Understanding transcriptional regulation by integrative analysis of transcription factor binding data

Cheng et al. 2012

Shu Yang

Feb. 21, 2013

Introduction

DNA-binding Proteins

- sequence-specific TFs (TFSS): MYC, MAX
- general or nonspecific TFs (TFNS): TBP (TATA-binding proteins)
- chromatin structure factors (ChromStr): CHD2
- chromatin remodeling factors (ChromRem)
- histone methyltransferases (HISase)
- Pol3-associated factors (Pol3F): POLR3A

Gene expression

- is the process of producing a specific amount of gene product in a spatiotemporal manner.
- is regulated in steps including: transcriptional regulation, splicing, end modification, export, and degradation.
- Transcriptional regulation can occur on both genetic and epigenetic levels.

Questions

- Is there much difference in the prediction accuracy of expression levels of TSSs captured by different technologies(CAGE,RNA-PET,RNA-seq)
- What is the effect of promoter CpG content on gene expression?
- Do TFs regulate alternative TSSs in the same mechanisms?
- Between two cell lines, can the difference of TF-binding signals precisely reflect the differential expression of TSSs?
- Can TF-binding signals predict histone modifications?

Data and Models

ENCODE data

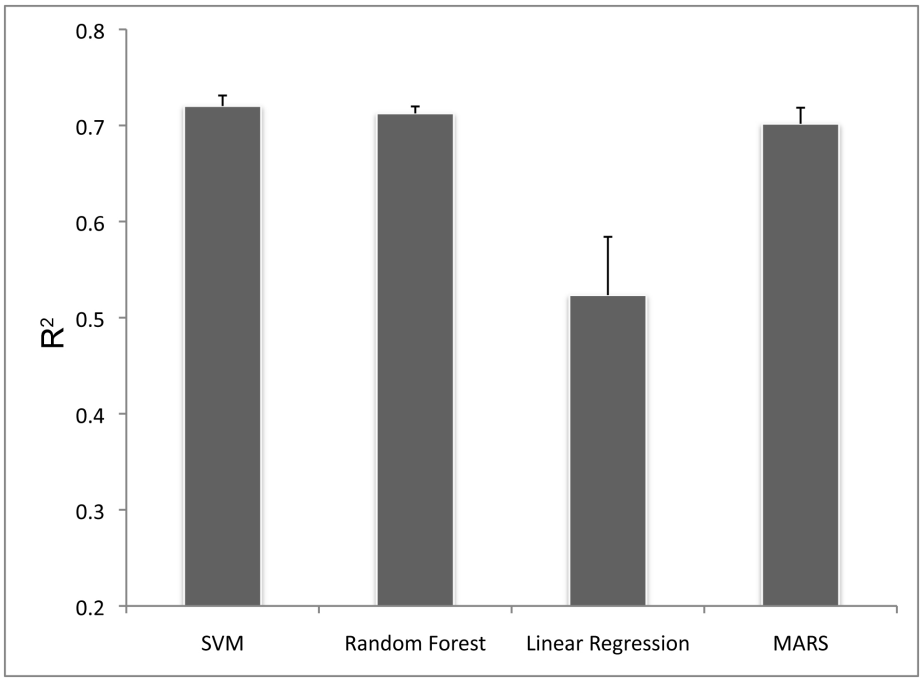
- Gene expression data (TSS):
  - >130,000 TSSs; 267 expression profiles; 12 cell lines (K562 and GM12878)
  - CAGE, RNA-seq, RNA-PET
- TF binding data:
  - >120 TFs; >400 binding profiles
  - ChIP-seq

Machine Learning Models

- Four methods:
  - multiple linear regression (MLR)
  - multivariate adaptive regression splines (MARS)
  - support vector regression (SVR): single predictor
  - random forest (RF): multiple predictors
- Evaluation:
  - Regression: R; R^2
  - Classification: AUC
  - 2000 promoters training; rest testing

Nonlinear relationship between TF binding and TSS expression

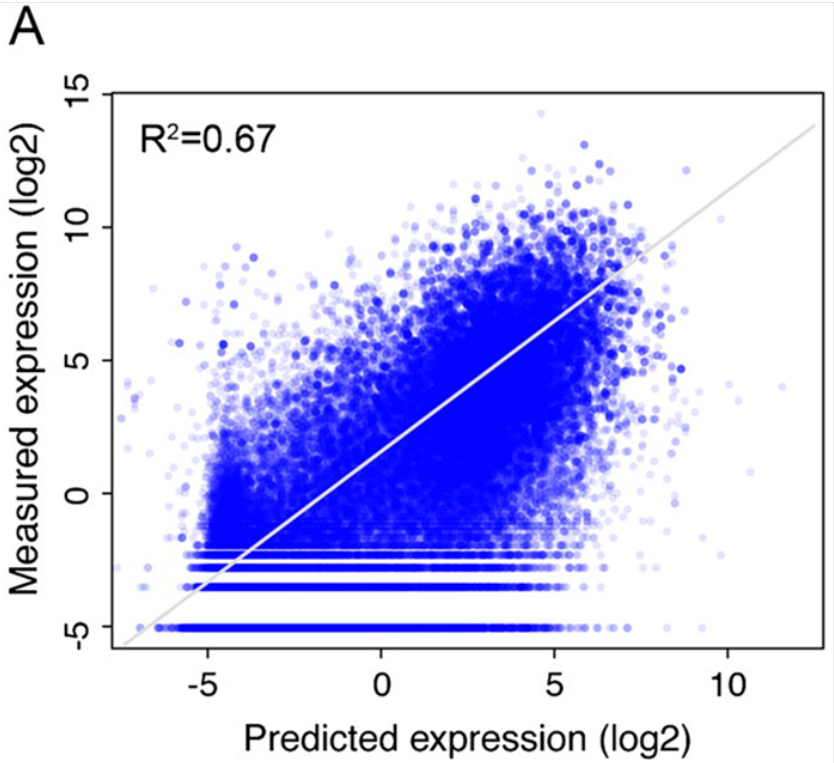
Nonlinear relationship between TF binding and TSS expression



- SVR: single predictor
- RF: multiple predictors
- CAGE ployA+ whole cell from K562 (Default)

Results

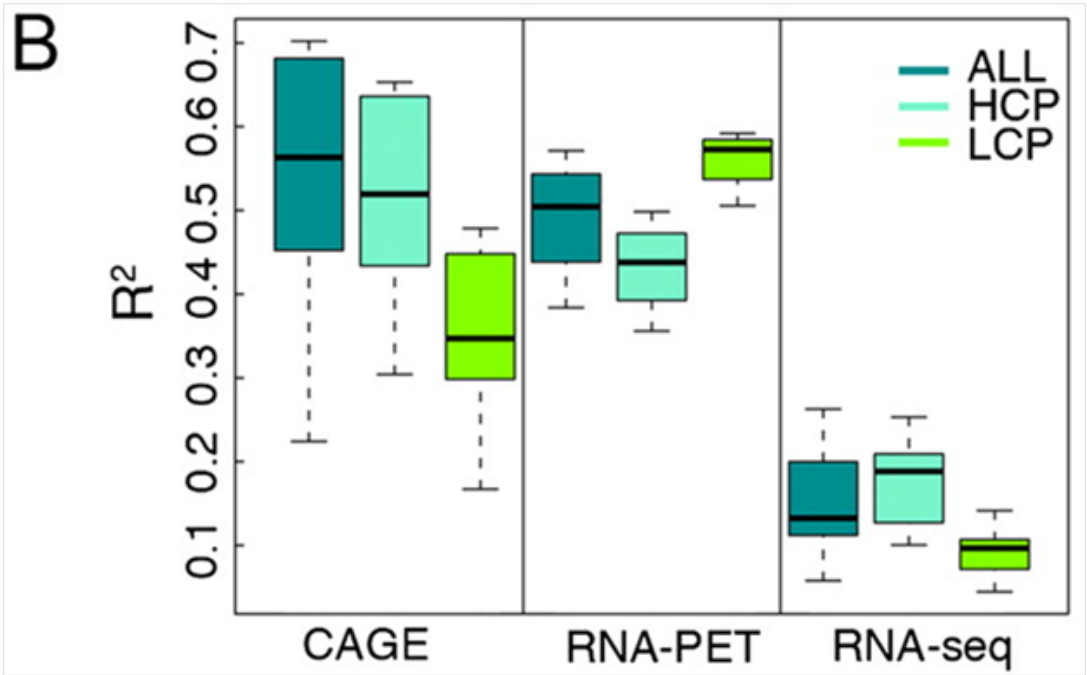
Accuracy of the TF model for predicting TSS expression levels



- Figure 1.A shows the consistency between predicted and actual expression levels of TSSs measured by CAGE of whole cell Poly A+ RNA in K562 cells. "Prediction accuracy" model.

Comparison of three different technologies

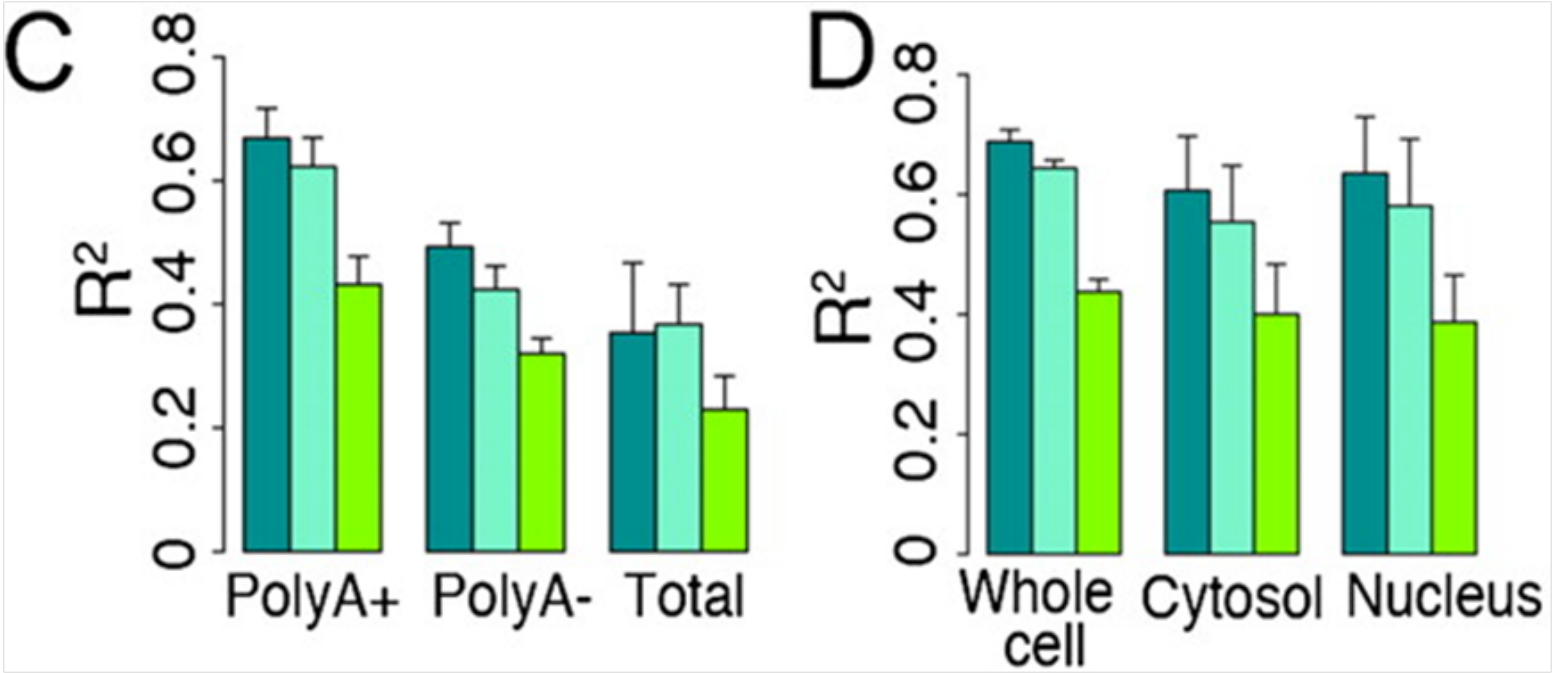
Comparison of three different technologies



- They used the binding signals of 40 TFs to predict each of the 57 K562 expression profiles
- The highest predictive accuracy was achieved from CAGE

Comparison of different RNA extraction protocols, different cellular components

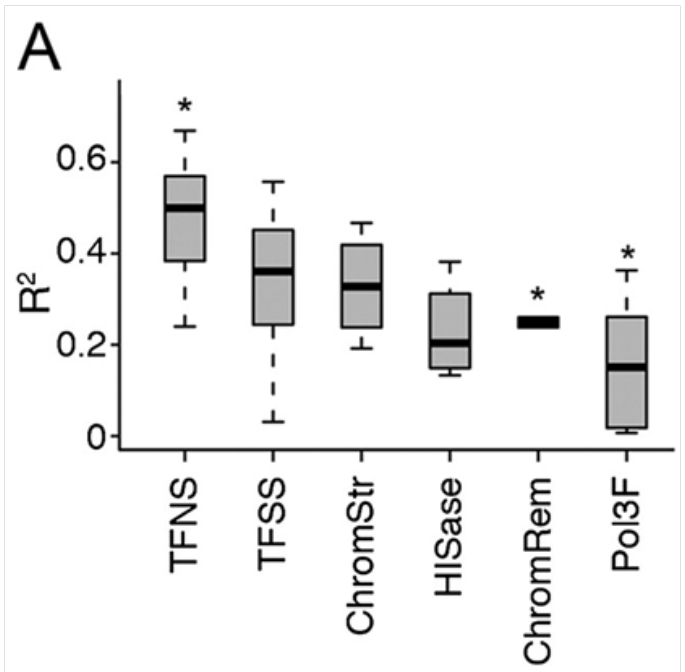
Comparison of different RNA extraction protocols, different cellular components



- protocols: Poly A+ >Poly A- >Total RNA
- cellular components: no obvious difference

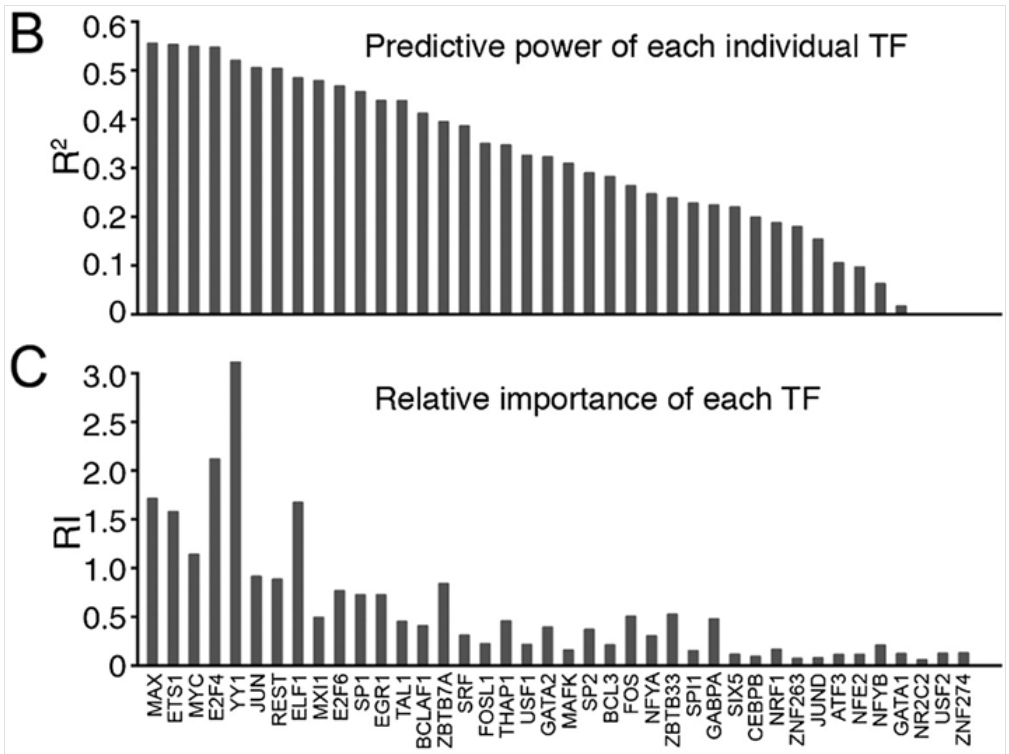
The capabilities of different TFs to predict TSS expression level

The capabilities of different TFs to predict TSS expression level



- TFNS TFs are the most predictive. (Binding of these TFs is essential for transcriptional initiation of most promoters)
- Pol3F are the least predictive. (RNA Pol III is involved in a small fraction of promoters)

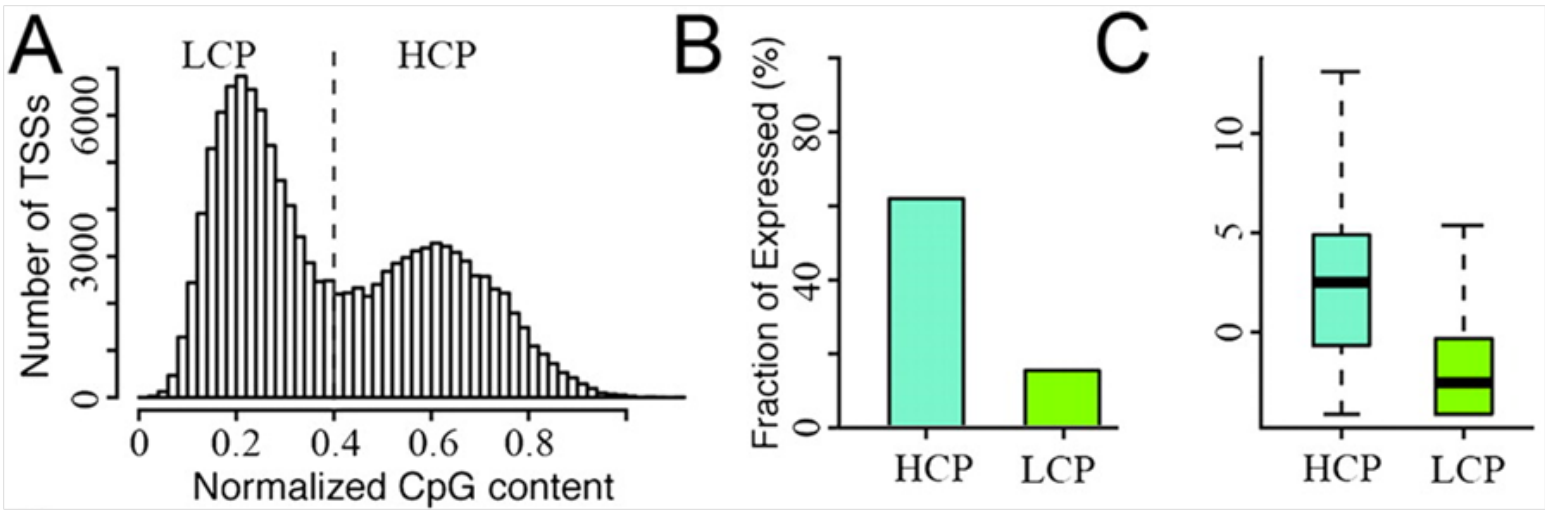
The capabilities of individual TFSS TF to predict TSS expression level



- R^2 for each TF is fairly high.
- RI (increase of MSE when testing data permuted)

The relationship between promoter CpG content and expression level

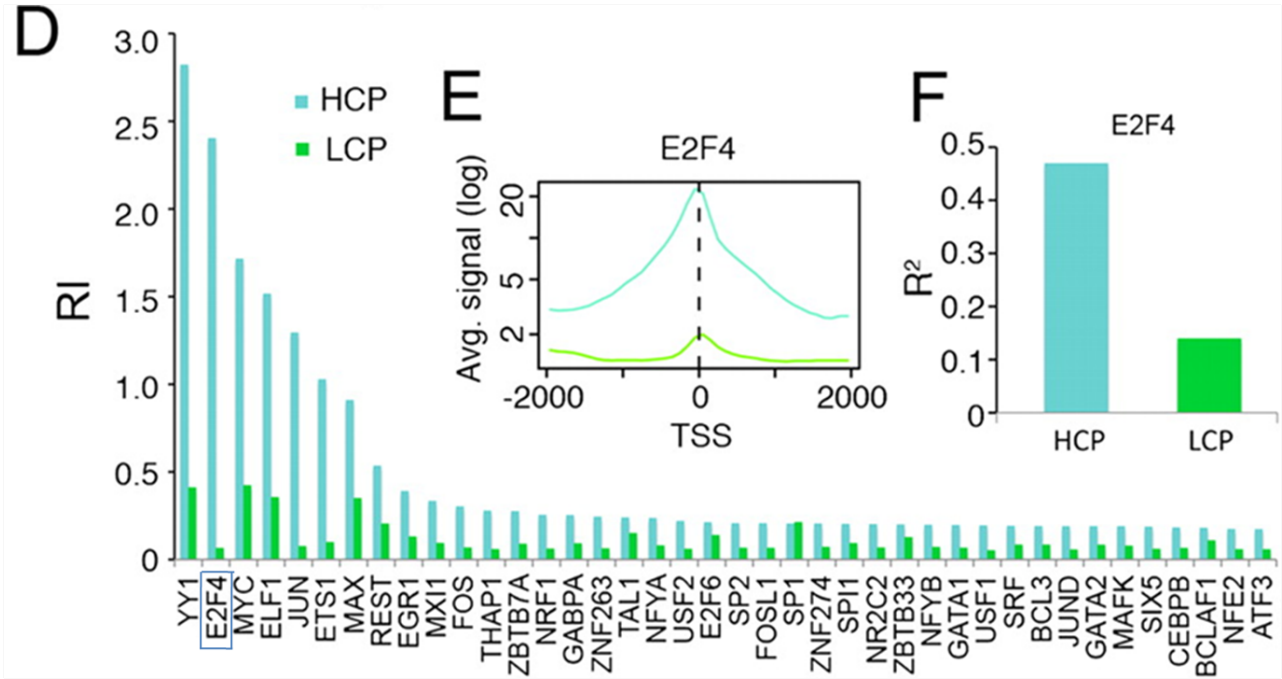
The relationship between promoter CpG content and expression level



- A: Bimodal distribution: LCP and HCP
- B: HCP are more highly expressed than LCP.
- C: Among expressed TSSs, expression level HCP >LCP

Relative Importance for each TF, HCP vs. LCP

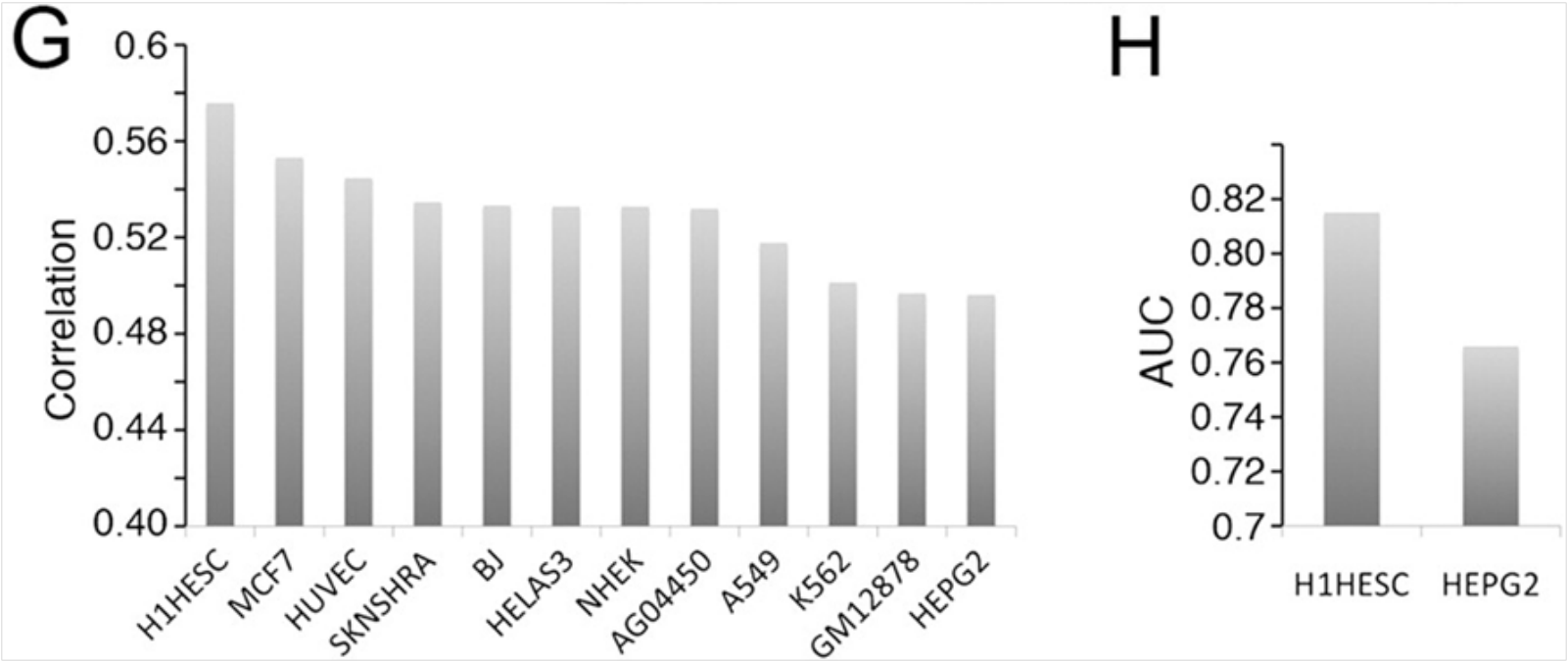
Relative Importance for each TF, HCP vs. LCP



- D: RI: HCP >LCP; E2F4: high RI for HCP but low for LCP
- E: Binding signal of E2F4: HCP >LCP
- F: R^2 of E2F4 (single predictor): HCP >LCP
- The regulation of E2F4 on gene expression might be affected by status of CpG sites.

Correlation between CpG and expression level in different cell lines

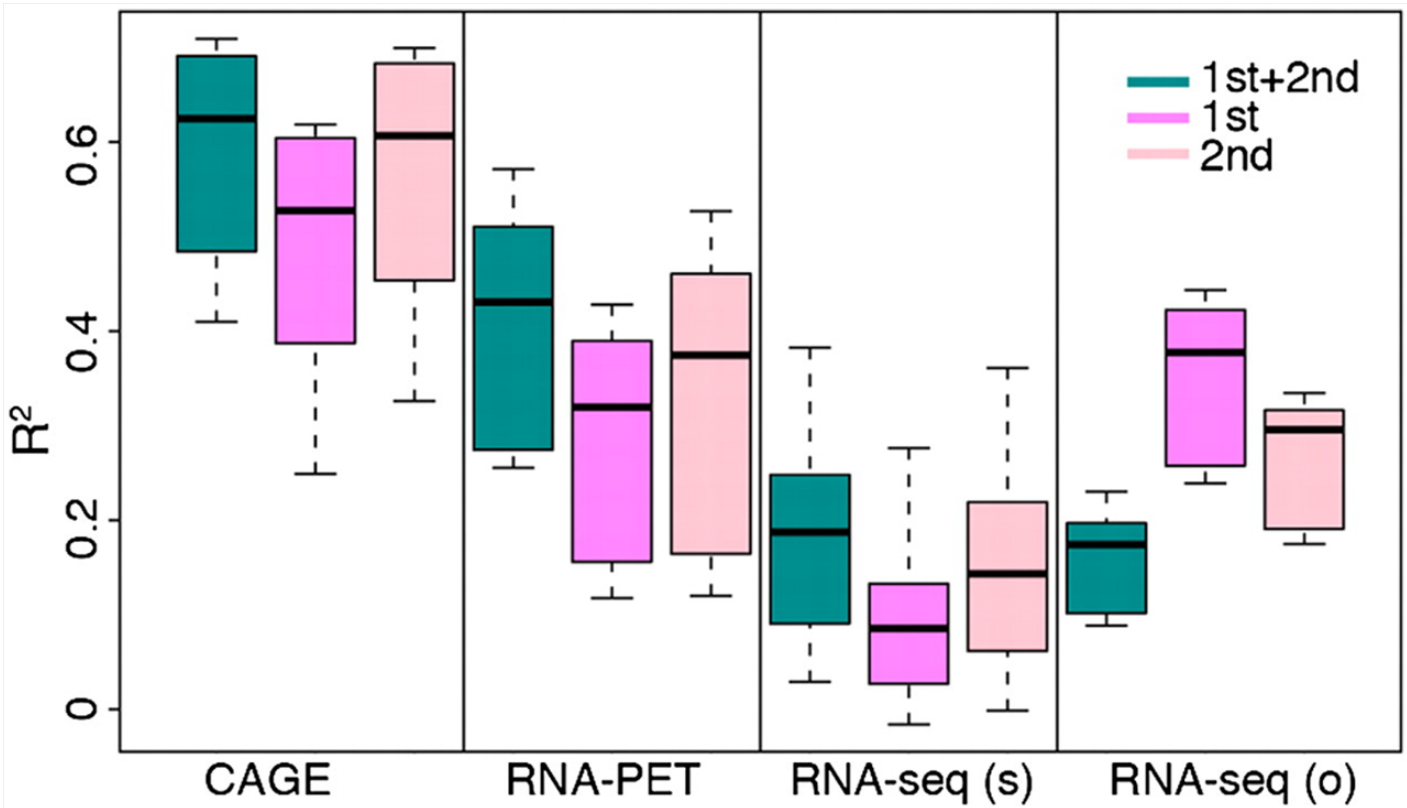
Correlation between CpG and expression level in different cell lines



- G: Best correlation: H1HESC (H1 human embryonic stem cell)
- High CpG to UpG rate for promoters repressed in germline cells or in early developmental stage. CpG ->methylation ->expression repressed ->mutation ->lower CpG content?
- H: CpG as classifier for expressed or nonexpressed promoters. High accuracy: AUC=0.82 in H1HESC

Regulation of alternative TSS by TFs

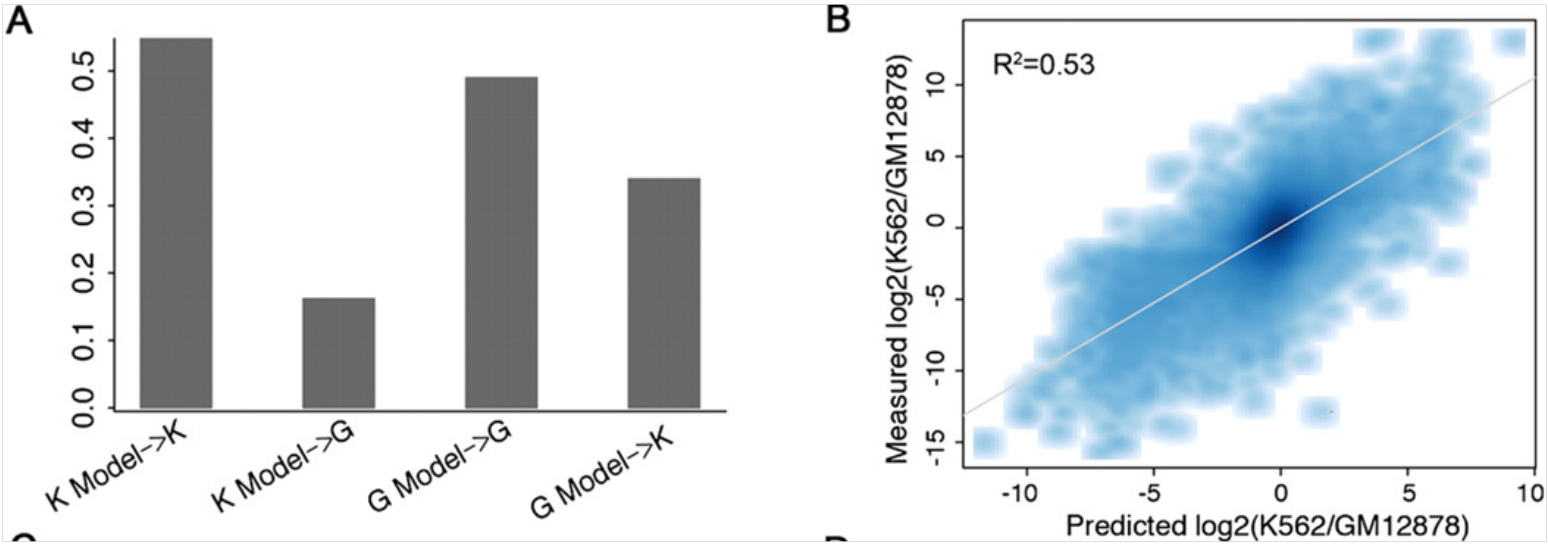
Regulation of alternative TSS by TFs



- Around 35% of GENCODE genes posses >1 TSS; compare 1st and 2nd TSS
- Higher predictive accuracy for 2nd TSS: CAGE, RNA-PET and RNA-seq(o)
- Expression levels of 1st and 2nd are similar ->2nd TSS rely more on TF regulation. Also different RIs.

Cell line specificity of the TF model 1

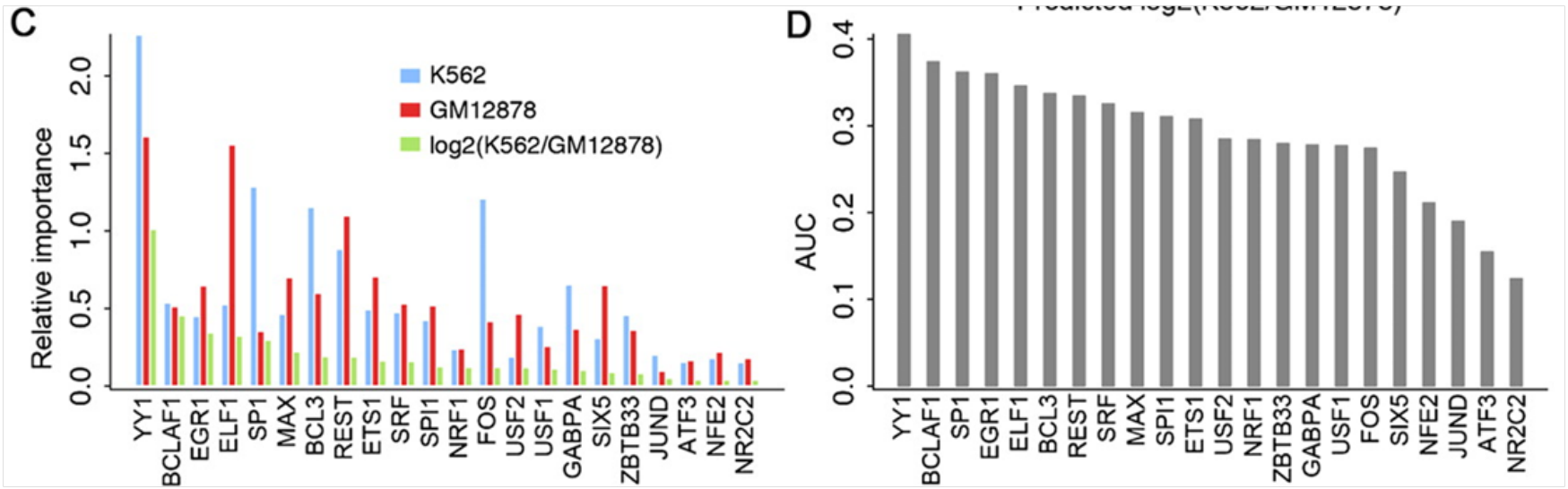
Cell line specificity of the TF model 1



- Cell line specific promoters (fourfold expression difference); 22 TFs
- A: K562 (erythroleukemia) and GM12878 (normal lymphoblastoid) independent models
- B: Using binding differences ($log_2(K562/GM12878)$) to predict expression difference of cell lines.

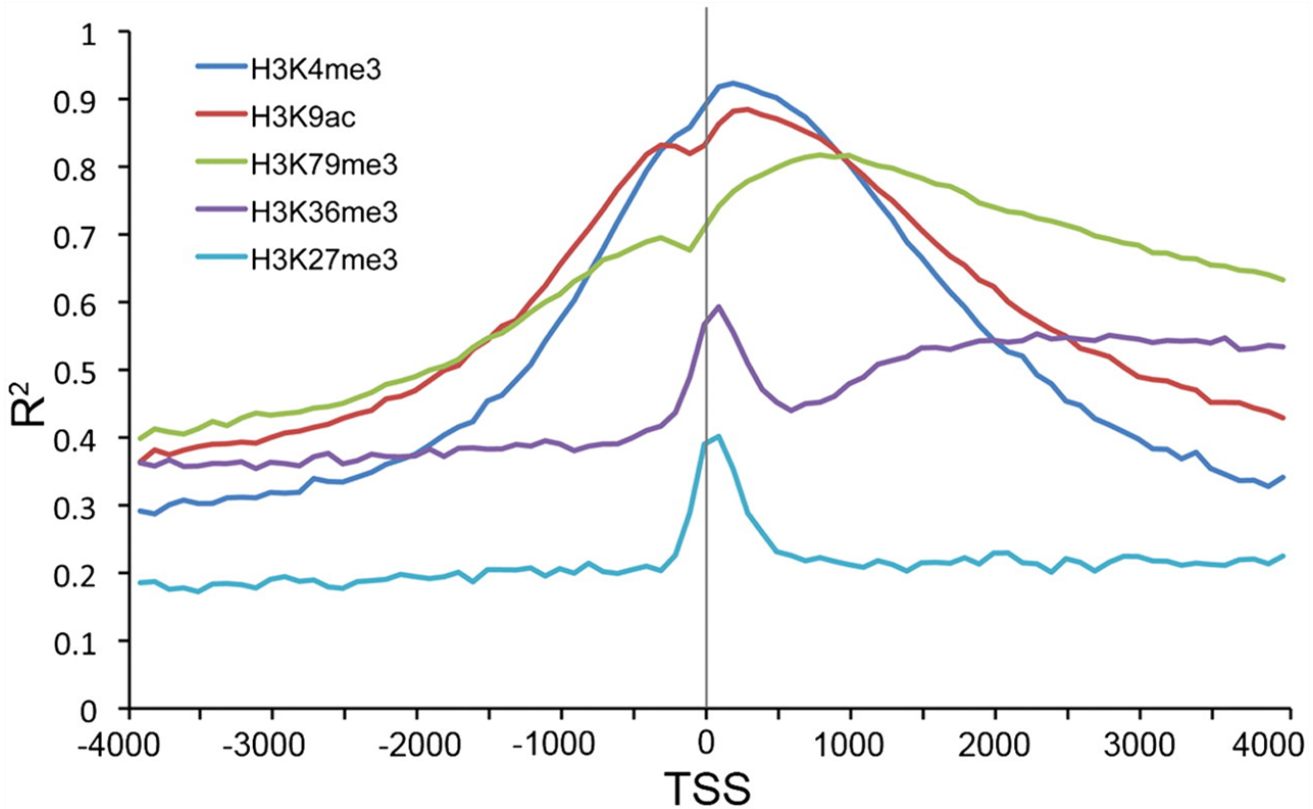
Cell line specificity of the TF model 2

Cell line specificity of the TF model 2



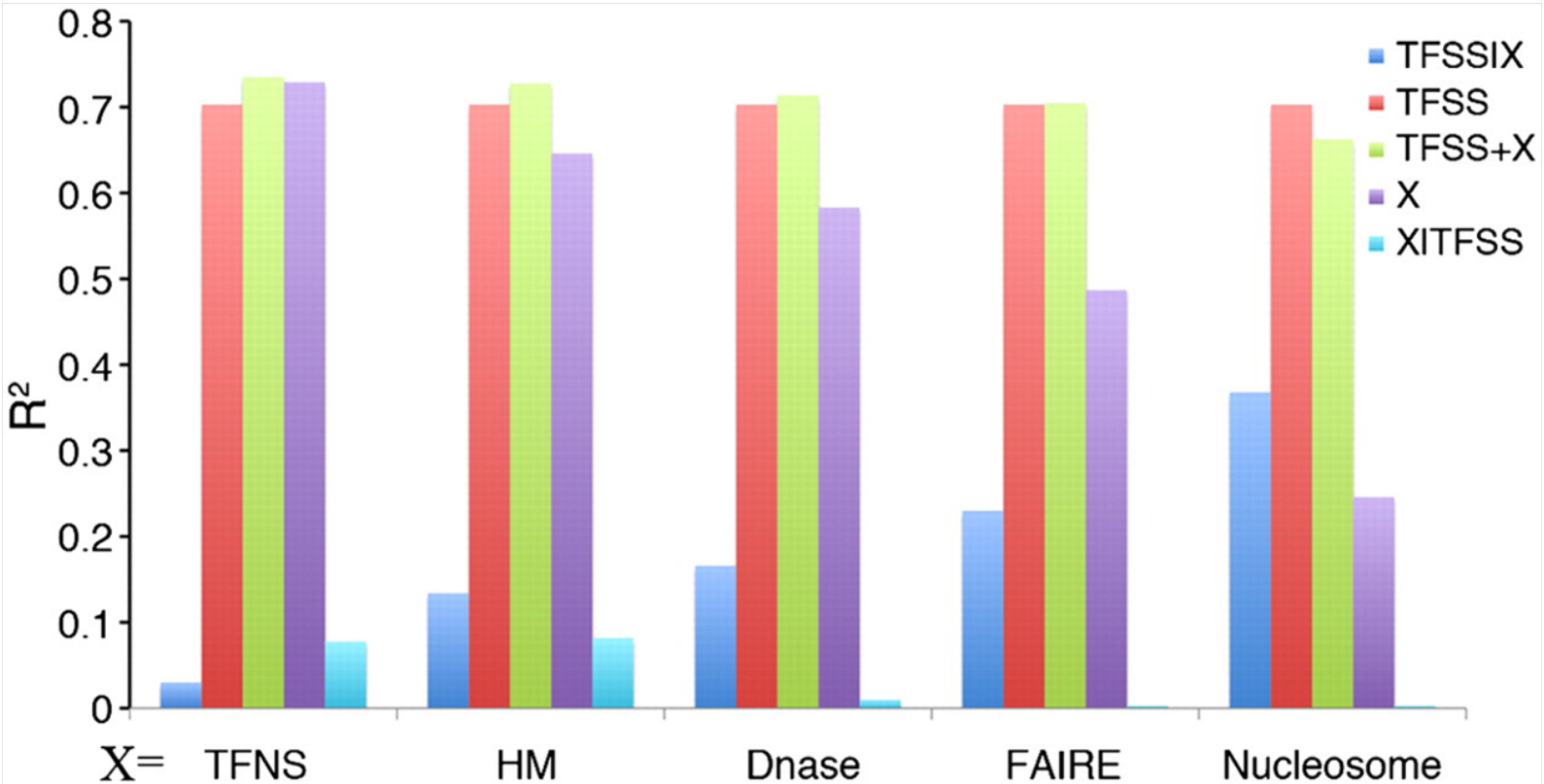
- C: Regression model, RIs of individual TF. Find that TFs with high RIs for differential expression model are TFs with high RIs in both K and G models.
- D: Classification model, using individual TF to classify TSSs in different cell lines. All of the TFs can classify with YY1 the best (AUC=0.86)

The capabilities of TFSS TFs to predict histone modification signals



- Histone modification can be predicted accurately by TF binding signals at TSS region (HsK4me3 $R^2 = 0.85$).
- TSS (-4kb, 4kb) region are divided into 80 bins, each 100bp. Predicting on each bin.

Interplay between TFSS TFs binding and other chromatin features for predicating promoter expression



- Other chromatin structure features: HM (histone modifications), Dnase (DNase hypersensitivity), FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements), and nucleosome occupancy.
- X|TFSS: X $\sim$ y-f(TFSS)
- TFSS+TFNS reaches the best (equals to full model with $R^2 = 0.74$).

Conclusion

Conclusions

- Notable difference in prediction accuracy of expression levels captured by different technologies and protocols
- The expression levels of TSSs with high CpG content are more predictable than those with low CpG content.
- For genes with alternative TSSs, the expression levels of downstream TSSs are more predictable than those of the upstream ones.
- Between two cell lines, the differential expression of TSS can be predicted by the different TF-binding signals.
- TF binding signals and other chromatin features regulate transcription in a coordinated manner.

Regulatory mechanism of TF binding, histone modification, and other chromatin features on gene expression.