease/trait and by SNP, is available at: http://regulome.stanford.edu/GWAS

There is a list of over 4700 SNPs associated with human traits and disease (arrow 1), as well as a list of over 470 human traits and diseases (arrow 2).



Linking Disease Associations with Regulatory Information in the Human Genome

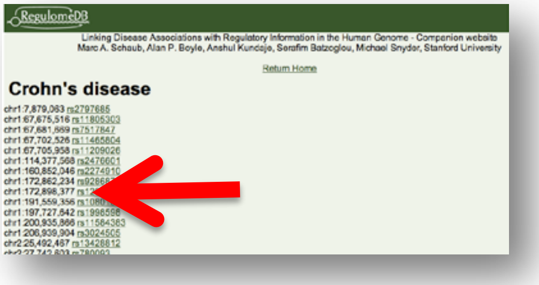
Companion website

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

Access the list of GWAS associations, and [illegible] esponding 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 carbamazepine
  - Age-related macular degeneration

Clicking on a trait/disease returns a list of SNPs that have been associated with that trait or disease:



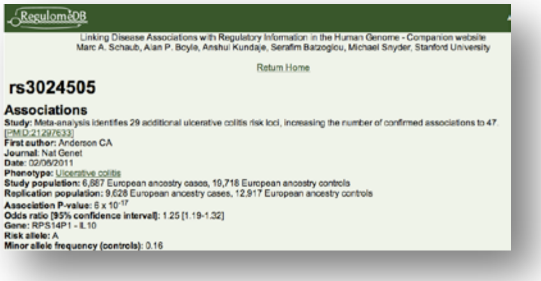
Linking Disease Associations with Regulatory Information in the Human Genome - Companion website
Marc A. Schaub, Alan P. Boyle, Anshul Kundaje, Serafim Batzoglou, Michael Snyder, Stanford University

Return Home

Crohn's disease

chr1:7,879,063 rs2797685
chr1:67,675,516 rs11805303
chr1:67,681,669 rs7517847
chr1:67,702,526 rs11465804
chr1:67,705,958 rs11209026
chr1:114,377,568 rs2476601
chr1:160,852,046 rs2274910
chr1:172,862,234 rs92868
chr1:172,898,377 rs1
chr1:191,559,356 rs1080
chr1:197,727,642 rs1998598
chr1:200,935,866 rs11584363
chr1:206,939,904 rs3024505
chr2:25,492,467 rs13428812

Clicking on a SNP (red arrow) returns the evidence for the association:



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

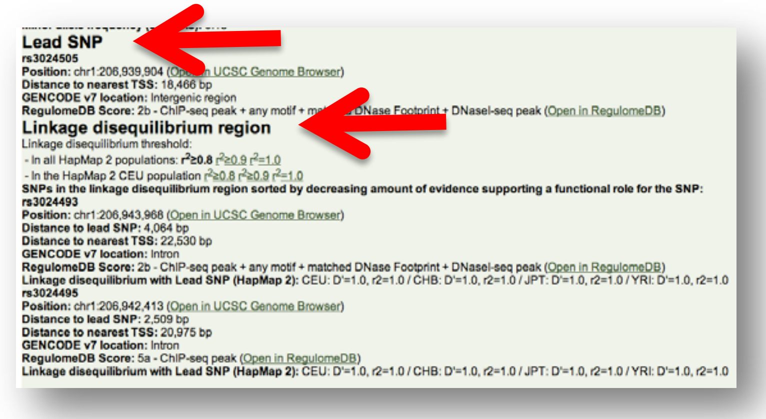
Return Home

rs3024505

Associations

Study: Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. [PMID:21297633]
First author: Anderson CA
Journal: Nat Genet
Date: 02/06/2011
Phenotype: Ulcerative colitis
Study population: 6,687 European ancestry cases, 19,718 European ancestry controls
Replication population: 9,628 European ancestry cases, 12,917 European ancestry controls
Association P-value: 6×10^{-17}
Odds ratio [95% confidence interval]: 1.25 [1.19-1.32]
Gene: RPS14P1 - IL10
Risk allele: A
Minor allele frequency (controls): 0.16

As well as the annotation for the lead SNP, and other SNPs in LD that, based on functional annotation, are candidates for the functional variant:



Lead SNP

rs3024505
Position: chr1:206,939,904 (Open in UCSC Genome Browser)
Distance to nearest TSS: 18,466 bp
GENCODE v7 location: Intergenic region
RegulomeDB Score: 2b - ChIP-seq peak + any motif + matched DNase Footprint + DNaseI-seq peak (Open in RegulomeDB)

Linkage disequilibrium region

Linkage disequilibrium threshold:
- In all HapMap 2 populations: $\mathbf{r^2 \geq 0.8}$ $r^2 \geq 0.9$ $r^2=1.0$
- In the HapMap 2 CEU population $r^2 \geq 0.8$ $r^2 \geq 0.9$ $r^2=1.0$

SNPs in the linkage disequilibrium region sorted by decreasing amount of evidence supporting a functional role for the SNP:
rs3024493
Position: chr1:206,943,968 (Open in UCSC Genome Browser)
Distance to lead SNP: 4,064 bp
Distance to nearest TSS: 22,530 bp
GENCODE v7 location: Intron
RegulomeDB Score: 2b - ChIP-seq peak + any motif + matched DNase Footprint + DNaseI-seq peak (Open in RegulomeDB)
Linkage disequilibrium with Lead SNP (HapMap 2): CEU: D'=1.0, r2=1.0 / CHB: D'=1.0, r2=1.0 / JPT: D'=1.0, r2=1.0 / YRI: D'=1.0, r2=1.0
rs3024495
Position: chr1:206,942,413 (Open in UCSC Genome Browser)
Distance to lead SNP: 2,509 bp
Distance to nearest TSS: 20,975 bp
GENCODE v7 location: Intron
RegulomeDB Score: 5a - ChIP-seq peak (Open in RegulomeDB)
Linkage disequilibrium with Lead SNP (HapMap 2): CEU: D'=1.0, r2=1.0 / CHB: D'=1.0, r2=1.0 / JPT: D'=1.0, r2=1.0 / YRI: D'=1.0, r2=1.0

One can follow the links to view the genomic annotation of these SNPs in the genome browser.