Identifying promoters and regulatory elements for DNA variation

Sarah Pendergrass, PhD MS
Center for Systems Genomics

Outline

- Why do we want to identify potential regulatory elements?
- What are regulatory elements?
- Tools/resources for annotation of your data

Rationale

- Why do we want to identify potential regulatory elements?
  - Much of the focus of Genome Wide Association Analyses (GWAS) analyses have been on protein coding regions
    - The idea “identify the genes SNPs are in or near”
    - Surely significant associations are due to modification of proteins affecting phenotypes?

Rationale

- NHGRI GWAS catalog keeps a record of highly statistically significant GWAS associations
    - Out of 8455 GWAS associations reporting SNPs within genes
        - 438 results were not within genes
    - A total of 6606 GWAS catalog records that reported an upstream gene, 6608 records reporting a downstream gene
    - A large proportion of SNPs reported to be upstream or downstream of specific genes are not actually in linkage disequilbrium (correlated) with SNPs within these reported genes when using HapMap

Rationale

- So via GWAS we keep finding SNPs that are not within genes, or not correlated with SNPs within genes
  - Many of the GWAS SNPs are not non-synonymous, or are intronic when they are in a gene
  - Time to look at other potential functionality of genetic variation
- Areas of the genome once considered "deserts" are being characterized at a fast rate

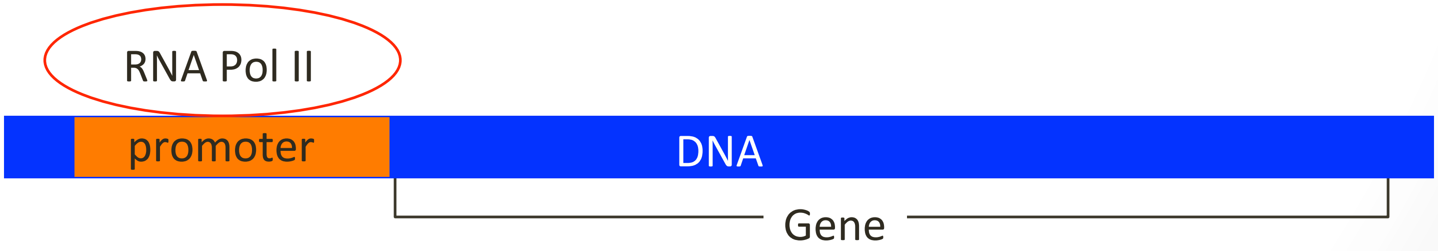
Rationale

- For example
  - You have performed a GWAS
  - There are 10 SNPs of interest passing your p-value cutoff
  - Looks like 3 of the SNPs are within protein coding regions so you looked up those genes and identified possibly interesting information
  - What about the other 7 SNPs?
    - Is there evidence they DO something?
- Or perhaps you have some low frequency variants you want to explore…

Rationale

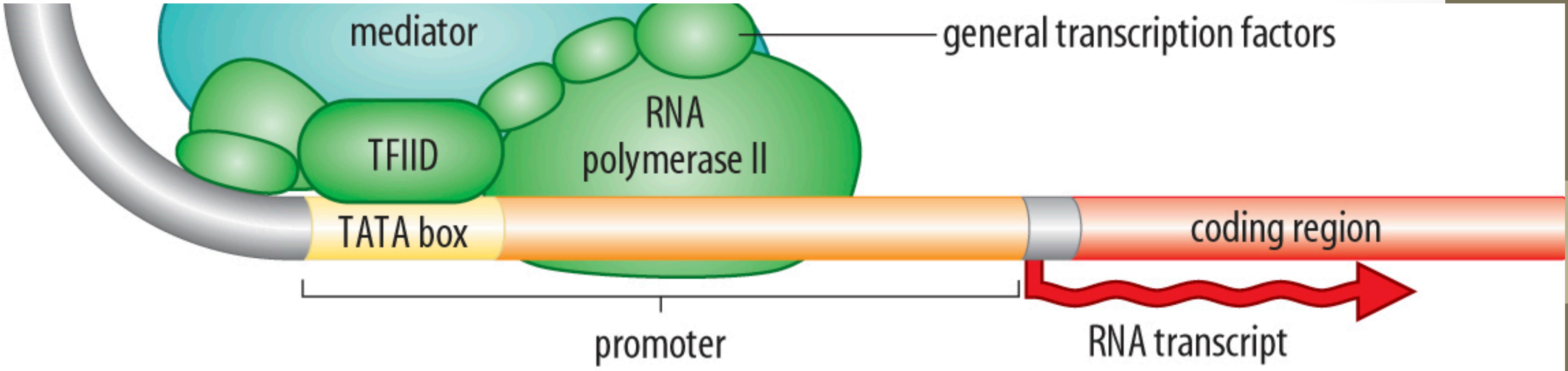
- What about outside protein coding regions?

- In gene transcription, RNA polymerase binds upstream of a gene to a promoter initiate transcription
  - But the process of gene expression is very spatially and temporally regulated
    - Changes from cell type to cell type
    - Many proteins involved
- Considering more of the regulation of transcription when evaluating genetic variants for functionality
  - Identification of other biology associated with phenotypic traits and outcomes



Regulatory Regions

- Promoter region (promoters)
  - Region of DNA before coding region
  - RNA polymerase II binds there
  - A series of general transcription factors also bind
    - Including Transcription Factor II D
    - Making up the RNA polymerase II pre-initiating complex
- But other transcription factors bind other places!



http://www.mun.ca/biology/desmid/brian/BIOL3530/DEVO_10/devo_10.html

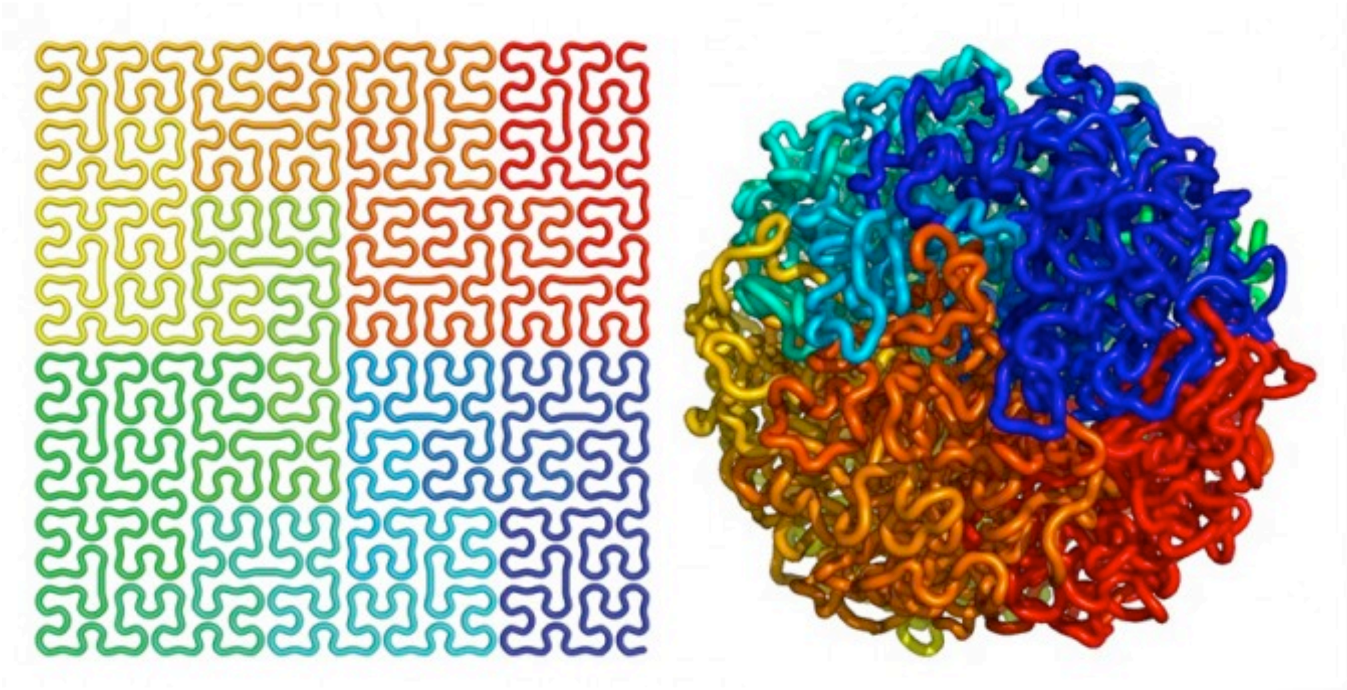
Regulatory Elements

- For transcription the following are required
  - Transcriptional region
    - Where transcription of gene takes place
  - Promoter region
    - Start of transcription
  - Regulatory regions
    - That enable or inhibit transcription
  - Proteins that bind to these promoters and regulatory regions
    - Transcription Factors (TFs)
  - Access to the transcriptional AND regulatory region(s)

  - *Genetic variation can affect all of the above, causing changes in proteins and/or the ability of proteins to bind to regions*

Rationale

- DNA in the nucleus is three dimensional
  - Densely packed
  - Some regions closer to others
  - No knots



http://www.wired.com/2009/10/fractal-genome/ Erez Lieberman Aiden

Transcription Factors

- Transcription factors (TFs)
  - Regulatory proteins

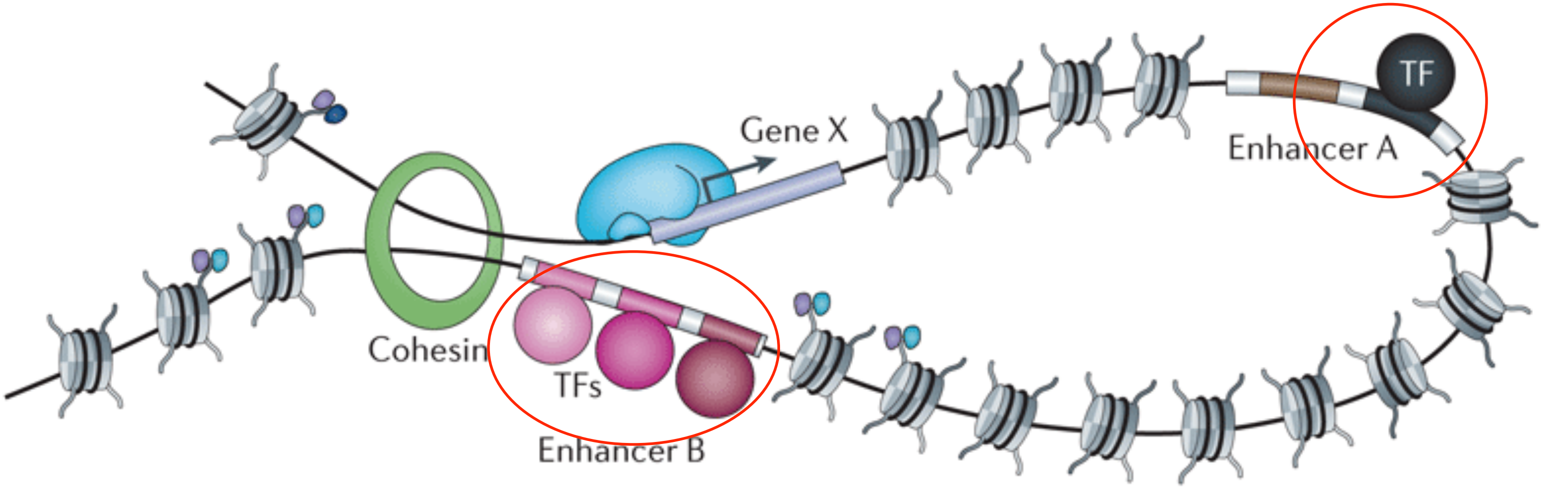
- Activate and sometimes inhibit transcription of DNA by binding to DNA sequences
  - There can be repressive TFs

- TFs bind to highly conserved sequences
  - These sequences have been used to categorize TFs in to "families"
  - TFs can also be classified by their 3-D structure

- Requires coordinated interactions of multiple proteins to regulate gene expression

Enhancers

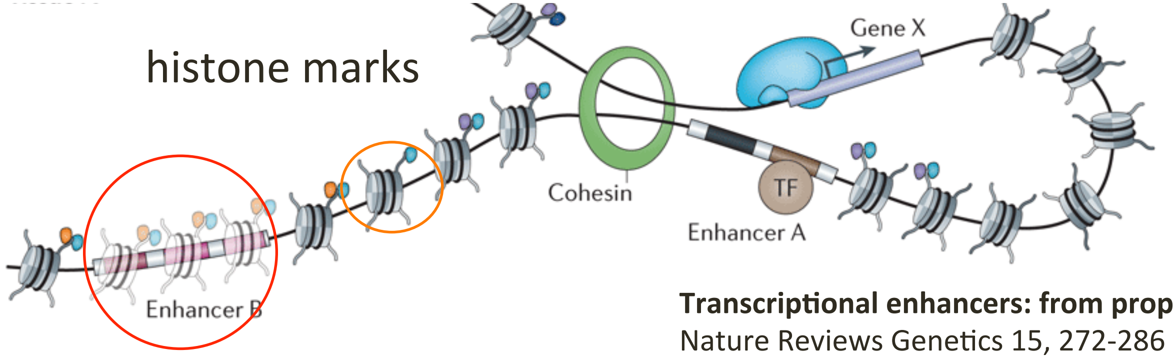
- Enhancers are short regions of DNA (< 10 bp)
- Bind TFs
- May be several to MANY kb distant from the gene
- DNA can be coiled so that enhancers interact to form a large protein complex
- Potentially increase concentration of activators near promoter



Transcriptional enhancers: from properties to genome-wide predictions
Nature Reviews Genetics 15, 272-286 doi:10.1038/nrg3682

Access to the Region

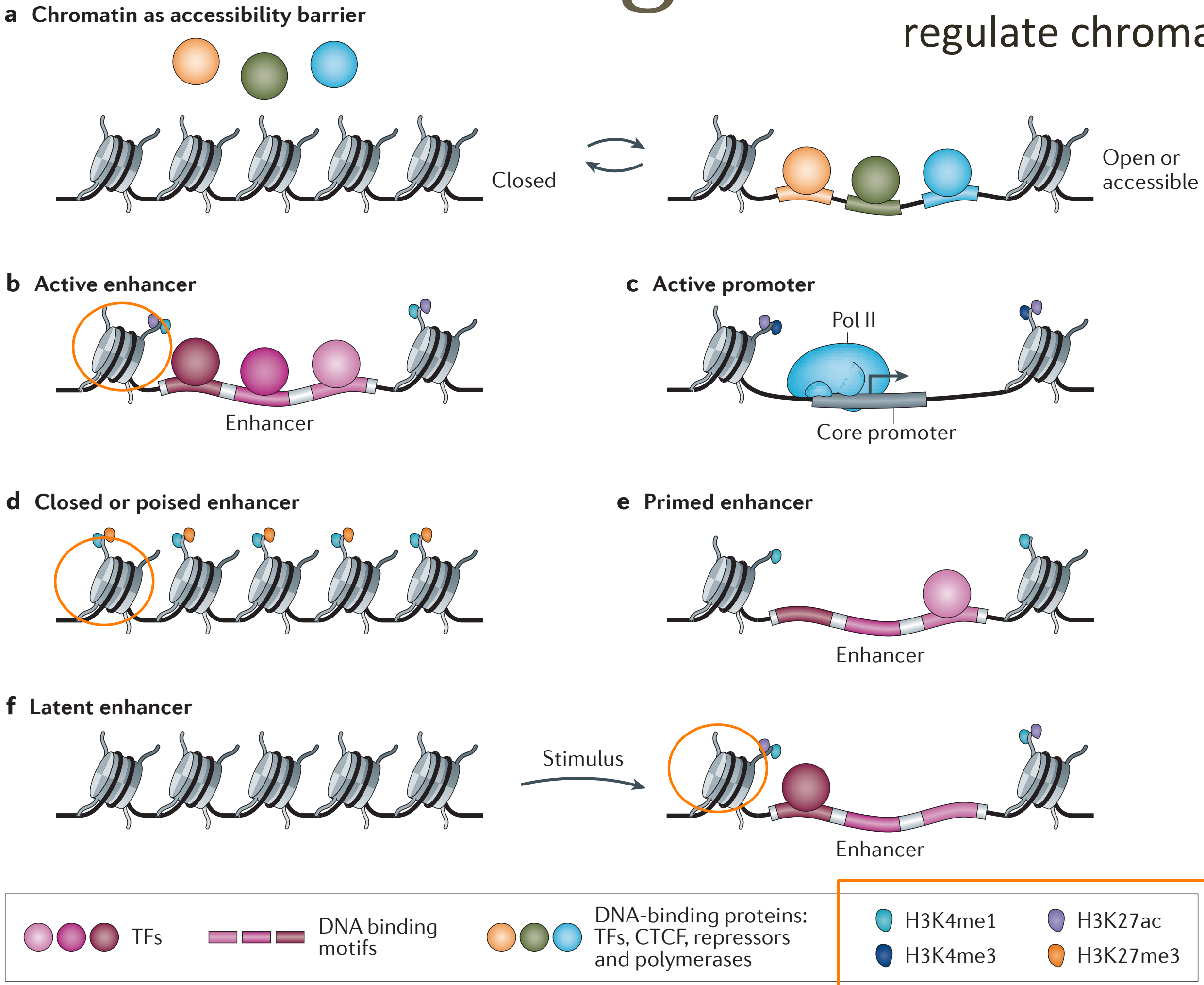
- Transcription factors can be present but no transcription
  - TFs must reach their target sequences
- DNA and histone proteins
  - Chromatin state – DNA wound around histones (nucleosome)
    - Can limit access of transcription factors and RNA polymerase to DNA promoters
  - Active promoters and enhancers are characterized by depletion of nucleosomes
  - Inactive promoters and enhancers might be silenced by histone marks or repressive TF binding


histone marks

Transcriptional enhancers: from properties to genome-wide predictions
Nature Reviews Genetics 15, 272-286 doi:10.1038/nrg3682

Access to the Region

Modifications on histones or on DNA recruit proteins that regulate chromatin function

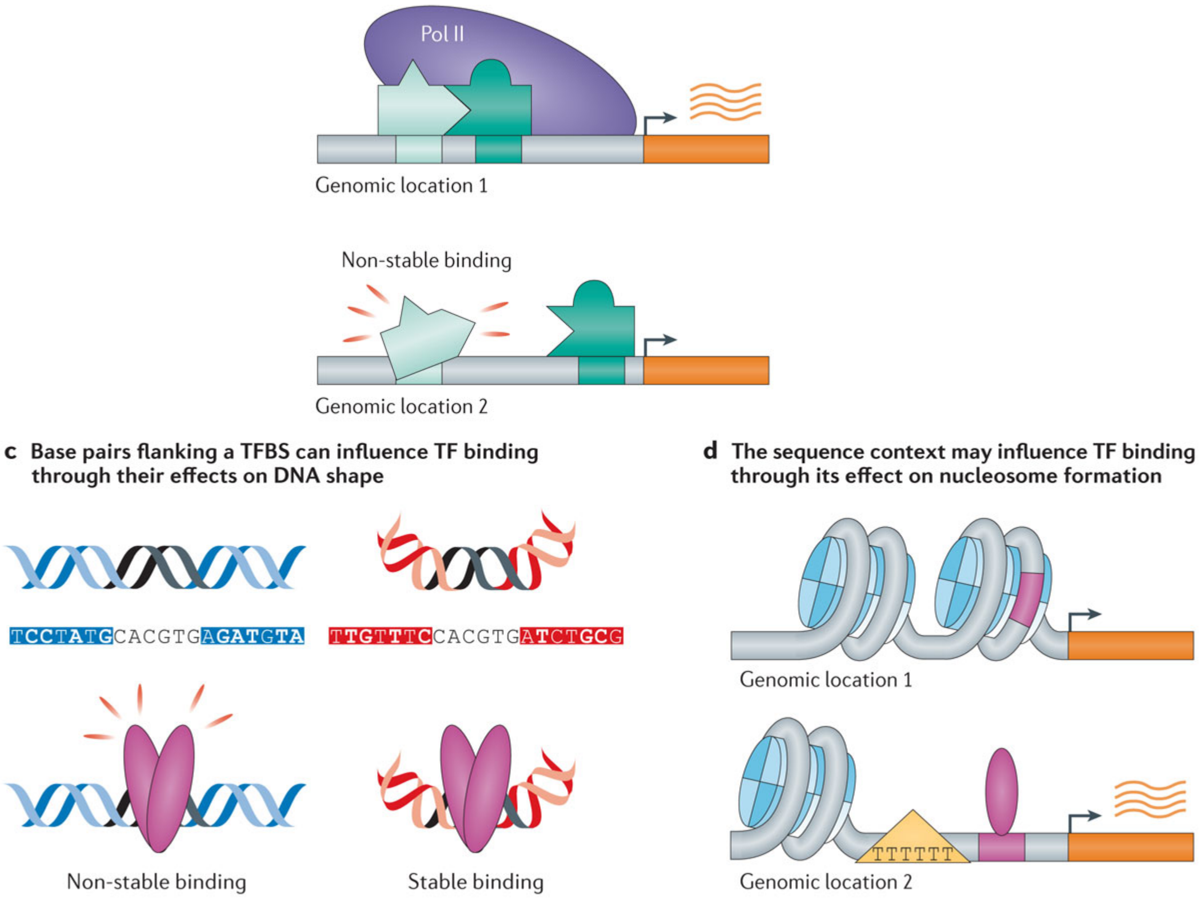


Histone Modifications

Transcriptional enhancers: from properties to genome-wide predictions
Nature Reviews Genetics 15, 272-286 doi:10.1038/nrg3682

SNPs Affecting Transcription

- Genetic variation can cause affects on transcription multiple ways



c Base pairs flanking a TFBS can influence TF binding through their effects on DNA shape

d The sequence context may influence TF binding through its effect on nucleosome formation

Nature Reviews | Genetics

In pursuit of design principles of regulatory sequences
Nature reviews. Genetics, 07/2014, Volume 15, Issue 7

Annotation

- So there is a vast region to explore – the affect of genetic variation on transcription

- What are useful sources for identifying regulatory elements?

- The GWAS example
  - You have performed a GWAS
  - There are 10 SNPs of interest passing your p-value cutoff
  - Looks like 3 of the SNPs are within protein coding regions so you were able to look up those genes and identify interesting features
    - But these SNPs don't seem to cause a change in the protein?
  - What about the other7 SNPs?
    - Is there evidence they DO something?

- Is there evidence this SNP changes gene expression?

SCAN Database

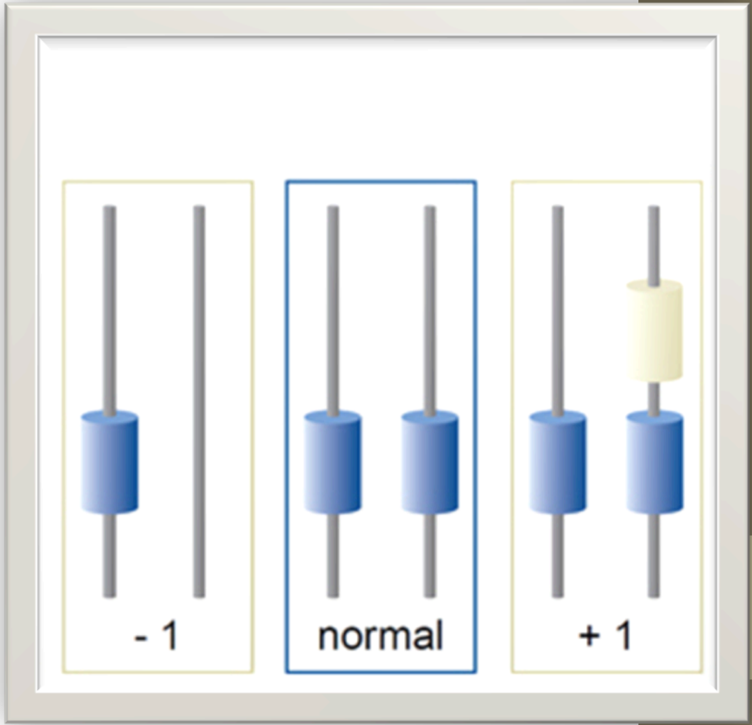
- eQTL experiment information that can be used to annotate SNPs

- What is an eQTL?
  - Expression quantitative trait loci
- Gene expression = relative mRNA abundance
  - Can be measured and used like a phenotypic trait
- The association between SNP variation and gene expression variation can be calculated
  - Also can evaluate Copy Number Variants (CNVs)

- Some genetic variation will have very statistically significant associations with changes in gene expression

SCAN Database

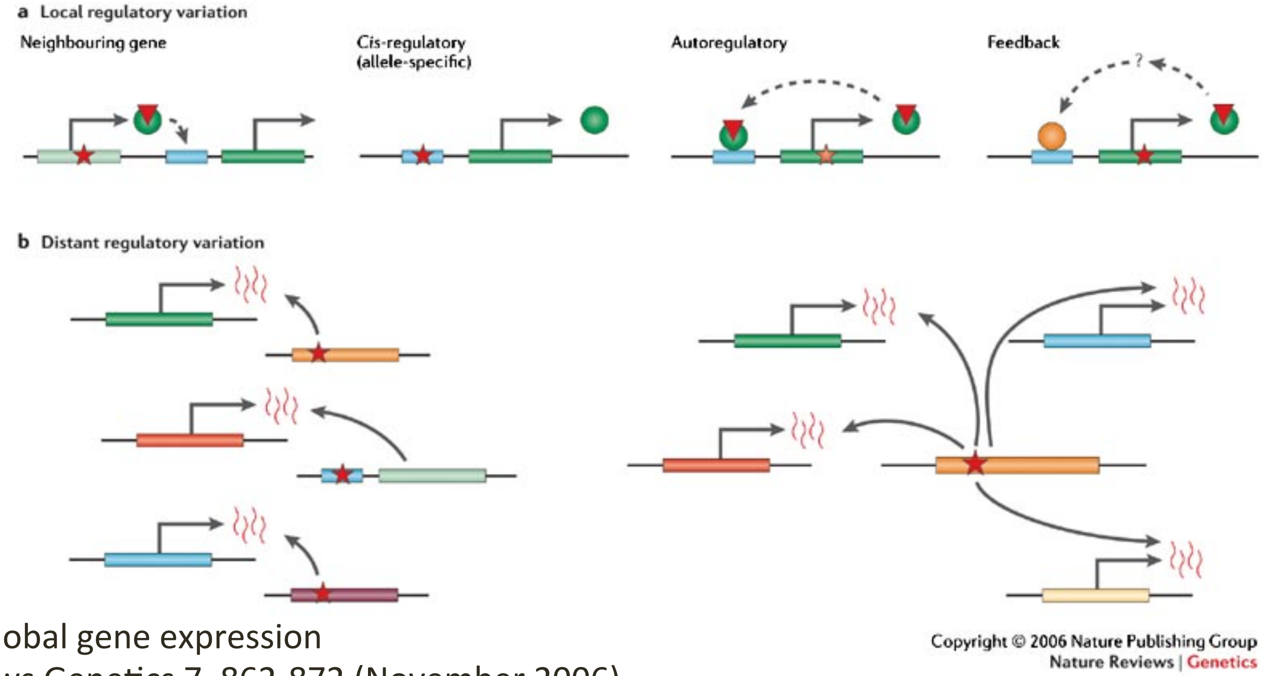
- Copy Number Variants (CNVs)
  - Sequences that differ in the total number of copies among individuals
  - Can be duplications or deletions
  - Can range in size from 10Kb to 1Mb



Copy Number Variation and Human Disease
Nature Education 1(3):1

SCAN Database

- Cis and Trans gene expression
  - Cis – the SNP changes the gene expression of the gene the SNP is in
    - SNPs located near or within a gene
  - Trans – the SNP changes the gene expression of a different gene
    - Any other SNPs


a Local regulatory variation
Neighbouring gene
Cis-regulatory (allele-specific)
Autoregulatory
Feedback
?
b Distant regulatory variation

Copyright © 2006 Nature Publishing Group
Nature Reviews | Genetics

Genetics of global gene expression
Nature Reviews Genetics 7, 862-872 (November 2006)

SCAN Database

- eQTLs potentially more effective than associations with complex traits
  - Gene expression is a complex trait
  - An intermediate phenotype between genetic loci and higher level cellular/clinical phenotypes
    - Disease risk
    - Drug response

Trait-associated SNPs are more likely to be eQTLs:
annotation to enhance discovery from GWAS
PLoS Genet. 2010 Apr 1;6(4)

SCAN Database

- Genetic variation contributes a great deal to natural variation in gene expression

- SNPs associated with complex human traits included in the NHGRI GWAS catalog are significantly enriched for eQTLs identified in lymphoblastoid cell lines (LCLs)

- SNPs are enriched for "master regulators"
  - eQTLs that predict transcript levels of 10 more genes

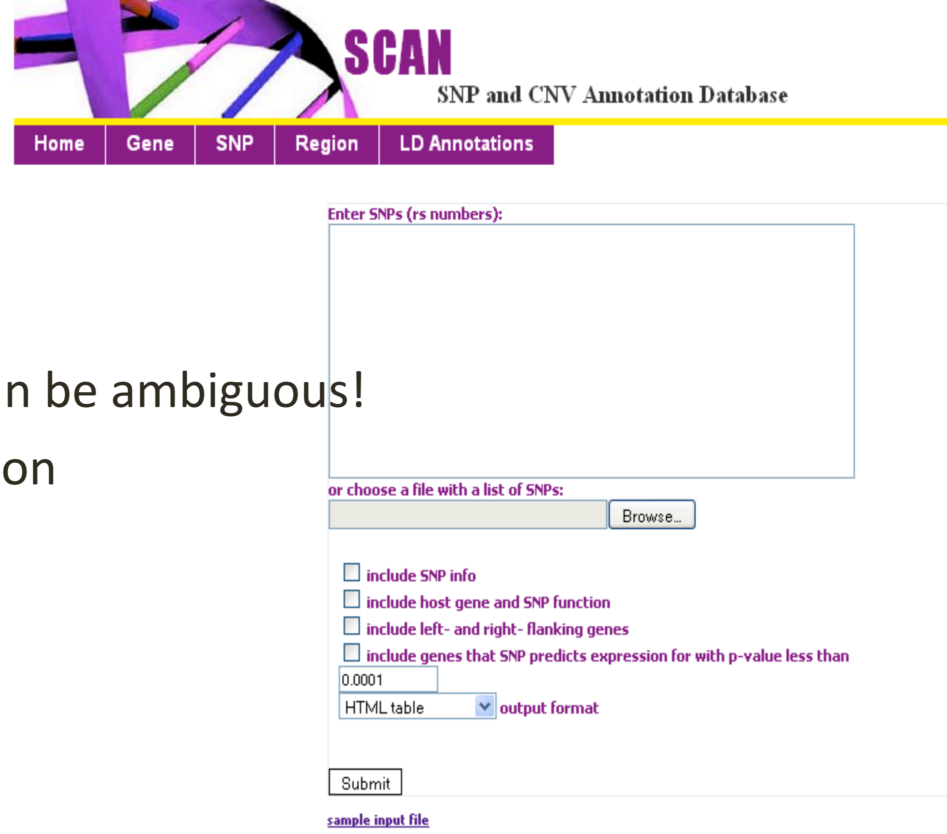
SCAN Database

- eQTL experiment information
- Used HapMap project data (http://www.hapmap.org/)
  - > 3.1 million SNPs
  - 27 lymphoblastoid cell lines (LCLs) from African, Asian and European ancestry

SCAN Database

- eQTL experiments behind SCAN
  - Used QTDT (Quantitative Transmission Disequilbrium Test)
    - Relatedness of the individuals in each ancestry
    - Trios (parents and child)
  - 13,000 transcripts with consistent expression signal in at least 80% of the samples
  - 2 million common SNPs with minor allele frequency > 5%
  - Stratified by ancestry
  - Little mention of how they calculated the CNV eQTL…

SCAN



Enter SNPs (rs numbers):

or choose a file with a list of SNPs:

□ include SNP info
□ include host gene and SNP function
□ include left- and right- flanking genes
□ include genes that SNP predicts expression for with p-value less than

0.0001

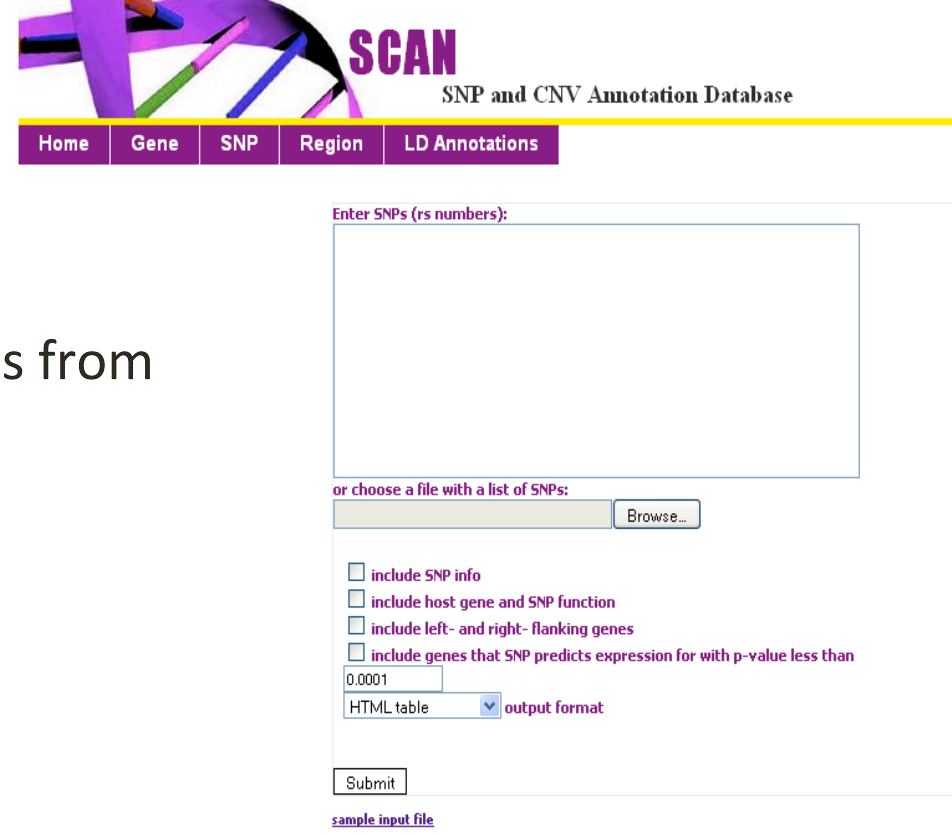
HTML table output format

sample input file

- Can query SCAN with rsID
  - Remember that rsID's for SNPs can be ambiguous!
  - Chromosome and base pair location
- Can include in output
  - Host gene
    - Genomic coordinates
  - SNP function
    - dbSNP's classification scheme
      - SNP represents coding change
  - Left and right flanking genes
- Can include P-value cutoff for eQTL of interest
- Output format of choice
- Note: can also explore Genes, SNPs, Regions, and LD Annotation

SCAN

- Ok, what if I give it my list of SNPs from my GWAS?



Enter SNPs (rs numbers):

or choose a file with a list of SNPs:

□ include SNP info

□ include host gene and SNP function

□ include left- and right- flanking genes

□ include genes that SNP predicts expression for with p-value less than 0.0001

HTML table output format

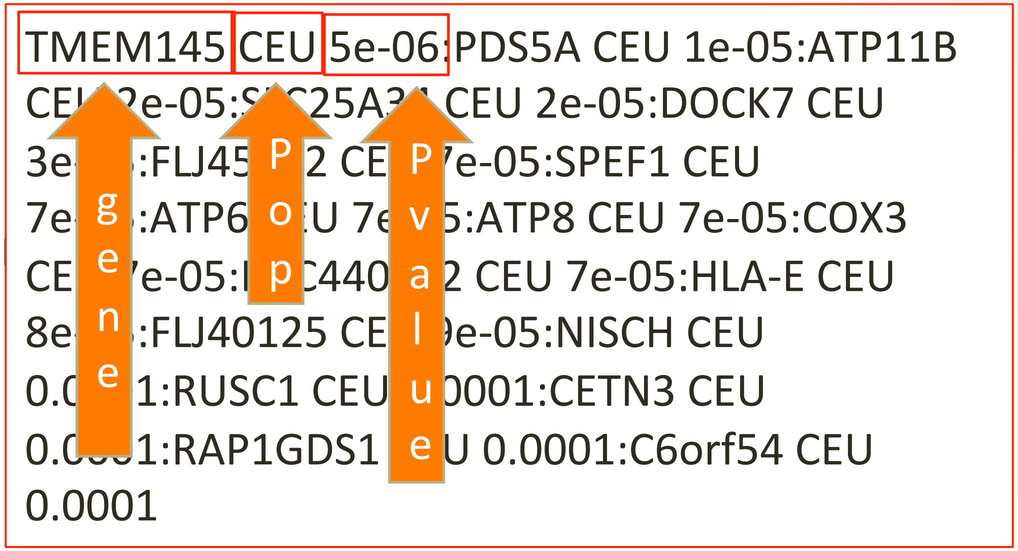
sample input file

SCAN

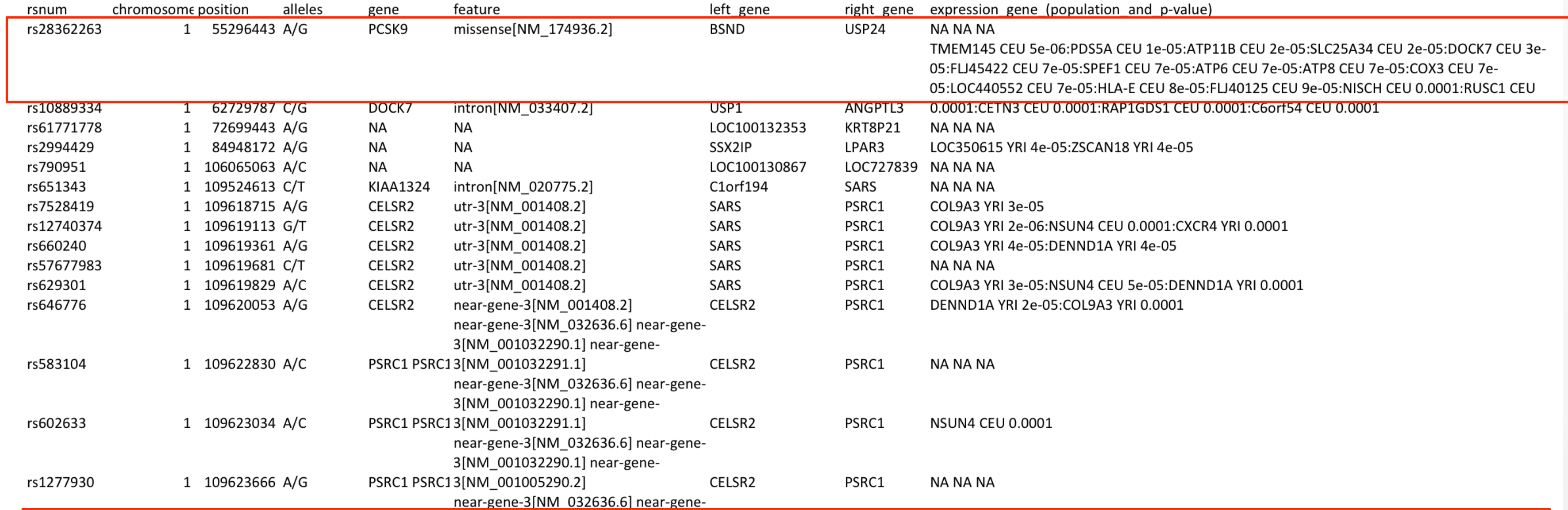
- HTML output an option
- Text also possible
- Ton of information per SNP
  - SNP
  - Gene
  - Function
  - Minor allele frequency

SCAN

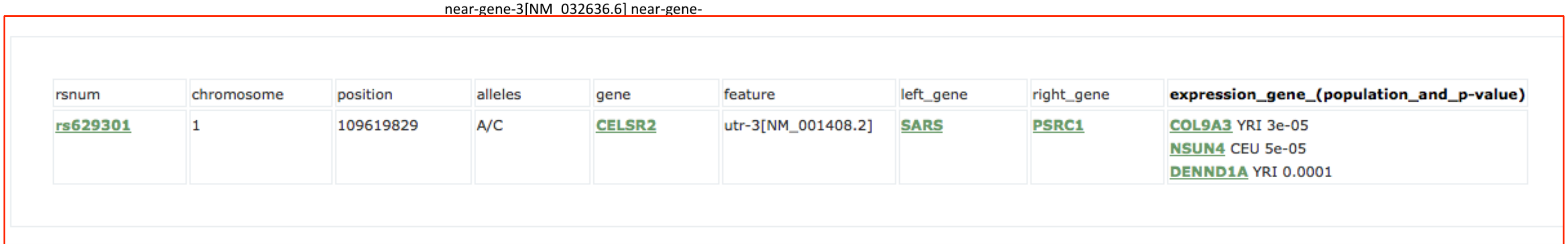
- HTML output an optio
- Text also possible
- Information per SNP



TMEM145 CEU 5e-06:PDS5A CEU 1e-05:ATP11B CEU 2e-05:SLC25A34 CEU 2e-05:DOCK7 CEU 3e-05:FLJ45422 CEU 7e-05:SPEF1 CEU 7e-05:ATP6 CEU 7e-05:ATP8 CEU 7e-05:COX3 CEU 7e-05:LOC440552 CEU 7e-05:HLA-E CEU 8e-05:FLJ40125 CEU 9e-05:NISCH CEU 0.0001:RUSC1 CEU 0.0001:CETN3 CEU 0.0001:RAP1GDS1 CEU 0.0001:C6orf54 CEU 0.0001



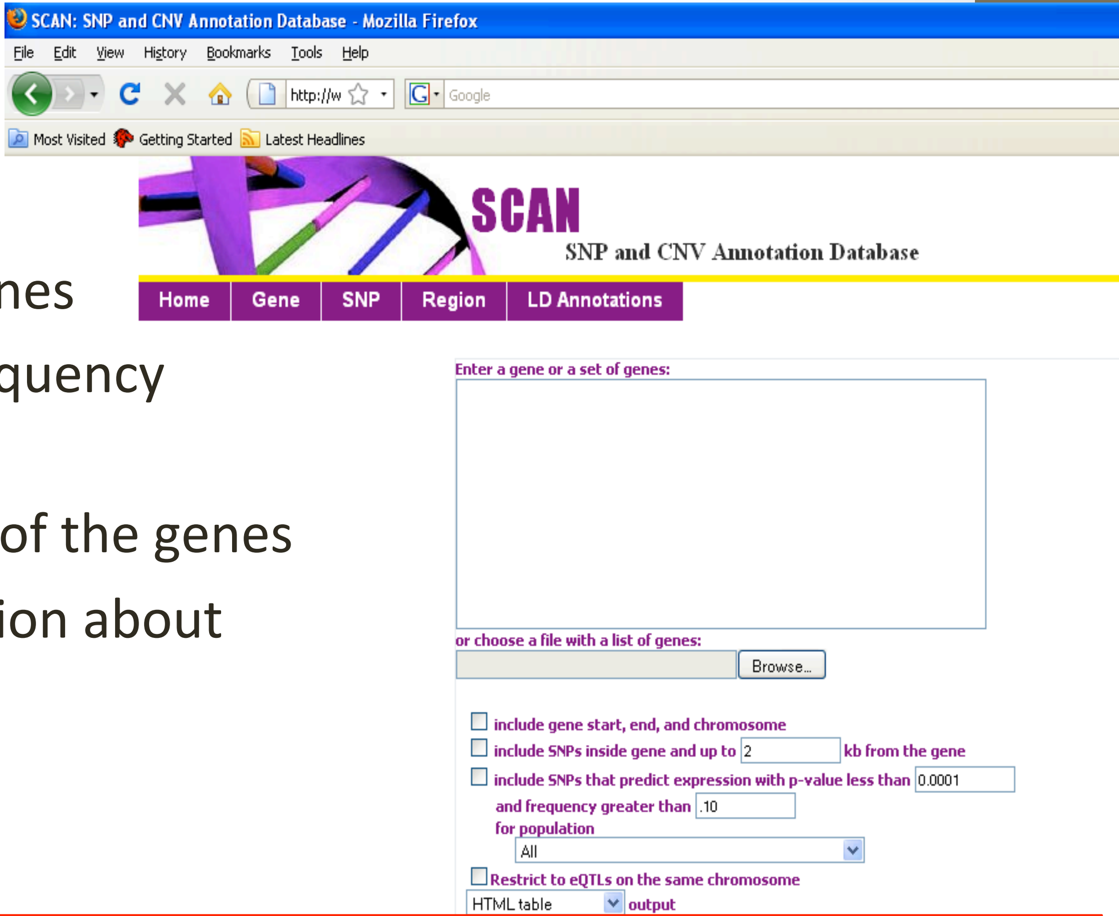
rsnum	chromosome	position	alleles	gene	feature	left_gene	right_gene	expression_gene_(population_and_p-value)
rs28362263	1	55296443	A/G	PCSK9	missense[NM_174936.2]	BSND	USP24	NA NA NA
rs10889334	1	62729787	C/G	DOCK7	intron[NM_033407.2]	USP1	ANGPTL3	TMEM145 CEU 5e-06:PDS5A CEU 1e-05:ATP11B CEU 2e-05:SLC25A34 CEU 2e-05:DOCK7 CEU 3e-05:FLJ45422 CEU 7e-05:SPEF1 CEU 7e-05:ATP6 CEU 7e-05:ATP8 CEU 7e-05:COX3 CEU 7e-05:LOC440552 CEU 7e-05:HLA-E CEU 8e-05:FLJ40125 CEU 9e-05:NISCH CEU 0.0001:RUSC1 CEU 0.0001:CETN3 CEU 0.0001:RAP1GDS1 CEU 0.0001:C6orf54 CEU 0.0001
rs61771778	1	72699443	A/G	NA	NA	LOC100132353	KRT8P21	NA NA NA
rs2994429	1	84948172	A/G	NA	NA	SSX2IP	LPAR3	LOC350615 YRI 4e-05:ZSCAN18 YRI 4e-05
rs790951	1	106065063	A/C	NA	NA	LOC100130867	LOC727839	NA NA NA
rs651343	1	109524613	C/T	KIAA1324	intron[NM_020775.2]	C1orf194	SARS	NA NA NA
rs7528419	1	109618715	A/G	CELSR2	utr-3[NM_001408.2]	SARS	PSRC1	COL9A3 YRI 3e-05
rs12740374	1	109619113	G/T	CELSR2	utr-3[NM_001408.2]	SARS	PSRC1	COL9A3 YRI 2e-06:NSUN4 CEU 0.0001:CXCR4 YRI 0.0001
rs660240	1	109619361	A/G	CELSR2	utr-3[NM_001408.2]	SARS	PSRC1	COL9A3 YRI 4e-05:DENND1A YRI 4e-05
rs57677983	1	109619681	C/T	CELSR2	utr-3[NM_001408.2]	SARS	PSRC1	NA NA NA
rs629301	1	109619829	A/C	CELSR2	utr-3[NM_001408.2]	SARS	PSRC1	COL9A3 YRI 3e-05:NSUN4 CEU 5e-05:DENND1A YRI 0.0001
rs646776	1	109620053	A/G	CELSR2	near-gene-3[NM_001408.2]	CELSR2	PSRC1	DENND1A YRI 2e-05:COL9A3 YRI 0.0001
rs583104	1	109622830	A/C	PSRC1 PSRC1	near-gene-3[NM_032636.6] near-gene-3[NM_001032290.1] near-gene-3[NM_001032291.1]	CELSR2	PSRC1	NA NA NA
rs602633	1	109623034	A/C	PSRC1 PSRC1	near-gene-3[NM_032636.6] near-gene-3[NM_001032290.1] near-gene-3[NM_001032291.1]	CELSR2	PSRC1	NSUN4 CEU 0.0001
rs1277930	1	109623666	A/G	PSRC1 PSRC1	near-gene-3[NM_032636.6] near-gene-3[NM_001032290.1] near-gene-3[NM_001005290.2]	CELSR2	PSRC1	NA NA NA
					near-gene-3[NM_032636.6] near-gene-			



rsnum	chromosome	position	alleles	gene	feature	left_gene	right_gene	**expression_gene_(population_and_p-value)**
rs629301	1	109619829	A/C	CELSR2	utr-3[NM_001408.2]	SARS	PSRC1	COL9A3 YRI 3e-05 NSUN4 CEU 5e-05 DENND1A YRI 0.0001

SCAN

- Can also query a list of genes
- Can include SNP allele frequency
  - Can choose population
- Can include SNPs outside of the genes
- Will also receive information about expression CNVs



Enter a gene or a set of genes:

or choose a file with a list of genes:

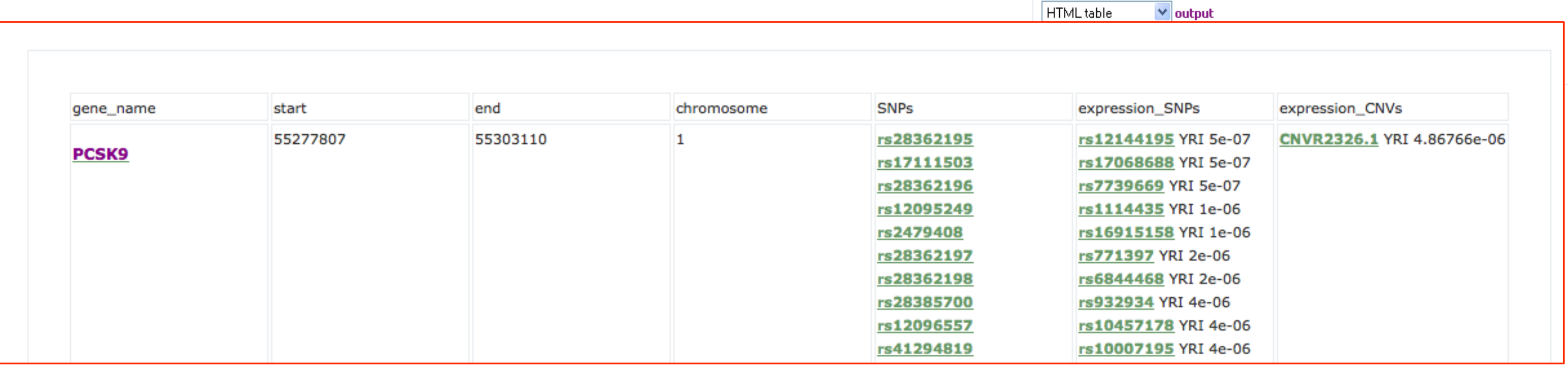
☐ include gene start, end, and chromosome

☐ include SNPs inside gene and up to 2 kb from the gene

☐ include SNPs that predict expression with p-value less than 0.0001 and frequency greater than .10 for population All

☐ Restrict to eQTLs on the same chromosome

HTML table output



gene_name	start	end	chromosome	SNPs	expression_SNPs	expression_CNVs
PCSK9	55277807	55303110	1	rs28362195 rs17111503 rs28362196 rs12095249 rs2479408 rs28362197 rs28362198 rs28385700 rs12096557 rs41294819	rs12144195 YRI 5e-07 rs17068688 YRI 5e-07 rs7739669 YRI 5e-07 rs1114435 YRI 1e-06 rs16915158 YRI 1e-06 rs771397 YRI 2e-06 rs6844468 YRI 2e-06 rs932934 YRI 4e-06 rs10457178 YRI 4e-06 rs10007195 YRI 4e-06	CNVR2326.1 YRI 4.86766e-06

Done

Exploring Results

- So back to the 10 SNPs you have from a GWAS
    - You know that 3 are within protein coding regions
    - Using SCAN, you identified that 4 of your SNPs seem to have some impact on gene expression
        - Some cis, some trans
        - Some of the genes that show marked changes in gene expression are interesting and related to your trait of interest (e.g. hypertension)
        - You have identified some interesting pathways these genes are in
    - What about other evidence that the SNPs of your study impact transcription?

https://ritchielab.psu.edu/software/biofilter-download

ENCODE

- Encyclopedia of DNA Elements
- Funded by the NHGRI
- The goal:
  - Build a comprehensive parts list of the functional elements of the human genome
- Nearly 99% of the ~3.3 billion nucleotides that constitute the human genome do not code for proteins
  - WHAT DO THEY DO???
- ENCODE and GENCODE are identifying and characterizing this “dark matter”

http://www.genome.gov/encode/

ENCODE

- ENCODE
  - Identifying genomic sequences
    - From which short and long RNAs, both nuclear and cytoplasmic, are transcribed
    - Occupied by sequence-specific transcription factors, cofactors, or chromatin regulatory proteins
    - Organized in accessible chromatin
    - Marked by DNA methylation or specific histone modifications
    - Physically brought together by long-range chromosomal interactions. GENCODE: in humans and mice

http://www.gencodegenes.org/

http://www.genome.gov/encode/

ENCODE

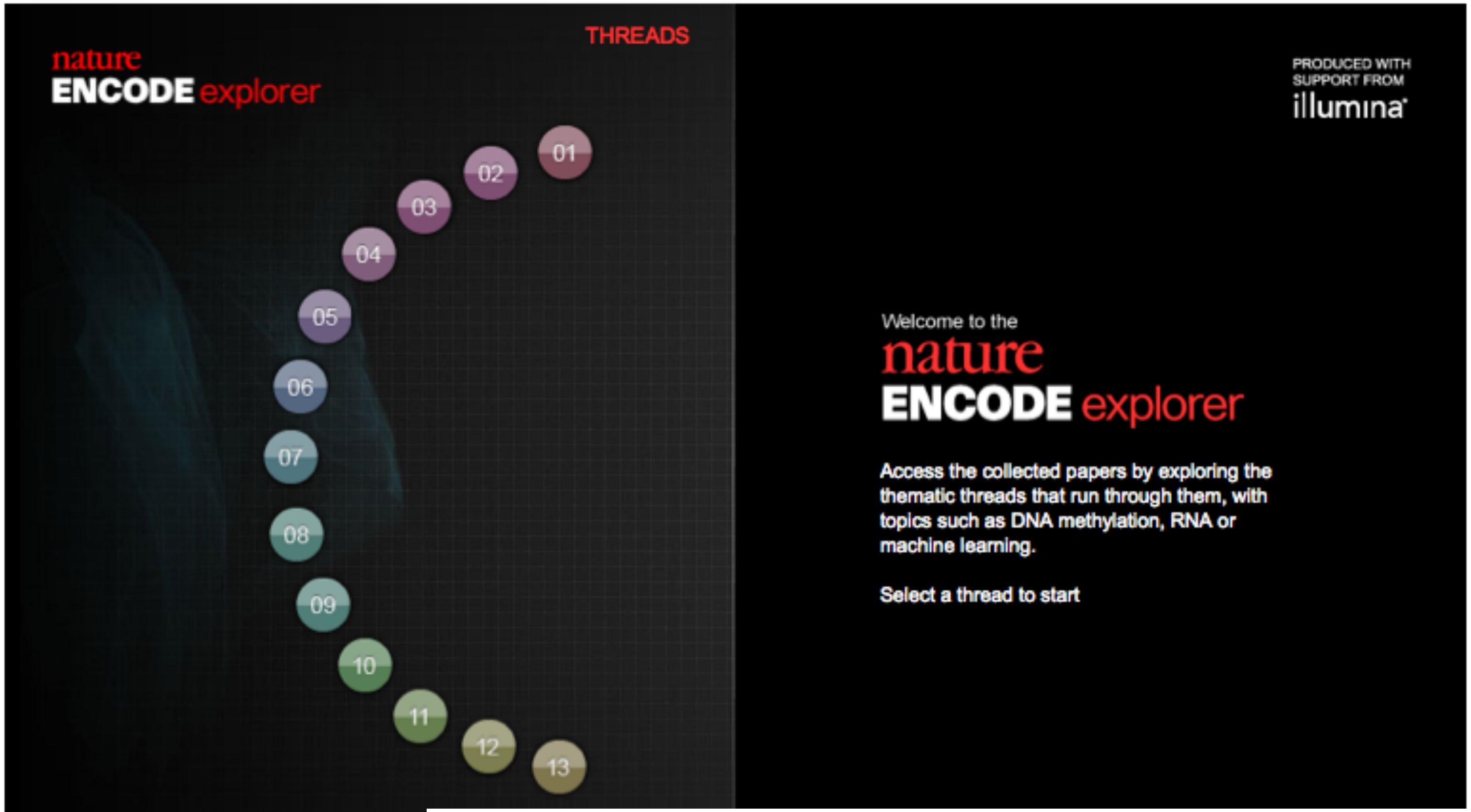
- ENCODE
  - Enhancing and extending annotation of all evidence-based gene features in the genome at a high accuracy
    - Protein-coding loci with alternatively spliced variants
    - Non-coding loci
      - Non-protein coding RNA for instance

http://www.gencodegenes.org/

http://www.genome.gov/encode/

ENCODE

80% of the components of the human genome now have at least one biochemical function associated with them



nature
ENCODE explorer

THREADS

PRODUCED WITH SUPPORT FROM illumina

Welcome to the
nature
ENCODE explorer

Access the collected papers by exploring the thematic threads that run through them, with topics such as DNA methylation, RNA or machine learning.

Select a thread to start

30 papers published across 3 different journals

http://www.nature.com/encode/#/threads

ENCODE

- ENCODE
  - How do I use this with my GWAS SNPs?
  - Lots of information, do I have to go look it up in each individual dataset out there?
  - Thankfully database resources exist!
    - Note, this data is being added to all the time

http://www.genome.gov/encode/

RegulomeDB

- Known and predicted regulatory DNA elements including
  - Regions of DNAase hypersensitivity
  - Binding sites of transcription factors
  - Promoter regions
  - All have been biochemically characterized
- Using an RSID, chromosome or base pair location, or a chromosomal region
  - BED files and VCF files can even be uploaded

- Note, unlike SCAN, information on cell type specificity but not ancestry

Annotation of functional variation in personal genomes using RegulomeDB
Genome Res. 2012 Sep;22(9):1790-7. doi: 10.1101/gr.137323.112.

RegulomeDB

- Known and predicted regulatory DNA elements including

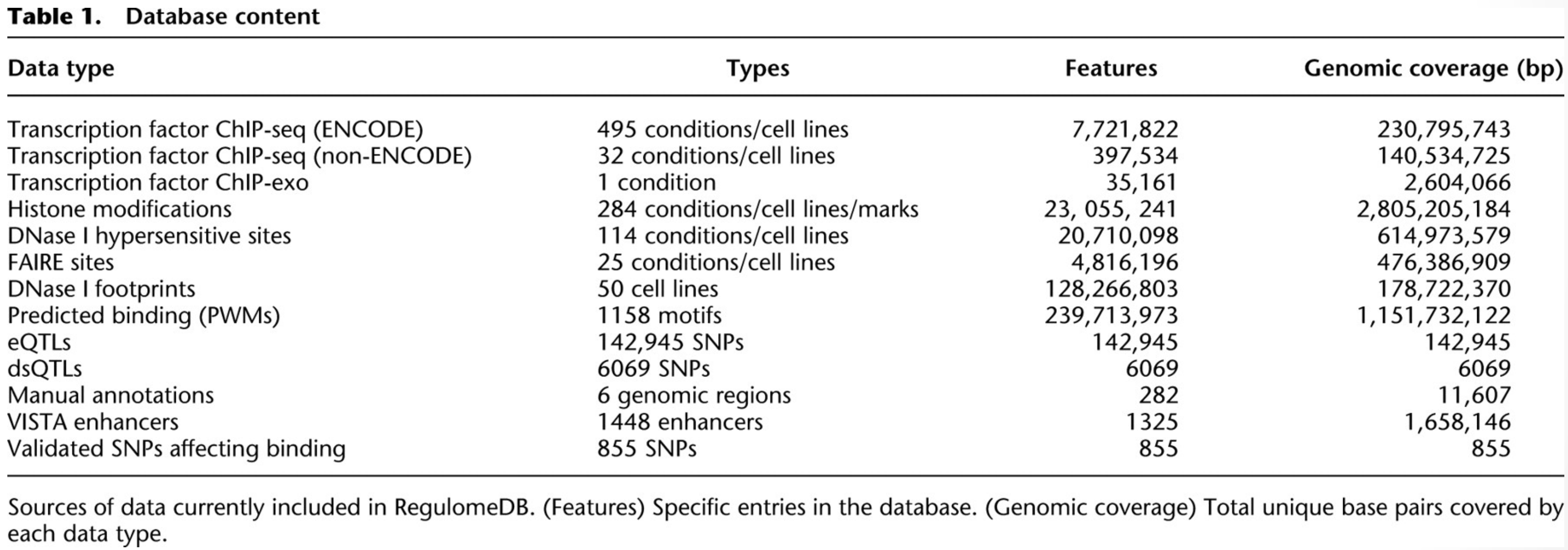


Table 1. Database content

Data type	Types	Features	Genomic coverage (bp)
Transcription factor ChIP-seq (ENCODE)	495 conditions/cell lines	7,721,822	230,795,743
Transcription factor ChIP-seq (non-ENCODE)	32 conditions/cell lines	397,534	140,534,725
Transcription factor ChIP-exo	1 condition	35,161	2,604,066
Histone modifications	284 conditions/cell lines/marks	23, 055, 241	2,805,205,184
DNase I hypersensitive sites	114 conditions/cell lines	20,710,098	614,973,579
FAIRE sites	25 conditions/cell lines	4,816,196	476,386,909
DNase I footprints	50 cell lines	128,266,803	178,722,370
Predicted binding (PWMs)	1158 motifs	239,713,973	1,151,732,122
eQTLs	142,945 SNPs	142,945	142,945
dsQTLs	6069 SNPs	6069	6069
Manual annotations	6 genomic regions	282	11,607
VISTA enhancers	1448 enhancers	1325	1,658,146
Validated SNPs affecting binding	855 SNPs	855	855

Sources of data currently included in RegulomeDB. (Features) Specific entries in the database. (Genomic coverage) Total unique base pairs covered by each data type.

Annotation of functional variation in personal genomes using RegulomeDB
Genome Res. 2012 Sep;22(9):1790-7. doi: 10.1101/gr.137323.112.

RegulomeDB

- This is a huge amount of information!
  - If 80% of the components of the human genome now have at least one biochemical function associated with them… how do I decide what might be important?

- Regulome DB uses a scoring system
  - The more pieces of evidence that a SNP is regulatory in some way, the higher the score
  - Increasing confidence that a variant lies in a functional location and likely results in a functional consequence

Annotation of functional variation in personal genomes using RegulomeDB
Genome Res. 2012 Sep;22(9):1790-7. doi: 10.1101/gr.137323.112.

- RegulomeDB uses a scoring system
  - The more pieces of evidence that a SNP is regulatory in some way, the higher the score
  - Increasing confidence that a variant lies in a functional location and likely results in a functional consequence

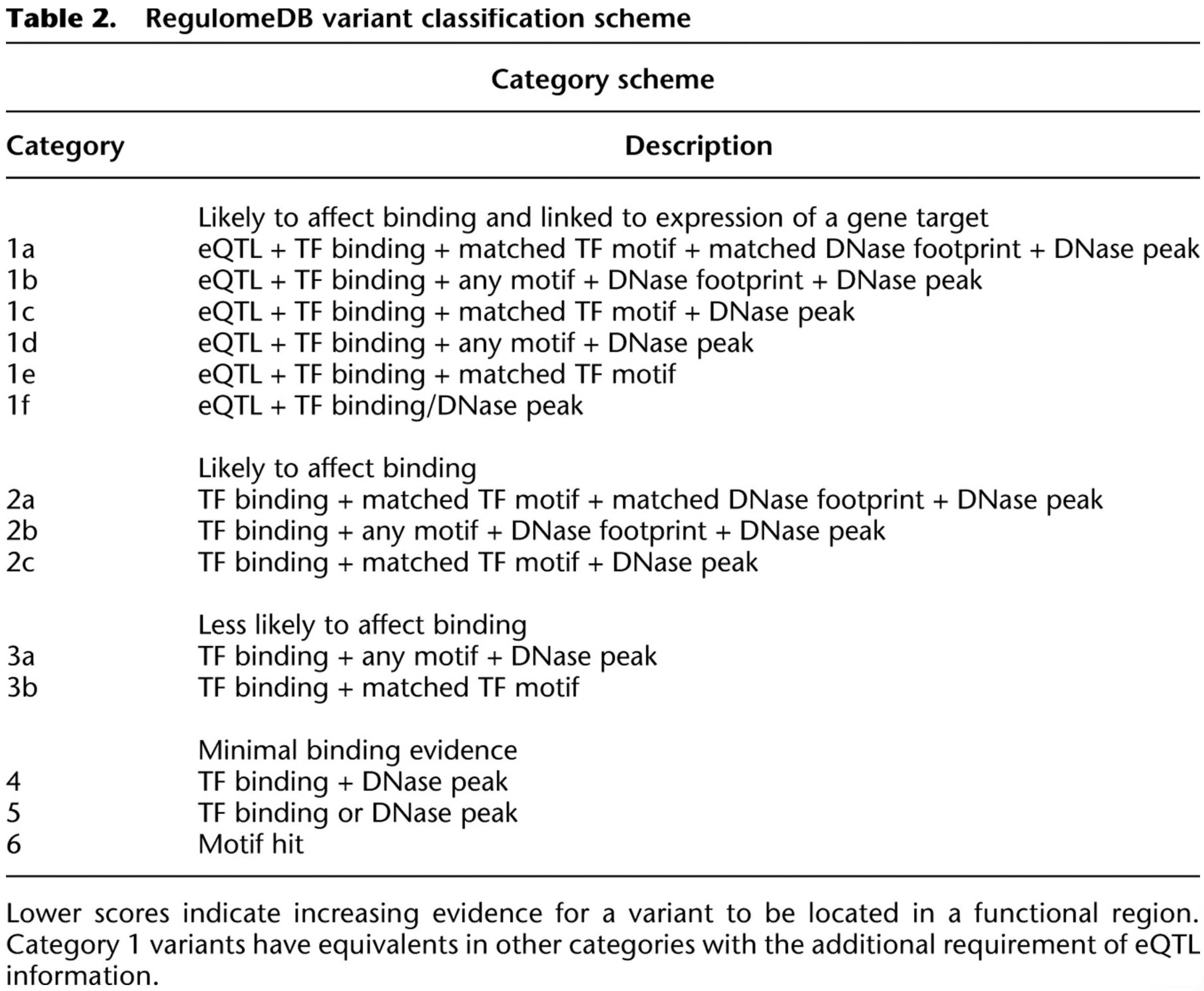


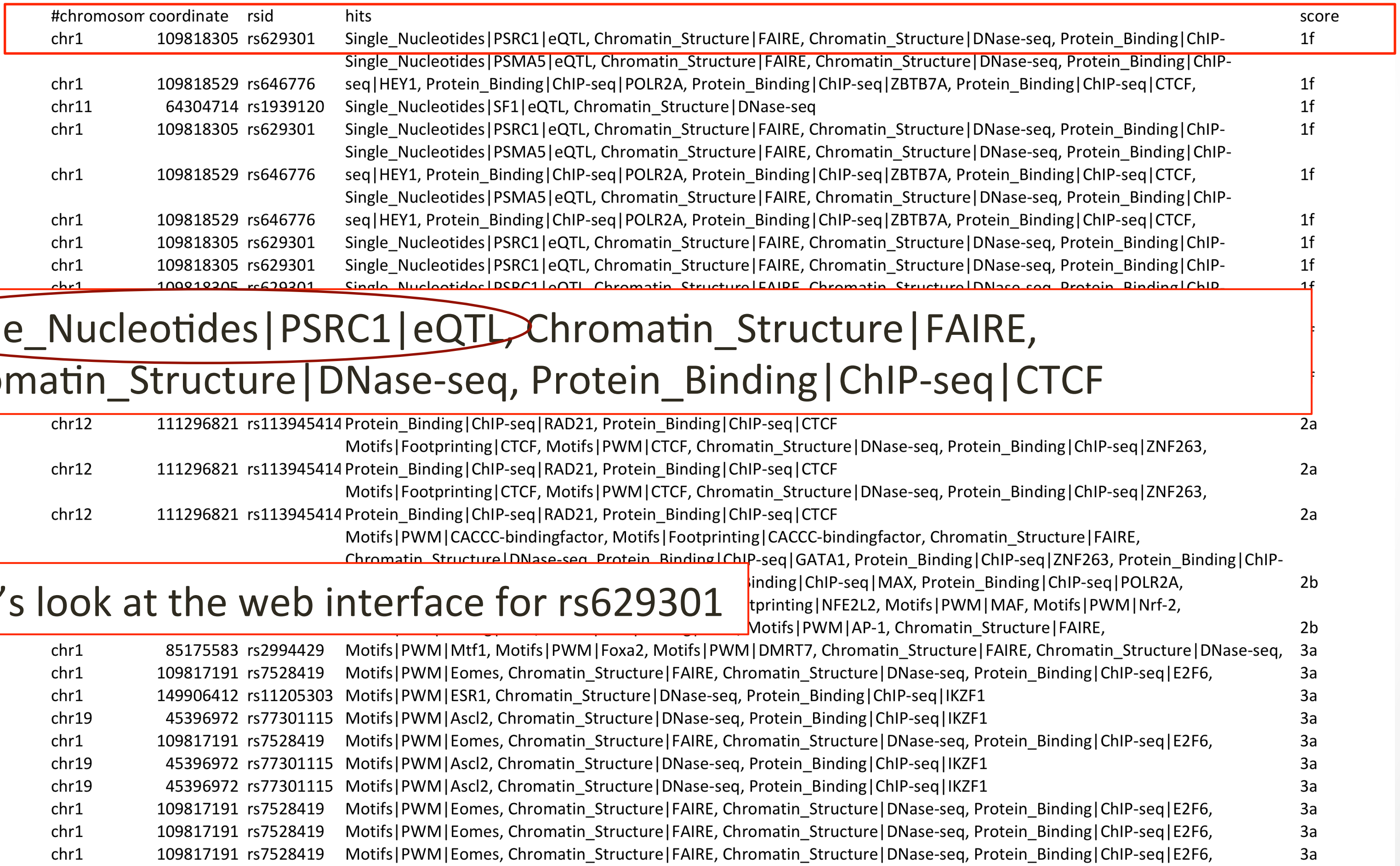
Table 2. RegulomeDB variant classification scheme

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

- So can provide my SNPs of interest and get annotation that looks like this:



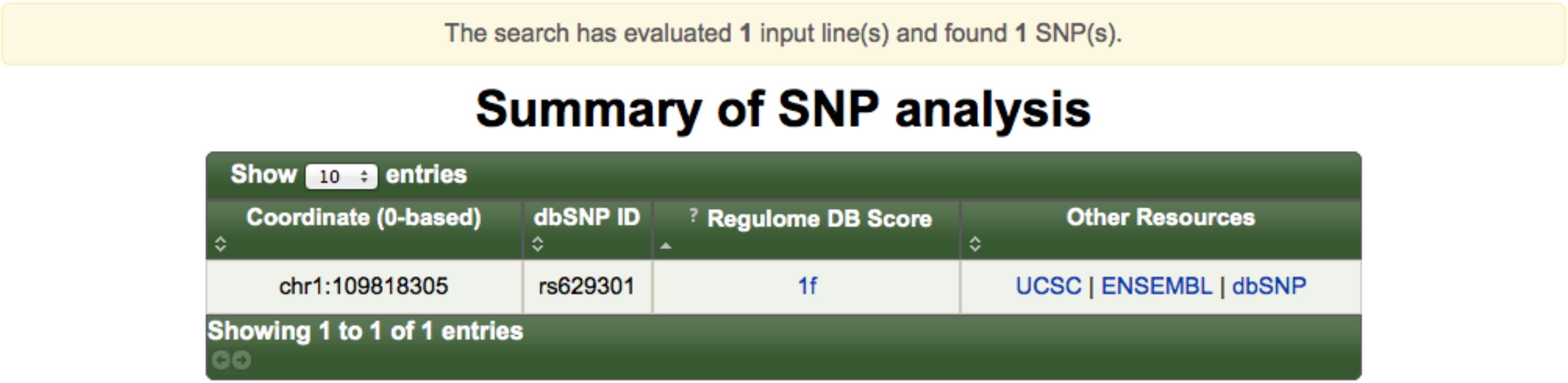
#chromosom	coordinate	rsid	hits	score
chr1	109818305	rs629301	Single_Nucleotides|PSRC1|eQTL, Chromatin_Structure|FAIRE, Chromatin_Structure|DNase-seq, Protein_Binding|ChIP-	1f
chr1	109818529	rs646776	Single_Nucleotides|PSMA5|eQTL, Chromatin_Structure|FAIRE, Chromatin_Structure|DNase-seq, Protein_Binding|ChIP- seq|HEY1, Protein_Binding|ChIP-seq|POLR2A, Protein_Binding|ChIP-seq|ZBTB7A, Protein_Binding|ChIP-seq|CTCF,	1f
chr11	64304714	rs1939120	Single_Nucleotides|SF1|eQTL, Chromatin_Structure|DNase-seq	1f
chr1	109818305	rs629301	Single_Nucleotides|PSRC1|eQTL, Chromatin_Structure|FAIRE, Chromatin_Structure|DNase-seq, Protein_Binding|ChIP-	1f
chr1	109818529	rs646776	Single_Nucleotides|PSMA5|eQTL, Chromatin_Structure|FAIRE, Chromatin_Structure|DNase-seq, Protein_Binding|ChIP- seq|HEY1, Protein_Binding|ChIP-seq|POLR2A, Protein_Binding|ChIP-seq|ZBTB7A, Protein_Binding|ChIP-seq|CTCF,	1f
chr1	109818529	rs646776	Single_Nucleotides|PSMA5|eQTL, Chromatin_Structure|FAIRE, Chromatin_Structure|DNase-seq, Protein_Binding|ChIP- seq|HEY1, Protein_Binding|ChIP-seq|POLR2A, Protein_Binding|ChIP-seq|ZBTB7A, Protein_Binding|ChIP-seq|CTCF,	1f
chr1	109818305	rs629301	Single_Nucleotides|PSRC1|eQTL, Chromatin_Structure|FAIRE, Chromatin_Structure|DNase-seq, Protein_Binding|ChIP-	1f
chr1	109818305	rs629301	Single_Nucleotides|PSRC1|eQTL, Chromatin_Structure|FAIRE, Chromatin_Structure|DNase-seq, Protein_Binding|ChIP-	1f
chr1	109818305	rs629301	Single_Nucleotides|PSRC1|eQTL, Chromatin_Structure|FAIRE, Chromatin_Structure|DNase-seq, Protein_Binding|ChIP-	1f
chr12	111296821	rs113945414	Protein_Binding|ChIP-seq|RAD21, Protein_Binding|ChIP-seq|CTCF	2a
chr12	111296821	rs113945414	Motifs|Footprinting|CTCF, Motifs|PWM|CTCF, Chromatin_Structure|DNase-seq, Protein_Binding|ChIP-seq|ZNF263, Protein_Binding|ChIP-seq|RAD21, Protein_Binding|ChIP-seq|CTCF	2a
chr12	111296821	rs113945414	Motifs|Footprinting|CTCF, Motifs|PWM|CTCF, Chromatin_Structure|DNase-seq, Protein_Binding|ChIP-seq|ZNF263, Protein_Binding|ChIP-seq|RAD21, Protein_Binding|ChIP-seq|CTCF	2a
			Motifs|PWM|CACCC-bindingfactor, Motifs|Footprinting|CACCC-bindingfactor, Chromatin_Structure|FAIRE, Chromatin_Structure|DNase-seq, Protein_Binding|ChIP-seq|GATA1, Protein_Binding|ChIP-seq|ZNF263, Protein_Binding|ChIP- inding|ChIP-seq|MAX, Protein_Binding|ChIP-seq|POLR2A,	2b
			tprinting|NFE2L2, Motifs|PWM|MAF, Motifs|PWM|Nrf-2, Motifs|PWM|AP-1, Chromatin_Structure|FAIRE,	2b
chr1	85175583	rs2994429	Motifs|PWM|Mtf1, Motifs|PWM|Foxa2, Motifs|PWM|DMRT7, Chromatin_Structure|FAIRE, Chromatin_Structure|DNase-seq,	3a
chr1	109817191	rs7528419	Motifs|PWM|Eomes, Chromatin_Structure|FAIRE, Chromatin_Structure|DNase-seq, Protein_Binding|ChIP-seq|E2F6,	3a
chr1	149906412	rs11205303	Motifs|PWM|ESR1, Chromatin_Structure|DNase-seq, Protein_Binding|ChIP-seq|IKZF1	3a
chr19	45396972	rs77301115	Motifs|PWM|Ascl2, Chromatin_Structure|DNase-seq, Protein_Binding|ChIP-seq|IKZF1	3a
chr1	109817191	rs7528419	Motifs|PWM|Eomes, Chromatin_Structure|FAIRE, Chromatin_Structure|DNase-seq, Protein_Binding|ChIP-seq|E2F6,	3a
chr19	45396972	rs77301115	Motifs|PWM|Ascl2, Chromatin_Structure|DNase-seq, Protein_Binding|ChIP-seq|IKZF1	3a
chr19	45396972	rs77301115	Motifs|PWM|Ascl2, Chromatin_Structure|DNase-seq, Protein_Binding|ChIP-seq|IKZF1	3a
chr1	109817191	rs7528419	Motifs|PWM|Eomes, Chromatin_Structure|FAIRE, Chromatin_Structure|DNase-seq, Protein_Binding|ChIP-seq|E2F6,	3a
chr1	109817191	rs7528419	Motifs|PWM|Eomes, Chromatin_Structure|FAIRE, Chromatin_Structure|DNase-seq, Protein_Binding|ChIP-seq|E2F6,	3a
chr1	109817191	rs7528419	Motifs|PWM|Eomes, Chromatin_Structure|FAIRE, Chromatin_Structure|DNase-seq, Protein_Binding|ChIP-seq|E2F6,	3a

Single_Nucleotides|PSRC1|eQTL, Chromatin_Structure|FAIRE, Chromatin_Structure|DNase-seq, Protein_Binding|ChIP-seq|CTCF

Let's look at the web interface for rs629301

RegulomeDB

Let's look at the web interface for rs629301 again



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 Score	Other Resources
chr1:109818305	rs629301	1f	UCSC | ENSEMBL | dbSNP

Showing 1 to 1 of 1 entries

RegulomeDB

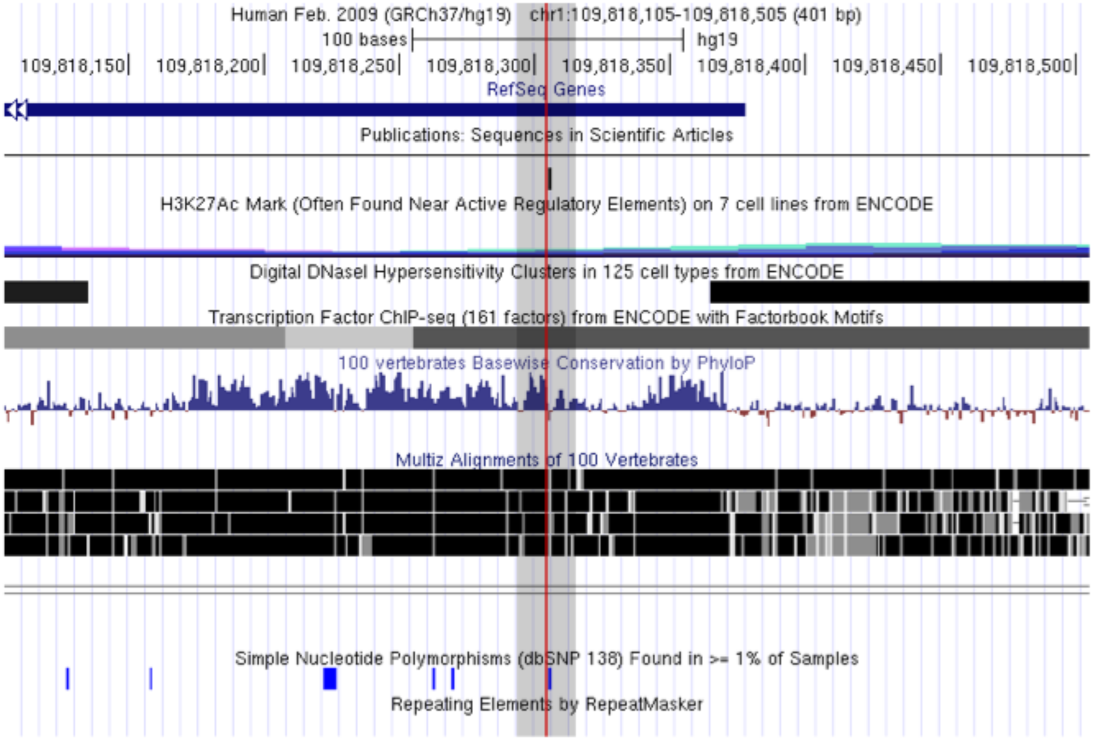
Let’s look at the web interface for rs629301

Can view UCSC genome browser information about the location of the SNP

Data supporting chr1:109818305 (rs629301)

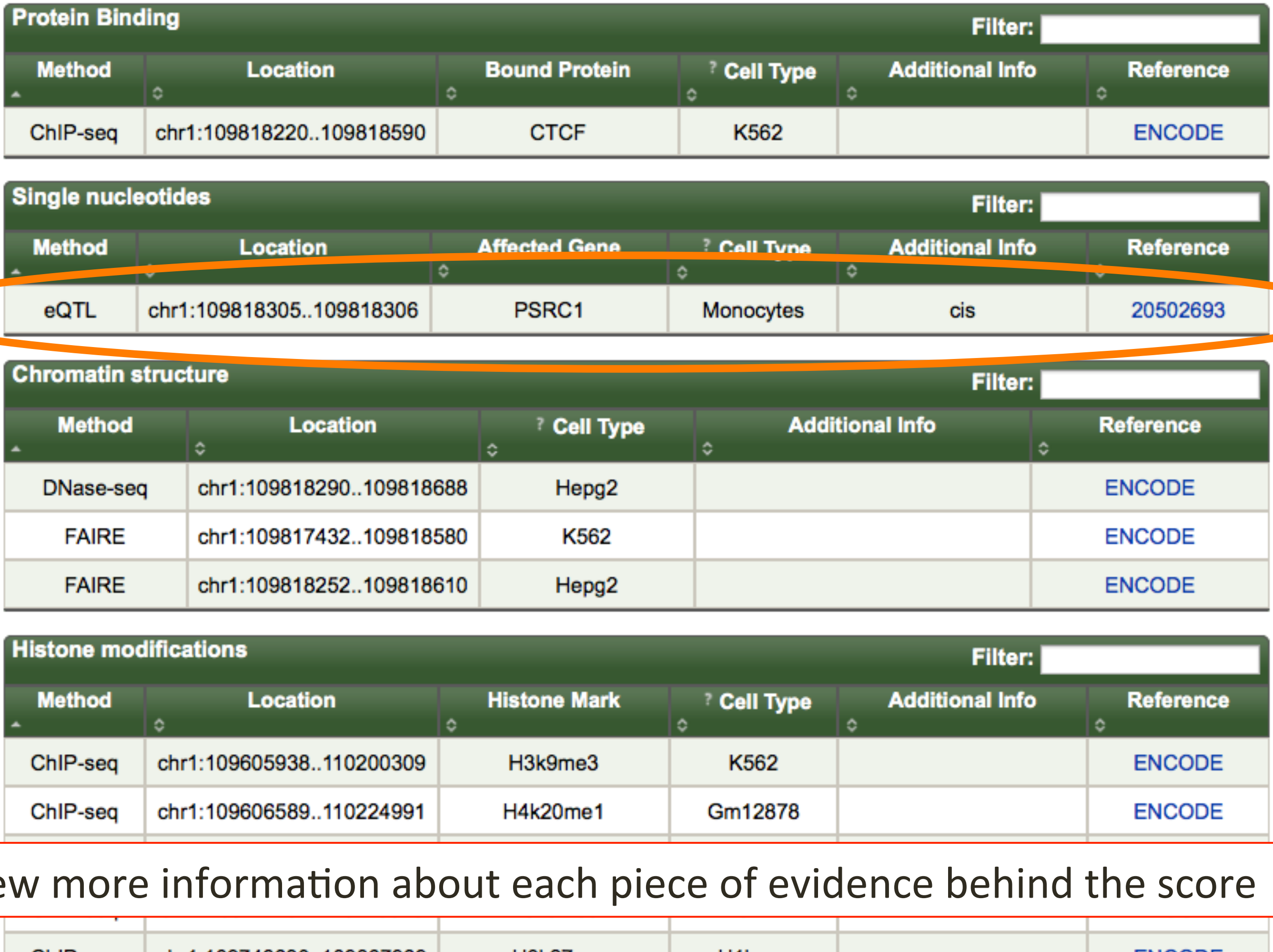
Score: 1f

Likely to affect binding and linked to expression of a gene target


Human Feb. 2009 (GRCh37/hg19) chr1:109,818,105-109,818,505 (401 bp)
100 bases
hg19
109,818,150
109,818,200
109,818,250
109,818,300
109,818,350
109,818,400
109,818,450
109,818,500
RefSeq Genes
Publications: Sequences in Scientific Articles
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Digital DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE
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 138) Found in >= 1% of Samples
Repeating Elements by RepeatMasker

RegulomeDB

Let’s look at the web interface for rs629301



Protein Binding Filter:

Method	Location	Bound Protein	Cell Type	Additional Info	Reference
ChIP-seq	chr1:109818220..109818590	CTCF	K562		ENCODE

Single nucleotides Filter:

Method	Location	Affected Gene	Cell Type	Additional Info	Reference
eQTL	chr1:109818305..109818306	PSRC1	Monocytes	cis	20502693

Chromatin structure Filter:

Method	Location	Cell Type	Additional Info	Reference
DNase-seq	chr1:109818290..109818688	Hepg2		ENCODE
FAIRE	chr1:109817432..109818580	K562		ENCODE
FAIRE	chr1:109818252..109818610	Hepg2		ENCODE

Histone modifications Filter:

Method	Location	Histone Mark	Cell Type	Additional Info	Reference
ChIP-seq	chr1:109605938..110200309	H3k9me3	K562		ENCODE
ChIP-seq	chr1:109606589..110224991	H4k20me1	Gm12878		ENCODE

Can view more information about each piece of evidence behind the score

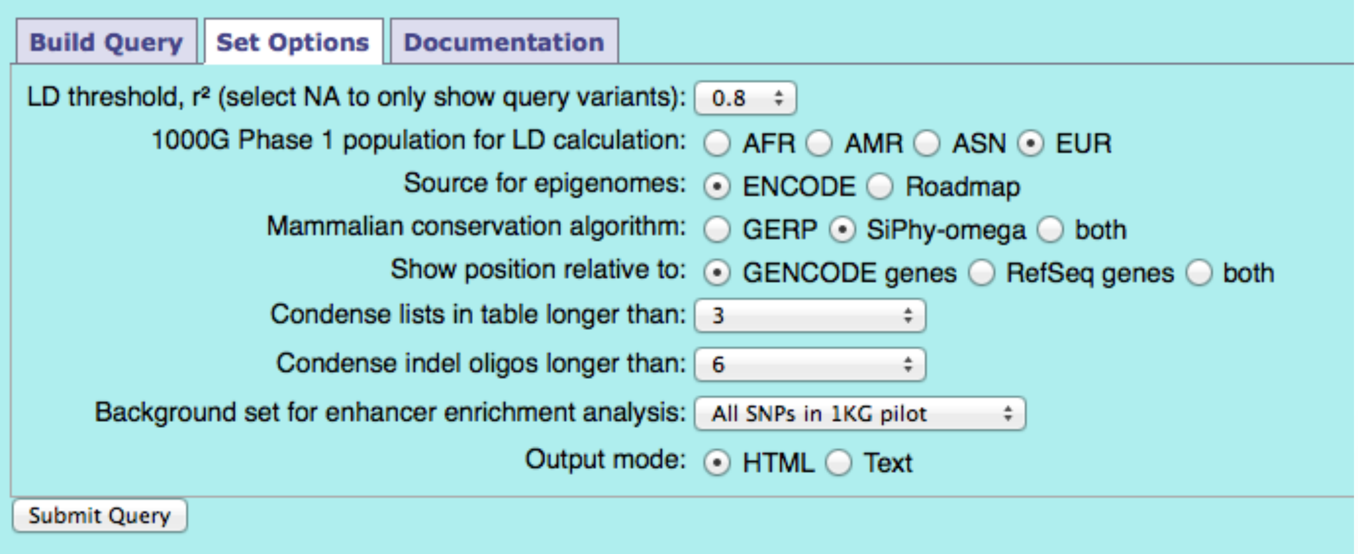
HaploReg

- Exploring annotations of noncoding genome for SNPs
  - A way to develop mechanistic hypothesis of non (protein) coding variants on phenotypic variaiton
- Provides LD information
  - 1000 Genomes Project
- Linked SNPs and small indels (insertions/deletions) can be visualized with predicted chromatin state
- Sequence conservation across mammals
- Effect on regulation
- New Version 2

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

HaploReg

- Enter a list of SNPs
  - We can enter our SNP rs629301
  - Do we see anything different from RegulomeDB?
    - They have a focus on LD
    - Can identify information about SNPs in linkage disequilibrium with your SNP(s) of interest
      - Based on 1000 Genomes populations



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: ● ENCODE ○ Roadmap

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

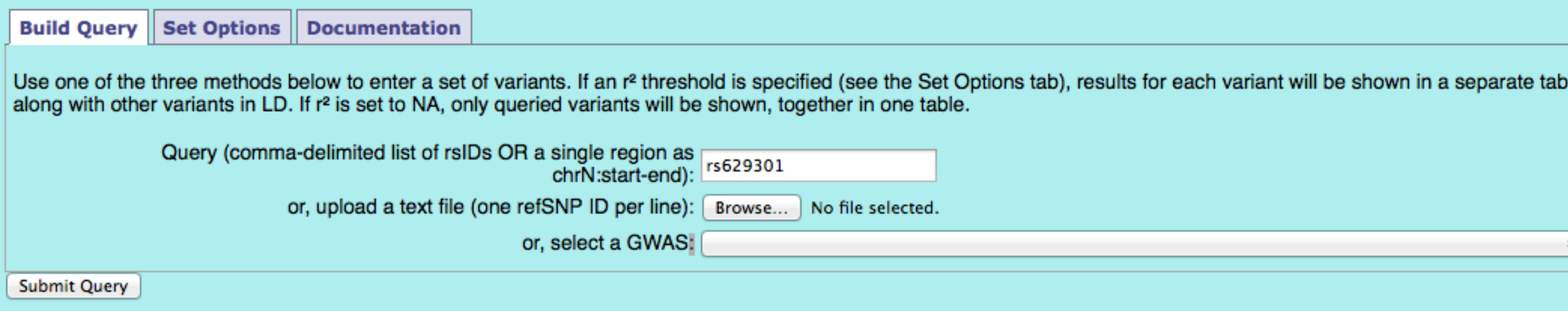
Background set for enhancer enrichment analysis: All SNPs in 1KG pilot

Output mode: ● HTML ○ Text

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

HaploReg

- Enter a list of SNPs
  - We can enter rs629301
  - Do we see anything different from RegulomeDB?



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 tab 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): rs629301

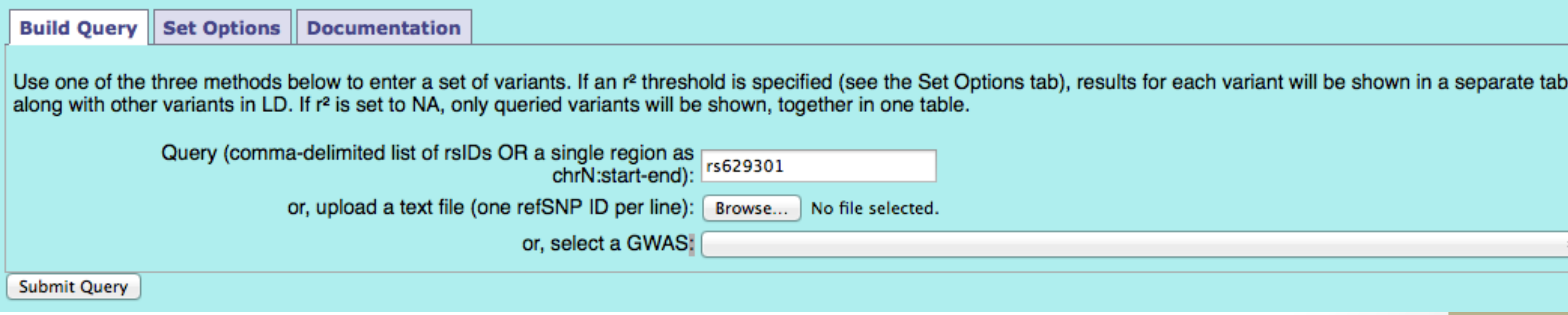
or, upload a text file (one refSNP ID per line): Browse... No file selected.

or, select a GWAS:

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

HaploReg

- Enter a list of SNPs
  - We can enter rs629301
  - Do we see anything different from RegulomeDB?



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 tab 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): rs629301

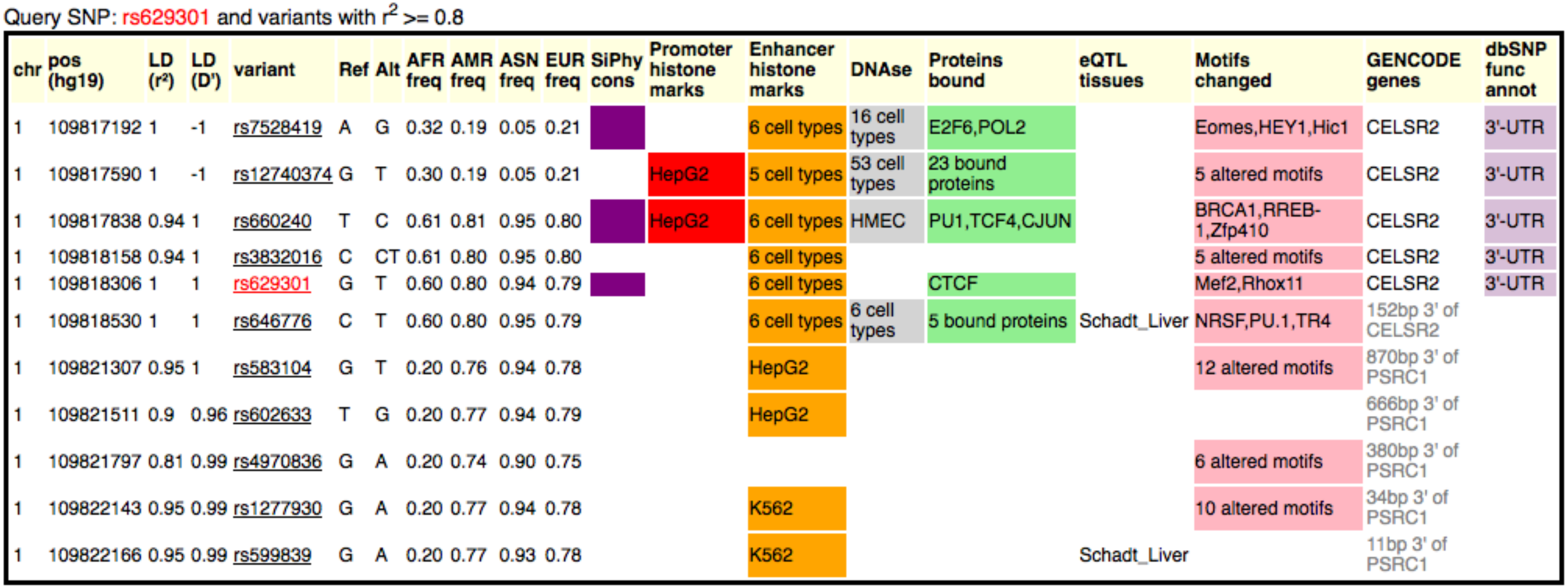
or, upload a text file (one refSNP ID per line): Browse... No file selected.

or, select a GWAS:

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

Using HaploReg

- So for our SNP
  - We have regulatory information for that SNP and nearby SNPs
  - It might be that a SNP in LD with the SNP you have identified in your GWAS is more likely functional…

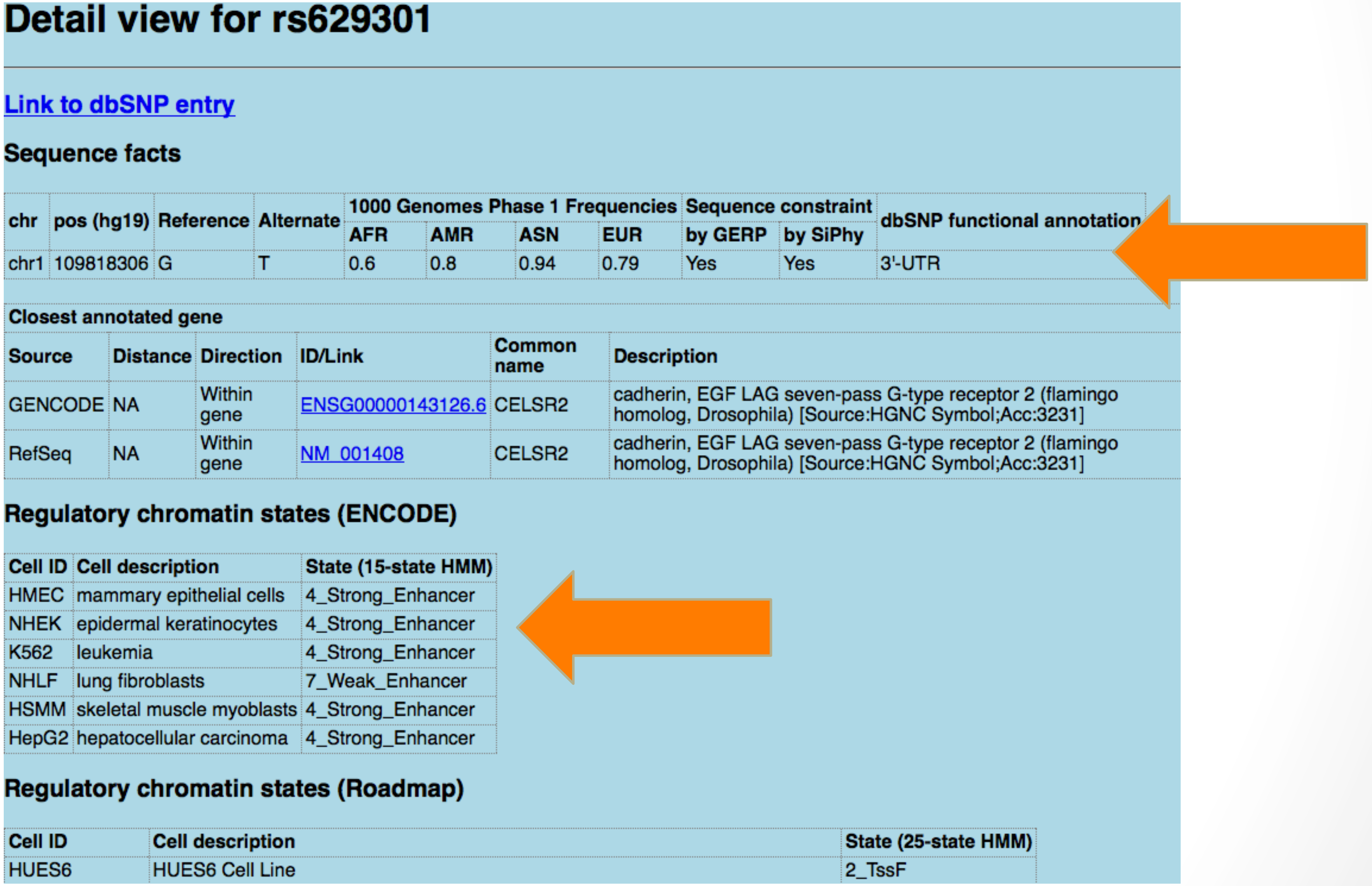


Query SNP: rs629301 and variants with r^2 >= 0.8

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
1	109817192	1	-1	rs7528419	A	G	0.32	0.19	0.05	0.21	🟪		6 cell types	16 cell types	E2F6,POL2		Eomes,HEY1,Hic1	CELSR2	3'-UTR
1	109817590	1	-1	rs12740374	G	T	0.30	0.19	0.05	0.21		HepG2	5 cell types	53 cell types	23 bound proteins		5 altered motifs	CELSR2	3'-UTR
1	109817838	0.94	1	rs660240	T	C	0.61	0.81	0.95	0.80	🟪	HepG2	6 cell types	HMEC	PU1,TCF4,CJUN		BRCA1,RREB-1,Zfp410	CELSR2	3'-UTR
1	109818158	0.94	1	rs3832016	C	CT	0.61	0.80	0.95	0.80			6 cell types				5 altered motifs	CELSR2	3'-UTR
1	109818306	1	1	rs629301	G	T	0.60	0.80	0.94	0.79	🟪		6 cell types		CTCF		Mef2,Rhox11	CELSR2	3'-UTR
1	109818530	1	1	rs646776	C	T	0.60	0.80	0.95	0.79			6 cell types	6 cell types	5 bound proteins	Schadt_Liver	NRSF,PU.1,TR4	152bp 3' of CELSR2	
1	109821307	0.95	1	rs583104	G	T	0.20	0.76	0.94	0.78			HepG2				12 altered motifs	870bp 3' of PSRC1	
1	109821511	0.9	0.96	rs602633	T	G	0.20	0.77	0.94	0.79			HepG2					666bp 3' of PSRC1	
1	109821797	0.81	0.99	rs4970836	G	A	0.20	0.74	0.90	0.75							6 altered motifs	380bp 3' of PSRC1	
1	109822143	0.95	0.99	rs1277930	G	A	0.20	0.77	0.94	0.78			K562				10 altered motifs	34bp 3' of PSRC1	
1	109822166	0.95	0.99	rs599839	G	A	0.20	0.77	0.93	0.78			K562			Schadt_Liver		11bp 3' of PSRC1	

Using HaploRegDB

- What if I look closer at that SNP rs629301?
  - Similar results to RegulomeDB, EXCEPT FOR



Detail view for rs629301

[Link to dbSNP entry]

Sequence facts

chr	pos (hg19)	Reference	Alternate	1000 Genomes Phase 1 Frequencies				Sequence constraint		dbSNP functional annotation
				AFR	AMR	ASN	EUR	by GERP	by SiPhy	
chr1	109818306	G	T	0.6	0.8	0.94	0.79	Yes	Yes	3'-UTR

Closest annotated gene

Source	Distance	Direction	ID/Link	Common name	Description
GENCODE	NA	Within gene	ENSG00000143126.6	CELSR2	cadherin, EGF LAG seven-pass G-type receptor 2 (flamingo homolog, Drosophila) [Source:HGNC Symbol;Acc:3231]
RefSeq	NA	Within gene	NM_001408	CELSR2	cadherin, EGF LAG seven-pass G-type receptor 2 (flamingo homolog, Drosophila) [Source:HGNC Symbol;Acc:3231]

Regulatory chromatin states (ENCODE)

Cell ID	Cell description	State (15-state HMM)
HMEC	mammary epithelial cells	4_Strong_Enhancer
NHEK	epidermal keratinocytes	4_Strong_Enhancer
K562	leukemia	4_Strong_Enhancer
NHLF	lung fibroblasts	7_Weak_Enhancer
HSMM	skeletal muscle myoblasts	4_Strong_Enhancer
HepG2	hepatocellular carcinoma	4_Strong_Enhancer

Regulatory chromatin states (Roadmap)

Cell ID	Cell description	State (25-state HMM)
HUES6	HUES6 Cell Line	2_TssF

Exploring Results

- So back to the 10 SNPs you have from a GWAS
  - Worth looking at SCAN, RegulomeDB, and HaploReg
  - Each source provides different key pieces of information

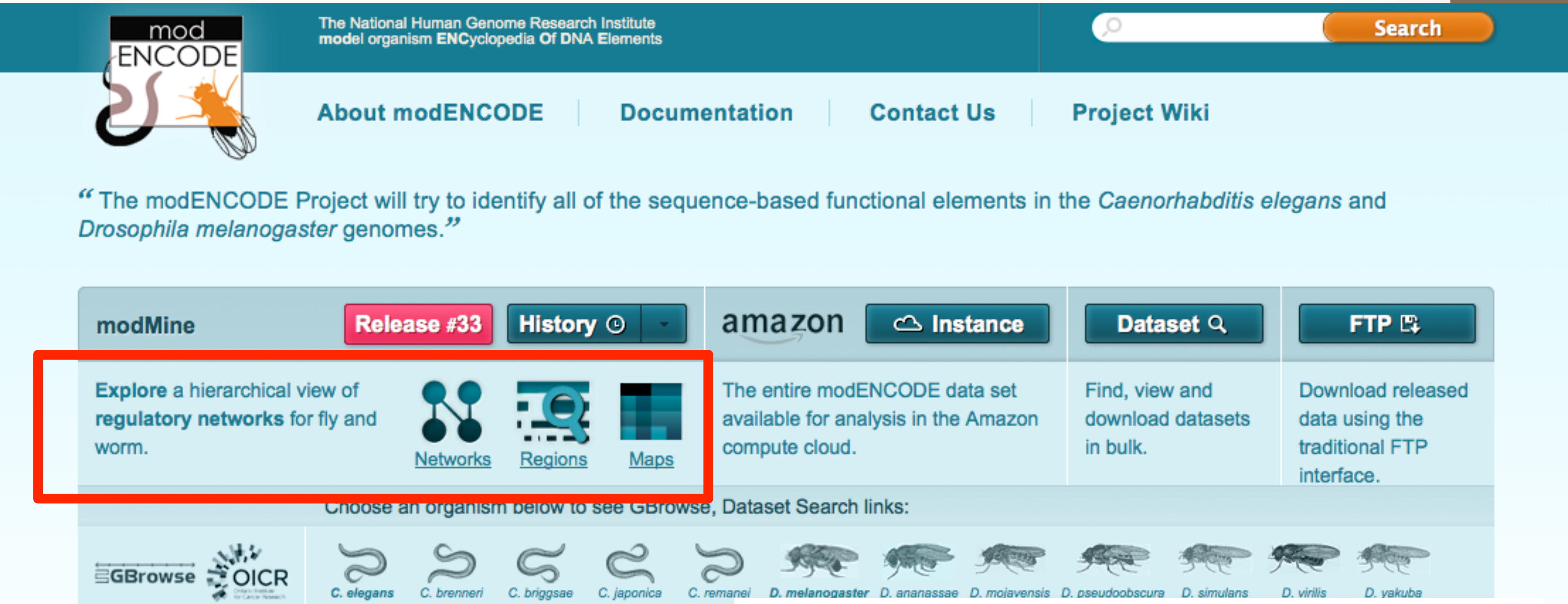
    - **SCAN:** Signs of being an eQTL
      - Target genes, p-values, and population

    - **RegulomeDB:** Information about multiple functional measures indicating the SNP is likely functional
      - Scoring system
      - Cell type specificity

    - **HaploRegDB:** Information about being a likely promoter or enhancer
      - Cell type specificity
      - Expansion to other SNPs based on LD for different ancestry groups

Model Organisms

- Have only discussed human based ENCODE
  - ModENCODE: trying to identify all sequence-based functional elements in *C. elegans* and *Drosophila melanogaster*



“The modENCODE Project will try to identify all of the sequence-based functional elements in the *Caenorhabditis elegans* and *Drosophila melanogaster* genomes.”

Explore a hierarchical view of **regulatory networks** for fly and worm.

The entire modENCODE data set available for analysis in the Amazon compute cloud.

Find, view and download datasets in bulk.

Download released data using the traditional FTP interface.

Choose an organism below to see GBrowse, Dataset Search links:

Mouse ENCODE too…

Explore hierarchical view of regulatory networks
Upload genetic regions and explore
Upload list of fly genes and explore in heatmap

Questions?