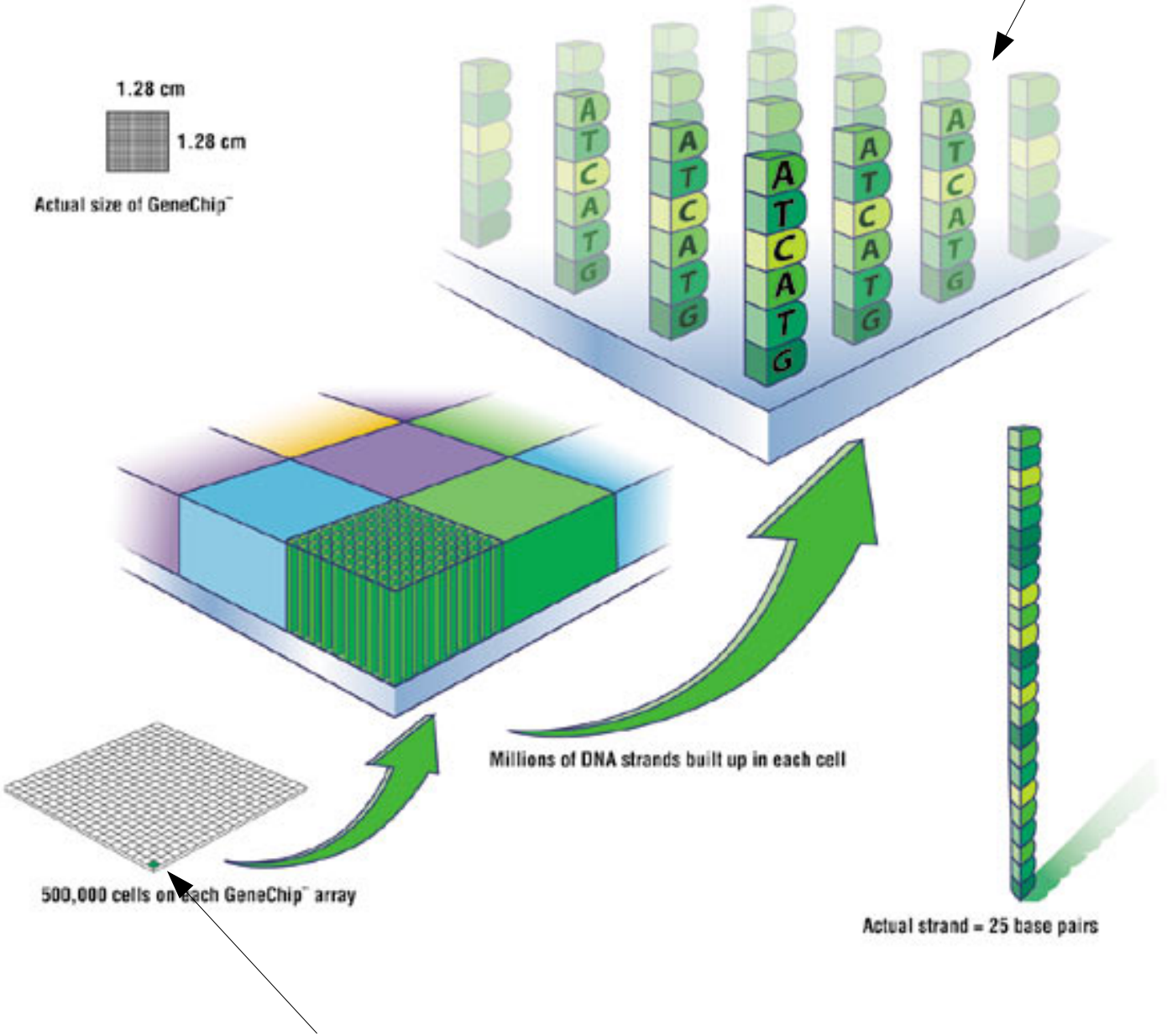
Models for Microarray Analysis: Sequence Effects and RNA Degradation

„All models are wrong, and increasingly you can succeed without them.”

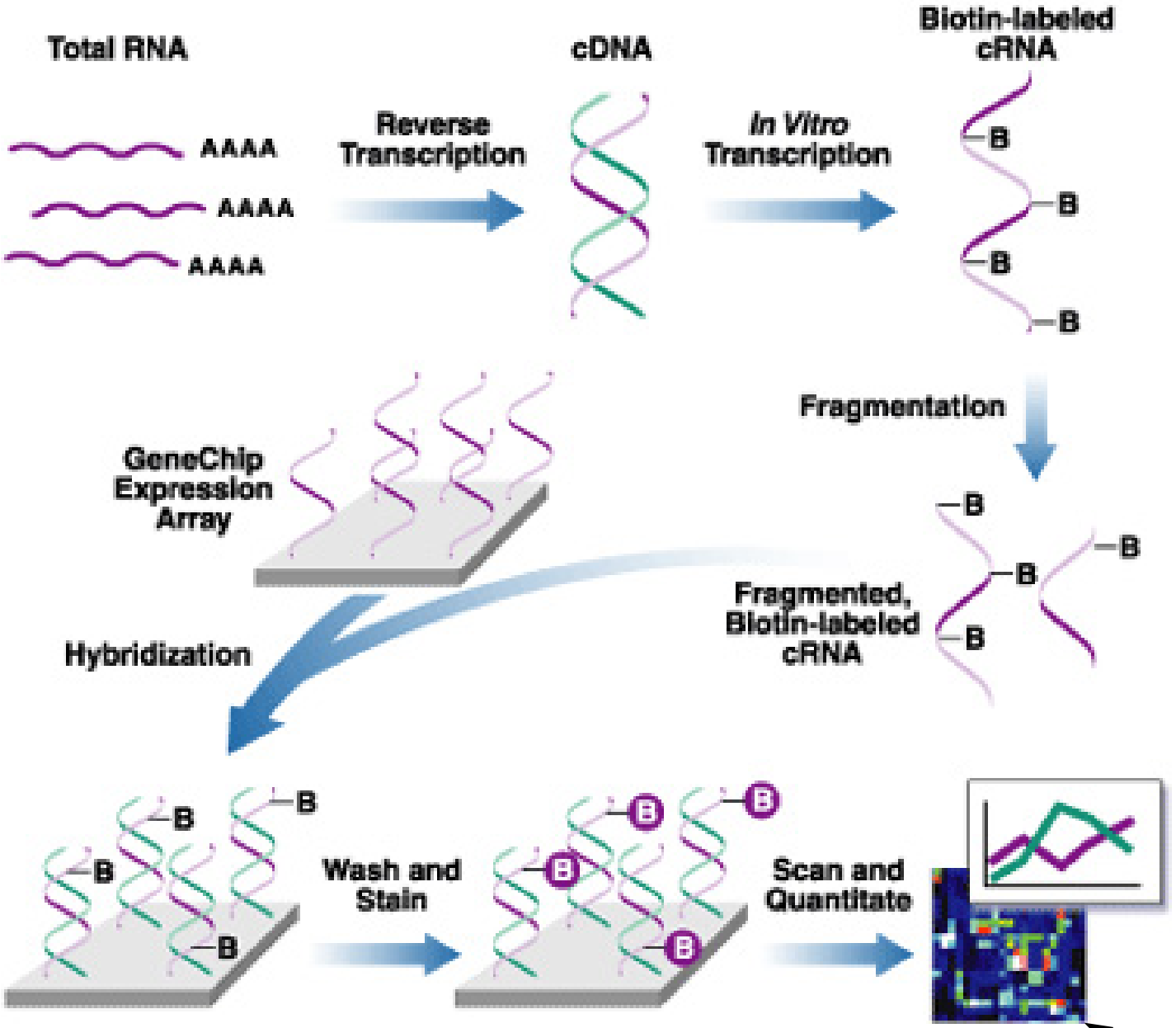
(Peter Norvig, Research Director Google Inc.)

Affymetrix Microarrays

Probe Sequence $\xi_p \in \{A, C, G, T\}^{25}$

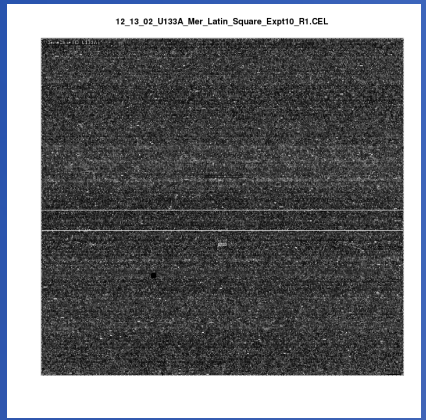


Probe Position on the Chip X_p, Y_p

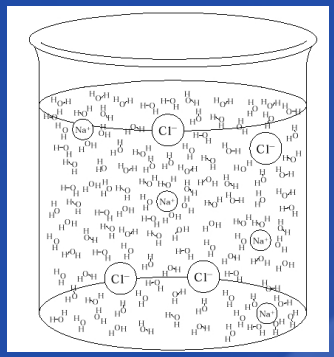


Probe Intensity I_p

Probe 1 I_p
Probe 2 I_p
Probe 3 I_p
...
Probe 11 I_p

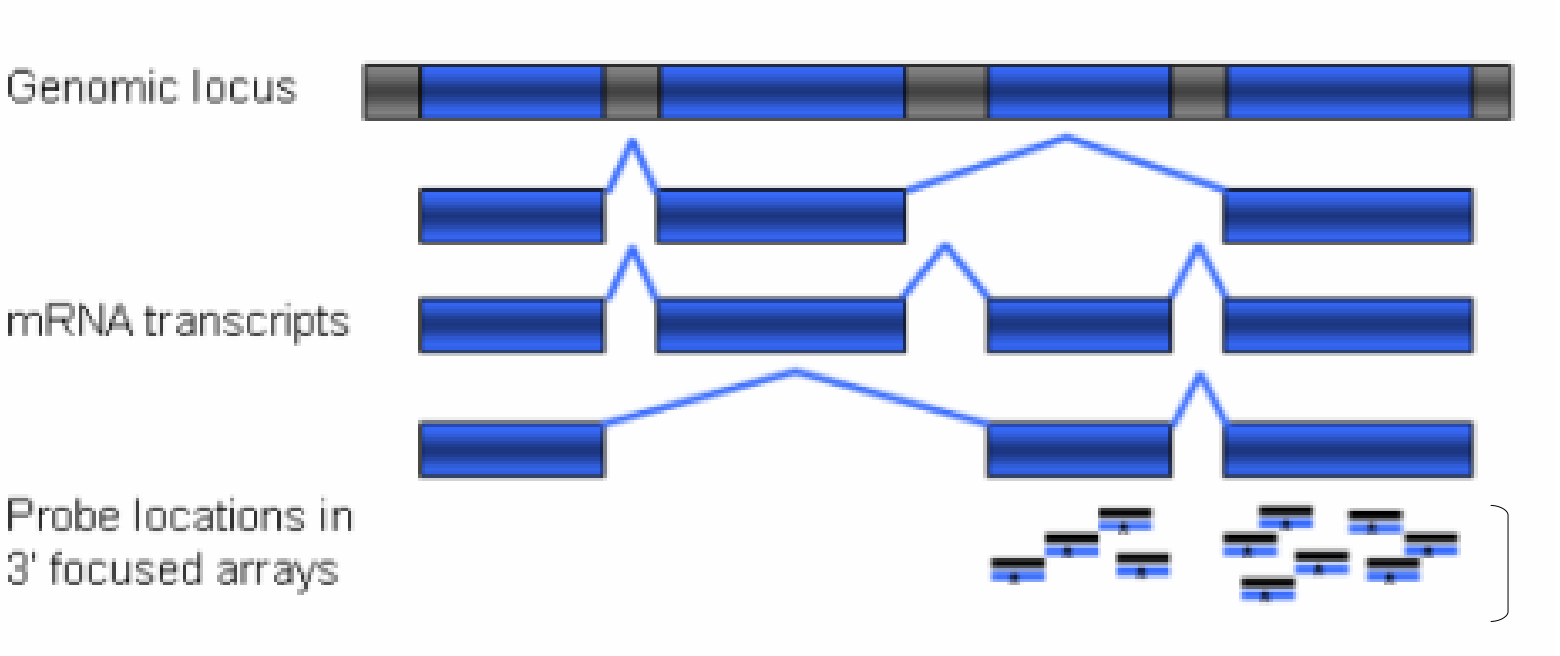


„Microarray Correction”



mRNA Concentration

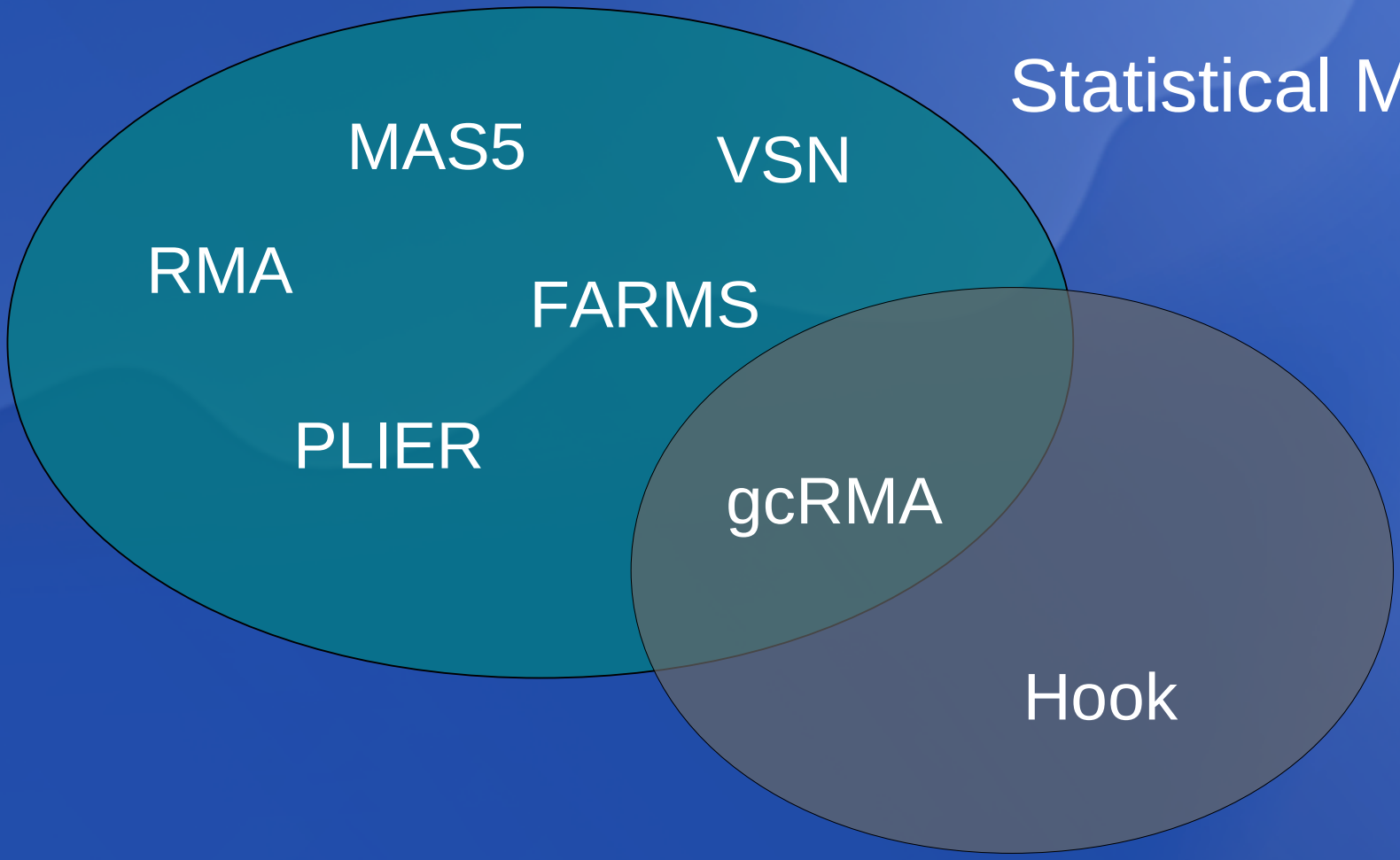
(Expression Measure)



Probeset

Probes are designed to match sequences towards the 3' end in expression arrays

Microarray Correction Methods



Comparison available at http://affycomp.biostat.jhsph.edu

The „Hook”-Method

Saturation

$$I_p = \frac{L_p}{1 + L_p / I^{max}} + I^{min}$$

The „Hook”-Method

Saturation

$$I_p = \frac{L_p}{1 + L_p / I^{max}} + I^{min}$$

Two-species: specific and non-specific binding [S]/[NS]

$$L_p = L_p^N + L_p^S$$

The „Hook”-Method

Saturation

$$I_p = \frac{L_p}{1 + L_p / I^{max}} + I^{min}$$

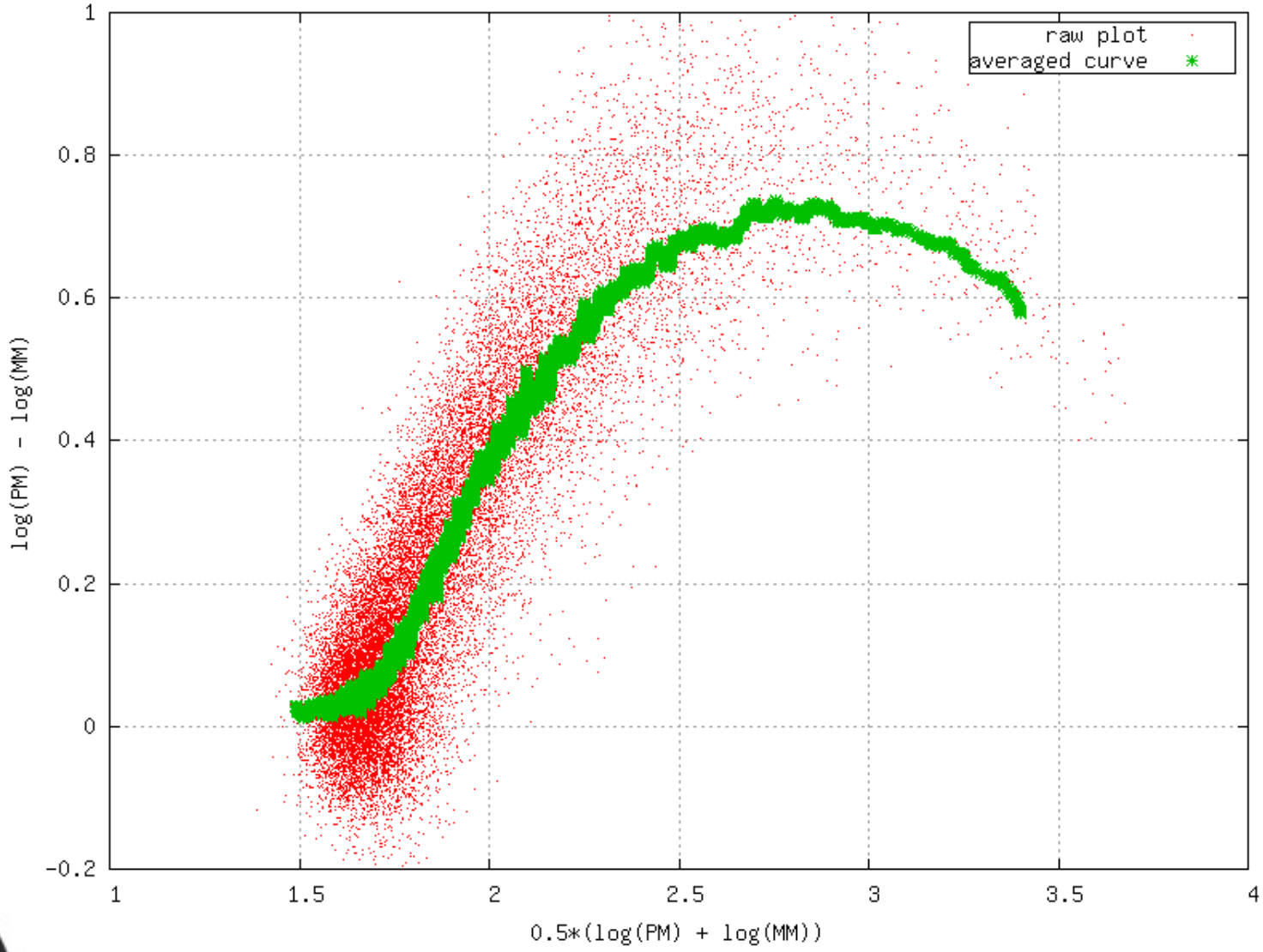
Two-species: specific and non-specific binding [S]/[NS]

$$L_p = L_p^N + L_p^S$$

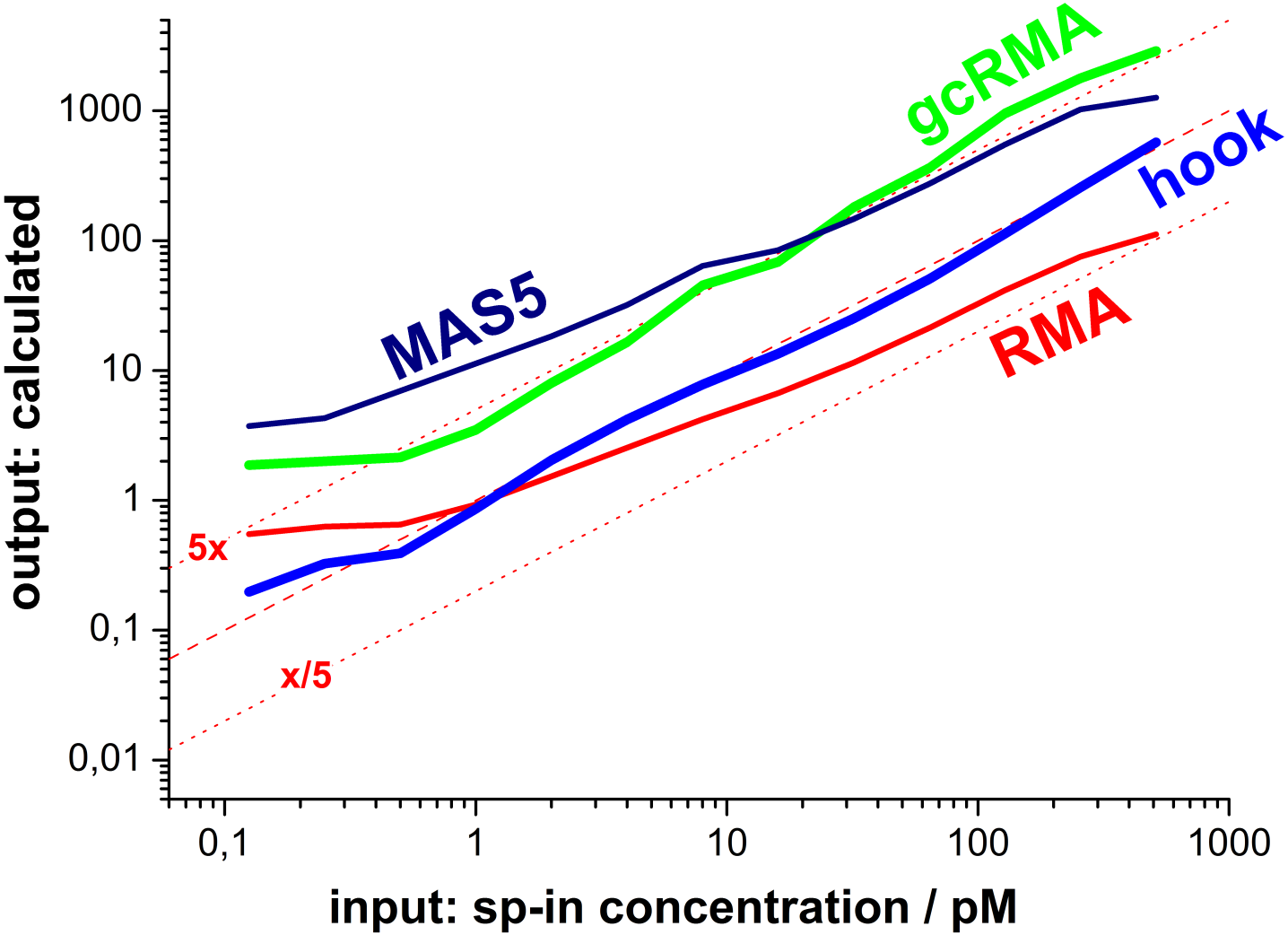
Sequence effects

$$\log L^S = \log[S] + \delta A(\xi)$$

$$\log L^N = \log[N] + \delta A(\xi)$$


raw plot
averaged curve
log(PM) - log(MM)
0.5*(log(PM) + log(MM))
1
0.8
0.6
0.4
0.2
0
-0.2
1
1.5
2
2.5
3
3.5
4

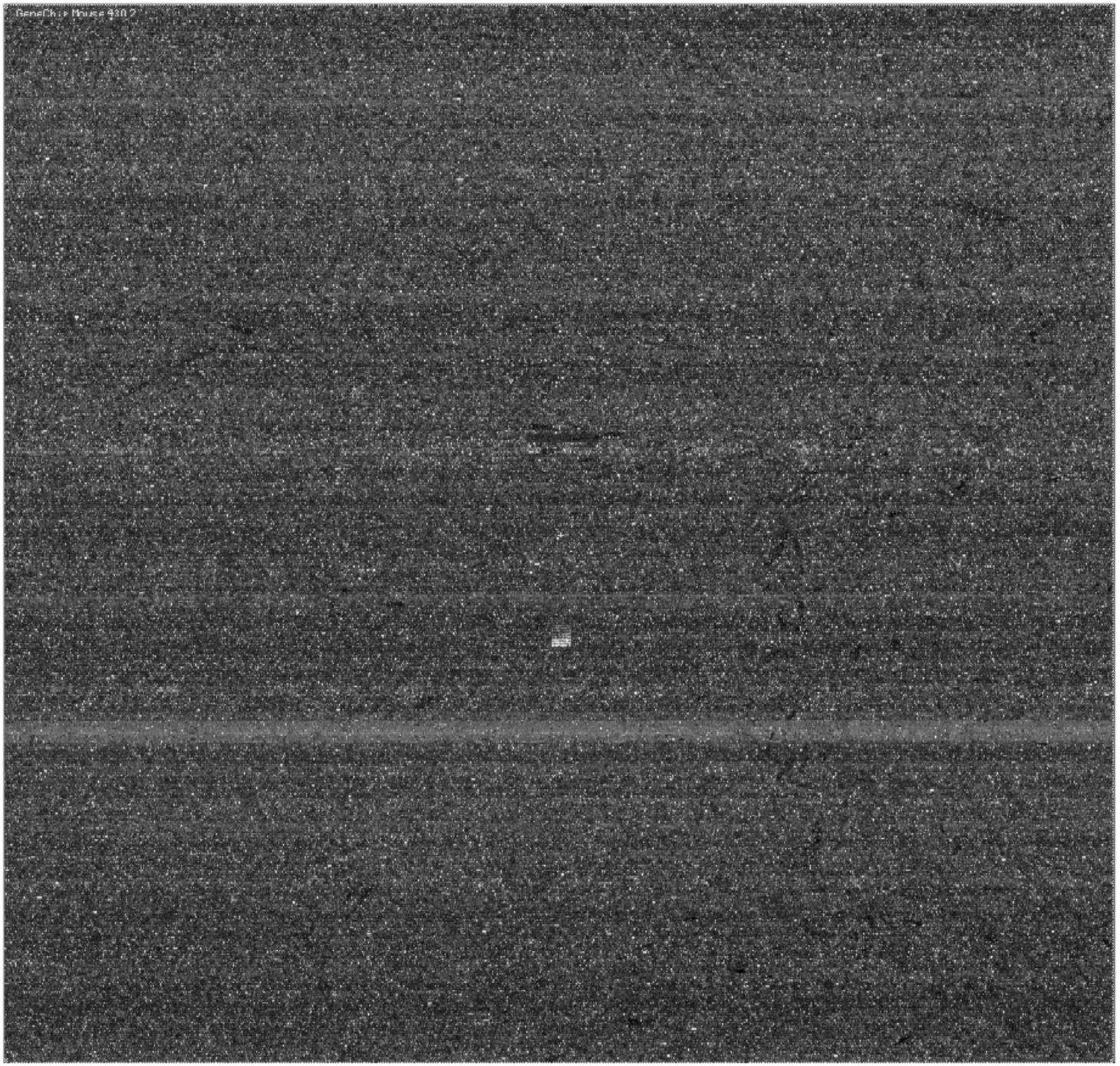
The „Hook”



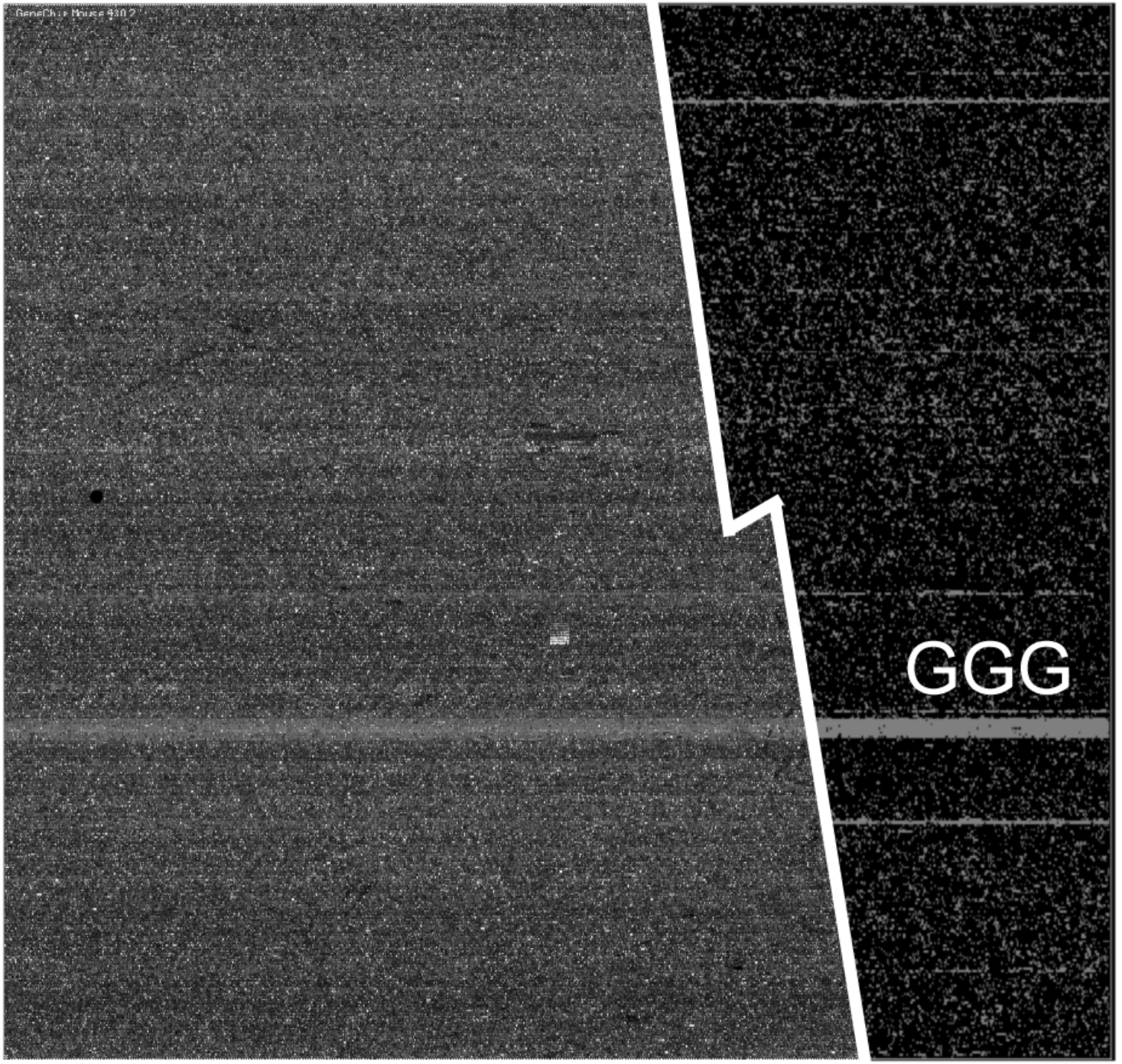
Hans Binder and Stephan Preibisch: "Hook"-calibration of GeneChip-microarrays: Theory and algorithm; *Algorithms for Molecular Biology, 2008*

http://www.izbi.de/downloads_links/programs/hook.php

dummy.cel



dummy.cel



Sequence Effects

$$\log L^N = \log[N] + \delta A(\xi)$$

$$\delta A(\xi) = \sum_{k=1}^{25-r+1} \sigma_k(\xi^{k,k+r-1})$$

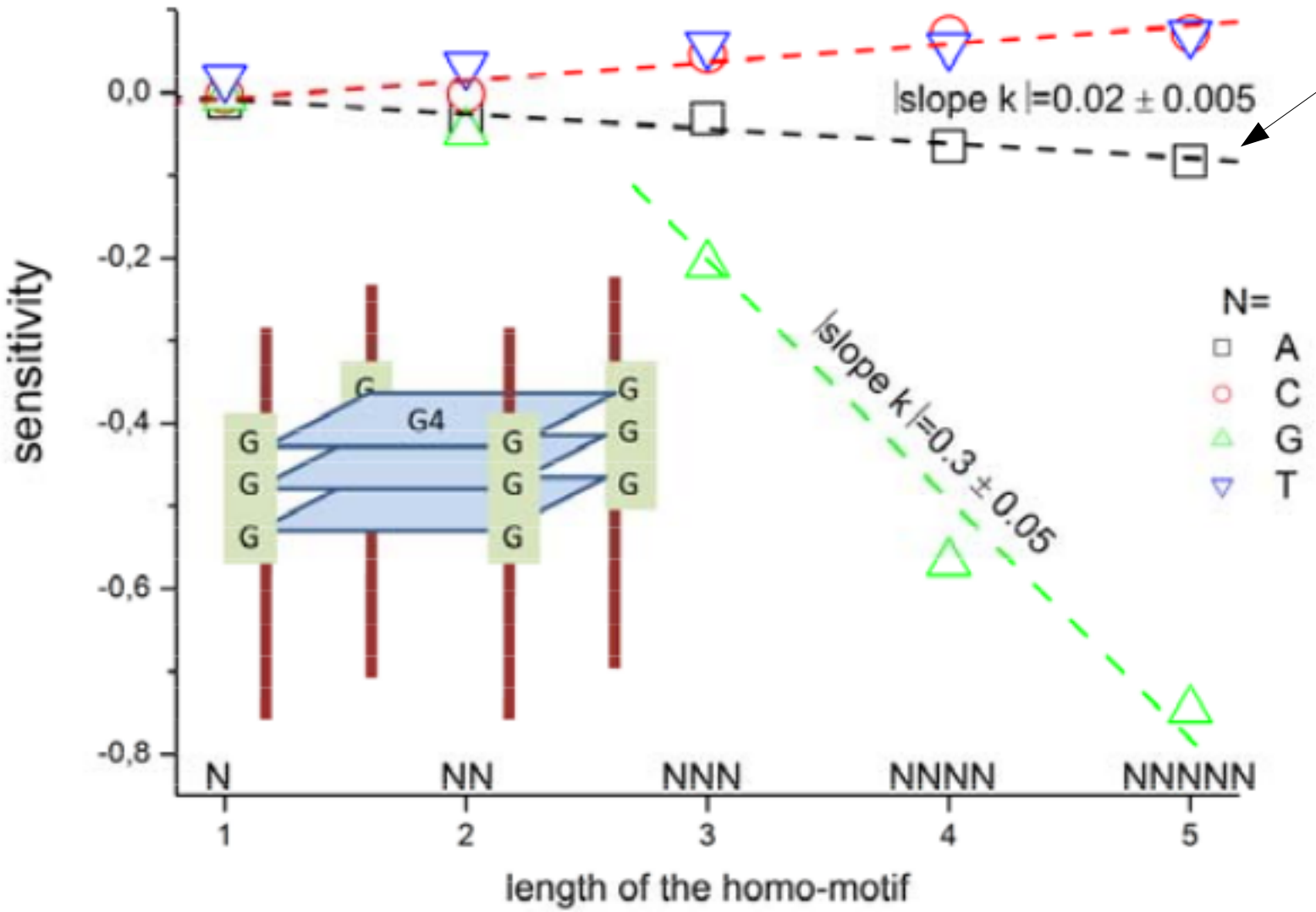
Example:

$r = 2$(Nearest-Neighbor-Model), Sequence $\xi = \mathrm{GTGACCGTTATCCA}$

$$\delta A(\xi) = \sigma_1(GT) + \sigma_2(TG) + \cdots + \sigma_{24}(CA)$$

Stacks of G-tedrads

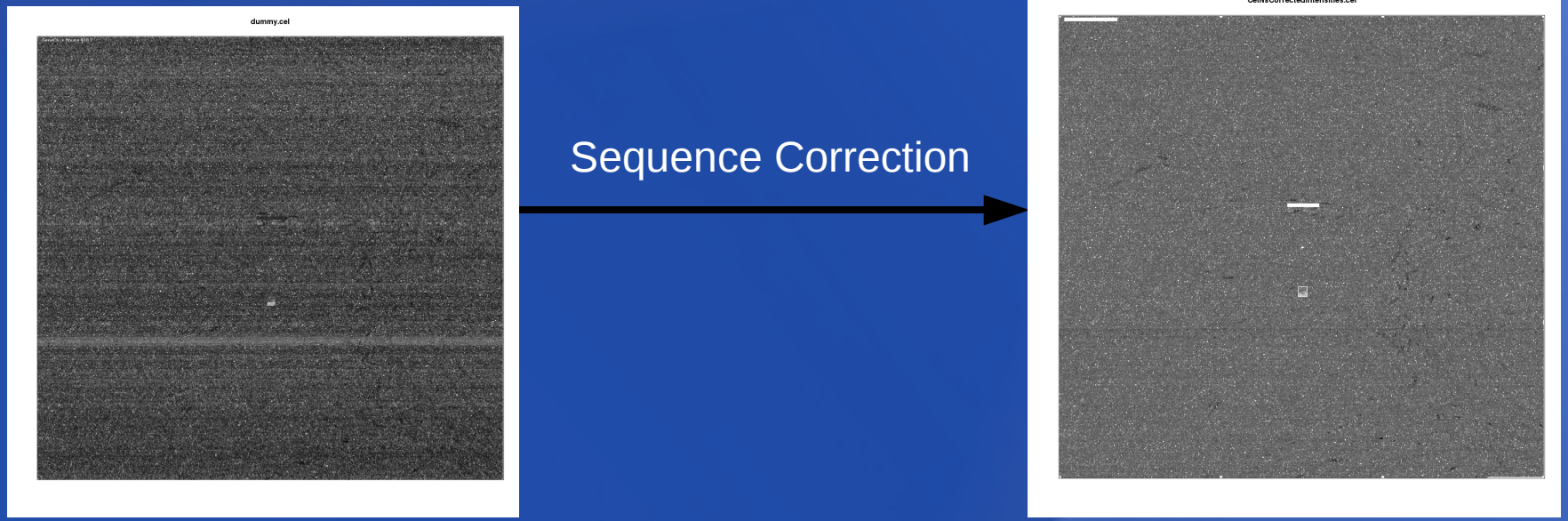
$$\sum_{k=1}^{20} \sigma_k(AAAAA)$$



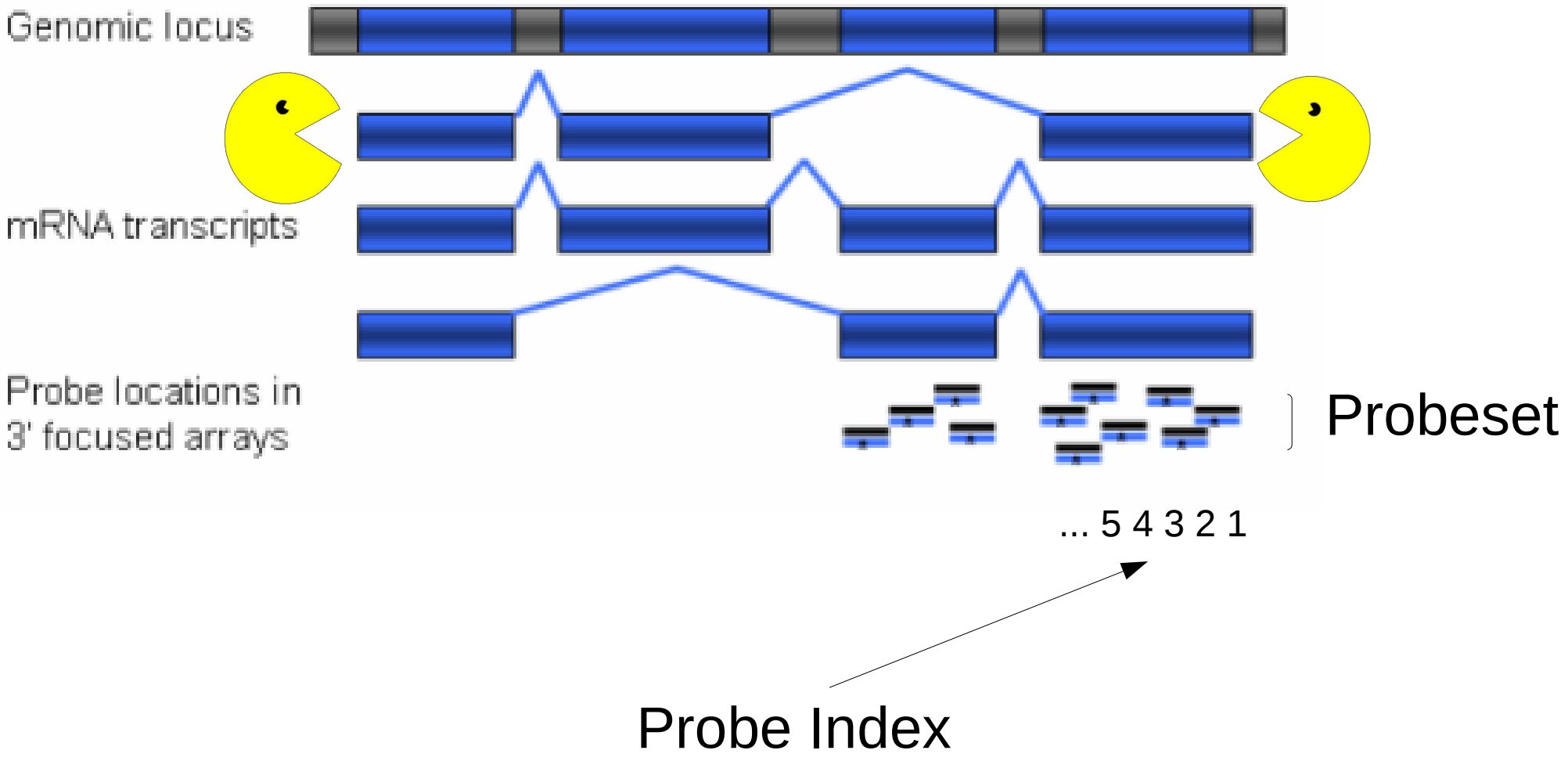
Correcting with the NN+GGG Sequence Affinity Model

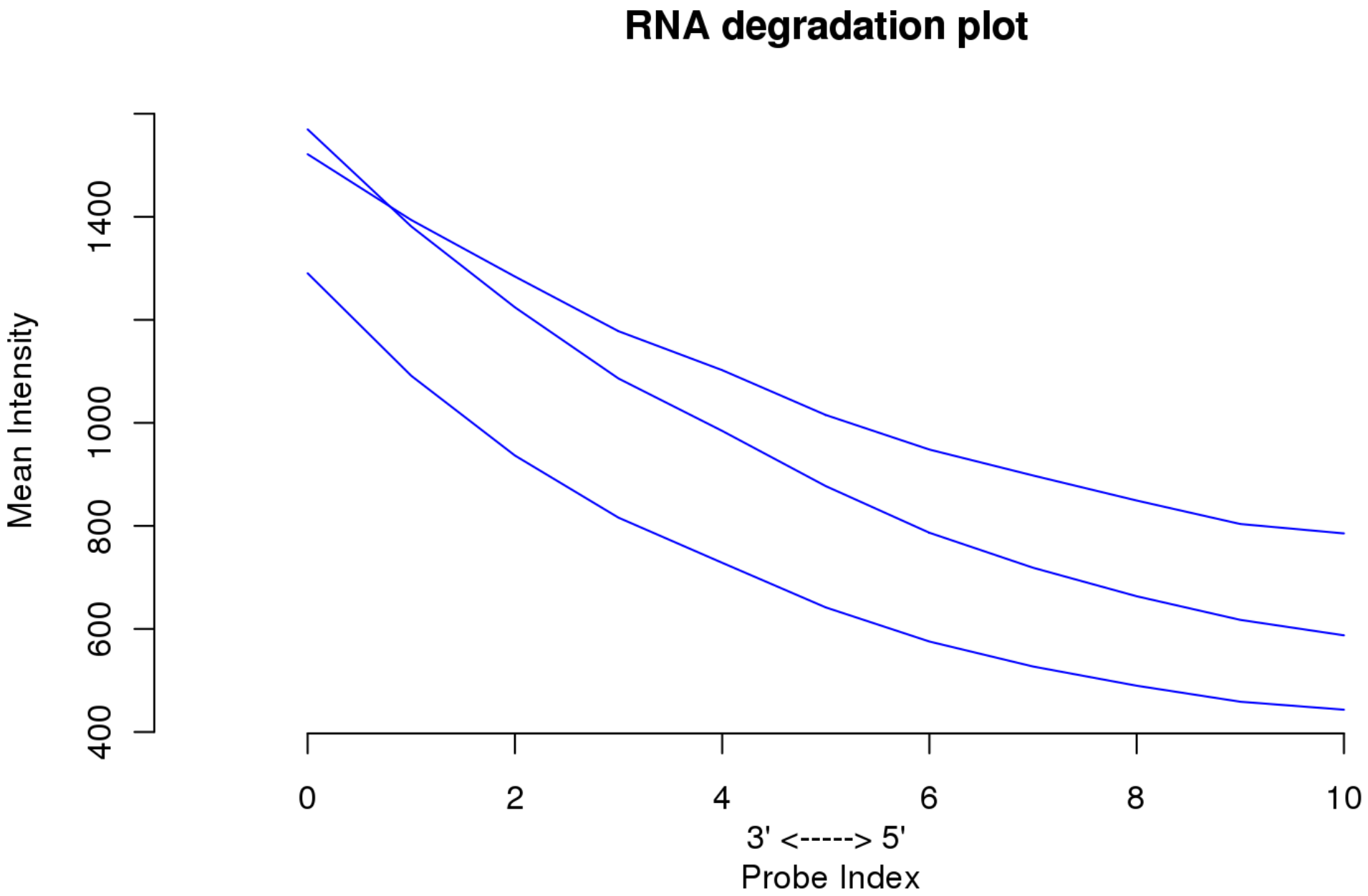
$$\delta A(\xi_p) = \sum_{i=1}^{24} \sigma_i(\xi_{p,i\ldots i+1}) + \sum_{i=1}^{23} \widetilde{\sigma}_i(\xi_{p,i\ldots i+2})$$

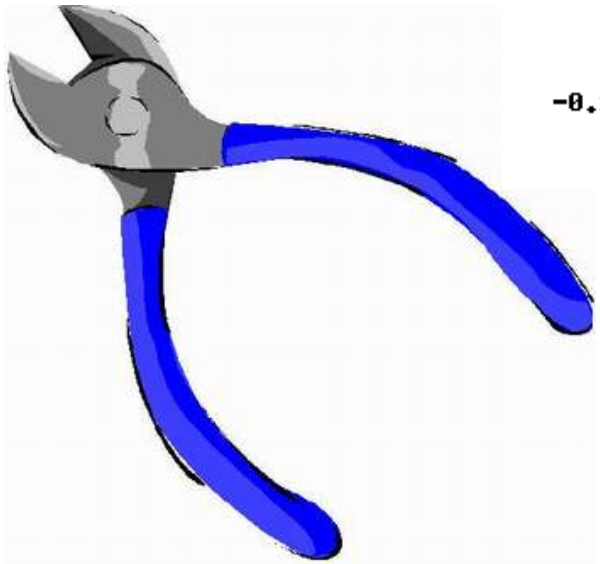
$$\widetilde{\sigma}_i(b) = 0 \text{ for } b \neq GGG$$

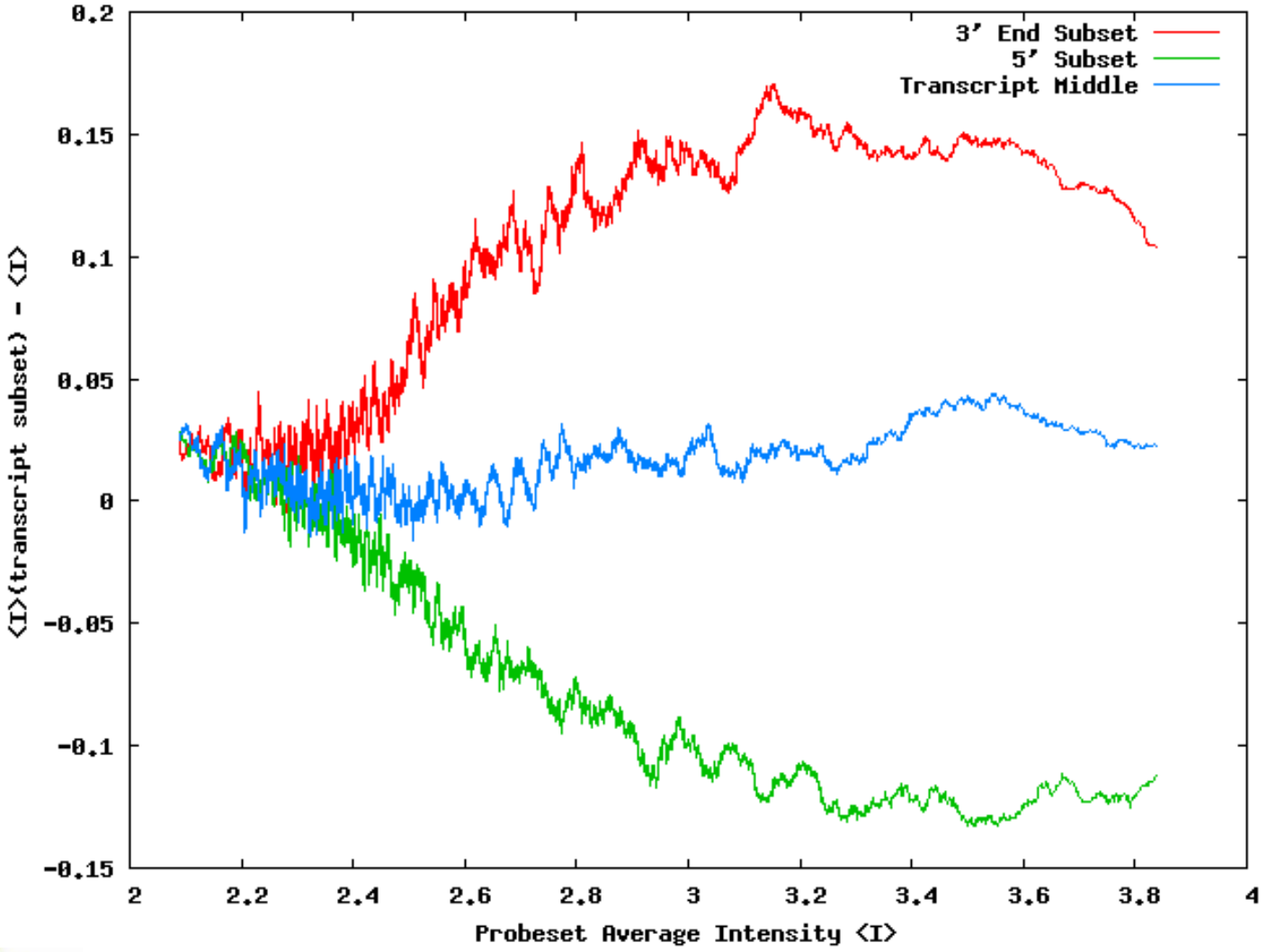


RNA Degradation




RNA degradation plot
Mean Intensity
400
600
800
1000
1400
0
2
4
6
8
10
3' <-----> 5'
Probe Index



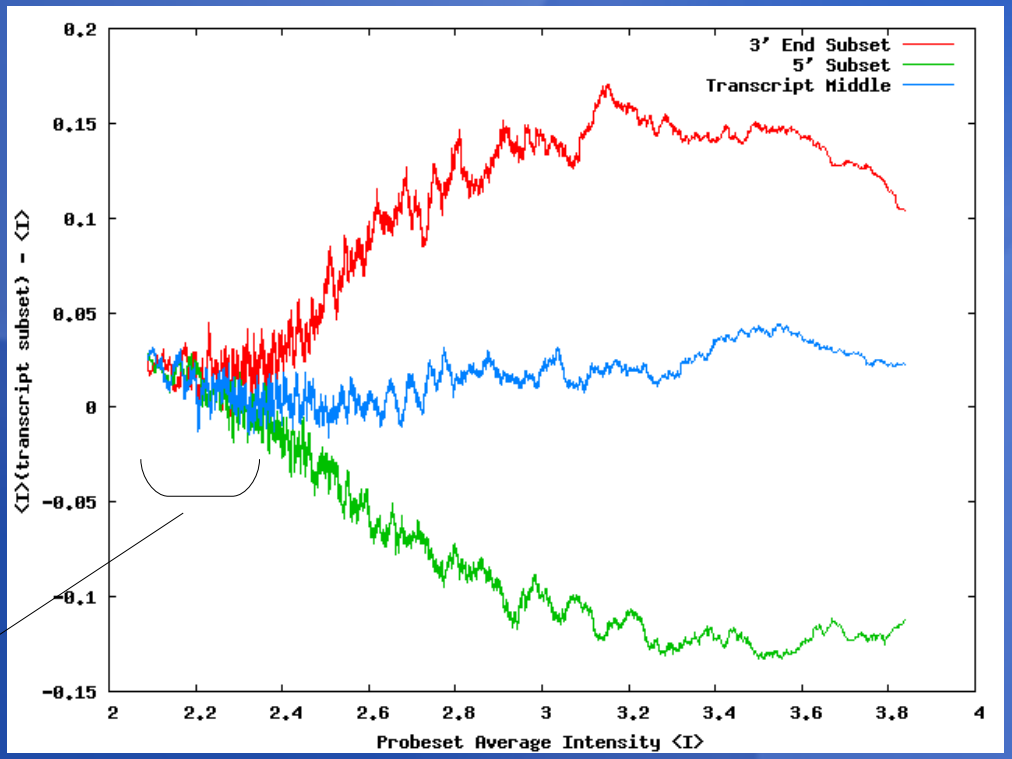

3’ End Subset
5’ Subset
Transcript Middle
0.2
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
2
2.2
2.4
2.6
2.8
3
3.2
3.4
3.6
3.8
4
<I>(transcript subset) - <I>
Probeset Average Intensity <I>

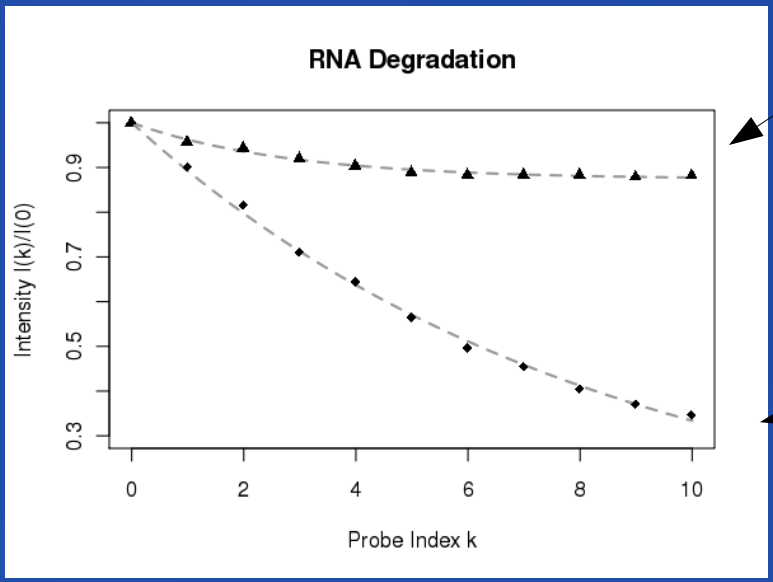
„Tongs”-Plot

RNA Degradation Depends on Probe Intensity

$$\langle I \rangle(k) = \langle I \rangle(0) * D(k)$$

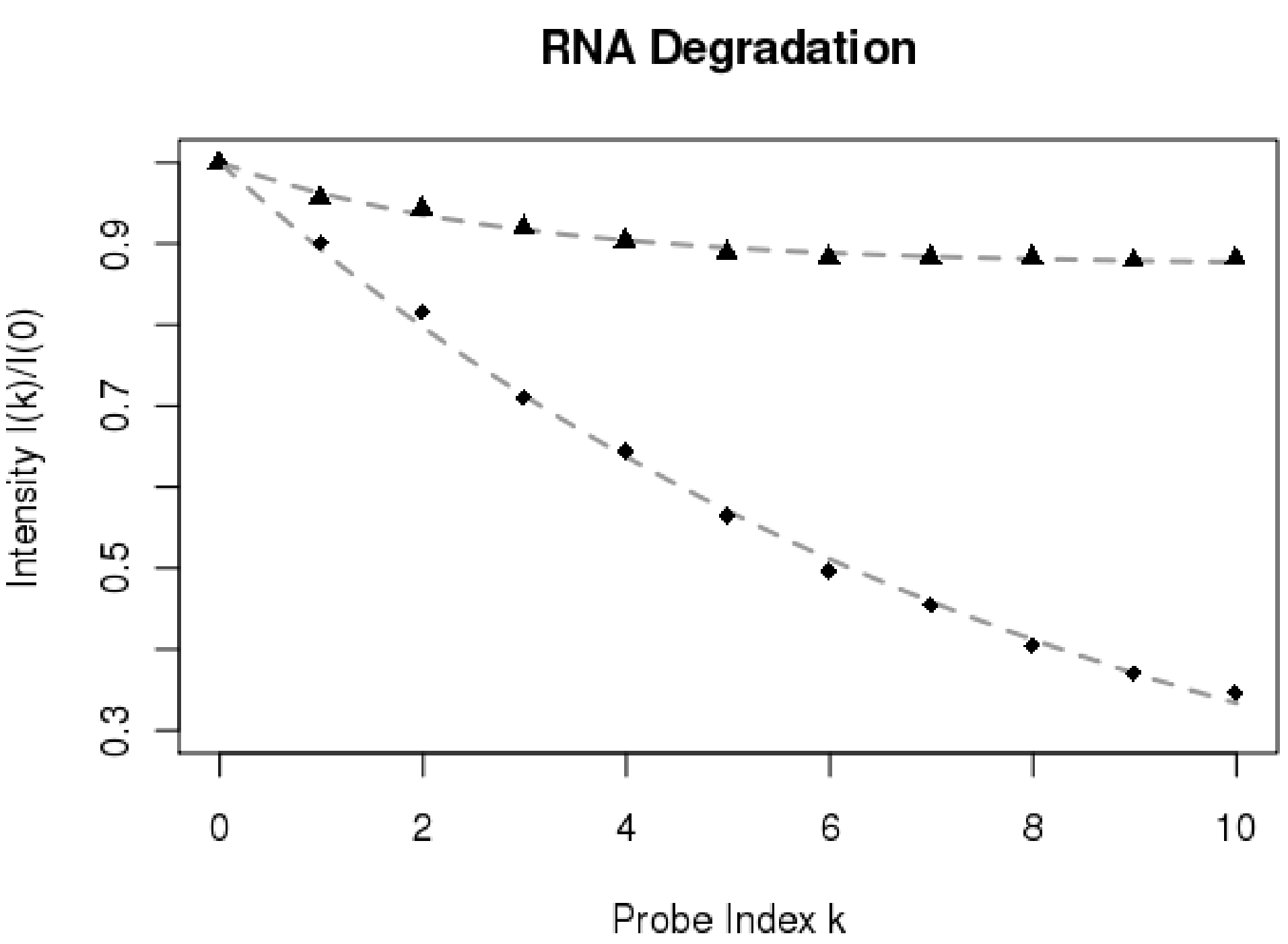
$$D(k) = (1 - d_1)e^{-d_2 k} + d_1$$

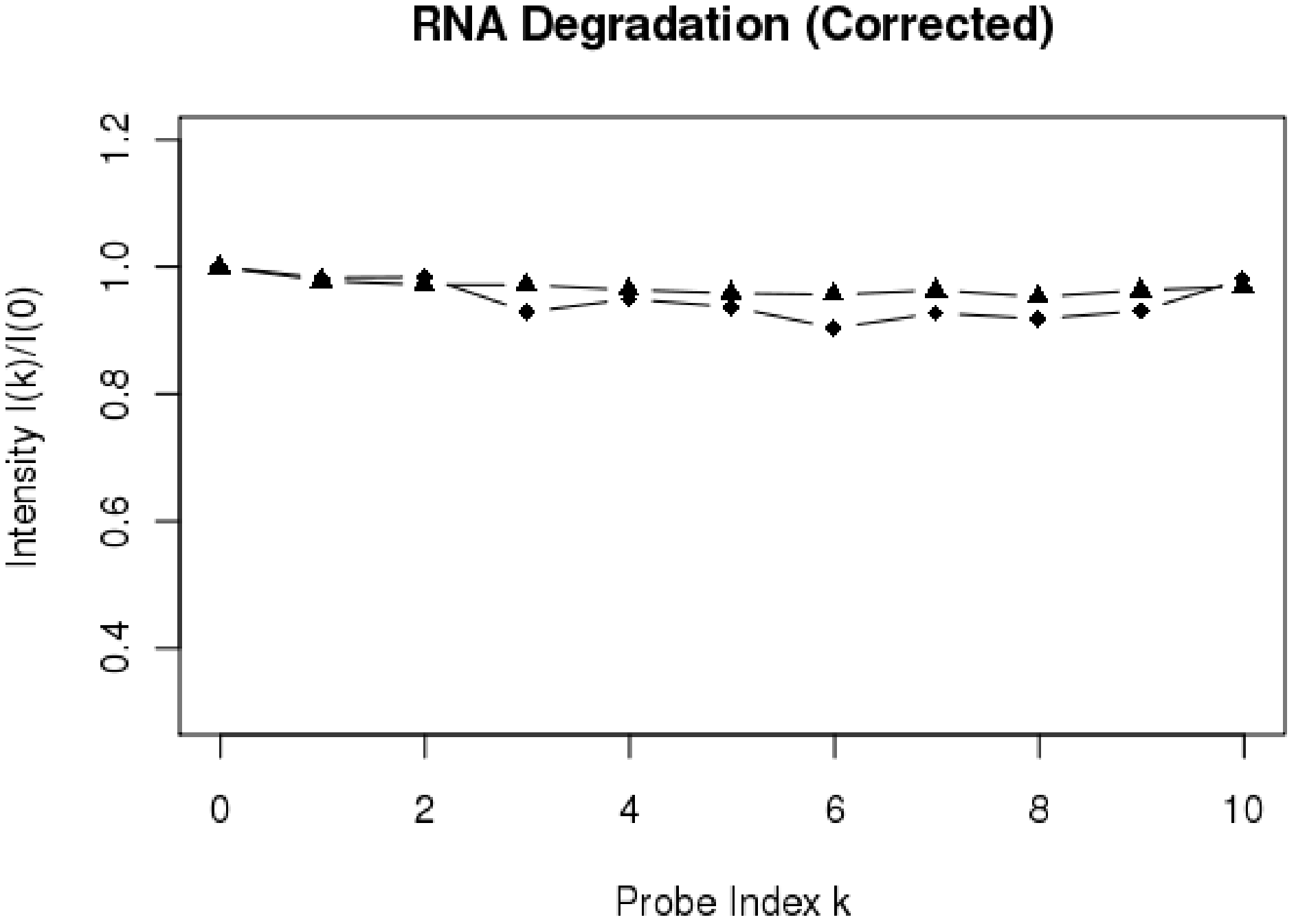


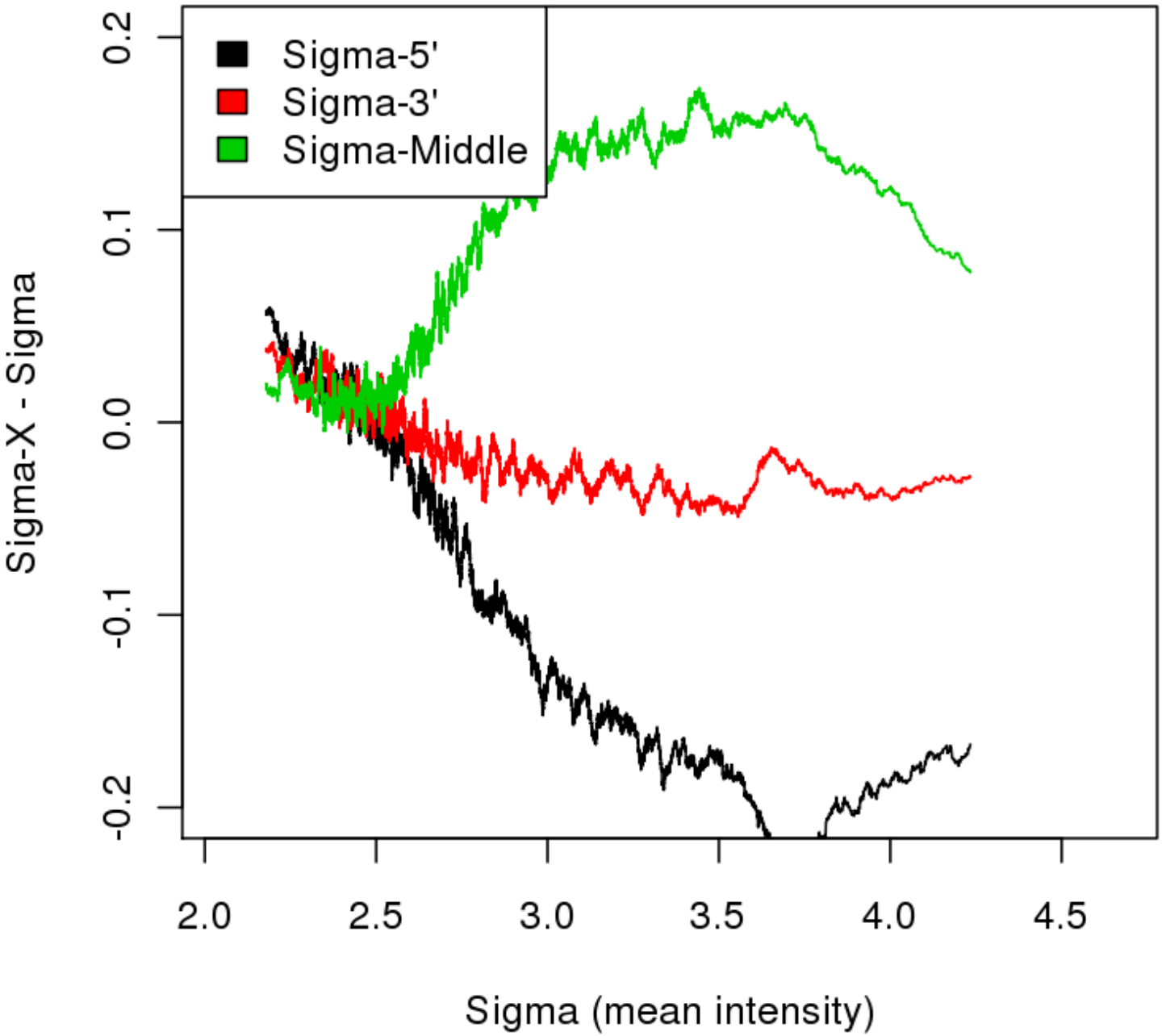


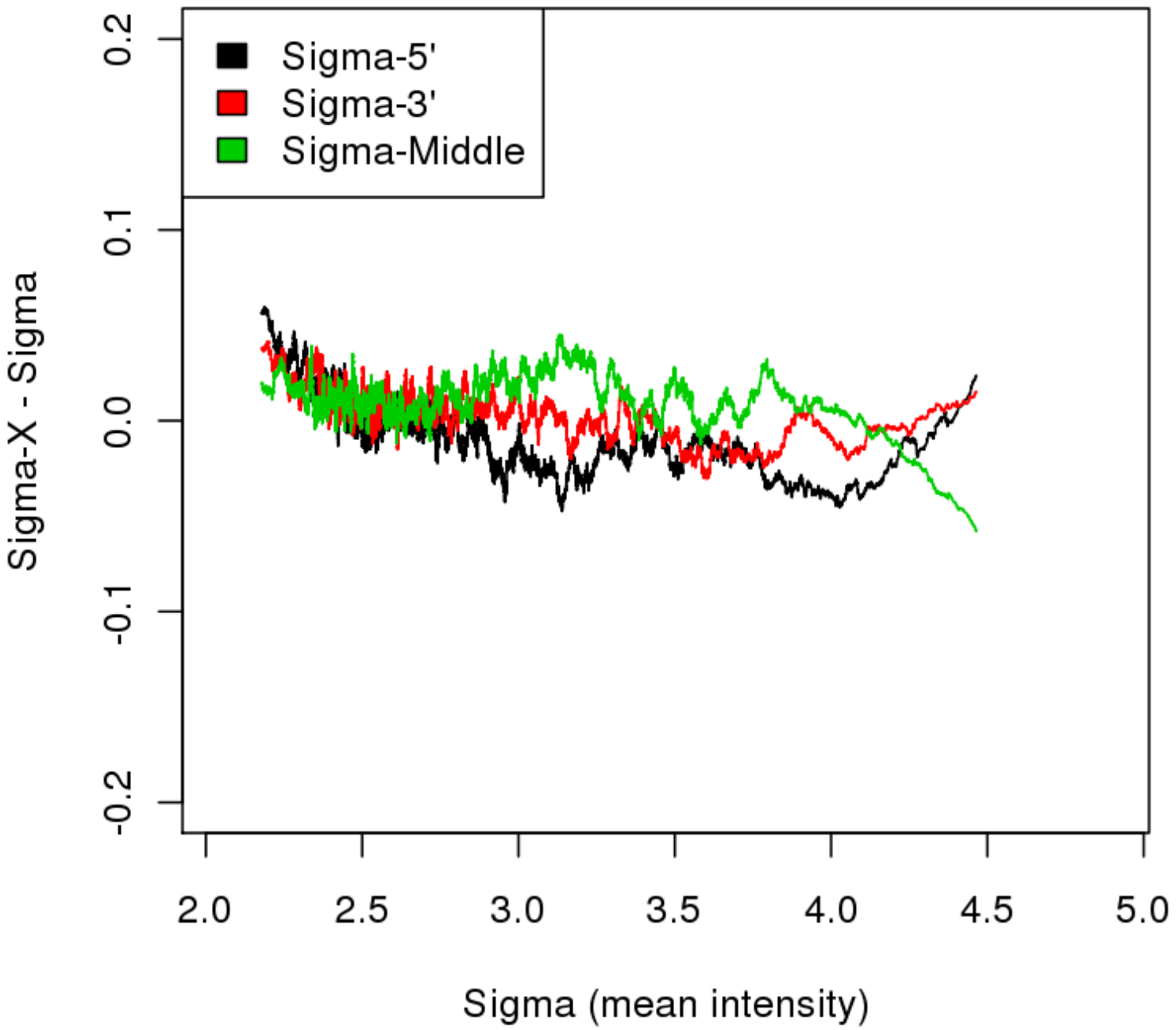
Non-specific binding

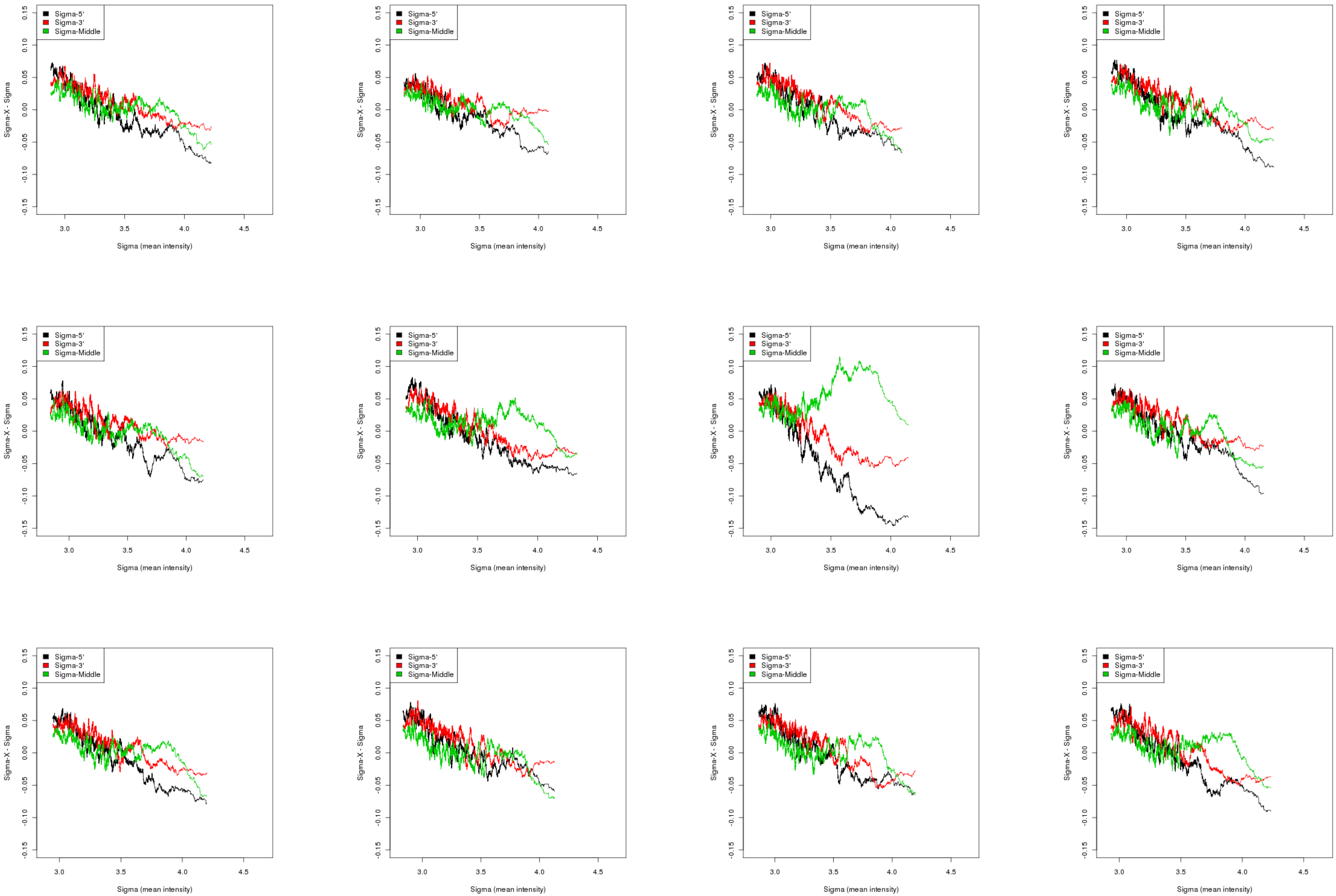
Specific binding


RNA Degradation
Intensity I(k)/I(0)
0.3
0.5
0.7
0.9
0
2
4
6
8
10
Probe Index k

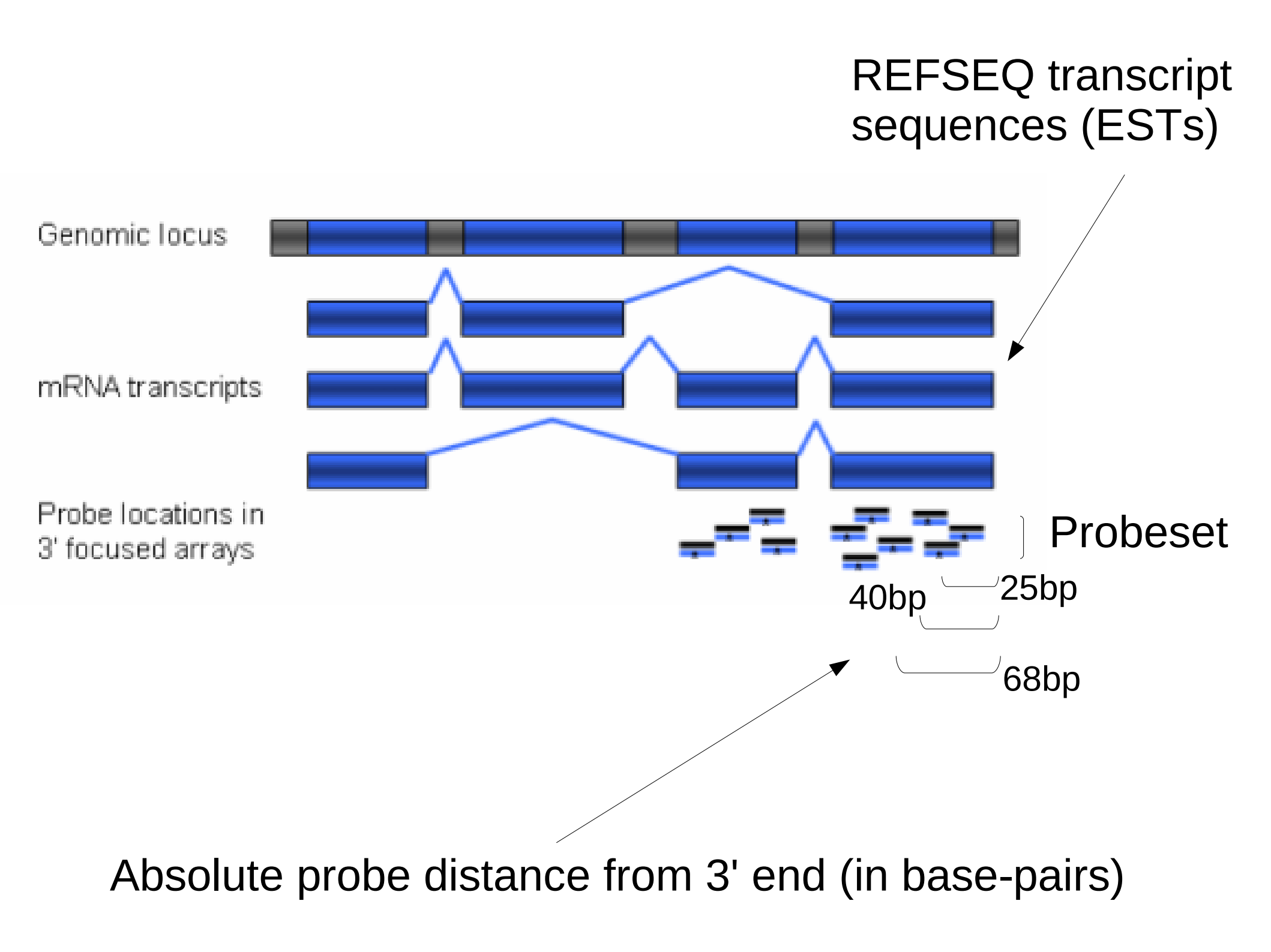

RNA Degradation (Corrected)
Intensity I(k)/I(0)
0.4
0.6
0.8
1.0
1.2
0
2
4
6
8
10
Probe Index k

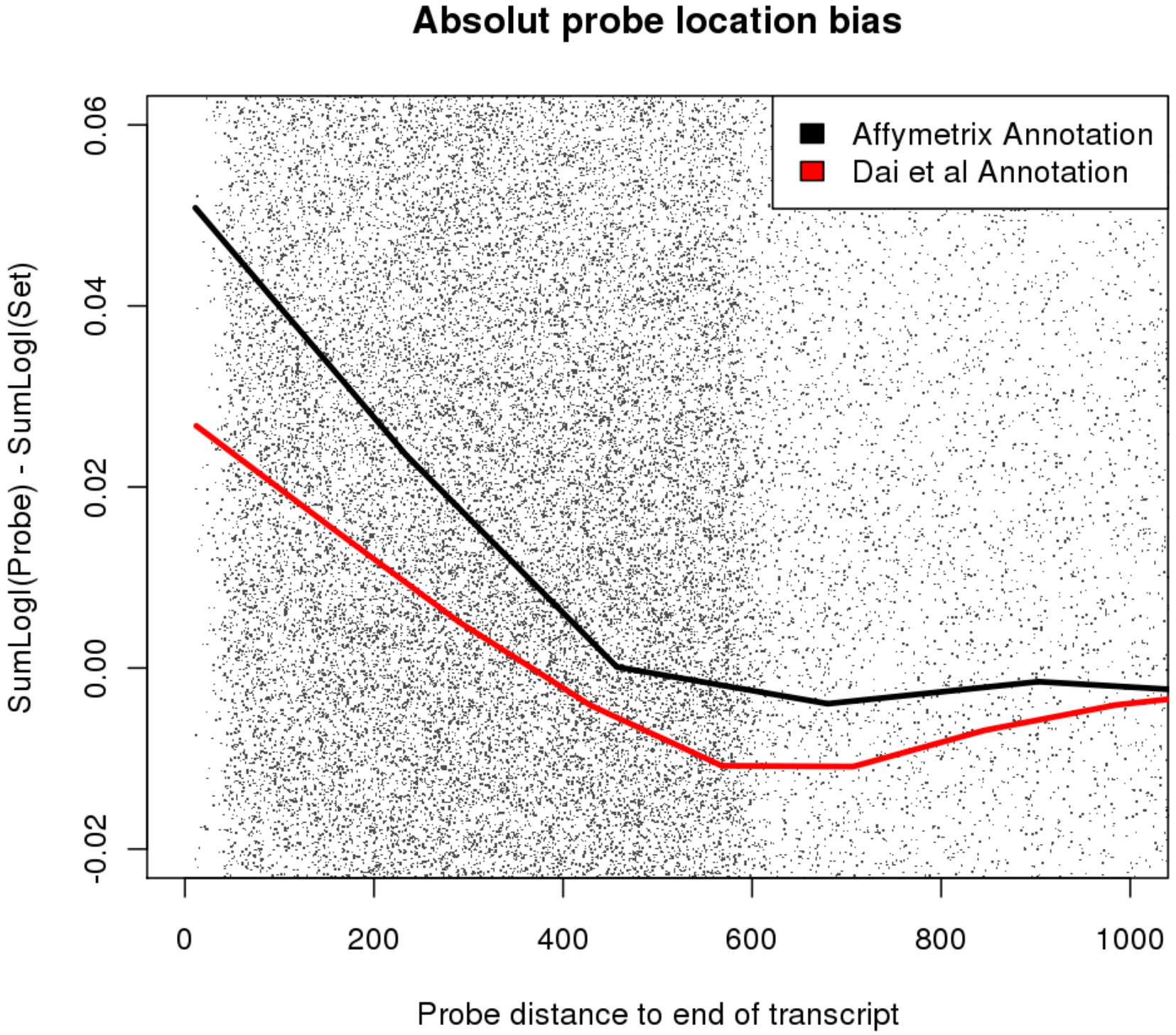

Sigma-5'
Sigma-3'
Sigma-Middle
Sigma-X - Sigma
0.2
0.1
0.0
-0.1
-0.2
2.0
2.5
3.0
3.5
4.0
4.5
Sigma (mean intensity)


Sigma-5'
Sigma-3'
Sigma-Middle
Sigma-X - Sigma
0.2
0.1
0.0
-0.1
-0.2
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Sigma (mean intensity)



Sigma-5'
Sigma-3'
Sigma-Middle
Sigma-X - Sigma
Sigma (mean intensity)
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
3.0
3.5
4.0
4.5


REFSEQ transcript sequences (ESTs)
Genomic locus
mRNA transcripts
Probe locations in 3' focused arrays
Probeset
40bp
25bp
68bp
Absolute probe distance from 3' end (in base-pairs)


Absolut probe location bias
Affymetrix Annotation
Dai et al Annotation
SumLogI(Probe) - SumLogI(Set)
0.06
0.04
0.02
0.00
-0.02
0
200
400
600
800
1000
Probe distance to end of transcript

„All models are wrong, but some are useful.”

(George Box)