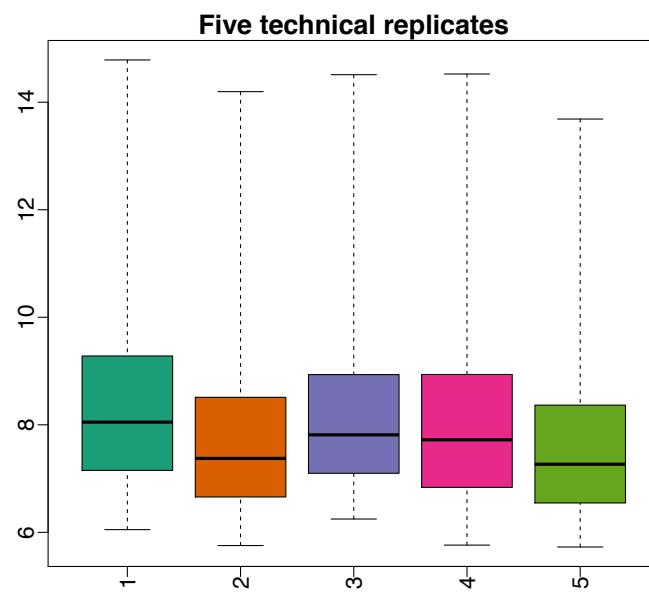


# Statistical Modeling 3

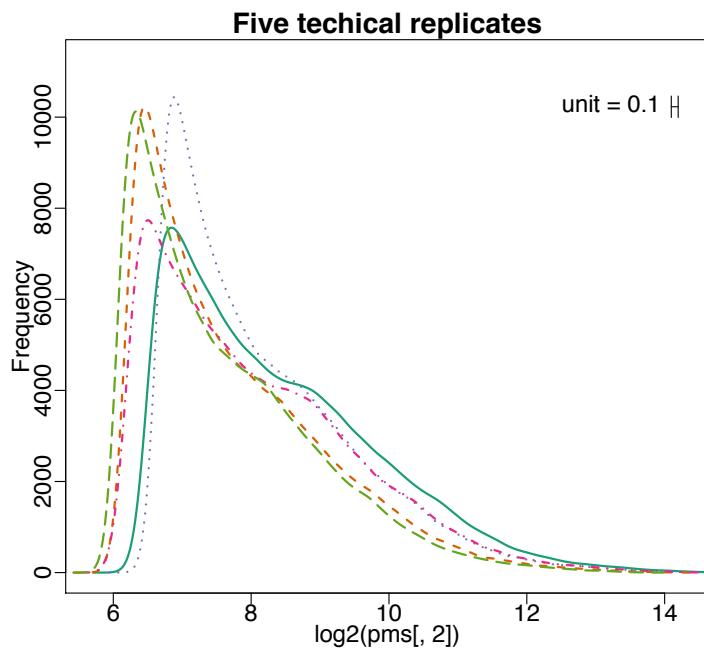
Bias correction and normalization

Ingo Ruczinski | Asian Institute in Statistical Genetics and Genomics | July 21-22, 2017



Ingo Ruczinski | Asian Institute in Statistical Genetics and Genomics | July 21-22, 2017

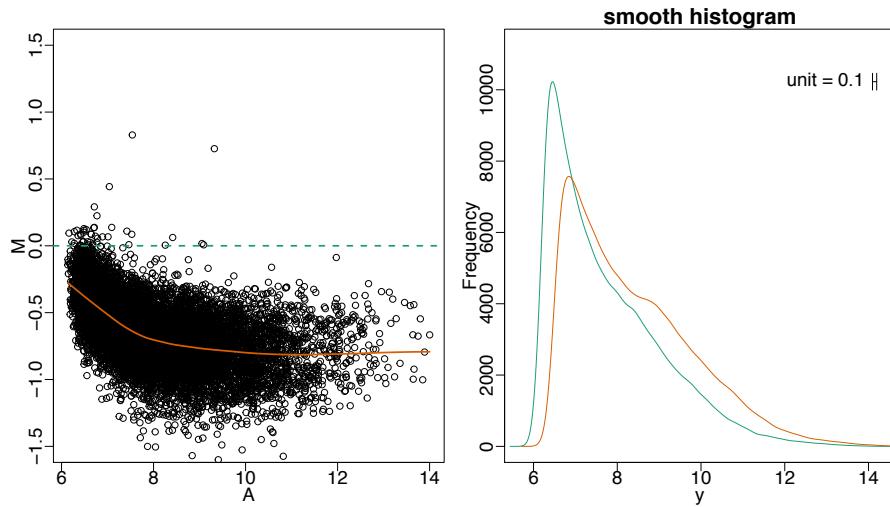
[ RI ]



Ingo Ruczinski | Asian Institute in Statistical Genetics and Genomics | July 21-22, 2017

[ RI ]

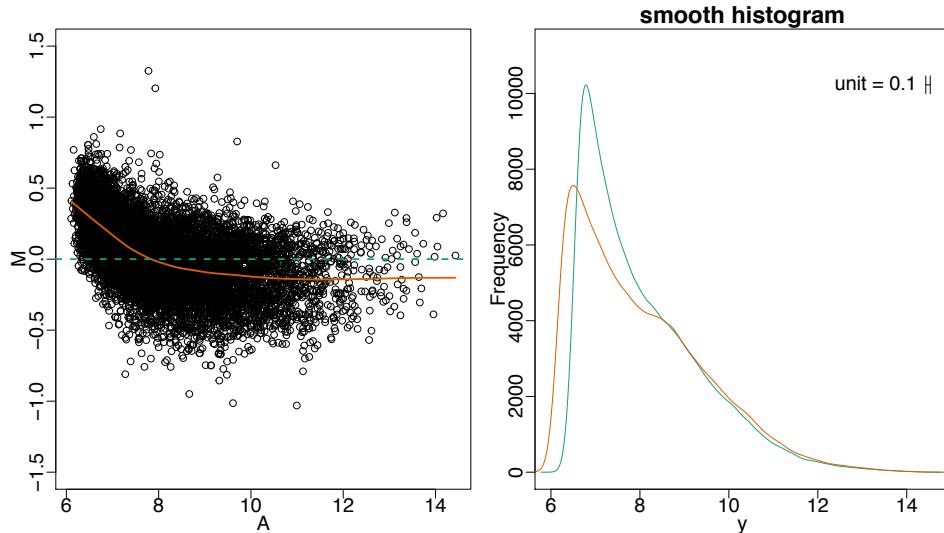
## More than location and scale changes!



Ingo Ruczinski | Asian Institute in Statistical Genetics and Genomics | July 21-22, 2017

[ RI ]

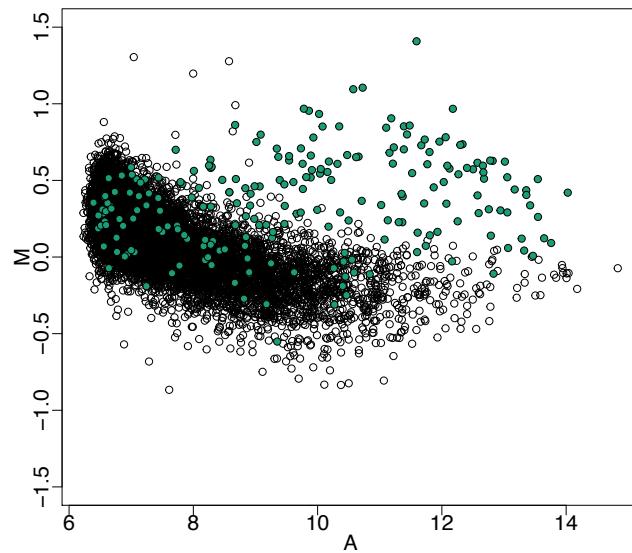
## Median shifts do not solve the problem!



Ingo Ruczinski | Asian Institute in Statistical Genetics and Genomics | July 21-22, 2017

[ RI ]

## There are non-linear effects!



Ingo Ruczinski | Asian Institute in Statistical Genetics and Genomics | July 21-22, 2017

[ RI ]

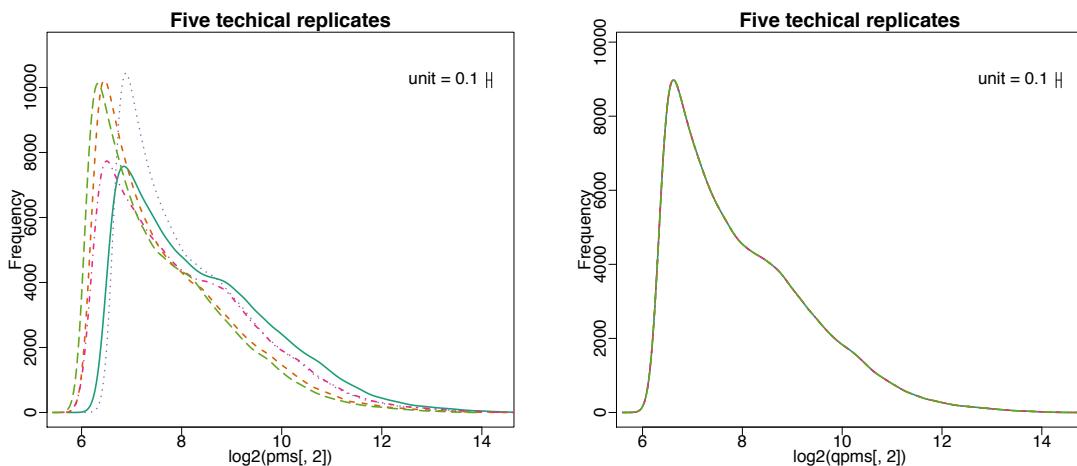
## Quantile normalization

Original				Order				Averaged				Re-order			
2	4	4	5	2	4	3	5	3.5	3.5	3.5	3.5	3.5	3.5	5.0	5.0
5	14	4	7	3	8	4	5	5.0	5.0	5.0	5.0	8.5	8.5	5.5	5.5
4	8	6	9	3	8	4	7	5.5	5.5	5.5	5.5	6.5	5.0	8.5	8.5
3	8	5	8	4	9	5	8	6.5	6.5	6.5	6.5	5.0	5.5	6.5	6.5
3	9	3	5	5	14	6	9	8.5	8.5	8.5	8.5	5.5	6.5	3.5	3.5

Ingo Ruczinski | Asian Institute in Statistical Genetics and Genomics | July 21-22, 2017

[ RI ]

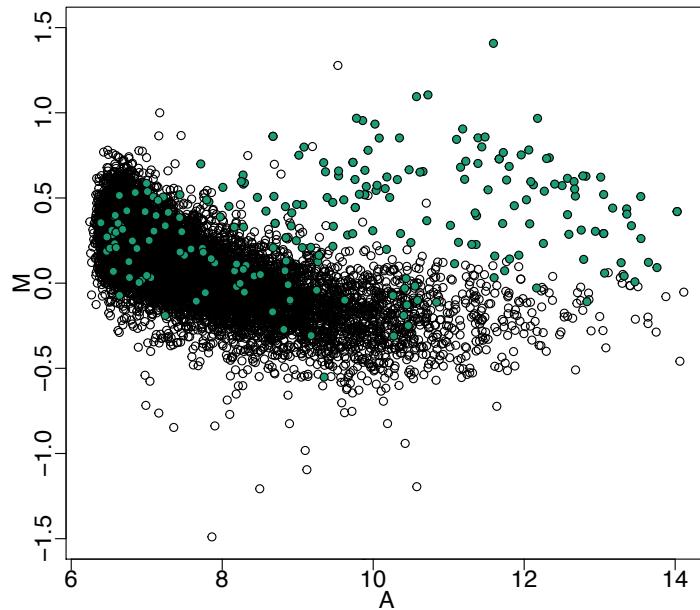
## Densities are forced to be identical



Ingo Ruczinski | Asian Institute in Statistical Genetics and Genomics | July 21-22, 2017

[ RI ]

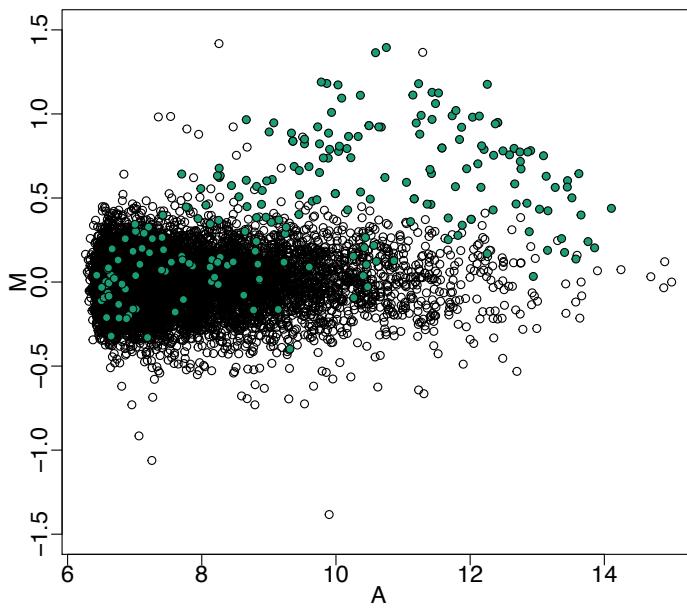
## Differential expression can be preserved



Ingo Ruczinski | Asian Institute in Statistical Genetics and Genomics | July 21-22, 2017

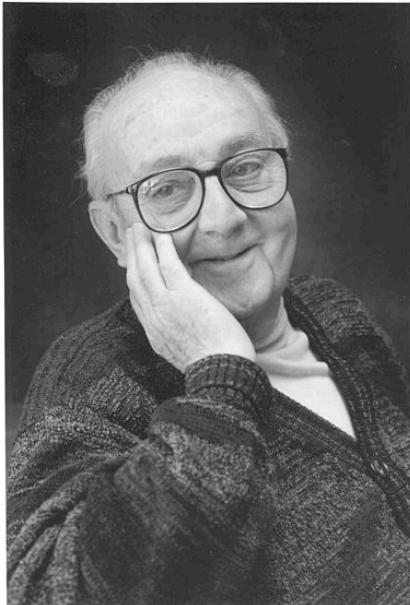
[ RI ]

## Differential expression can be preserved



Ingo Ruczinski | Asian Institute in Statistical Genetics and Genomics | July 21-22, 2017

[ RI ]



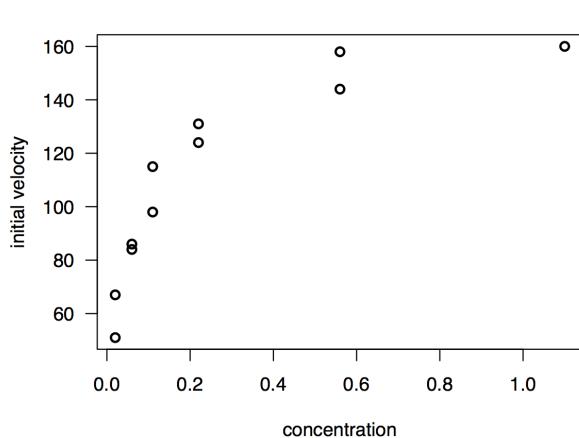
“Essentially, all models are wrong,  
but some are useful”

*George E.P. Box*

Ingo Ruczinski | Asian Institute in Statistical Genetics and Genomics | July 21-22, 2017

[ RI ]

## A biochemical experiment



Michaelis-Menten equation

$$V = \frac{V_{\max} \times C}{K + C}$$

$V$  = initial velocity

$C$  = concentration

$V_{\max}$  = maximum velocity

$K$  = rate constant

Ingo Ruczinski | Asian Institute in Statistical Genetics and Genomics | July 21-22, 2017

[ 140.615 ]

## A biochemical experiment

$$V = \frac{V_{\max} \times C}{K + C}$$

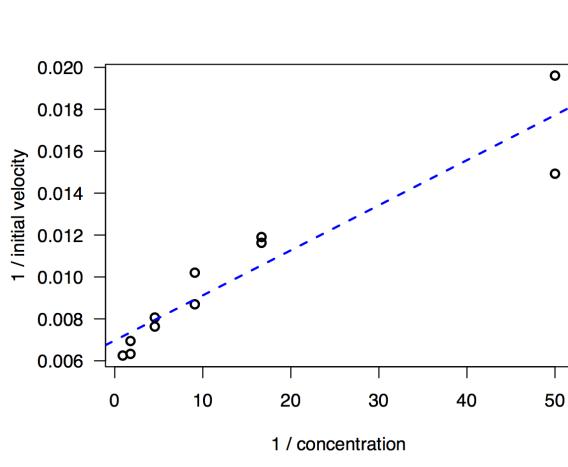
$$\begin{aligned}\Rightarrow \frac{1}{V} &= \frac{K + C}{V_{\max} \times C} \\ &= \frac{K}{V_{\max} \times C} + \frac{1}{V_{\max}}\end{aligned}$$

$$\Rightarrow \frac{1}{V} = \left( \frac{1}{V_{\max}} \right) + \left( \frac{K}{V_{\max}} \right) \times \left( \frac{1}{C} \right)$$

Ingo Ruczinski | Asian Institute in Statistical Genetics and Genomics | July 21-22, 2017

[ 140.615 ]

## A biochemical experiment



Model:

$$\frac{1}{V} = \beta_0 + \beta_1 \left( \frac{1}{C} \right) + \text{error}$$

$$\begin{aligned}\text{Intercept} &= 0.00697 \\ \text{Slope} &= 0.00022\end{aligned}$$

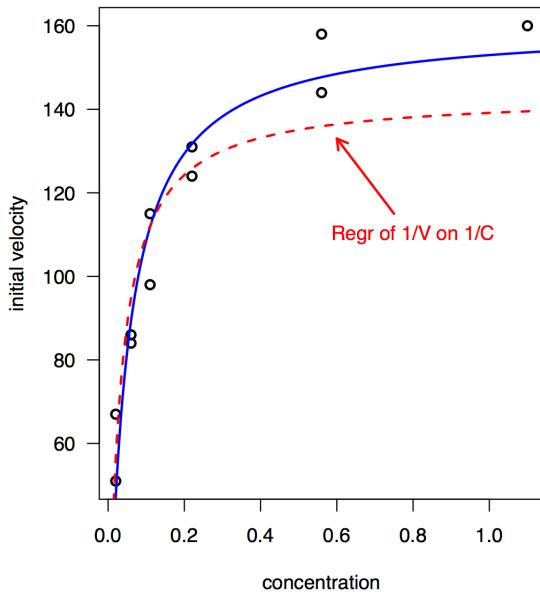
$$\hat{V}_{\max} = 1 / \text{Intercept} = 1 / 0.00697 = 143$$

$$\hat{K} = \text{Slope} \times \hat{V}_{\max} = 0.031$$

Ingo Ruczinski | Asian Institute in Statistical Genetics and Genomics | July 21-22, 2017

[ 140.615 ]

## A biochemical experiment



Which is more reasonable?

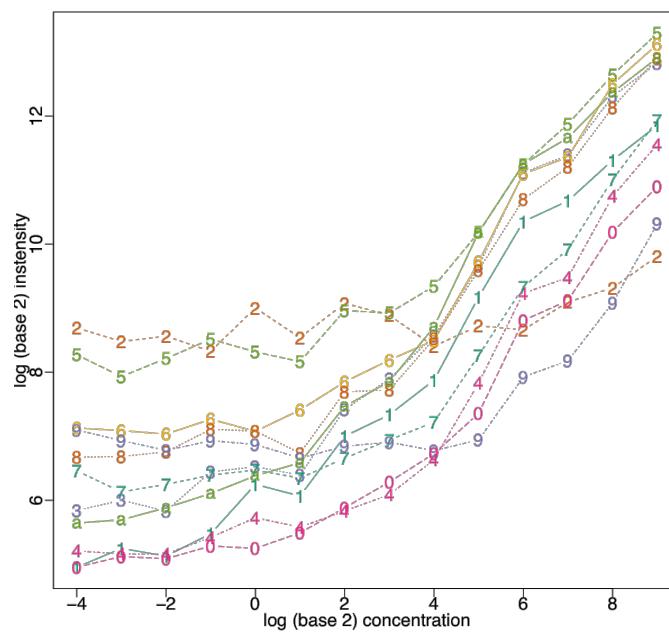
$$\frac{1}{V} = \beta_0 + \beta_1 \left( \frac{1}{C} \right) + \text{error}$$

$$V = \frac{V_{\max} \times C}{K + C} + \text{error}$$

Ingo Ruczinski | Asian Institute in Statistical Genetics and Genomics | July 21-22, 2017

[ 140.615 ]

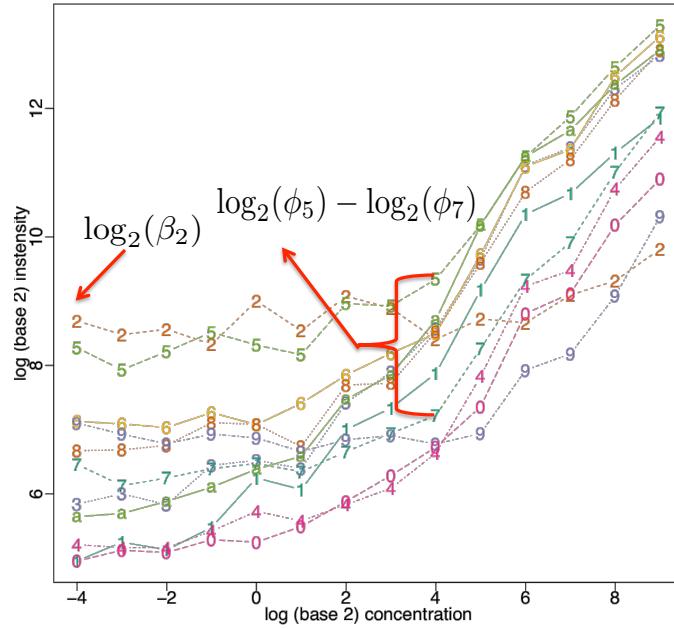
## Eleven probes from one spiked-in gene



Ingo Ruczinski | Asian Institute in Statistical Genetics and Genomics | July 21-22, 2017

[ RI ]

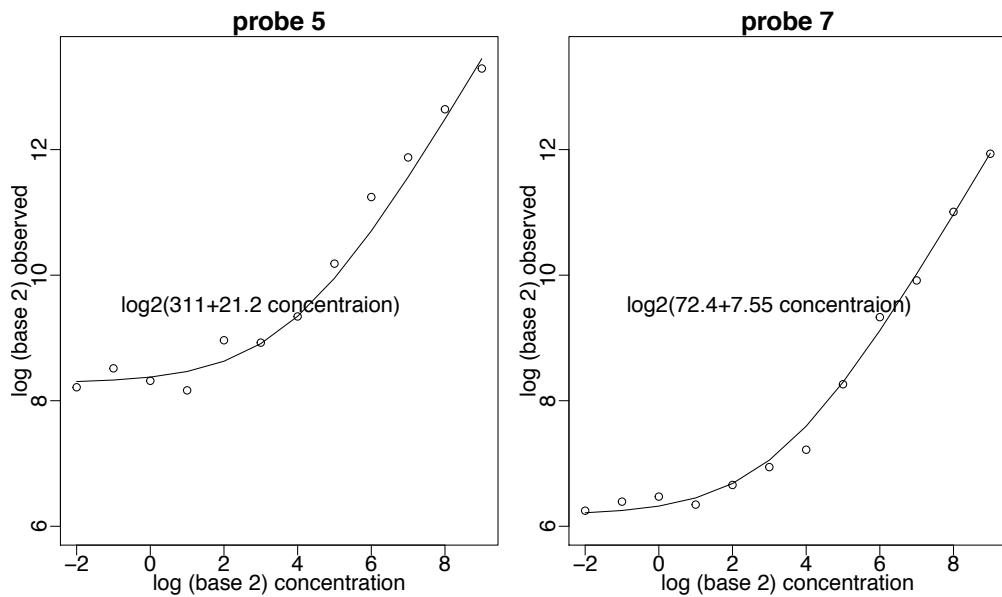
$$Y_{ij} = \beta_j + \theta_i \phi_j + \varepsilon_{ij} \quad \text{var}(\varepsilon_{ij}) \propto \theta_i \phi_j$$



Ingo Ruczinski | Asian Institute in Statistical Genetics and Genomics | July 21-22, 2017

[ RI ]

## Model fit to two probes



Ingo Ruczinski | Asian Institute in Statistical Genetics and Genomics | July 21-22, 2017

[ RI ]