Comparison of Normalization Methods for cDNA Microarrays

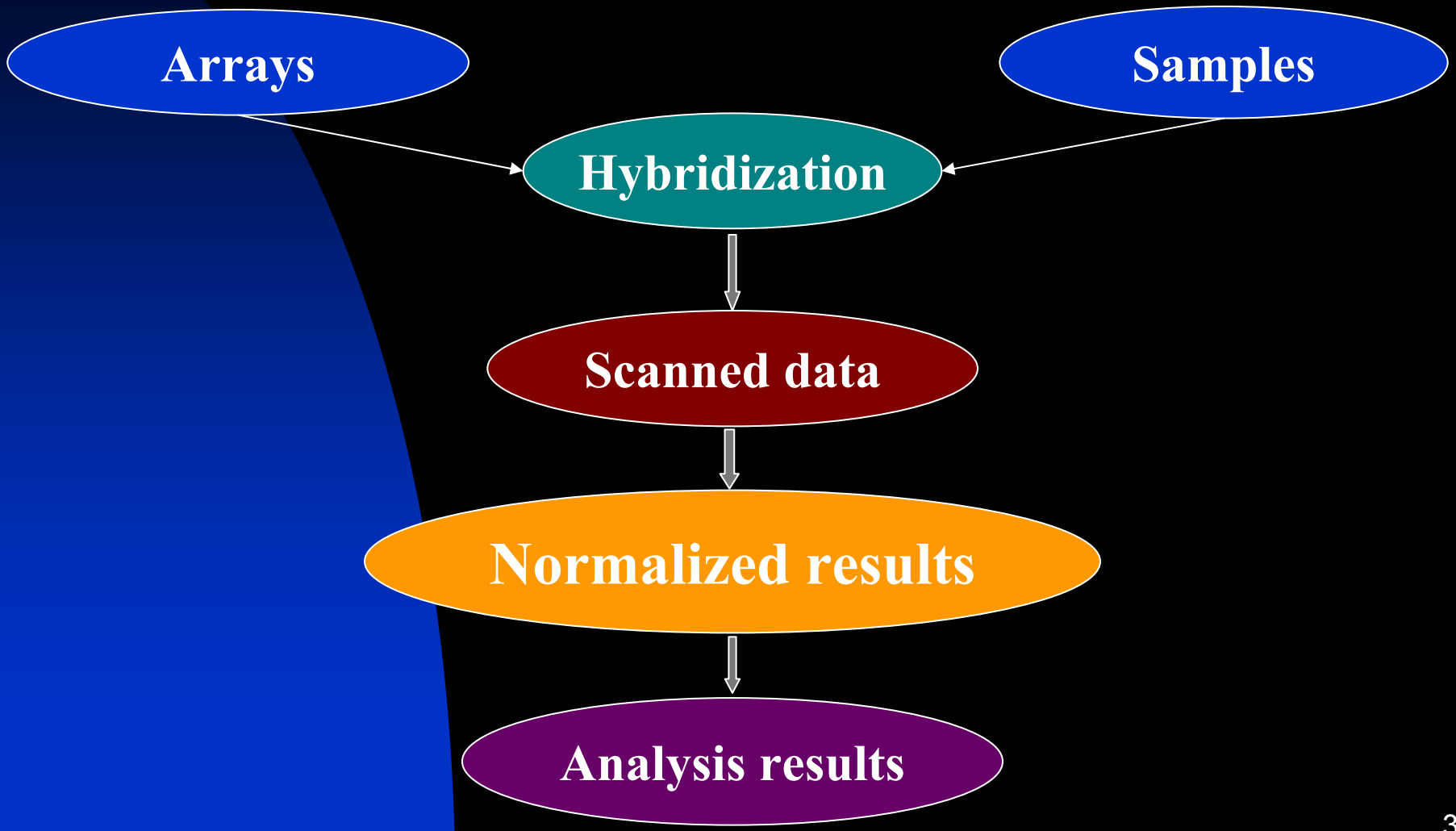
Liling Warren, Ben Hui Liu

Bioinformatics Program, NCSU, Raleigh, NC
Bio-informatics Group, Inc. Cary, NC

Topics of Discussion

- Data flow in a microarray experiment
- Describe different normalization methods
- Evaluate different normalization methods
- To normalize or not to normalize
- Data quality
- Experimental design
- Conclusions

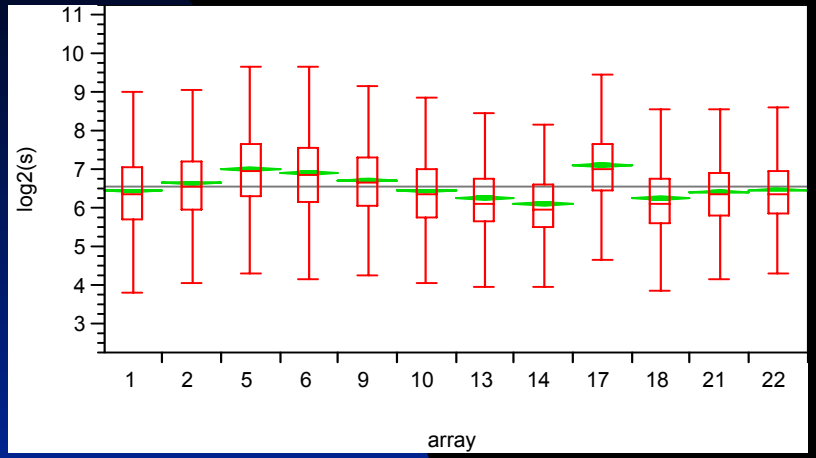
Data flow in a microarray experiment

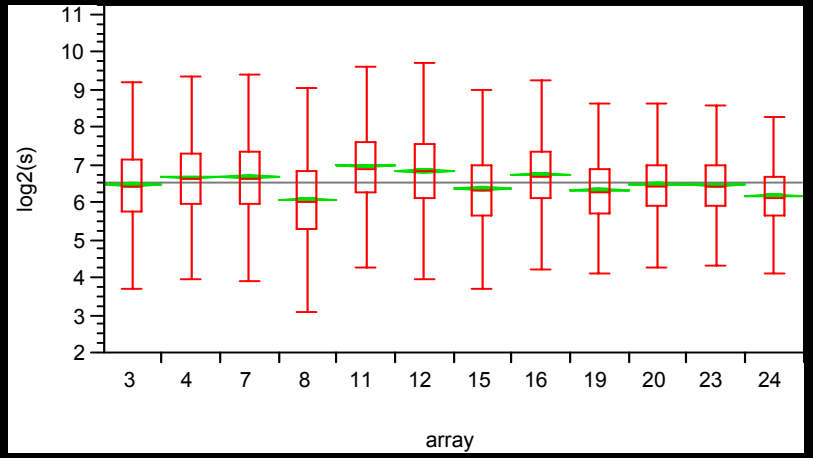


Purpose of Data Normalization

- To remove **systematic** errors introduced at various stages of a microarray experiment.
- Systematic effects include:
  - Array effect
  - Pin/block effect
  - Dye effect (Cy3/Cy5)
  - mRNA extraction effect
  - Dye labeling effect

Systematic Errors – Array Effect





Analysis of Variance

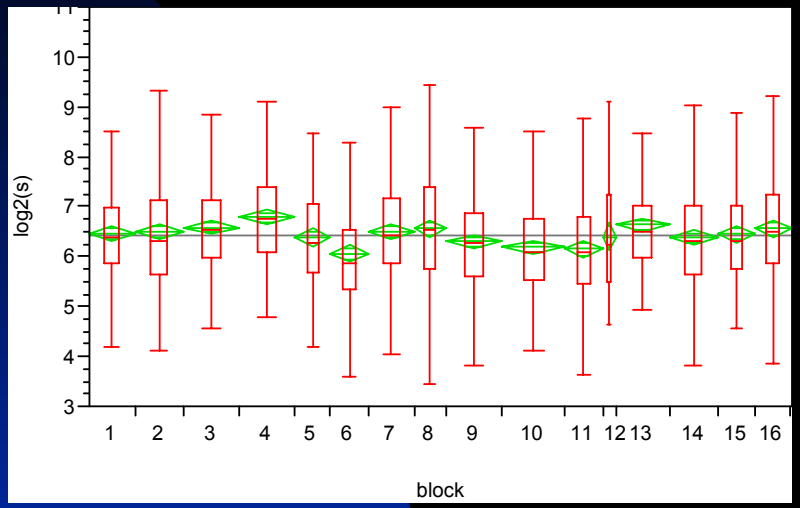
Source	DF	Sum of Square	Mean Squar	F Ratio	Prob > F
array	11	2977.951	270.723	288.6899	0.0000
Error	32757	30718.314	0.938		
C. Total	32768	33696.265			

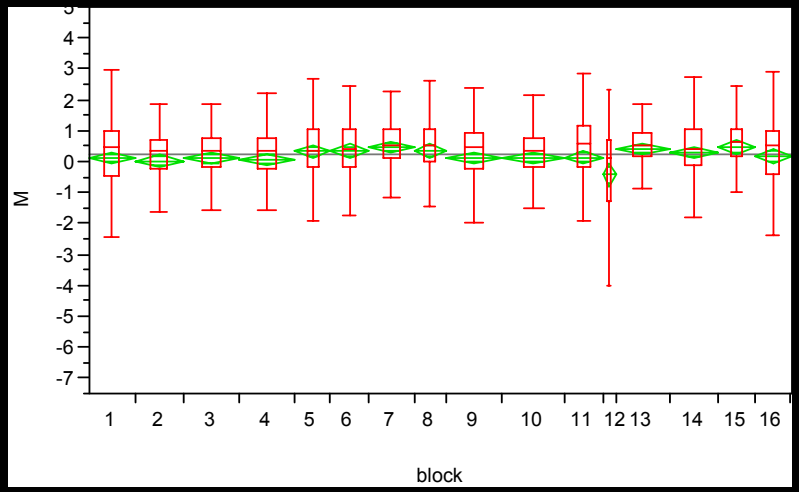
Analysis of Variance

Source	DF	Sum of Square	Mean Squar	F Ratio	Prob > F
array	11	2067.326	187.939	181.4893	0.0000
Error	32754	33917.943	1.036		
C. Total	32765	35985.269			

Box plots and ANOVA tests show that between array variation is highly significant with a P value < 0.0001 using either log ratios or log signal intensity

Systematic Errors – Block Effect





Analysis of Variance

Source	DF	Sum of Square	Mean Squa	F Rati	Prob >
block	15	98.3035	6.55357	6.1835	<.000
Error	2715	2877.4692	1.05984		
C. Total	2730	2975.7727			

Analysis of Variance

Source	DF	Sum of Square	Mean Squa	F Rati	Prob >
block	15	77.2856	5.15237	2.8905	0.0002
Error	2715	4839.5028	1.78251		
C. Total	2730	4916.7885			

Box plots and ANOVA tests show that between block variation is highly significant with a P value < 0.0001 using either log ratios or log signal intensity

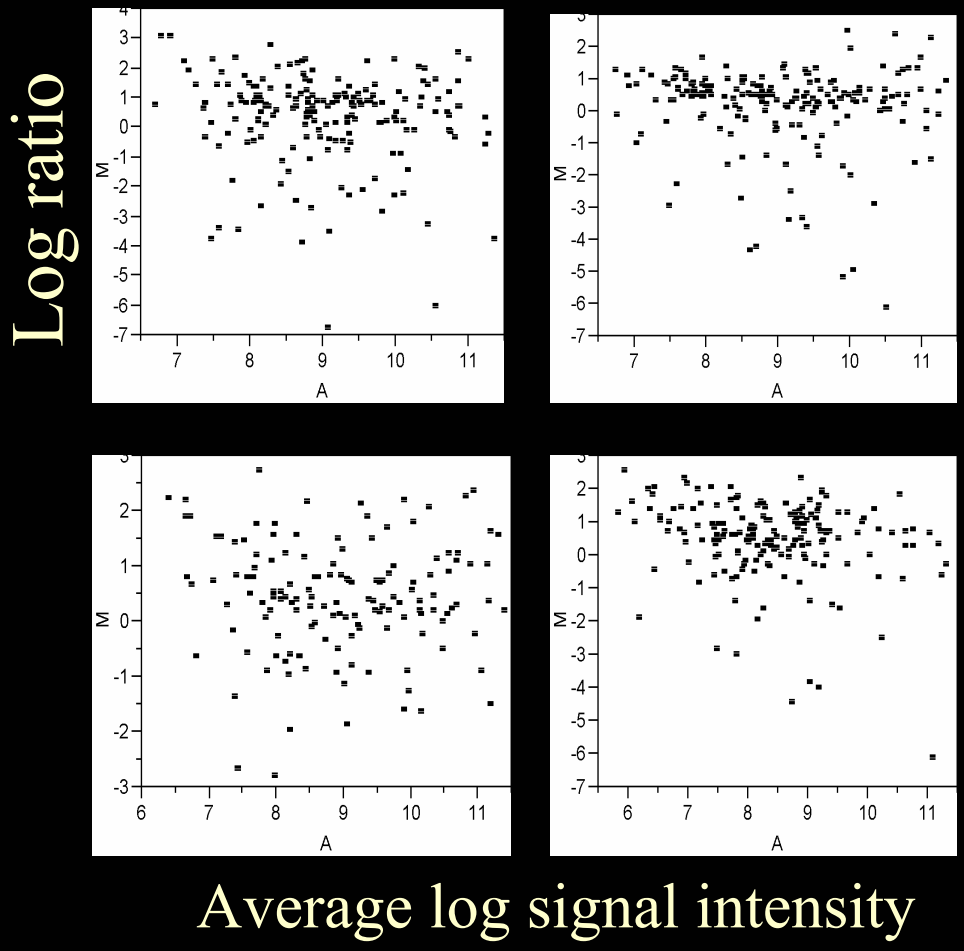
Systematic Errors – Dye Effect

- M vs. A plots for block 1, 2, 5,6 in array 1 of Kidney data

$$M = \log_2(R) - \log_2(G)$$

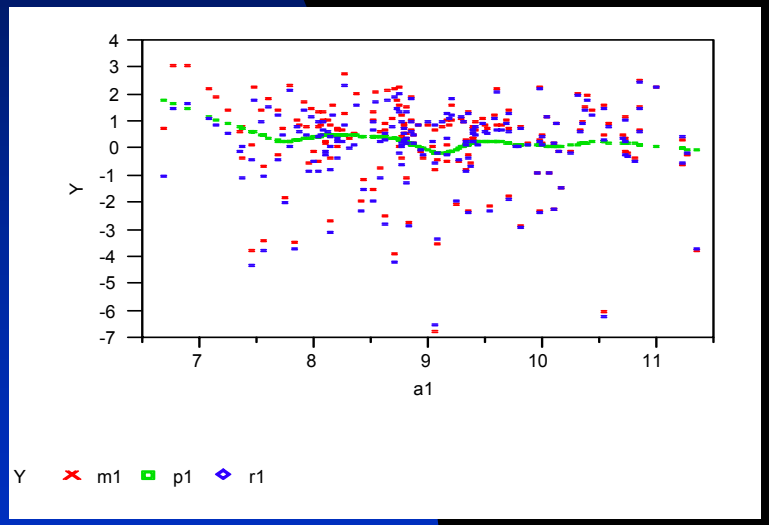
$$A = 1/2(\log_2(R) + \log_2(G))$$

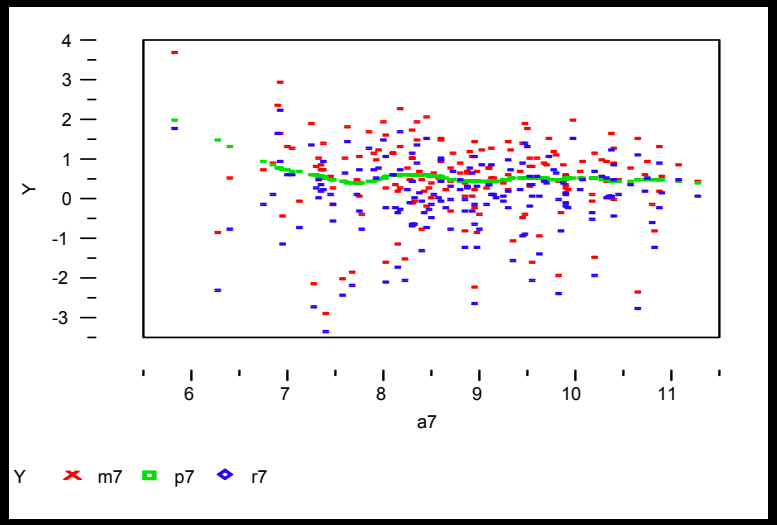
- M vs. A plots reveal the dependency of log ratios on average signal intensity



Comparing Normalization Methods

- Method #1: log ratio based, local smoothing method using loess function





Red: M values; Green:Predicted values
Blue:Residual values

Comparing Normalization Methods

- Method #2: log ratio based, block-specific global normalization

$$\tilde{y}_{ijk} = (y_{ijk} - \bar{y}_{ij}) / s_{ij}$$

where i=1,…,24; j=1, …, 16; k=1, .., n_{ij}, and

$\bar{y}_{ij}$: block-specific mean

s_{ij} : block-specific standard deviation

Comparing Normalization Methods

- Method #3: log ratio based, ANOVA normalization

$$y_{ijklm} = \mu + A_i + B_j + M_k + D_l + (AB)_{ij} + \varepsilon_{ijklm}$$

-- Random effects: A, B, AB

-- Fixed effects: M, D

-- Residuals are subsequently used as input for gene-based ANOVA model

Methods Omitting Normalization

- **Method #4: gene-based ANOVA, omitting normalization, using log ratios**

$$y_{ijk} = \mu + m_i + d_j + (md)_{ij} + \varepsilon_{ijk}$$

- **Method #5: gene-based Analysis of Covariance, omitting normalization, using log signal intensity**

$$y_{ijk} = \mu + m_i + d_j + (md)_{ij} + \beta(x_{ijk} - \bar{x}_{...}) + \varepsilon_{ijk}$$

y_{ijk} **: log signal intensity from test sample;**

x_{ijk} **: log signal intensity from reference sample**

“project normal” Data Analysis

- **Gene based ANOVA model:**

$$y_{ijk} = \mu + m_i + d_j + (md)_{ij} + \varepsilon_{ijk}$$, i=1 to 6, j=1, 2 and k=1 to 4.

Source	*df*	*MS*	*EMS*
Mouse	5	*MSM*	$\sigma^2 + 4\tau_m^2$
Dye	1	*MSD*	$\sigma^2 + 20\tau_d^2$
Mouse*Dye	5	*MS(MD)*	$\sigma^2 + 4\tau_{md}^2$
Error	12	*MSE*	σ^2

The null hypothesis of no mouse effect is tested with

$F_o = MSM / MSE$, with df1=5 and df2=12.

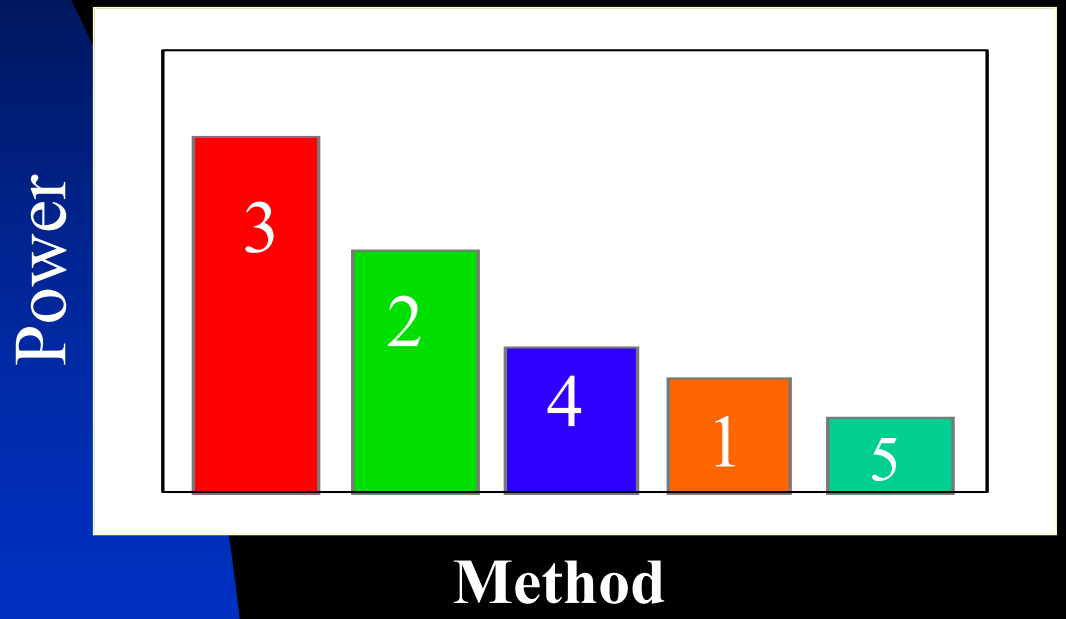
Comparison Results

	Method 1	Method 2	Method 3	Method 4	Method 5
Method 1	129	315	451	243	182
Method 2	89	275	522	362	318
Method 3	80	155	402	409	410
Method 4	51	78	158	165	174
Method 5	32	42	77	76	85

On the diagonal: numbers of genes detected by the specific method;
Upper triangle: detected by either of the two corresponding methods;
Lower triangle: detected by both methods (Kidney data).

Power Comparison

- Power rank:



- Pair-wise power comparison - McNemar's Test

McNemar's Test

First Method	Second Method		
	Reject	Accept	Total
Reject	n_{11}	n_{12}	$n_{1.}$
Accept	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	N

Under H_0: $\pi_{1+} = \pi_{+1}$

Test statistic: $\chi_1^2 = (n_{12} - n_{21})^2 / (n_{12} + n_{21})$

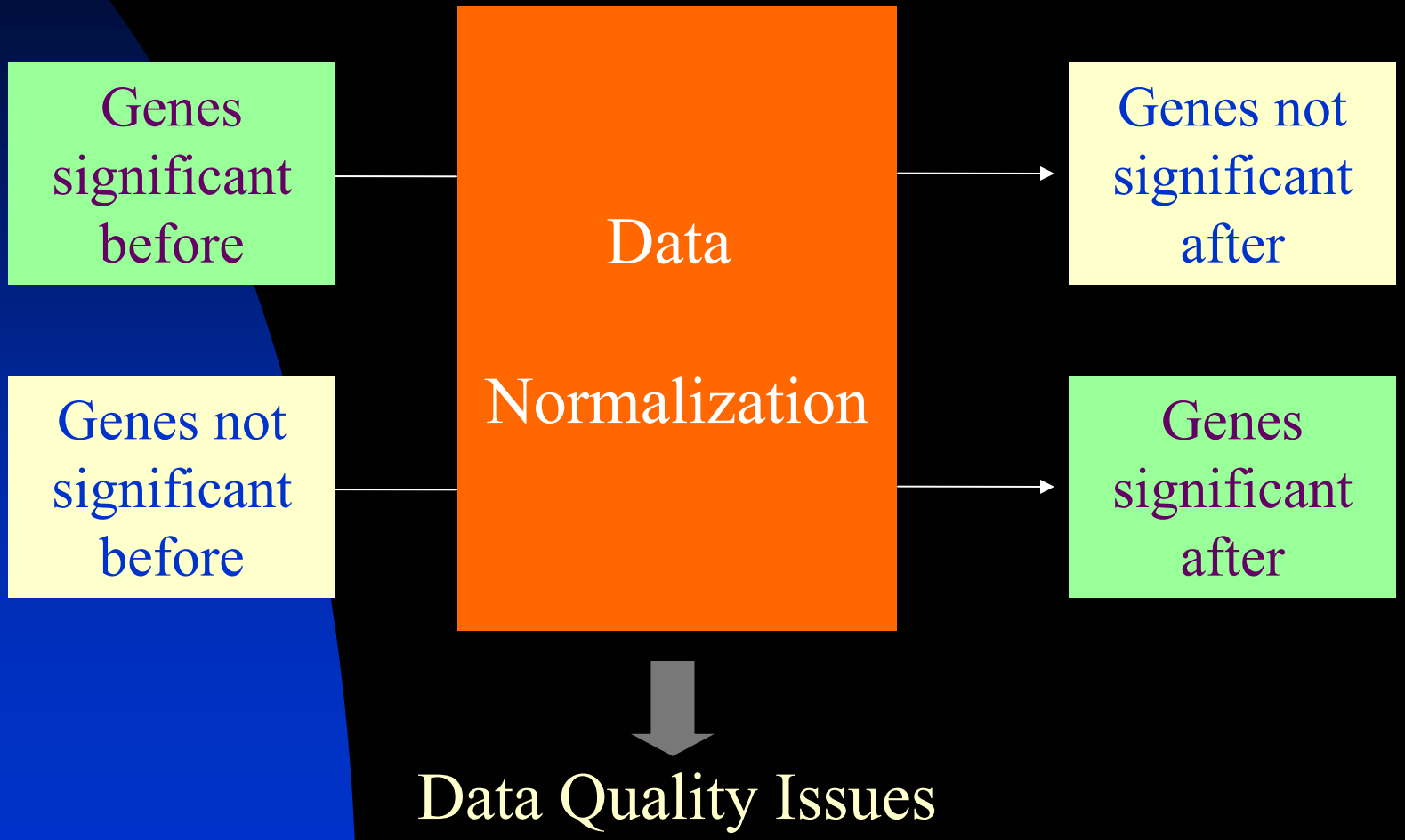
Reject H_0 if $\chi_1^2 > 3.84$ **at** $\alpha = 0.05$

McNemar's Test Results

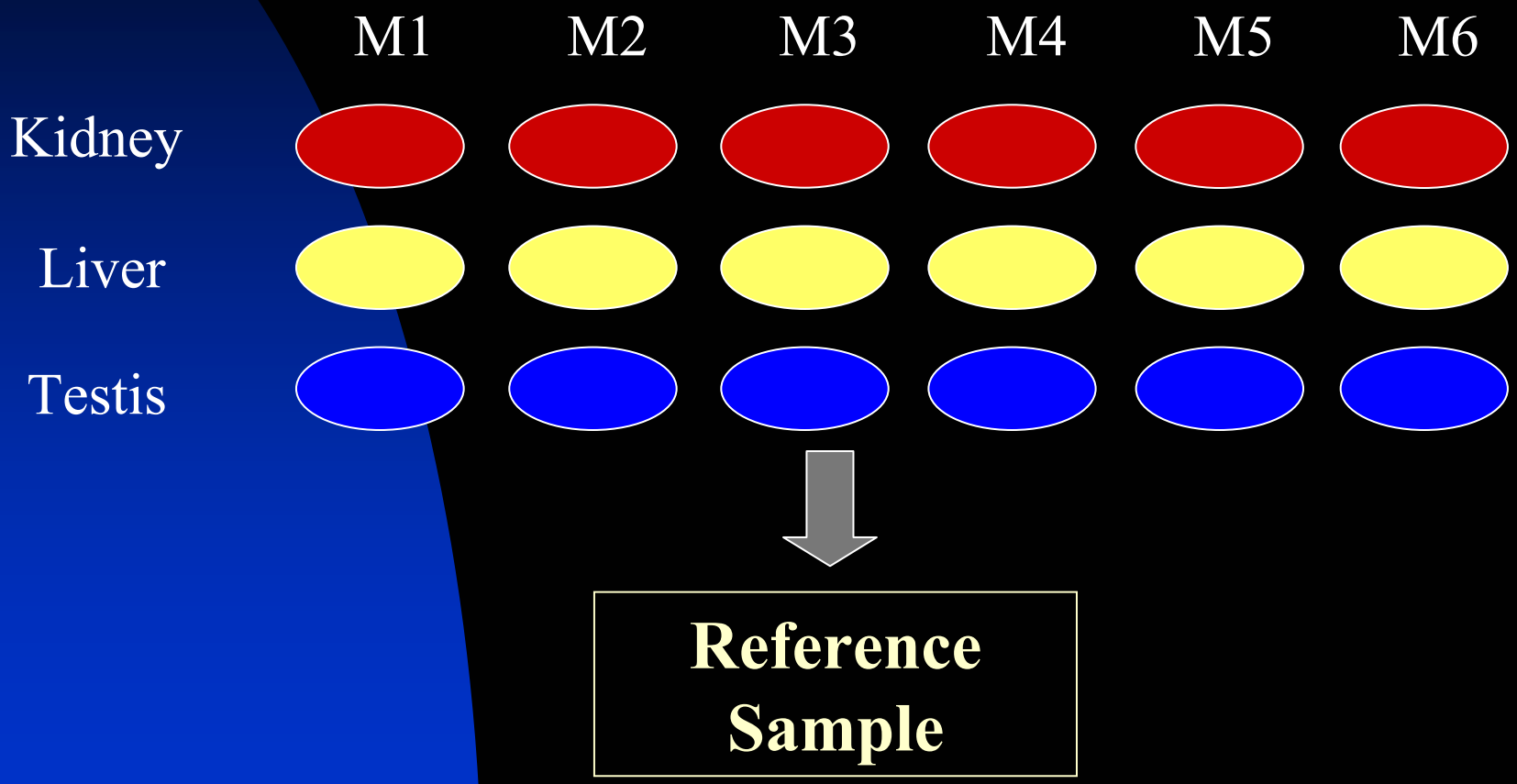
	Method 1	Method 2	Method 3	Method 4
Method 2	94.3			
Method 3	12	22.4		
Method 4	6.8	42.6	223.8	
Method 5	12.91	130.8	301.77	65.31

Pair-wise power comparisons show all pairs of methods have significantly different power in detecting mouse effect

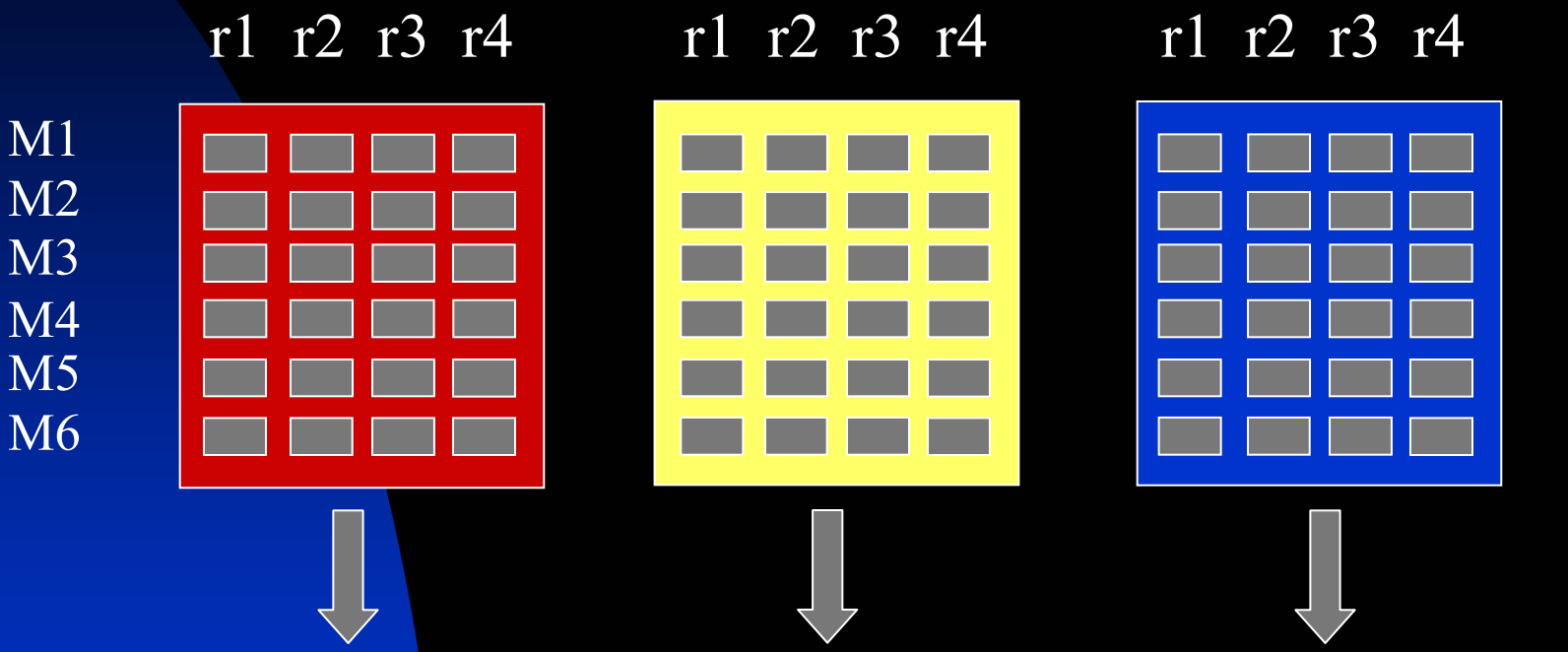
Why Do They Differ



Assessing Data Quality



Assessing Data Quality



Assessing Data Quality

On a gene-by-gene basis

Test Samples (72)

x_{ijk}

y_{ijk}

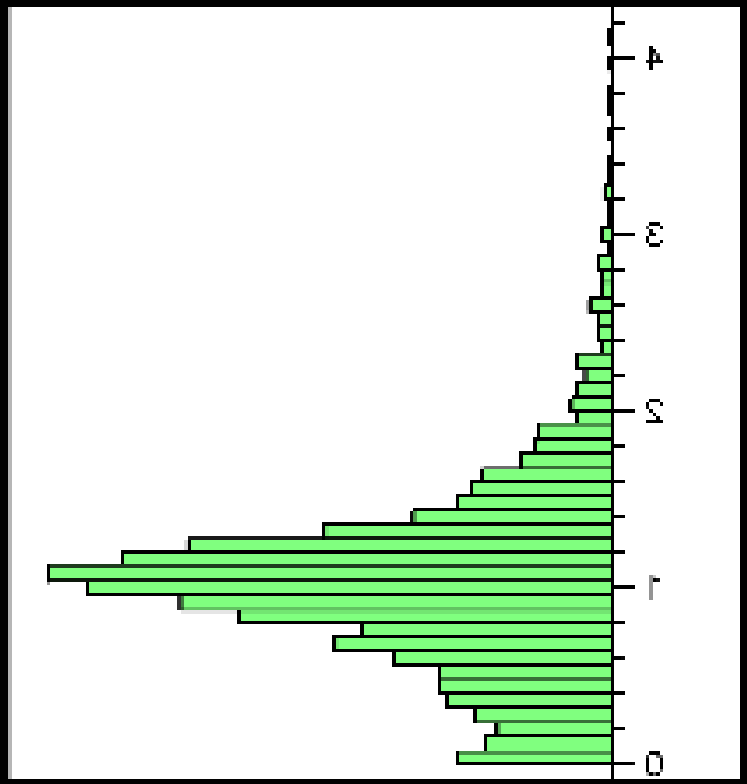
Reference Samples (72)

$$\sum_{i=1}^{3}\sum_{j=1}^{6}\sum_{k=1}^{4} x_{ijk} = \sum_{i=1}^{3}\sum_{j=1}^{6}\sum_{k=1}^{4} y_{ijk}$$

Assessing Data Quality

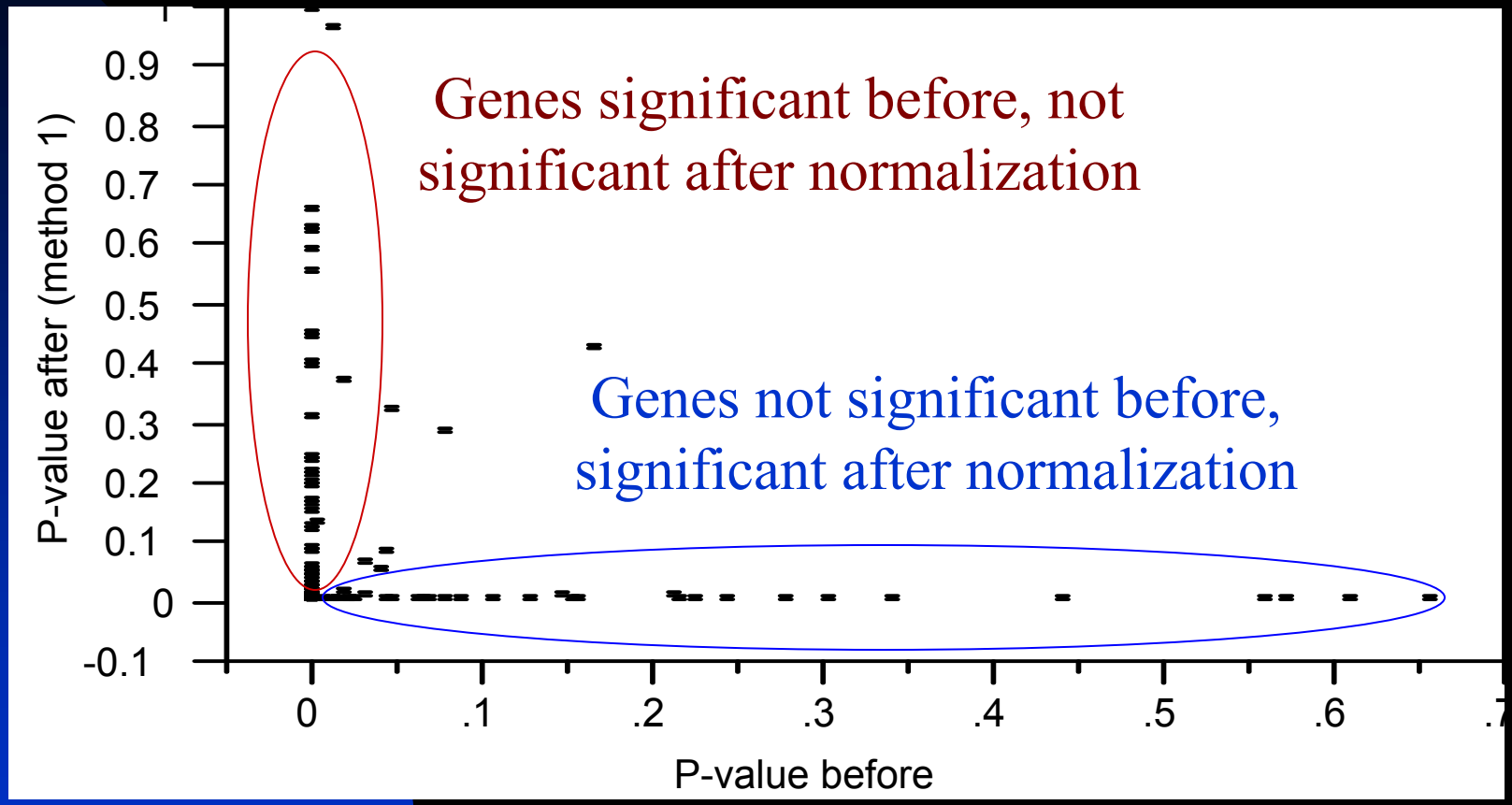
- Let r = $\sum_{i=1}^{3}\sum_{j=1}^{6}\sum_{k=1}^{4} x_{ijk} \Big/ \sum_{i=1}^{3}\sum_{j=1}^{6}\sum_{k=1}^{4} y_{ijk}$
- Examine normalization effect within the set of genes where 1) r<0.5

  2) r>2
- 388 genes in Kidney are significant by at least one method, among which 156 genes have r<0.5 or r >2.



Histogram of r for all genes

Normalization Effect



P-values before and after normalization method 1

Assessing Data Quality

- When (foreground – background) < 0
  ➔ no hybridization
- How about (foreground – background) >0, but < 100?
- 388 genes in Kidney are significant by at least one method, among which 124 genes have (foreground – background) < 100.
- Effect of normalization is examined in these genes.

Assessing Data Quality

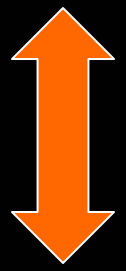
Rep #	Signal(test)	Signal(ref)
1	115	152
2	1077	1476
3	**58**	**60**
4	409	425
1	965	3453
2	865	2243
3	**94**	1471
4	407	2194

- **Low signal intensity**
- **Low mRNA copy number?**
- **Failed hybridization?**
  - **Due to Spotting (if both numbers are small)**
  - **Due to Labeling (one number is small)**

Assessing Data Quality

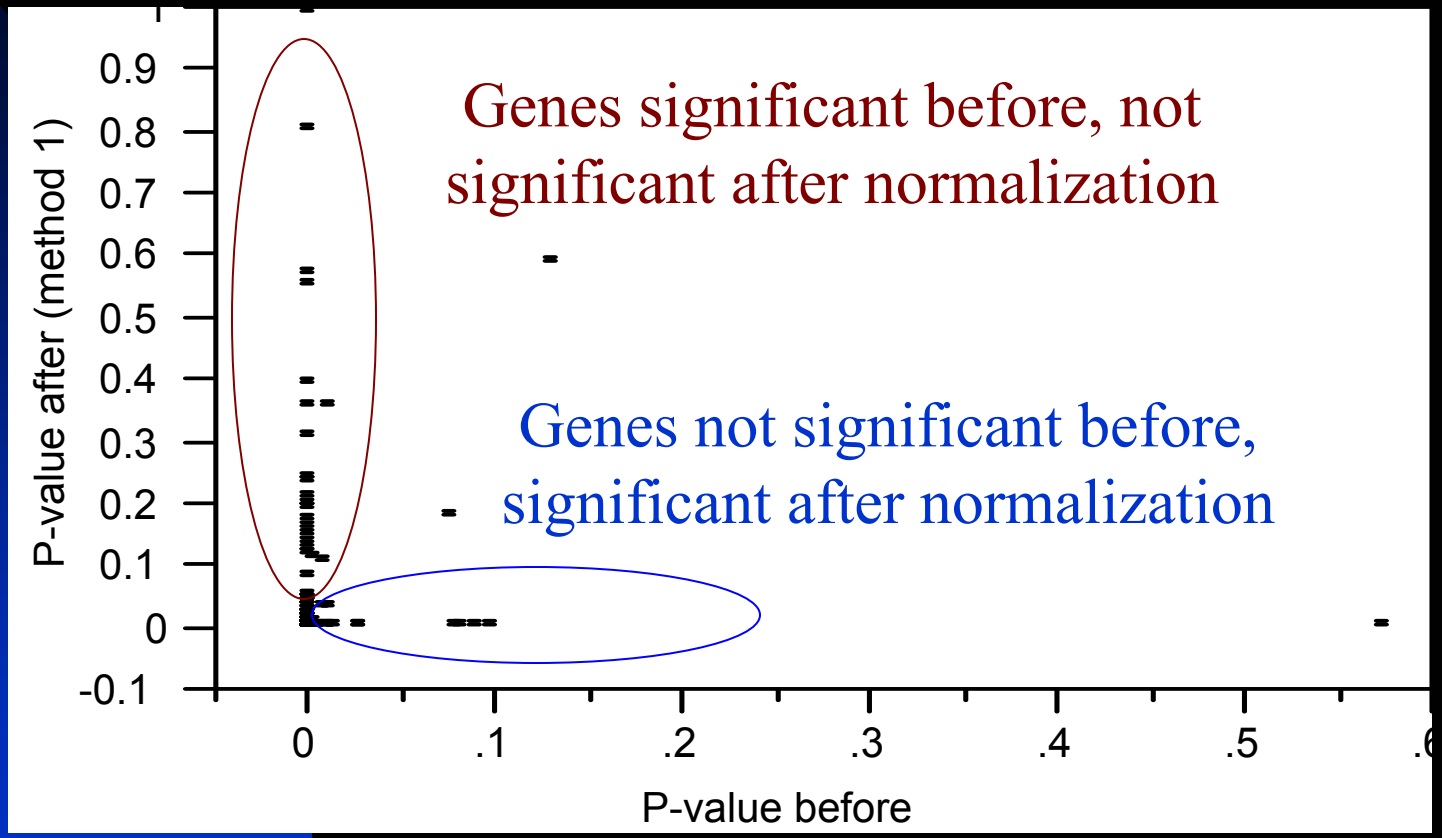
Rep #	Signal (test)	Signal (ref)
1	610	151
2	575	145
3	**10**	**6**
4	365	364
1	525	426
2	**84**	**77**
3	596	753
4	572	753

These genes are affected by normalization



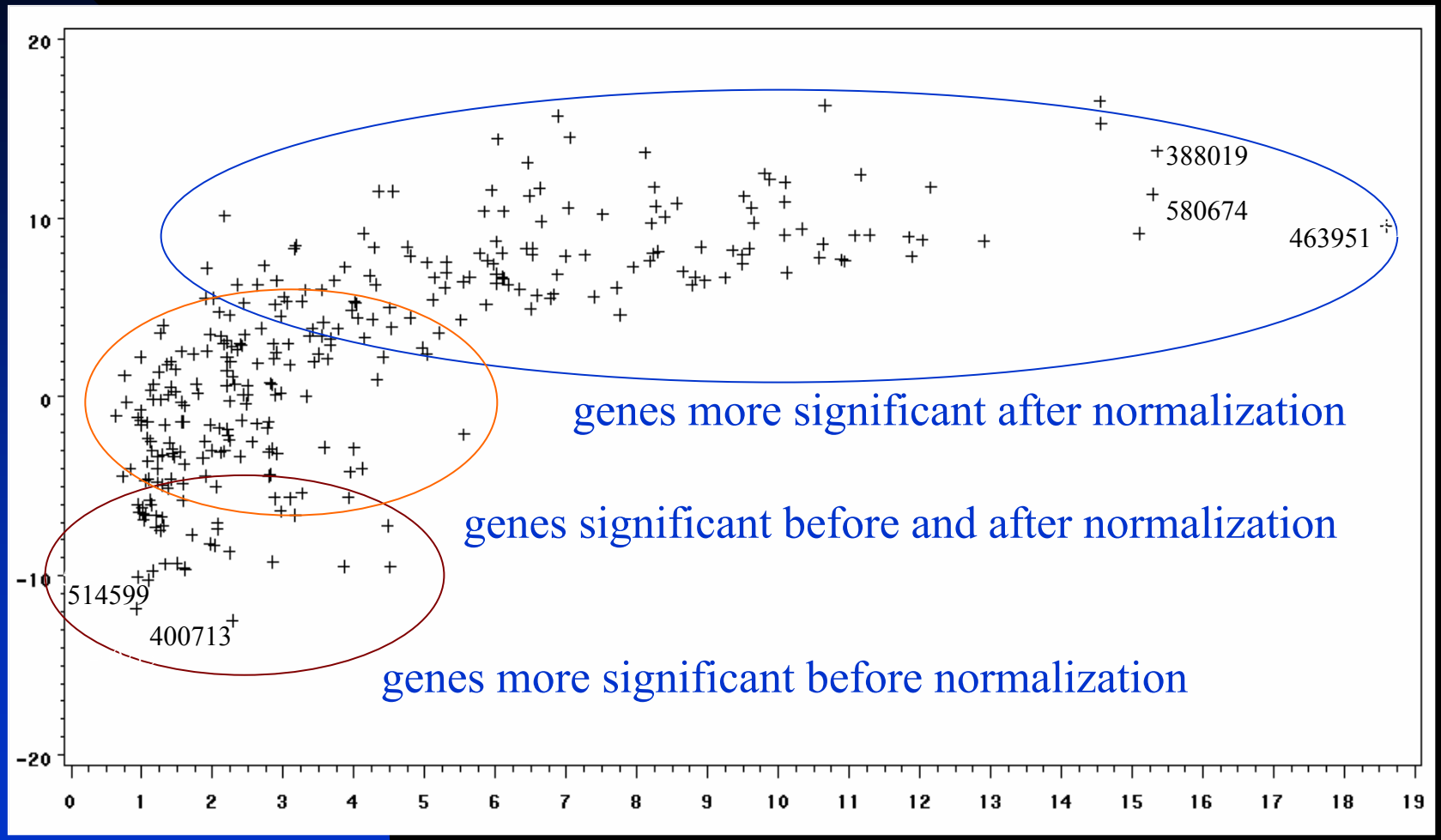
Affecting other genes in the same normalization group

Normalization Effect



P-values before and after normalization method 1

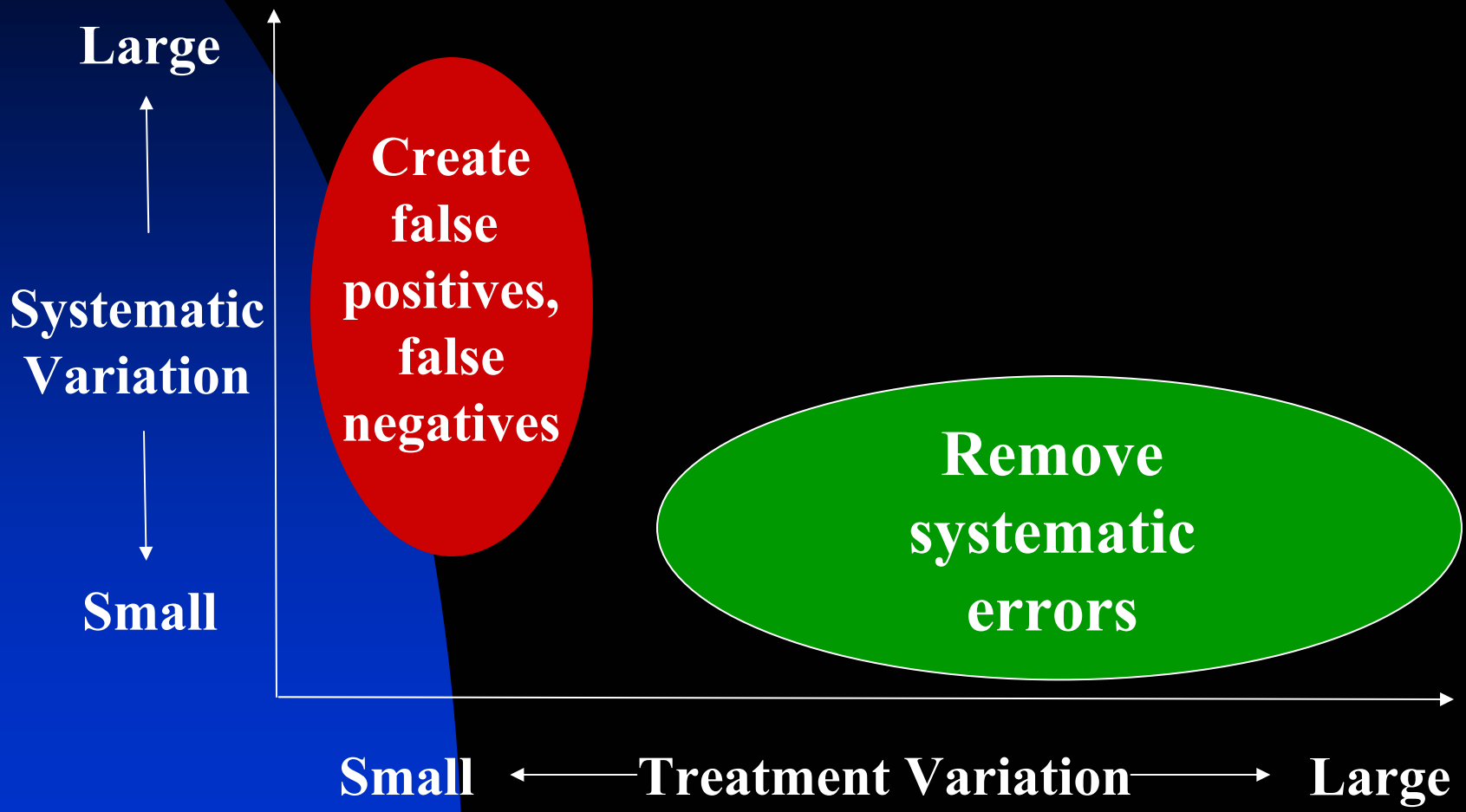
Examine Normalization Effect



Examine Normalization Effect

cDNA	P-value (before)	P-value (after)	STD among mice	STD within block	STD within block/ STD among mice
514599	<0.000001	0.03125	1.40	1.28	0.91
400713	<0.000001	0.06612	1.17	1.11	0.95
463951	0.7182	0.00005	0.07	1.28	18.29
580674	0.6563	0.000008	0.08	1.26	15.75
388019	0.5597	0.000001	0.08	1.26	15.75

Examine Normalization Effect

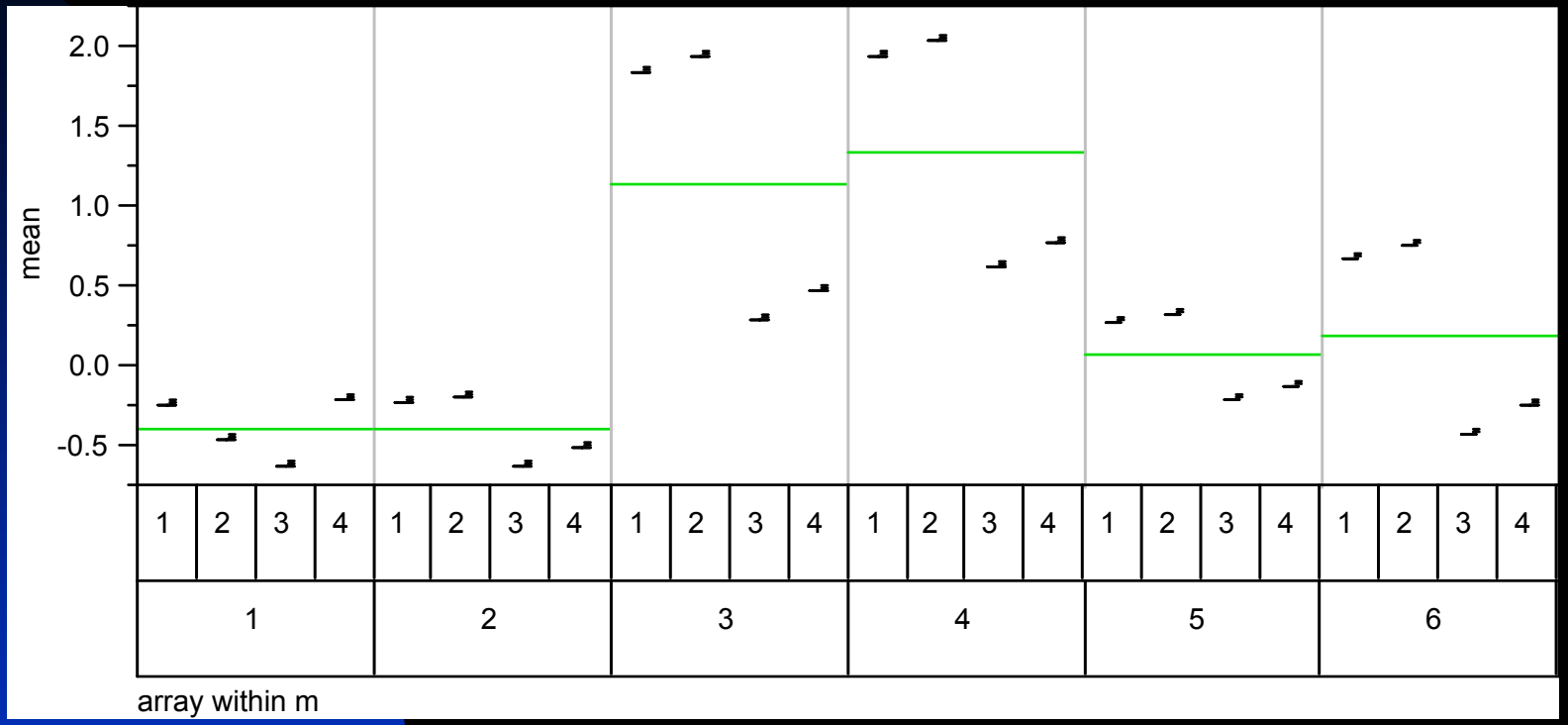


Method Comparison in Three Tissues

Tissue	Criteria	method1	method 2	method 3	method 4	method 5
	Raw_P	936	1440	1808	1253	1057
Kidney	Bonf_P	63	114	196	73	27
	FDR_P	488	1109	1551	757	441
	Raw_P	464	867	809	705	654
Liver	Bonf_P	12	31	25	0	1
	FDR_P	56	328	265	4	1
	Raw_P	853	966	825	3090	3042
Testis	Bonf_P	35	25	24	1956	1163
	FDR_P	272	407	232	3089	3038

Array Means in Testis Tissue

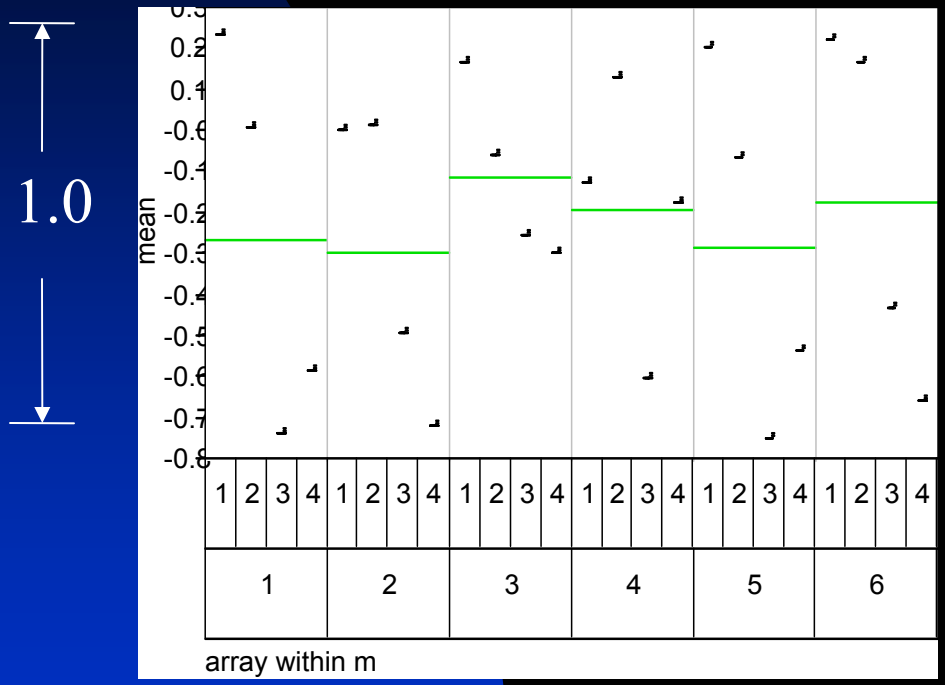
2.5



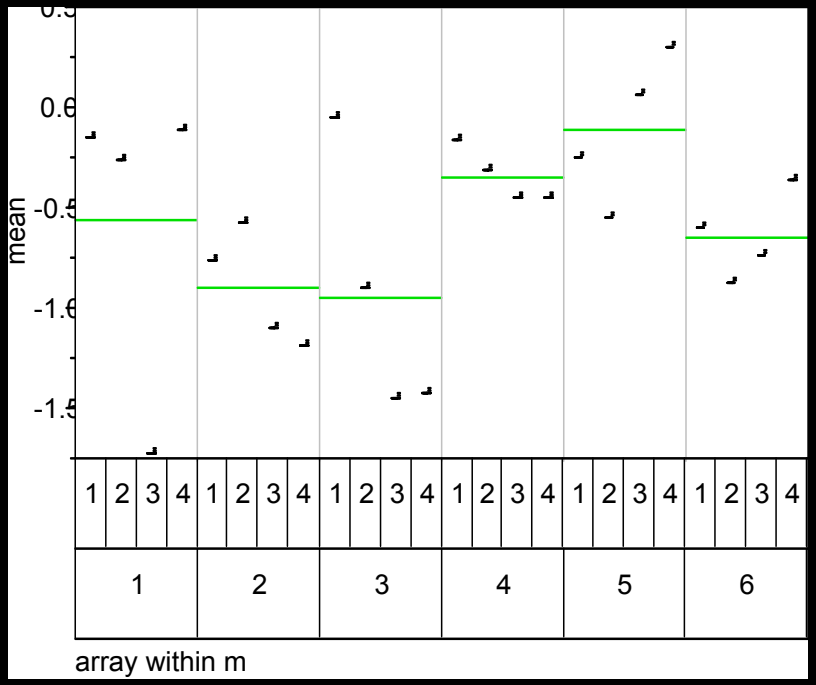
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
m	5	11.150598	2.23012	7.1844	0.0007
Error	18	5.587398	0.31041		
C. Total	23	16.737997			

Array Means in Liver and Kidney



Kidney



Liver

Normalization: to do or not to do

- There exists significant systematic errors
- Normalization aims at removing such systematic errors
- To normalize: added noise can create false positives and false negatives
- Not to normalize: systematic errors can create false positives and false negatives

Some Possible Solutions

- Quality Control
  - Ensure data quality by using both positive and negative controls
  - Perform multiple independent labeling reactions
- Experimental Design
  - Replicate genes within and among blocks such that block effect can be fit into gene-based ANOVA models
  - Two-stage experiment:
    - Pilot study – estimate variances; conduct power analysis to determine how many replicates, how many samples, etc. for the experiment
    - Large scale experiment

Conclusions

- Normalization is an important step to remove systematic effects before data analysis.
- Effective normalization needs to be done after data quality is ensured. QC standards need to be established for large scale microarray experiments to ensure data quality.
- Experimental design plays a crucial role in both data analysis and making normalization effective. Labeling effect needs to be incorporated into the design.
- Genes can have normal baseline variations - different positive controls need to be incorporated into experimental designs.

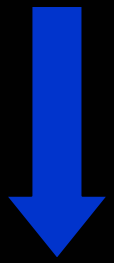
Conclusions

Stage I

(hypothesis generating)

Large number of genes, small number of samples

fully balanced



experimental designs

Stage II

(hypothesis testing)

Small number of genes, large number of samples

Acknowledgements

- Dr. Bruce Weir
- Dr. Ross Whetten
- Dr. Yinghsuan Sun