ENCODE

Tim Hubbard @timjph
King's College London, King's Health Partners
Genomics England

Bioinformatics, Interpretation and Data Quality in Genome Analysis
MSc in Genomics Medicine
15th February 2016

Questions about a variant

- Has it been seen before?
- How common is it?
- What is known about its effect on function?

- Rare diseases
  - Variant shouldn’t be common
  - Might have already been linked to a disease

Questions about a variant

- Has it been seen before?
- How common is it?
- What is known about its effect on function?

- Rare diseases
  - Variant shouldn't be common
  - Might have already been linked to a disease

Annotating Genes

Vertebrate annotation timeline

<1995: RefSeq Human cDNA curation at NCBI
HPG clone curation at Sanger, submitted to EMBL

1999: Ensembl (automatic gene annotation)
First curated HGP chromosome in Nature (chr22)

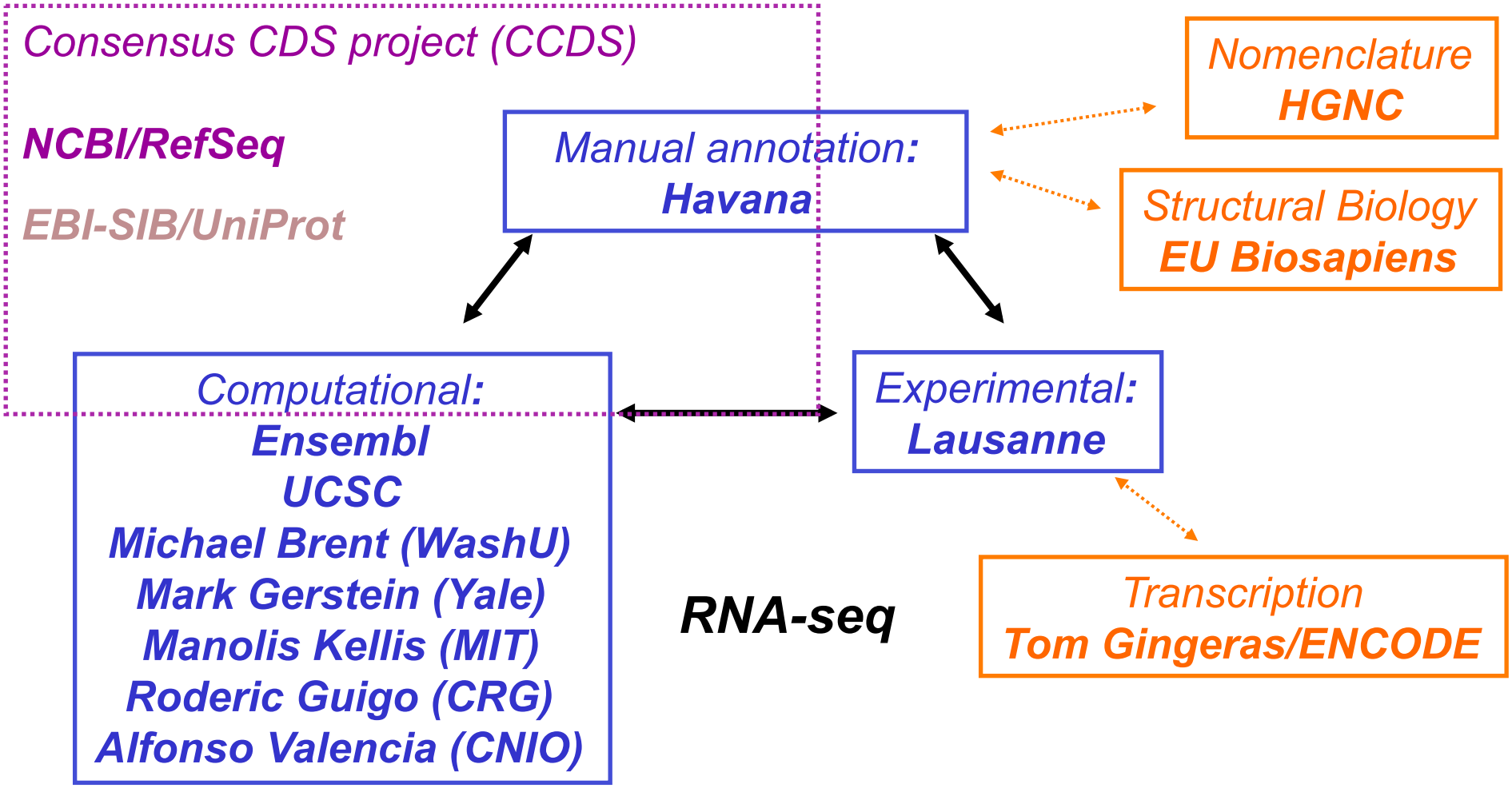
2002: HGP agreement: each human chromosome to be annotated, deposited (Vega), published

2003: Creation of CCDS collaboration: Hinxton/NCBI/UCSC; ENCODE pilot starts

2004: GENCODE consortium for geneset for ENCODE pilot

2005: EGASP “competition” meeting organised by GENCODE

GENCODE: Consensus human gene set



GENCODE = ENSEMBL human geneset

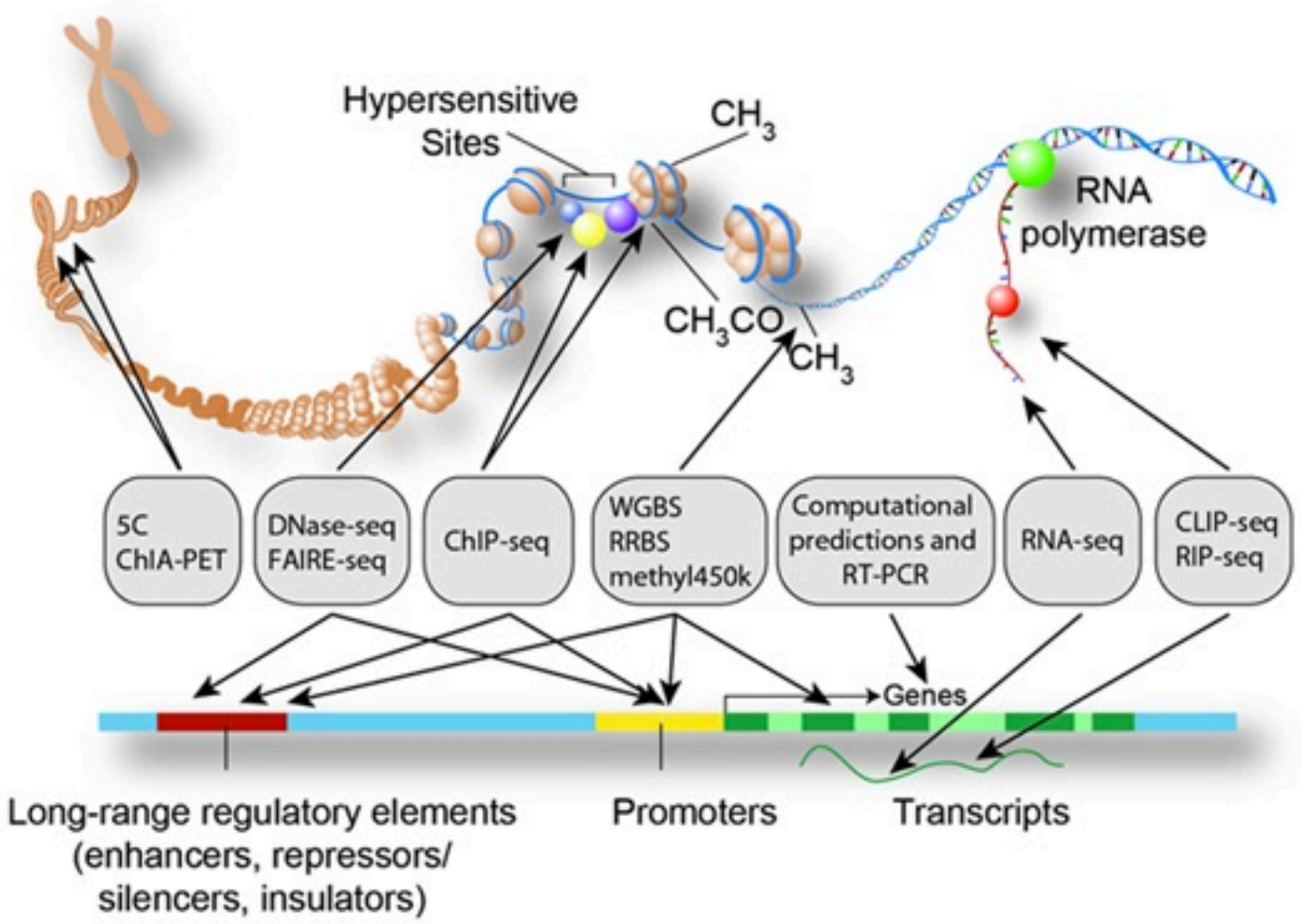
GENCODE15 = Ensembl v70 – http://www.gencodegenes.org/

Highlights from GENCODE project

- Still ~20,000 protein coding genes, but >10,000 non coding RNA genes (ncRNAs)
- 10,000 pseudo genes, but some are really polymorphic genes; some are expressed; some show translation products
- Still many uncertainties:
  - How many ncRNA genes?
  - How many apparently ncRNA genes contain short CDS regions
  - How many more alternative splice forms are there?

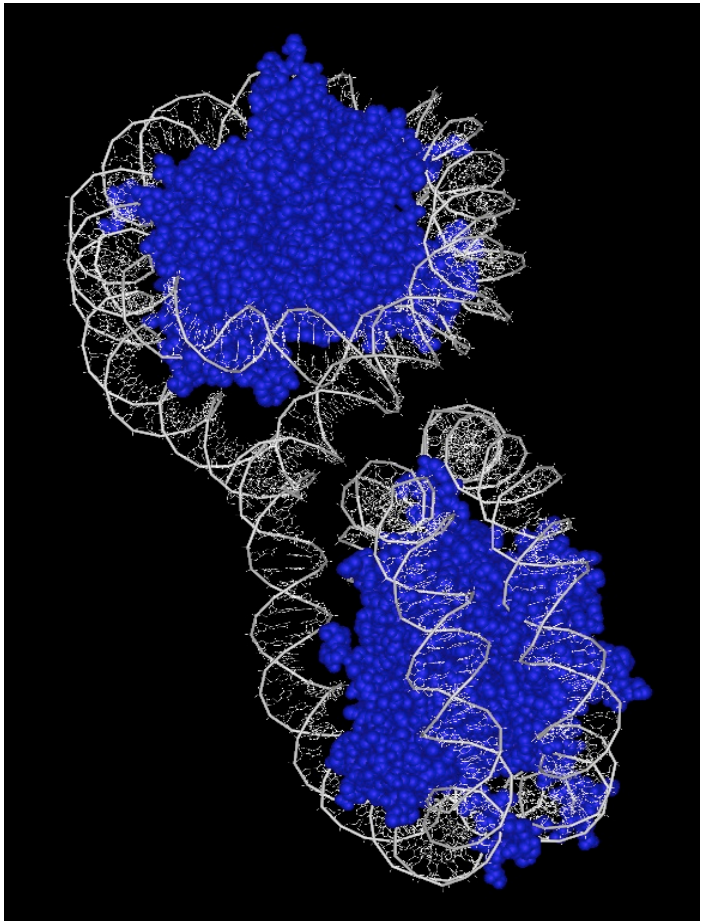
Annotating the rest of the Genome

ENCODE experiments



ENCODE Chromatin Structure Data

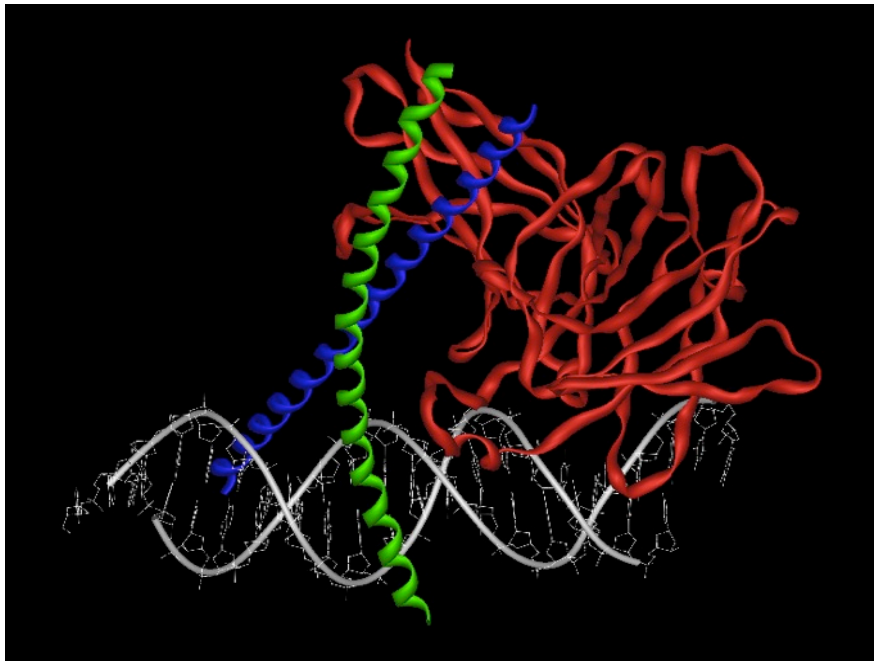
- DNase
  - Enhancers
  - Promoters
  - Transcription Factor Footprints
- Histone Modifications
  - Promoters
  - Enhancers
  - Transcribed regions
  - Active and repressed regions



Richmond, PDB 1ZBB

ENCODE Nucleic Acid Binding Data

- Transcription factor binding
  - Activators
  - Repressors
  - Remodelers
  - RNA Polymerases
- RNA binding proteins
  - RNA Splicing
  - Translation
  - RNA Stability
  - RNA Localization



Harrison, PDB_1A02

ENCyclopedia Of DNA Elements

- ENCODE 1% pilot (2003-7)
  - array based assays
- ENCODE production (2007-12)
  - array based assays; standardized cell lines
  - from 2008, sequencing based assays

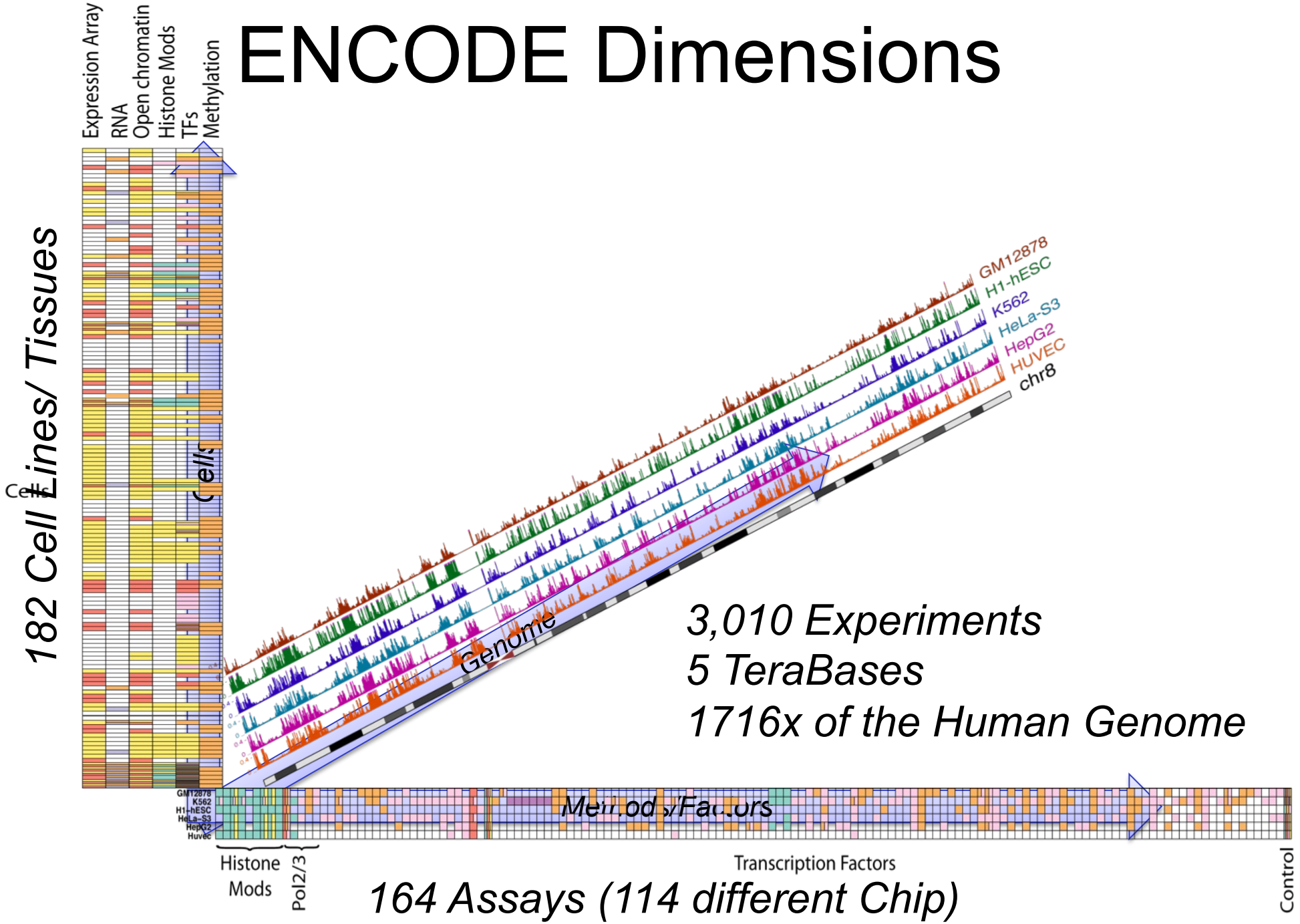
ENCODE Project Consortium et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447, 799–816 (2007).

ENCyclopedia Of DNA Elements

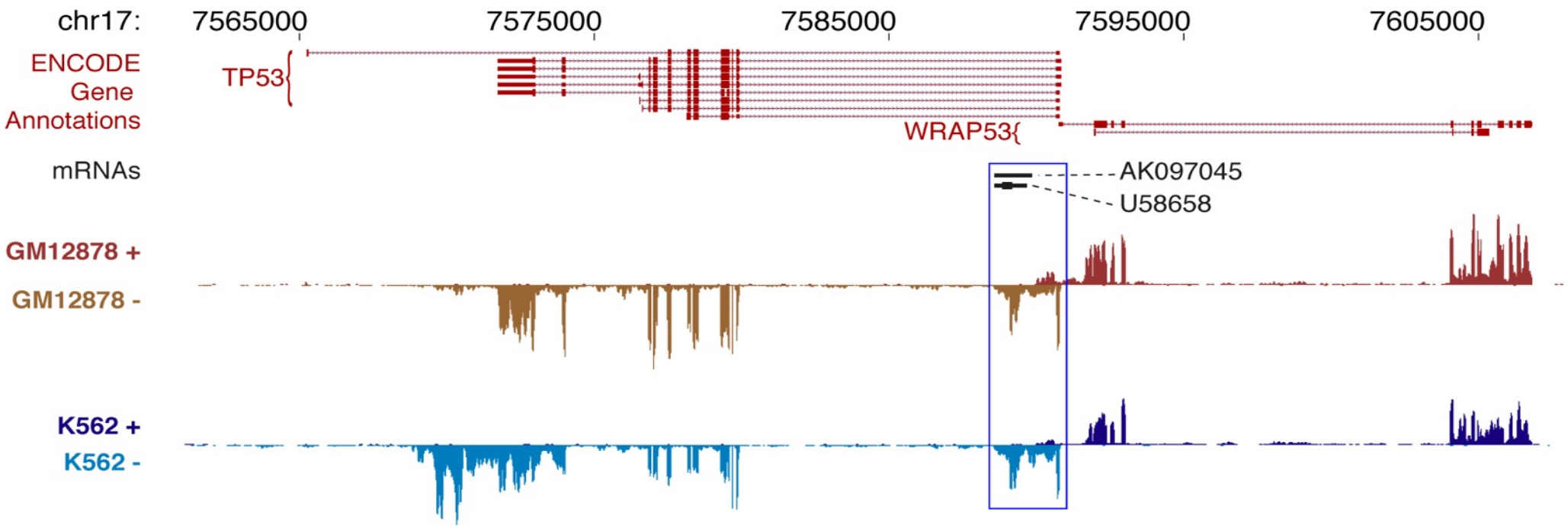
- ENCODE 1% pilot (2003-7)
  - array based assays
- ENCODE production (2007-12)
  - array based assays; standardized cell lines
  - from 2008, sequencing based assays
- ENCODE3 (2013-17)
  - Participants: http://www.genome.gov/26525220

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

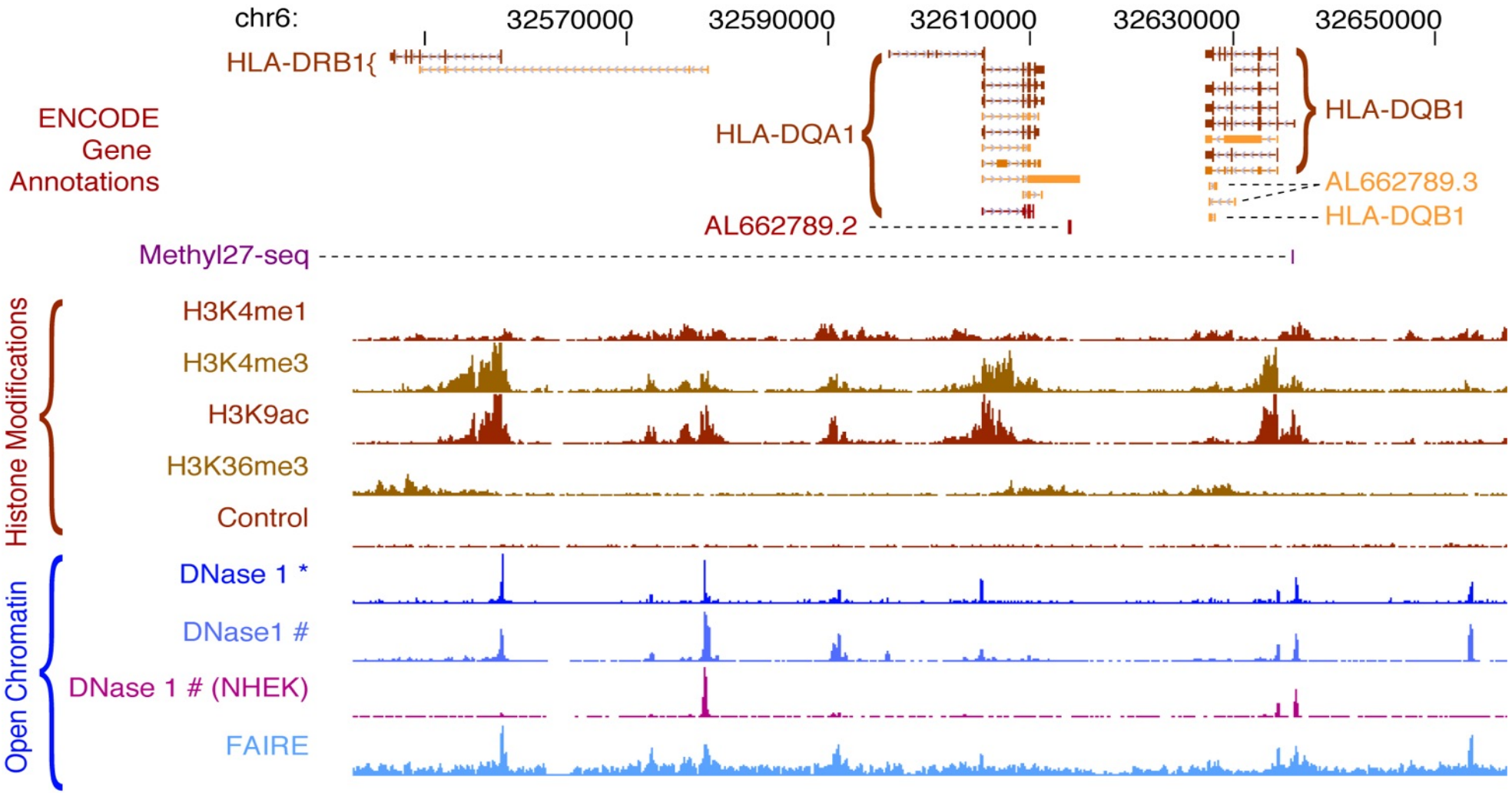
ENCODE Dimensions



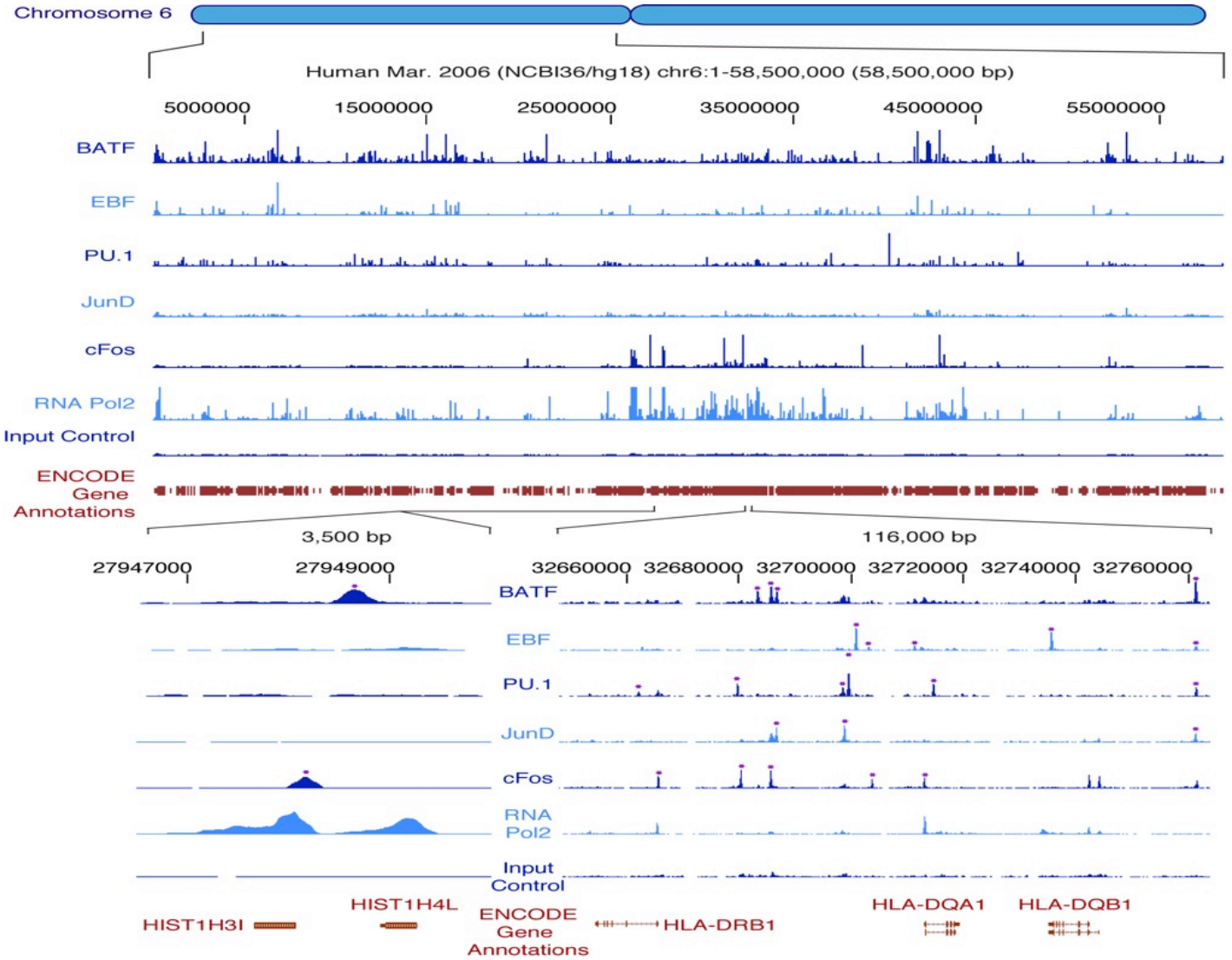
ENCODE Data: RNA-seq



ENCODE Data: Chromatin Assays

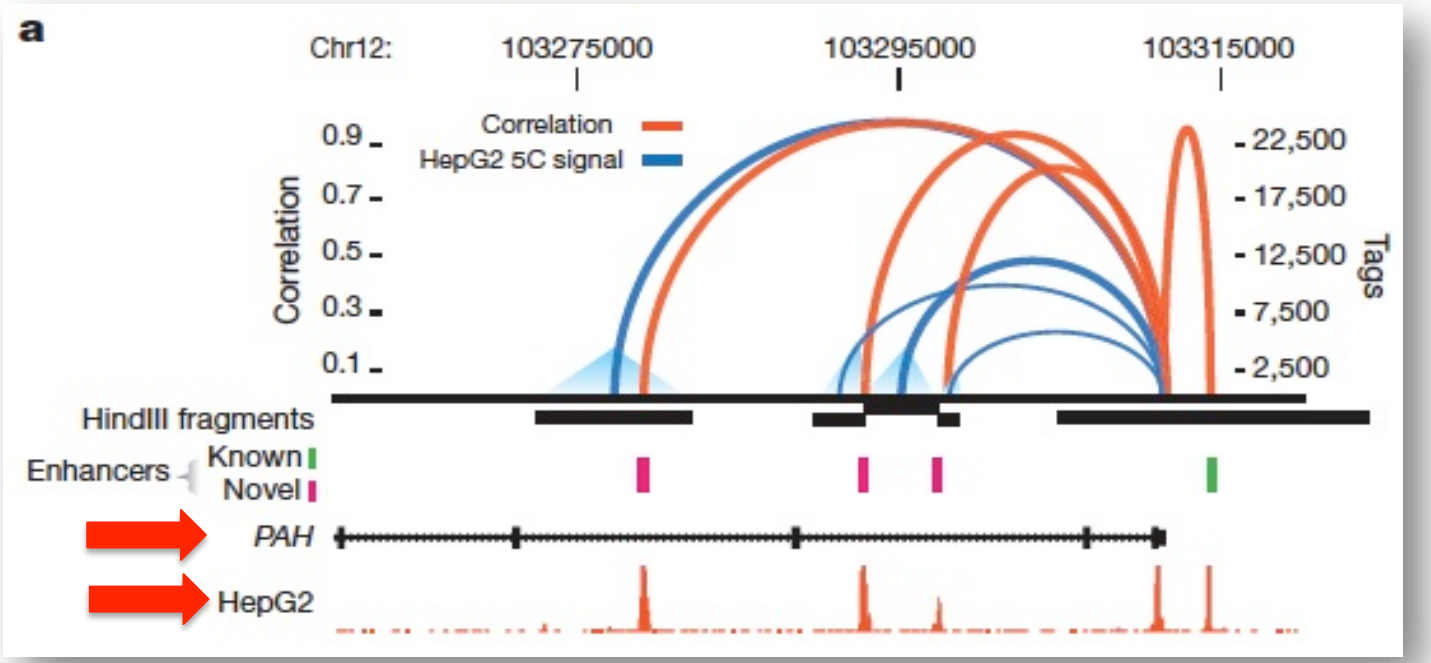


ENCODE production: TF ChIP-seq



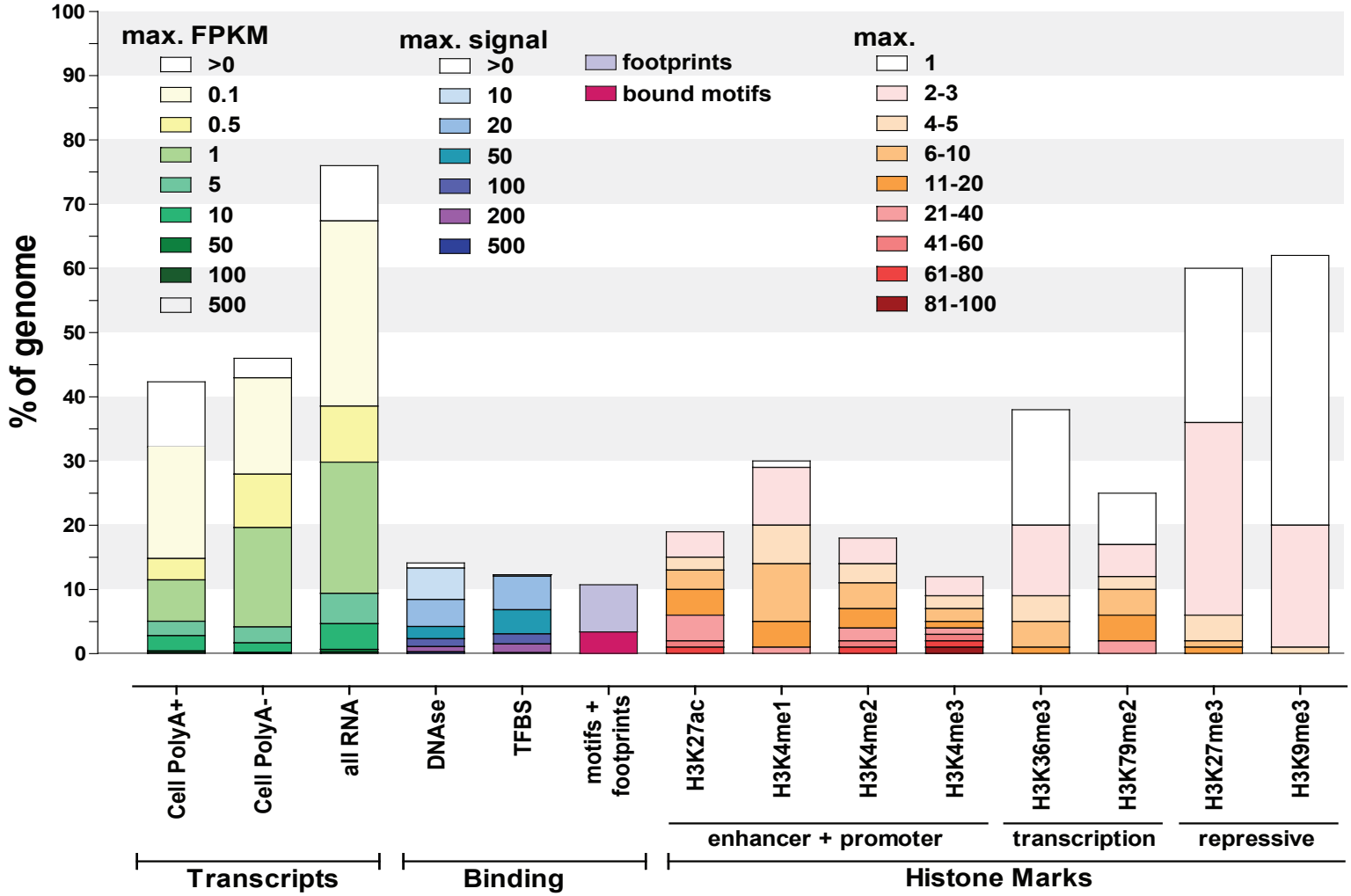
Information From ENCODE Data Integration

- Linkage of functional elements
  - Correlation of DNase signals
  - Long-range assays (5C and ChIA-PET)



Stamatoyannopoulos, Crawford, Nature 489-75, 2012

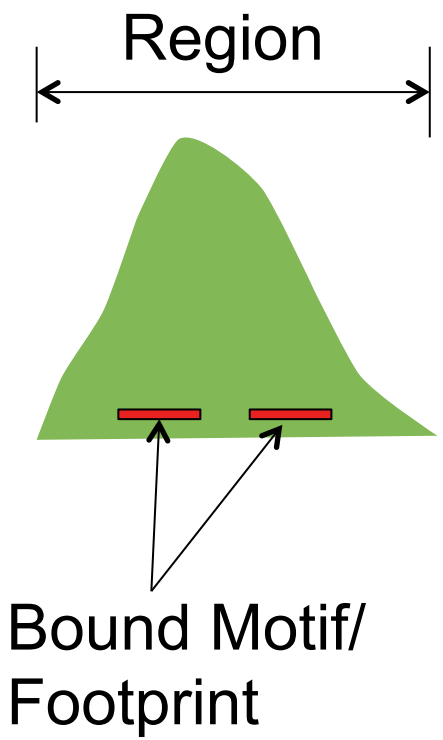
Genome Coverage of ENCODE Elements



Raw genome coverage of elements

Element Type	Coverage	Cumulative Coverage
Exons	3%	3%
Chip-seq bound motifs	4.5%	5%
DNaseI Footprints	5.7%	9%
Chip-seq bound regions	8.1%	12%
DNaseI HS regions	15.2%	19.4%
Histone Modifications (*)	44%	49%
RNA	62%	80%
(* excluding broad marks)		

(Union over all experiments and cell types)

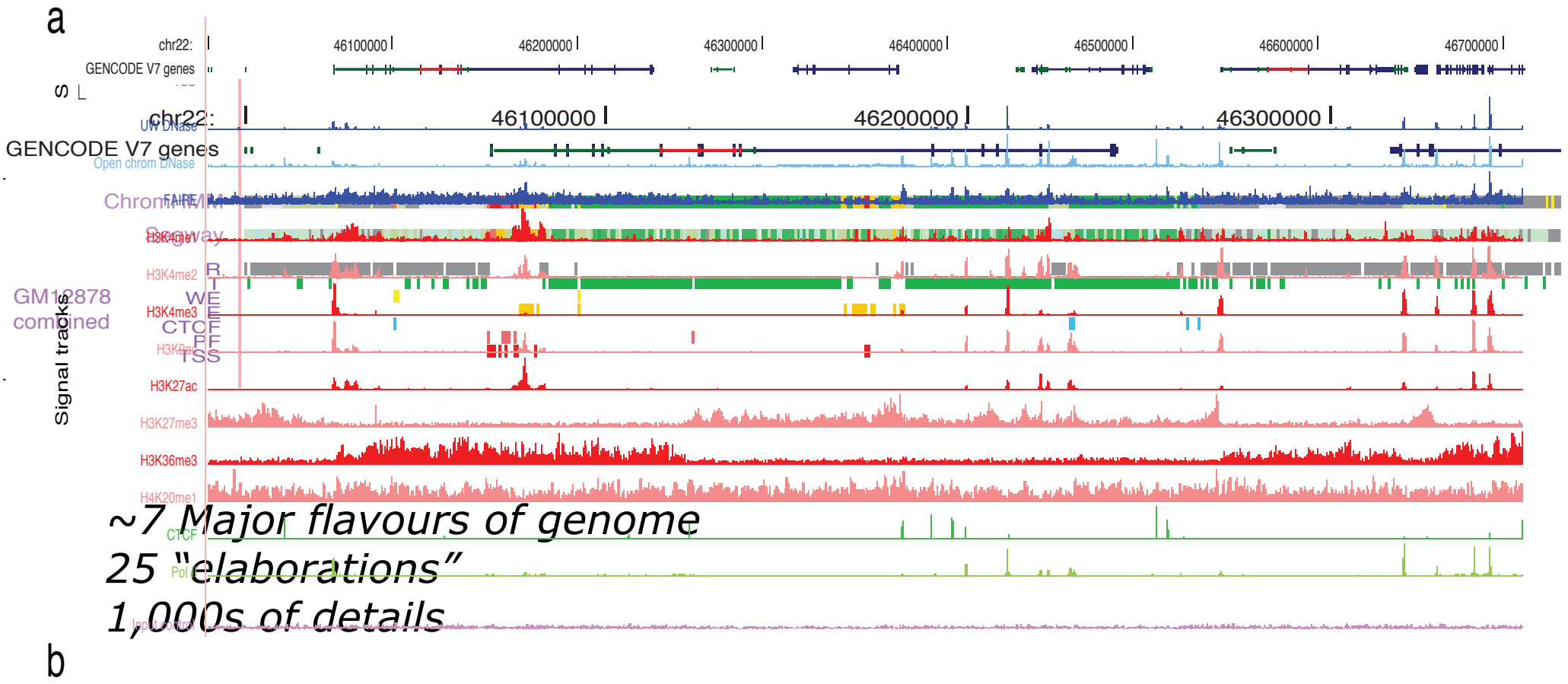


Evenly spaced over the genome

99% of the genome is within 1.7 KB of a biochemical event

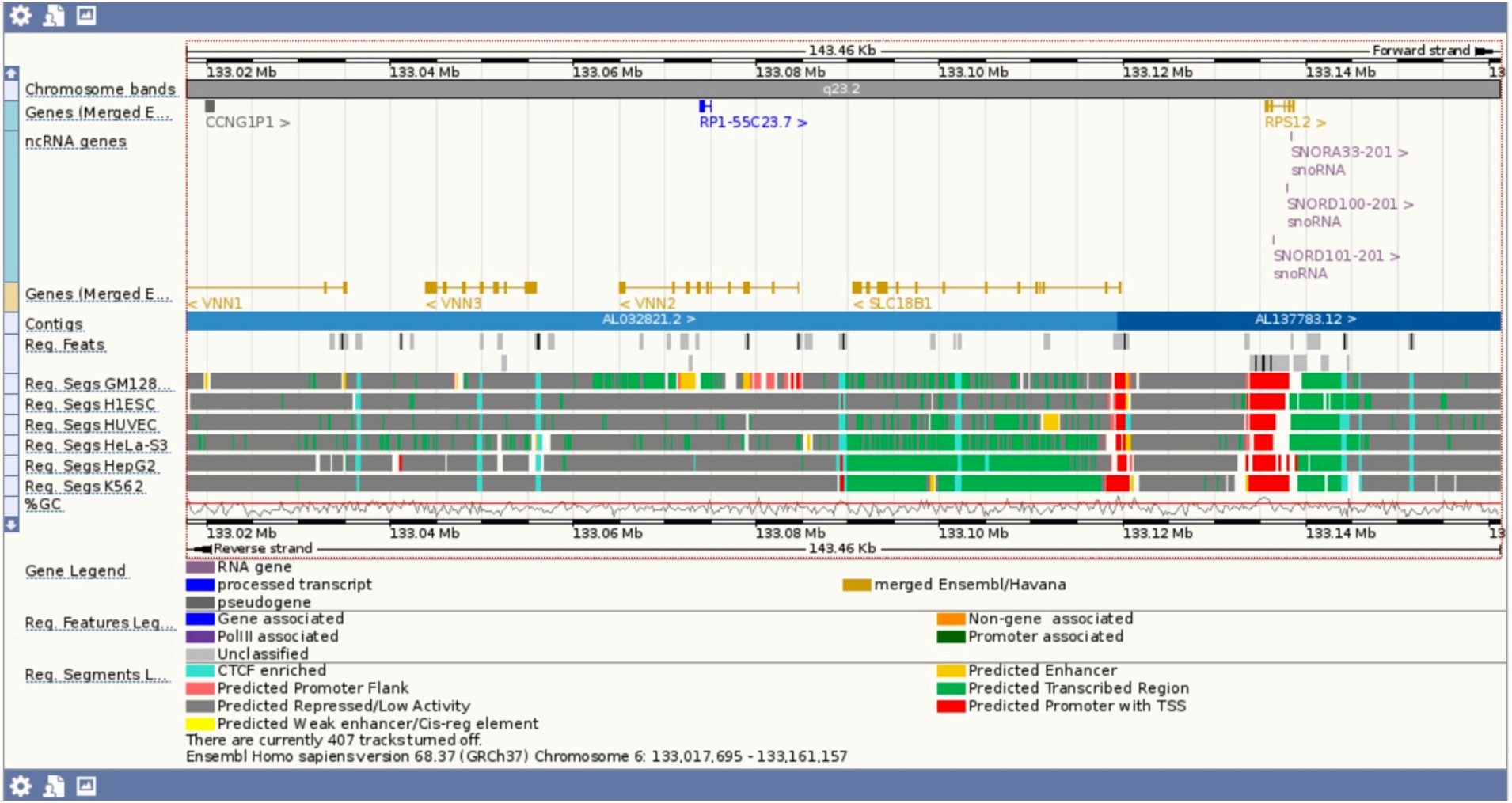
95% of the genome is within 8 KB of a bound motif or footprint

Genome Segmentation



~7 Major flavours of genome
25 "elaborations"
1,000s of details

Genome Segmentation



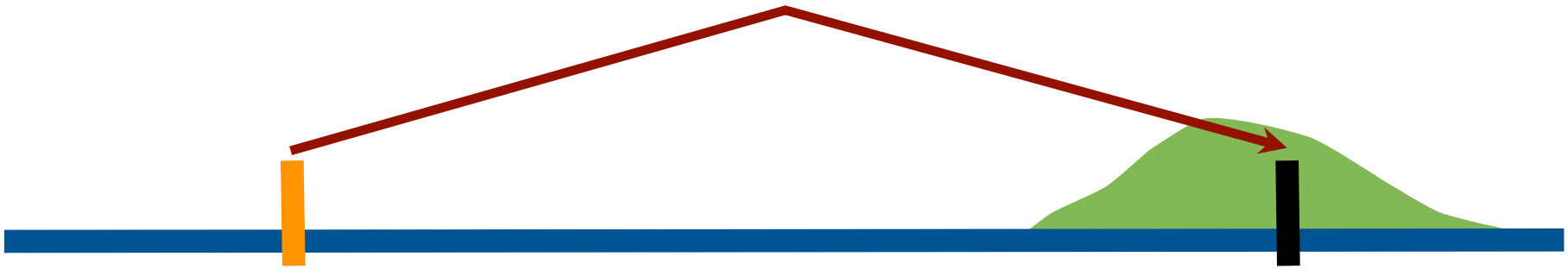
Functional SNPs

Belinda Giardine, Marc Shaub, Ross Hardison, Mike Snyder, John Stam.

Genome Wide Association Studies (GWAS) Results

Linkage Disequilibrium

ENCODE Functional Region



Reported SNP

Statistically associated with the phenotype

fSNP

✔ Associated with the phenotype
✔ In a functional region

Evidence That Non-coding DNA Is Important

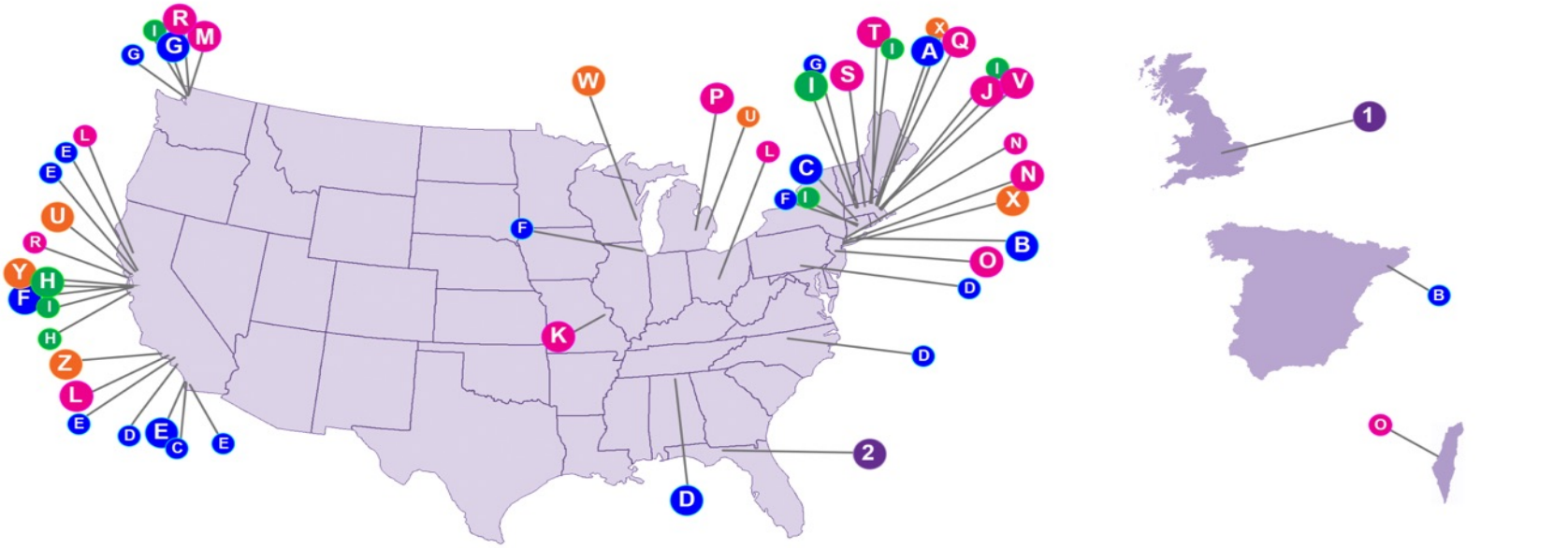
Non-coding DNA variants are known to cause human diseases

- Prostate Cancer
- Thyroid Carcinoma
- Hereditary mixed polyposis syndrome
- Thalassemias
- Fragile X Syndrome
- Alzheimers, ALS
- Diabetes

High-Level Findings

- Very large fraction of the genome is biochemically active
  - 80% of the genome has an ENCODE annotation in at least one cell type
  - 99% of genome is within 1.7kb of at least one ENCODE-measured biochemical event
- GWAS SNPs are enriched within non-coding functional elements
  - >50% of non-coding GWAS SNPs are near ENCODE-defined regions
  - In many cases, disease phenotypes can be associated with a specific cell type or transcription factor.
- Segmenting the genome into 7 chromatin states predicts ~400,000 enhancers and ~70,000 promoters as well as 1000s of quiescent states

ENCODE Consortium



Production Groups

- A **Broad Institute**
- B **Cold Spring Harbor;** Centre for Genomic Regulation (CRG);
- C **University of Connecticut Health Center**; UCSD
- D **HudsonAlpha;** Pennsylvania State; UC Irvine; Duke; Caltech
- E **UCSD**; Salk Institute ; Joint Genome Institute; Lawrence Berkeley National Laboratory; UCSD
- F **Stanford**; University of Chicago; Yale
- G **University of Washington**; Fred Hutchinson Cancer Research Center; University of Massachusetts Medical School

Data Coordination Center

- H **Stanford**; UCSC

Data Analysis Center

- I **University of Massachusetts Medical School**; Yale; MIT; Stanford; Harvard; University of Washington

Technology Development Groups

- J **MIT**
- K **Washington University, St. Louis**
- L **USC**; Ohio State University; UC, Davis
- M **University of Washington**
- N **Sloan-Kettering**; Weill Cornell Medical College
- O **Princeton;** Weizmann
- P **University of Michigan**
- Q **Broad Institute**
- R **University of Washington**; UCSF
- S **Advanced RNA Technologies, LLC**
- T **Harvard**

Computational Analysis Groups

- U **Berkeley**; Wayne State University
- V **MIT**
- W **University of Wisconsin**
- X **Sloan-Kettering**; Broad Institute
- Y **Stanford**
- Z **UCLA**

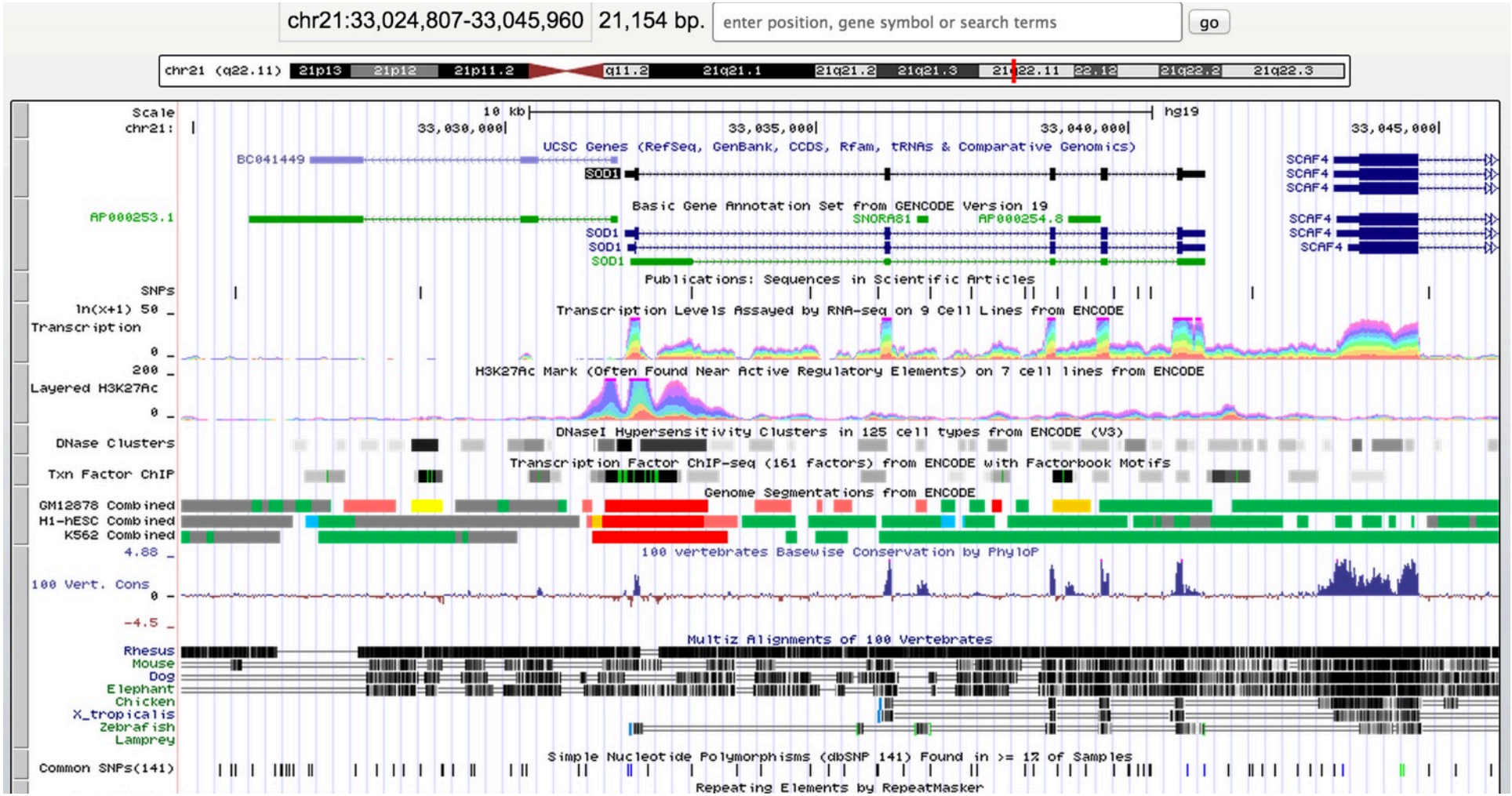
Affiliated Groups

- 1 **Wellcome Trust Sanger Institute**
- 2 **Florida State University**

ENCODE vs. NIH Roadmap Epigenomics Mapping Centers

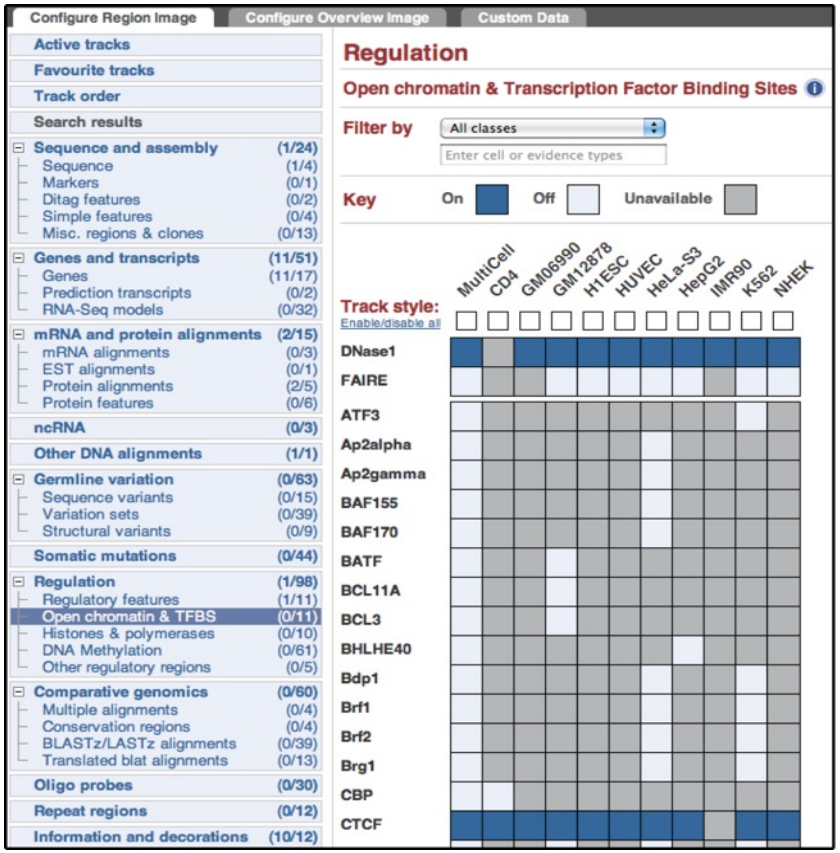
- REMCs
  - more focused on primary tissues
  - more emphasis on "epigenome maps" i.e., same marks on each cell type
  - more emphasis on DNA methylation
- ENCODE
  - more in depth transcriptome analysis
  - map transcription factor and RNA binding protein binding sites
  - genome annotation

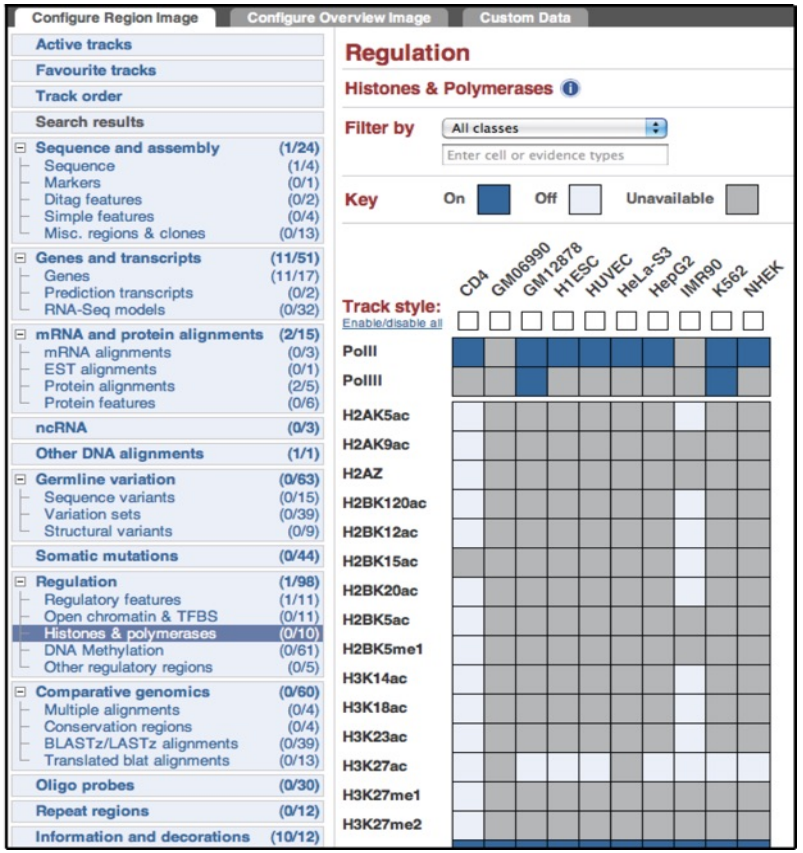
Access to ENCODE through UCSC https://genome.ucsc.edu/ENCODE/



Access to ENCODE through Ensembl http://www.ensembl.org/info/website/tutorials/

- *Use of cell type specific functional datasets in regulatory build*
  - *Matrix configuration*



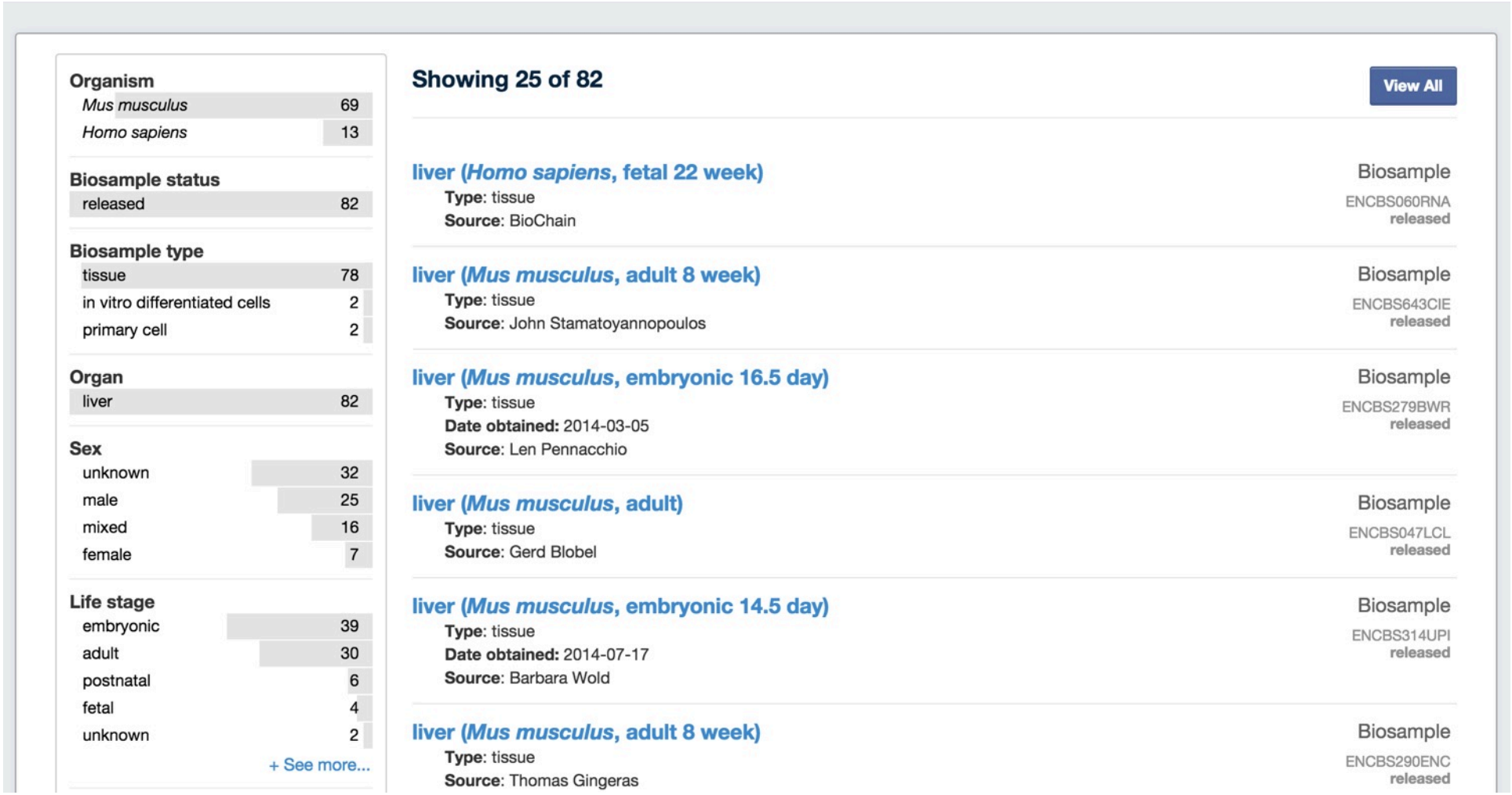


EMBL-EBI

wellcome trust sanger institute

Biodalliance – Javascript browser http://www.biodalliance.org/

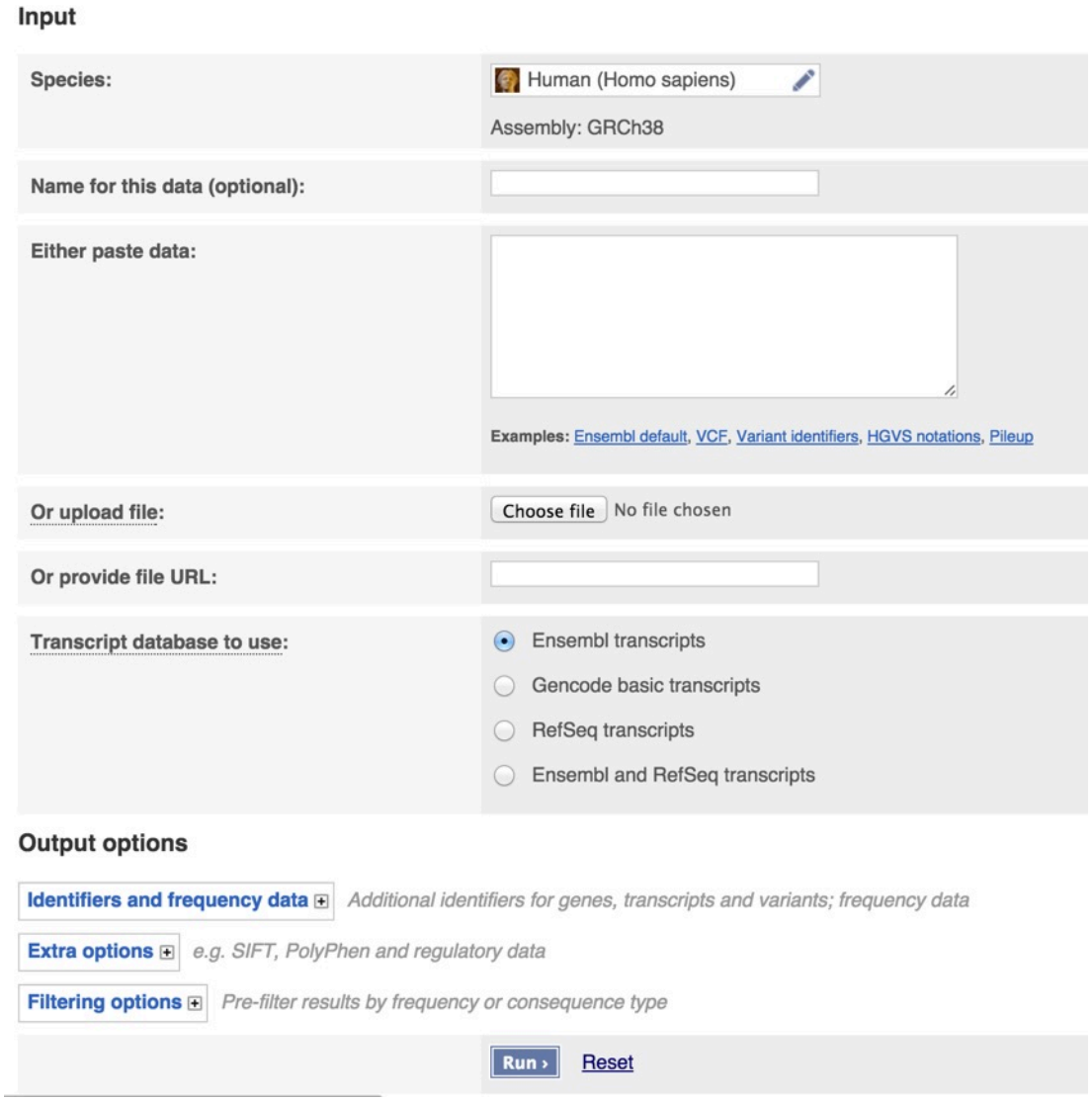
Access to ENCODE3 data
https://www.encodeproject.org/search/



Example tools that integrate data:

- Ensembl VEP (Variant Effect Predictor)
- RegulomeDB
- HaploReg

Variation annotation- Variant Effect Predictor



What can Ensembl tell me about my allele change?

All species

- Ensembl: 50+ species
- Ensembl genomes: 300+

Support for multiple file formats:

- VCF, Pileup, HGVS, rsIDs

Filter input by frequency data

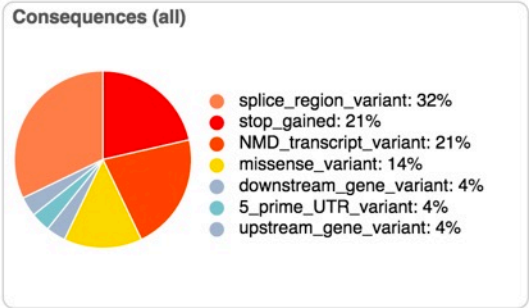
Variation annotation – VEP output

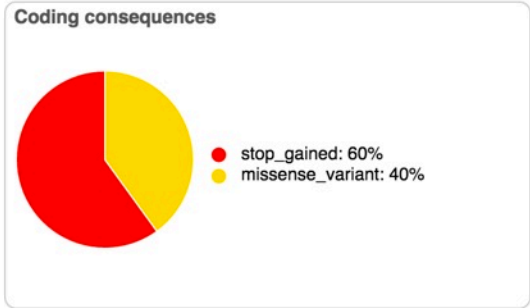
Variant Effect Predictor results

Job details

Summary statistics:

Category	Count
Variants processed	3
Variants remaining after filtering	3
Novel / existing variants	0 (0.0%) / 3 (100.0%)
Overlapped genes	4
Overlapped transcripts	12
Overlapped regulatory features	-





Results preview

Navigation — Showing 13 results for variants 1-3 of 3 | Show 1 All |

Filters — Uploaded variation | is | defined | Add

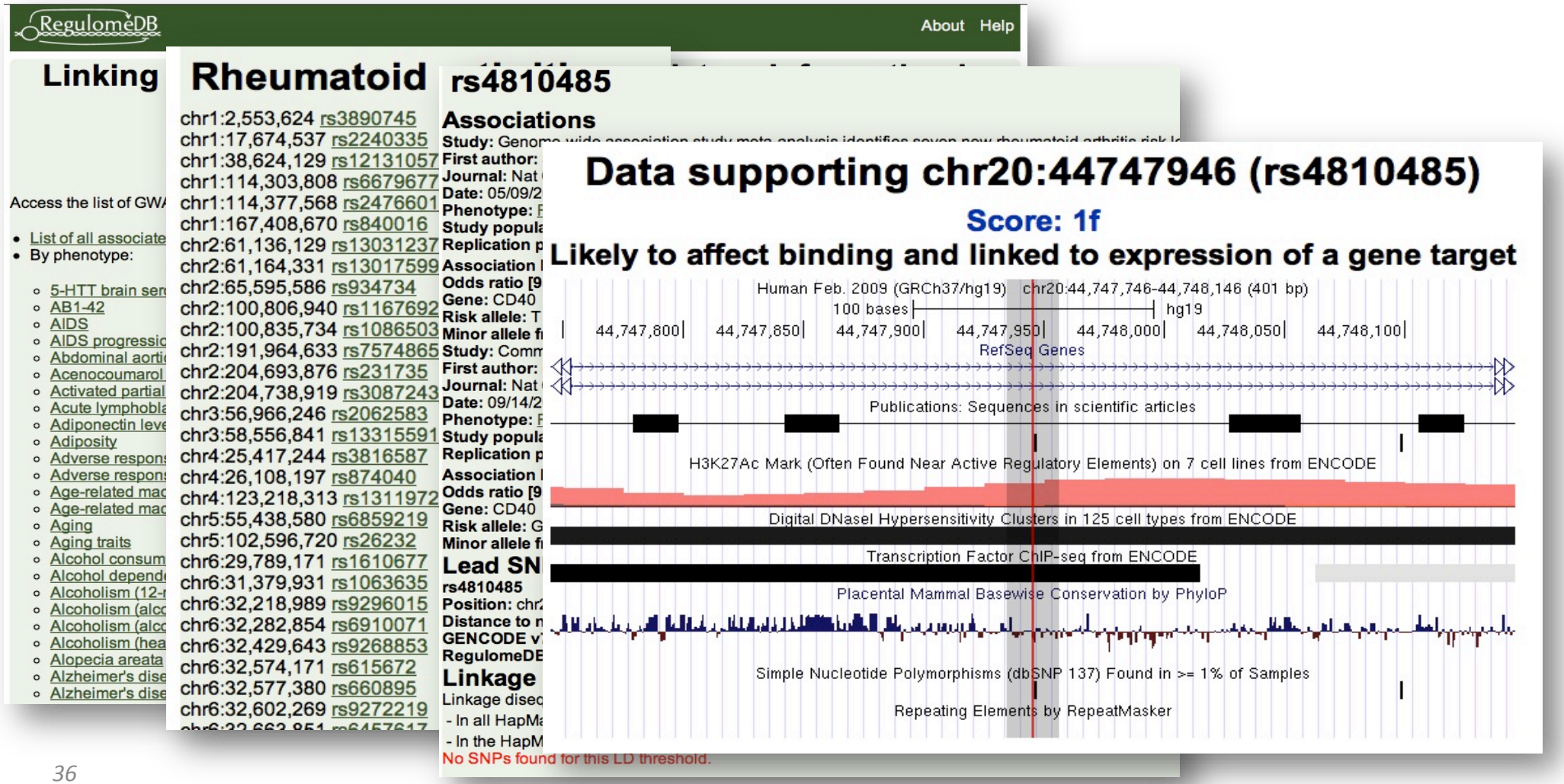
Download — All VCF VEP TXT; BioMart Variants Genes

Show/hide columns

Uploaded variation	Location	Allele	Gene	Feature	Feature type	Consequence	cDNA position	CDS position	Protein position	Amino acids	Codons	Existing variation	Distance to transcript	Feature strand	Symbol
rs144678492	1:230704270	A	ENSG00000135744	ENST00000366667	Transcript	stop_gained	1407	1192	398	Q/*	CAG/TAG	rs144678492	-	-1	AGT
rs699	1:230710048	G	ENSG00000135744	ENST00000366667	Transcript	missense_variant	1018	803	268	M/T	ATG/ACG	rs699, CM920010, COSM425562	-	-1	AGT
rs699	1:230710048	G	ENSG00000244137	ENST00000412344	Transcript	downstream_gene_variant	-	-	-	-	-	rs699, CM920010, COSM425562	650	-1	RP11-99J16__A.2
COSM354157	20:31898474	T	ENSG00000199497	ENST00000362627	Transcript	upstream_gene_variant	-	-	-	-	-	COSM354157, COSM354156	2780	1	RNU1-94P
COSM354157	20:31898474	T	ENSG00000131044	ENST00000375938	Transcript	stop_gained, splice_region_variant	368	115	39	E/*	GAG/TAG	COSM354157, COSM354156	-	1	TTLL9

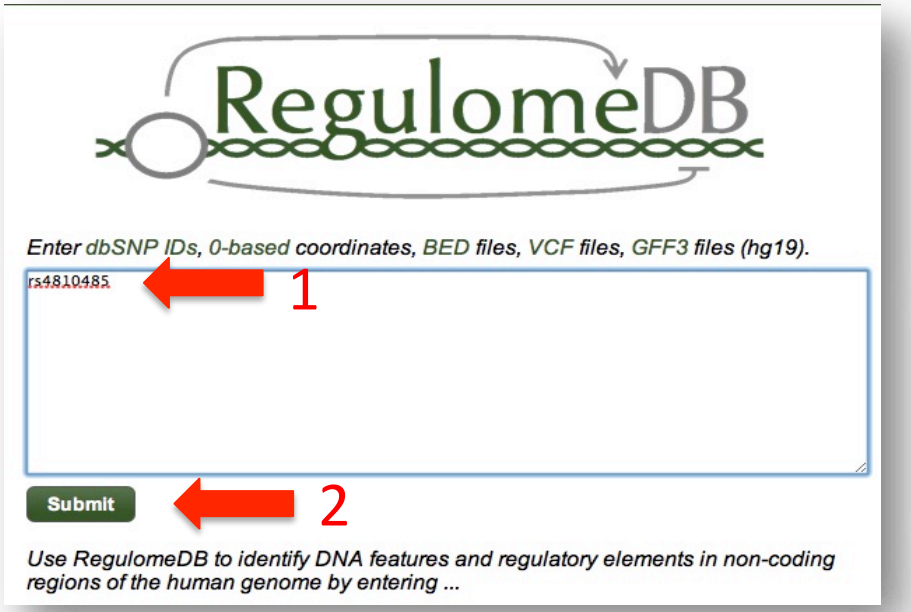
RegulomeDB Disease Database

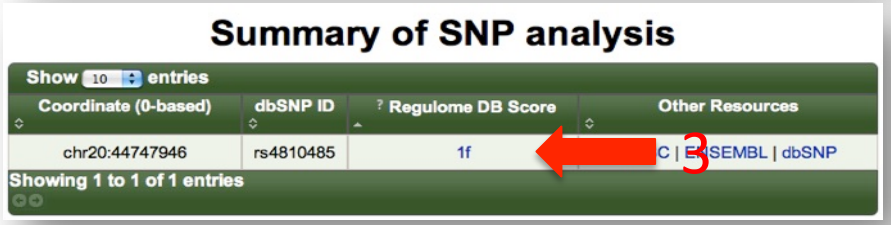
http://regulome.stanford.edu/GWAS; Cherry, Snyder, Genome Research 22-1748,2012

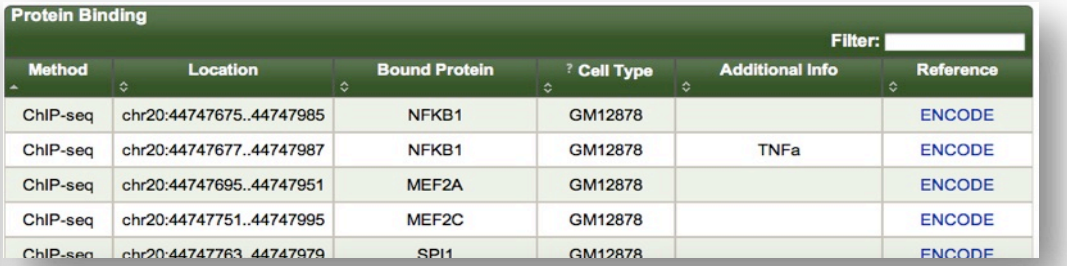


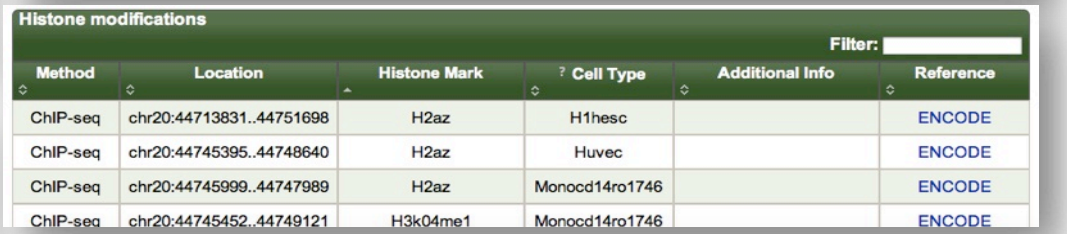
ENCODE Data From RegulomeDB

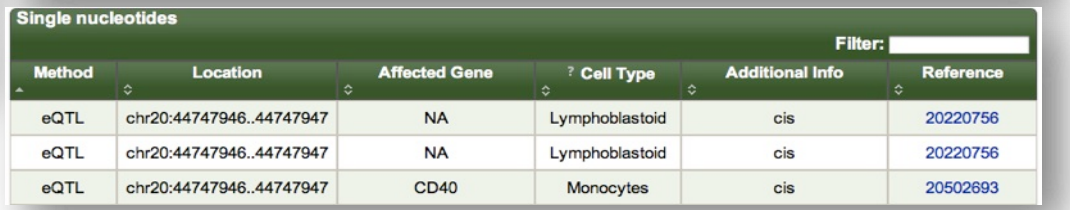
http://regulome.stanford.edu/ ; Cherry, Snyder, Genome Research 22-1790,2012







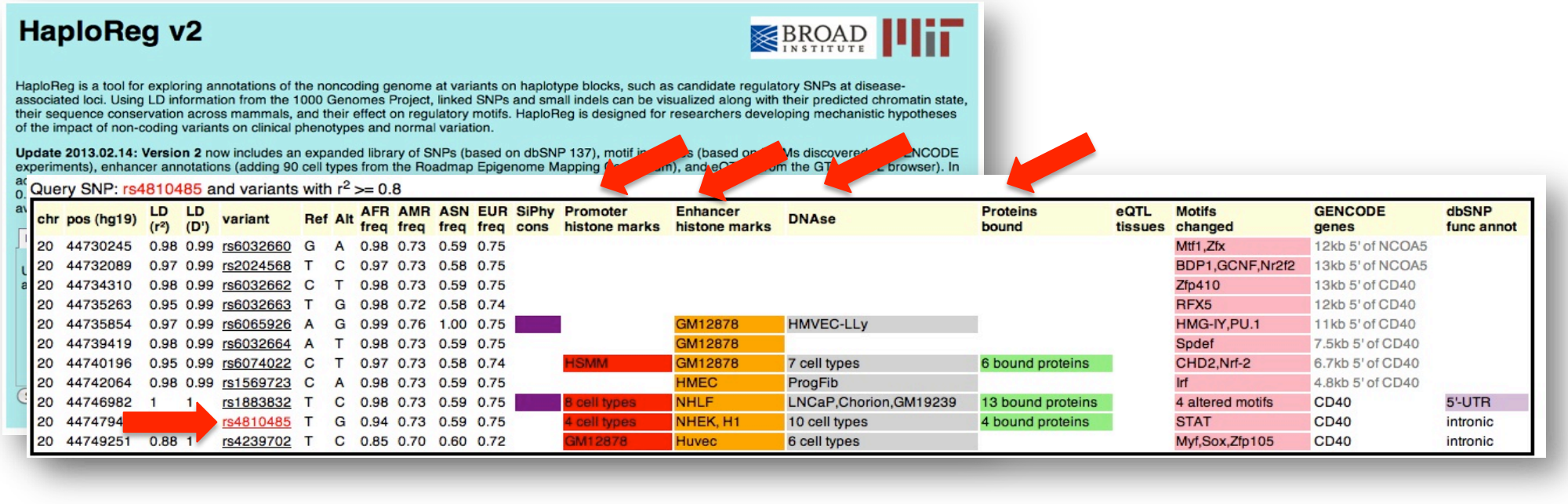




ENCODE/Epigenomics Data From HaploReg v2

www.broadinstitute.org/mammals/haploreg/

Ward and Kellis, Nucleic Acids Research 40-D930, 2011



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 i… (based o… Ms discovered… ENCODE experiments), enhancer annotations (adding 90 cell types from the Roadmap Epigenome Mapping … m), and eQ… om the GT… browser). In a… 0. a…

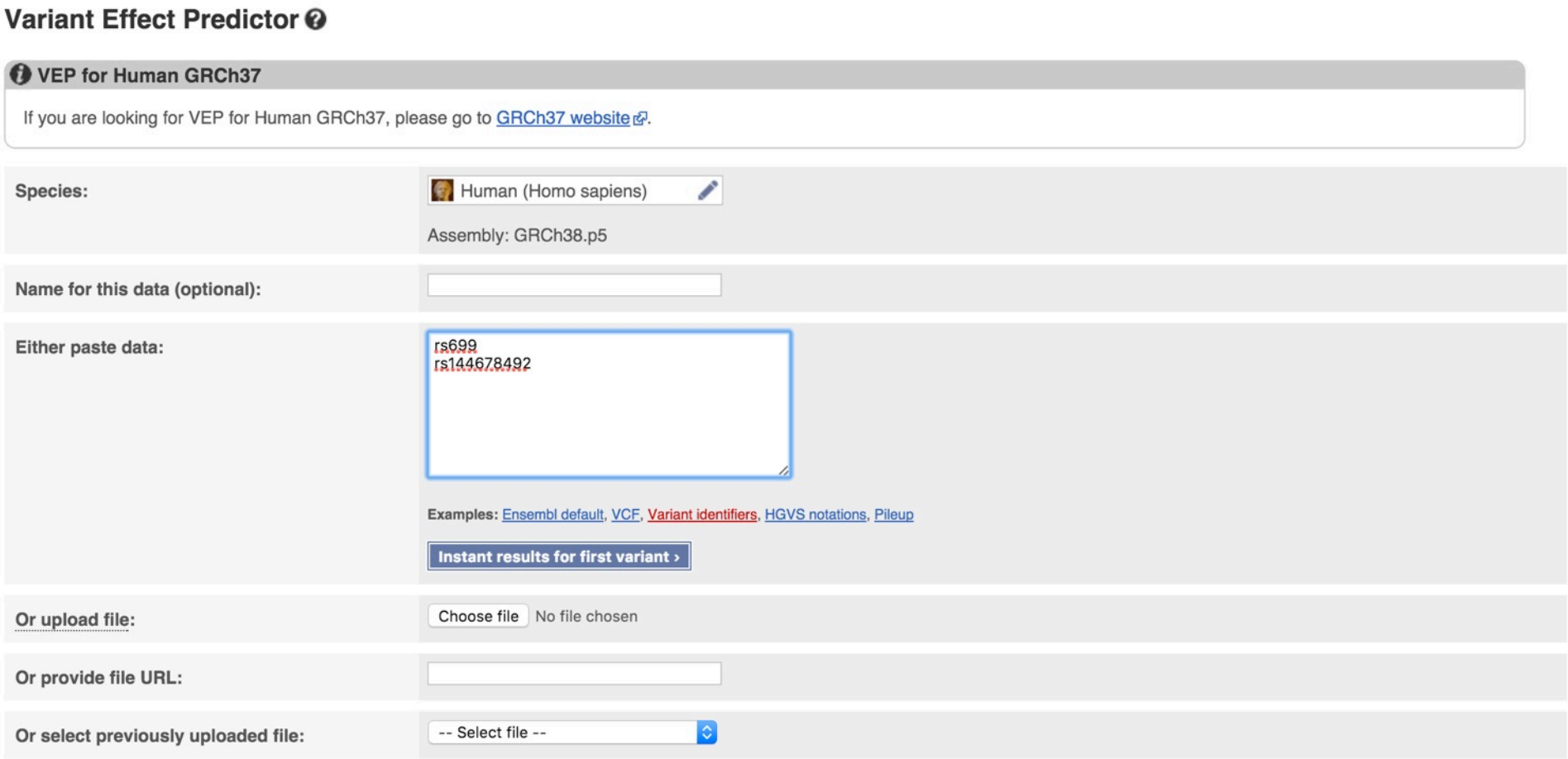
Query SNP: rs4810485 and variants with r^2 >= 0.8

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
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	4474794…			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

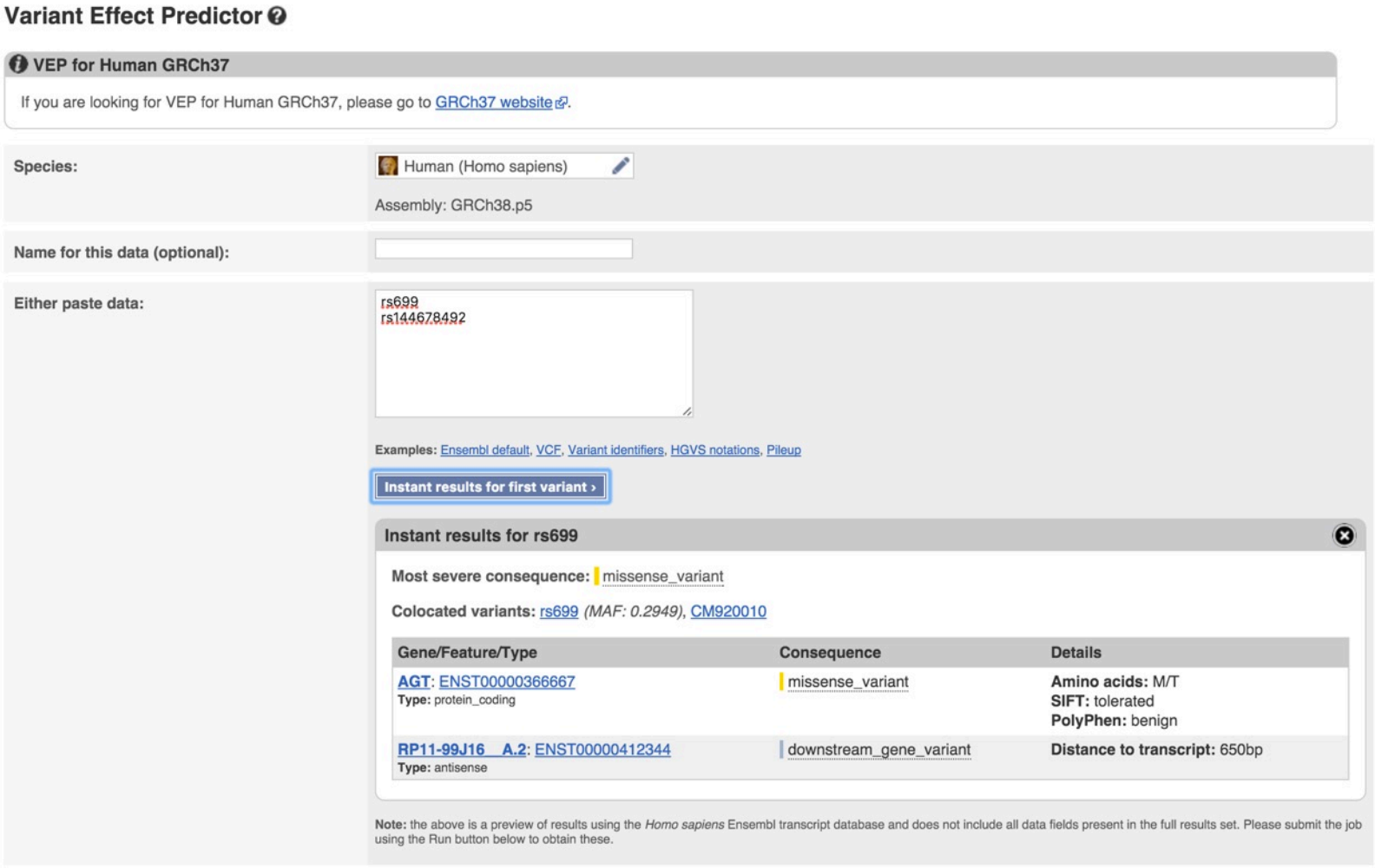
Example tools for variant rs699:

- Ensembl VEP (Variant Effect Predictor)
- RegulomeDB
- HaploReg

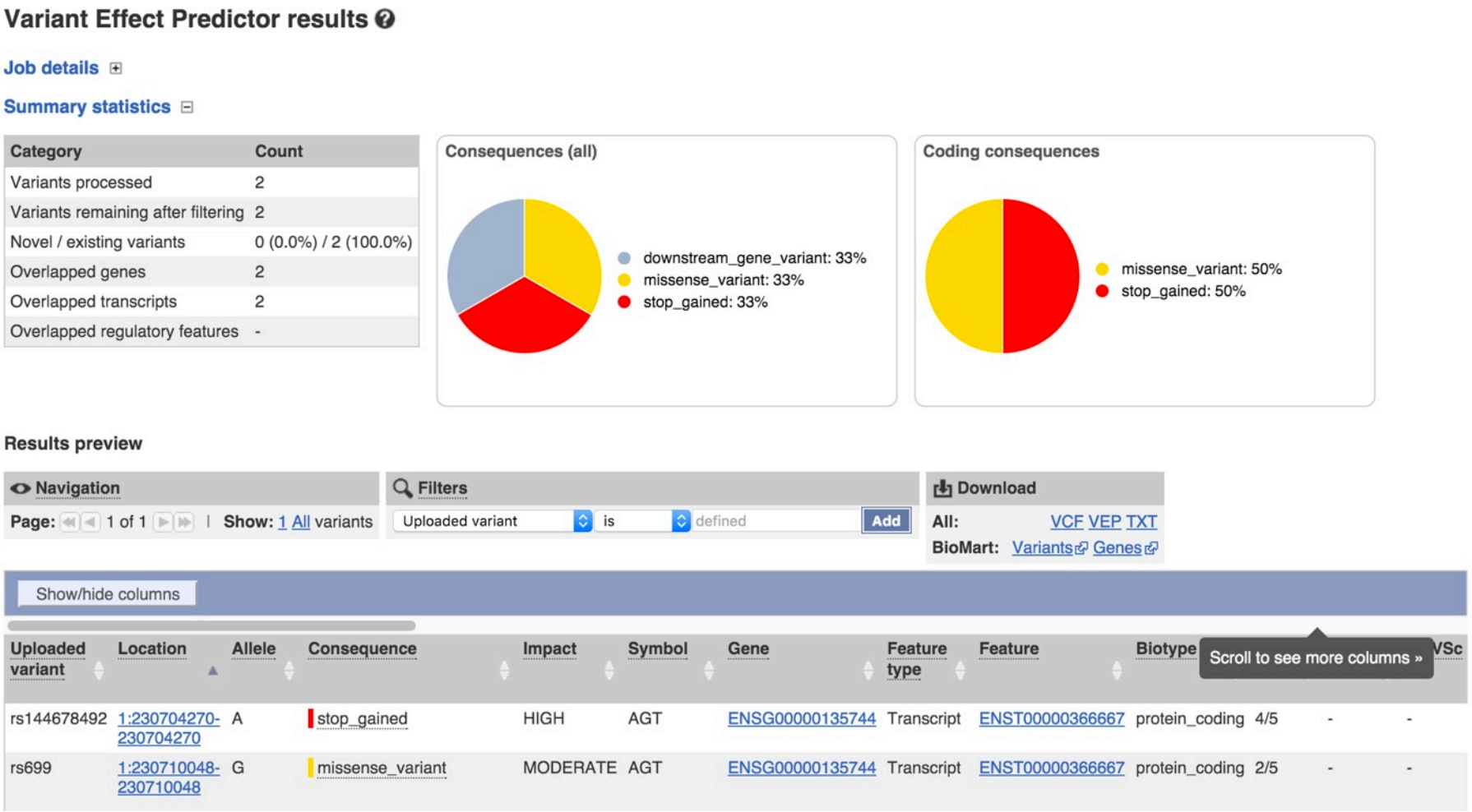
Ensembl VEP
http://www.ensembl.org/Tools/VEP



Ensembl VEP
http://www.ensembl.org/Tools/VEP



Ensembl VEP
http://www.ensembl.org/Tools/VEP



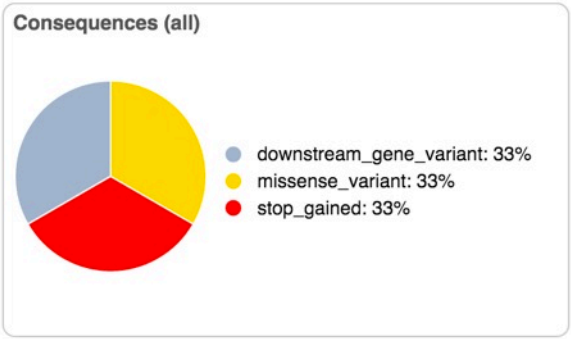
Ensembl VEP
http://www.ensembl.org/Tools/VEP

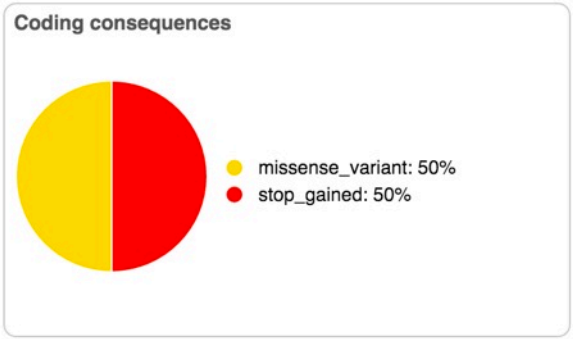
Variant Effect Predictor results

Job details ⊞

Summary statistics ⊟

Category	Count
Variants processed	2
Variants remaining after filtering	2
Novel / existing variants	0 (0.0%) / 2 (100.0%)
Overlapped genes	2
Overlapped transcripts	2
Overlapped regulatory features	-





Results preview

Navigation — Page: 1 of 1 | Show: 1 All variants

Filters — Uploaded variant | is | defined | Add

Download — All: VCF VEP TXT; BioMart: Variants Genes

Show/hide columns

Uploaded variant	Location	Allele	Consequence	Impact	Symbol	Gene	Feature type	Feature	Biotype	Exon	Intron	HGVSc
rs144678492	1:230704270-230704270	A	stop_gained	HIGH	AGT	ENSG00000135744	Transcript	ENST00000366667	protein_coding	4/5	-	-
rs699	1:230710048-230710048	G	missense_variant	MODERATE	AGT	ENSG00000135744	Transcript	ENST00000366667	protein_coding	2/5	-	-

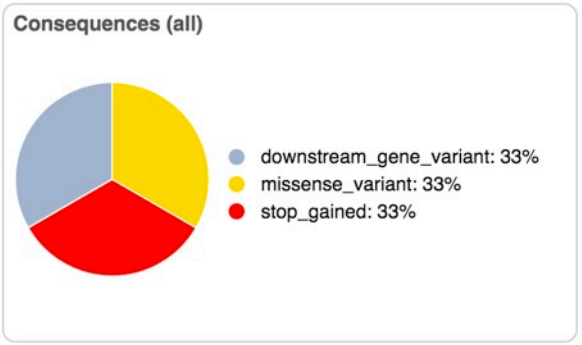
Ensembl VEP
http://www.ensembl.org/Tools/VEP

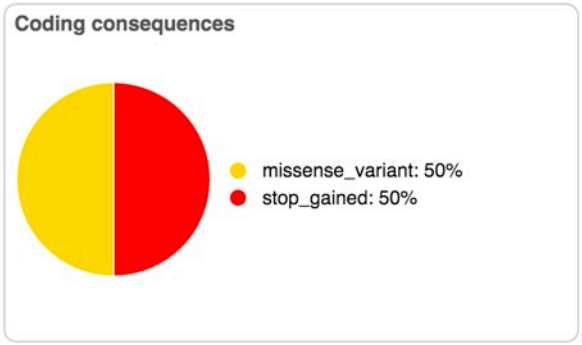
Variant Effect Predictor results

Job details

Summary statistics

Category	Count
Variants processed	2
Variants remaining after filtering	2
Novel / existing variants	0 (0.0%) / 2 (100.0%)
Overlapped genes	2
Overlapped transcripts	2
Overlapped regulatory features	-





Results preview

Navigation — Page: 1 of 1 | Show: 1 All variants

Filters — Uploaded variant | is | defined | Add

Download — All: VCF VEP TXT; BioMart: Variants Genes

Show/hide columns

HGVSp	cDNA position	CDS position	Protein position	Amino acids	Codons	Existing variant	Distance to transcript	Feature strand	Symbol source	HGNC ID	Transcript support level	APPRIS	SIFT	Pol
-	1407	1192	398	Q/*	CAG/TAG	rs144678492	-	-1	HGNC	HGNC:333	1	P1	-	-
-	1018	803	268	M/T	ATG/ACG	rs699, CM920010	-	-1	HGNC	HGNC:333	1	P1	1	

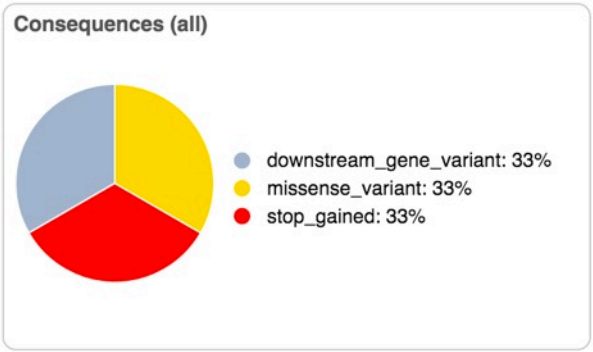
Ensembl VEP
http://www.ensembl.org/Tools/VEP

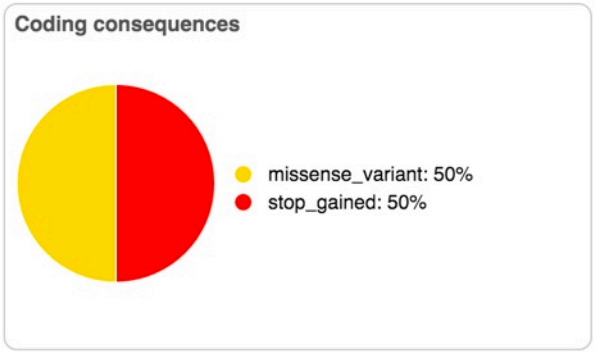
Variant Effect Predictor results

Job details

Summary statistics

Category	Count
Variants processed	2
Variants remaining after filtering	2
Novel / existing variants	0 (0.0%) / 2 (100.0%)
Overlapped genes	2
Overlapped transcripts	2
Overlapped regulatory features	-





Results preview

Navigation | Filters | Download

Page: 1 of 1 | Show: 1 All variants

Uploaded variant is defined Add

All: VCF VEP TXT

BioMart: Variants Genes

Show/hide columns

SIFT	PolyPhen	GMAF	AFR MAF	AMR MAF	EAS MAF	EUR MAF	SAS MAF	AA MAF	EA MAF	ExAC MAF	ExAC Adj MAF	ExAC AFR MAF	ExAC AMR MAF	ExAC EAS MAF	ExAC FIN MAF	ExAC NFE MAF	ExAC OTH MAF	ExA SAS MA
-	-	-	-	-	-	-	-	A:0	A:0.0001	A:8.236e-06	A:8.25e-06	A:0	A:0	A:0	A:0	A:1.5e-05	A:0	A:0
1	0	A:0.2949	G:0.9032	G:0.6354	G:0.8532	G:0.4115	G:0.6360	-	-	G:0.548	G:0.5484	G:0.8447	G:0.7416	G:0.8378	G:0.4356	G:0.4243	G:0.5187	G:0

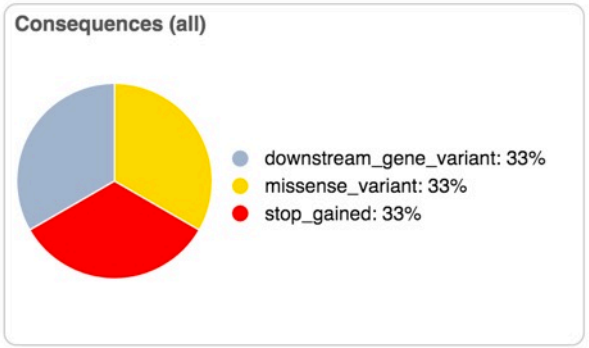
Ensembl VEP
http://www.ensembl.org/Tools/VEP

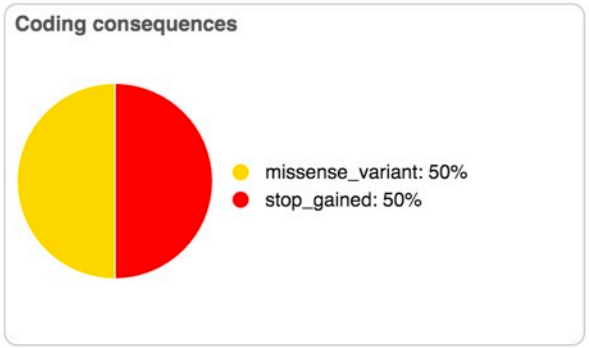
Variant Effect Predictor results

Job details

Summary statistics

Category	Count
Variants processed	2
Variants remaining after filtering	2
Novel / existing variants	0 (0.0%) / 2 (100.0%)
Overlapped genes	2
Overlapped transcripts	2
Overlapped regulatory features	-





Results preview

Navigation — Page: 1 of 1 | Show: 1 All variants

Filters — Uploaded variant | is | defined | Add

Download — All: VCF VEP TXT; BioMart: Variants Genes

Show/hide columns

	ExAC Adj MAF	ExAC AFR MAF	ExAC AMR MAF	ExAC EAS MAF	ExAC FIN MAF	ExAC NFE MAF	ExAC OTH MAF	ExAC SAS MAF	Clinical significance	Somatic status	Phenotype or disease	Pubmed	Motif name	Motif position	High info position	Motif score change
36e-	A:8.25e-06	A:0	A:0	A:0	A:0	A:1.5e-05	A:0	A:0	-	-	-	-	-	-	-	-
48	G:0.5484	G:0.8447	G:0.7416	G:0.8378	G:0.4356	G:0.4243	G:0.5187	G:0.6229	risk_factor	-	1, 1	18513389, 19131662,	-	-	-	-

RegulomeDB
http://regulomedb.org/

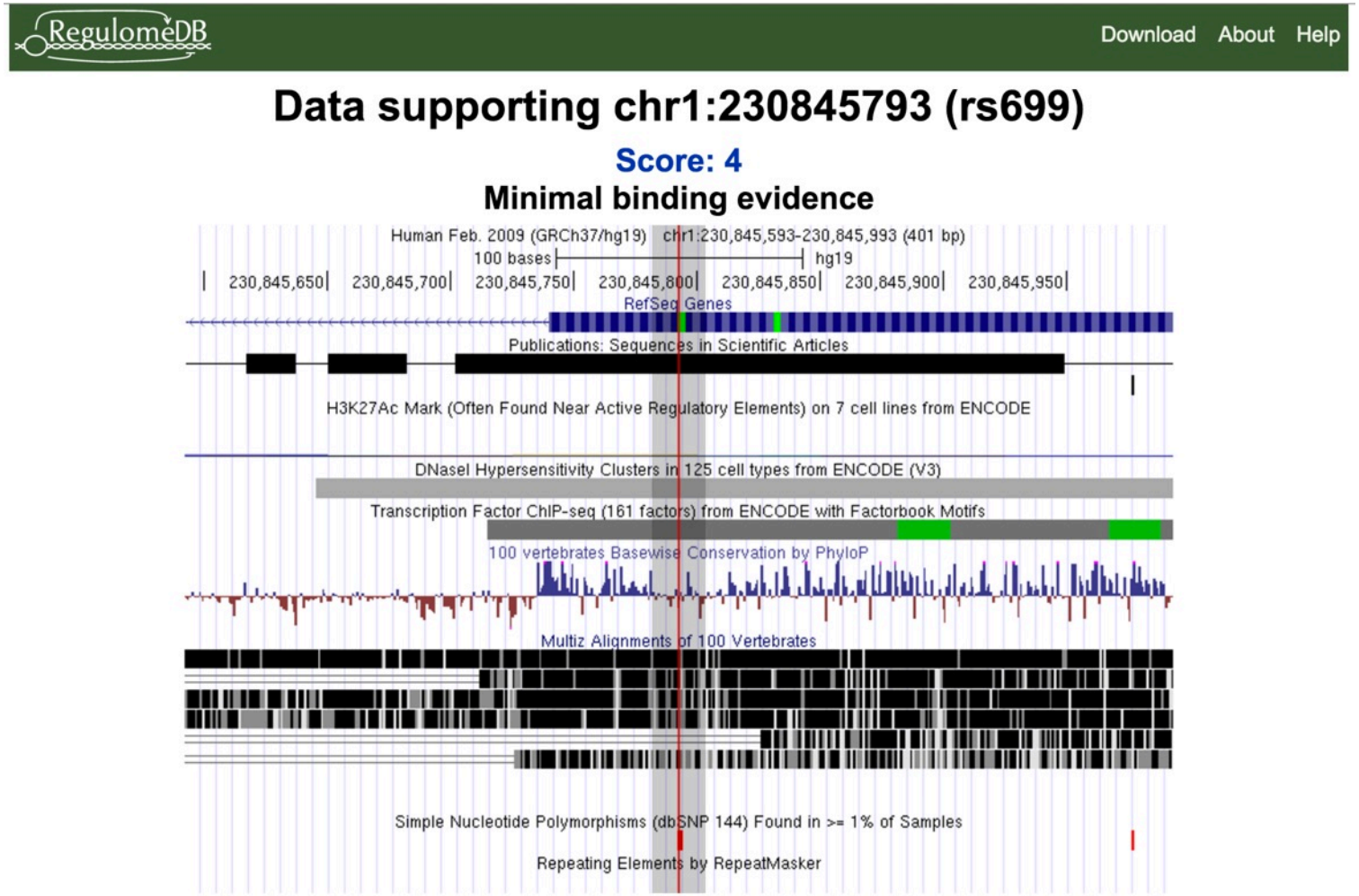
RegulomeDB has been updated to Version 1.1. This includes bringing our database up-to-date with current ENCODE releases: Xie et al. (2013) and Boyle et al. (2014). We have also added Chromatin States from the Roadmap Epigenome Consortium (unpublished) as well as updates to DNase footprinting, PWMs, and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

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

rs699
rs144678492

Submit

RegulomeDB
http://regulomedb.org/



RegulomeDB
http://regulomedb.org/

Protein Binding Filter:

Method	Location	Bound Protein	? Cell Type	Additional Info	Reference
ChIP-seq	chr1:230845715..230846219	REST	A549	02pct	ENCODE
ChIP-seq	chr1:230845750..230845994	EGR1	K562		ENCODE
ChIP-seq	chr1:230845779..230846235	SIN3A	HepG2		ENCODE

Chromatin structure Filter:

Method	Location	? Cell Type	Additional Info	Reference
DNase-seq	chr1:230845645..230845795	Fibrobl		ENCODE
DNase-seq	chr1:230845645..230845795	8988t		ENCODE
DNase-seq	chr1:230845647..230846135	K562		ENCODE
DNase-seq	chr1:230845665..230845855	Huh7		ENCODE
DNase-seq	chr1:230845697..230845965	T47d	Est10nm30m	ENCODE
DNase-seq	chr1:230845697..230845965	T47d		ENCODE
DNase-seq	chr1:230845705..230845855	Huh75		ENCODE
DNase-seq	chr1:230845738..230845918	Hmec		ENCODE
DNase-seq	chr1:230845740..230845890	K562	Znfa41c6	ENCODE
DNase-seq	chr1:230845750..230845797	Cerebellumoc		ENCODE
DNase-seq	chr1:230845785..230846165	Medullo		ENCODE

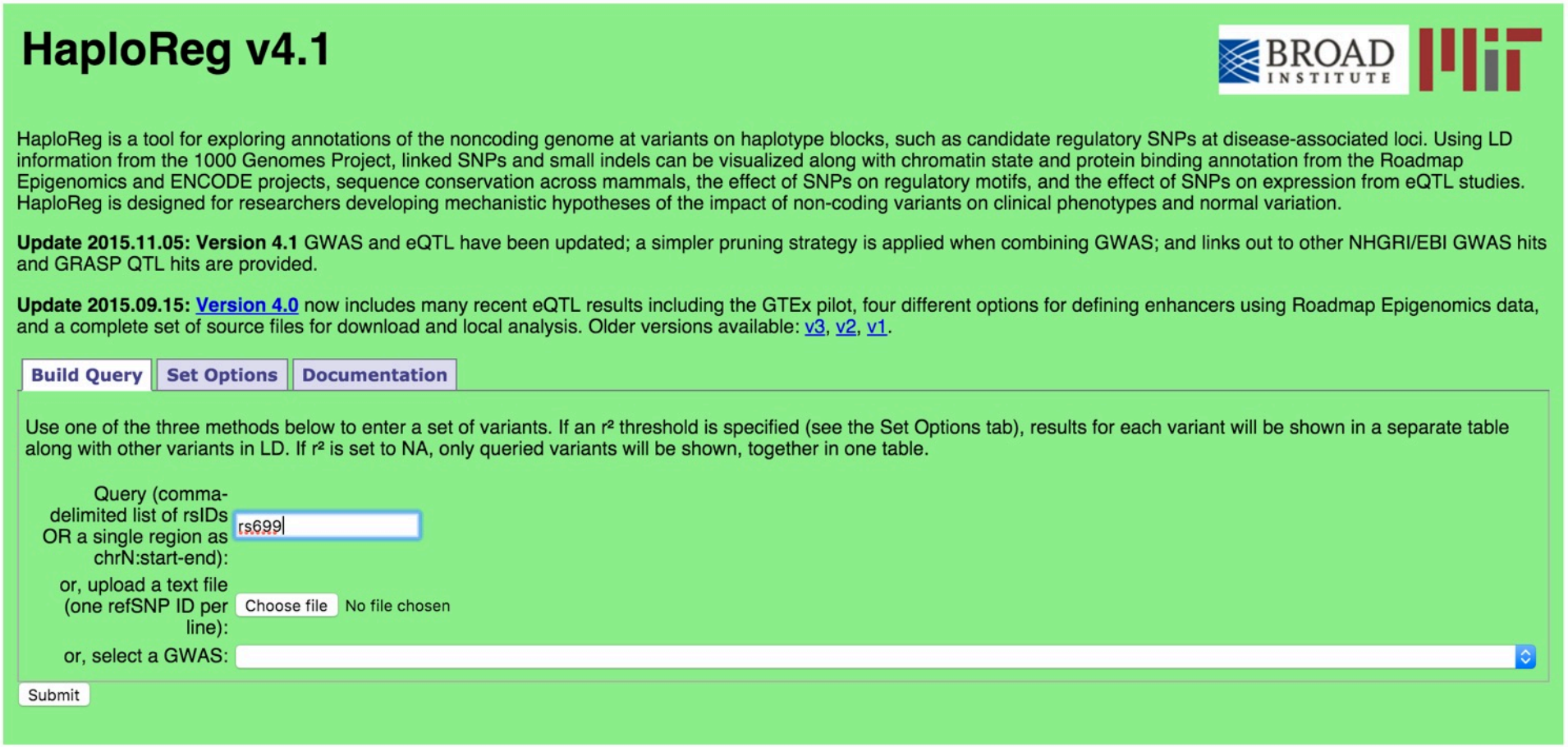
RegulomeDB
http://regulomedb.org/

Histone modifications Filter:

Method	Location	Chromatin State	Tissue Group	Tissue	Reference
ChromHMM	chr1:230808200..230887000	Quiescent/Low	ES-deriv	H1 BMP4 Derived Mesendoderm Cultured Cells	REMC
ChromHMM	chr1:230827600..230849800	Quiescent/Low	ESC	ES-WA7 Cell Line	REMC
ChromHMM	chr1:230829800..230868000	Quiescent/Low	Blood & T-cell	Primary T CD8+ memory cells from peripheral blood	REMC
ChromHMM	chr1:230829000..230846200	Weak transcription	Muscle	Psoas Muscle	REMC
ChromHMM	chr1:230832600..230883400	Weak Repressed PolyComb	Epithelial	Foreskin Keratinocyte Primary Cells skin02	REMC
ChromHMM	chr1:230838400..230884800	Quiescent/Low	Other	Fetal Kidney	REMC
ChromHMM	chr1:230845000..230882800	Quiescent/Low	Other	Placenta Amnion	REMC
ChromHMM	chr1:230830400..230866200	Quiescent/Low	Blood & T-cell	Primary T CD8+ naive cells from peripheral blood	REMC
ChromHMM	chr1:230830400..230860400	Quiescent/Low	Blood & T-cell	Primary T helper memory cells from peripheral blood 1	REMC
ChromHMM	chr1:230832200..230849400	Quiescent/Low	Sm. Muscle	Colon Smooth Muscle	REMC
ChromHMM	chr1:230832400..230849400	Quiescent/Low	Epithelial	Foreskin Melanocyte Primary Cells skin03	REMC
ChromHMM	chr1:230831000..230847600	Quiescent/Low	Blood & T-cell	Primary T helper cells PMA-I stimulated	REMC
ChromHMM	chr1:230831200..230849600	Quiescent/Low	ESC	HUES6 Cell Line	REMC
ChromHMM	chr1:230831400..230913200	Quiescent/Low	Blood & T-cell	Primary T helper memory cells from peripheral blood 2	REMC

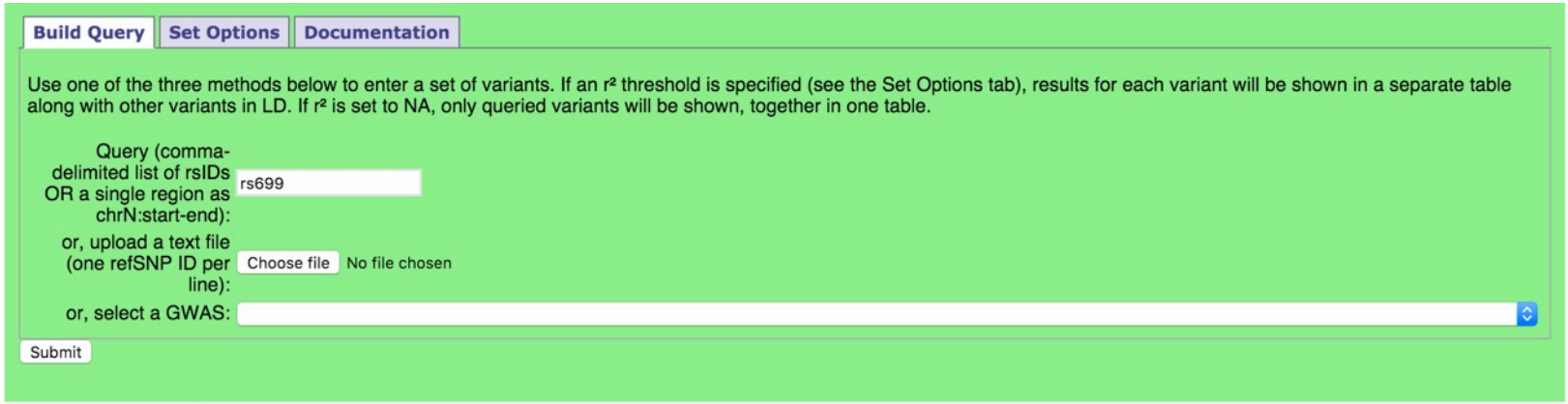
HaploReg v4.1

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



HaploReg v4.1

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



Query SNP: rs699 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	Motifs changed	NHGRI/EBI GWAS hits	GRASP QTL hits	Selected eQTL hits	GENCODE genes	dbSNP func annot
1	230694922	0.86	-0.94	rs2493126	G	C	0.12	0.36	0.13	0.60			6 tissues			AP-2rep,Pou2f2,Zbtb3			9 hits	939bp 3' of COG2	
1	230696685	0.84	-0.93	rs2493128	G	T	0.27	0.38	0.15	0.60			4 tissues			4 altered motifs			7 hits	2.7kb 3' of COG2	
1	230697920	0.84	-0.92	rs2006765	G	A	0.11	0.35	0.16	0.59			LIV, MUS			9 altered motifs			9 hits	3.9kb 3' of COG2	
1	230701298	0.87	-0.96	rs943580	G	A	0.12	0.36	0.15	0.60			LIV	BRN		NF-Y		3 hits	17 hits	1.2kb 3' of AGT	
1	230708527	0.98	0.99	rs2493133	T	C	0.84	0.62	0.84	0.41		BRN	11 tissues	GI,GI		7 altered motifs		3 hits	10 hits	AGT	intronic
1	230708564	0.98	0.99	rs2478543	T	C	0.84	0.63	0.84	0.41		BRN	11 tissues	GI,GI,LIV	POL2	IRC900814,SREBP		3 hits	10 hits	AGT	intronic
1	230709026	1	1	rs2478539	G	T	0.83	0.63	0.84	0.41		SKIN, BRN, GI	12 tissues	5 tissues		Spz1		3 hits	8 hits	AGT	intronic
1	230709246	0.8	1	rs6687360	C	T	0.57	0.45	0.69	0.36		BRN	7 tissues	GI,GI,LIV		HNF1			10 hits	AGT	intronic
1	230710048	1	1	rs699	A	G	0.87	0.64	0.84	0.41		BRN	GI, MUS, HRT	LIV	EGR1			5 hits	10 hits	AGT	missense
1	230713422	0.8	1	rs7539013	C	T	0.54	0.45	0.70	0.36		6 tissues	7 tissues						9 hits	AGT	intronic
1	230713444	0.8	1	rs7539020	C	T	0.54	0.45	0.70	0.36		6 tissues	7 tissues			SP1			9 hits	AGT	intronic
1	230713613	1	1	rs2493134	T	C	0.86	0.64	0.84	0.41		5 tissues	11 tissues	MUS		ERalpha-a		4 hits	9 hits	AGT	intronic
1	230714053	0.83	0.92	rs2148582	A	G	0.85	0.64	0.84	0.41		9 tissues	14 tissues	13 tissues	HEY1,POL2,TBP	p53		4 hits	11 hits	AGT	intronic
1	230714126	0.93	0.97	rs5051	C	T	0.88	0.64	0.83	0.41		9 tissues	17 tissues	23 tissues	4 bound proteins	5 altered motifs		1 hit	9 hits	AGT	5'-UTR
1	230715682	0.98	1	rs2493135	C	G	0.87	0.64	0.84	0.40			LIV, GI	ESC,LNG	4 bound proteins	Lmo2-complex,TCF12			8 hits	RP11-99J16__A.2	
1	230715790	0.99	0.99	rs2493136	C	T	0.87	0.64	0.84	0.41			LIV, GI	LIV	4 bound proteins	Hbp1,Pax-8			8 hits	RP11-99J16__A.2	

Summary

- Reference gene set still being expanded
  - Still ~20K protein coding genes, but many more non-coding genes, many more transcripts.
- Reference annotations of cellular state
  - Epigenetic data summarizes cellular state and is informative for genetic analysis

ENCODE Consortium acknowledgements

Brad Bernstein (Eric Lander, Manolis Kellis, Tony Kouzarides)

Ewan Birney (Jim Kent, Mark Gerstein, Bill Noble, Peter Bickel, Ross Hardison, Zhiping Weng)

Greg Crawford (Ewan Birney, Jason Lieb, Terry Furey, Vishy Iyer)

Jim Kent (David Haussler, Kate Rosenbloom)

John Stamatoyannopoulos (Evan Eichler, George Stamatoyannopoulos, Job Dekker, Maynard Olson, Michael Dorschner, Patrick Navas, Phil Green)

Mike Snyder (Kevin Struhl, Mark Gerstein, Peggy Farnham, Sherman Weissman)

Rick Myers (Barbara Wold)

Scott Tenenbaum (Luiz Penalva)

Tim Hubbard (Alexandre Reymond, Alfonso Valencia, David Haussler, Ewan Birney, Jim Kent, Manolis Kellis, Mark Gerstein, Michael Brent, Roderic Guigo)

Tom Gingeras (Alexandre Reymond, David Spector, Greg Hannon, Michael Brent, Roderic Guigo, Stylianos Antonarakis, Yijun Ruan, Yoshihide Hayashizaki)

Zhiping Weng (Nathan Trinklein, Rick Myers)

Additional ENCODE Participants: Elliott Marguiles, Eric Green, Job Dekker, Laura Elnitski, Len Pennachio, Jochen Wittbrodt

.. and many senior scientists, postdocs, students, technicians, computer scientists, statisticians and administrators in these groups

NHGRI: Elise Feingold, Mike Pazin, Peter Good

Acknowledgements

WTSI Vertebrate Genome Annotation:
Jen Harrow & Steve Searle

Havana:
Adam Frankish
Jane Loveland
Jonathan Mudge
Charles Steward
Laurens Wilming
Clara Amid
Veronika Boychenko
If Barnes
Alex Bignell
Denise Carvalho-Silva
Gloria Despacio-Reyes
Sarah Donaldson
Gabriella Frigerio
Toby Hunt
Mike Kay
Gavin Laird
David Lloyd
Gaurab Mukherjee
Jeena Rajan
Gary Saunders
Harminder Sehra
Catherine Snow
Emma Kenyon
Marie-Marte Suner
Mark Thomas

Anacode:
James Gilbert
Matthew Astley
Michael Gray
Jeremy Henty

Acedb/zmap:
Ed Griffiths
Malcolm Hinsley
Gemma Barston

Ensembl strategy Includes:
Richard Durbin
Ewan Birney
Tim Hubbard

WTSI Ensembl
Bronwen Aken
Anne Parker
Stephen Trevanion
Amonida Zadissa
Mohammad Amode
Simon Brent
Susan Fairley
Carlos Garcia Giron
Thibaut Hourlier
Rishi Nag
Harpreet Riat
Magali Ruffier
Daniel Sheppard
Bethan Yates
Simon White

ISG:
Guy Coates

CCDS:
NCBI(Refseq)
UCSC

Gencode
Jose Gonzalez
Ilektra Tapanari
Daniel Barrell
CRG
Lausanne
UCSC
WashU
MIT
Yale
CNIO

GRC
Kerstin Howe
WashU
NCBI
EBI

EBI Ensembl
Paul Flicek

Fiona Cunningham
Javier Herrero
Ian Dunham
Rhoda Kinsella
Giulietta Spudich
Andy Yates
Kathryn Beal
Yuan Chen
Stephen Fitzgerald Leo Gordon
Andy Jenkinson
Nathan Johnson
Andreas Kahari
Damian Keefe
Ian Longden
Will McLaren
Bert Overduin
Daniel Rios
Guy Slater
Michael Schuster
Albert Vilella
Steven Wilder