Enhancer Prediction

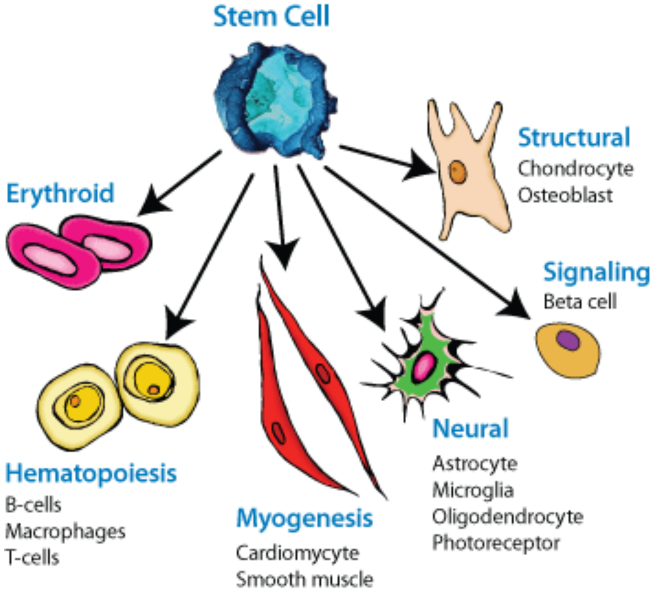
Albert Kuo

July 17, 2015

Background

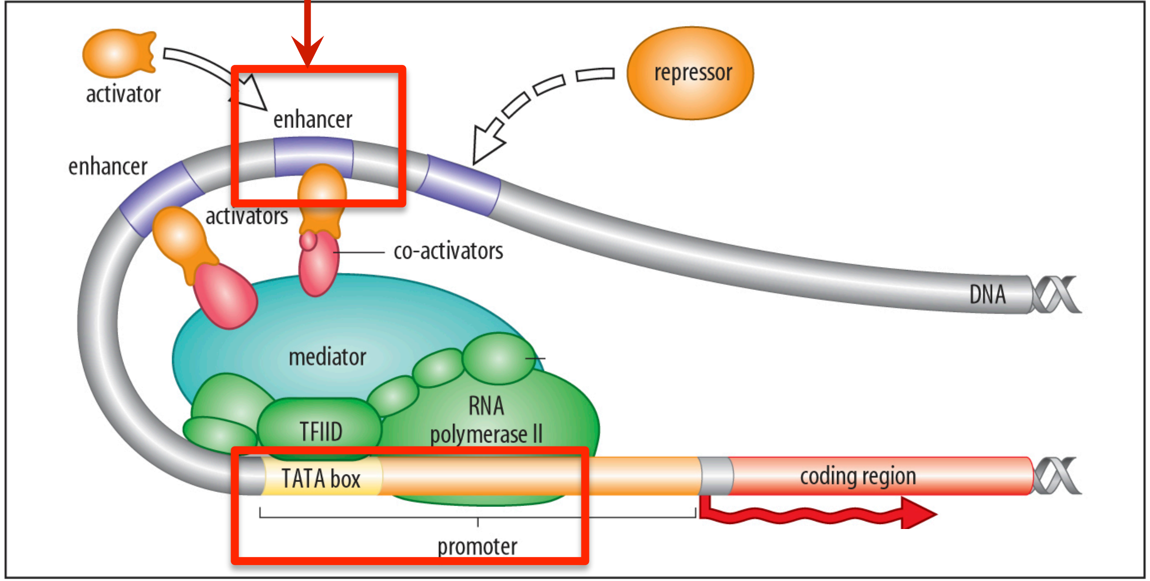
- What are enhancers?
- How can we identify enhancers?
  - Histone modifications
  - p300
  - eRNA

Enhancers



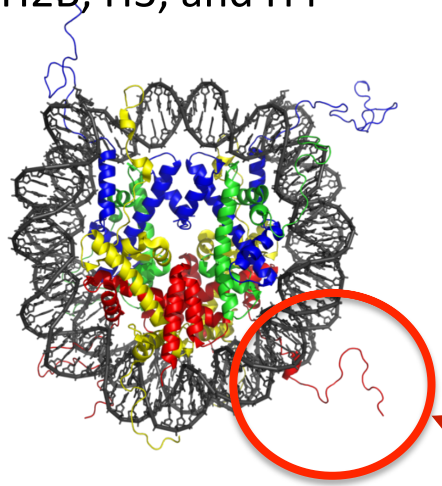
Cell-type specificity

Enhancers are short regions of DNA that can increase the transcription of genes.



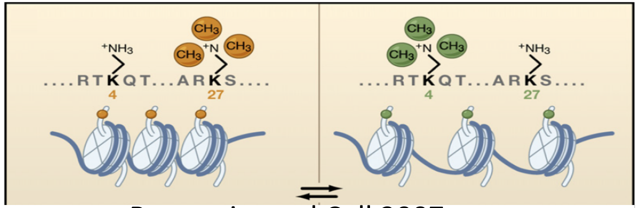
Histone modifications

Nucleosome consists of two copies each of H2A, H2B, H3, and H4



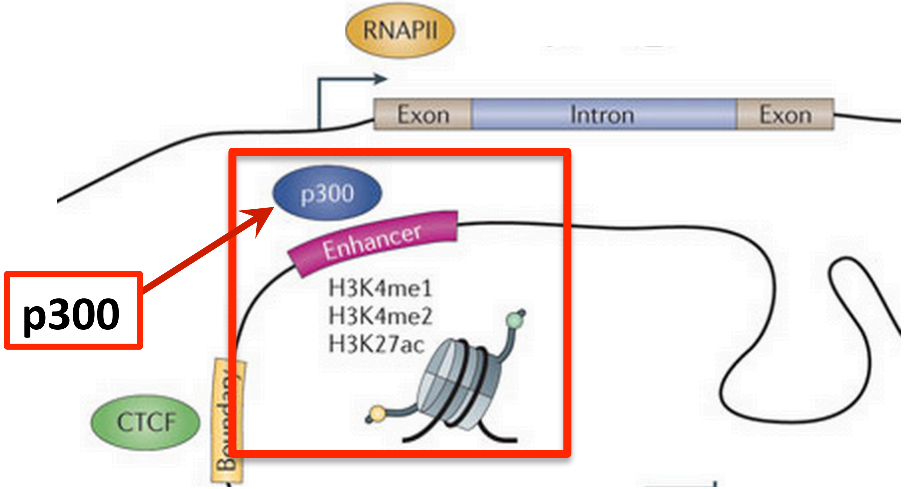
Histone tail

- Addition of a chemical group in an amino acid in a histone tail
- Histone modification mark:
  H3K4me1 - mark of enhancers
  H3K4me3 - mark of active promoters and TSS



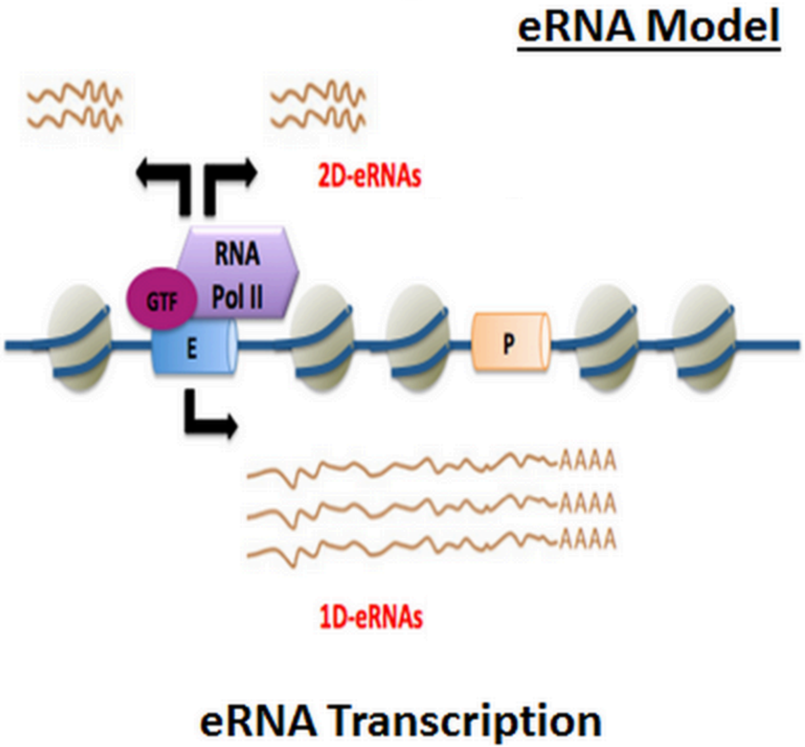
Bernstein et al Cell 2007

p300



- Binding protein
- Coactivator that increases gene expression
- Interacts with transcription factors
- Associated with enhancers

Enhancer RNAs (eRNAs)



- Non-coding RNA transcribed from enhancer regions
- eRNAs remain in nucleus and are quickly degraded
- Transcripts per million (TPM) reads measure RNA abundance

Recap: Identifying enhancers

1. Histone marks (e.g. H3K4me1)
2. p300, coactivator protein
3. eRNAs

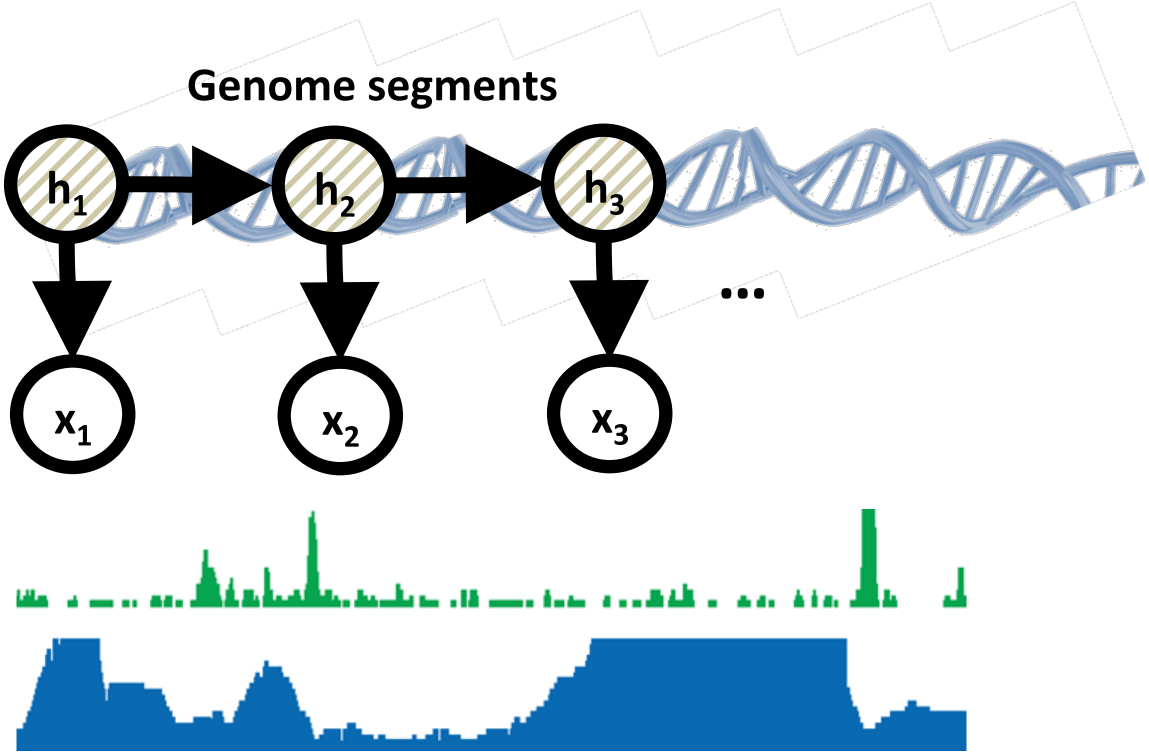
Dataset

- **Histone marks** from ChIP-seq data from ENCODE
  - H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac, H3K27me3, H3K36me3, H4K20me1
- **p300** from ENCODE
- **eRNA** expression from FANTOM
- Cell line
  - GM12878, lymphoblastoid cell line

Hidden Markov model (HMM) for chromatin state annotation

Hidden state ***h***$_i$
(Chromatin state)

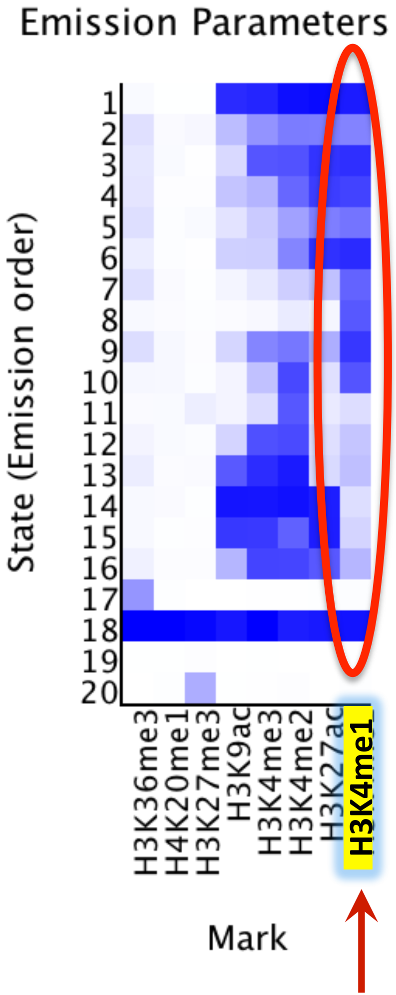
Observed state ***x***$_i$
(Combinations of histone modification marks on/off)



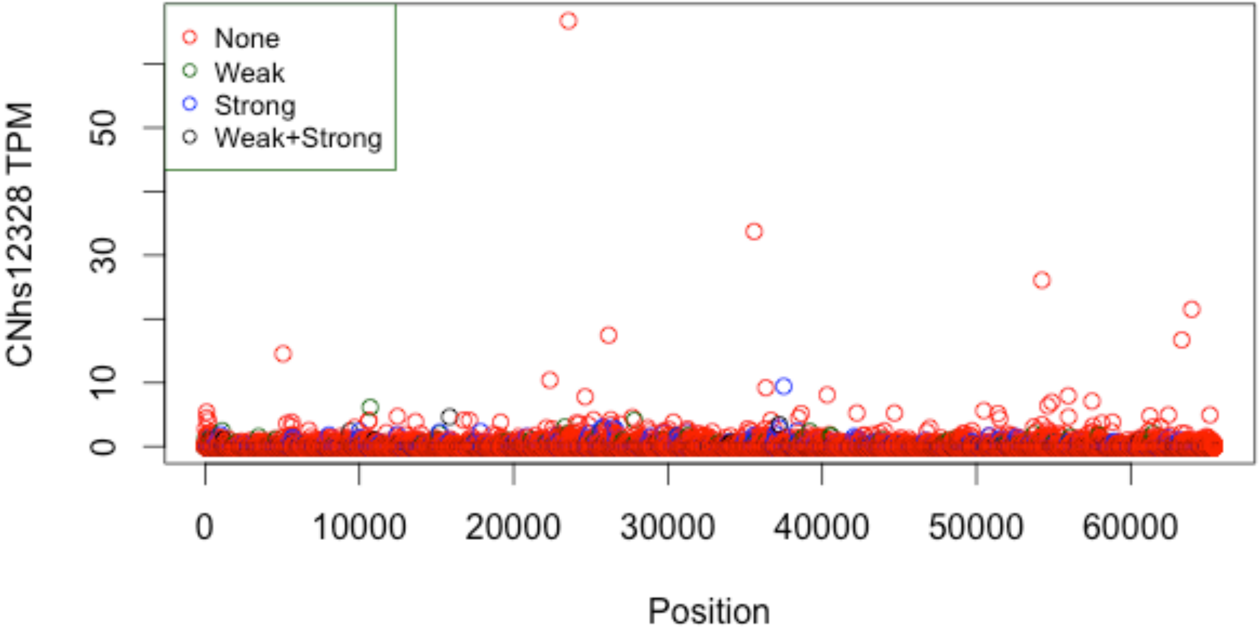
Histone modification marks

Spectacle

- Spectacle implements HMM
- Find states based on histone marks to make enhancer predictions
  - Level of H3K4me1 indicates weak enhancer state or strong enhancer state



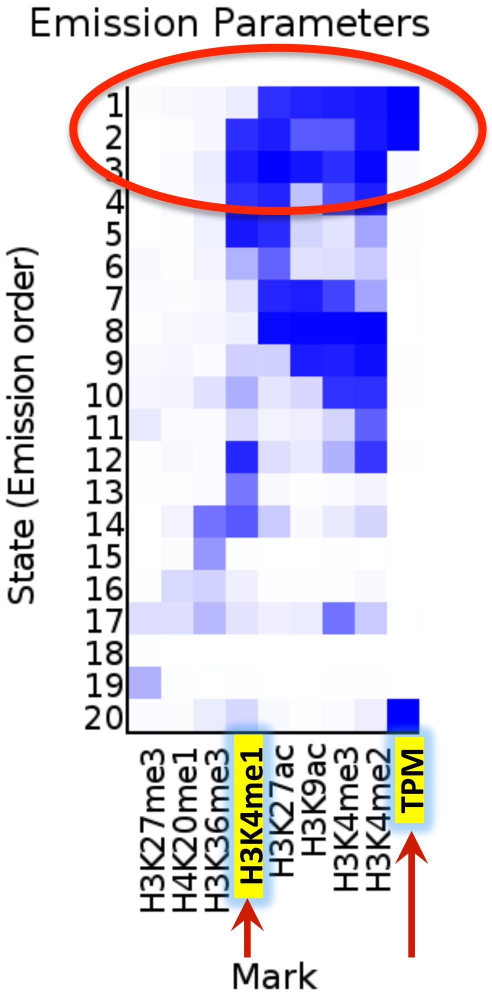
Enhancer predictions and eRNA



Group	Mean
None	0.032
Weak	0.034
Strong	0.055
Weak + Strong	0.034

Spectacle with eRNA

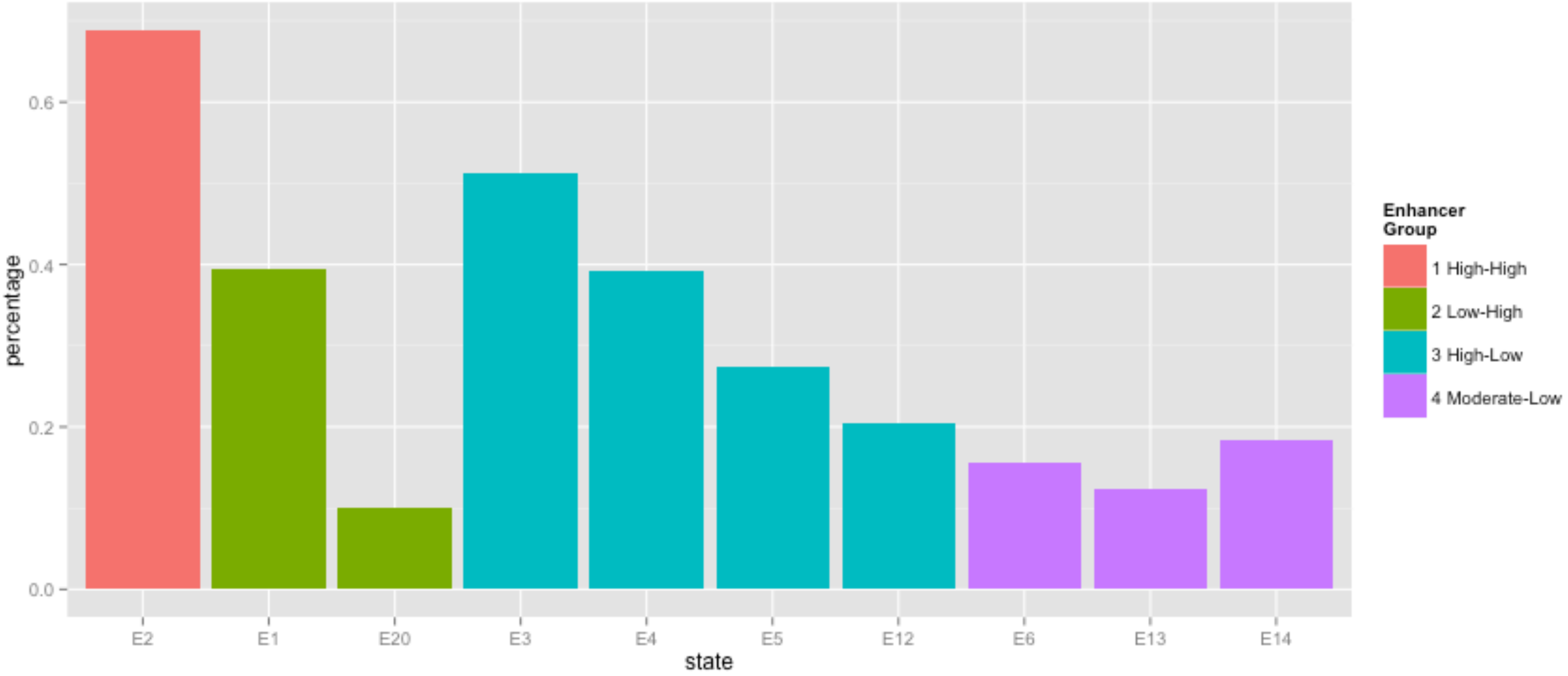
- Binary eRNA (threshold is TPM> 0)
- Different predictions based on histone marks versus eRNA expression
- Suggests eRNA expression level is not redundant information



Enhancer Groups

Group	State (# of sections)	Features
1 High-High	State 2 (4,791)	- High H3K4me1 - High TPM
2 Low-High	State 1 (1,917) State 20 (4,481)	- Low H3K4me1 - High TPM
3 High-Low	State 3 (12,626) State 4 (15,748) State 5 (21,871) State 12 (34,262)	- High H3K4me1 - Low TPM
4 Moderate-Low	States 6 (9,153) State 13 (7,8477) State 14 (8,534)	- Moderate H3K4me1 - Low TPM

Percentage Overlap with p300



Future Work

- Support Vector Machine (SVM) to predict enhancers based on histone marks and eRNA
  - Supervised learning model
  - Training set is p300 overlap
- Integration with other biological datasets
- Spectacle with eRNA and p300

Acknowledgments: Dr. Kevin Chen, Jimin Song