Chromatin Model for Predicting Transcription Factor Binding and Gene Expression

Chao Cheng

Group meeting

Feb 16, 2010

Why chromatin features?

- Driven by data

  - encode; modencode; published data

- Important in biological research

  - transcription regulation

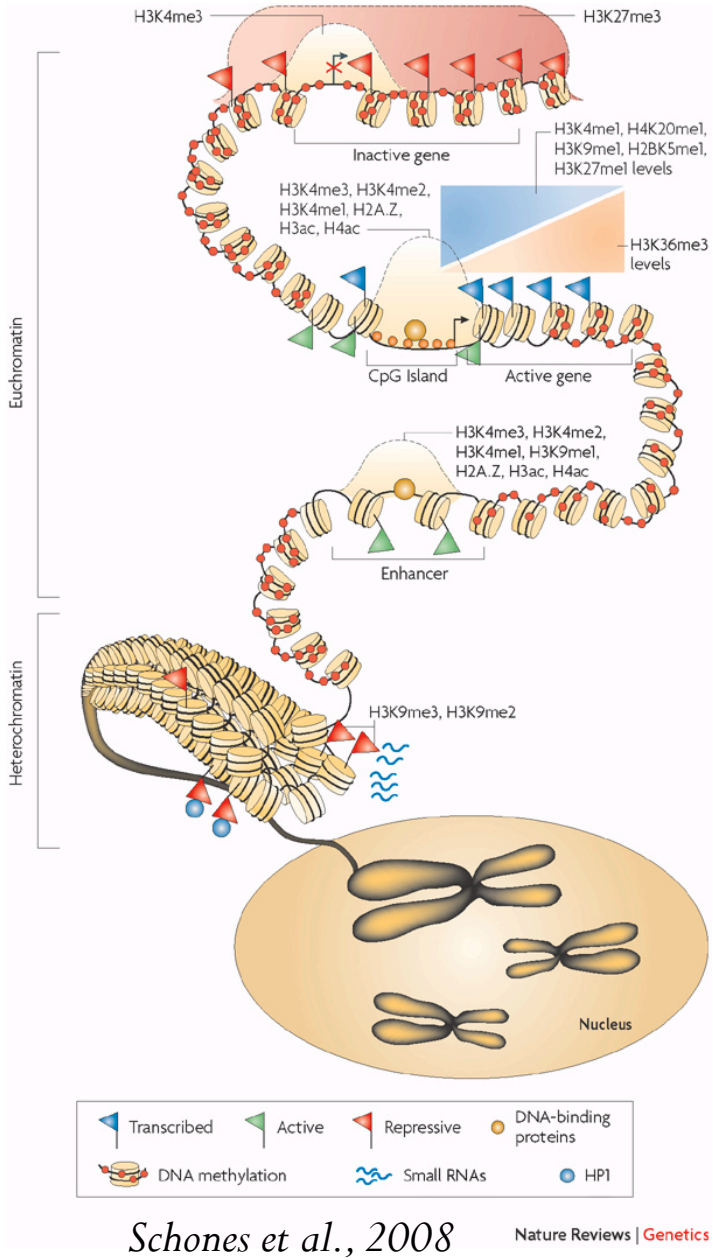
  - development

  - cancer

  - others

How chromatin modifications regulate gene expression?

- Modulate DNA accessibility
- Recruit transcription regulators

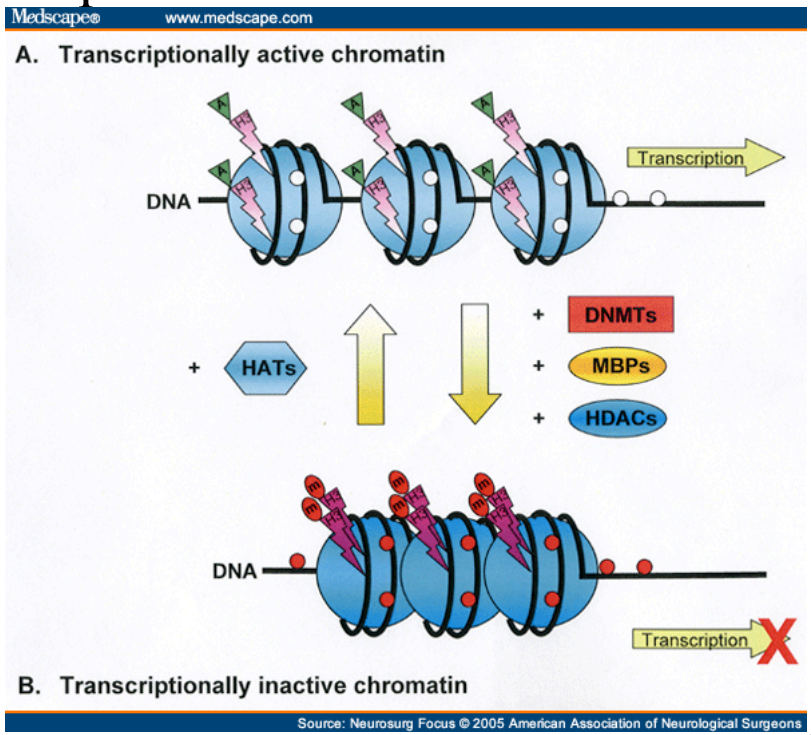


Schones et al., 2008

Nature Reviews | Genetics

What have we done?

- Part I: a chromatin model for gene expression prediction
  - human K562 and GM12878 cell lines
  - worm EEMB and L3
- Part II: a two-step method for TFBS prediction
  - yeast
  - human
  - worm



Part I: a chromatin model for gene expression prediction

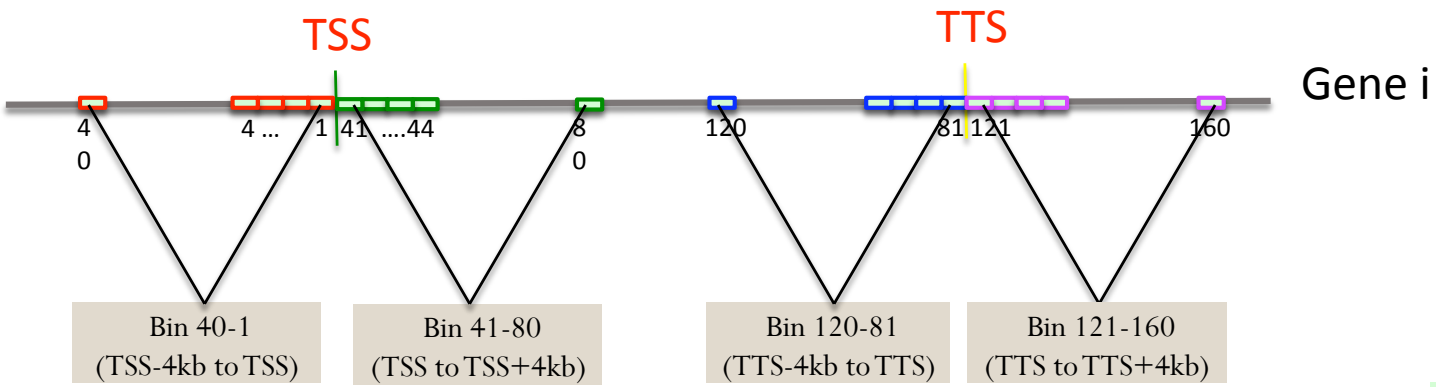
-Individual: correlation of each chromatin features with gene expression

-Collective: hierarchical clustering of genes based on chromatin features

-Integrative: supervised model for gene expression prediction

- Note: Relative importance of each chromatin feature for transcription regulation depends on its position relative to TSS: divide chromosome into small bins

Divide and compare strategy

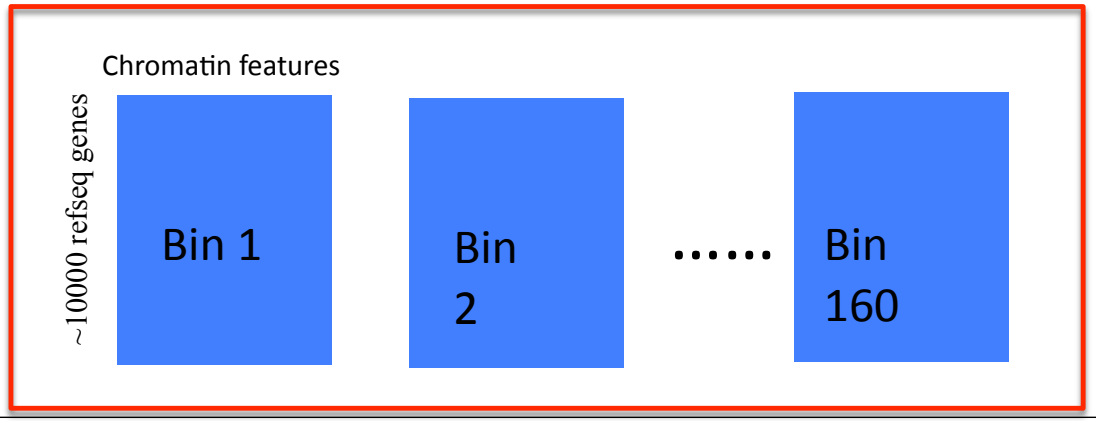


Chromatin features:
Histone methylation
Pol II binding

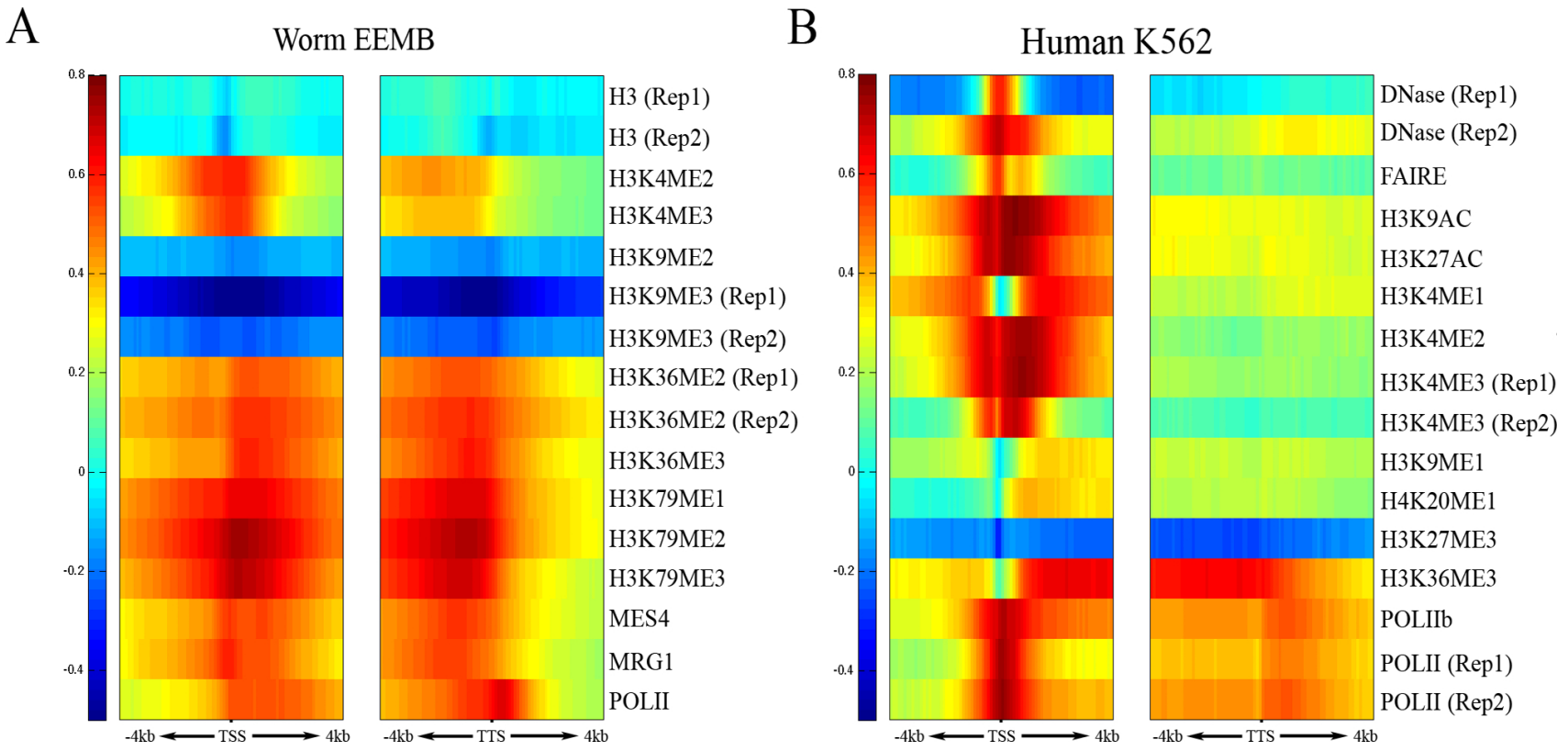
RNA-Seq data

Prediction target:
Gene expression level

Predictors



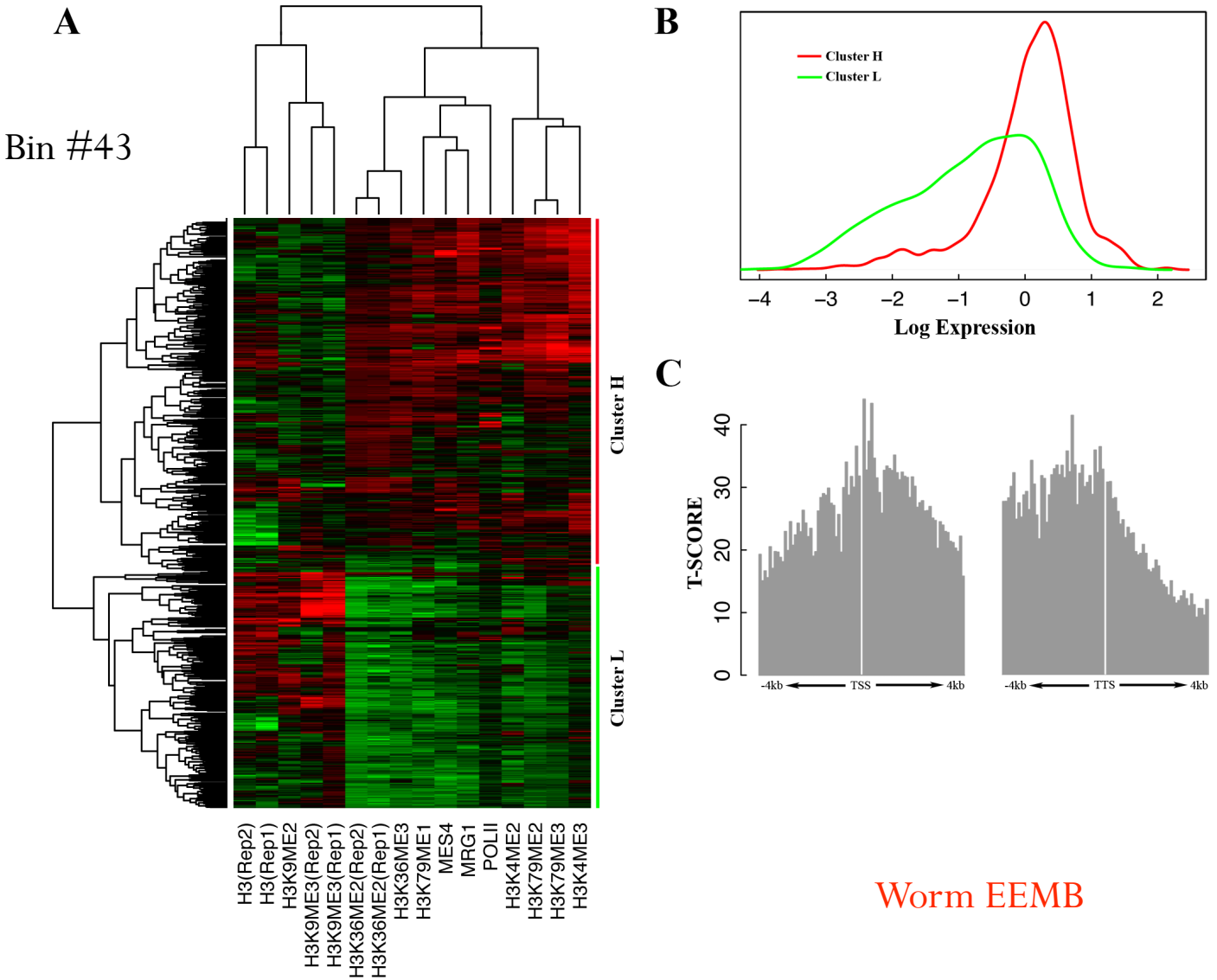
Correlation pattern of chromatin features with gene expression



Human: ChIP-seq, real transcription start and terminal sites

Worm: ChIP-chip, translation start and terminal sites

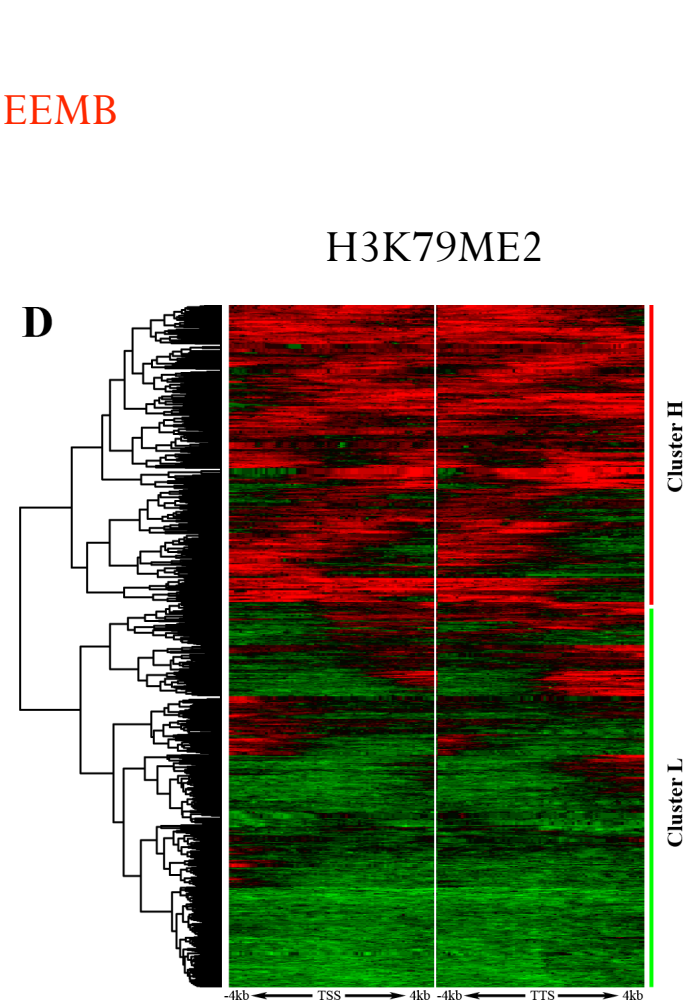
Clustering on chromatin profile

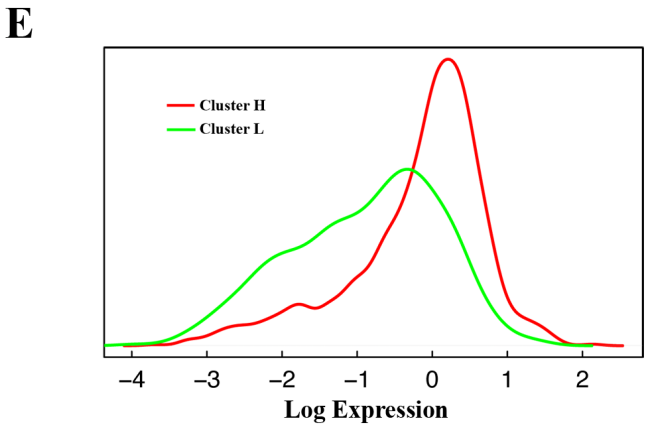


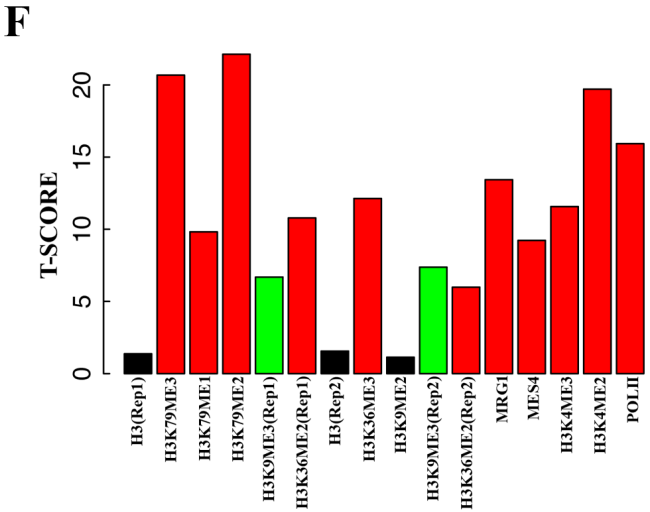
Worm EEMB

Clustering on bin profile

Worm EEMB

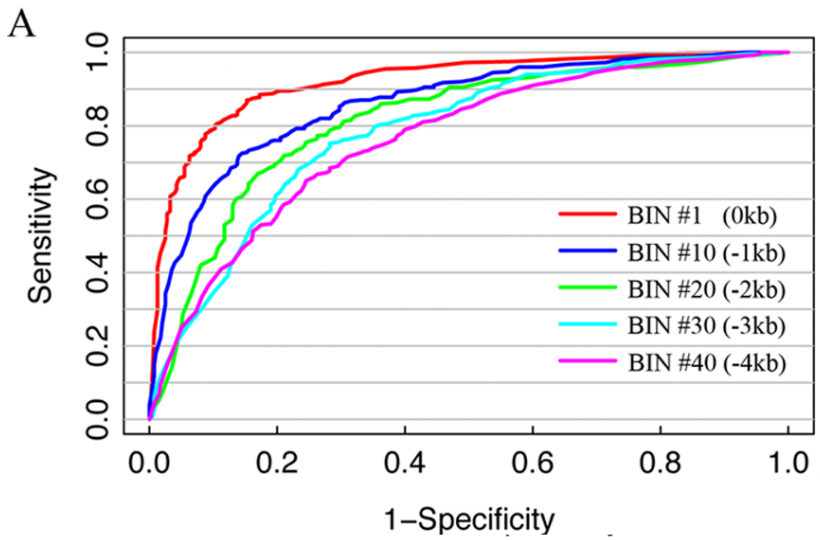


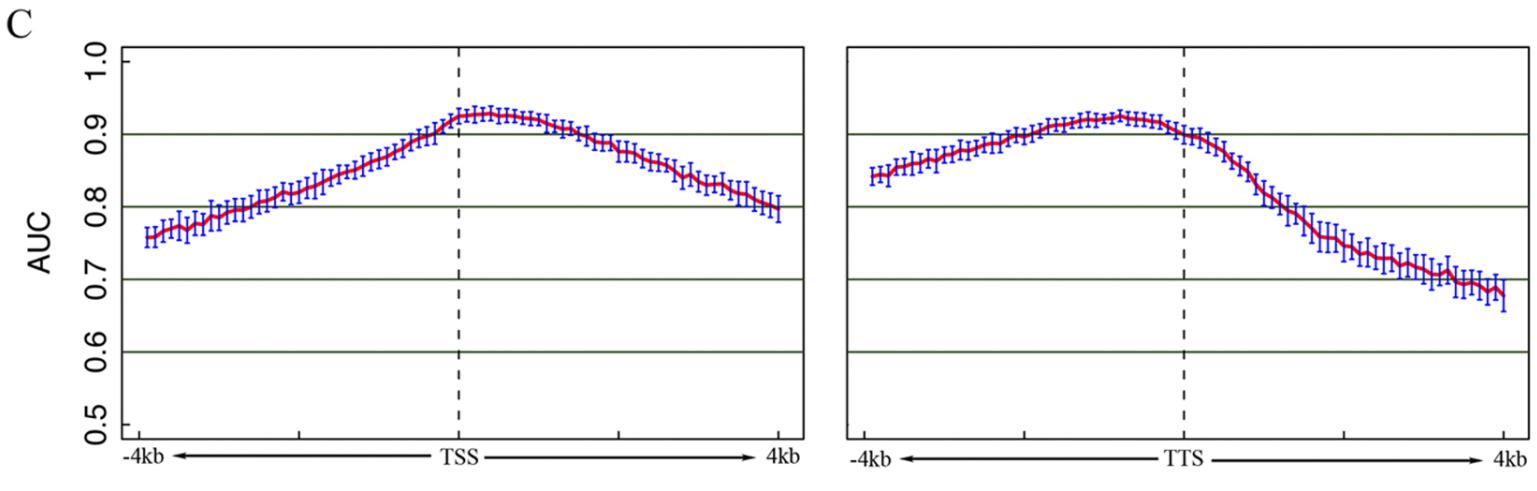




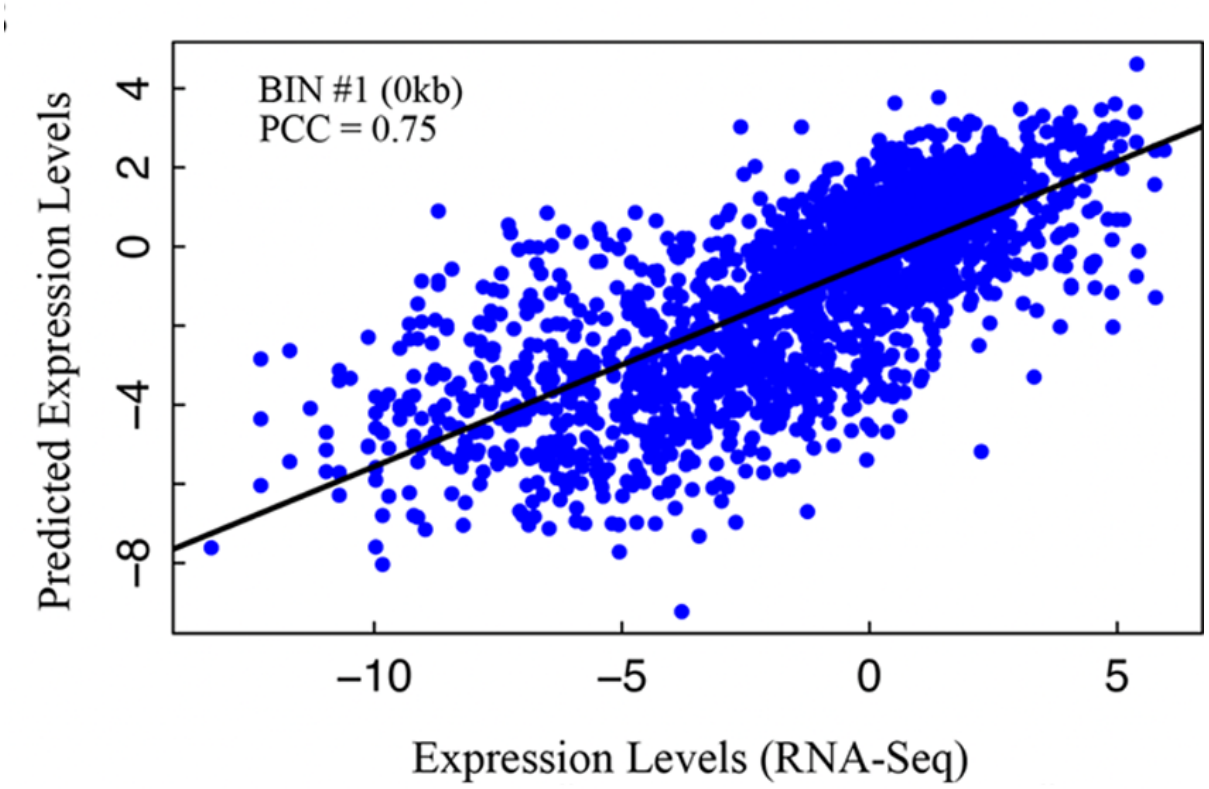
Predict gene expression: (SVM classification)

Worm EEMB



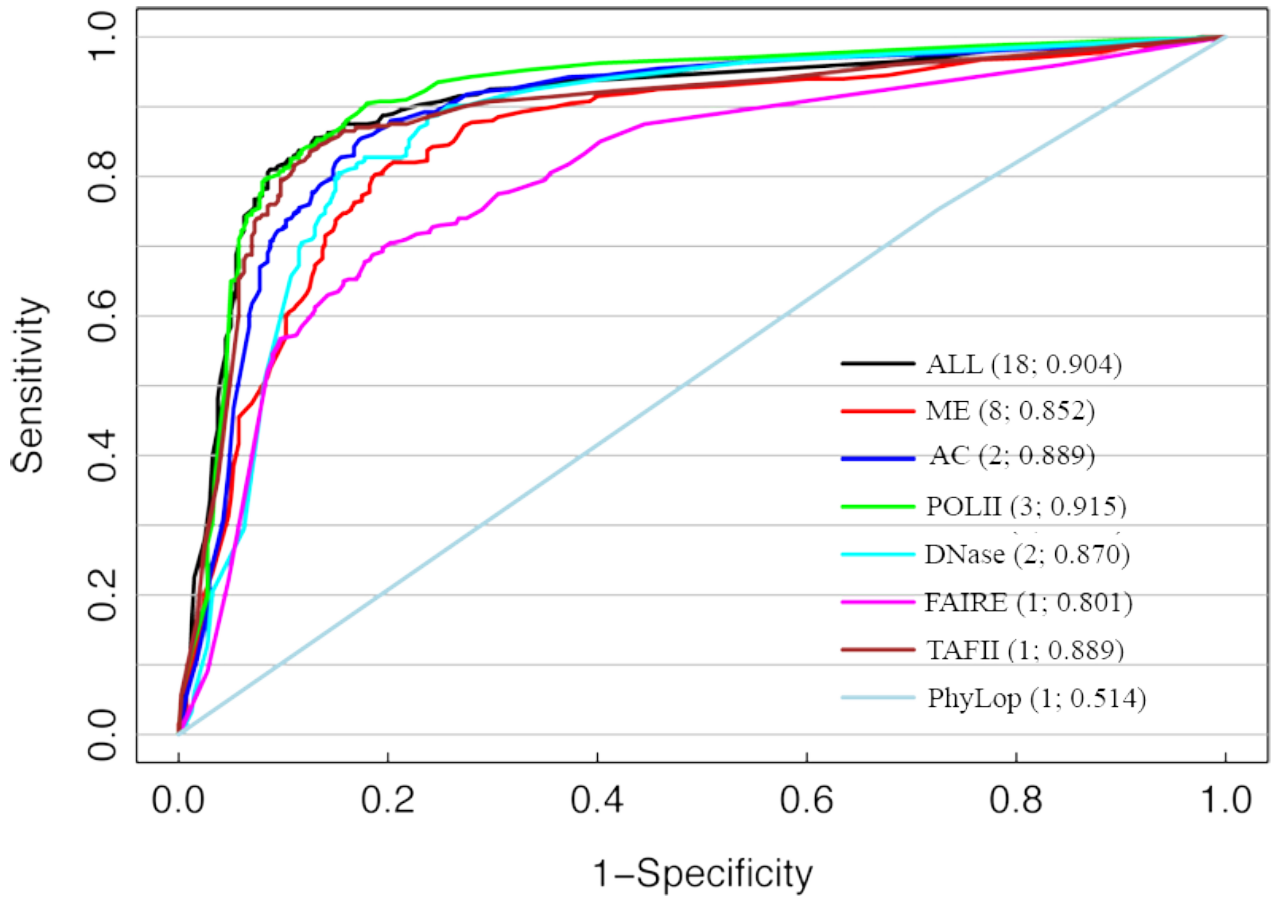


Predict gene expression: SVR regression

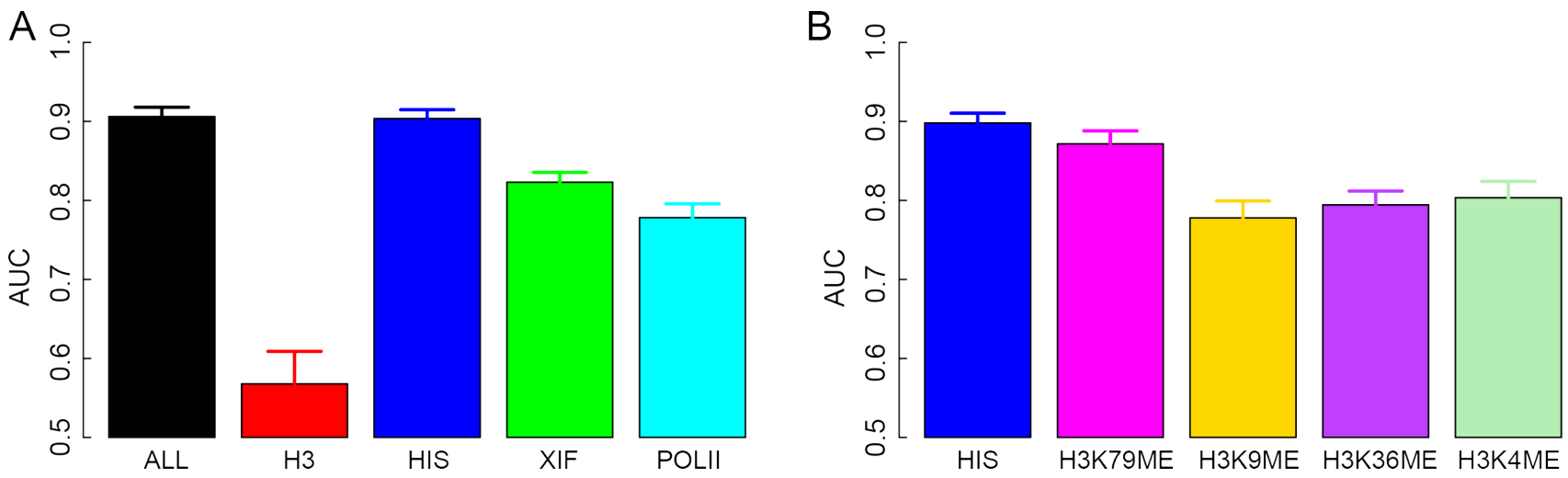


Relative contribution of each chromatin features (human)

Human K562

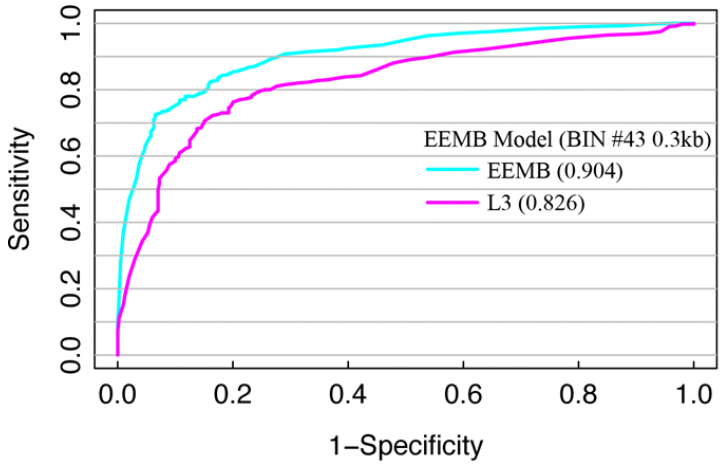


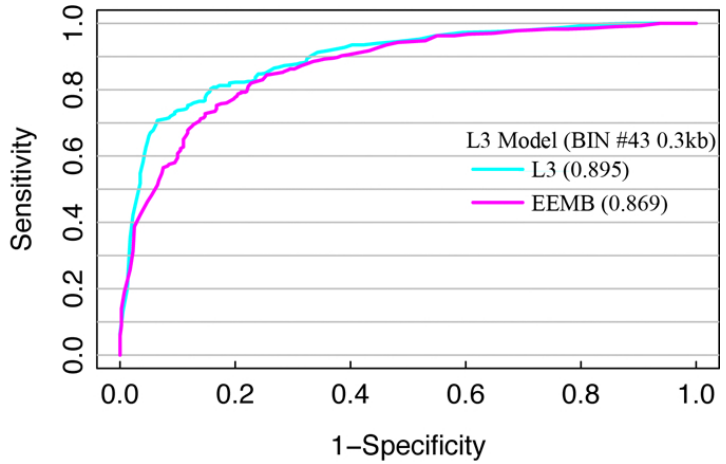
Relative importance of chromatin features (worm)



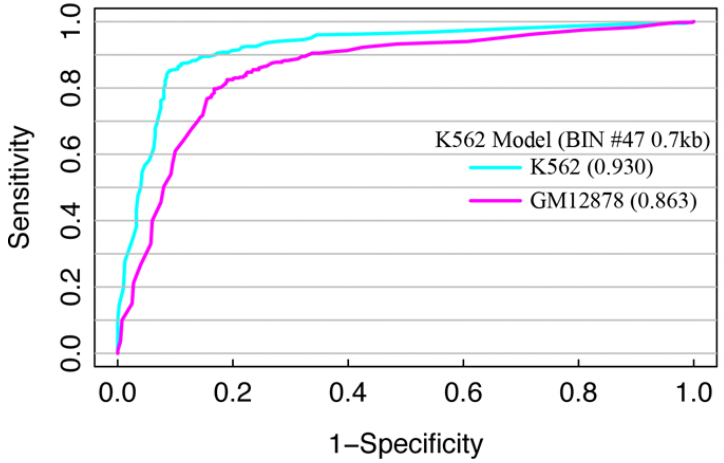
Chromatin model is tissue/cell line specific and development stage specific

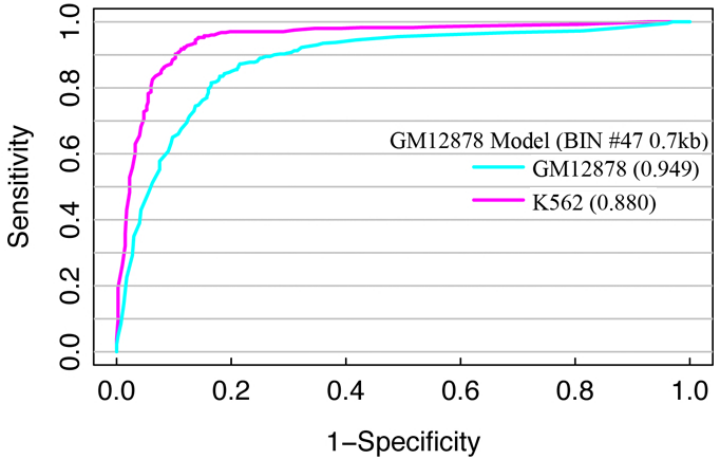
Worm



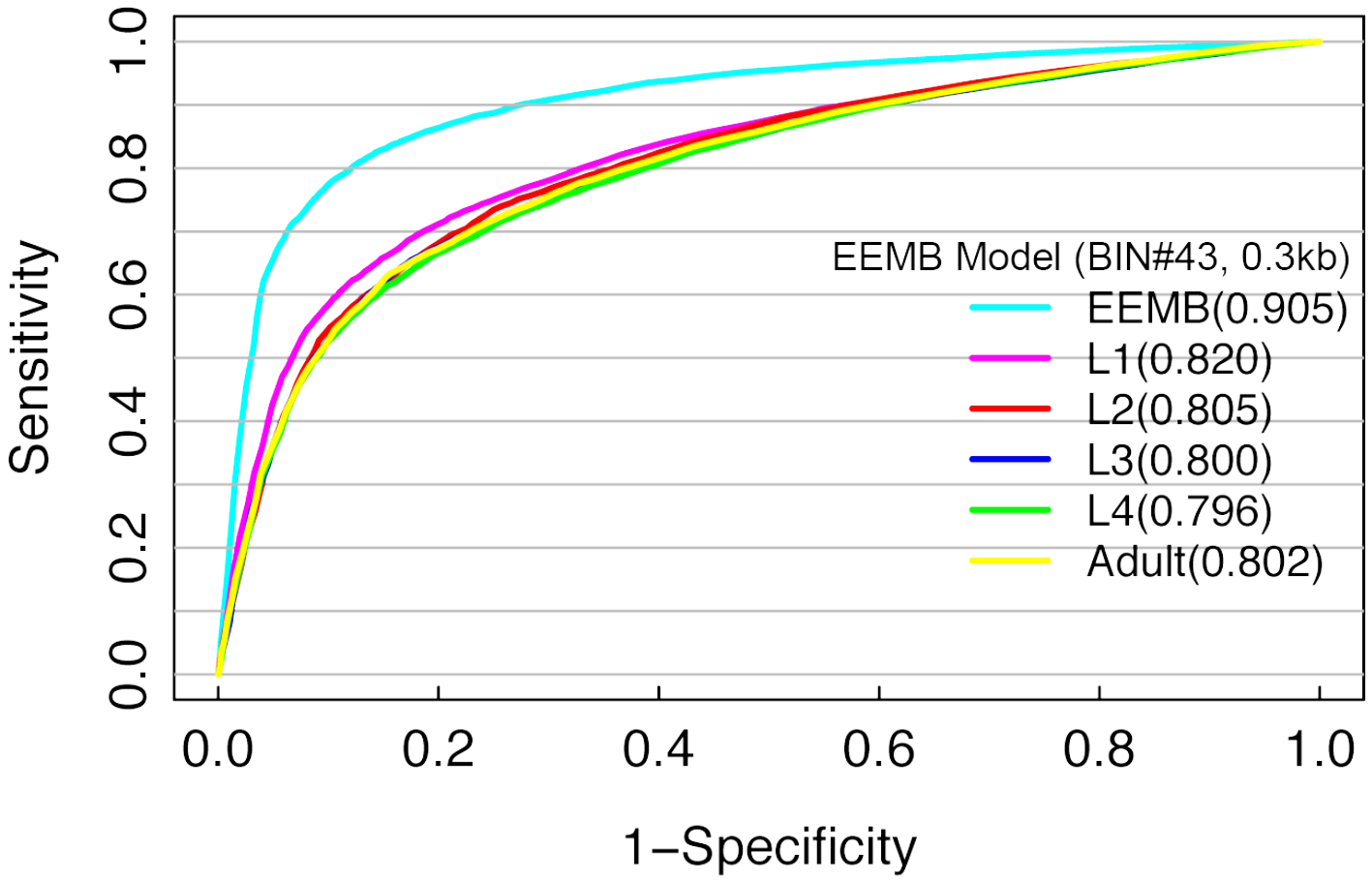


Human

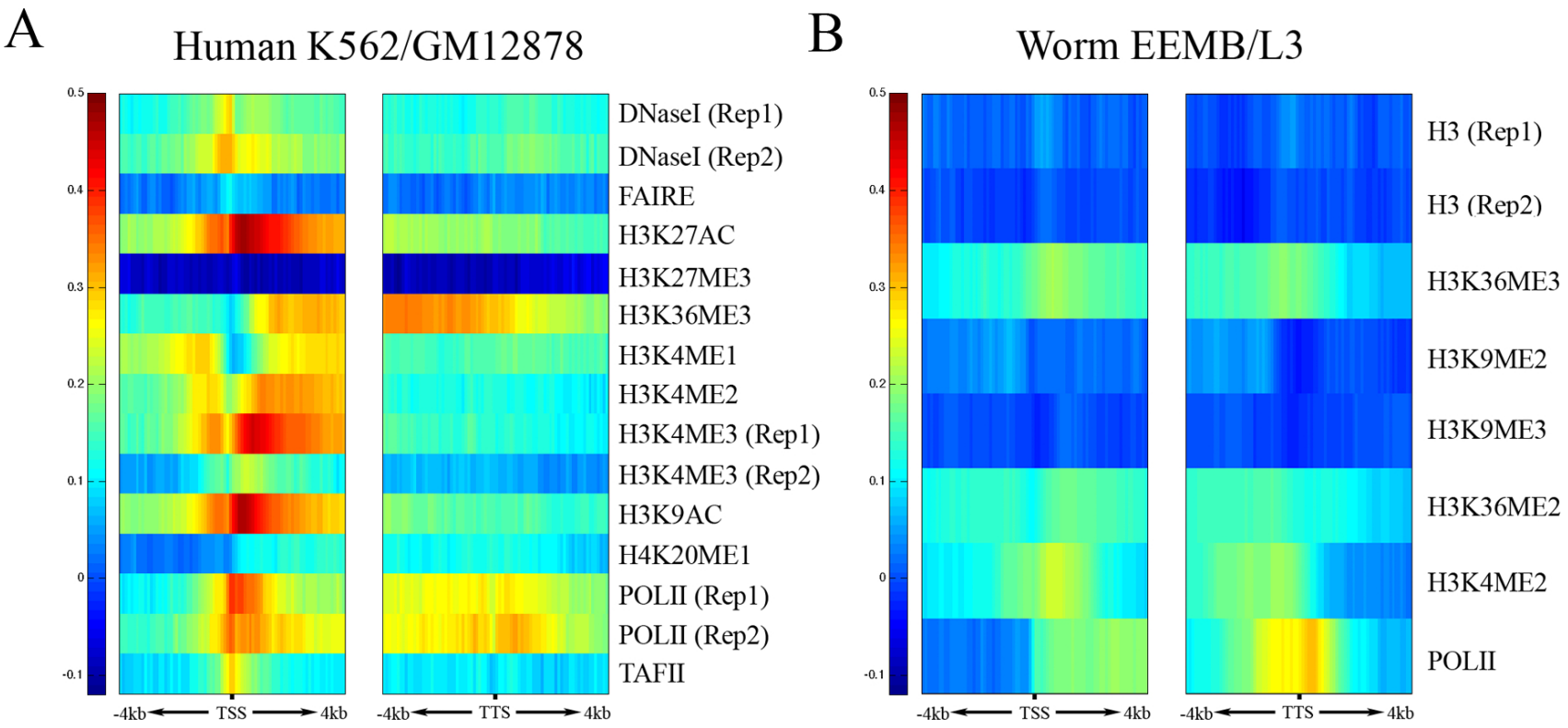




Development stage specificity: worm



Predicting differential expression

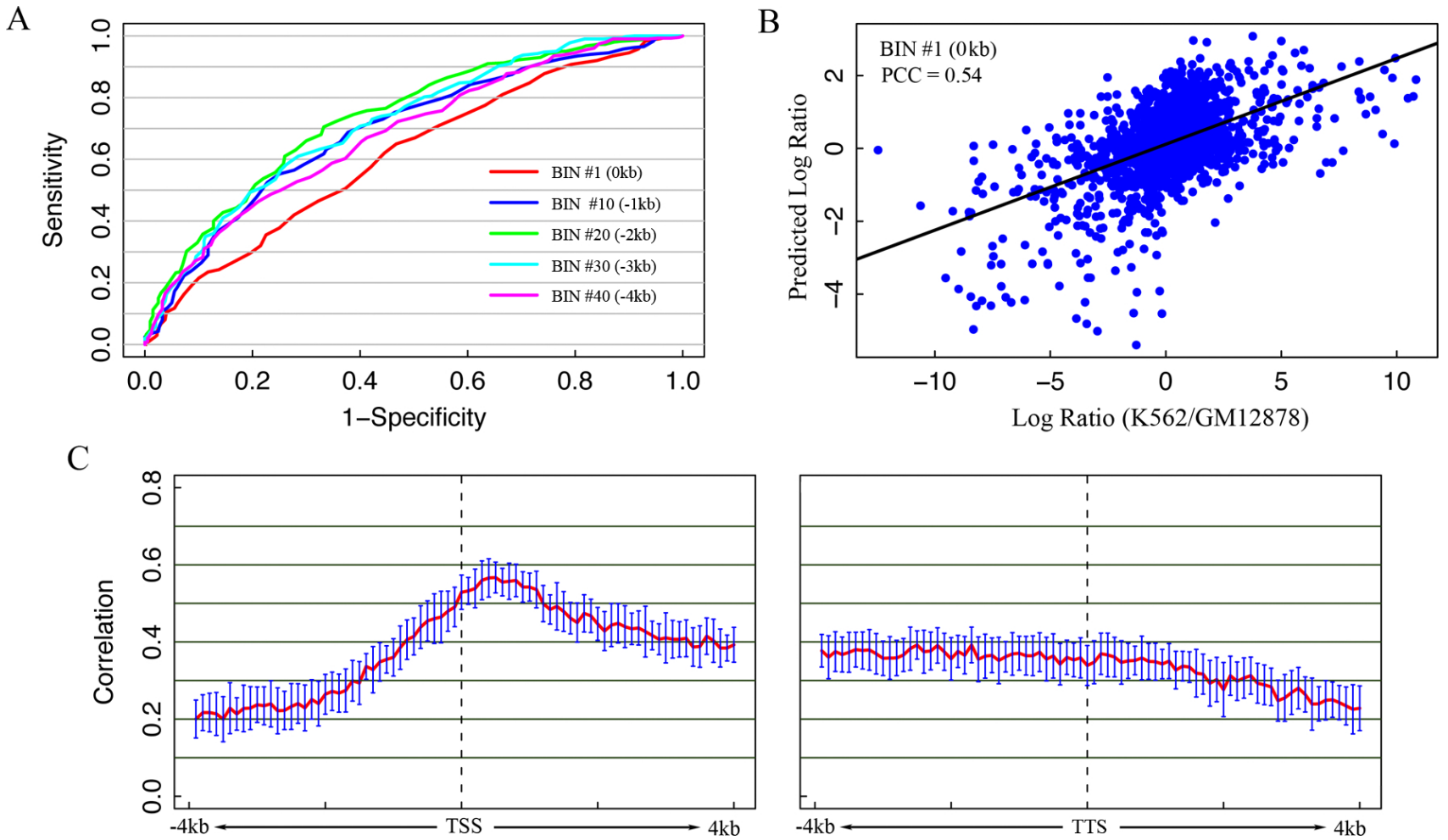


Yes

No

Predict differential expression

Human K562/GM12878



Predicting microRNA expression

- 162 worm microRNAs from miRBASE

  - ~100 bp corresponding to the pre-microRNAs

- Calculate the signal of all chromatin features in their genomic location
- Apply the SVM model to predict their expression

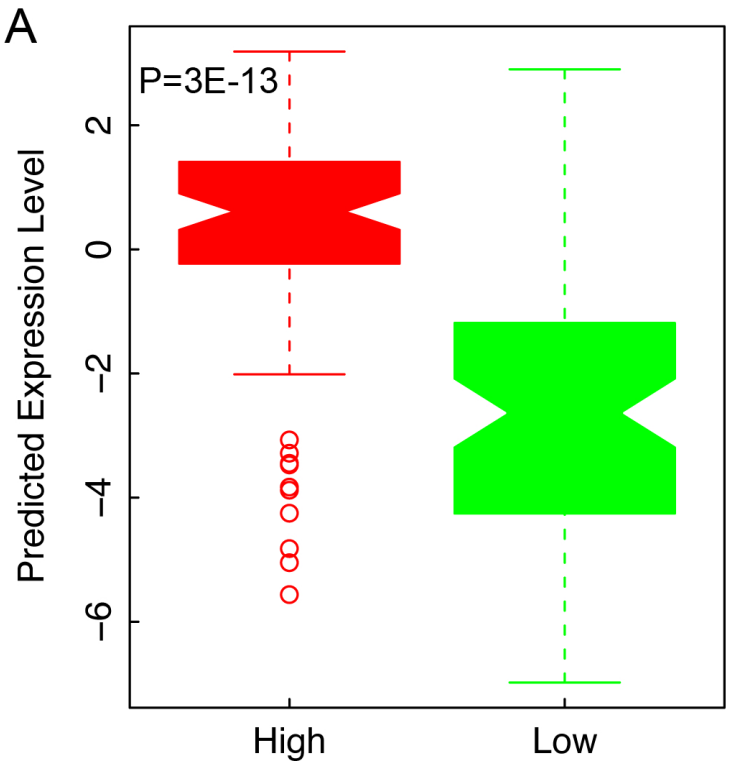
  - NOTE: the model is trainned using protein-coding genes

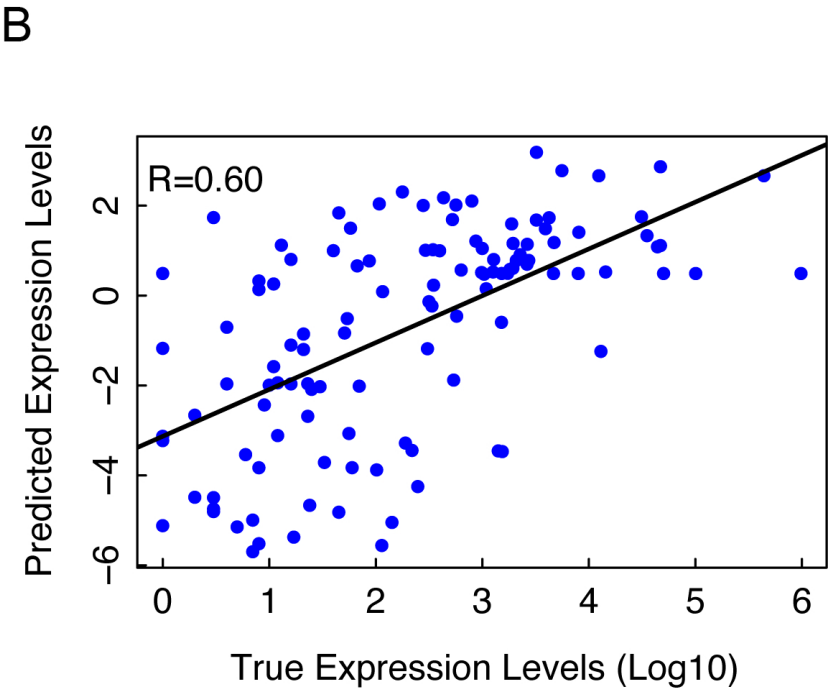
- Validate prediction results from experimental data

  - Kato et al., small RNA-seq data

- Worm and human

Predicting worm microRNA expression





Conclusion

- Chromatin features can accurately predict gene expression
- Chromatin model is tissue/cell line/development stage specific
- Chromatin feature are highly redundant in terms of expression prediction
- Chromatin feature can predict differential gene expression
- ChIP-seq has higher resolution than ChIP-chip
- Chromatin models are valid for protein-coding genes and microRNAs.

Part II: A two-step approach for predicting TF binding sites

-yeast: the merit of small compact genome

-human: the merit of being human

-worm: where to find the TF PWMs?

Predicting TF targets genes of yeast

- TF binding data:

  - ChIP-chip for 203 yeast TFs

  - for each TF-gene pair, a P-value is calculated

- Chromatin modification profile data

  - ChIP-chip, Pokholok et al. 2005

  - 14 profiles: H3(YPD), H4(YPD), H3(H2O2), H3K9AC/H3(YPD), H3K14AC/H3(YPD), H3K14AC/WCE(YPD), H3K14AC/H3(H2O2), H4AC/H3(YPD), H4AC/H3(H2O2), H3K4ME1/H3(YPD), H3K4ME2/H3(YPD), H3K4ME3/H3(YPD), H3K36ME3/H3(YPD), H3K79ME3/H3(YPD)

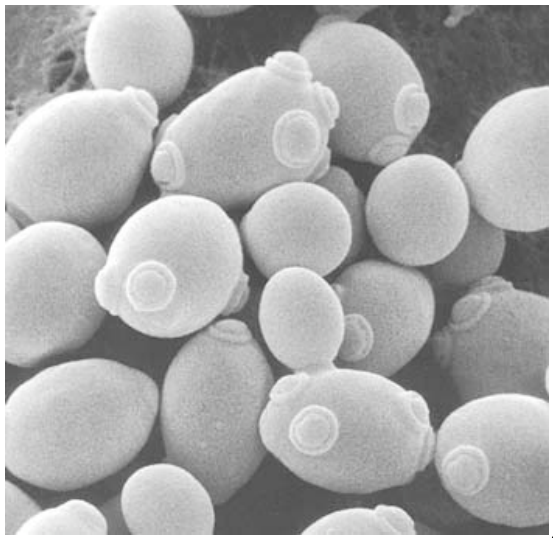
- Positional weighted matrices for yeast TFs

  - sequence analysis: motifs enriched in yeast promoters

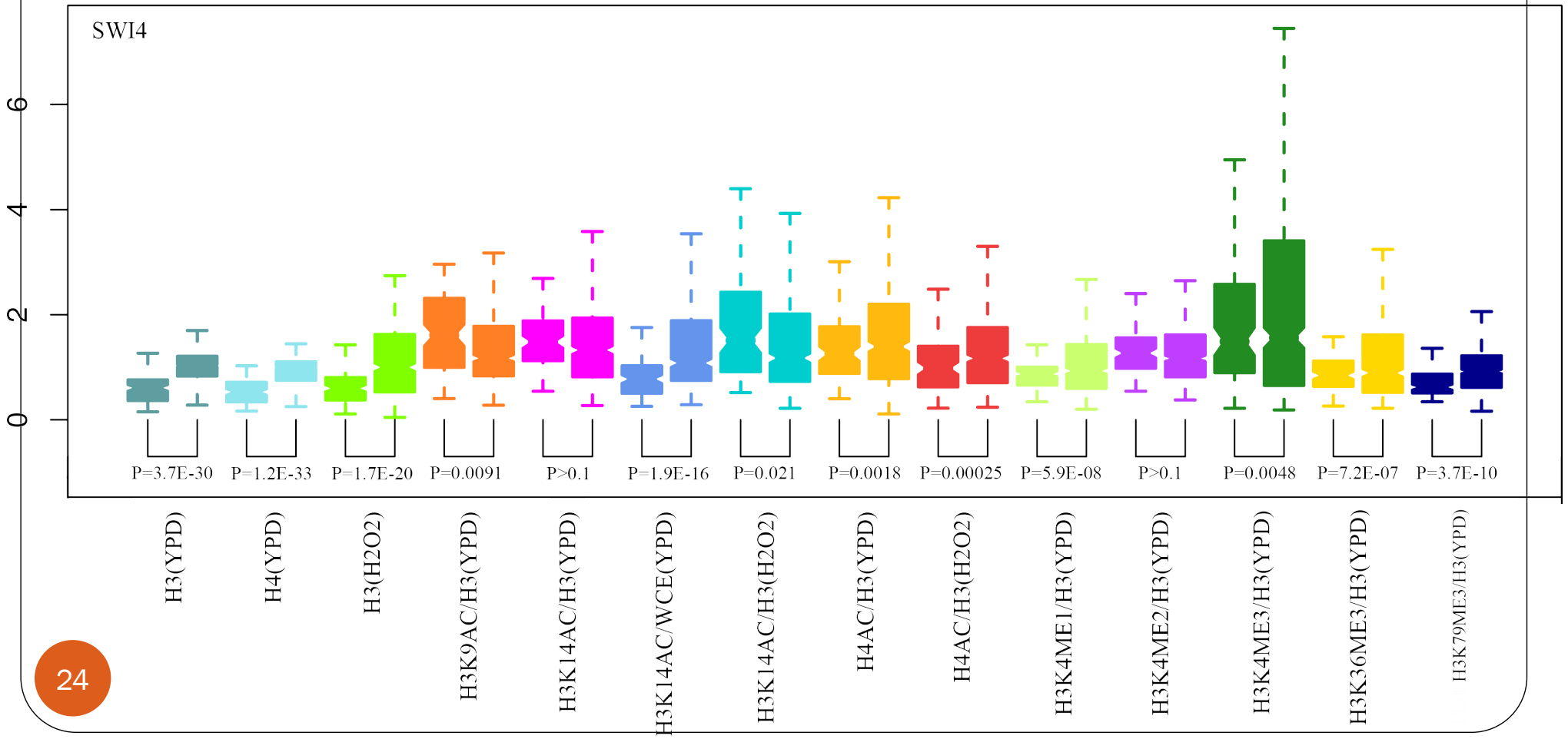
  - TF binding data:

Differential chromatin features between functional TFBS and non-functional TFBS

- SWI4 as example
- Functional binding sites: SGD, 99
- Non-functional binding sites: 485, have PWMs but not bound by SWI4
- For each site, calculate the average signal over all probes within 100bp for each chromatin feature



Differential chromatin features between functional TFBS and non-functional TFBS



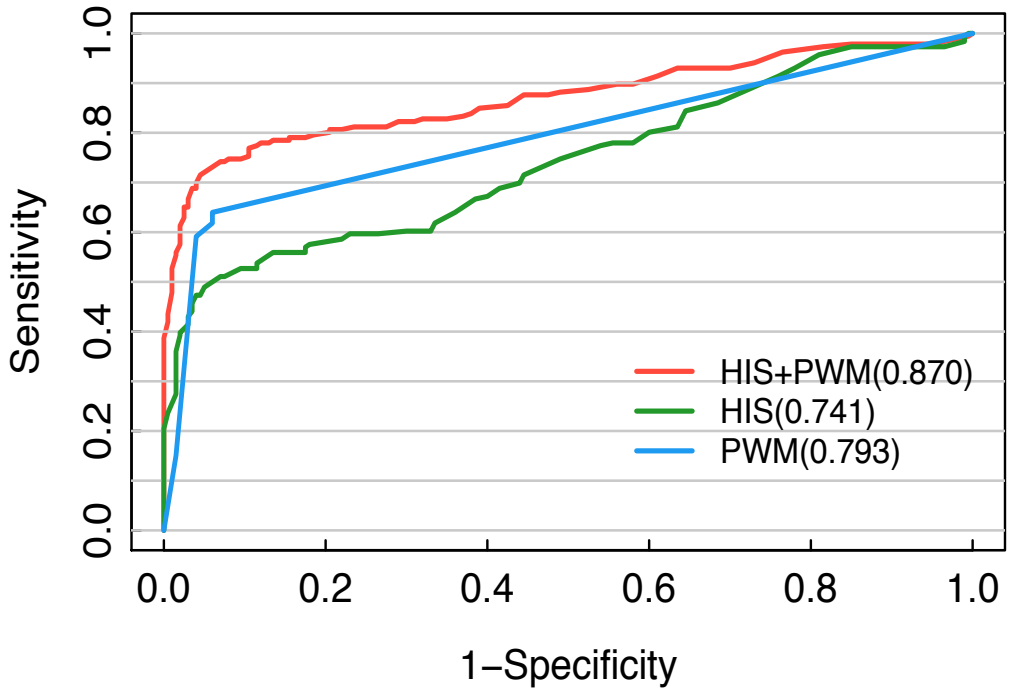
Final Data

Gene	TF target ?	PWM Score	H3 (up stream)	……	H3 (down-stream)	……
YAL001C	Yes	0.9945	1.555		1.222	
YAL056W	No	0	2.80		0.899	
……						
	Yes	0.2045	0.87		3.245	

Up-stream signal

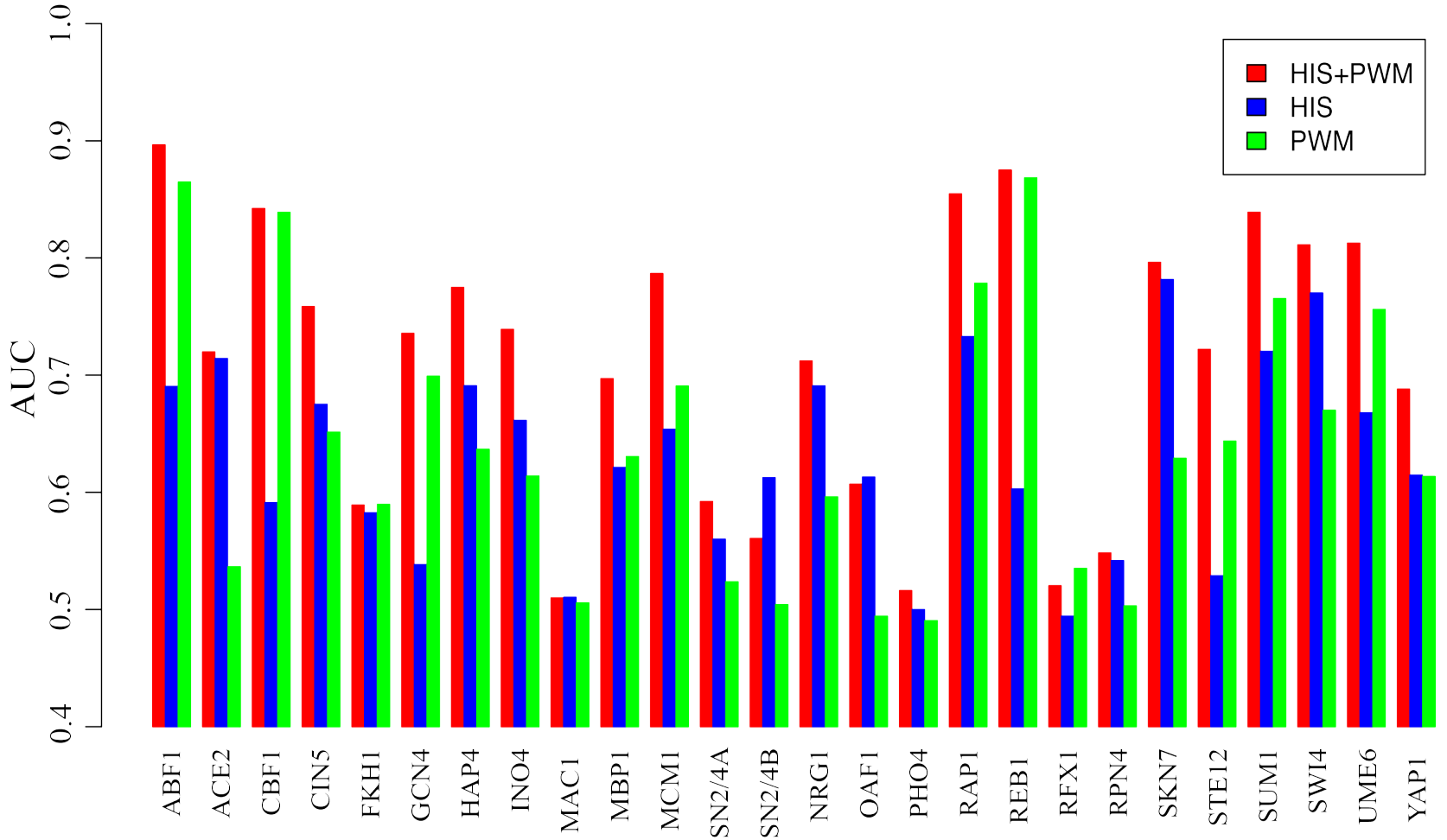
Down-stream signal

Chromatin feature improves target gene prediction

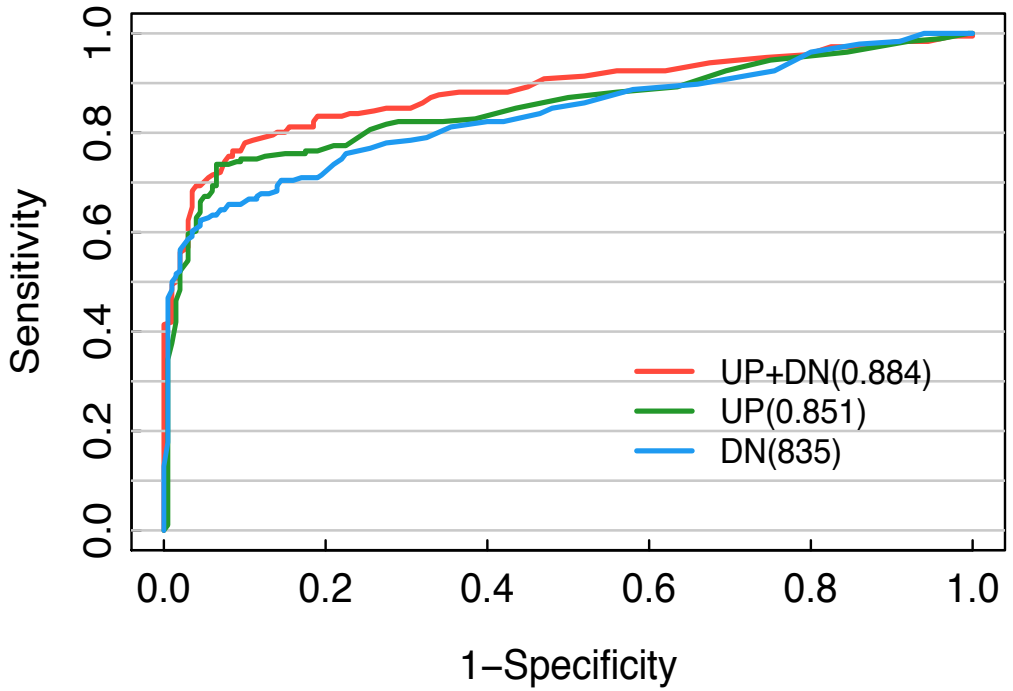


RAP1 (YPD), P<0.01

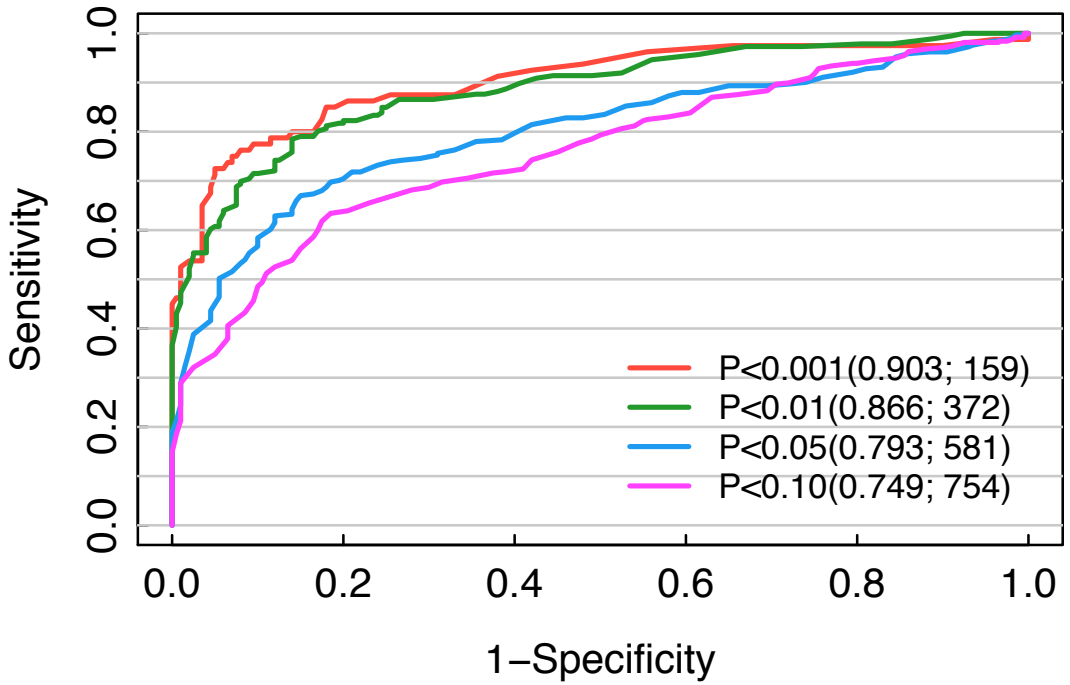
PWM I


AUC
0.4
0.5
0.6
0.7
0.8
0.9
1.0
HIS+PWM
HIS
PWM
ABF1
ACE2
CBF1
CIN5
FKH1
GCN4
HAP4
INO4
MAC1
MBP1
MCM1
MSN2/4A
MSN2/4B
NRG1
OAF1
PHO4
RAP1
REB1
RFX1
RPN4
SKN7
STE12
SUM1
SWI4
UME6
YAP1

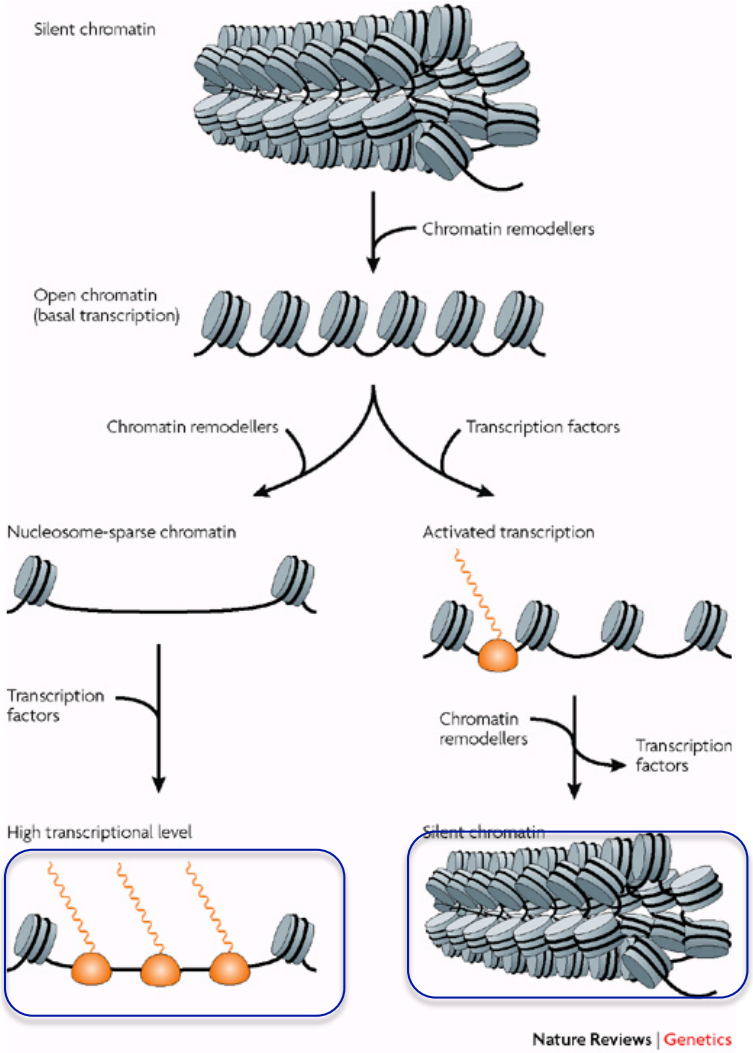
Up vs. down-stream signal



Different thresholds for target gene



Two-step approach for TFBS prediction in human and worm



- A supervised model that integrates all chromatin features to predict open chromatin regions in genome
  - Binding Active Region (BAR+)
  - confer tissue specificity
- Search for PWM
  - confer TF specificity

Report BAR+PWM+ as TFBS

Chromatin model

ChIP-seq for 15 human TFs

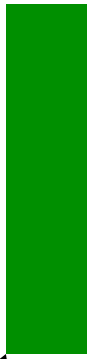
Encode data:
histone modification, DNA hypersensitivity,
TF binding, …

Calculate the average signal for each bin

30226454 bins

Predictor Matrix
32 experiments

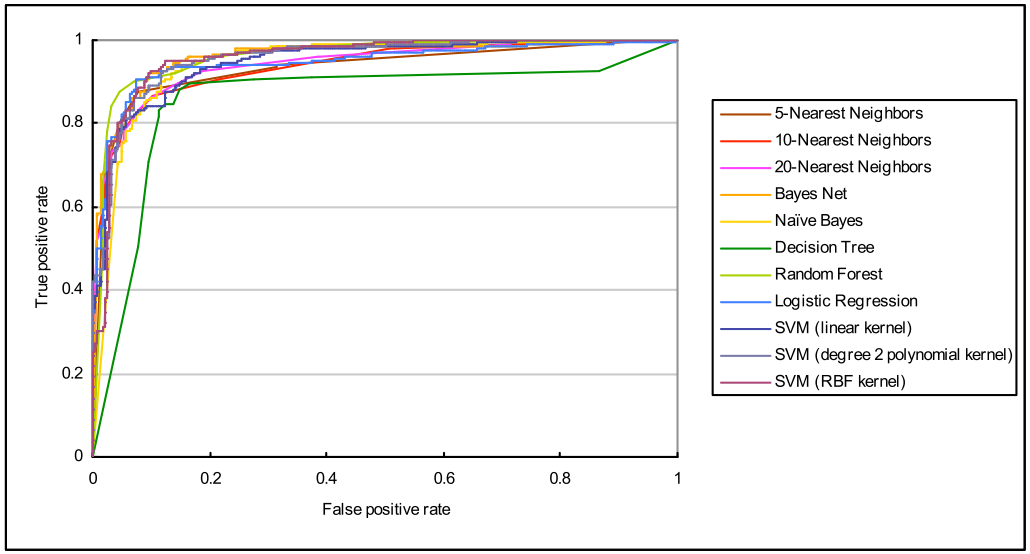
Response Y
Bound by any TF

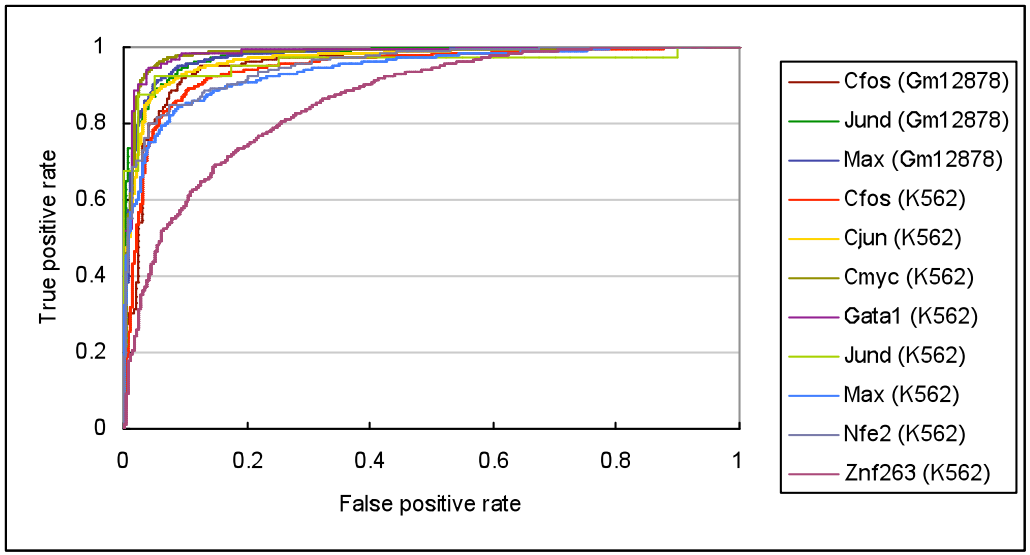


1: Positive bins (overlap with a peak: q-value>5%)

0: negative bins

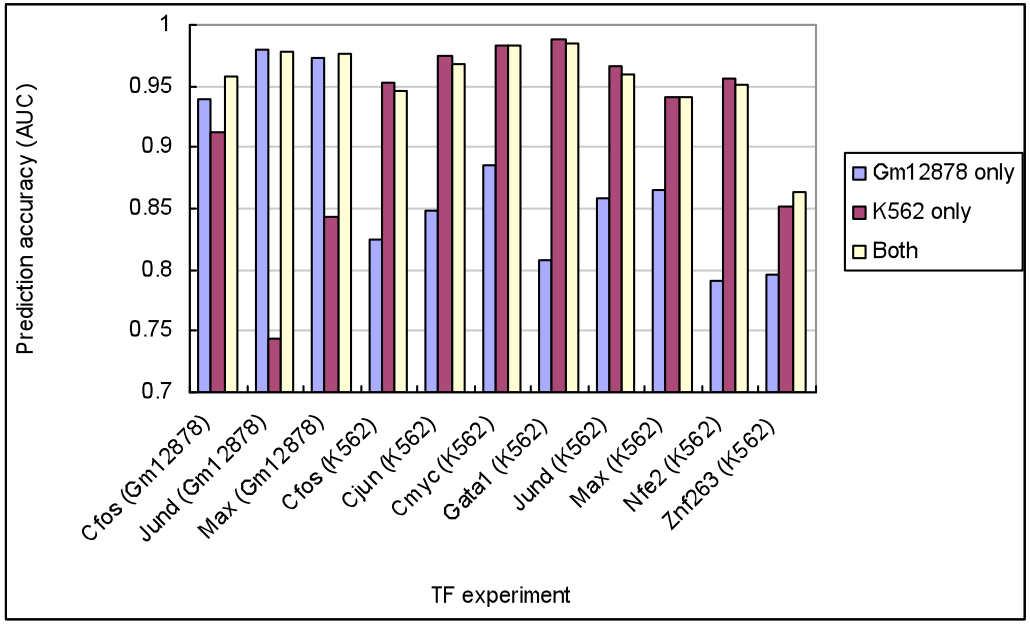
Chromatin features can accurately predict TF binding bins

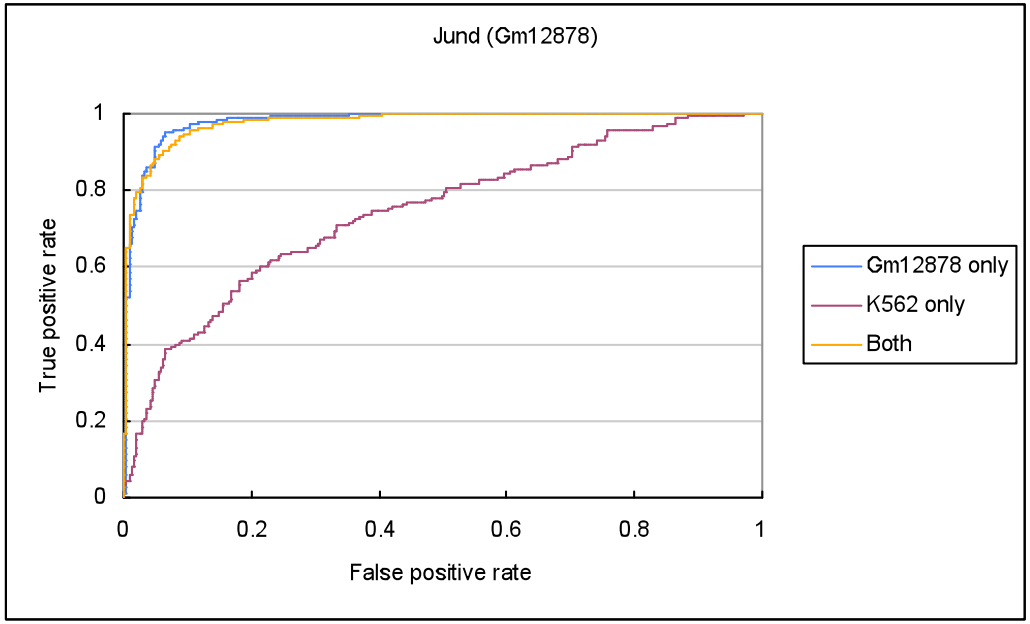




Human

Prediction model is cell line specific





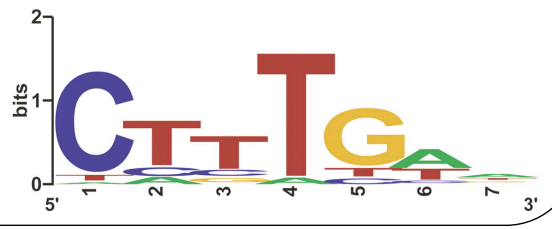
Chromatin model is not TF specific

- The chromatin model largely predicts the chromatin structure, the accessibility of a DNA region and thereby not TF specific.
- PWM in BAR+ regions are more likely to be functional TFBS.
- Previous TFBS prediction method based on PWM only is not effective in practice due to high positive rate.
- BAR+PWM+ reduce false positive rates

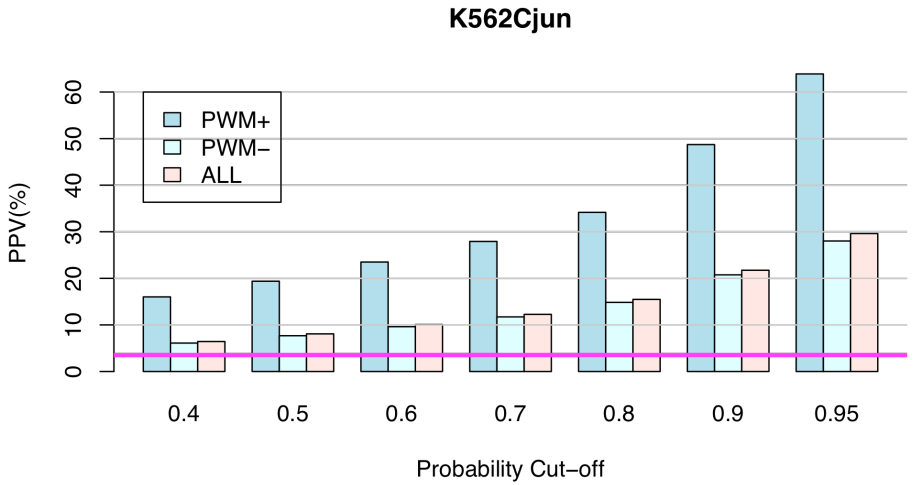
A

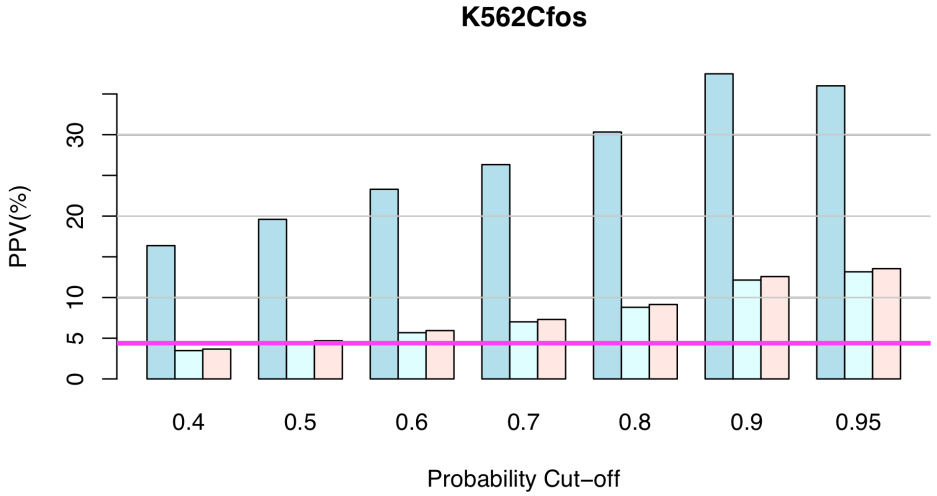
	1	2	3	4	5	6	7
A	1	4	1	2	0	17	13
C	28	5	5	0	3	3	2
G	0	0	4	0	25	1	7
T	2	22	21	29	4	10	9

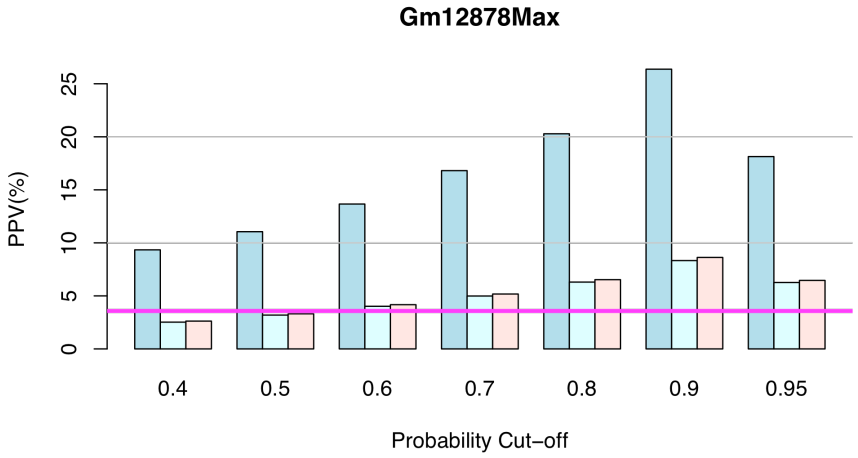
B



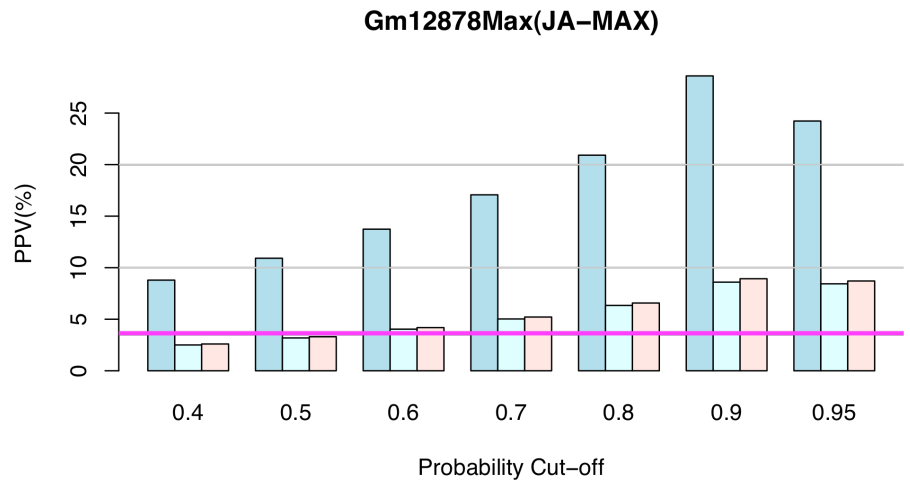
PWM enhances prediction precision

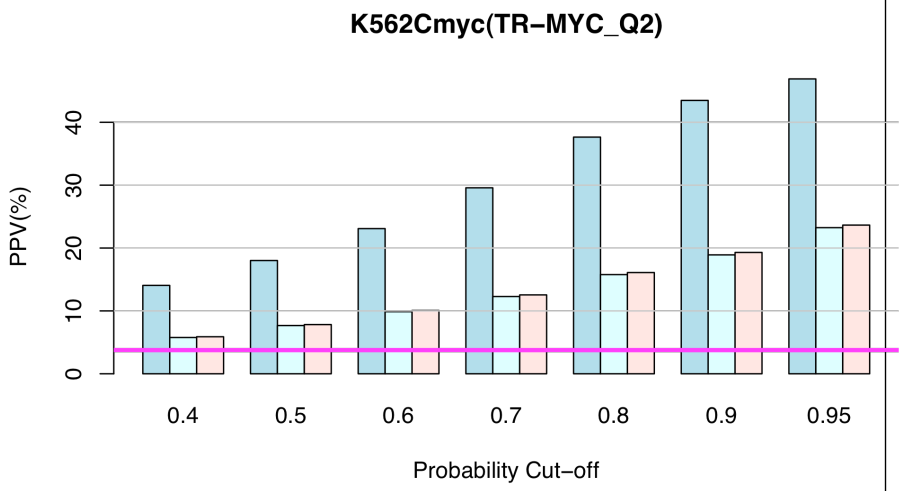


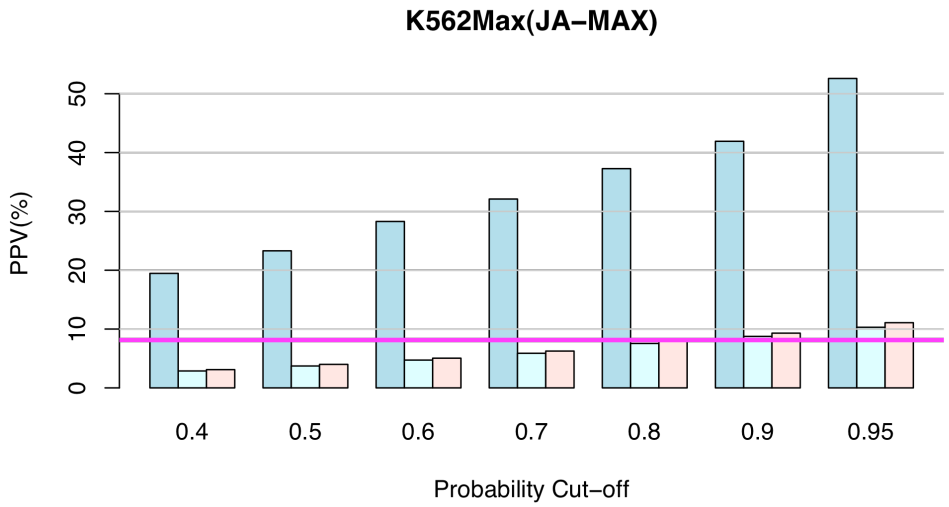


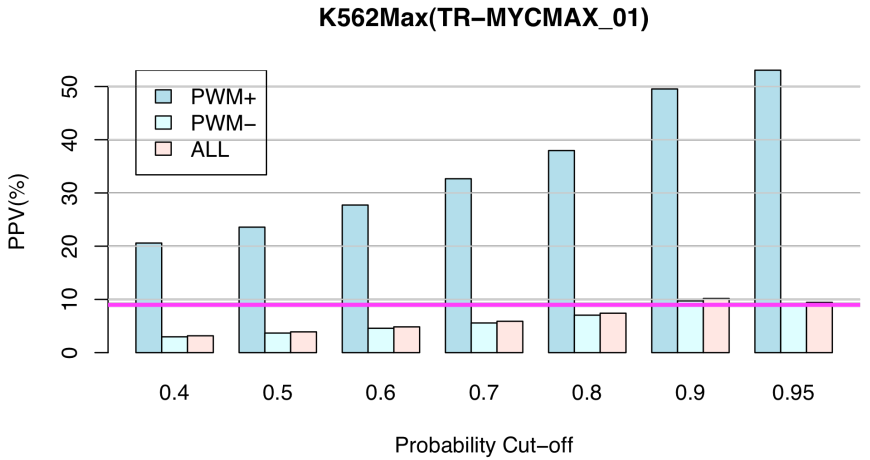


Using PWMs from database





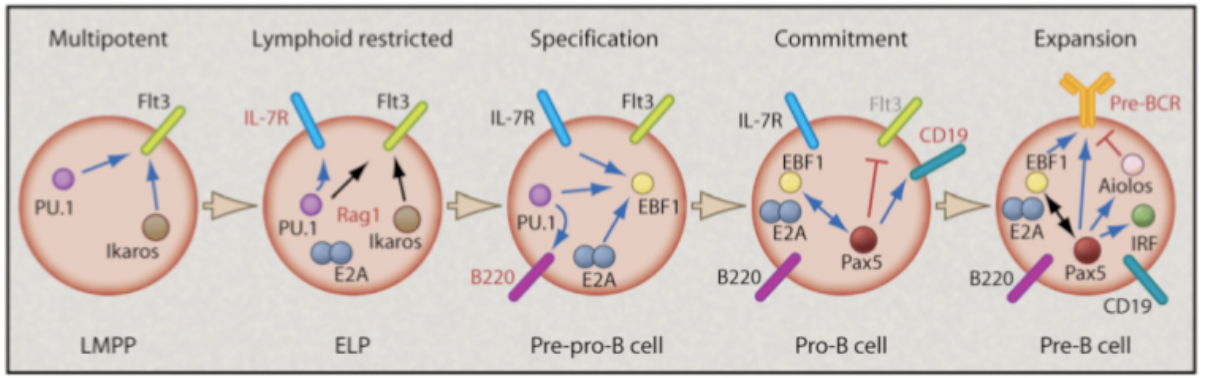




Applications of TFBS prediction method

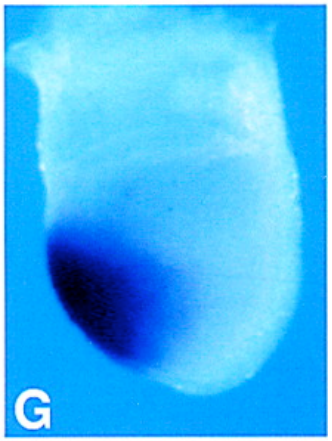
- Identify TFs that have differential activity between leukemia and normal blood cell.

  - Ikaros

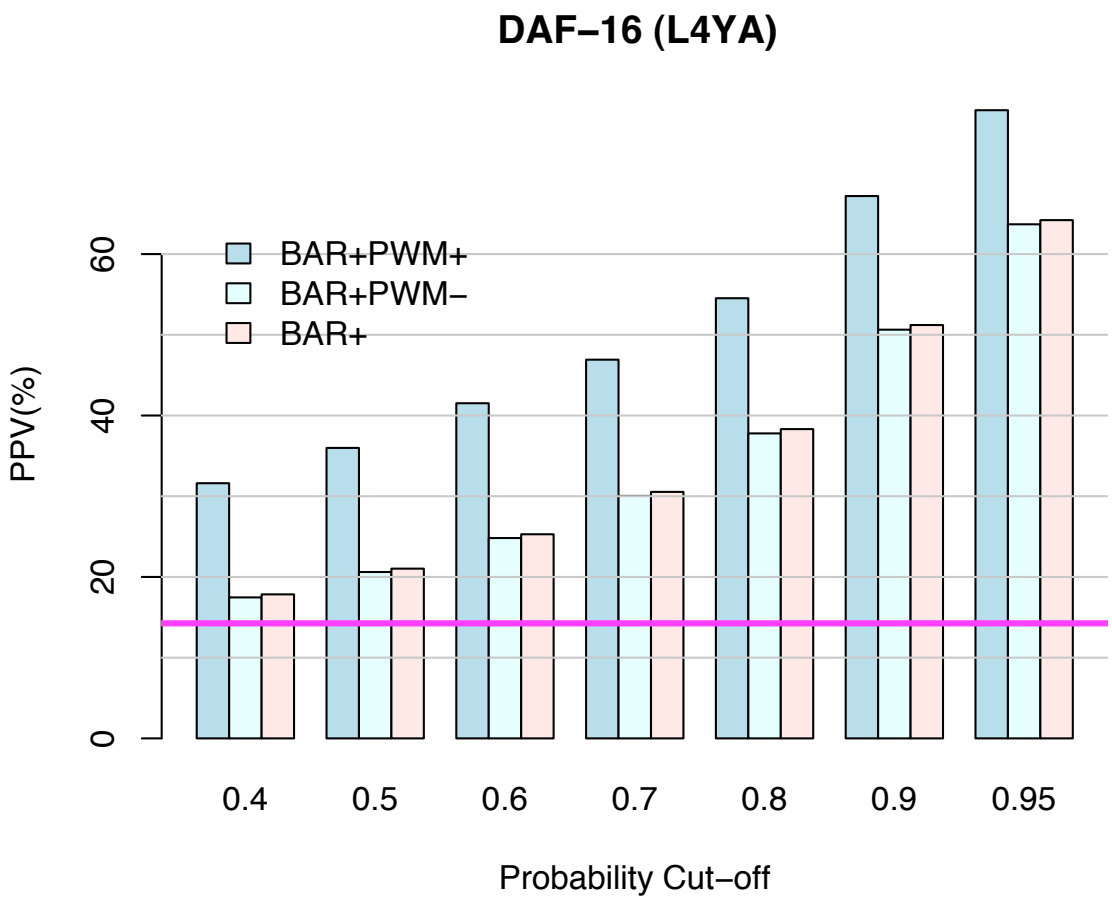


- Predict human enhancers

  - 50 enhancers, 4 have been experimentally tested in mouse embryo cells, among which 3 proved to be true positive

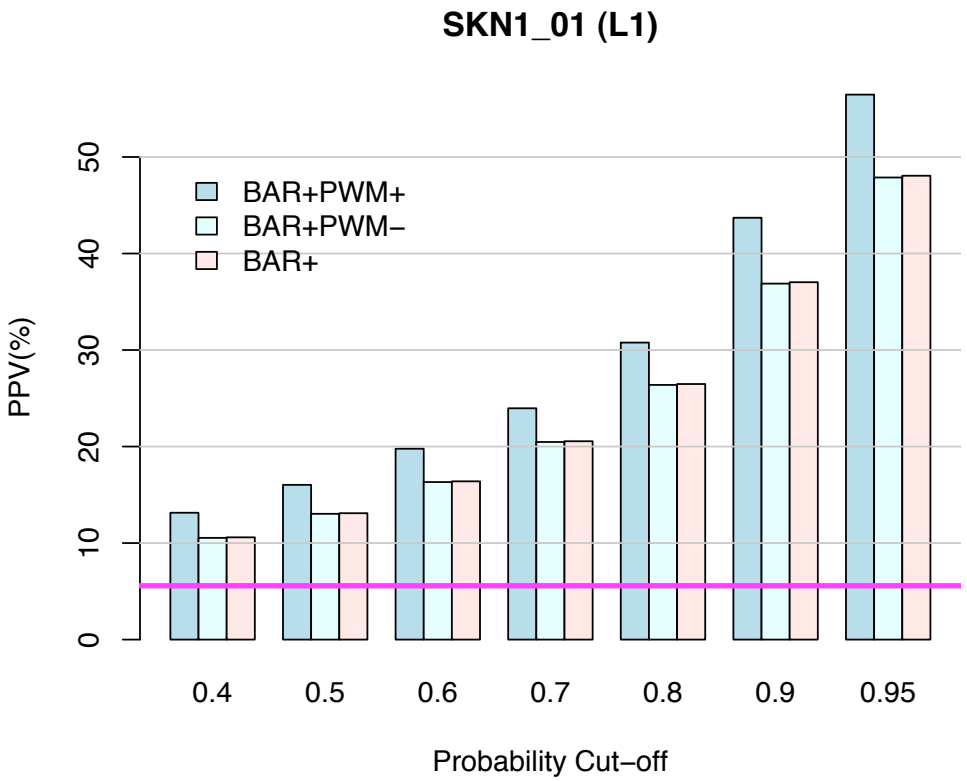


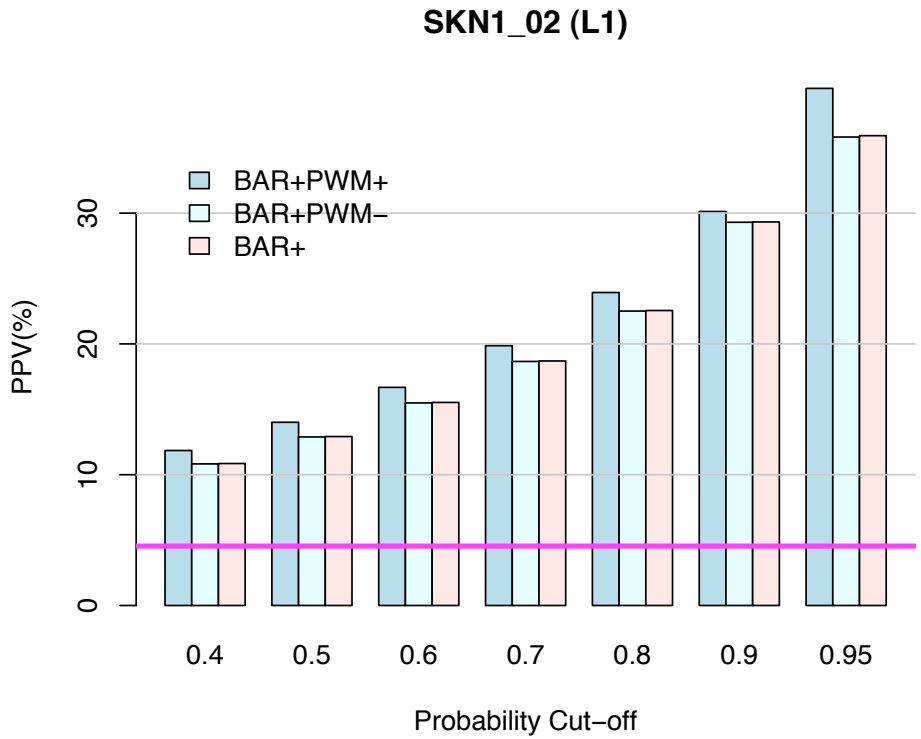
Predicting TFBS in *C.elegans*

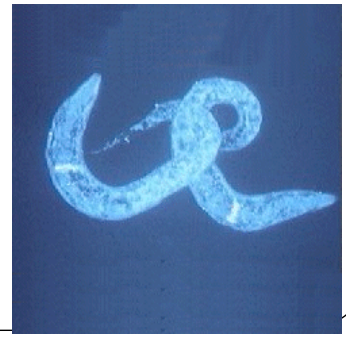


worm

SKN-1


SKN1_01 (L1)
BAR+PWM+
BAR+PWM−
BAR+
PPV(%)
0
10
20
30
40
50
0.4
0.5
0.6
0.7
0.8
0.9
0.95
Probability Cut−off


SKN1_02 (L1)
BAR+PWM+
BAR+PWM−
BAR+
PPV(%)
0
10
20
30
0.4
0.5
0.6
0.7
0.8
0.9
0.95
Probability Cut−off



Acknowledgements

- Kevin Yip
- Koon-Kiu Yan
- Joel Rozowsky
- Chong Shou
- Roger Alexander
- Mengjie Chen
- Genome Tech® & Nets®
- Mark Gerstein