Gene Expression Prediction using Chromatin Models

ENCODE AWG
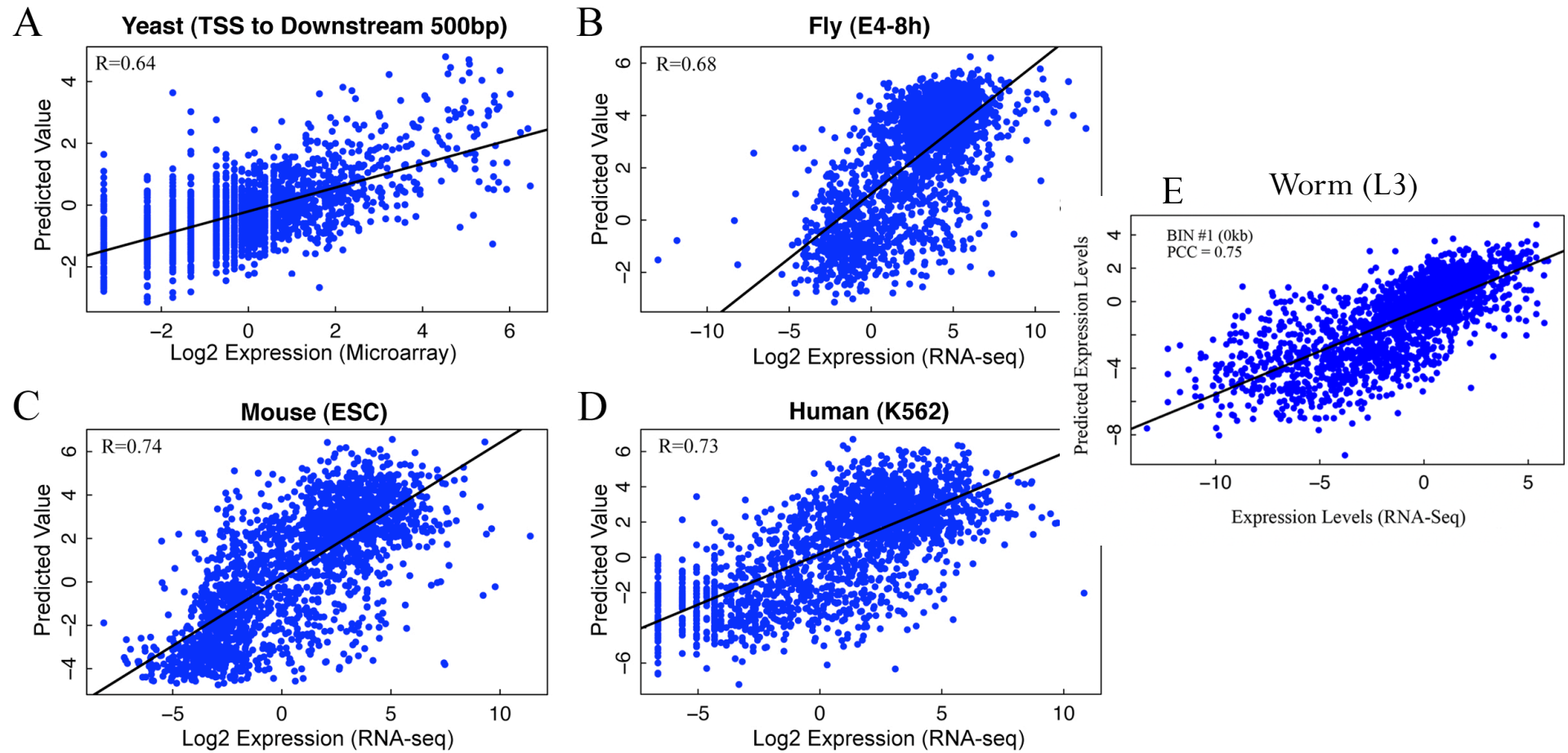
Chao Cheng (the Gerstein Lab)

Jan 20, 2011
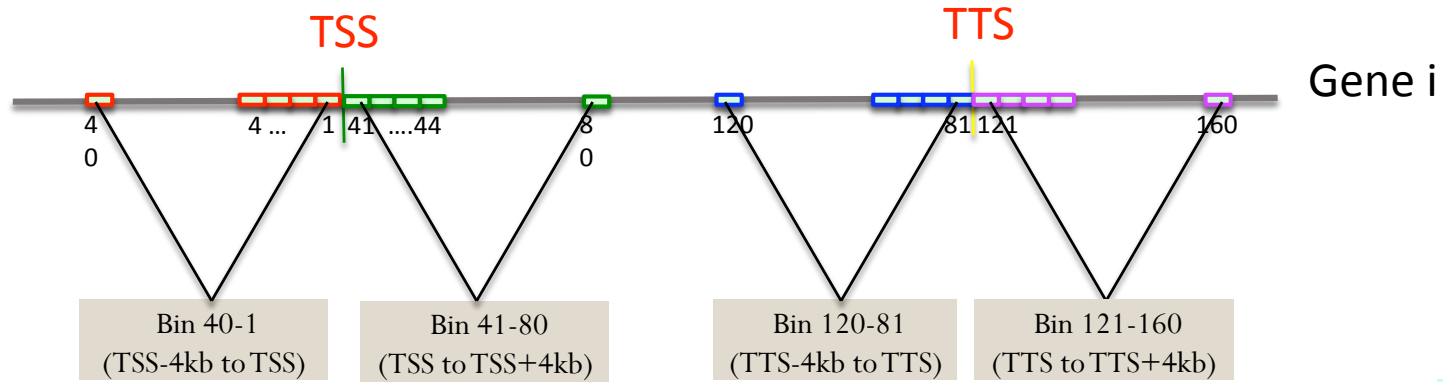
Outline

- Chromatin model for predicting gene expression levels based on integrative models

  Support Vector Machine (SVM) – classify highly and lowly expressed gens

  Support Vector Regression (SVR) – predict gene expression levels

- Tissue/cell specificity of the chromatin model

- Predict differentially expressed genes (K562/GM12878)

- Chromatin model for predicting highly and lowly expressed miRNAs

- Transcription factor (TF) 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
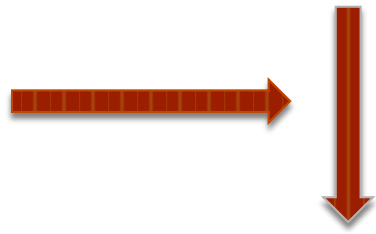
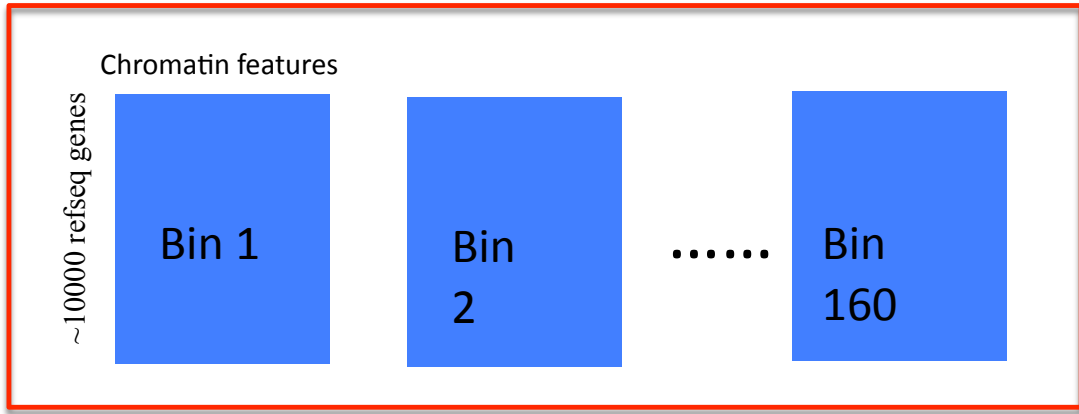# Application of our integrative model in 5 species

Divide and compare strategy

TSS TTS

Gene i

4
0

4 ...   1 41 44   8
0

120   81 121   160

Bin 40-1
(TSS-4kb to TSS)

Bin 41-80
(TSS to TSS+4kb)

Bin 120-81
(TTS-4kb to TTS)

Bin 121-160
(TTS to TTS+4kb)

RNA-Seq data

Chromatin features:
Histone methylation
Pol II binding

Predictors

Prediction target:
Gene expression level

Chromatin features

~10000 refseq genes

Bin 1   Bin 2   ......   Bin 160

Correlation pattern of chromatin features with gene expression

A Worm EEMB

H3 (Rep1)
H3 (Rep2)
H3K4ME2
H3K4ME3
H3K9ME2
H3K9ME3 (Rep1)
H3K9ME3 (Rep2)
H3K36ME2 (Rep1)
H3K36ME2 (Rep2)
H3K36ME3
H3K79ME1
H3K79ME2
H3K79ME3
MES4
MRG1
POLII

-4kb ← TSS → 4kb   -4kb ← TTS → 4kb

B Human K562

DNase (Rep1)
DNase (Rep2)
FAIRE
H3K9AC
H3K27AC
H3K4ME1
H3K4ME2
H3K4ME3 (Rep1)
H3K4ME3 (Rep2)
H3K9ME1
H4K20ME1
H3K27ME3
H3K36ME3
POLIIb
POLII (Rep1)
POLII (Rep2)

-4kb ← TSS → 4kb   -4kb ← TTS → 4kb

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

Predict gene expression class: (SVM classification)
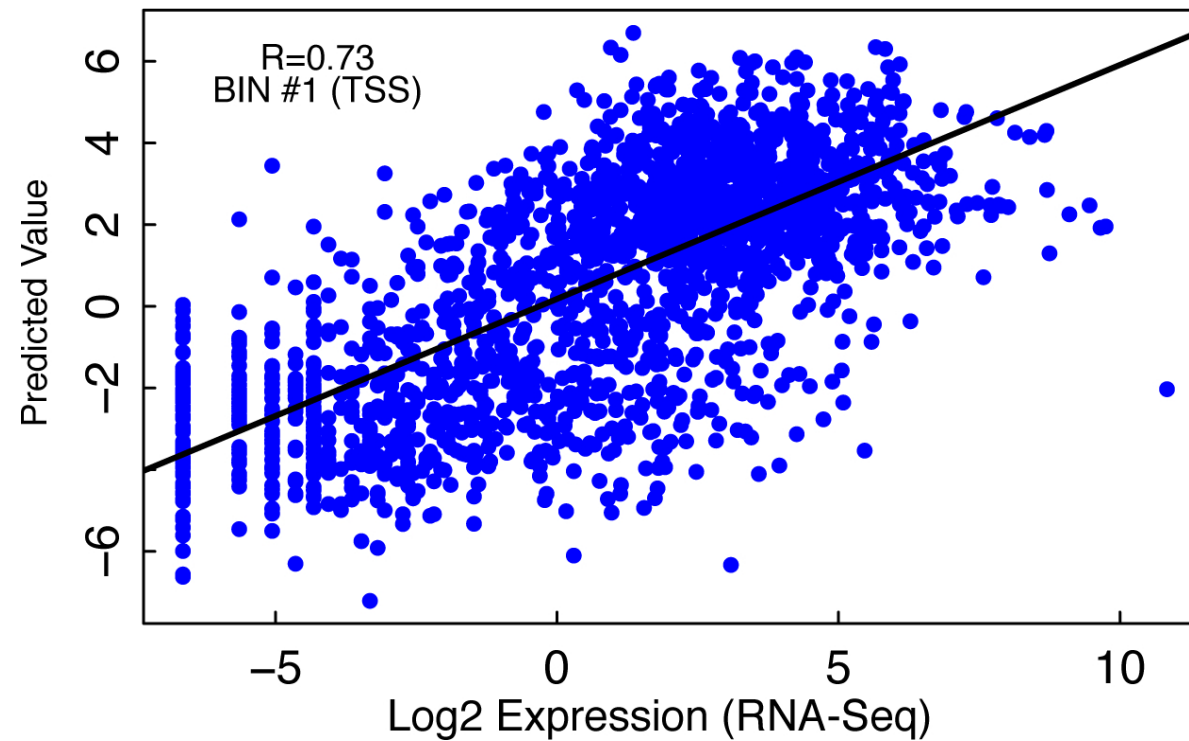
Highly expressed genes
Lowly expressed genes
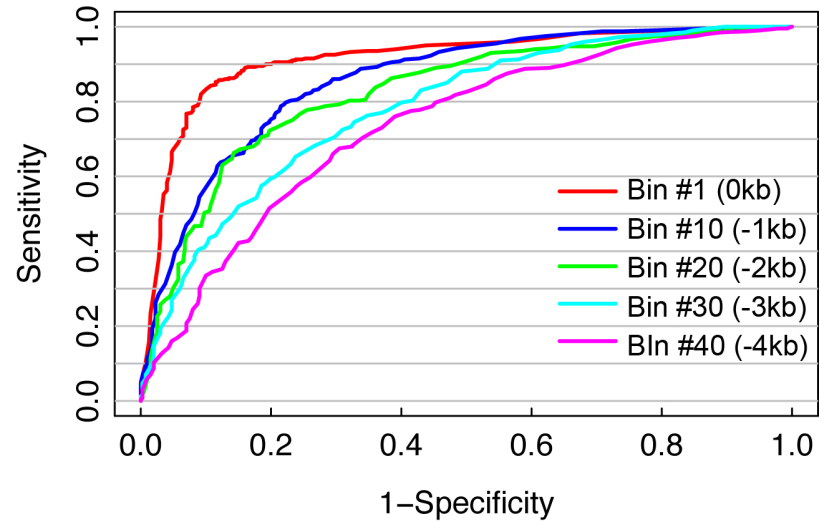
Human K562 using 13 chromatin features
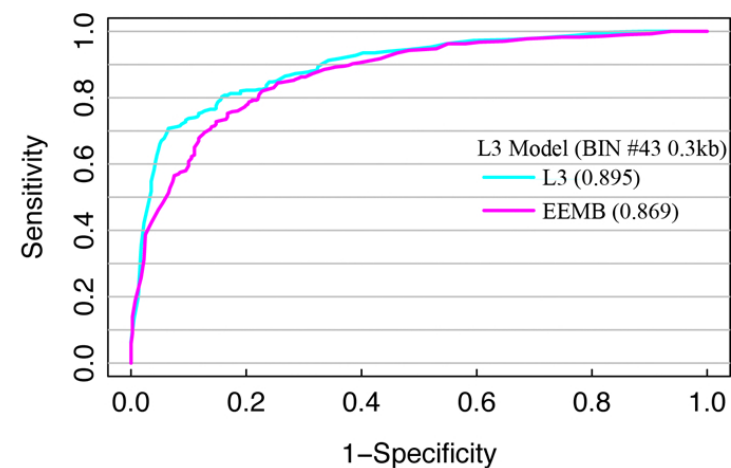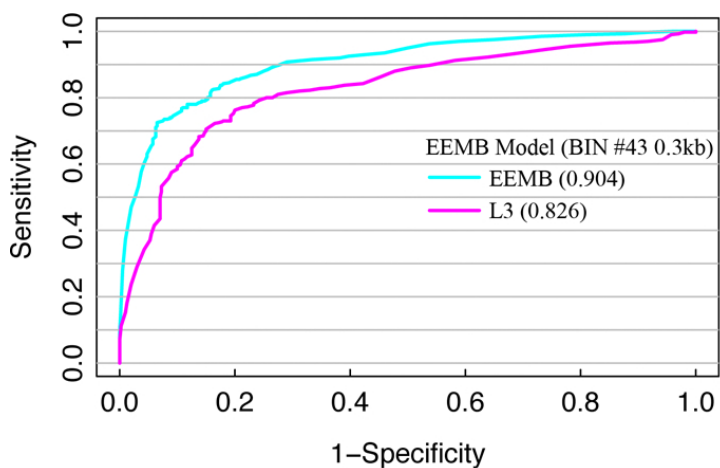

AUC for BIN i model
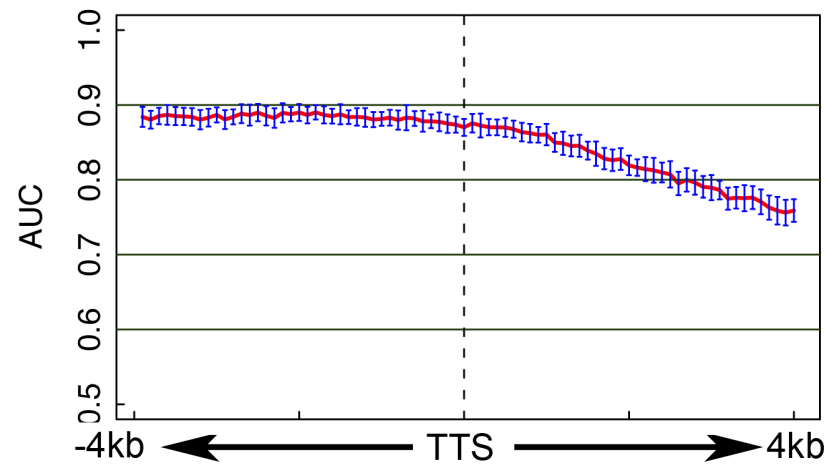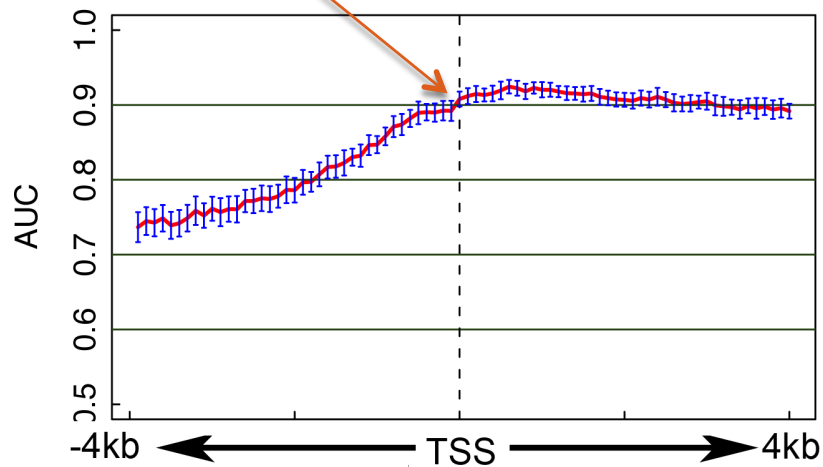
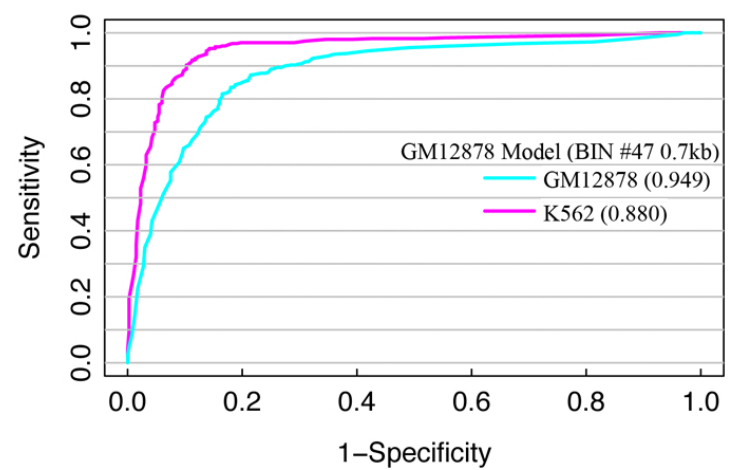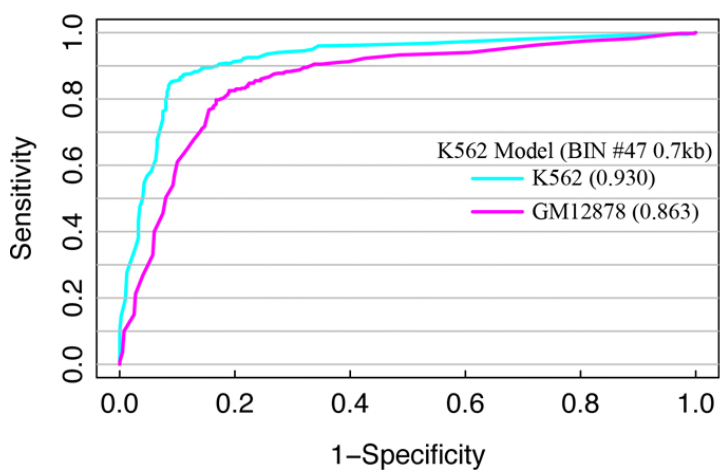# Predict gene expression levels: Support Vector Regression (SVR)

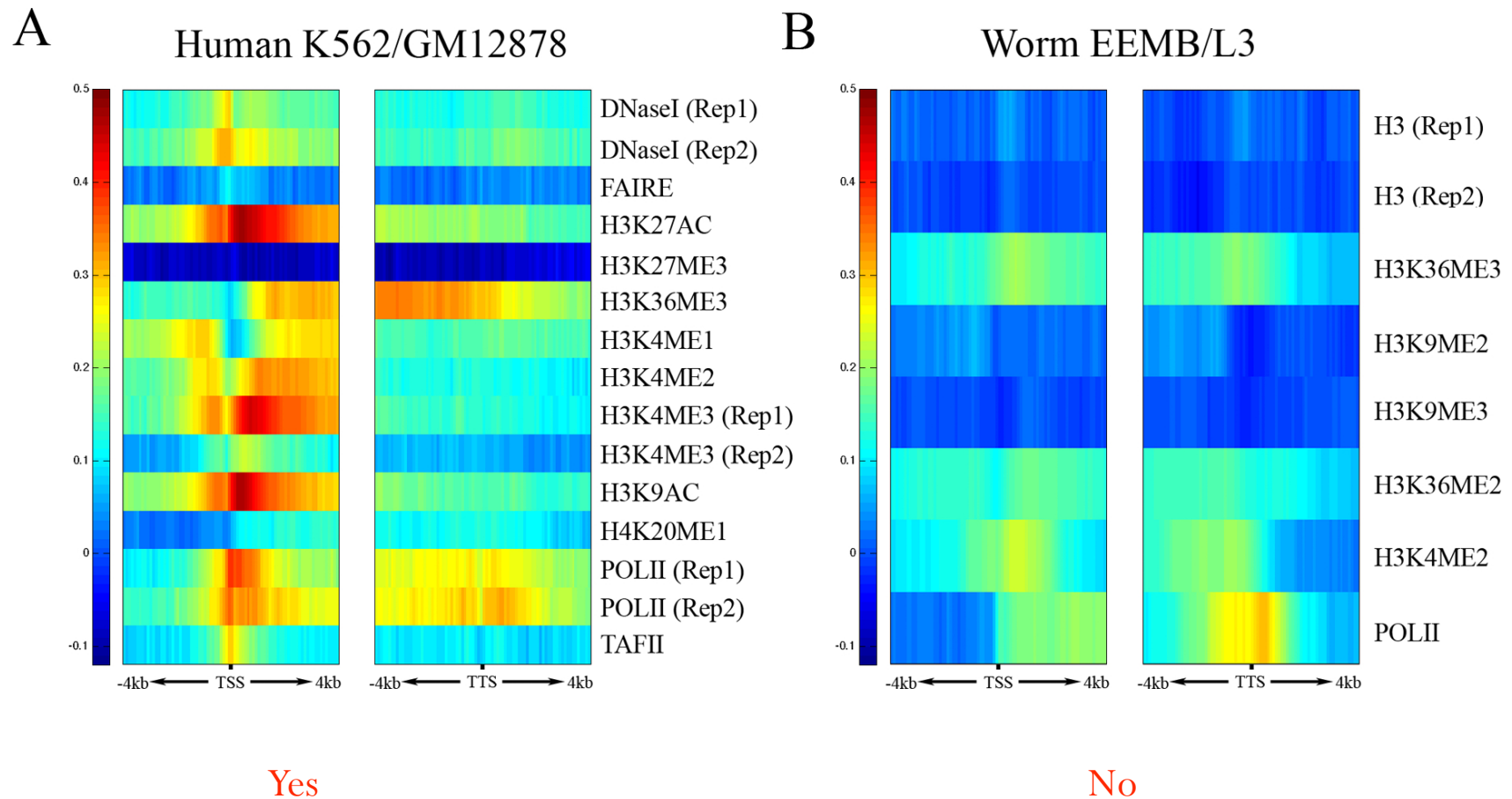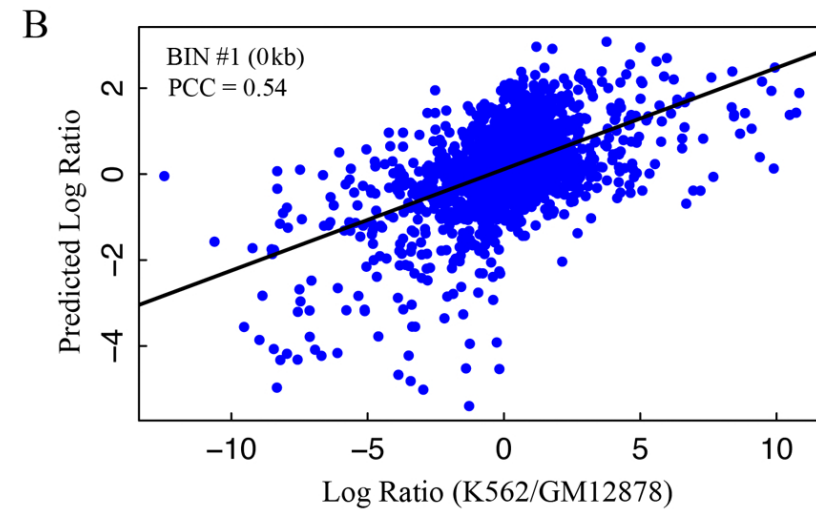# Chromatin model is tissue/cell line specific and development stage specific

Predicting differential expression

A   Human K562/GM12878

B   Worm EEMB/L3

Yes
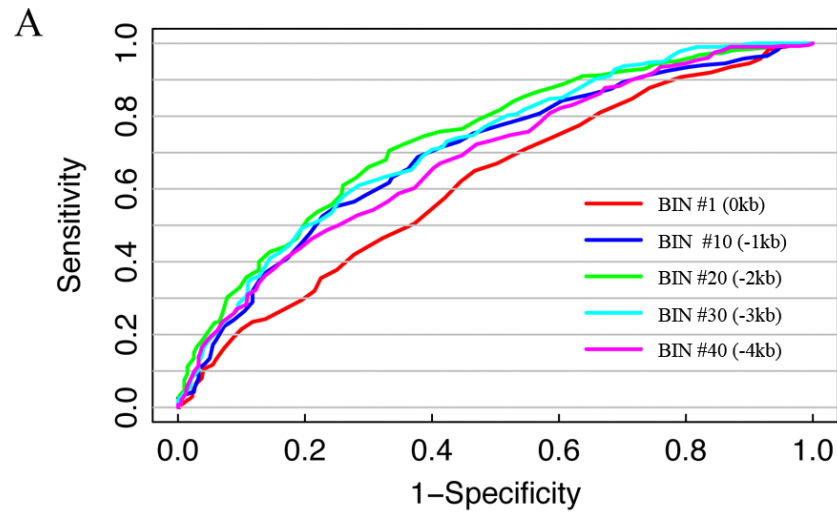
No

9

Predict differential expression

Human K562/GM12878
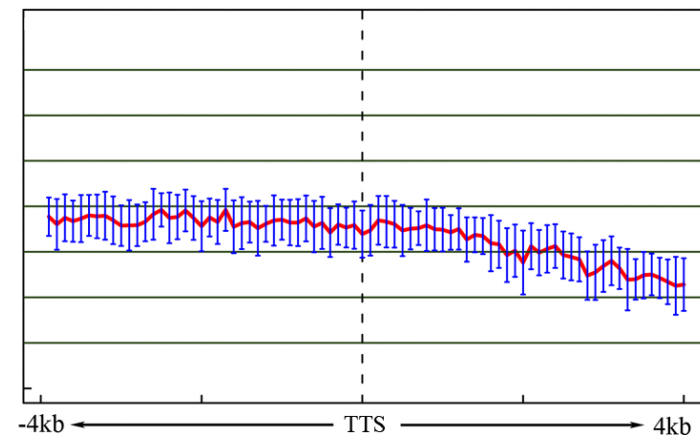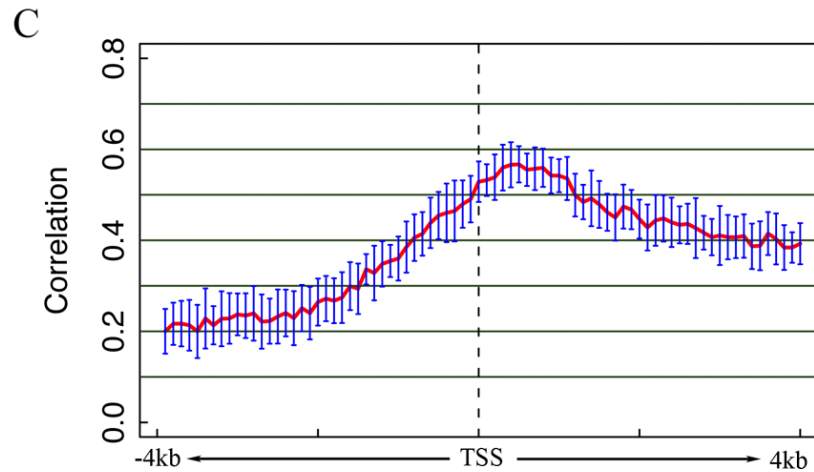
Predicting human microRNA expression

- 730 human 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 trained using protein-coding genes

- Validate prediction results from experimental data
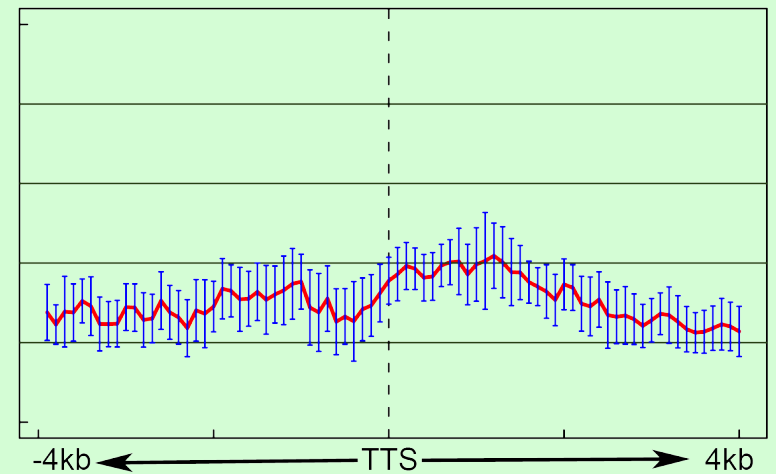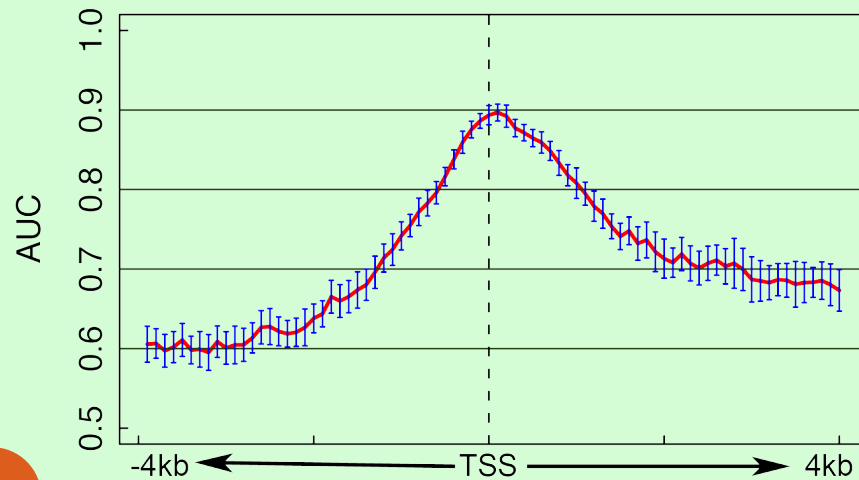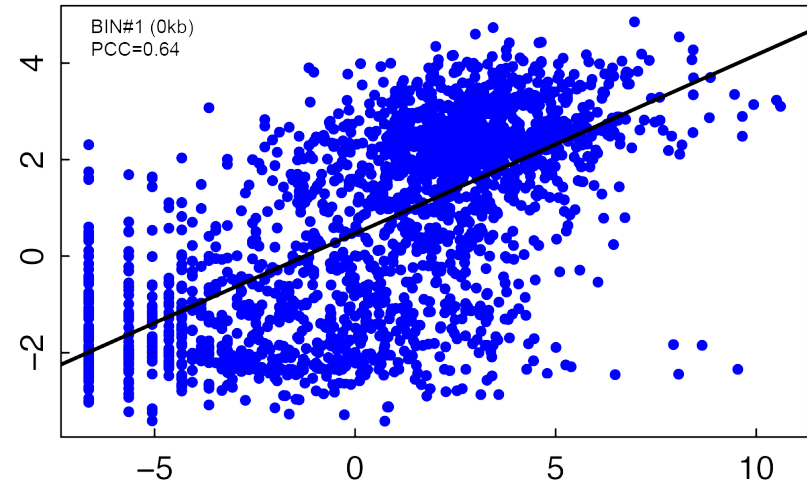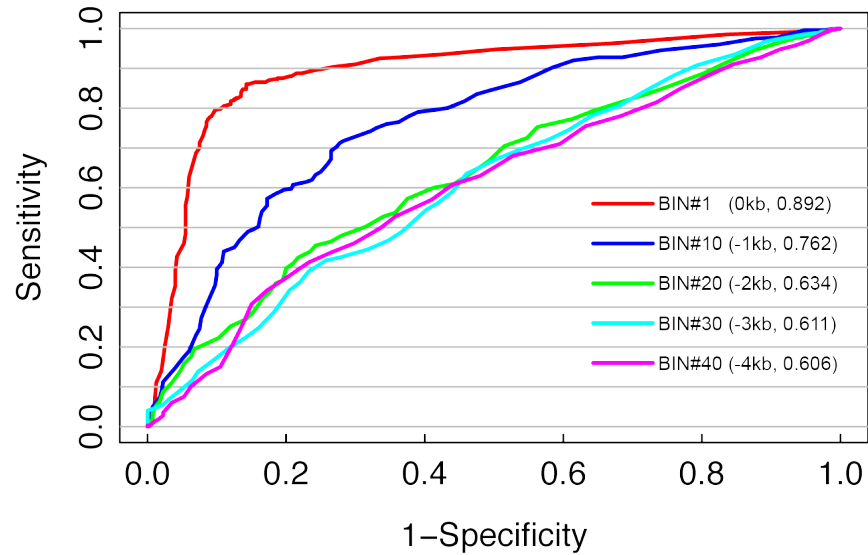
  - microRNA expression in cellular components

Predict miRNA expression
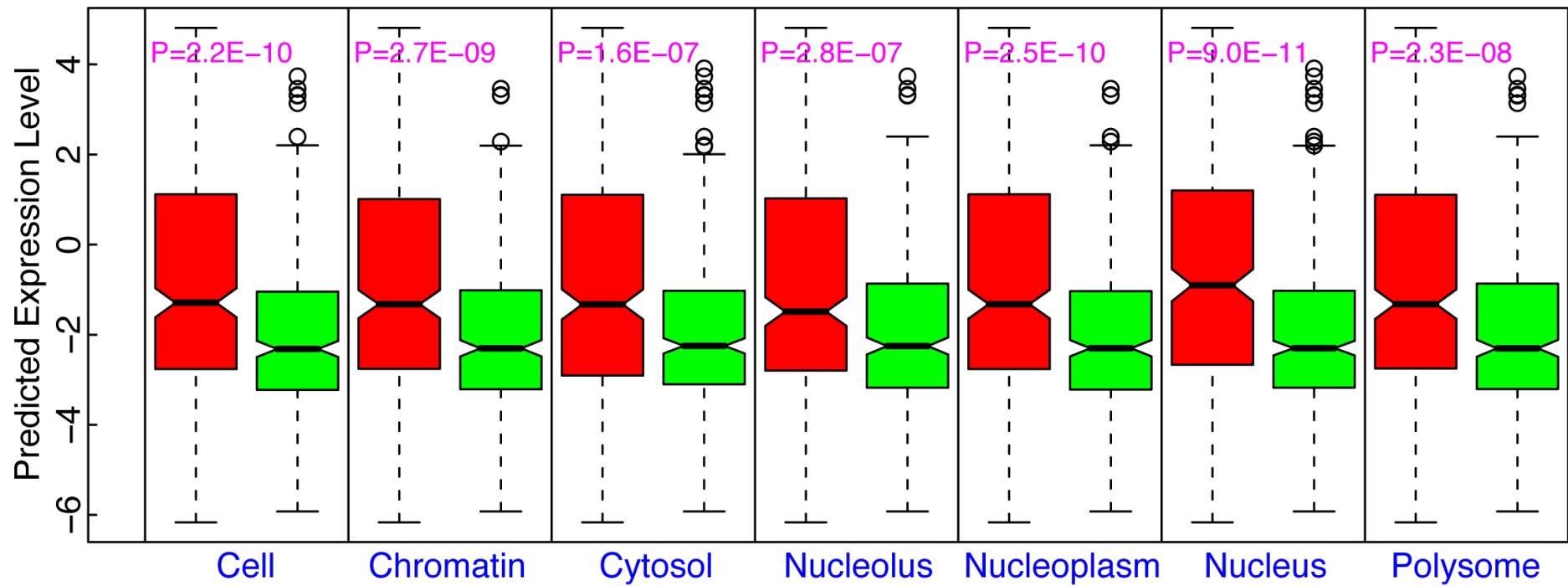
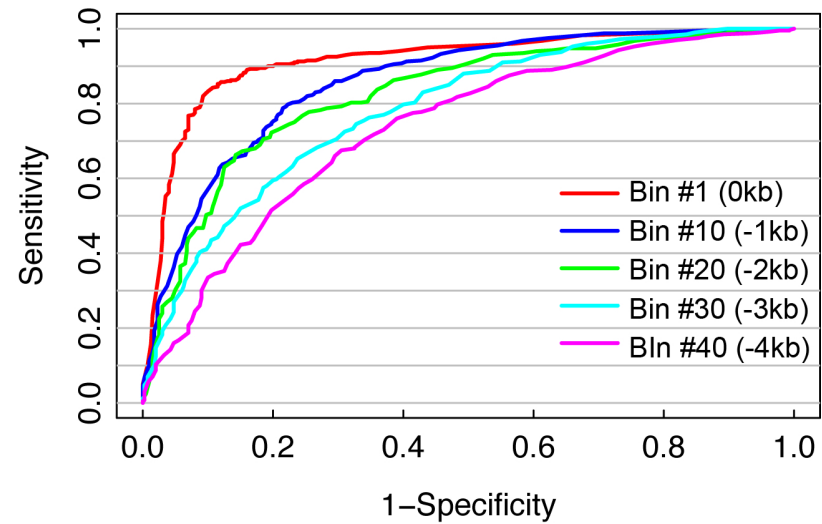# Correlation patterns of TFs

Correlation Pattern
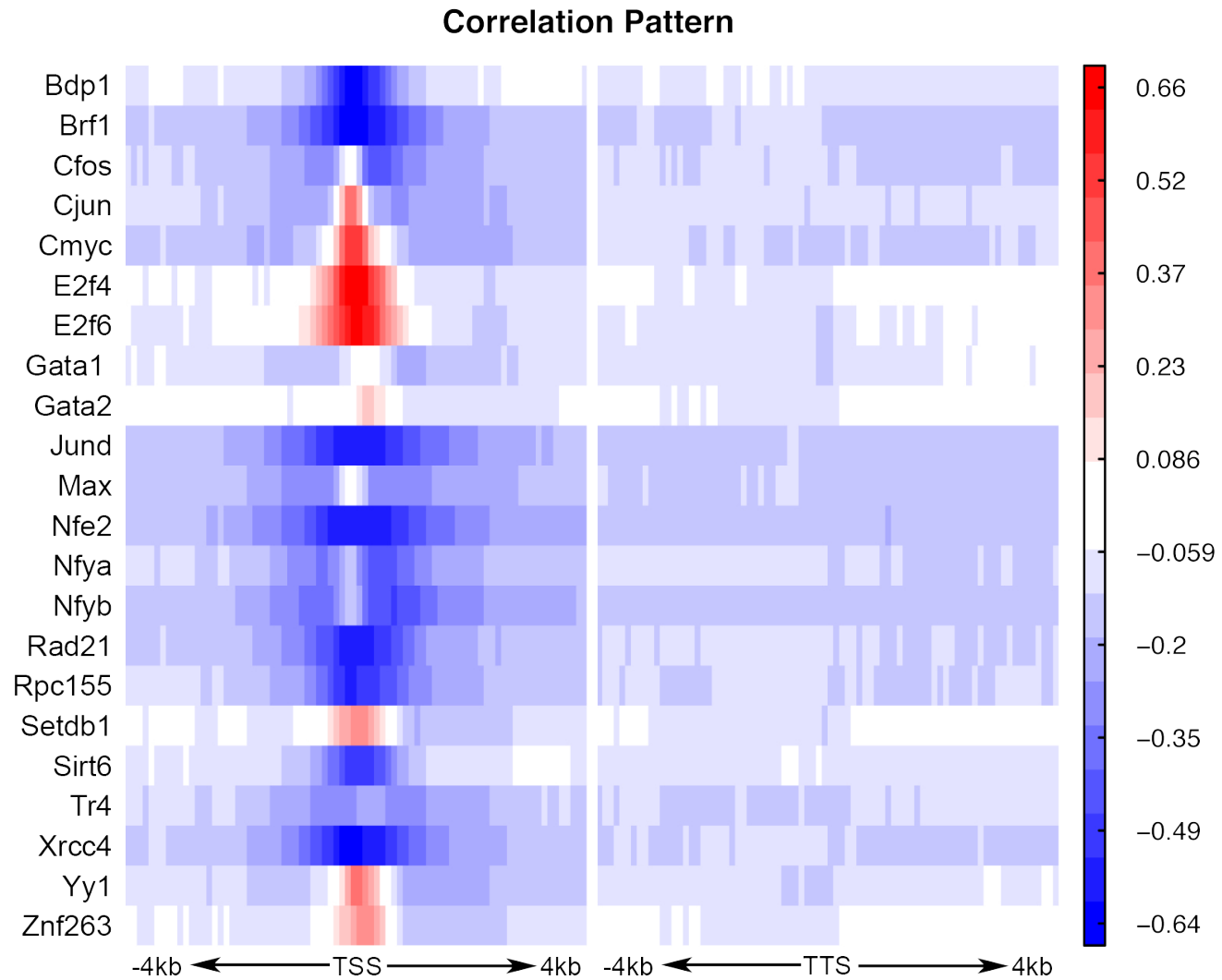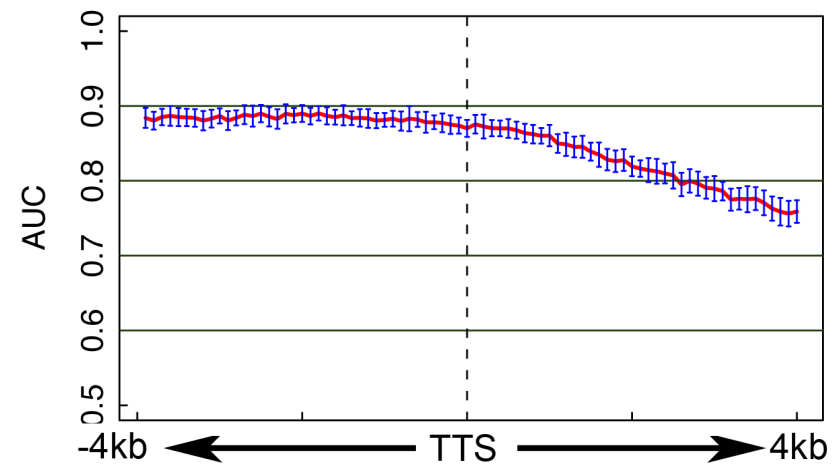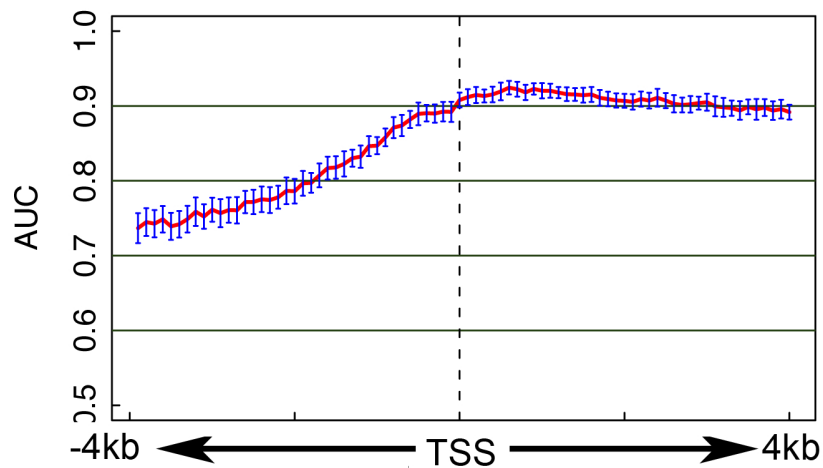
TF model: prediction power

Histone Model: prediction power

Different with TF model

Conclusion

- Chromatin features can accurately predict gene expression
- Chromatin model is tissue/cell line/development stage specific
- 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.
- TF model is also predictive of gene expression
- HM and TF models are different in their prediction pattern

Acknowledgement

- Koon-Kiu
- Mark Gerstein
- Michael Snyder
- Kevin Yip
- Jason Lieb
- Shirley Liu
- Roderic Guigo
- Tom Gingeras
- ENCODE & modENCODE consortium