Tiling arrays, ChIP, and localizing protein-DNA interaction

Statistics 246

April 6, 2006

Richard Bourgon

bourgon@stat.berkeley.edu

Protein-DNA interaction

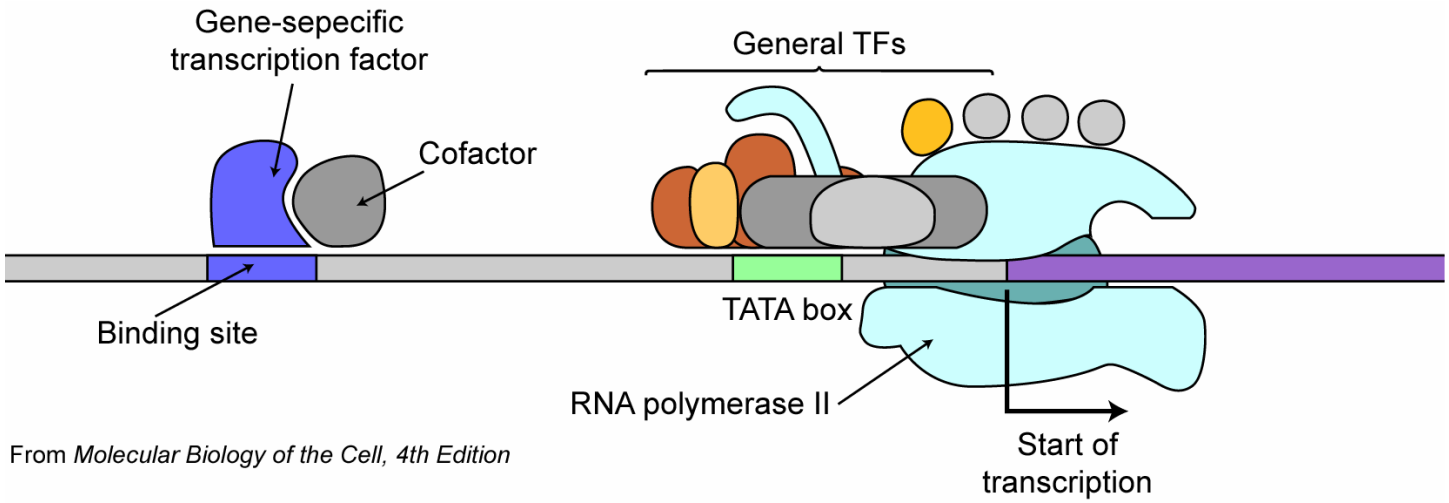
Protein-DNA interaction

Proteins interact with DNA to

- Carry out transcription of “activated” genes.
- Carry out DNA replication.
- Repair damaged DNA.
- Mediate recombination in meiosis.
- Modify or “remodel” the chromatin.
- Enhance or suppress gene transcription.
- Etc.

Transcription factor (TF) proteins regulate gene expression, and recognize short, degenerate motifs in the DNA.

ChIP-chip and protein-DNA interaction



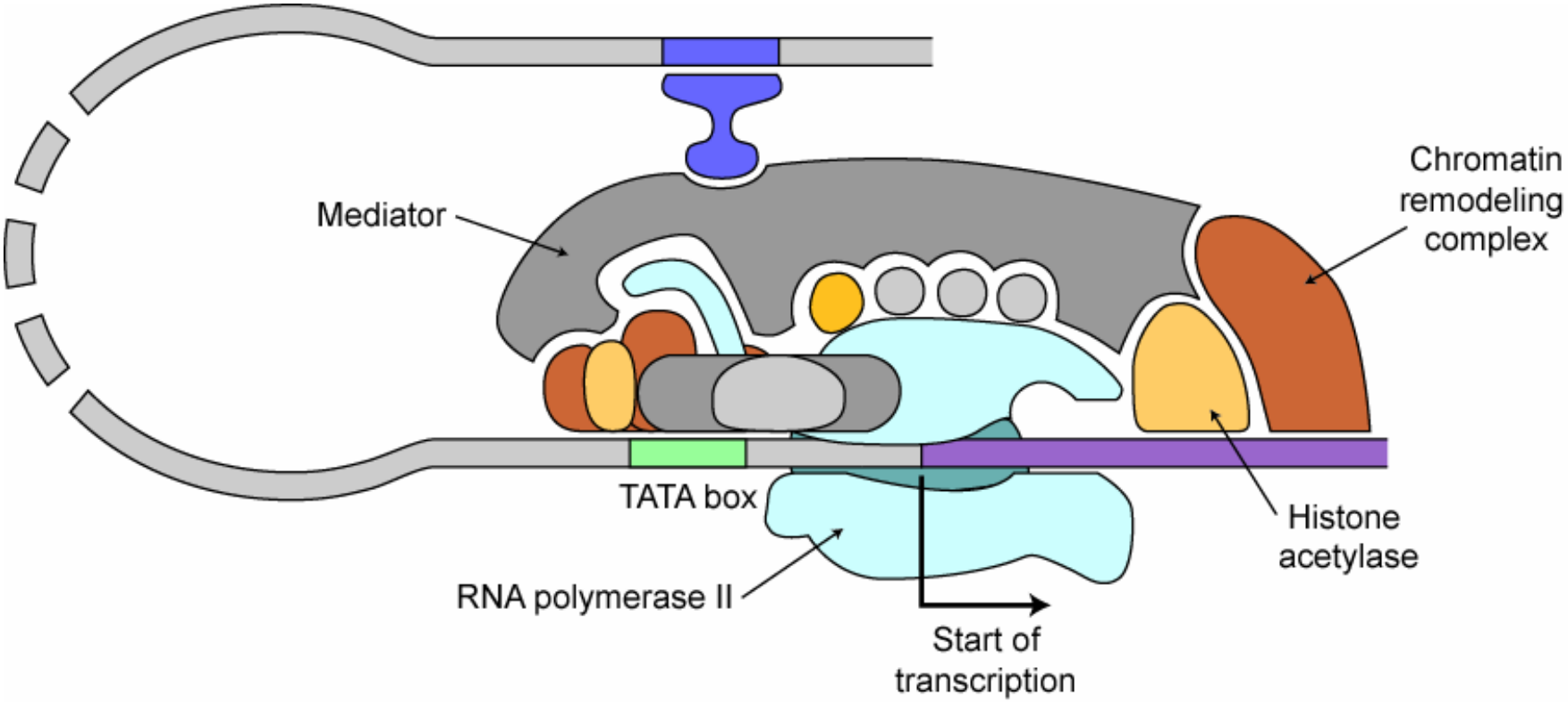
From *Molecular Biology of the Cell, 4th Edition*

ChIP-chip permits *in vivo*, genome-wide localization of transcription factor binding sites.

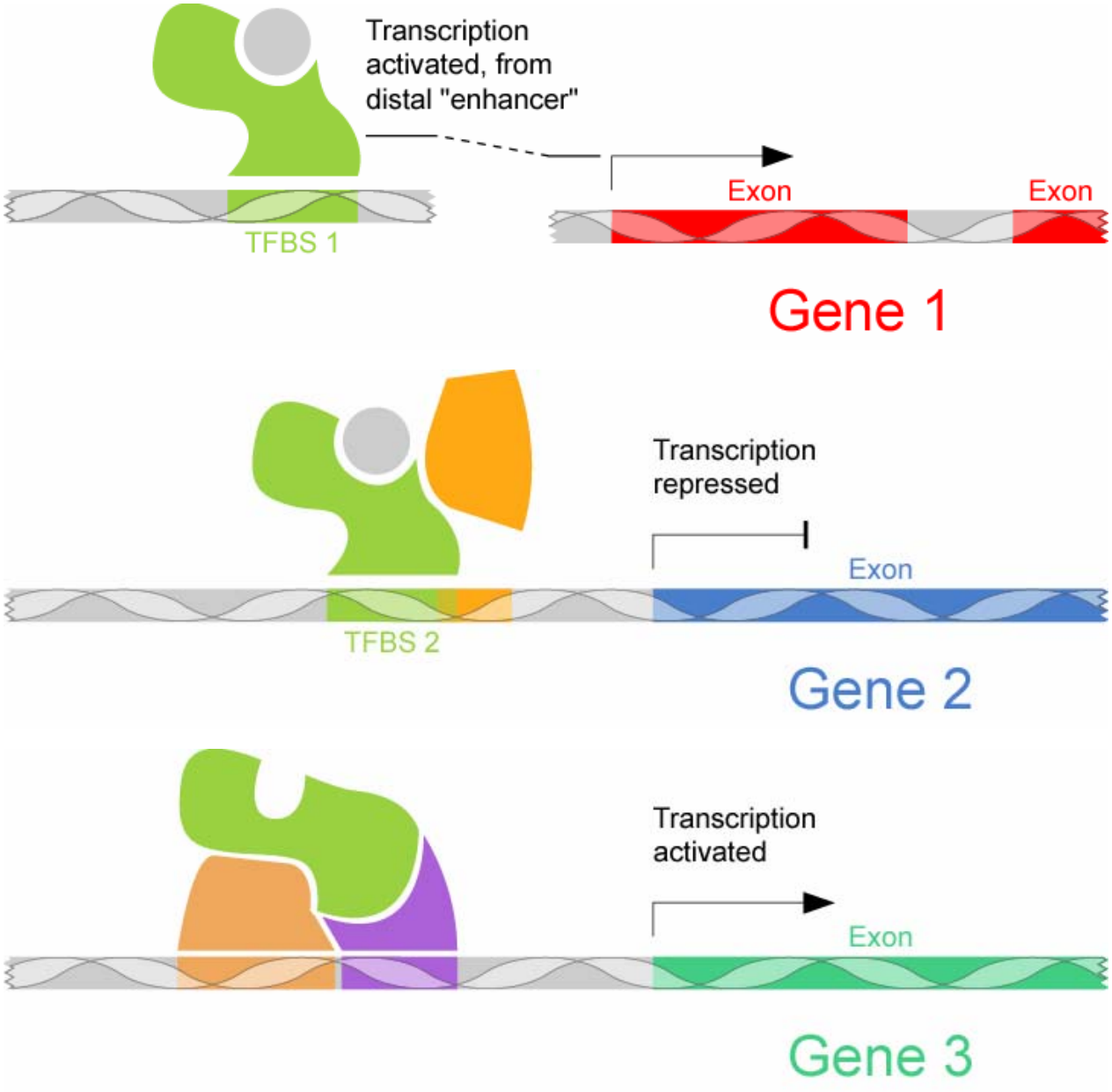
Other applications:

- Localization of transcriptional machinery.
- Histone modifying or chromatin remodeling proteins, or the modified (e. g., methylated) forms themselves.
- Origin recognition complexes.

Enhancers


Mediator
Chromatin
remodeling
complex
TATA box
Histone
acetylase
RNA polymerase II
Start of
transcription

Transcription factors and expression regulation

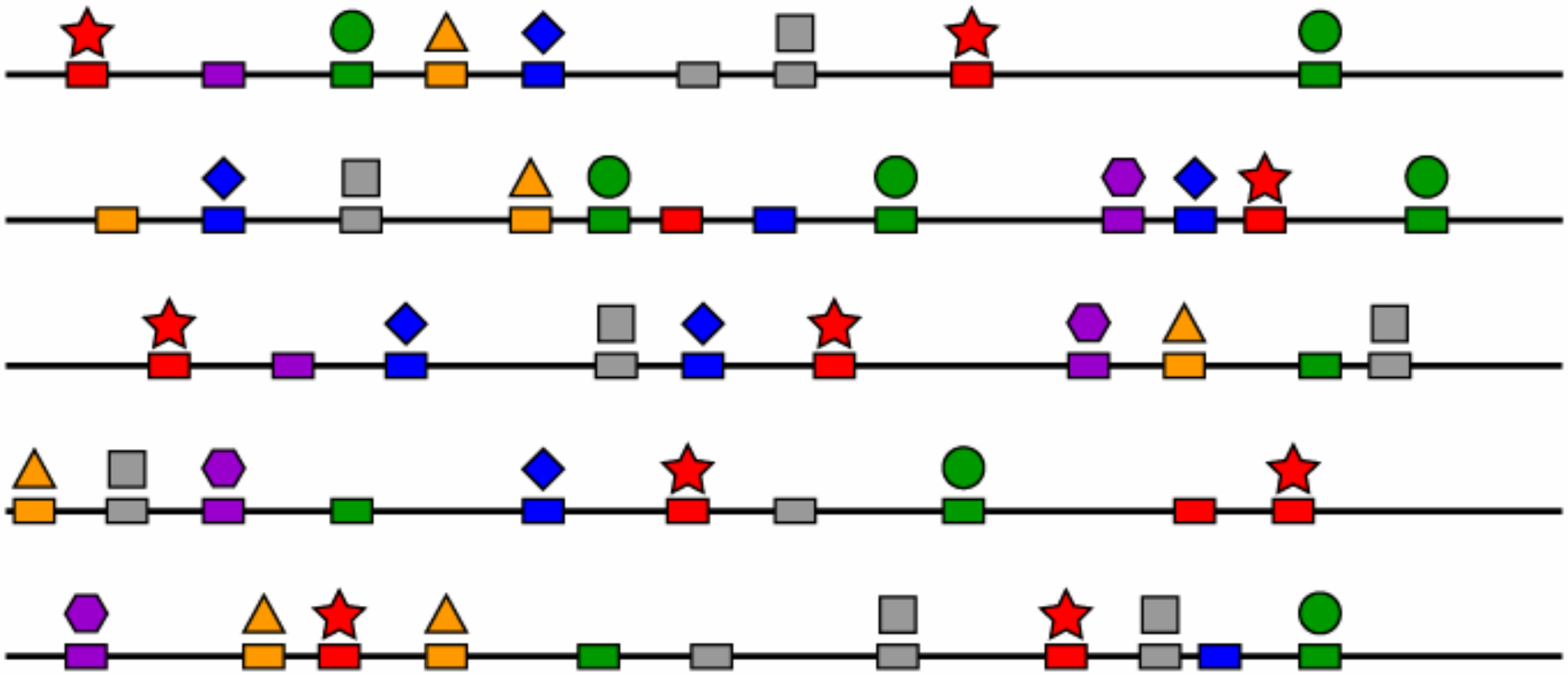


Identification of TF binding sites

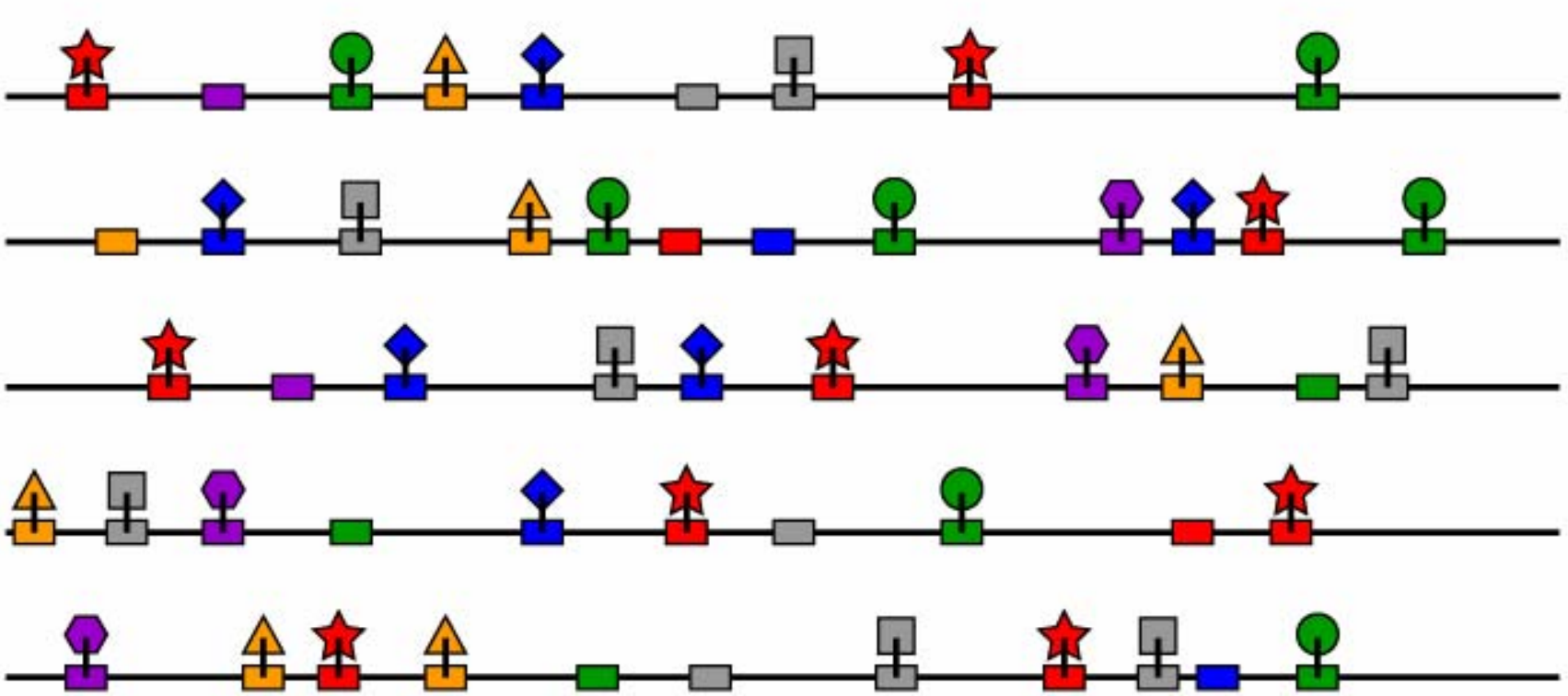
- *In vitro*?
  - Oligo-selection or gel-shift assays are often poor predictors of *in vivo* binding.
- With expression arrays?
  - Change in expression may be through intermediaries.
  - If required co-factors aren't present, genes which are direct targets may not exhibit differential expression.
- *In silico*?
  - Consensus sites appear far too often.
  - Motifs are degenerate.

Chromatin immunoprecipitation

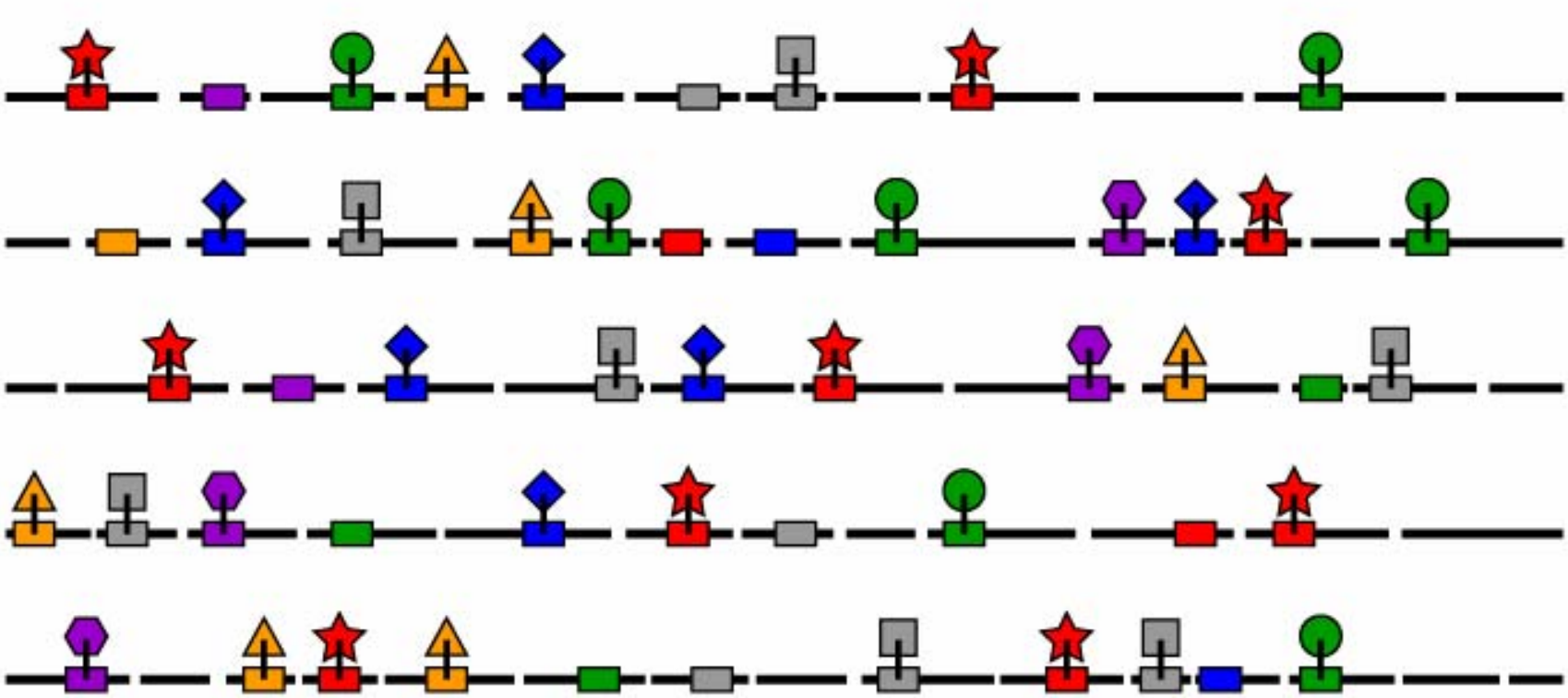
TFs associate with binding sites *in vivo*



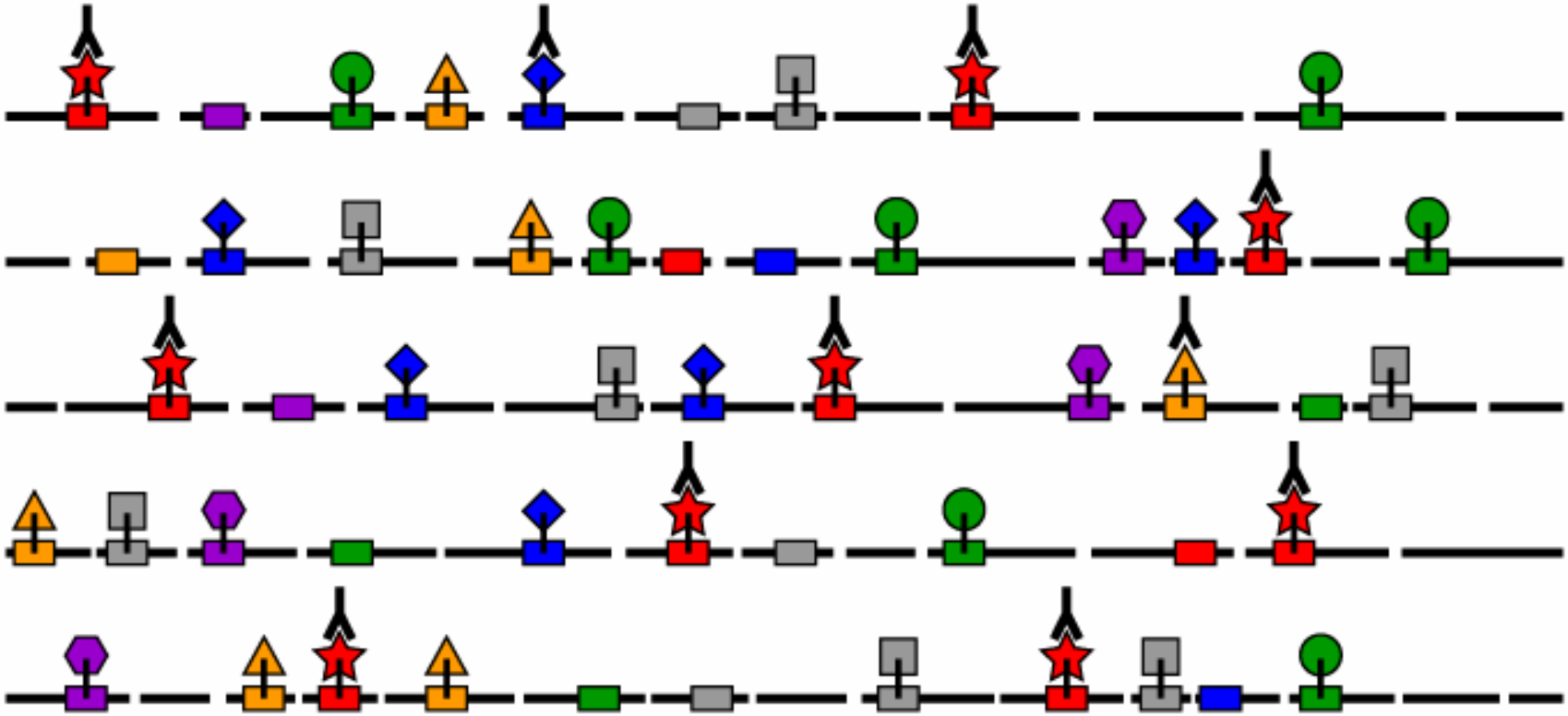
TF/DNA crosslinking *in vivo*



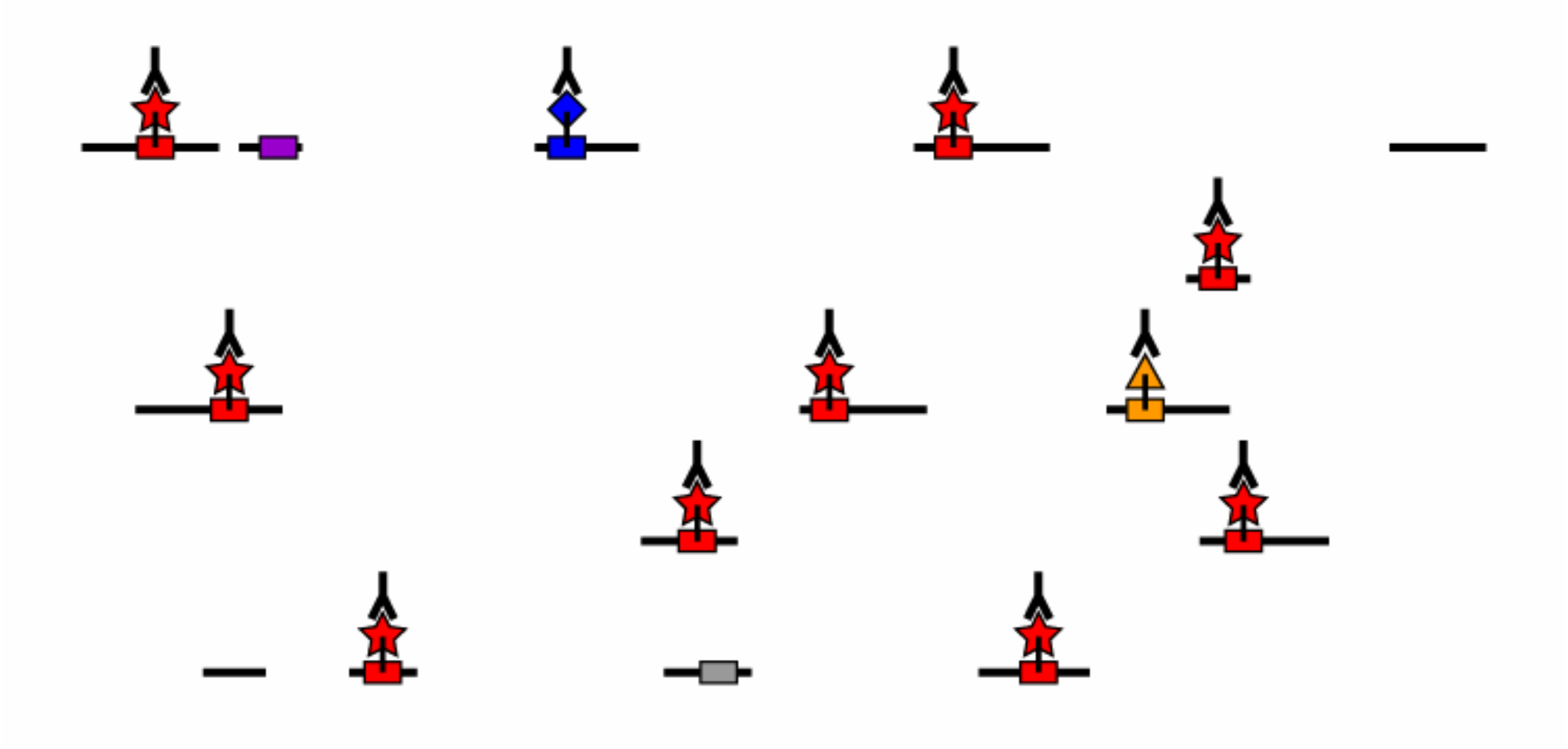
Sonication



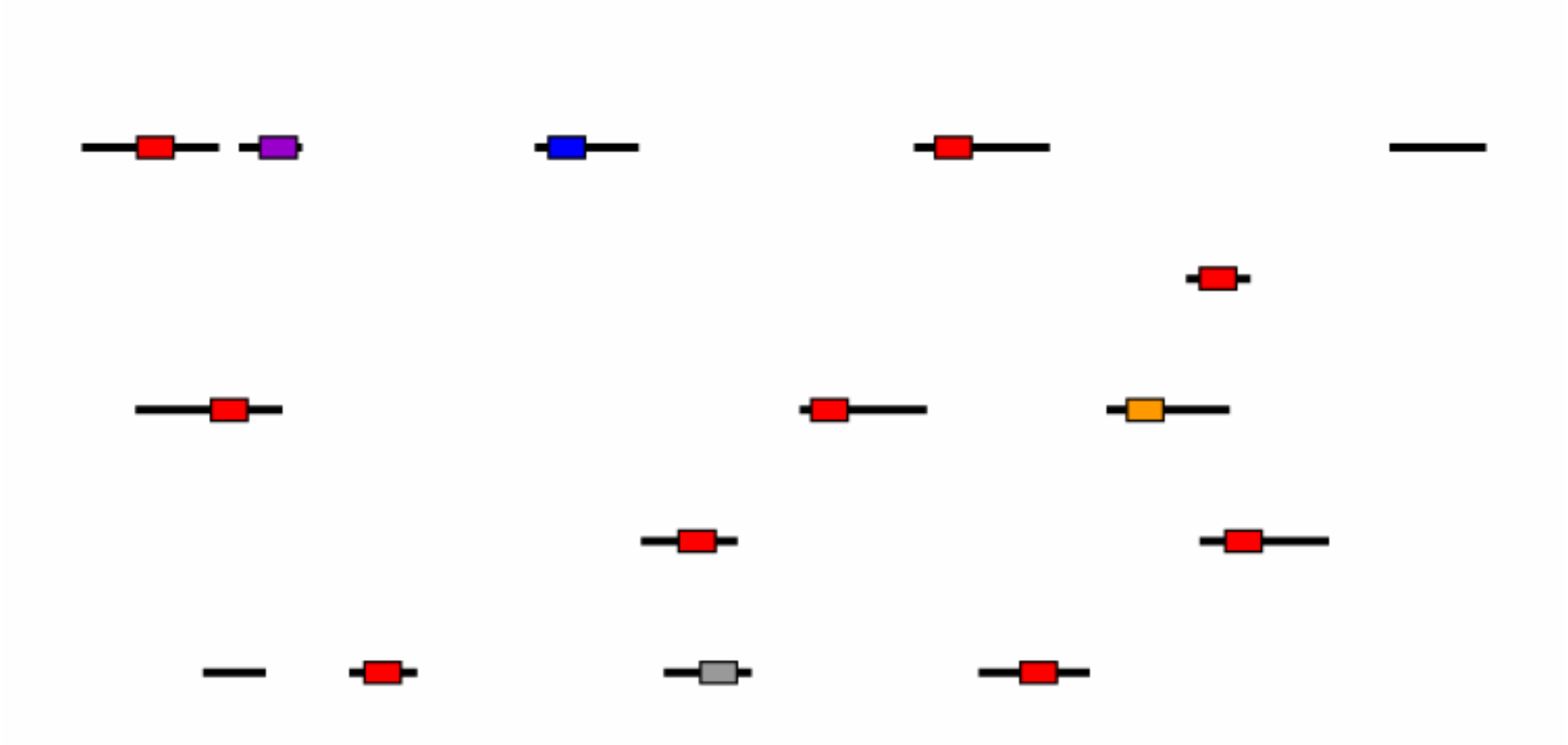
TF-specific antibody



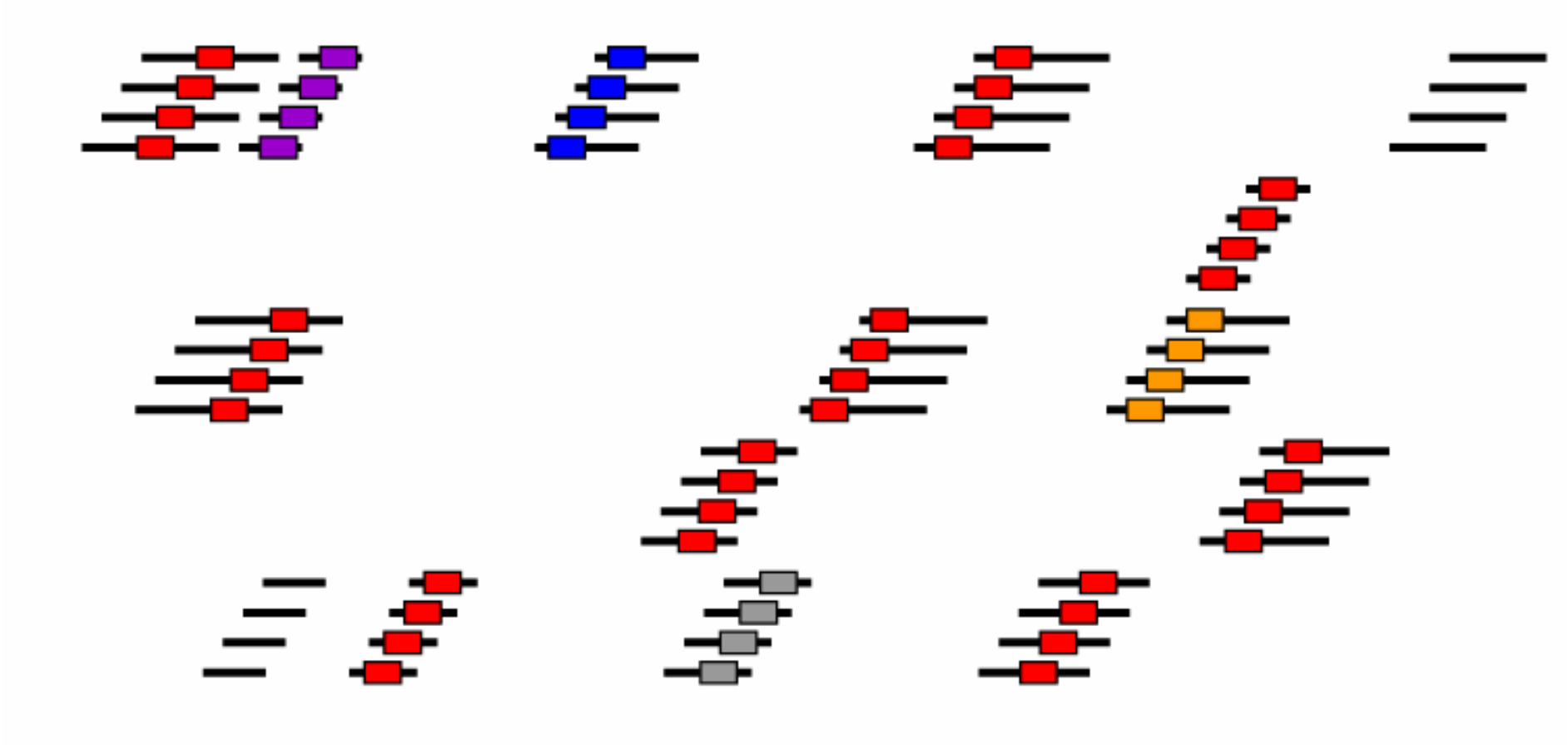
Immunoprecipitation



Crosslink reversal and DNA purification



Amplification



Chromatin immunoprecipitation

Summary of ChIP assay:

- Cross-link all proteins to genomic DNA.
- Fragment DNA (with cross-linked proteins still attached).
- ChIP: enrichment of TF-associated fragments.
- Reverse cross-linking and purify DNA.

Identification of enriched DNA regions:

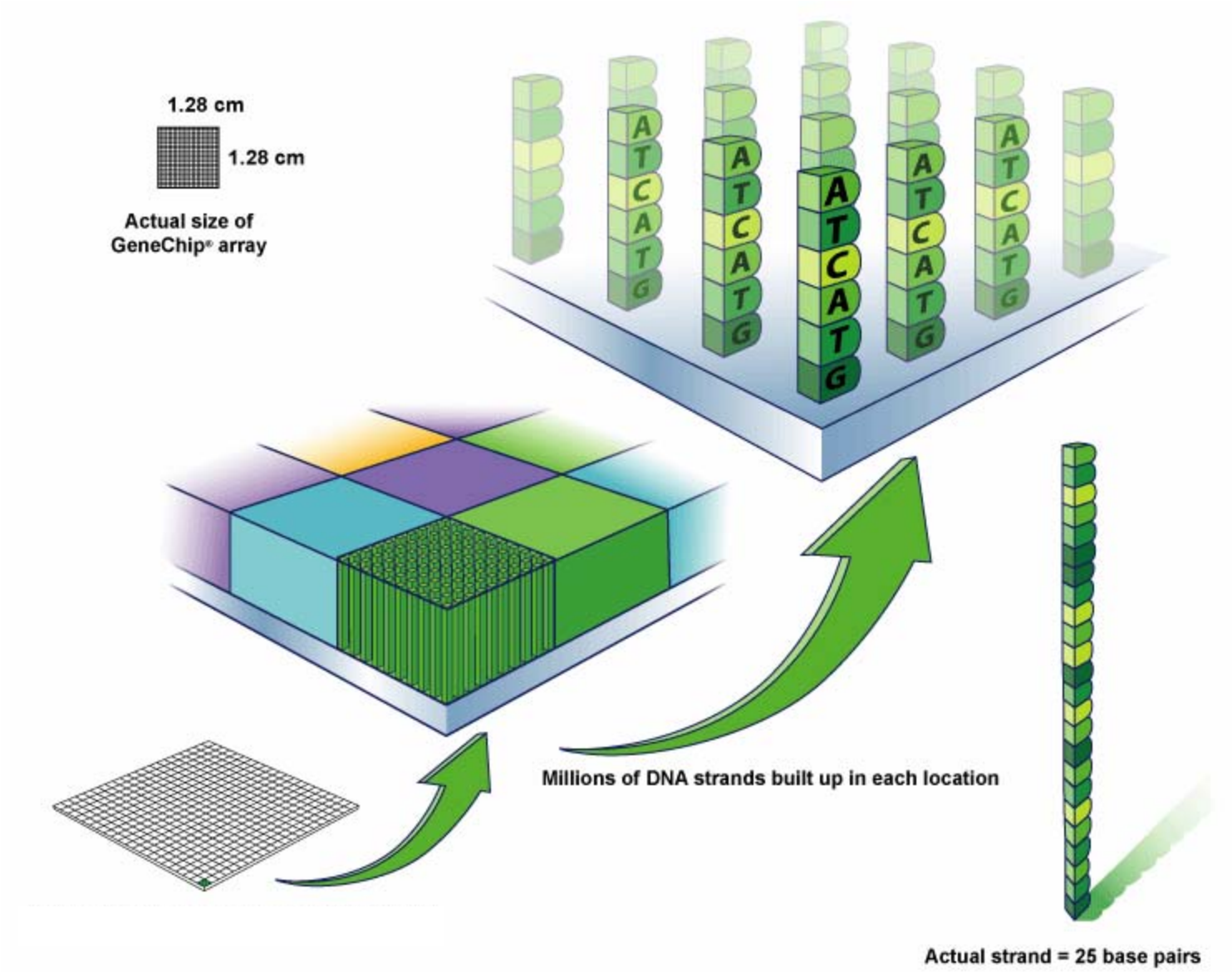
- Sequence and map back to the genome
- For anticipated binding sites, PCR with specific primers
- Amplify non-specifically, then hybridize to tiling array.

ChIP and tiling arrays

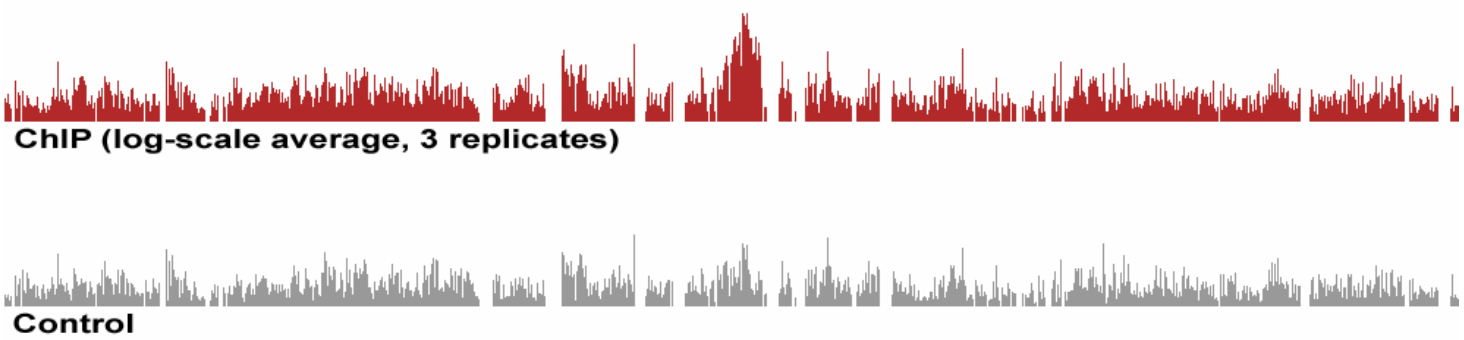
Early human ChIP-chip studies: spotted arrays

Paper	TF	Array
Ren *et al.* ’02	E2F1 E2F4	PCR products for 1.4K 5’-proximal promoters (≈1000bp)
Trinklein *et al.* ’04	HSF1	786 promoter clones, many with HSE motif
Mao *et al.* ’03	c-Myc	7.8K CpG island clones
Martone *et al.* ’03	NF-κB	Chr 22, 21K PCR products
Euskirchen *et al.* ’04	CREB	″

Short oligonucleotide microarrays

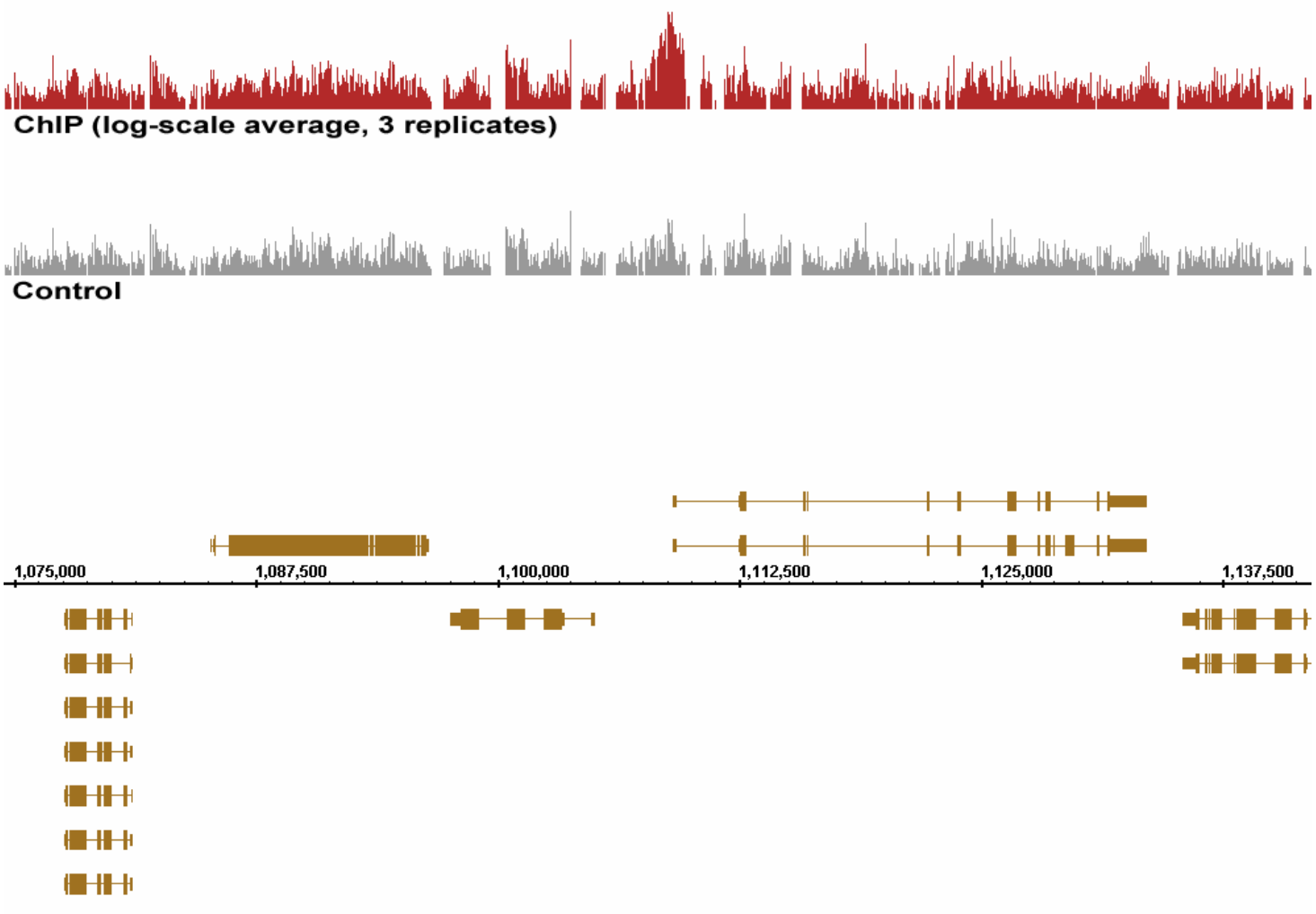


D. melanogaster tiling arrays

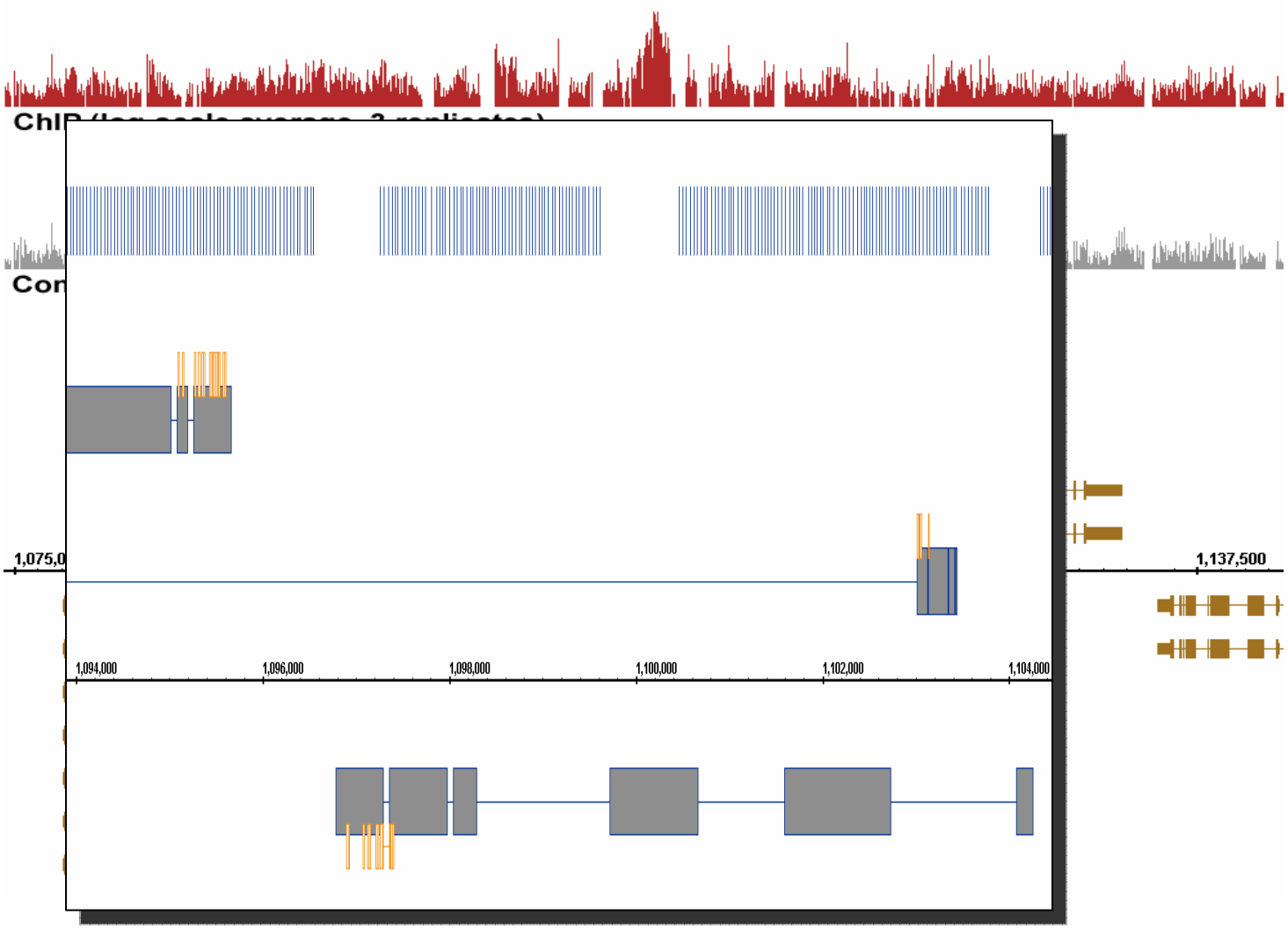


- 6M *in situ* synthesized 25-mer oligo probes.
- PM/MM pairs which differ by one base.
- Median distance between probe starts: 36 bp.
- Repetitive sequence is omitted.
- Probes with expected hybridization or synthesis problems are omitted.

D. melanogaster tiling arrays



3’-biased expression arrays vs. tiling arrays



Target abundance and measured fluorescence intensity

Definitions

Target abundance (A_{ij})

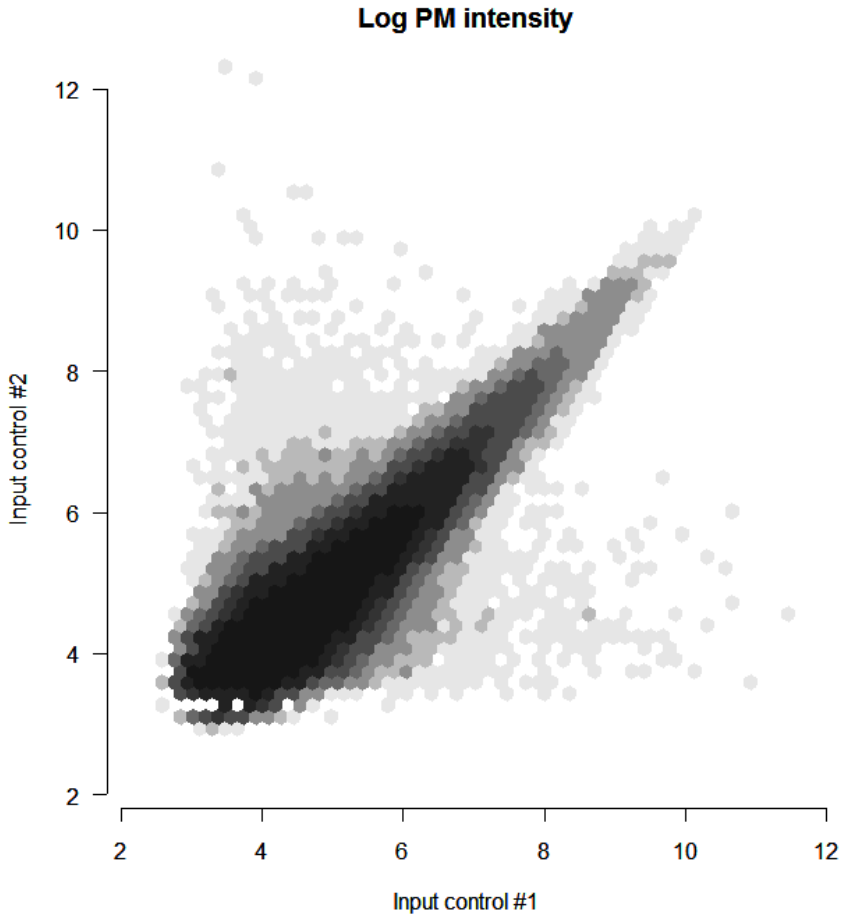
- The **unobservable** number of DNA fragments in sample *j* which contain sequence complementary to the probes in feature *i*.

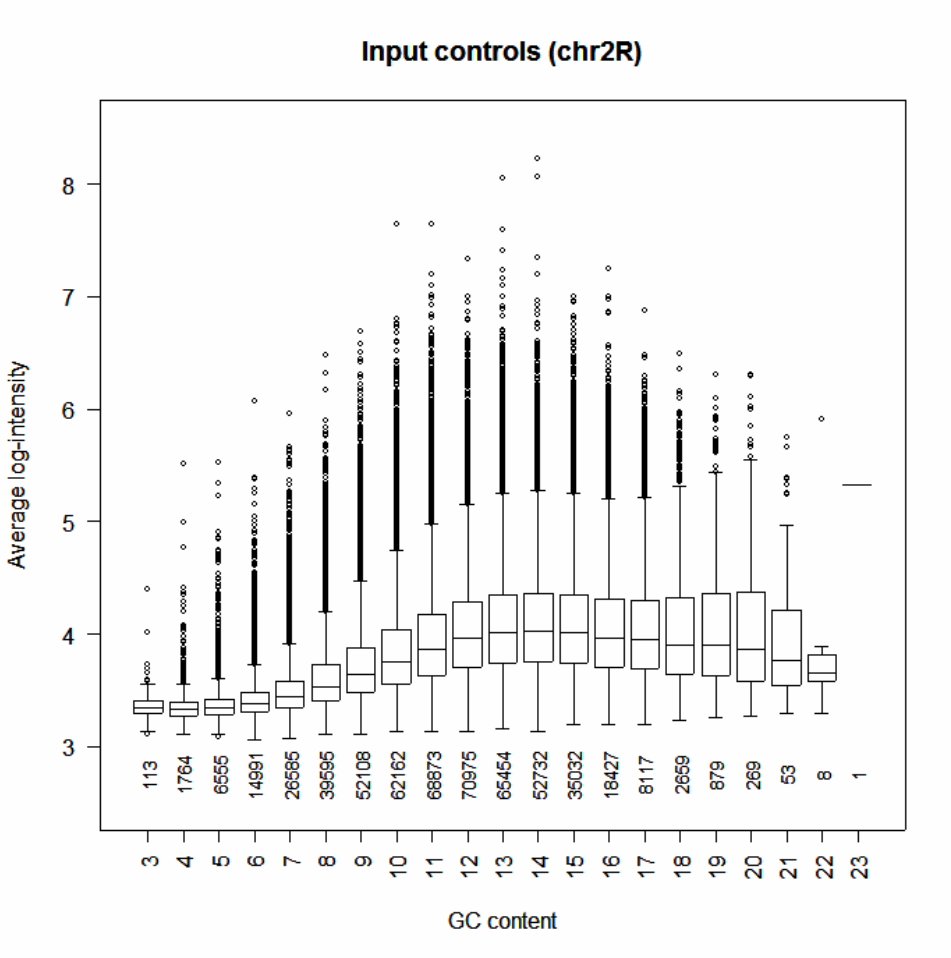
Fluorescence intensity (I_{ij})

- The **observable**, scanned intensity reading for feature *i*, sample *j*.

Abundance and intensity are related, but not in a simple way…

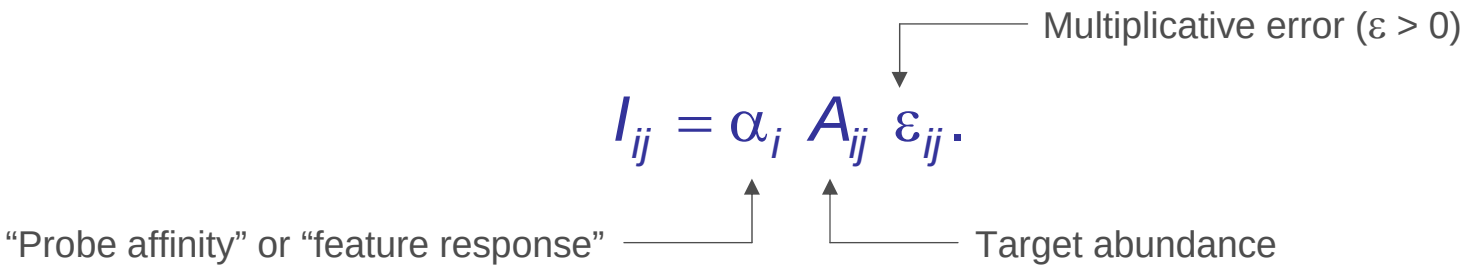
Probe affinity effects





The simplest model…

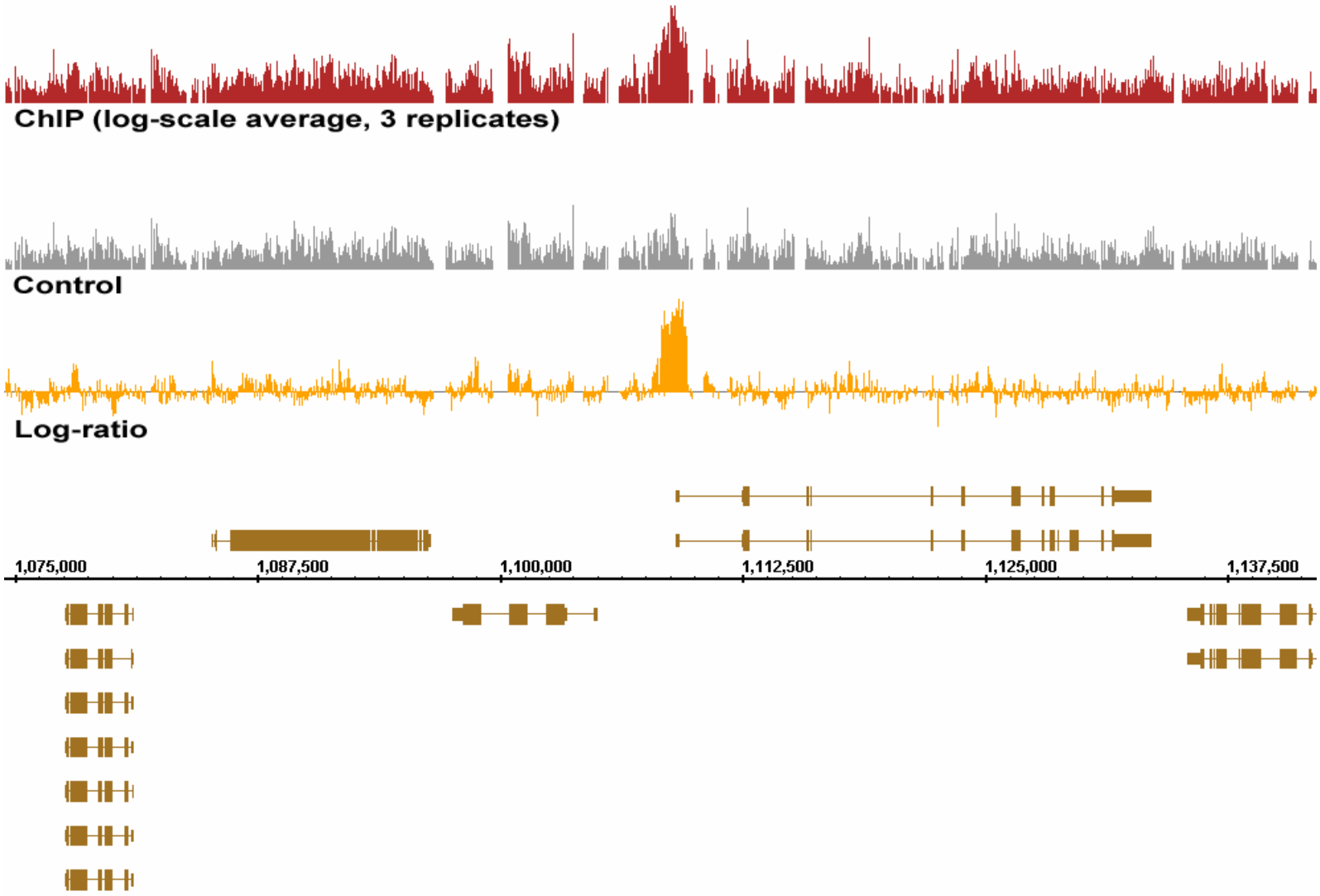
For probe *i* of sample *j*, assume that



When control data are available, we can eliminate the probe affinity effects with a ratio of intensities:

$$LR_i = \log A_i^T - \log A_i^C + \delta_i$$

Improved signal-to-noise ratio



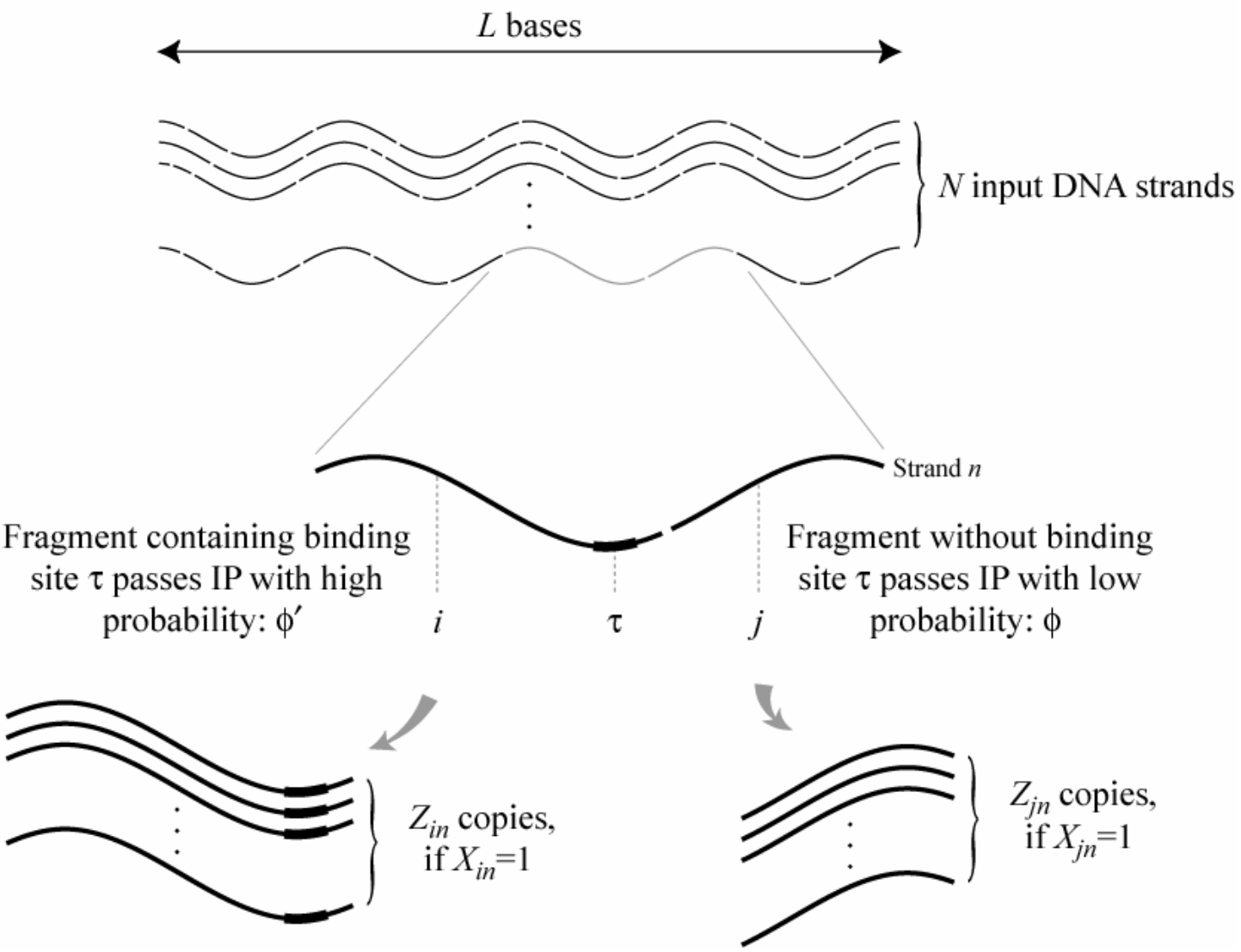
A statistical model

Abundance, intensity, and derived statistics

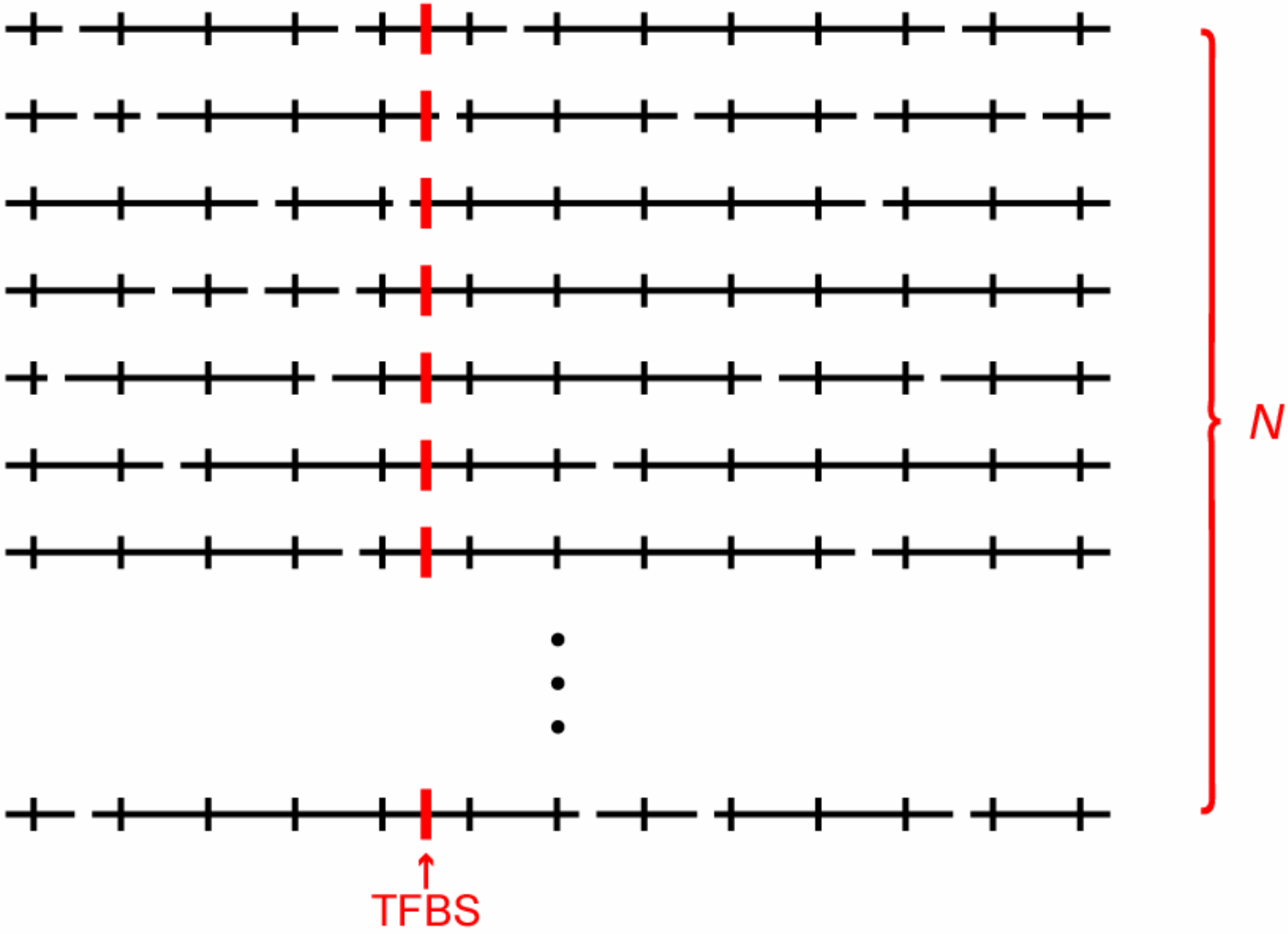
A model for the assay

Step	*Model*
Source material	N strands of extracted DNA.
Sonication	Uniform fragmentation of chromatin, with no interference. Probability of a break at any base is θ.
IP	Fragments with no binding site pass with probability ϕ; fragments with a binding site pass with probability ϕ', and $\phi' \gg \phi$.
Amplification	Z, a random multiplier for each fragment passing IP. (For PCR, Z is a branching process with t cycles and efficiency p.)

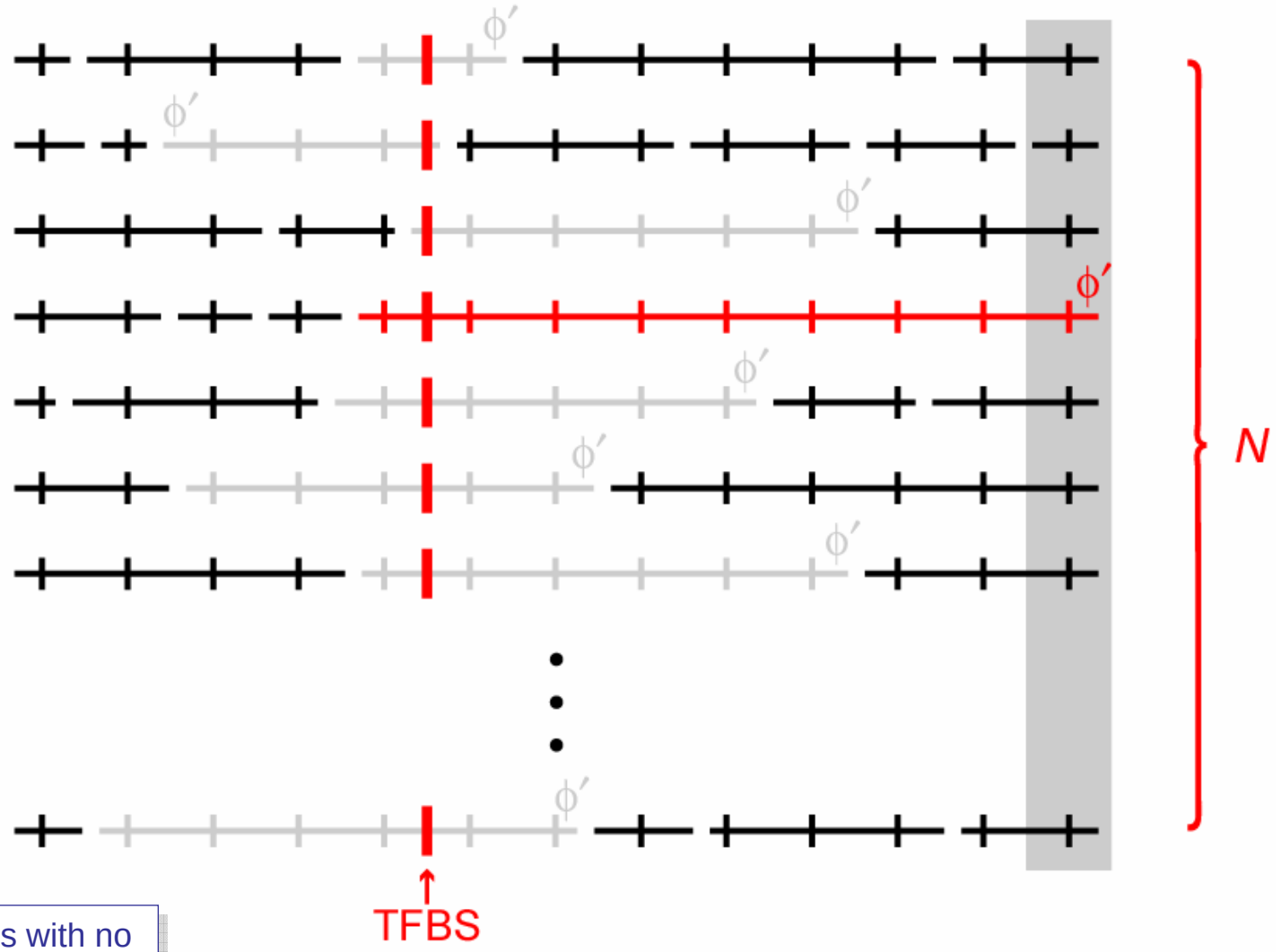
A model for the assay



Sonication

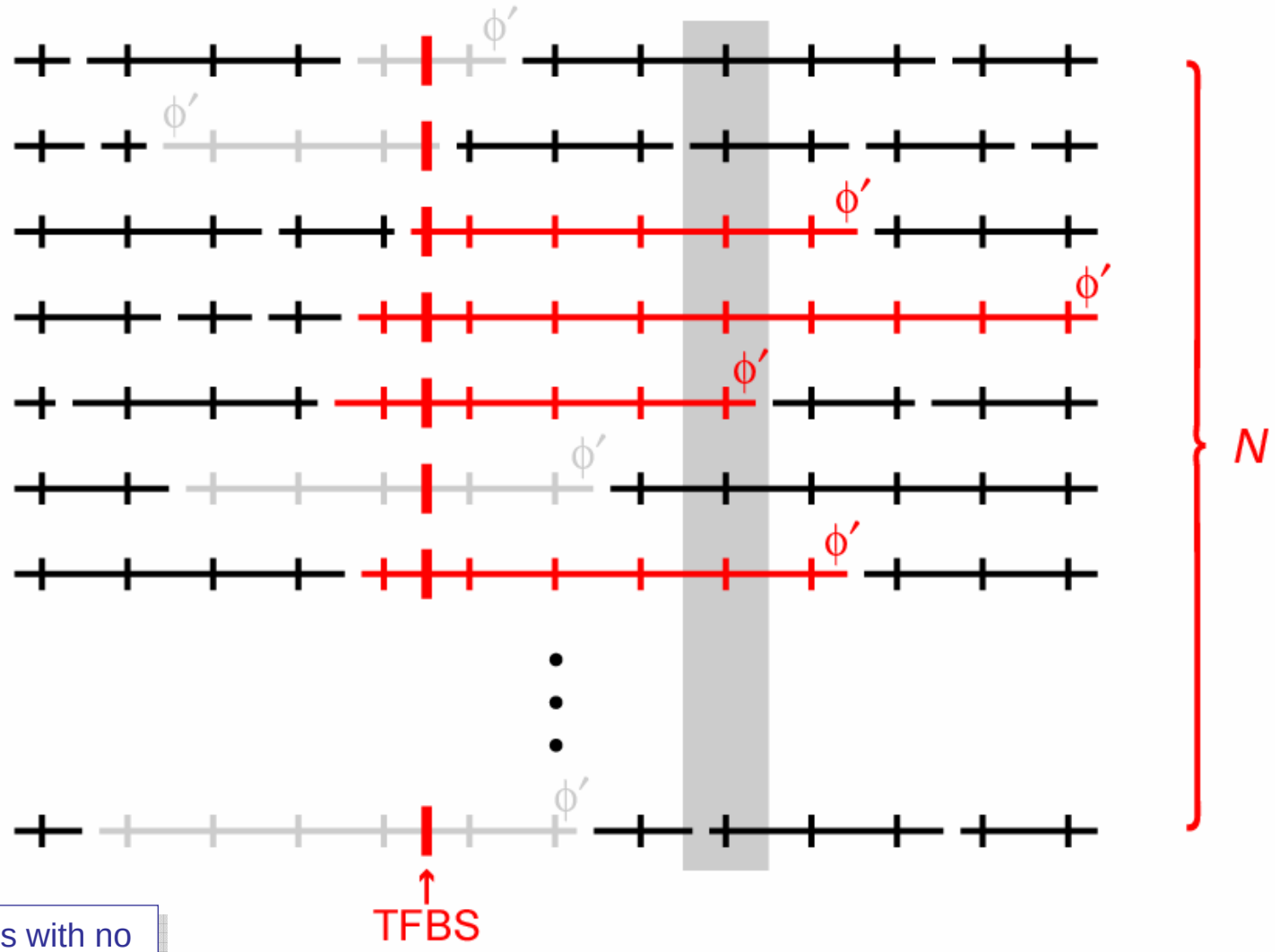


Immunoprecipitation



All fragments with no TF binding site pass with probability ϕ

Immunoprecipitation (closer to binding site)



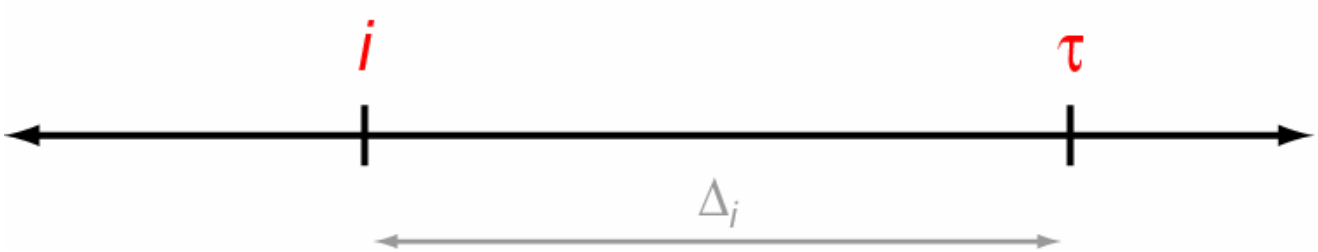
All fragments with no TF binding site pass with probability ϕ

Expected fragment size

For a single strand of length L, assume M breaks, making fragments of length $F_1, \ldots, F_M$, with $\sum_m F_m = L$.

$$\begin{aligned}
\mathbb{E}\bar{F} &= \mathbb{E}\left(\mathbb{E}(\bar{F} \mid M)\right) \\
&= \mathbb{E}\left(\frac{L}{M+1}\right) \\
&= \sum_{m=0}^{L} \frac{L}{m+1}\binom{L}{m}\theta^m(1-\theta)^{L-m} \\
&= \frac{1}{\theta}\frac{L}{L+1}\left(1-(1-\theta)^{L+1}\right) \\
&\approx \frac{1}{\theta}.
\end{aligned}$$

Expected target abundance



Under the model, the abundance of fragments available for hybridization to probe *i* is…

$$A_i = \sum_{n=1}^{N} X_{in} Z_{in}$$

X_{in}: 0/1 indicator: does fragment *n* pass IP?

Z_{in}: Amplification multiplier

If probe *i* is Δ_i bases from a binding site τ,

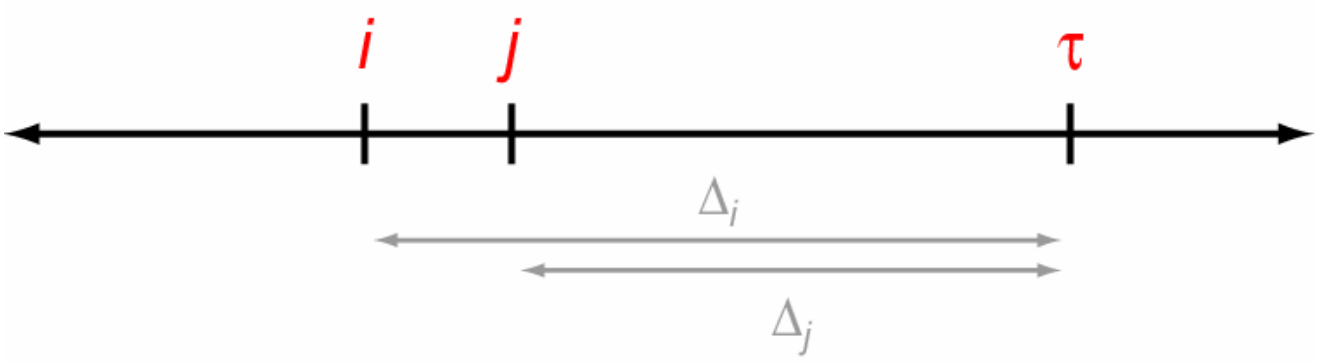
$$\mathbb{P}(X_{in} = 1) = (1-\theta)^{\Delta_i}\phi' + \left(1-(1-\theta)^{\Delta_i}\right)\phi$$

$$\equiv \pi(\Delta_i),$$

and then…

$$\mathbb{E}A_i = N\ \pi(\Delta_i)\ \mathbb{E}Z$$

Spatial correlation in target abundance



Now consider two probes, *i* and *j*. Away from binding sites, $\pi(\Delta_i) \approx \pi(\Delta_j) \approx \phi$ because $\pi(\Delta)$ quickly decays from ϕ' to ϕ. Under the model,

$$\mathrm{Corr}(A_i, A_j) = (1-\theta)^{d(i,j)} \sqrt{\frac{\pi(\Delta_j)\left(1-\pi(\Delta_j)+\mathrm{CV}^2(Z)\right)}{\pi(\Delta_i)\left(1-\pi(\Delta_i)+\mathrm{CV}^2(Z)\right)}}$$

$$\approx (1-\theta)^{d(i,j)}.$$

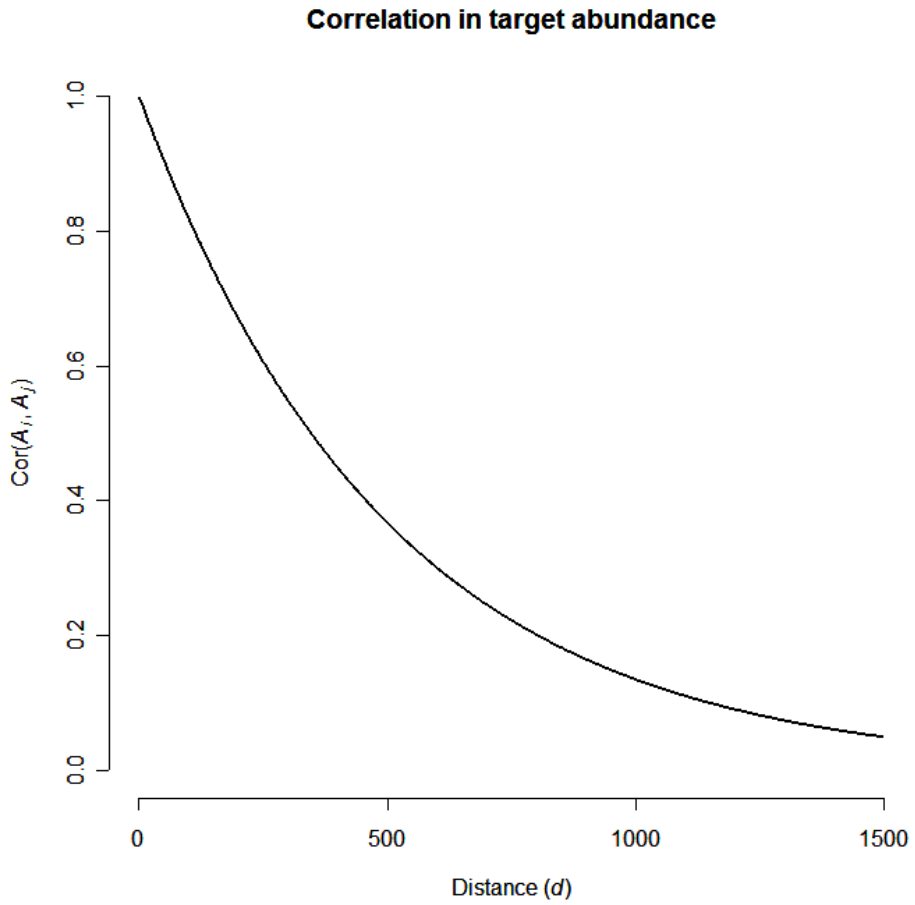
Spatial correlation in target abundance

Under the model, the expected fragment length after sonication is $1/\theta$.

For an average fragment length of 500 bases,

$$\mathrm{Corr}(A_i, A_j) \approx (1-\theta)^{d(i,j)}$$

is appreciable over a large number of probes in the tiling.



Intensity and log-ratio

Near a binding site, target abundance varies regularly with Δ_i, but intensity doesn't:

$$\mathbb{E}I_i = \alpha_i \mathbb{E}A_i = \alpha_i N\ \pi(\Delta)\ \mathbb{E}Z.$$

If the multiplicative intensity model is approximately correct, expansion of $\log\left(A_i / \mathbb{E}A_i\right)$ around 1 gives:

$$\begin{aligned}\mathbb{E}LR_i &= \mathbb{E}\log A_i^T - \mathbb{E}\log A_i^C + \mathbb{E}\delta_i \\ &\approx \log\frac{\pi^T(\Delta_i)}{\pi^C(\Delta_i)} + \log\frac{\mathbb{E}Z^T}{\mathbb{E}Z^C} + \mathbb{E}\delta_i .\end{aligned}$$

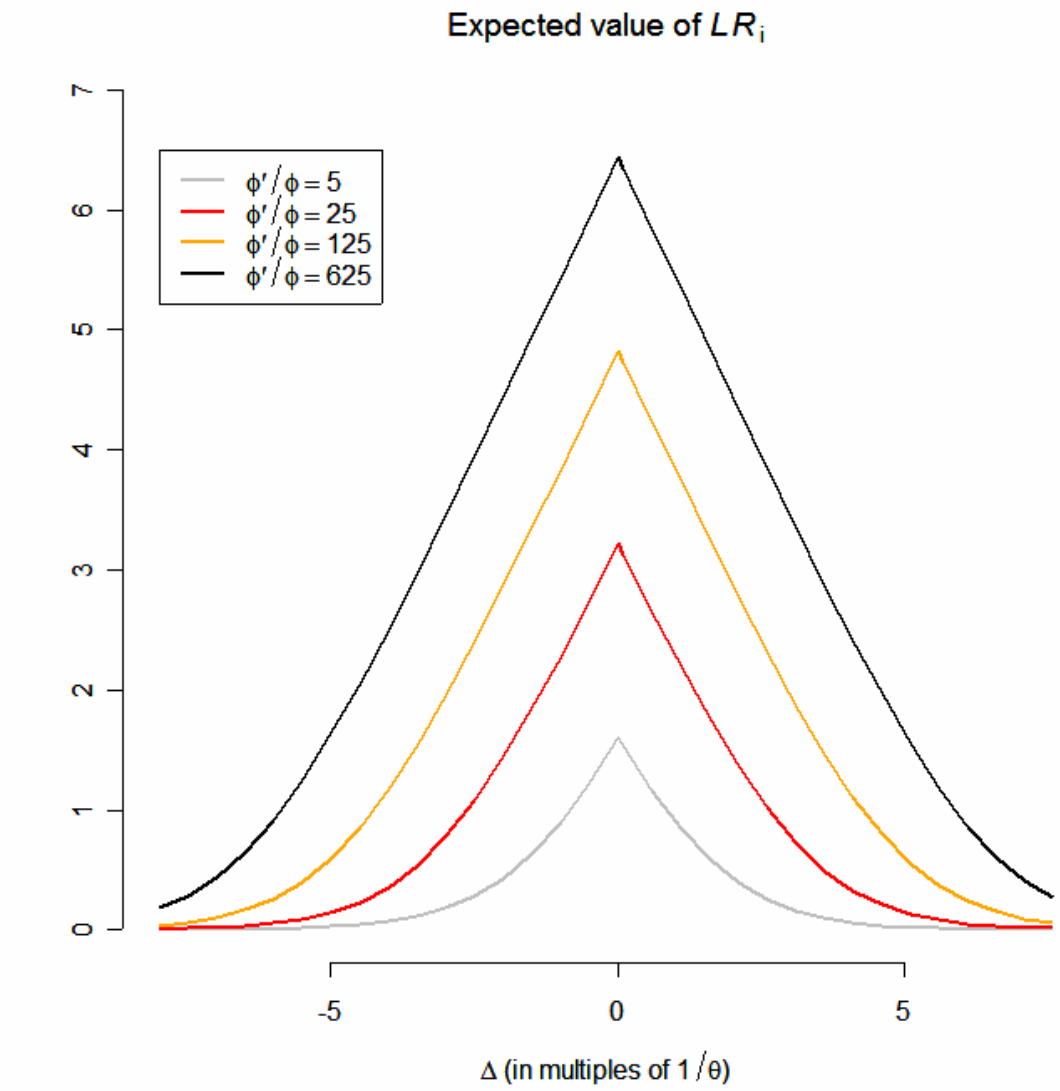
Expected log-ratio

Normalization generally cleans up the constants.

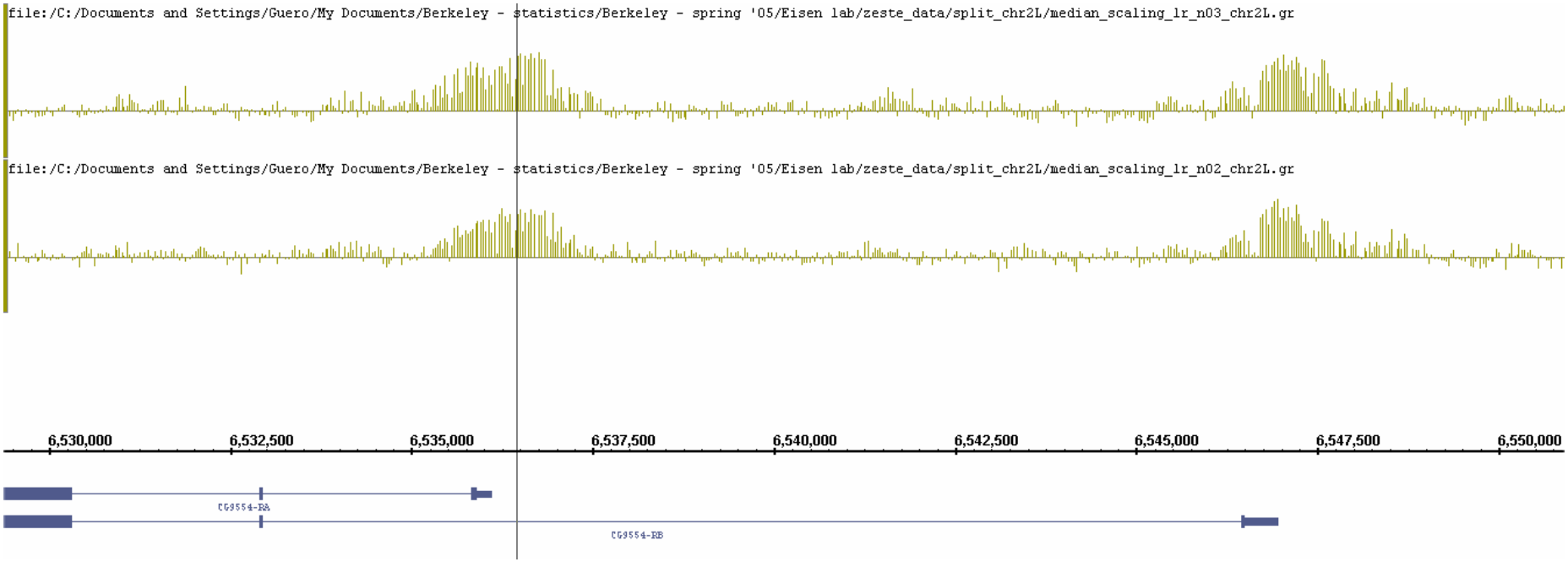
Peak amplitude *and* width depend on the efficiency ratio:

$$\frac{\phi'}{\phi}.$$

Note: ϕ' is binding-site specific.

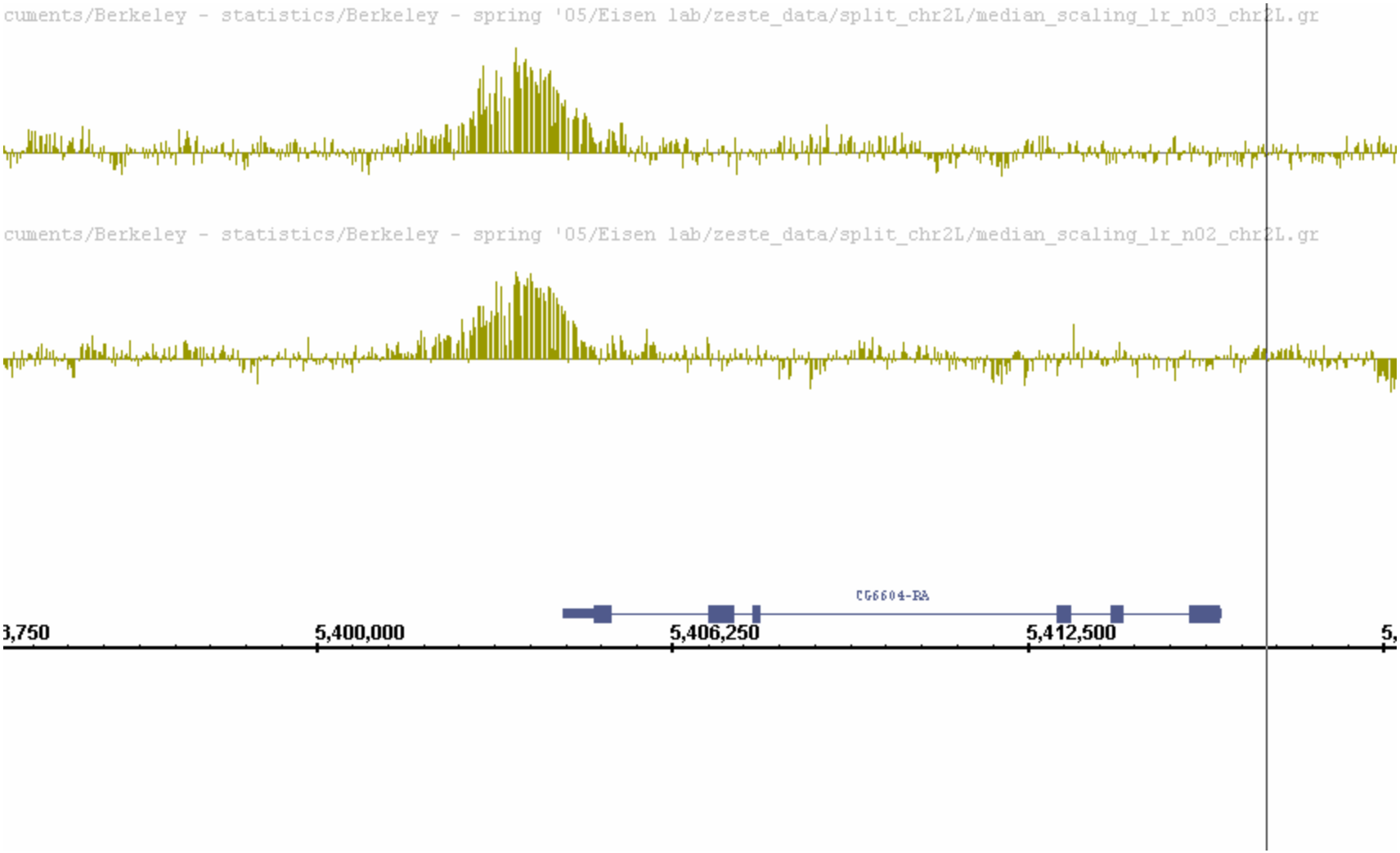


CG69554-RA and RB

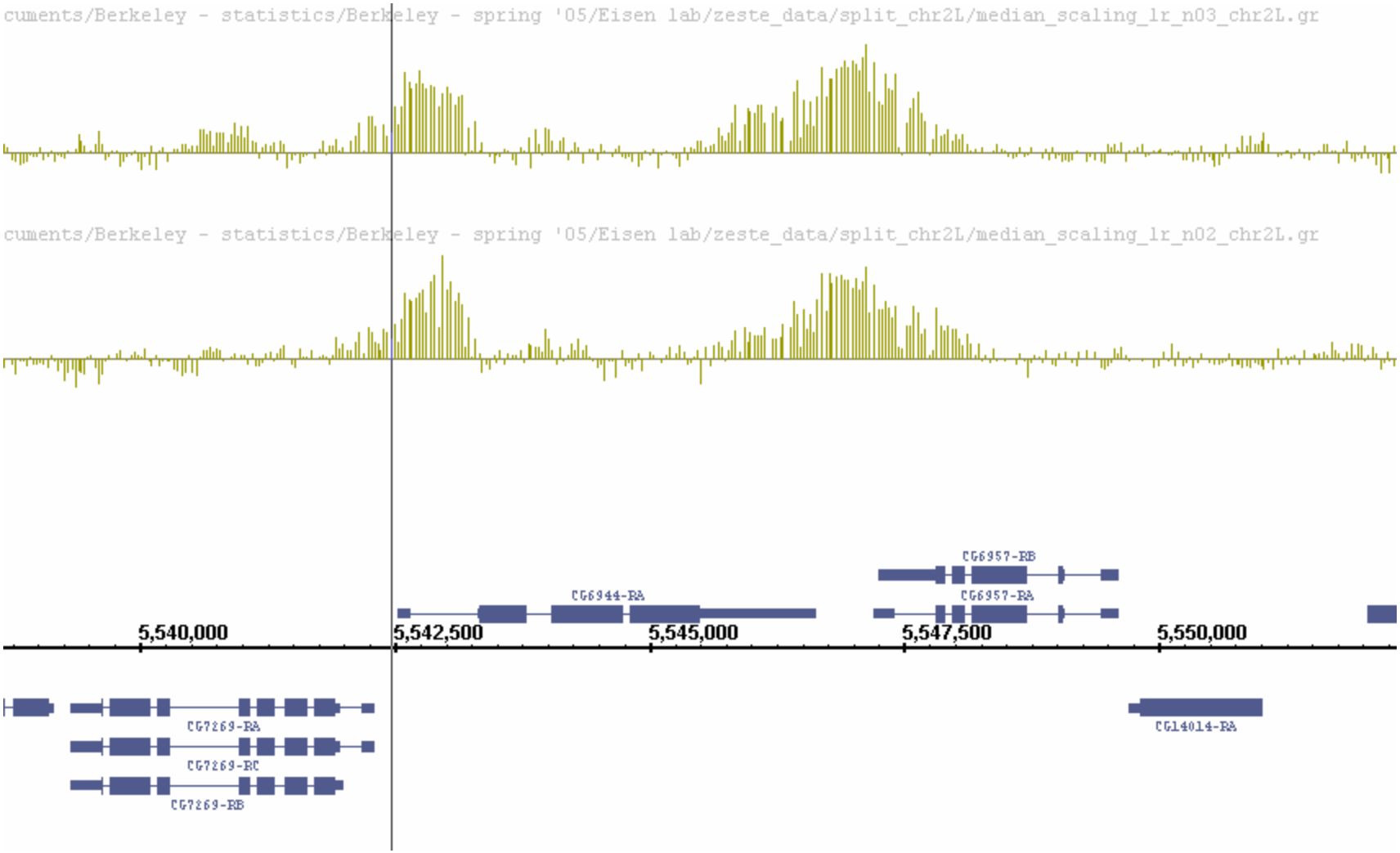


- *D. melanogaster* chromosome 2L
- Log ratios (unsmoothed) from 3 vs. 3 comparisons, two different IP/PCR/hybridization groups.

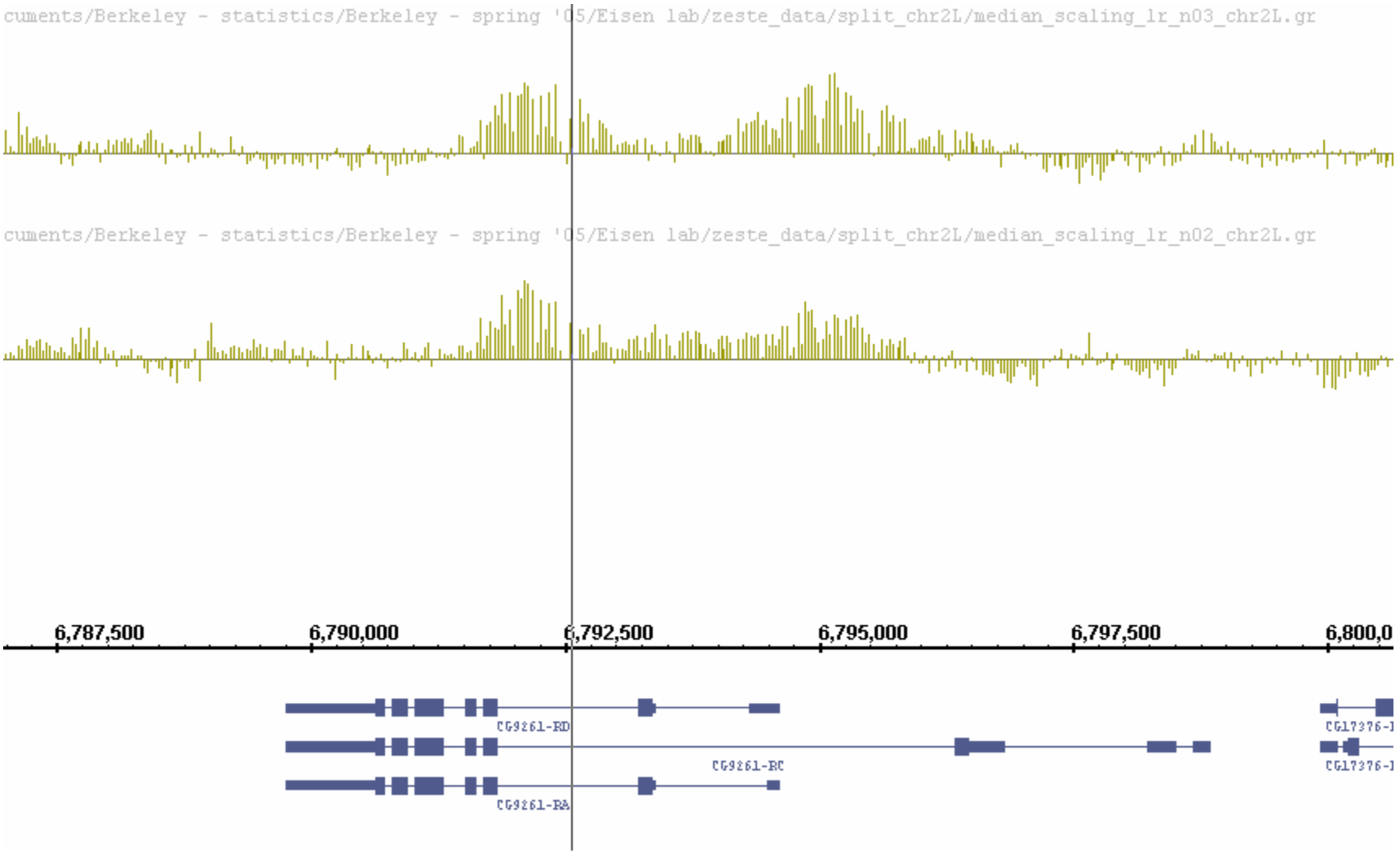
CG6604


cuments/Berkeley - statistics/Berkeley - spring '05/Eisen lab/zeste_data/split_chr2L/median_scaling_lr_n03_chr2L.gr
cuments/Berkeley - statistics/Berkeley - spring '05/Eisen lab/zeste_data/split_chr2L/median_scaling_lr_n02_chr2L.gr
CG6604-RA
3,750
5,400,000
5,406,250
5,412,500
5,

CG6944 or CG7269, CG6957



CG9261


cuments/Berkeley - statistics/Berkeley - spring '05/Eisen lab/zeste_data/split_chr2L/median_scaling_lr_n03_chr2L.gr
cuments/Berkeley - statistics/Berkeley - spring '05/Eisen lab/zeste_data/split_chr2L/median_scaling_lr_n02_chr2L.gr
6,787,500
6,790,000
6,792,500
6,795,000
6,797,500
6,800,0
CG9261-RD
CG9261-RC
CG9261-RA

Spatial correlation in log-ratio

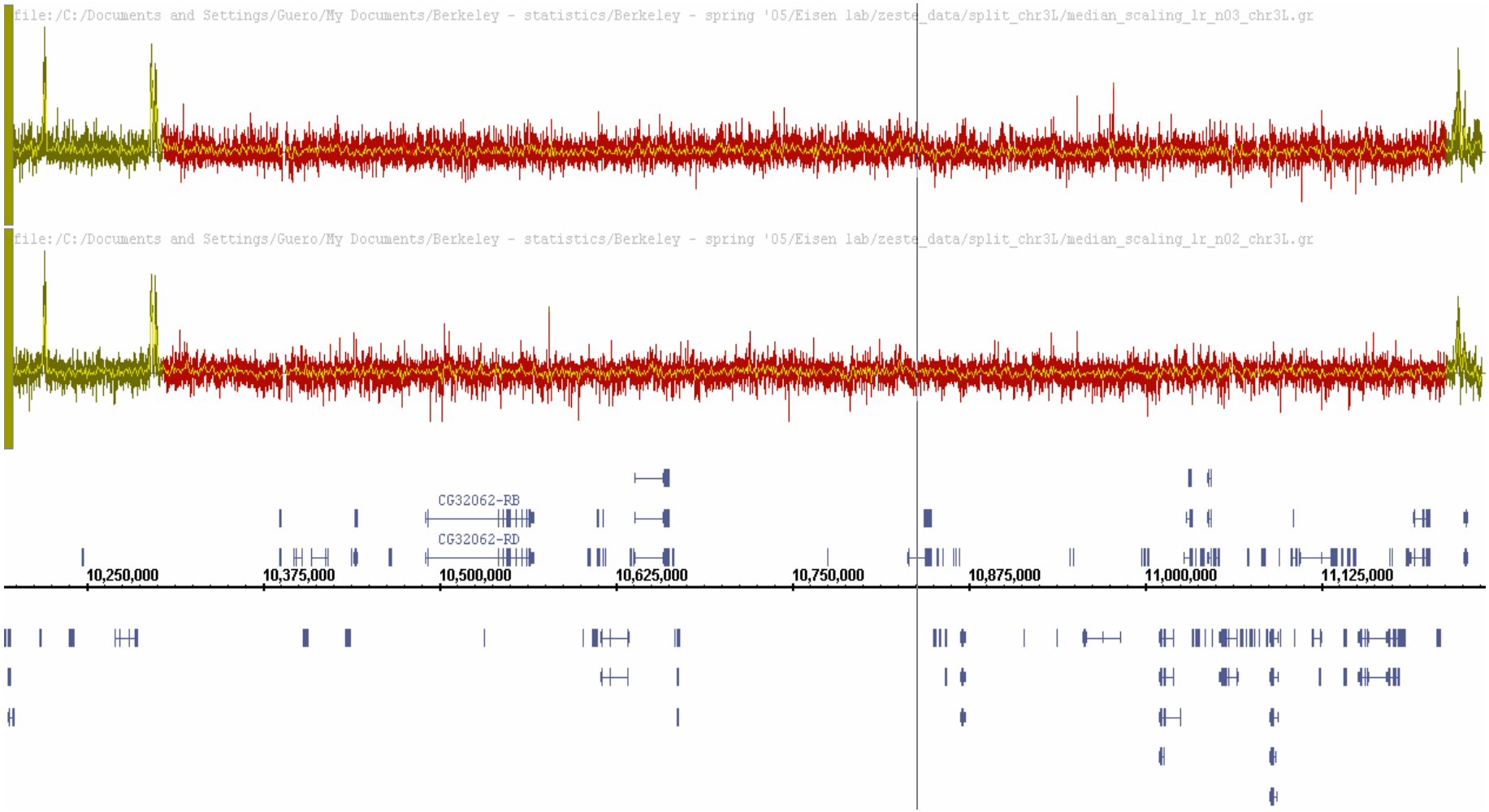
Again, abundance is not observable but the log-scale intensity ratios are:

$$LR_i = \log I_i^T - \log I_i^C = \log A_i^T - \log A_i^C + \delta_i$$

Approximating $\log\left(A_i / \mathbb{E}A_i\right)$ as before, it follows that

$$\mathrm{Corr}(LR_i, LR_j) \approx (1-\theta)^{d(i,j)} \left(1 + \frac{\mathrm{Var}\ \delta}{\mathrm{Var}(\log A^T) + \mathrm{Var}(\log A^C)}\right)^{-1}.$$

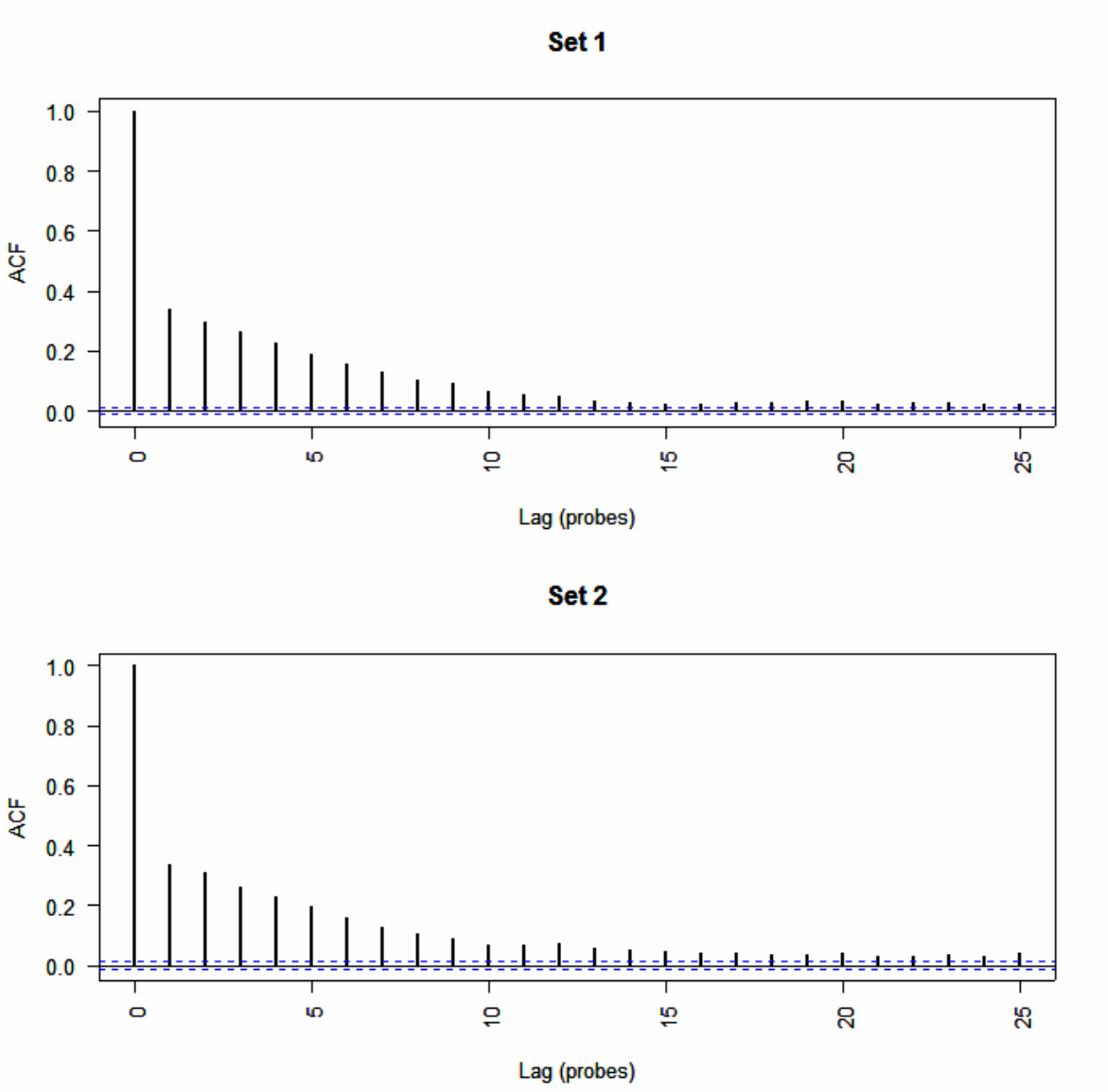
Spatial correlation in log-ratio



Spatial correlation in log-ratio

- For simplicity, ignore irregularity of probe spacing.
- Compute auto-correlation at various lags.

For both data sets, there is statistically significant auto-correlation up to a lag of ≈15 positions.



Summarizing…

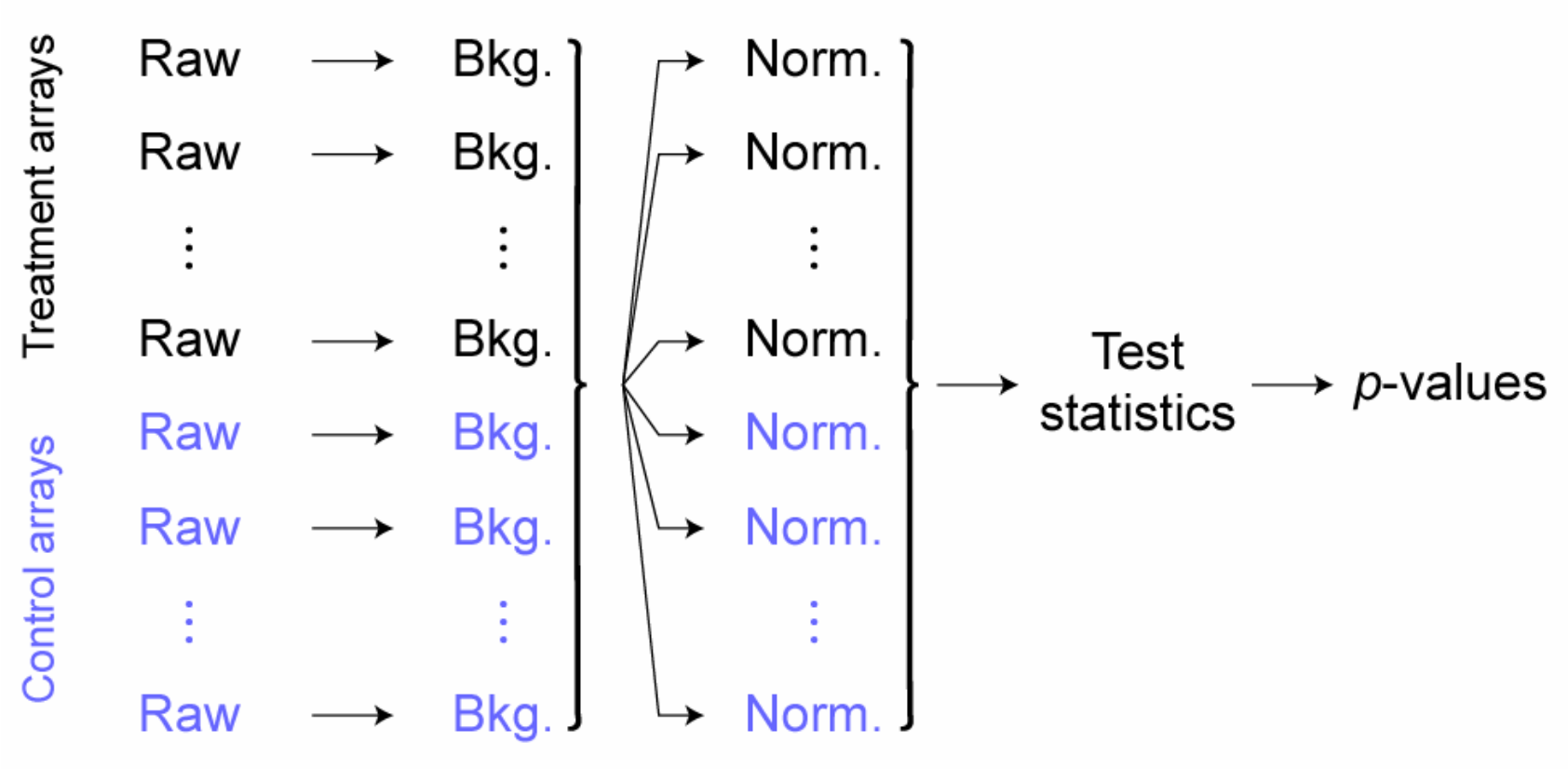
- The model predicts peak-like signal response near binding sites.
- The model predicts spatial correlation in both target abundance and log-ratio. Target sequence for neighboring probes *tends to end up on the same fragment.* IP and amplification take place at the fragment level.
- Data are consistent with these predictions.

Data pre-processing and analysis

Background correction, normalization

summary statistics, and significance

A common framework for current methods



Additive background

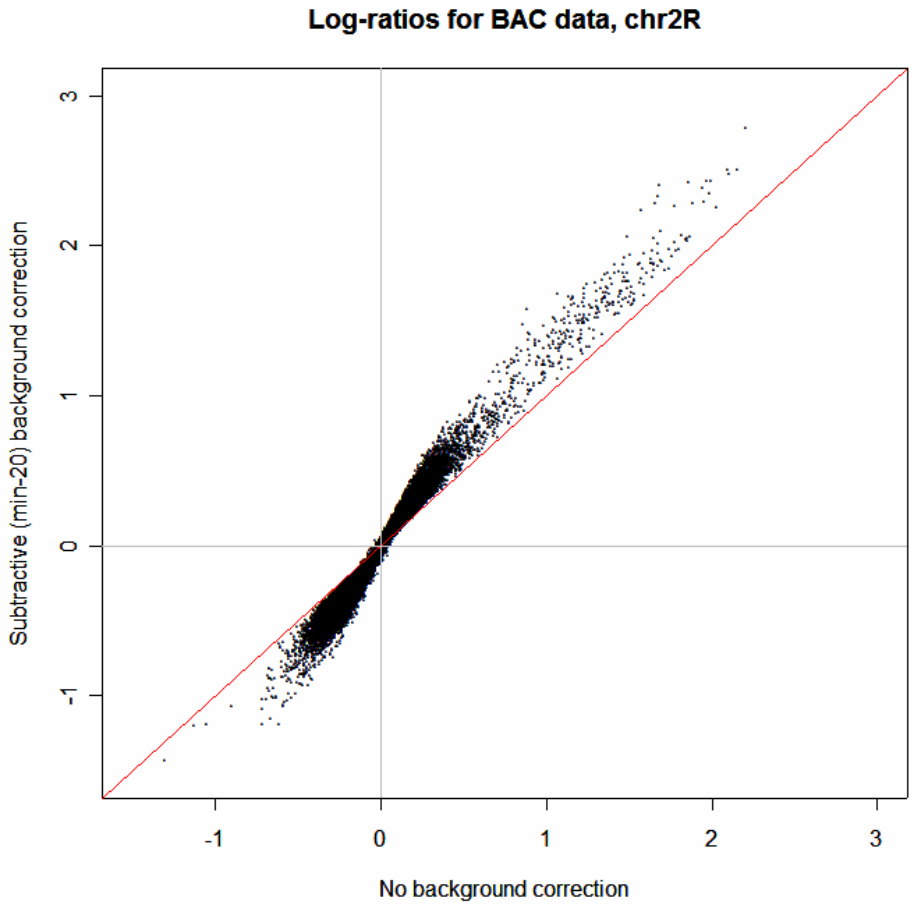
A better (though still imperfect!) model for intensity:

$$I_{ij} = \alpha_i \, A_{ij} \, \varepsilon_{ij} + B_{ij}.$$

Additive background

$$LR_i = \log\left(\frac{\alpha_i A_i^T \varepsilon_i^T + B_i^T}{\alpha_i A_i^C \varepsilon_i^C + B_i^C} \right)$$

In the presence of positive background, the log-ratio will be biased towards 0.



Background correction methods

Approach	*Examples*
Do nothing.	TiMAT (from BDTNP)
Mismatch subtraction, then set non-positive values to 1 before taking logs.	Cawley et al., *Cell*, '04; G-TRANS
Mismatch subtraction, transformed by a family of "generalized log" functions.	PLIER
Global intensity-based correction.	RMA convolution method
Sequence-based background estimation.	GC-RMA "affinities" method; MAT
Hybrid: smoothly combine MM subtraction and sequence-based estimates.	GC-RMA "full model"

Unnormalized log-intensities

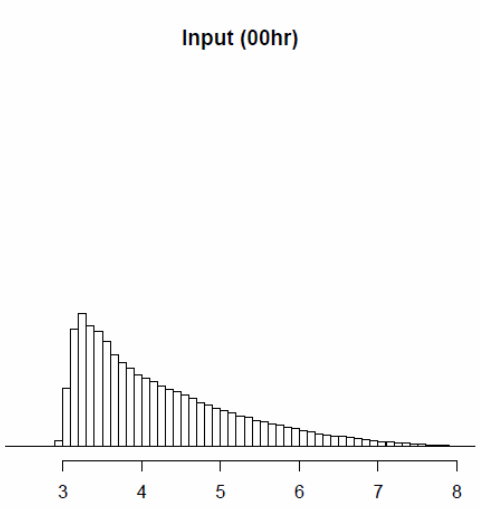
Common ChIP-chip normalization scheme:

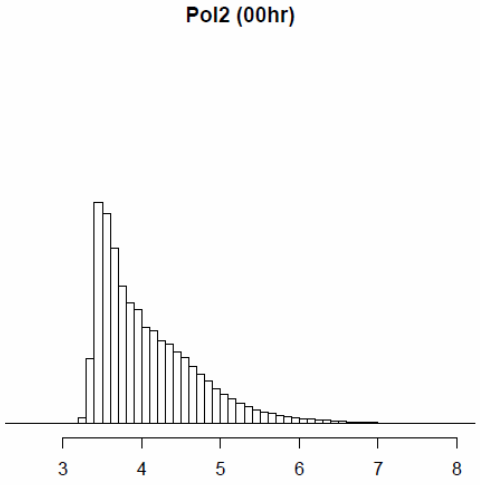
- Quantile within treatment condition.
- Median scaling between treatment and control.

Here, distributional differences at 00hr are much stronger than at 02hr.

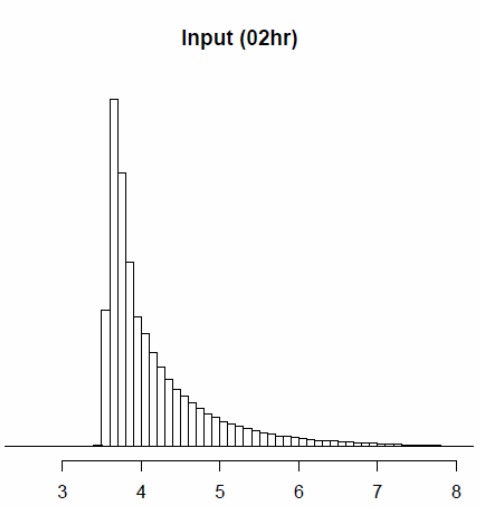
ENCODE Pol2:
B1 (2x), B2 (2x), B3 (2x).

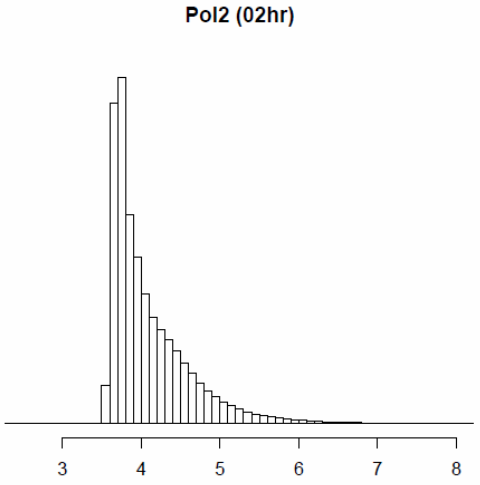
Replicate 1





Replicate 2



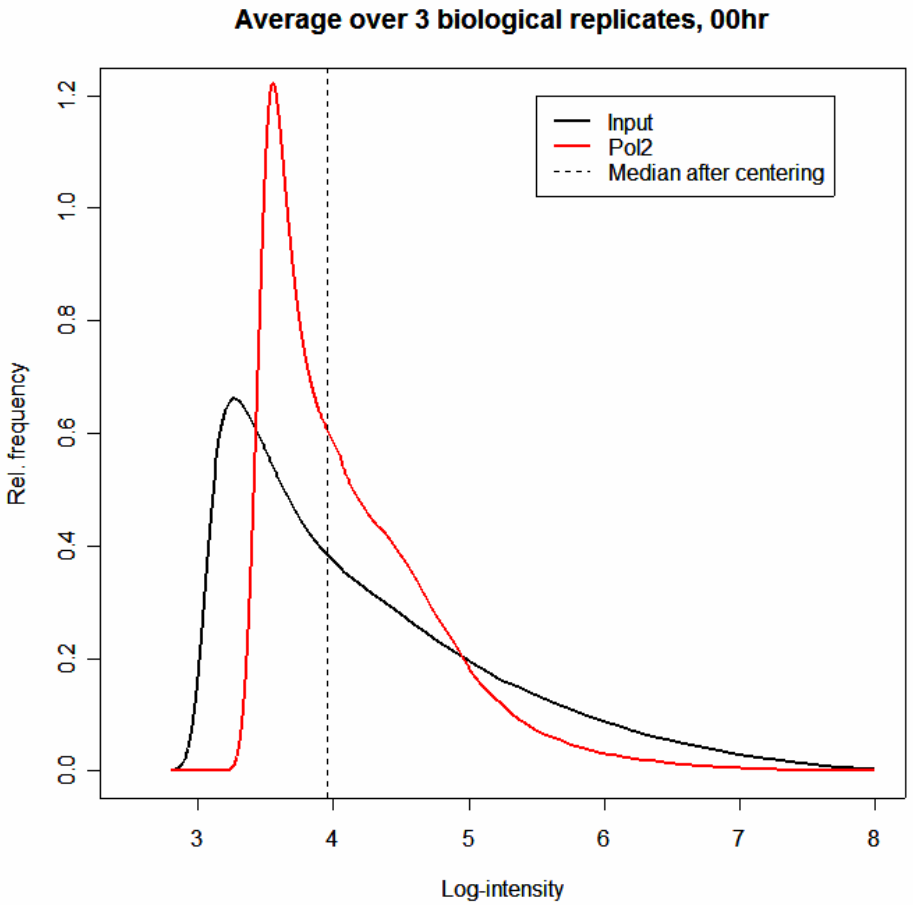


Unnormalized log-intensities

Common ChIP-chip normalization scheme:

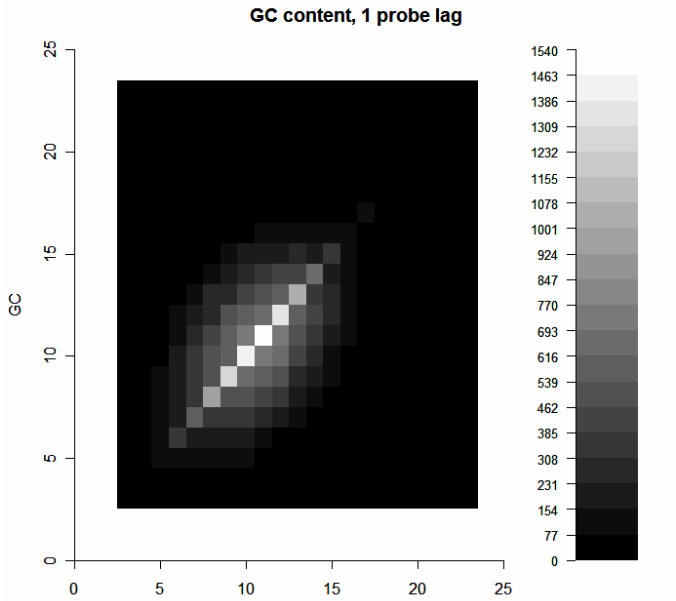
- Quantile within treatment condition.
- Median scaling between treatment and control.

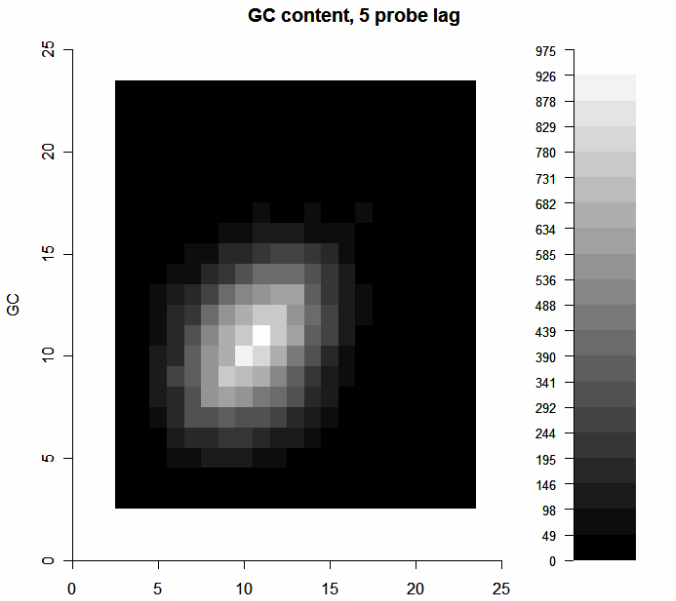
Here, distributional differences at 00hr are much stronger than at 02hr.

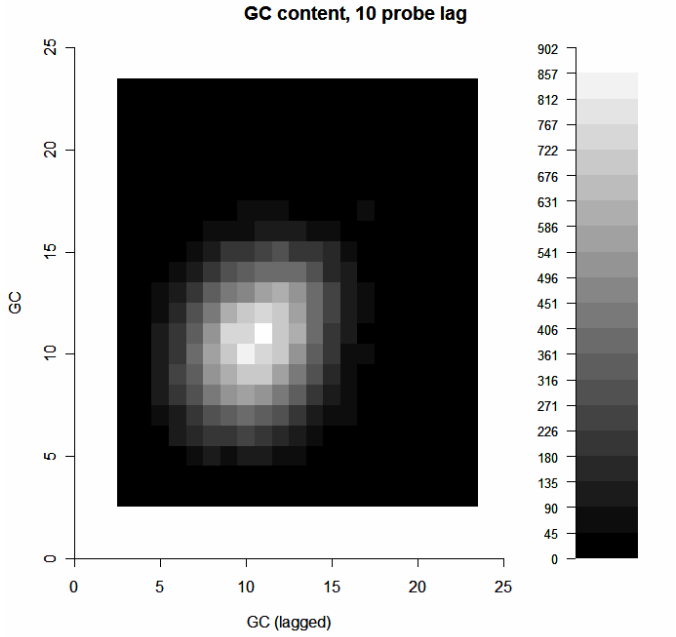


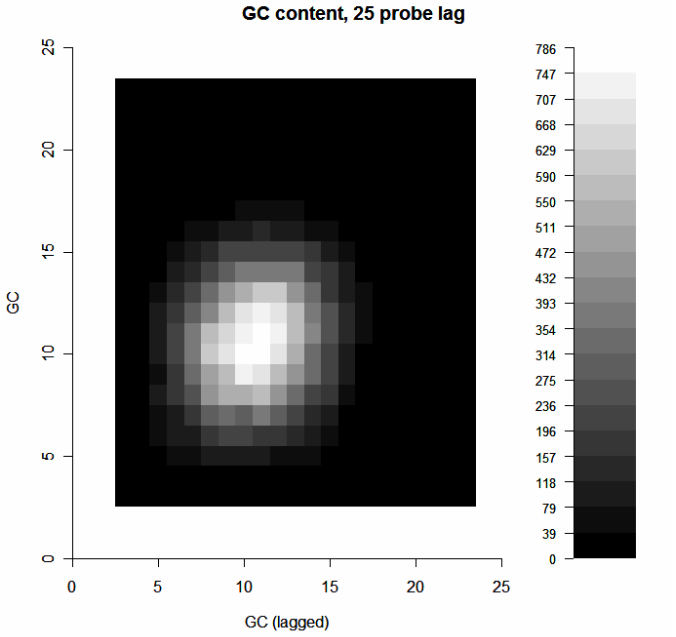
ENCODE Pol2:
B1 (2x), B2 (2x), B3 (2x).

Spatial correlation in probe GC content









Normalization methods

Approach	*Examples*
Do nothing.	MAT
Scaling.	—
Quantile normalization within condition, scaling across conditions.	Cawley et al., *Cell*, ’04; G-TRANS
Full quantile normalization.	TileMap; Li, Meyer and Liu, *Bioinformatics* ’05; G-TRANS

Combining probe-level statistics

- Actual binding site signal spans multiple positions.
- Single probes are…
  - …prone to gross error.
  - …frequently either lazy or promiscuous hybridizers.

Statistical approaches:

- Two-state hidden Markov models.
  Li, Meyer and Liu, *Bioinformatics*, '05; TileMap
- Smoothed or windowed probe-level statistics
  Cawley et al., *Cell*, '04; Keles et al., '04; MAT;
  Buck, Nobel and Lieb, *Genome Biology* '05
- *Ad hoc* post-processing of probe-level calls

- Peak fitting

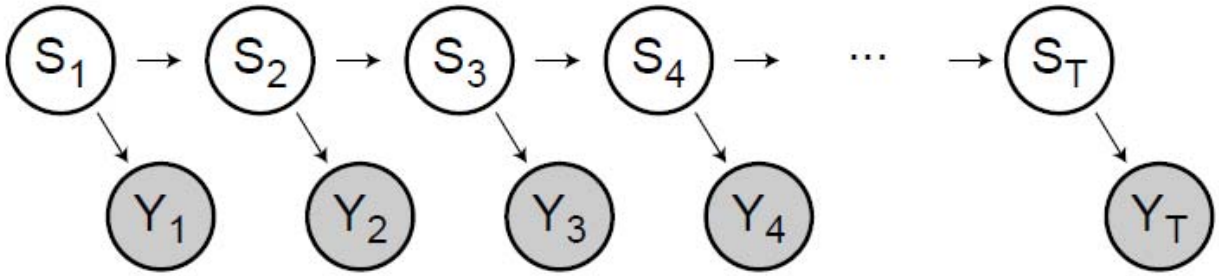
Assessing significance

- HMMs: posterior distributions on states
- Windowed methods:

Approach	*Examples*
Non-parametric: Wilcoxon rank sum	Cawley et al., *Cell*, ’04; G-TRANS (now “TAS”?)
A global null distribution: assume common variance for window-level statistics.	TiMAT; Buck, Nobel and Lieb, *Genome Biology* ’05
Standard *t* statistics.	Keles et al., ’04
Moderated *t* statistics via empirical Bayes.	TileMap, `limma` in R
Moderated *t* statistics via direct estimation.	MAT

Standard HMMs

In a traditional hidden Markov model, hidden multinomial state variables $S_1, \ldots, S_T$ lead to observable "emitted values."

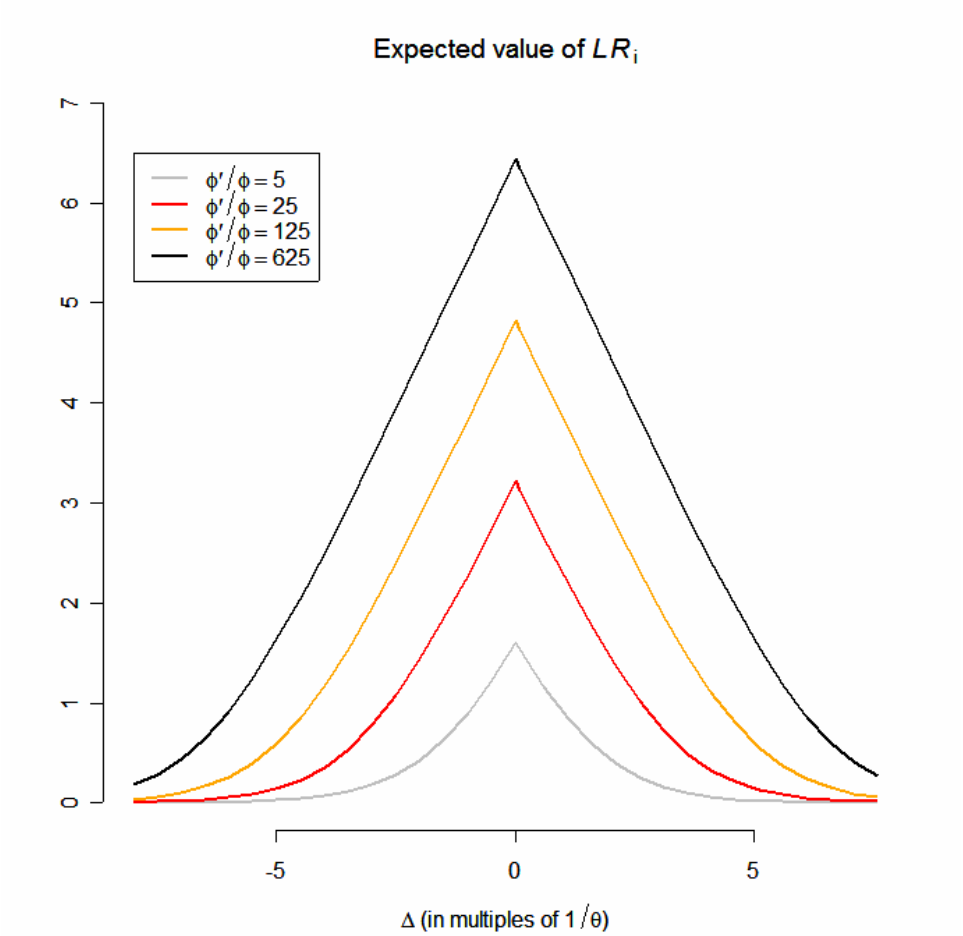


- The hidden states $\{S_t\}_{t=1}^T$ form a Markov chain on n states, with transition matrix $A \equiv (a_{ij})_{i,j=1}^n$.

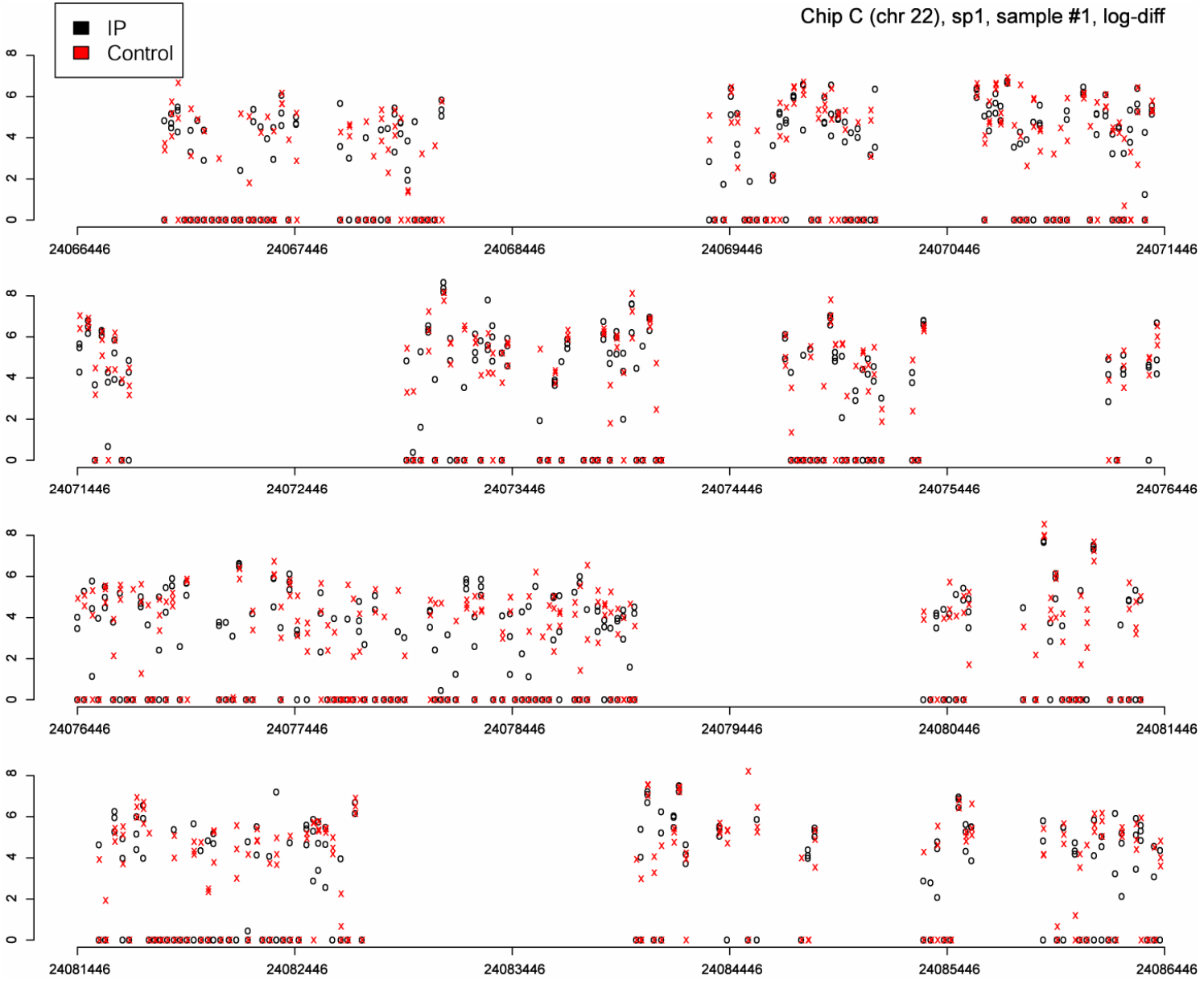
- The distribution of the observable variable depends on the hidden state: $Y_t|_{S_t} \sim p_{S_t}$.

Standard HMMs miss the mark…

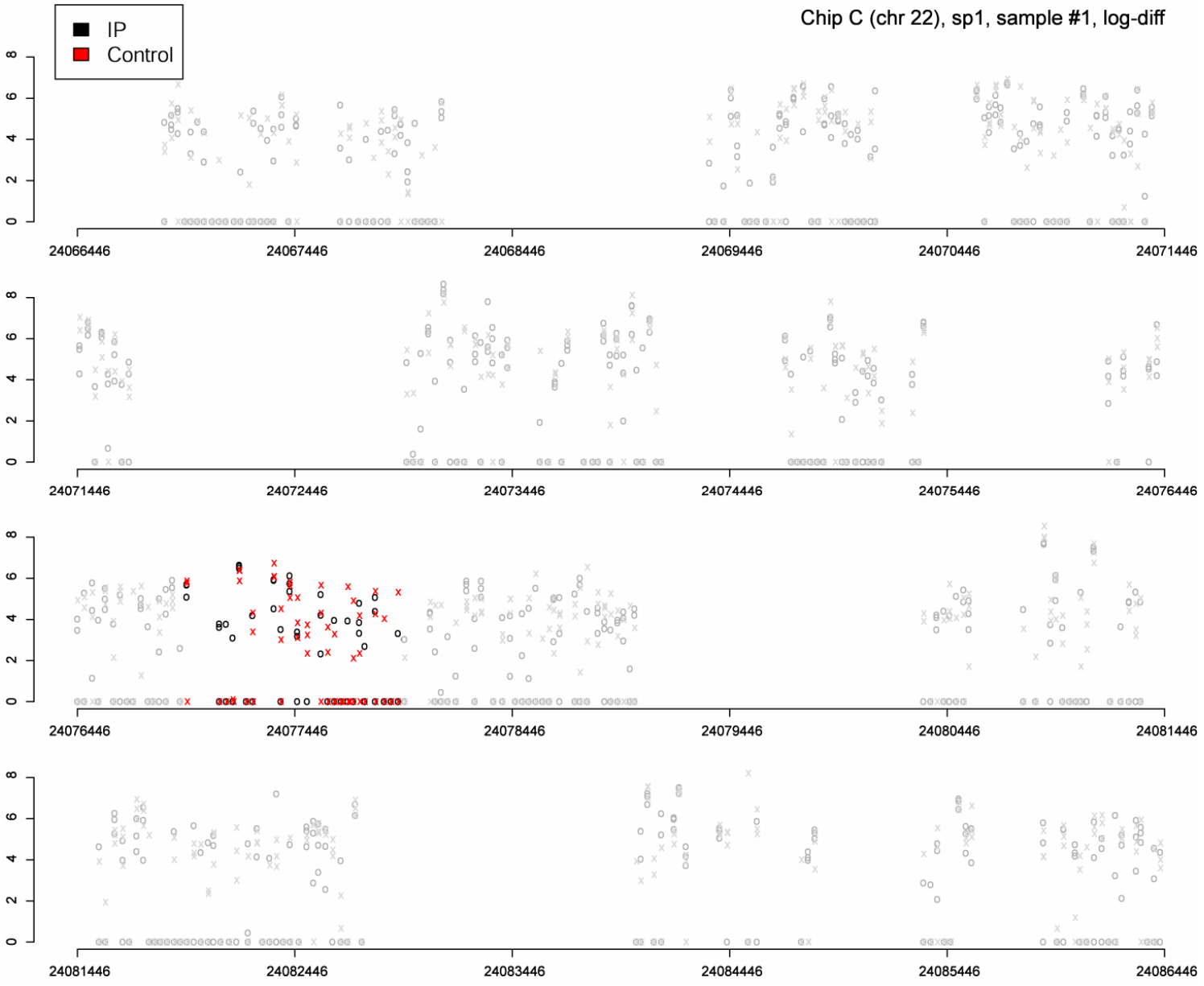
- We don't expect sharp transitions.
- A single "enriched" state ignores real variety.
- Forced geometric state duration distributions.
- Ignores expected spatial structure.
- Conditional independence of observations given the sequence of hidden states?



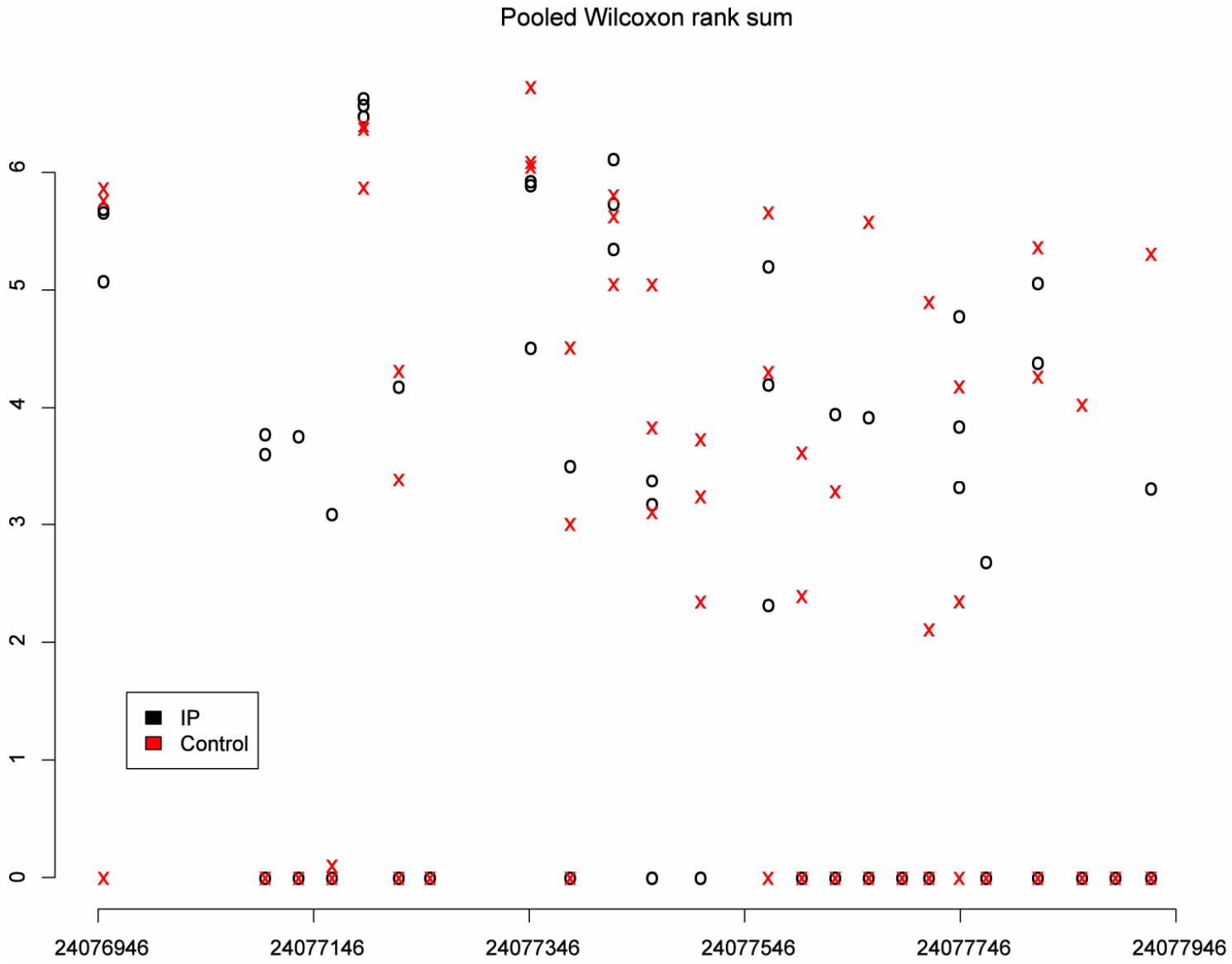
Affy/G-TRANS Wilcoxon rank sum



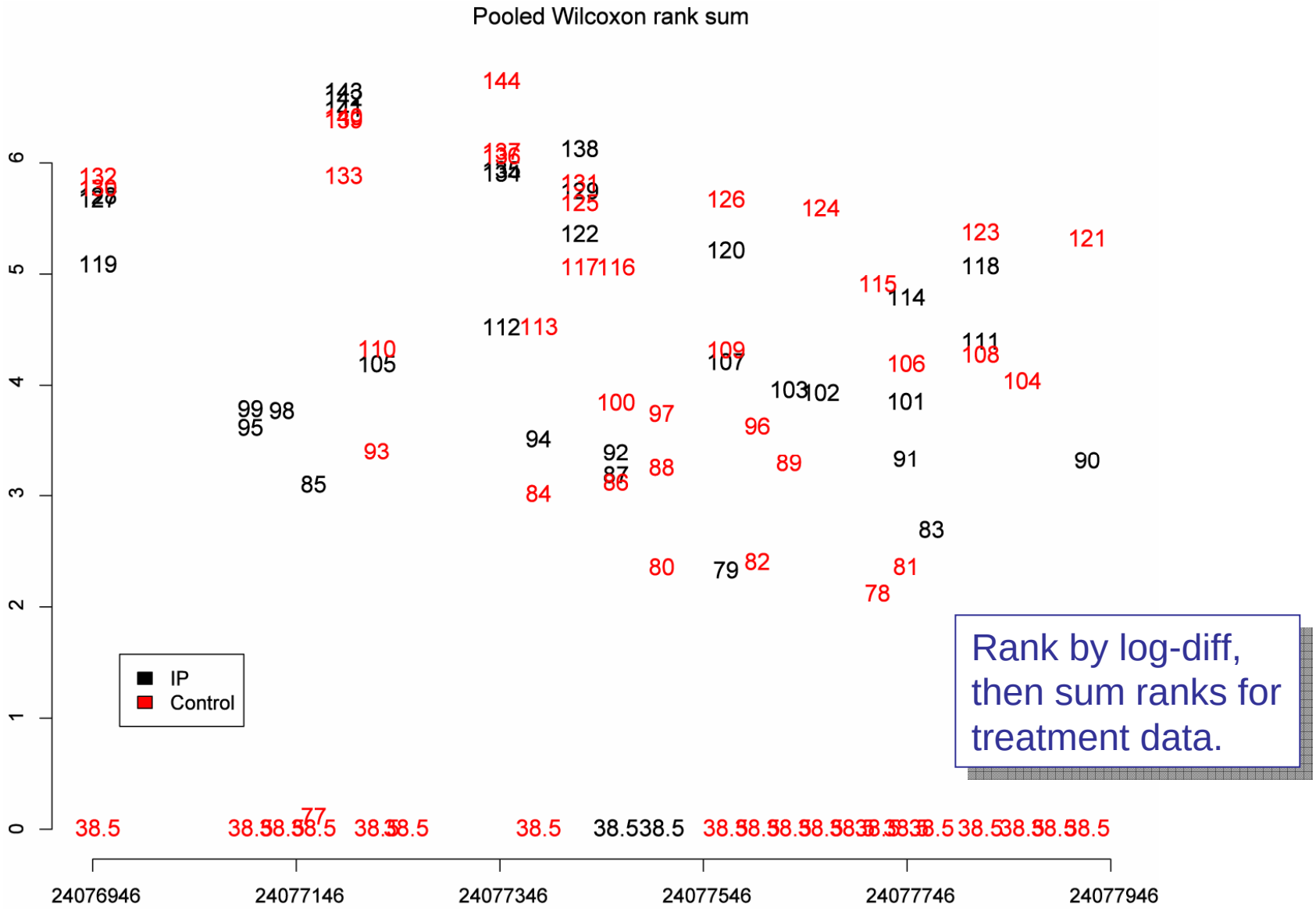
Affy/G-TRANS Wilcoxon rank sum



Pooled Wilcoxon rank sum



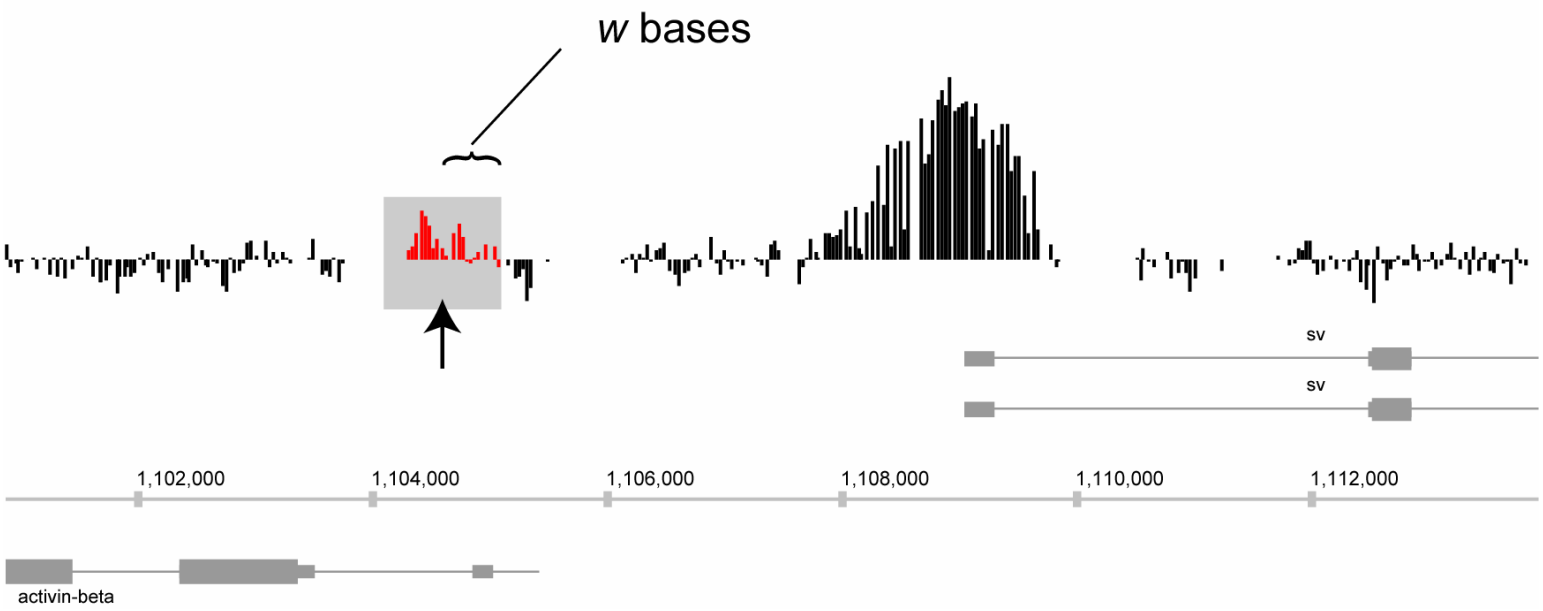
Pooled Wilcoxon rank sum



Rank by log-diff, then sum ranks for treatment data.

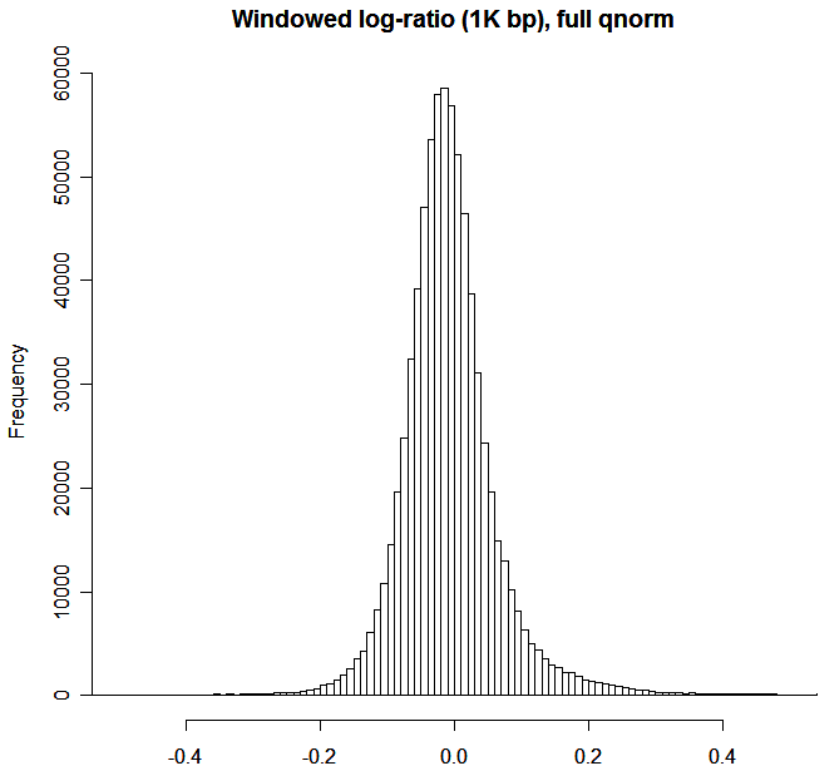
Windowing probe-level statistics

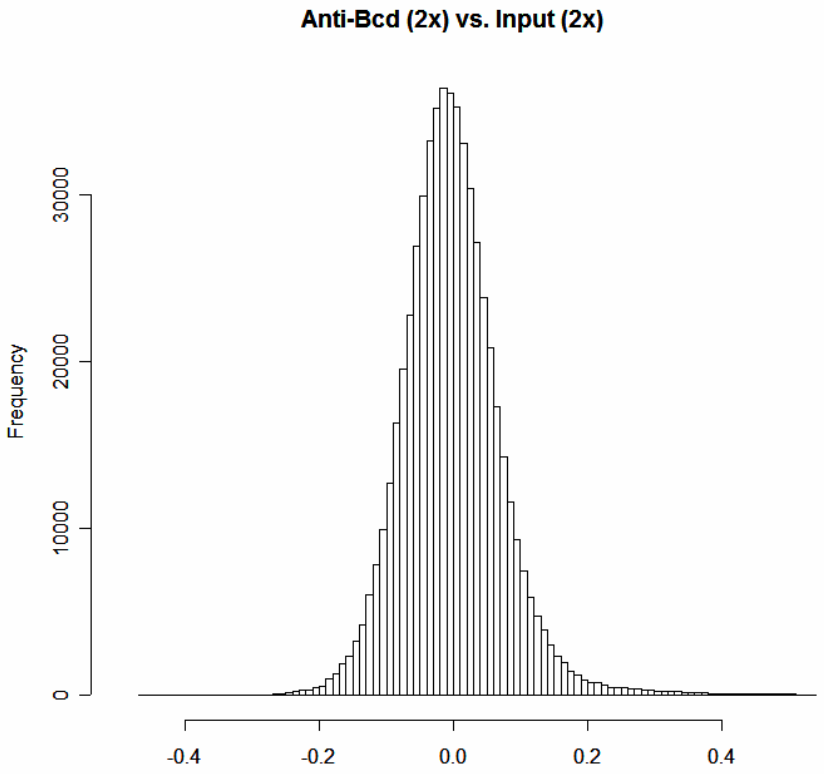
$$T_i \equiv T(i;w,f) = f(\{X_j : |g_i - g_j| \le w\}).$$



Apply some function *f* (e.g., mean, median), to all scores within *w* bases of probe *i*, and call the result T_i. Typically rescale by $\sqrt{n_i}$.

Windowed log-ratios

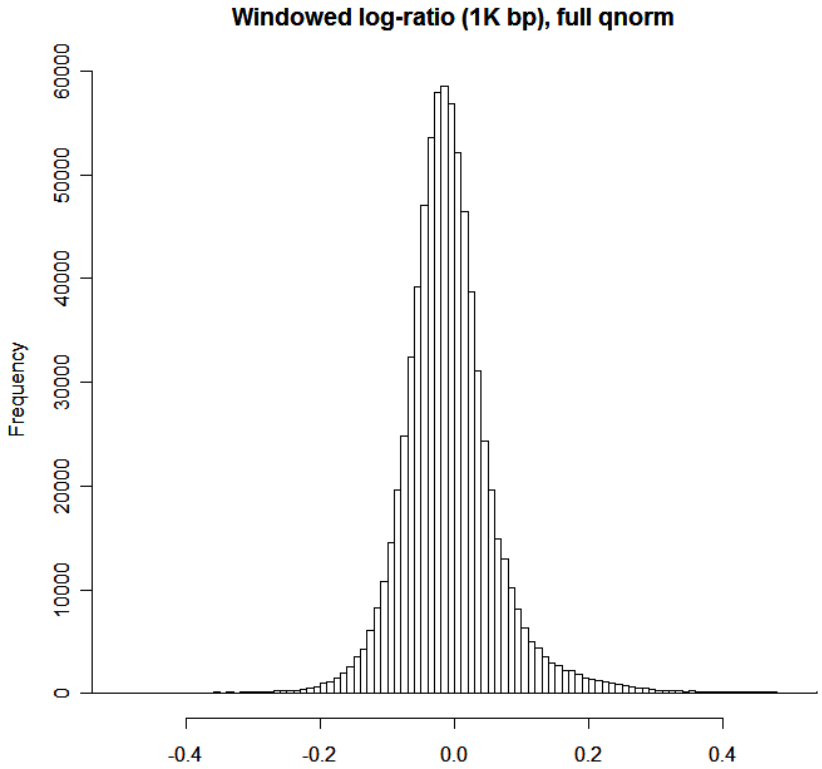


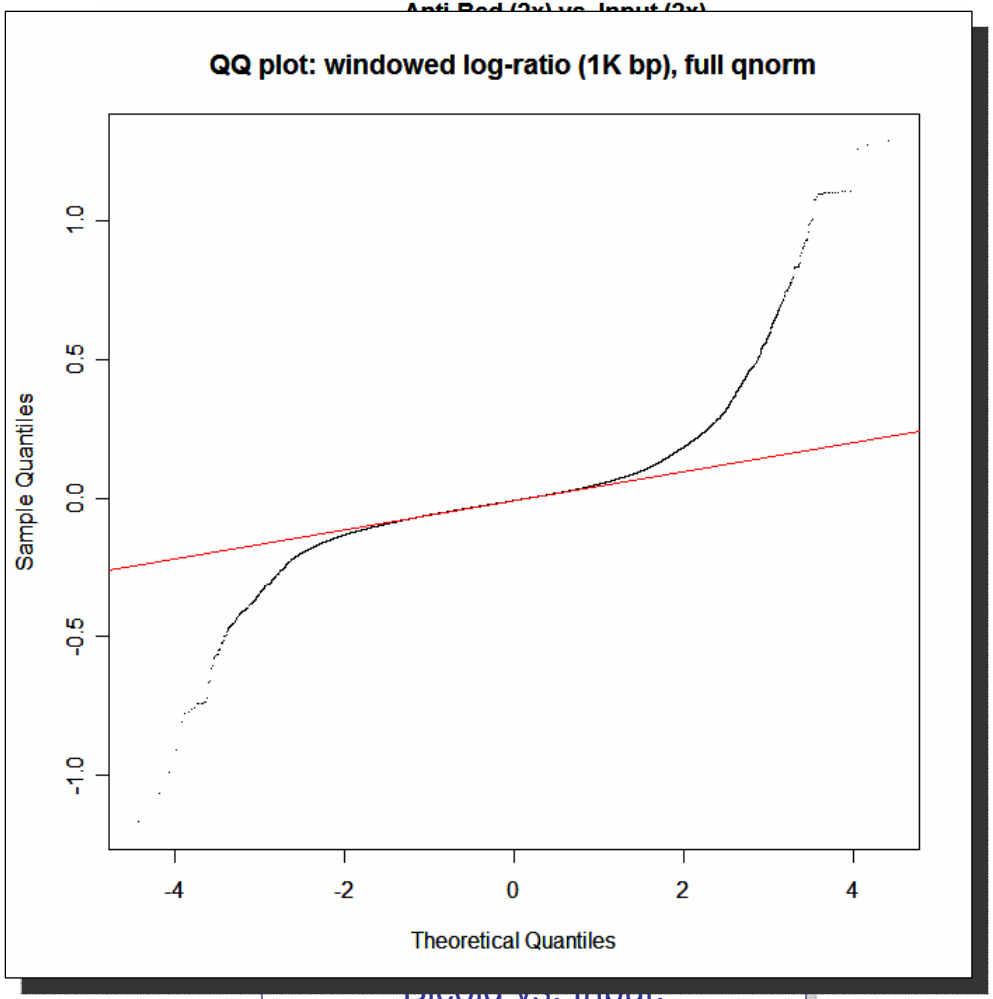


ENCODE data: Pol2, 00hr,
B1 vs. B1,4,5 pooled.

Bicoid vs. Input,
2x each

Windowed log-ratios

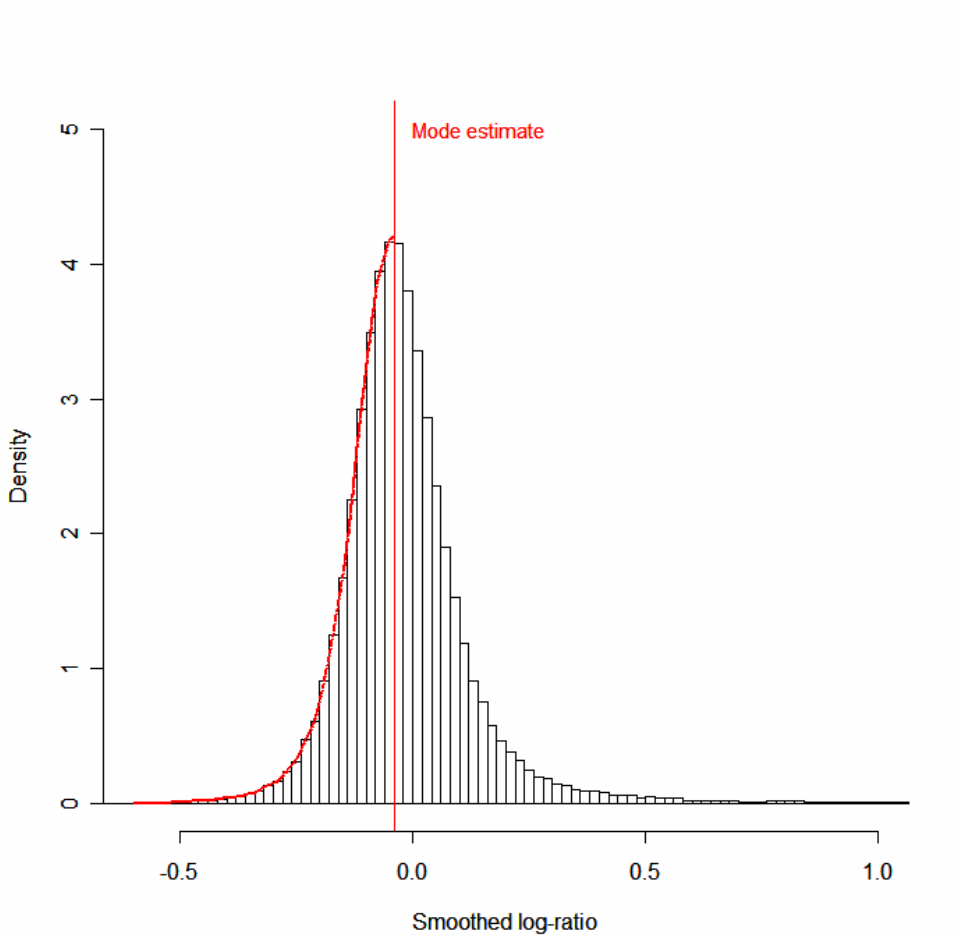




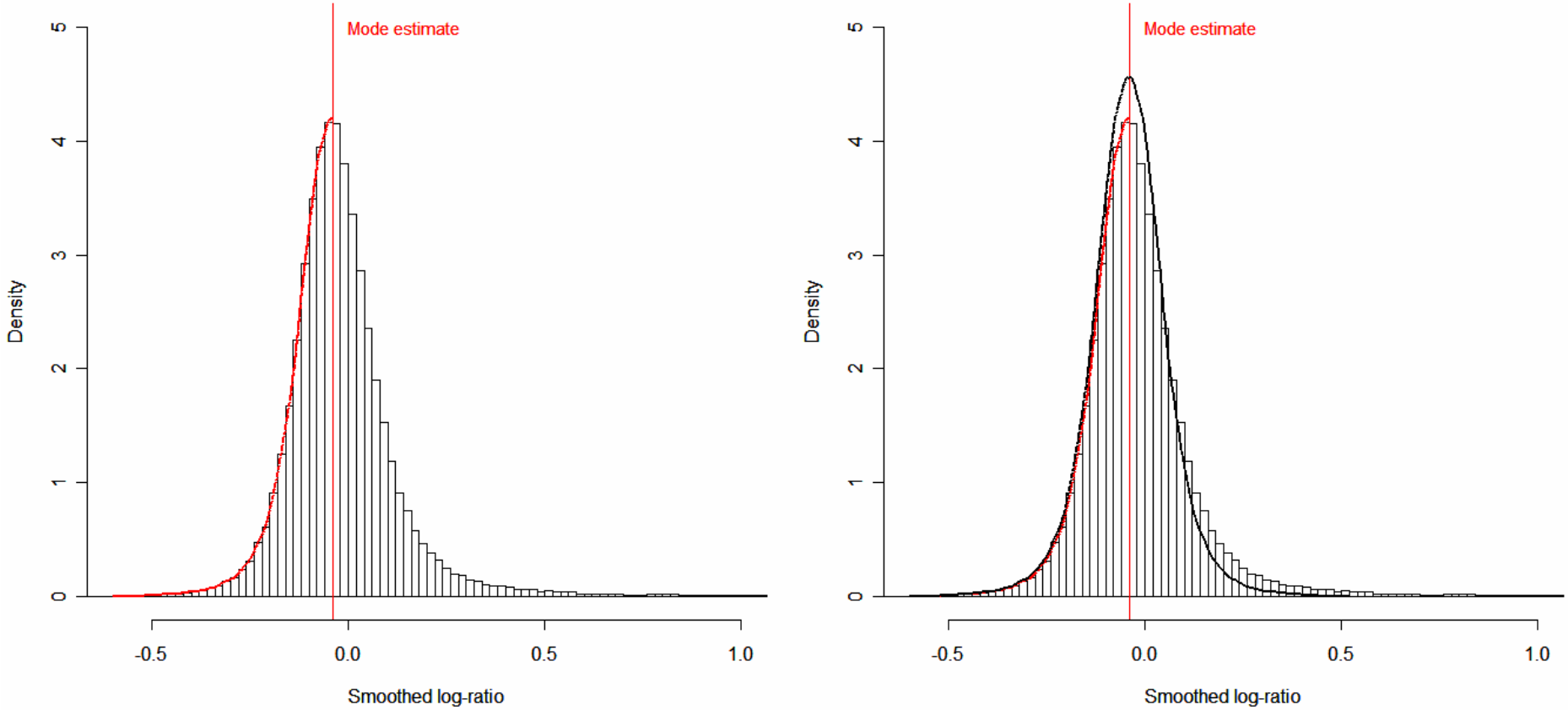
ENCODE data: Pol2, 00hr,
B1 vs. B1,4,5 pooled.

2x each

Non-parametric *p*-values



Non-parametric *p*-values



Gibbons et al., *Genome Biology*, 2005 propose a parametric version which is similar. Also see Efron, *JASA*, 2004.

What works best (so far!)

Background correction

- If skipped, there is some bias in probe-level statistics — towards high-affinity probes and away from low-affinity.
- Using MM adds a lot of noise, probably offsetting any gains in dynamic range and unbiasing. (Moderated MM methods, e.g. look better.)

Normalization

- Yes! Full quantile normalization is often advantageous, and doesn't seem to hurt too much even when unnecessary.

Statistics and significance

- Moderated *t*-statistics. Moving windows aren't perfect, but seem to perform well.
- Normality is probably not always a safe assumption.

Work to be done…

- A better understanding of hybridization dynamics, permitting more effective, sequence-specific background correction and calibration.
- A model which more naturally accommodates signal runs of varying size. (HSMM?)
- Integration with…
  - …expression arrays.
  - …algorithms for binding site motif identification.
- In multi-factor experiments, methods for assessing coordinated binding.

Acknowledgments

- U.C. Berkeley

  Terry Speed, support from VIGRE (NSF).

- LBNL

  Mike Eisen, Mark Biggin, Xiaoyong Li.

- Affymetrix

  Simon Cawley, Tom Gingeras, Antonio Piccolboni, Stefan Bekiranov, Srinka Ghosh, David Nix.