

# Using the MAT Algorithm for ChIP-chip Data Analysis

*Cliff Meyer*

*Dept of Biostatistics and Computational Biology*

*Dana-Farber Cancer Institute*

*Harvard School of Public Health*

# Outline

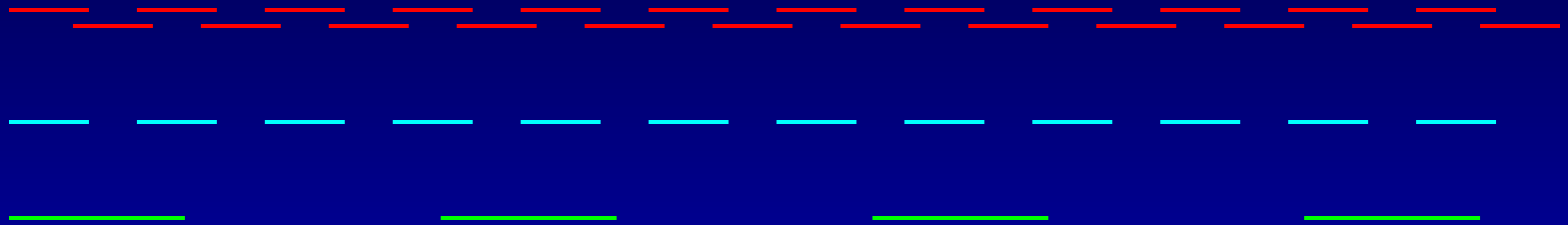
- **ChIP-chip on genome tiling microarrays**
- **MAT: Model-based Analysis of Tiling arrays**
- **How to use MAT**
- **Interpreting the results**

# Genome Tiling Microarrays

Genomic DNA on the chromosome

---

Tiling  
Probes

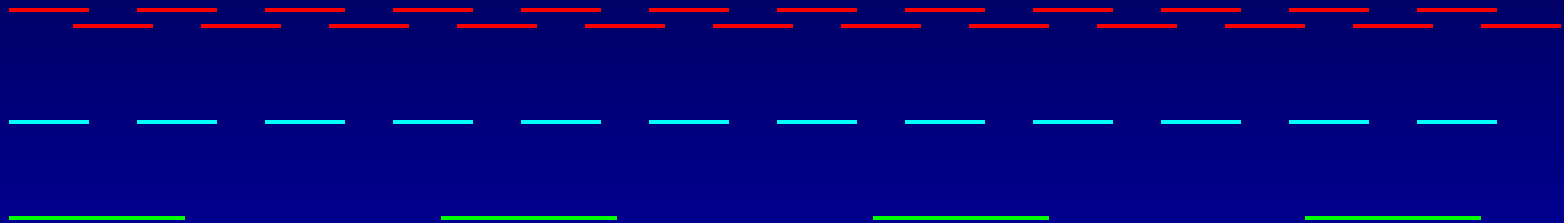


# Genome Tiling Microarrays

Genomic DNA on the chromosome

---

Tiling  
Probes



- **Affymetrix whole human or mouse genome tiling**
  - \$2,065 = arrays + reagent + processing
  - 7 arrays  $\times$  6 million probes = 42 million data points
  - 25 bp probe tiled at 35 bp resolution



slide by Shirley Liu

# Gene Transcription Regulation

- Human genome: ~2% coding, 98% “junk” contains regulatory elements



# Gene Transcription Regulation

- Human genome: ~2% coding, 98% “junk” contains regulatory elements



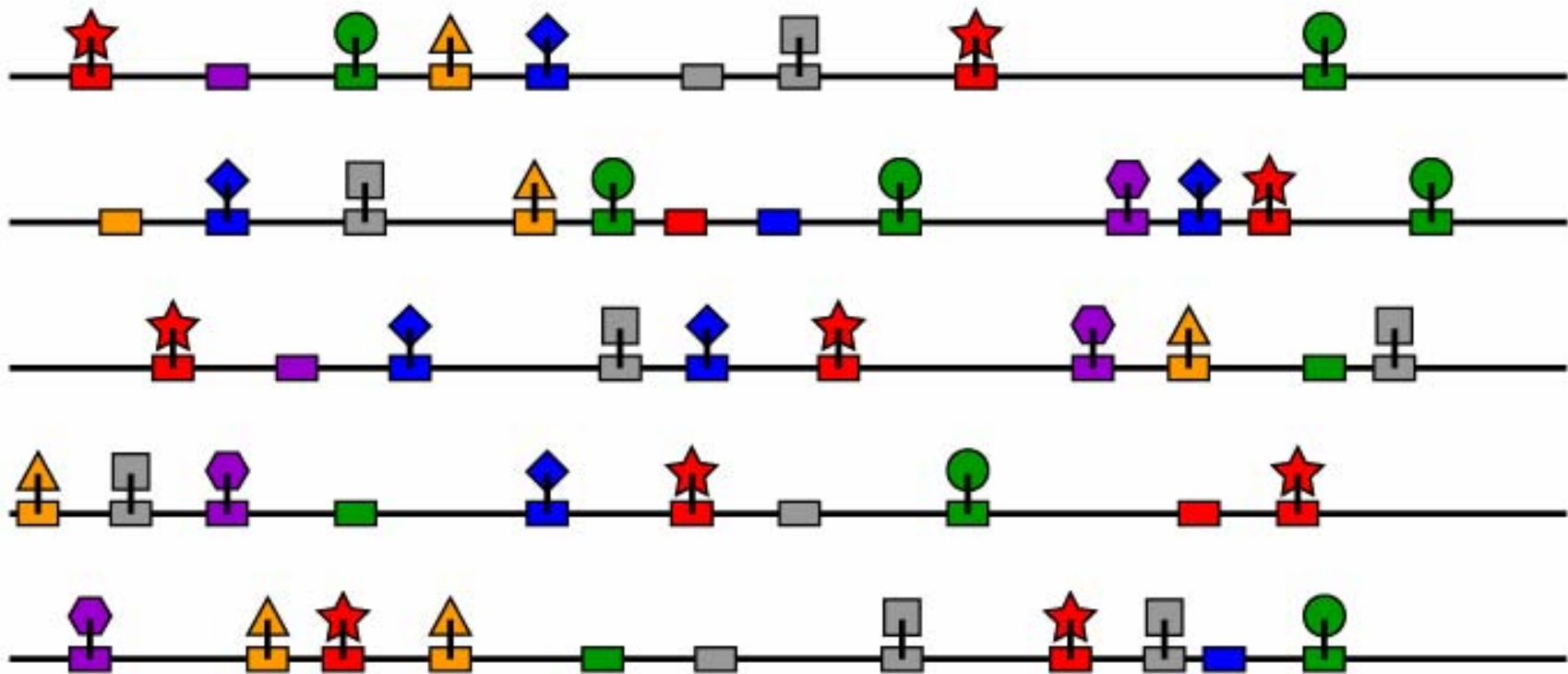
# Gene Transcription Regulation

- **Human genome: ~2% coding, 98% “junk” contains regulatory elements**

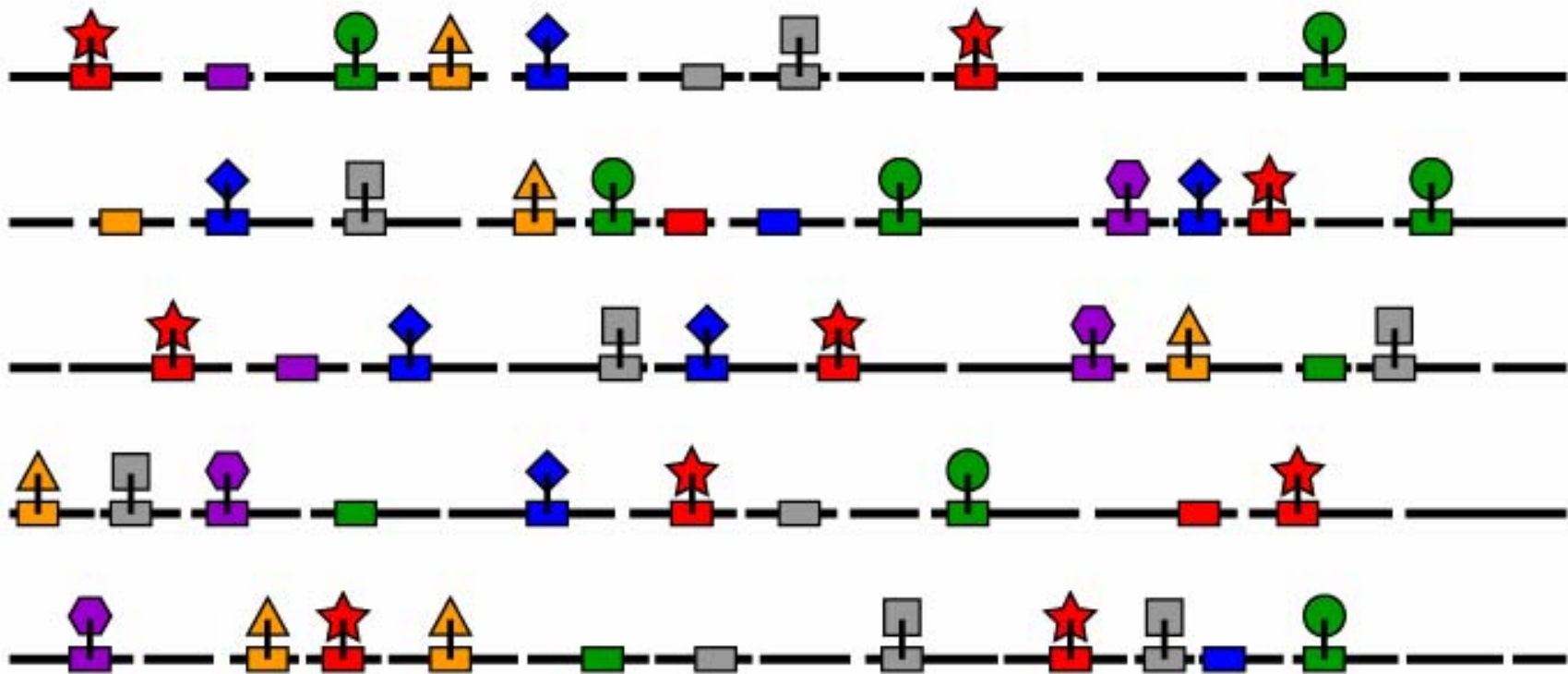


- **Locating the (disease/developmental) factors genome-wide is essential to understanding their targets and regulatory mechanism**

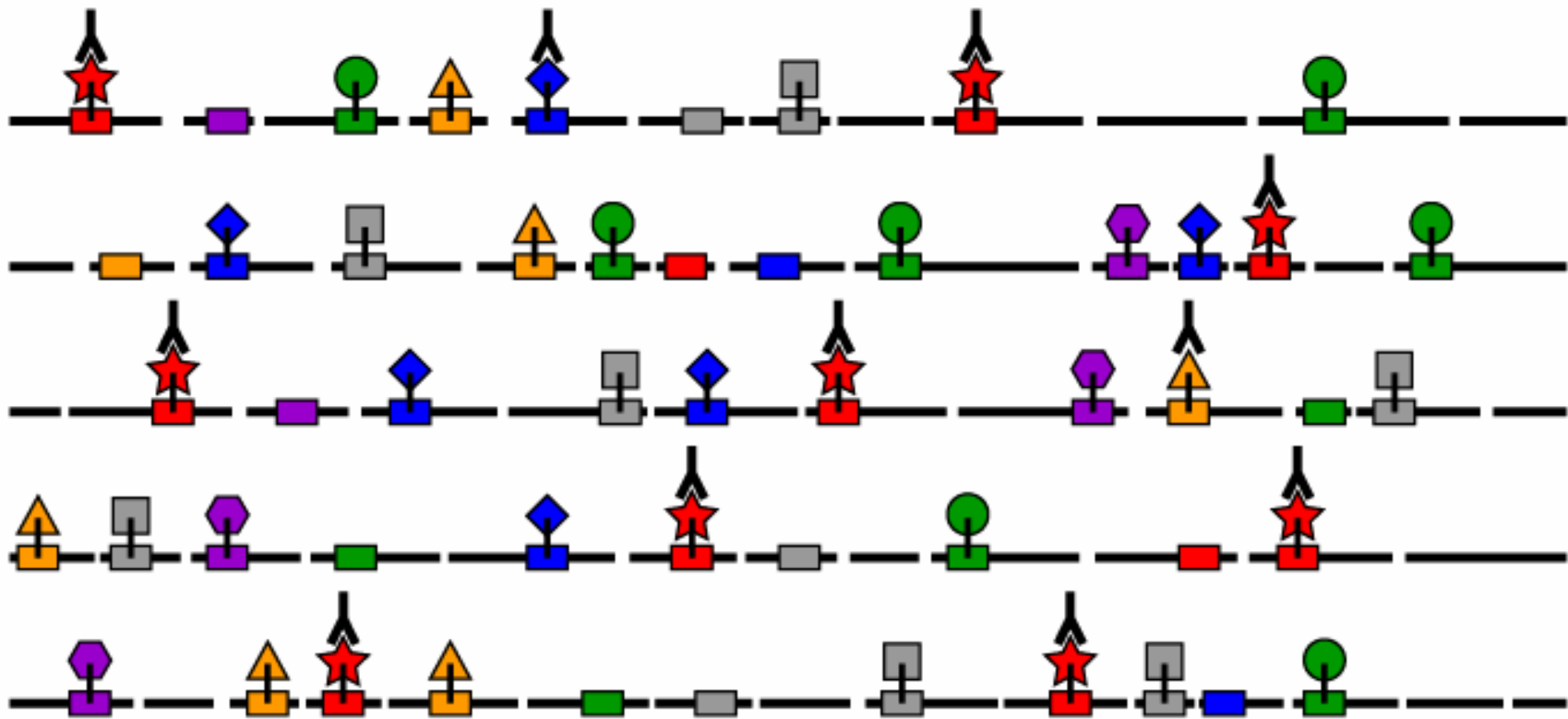
# Protein-DNA Interaction *in vivo*



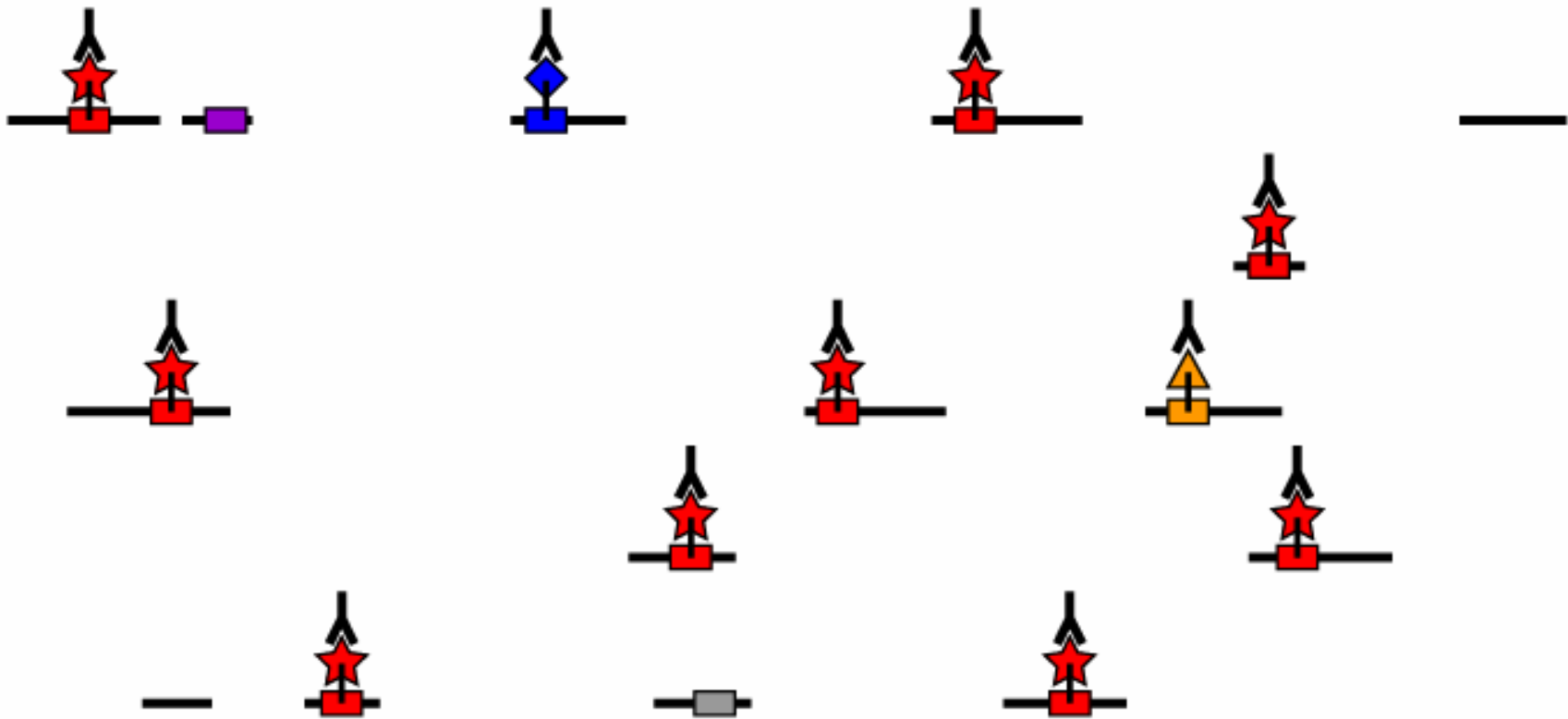
# Sonication (~500bp)



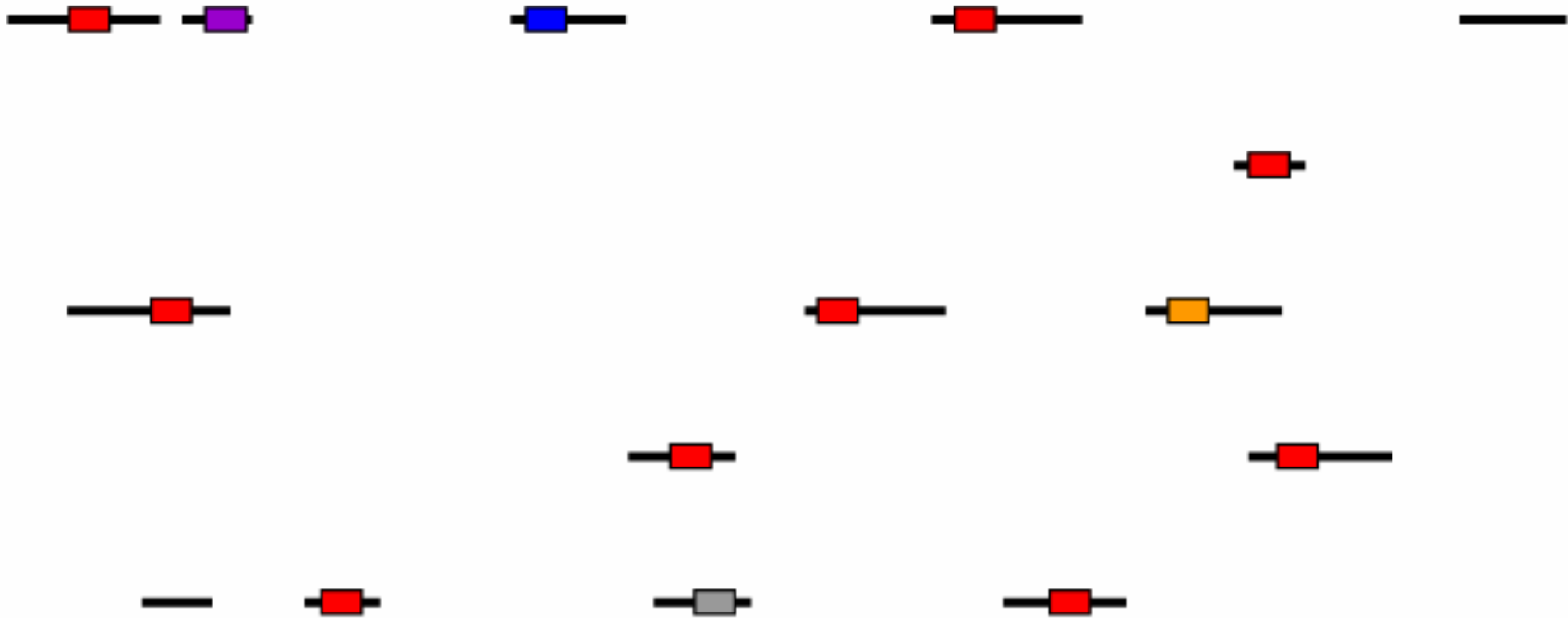
# Factor-specific Antibody



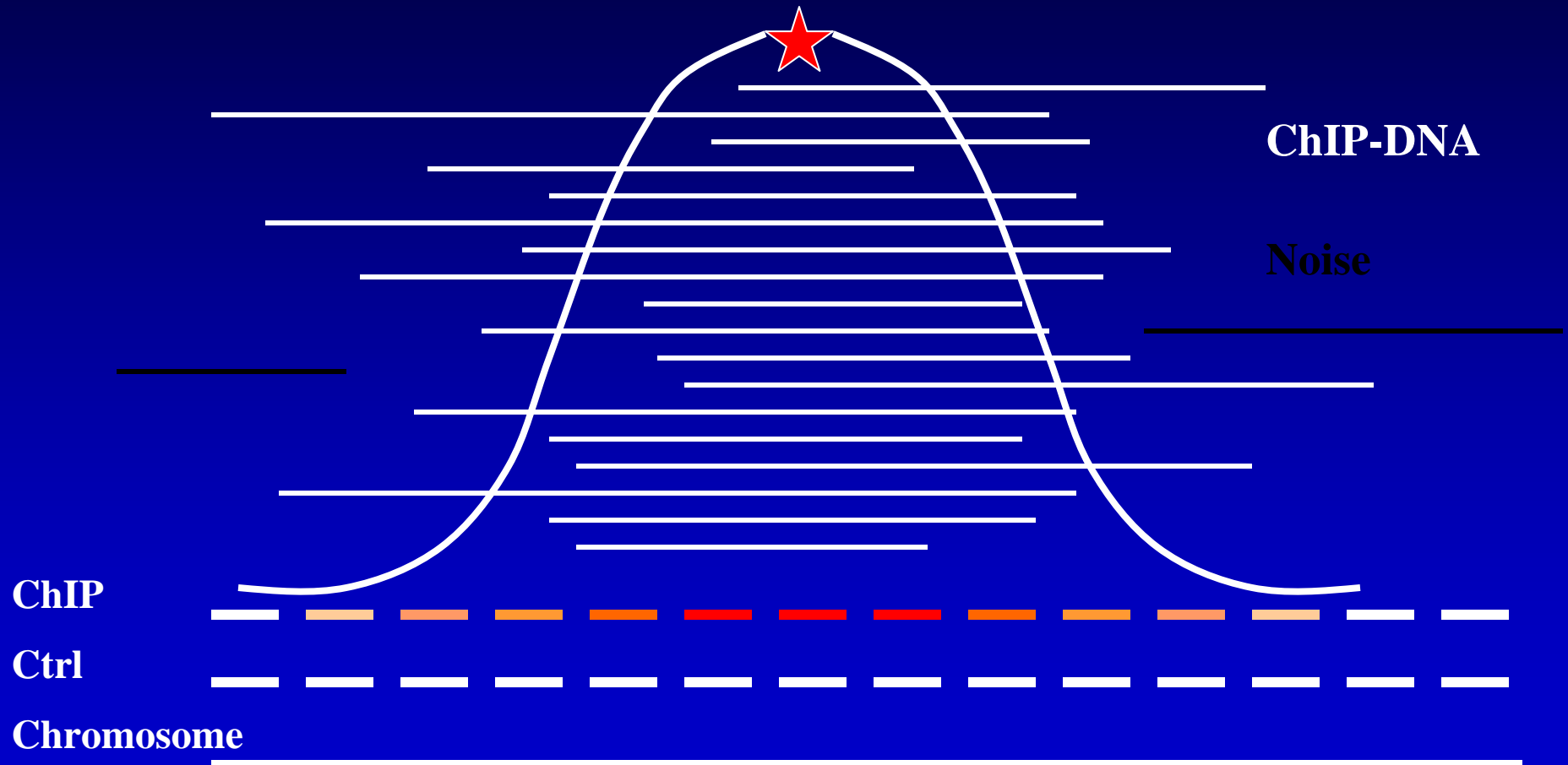
# Immunoprecipitation



# DNA Purification



# ChIP-chip Array Hybridization



# Mann-Whitney Rank Sum Test for ChIP-region Detection

- Affy TAS, Cawley et al (*Cell* 2004):
  - Each probe: rank probes within [-500bp, +500bp] window
  - Check whether sum of ChIP ranks is much smaller

	ctrl 1	ctrl 2	ChIP 1	ChIP 2		ctrl 1	ctrl 2	ChIP 1	ChIP 2
probe 1	1.71	2.23	3.02	2.25	probe 1	17	15	13	14
probe 2	4.27	3.10	3.86	4.70	probe 2	6	12	10	3
probe 3	4.06	3.67	4.03	4.74	probe 3	7	11	8	2
probe 4	1.20	0.40	1.31	1.85	probe 4	19	20	18	16
probe 5	4.29	3.95	4.56	4.76	probe 5	5	9	4	1

# Outline

- ChIP-chip on genome tiling microarrays
- **MAT: Model-based Analysis of Tiling arrays**
- How to use MAT
- Interpreting the results

# Data Analysis Challenges

- Large quantity of data (42 M / sample).
- Many, many hypotheses.
- Probe values noisy.
- Probe specific behaviour.
- Genomic artifacts.
- Good results from ChIP-chip rare on first try.
- Interpretation of “ChIP-enriched” regions.

# MAT Features

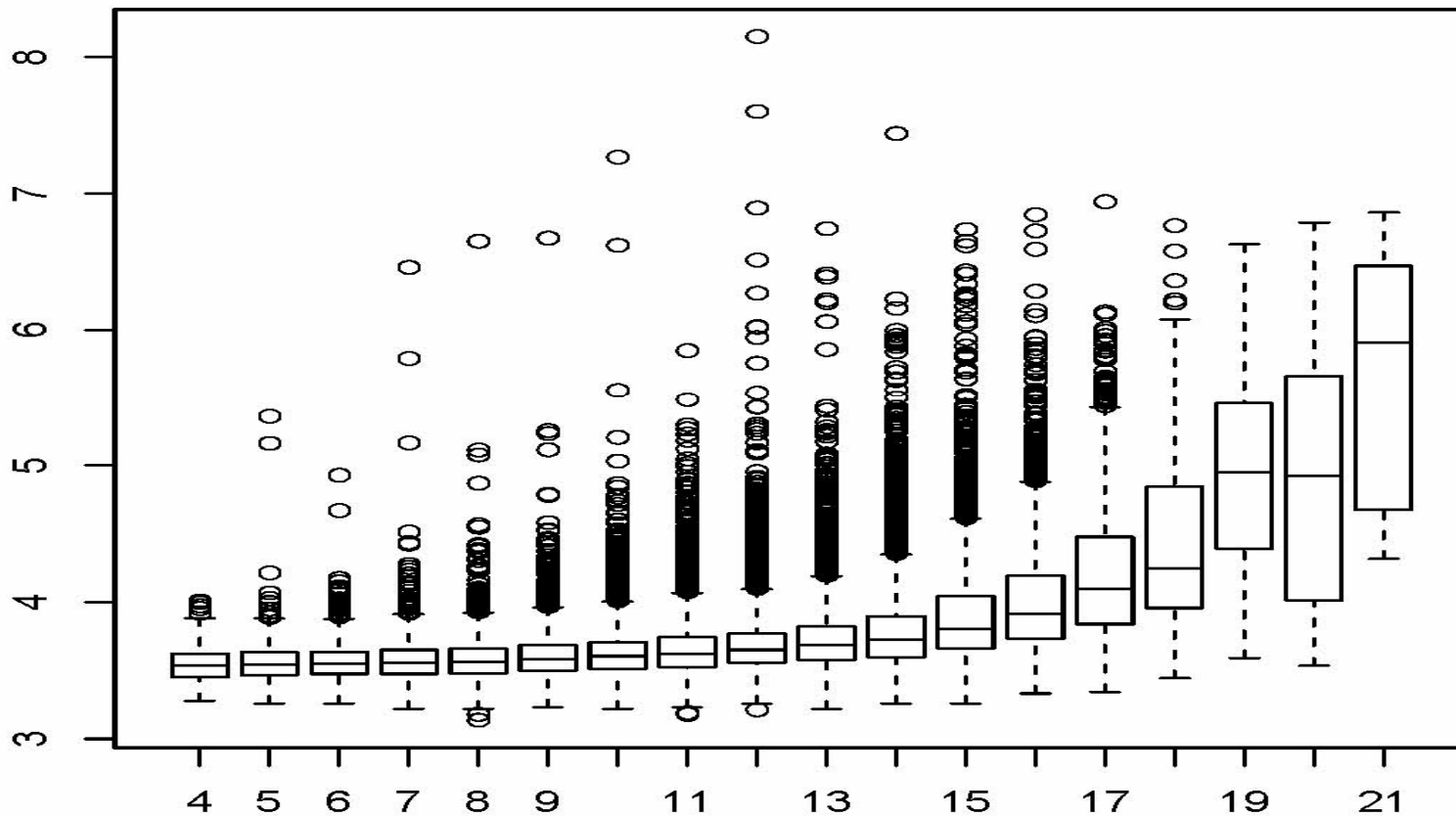
*MAT: Johnson et al, PNAS 2006*

*<http://chip.dfci.harvard.edu/~wli/MAT>*

- **Flexible, accurate method works well for with single ChIP.**
- **Higher sensitivity, specificity with multiple ChIPs and controls.**
- **Careful treatment of genomic artifacts.**
- **Simple to use.**
- **Output convenient for use in UCSC genome browser or Affymetrix Integrated Genome Browser (IGB).**
- **Diagnostic measures of data quality.**

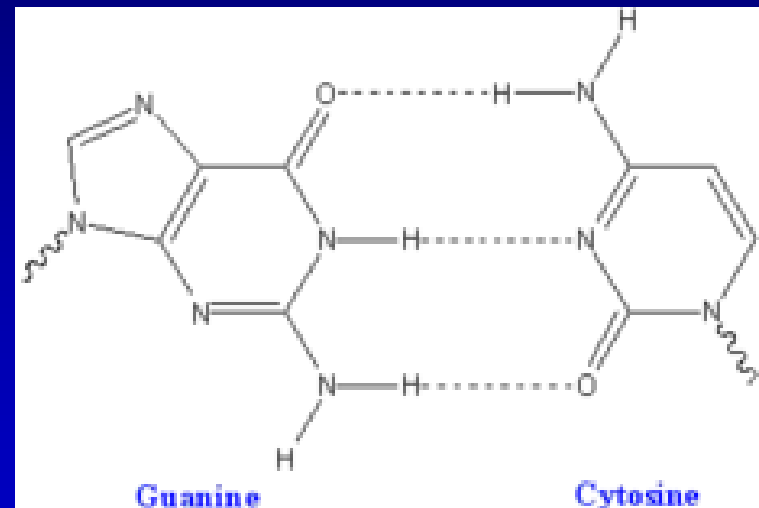
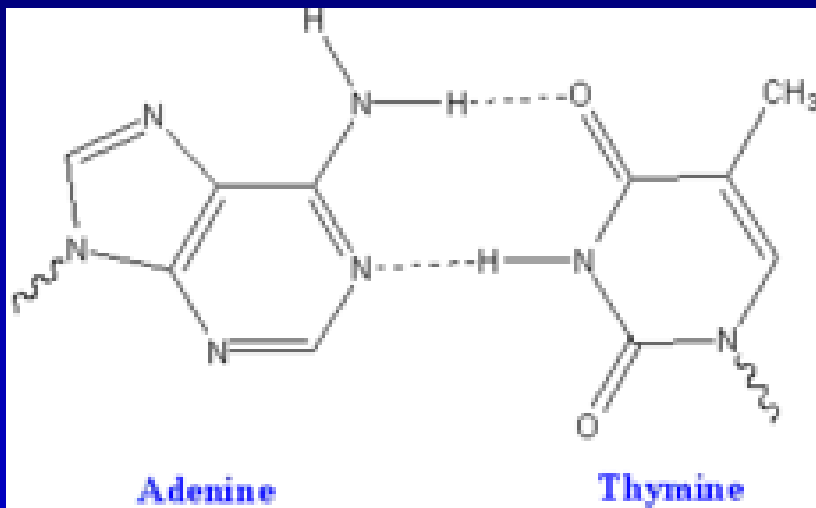
# Probe Sequence Effect on Signal Intensity

5ug log(PM) by GC bin

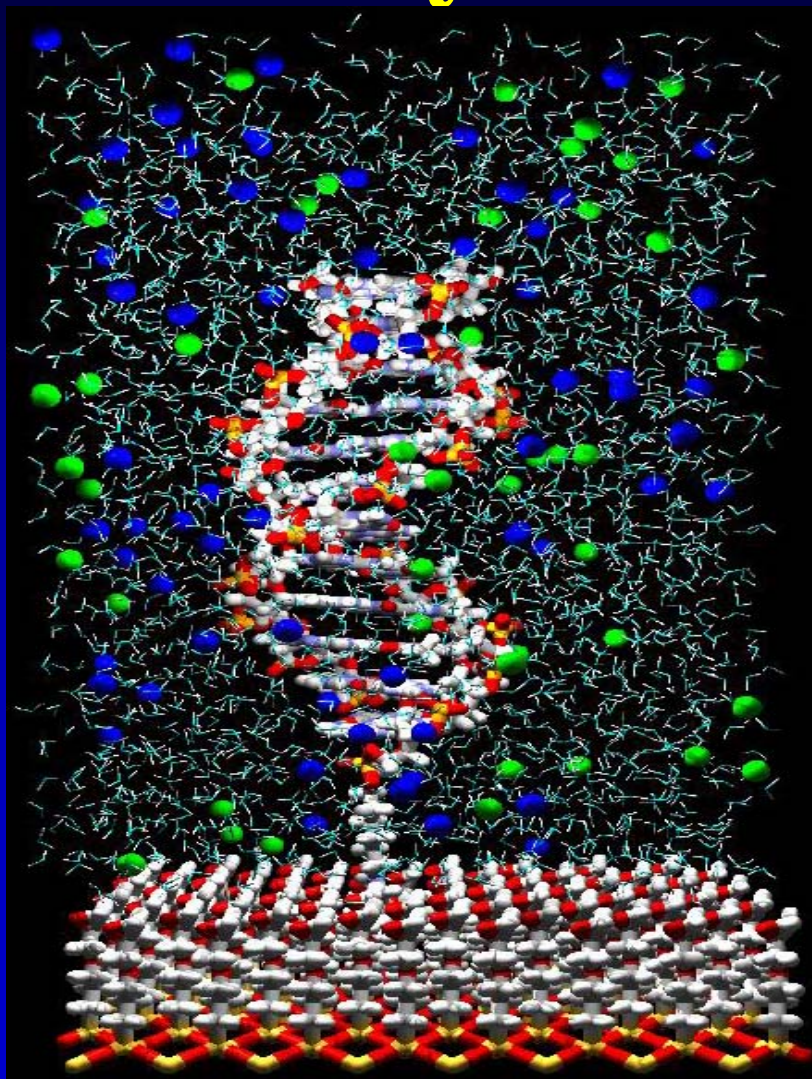


# Hybridization Thermodynamics

## Describe Probe Behaviour



# Hybridization Thermodynamics



# MAT

- **Empirical estimate of probe behaviour.**
- **Most of the probes in ChIP-chip measure non-specific hybridization.**
- **Estimate probe behaviour by checking other probes with similar sequence on the same array.**

# MAT Probe Behaviour Model

$$\text{Log}(PM_i) = \alpha n_{iT} + \sum_{j=1}^{25} \sum_{k=A,C,G} \beta_{jk} I_{ijk} + \sum_{l=A,C,G,T} \gamma_l n_{il}^2 + \delta \text{Log}(c_i) + \varepsilon_i$$

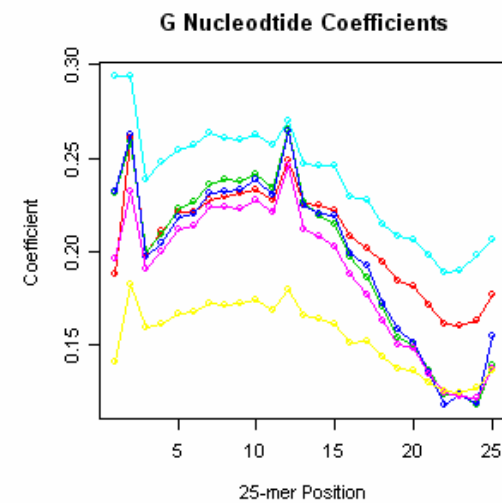
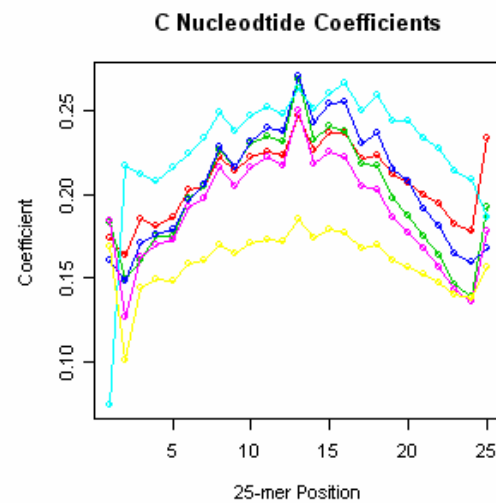
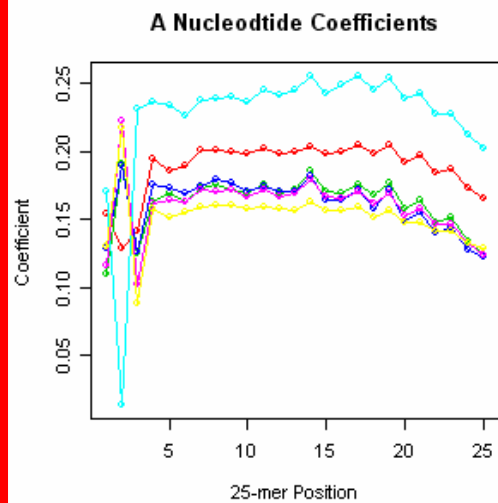
Probe  
signal

# of T's  
intercept

Position-specific  
A, C, G effect

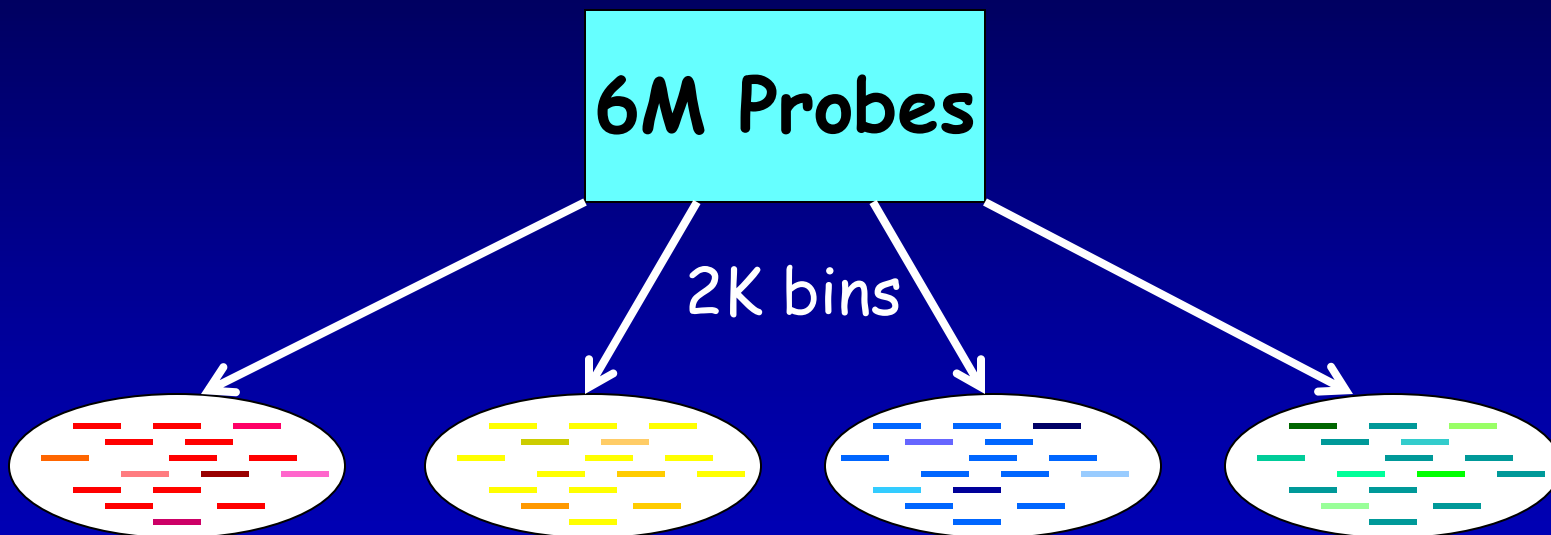
A,C,G,T count  
squared

25-mer copy  
number



# Probe Standardization

- Fit the probe model array by array



- Standardize probe behavior within each bin
- After standardization: different probes in different samples are ~ comparable

## Raw probe values at two spike-in regions with concentration 2X

2X

2X

Spike-in (ChIP) raw data



Ctrl raw data



## Sequence-based probe behavior standardization

ChIP standardized



Ctrl standardized



## Window-based neighboring probe combination for ChIP-region detection

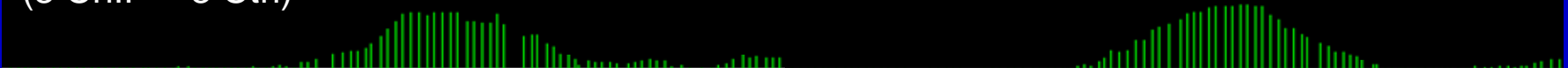
ChIP Window



(ChIP - Ctrl)



(3 ChIP - 3 Ctrl)



126,589,000

126,590,000

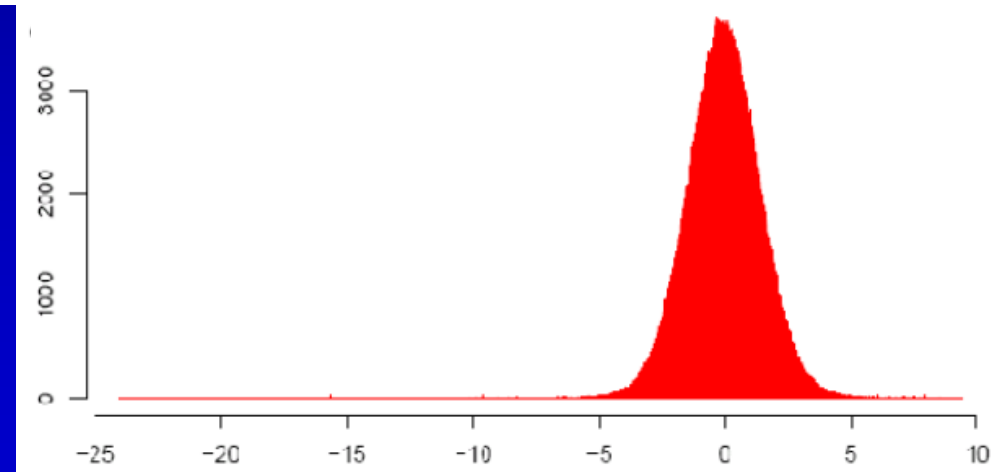
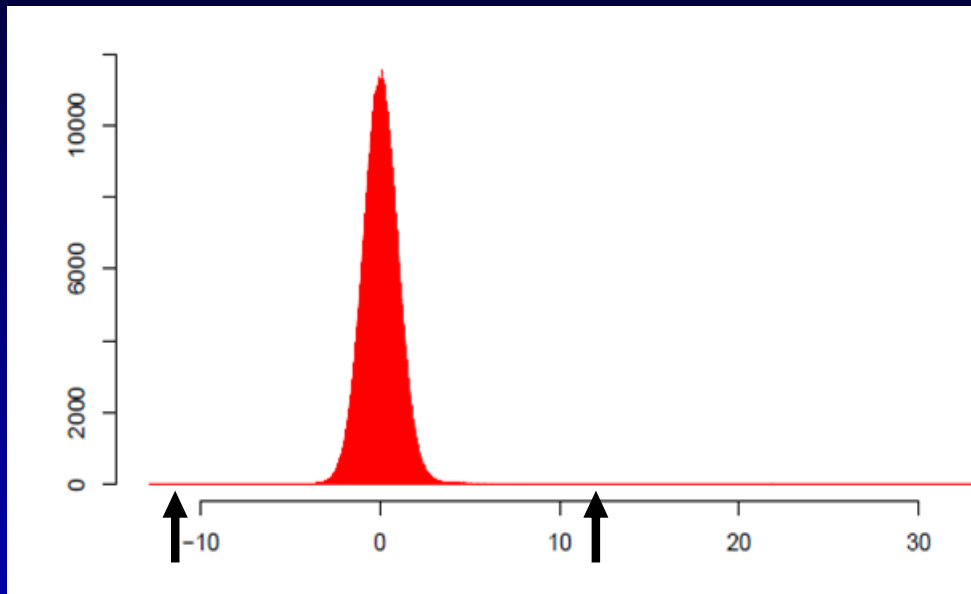
1

114,176,000

114,177,000

# Statistical Significance of Hits

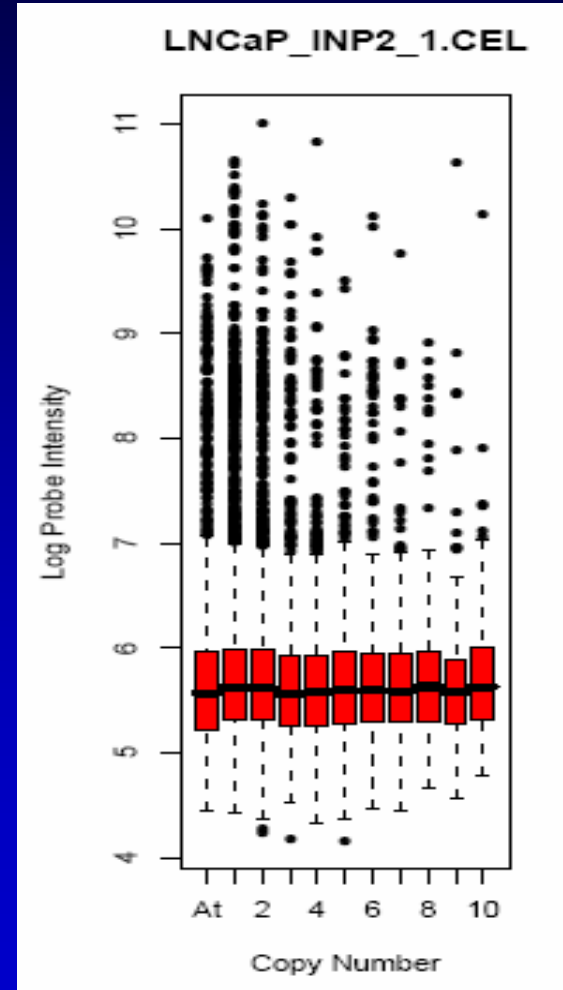
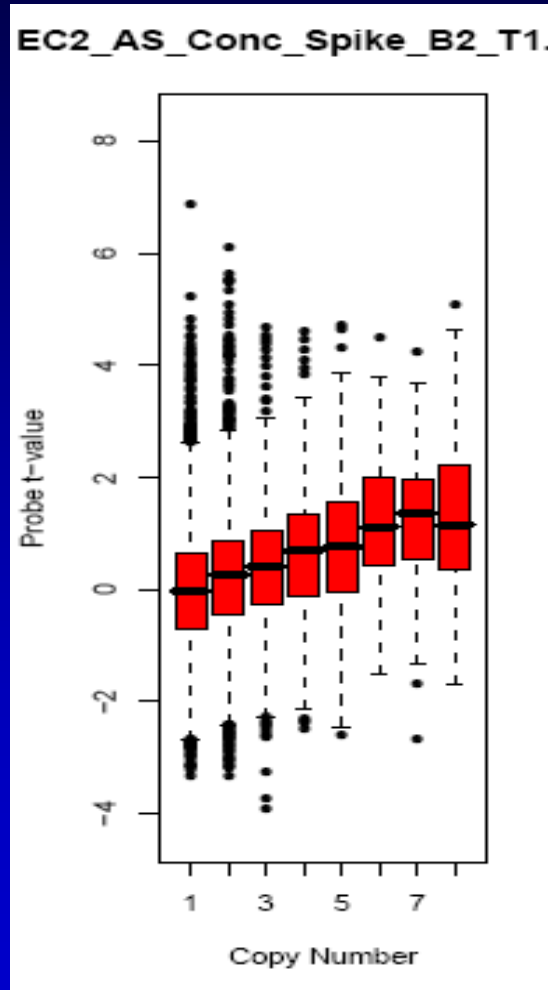
- FDR estimated based on negative peaks
- Can find failed samples



slide by Shirley Liu

# Internal Control of Labeling and Hybridization

- Collect probes that map to 0, 1, 2,...10 copies in the genome and see their average behavior
- Can tell failed hybridization



slide by Shirley Liu

# Benchmark for ChIP-chip Target Detection

- **ENCODE Spike-in experiment:**  
both amplified and un-amplified

## ChIP

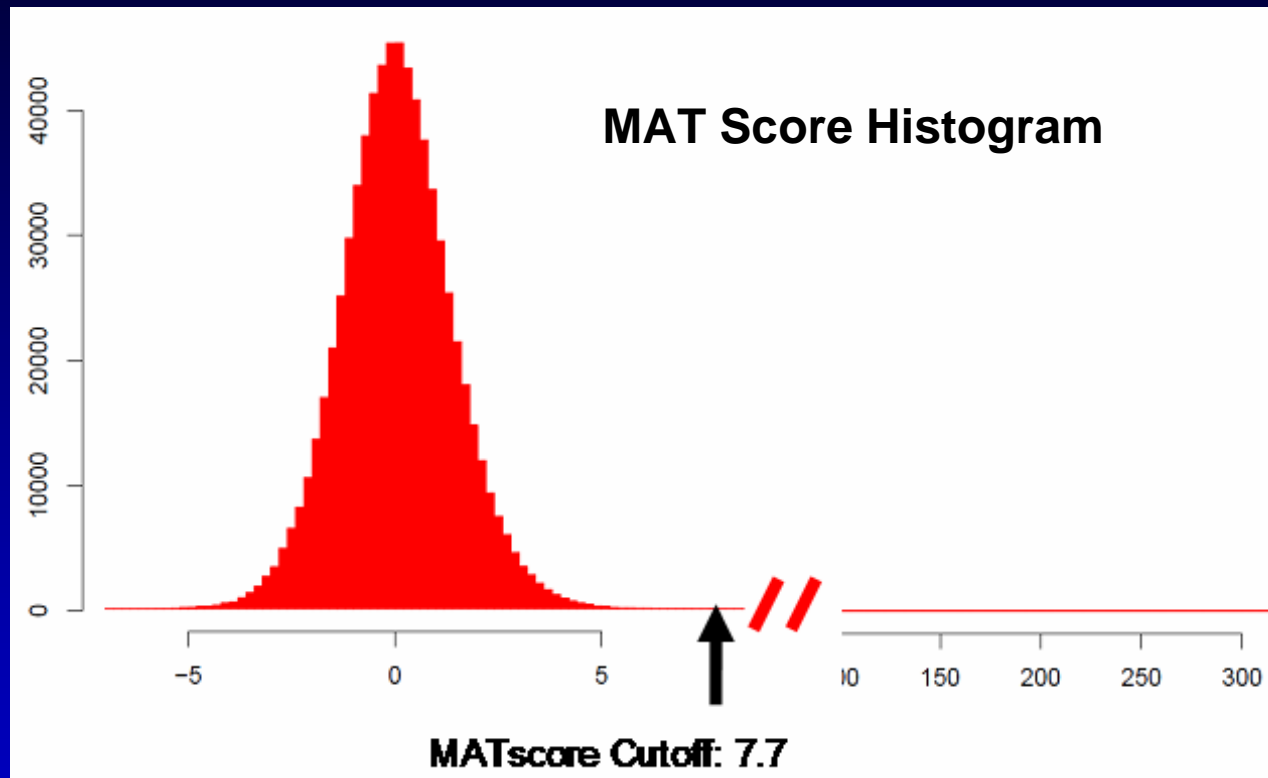
96 ENCODE clones,  
**2,4,8,...,256X** enrichment +  
total chromatin DNA

## Input

total genomic DNA

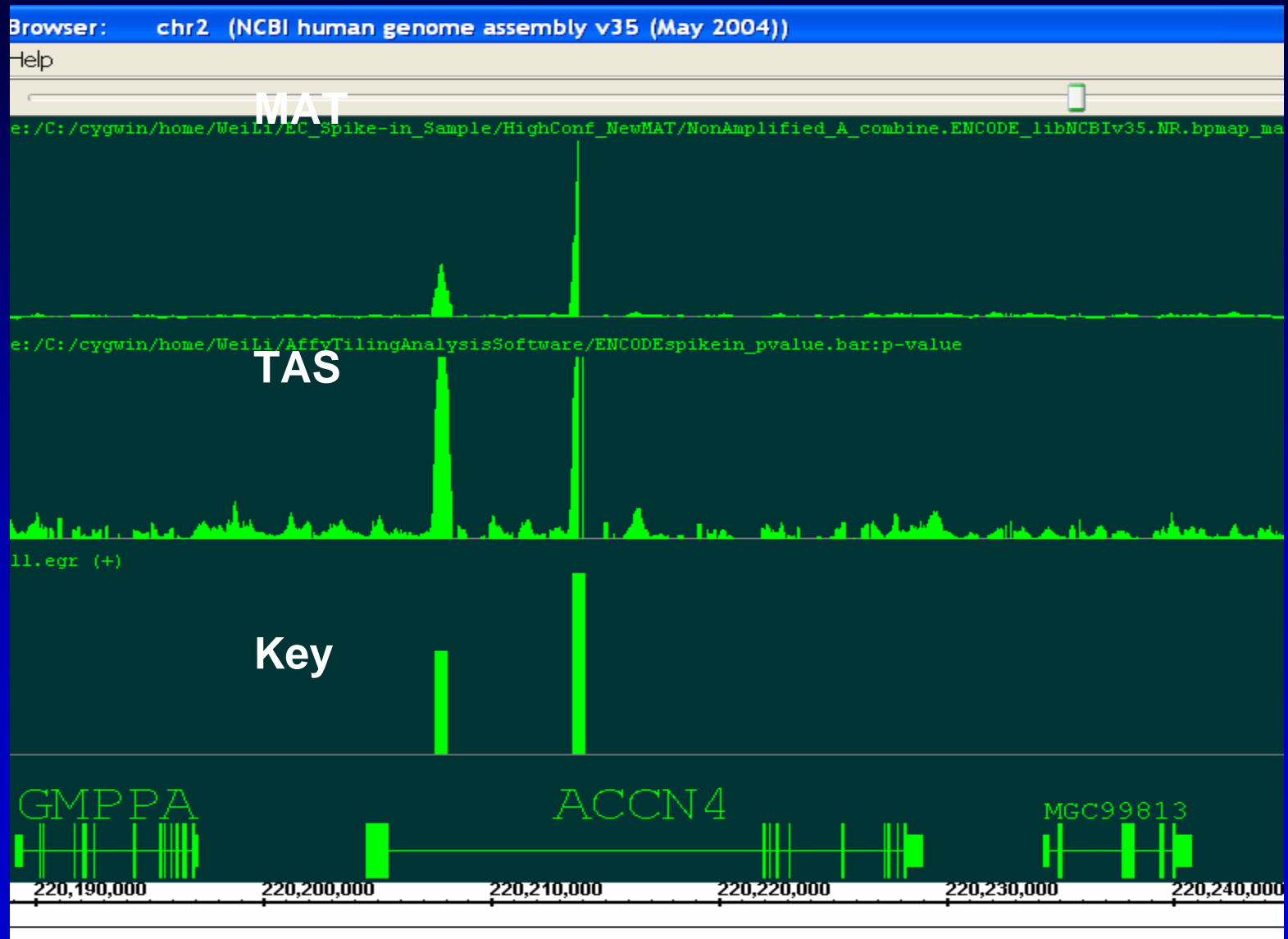
- **Blind test:**  
Samples hybridized to different tiling arrays,  
predictions made before the key was released

# ENCODE Spike-in



- **MAT predicts 70% targets with single spike-in (no ctrl)**
- **100% target detection accuracy from 5 ChIP / 5 Ctrl**

# MAT Scores are Quantitative

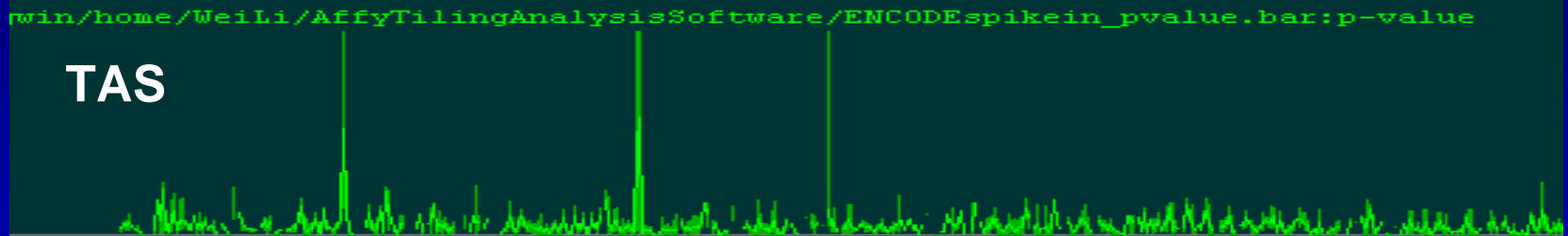


chr1 (NCBI human genome assembly v35 (May 2004))

MAT



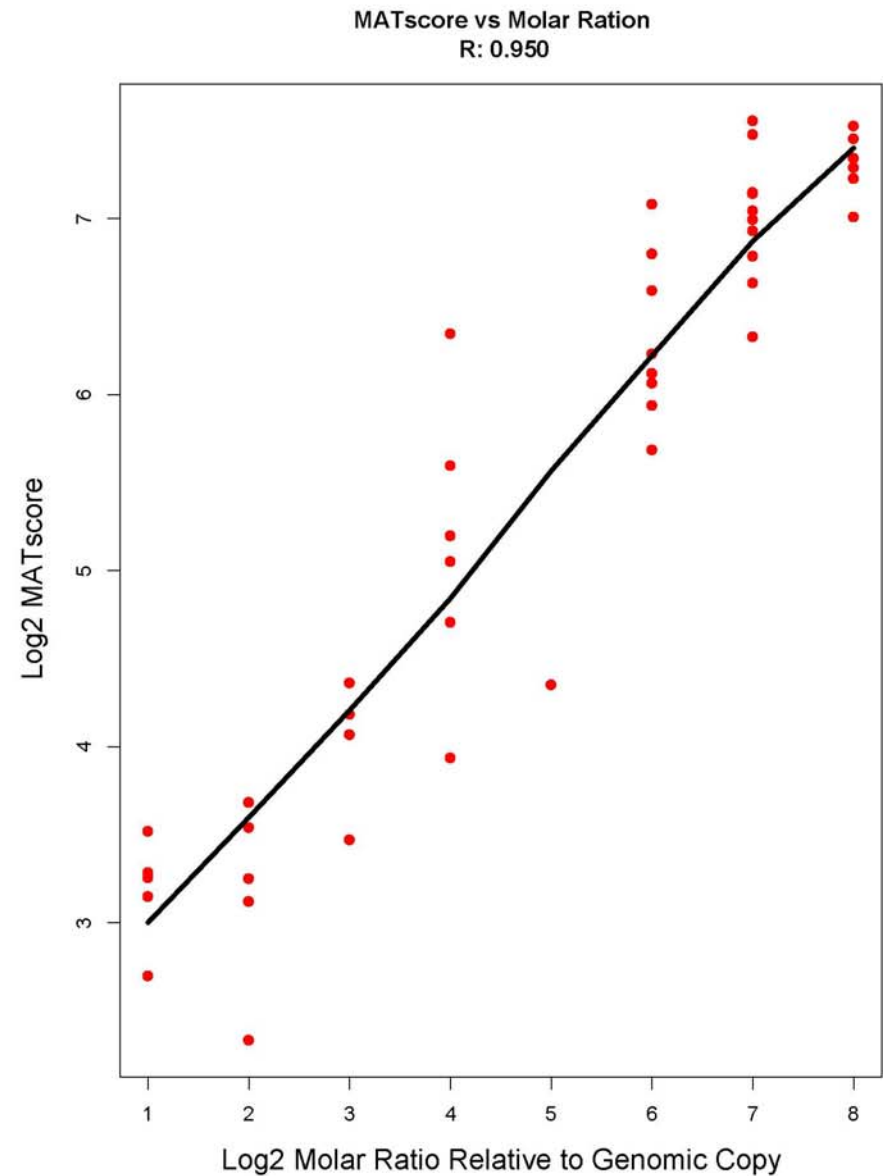
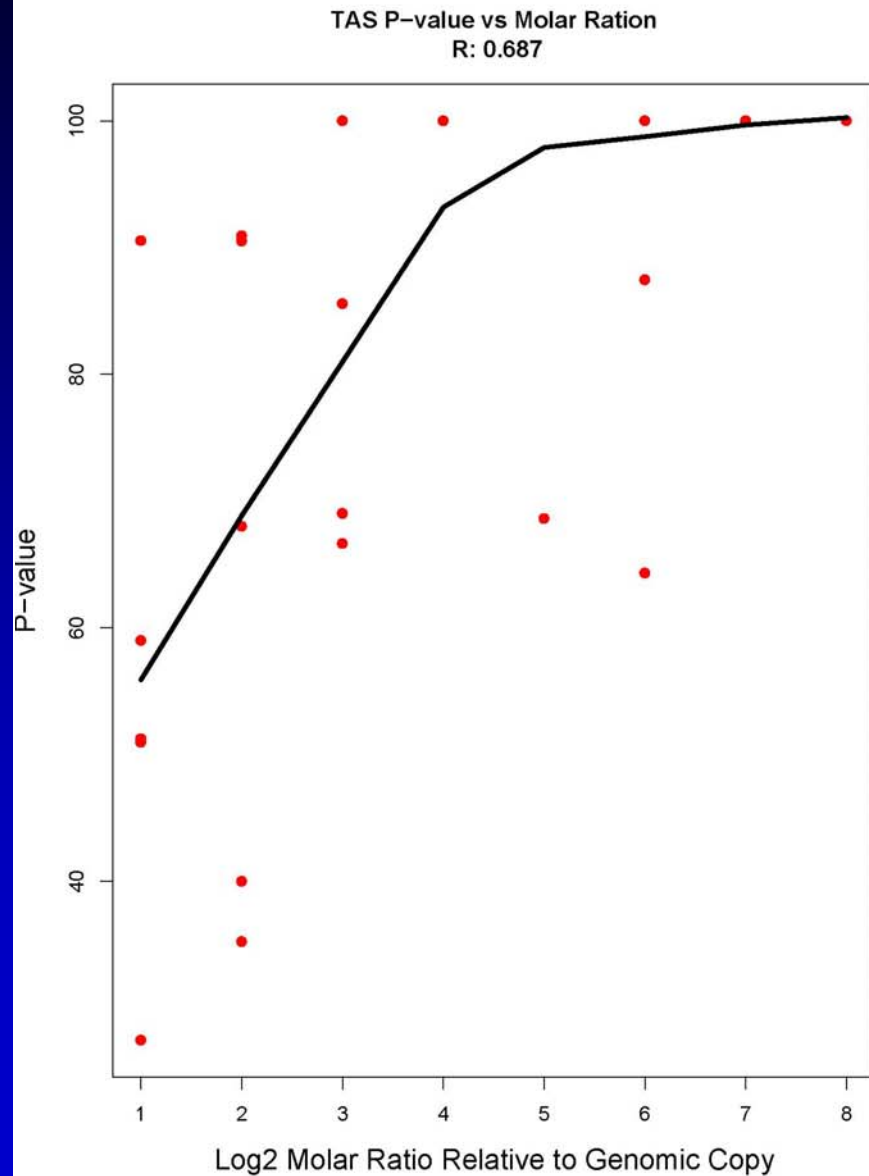
TAS



Key



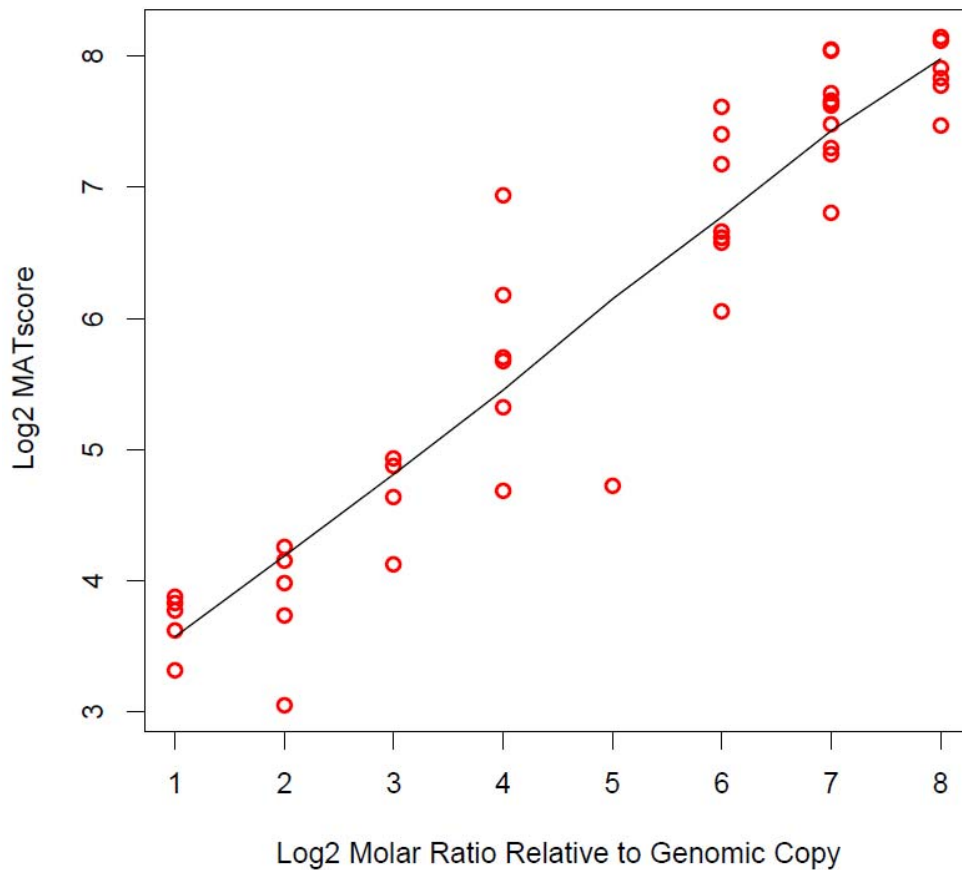
# Quantitative Assessment of Predictions



# Amplification Effects

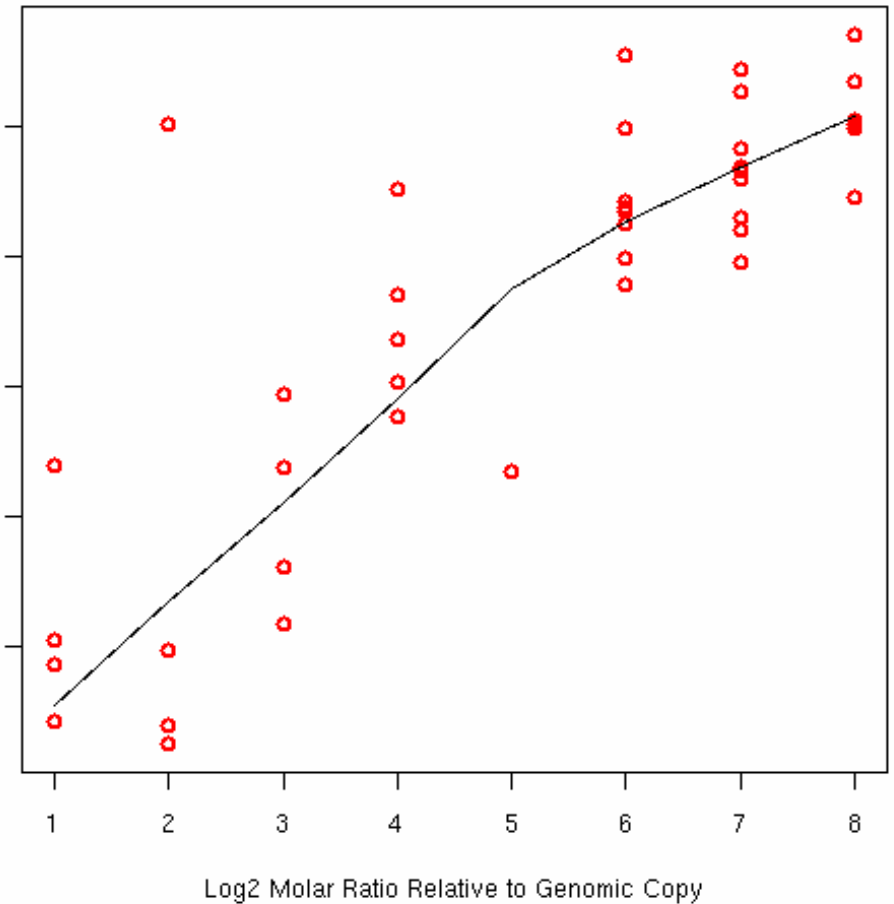
## Non-Amplified

MATscore vs Molar Ratio  
R: 0.950



## PCR Amplified

MATscore vs Molar Ratio  
R: 0.844



# Outline

- ChIP-chip on genome tiling microarrays
- MAT: Model-based Analysis of Tiling arrays
- **How to use MAT**
- Interpreting the results

# How to Use MAT

- Download MAT software and install.

<http://chip.dfci.harvard.edu/~wli/MAT/index.html>

- Download library files.

- Using text editor create .tag file: *my\_experiment.tag*

- Run MAT:

*MAT my\_experiment.tag*

- Examine signal using Integrated Genome Browser.

- Use .bed file for downstream analysis.

**[data]**

BpmapFolder = /home/jjoyce/database/bpmap

CelFolder = /home/jjoyce/celfiles

GenomeGrp =

RepLib = /home/jjoyce/Humanhg17Rep.lib

Group = 111000

**[bpmap]**

1 = P1\_CHIP\_A.Anti-Sense.hs.NCBIv35.NR.bpmap

2 = P1\_CHIP\_B.Anti-Sense.hs.NCBIv35.NR.bpmap

**[cel]**

1 = IP1.A.cel IP2.A.cel IP3.A.cel input1.A.cel input2.A.cel input3.A.cel

2 = IP1.B.cel IP2.B.cel IP3.B.cel input1.B.cel input2.B.cel input3.B.cel

**[intensity analysis]**

BandWidth = 300

MaxGap = 300

MinProbe = 10

Var = 0

Tvalue = 0

**[interval analysis]**

Pvalue = 1e-5

**[log]**

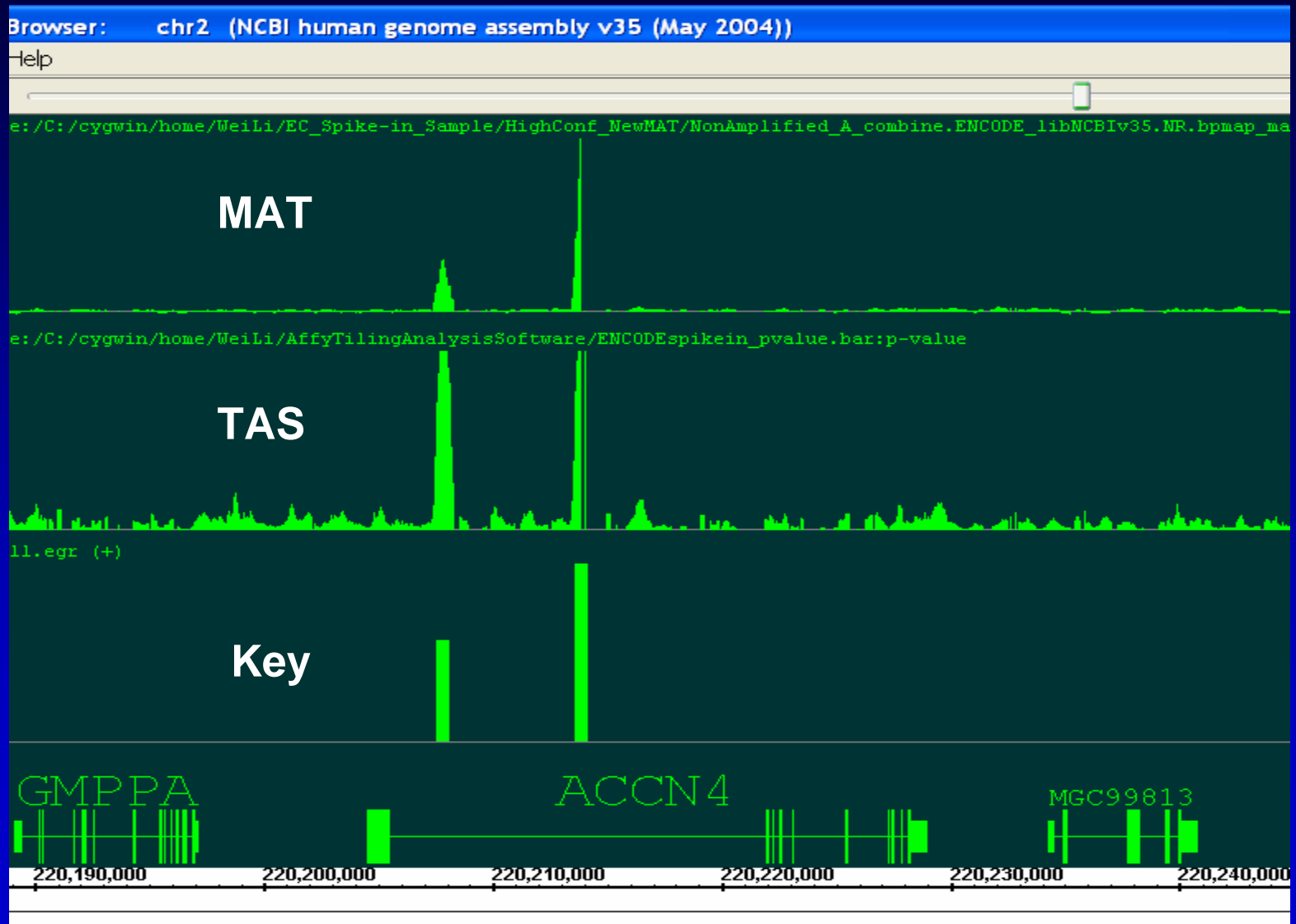
log = /home/jjoyce/test.log

# .bed File Output

browser position chr21:9932928-31824417 track  
name="ER\_1" description="MATscore=4.24 Pvalue=1.00e-  
05 Qvalue=5.15e-03"

chr21	9932928	9933552	ER_1_R_Se5	51.60
chr21	10048111	10049111	ER_2_R_Se4	86.22
chr21	10203698	10204322	ER_3_R_Se2	50.88
chr21	13914938	13915562	ER_4_R_Se9	56.78
chr21	14599949	14601069	ER_5	178.71
chr21	15171511	15172634	ER_6	142.46
chr21	40681728	40682542	ER_7_R	54.75
chr21	45757427	45758544	ER_8	107.12
chr21	45853089	45854179	ER_9_R_Si	132.18
chr21	46432429	46433247	ER_10	62.61

# .bar File Output in IGB



# Outline

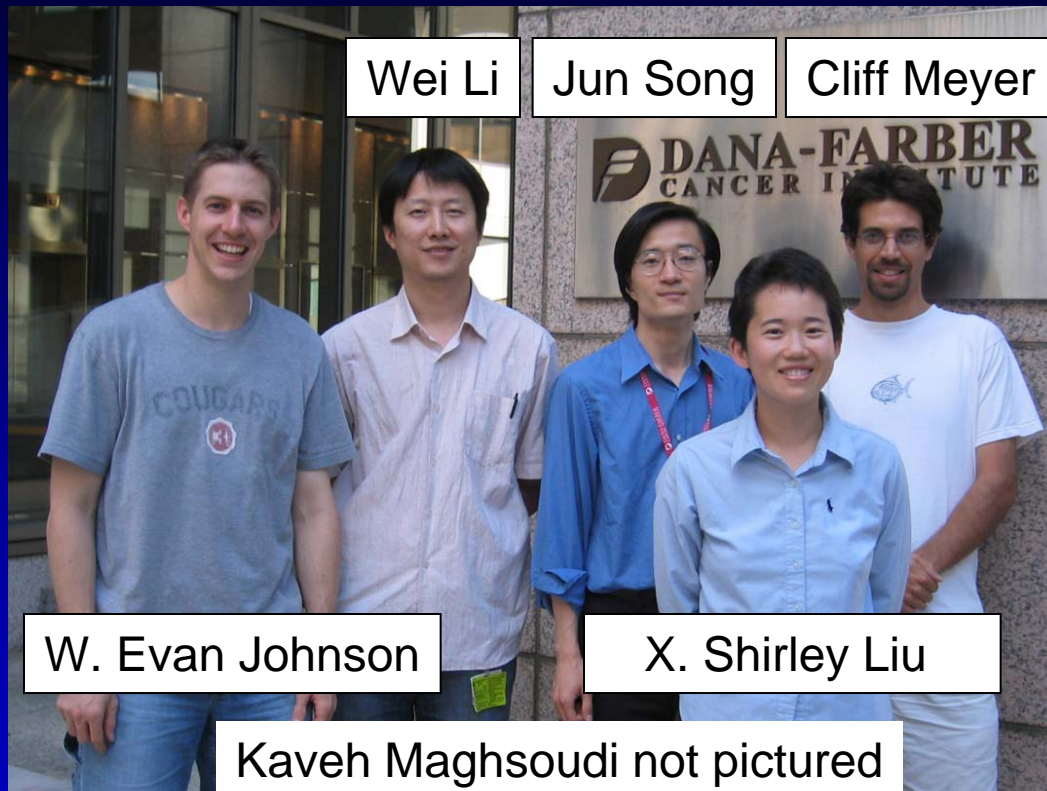
- ChIP-chip on genome tiling microarrays
- MAT: Model-based Analysis of Tiling arrays
- How to use MAT
- **Interpreting the results**

# Interpretation



- **Linking binding sites to annotated genomic loci.**
- **Integrating ChIP-chip with gene expression.**
- **Comparison of ChIP-chip experimental results.**
- **DNA sequence analysis.**
- **Comparative genomics analysis.**

# Acknowledgements



## Myles Brown Lab

Jason Carroll

Qianben Wang

Tim Geistlinger

- **Ed Fox**
  - Giles Hall
- **David Harrington**

## Funding

- NIH T90 DK070078-01