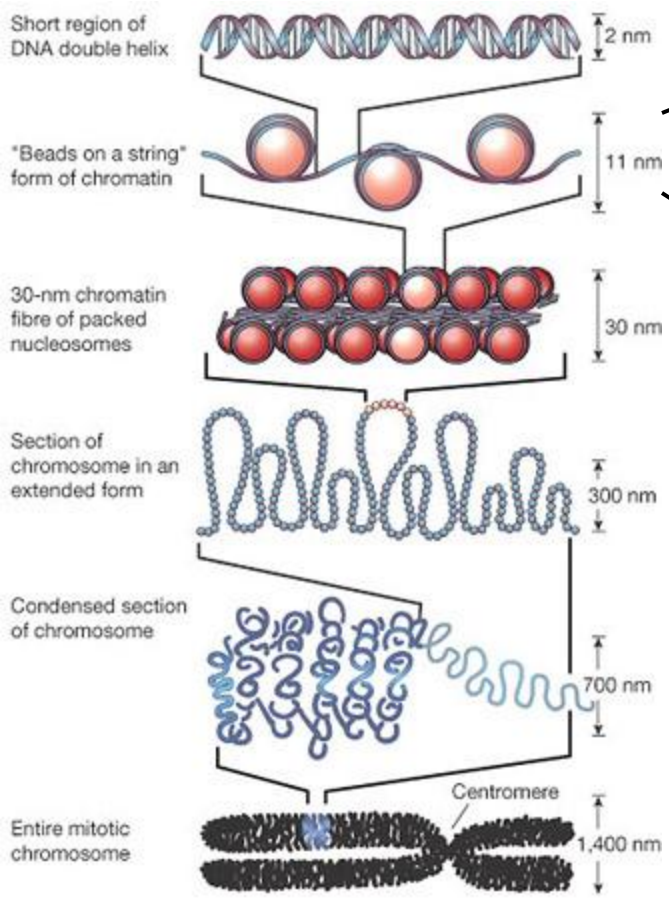
Modeling gene expression using chromatin features in various cellular contexts

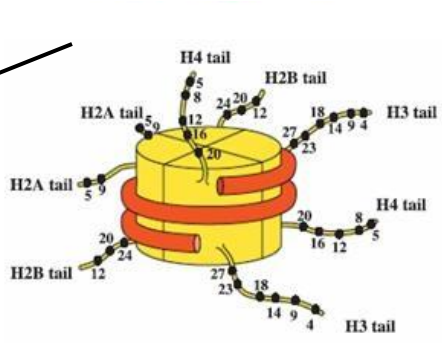
Dong X, et al

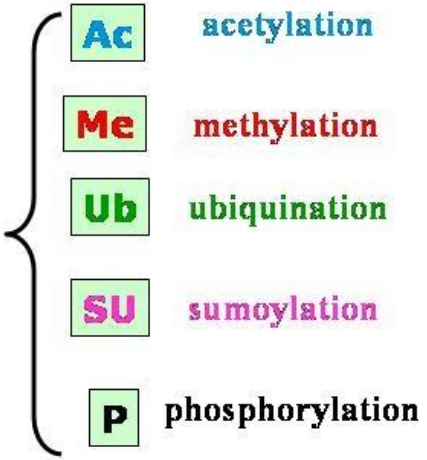
Questions

- Can we reproduce the quantitative relationship between gene expression levels and histone modifications?
- Does the relationship hold across different human cell lines and between different groups of genes?
- Do the most predictive chromatin features differ depending on expression quantification used?

nucleosome: octameric complex of histone proteins





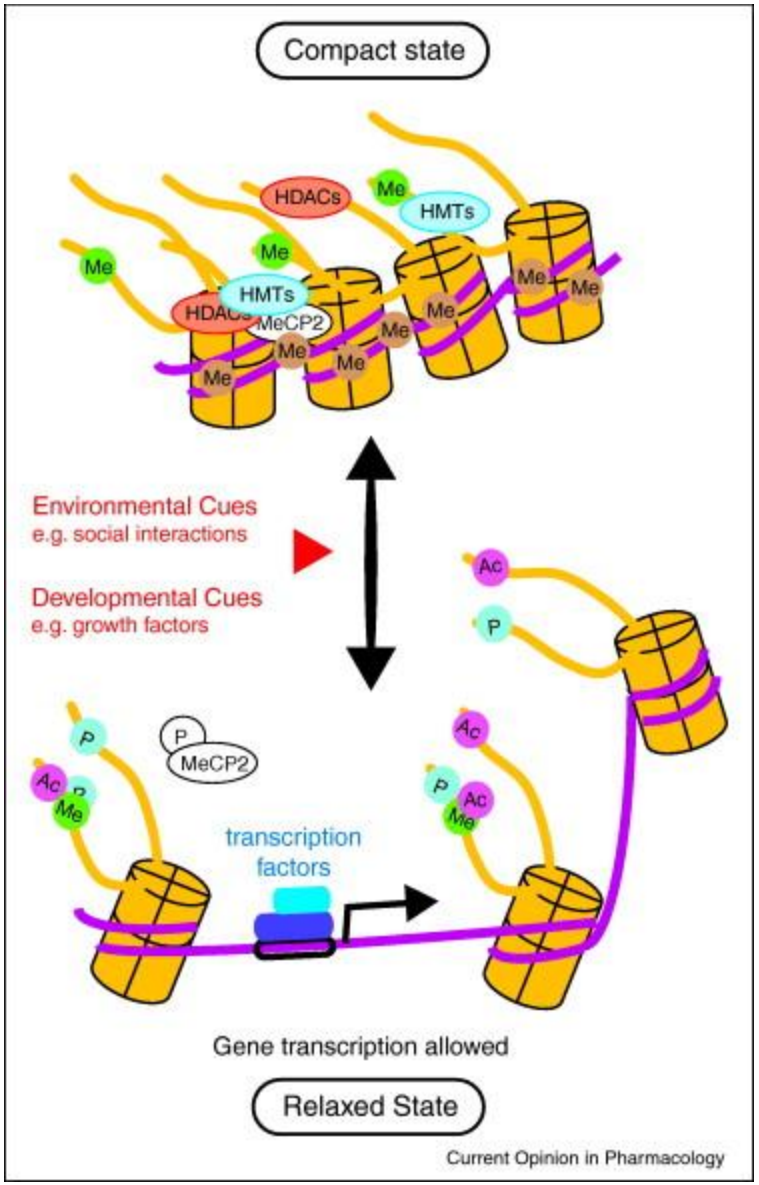


e.g. H3 K4 me2

name of histone

aa position in protein

type and number of modification(s)



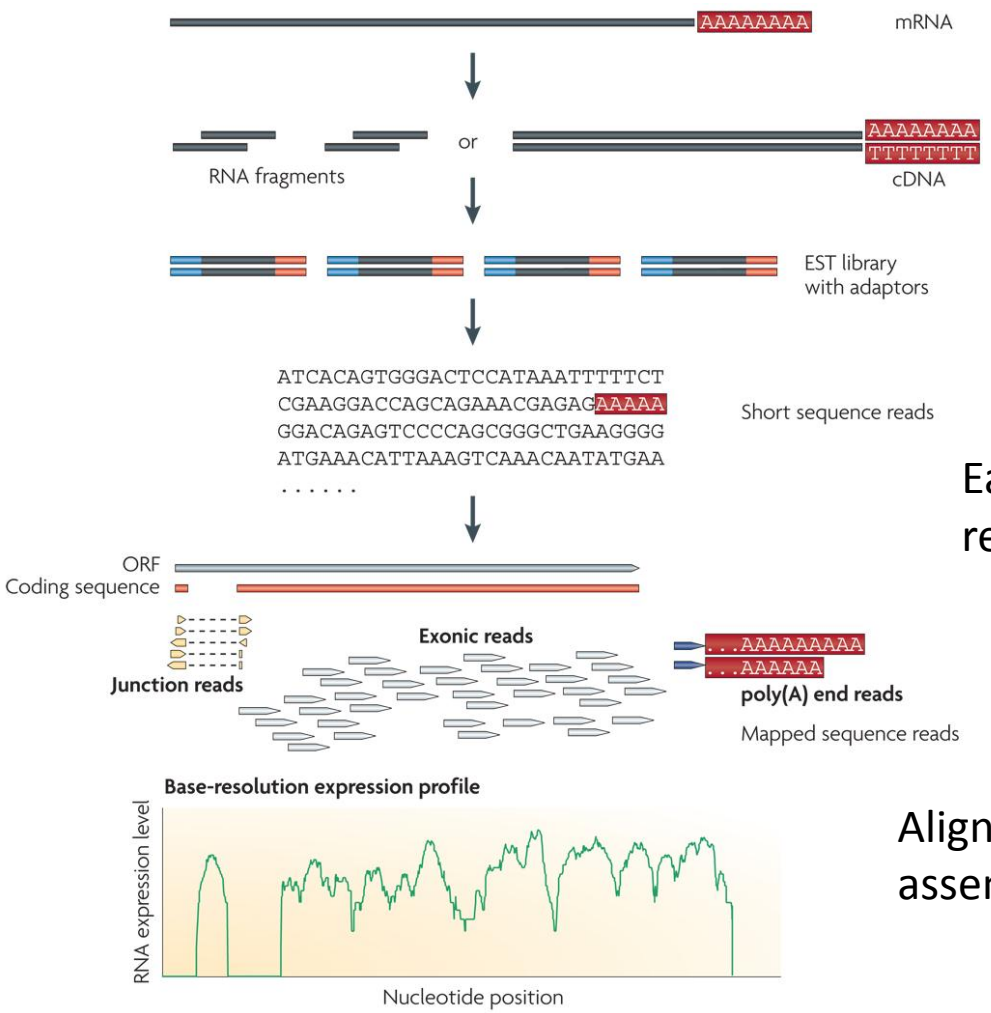
Geranton SM, 2012

11 histone modifications

- H3K4me1 (distal/other)
- H3K4me2 (promoter mark)
- H3K4me3 (promoter mark)
- H3K27me3 (repression)
- H3K36me3 (structural mark)
- H3K79me2 (structural mark)
- H3K9me1 (distal/other)
- H3K9me3 (repression)
- H4K20me1 (distal/other)
- H3K9ac (promoter mark)
- H3K27ac (promoter mark)
- H2A.Z (promoter mark)

- Seven human cell lines
- Expression quantification
  - RNA-Seq (transcript-based)
  - CAGE and RNA-PET (TSS-based)
- Chromatin feature data
  - CHiP-Seq for 11 histone modifications
  - DNaseI hypersensitive sites

RNA-Seq



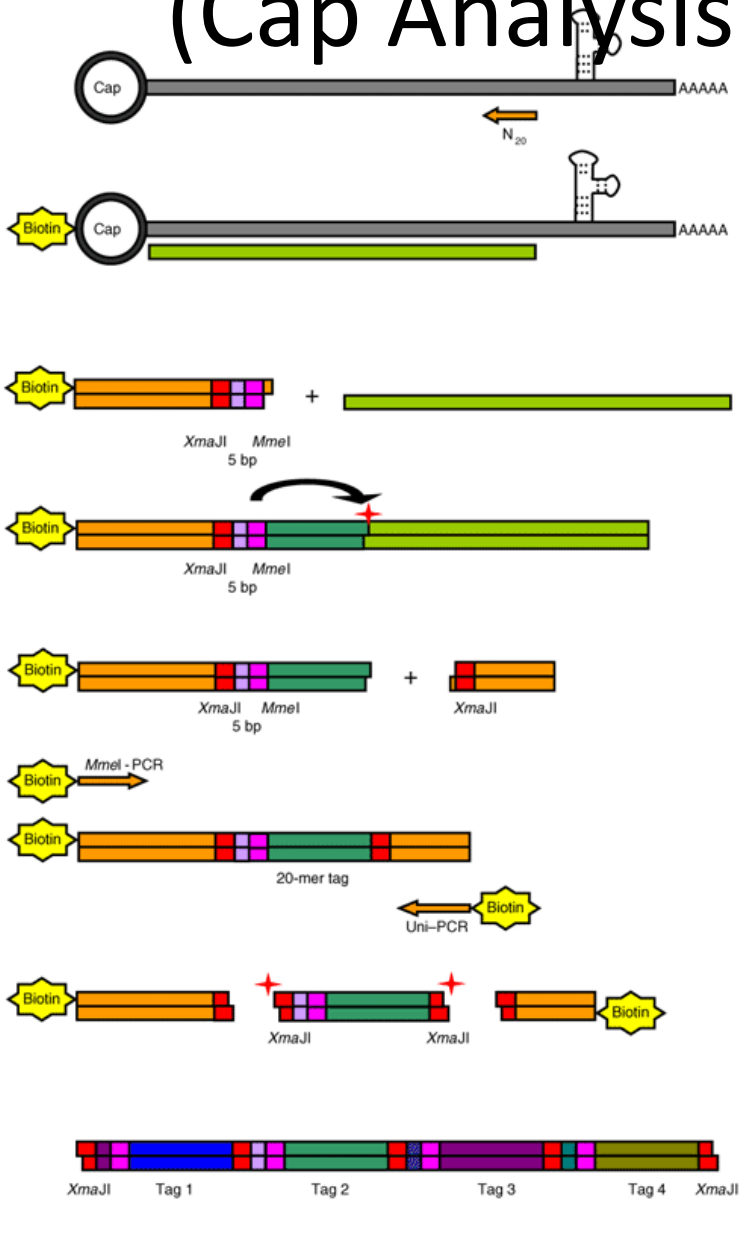
PolyA+ RNA converted to a library of cDNA fragments with adaptors

Each molecule is sequenced; short reads 30-400 bp

Aligned to a reference genome or assembled de novo

Wang et al, 2009

CAGE (Cap Analysis of Gene Expression)



- high-throughput identification of sequence tags corresponding to 5’ ends of mRNA at the cap sites and the identification of the TSS

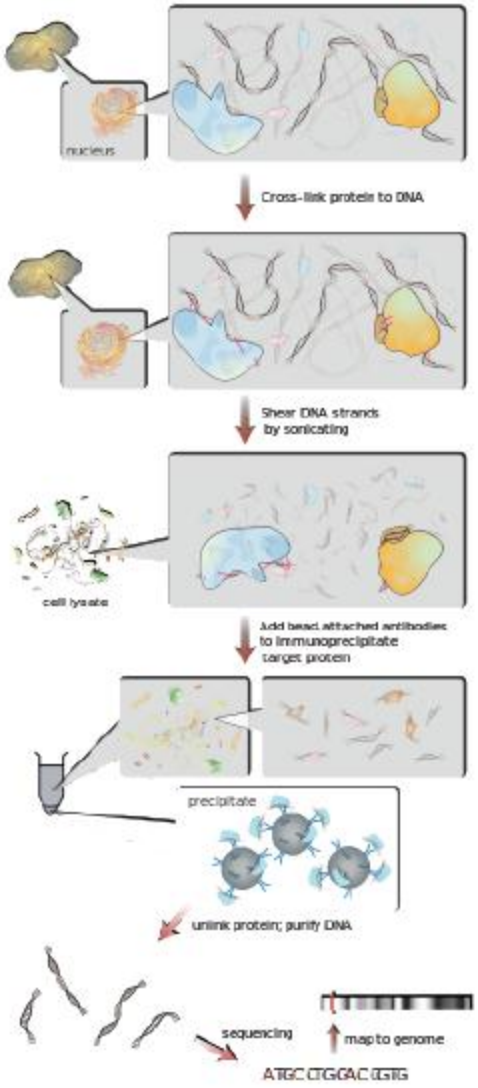
Linkers attached to 5’ end

Cleavage of first 20bp by class II RE

RNA-PET (Paired-End-Tag nextgen sequencing)

- Captures and sequences the 5’ and 3’ end tags of full length cDNA fragments of all expressed genes
  - Demarcate the boundaries of transcription units
- For this study only 5’ tags were used to capture TSS

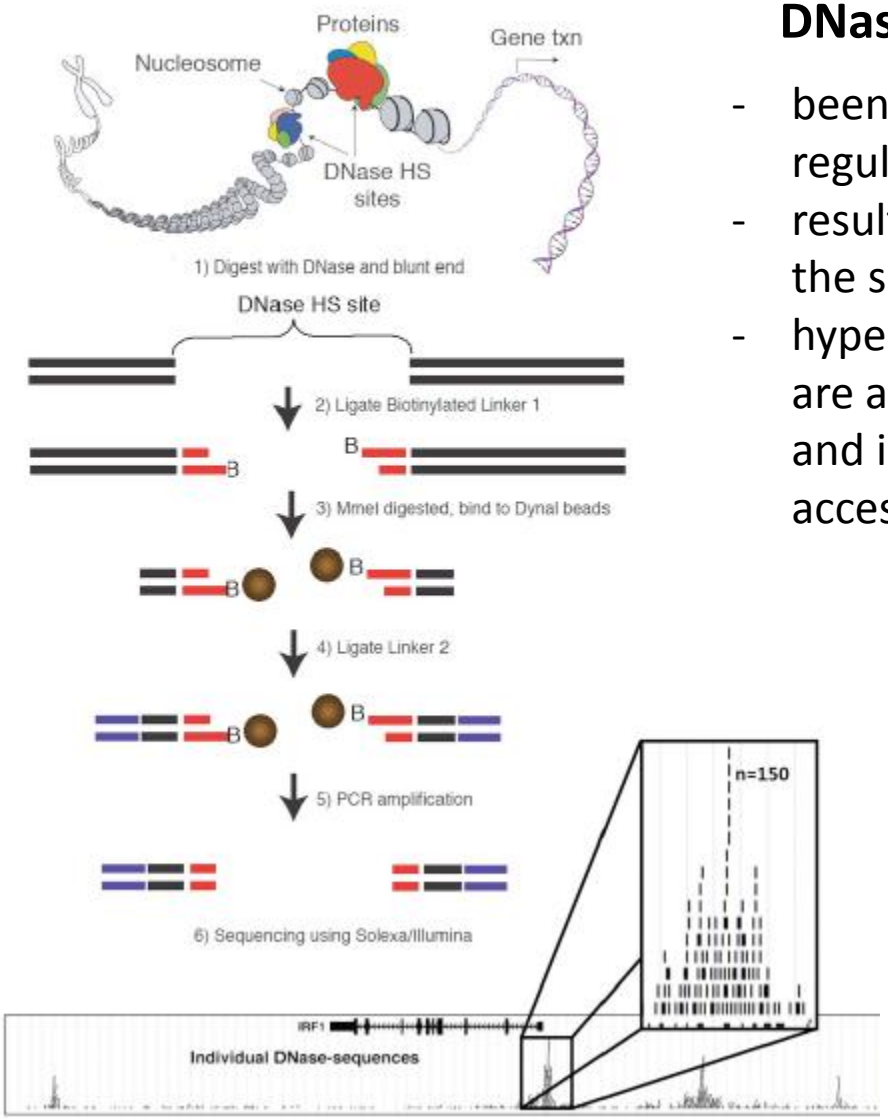
CHiP-Seq



Goal:

To map the binding sites of a target protein (modifies histones) with maximal signal-to-noise ratio and completeness across the genome

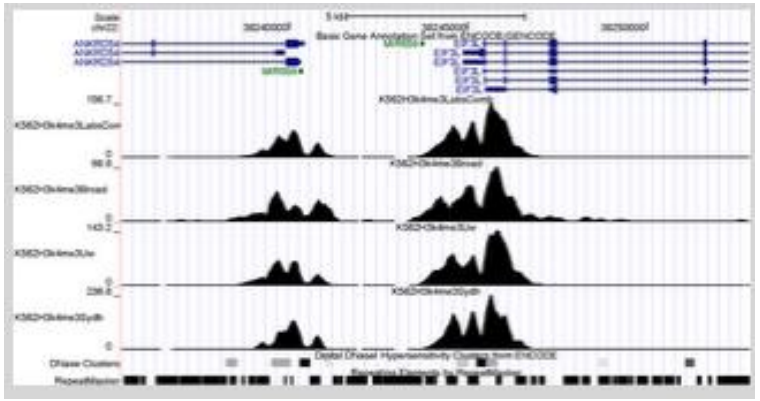
DNase-Seq



DNaseI Hypersensitive Sites

- been shown to be associated with all types of regulatory elements (i.e. promoters, enhancers)
- result from the binding of trans-acting factors at the site of canonical
- hypersensitivity is an indication that nucleosomes are absent or that chromatin structure is loose, and is a reflection of chromatin openness and accessibility

- Datasets downloaded as signal tracks in bigwig format



- Defining ‘bestbin’ of chromatin feature density
  - ‘bestbin’= bin with the highest correlation with gene expression level
  - Mean density of chromatin features in each bin using bigwig summary

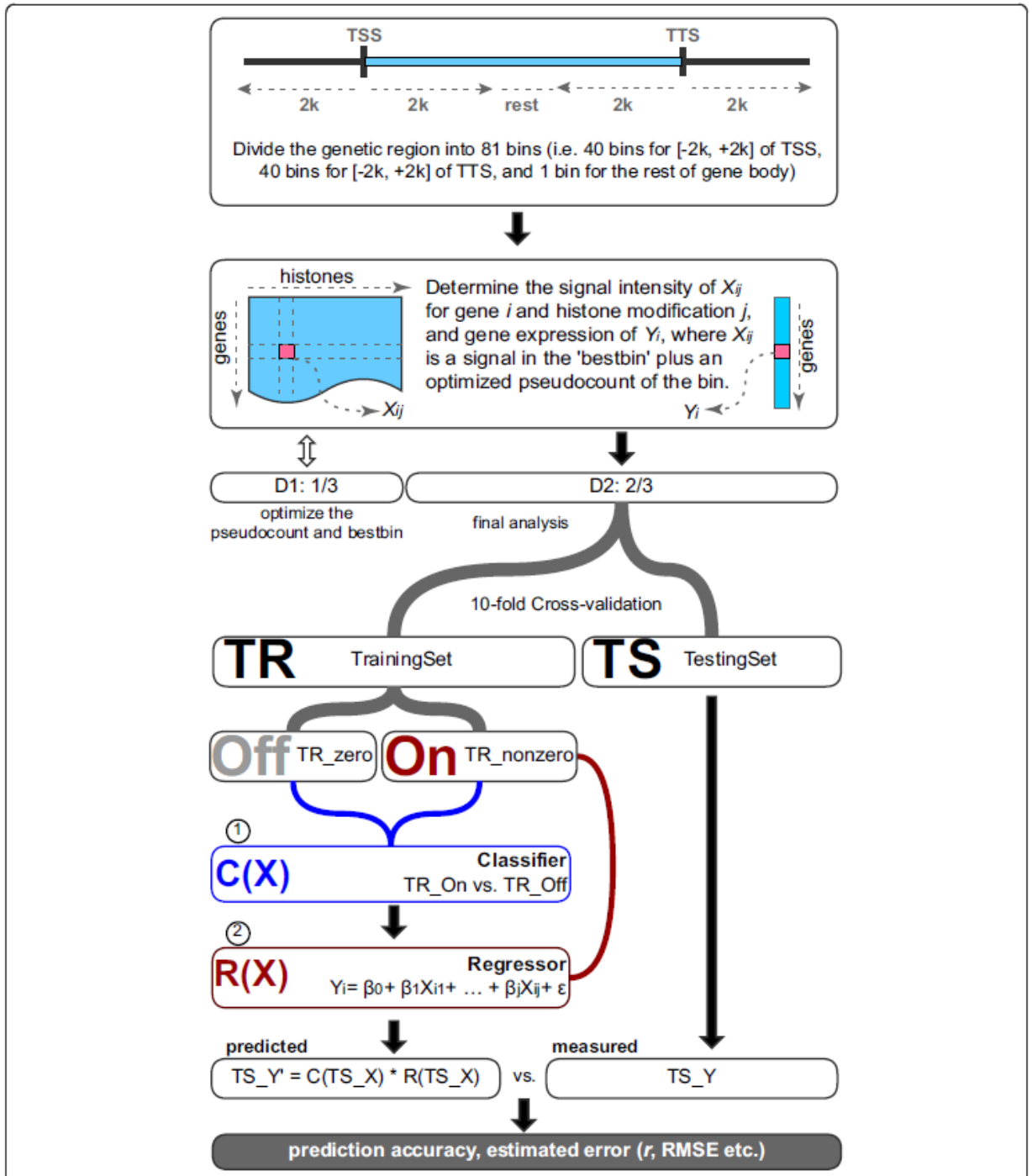


Figure 1 Modeling pipeline. Genes longer than 4,100 bp were extended and divided into 81 bins. The chromatin feature density in each bin is

2-step model to predict the expression levels of GENCODE genes

- Random forest classification: to predict whether the promoter is expressed
- Regression model: to predict expression level of promoter
- Performance was evaluated based on ten-fold cross-validation
  - Each dataset divided into training genes (1/3) and test set (2/3)
  - AUC to measure accuracy of classification; PCC to measure predictive accuracy of regression model

CAGE on long cytosolic PolyA+; K562 cells

~6000 genes correctly classified as unexpressed

Most experiments show a strong correlation between predicted and measured expression levels

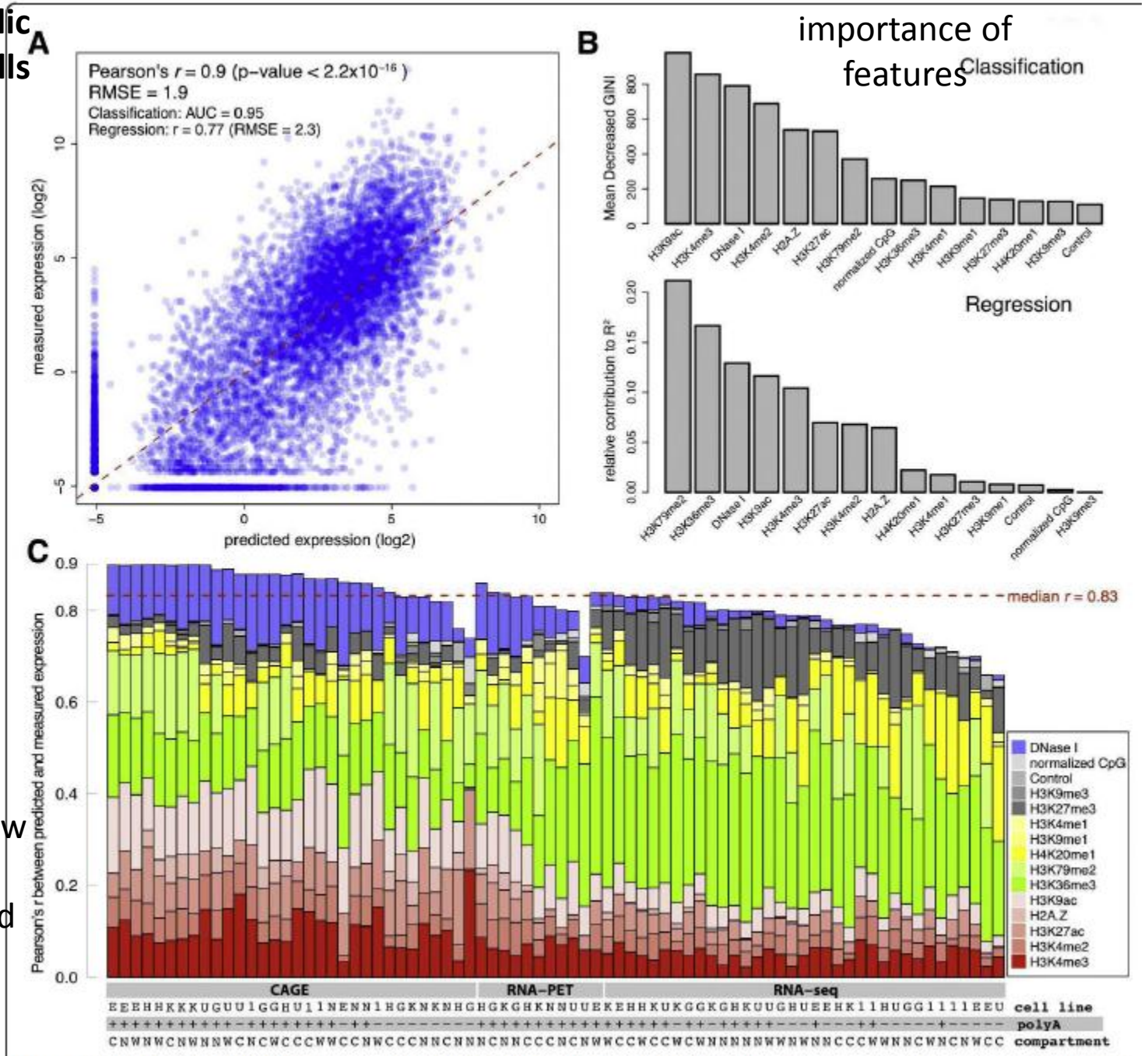


Figure 2 Quantitative relationship between chromatin feature and expression. (a) Scatter plot of predicted expression values using the

- 3 Randomization tests: no inherent structures leading to ‘easy’ prediction
  - Randomly shuffling expression values of genes
  - Shuffle each chromatin feature independently
  - Swapping the x labels? Before applying models to the testing set

Comparison of different techniques

Whole cell is more similar to cytosol →

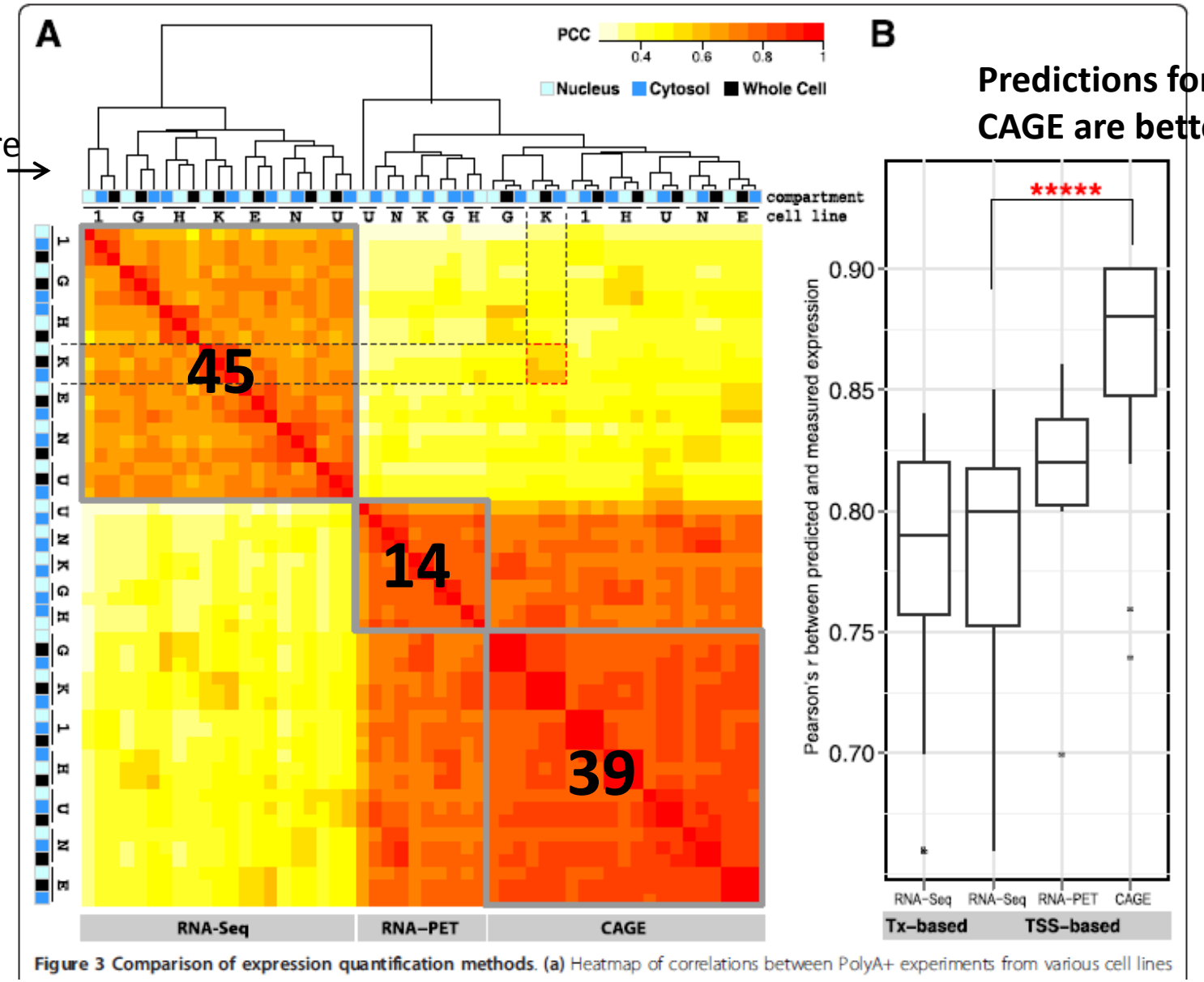


Figure 3 Comparison of expression quantification methods. (a) Heatmap of correlations between PolyA+ experiments from various cell lines

Prediction across different cell lines

Cross-cell line prediction (keeping technique and compartment constant)

Relatively low for hESC line

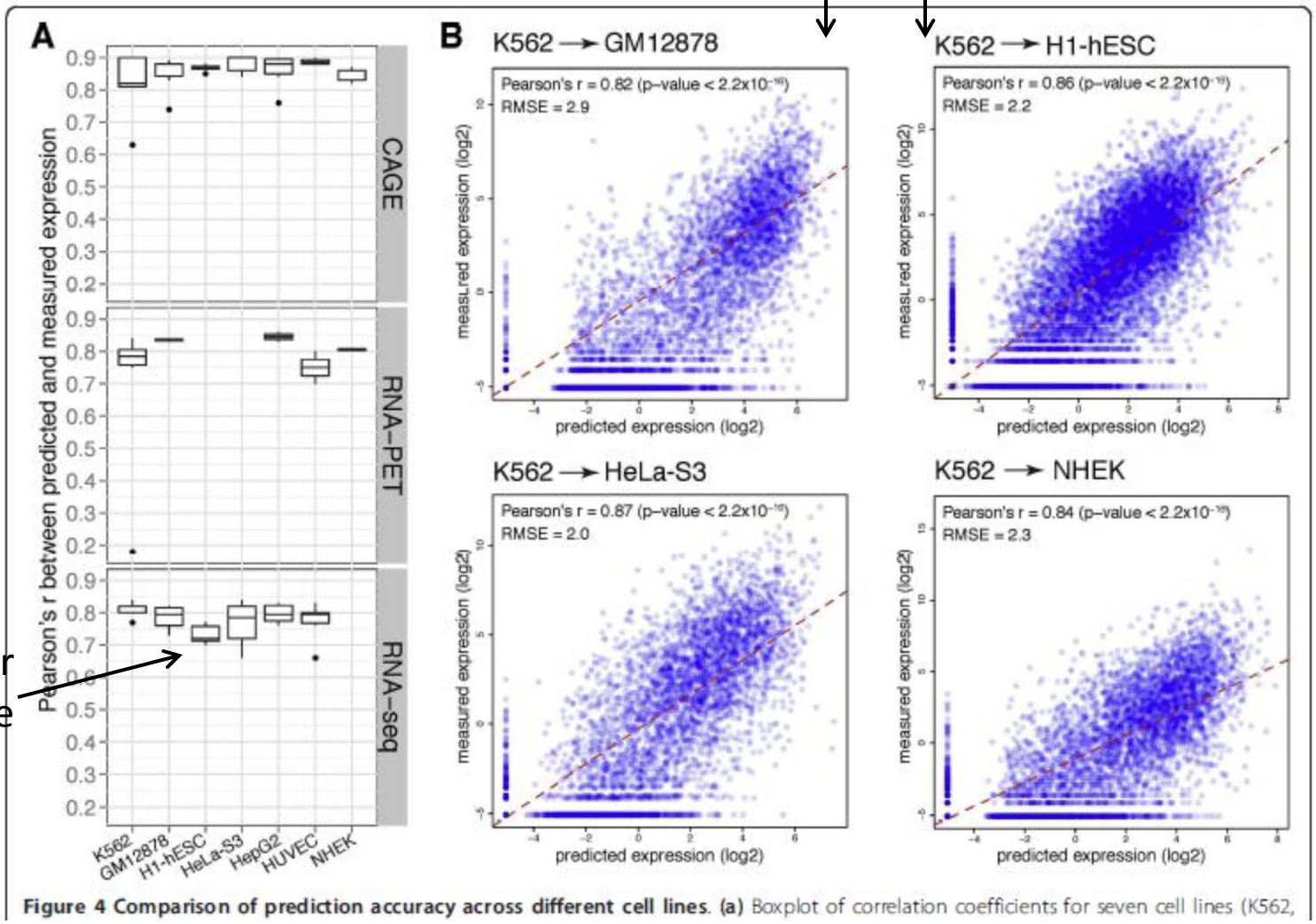
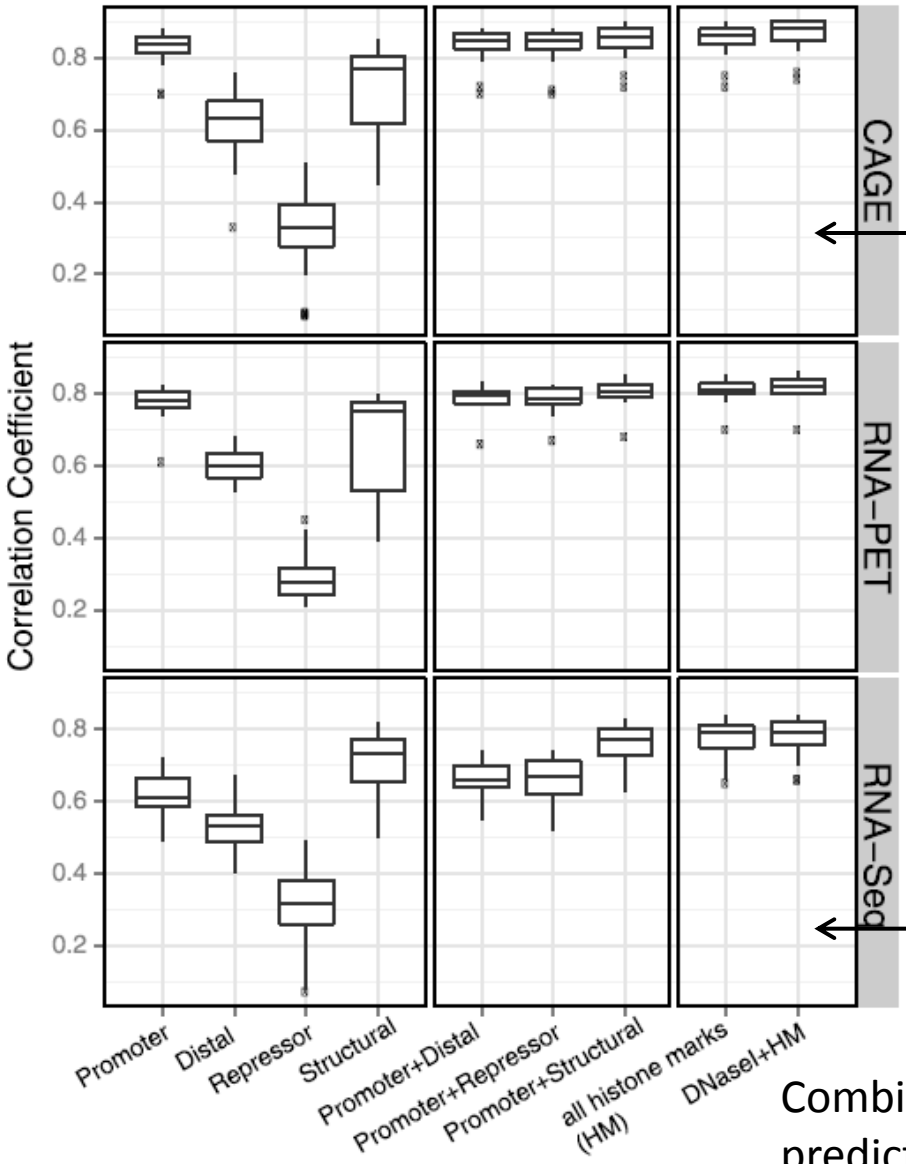


Figure 4 Comparison of prediction accuracy across different cell lines. (a) Boxplot of correlation coefficients for seven cell lines (K562,

Transcription initiation and elongation are reflected by different chromatin features



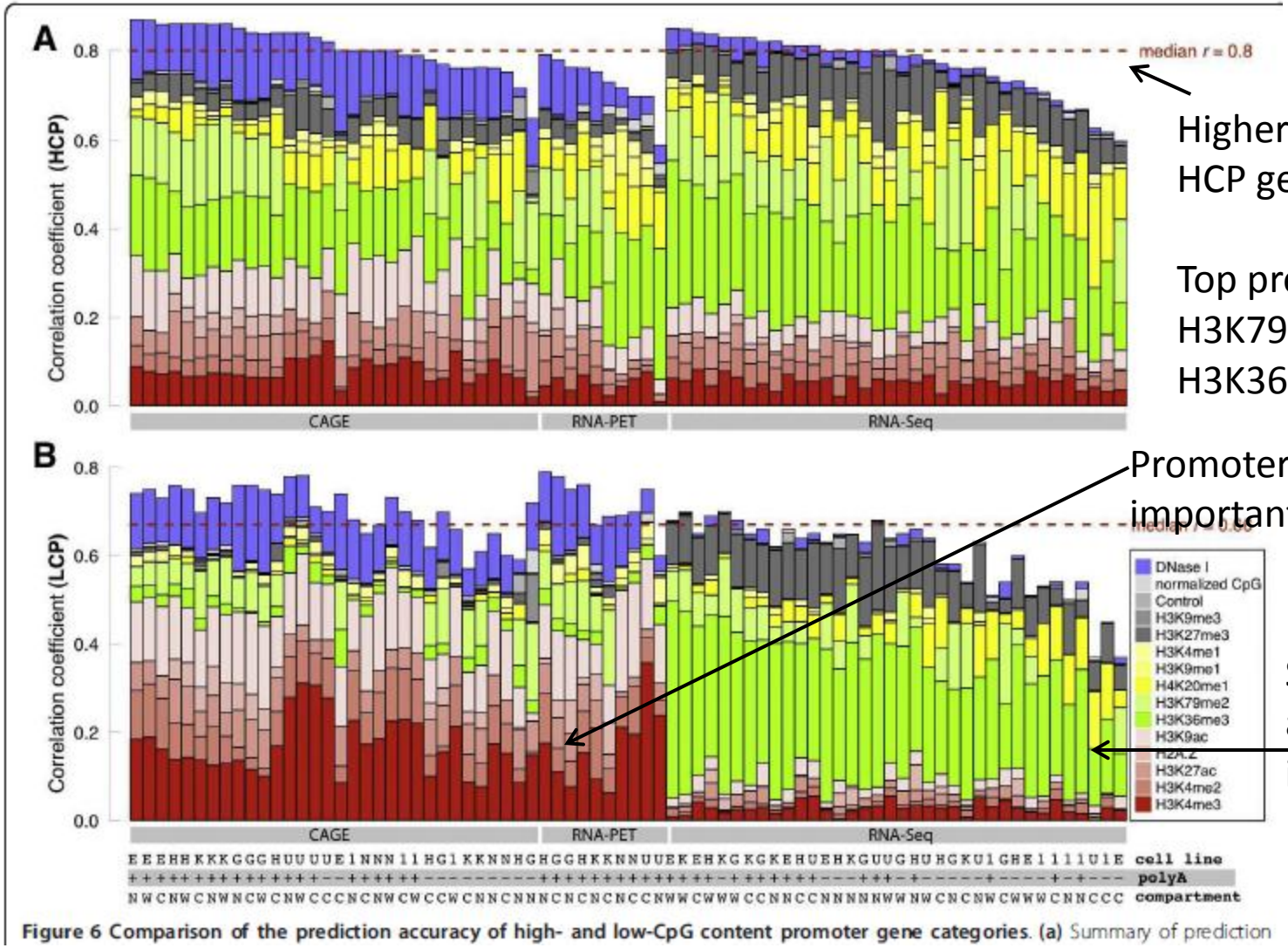
Promoter marks are most predictive

Structural marks are most predictive

Combined groups also give high prediction accuracy

Figure 5 Comparison of groups of chromatin features. Twelve chromatin features are grouped into four categories according to their known

Genes with different promoter CpG content



Higher prediction with HCP genes

Top predictors: H3K79me2 and H3K36me3

Promoter marks are more important

Structural marks are more important

Figure 6 Comparison of the prediction accuracy of high- and low-CpG content promoter gene categories. (a) Summary of prediction

Different RNA types and different cell compartments

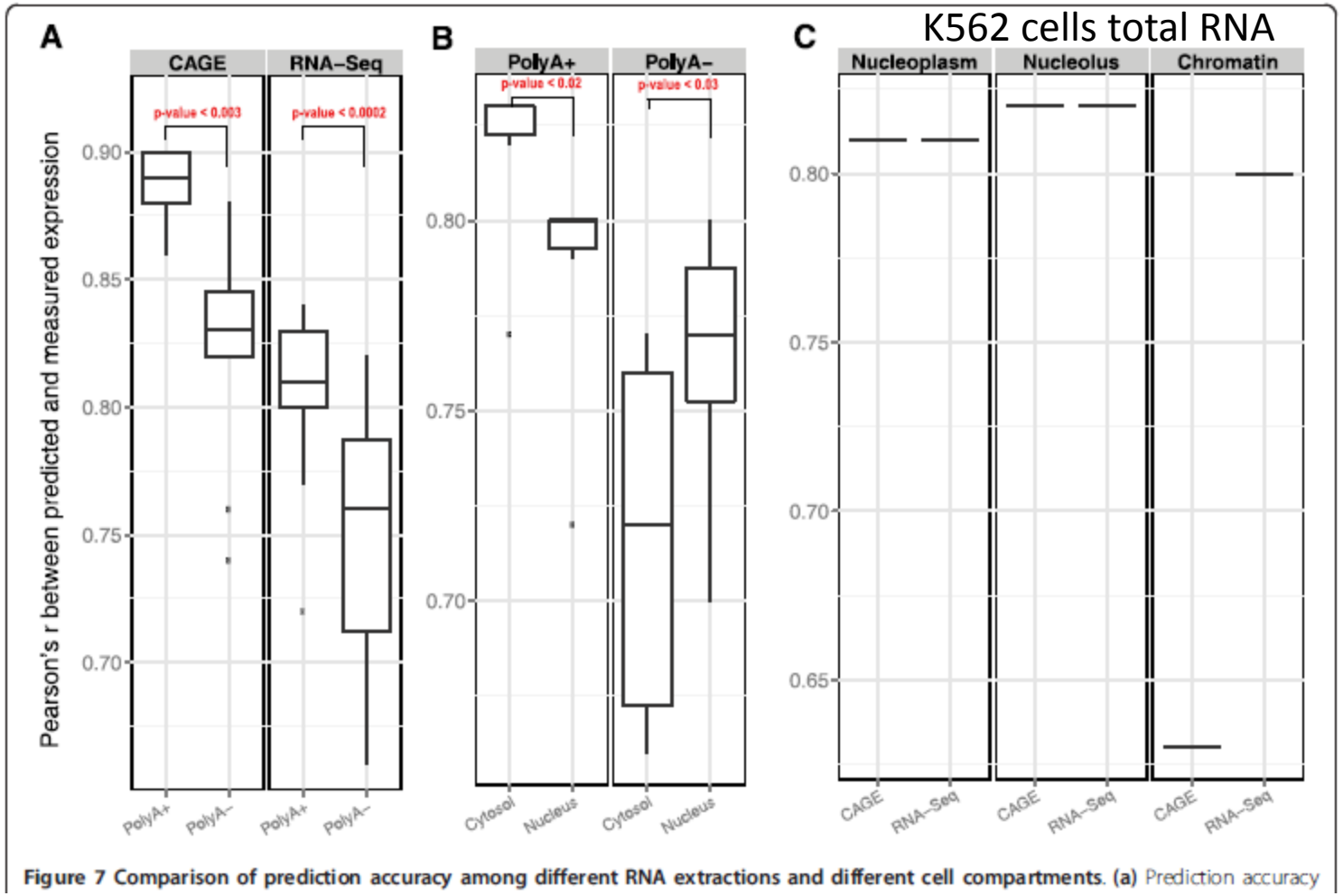


Figure 7 Comparison of prediction accuracy among different RNA extractions and different cell compartments. (a) Prediction accuracy

Summary

- Confirming pre-existing studies
  - Strong correlation between gene expression and chromatin features
  - Transcription initiation and elongation are represented by different sets of chromatin features
  - PolyA RNAs might be regulated by different mechanisms than non-PolyA RNAs
- Contributions
  - Wide range of ENCODE datasets
  - Novel two-step model
  - Model preforms well in predicting expression level

Limitations

- Histone modifications is a dynamic process; chromatin features may work combinatorially
  - Interaction terms rather than grouping
- Multiple transcripts and differential chromatin regulation
- Genes with zero expression or repressed
- Only transcripts longer that 4100 bp