Model Selection in Machine Learning

+

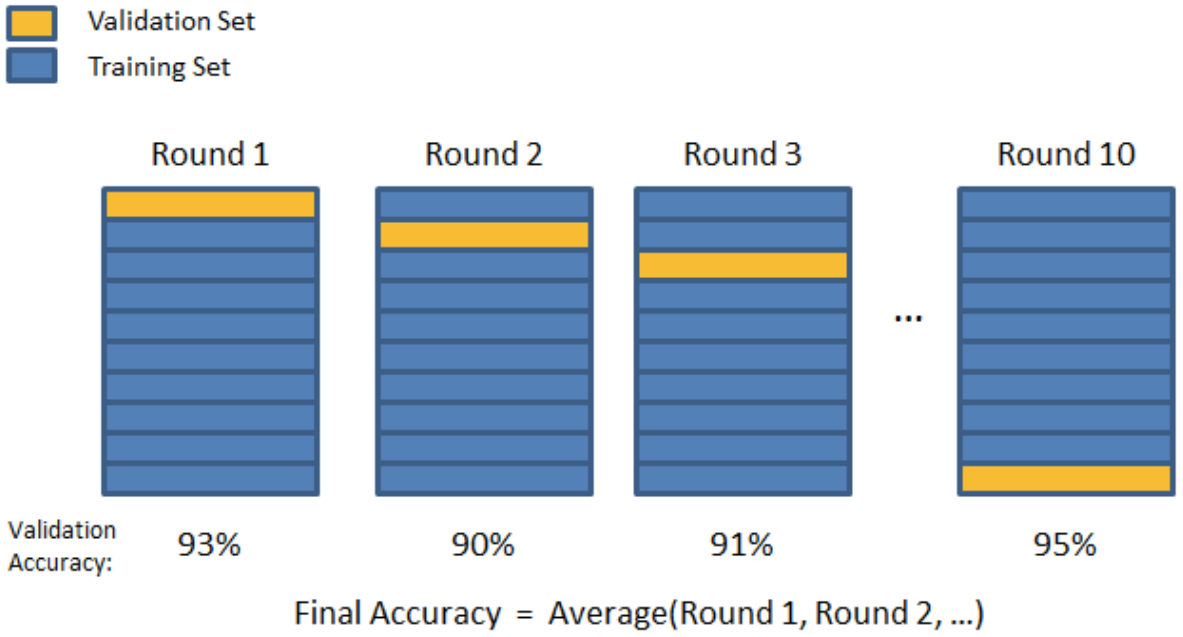
Predicting Gene Expression from ChIP-Seq signals

https://twitter.com/nature

Review and warm-up questions

Training vs. Cross-Validation

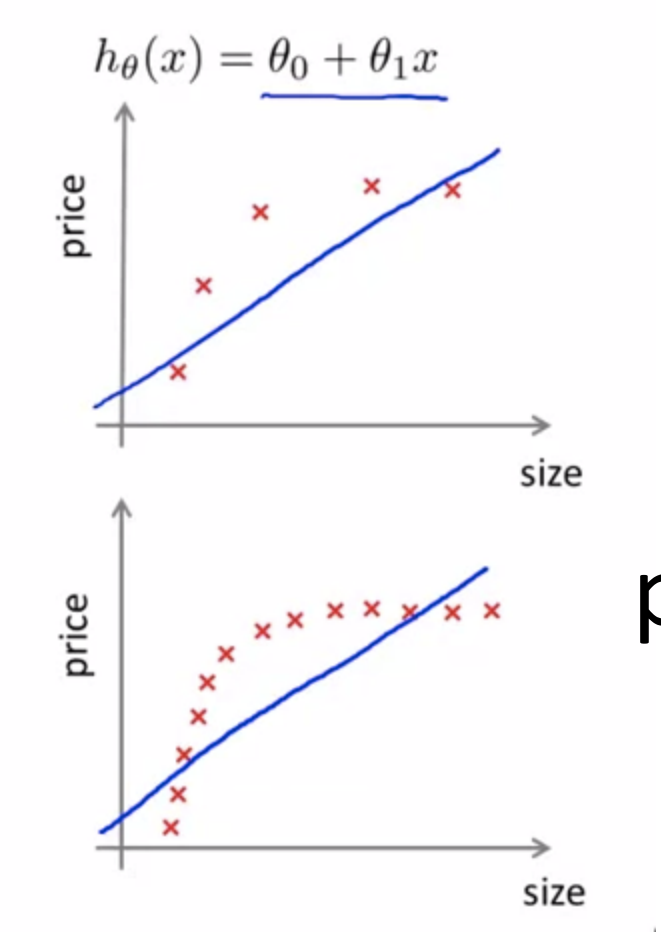
- Fit model to example data points
- Evaluate model on *separate* set of data points



https://chrisjmccormick.wordpress.com/2013/07/31/k-fold-cross-validation-with-matlab-code/

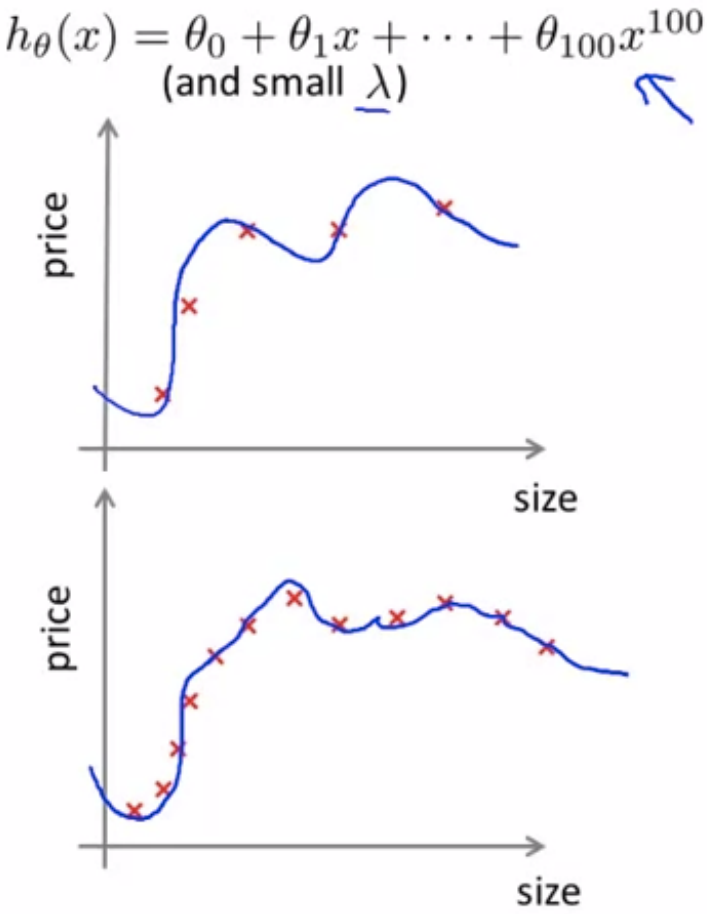
Bias vs. Variance

Model too simple



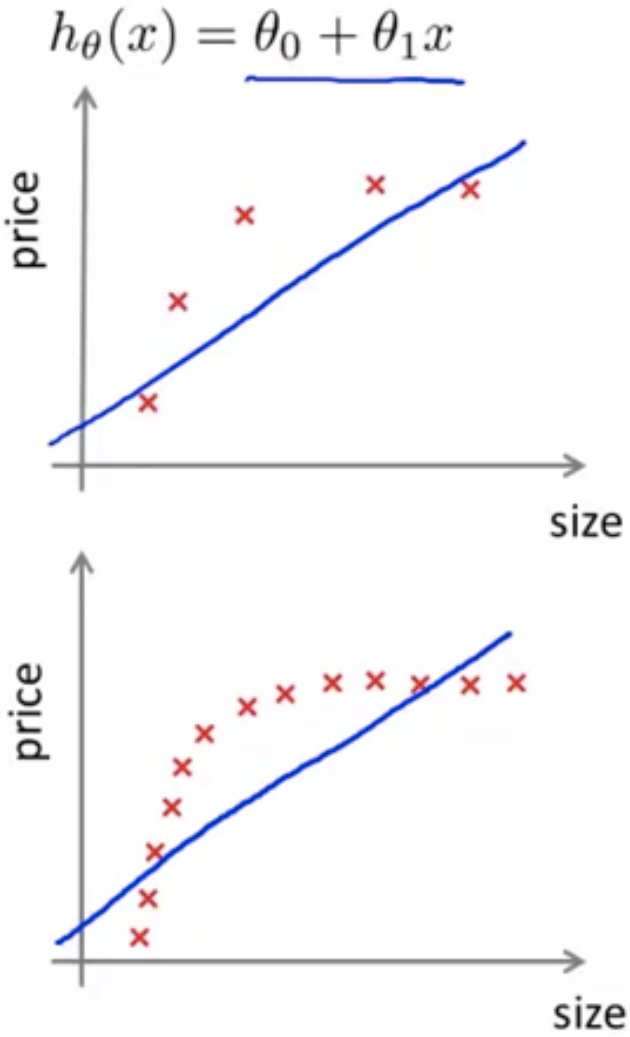
What's the problem?

Model too complex



Adapted from Andrew Ng – https://www.youtube.com/watch?v=DYCv5e0Isow, http://see.stanford.edu/materials/aimlcs229/ML-advice.pdf

What's the problem? – Bias



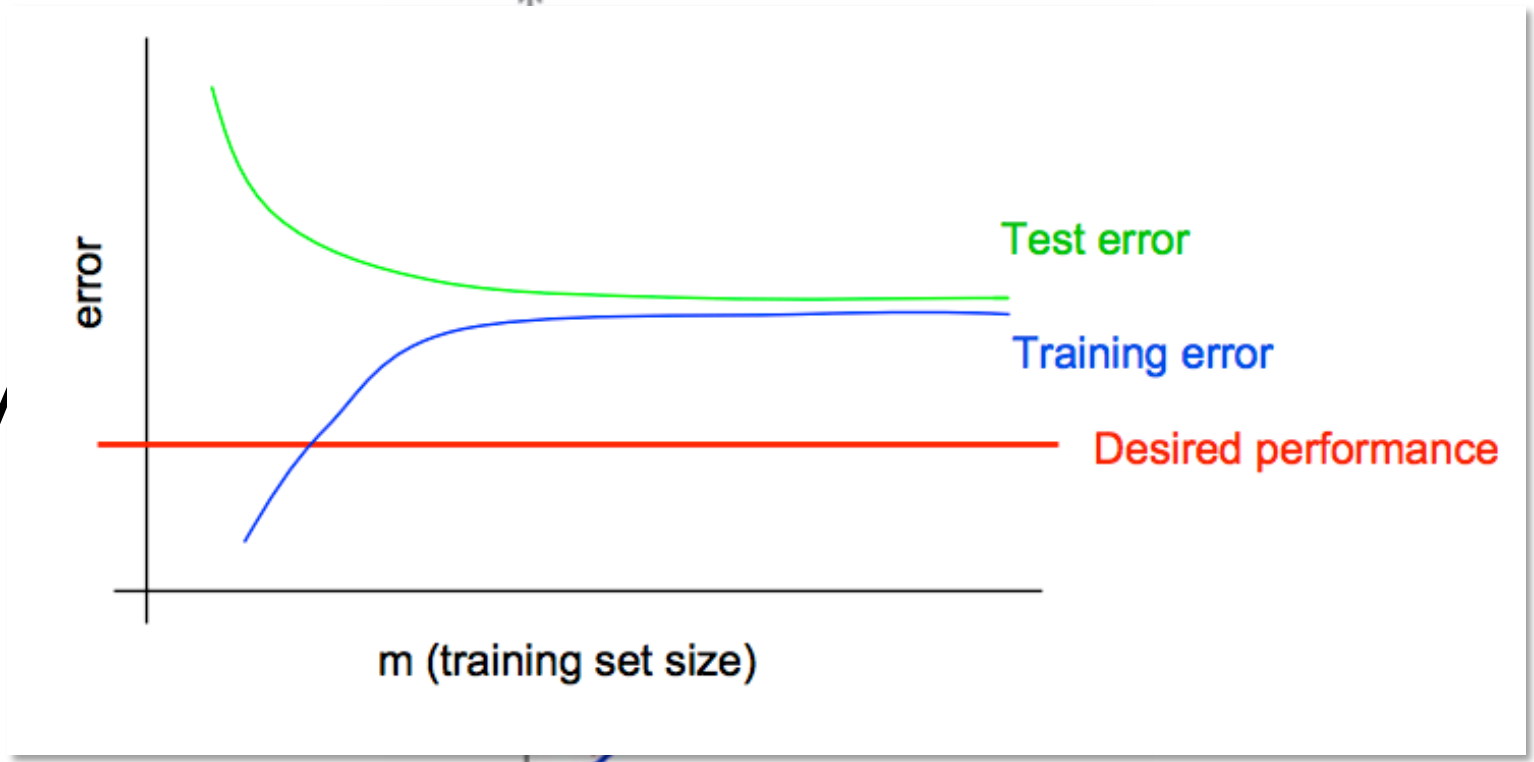
Adapted from Andrew Ng – https://www.youtube.com/watch?v=DYCv5e0Isow, http://see.stanford.edu/materials/aimlcs229/ML-advice.pdf

Learning curve – Bias

$$h_\theta(x) = \theta_0 + \theta_1 x$$

B

M

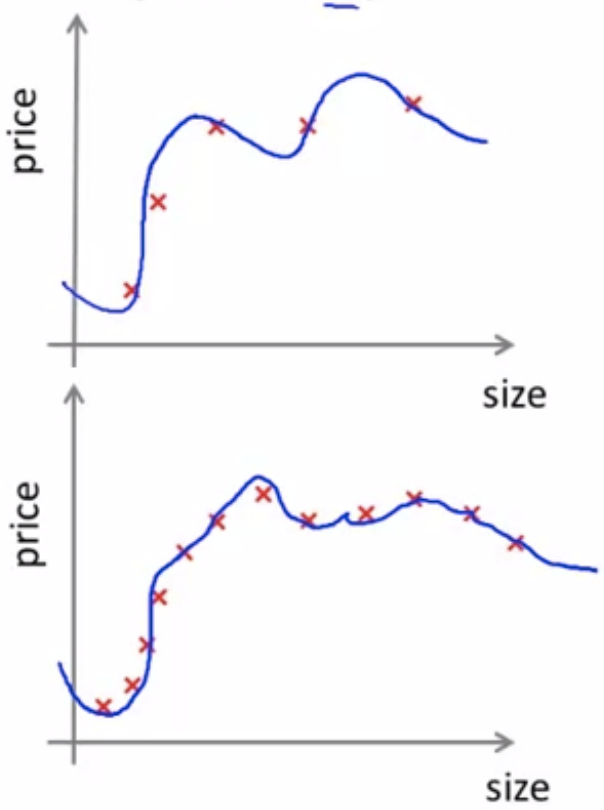


Adapted from Andrew Ng – https://www.youtube.com/watch?v=DYCv5e0Isow, http://see.stanford.edu/materials/aimlcs229/ML-advice.pdf

What's the problem? – Variance

$$h_\theta(x) = \theta_0 + \theta_1 x + \cdots + \theta_{100} x^{100}$$

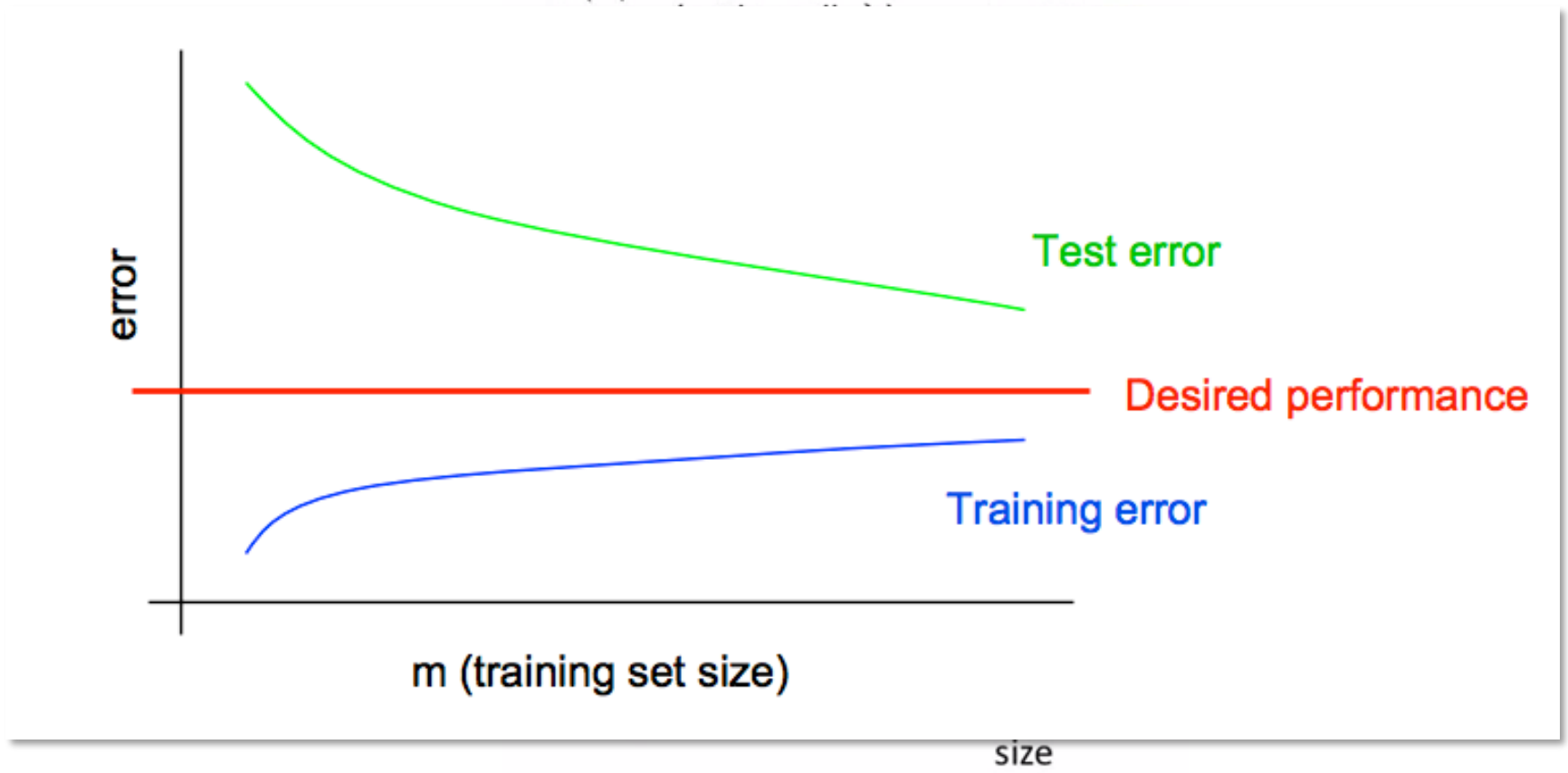
(and small λ)



Adapted from Andrew Ng – https://www.youtube.com/watch?v=DYCv5e0Isow, http://see.stanford.edu/materials/aimlcs229/ML-advice.pdf

Learning curve – Variance

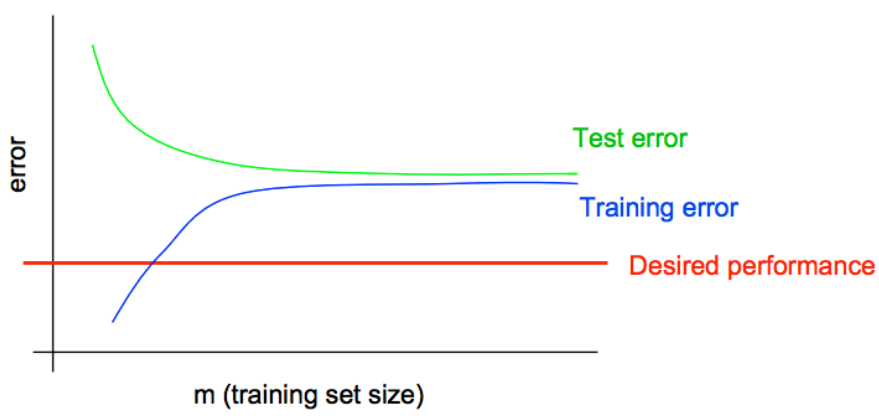
$$h_\theta(x) = \theta_0 + \theta_1 x + \cdots + \theta_{100} x^{100}$$



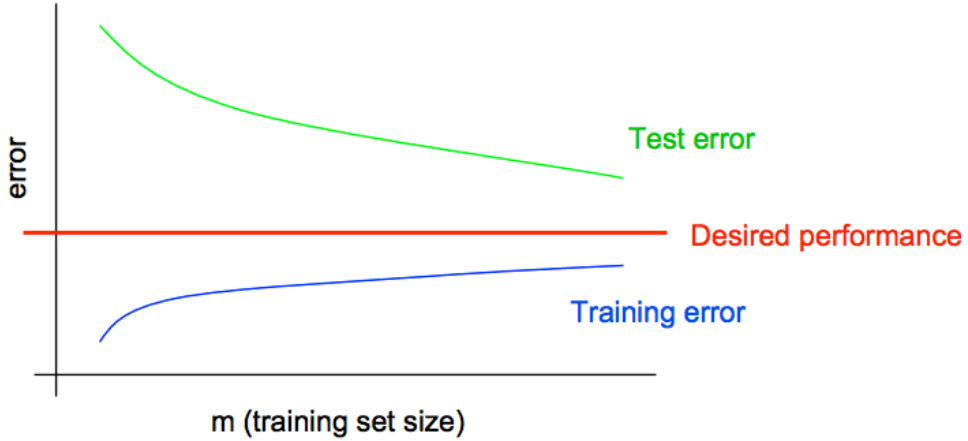
Adapted from Andrew Ng – https://www.youtube.com/watch?v=DYCv5e0Isow, http://see.stanford.edu/materials/aimlcs229/ML-advice.pdf

What is the next step?

Bias



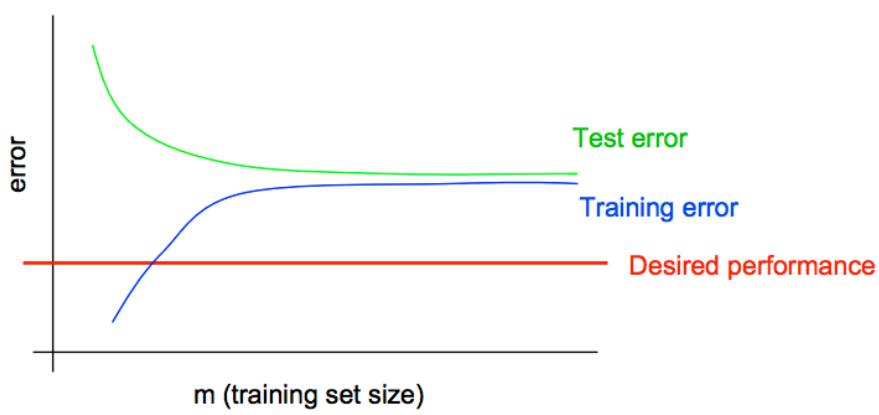
Variance



Adapted from Andrew Ng – https://www.youtube.com/watch?v=DYCv5e0Isow, http://see.stanford.edu/materials/aimlcs229/ML-advice.pdf

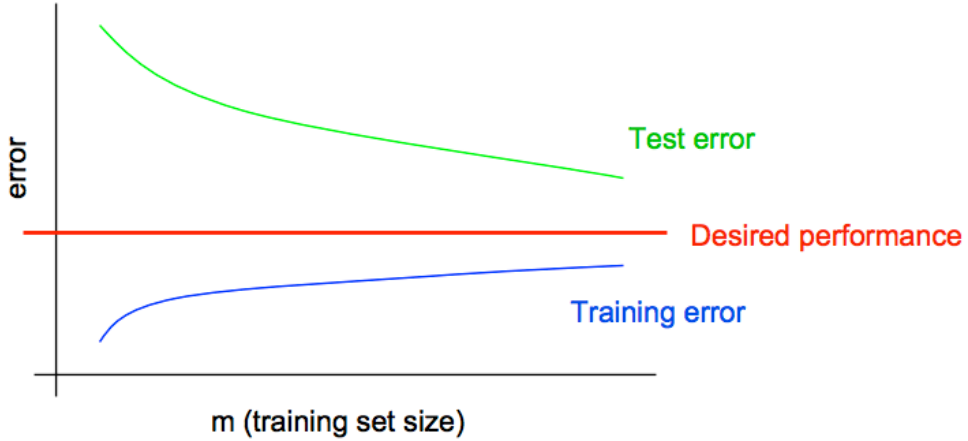
What is the next step?

Bias



More training **features**
Train more complicated model

Variance

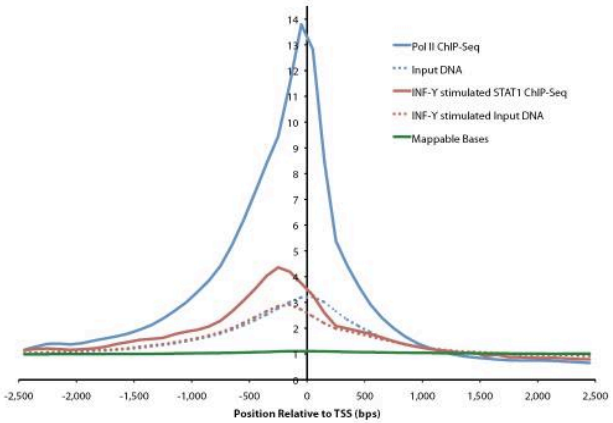


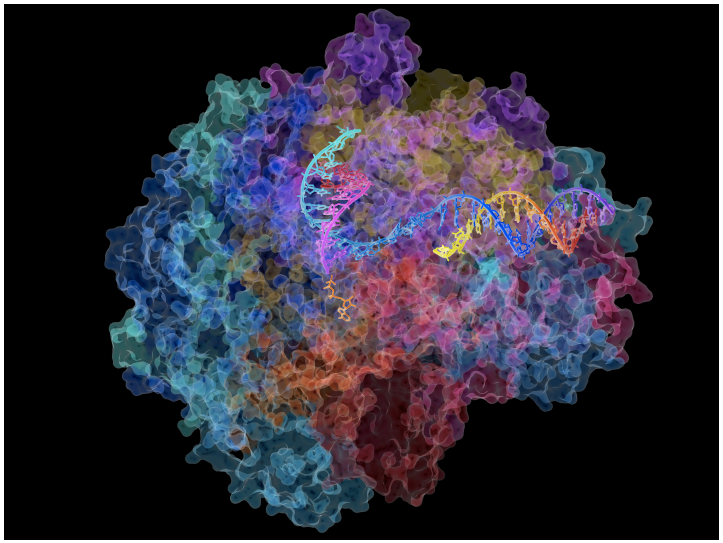
More training **examples**
Try fewer features → Dimension Reduction
Simplify model

Adapted from Andrew Ng – https://www.youtube.com/watch?v=DYCv5e0Isow, http://see.stanford.edu/materials/aimlcs229/ML-advice.pdf

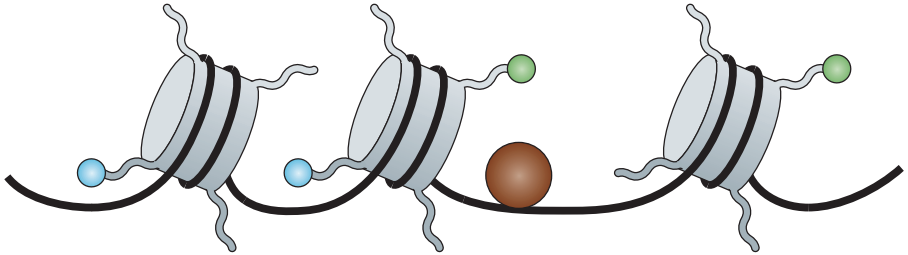
Practical Application:

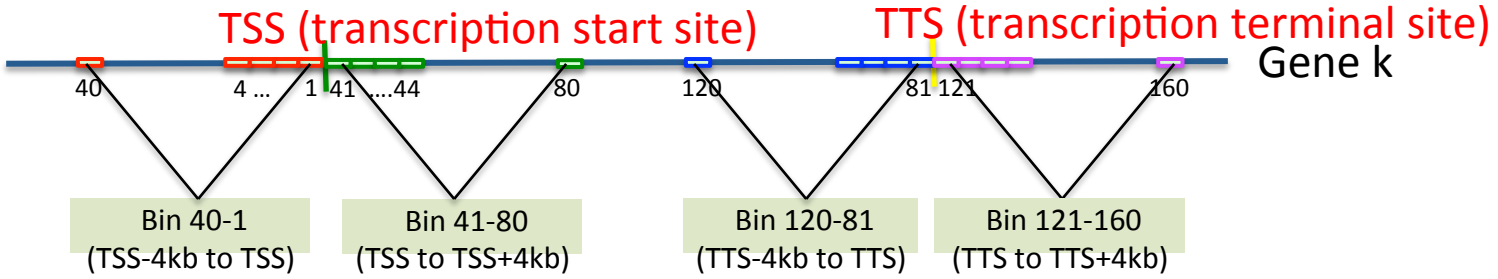
Predicting gene expression from ChIP-Seq signals





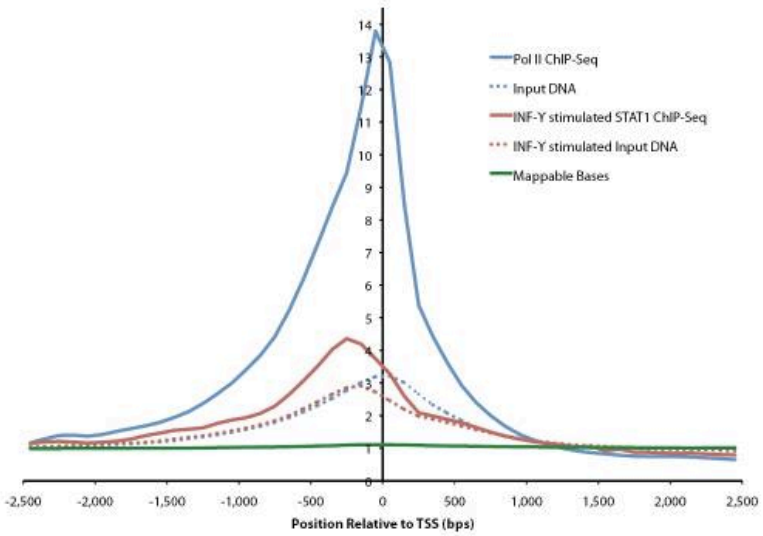
Where should we start?





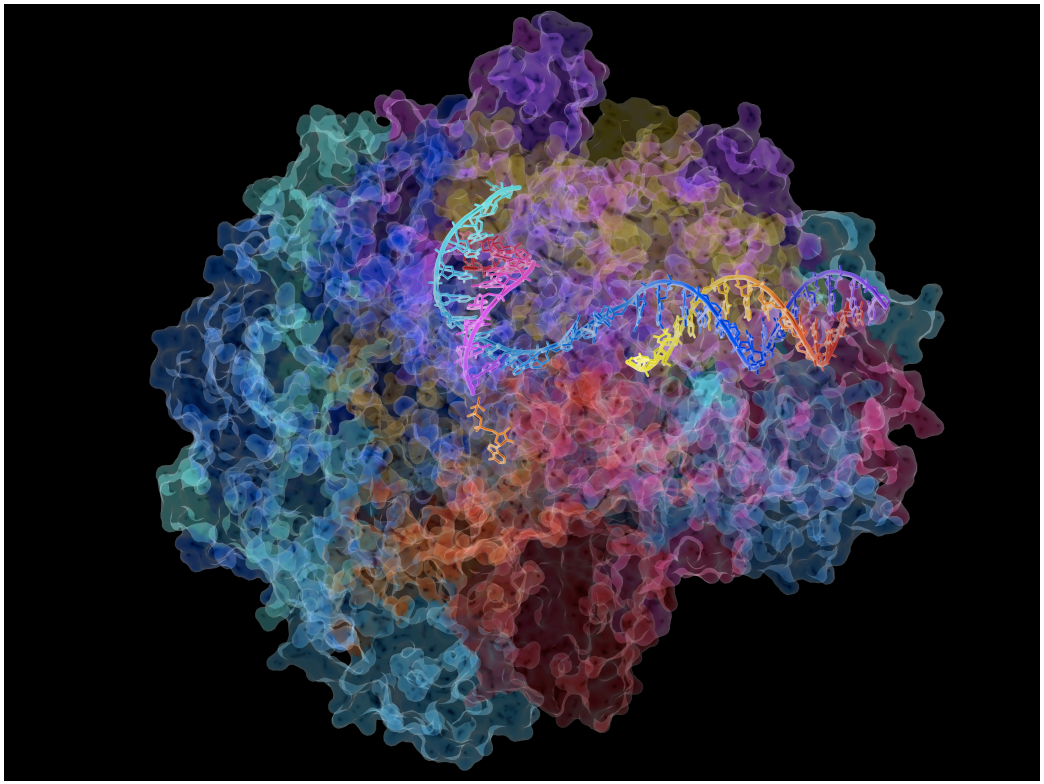
Park *et al Nature Reviews Genetics* 2009, Rozowsky *et al Nature Biotech* 2009, http://images.nigms.nih.gov/imageRepository/2484/RNA_pol_II_medium.jpg, Cheng et *al Genome Biology 2011*

RNA is transcribed by RNA polymerase



RNA pol II ChIP at
Transcription Start Sites

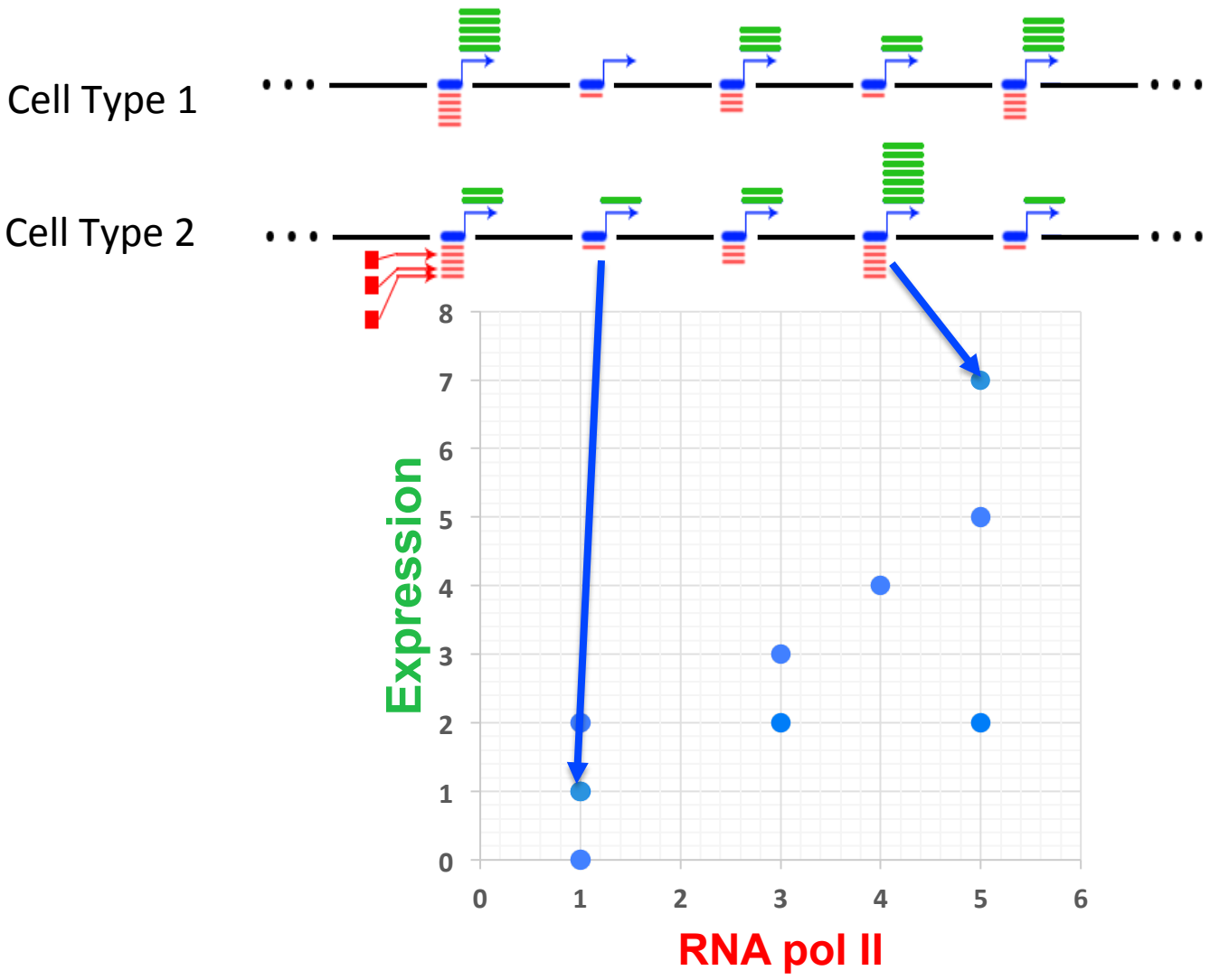
Rozowsky *et al. Nature Biotech* 2009



RNA polymerase II – Crystal structure
Roger Kornberg Nobel Prize

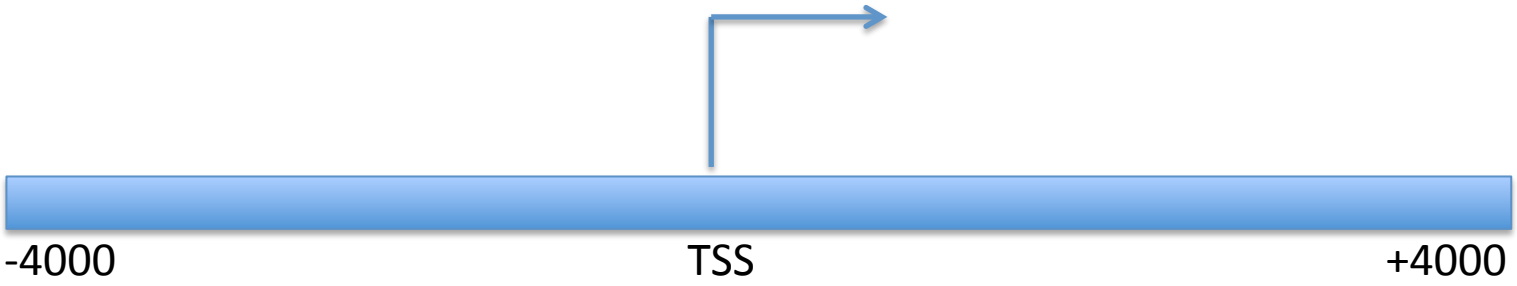
http://images.nigms.nih.gov/imageRepository/2484/RNA_pol_II_medium.jpg

Relating Genomic Inputs to Outputs



Mark Gerstein/Mengting Gu

Initial model



data from K562 cell line, ENCODE consortium

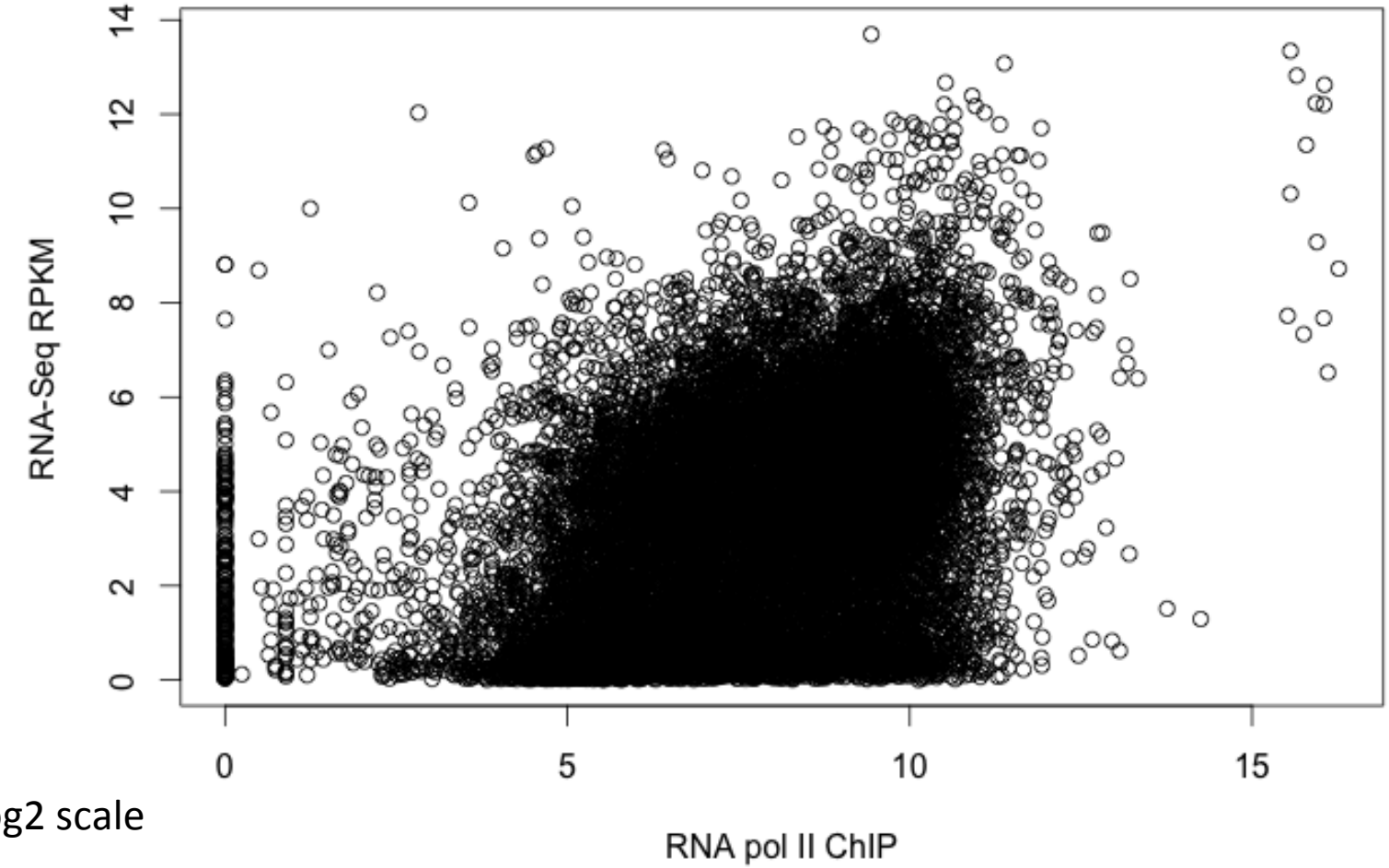
Sum pol II ChIP signal across 8000 bp centered around transcription start site.

$$\log_2(\text{RNA-Seq RPKM}) = a + b*\log_2(\text{RNA Pol II ChIP})$$

Why log scale??

Can we do better than this?

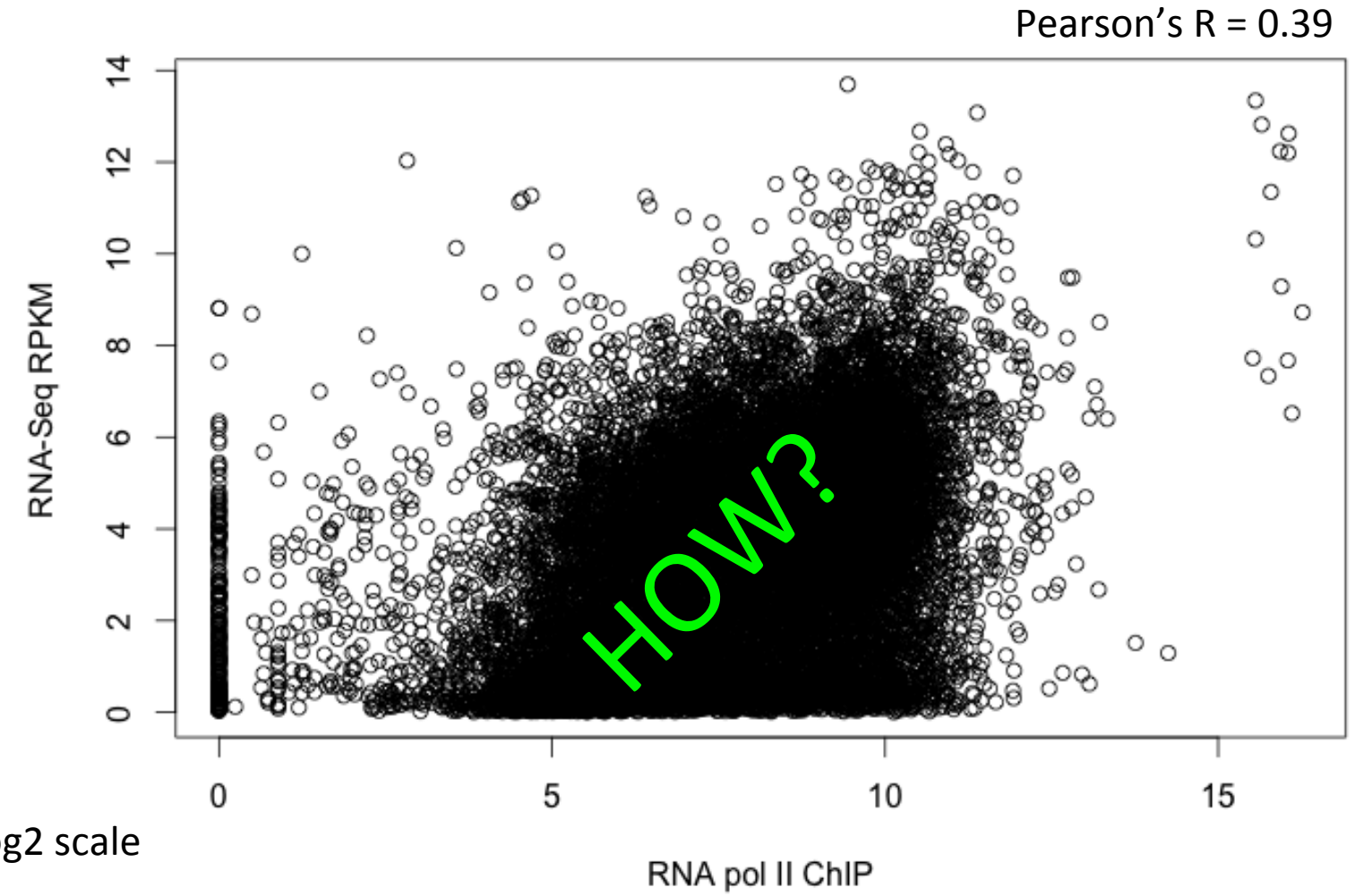
Pearson's R = 0.39



log2 scale

K562 Cell Line, ENCODE data

Can we do better than this?

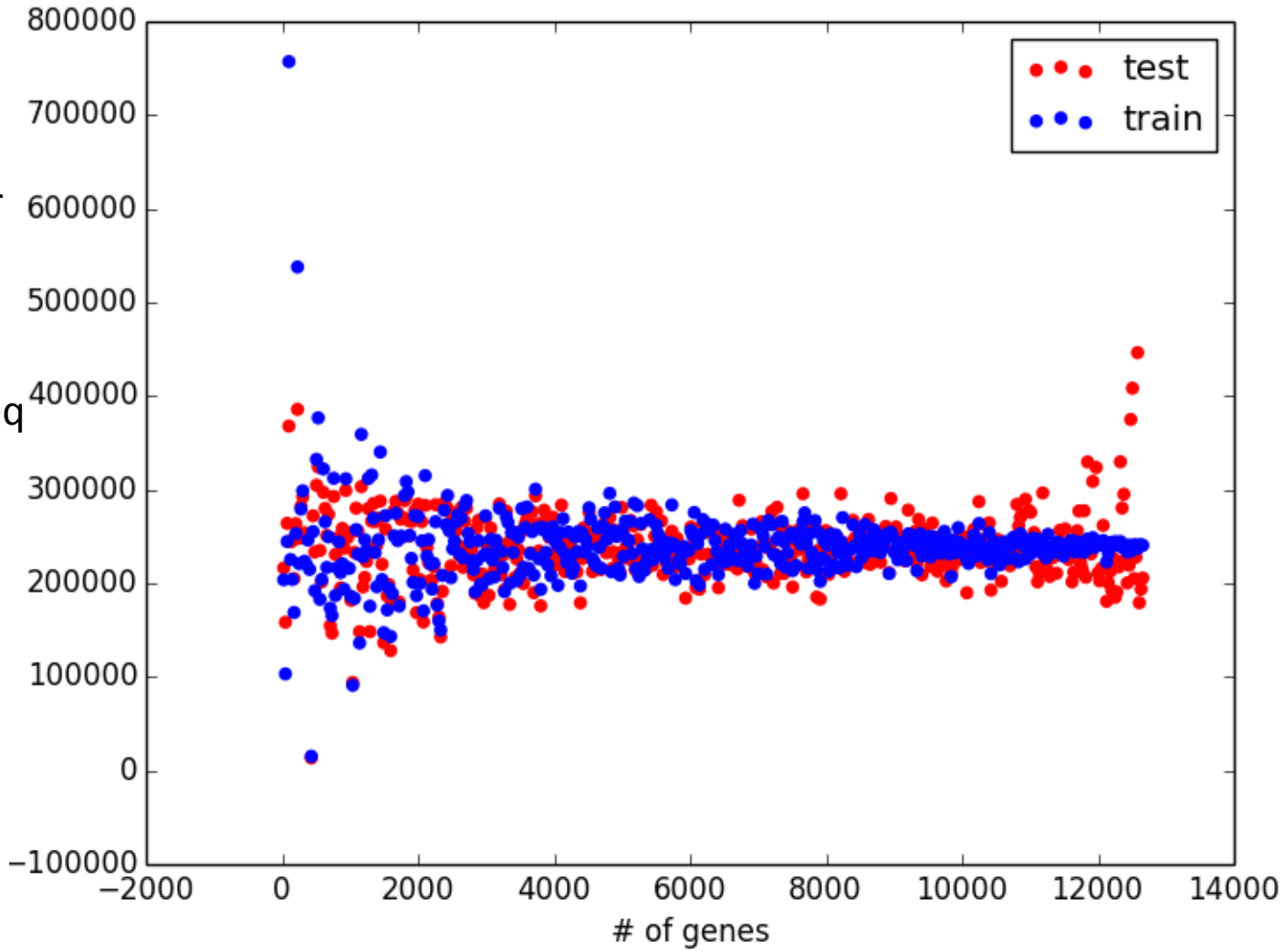


log2 scale

K562 Cell Line, ENCODE data

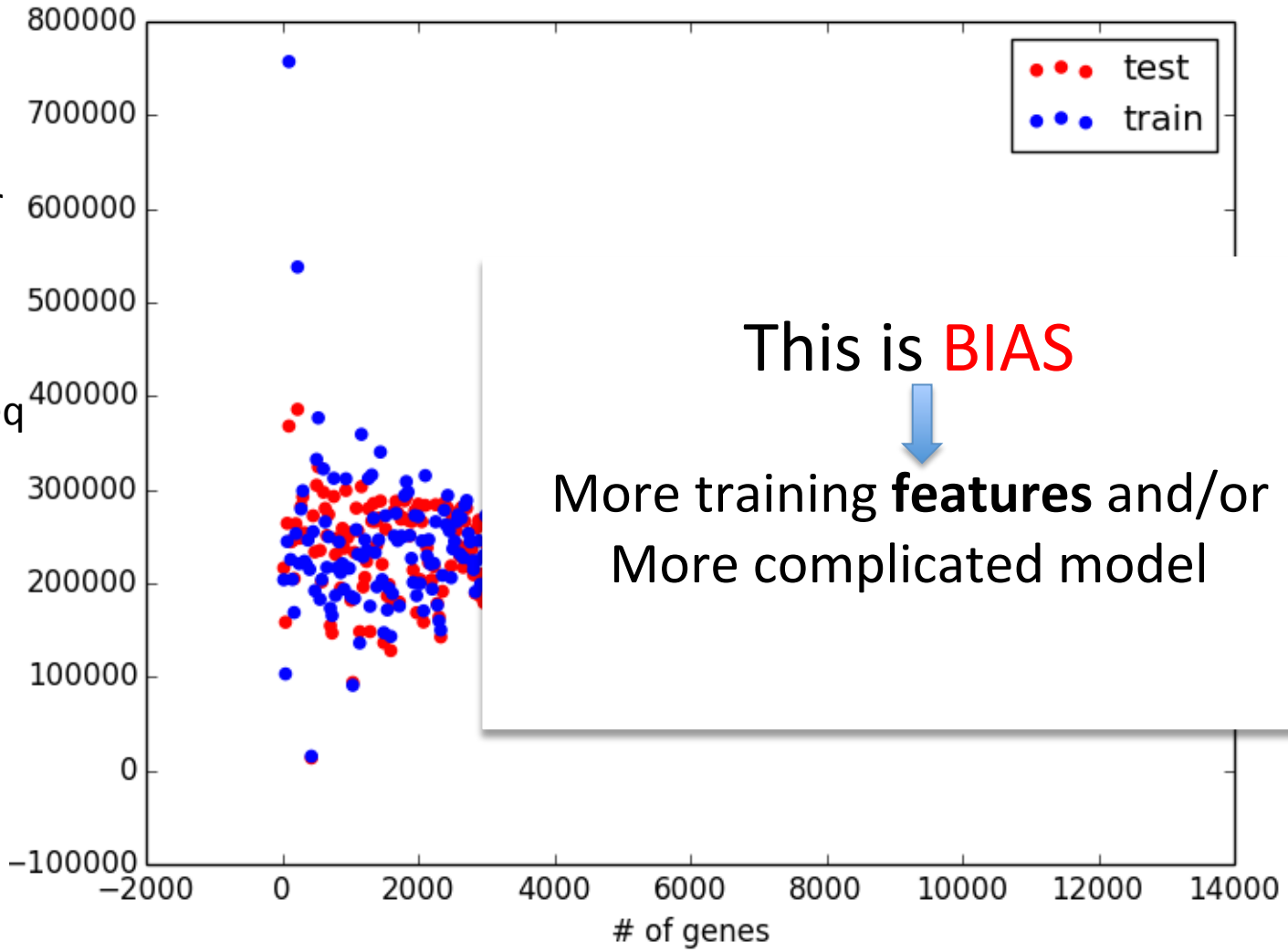
Learning curve: What's our problem?

Mean squared error
of
Linear Regression
for
pol II ChIP vs. RNA-Seq



Learning curve: What's our problem?

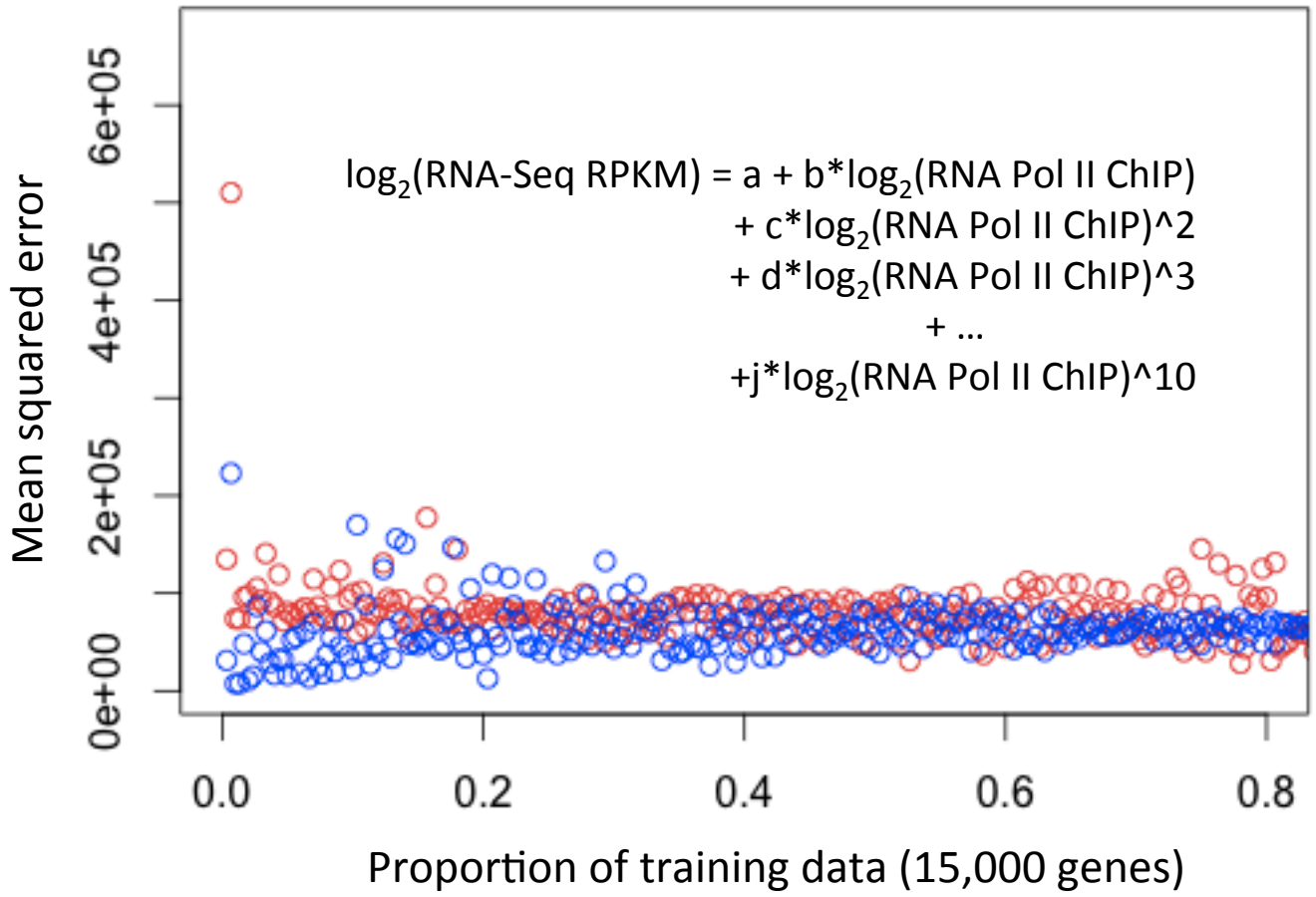
Mean squared error
of
Linear Regression
for
pol II ChIP vs. RNA-Seq



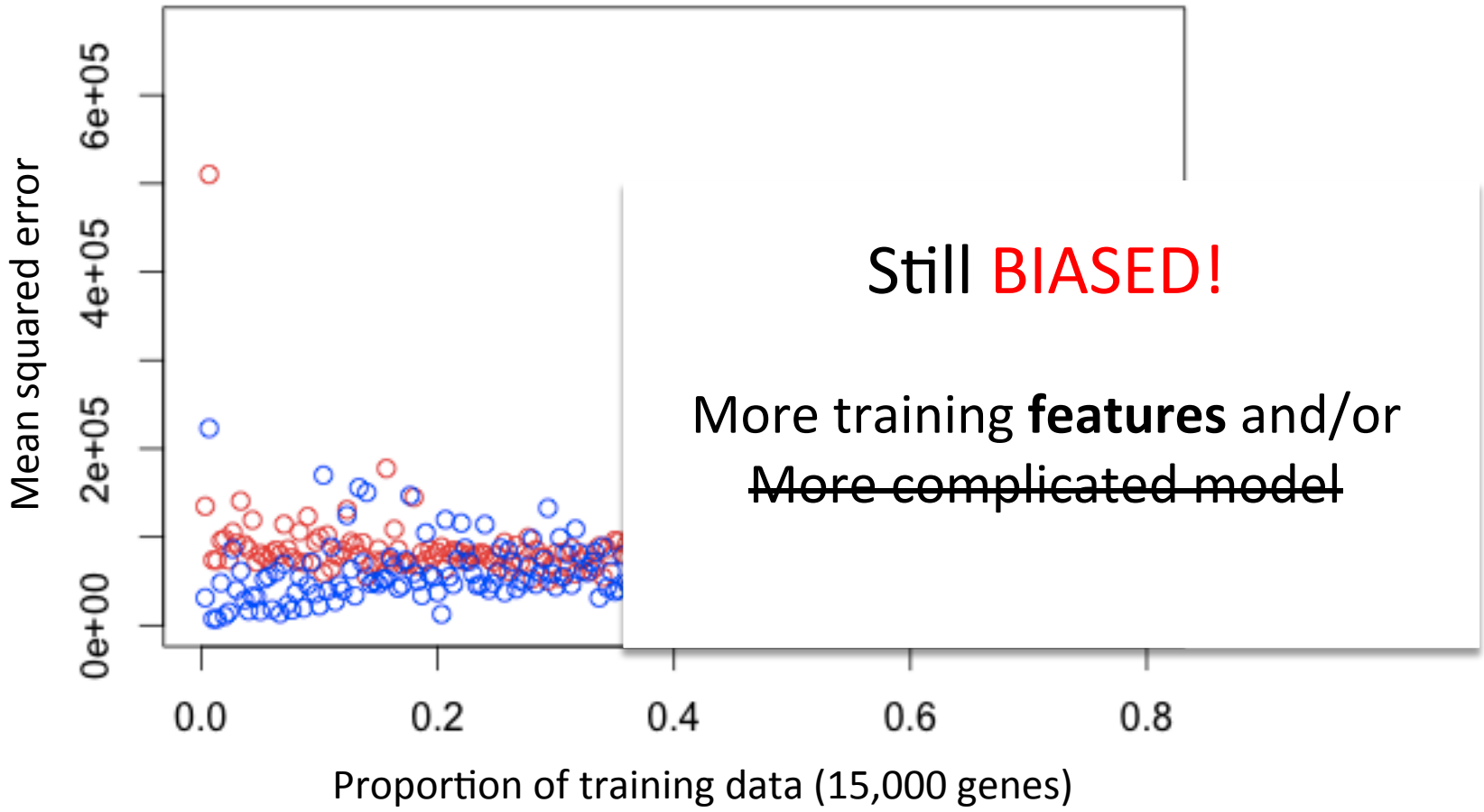
This is BIAS

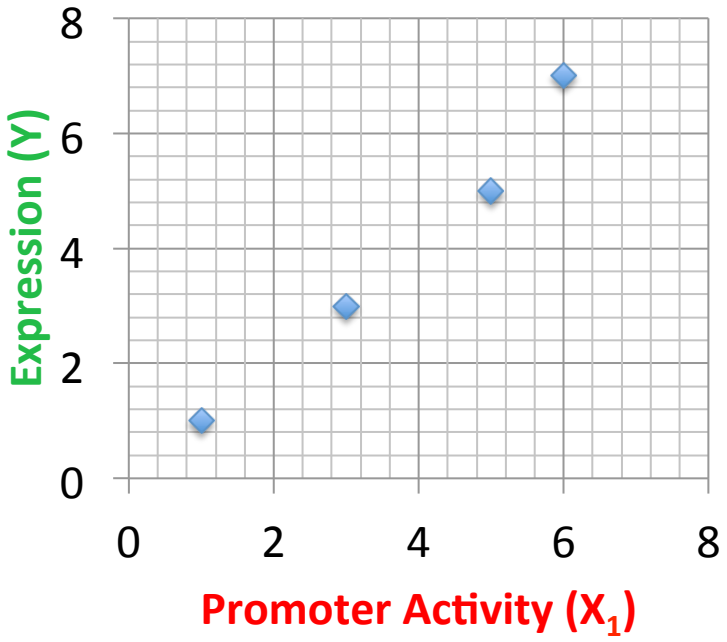
More training **features** and/or
More complicated model

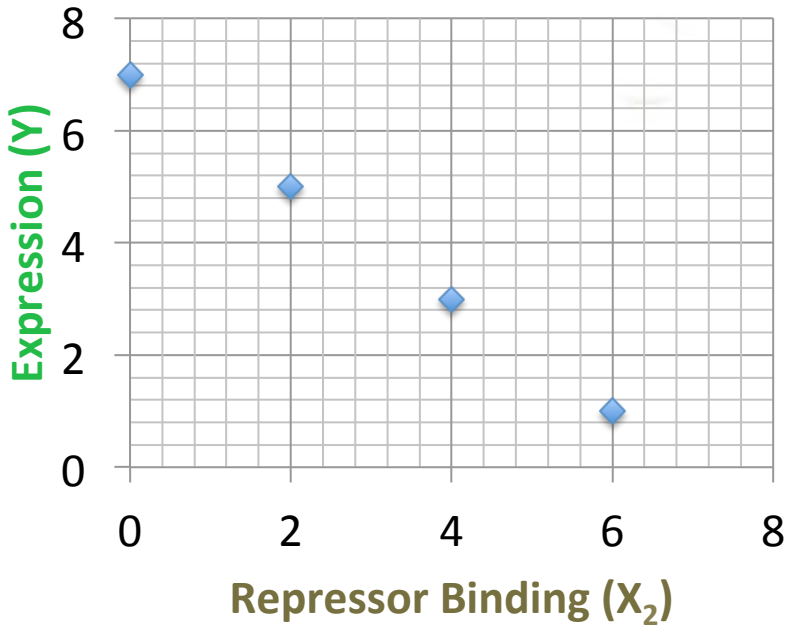
Polynomial regression: polII ChIP vs. RNA-Seq



Polynomial regression: polII ChIP vs. RNA-Seq



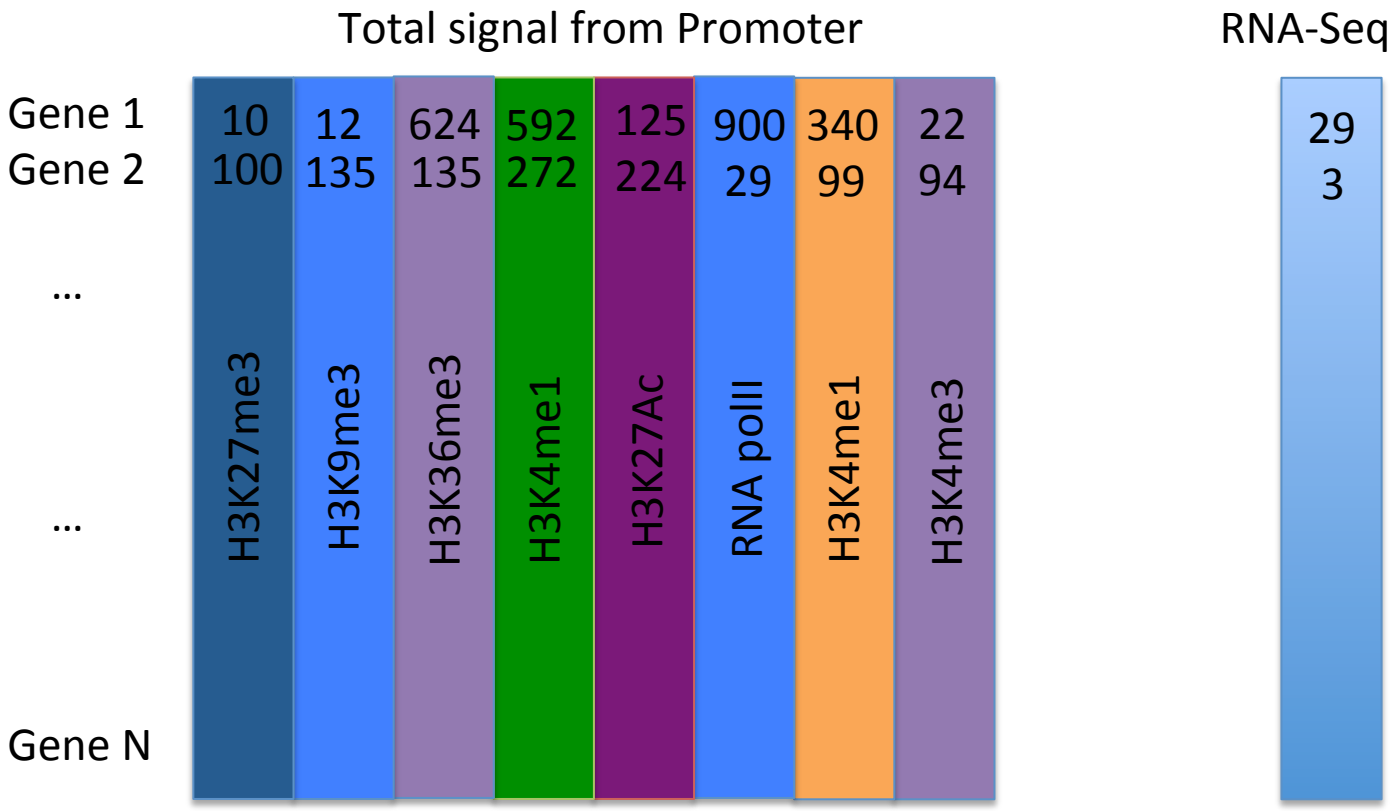




$$Y = aX_1 + bX_2 + c$$

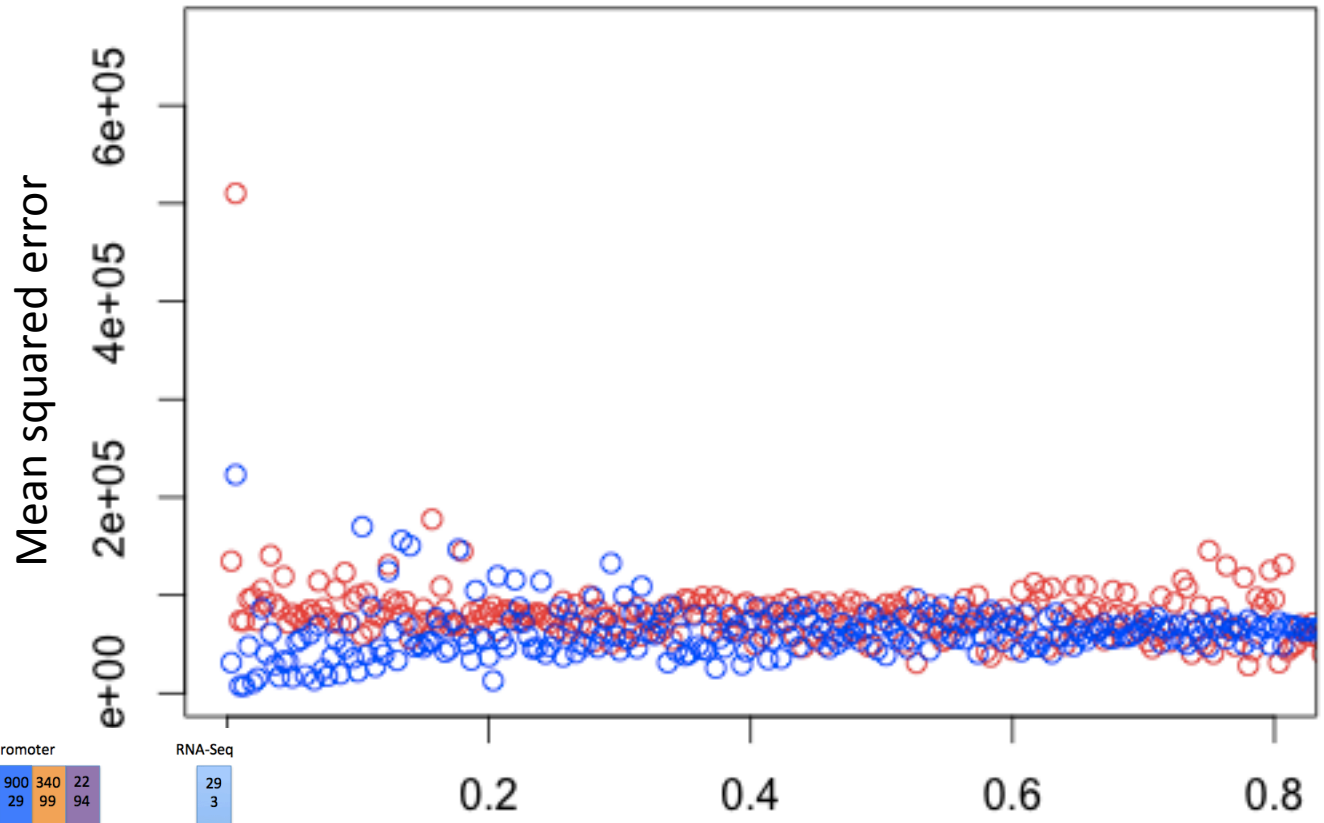
Mark Gerstein/Mengting Gu

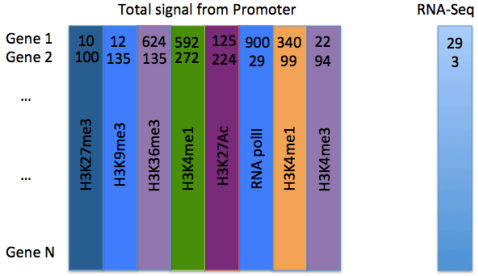
Adding more signals



Take $\log_2$ of each element

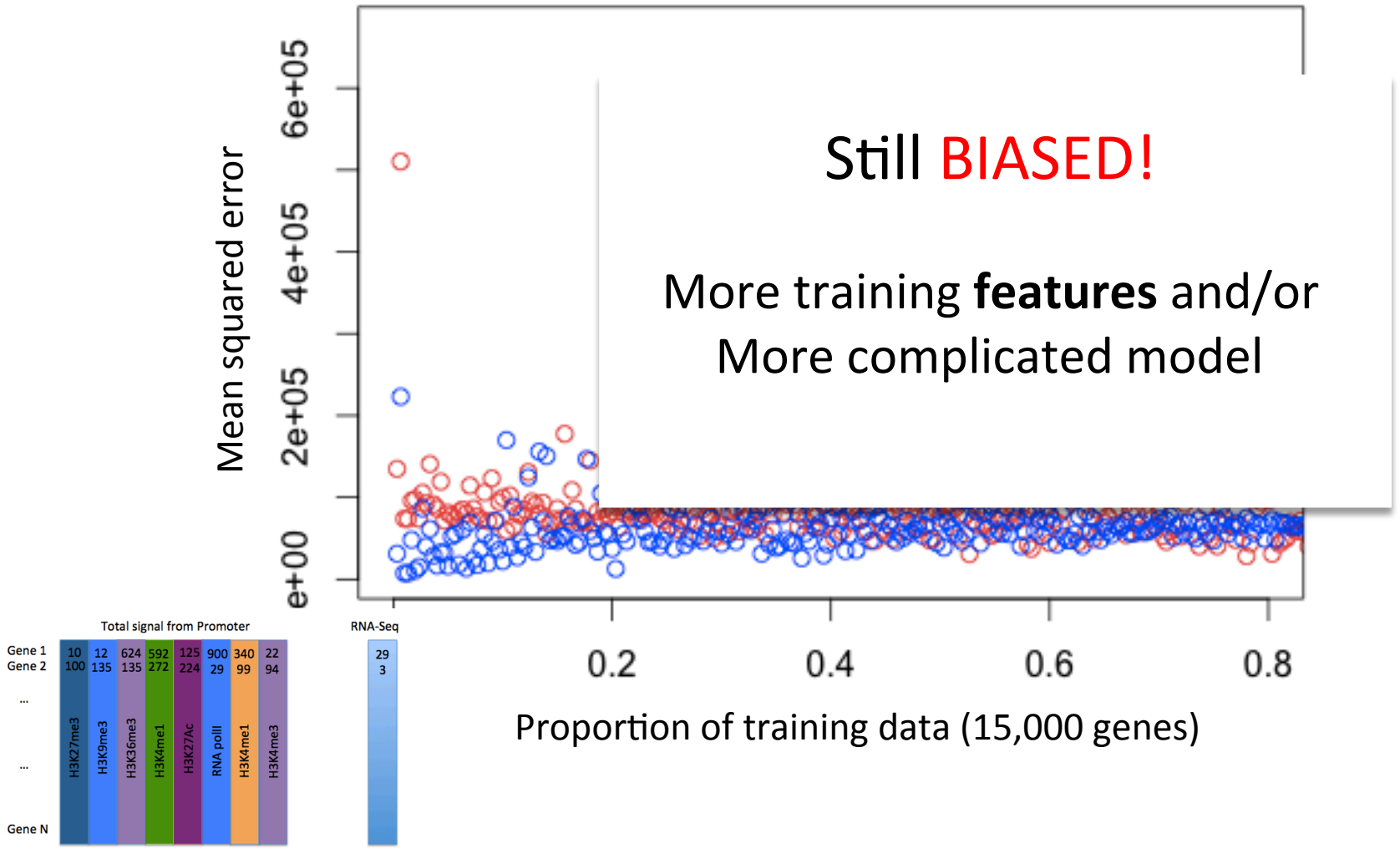
Multiple Linear Regression





$$Y = aX_1 + bX_2 + c + \ldots$$

Multiple Linear Regression



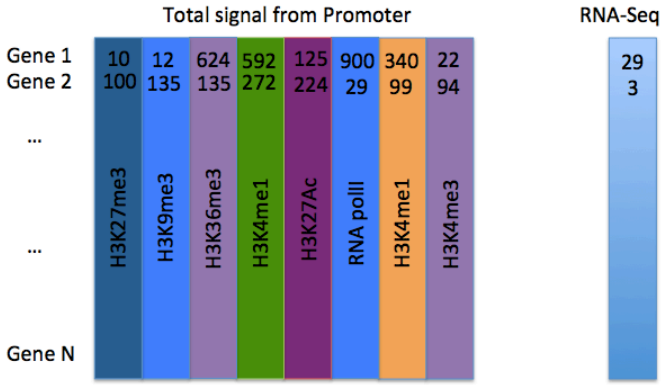
Random Forest Regression

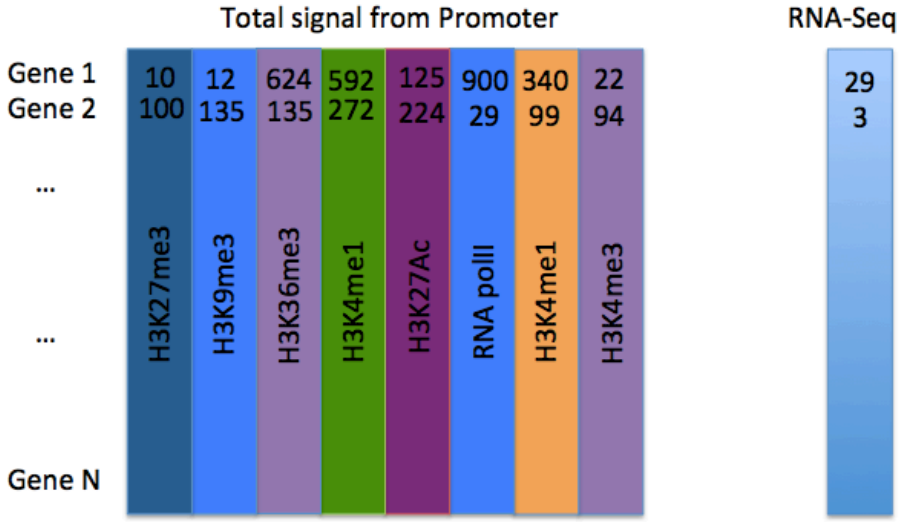
Log-transformed

Training correlation: 0.93
Test correlation: 0.52

Not Log-transformed

Training correlation: 0.95
Test correlation: 0.16



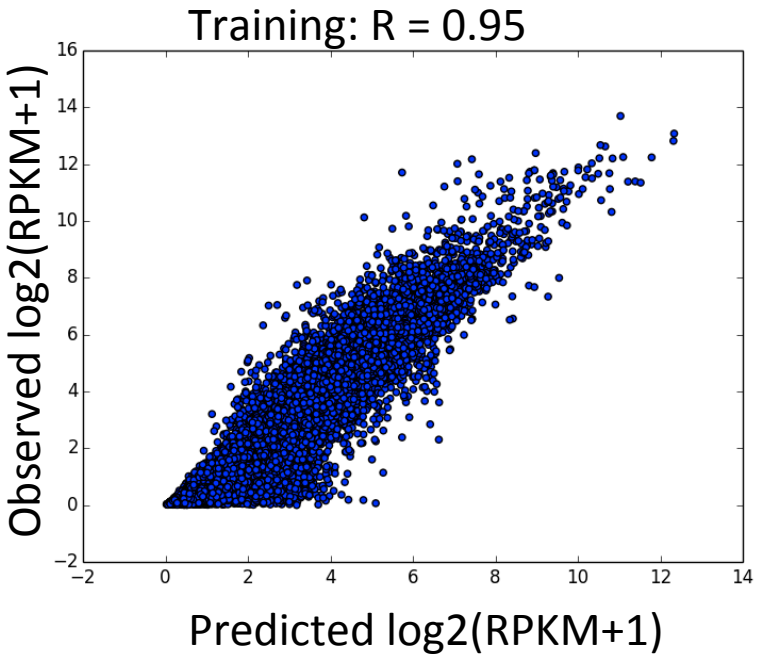


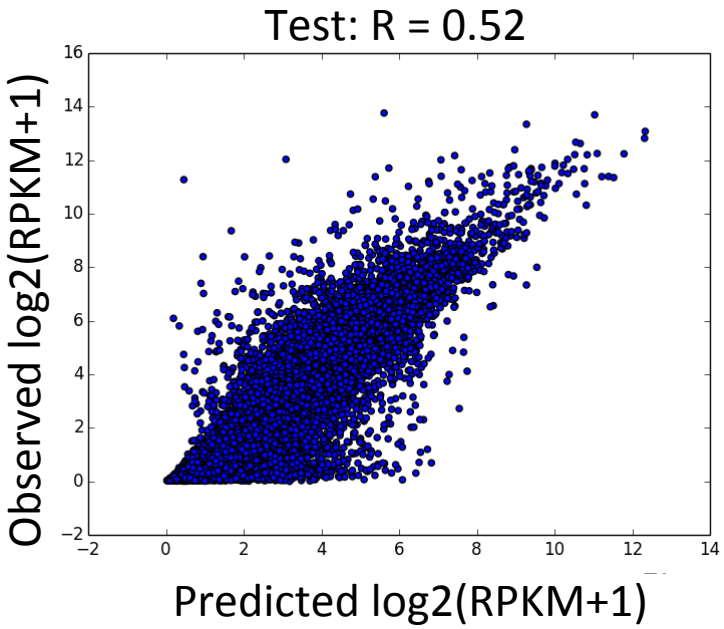
Random Forest Regression

Candy:

What's the problem – Bias or Variance?

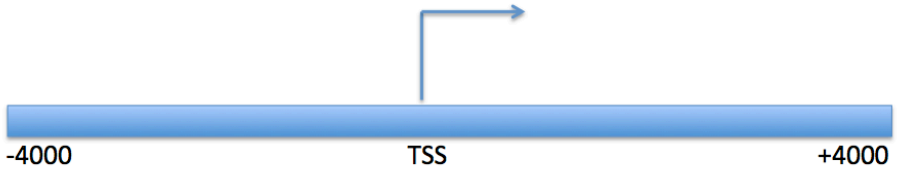
What should we do now?





What's the best model setup?

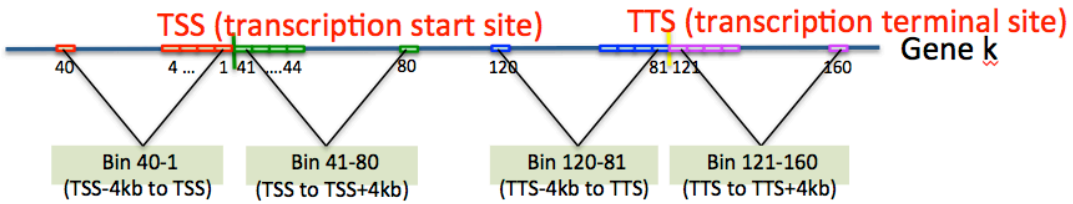
One bin around TSS



Candy: Which setup do we expect to perform better?

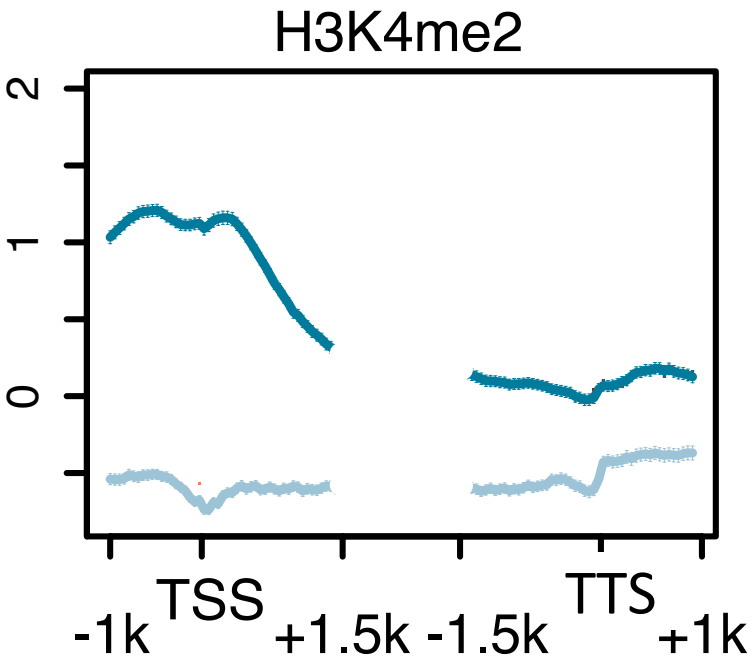
Vs.

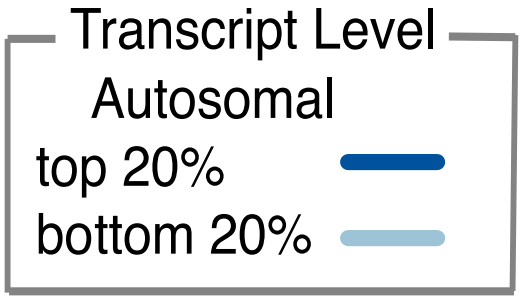
80 bins around TSS + 80 bins around TTS



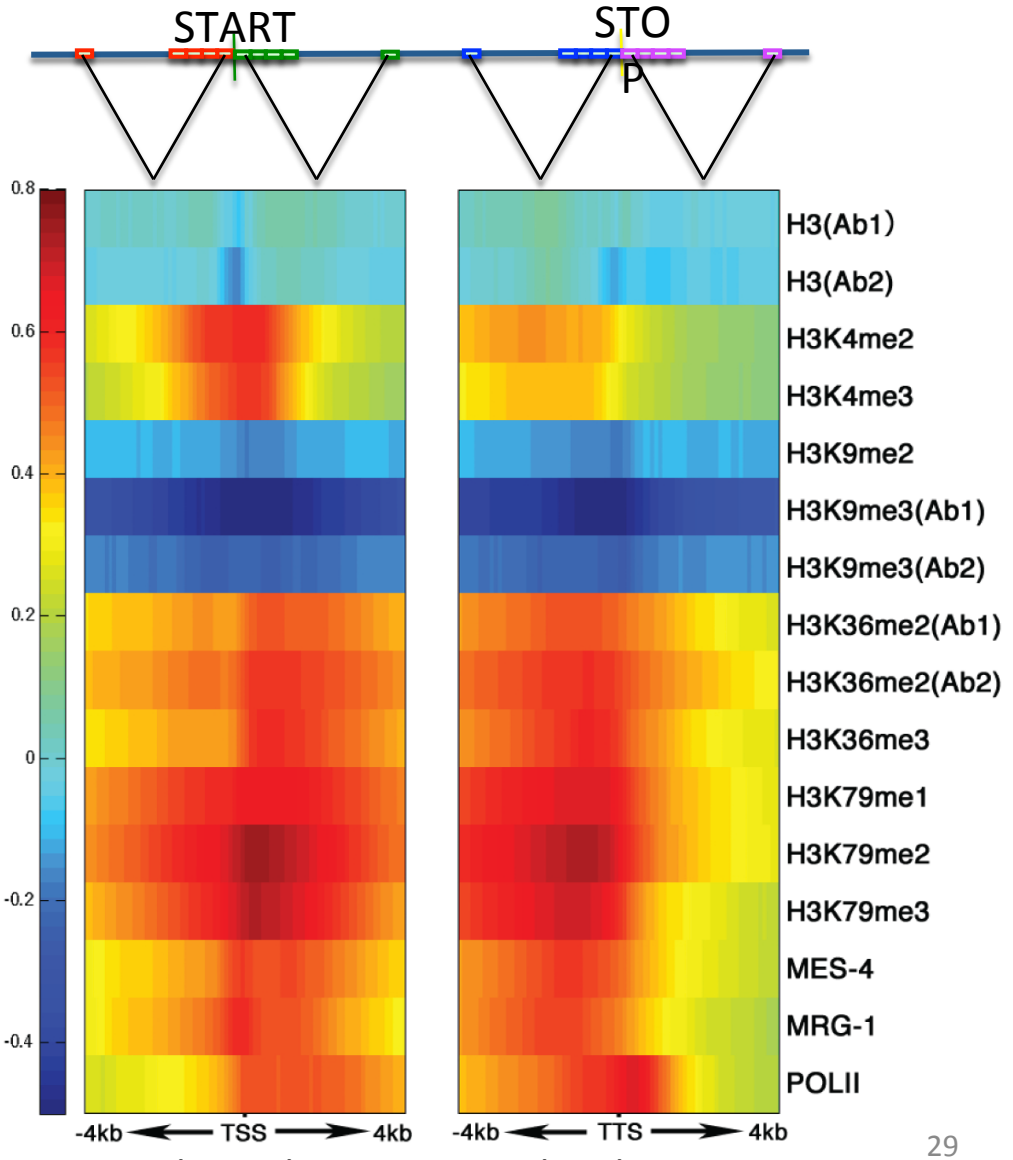
Cheng *et al*. *Genome Biology* 2011

Effects of signal depend on Location!





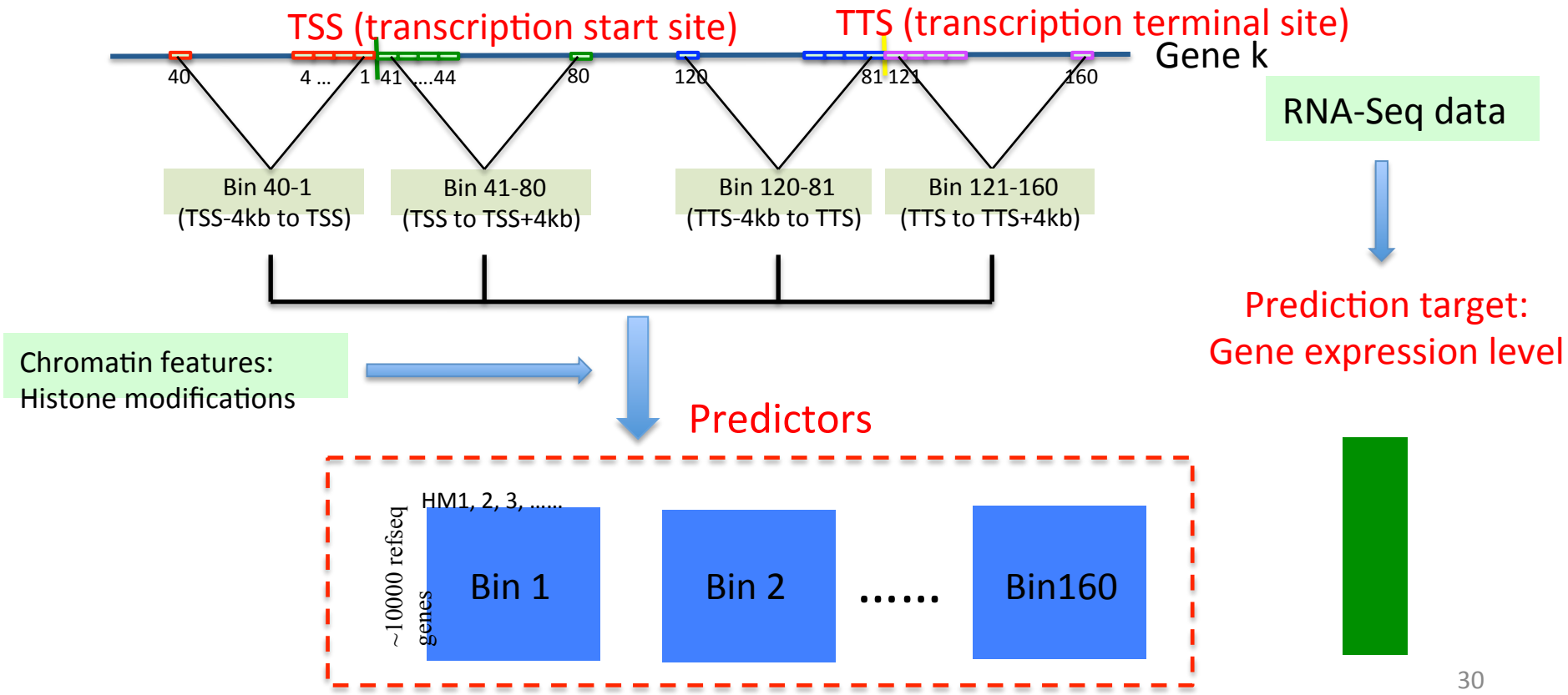
Gerstein,…, Cheng* et al. 2010, Science*



Correlation between Signal and expression

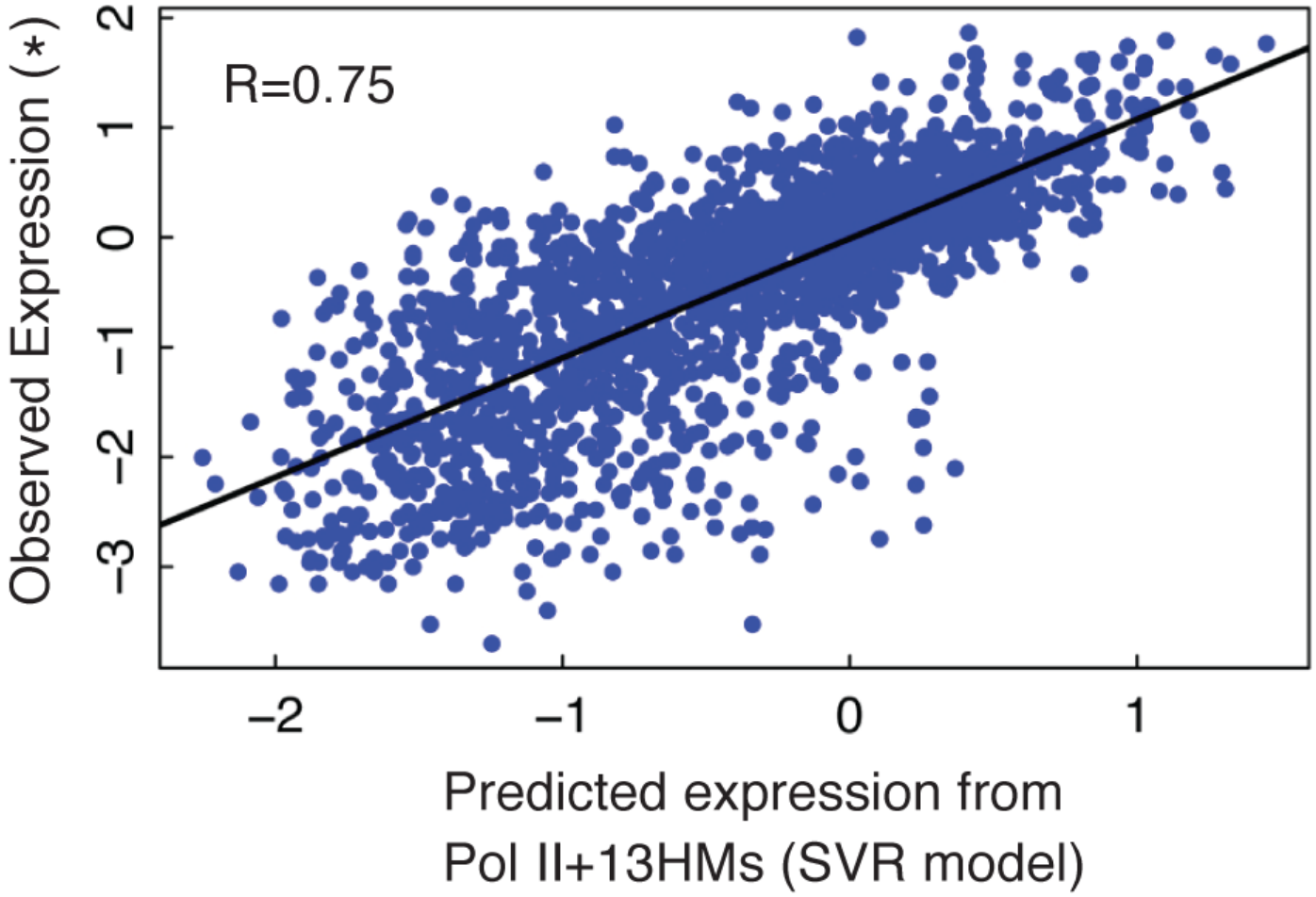
Slide by Chao Cheng

Setting up the model



Slide by Chao Cheng

Support vector regression to predict gene expression levels



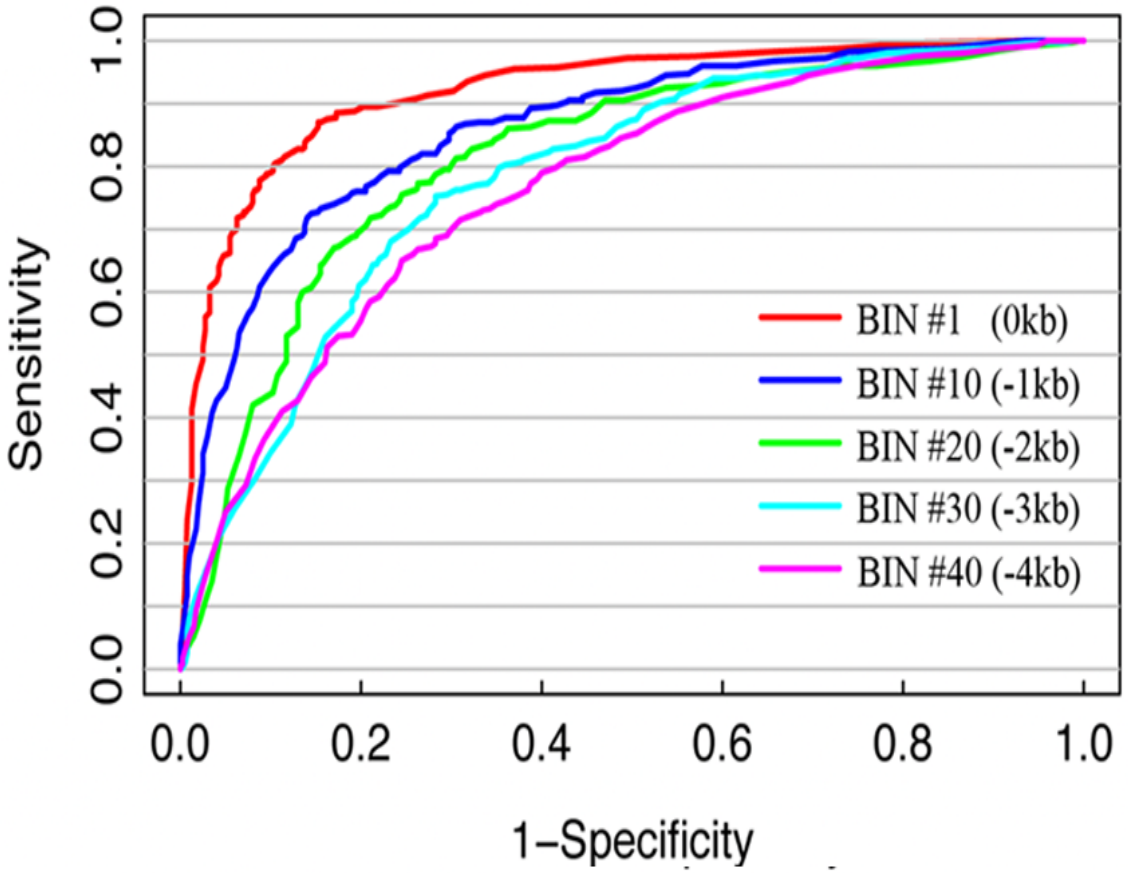
Slide by Chao Cheng

Context (TA bias)

- My implementation:
  - Train correlation = 0.95
  - Test correlation = 0.58

Support vector machine to classify genes with high, medium and low expression

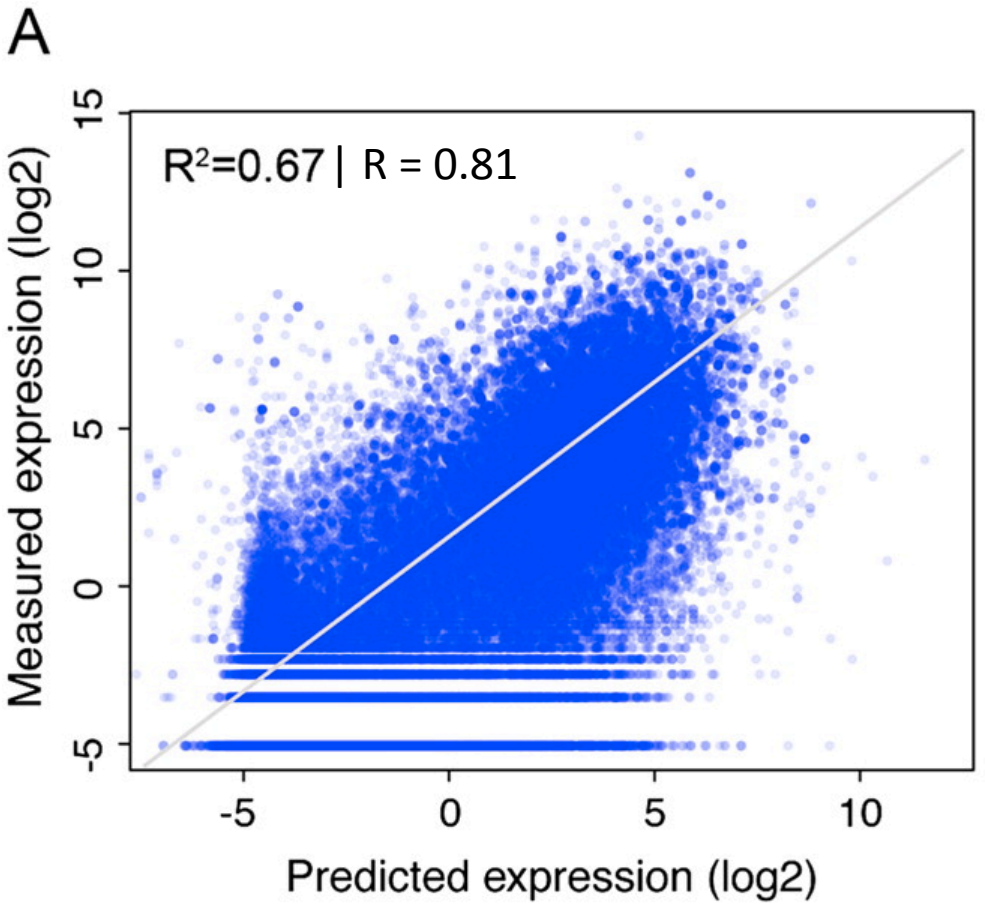
Candy: Describe a ROC curve!



✓Areas close to TSS predict expression better

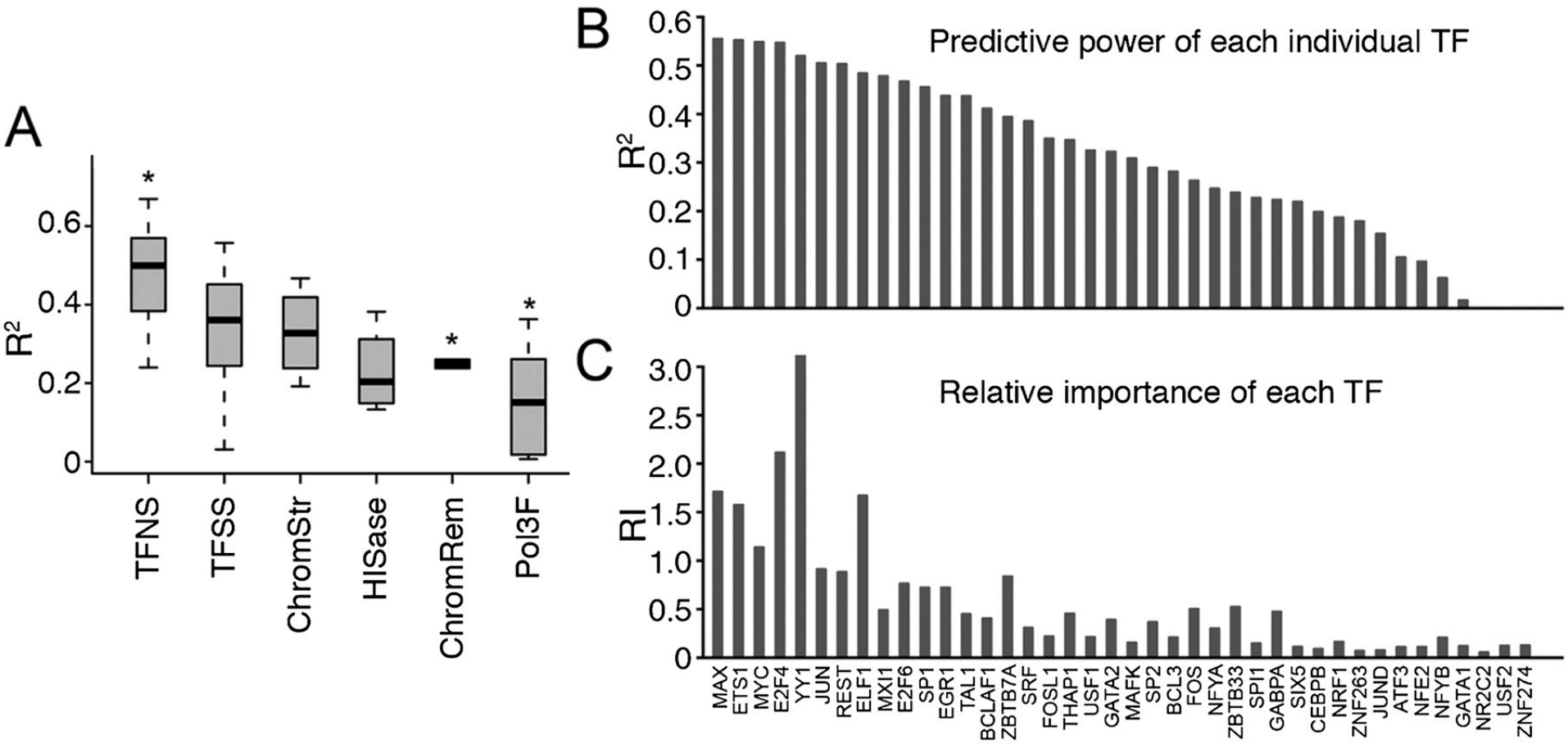
Adapted from Chao Cheng

Predicting Gene Expression with Transcription Factor ChIP-Seq signals



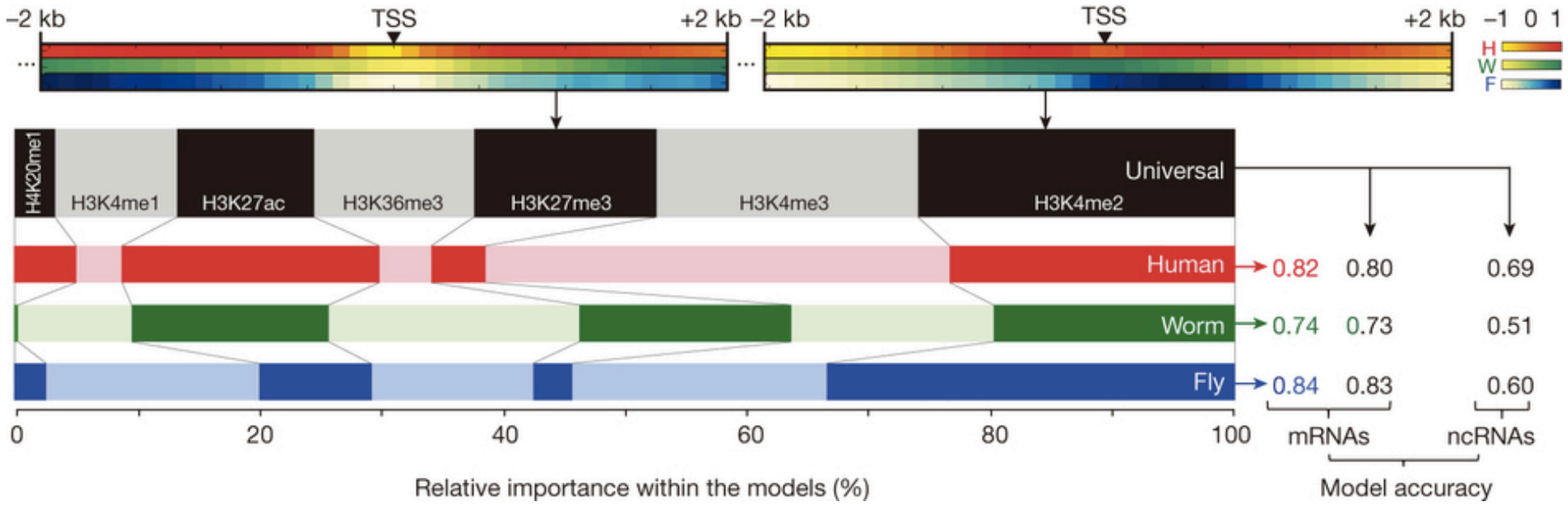
Cheng *et al*. *Genome Research* 2012

Predicting Gene Expression with Transcription Factor ChIP-Seq signals



Cheng *et al*. *Genome Research* 2012

Modeling Transcription Between Organisms



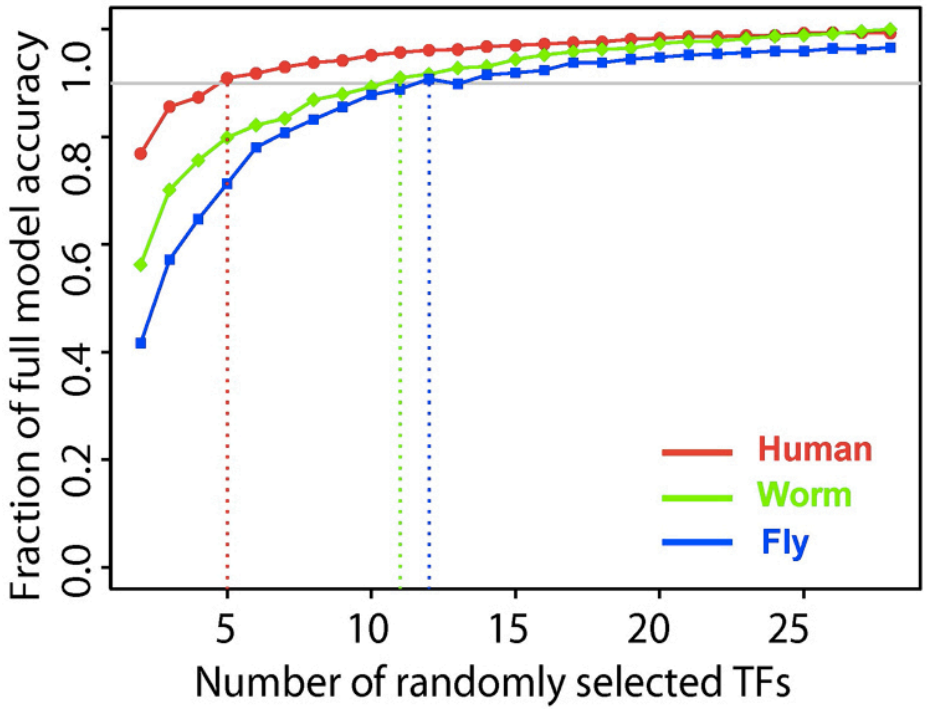
Gerstein *et al*. *Nature* 2014

Why do we care?

- What are the benefits of a quantitative model?
- Does this model help us understand the mechanism of transcription?

For discussion

- Will the prediction model perform accurately in cells with a transcription factor knocked out?



Gerstein *et al*. *Nature* 2014