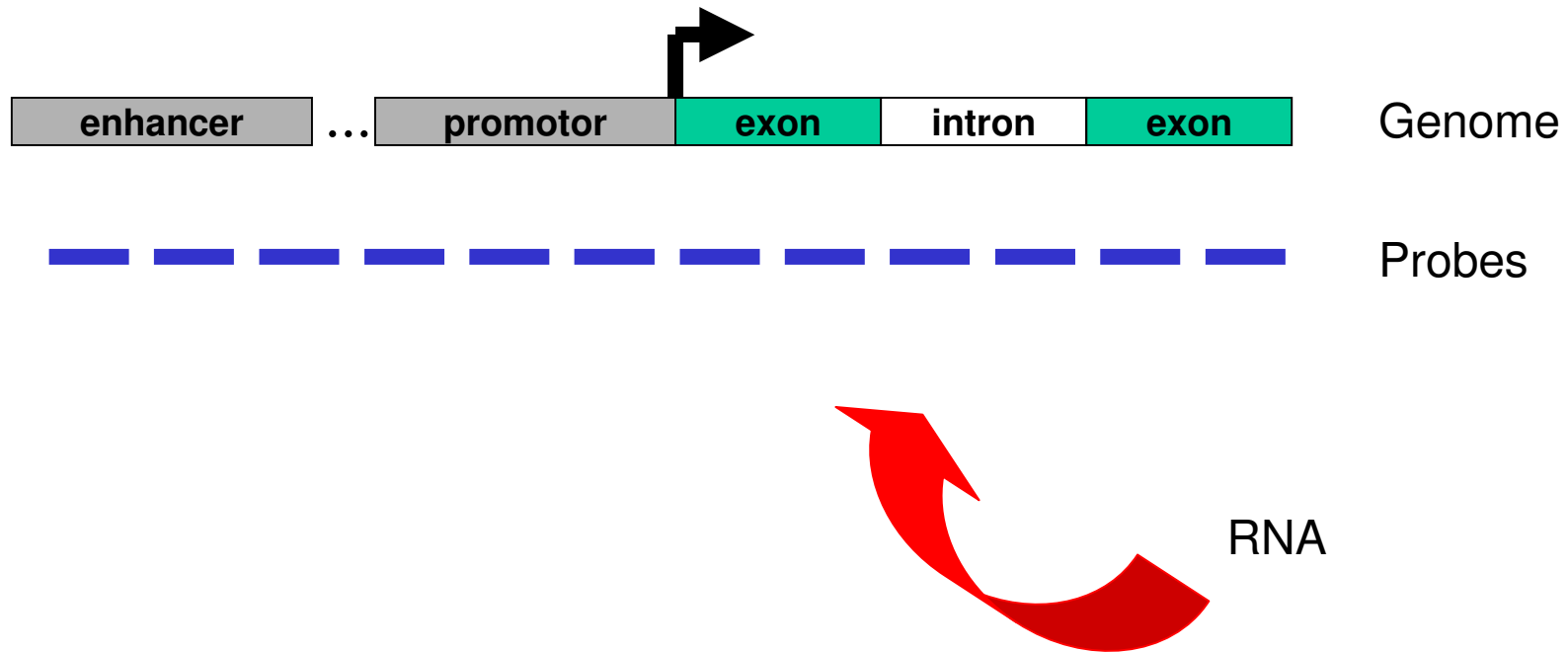


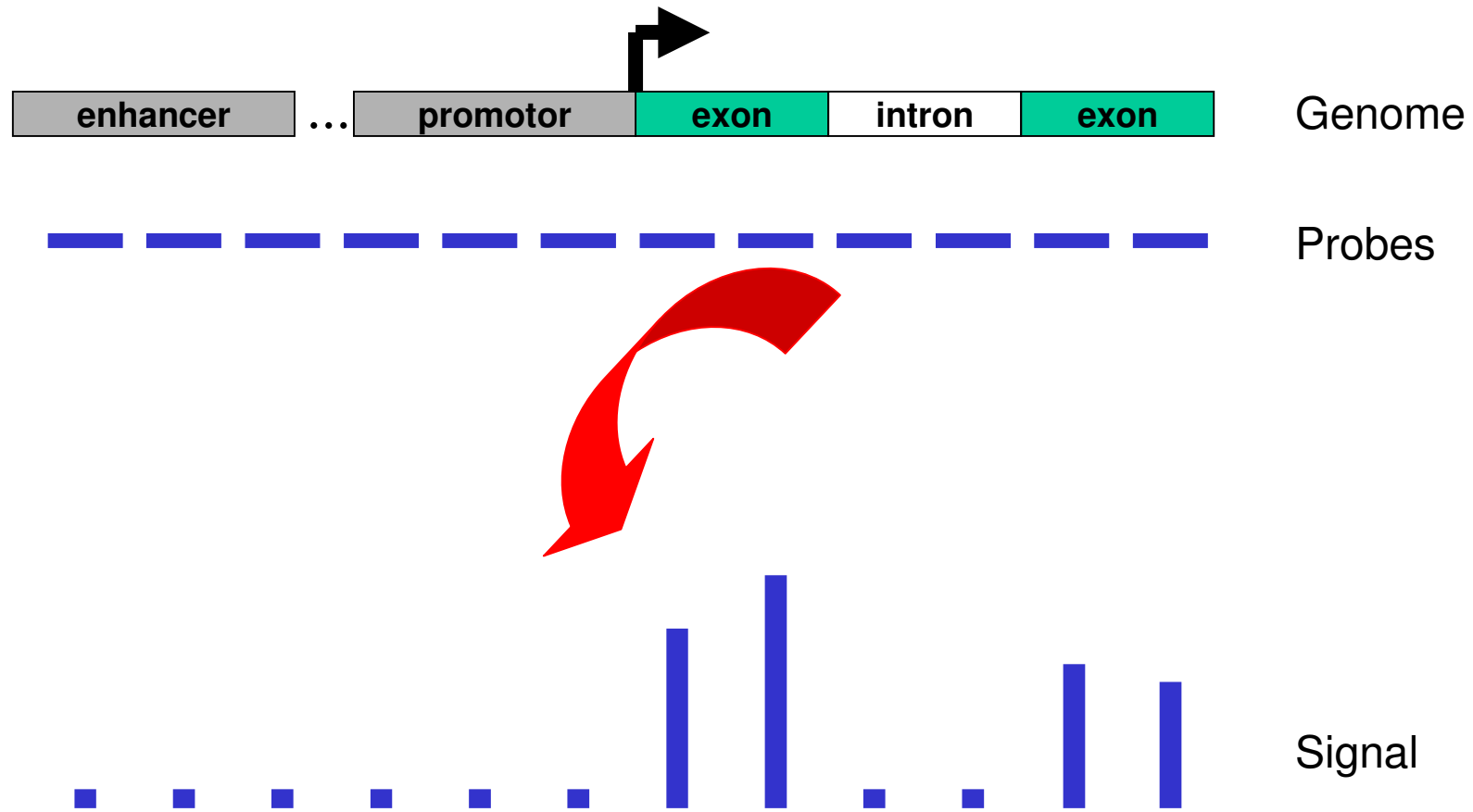
# Tiling Arrays transcript annotation

Ho-Ryun Chung

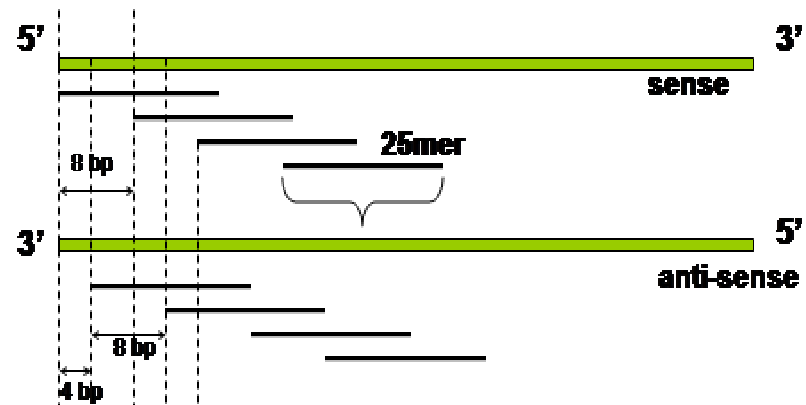
# ... tiling array



# ... tiling array

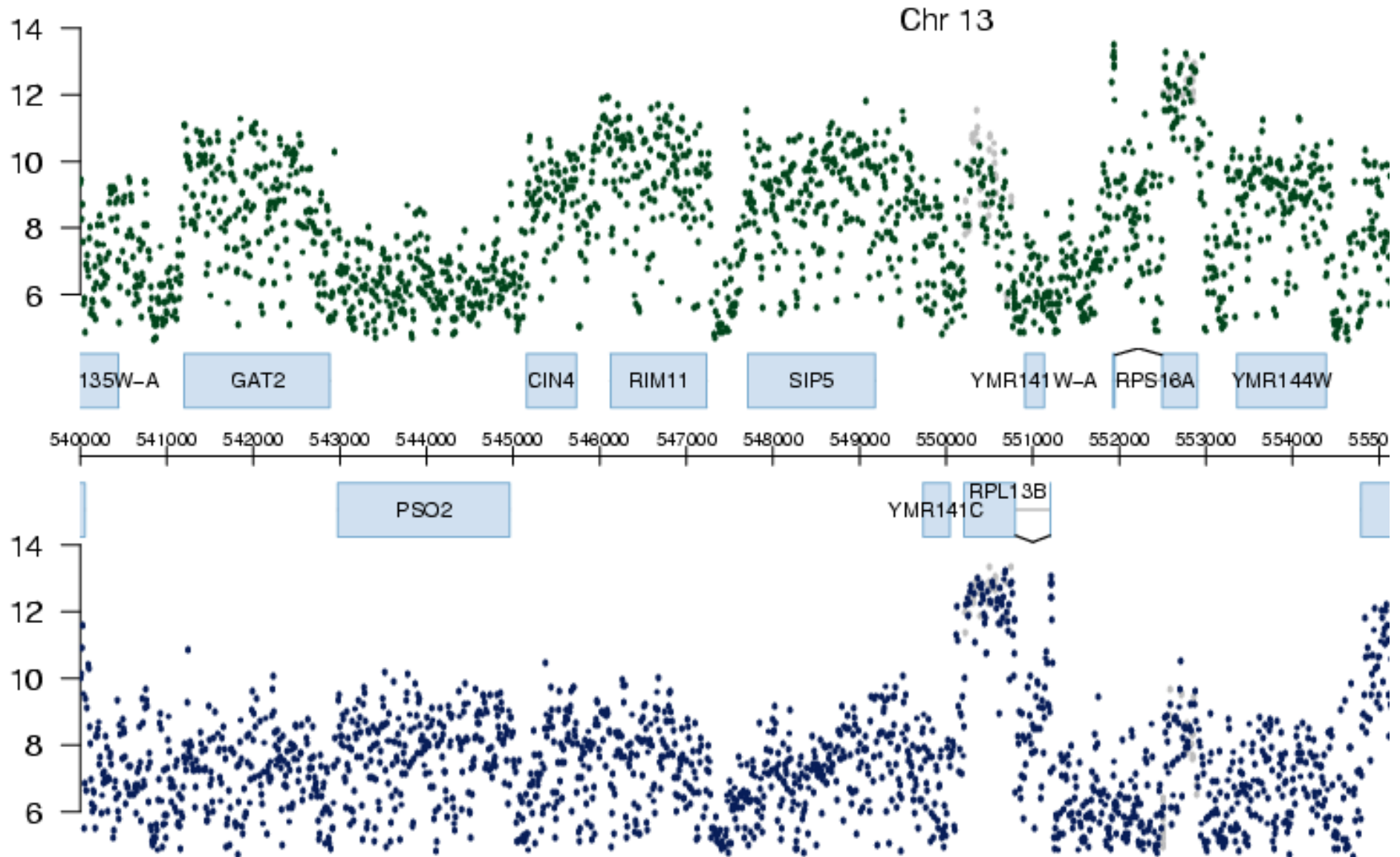


# ... *S. cerevisiae* affymetrix tiling array

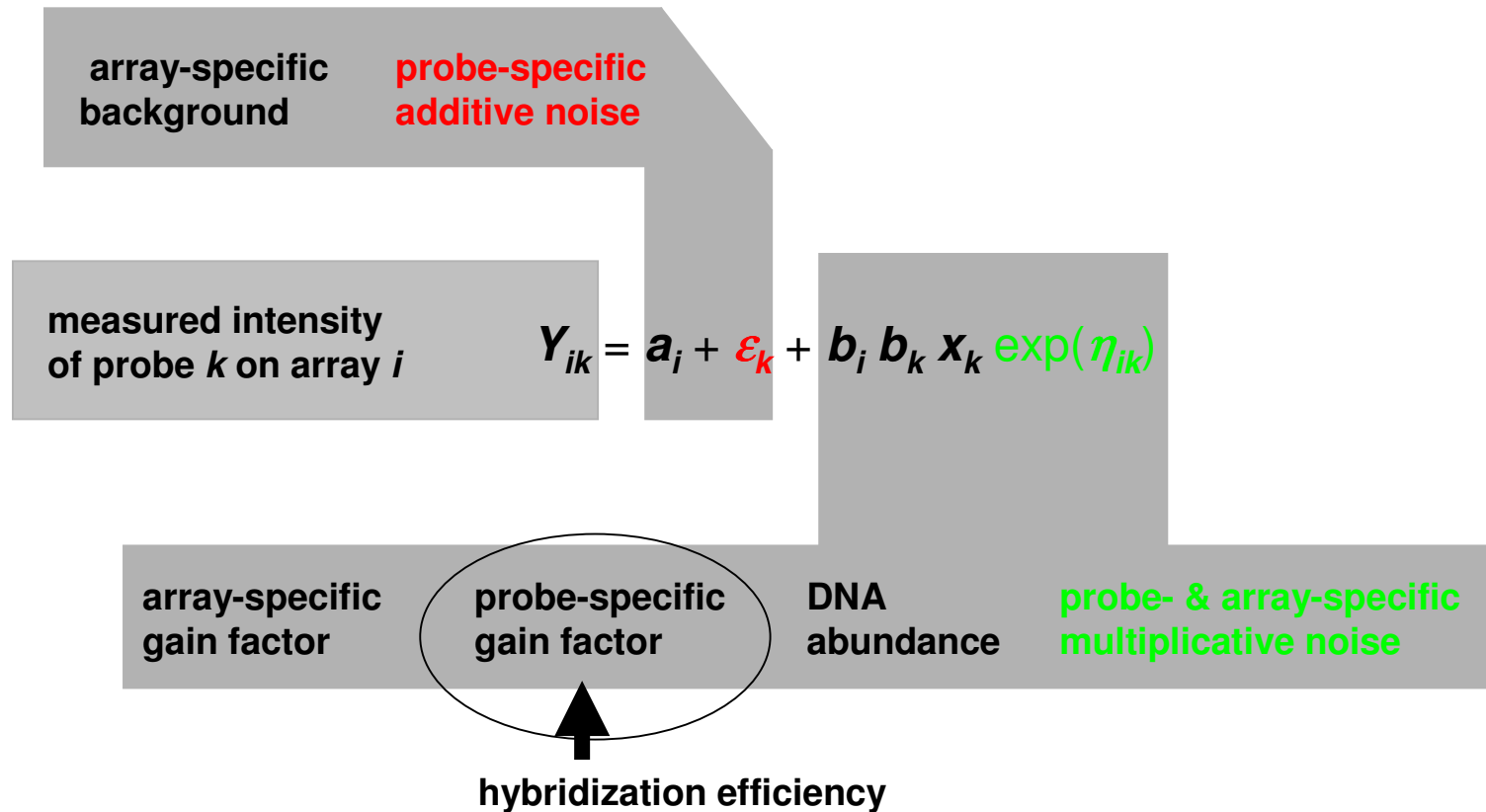


**+ RNA**

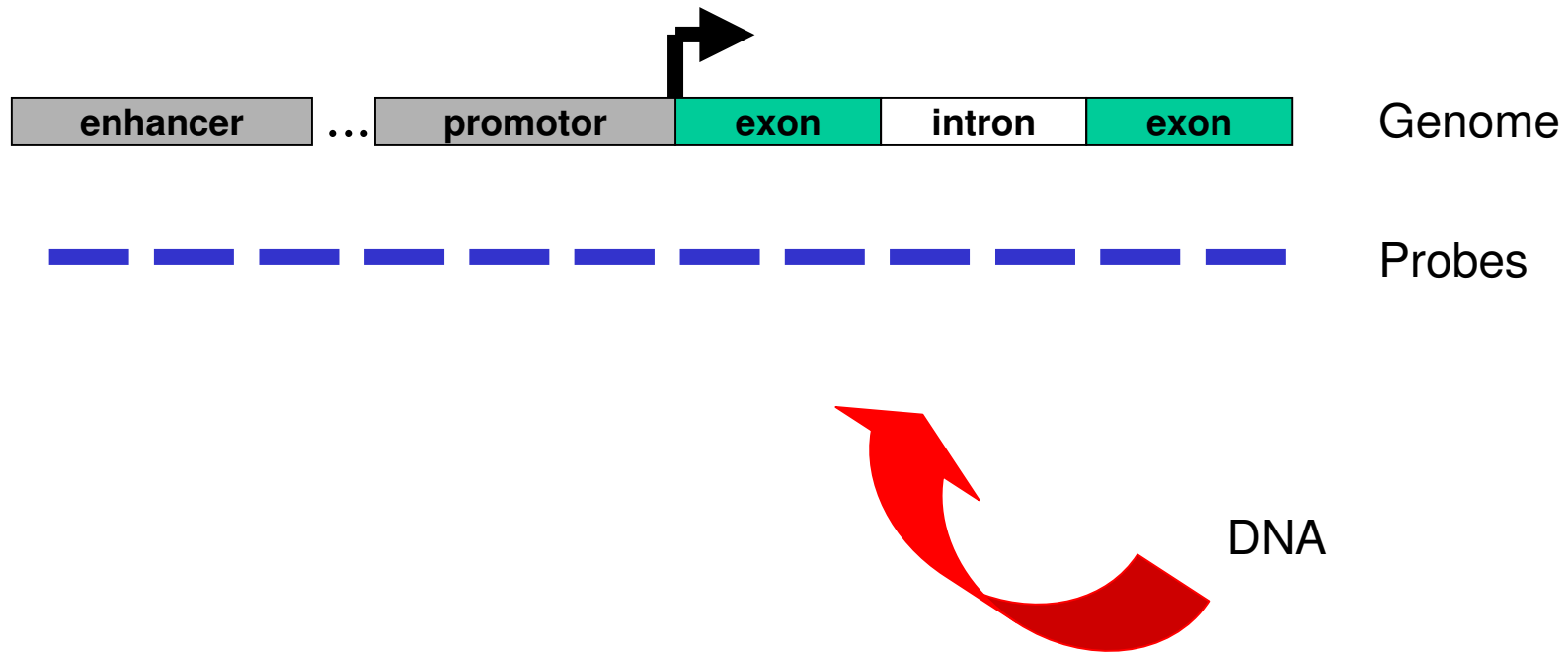
# ... the raw signal (log-scale)



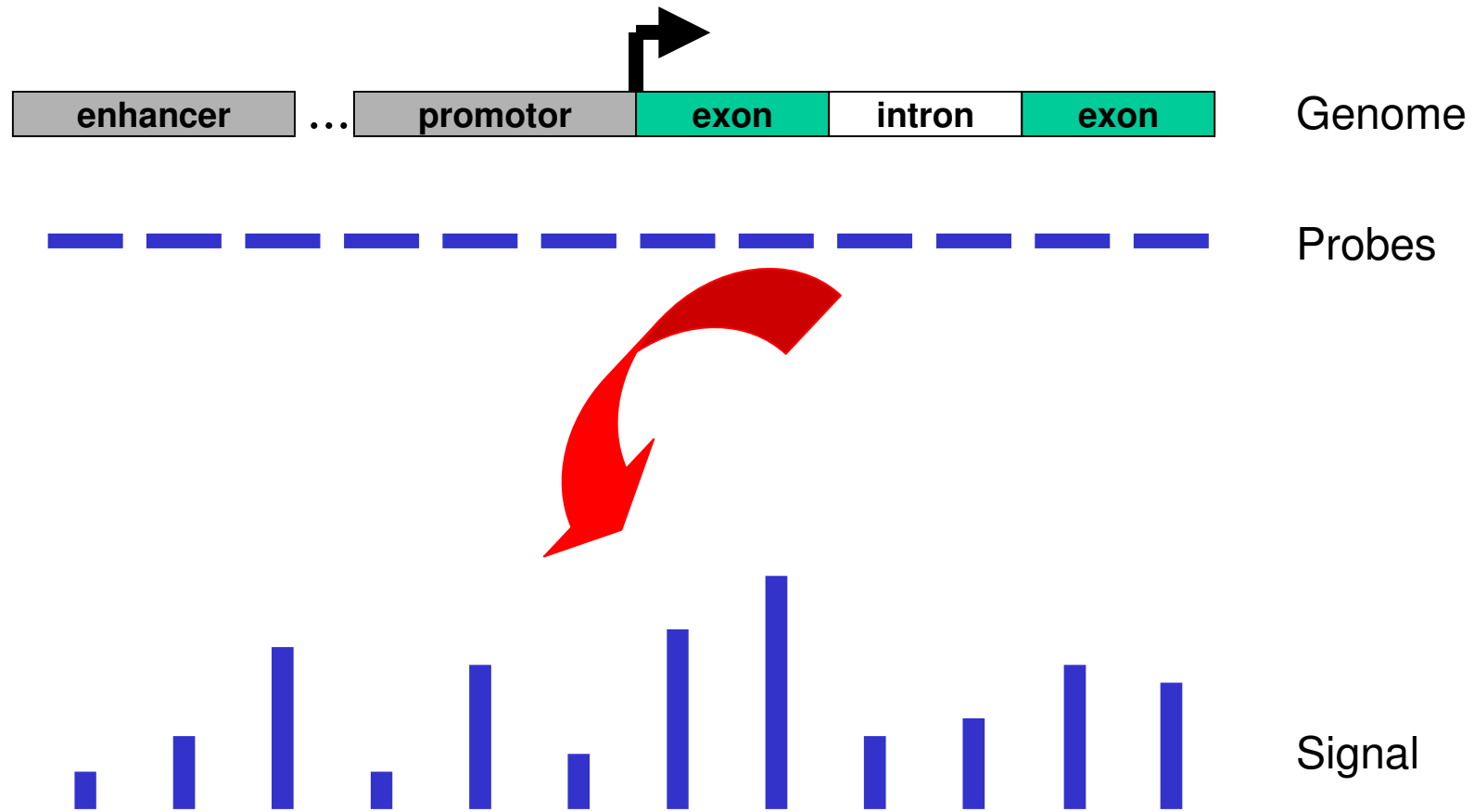
# ... again probe-specific effects



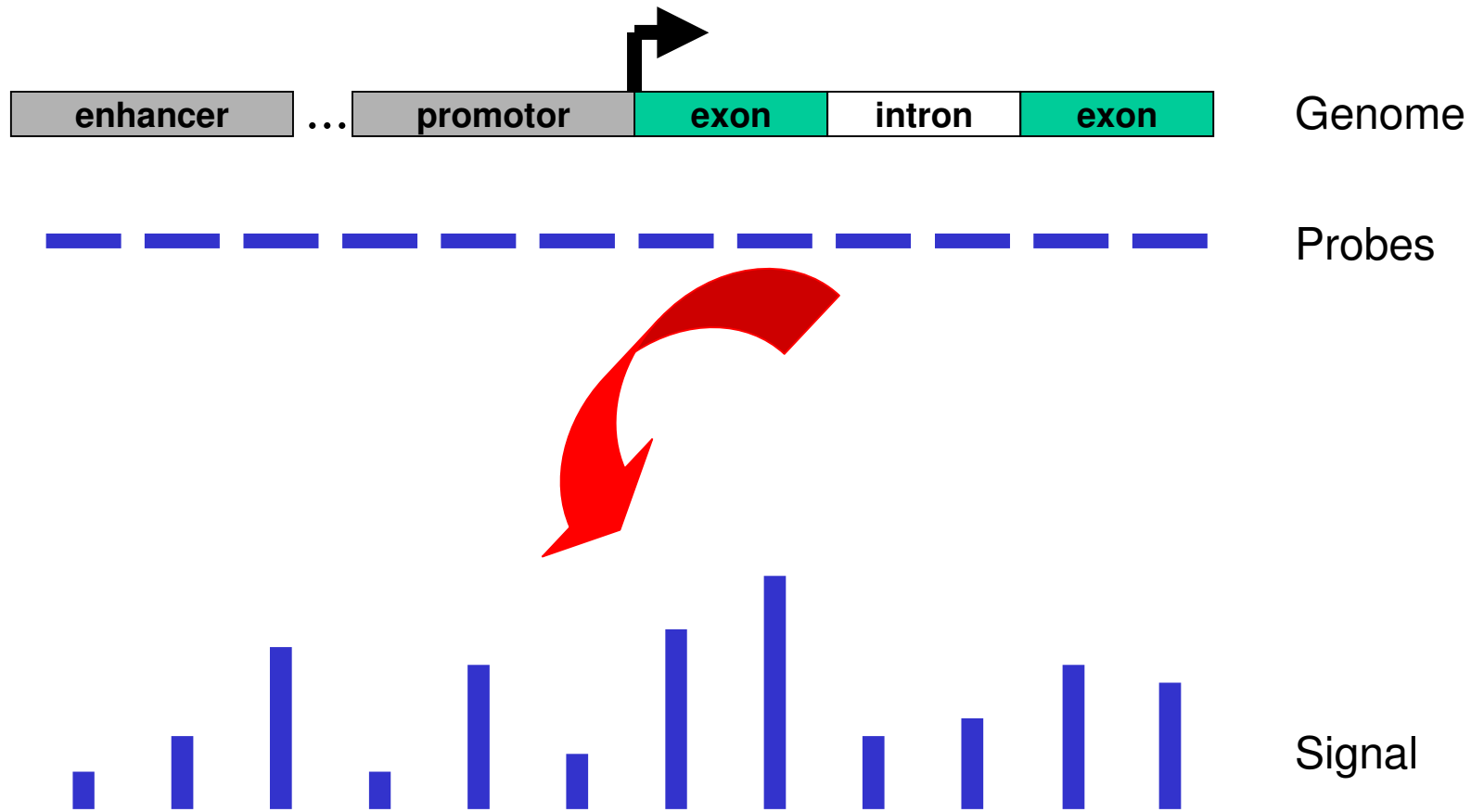
# ... tiling array



# ... tiling array



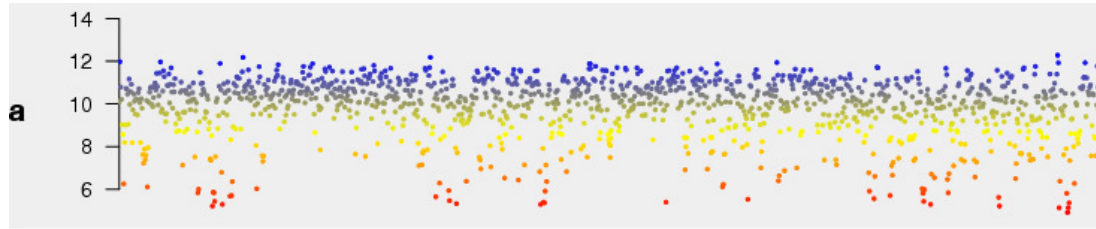
# ... tiling array



**variation due to probe-specific effect**

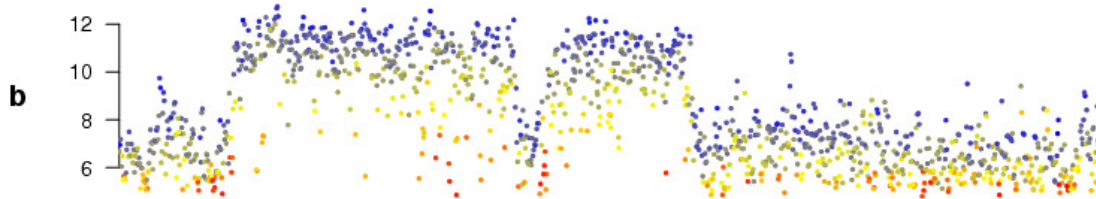
Probe  
specific  
response  
normali-  
zation

$$\log_2 s_i$$



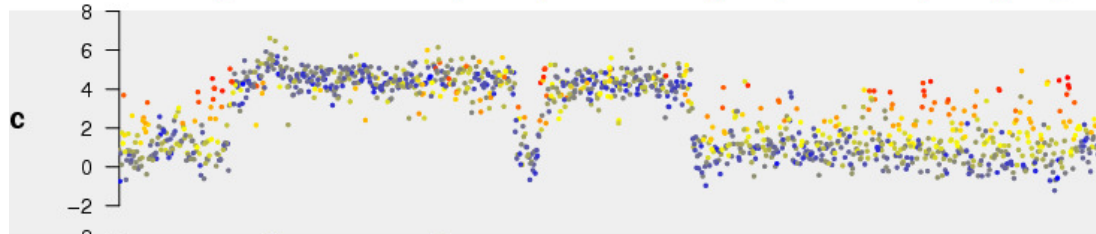
**S/N**

$$\log_2 y_i$$



**3.22**

$$q_i = \log_2 \frac{y_i}{s_i}$$



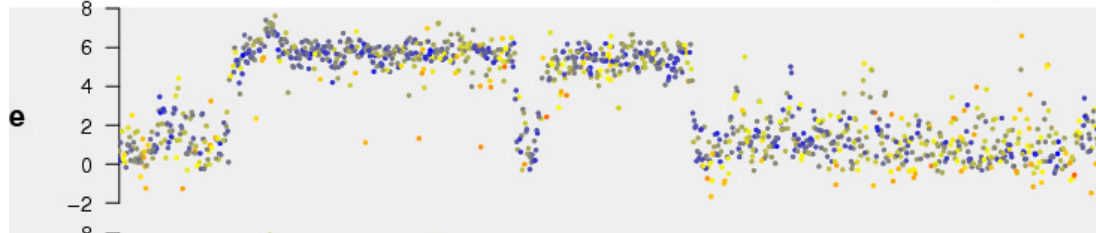
**3.47**

$$q_i = g \log_2 \frac{y_i - b(s_i)}{s_i}$$



**4.04**

remove 'dead' probes



**4.58**

$$q_i = g \log_2 \frac{PM_i - MM_i}{s_i}$$



**4.36**



# Probe-specific response normalization

$$q_i = \text{glog}_2 \frac{y_i - b(s_i)}{s_i}$$

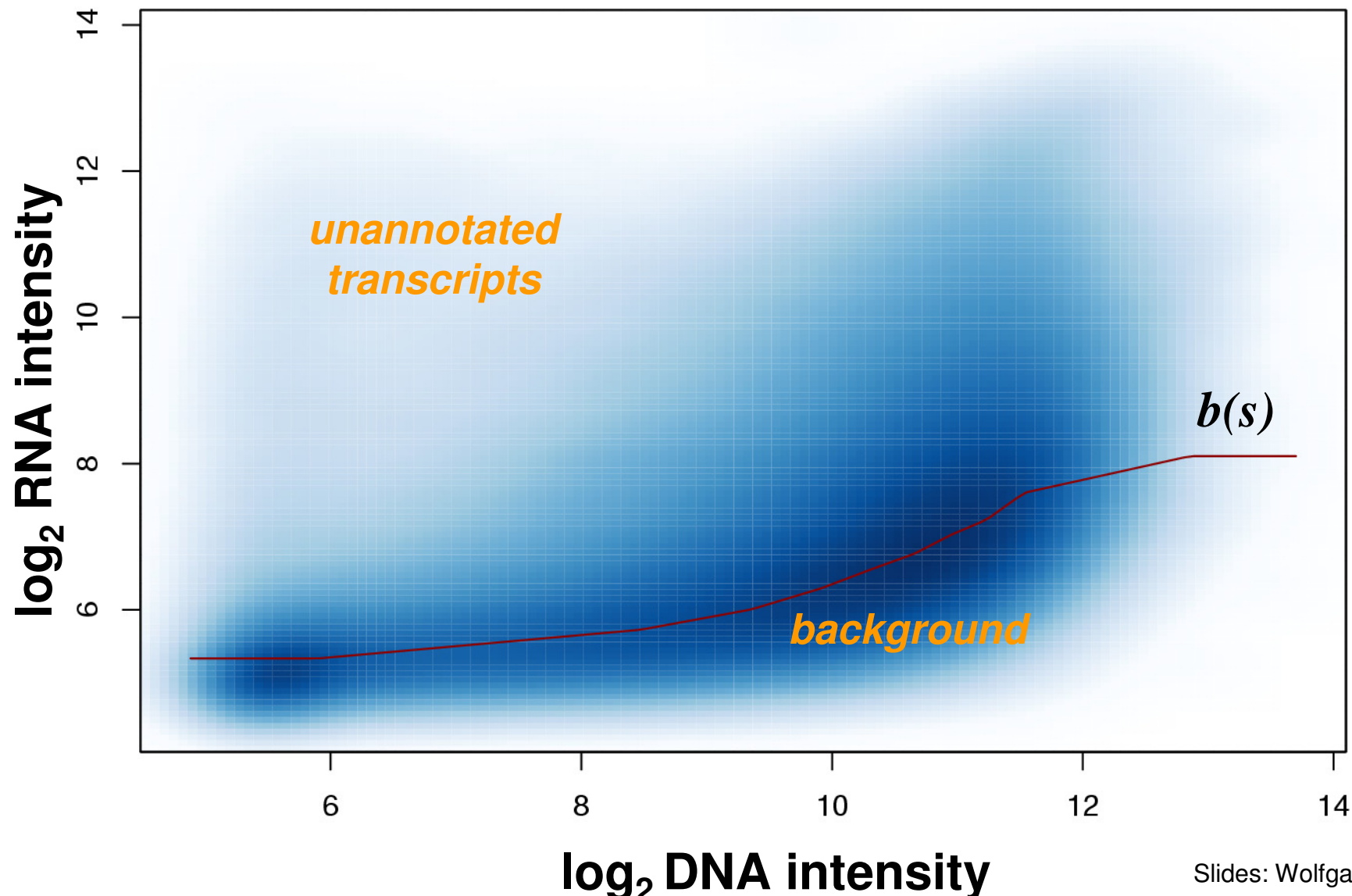
**$s_i$  probe specific response factor.**

**Estimate taken from DNA hybridization data**

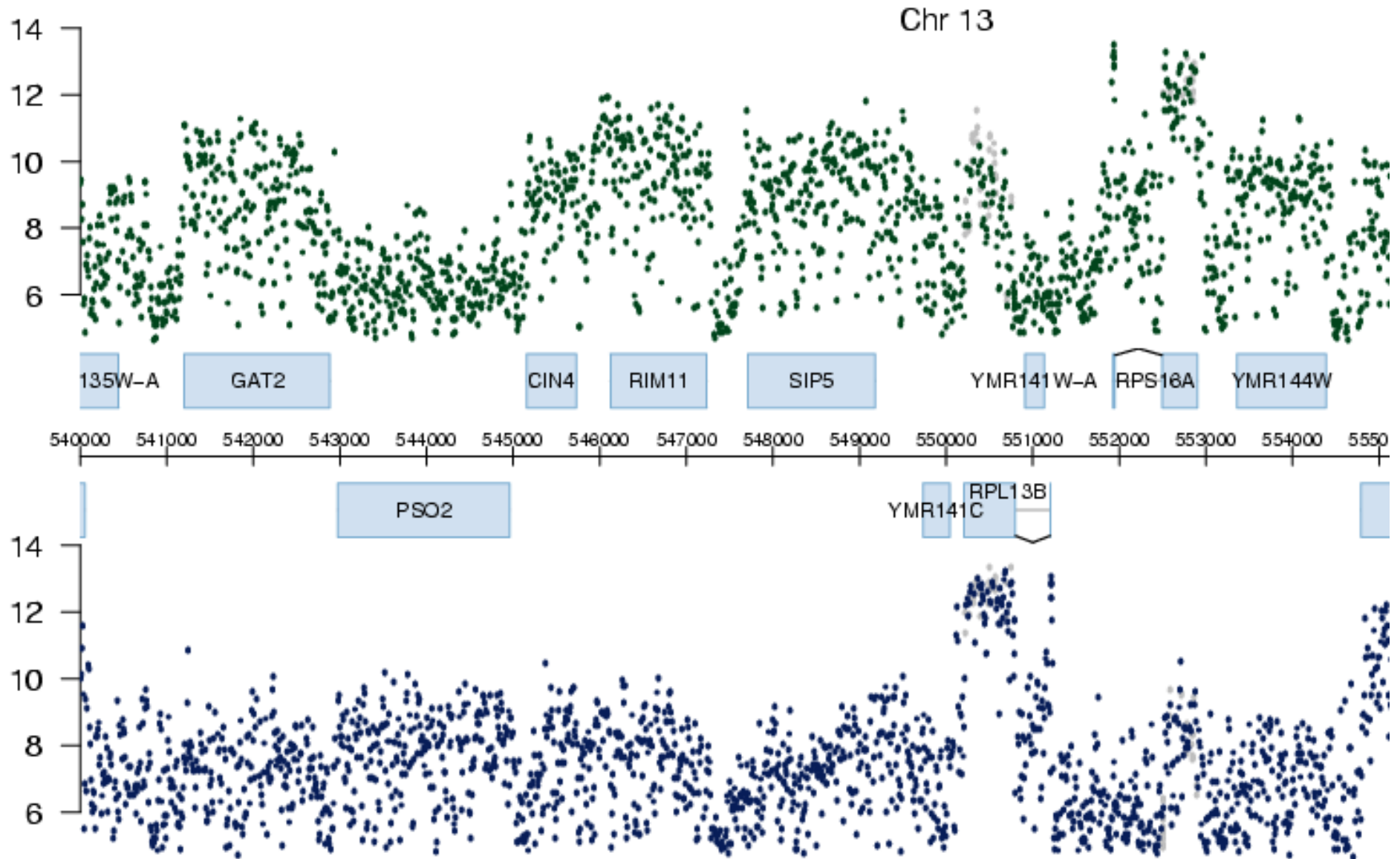
**$b_i = b(s_i)$  probe specific background term.**

**Estimation: for strata of probes with similar  $s_i$ , estimate  $b$  through location estimator of distribution of intergenic probes, then interpolate to obtain continuous  $b(s)$**

# Estimation of $b$ : joint distribution of (DNA, RNA) values of intergenic PM probes

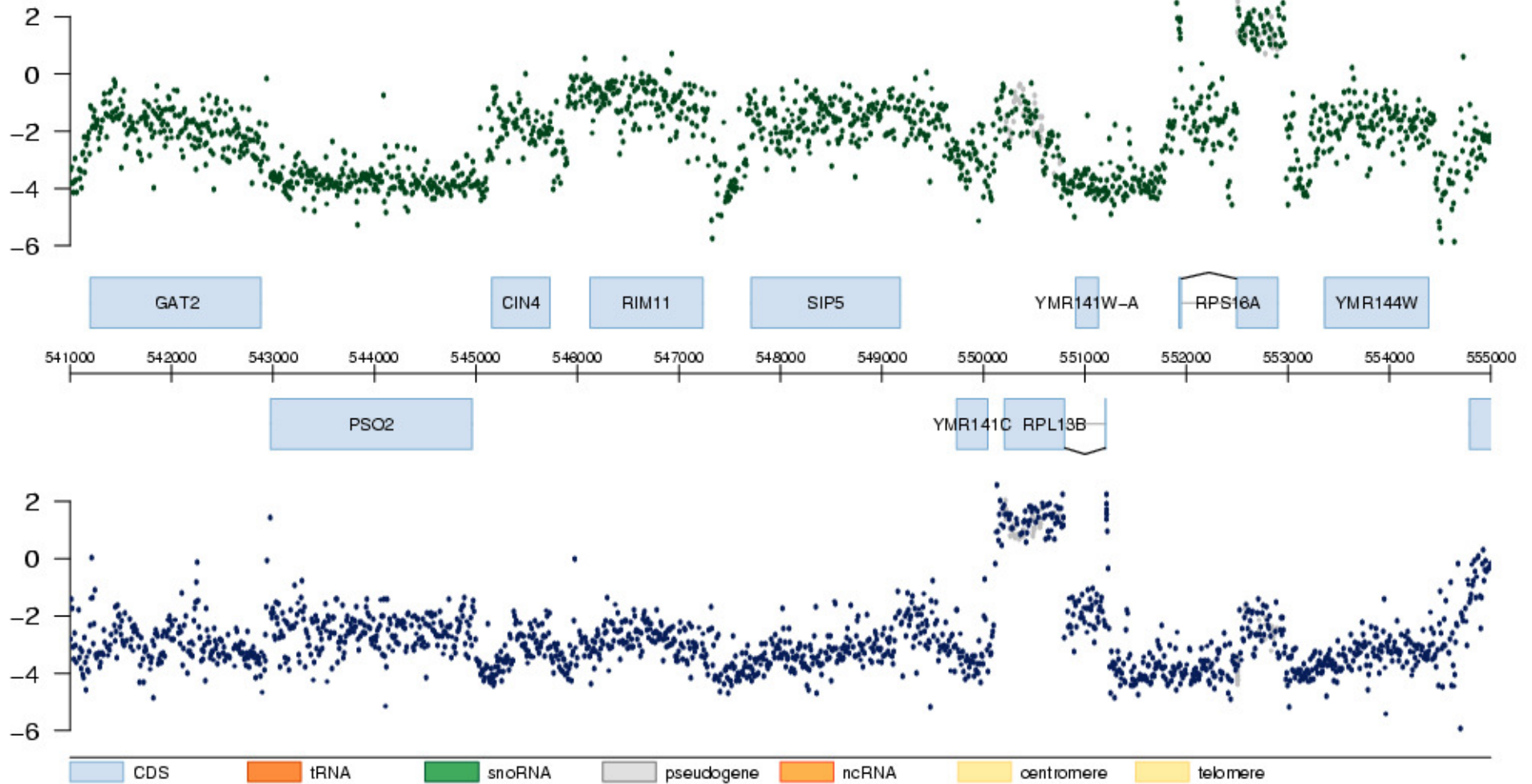


# ... before normalization



# ... after normalization

Chr 13



# ... segmentation

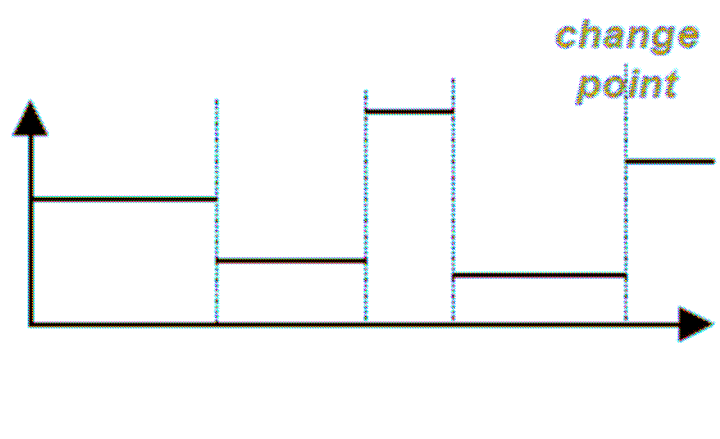
## Two obvious options:

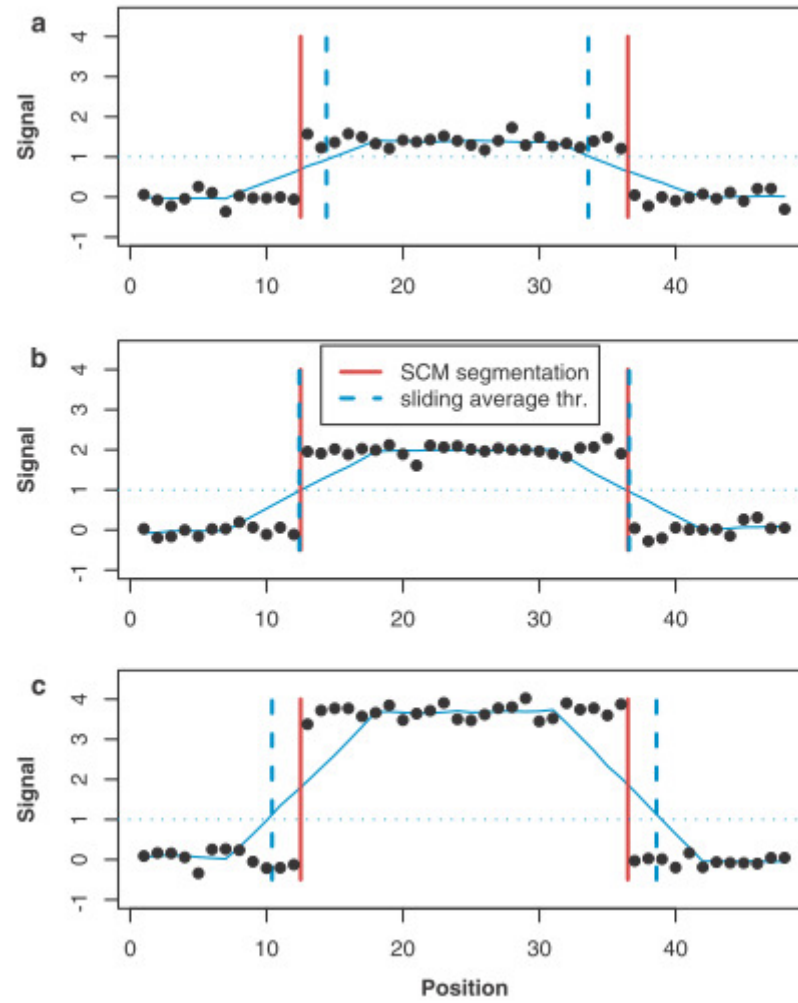
*Smoothing* and thresholding: simple, but estimates of transcript boundaries will be **biased** and depend on expression level

*Hidden Markov Model (HMM)*: but our “states” come from a continuum, unclear how to discretize

## The solution:

Fit a piecewise constant function





... Structural change model (SCM):  
piecewise constant functions

$$\forall x \in [t_{k-1}, t_k]:$$

$$Y(x) = \mu_k + \varepsilon(x)$$

$t_1, \dots, t_S$ : change points  
 $Y$ : normalized intensities  
 $x$ : genomic coordinates

$\mu_k$ : level of k-th segment

... model fitting

**Minimize**

$$G(t_1, \mathbf{K}, t_S) = \sum_{s=1}^S \sum_{j=1}^J \sum_{i \geq t_s}^{i < t_{s+1}} \left( y_{ij} - \bar{y}_{sj} \right)^2$$

$t_1, \dots, t_S$ : change points

$J$ : number of replicate arrays

# ... optimization

Naïve optimization has complexity  $n^s$ , where  $n \approx 10^5$  and  $s \approx 10^3$ .

Fortunately, there is a **dynamic programming** algorithm with complexity  $O(n^2)$ , and good heuristic  $O(n)$ :

$$k = 0, \quad \forall 0 \leq i < j \leq n \quad \hat{J}_1(i, j) = \sum_{x=i+1}^j \left\{ \log(2\pi \times \hat{\sigma}_1^2) + \left[ \frac{y(x) - \hat{\mu}_1}{\hat{\sigma}_1} \right]^2 \right\}$$
$$\forall k \in [1, K_{max}] \quad \hat{J}_{k+1}(1, j) = \min_h \left\{ \hat{J}_k(1, h) + \hat{J}_1(h+1, j) \right\}$$

F. Picard, S. Robin, M. Lavielle, C. Vaisse, G. Celeux, JJ Daudin, BMC Bioinformatics (2005)

Bai+Perron, Journal of Applied Econometrics (2003)

Software: W. Huber, package `tilingArray`, [www.bioconductor.org](http://www.bioconductor.org)

A. Zeileis, package `strucchange`, CRAN

... result

