

## RegulomeDB and HaploReg Exercises

---

### Exercise #1

rs2816316 has been associated with Celiac Disease in the European population by two studies (Hunt, ..., van Heel (2008) *Nature Genetics* and Dubois, ..., van Heel (2010) *Nature Genetics*). rs2816316 lies thousands of base pairs upstream of protein coding gene RGS1 in an intergenic region of the genome.

You decide to further investigate this SNP using RegulomeDB and HaploReg.

1. What score does RegulomeDB assign to rs2816316? Is this SNP likely to affect transcription factor binding?
  2. Using HaploReg, determine if there are there any SNPs in high LD with rs2816316. Are any of these SNPs more likely to be causal?
  3. Using RegulomeDB, determine the scores for each of the SNPs in LD with rs2816316 that you think may be casual. Is there a SNP that is likely to affect transcription factor binding? Which SNP(s) would you further investigate?
- 

### Exercise #2

You are interested in studying genetics variants associated with Amyotrophic lateral sclerosis (ALS), which causes muscle atrophy due to the degeneration of motor neurons. Eleven studies have reported 66 SNPs associated with ALS. Since little is known about the disease, you decided to investigate these genetic variants.

1. Using HaploReg, determine if there are enrichments for enhancers in any ENCODE cell types for these ALS SNPs. Are there enrichments in DNase regions?
  2. Perform the same analysis using Roadmap epigenomes. Are disease relevant tissue and cell types enriched?
- 

### Exercise #3

rs6774494 has been associated with Nasopharyngeal carcinoma in the Chinese population (Bei, ..., Zeng (2010) *Nature Genetics*)

1. What score does RegulomeDB assign to rs6774494? Is this SNP likely to affect transcription factor binding?

2. Using HaploReg, determine if there are there any SNPs in high LD with rs2816316. Are any of these SNPs more likely to be causal?
  3. How would your results change if you used the default settings for HaploReg (i.e. European LD?)
-

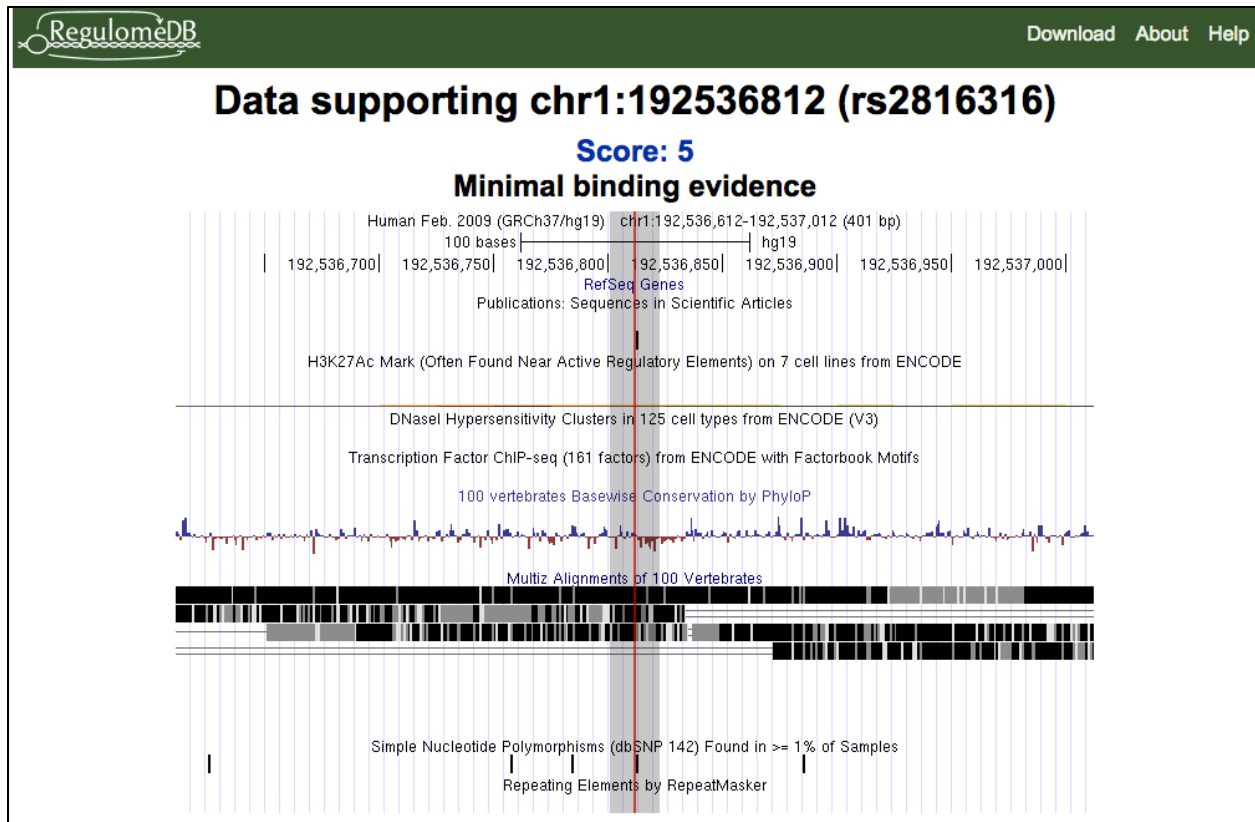
## SOLUTIONS

### Exercise #1

rs2816316 has been associated with Celiac Disease in the European population by two studies (Hunt, ..., van Heel (2008) *Nature Genetics* and Dubois, ..., van Heel (2010) *Nature Genetics*). rs2816316 lies thousands of base pairs up stream of protein coding gene RGS1 in an intergenic region of the genome.

You decide to further investigate this SNP using RegulomeDB and HaploReg.

1. What score does RegulomeDB assign to rs2816316? Is this SNP likely to affect transcription factor binding?  
RegulomeDB assigns rs2816316 a score of 5, which means that there is minimal binding evidence. This SNP is not likely to affect TF binding



2. Using HaploReg, determine if there are there any SNPs in high LD with rs2816316. Are any of these SNPs more likely to be causal?  
There are 25 SNPs in LD ( $r^2 > 0.8$ ) with rs2816316. There are three SNPs that overlap TF binding sites: rs2816305, rs2984920 and rs7535818. These SNPs also overlap DHSs, promoter marks, and enhancer marks for several cells lines.

Query SNP: rs2816316 and variants with r<sup>2</sup> >= 0.8

chr	pos (hg19)	LD (r <sup>2</sup> )	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		
1	192519865	0.92	0.97	rs1323298	T	C	0.97	0.80	0.78	0.81			5 cell types					5 altered motifs	RPS-1011O1.2		
1	192523556	0.81	0.98	rs9903778	T	C	0.77	0.77	0.76	0.78			GM12878					12 altered motifs	RPS-1011O1.2		
1	192523571	0.84	0.98	rs9804163	C	T	0.78	0.77	0.76	0.79			GM12878					Ets, GZF1	RPS-1011O1.2		
1	192524269	0.96	0.99	rs2815326	C	T	0.87	0.78	0.77	0.81			GM12878	5 cell types	8 bound proteins			LEP-1, NF-1, Ptx-8	RPS-1011O1.2		
1	192524885	0.96	0.99	rs4658064	G	T	0.80	0.78	0.76	0.81			GM12878					6 altered motifs	RPS-1011O1.2		
1	192524993	0.97	0.99	rs3011685	C	T	0.97	0.79	0.76	0.81			GM12878					Foxl1, Foxp3	RPS-1011O1.2		
1	192525276	0.96	0.99	rs2760521	G	A	0.87	0.78	0.76	0.81			GM12878					4 altered motifs	RPS-1011O1.2		
1	192525984	0.94	0.97	rs9427262	A	C	0.88	0.78	0.76	0.81								CEBPB	RPS-1011O1.2		
1	192529414	0.96	0.98	rs2760522	G	G	0.97	0.79	0.77	0.81			5 cell types	HRGEC				Cdx2, Hoxa10, Hoxa9	RPS-1011O1.2		
1	192530548	0.96	0.98	rs2760524	A	G	0.97	0.79	0.77	0.81			NHLF, Huvec	Medullo					RPS-1011O1.2		
1	192530967	0.95	0.97	rs1070921	GT	G	0.80	0.77	0.77	0.81			GM12878, NHLF					Foxp3, Pou2f2	RPS-1011O1.2		
1	192535553	0.96	1	rs1358063	G	A	0.33	0.69	0.59	0.81								Gli1b, NF-kappaB	RPS-1011O1.2		
1	192536385	1	1	rs2815317	G	A	0.83	0.78	0.76	0.81								GR, Nimg	37bp 5' of RPS-1011O1.2		
1	192536451	0.98	1	rs4658437	G	T	0.40	0.75	0.69	0.81								4 altered motifs	102bp 5' of RPS-1011O1.2		
1	192536758	1	1	rs2760527	G	A	0.83	0.78	0.76	0.81				H7-HESC				410bp 5' of RPS-1011O1.2			
1	192536785	1	1	rs2760528	G	A	0.83	0.78	0.76	0.81				H7-HESC				Pou2f2, STAT	437bp 5' of RPS-1011O1.2		
1	192536813	1	1	rs2815318	C	A	0.74	0.77	0.76	0.81								Nix2, Nix3, STAT	485bp 5' of RPS-1011O1.2		
1	192536886	1	1	rs2815315	C	T	0.83	0.78	0.76	0.81									538bp 5' of RPS-1011O1.2		
1	192537400	1	1	rs1323297	T	C	0.83	0.78	0.76	0.81				H7-HESC				ERalpha-a	1.1kb 5' of RPS-1011O1.2		
1	192537508	0.97	1	rs1323296	A	G	0.83	0.78	0.76	0.81								7 altered motifs	1.2kb 5' of RPS-1011O1.2		
1	192541021	0.98	1	rs1323292	G	A	0.98	0.79	0.76	0.81								GATA, Hoxa7	3.8kb 5' of RGS1		
1	192541172	0.98	1	rs2884912	A	T	0.83	0.78	0.76	0.81									3.7kb 5' of RGS1		
1	192541472	0.98	1	rs1358062	C	G	0.58	0.76	0.76	0.81									6 altered motifs	3.4kb 5' of RGS1	
1	192543837	1	1	rs1547624	A	T	0.67	0.77	0.77	0.81			GM12878	Wt-38				4 altered motifs	1kb 5' of RGS1		
1	192544795	0.99	1	rs2884920	A	G	0.97	0.79	0.77	0.81			GM12878, NHLF	21 cell types	16 bound proteins				lrf, PU.1	81bp 5' of RGS1	
1	192545099	0.95	0.99	rs753818	G	A	0.33	0.69	0.59	0.81			GM12878, NHLF	Th2	5 bound proteins				6 altered motifs	RGS1	intronic

rs2984920 is a strong candidate as it overlaps regulatory marks in the most cell lines. It also disrupts a PU.1 motif (Log odds drop from 14.5 to 2.9) and overlaps a PU.1 binding site. It is also in the promoter of RGS1.

Regulatory motifs altered				
PWM	Strand	Ref	Alt	Match on:
				Ref: CAAGATACACGTCACAGCACACCAAGAAAAGGGGAACTTCCAGTGTCTGTGGTAAACATC Alt: CAAGATACACGTCACAGCACACCAAGAAAAGGGGAACTTCCAGTGTCTGTGGTAAACATC
lrf_known2	+	-0.8	-12.7	GGAAAAGYGAAS
PU.1_disc1	-	14.5	2.9	AWGRGGAAGT
PU.1_known3	+	13.7	12.2	NDWDRVGAASDTN

rs7535818 primarily overlaps POL2 binding sites which suggests it would not affect regulation but is instead in an actively transcribed region. It is also in the promoter/first intron of RGS1.

Proteins bound by ChIP (ENCODE)	
Cell ID	Protein
GM12878	POL2
GM12878	POL24H8
GM12878	SMC3
GM12878	USF2
GM12878	WHIP
GM12891	POL2
GM12891	POL2
GM12891	POL24H8
GM12892	POL2
GM12892	POL2
GM12892	POL24H8
GM12892	POL24H8
GM12892	POL24H8
GM18951	POL2
GM19099	POL2
PBDE	POL2

[rs2816305](#) also overlaps regulatory regions and TFs. It overlaps some motifs but not those corresponding to TFs with overlapping binding sites. However, it is important to remember that not all TFs are surveyed by ENCODE.

### Proteins bound by ChIP (ENCODE)

Cell ID	Protein
GM12878	BCL11A
GM12878	OCT2
GM12878	PBX3
GM12878	POL24H8
GM12878	POU2F2
GM12878	TBP

### Regulatory motifs altered


PWM	Strand	Ref	Alt	Match on:
				Ref: GAGCTGCTCTTTGCTCTGACAGTGACGCAGTGGTTGACGCTAGACTAGACGGTTGACCA Alt: GAGCTGCTCTTTGCTCTGACAGTGACGCATTGGTTGACGCTAGACTAGACGGTTGACCA
LBP-1_3	+	13.2	1.4	RCTGGKTHKVCYDS
NF-Y_known2	-	-5.7	6.3	CTCATTGGCTG
Pax-8_1	+	6.2	10.4	DSHBHNWHEGMRDGRDNDHV

- Using RegulomeDB, determine the scores for each of the SNPs in LD with rs2816316 that you think may be casual. Is there a SNP that is likely to affect transcription factor binding? Which SNP(s) would you further investigate?

[rs2816305](#) = 1d

[rs2984920](#) = 2a

[rs7535818](#) = 3a


Download About Help

The search has evaluated 3 input line(s) and found 3 SNP(s).

### Summary of SNP analysis

Show 10 entries

Coordinate (0-based)	dbSNP ID	Regulome DB Score	Other Resources
chr1:192524268	rs2816305	1d	<a href="#">UCSC</a>   <a href="#">ENSEMBL</a>   <a href="#">dbSNP</a>
chr1:192544794	rs2984920	2a	<a href="#">UCSC</a>   <a href="#">ENSEMBL</a>   <a href="#">dbSNP</a>
chr1:192545098	rs7535818	3a	<a href="#">UCSC</a>   <a href="#">ENSEMBL</a>   <a href="#">dbSNP</a>

Showing 1 to 3 of 3 entries

Download
BED
GFF
Full Output

It would be worthwhile to further investigate both rs2984920 and rs2816305. rs2816305 has the lowest RegulomeDB score since it was reported to be a eQTL for RGS1. It does not overlap a motif corresponding to a bound TF but is in a regulatory region. rs2984920 lies in the promoter of RGS1 and overlaps motifs for several bound TFs including PU.1 and NFkB (discovered by RegulomeDB). rs2984920 and rs2816305 are also in LD, so the eQTL signal from rs2816305 could be due to rs2984920. Both SNPs would be worth investigating further to determine the casual variant.

## Exercise #2

You are interested in studying genetics variants associated with Amyotrophic lateral sclerosis (ALS), which causes muscle atrophy due to degeneration of motor neurons. Eleven studies have reported 66 SNPs associated with ALS. Since little is known about the disease, you decided to investigate these genetic variants.

- Using HaploReg, determine if there are enrichments for enhancers in any ENCODE cell types for these ALS SNPs. Are there enrichments in DNase regions?

HepG2 – Strong Enhancers, HMEC – Strong Enhancers, GM12878 – All Enhancers & Strong Enhancers

DNase: HFF-Myc, HA-sp, Th2, and GM18507

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 (comma-delimited list of rsIDs OR a single region as chrN:start-end): <input type="text"/>										
or, upload a text file (one refSNP ID per line): <input type="button" value="Choose File"/> No file chosen										
or, select a GWAS: <input type="text" value="Amyotrophic lateral sclerosis (11 studies combined), 66 SNPs"/>										
<input type="button" value="Submit"/>										
Enhancer enrichment analysis										
Cell type	ID	Description	All enhancers				Strongest enhancers			
			Obs	Exp	Fold	p	Obs	Exp	Fold	p
	HepG2	hepatocellular carcinoma	3	2.4	1.2	0.441866	3	0.8	3.7	0.047749
	HMEC	mammary epithelial cells	6	3.9	1.5	0.199915	5	1.5	3.3	0.017548
	GM12878	B-lymphocyte, lymphoblastoid	8	3.1	2.6	0.011706	4	1.1	3.6	0.026063
DNase enrichment analysis										
Cell type	ID	Description	Treatment	Production center	DNase					
					Obs	Exp	Fold	p		
	HFF-Myc	foreskin fibroblast cells expressing canine cMyc	None	UW	3	0.8	3.8	0.044225		
	HA-sp	astrocytes spinal cord	None	UW	4	0.7	5.8	0.005086		
	Th2	primary Th2 T cells	None	UW	2	0.3	6.2	0.041565		
	GM18507	B-lymphocyte, lymphoblastoid	None	Duke	3	0.4	7	0.009259		

- Perform the same analysis using Roadmap epigenomes. Are these disease relevant tissue and cell types enriched?

Colon: All Enhancers, Penis Foreskin: Strong Enhancers, Brain Substantia Nigra: All Enhancers, Brain Inferior Temporal lobe: All Enhancers, Brain Cingulate Gyrus: All Enhancers, Skeletal Muscle: Strong Enhancers

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

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

or, select a GWAS:

**Enhancer enrichment analysis**

Cell type	Description	All enhancers				Strongest enhancers			
		Obs	Exp	Fold	p	Obs	Exp	Fold	p
COL.SMUS	Colon Smooth Muscle	6	2.8	2.3	0.04766	1	0.8	1.7	0.444256
PFM.3	Penis Foreskin Melanocyte Primary Cells Donor skin03	6	3.7	1.6	0.169735	5	1.4	3.6	0.013261
BN.SN	Brain Substantia Nigra	7	3.2	2.2	0.039285	3	1.8	1.7	0.271144
BN.ITL	Brain Inferior Temporal Lobe	9	3.5	2.6	0.008009	5	2.2	2.3	0.066668
BN.CC	Brain Cingulate Gyrus	7	3.4	2.1	0.049926	4	2.1	1.9	0.158016
SK.MUS	Skeletal Muscle	4	3	1.3	0.363753	4	1.2	3.2	0.03647

### Exercise #3

rs6774494 has been associated with Nasopharyngeal carcinoma in the Chinese population (Bei, ..., Zeng (2010) *Nature Genetics*)

1. What score does RegulomeDB assign to rs6774494? Is this SNP likely to affect transcription factor binding?

There is no data for rs6774494 in RegulomeDB so no score is assigned. This SNP is unlikely to affect TF binding.

RegulomeDB Download About Help

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

### Summary of SNP analysis

Show  entries

Coordinate (0-based)	dbSNP ID	Regulome DB Score	Other Resources
chr3:169082632	rs6774494	No Data	<a href="#">UCSC</a>   <a href="#">ENSEMBL</a>   <a href="#">dbSNP</a>

Showing 1 to 1 of 1 entries

2. Using HaploReg, determine if there are there any SNPs in high LD with rs2816316. Are any of these SNPs more likely to be causal?

There are ten SNPs in LD with rs2816316. rs9869781 and rs13322424 overlap enhancer marks, promoter marks, DNase peaks and TF binding sites and would be worth investigating further.

Query SNP: rs6774494 and variants with  $r^2 \geq 0.8$

chr	pos (hg19)	LD (r <sup>2</sup> )	LD (D')	variant	Ref	Alt	AFR	AMR	ASN	EUR	SIphy	Promoter	Enhancer	DNase	Proteins	eQTL	Motifs	GENCODE	dbSNP
							freq	freq	freq	freq	cons	histone marks	histone marks		bound	tissues <td>changed</td> <td>genes</td> <td>func annot</td>	changed	genes	func annot
3	169042446	0.92	0.98	rs9869781	T	C	0.43	0.40	0.56	0.24		11 cell types	13 cell types	MCF-7, Ishikawa, SAEC	ERALPHA_A		MAZ, Rad21, SMC3	MECOM	intronic
3	169057657	0.98	0.99	rs2106123	G	A	0.63	0.43	0.58	0.27							4 altered motifs	MECOM	intronic
3	169058009	0.98	0.99	rs2106124	G	A	0.63	0.45	0.58	0.27							GR	MECOM	intronic
3	169063121	0.95	1	rs13322424	C	T	0.63	0.43	0.59	0.27								MECOM	intronic
3	169063223	0.95	1	rs9863103	T	C	0.64	0.43	0.59	0.27								MECOM	intronic
3	169064503	0.92	0.97	rs766283	T	A	0.62	0.44	0.58	0.27								MECOM	intronic
3	169066232	0.95	1	rs13322424	A	G	0.64	0.43	0.59	0.27		OD4, MBP1308, OD4, MBP1362	13 cell types				TCF12	MECOM	intronic
3	169075999	0.95	1	rs1530438	A	G	0.64	0.43	0.59	0.27		13 cell types					BAF155, SP1, TFIIA	MECOM	intronic
3	169075078	0.98	0.99	rs1522272	G	A	0.64	0.47	0.57	0.32							4 altered motifs	MECOM	intronic
3	169082633	1	1	rs6774494	G	A	0.65	0.47	0.58	0.32							ERalpha-a, Nr22, TLX1, NFIC	MECOM	intronic
3	169085277	0.94	1	rs2188132	C	A	0.65	0.47	0.59	0.32								MECOM	intronic

3. How would your results change if you used the default settings for HaploReg (i.e. European LD?)

If you use European LD, there are only two SNPs in LD with rs6774494 and neither are as likely to affect TF binding as rs9869781 and rs1332242.

Query SNP: rs6774494 and variants with $r^2 \geq 0.8$																			
chr	pos (hg19)	LD (r <sup>2</sup> )	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
3	169075078	1	1	rs1522278	G	A	0.64	0.47	0.57	0.32			HD,CD184EC, CD34,MBP1562				4 altered motifs	MECOM	intronic
3	169082833	1	1	rs6774494	G	A	0.65	0.47	0.58	0.32			GAS				E2alpha, Nr2f2, TLX1, NFIC	MECOM	intronic
3	169085277	1	1	rs2188132	C	A	0.65	0.47	0.59	0.32			HD,CD184EC, HUES48					MECOM	intronic