Peaks

DNAse I-seq

CHIP-seq

SIGNAL PROCESSING FOR NEXT-GEN SEQUENCING DATA

Gene models

RNA-seq

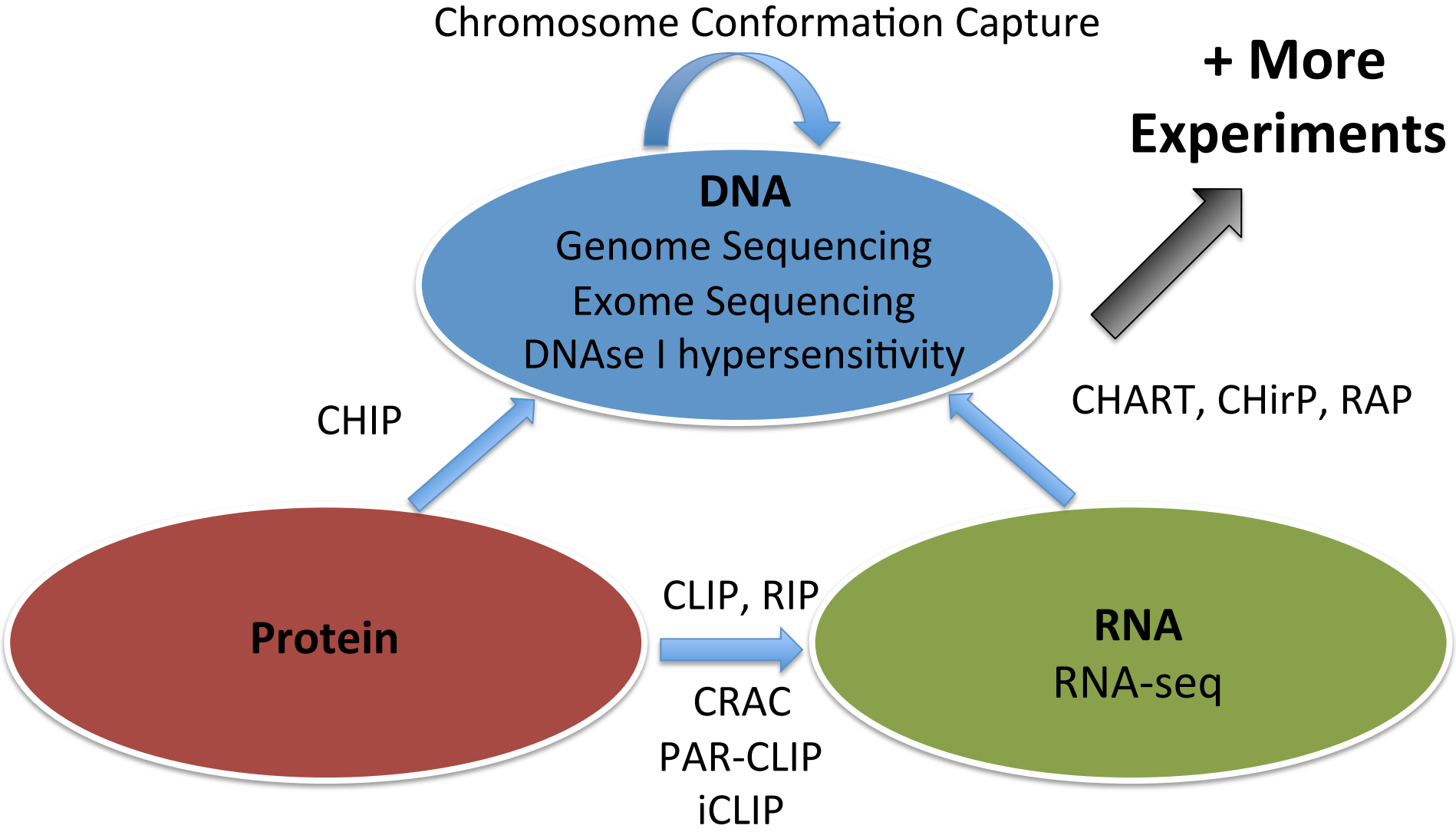
RIP/CLIP-seq

Binding sites

Transcripts

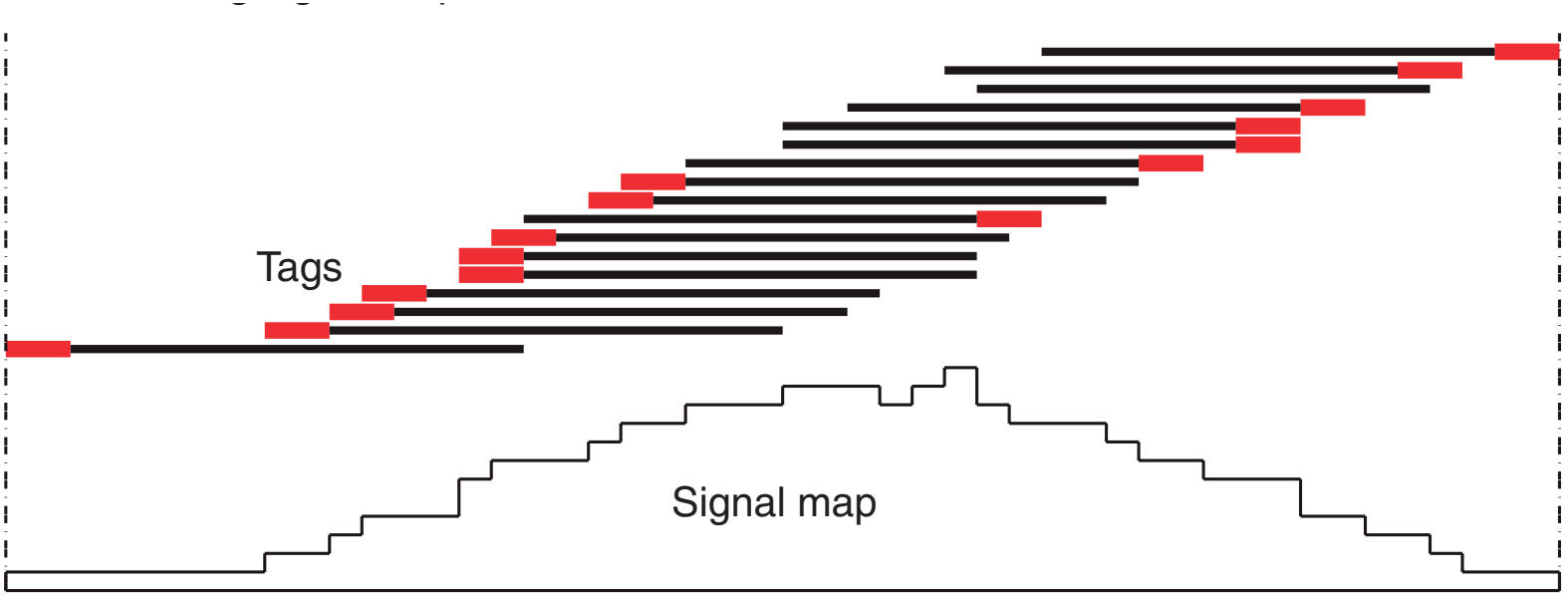
FAIRE-seq

The Power of Next-Gen Sequencing



For more Seq technologies, see: http://liorpachter.wordpress.com/seq/

Next-Gen Sequencing as Signal Data

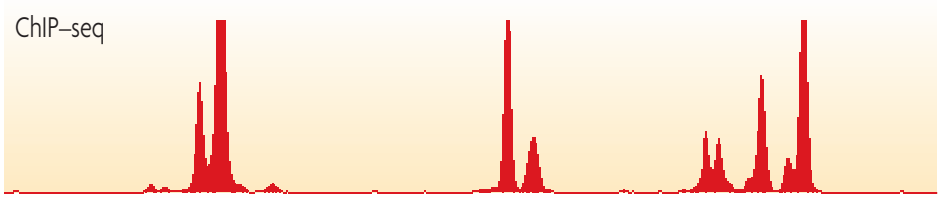


- ✓ Map reads (red) to the genome. Whole pieces of DNA are black.
- ✓ Count # of reads mapping to each DNA base → signal

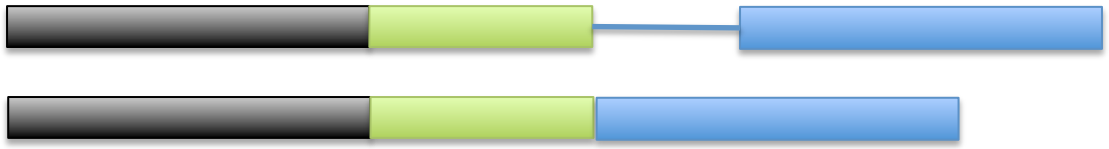
Rozowsky *et al*. 2009 *Nature Biotech*

Outline

- **Read mapping:** Creating signal map
- Finding **enriched regions**
  - **CHIP-seq**: peaks of protein binding



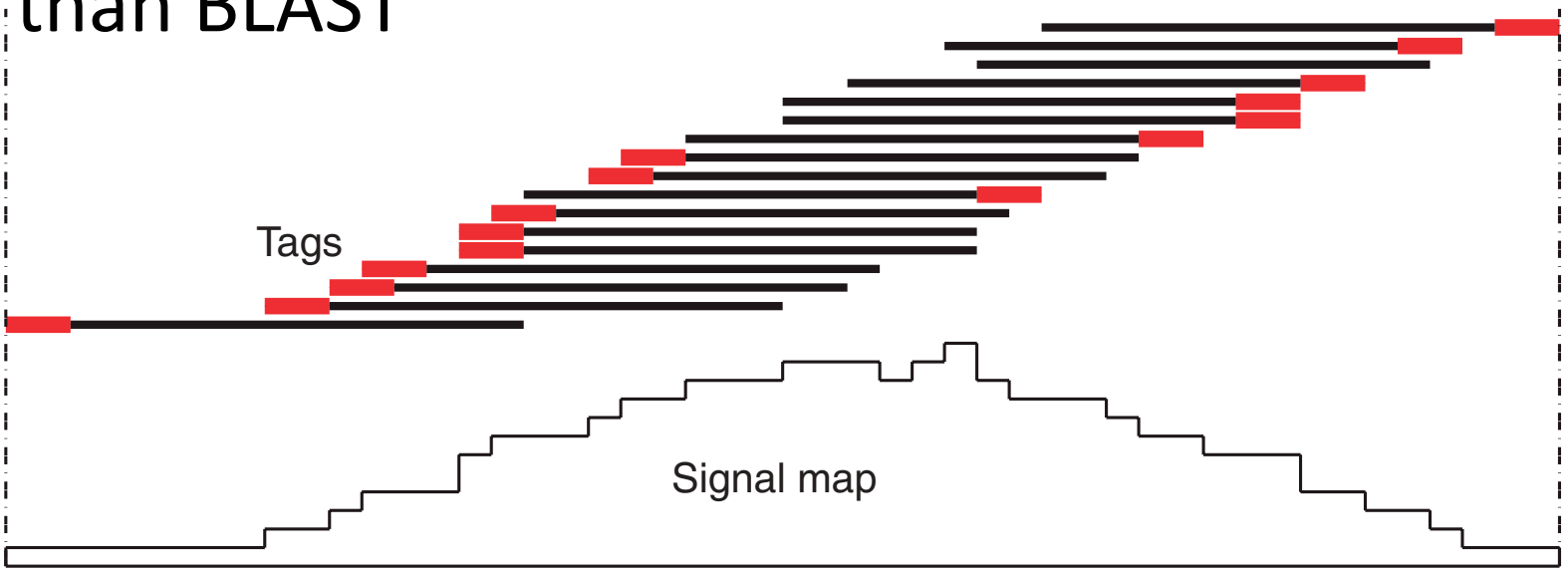
- **RNA-seq**: from enrichment to transcript quantification



- Application: **Predicting gene expression** from transcription factor and histone modification binding

Read mapping

- Problem: match up to a **billion** short sequence reads to the genome
- Need sequence alignment algorithm faster than BLAST



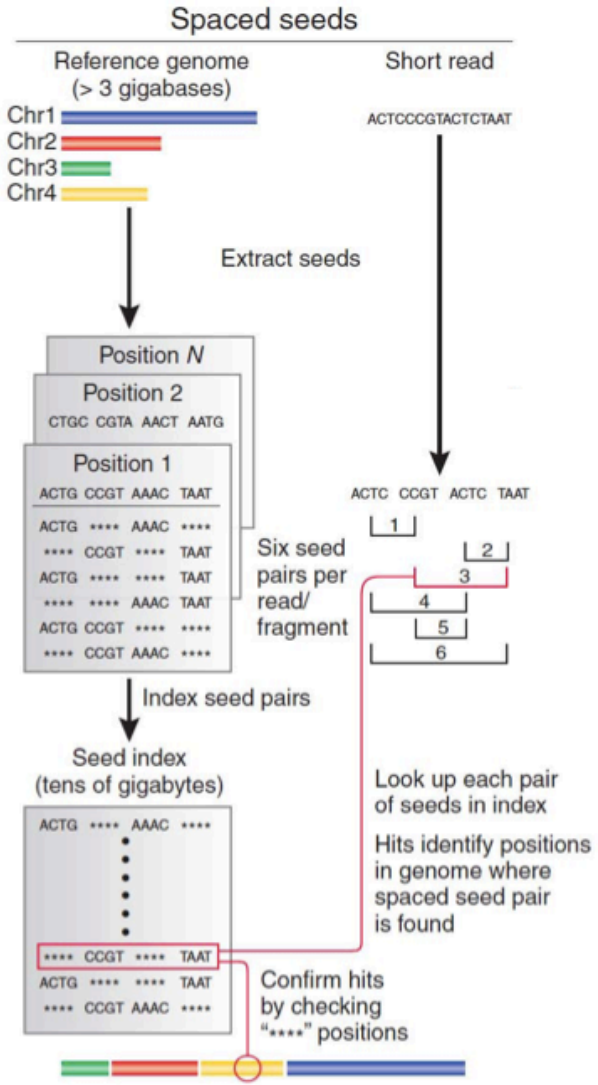
Rozowsky *et al*. 2009 *Nature Biotech*

Read mapping (sequence alignment)

- Dynamic programming
  - Optimal, but SLOW
- BLAST
  - Searches primarily for close matches, still too slow for high throughput sequence read mapping
- Read mapping
  - Only want very close matches, must be super fast

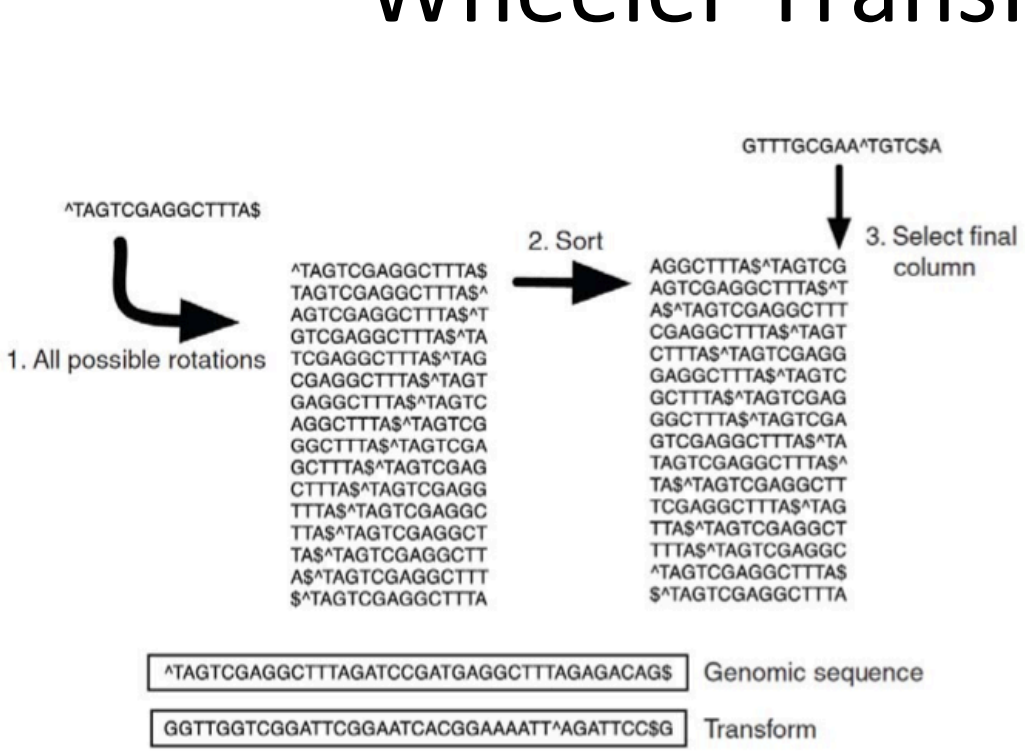
Index-based short read mappers

- Similar to BLAST
- Map all genomic locations of all possible short sequences in a hash table
- Check if read subsequences map to adjacent locations in the genome, allowing for up to 1 or 2 mismatches.
- Very memory intensive!

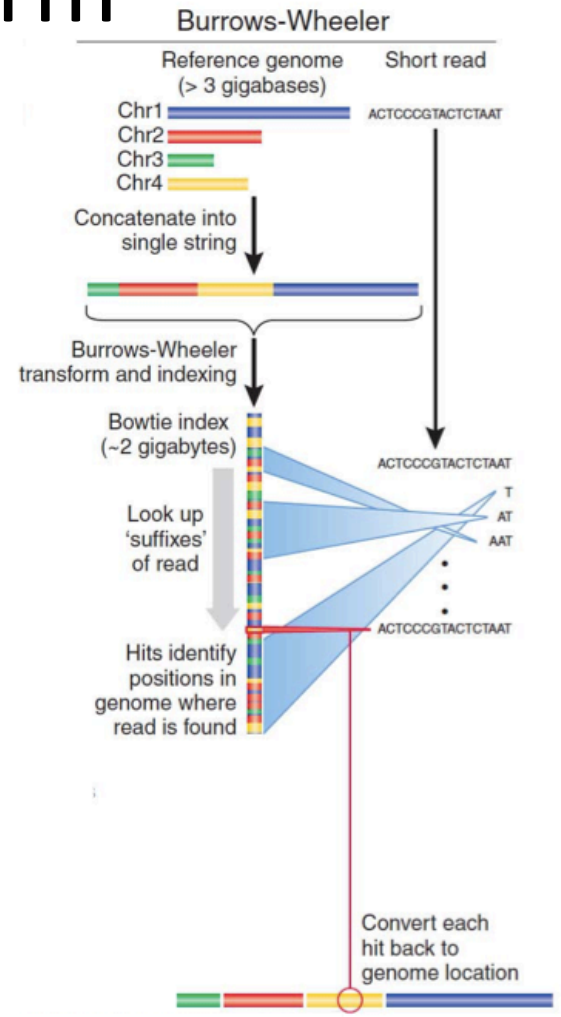


Trapnell and Salzberg 2009, Slide adapted from Ray Auerbach

Read Alignment using Burrows-Wheeler Transform



Adapted from Flicek, P. & Birney, E. Sense from sequence reads: methods for alignment and assembly. *Nat Meth* **6**, S6-S12 (2009).



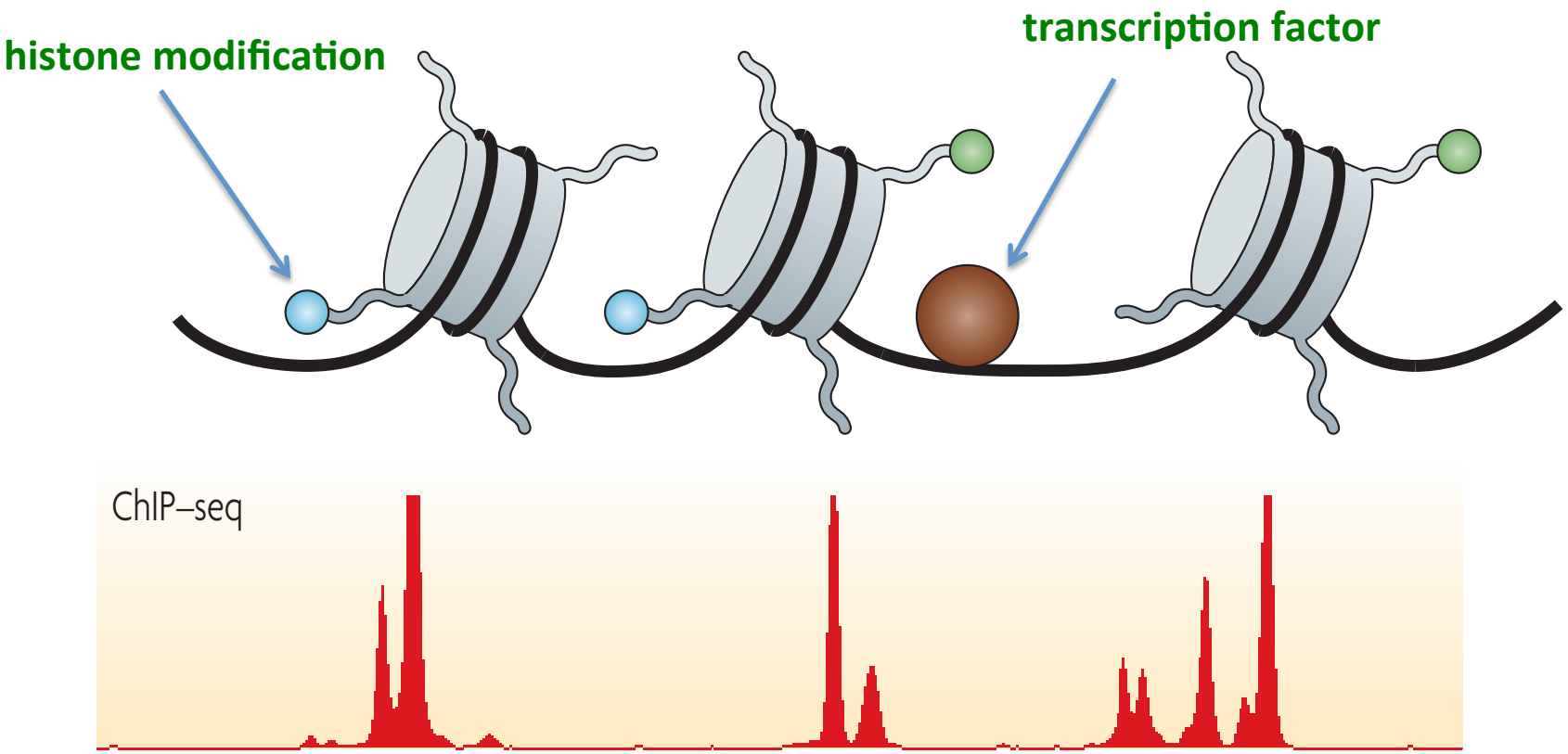
- Used in Bowtie, the current most widely used read aligner
- Described in Coursera course: Bioinformatics Algorithms (part 1, week 10)

Trapnell and Salzberg 2009, Slide adapted from Ray Auerbach

Read mapping issues

- Multiple mapping
- Unmapped reads due to sequencing errors
- VERY computationally expensive
  - Remapping data from The Cancer Genome Atlas consortium would take 6 CPU years[1]
- Current methods use heuristics, and are not 100% accurate
- These are open problems

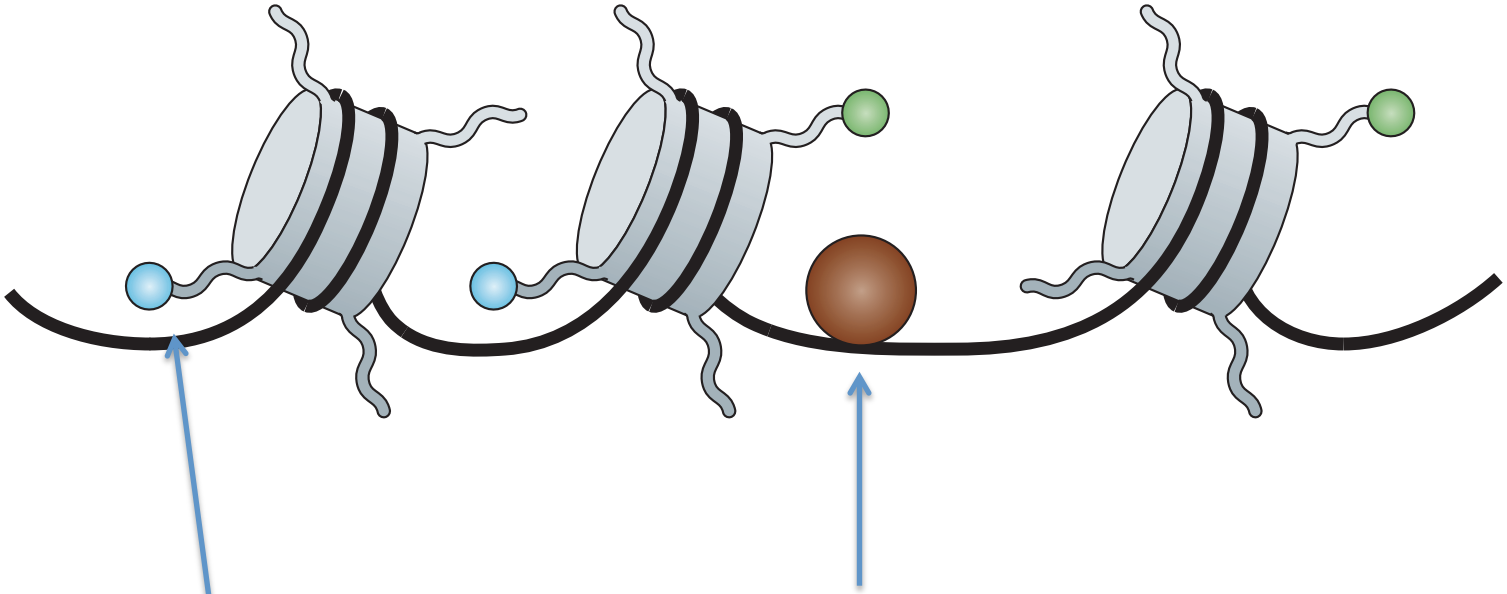
[1] https://twitter.com/markgerstein/status/396658032169742336



FINDING ENRICHED REGIONS: CHIP-SEQ DATA ANALYSIS

Park 2009 *Nature Reviews Genetics*

CHIP-seq Intro

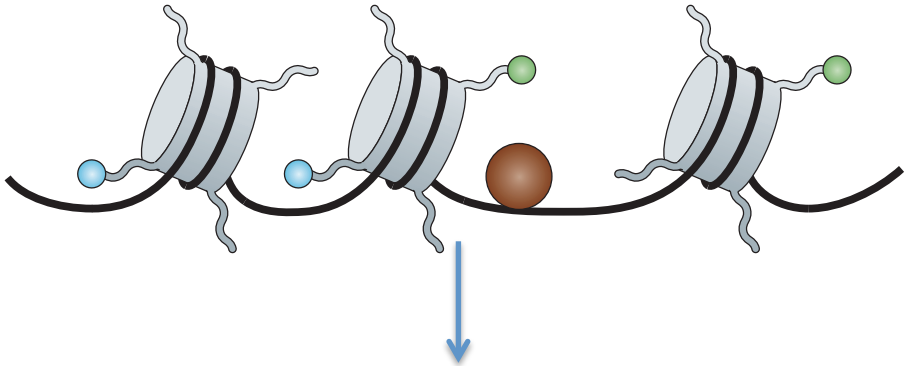


- Determine locations of **transcription factors** and **histone modifications**.
- The binding of these factors is what regulates whether genes get transcribed.

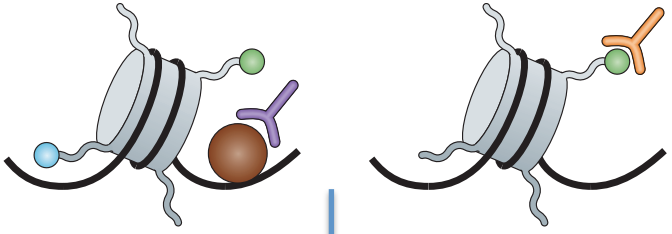
Park 2009 *Nature Reviews Genetics*

CHIP-seq protocol

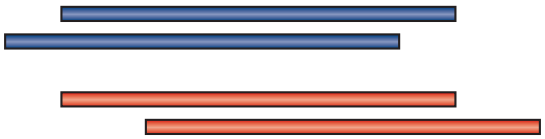
DNA bound by histones and transcription factors



Target protein of interest with Antibody

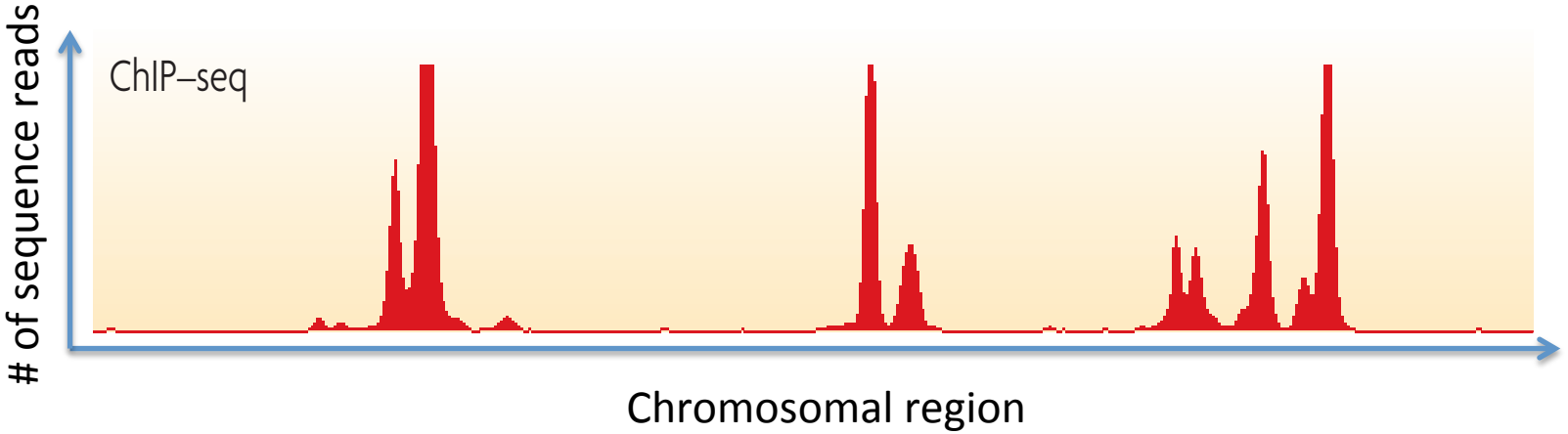


Sequence DNA bound by protein of interest



Park 2009 *Nature Reviews Genetics*

CHIP-seq Data

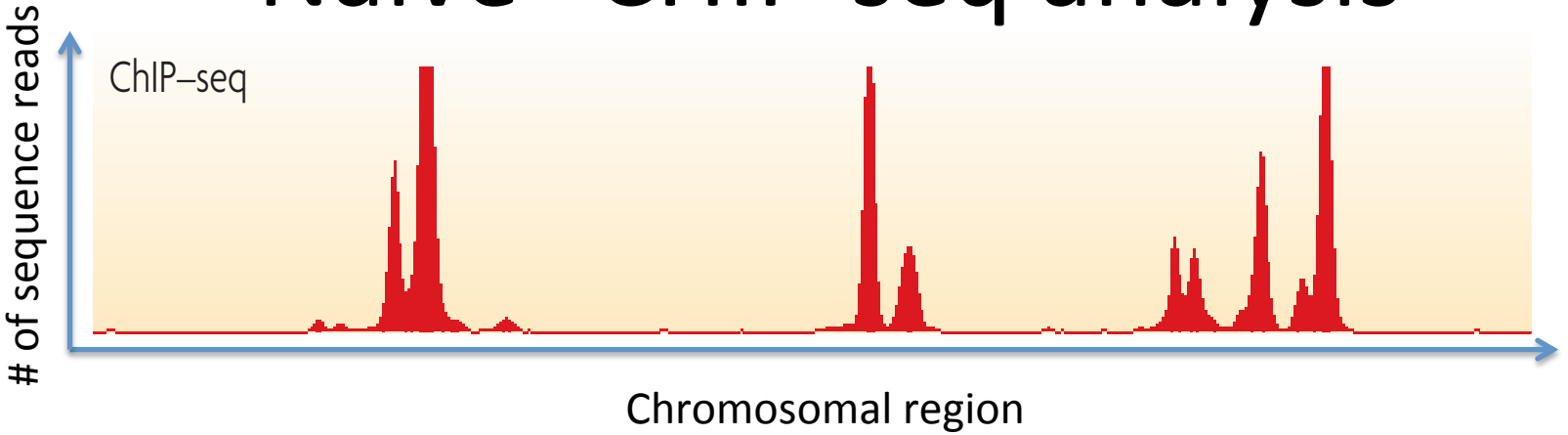


Basic interpretation: Signal map to represents binding profile of protein to DNA

How do we identify binding sites from CHIP-seq signal "peaks"?

Park 2009 *Nature Reviews Genetics*

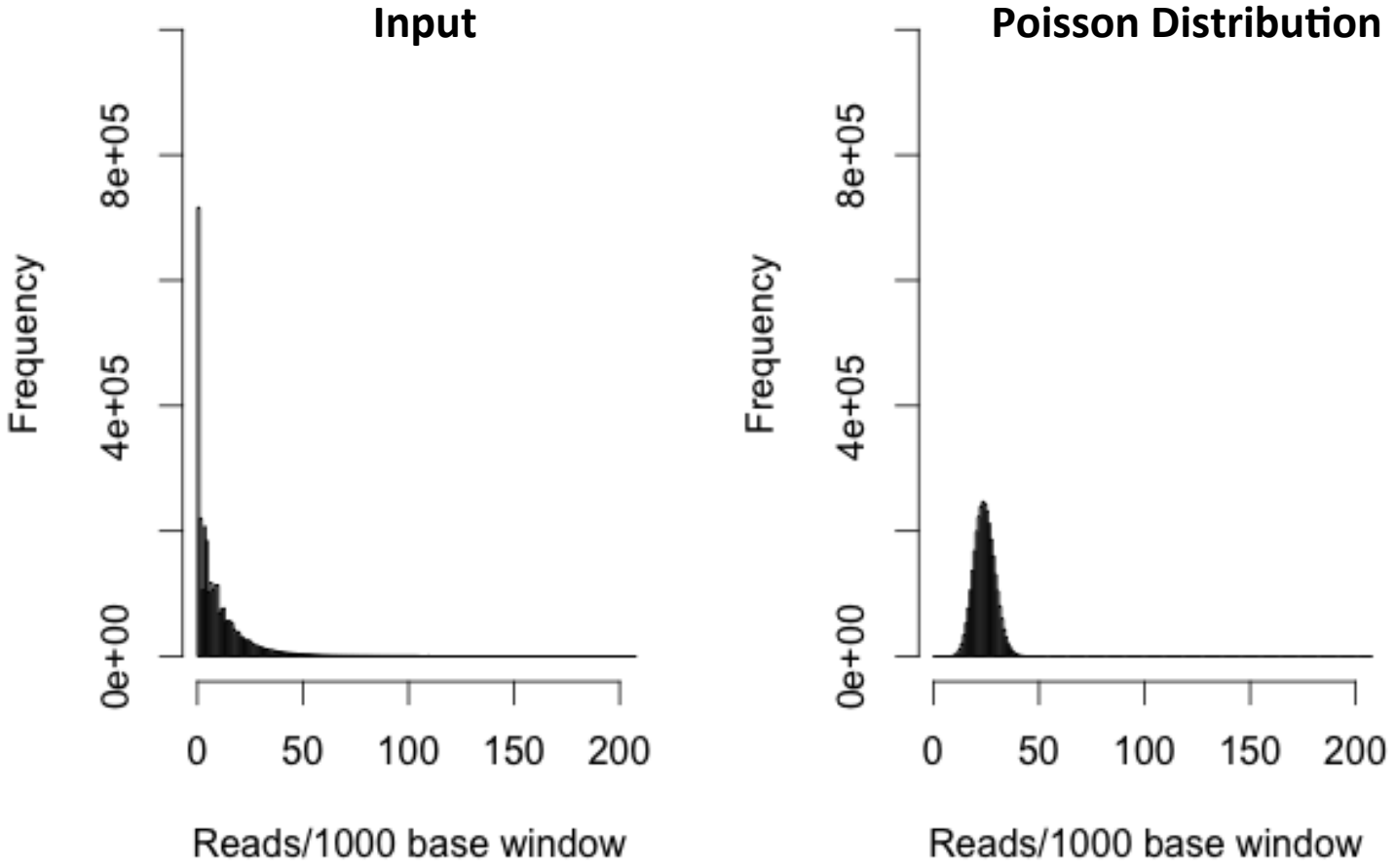
“Naïve” CHIP-seq analysis



- Background assumption: all sequence reads map to random locations within the genome
- Divide genome into bins, distribution of expected frequencies of reads/bin is described by the Poisson distribution.
- Assign p-value based on Poisson distribution for each bin based on # of reads

Rozowsky *et al*. 2009 *Nature Biotech,* Park 2009 *Nature Reviews Genetics*

Is a Poisson background reasonable for CHIP-seq data?

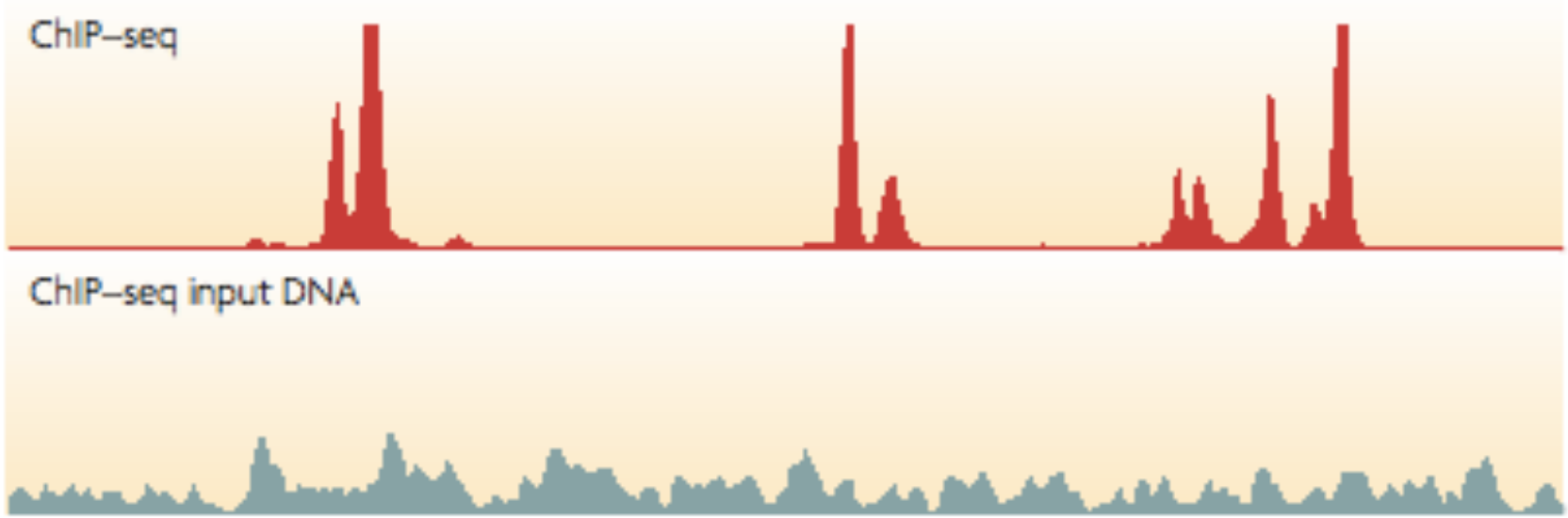


- "Input" is from a CHIP-seq experiment using an antibody for a non-DNA binding protein

ENCODE NF-Kb CHIP-seq data

Is a Poisson background reasonable for CHIP-seq data?

- "Input" experiment: Do CHIP-seq using an antibody for a protein that doesn't bind DNA



- There are also "peaks" in the input!

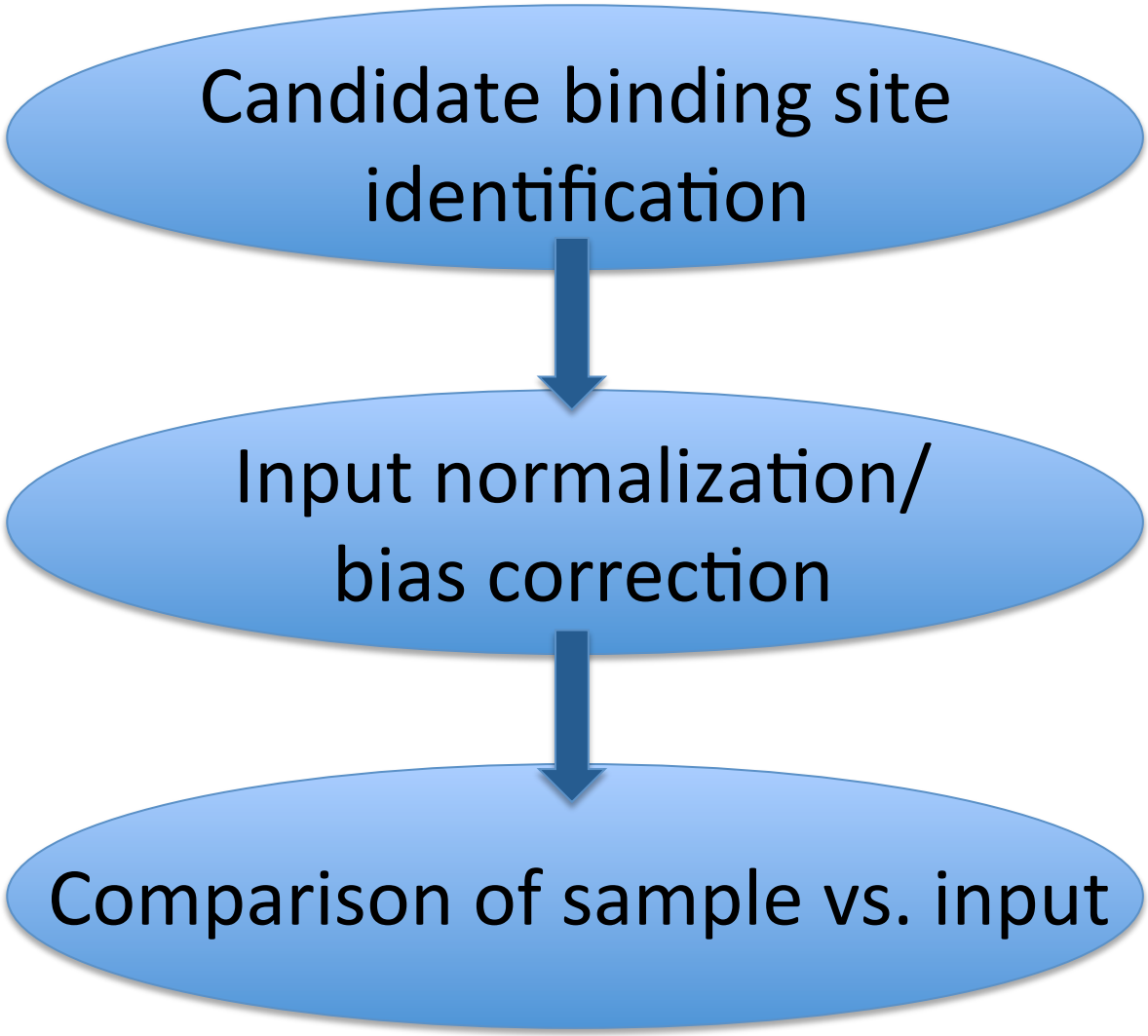
Park 2009 *Nature Reviews Genetics*

Peakseq

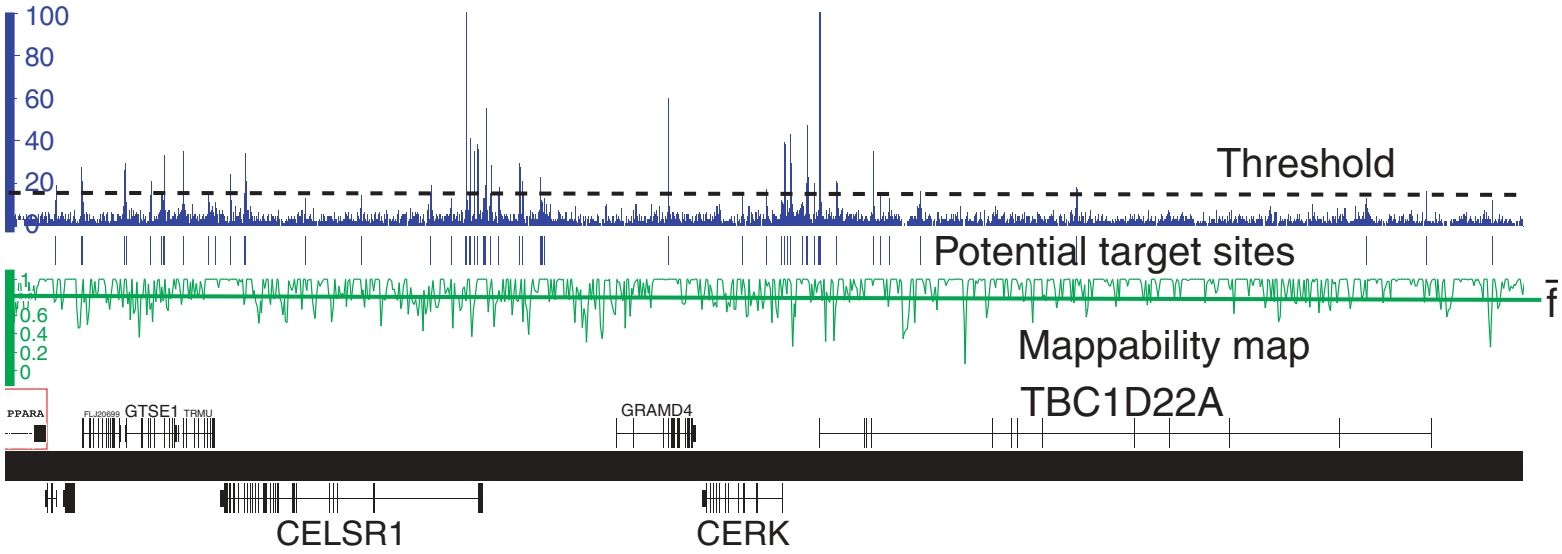
Rozowsky *et al*. 2009 *Nature Biotech*

Gerstein Lab

Determining protein binding sites by comparing CHIP-seq data with input



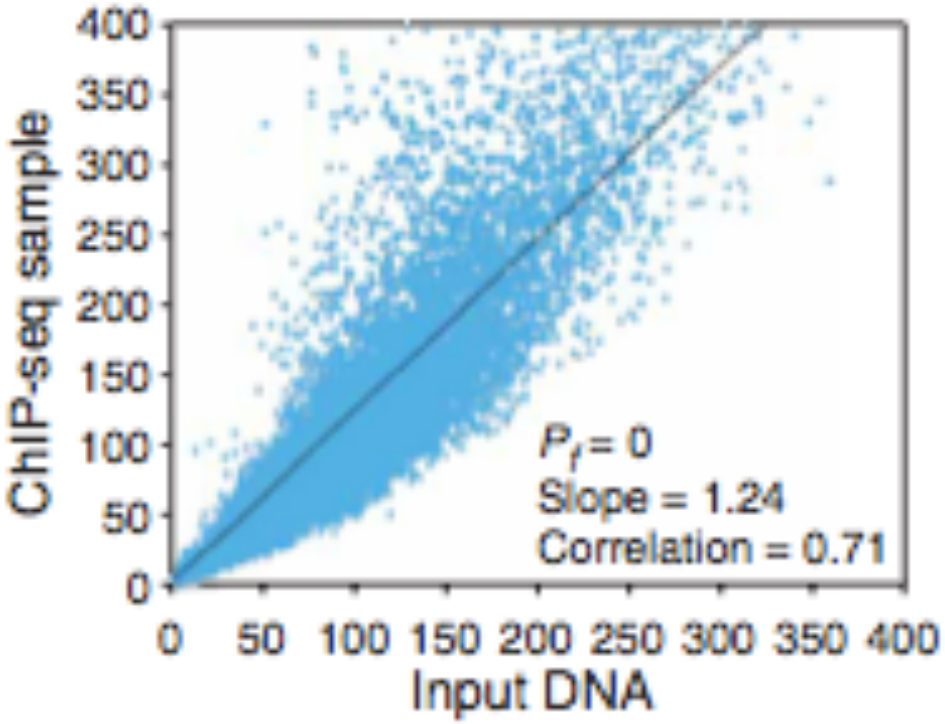
Candidate binding site identification



- Use Poisson distribution as background, as in the "naïve" analysis discussed earlier
- Normalize read counts for mappability (uniqueness) of genomic regions
- Use large bin size, finer resolution analysis later

Rozowsky *et al*. 2009 *Nature Biotech*

Input normalization



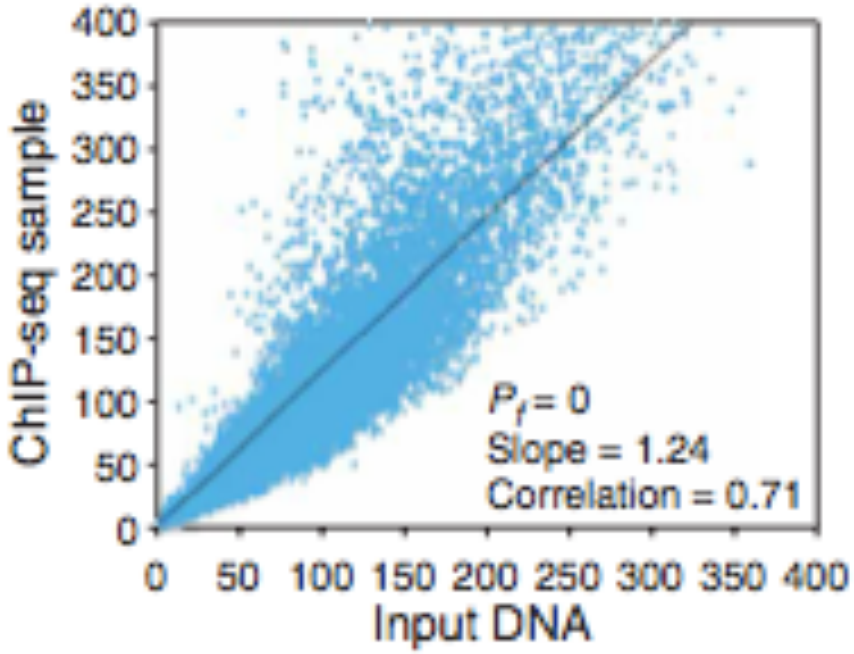
✓Normalize based on slope of least squares regression line.

Normalized reads = CHIP-seq reads/(slope*input reads)

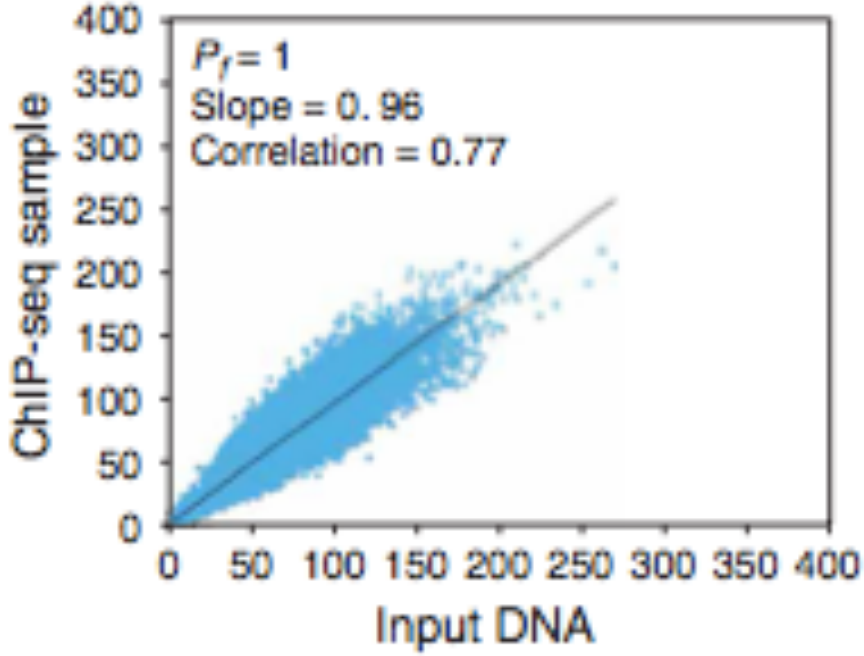
Rozowsky *et al*. 2009 *Nature Biotech*

Input normalization

All data points



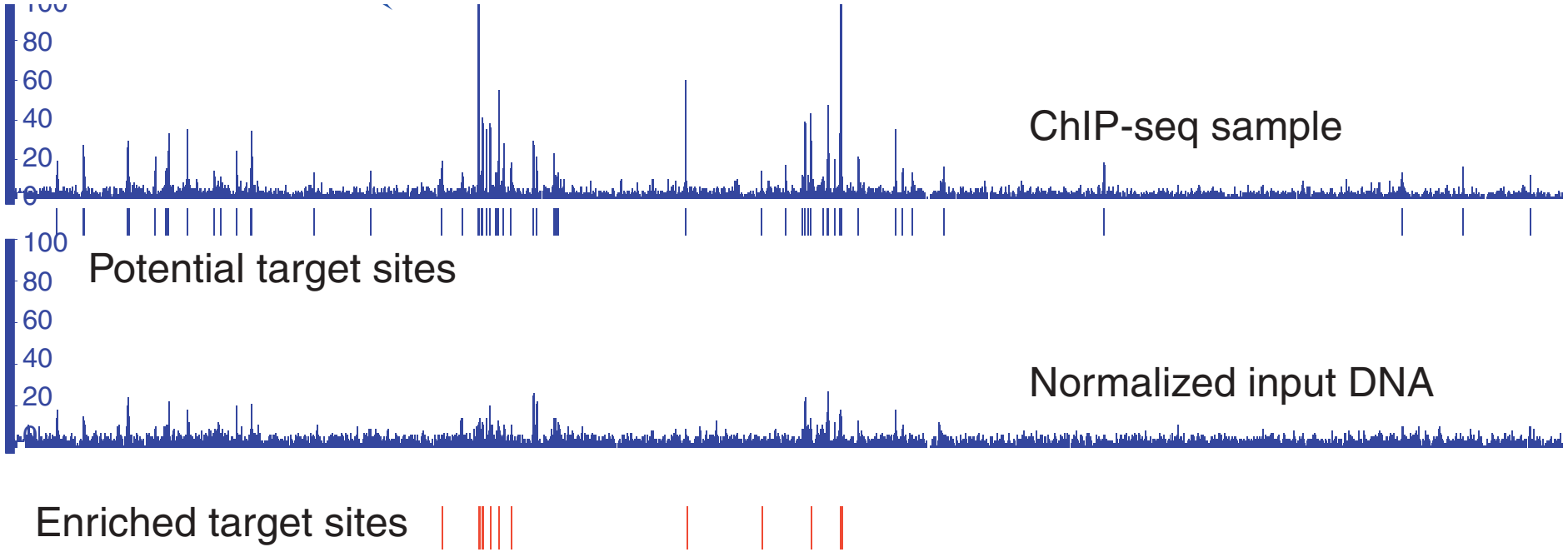
Candidate peaks removed



✓Using regression based on all data points (including candidate peaks) is overly conservative.

Rozowsky *et al*. 2009 *Nature Biotech*

Calling peaks vs. input



- Binomial distribution
  - Each genomic region is like a coin
  - The combined number of reads is the # of times that the coin is flipped
  - Look for regions that are "weighted" toward sample, not input

ENCODE NF-Kb CHIP-seq data

Multiple Hypothesis Correction

- Millions of genomic bins → expect many bins with p-value < 0.05!
- How do we correct for this?

Multiple Hypothesis Correction

- Bonferroni Correction
  - Multiply p-value by number of observations
  - Adjusts p-values → expect up to 1 false positive
  - Very conservative

Multiple Hypothesis Correction

- False discovery rate (FDR)
  - Expected number of false positives as a percentage of the total rejected null hypotheses
  - Expectation[false positives/(false positives+true postives)]
- q-value: maximum FDR at which null hypothesis is rejected.
- Benjamini-Hochberg Correction
  - q-value = p-value*# of tests/rank

Is PeakSeq an optimal algorithm?

Many other CHIP-seq “peak”-callers

Program	Reference	Version	Graphical user interface?	Window-based scan	Tag clustering	Gaussian kernel density estimator	Strand-specific scoring	Peak height or fold enrichment (FE)	Background subtraction	Compensates for genomic duplications or deletions	False Discovery Rate	Compare to normalized control data (FE)	Compare to statistical model fitted with control data	Statistical model or test
CisGenome	28	1.1	X*	X				X	X		X		X	conditional binomial model
Minimal ChipSeq Peak Finder	16	2.0.1			X			X				X		
E-RANGE	27	3.1			X			X				X	X	chromsome scale Poisson dist.
MACS	13	1.3.5		X				X			X		X	local Poisson dist.
QuEST	14	2.3				X		X			X**		X	chromsome scale Poisson dist.
HPeak	29	1.1		X				X					X	Hidden Markov Model
Sole-Search	23	1	X	X				X		X			X	One sample t-test
PeakSeq	21	1.01			X			X					X	conditional binomial model
SISSRS	32	1.4		X			X					X		
spp package (wtd & mtc)	31	1.7		X			X		X	X'	X			
				Generating density profiles			Peak assignment		Adjustments w. control data		Significance relative to control data			

X* = Windows-only GUI or cross-platform command line interface

X** = optional if sufficient data is available to split control data

X' = method excludes putative duplicated regions, no treatment of deletions

Wilbanks EG, Facciotti MT (2010) Evaluation of Algorithm Performance in ChIP-Seq Peak Detection. PLoS ONE 5(7): e11471. doi:10.1371/journal.pone.0011471
http://www.plosone.org/article/info:doi/10.1371/journal.pone.0011471

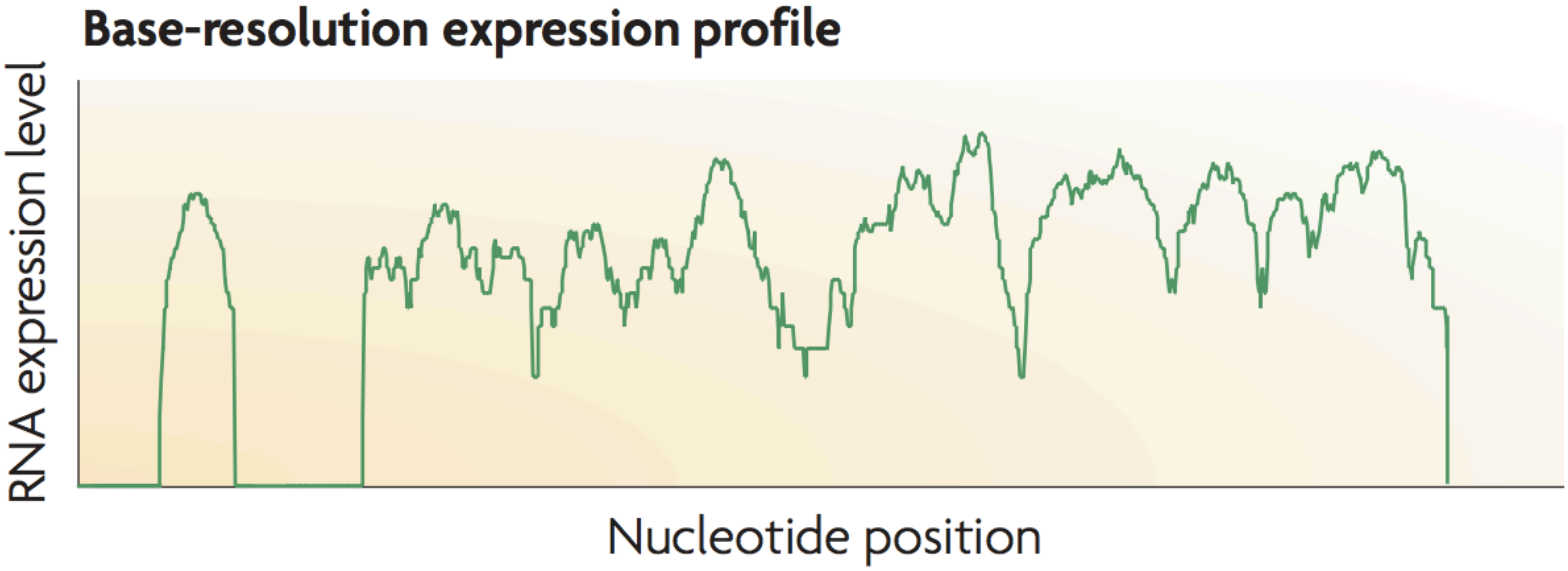
CHIP-seq summary

- Method to determine **DNA binding sites** of **transcription factors** or locations of **histone modifications**
- Must normalize sequence reads to experimental input
- Search for signal enrichment to find **peaks**
  - Peakseq: binomial test + Benjamini-Hochberg correction
  - Many other methods

RNA-SEQ: GOING BEYOND ENRICHMENT

RNA-seq

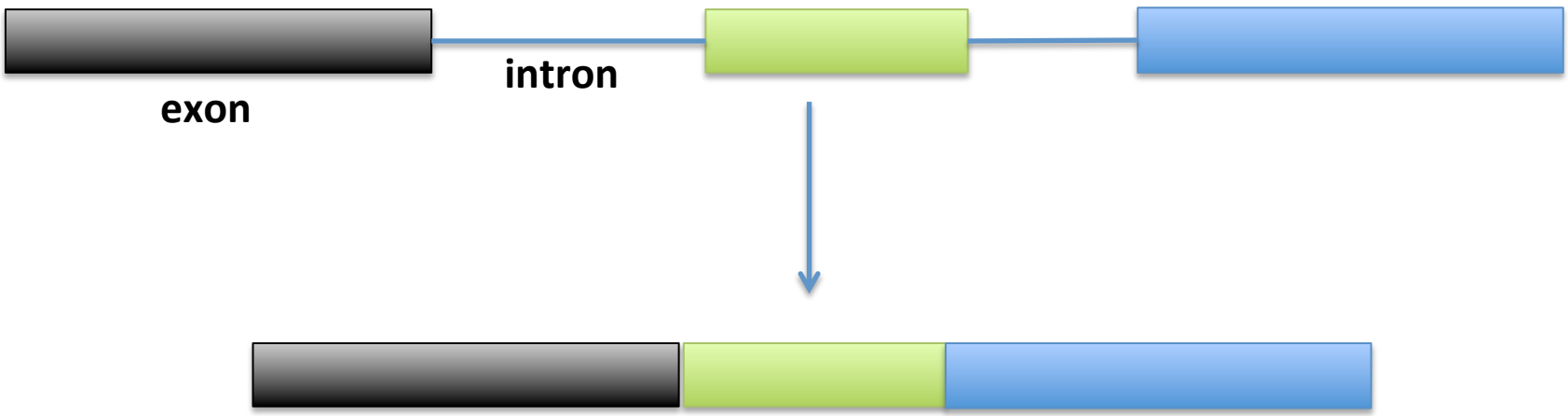
- Searching for “peaks” not enough:



- Are these “peaks” part of the same RNA molecule?
- How much of the RNA is really there?

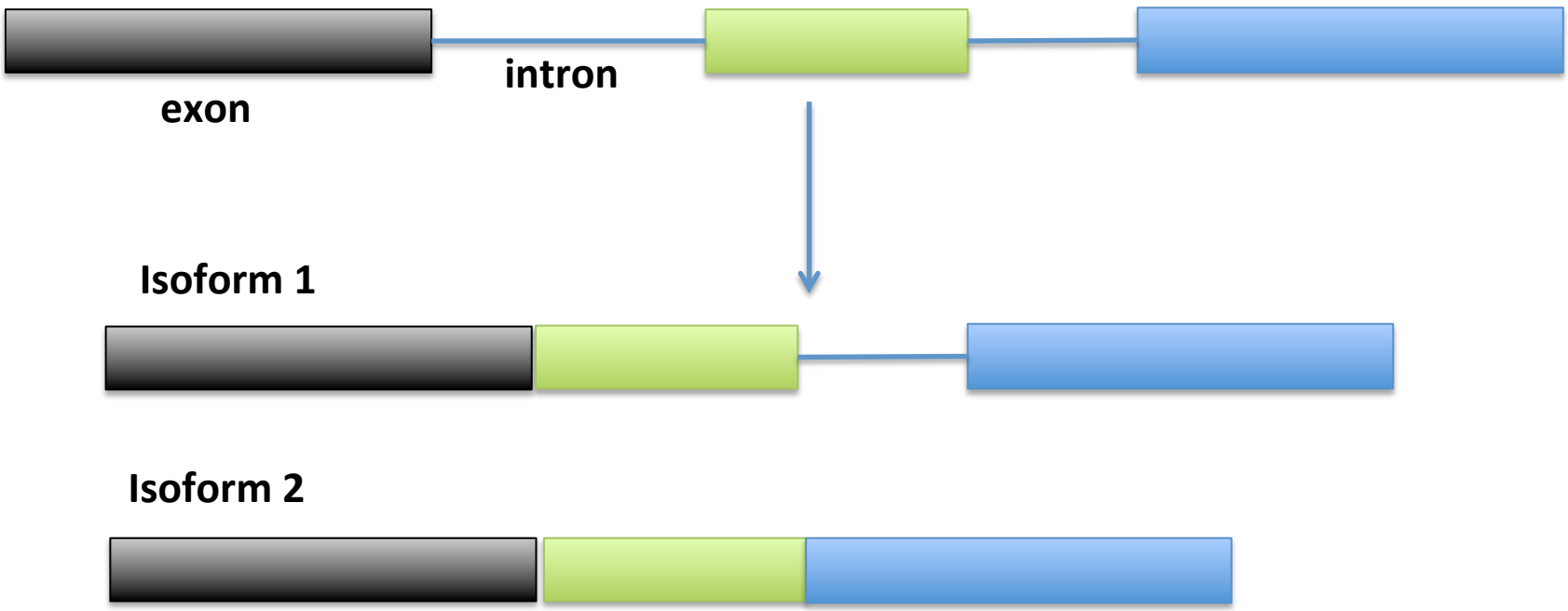
Wang *et al Nature Reviews Genetics* 2009

Background: RNA splicing



- **pre-mRNA** must have **introns** *spliced* out before being *translated* into **protein**.
- The components that are retained in the mature **mRNA** are called **exons**

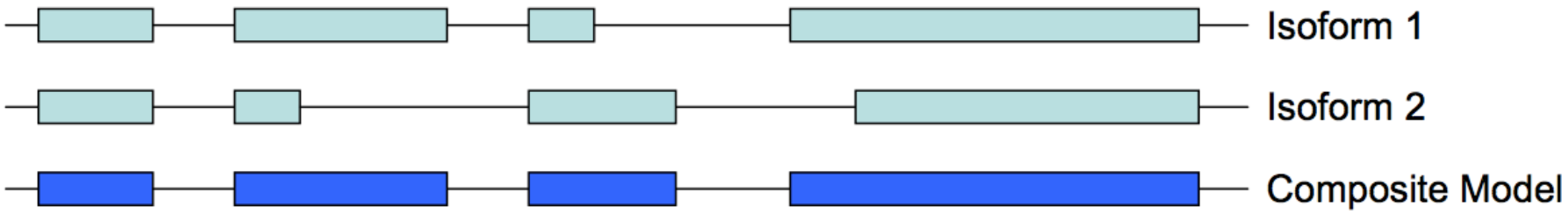
Background: alternative splicing



- *Alternative splicing* leads to creation of multiple RNA **isoforms**, with different component exons.
- Sometimes, **exons** can be *retained*, or **introns** can be *skipped*.

Simple quantification

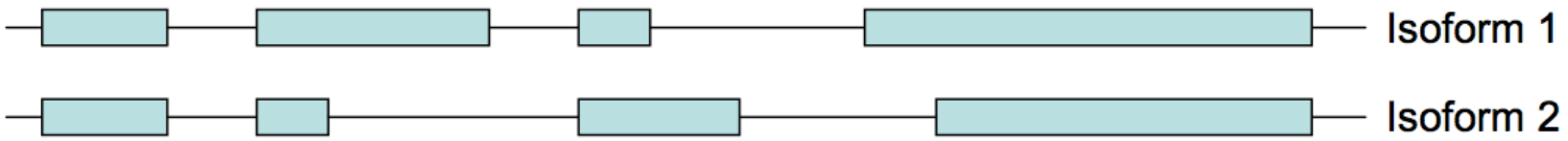
- Count reads overlapping annotations of known genes
- Simplest method: Make composite model of all isoforms of gene



- Quantification: Reads per kilobase per million reads (RPKM)

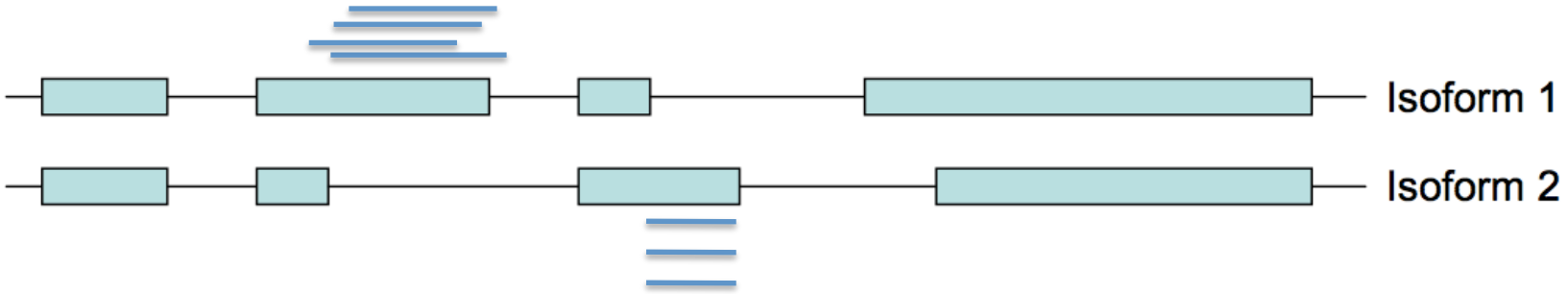
Isoform Quantification

- Map reads to genome
- How do we assign reads to overlapping transcripts?



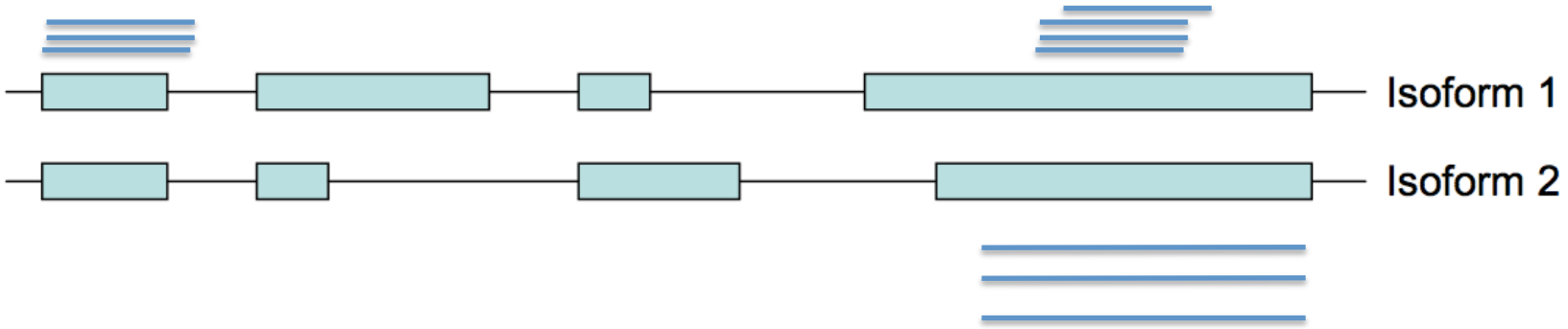
Isoform Quantification

- Simple method: only consider unique reads



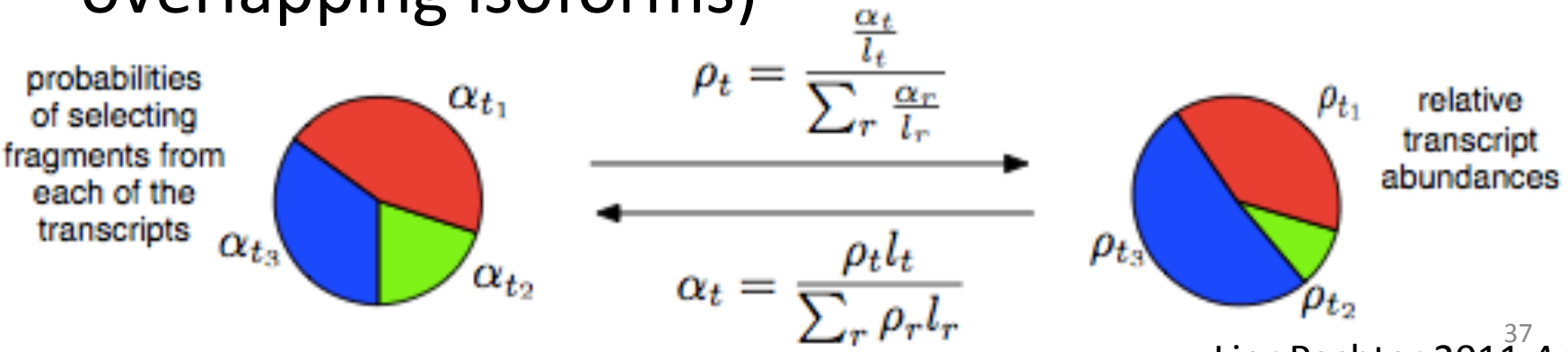
Isoform Quantification

- Simple method: only consider unique reads
- Problem: what about the rest of the data?



Expectation Maximization Algorithm

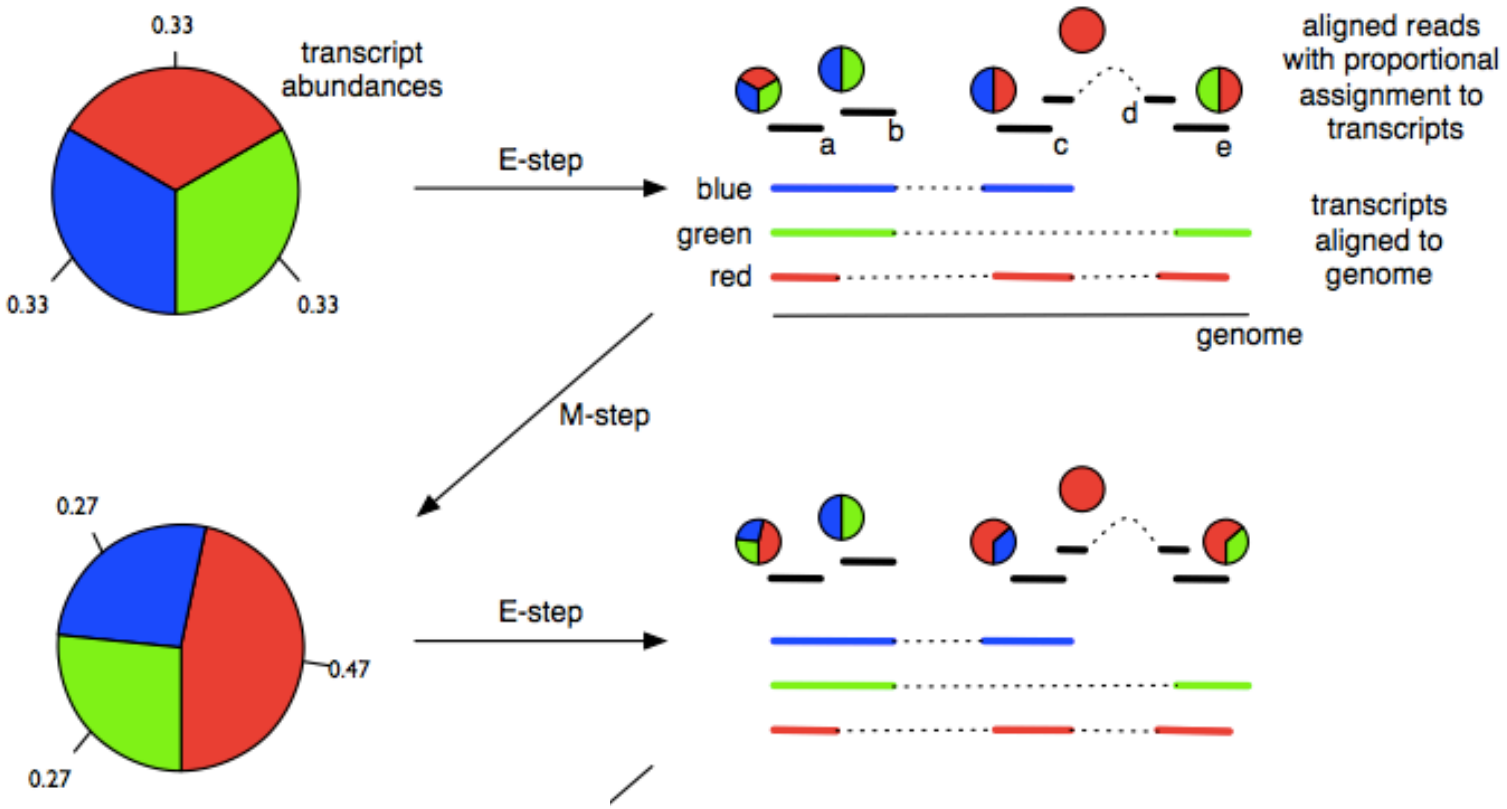
- Assign reads to isoforms to **maximize likelihood** of generating total pattern of observed reads.
- 0. **Initialize** (expectation): Assign reads randomly to isoforms based on naïve (length normalized) probability of the read coming from that isoform (as opposed to other overlapping isoforms)



Lior Pachter 2011 *ArXiv*

Expectation Maximization Algorithm

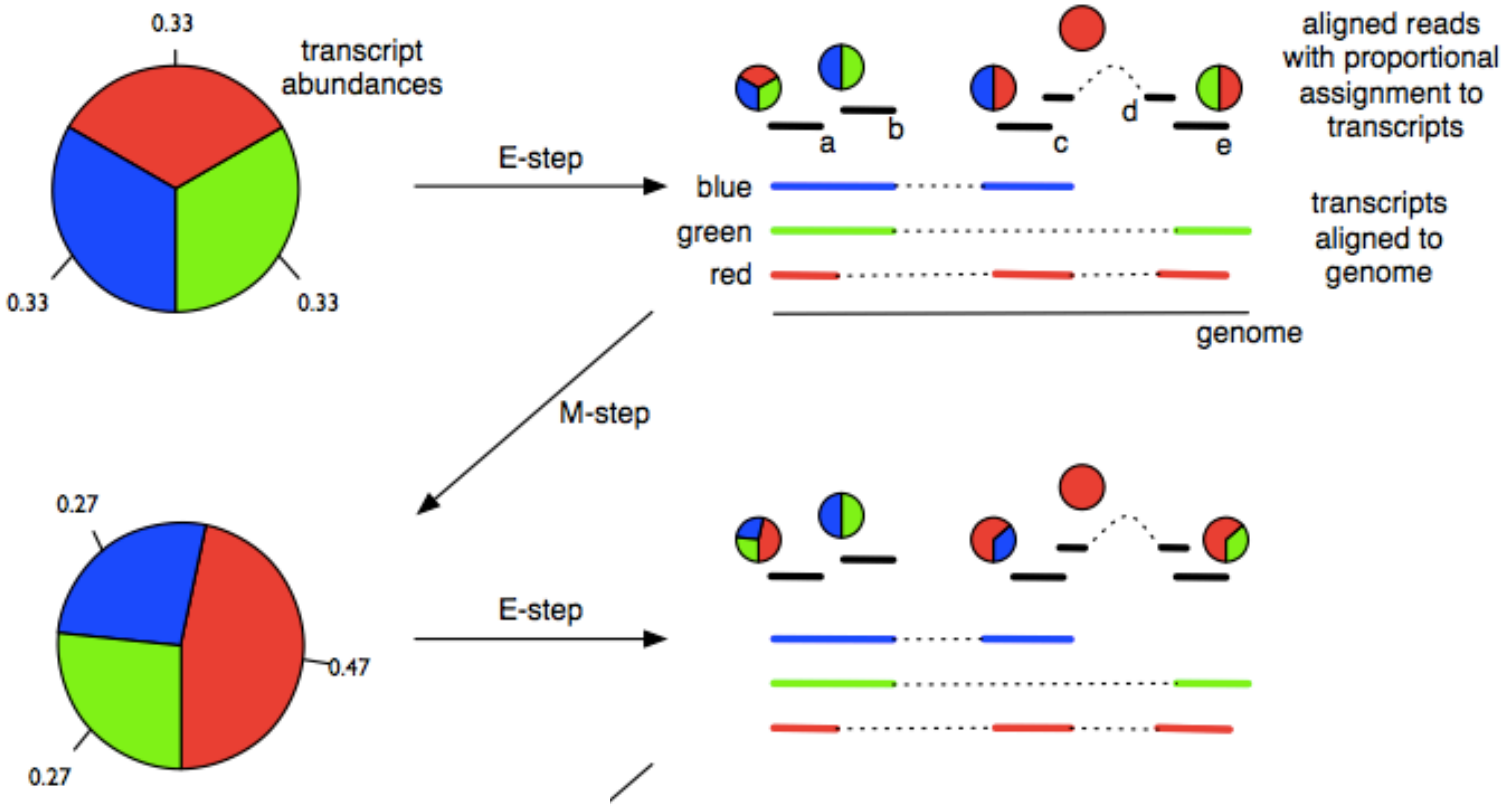
- 1. **Maximization:** Choose transcript abundances that maximize likelihood of the read distribution (Maximization).



Lior Pachter 2011 *ArXiv*

Expectation Maximization Algorithm

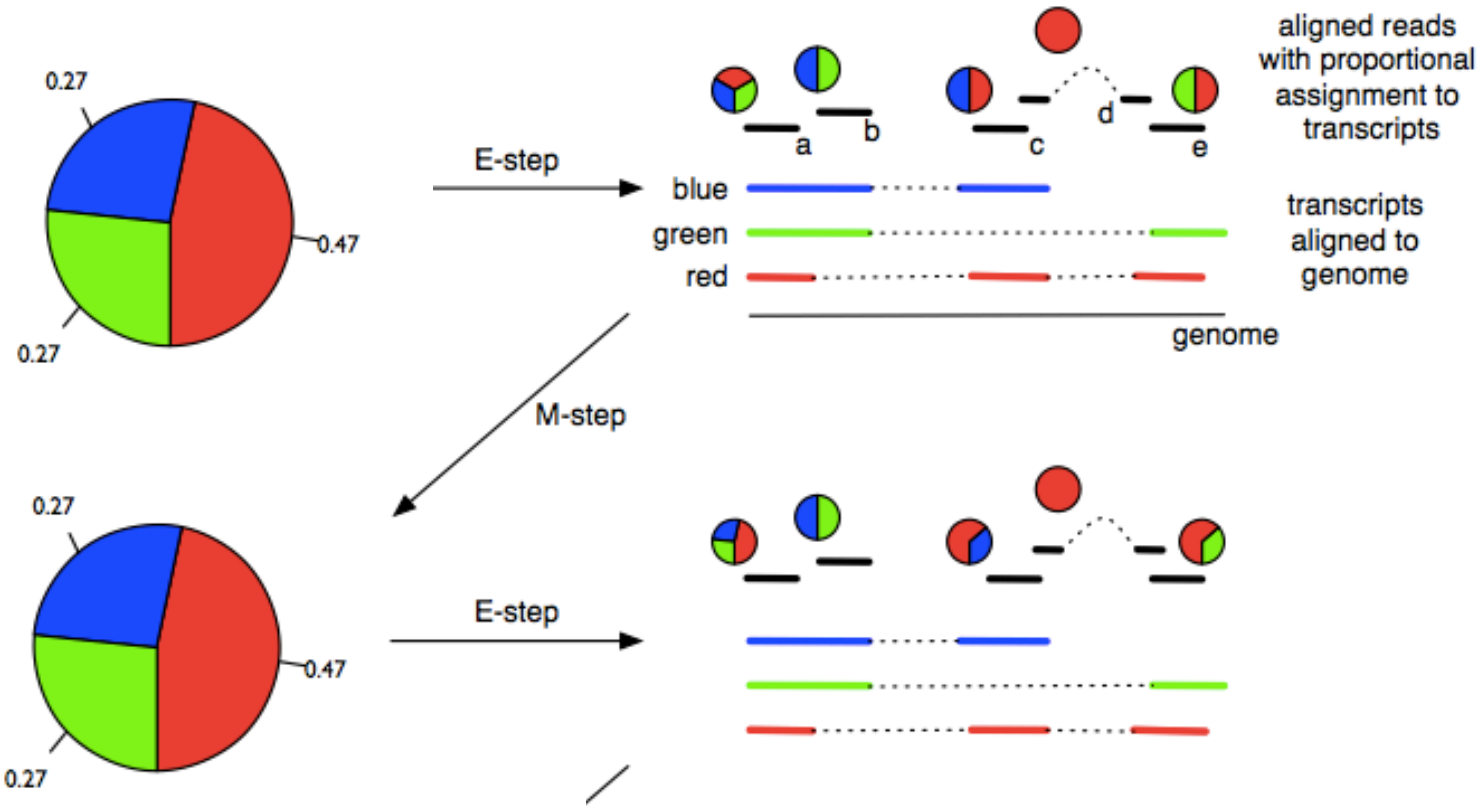
- 2. **Expectation:** Reassign reads based on the new values for the relative quantities of the isoforms.



Lior Pachter 2011 *ArXiv*

Expectation Maximization Algorithm

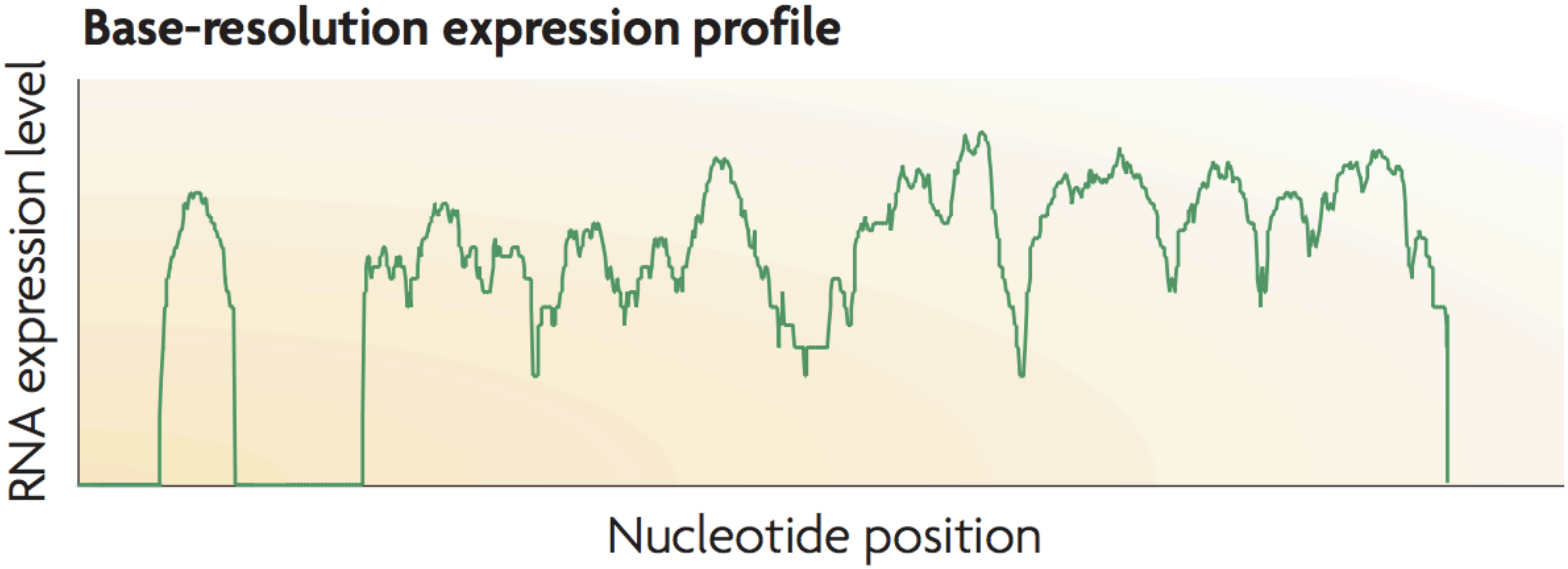
- 3. Continue **expectation** and **maximization** steps until isoform quantifications converge (it is a mathematical fact that this will happen).



Lior Pachter 2011 *ArXiv*

Detecting new transcripts

- Search for "peaks":



- Reads that overlap splice junctions→peaks part of same transcript
- Special sequencing techniques to find ends of transcripts
- **Still a major open area of research**

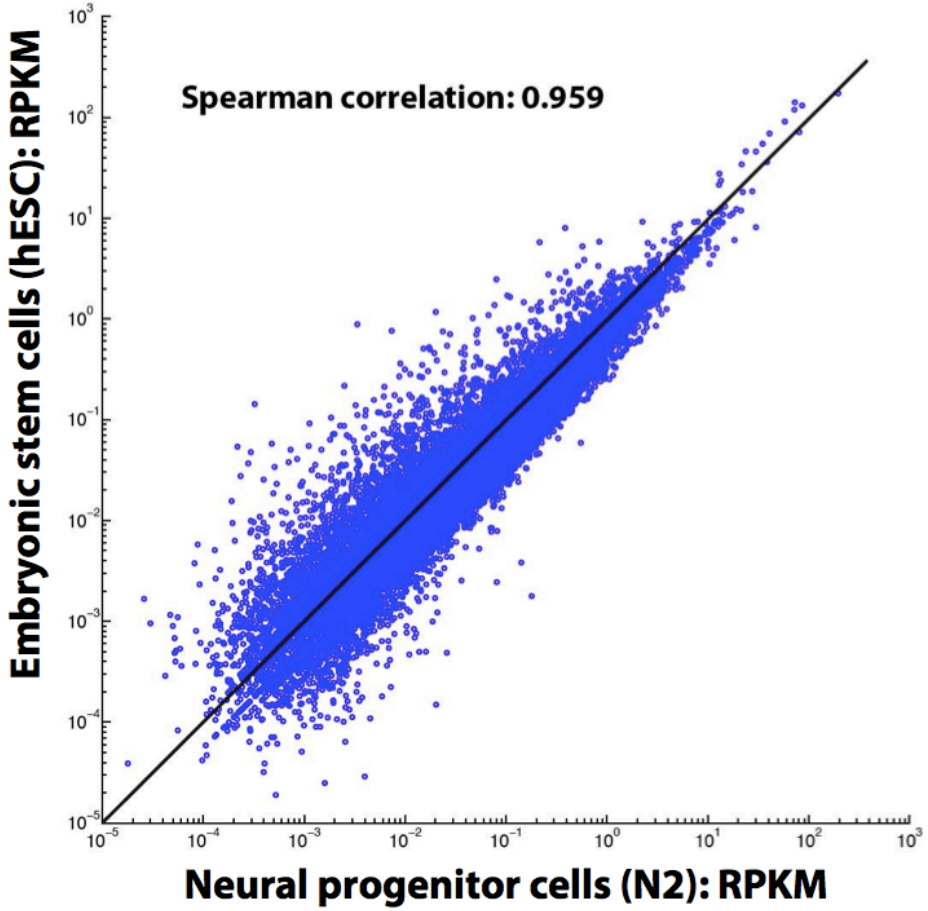
Wang *et al Nature Reviews Genetics* 2009

RNA-Seq conclusions

- RNA-Seq is a powerful tool to identify new transcribed regions of the genome and compare the RNA complements of different tissues.
- Quantification harder than CHIP-seq because of RNA splicing
- Expectation maximization algorithm can be useful for quantifying overlapping transcripts

COMPARING GENOME-WIDE SIGNALS

RNA-seq Expression Correlation

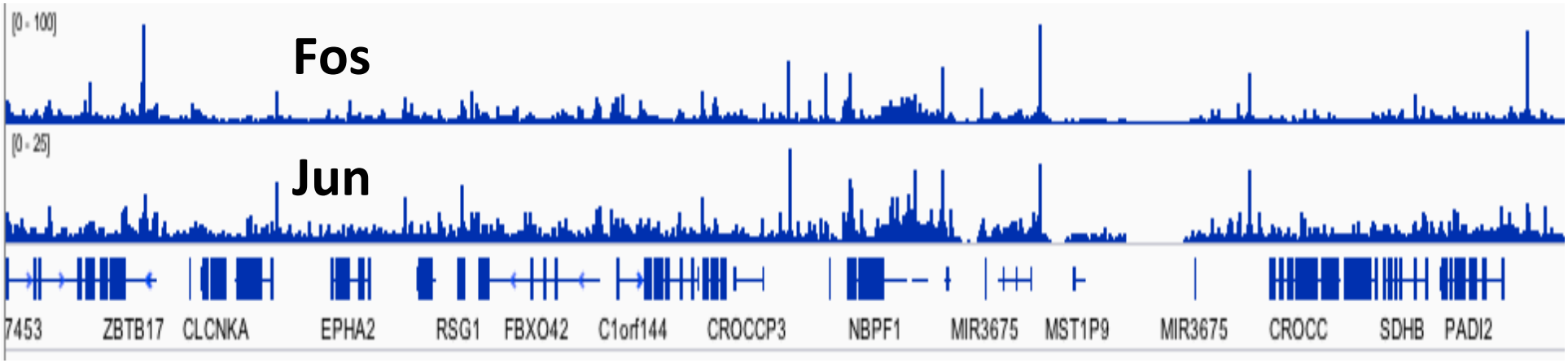


- Correlate expression between tissues

Adapted from Wu, J.Q., Habegger, L. et al. Dynamic transcriptomes during neural differentiation of human embryonic stem cells revealed by short, long, and paired-end sequencing. *Proceedings of the National Academy of Sciences* 107, 5254-5259 (2010).

Slide adapted from L Habegger

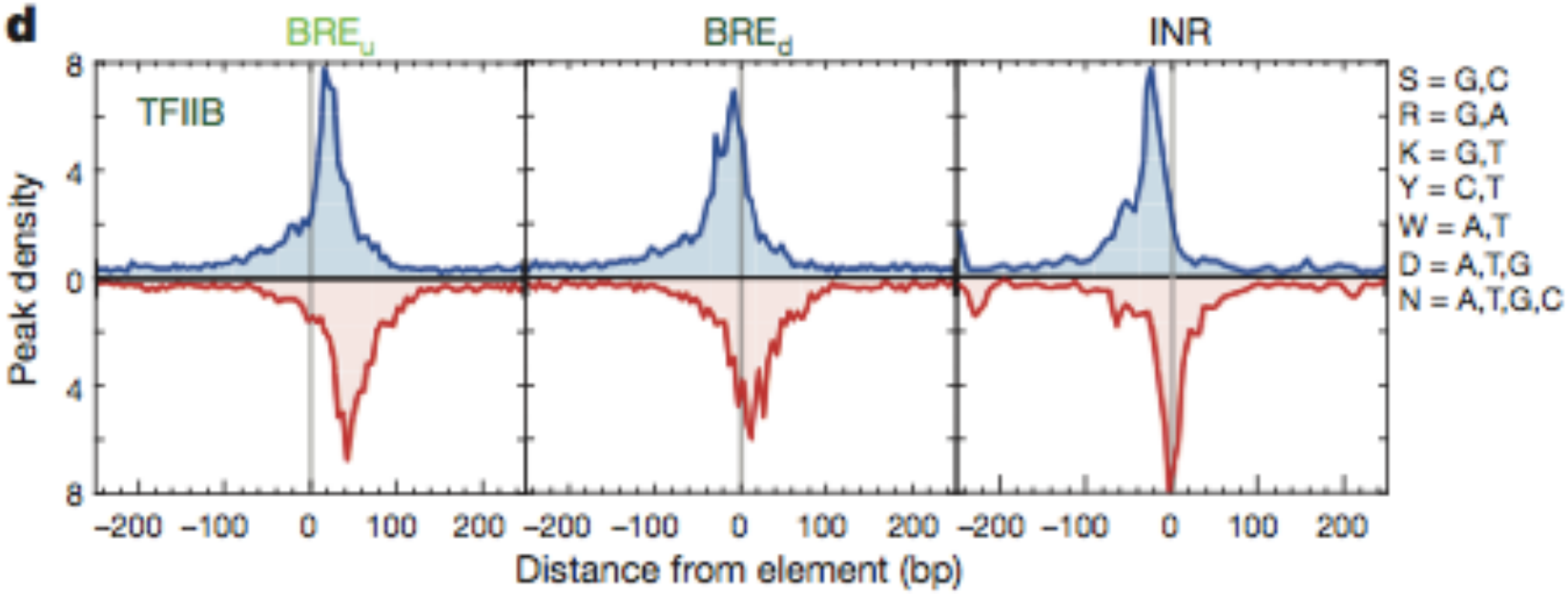
CHIP-seq signals of interacting proteins



- Fos and Jun, which interact physically, have similar binding profiles at many genomic loci.

Data from ENCODE Consortium, figure by me

Signal aggregation

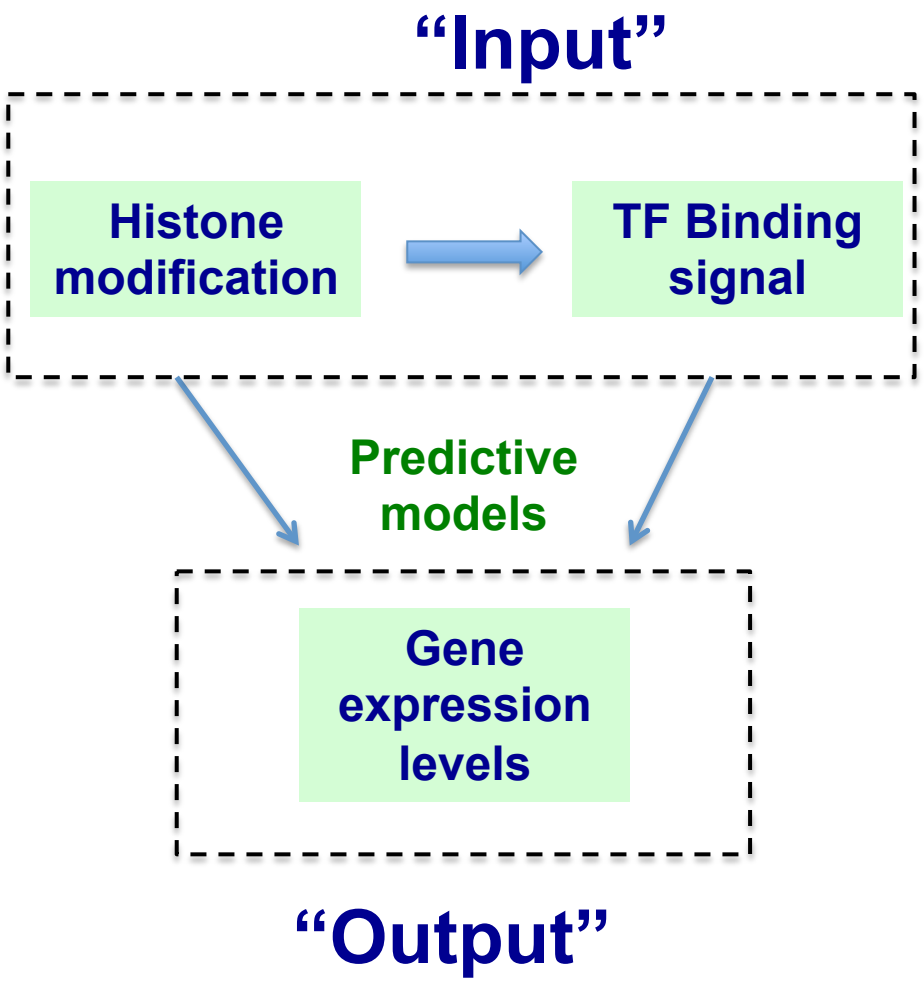


- Sum signals from all genomic locations of a certain type
  - Here: CHIP seq signal at fixed distances from protein binding motif

Venters and Pugh *Nature* 2013

PREDICTING GENE EXPRESSION WITH CHIP-SEQ DATA

RELATING GENE EXPRESSION WITH HISTONE MODIFICATION AND TF BINDING SIGNALS

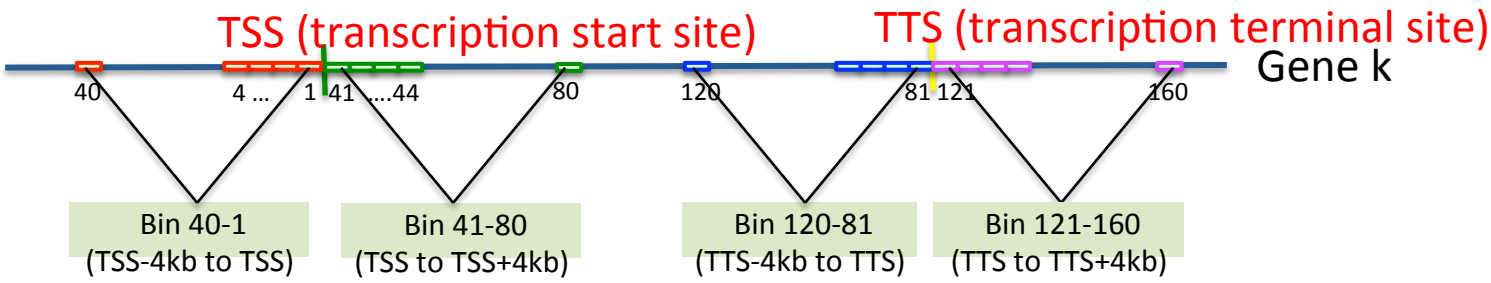


To what extent the gene expression levels are determined by TF binding/ HM modification?

Slide by Chao Cheng

Setting up the model

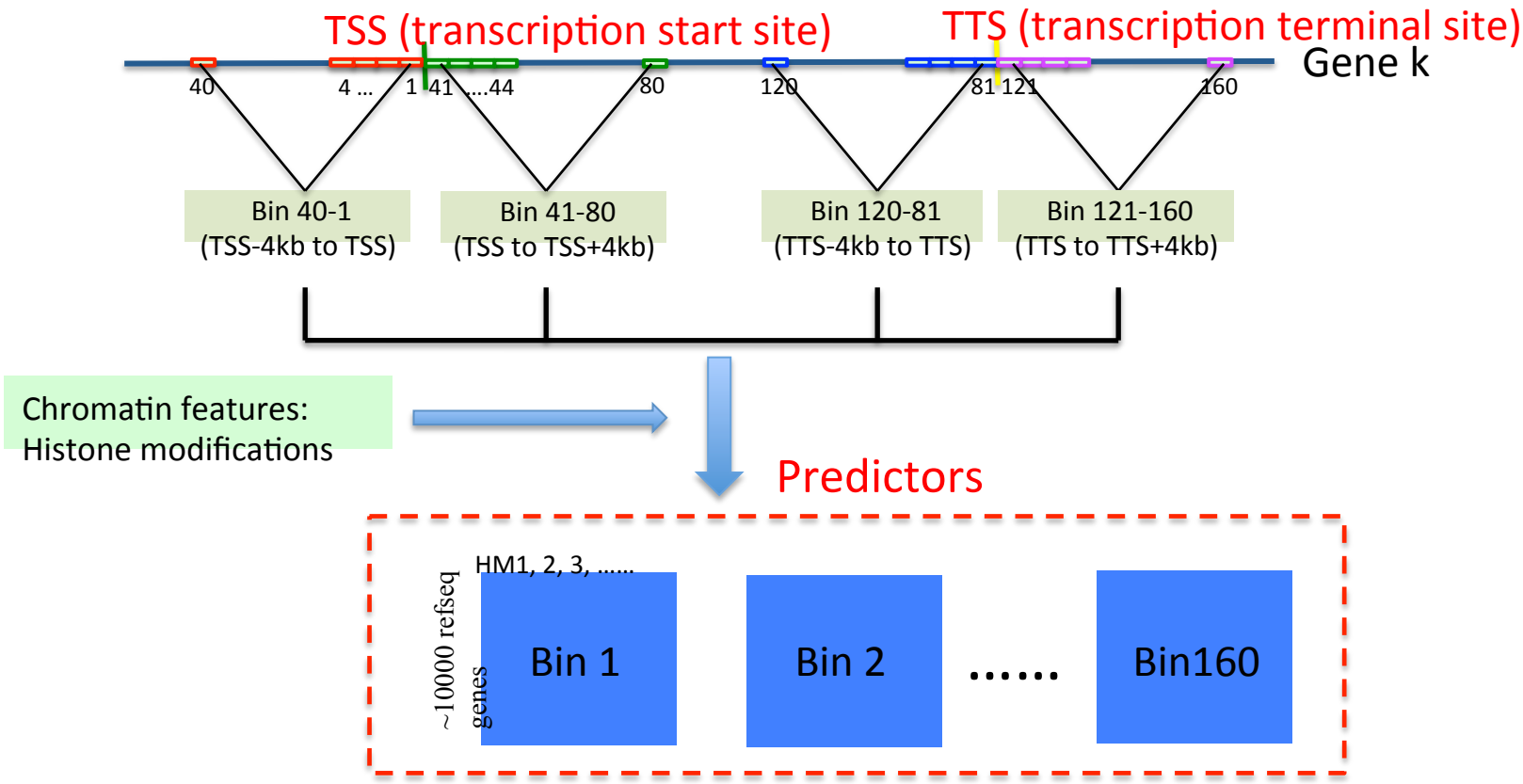
1. Divide area around gene into bins according to distance to trascription start and end sites



Slide by Chao Cheng

Setting up the model

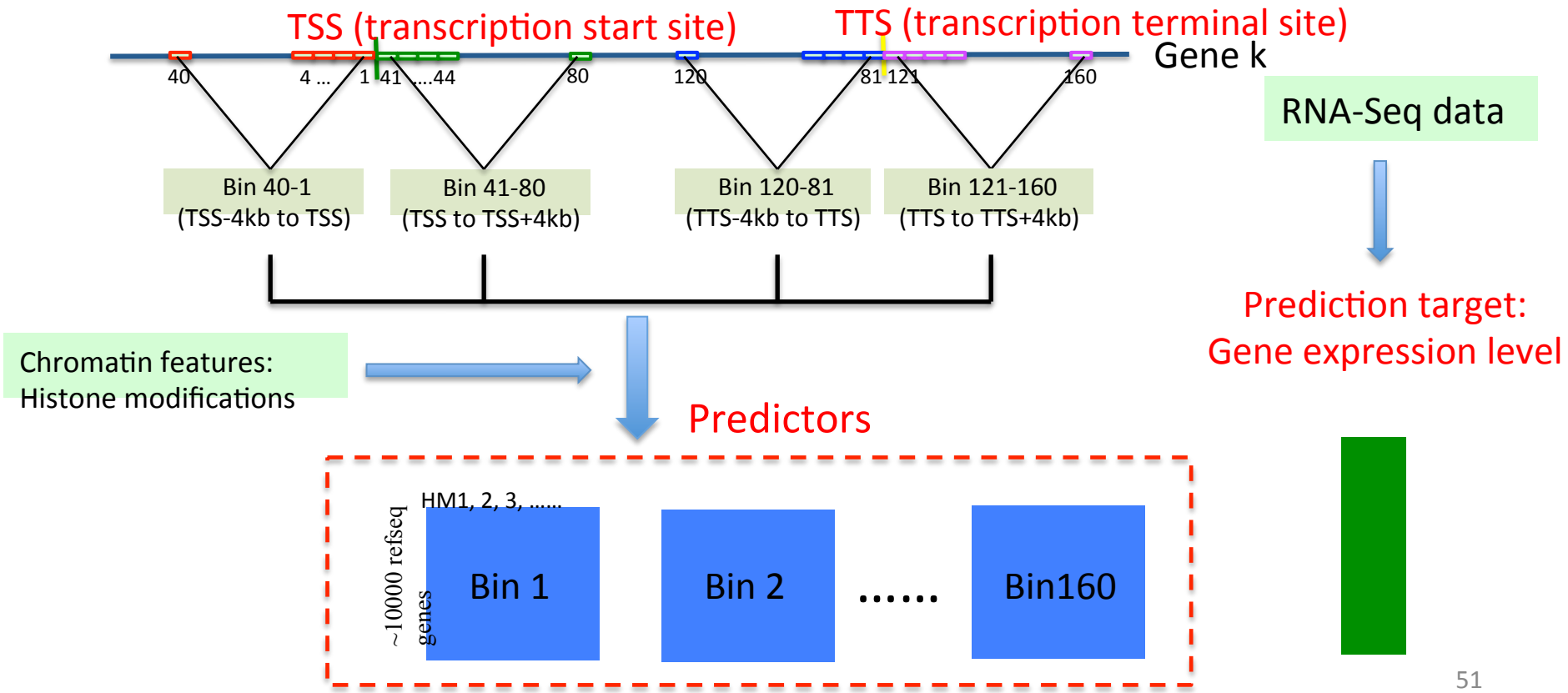
2. Collect histone modification data for each bin, and for each gene



Slide by Chao Cheng

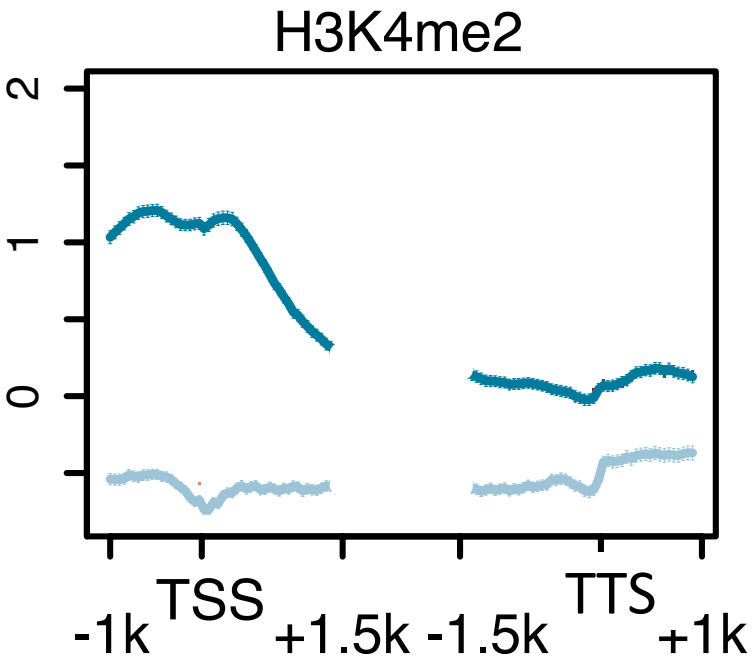
Setting up the model

3. Train model to "learn" relationship between CHIP-seq and RNA-seq data.

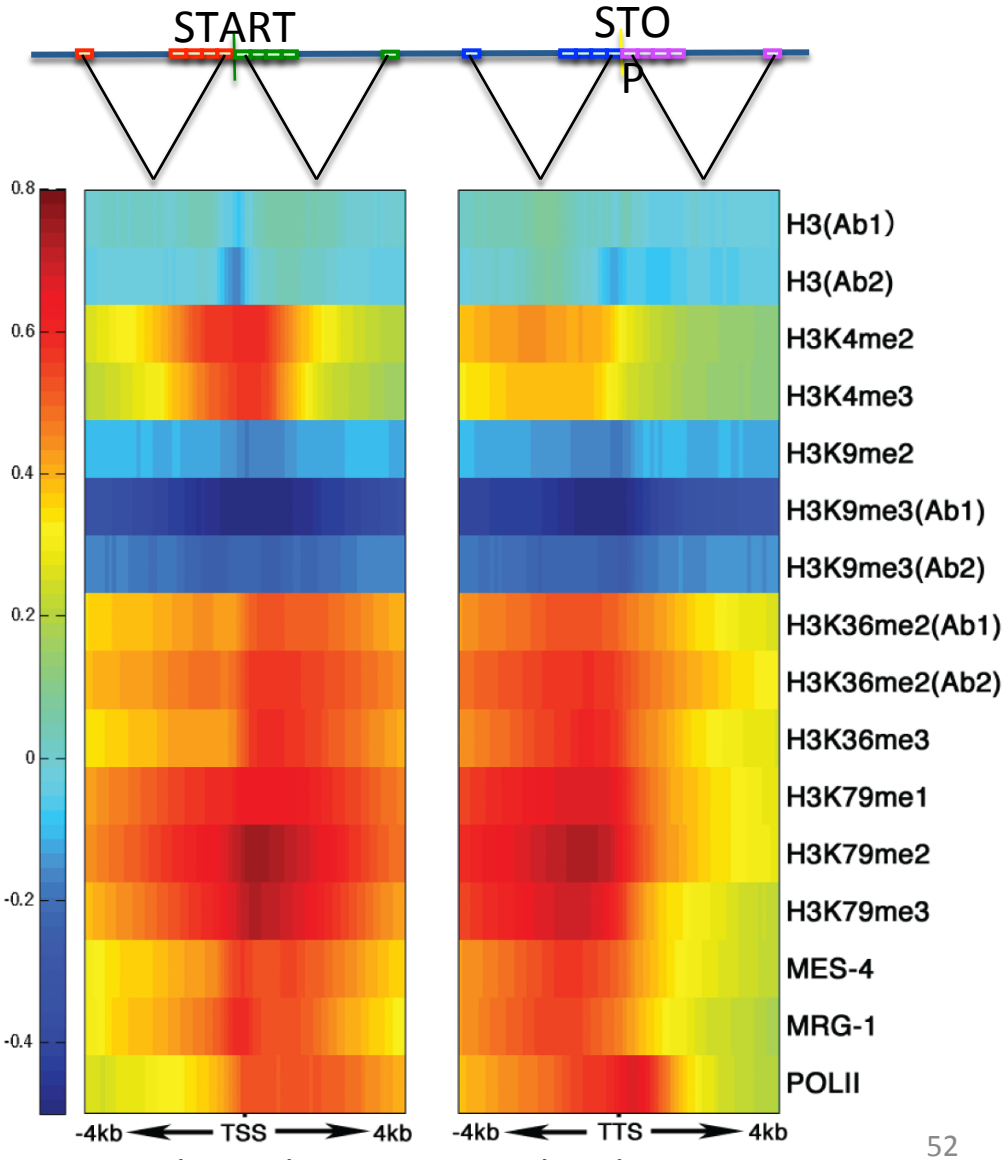


Slide by Chao Cheng

His. mods around TSS & TTS are clearly related to level of gene expression, in a position-dependent fashion



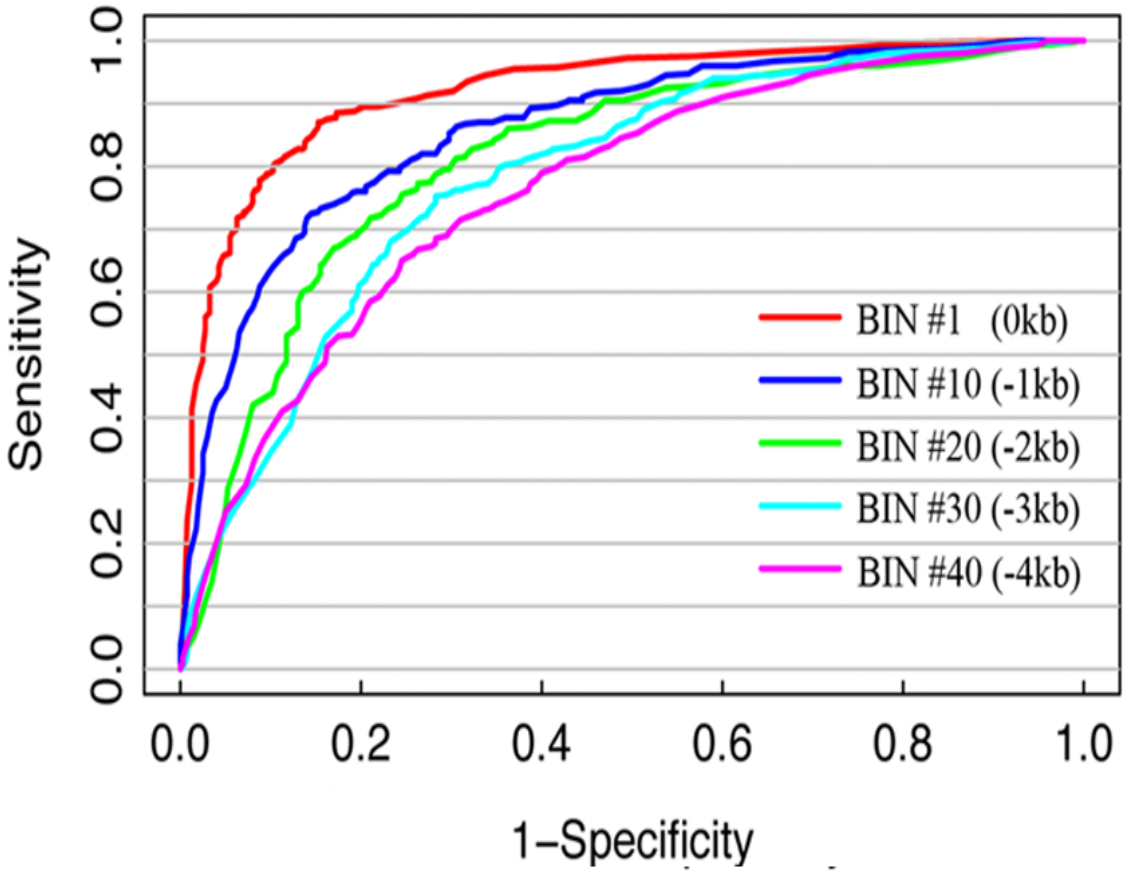
Gerstein,…, Cheng* et al. 2010, Science*



Correlation between Signal and expression

Slide by Chao Cheng

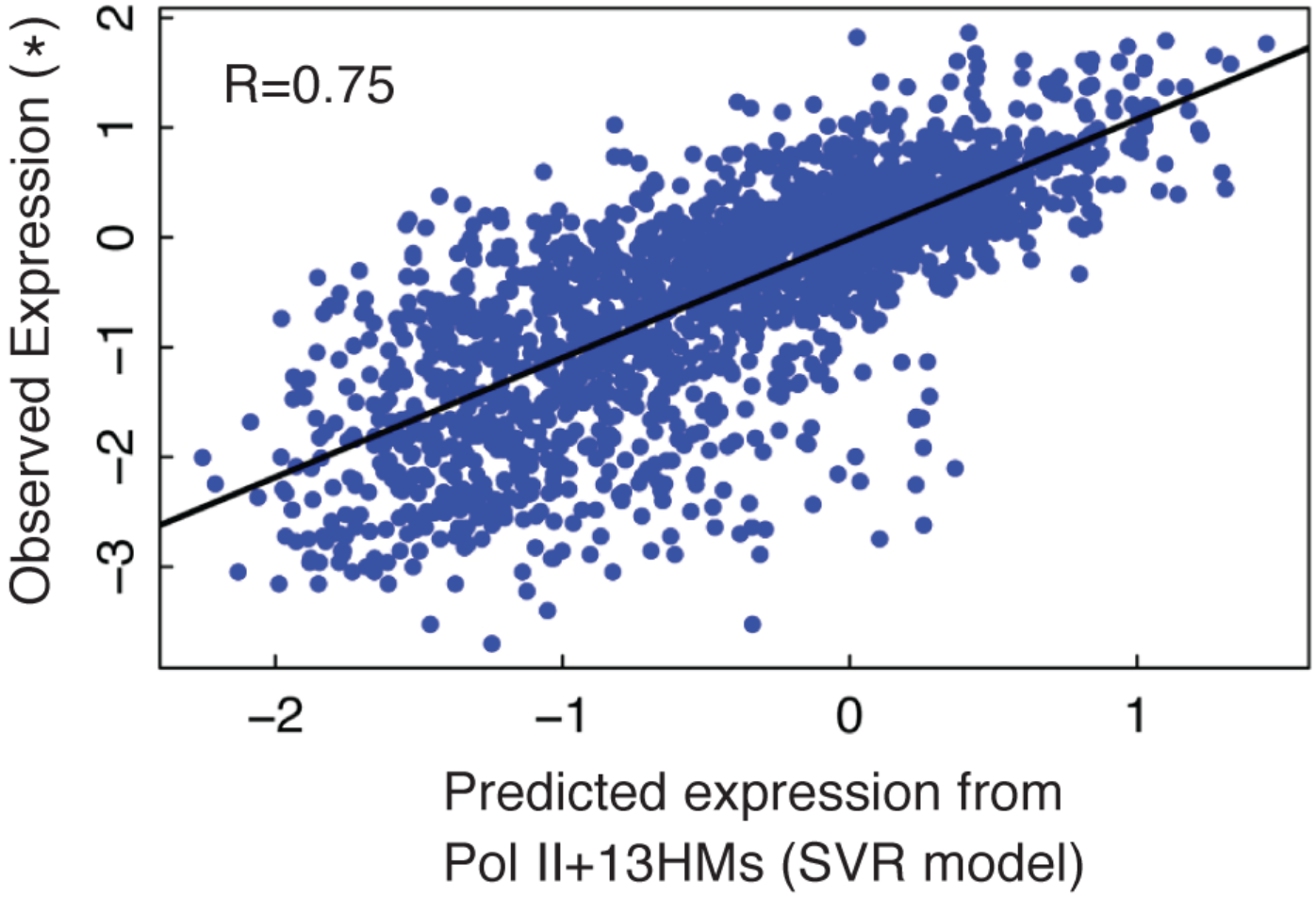
Support vector machine to classify genes with high, medium and low expression



✓Areas close to gene predict expression better

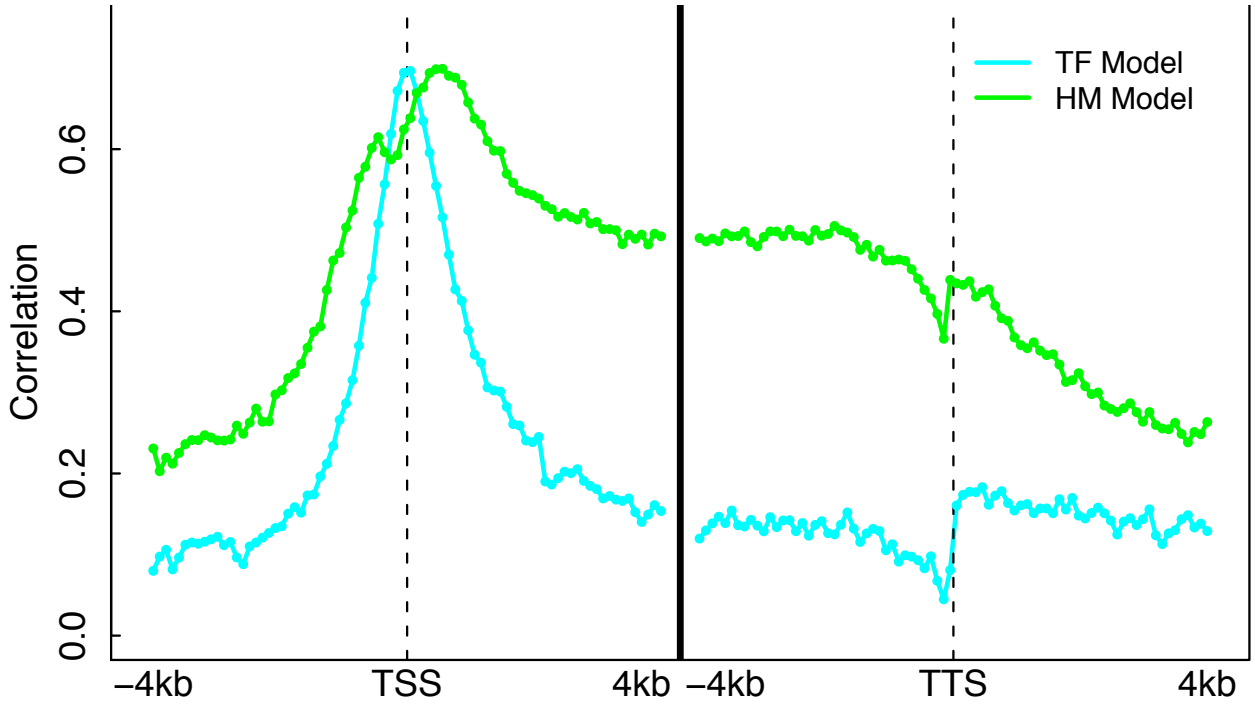
Slide by Chao Cheng

Support vector regression to predict gene expression levels



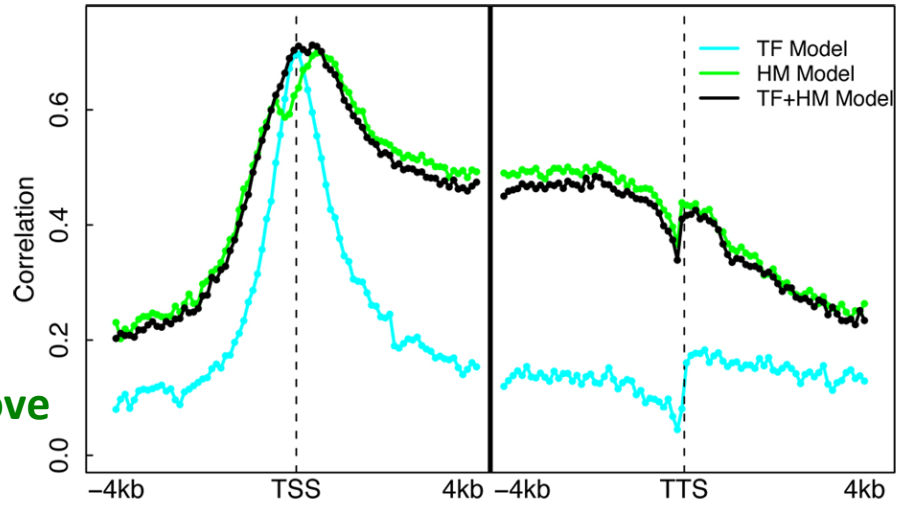
Slide by Chao Cheng

Mouse ESC Models Illuminates Different Regions of Influence for TFs vs HMs



- Datasets
  - ChIP-Seq for 12 TFs (Chen et al. 2008)
  - ChIP-Seq for 7 HMs (Meissner et al.'08; Mikkelsen et al. '07)
  - RNA-Seq (Cloonan et al. 2008)

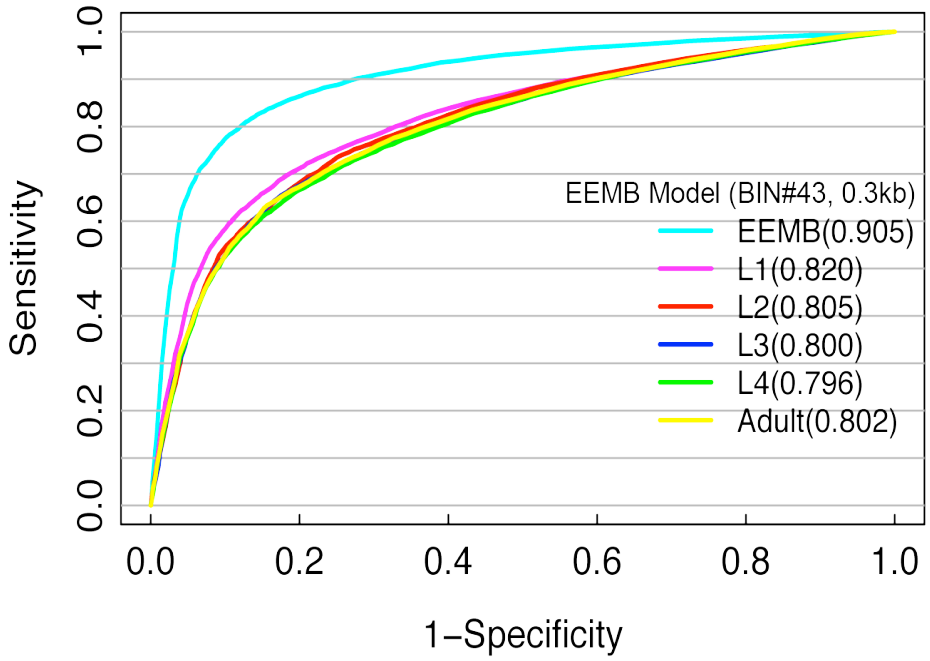
A TF+HM model that combine TF and HM features does NOT improve accuracy!



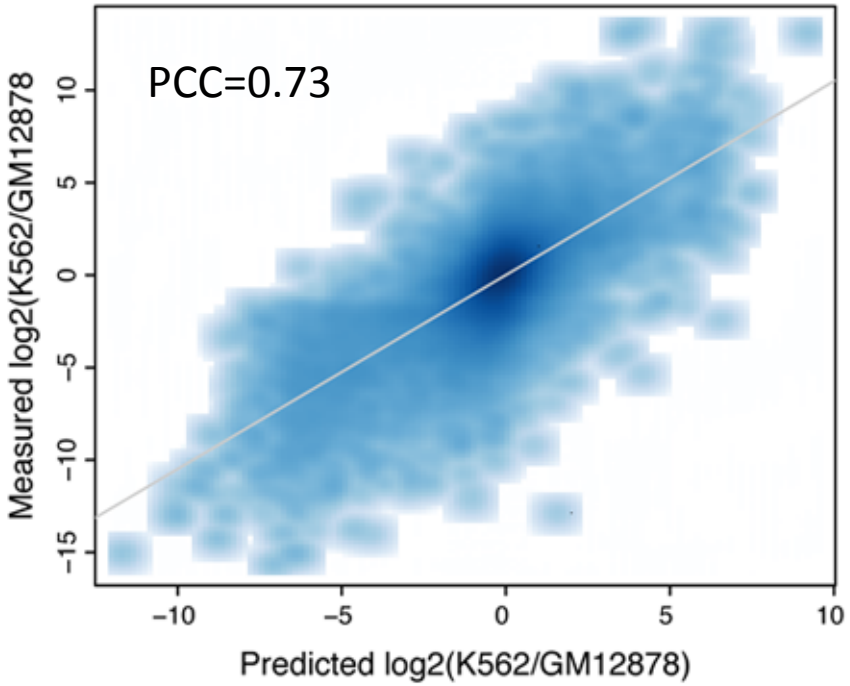
Cheng et al. 2011, Nucleic Acids Res.

Slide by Chao Cheng

TF and HM models are tissue specific



HM model-- Best prediction is achieved by using histone modification and expression data from the same developmental stage



TF model– differential TF binding signals are predictive of differential expression levels between two human cell lines

Cheng et al. 2011, Genome Biology (a)

Slide by Chao Cheng

Summary: relate TF/HM signals with expression

- TF/HM signals are highly **predictive** to gene expression
- TF and HM signals are **redundant** for 'predict' gene expression
- TF and HM models are tissue/cell line **specific**
- **microRNA** expression can also be predicted

Gerstein et al. 2010, Science *Cheng et al. 2011, Nucleic Acids Res.* *Cheng et al. 2011, Genome Biology (a)*

Slide by Chao Cheng

Conclusions

- Diverse sequencing experiments have common analysis elements, based on **signal processing**.
- Proper **statistics** key to making claims about NGS data.
- Integrating many genome-wide experiments through **machine learning** can yield useful inferences about **biology**.