Group meeting

Lou Shaoke

Department of Molecular Biophysics and Biochemistry

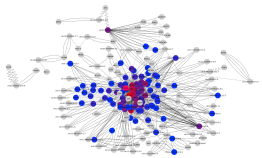
loushaoke@gmail.com

October 26, 2017

Yale

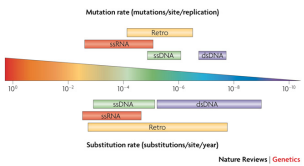
Conserv

Network

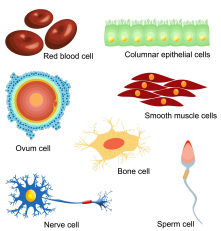


Motif

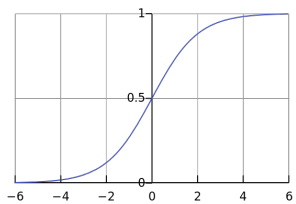
Mutation



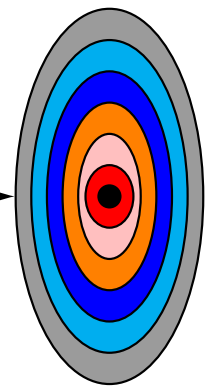
Cell specifity



Funseq and Funseq2: Priorizing the variants and evaluate the deleterious effect is challenging; Cannot be experimentally evaluated.



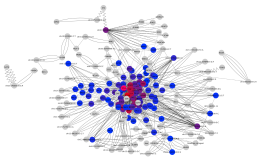
Model



Target
deleterious effect

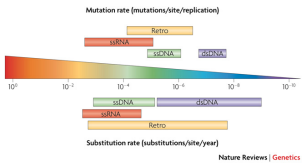
Conserv

Network

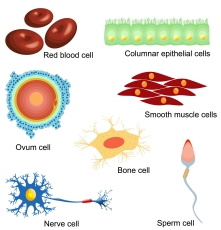


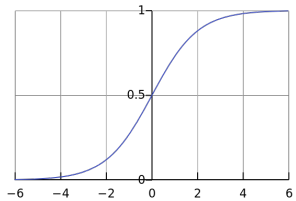
Motif

Mutation



Cell specifity

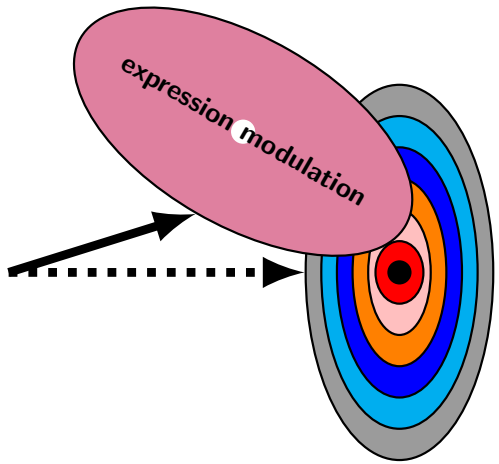




Model

Deleterious SNVs tend to affect the expression of its target gene by changing the TF binding affinity

Luciferase Assay (or alike assays) is the experimental way to evaluate **expression modulations**



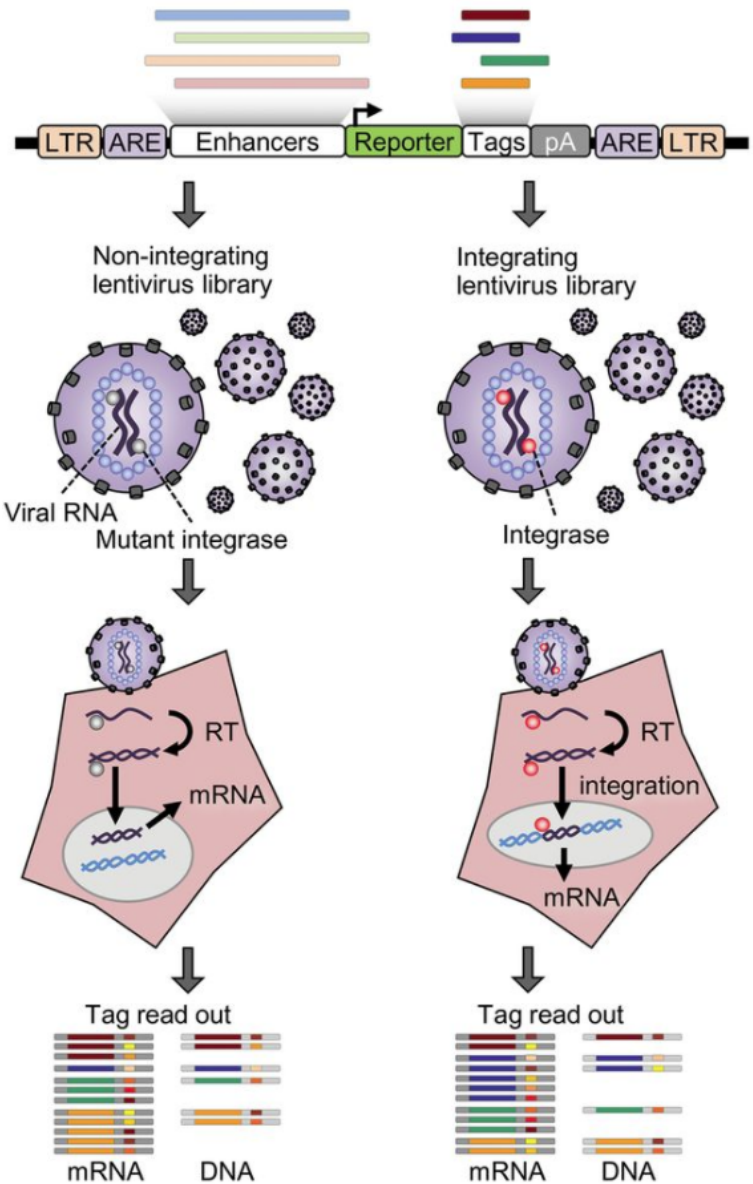
Target

deleterious effect

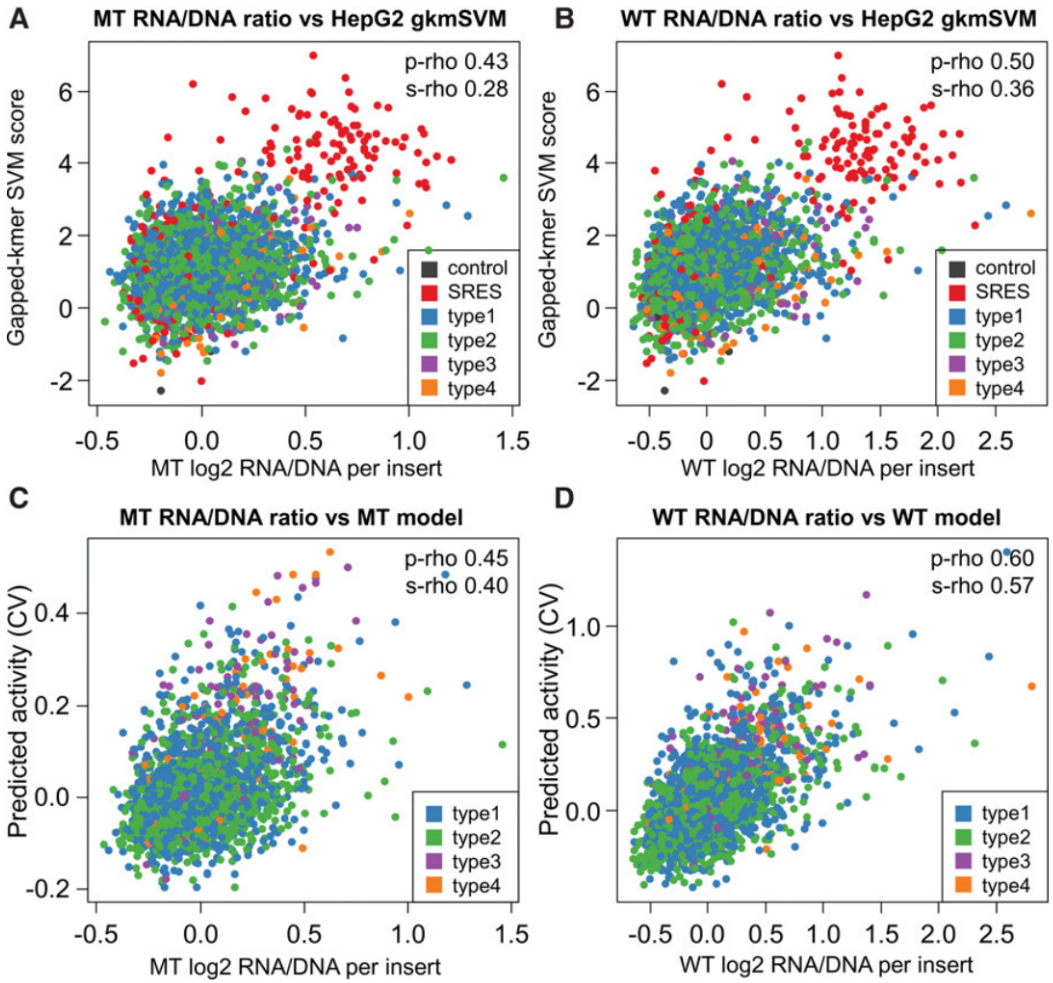
Reported Assay

its high-throughput version: MPRA (multiplex

Lenti MPRA


LTR
ARE
Enhancers
Reporter
Tags
pA
ARE
LTR
Non-integrating lentivirus library
Integrating lentivirus library
Viral RNA
Mutant integrase
Integrase
RT
mRNA
RT
integration
mRNA
Tag read out
Tag read out
mRNA
DNA
mRNA
DNA

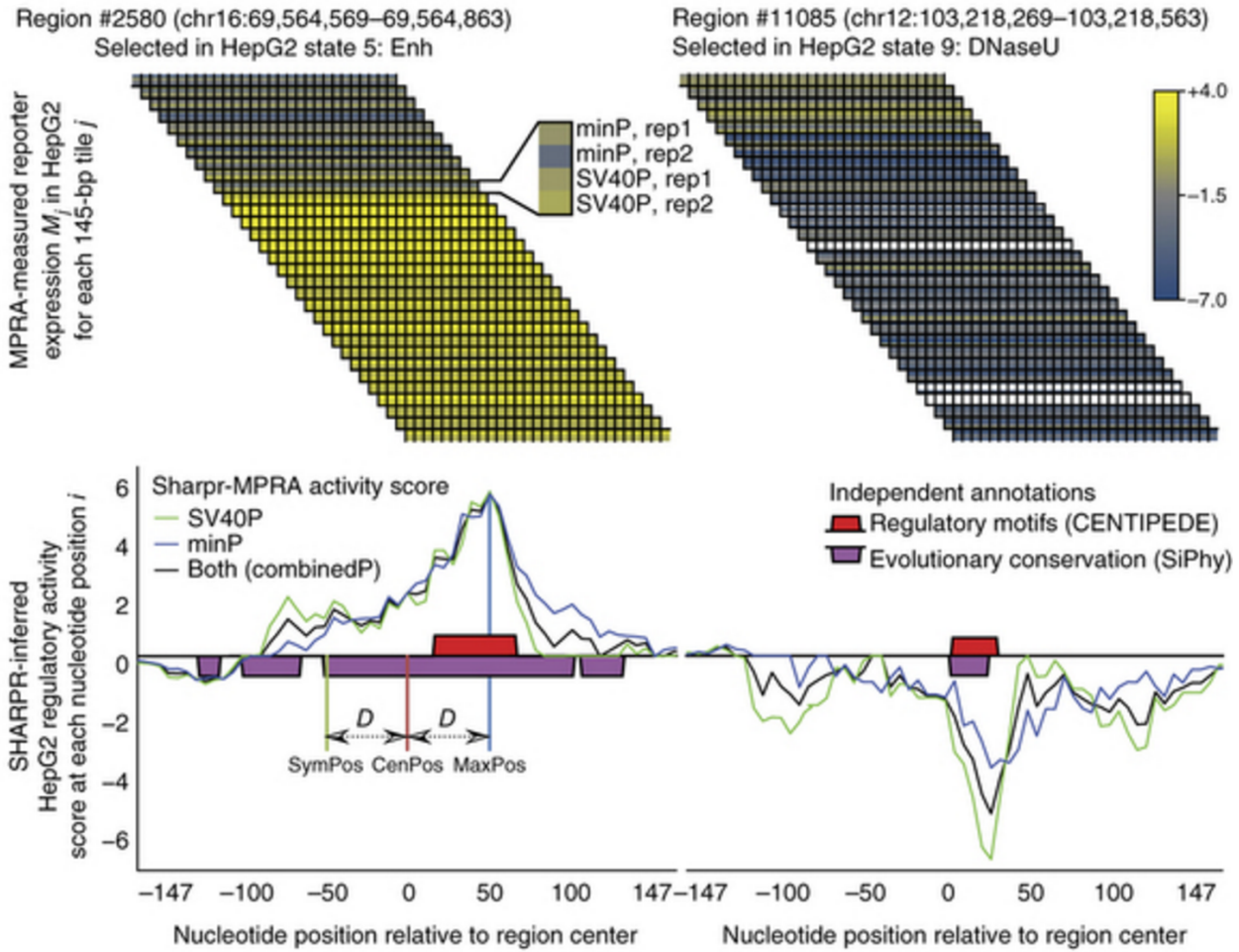
LentiMPRA compares differences with/without genomic context



Inoue et al Genome Res

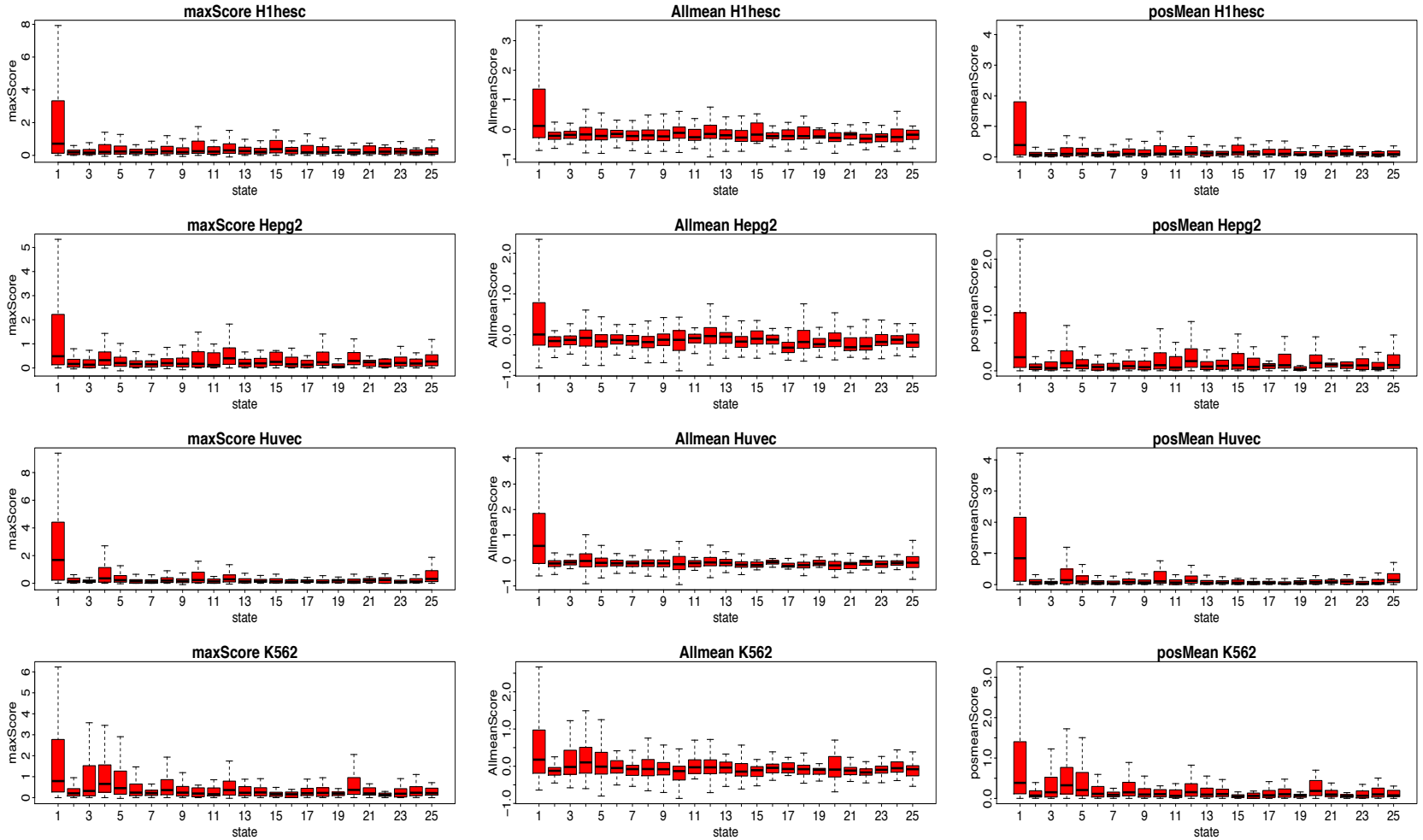
Nature Biotech: MPRA to test enhancer regions

MPRA in Nature Biotech paper. The regions tested in the paper based on ChromHMM segmentations



Manolis Nature Biotech 2016 The core regions of active element is about 150bp

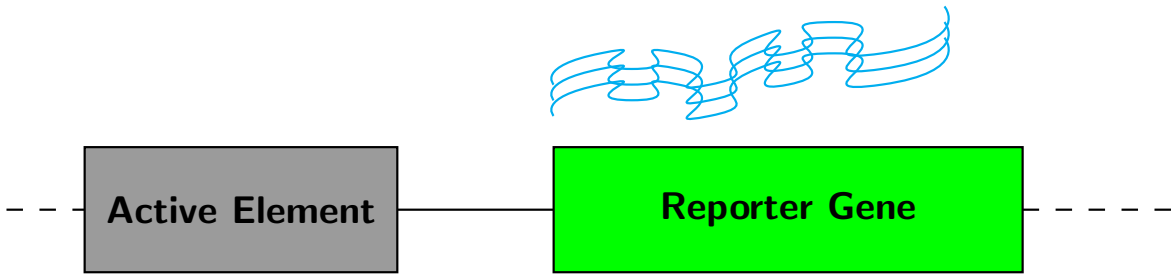
New viewpoint about MPRA/Luciferase Assay



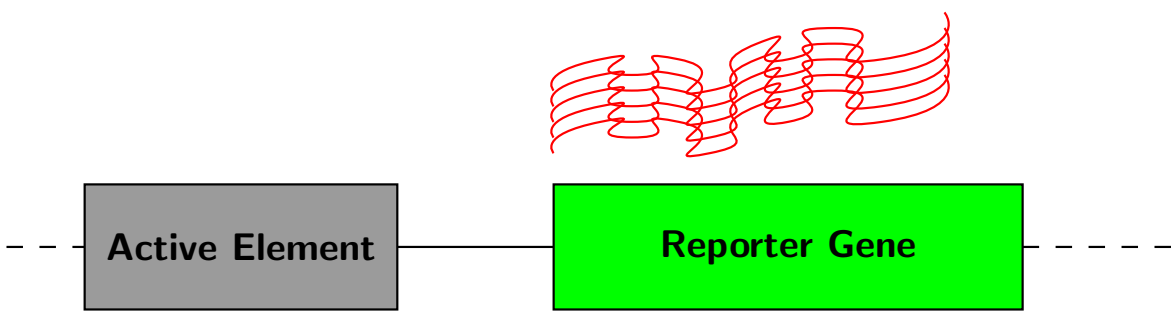
Weak cell-specific effect and Weak chromatin stat effect;

What we want to model

Plasmid Vector in luciferase assay



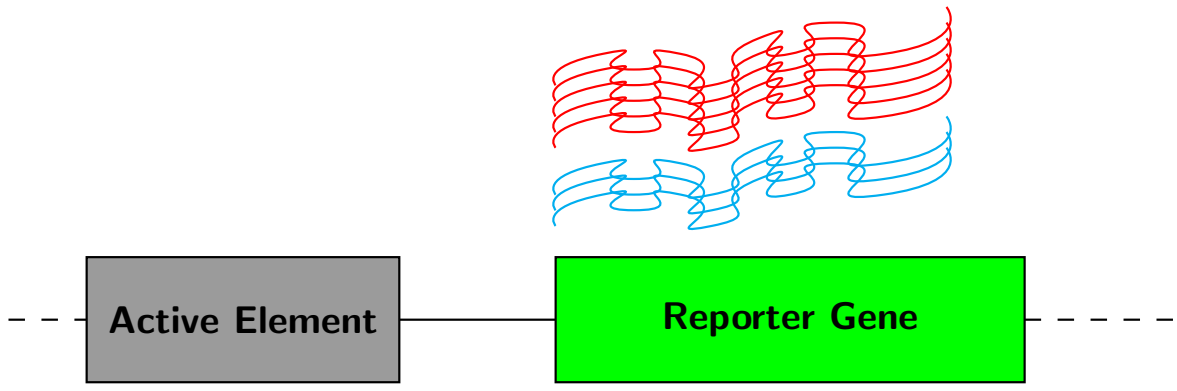
Allele1 ATGCAGCTT



Allele2 ATGCGGCTT

What we want to model

Plasmid Vector in luciferase assay



vs ATGCAGCTT
ATGCGGCTT

logSkew=$log(\frac{Expr_{allele1}}{Expr_{allele2}})$

Constrains: active element and ref allele has expr regulatory effect: more reads count for vector with ref allele active element than (Cell paper: either of alllele has expr regulatory effect)

Target variable Y =Significant regulatory change(expression) for Ref/Alt allele (logSkew).

Dataset

Dataset ever tried:

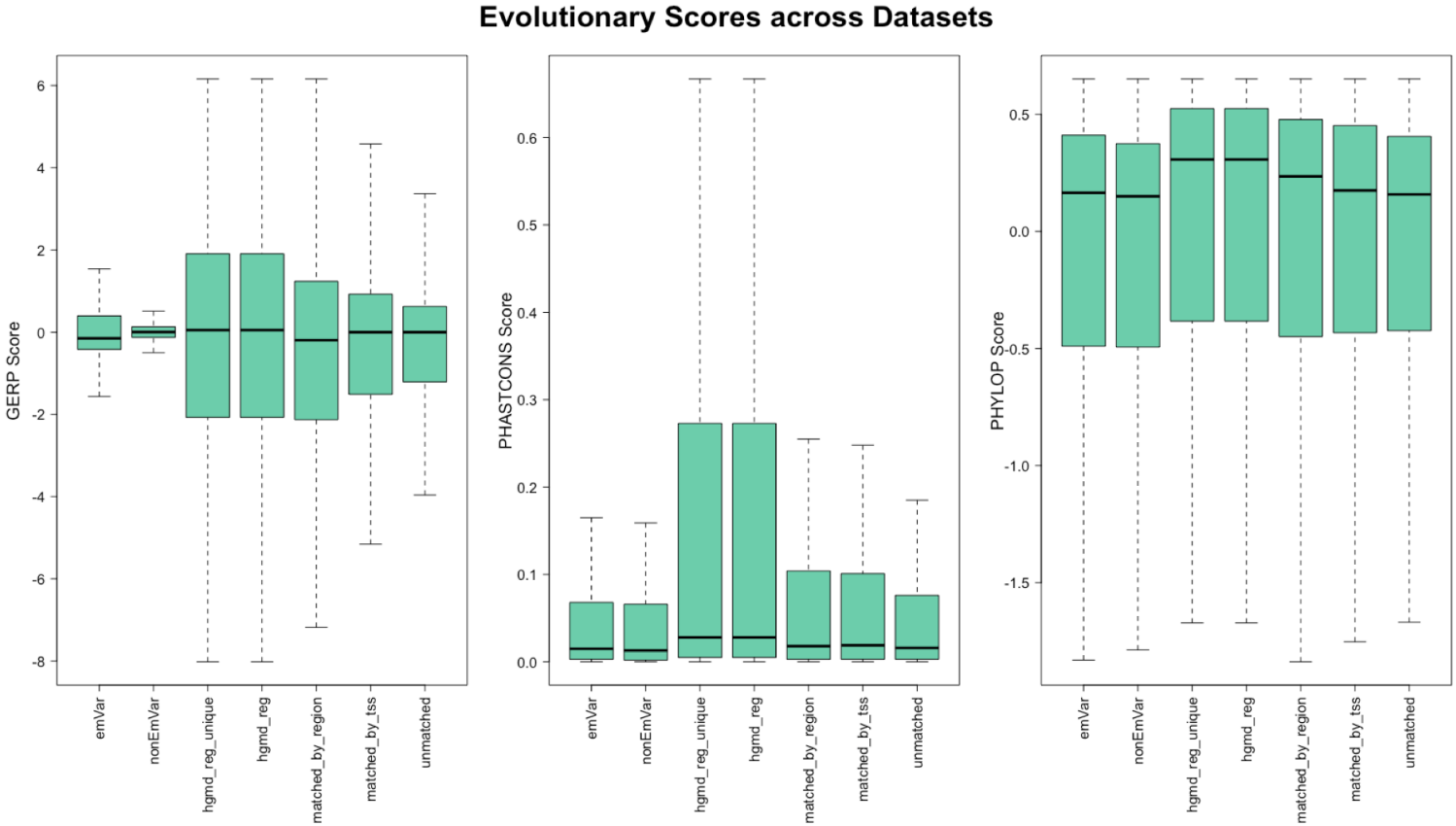
- Ryan cell paper expression-modulating variants(emVar) **VS** all non-emVar

All dataset depend on Ryan Cell paper, currently only SNVs(27005) are considered. It also require the reference sequence has potential expression modulation activity, which required more reads count for either ref/alt allele than that from a control vector;

The all SNVs contains unknown state (NA) variants and after filtering, only 4.5k SNVs left with significance estimation.

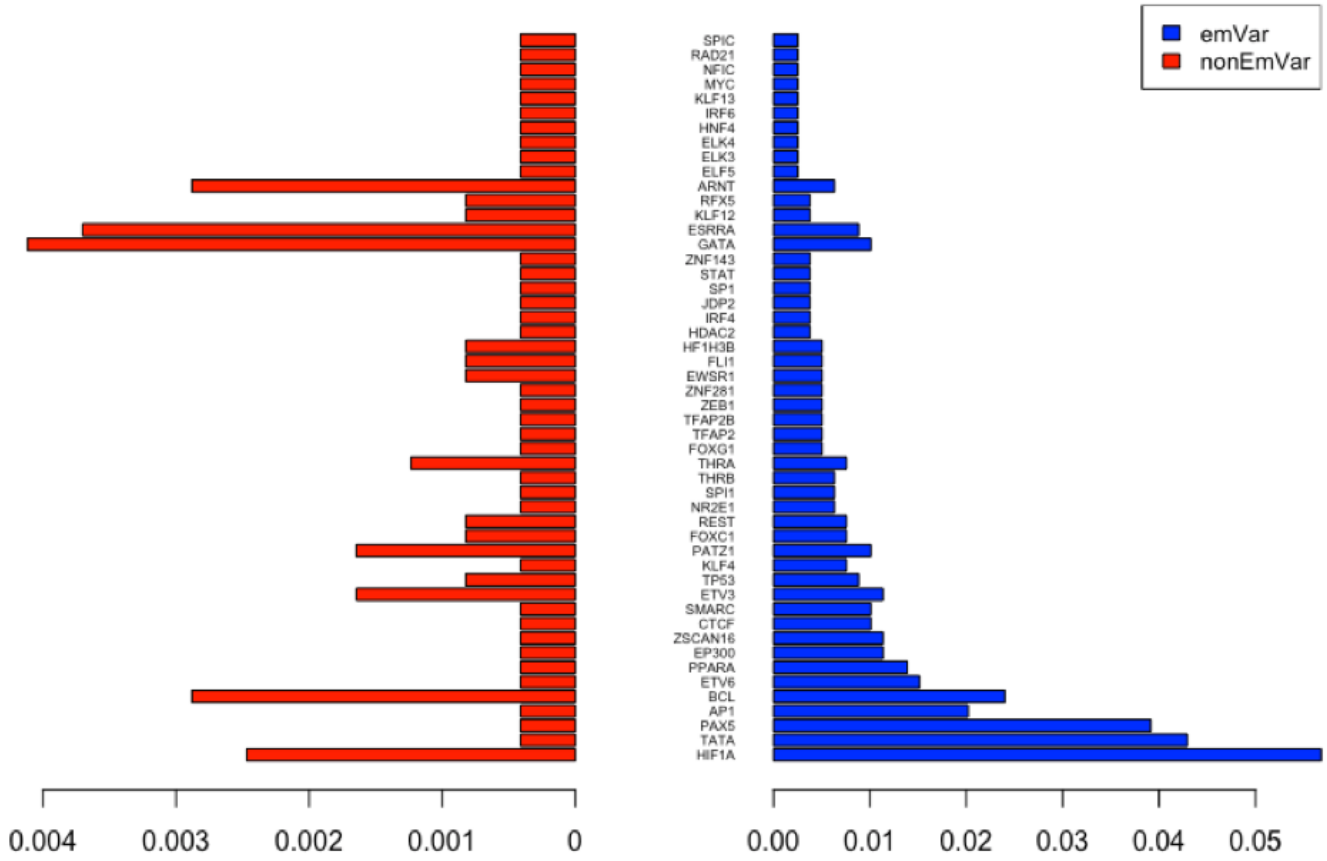
The SNV without overlapping with any tested histone and tf peaks were removed and 3k+ variants left

Part 1: Dataset exploration



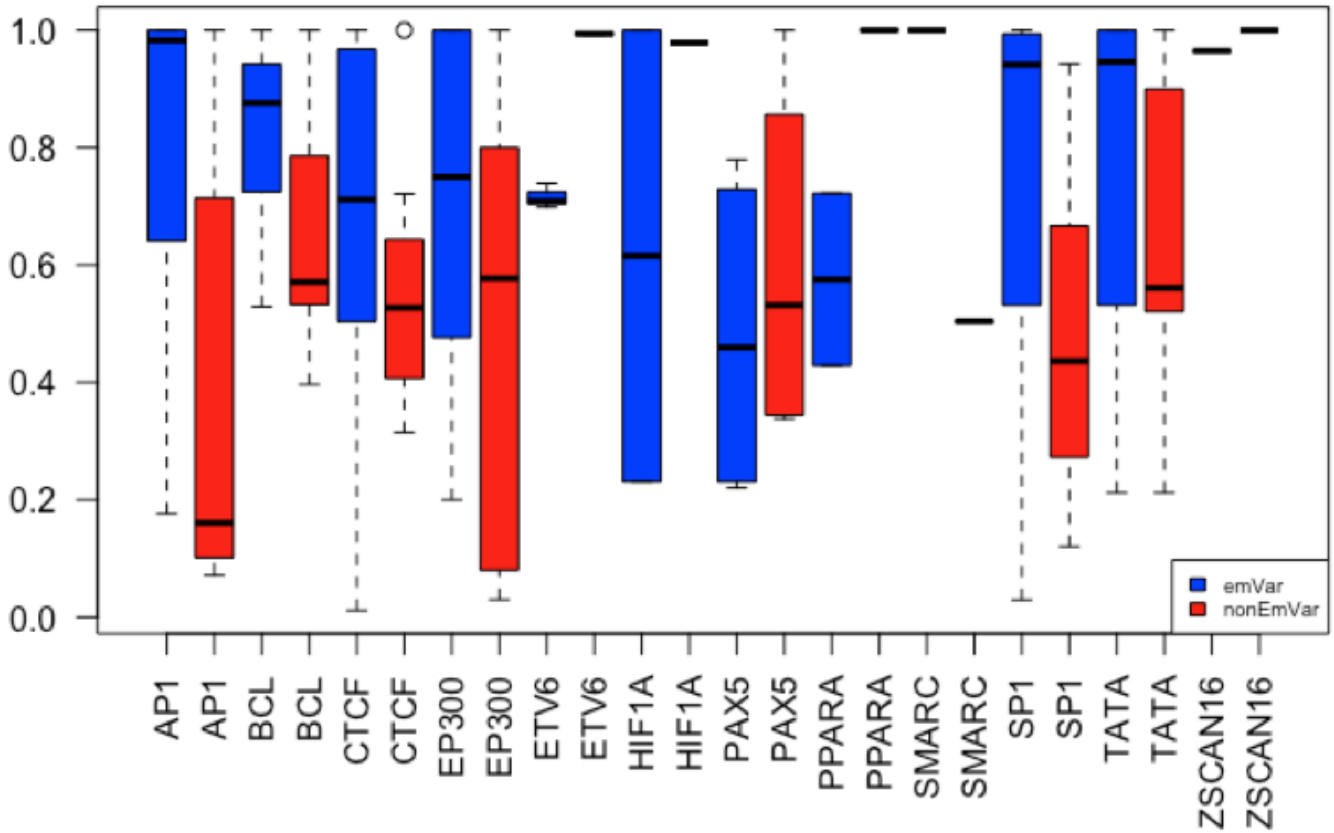
Evolution features: GERP, Phastcons and Phylop. Inter dataset difference is larger than intra pos and negative dataset.

Part 1: Dataset exploration



More motif binding event enriched in emVar.

Part 1: Dataset exploration

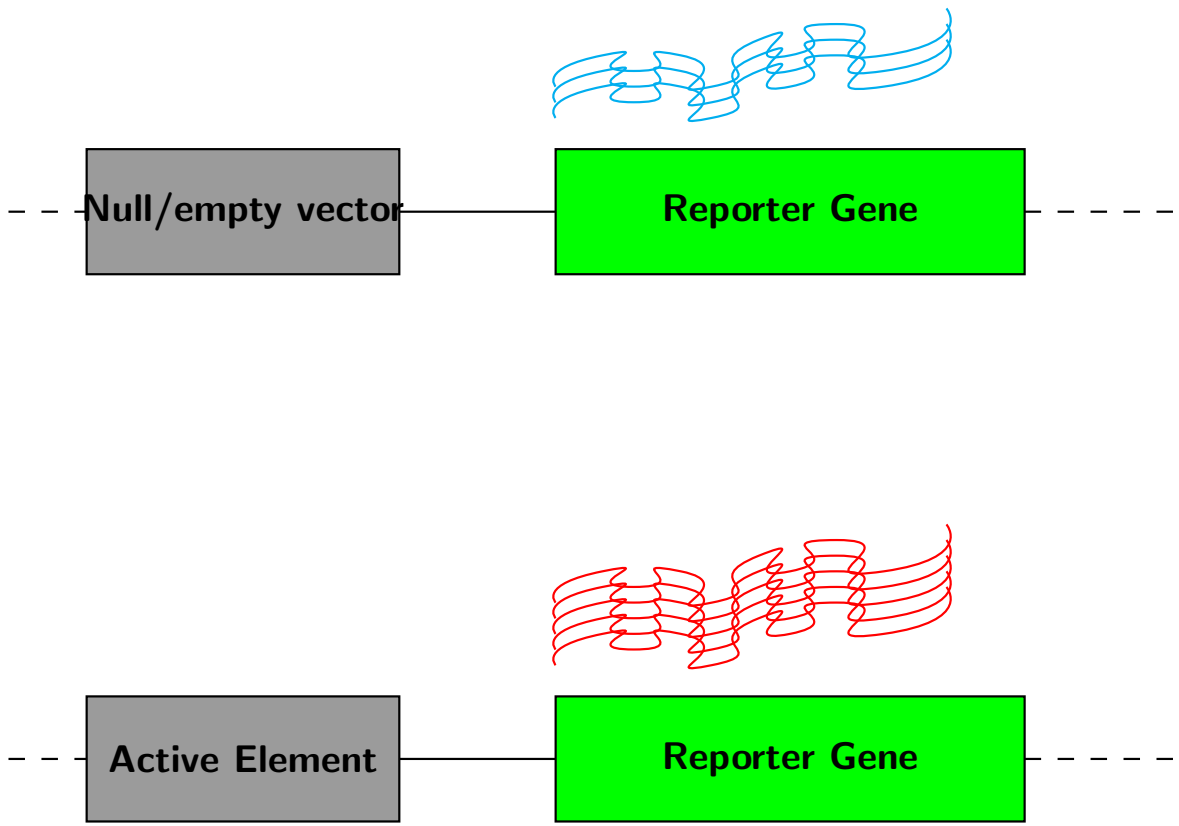


The motif break score in emVar group is larger than non-emVar group

Part 2: predict LogSkew using regression methods

Regression problem

(target variable: $LogFC\,\frac{\text{ref or alt}}{null}$, $orLogFC\,\frac{ref}{alt}$)

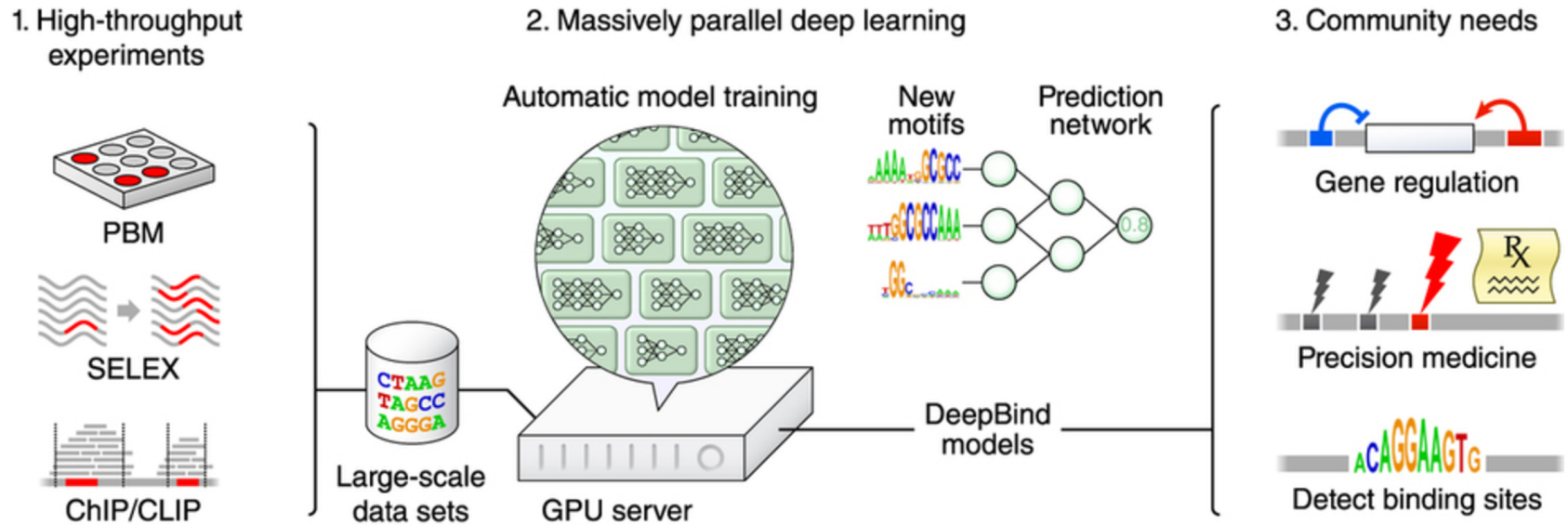


Allele1 ATGCGGCTT

Predictors:DeepBind Profile (DP)

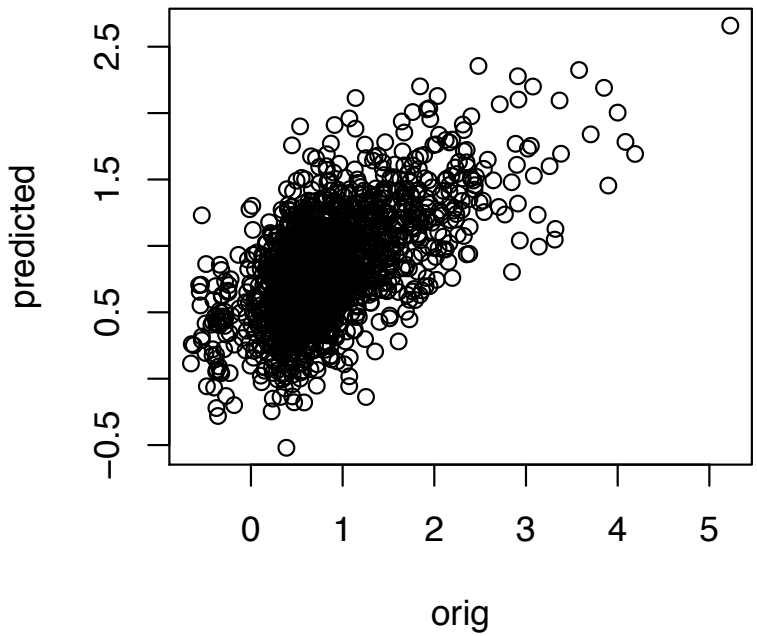
DeepBind profile score(DP) from Deepbind, which learns binding preferences from SELEX, ChIP−Seq.

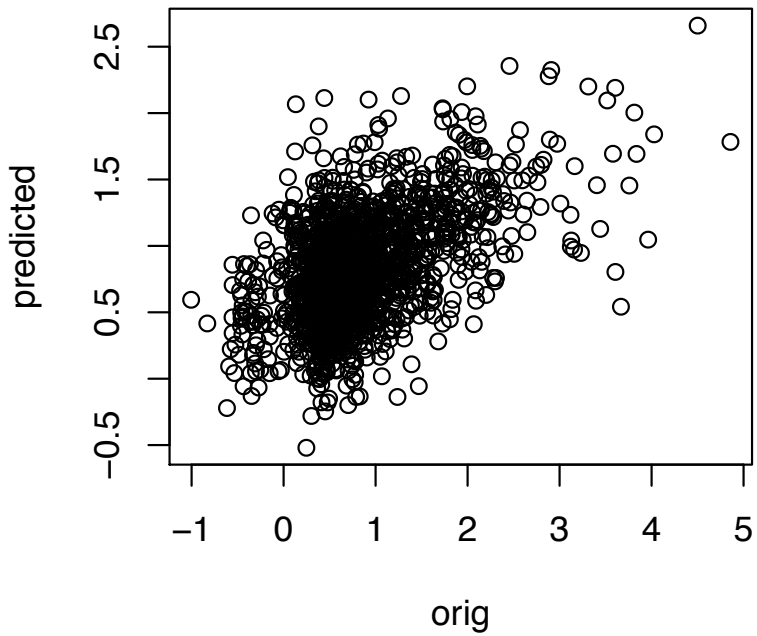
515 features in total was used to learn.



Linear regression, $Y : LogFC^{\frac{\text{ref or alt}}{null}} \sim DP_{\text{ref or alt}}$

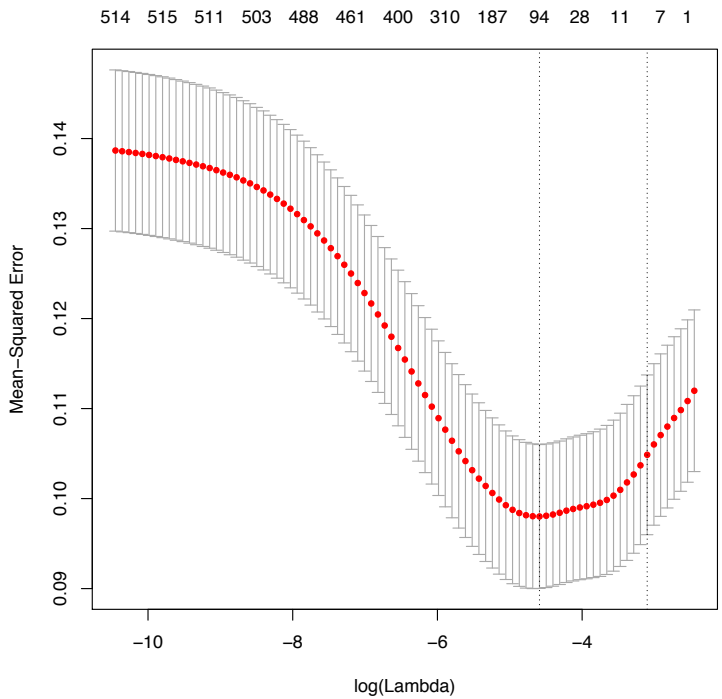
Train using ref allele information, then test on alt allele

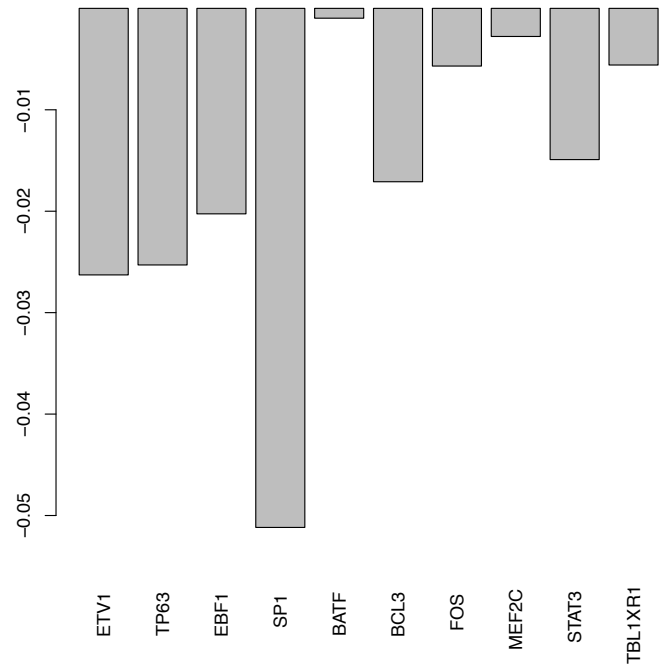




Pearson.cor=0.62(train), 0.51(test)
Spearman.Cor=0.55, 0.42

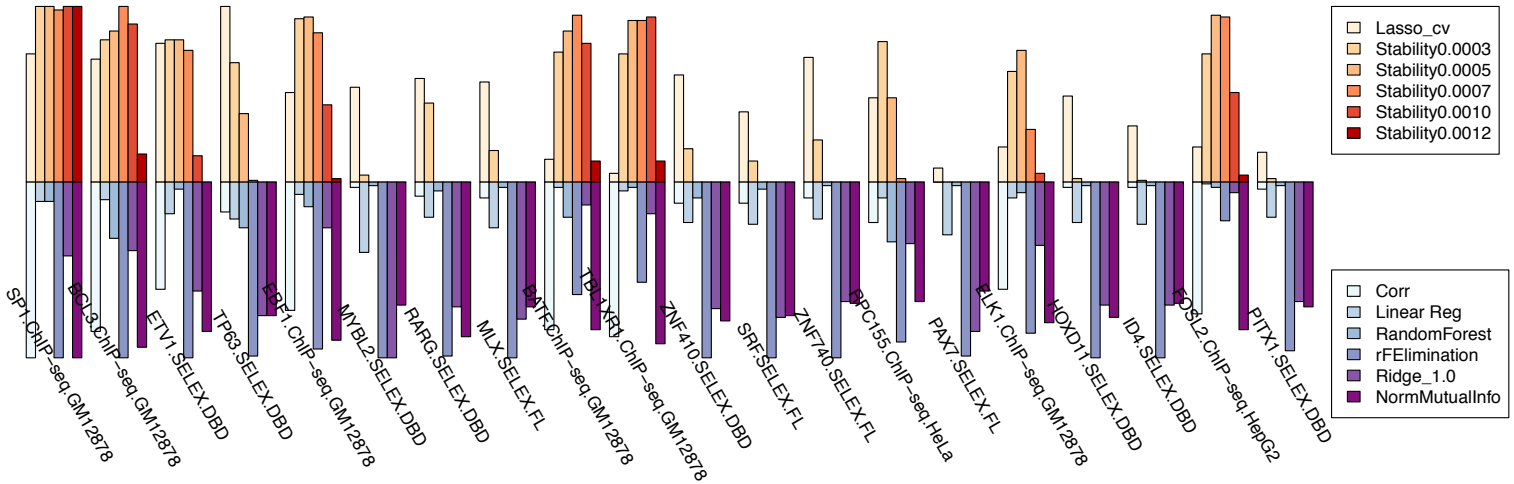
Lasso regression $Y : LogFC\frac{ref}{alt} \sim DP_{ref} - DP_{alt}$





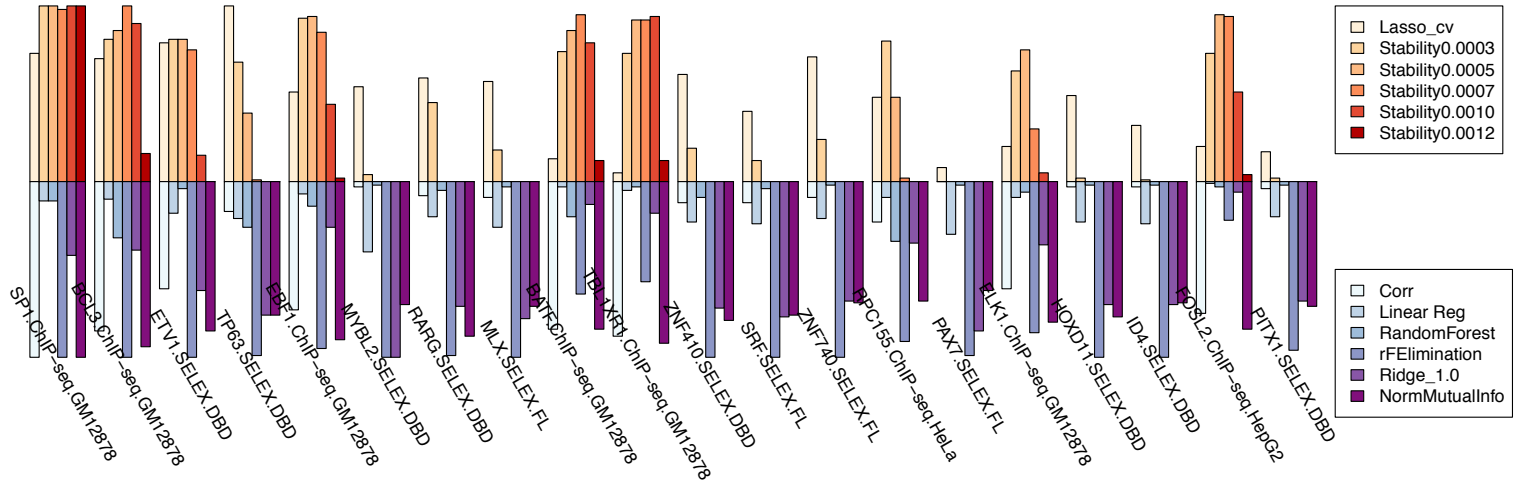
R^2: 0.29(DP), 0.39(DP+Hist+TF+CAGE+GERP)

Feature selection



Top 20 factors in all the feature selection frameworks, sorted by the average value.

- **ChIP-Seq TF binding features are cell-specific. that will limite the application in other cell lines.**

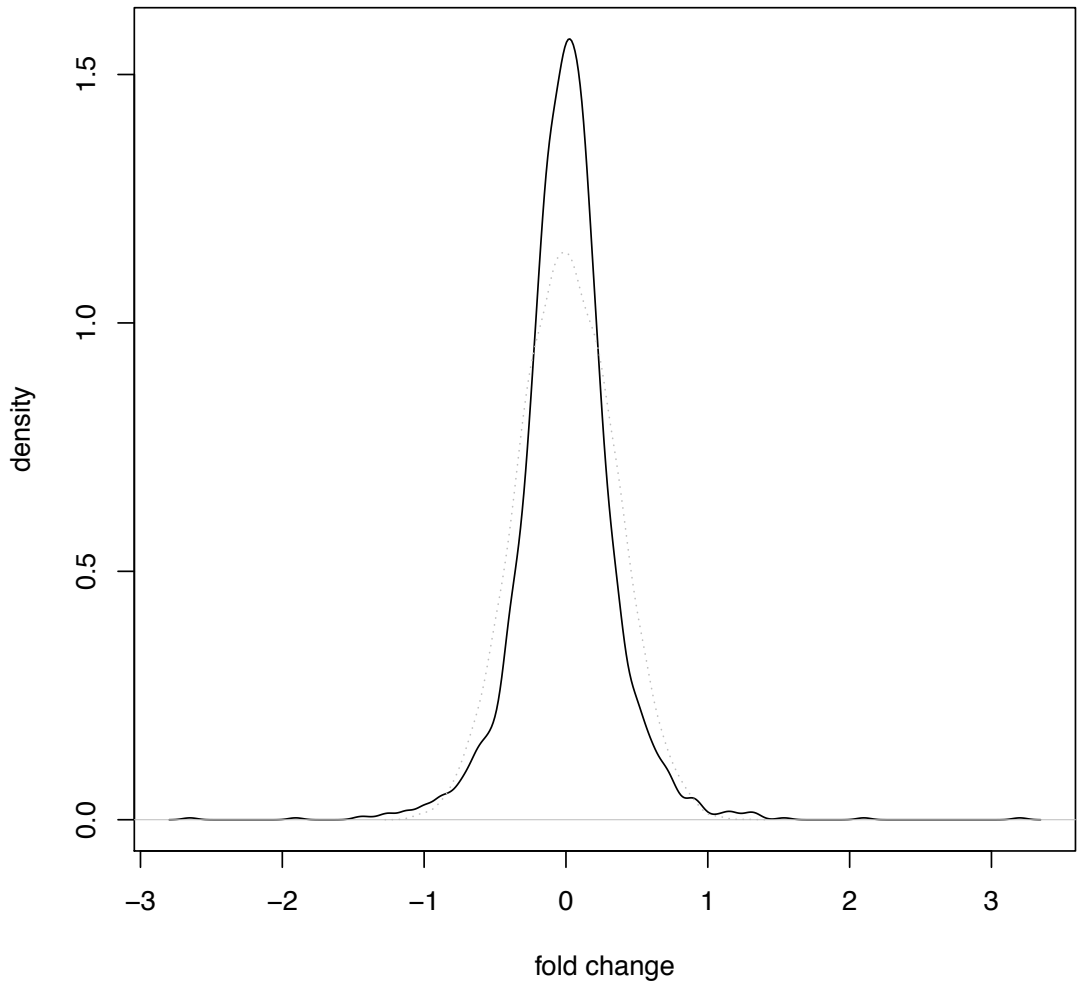


Top 20 factors in all the feature selection frameworks, sorted by the average value.

- **In vitro binding potential (SELEX) features don't include cell line information**

MSE	**Lasso 1se**	**SVR**	**RandomForest**
ChIPseq + SELEX	0.106	0.105	0.102
SELEX	0.111	0.108	0.107

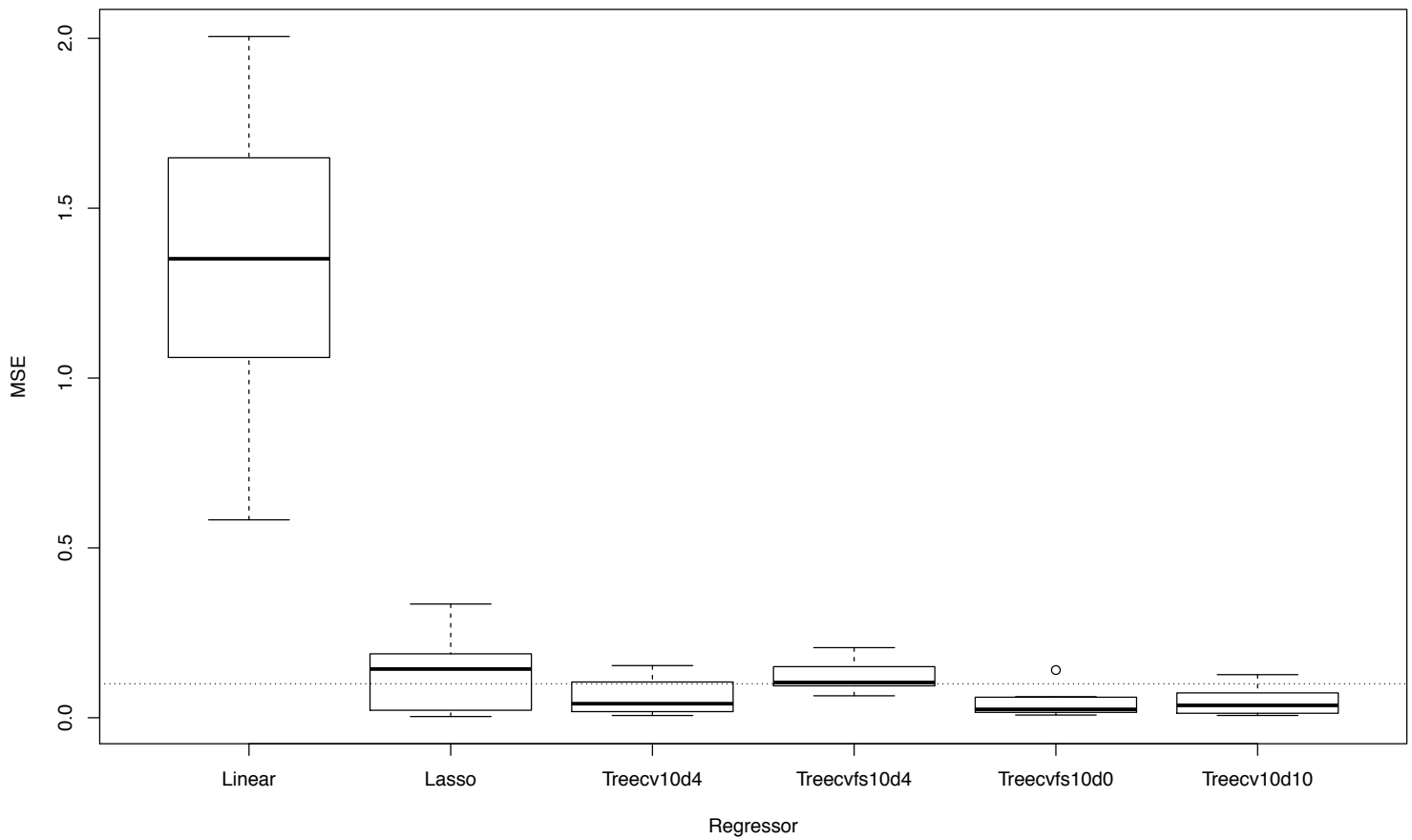
However, we need outlier sensitive regressor:



the target variable: log2 based fold change between mut and ref allele. In the cell paper, The lowest —log2skew— in emVar is 0.11. (or $|skew| = 1.08$ will be significant

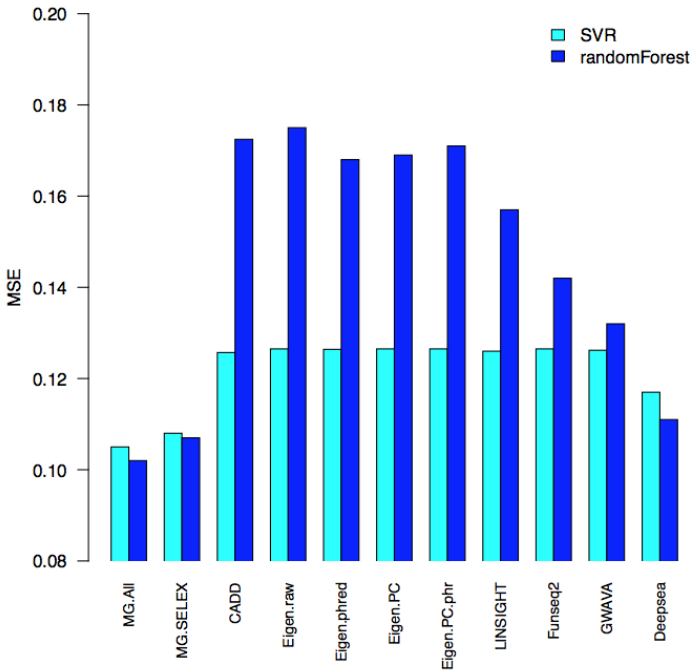
Adaboost

Adaboost are more outlier sensitive, ensemble a series of week regressors. The overfitting problem of tree-based algorithm caused: too many features and depth of tree, we tried forward selection using SVR: **ELK1 ,CREB3, IRF5, NKX6-1, SRF, H3k27me3, FEV, NHLH1, TEAD1**

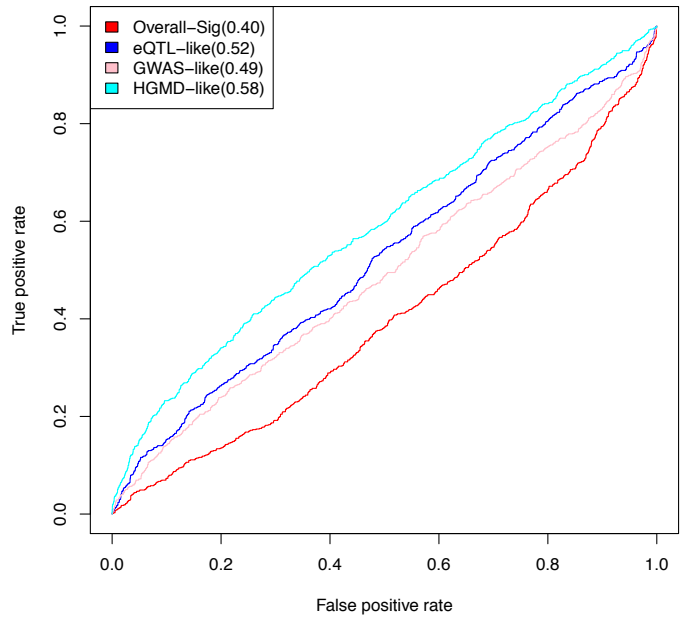


Comparison with other tools

No direct way to compare, train a 10-fold cross-validation SVR and RandomForest model using output from different tools

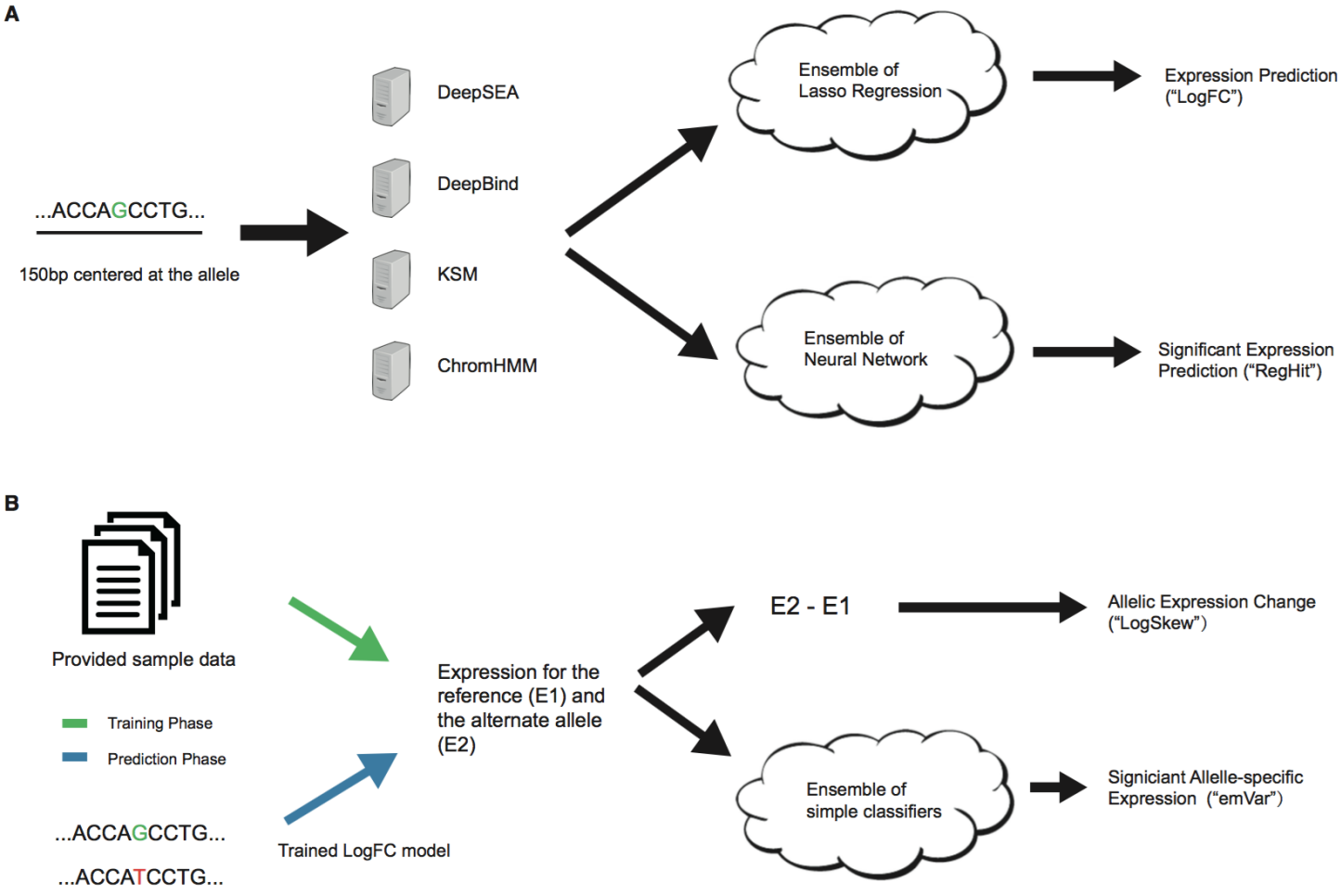


smaller is better



Part 3 Classification

State-of-the-art using the same dataset (CAGI 4)



State-of-the-art using the same dataset (CAGI 4)

Participant (Lab-Submission)	LogSkew Spearman corr.	emVar auPRC	emVar auROC
4 (EnsembleExpr)	**0.449760**	**0.452561**	**0.655261**
5-1	0.333893	0.409730	0.626850
Published state-of-the-art	Not Applicable	0.389	0.589
5-2	0.342004	0.369083	0.577220
7	0.007343	0.431639	0.562854
6-1	0.217845	0.345064	0.561953
6-2	0.190123	0.354726	0.561776
1-3	NaN*	0.311243	0.556499
1-1	NaN*	0.305258	0.550820
1-2	0.030243	0.295886	0.550048
2-3	-0.015476	0.303051	0.545206
1-5	0.056143	0.284863	0.541216
1-4	0.079049	0.293321	0.530856
3	0.030049	0.284356	0.511181
1-6	0.105376	0.286584	0.510103
2-2	-0.007377	0.249473	0.479746
2-1	-0.024347	0.234723	0.477301
2-5	-0.023092	0.233144	0.472651
2-6	-0.023092	0.233144	0.472651
2-4	-0.023092	0.233144	0.472651

*: every variant was assigned the same score, leading to incalculable Spearman correlations

A novel k-mer set memory (KSM) motif representation improves regulatory variant prediction

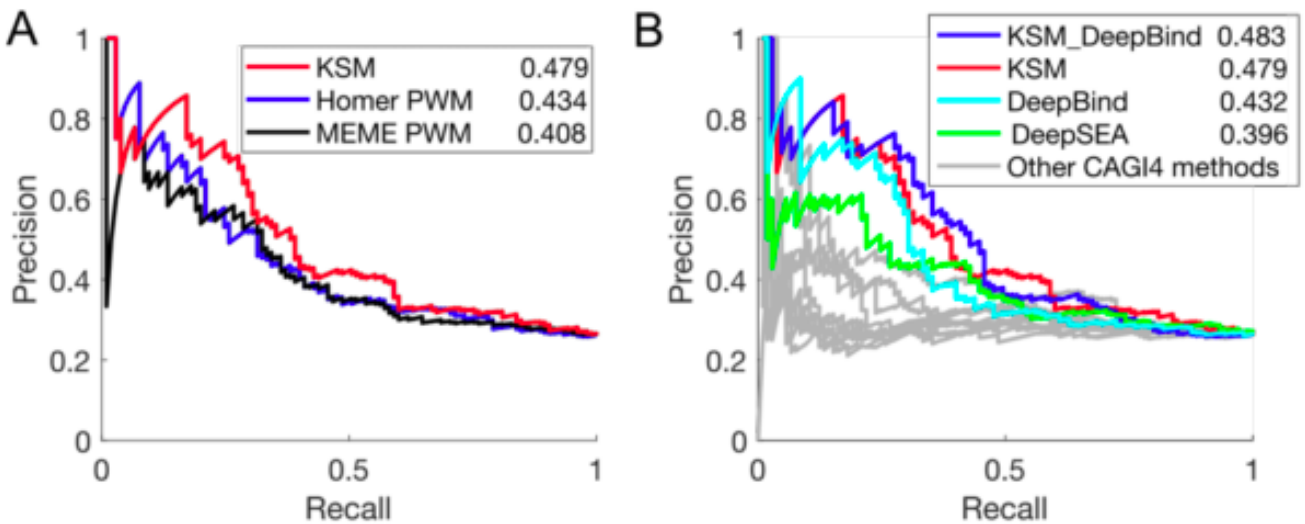
Yuchun Guo[1], Kevin Tian[1], Haoyang Zeng[1], Xiaoyun Guo[1], David K. Gifford[1]*

[1]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

* Corresponding author, gifford@mit.edu

Additional Footnotes:

Present address for Kevin Tian: Department of Computer Science, Stanford University, Stanford, CA 94305



AUROC = 0.668, AUPRC=0.479

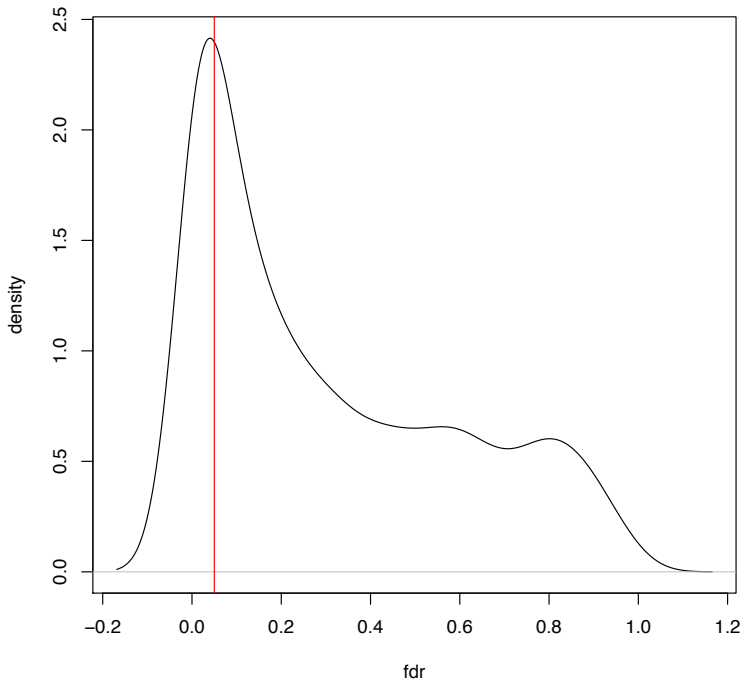
Features

Y ~ **GERP** + **TF+Hist** + **DHS** + **CAGE** + **Motif**

Significant

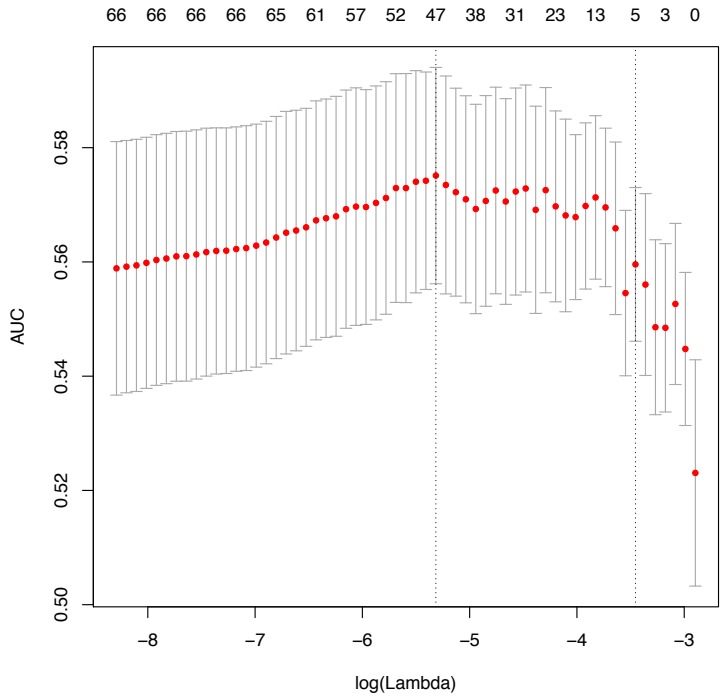
logSkew(Alt/Ref)

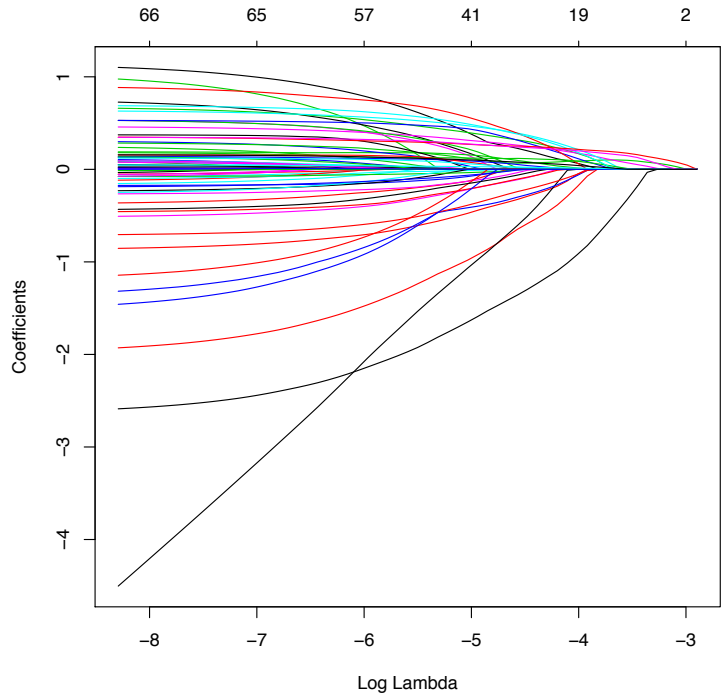
Tissue specific binary feature



FDR distribution

LASSO and Logistic Regression directly using features

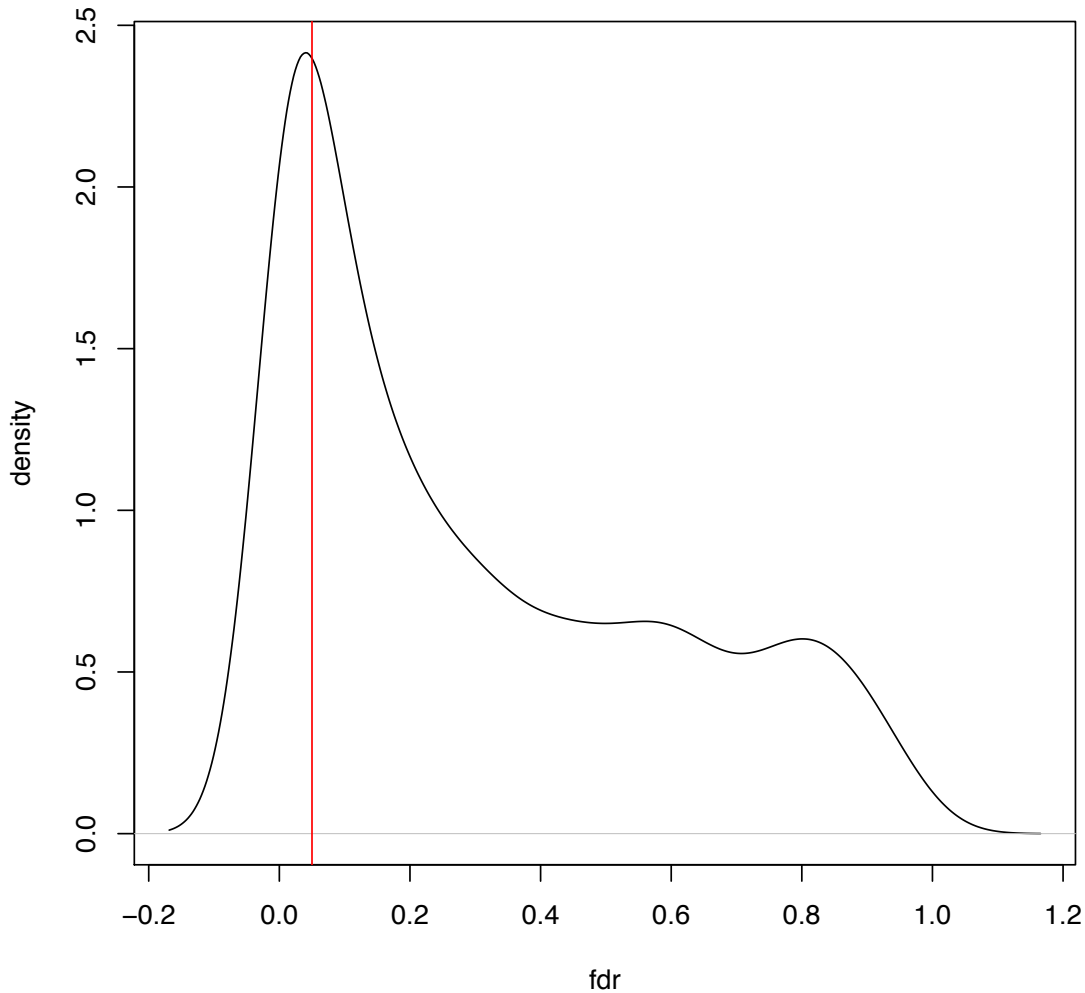




The AUC for classification is **0.5-0.6**

Definition of Positive and Negative dataset (PN learning)

FDR distribution of Log skewness for Ref/Alt
Positive dataset: FDR $<= 0.05$; and Negative dataset: FDR > 0.1



Transductive SVM

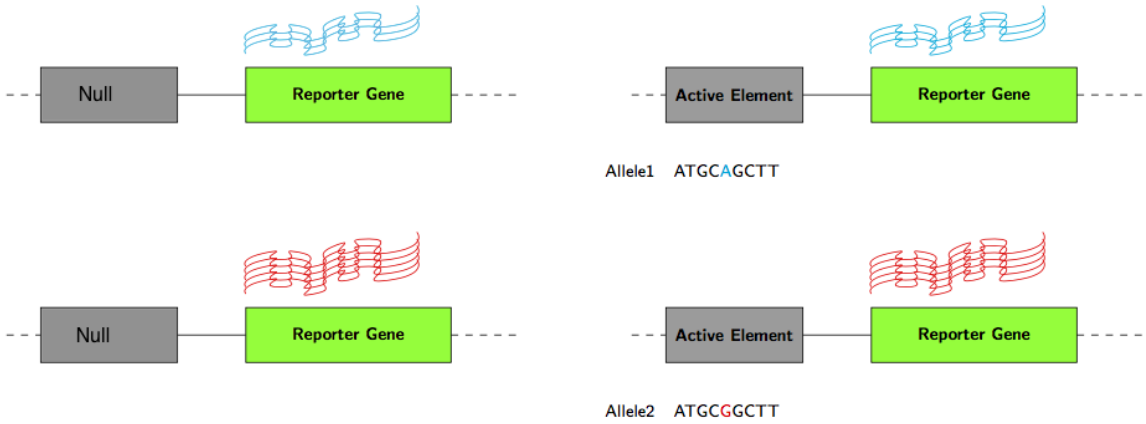
INPUT: $\mathcal{P}, \mathcal{U}, K =$ size of bootstrap samples, $T =$ number of bootstraps
OUTPUT: a score $s : \mathcal{U} \to \mathbb{R}$
Initialize $\forall x \in \mathcal{U},\ n(x) \leftarrow 0,\ f(x) \leftarrow 0$
for $t = 1$ to T **do**
Draw a bootstrap sample $\mathcal{U}_t$ of size K in $\mathcal{U}$.
Train a classifier f_t to discriminate $\mathcal{P}$ against $\mathcal{U}_t$.
For any $x \in \mathcal{U} \setminus \mathcal{U}_t$, update:

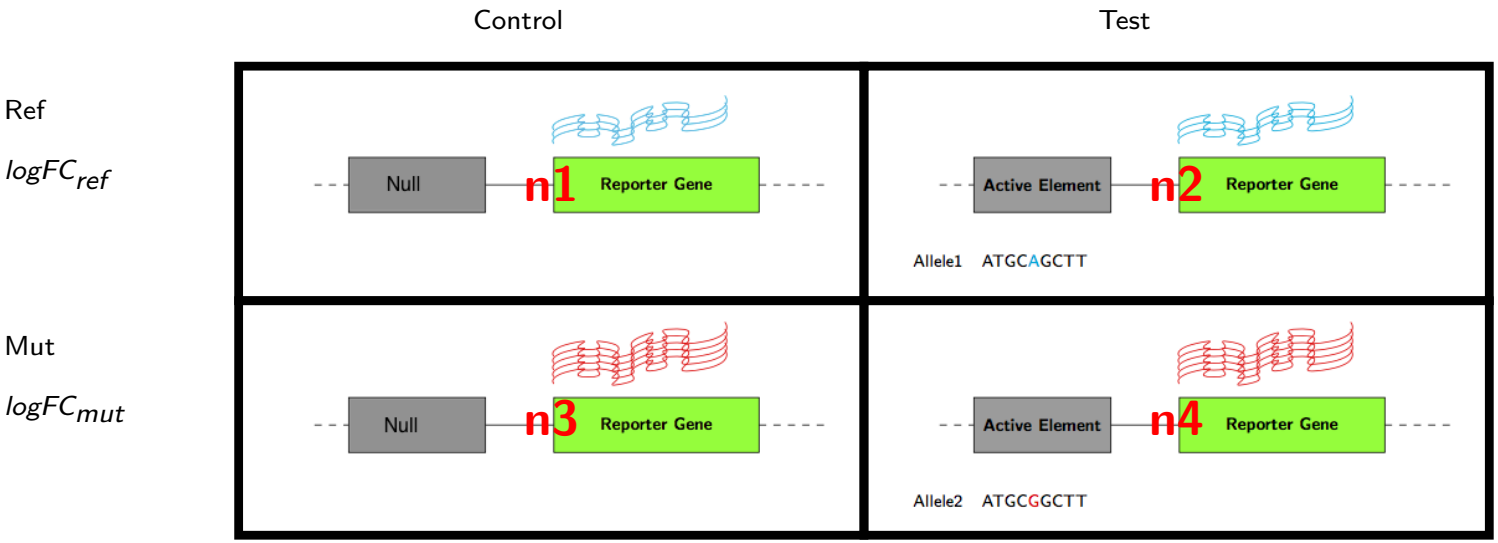
$$f(x) \leftarrow f(x) + f_t(x),$$

$$n(x) \leftarrow n(x) + 1.$$

end for
Return $s(x) = f(x)/n(x)$ for $x \in \mathcal{U}$

AUC: 0.6158824


Null
Reporter Gene
Active Element
Reporter Gene
Allele1 ATGCAGCTT
Null
Reporter Gene
Active Element
Reporter Gene
Allele2 ATGCGGCTT



In a 2x2 categorical analysis:

$$logSkew = log(odds) \approx Norm(log(odds), var(log(odds)))$$

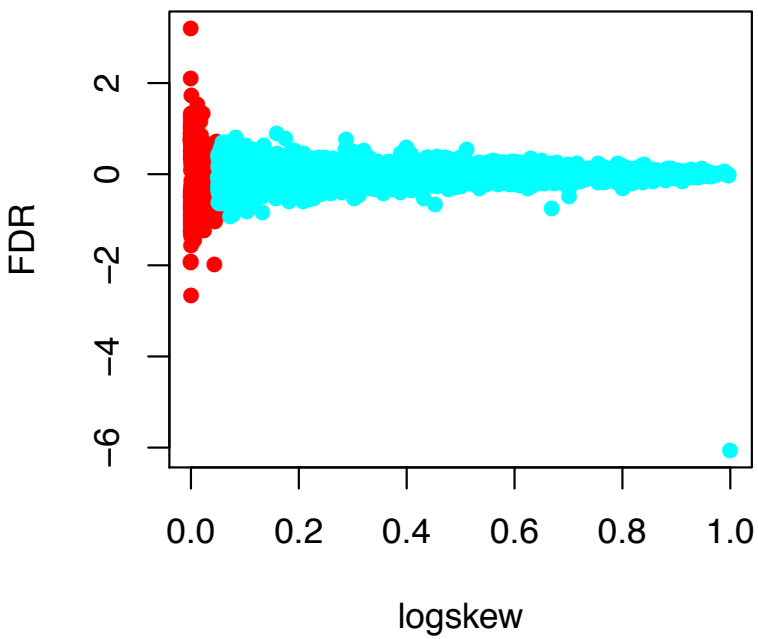
$$log(odds) = logFC_{mut} - logFC_{ref} = \log(\frac{n2}{n1} / \frac{n4}{n3})$$

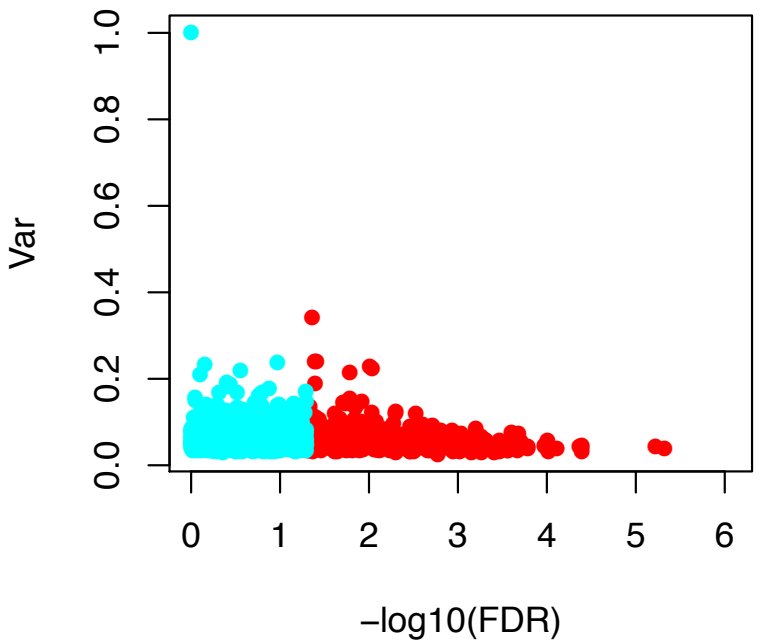
$$var(log(odd)) = \sqrt{\frac{1}{n1} + \frac{1}{n2} + \frac{1}{n3} + \frac{1}{n4}}$$

log FC is directly calculated from experiment count; log Skew rely on logFC

Both the log Skew and Var(log Odds) associate with the positive and negative dataset

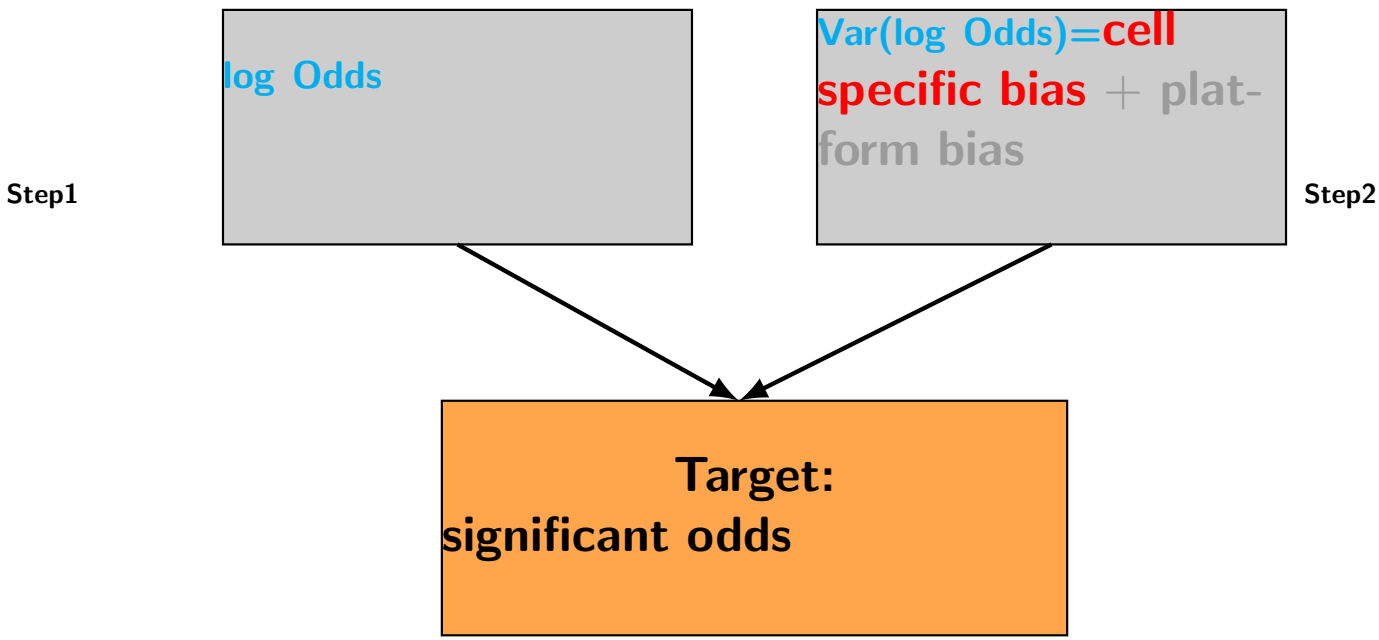
The original paper use DESeq2 to correct experiment count and get Log FC and then use Wald test to define emVar and non-emVar. The definition of positive and negative set is dispersion-awared



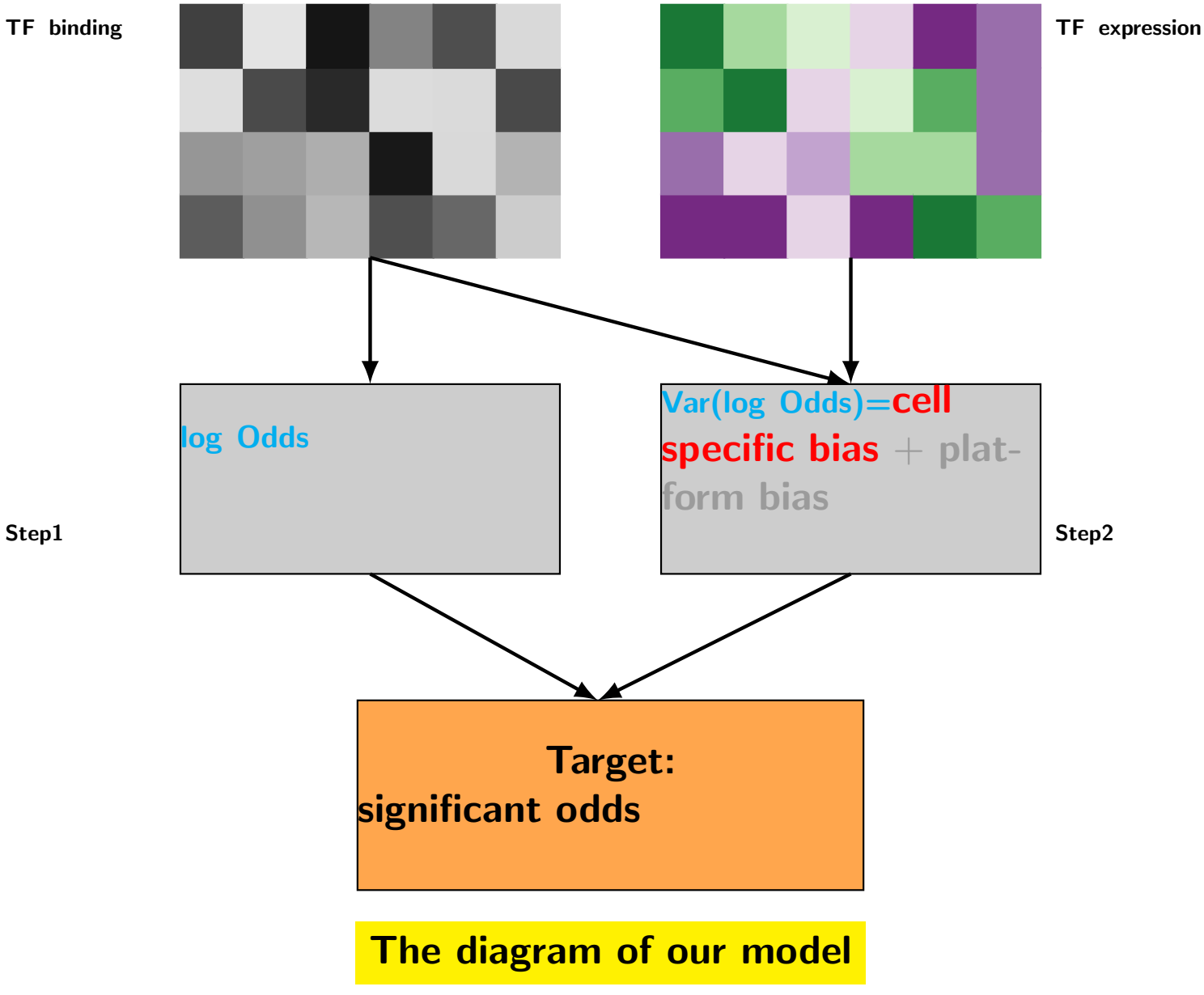


Target:
significant odds

The diagram of our model



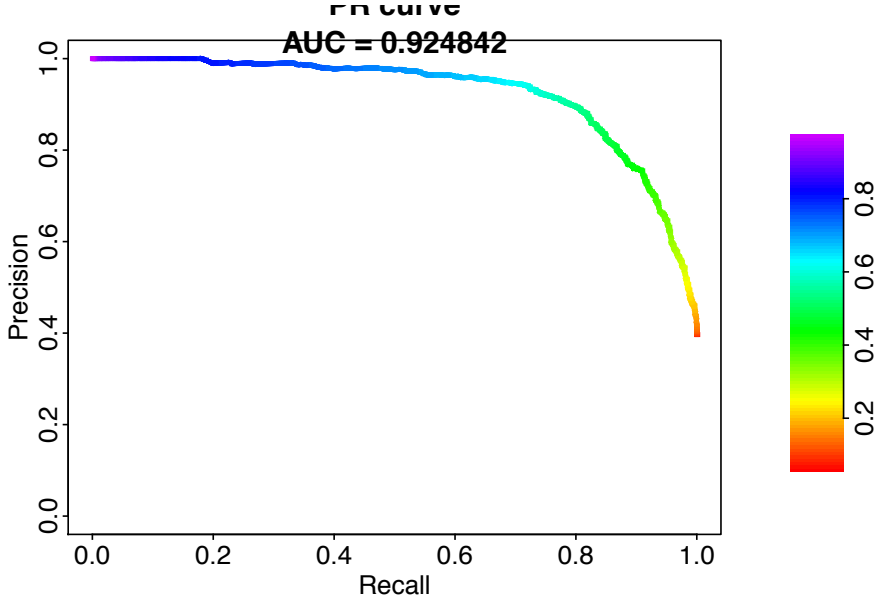
The diagram of our model

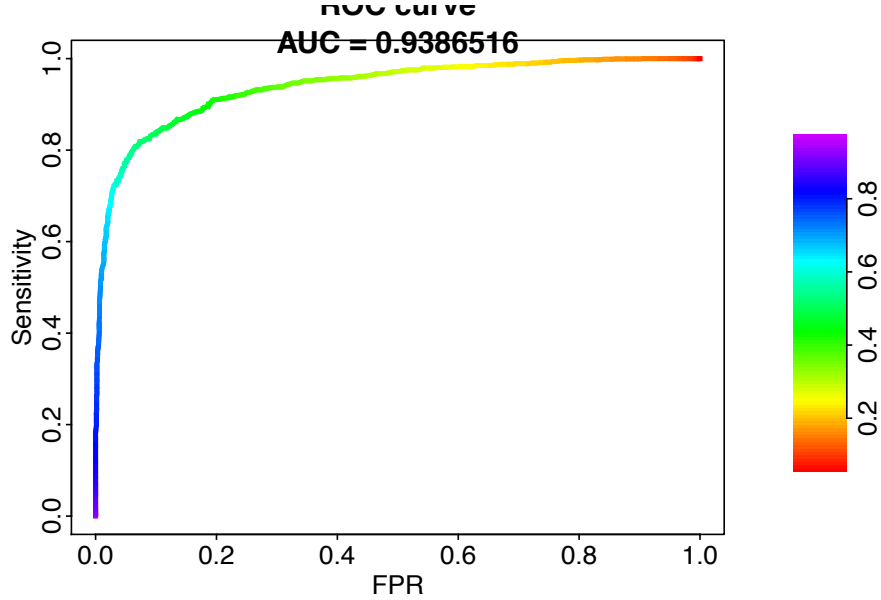


The diagram of our model

Step 1: The log FC classification

Directly from logFC, not log Odds (log Skew)



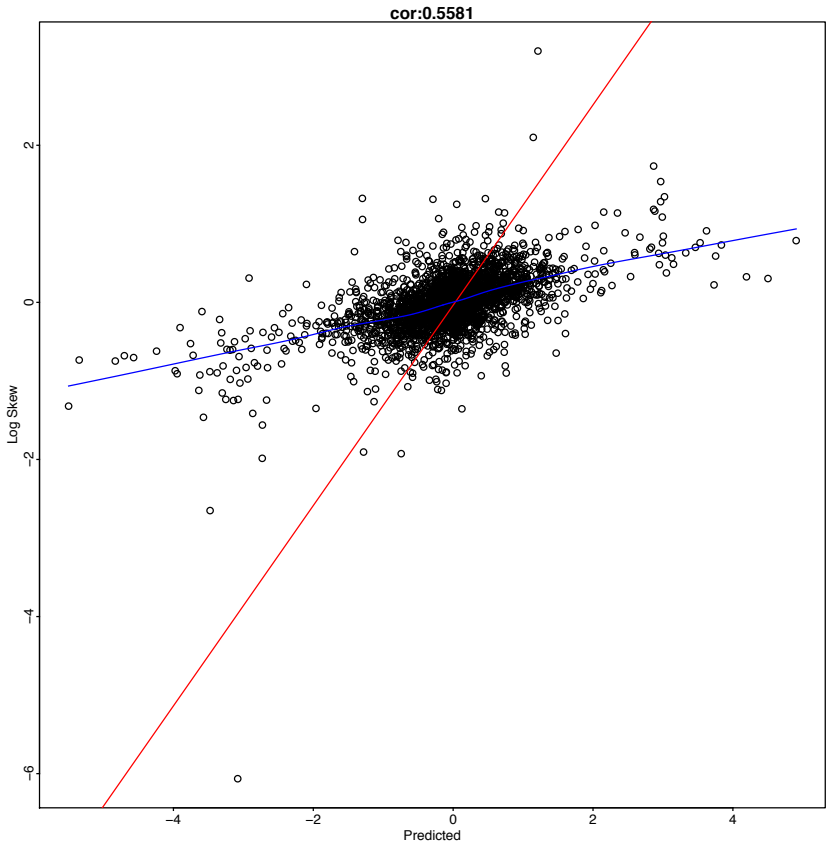


Define positive and negative using Log2FC for wild type and mutant element. Then train model to do classification.

The motif binding profile can easily identify the elements with high expression regulation effect with very high AUC and AUPR (10 fold cv).

logSkew correlate with predicted log Odds

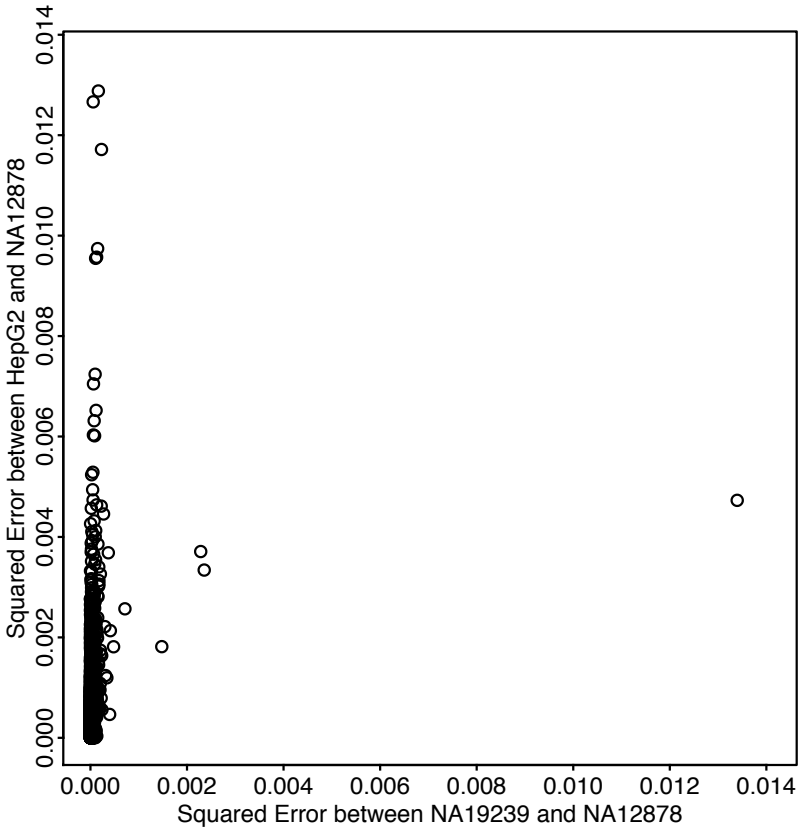
The predicted log Odds is defined as: $log2odds = log2(\frac{p_{mut}}{1-p_{mut}}) - log2(\frac{p_{ref}}{1-p_{ref}})$



Step2: Cell specific Bias (CSB)

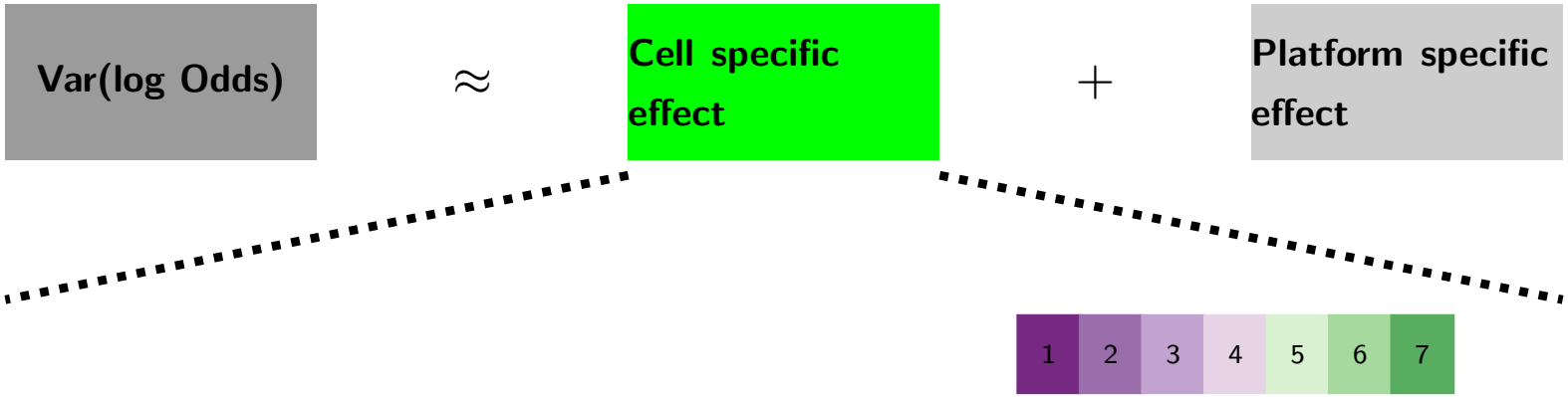
We define a binding effectaware cell specific bias feature (CSB):

Var(log Odds) $\approx$ **Cell specific effect** $+$ **Platform specific effect**



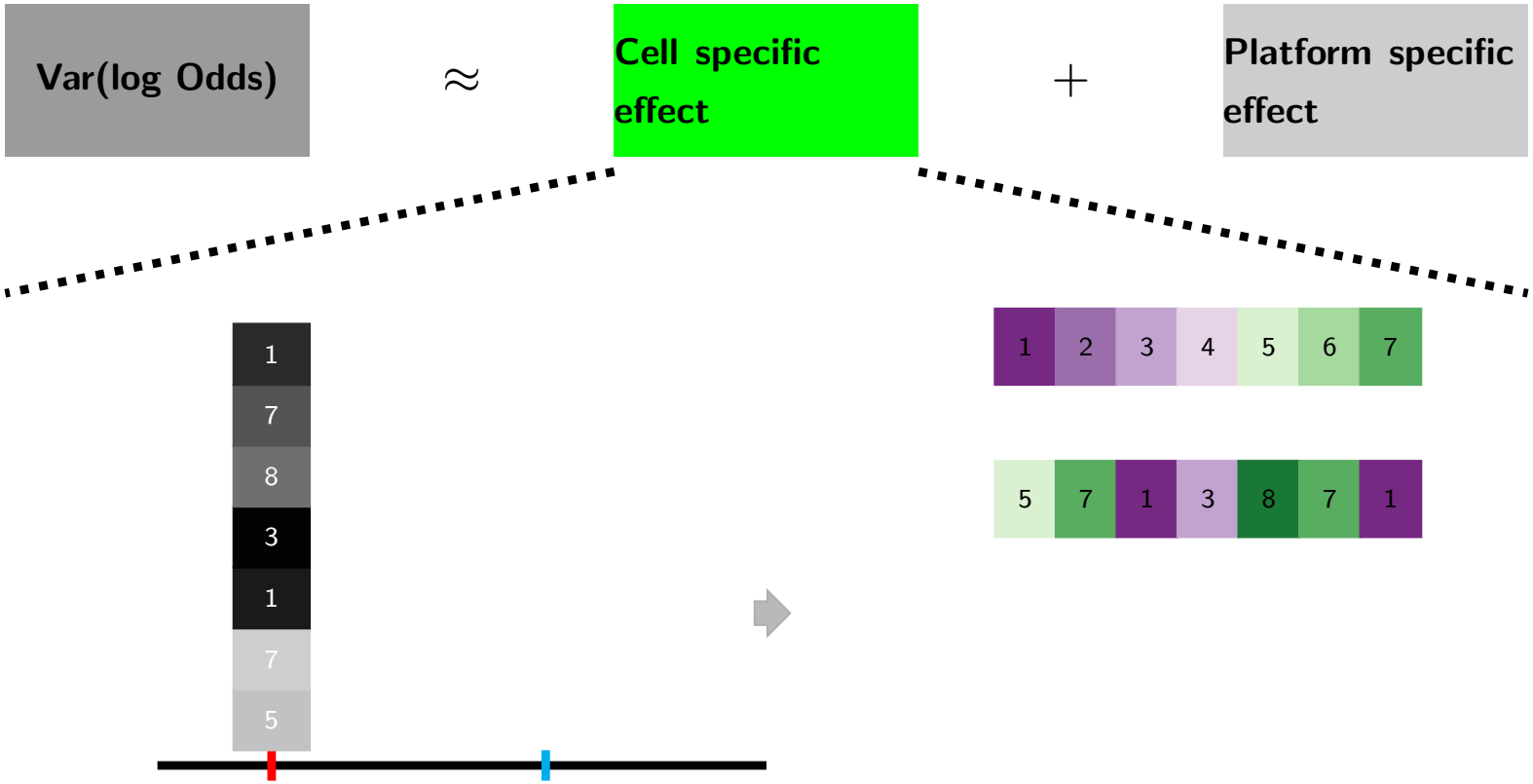
Step2: Cell specific Bias (CSB)

We define a binding effectaware cell specific bias feature (CSB):



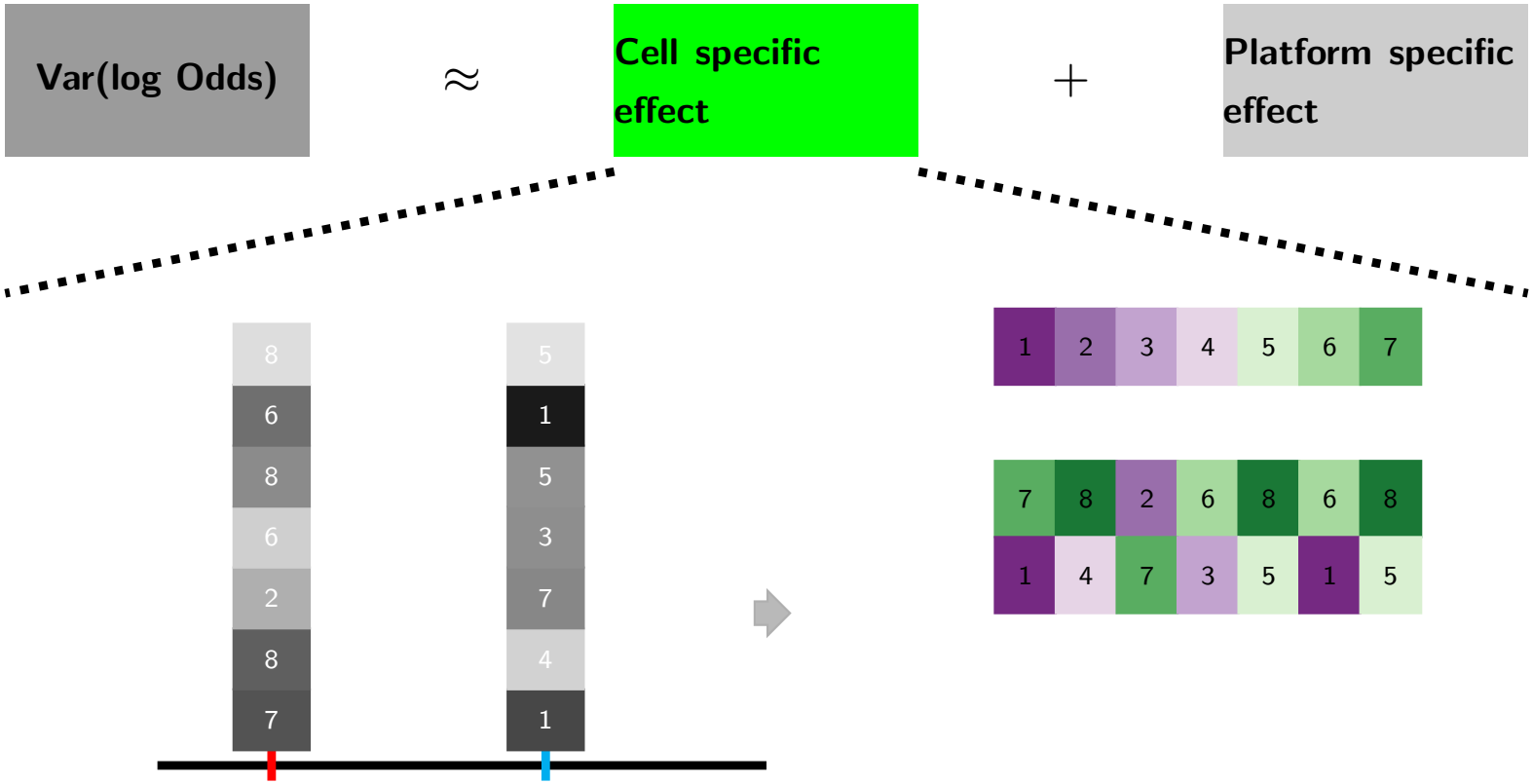
Step2: Cell specific Bias (CSB)

We define a binding effectaware cell specific bias feature (CSB):



Step2: Cell specific Bias (CSB)

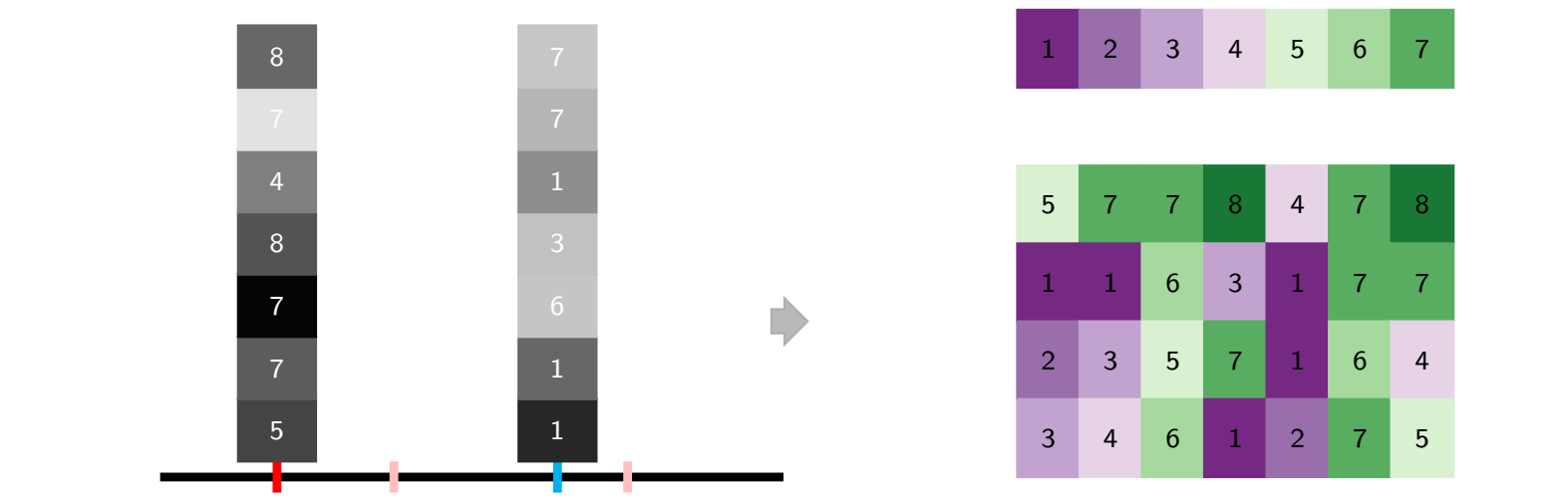
We define a binding effectaware cell specific bias feature (CSB):



Step2: Cell specific Bias (CSB)

We define a binding effectaware cell specific bias feature (CSB):

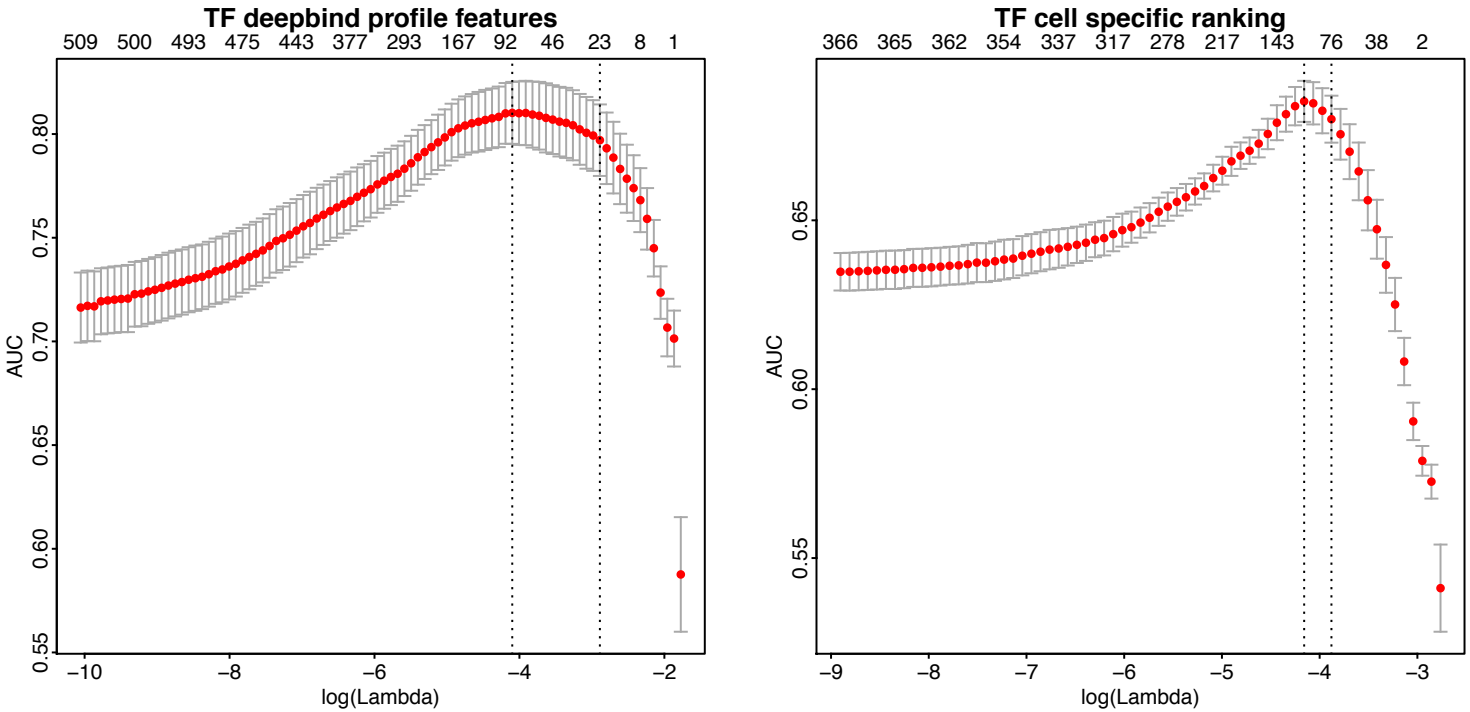
Var(log Odds) $\approx$ **Cell specific effect** $+$ **Platform specific effect**



TF binding $+$ TF expression

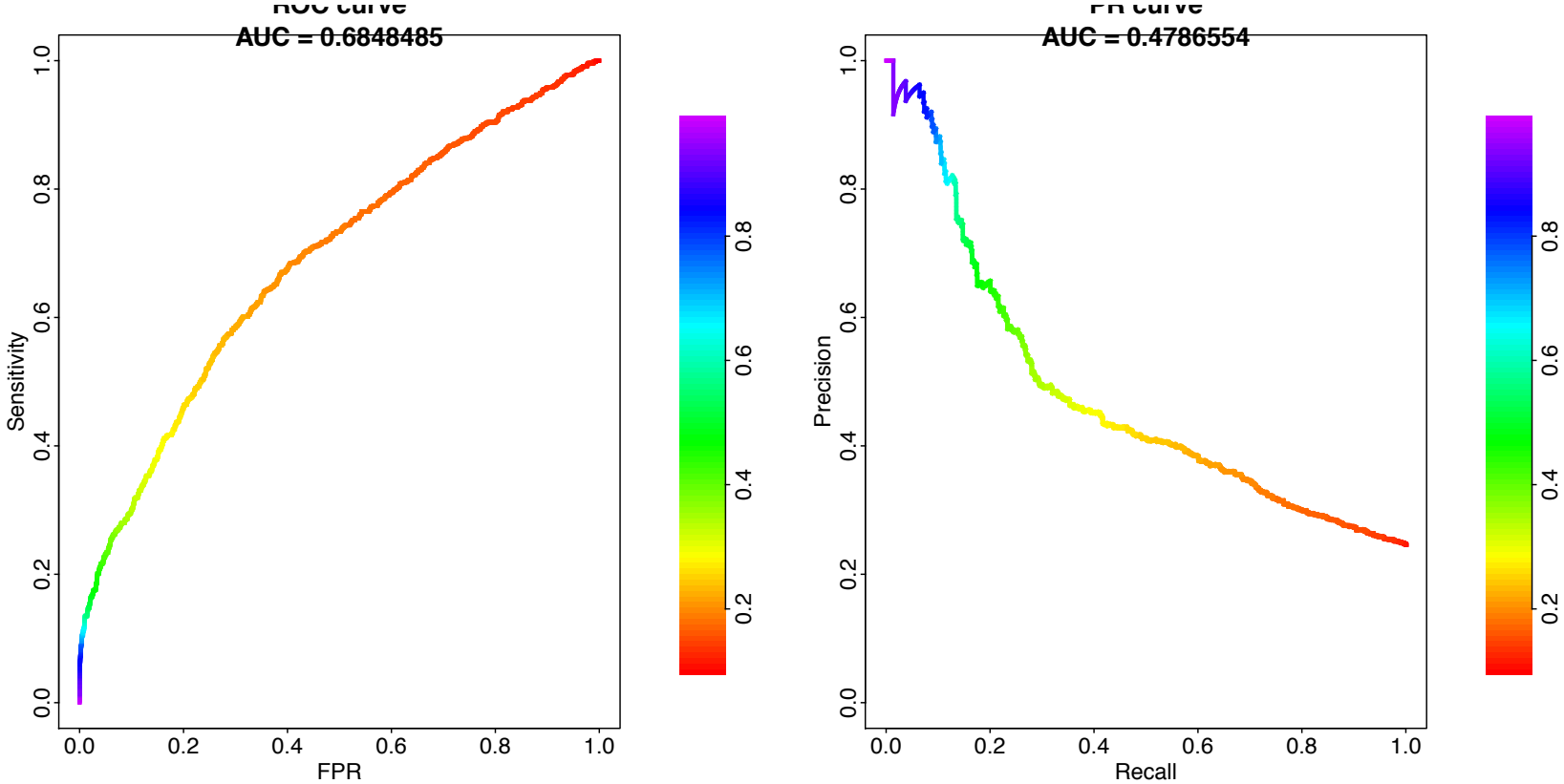
Learning Cell specific bias using TF binding and expression features

Both regression and classification were tried, but we use classification by taking out the two extreme quantile of response to define positive and negative dataset.



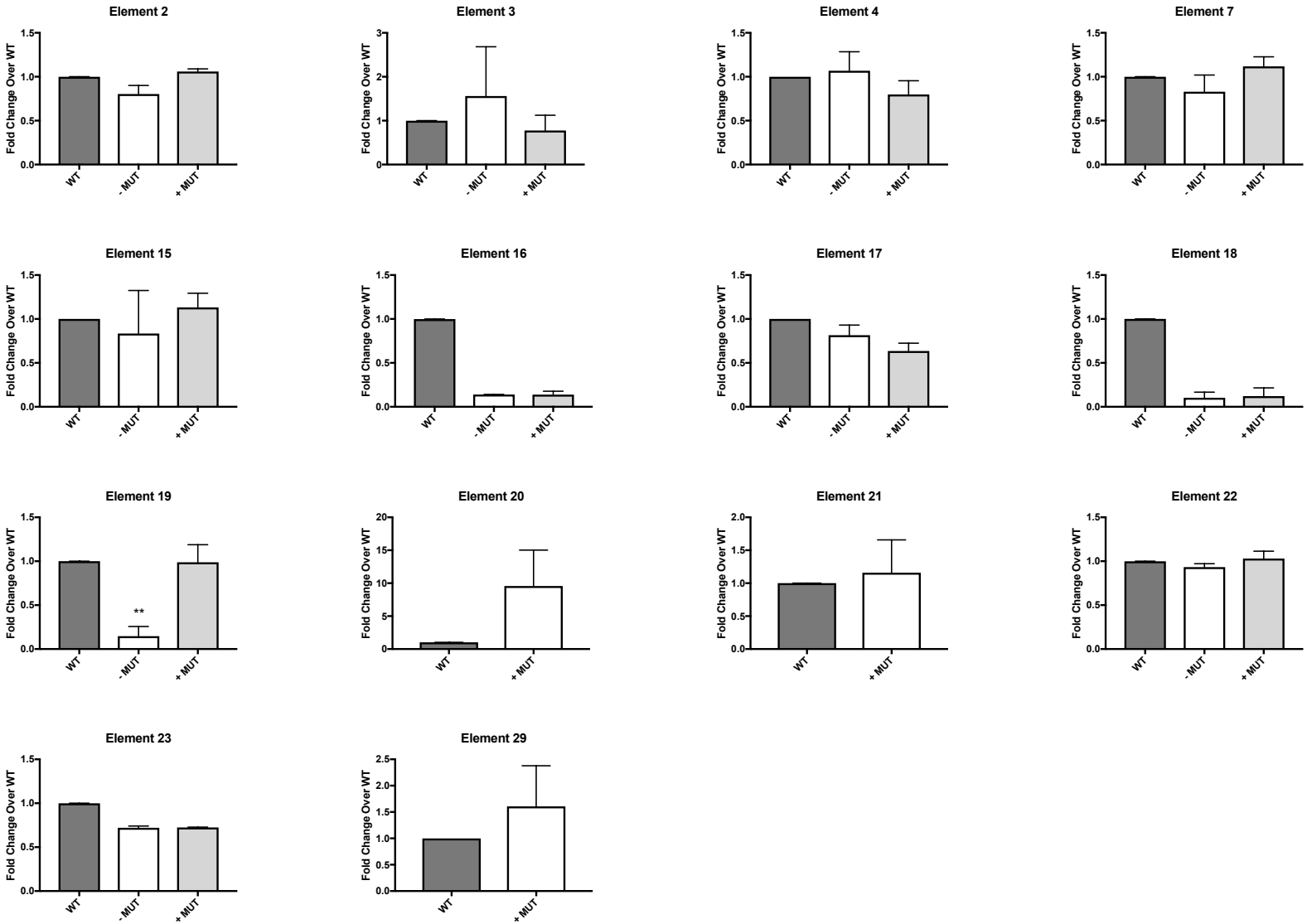
TF binding motif still be the best predictor, TF expression ranking is better then random.

Last Step: Lasso using predicted log Odds and CSB



AUCROC: 0.685, AUPRC=0.479 better than the-state-of-the-art (AUROC = 0.668, AUPRC=0.479)

Part 4: Experiment validation



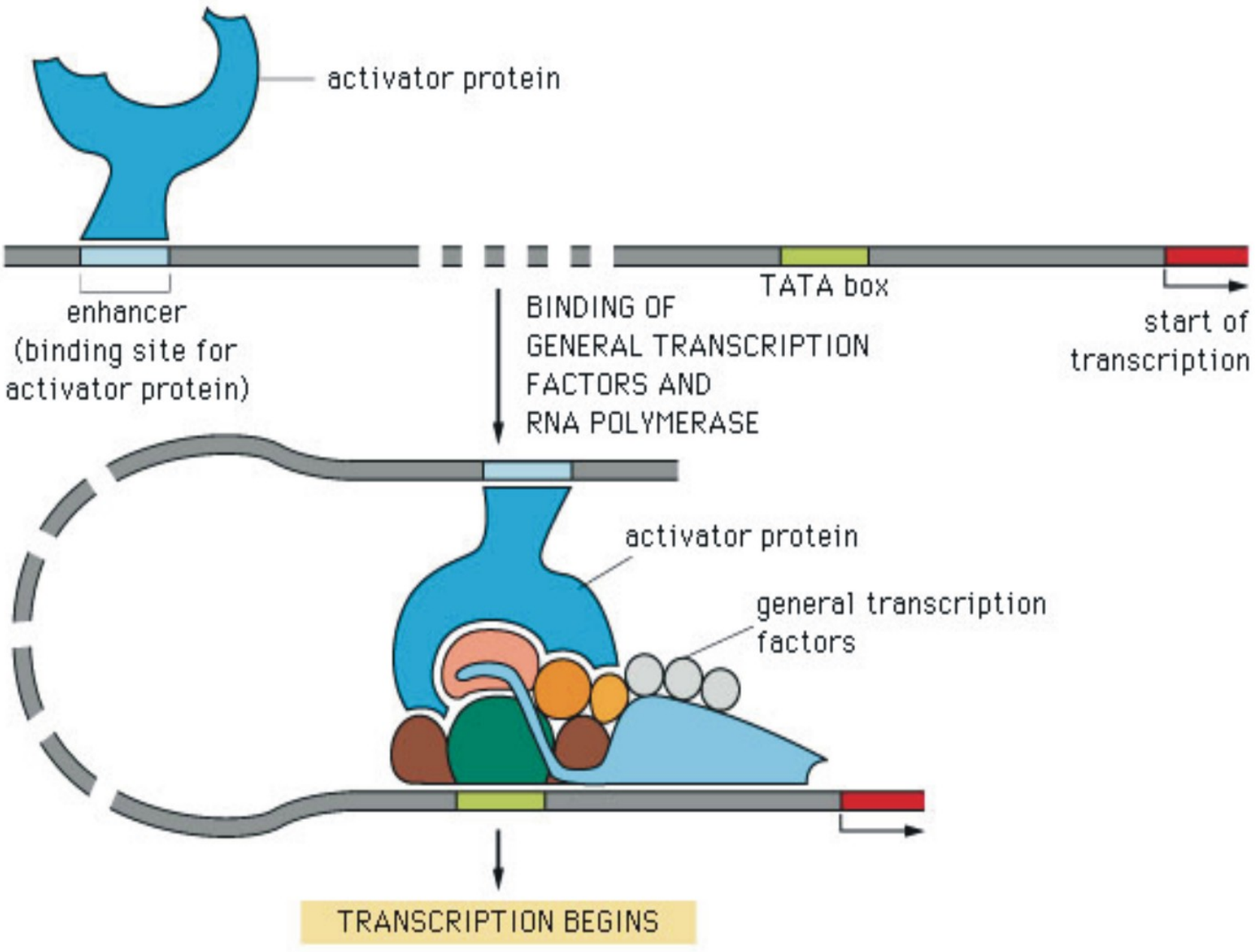
The prediction of Log FC has high accuracy but log2 Odds is not well predicted. Morevoer, the luciferase assay results have very high noise and dispersion.

Conclusion

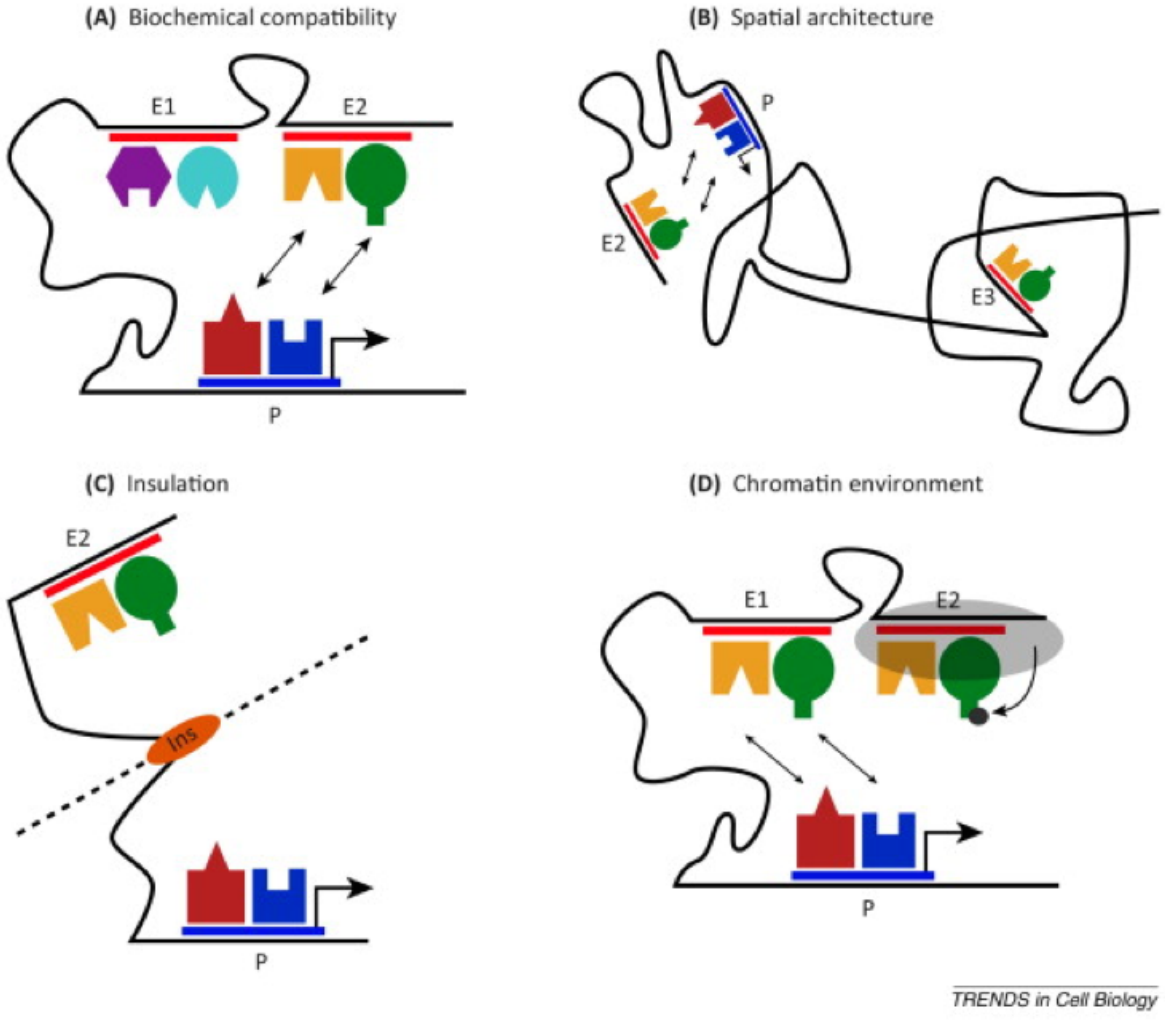
- Transcription factor binding is the most important feature in both regression and classification models
- The experimental precedure of reporter gene assay indicates the genomic context including chromatin status might not play indispensible role in the regulatory results, but cell specific TF binding and expression still have contributions.
- Just use TF binding can preciesely predict LogFC.
- The target variable for classification (significant change between mut_ref, need statistics analysis and cutoff) is not directly reflected from the experiment but some statitical analysis that may further introduce bias.
- Another dataset issue is the training set is not representable for the popluation set.

If still have time, then go to **ENGINE**

Biological of enhancer gene linkage

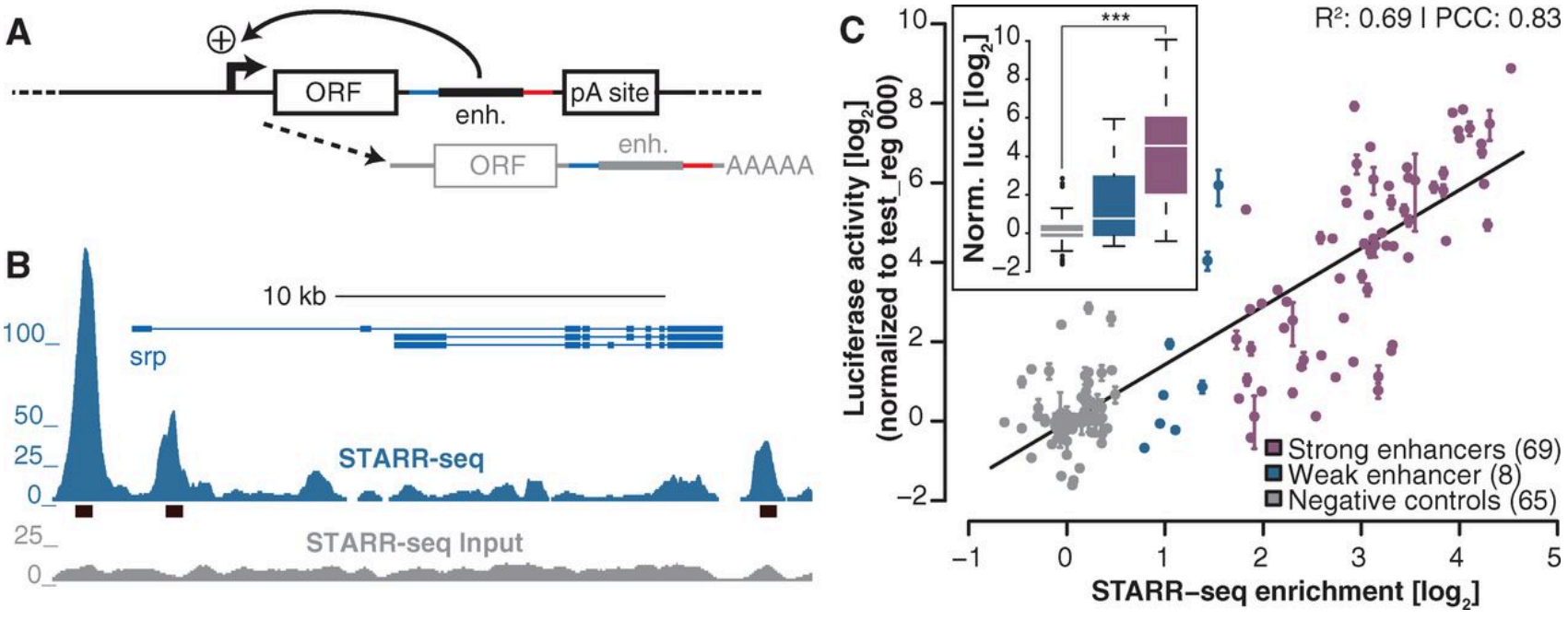


Biological of enhancer gene linkage



Classic problem: enhancer-promoter interaction. Biological compatiblity(sequence feature and motif); spatial compatibility (3d interaction); local environment (epigenomic marks)

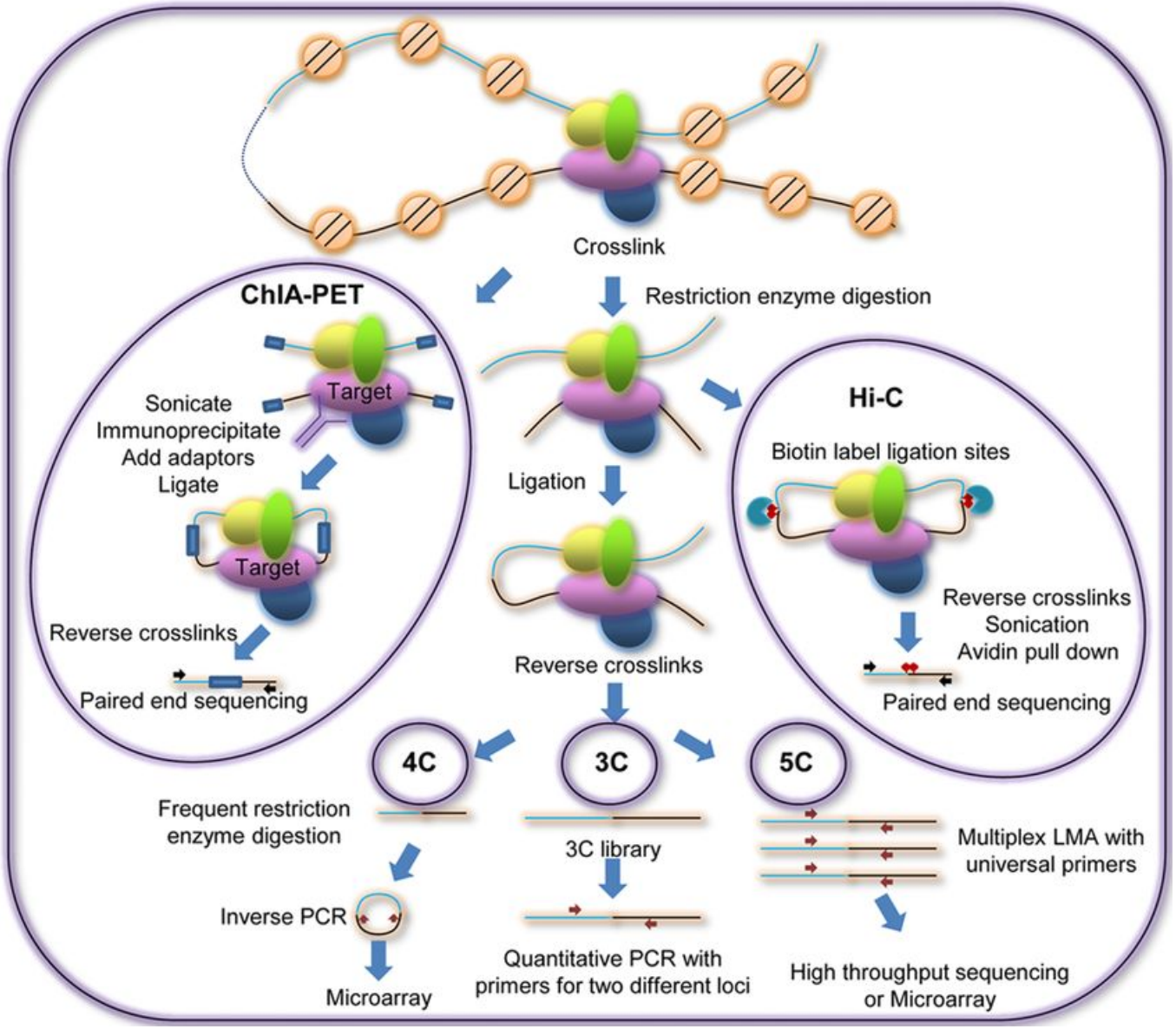
STARR-Seq: **enhancers can function independently of their relative positions.**



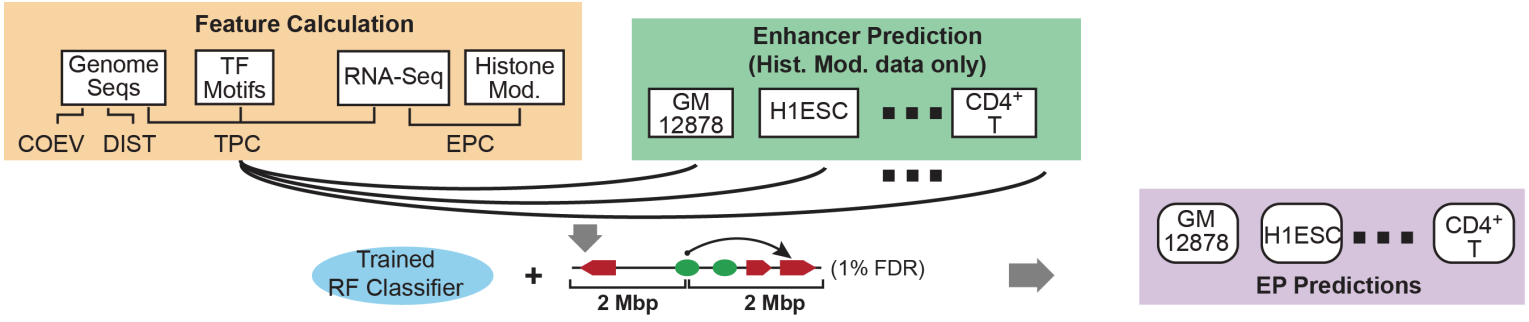
Enhancer can be very close to a gene(target)/in a gene, and also can be far from a target gene(distal enhancer), how to know their target?

3D genome techniques

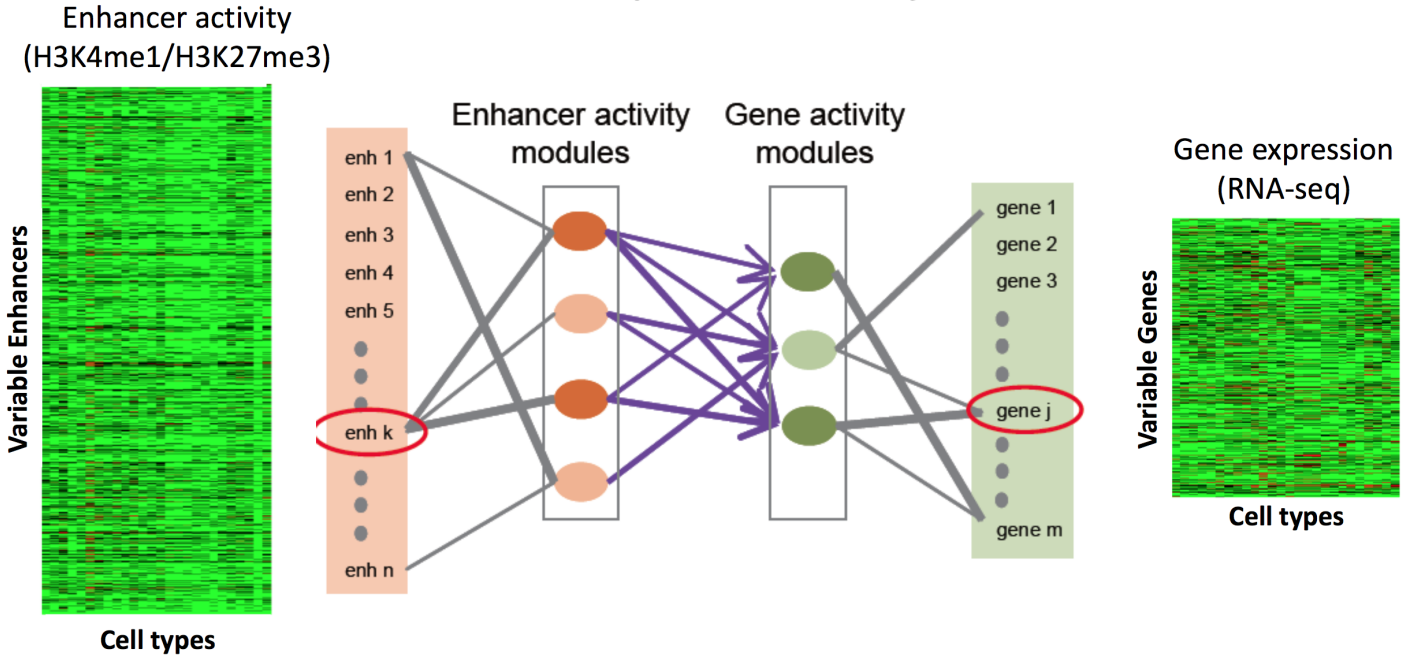
Enhancers, esp. distal enhancers, may need 3d chromosome structure to activate its target gene.



IM-PET: Consider information from 3D gnome interactions, DIST(distance) constrain is a triky feature, boosting AUC from 0.7+ to 0.9+.



LDA:a mixed membership method, didn't use information from 3d genome interaction, and reply on predifined enhancer region.



Sequence-based PromoterEnhancer Interaction with Deep learning(SPEID)

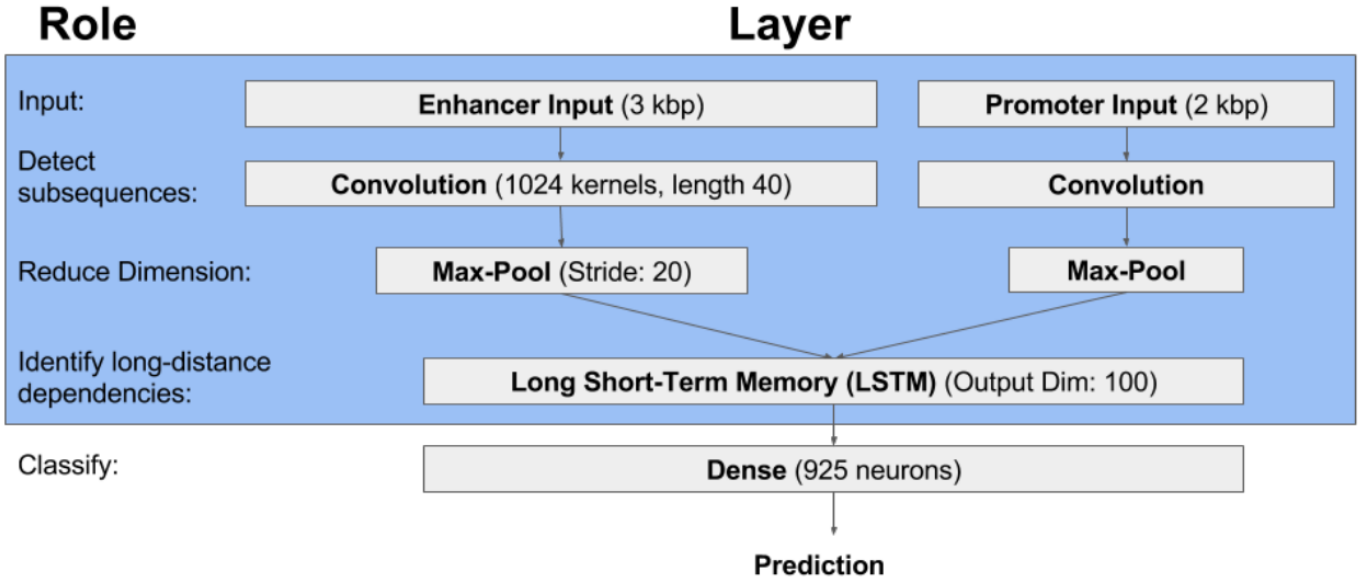
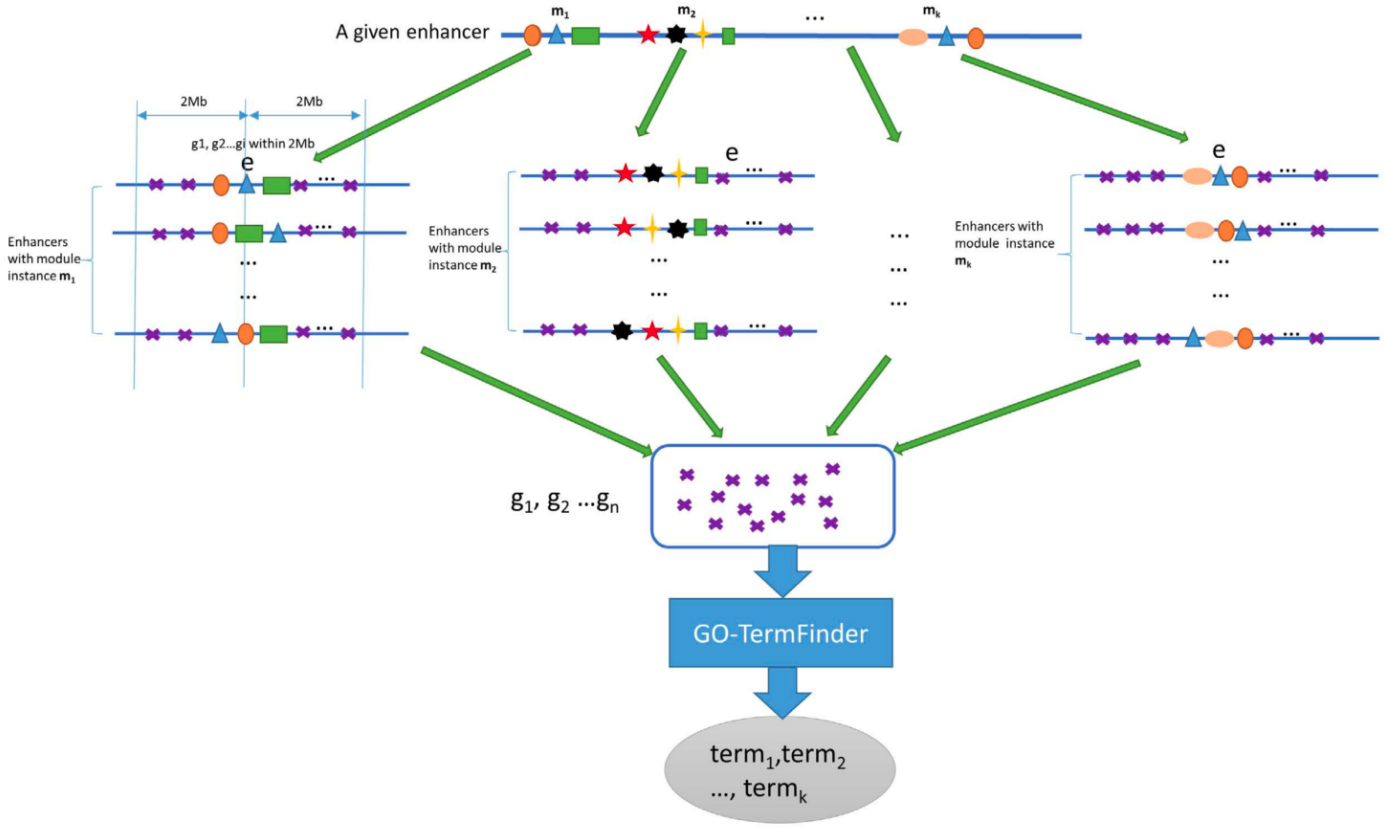


Figure 1: Diagram of our deep learning model SPEID.

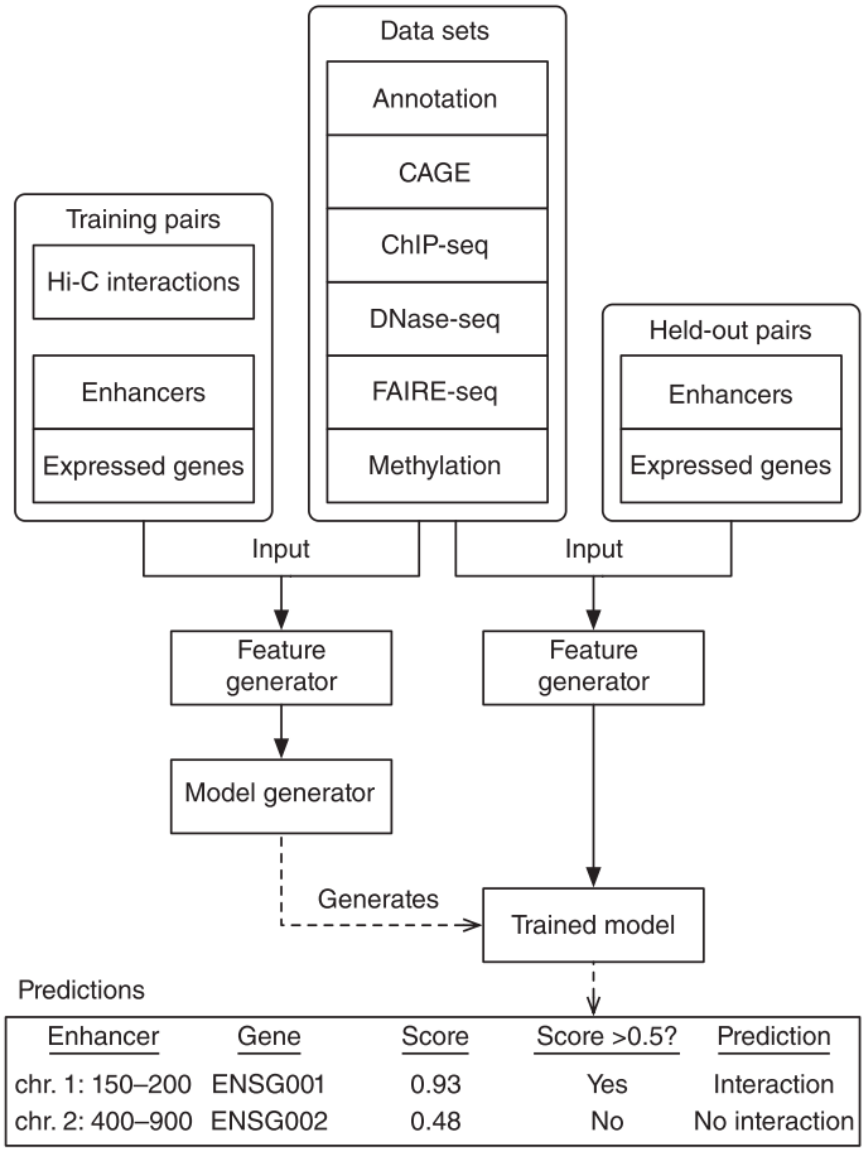
Sequence information alone can do prediction very well

PETModule: a motif module based approach for enhancer target gene prediction



Distance is the most important.

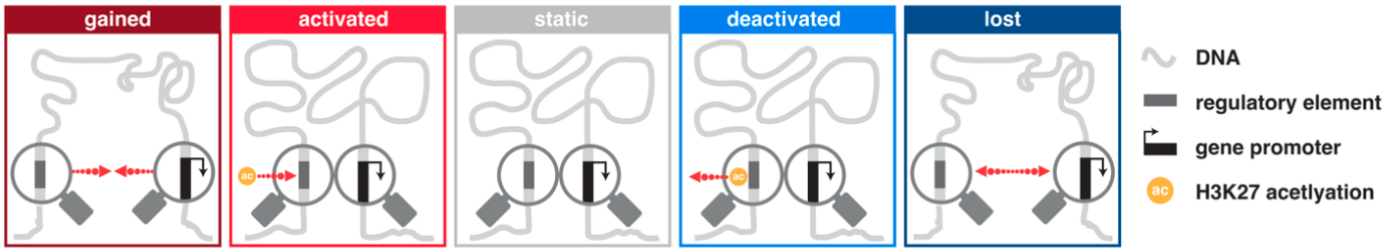
Enhancer promoter interactions are encoded by complex genomic signatures on looping chromatin



Data source in summary:

name	Source code	Enh-pair train	prediction	year
IM-PET	yes	No, but desc	yes	2014
PETModule	yes	No	Yes(in 4d genome)	2016
PreSTIGE	No	No	yes	2014
TargetFinder	Yes	Yes	Yes	2016
SPEID	yes	Raw, same as TF	weight	2016
JEME	Yes	K562	Yes	2017
LDA??	Yes	No	Yes	??

Dynamics



Expression and function of genes correlate with dynamic loop type and distal chromatin state

Motivations and structures

1. **Data source cross validation and comparison for positive and negative dataset.**
Papers use HiC, or ChIA-PET, or Fantome, or kind of combined to define positive dataset, which will affect the negative set definition. How to utilize and combine these dataset to get a reliable (related to positive dataset) and complete(close to complete, related to negative dataset)

2. **Connection between 3D interaction and enhancer target regulation.** Some interactions related to enhancer target regulation, the others not, what is the connection between 3d interaction and distal enhancer target regulation? We will focus on the comparison and explore the differences between structural and regulation interaction, and stable and dynamics.

3. **MultiClass learning and comparison with the-State-of-the-art.**
The traditional way is to define negative dataset from all position non-interacting pairs and it is limited: 1) the interaction is not randomly happened but dynamics, from the practical, it is not just tell the positive dataset from negative dataset from the genomic context; 2) machine learning can only learn the largest deviation between positive and negative dataset, which is bias if we will not known the machanism, and how many parts or elements get involved. So here need a multiple class learning, not only include a positive and negative dataset

4. **Reinforcement learning to study the possible enhancer target dynamics.** Even we have a multiclass learning, there is have a question left, how this happened and how the dynamics happened? Loop extrusion? We setup a deep reinformce learning algorithms to study the potential mechanism underling

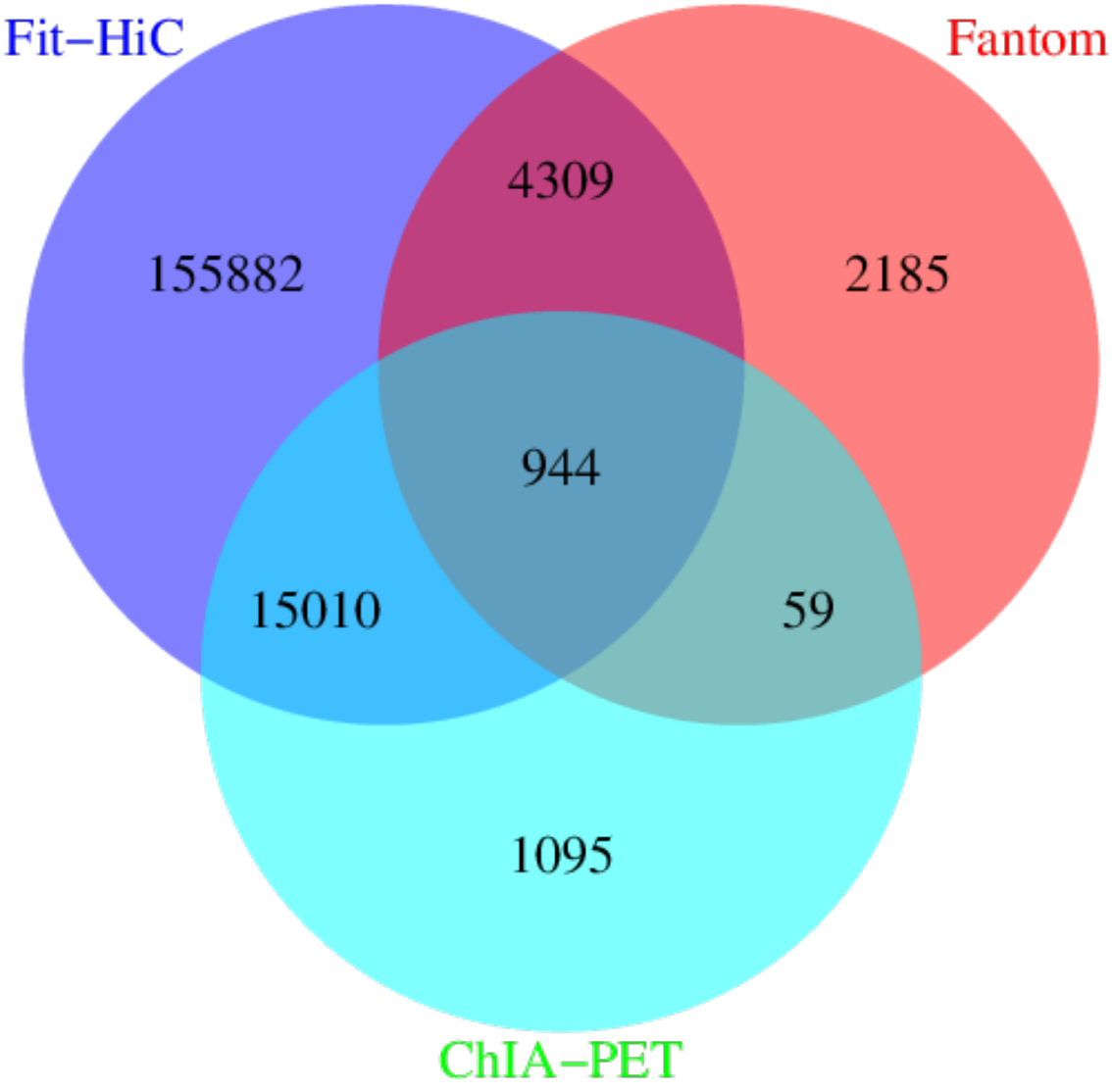
5. **Downstream analysis.**
Given the above, we want to further investigate more in the downstream analysis including network analysis, cellcycle or differential, supper enhancer , variants or other related analysis

Part 1

Papers use HiC, or ChIA-PET, or Fantome (correlation), or kind of combined to define positive dataset, which will affect the negative set definition. How to utilize and combine these dataset to get a reliable (related to positive dataset) and complete(close to complete, related to negative dataset?

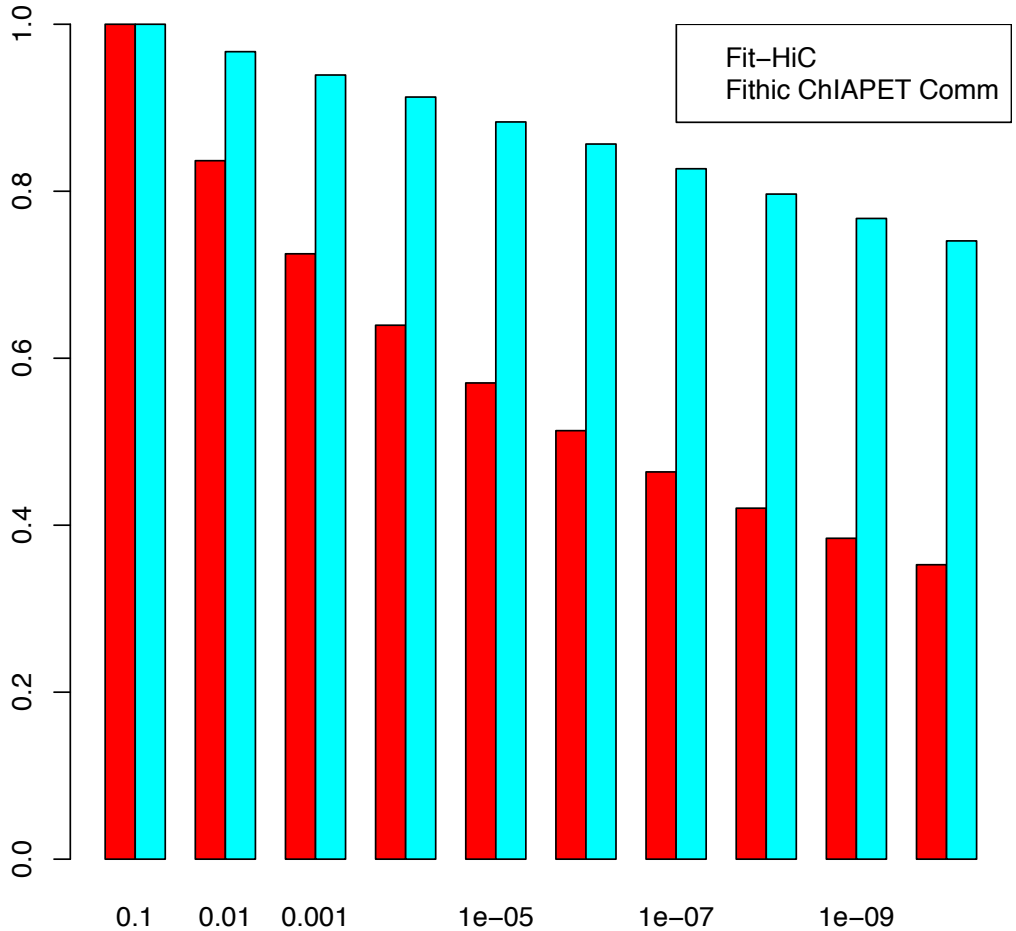
Summary of Hic, ChIA-PET and Fantom

Overlap of Enhancer promoter pair using different datasets

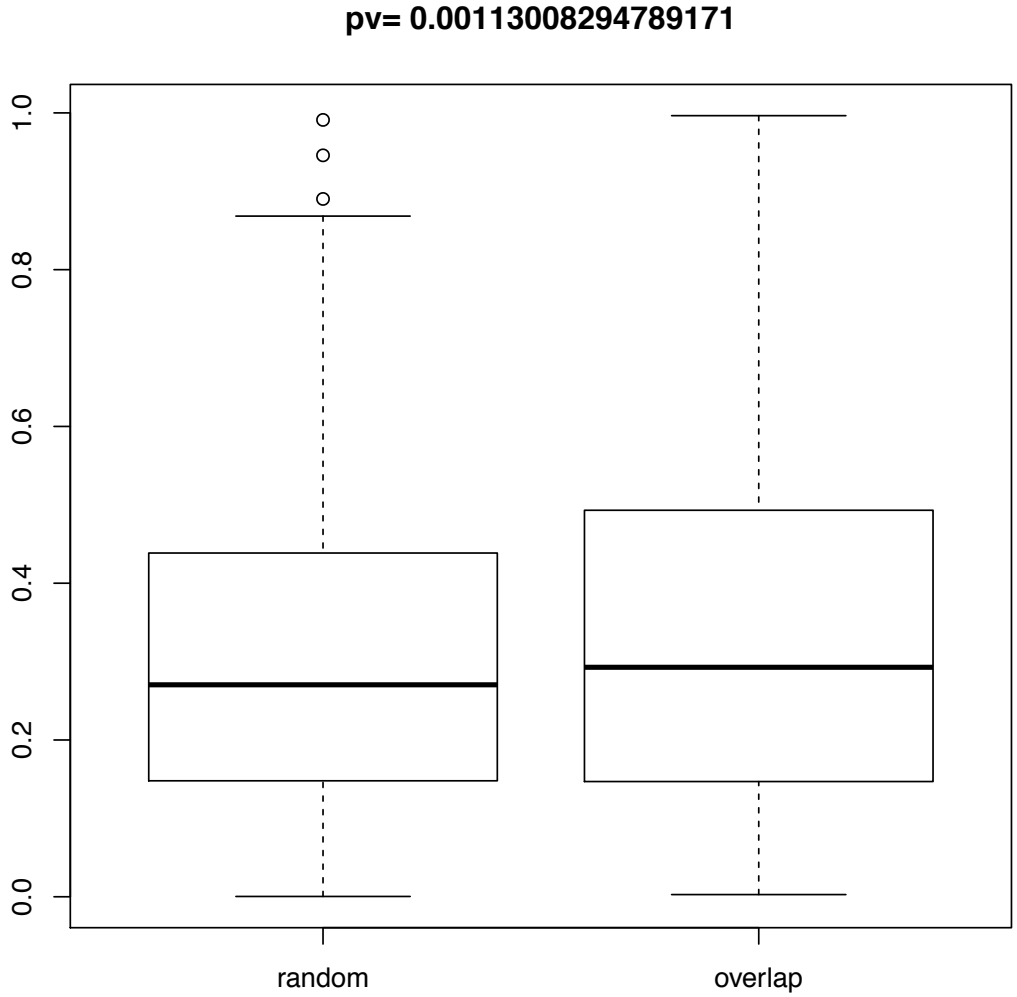


The number of Gm12878 active enhancer and promoter pairs overlapped with Hic, Fantom and ChIA PET.

The high quality of EP pair tend to enriched in the intersection of Hic and ChIAPET dataset

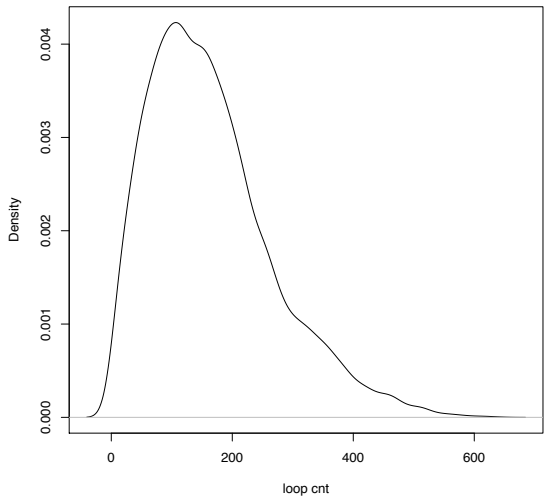


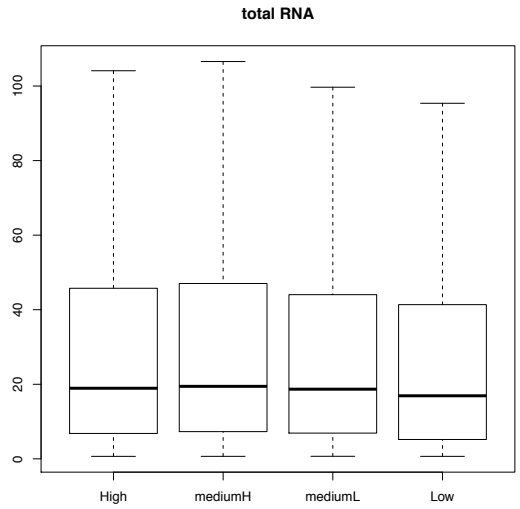
The genes tend to have relative higher correlation in fantom specific EP pair if they shared a same enhancer with the genes in the intersection set of EP pairs with Hic

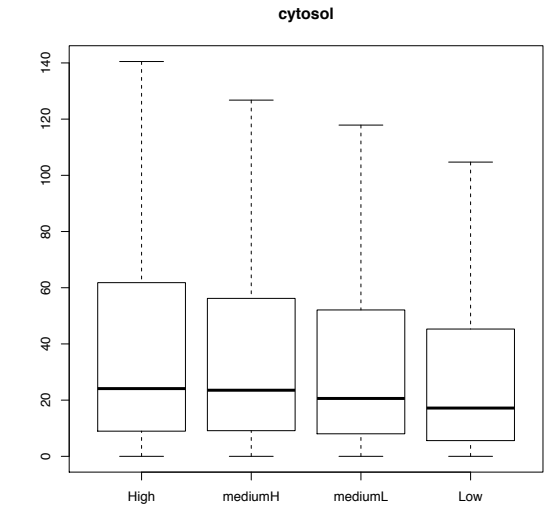


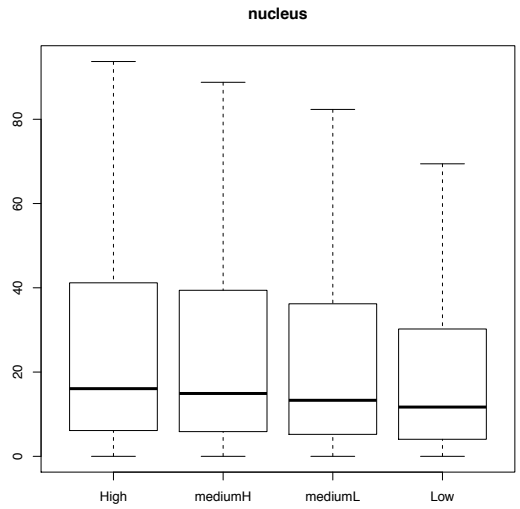
This indicate the potential problem of fantom dataset is the coexpression of genes will affect enhancer promoter target definition

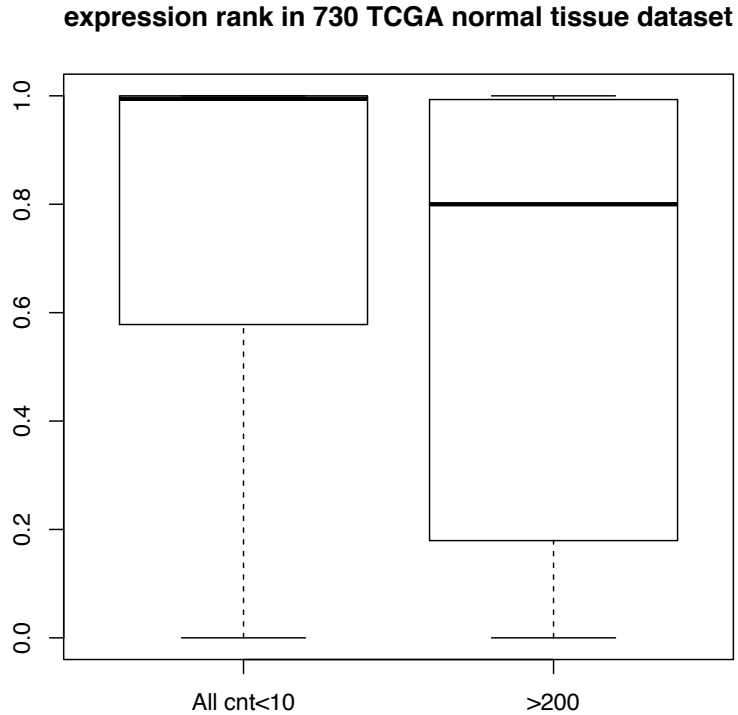
How the loop number close to a enhancer promoter pair affect the expression level of target genes?

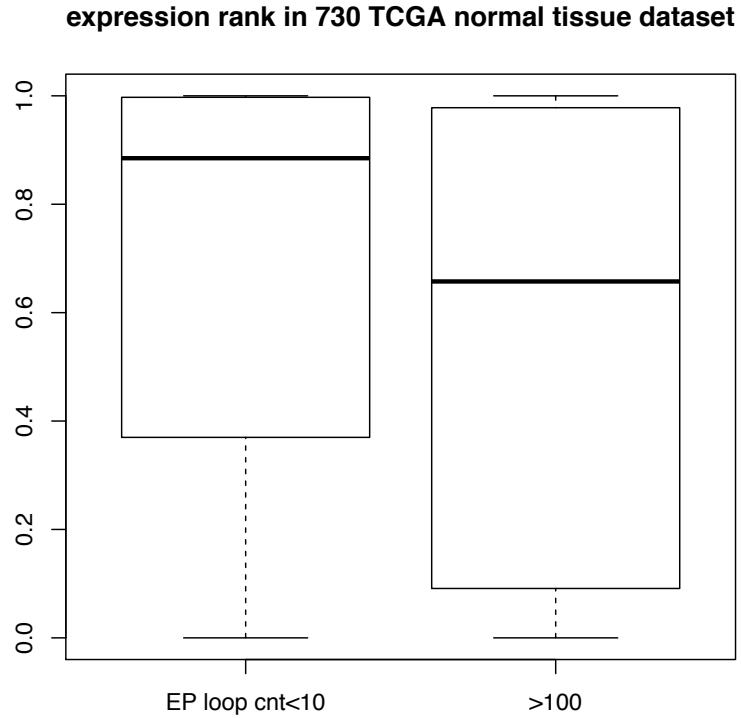












Gene expression activation negatively correlate with EP loop count.

Task2

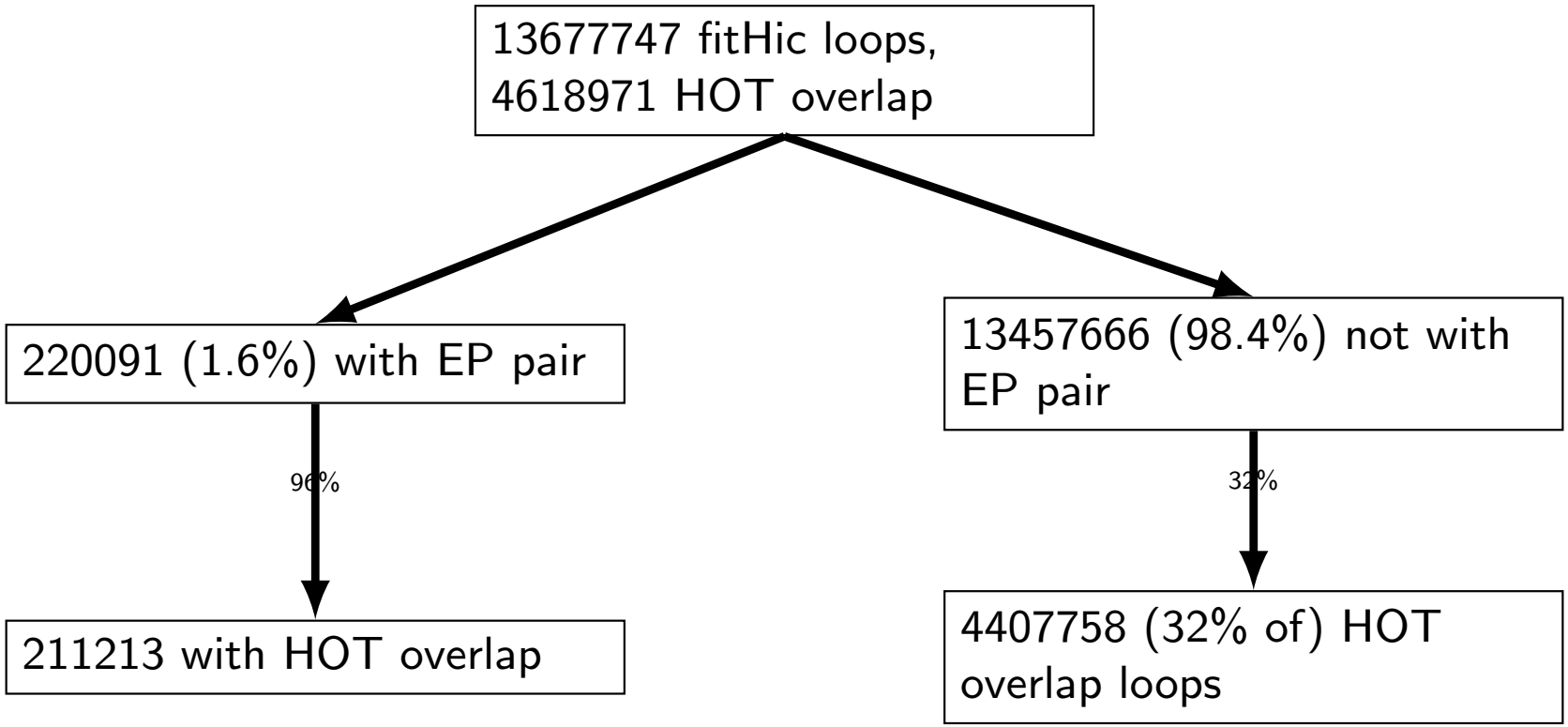
2. **Connection between 3D interaction and enhancer target regulation.** Some interactions related to enhancer target regulation, the others not, what is the connection between 3d interaction and distal enhancer target regulation? We will focus on the comparison and explore the differences between structural and regulation interaction, and stable and dynamics. (SKL, paritially)

Questions: Why loops have small fraction of EP pair (EP loop), mostly are non EP loop?

- Because arbitrary cutoff, such as anchor size(Hi-C resolution), or Loop qvalue cutoff?
- Different pattern of EP loops or nonEP loops? functional difference? (Functional vs structural)
- Dynamics and stable for EP loops and nonEP loops?
- From Hierachical structure, relationship of EP and nonEPloops.

keywords: Anchor, EP loop, nonEP loop, HOT region, Distance to Anchor

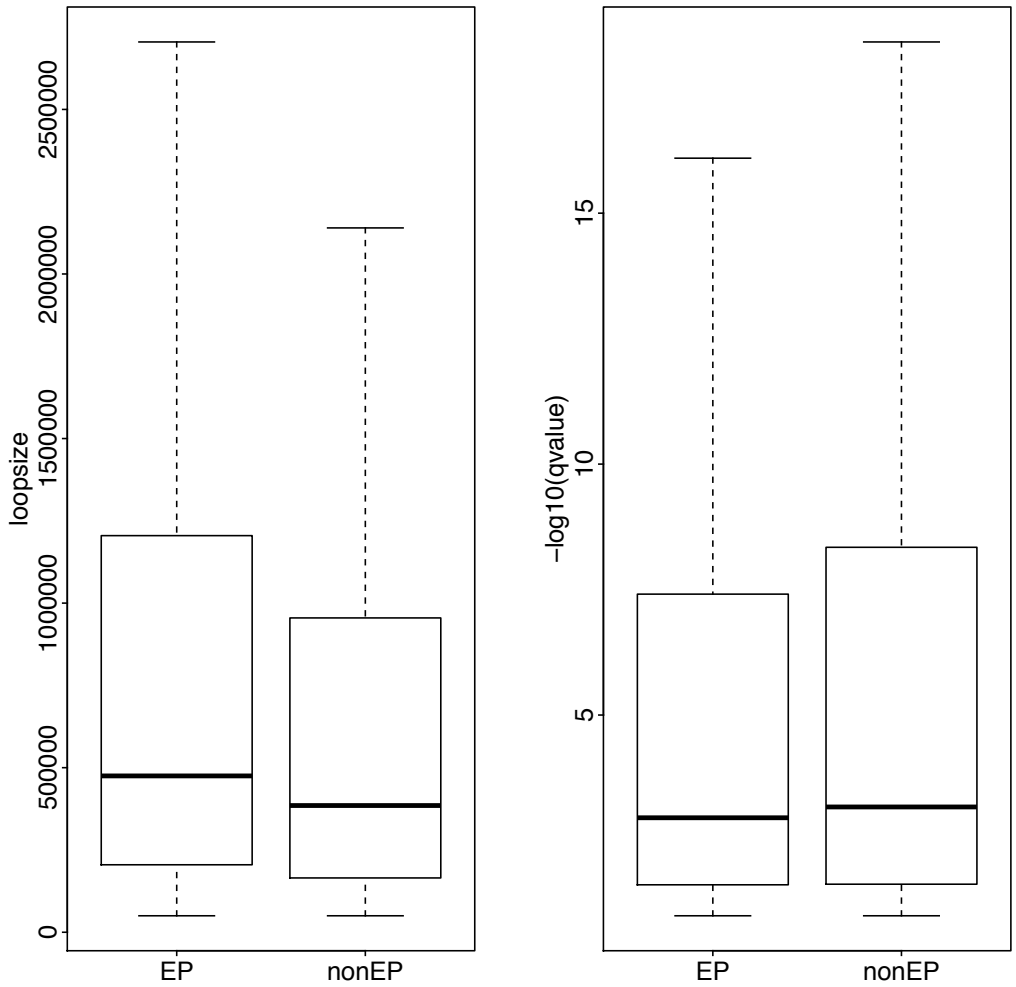
Summarieqs



14635 EP loop and 183516 No EP loop with one hot region for each anchor

Loop distance vs EP to loop anchor distance

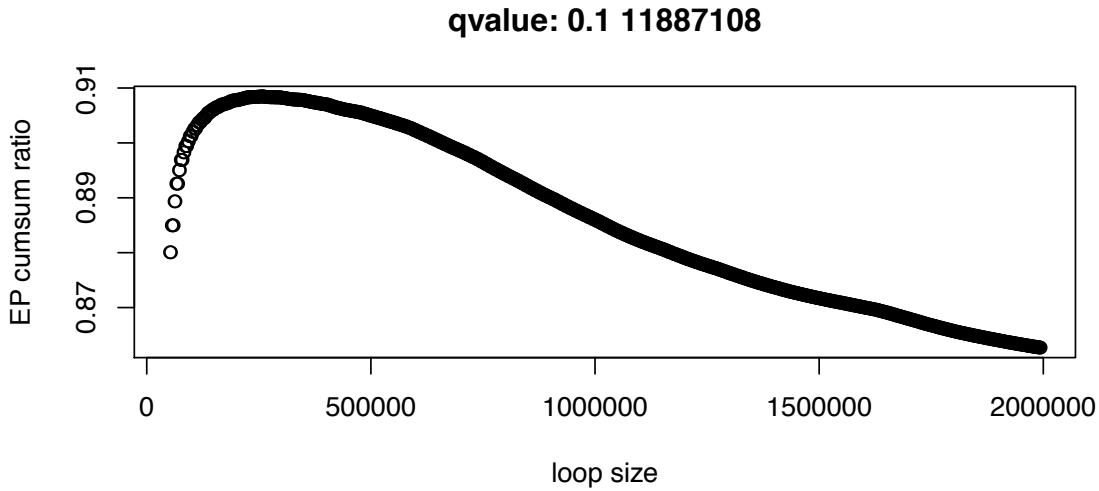
Compare all the HOT overlap loops, loopsize(inner size) and qvalue (related to contact frequency):

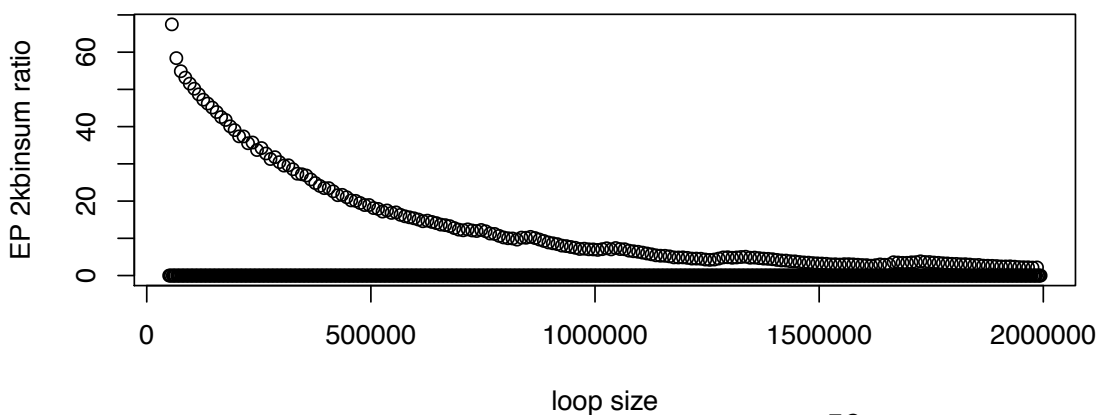


For HOT overlapped loops, EP loops tend to have larger loopsize and lower qvalue(more dynamic)

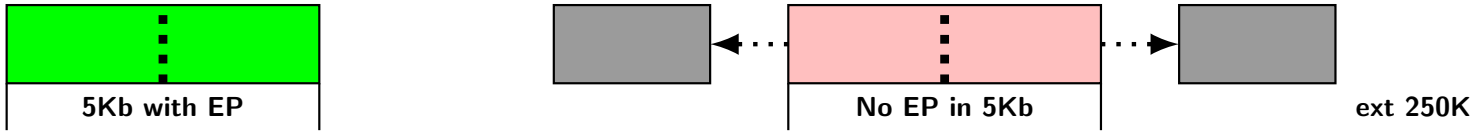
Loop size versus EP pair

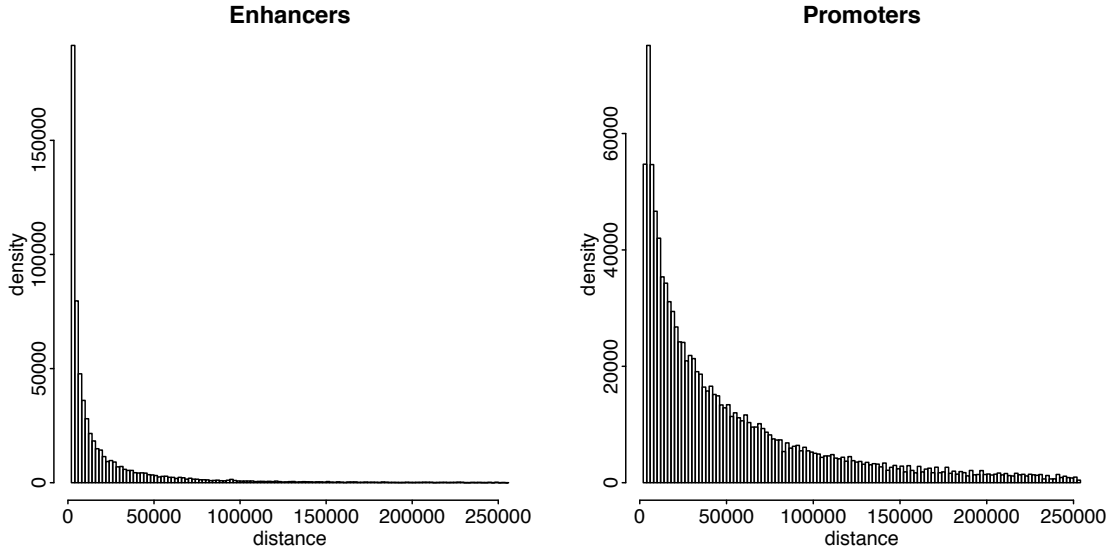
For all the loops with qvalue less than 0.1, extend both sides with 250k, find the closest EP pair, and the distance to anchor is the average of E or P center to anchor center.

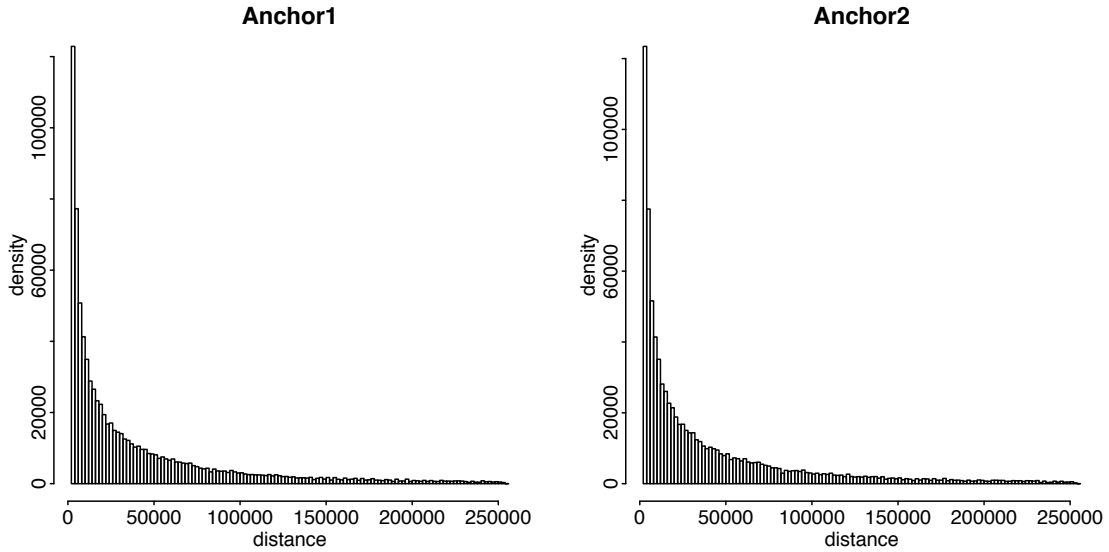




Cutoff bias: closest distance distribution of EP pair

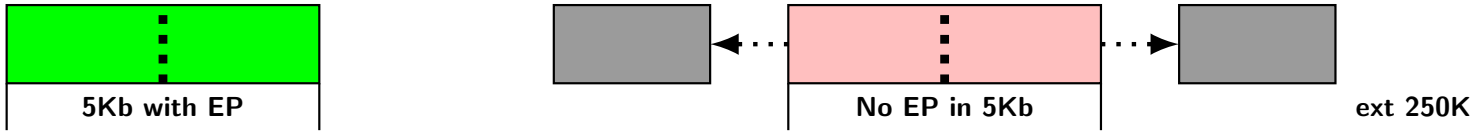


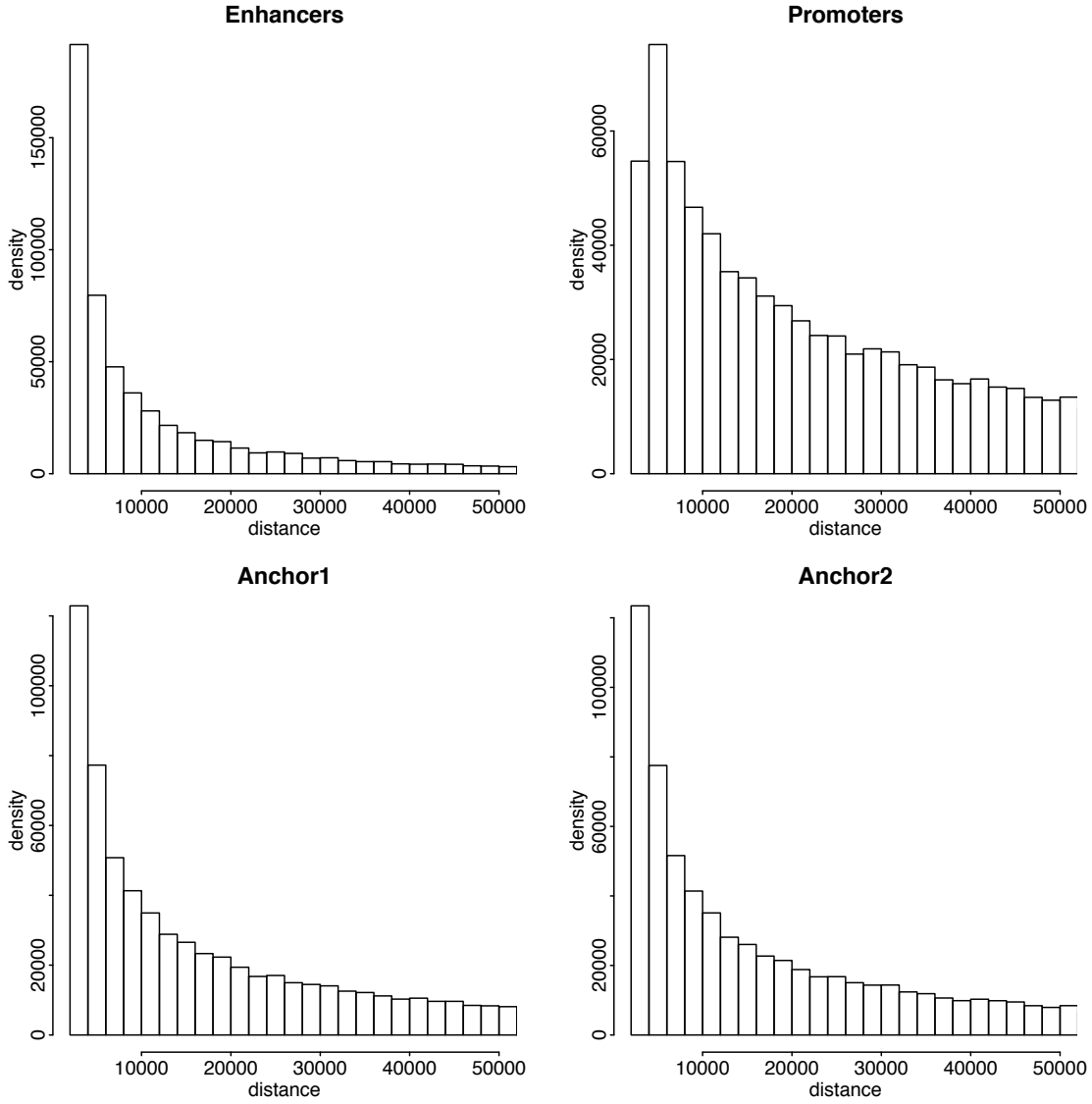




FitHic loop without EP pairs within 5kb bins

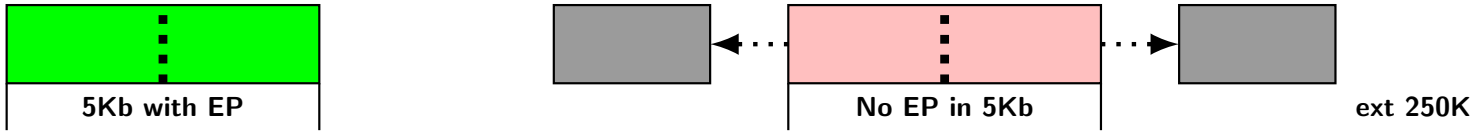
Cutoff bias: closest distance distribution of EP pair

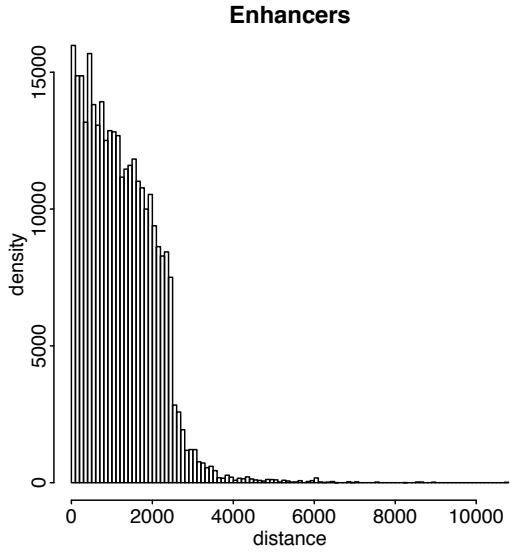


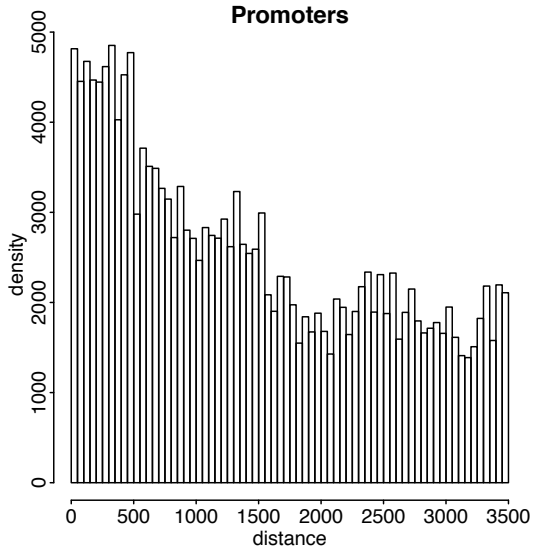


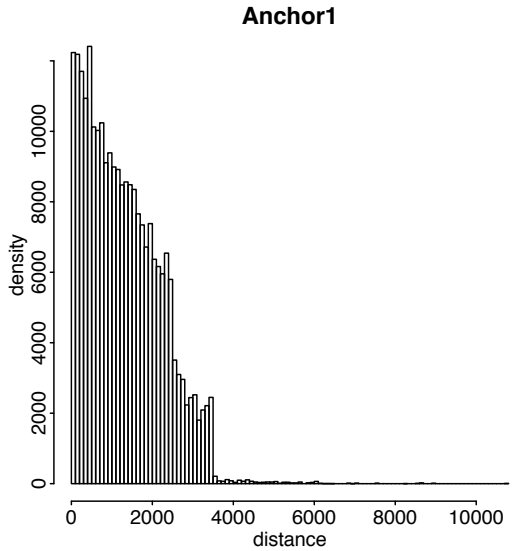
FitHic loop without EP pairs within 5kb bins, [2.5k50k]

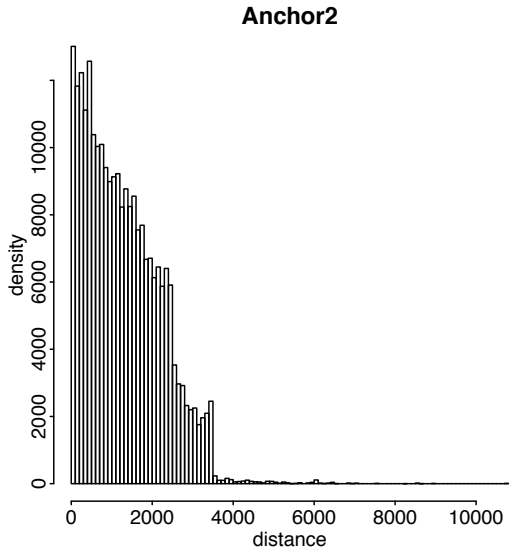
Cutoff bias: closest distance distribution of EP pair





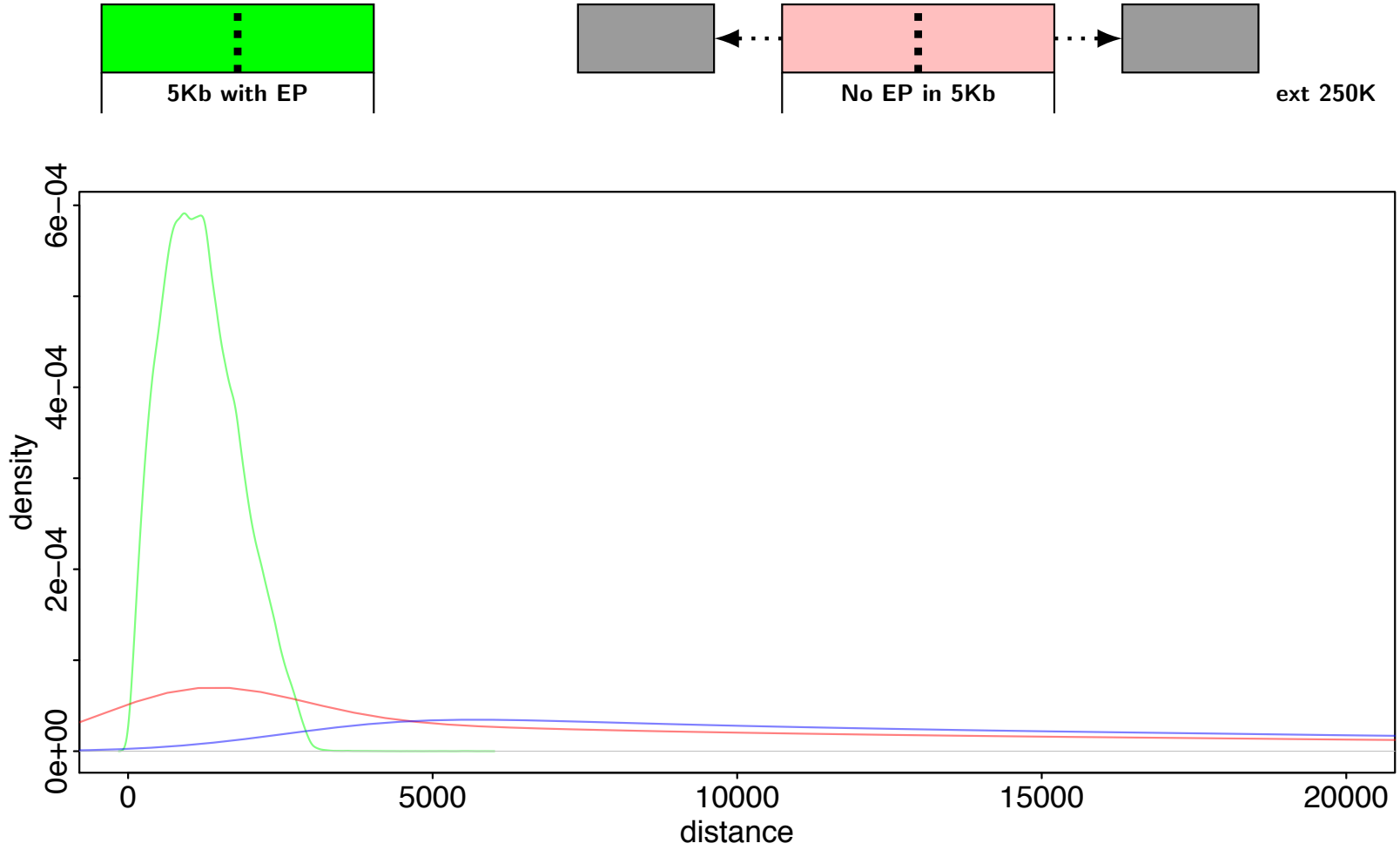






FitHic loop with EP pairs within 5kb bins

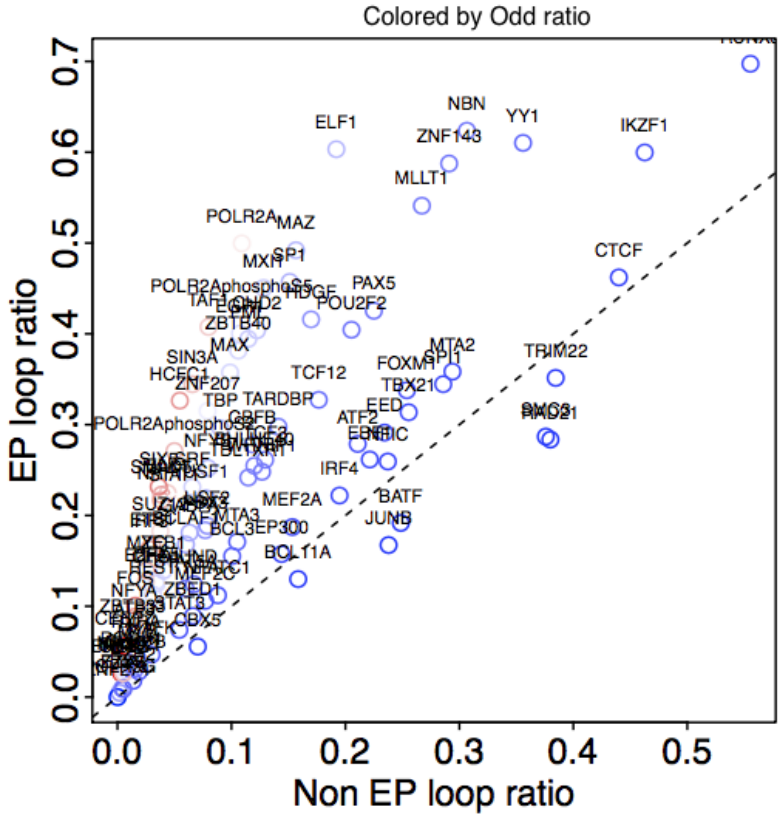
Cutoff bias: closest distance distribution of EP pair

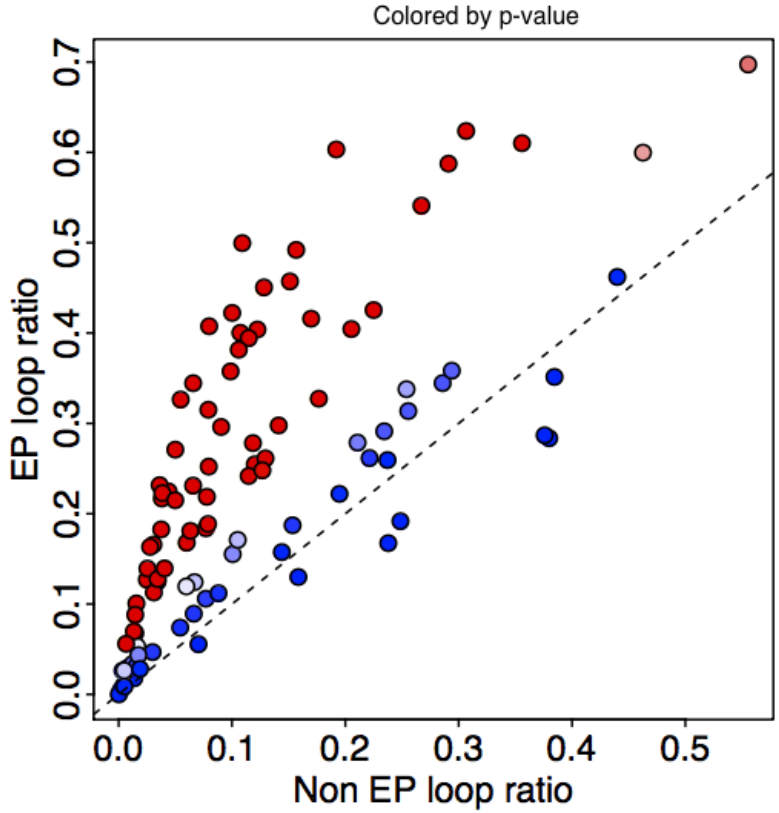


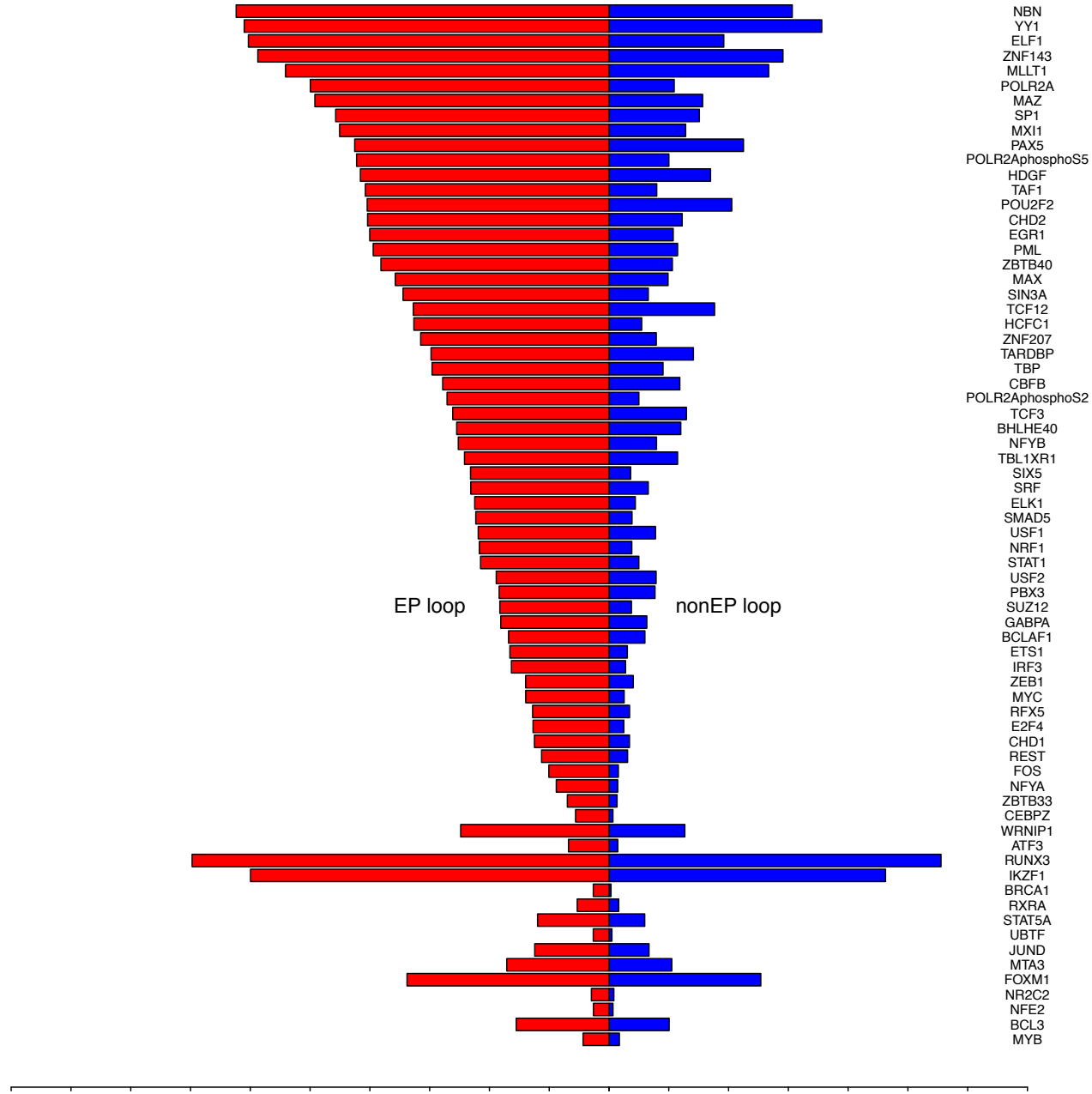
FitHic loop with EP pairs within 5kb bins, vs loops without EP pairs

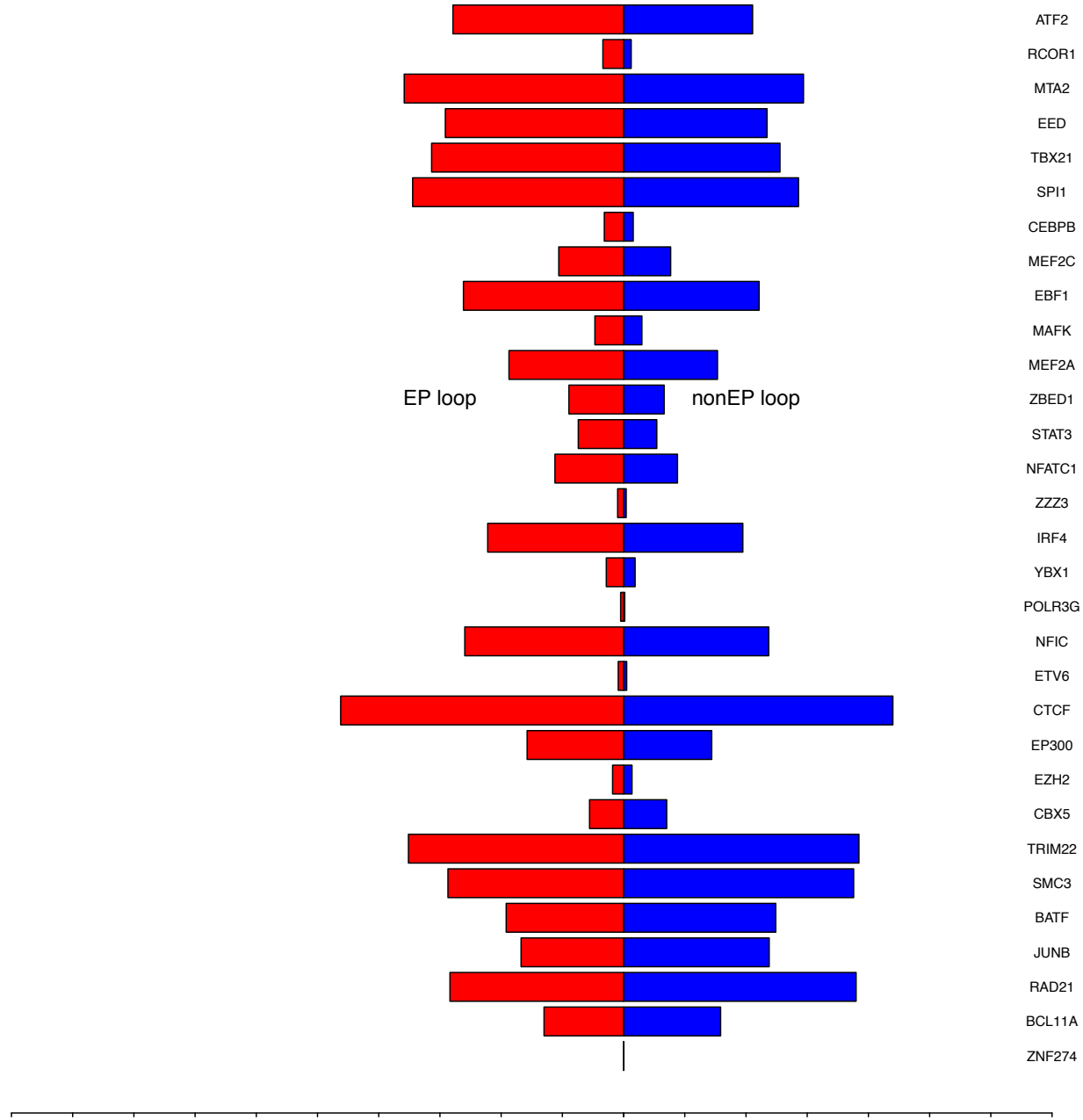
TF peak overlap

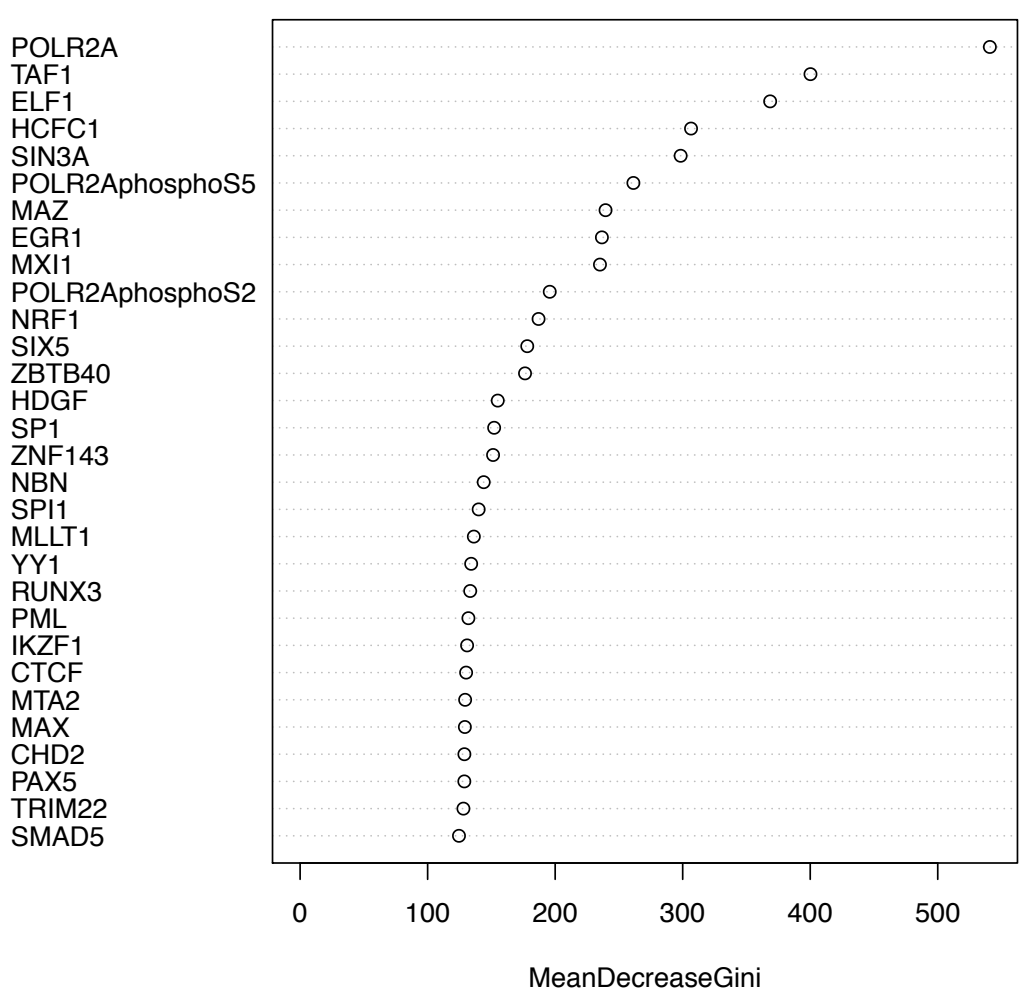
Take loop with one HOT for both side:






EP loop
nonEP loop
NBN
YY1
ELF1
ZNF143
MLLT1
POLR2A
MAZ
SP1
MXI1
PAX5
POLR2AphosphoS5
HDGF
TAF1
POU2F2
CHD2
EGR1
PML
ZBTB40
MAX
SIN3A
TCF12
HCFC1
ZNF207
TARDBP
TBP
CBFB
POLR2AphosphoS2
TCF3
BHLHE40
NFYB
TBL1XR1
SIX5
SRF
ELK1
SMAD5
USF1
NRF1
STAT1
USF2
PBX3
SUZ12
GABPA
BCLAF1
ETS1
IRF3
ZEB1
MYC
RFX5
E2F4
CHD1
REST
FOS
NFYA
ZBTB33
CEBPZ
WRNIP1
ATF3
RUNX3
IKZF1
BRCA1
RXRA
STAT5A
UBTF
JUND
MTA3
FOXM1
NR2C2
NFE2
BCL3
MYB
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7


EP loop
nonEP loop
ATF2
RCOR1
MTA2
EED
TBX21
SPI1
CEBPB
MEF2C
EBF1
MAFK
MEF2A
ZBED1
STAT3
NFATC1
ZZZ3
IRF4
YBX1
POLR3G
NFIC
ETV6
CTCF
EP300
EZH2
CBX5
TRIM22
SMC3
BATF
JUNB
RAD21
BCL11A
ZNF274
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7


ds.cbind.rf
POLR2A
TAF1
ELF1
HCFC1
SIN3A
POLR2AphosphoS5
MAZ
EGR1
MXI1
POLR2AphosphoS2
NRF1
SIX5
ZBTB40
HDGF
SP1
ZNF143
NBN
SPI1
MLLT1
YY1
RUNX3
PML
IKZF1
CTCF
MTA2
MAX
CHD2
PAX5
TRIM22
SMAD5
0
100
200
300
400
500
MeanDecreaseGini