

CAGI slide pack guidelines

I agree to my slides being used according to the following terms:

Alan Boyle

- Authors: please include THIS slide at the beginning of your presentation to indicate that you agree with reuse of your slides in CAGI presentations, according to the terms below, and digitally sign by writing your name above.
- To disseminate the CAGI results, I encourage all CAGI participants to remix these CAGI slides for inclusion in their oral and poster presentations
- CAGI participants may use or publish any content of these slides only in compliance with the CAGI Data Use Agreement
- CAGI participants may not cite or publish any content of these slides except with the written permission of the originating author or in the primary CAGI publications
- CAGI participants must acknowledge the original authors of these slides and CAGI 2013. Include the acknowledgement banner across the bottom
- CAGI participants must include the CAGI credits slide in their presentations
- Slides with a red slash (no remix) may not be included in a presentation unless all attendees are registered CAGI participants who signed the data use agreement
- Slides with a red X over the entire slide may not be used in any circumstances
- Slides with a blue slash (embargoed) may not be used unless the embargo is explicitly lifted
- Slides without slash or X may be tweeted unless flagged with the “No tweet” icon.
- CAGI participants using these slides must explain the “No tweet” icon, which means that the slide content should not be disseminated outside the presenting conference hall. If the presentation venue permits pervasive tweeting, you may not include these slides in your talk
- Note for slide authors: we expect all CAGI participants to abide by these restrictions, and will make best effort to ensure they are followed. However, but bear in mind that we have limited means of enforcing them, and therefore the restrictions cannot be guaranteed.



Prediction of regulatory SNPs affecting transcription factor binding using functional genomics data

Alan Boyle
University of Michigan

July 6, 2018
CAGI5

the **BOYLE** **LAB**

Jessica Switzenberg

Adam Diehl

Sierra Nishizaki

Torrin McDonald

Shengcheng Dong

Ningxin Ouyang

Chris Castro

Haley Amemiya

Melissa Englund

Samuel Zhao



Cherry lab (Stanford)

Ben Hitz

Tim Dreszer

Yunhai Luo

<http://BoyleLab.org>



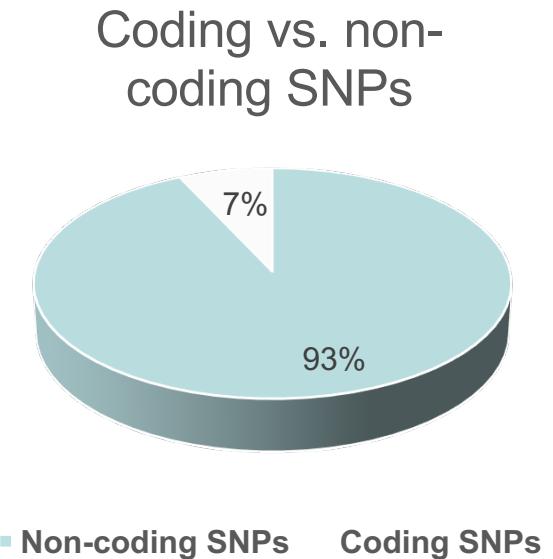
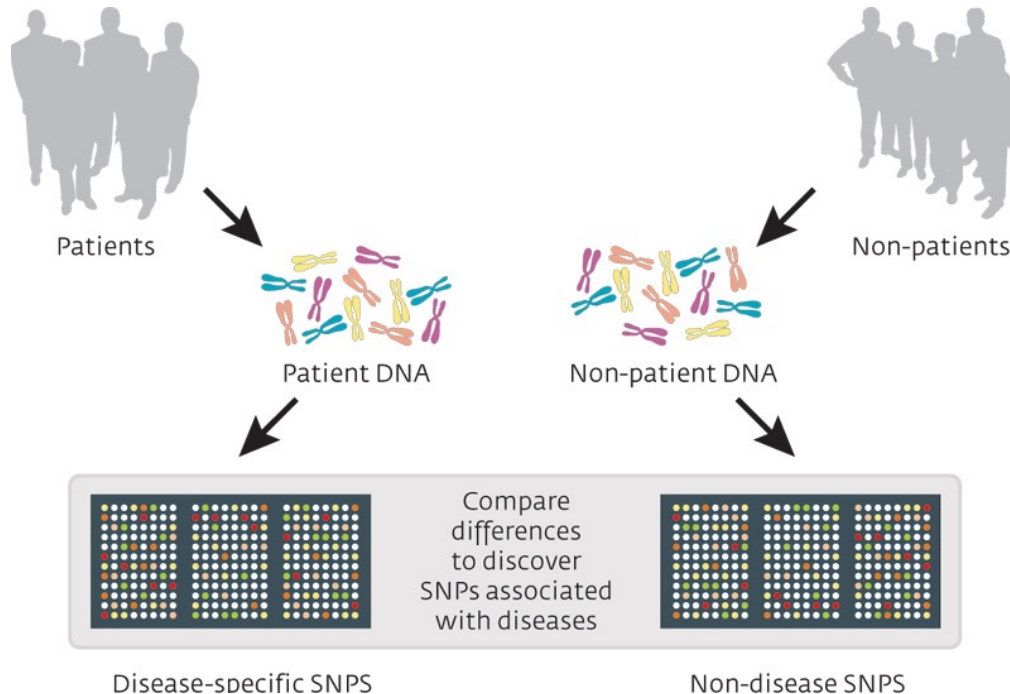
Alfred P. Sloan
FOUNDATION



**National Human
Genome Research
Institute**

Eleanor and Larry Jackier U-M/Technion and Weizmann Collaborative Research Grant

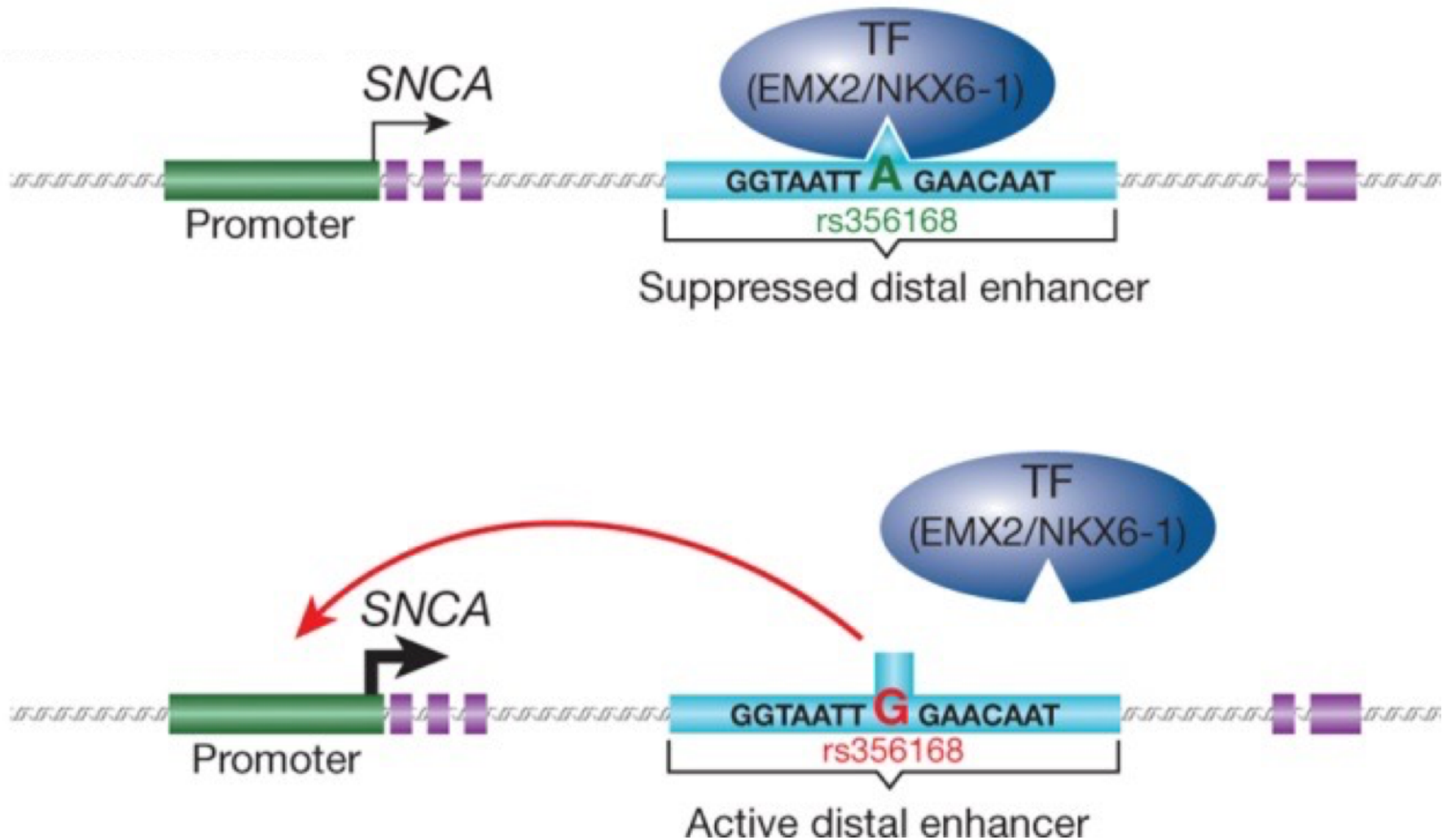
Genome wide association studies (GWAS) has identified variants associated with disease



• Challenge:

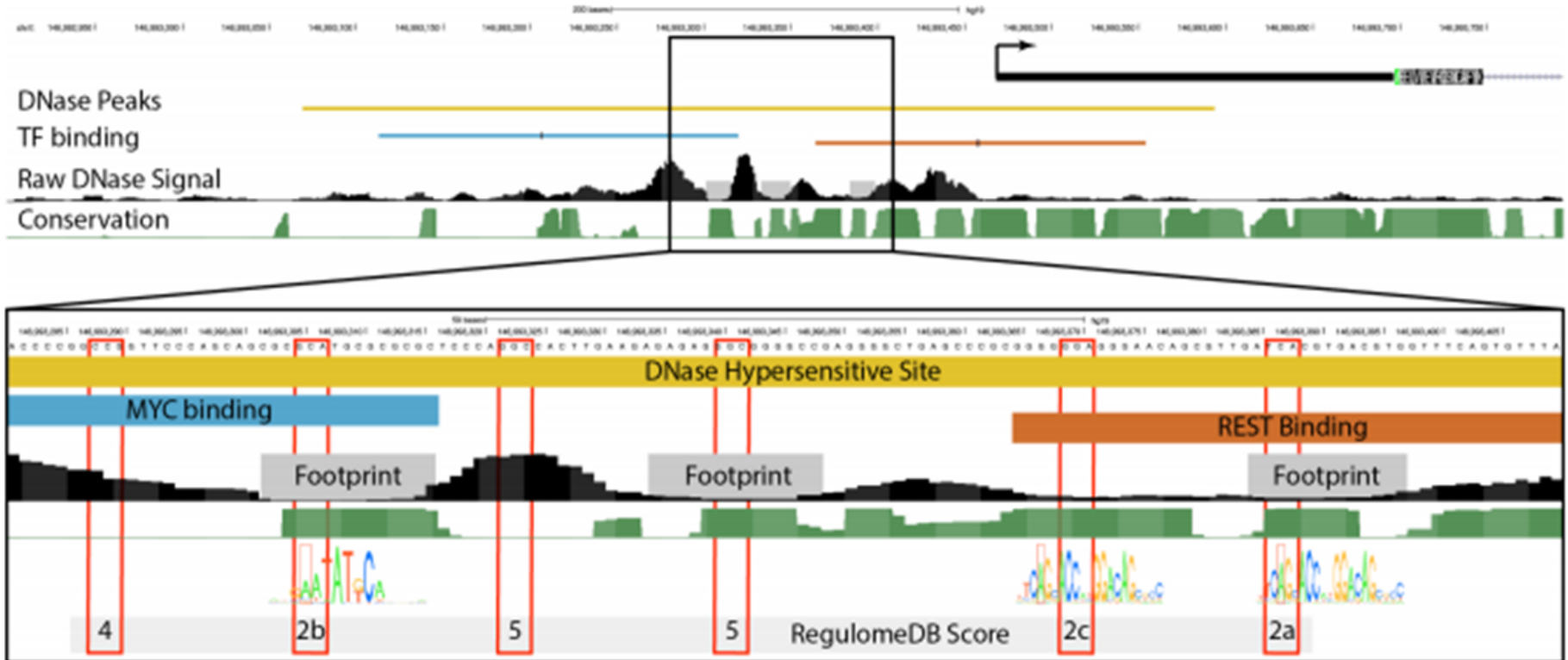
Understand the biological impact of disease-associated SNPs in non-coding regions

Disease-associated SNP affecting TF binding in non-coding region

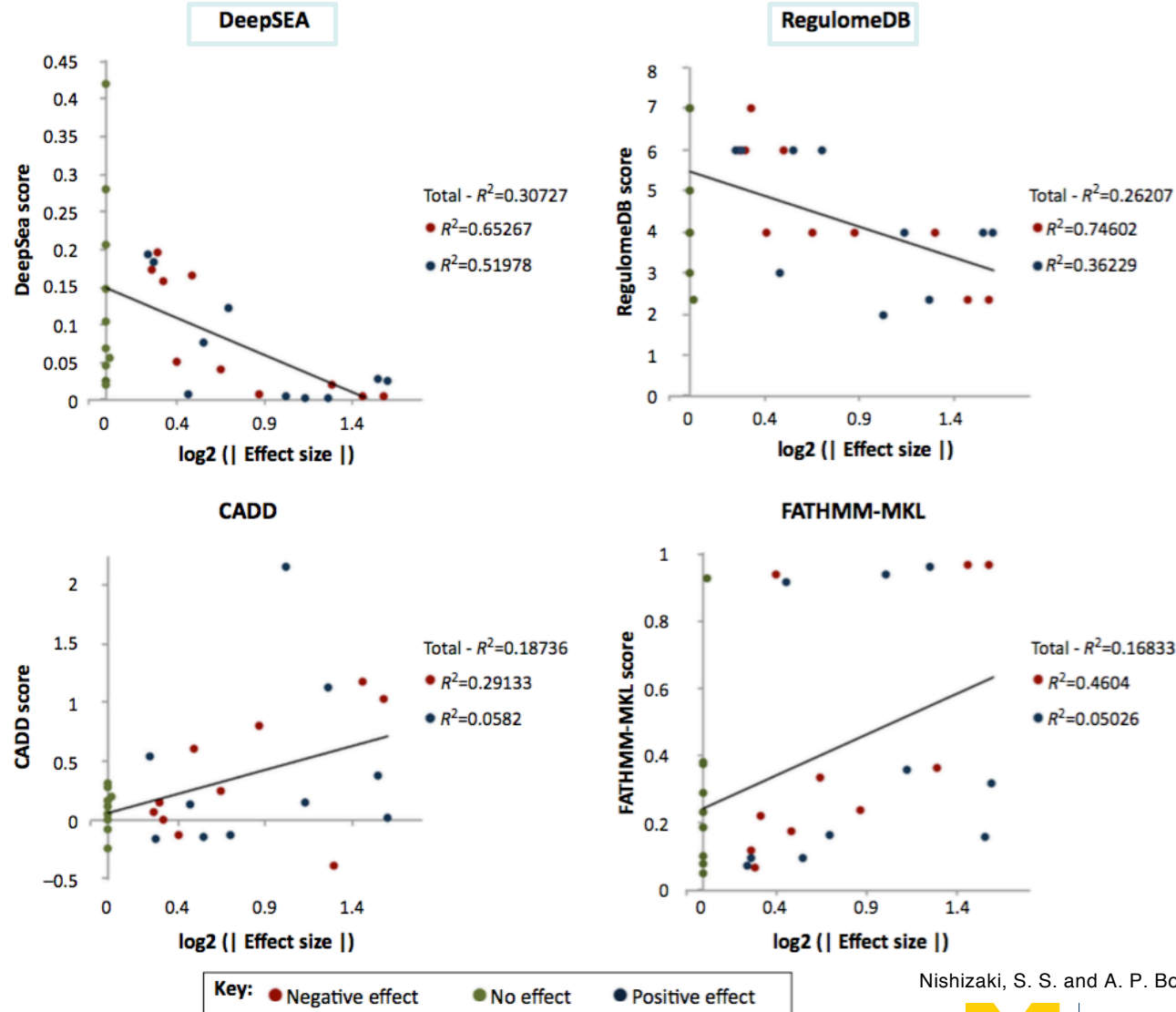


Soldner, F., et al. Nature 2016

Combined functional annotations increases confidence in a variant being functional



Evaluation of tools based on saturation mutagenesis massively parallel report assay



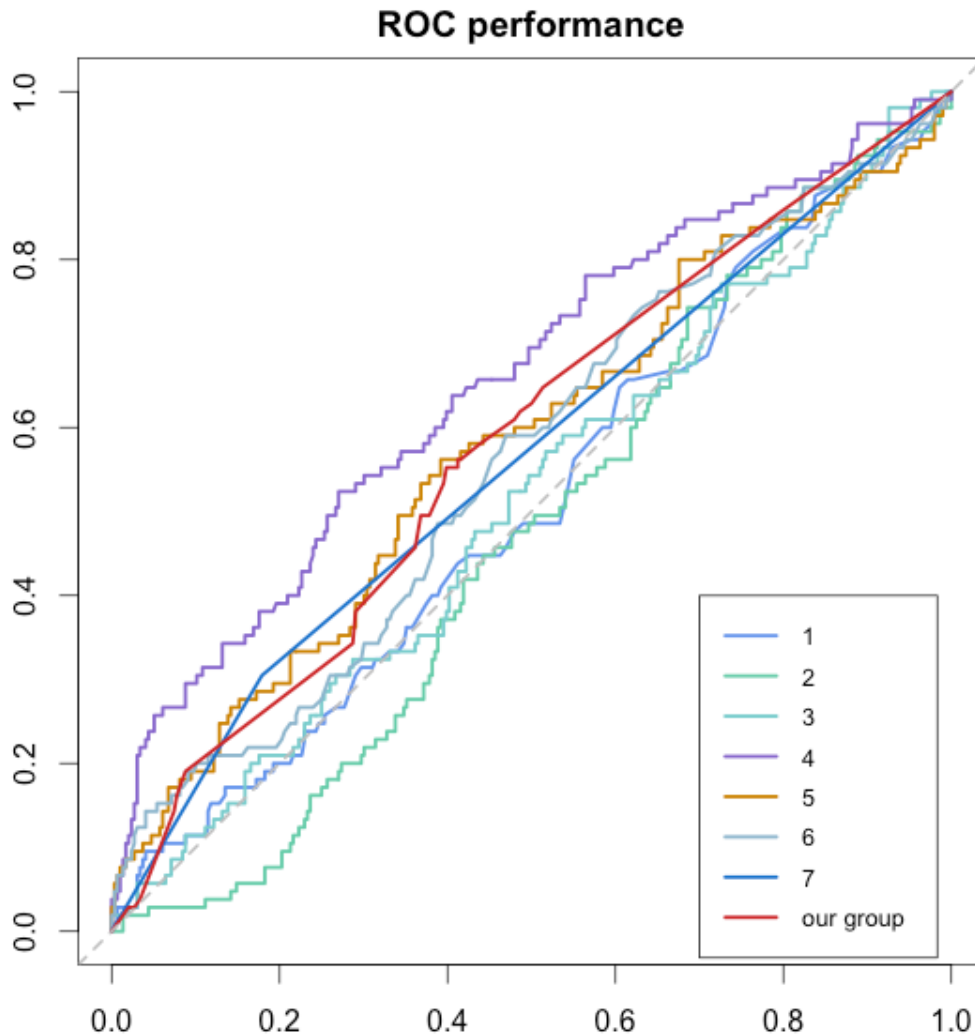
Nishizaki, S. S. and A. P. Boyle Trends Genet 2017

CAGI 2016 - Features used in each group

Group:	1	2	3	4 (*)	5	6	7
Sequence features (k-mers)			✓		✓		
TF binding Motif [transfer learning]	✓			✓		✓	
DeepBind [transfer learning] features				✓		✓	
DeepSEA [transfer learning] features				✓			✓
Basset [transfer learning] features		✓					
Conservation features	✓						
ENCODE track features	✓	✓				✓	
Genome segmentation features				✓			
Multiple prediction models	✓	✓		✓			

Slide from Haoyang Zeng CAGI 2016 presentation

ROC performance on test data from CAGI 2016



Task 2	LogSkew Spearman corr.	emVar AUROC
group		
4	0.449760	0.655261
5	0.342004	0.577220
7	0.007343	0.562854
6	0.190123	0.561776
3	0.030049	0.511181
1	0.105376	0.510103
2	-0.023092	0.472651

0.5798

CAGI challenge 2017

- Predicting individual non-coding variant effects in disease associated promoter and enhancer elements
- 17,500 single nucleotide variants (SNVs) in 5 human disease associated enhancers and 9 promoters
200~600 bases in each regulatory region
(25% for training, 75% for prediction)
- Assessed in a saturation mutagenesis massively parallel reporter assay
(effect size & confidence score)

[	chrX	138612668	138612669	T	A	-0.17	0.07	F9
	chrX	138612668	138612669	T	C	-0.26	0.24	F9
	chrX	138612668	138612669	T	G	0.34	0.05	F9
[	chrX	138612669	138612670	A	C	0	0	F9
	chrX	138612669	138612670	A	G	0.22	0.2	F9
	chrX	138612669	138612670	A	T	0.12	0.03	F9
[	chrX	138612670	138612671	T	A	0.06	0.02	F9
	chrX	138612670	138612671	T	C	0.26	0.21	F9
	chrX	138612670	138612671	T	G	0.08	0.01	F9

<https://genomeinterpretation.org/content/expression-variants>

Data source: massively parallel reporter assays (MPRA)

Table 1: Tested promoter sequences.

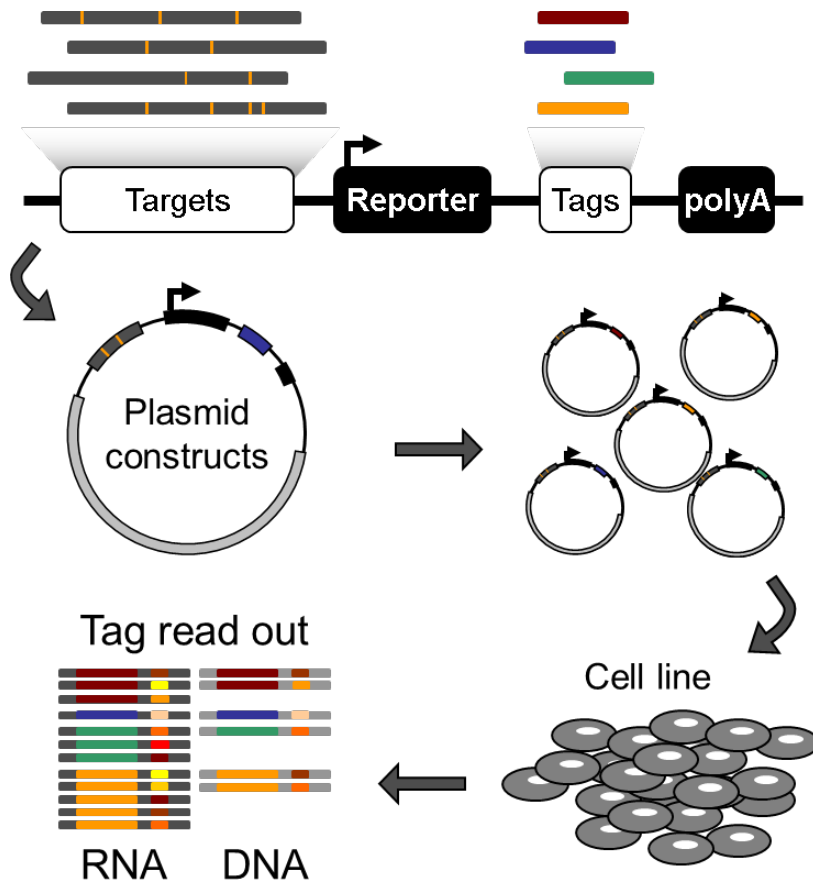
Promoter	Genomic coordinates (GRCh38/hg19)	Associated Phenotype	Cell line	Construct size (bp)
F9	chrX:138,612,624-138,612,923	Hemophilia B	HepG2	300
GP1BB	chr22:19,710,790-19,711,173	Bernard-Soulier Syndrome	HEL 92.1.7	384
HBB	chr11:5,248,252-5,248,438	Thalassemia	HEL 92.1.7	187
HBG	chr11:5,271,035-5,271,308	Hereditary persistence of fetal hemoglobin	HEL 92.1.7	274
HNF4A (P2)	chr20:42,984,160-42,984,444	Maturity-onset diabetes of the young (MODY)	HEK293T	285
LDLR	chr19:11199907-11200224	Familial hypercholesterolemia	HepG2	318
MSMB	chr10:51,548,988-51,549,578	Prostate cancer	HEK293T	591
PKLR	chr1:155271187-155271655	Pyruvate kinase deficiency	K562	469
TERT	chr5:1,295,105-1,295,362	Various types of cancer	HEK293T, GBM	258

Table 2: Tested enhancer sequences.

Enhancer	Genomic coordinates (hg19)	Phenotype caused by mutations	Cell line	Construct size (bp)
IRF4	chr6:396143-396593	Human pigmentation	SK-MEL-28	451
IRF6	chr1:209989135-209989734	Cleft lip	HaCaT	600
MYC	chr8:128,413,074-128,413,673	Various types of cancer (<i>rs6983267</i>)	HEK293T	600
SORT1	chr1:109,817,274-109,817,873	Plasma low-density lipoprotein cholesterol & myocardial infraction	HepG2	600
ZFAND3	chr6:37,775,276-37,775,853	Type 2 diabetes	MIN6	578

<https://genomeinterpretation.org/content/expression-variants>

Data source: massively parallel reporter assays (MPRA)



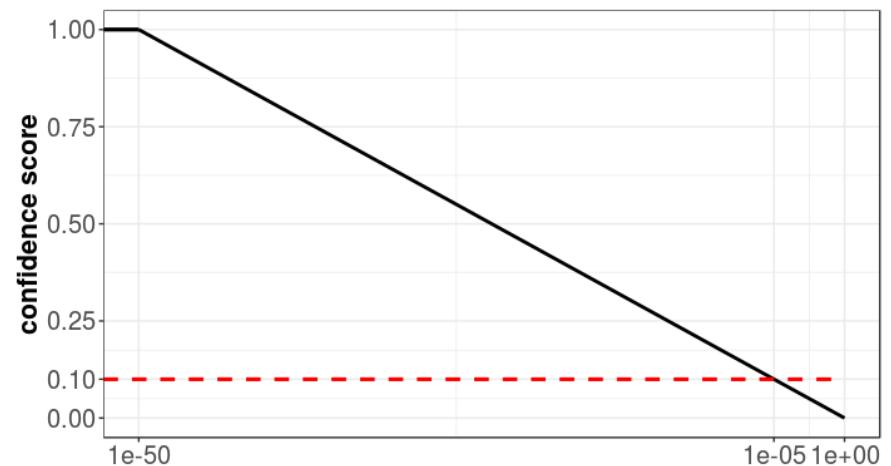
Multiple linear regression model
 $\log_2(\text{RNA}) \sim \log_2(\text{DNA}) + N + \text{offset}$

$$n \text{ tags} \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} \sim \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} + \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} \times \begin{pmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_m \end{pmatrix} + \epsilon$$



m Coefficients:
Effects of each
SNV

Confidence score:
derived from the p-value of regression



Prediction task

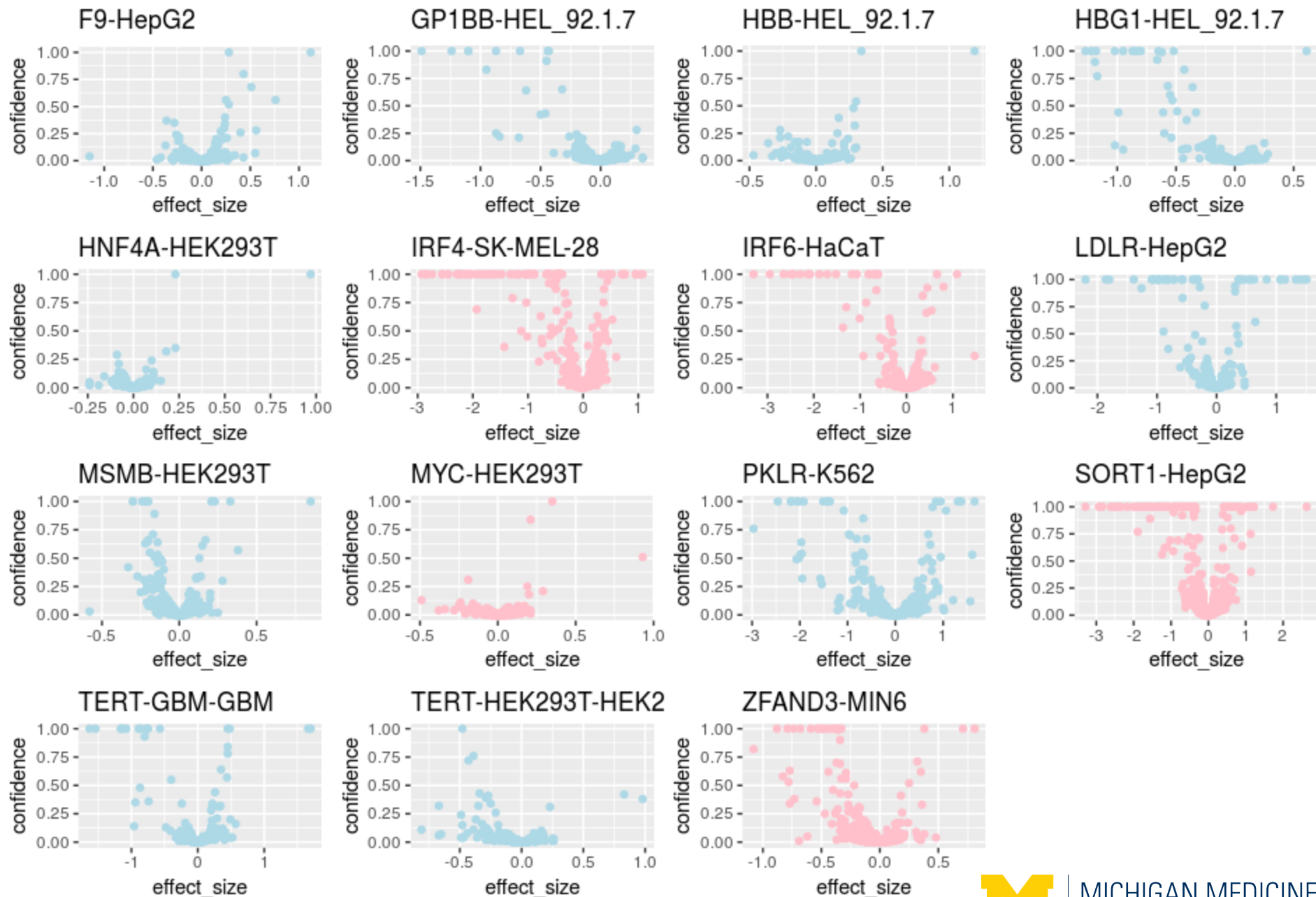
Training set format:

1. Chrom
2. Pos
3. Ref
4. Alt
5. Value - Estimated variant effects
6. Confidence - Confidence score derived from the p-value of the regression

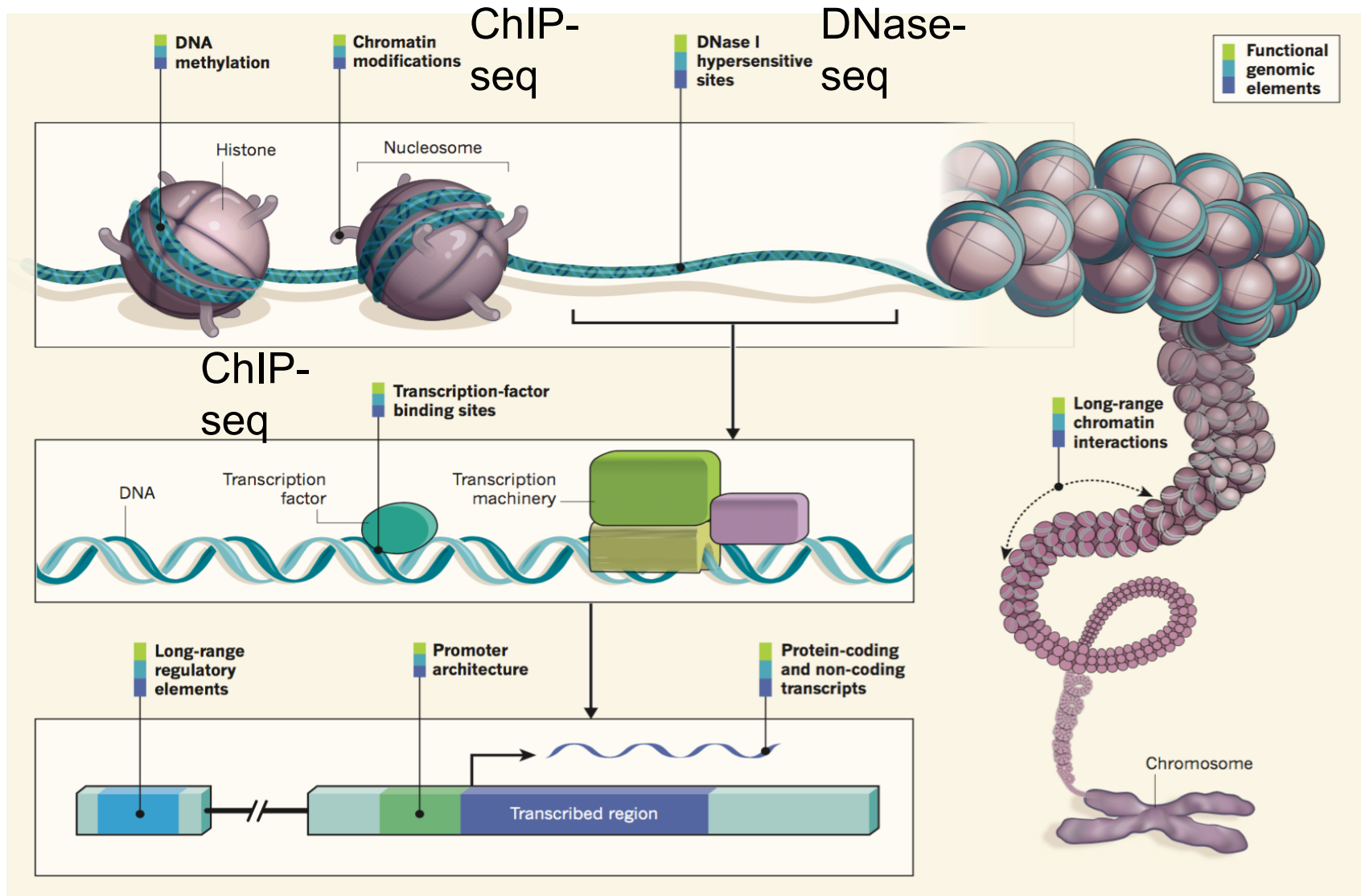
Prediction submission format:

1. Chrom
2. Pos
3. Ref
4. Alt
5. Direction - Repressive = submitted value of -1; activating = submit a value of 1; no effect = submit a value 0
6. P_Direction - Probability of a correct assignment in the previous field
7. Confidence - Prediction of the Kircher et al's confidence score
8. SD - Standard error of the value in the preceding field

Effect size – Confidence for each experiment



Functional genomics data for annotating SNVs



Ecker et al., 'Genomics: ENCODE explained' Nature (2012)

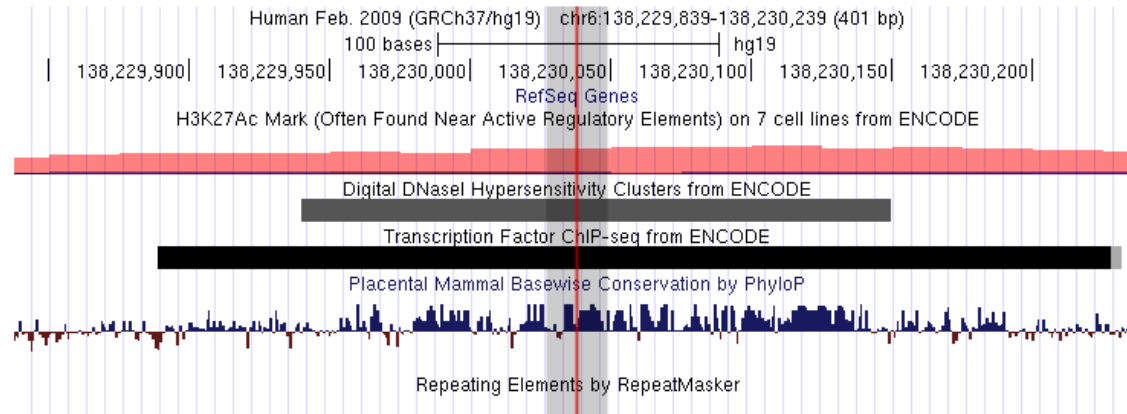
RegulomeDB

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

Data supporting chr6:138230039 (rs117480515)

Score: 2b

Likely to affect binding



Protein Binding					
Method	Location	Bound Protein	Cell Type	Additional Info	Reference
ChIP-seq	chr6:138229944..138230228	NFKB1	GM10847		ENCODE
ChIP-seq	chr6:138229960..138230196	BCL11A	GM12878		ENCODE
ChIP-seq	chr6:138229953..138230203	BCLAF1	GM12878		ENCODE
ChIP-seq	chr6:138229982..138230195	EBF1	GM12878		ENCODE
ChIP-seq	chr6:138230003..138230176	EGR1	GM12878		ENCODE
ChIP-seq	chr6:138229934..138230204	MEF2A	GM12878		ENCODE
ChIP-seq	chr6:138229970..138230185	MEF2C	GM12878		ENCODE

Experiment Summary



Assay category		Assay		Target of assay		Available data		Date released		Date Submitted	
DNA binding	8969	ChIP-seq	8969	transcription factor	4000	fastq	13803	July, 2013	2761	2013-07-31	2156
Transcription	3331	DNase-seq	873	histone	3102	bam	12815	March, 2014	887	2014-03-31	826
DNA accessibility	1132	polyA RNA-seq	817	control	2597	bigWig	11772	July, 2016	614	2010-11-16	258
DNA methylation	866	shRNA RNA-seq	533	broad histone mark	1727	bed narrowPeak	7354	May, 2016	569	2011-01-11	210
RNA binding	634	total RNA-seq	413	RNA binding protein	1466	bigBed narrowPeak	7343	October, 2016	485	2011-07-18	202
+ See more...		+ See more...		+ See more...		+ See more...		+ See more...		+ See more...	

Experiment status

released	14164
archived	1030
revoked	231

Project

ENCODE	9701
Roadmap	2761
modENCODE	1330
modERN	1165
GGR	461

+ See more...

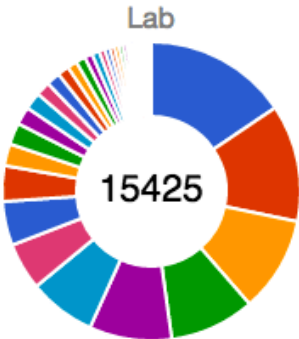
RFA

ENCODE3	6291
ENCODE2	2805
Roadmap	2761
modENCODE	1330
modERN	1165

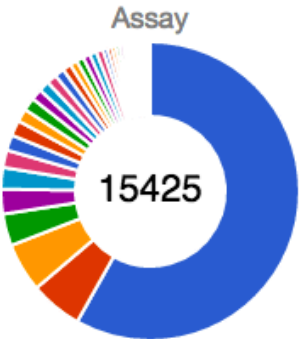
+ See more...

Genome assembly (visualization)

hg19	7401
------	------

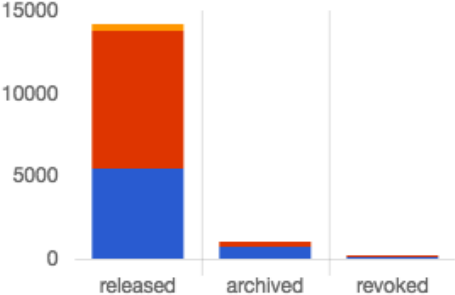


- Michael Snyder, Stanford
- Bradley Bernstein, Broad
- John Stamatoyannopoulos, UW
- Richard Myers, HAIB
- Bing Ren, UCSD
- Kevin White, UChicago
- Brenton Graveley, UConn
- Thomas Gingeras, CSHL
- Joseph Costello, UCSF
- Gene Yeo, UCSD
- Michael Snyder, Stanford



- ChIP-seq
- DNase-seq
- polyA RNA-seq
- shRNA RNA-seq
- total RNA-seq
- eCLIP
- RNA microarray
- DNAME array
- WGBS
- RRBS
- small RNA-seq

Status



Replication

- unreplicated
- isogenic
- anisogenic

DeepSEA (Deep Learning)

Zhou, J. and O. G.
Troyanskaya Nat
Methods 2015

Output:
variant functionality
prediction

Output:
predicted chromatin
effect

Output:
predicted allele-
specific chromatin
profile

919 chromatin features

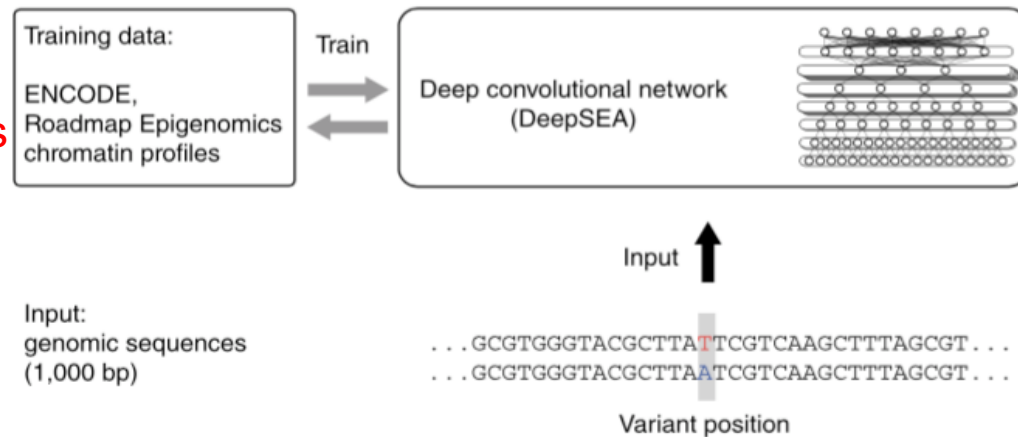
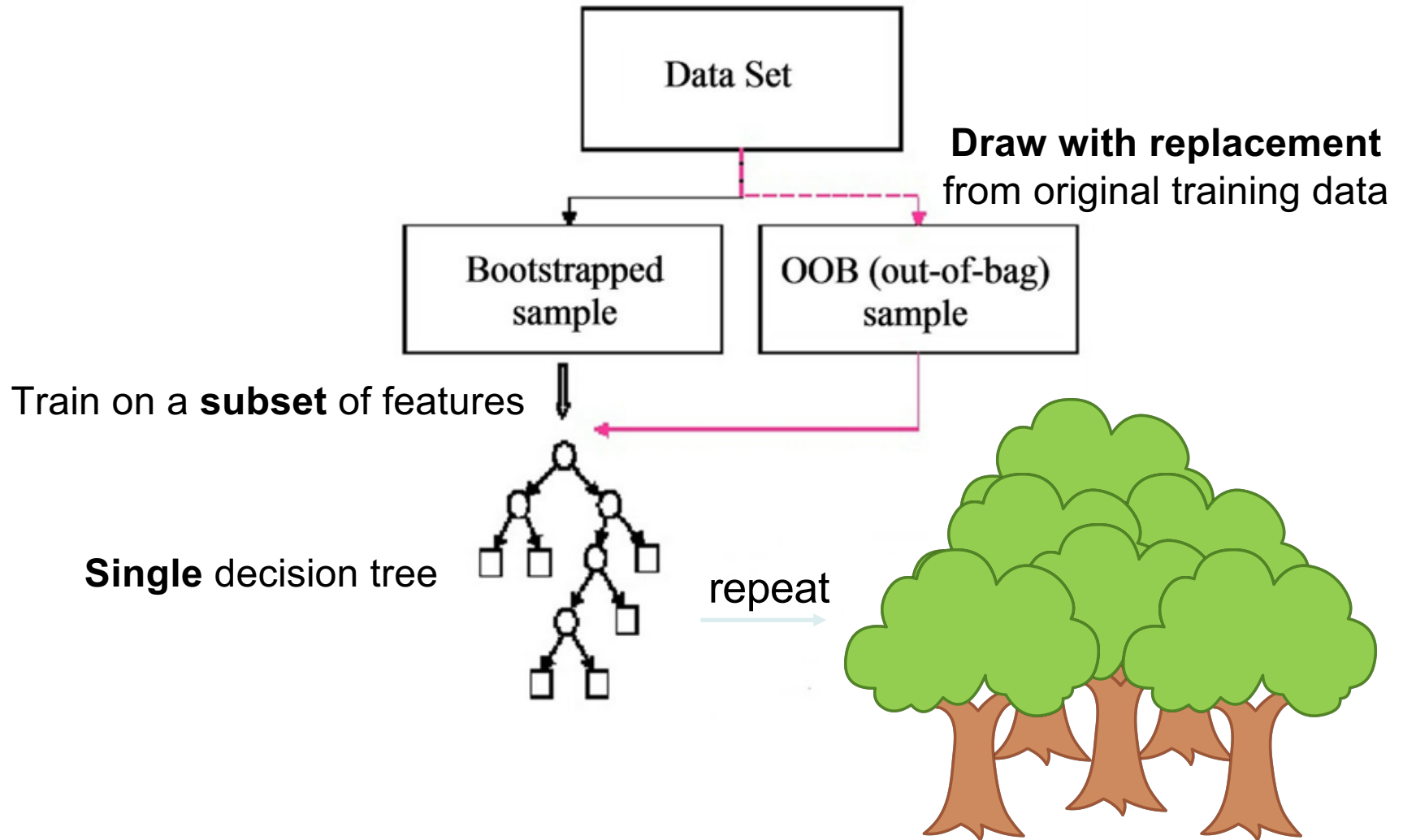


Figure 1 | Schematic overview of the DeepSEA pipeline, a strategy for predicting chromatin effects of noncoding variants.

Random Forest



Input features

Binary variables

CHIP	TF binding from ChIP-seq
DNASE	DNase I hypersensitive sites from DNase-seq
FOOTPRINT	DNase footprint
FOOTPRINT_matched	DNase footprint + matched TF ChIP peak
PWM	TF motif matches from JASPAR, TRANSFAC, UniPROBE
PWM_matched	Motif match + matched TF ChIP peak

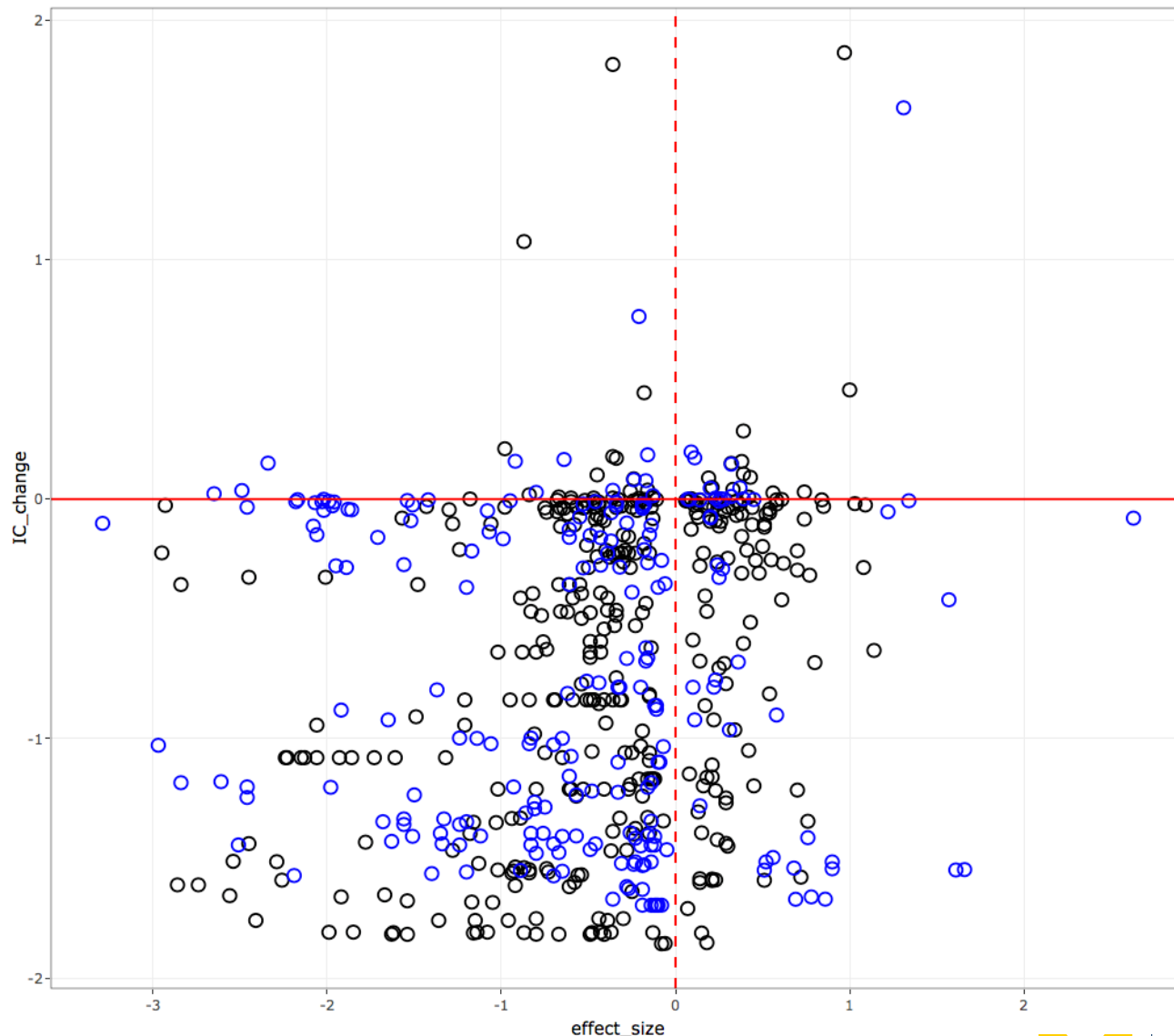
Numeric variables

IC_change_max	Maximum information content change
IC_change_matched	Maximum information content change in PWM with matched TF ChIP peak
CHIP_sig	ChIP-seq signal
IC_change_all (3,6)	Information content change for all PWMs (163 TFs)
DeapSEA scores (1,4)	Allele binding difference of chromatin features from DeepSEA + significance score (920 features)
DeepBind scores (2,5)	Allele binding difference from (515 TFs)

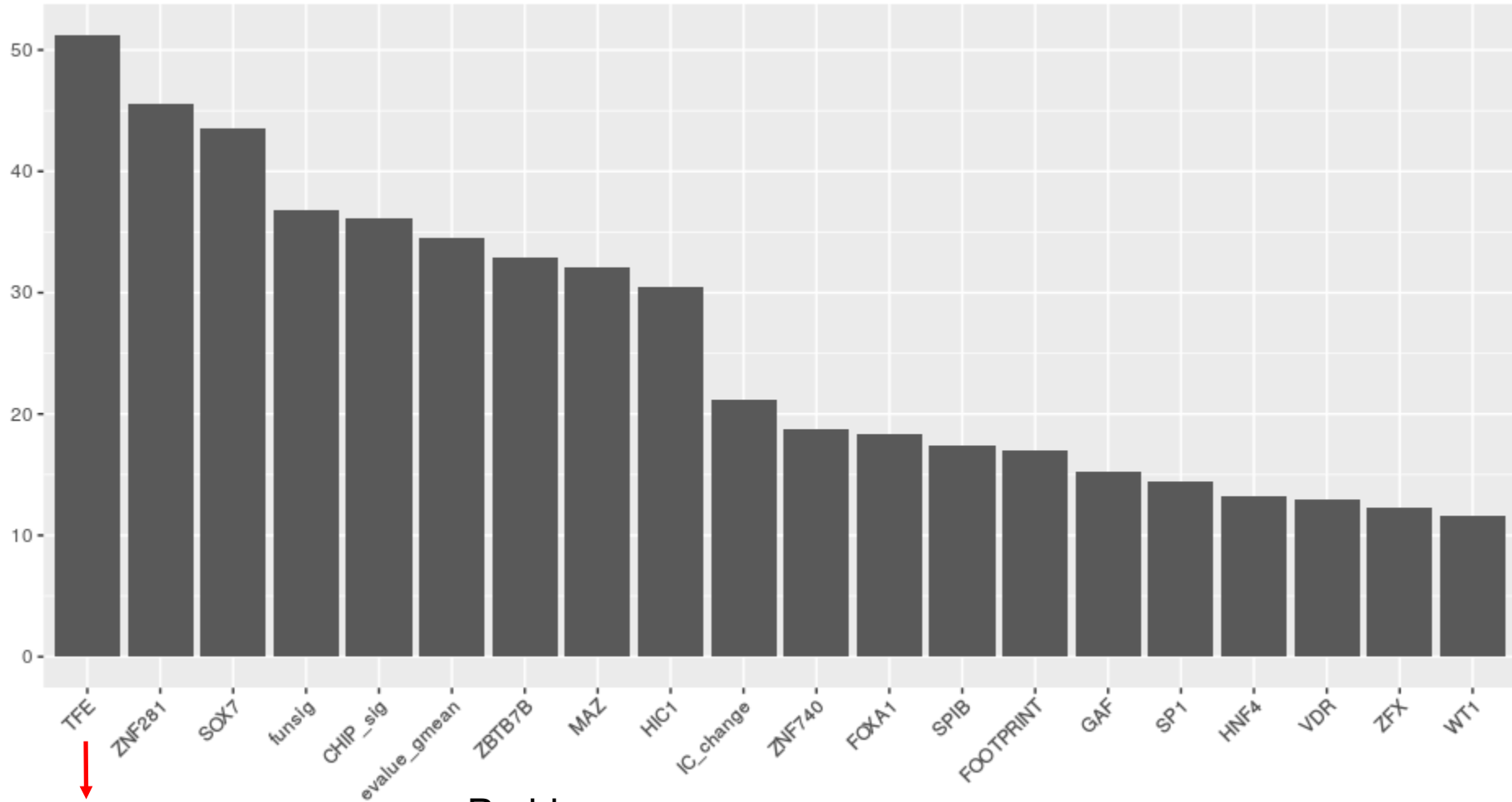
Input features

Submission	Feature	Number of features
1, 4	RegDB + DeepSEA	929
2, 5	RegDB + DeepBind	524
3, 6	RegDB + PWM changes	172

Labeling with GO annotation: negative regulation of transcription by RNA polymerase II



Random forest classifier (confidence ≥ 0.1 versus confidence < 0.1)



Alan P Boyle – CAGI5

Problems:

- Not enough data
- Test data have binding of TFs unobserved in training set



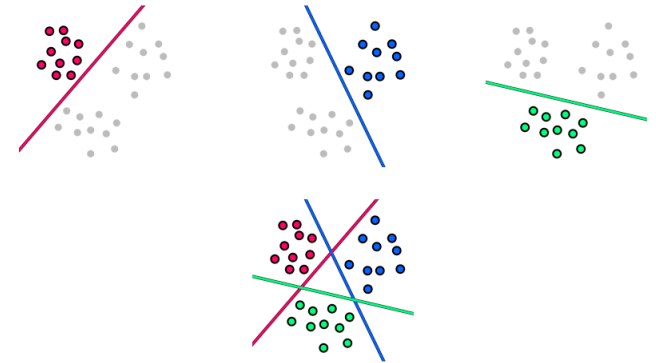
MICHIGAN MEDICINE
UNIVERSITY OF MICHIGAN

Multiclass classification

A random forest model was trained to make predictions for both direction of effects and confidence scores (ntree=500).

4,650 measurements classified for training:

- Confidence ≥ 0.1 & effect > 0 : **374 (1)**
- Confidence ≥ 0.1 & effect < 0 : **736 (-1)**
- Confidence < 0.1 : **3,540 (0)**



We trained three binary classifiers for each label with random forest model, and predicted the label with the highest probability.

For confidence prediction, we trained a random forest regression model on confidence scores and calculated the standard deviation of predictions from 500 trees in “SD” column.

Three submissions

	Binary classifier	AUC from 10-fold CV	Prediction (total: 13,636)
1	-1 / +1,0	0.858	N: 679 P: 138
	+1 / -1,0	0.78	
	0 / -1,+1	0.819	
2	-1 / +1,0	0.876	N: 399 P: 69
	+1 / -1,0	0.766	
	0 / -1,+1	0.837	
3	-1 / +1,0	0.712	N: 441 P: 35
	+1 / -1,0	0.58	
	0 / -1,+1	0.705	