Design and Analysis of Gene Expression Microarray Experiments

M. Kathleen Kerr

Gary A. Churchill

The Jackson Laboratory

OUTLINE

Experimental Design for Microarrays

Analysis of Variance for Microarrays

Identifying Differentially Expressed Genes

Corning Data Analysis

Synteni Data Analysis

Chu et al. Data Analysis

"Statistical procedure and experimental design are only two different aspects of the same whole, and that whole comprises all the logical requirements of the complete process of adding to natural knowledge by experimentation"

-Fisher 1935

"It is possible, and indeed it is all too frequent, for an experiment to be so conducted that no valid estimate of error is available"

-Fisher 1935

"A well–designed experiment will usually allow its conclusions to be easily obtained, wheras no computations, however industriously or ingeneously p rformed, can produce entirely satisfactcry conclusions from an ill–designed one.

Considerable tact is needed in discussion of these matters."

Finney 1953

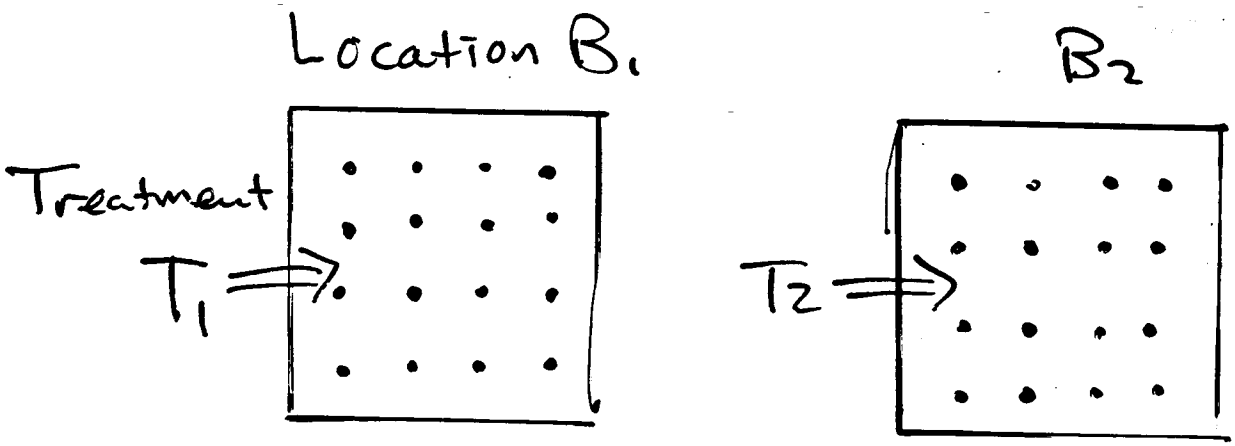
The "design" of an experiment is

1. the set of treatments selected for comparison

2. the specification of the units to which the treatments will be applied

3. the rules by which the treatments are allocated to the experimental units

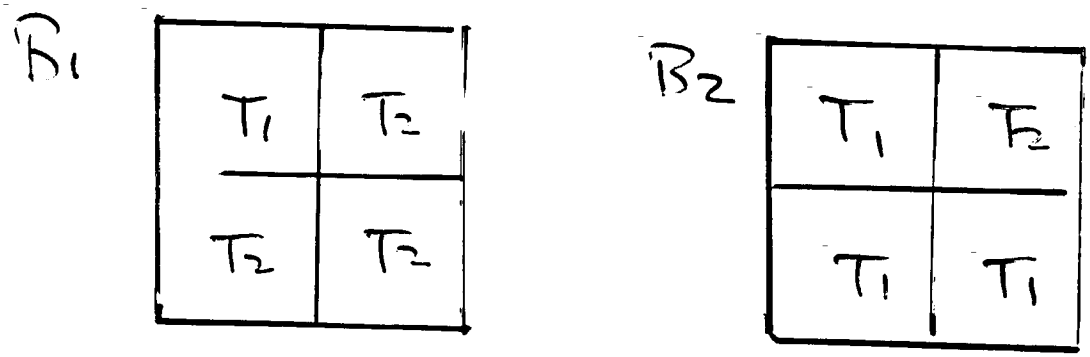
4. the specification ofthe measurements to be made.



Confounding

$Var(T_1 - T_2) = 2\sigma^2$

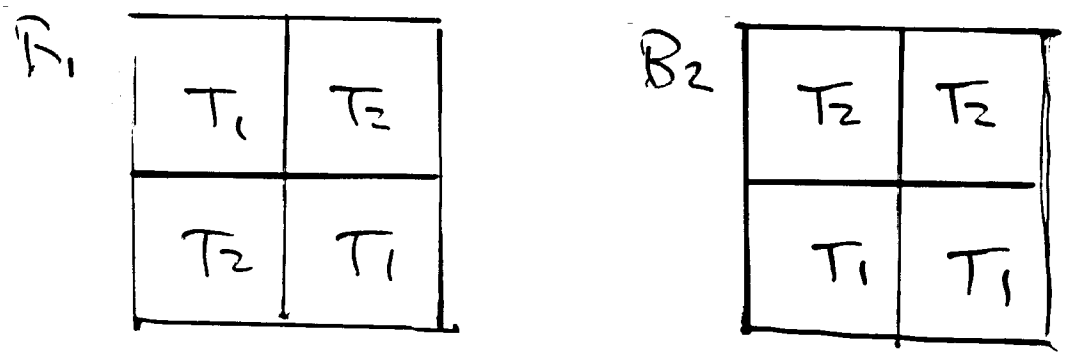
error $dF = 0$



Partial Confounding

$Var(T_1 - T_2) = 2/3\sigma^2$

error $dF = 6$



Orthogonality

$Var(T_1 - T_2) = 1/2\sigma^2$

error $dF = 6$

Relative EFFciency

unbalanced

B_1	B_2
A	A
B	A
B	B
B	A

balanced

B_1	B_2
A	B
B	B
A	A
B	A

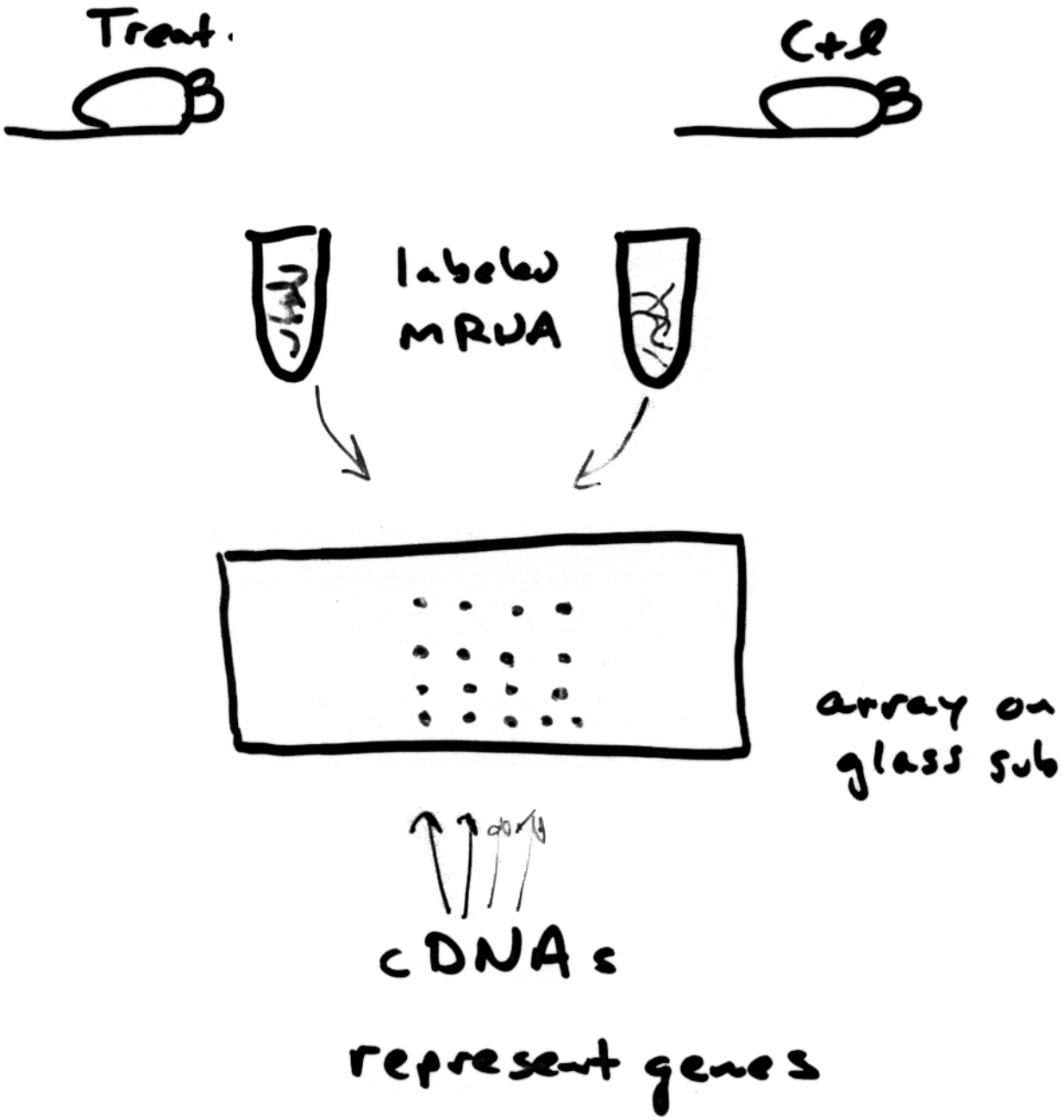
$$X = \begin{bmatrix} & 1 \\ & 1 \\ & 1 \\ & 1 \\ -1 & 1 \\ -1 & 1 \\ & 1 \\ -1 & 1 \end{bmatrix} \qquad \begin{bmatrix} & 1 \\ & 1 \\ -1 & 1 \\ -1 & 1 \\ -1 & 1 \\ -1 & 1 \\ & 1 \\ & 1 \end{bmatrix}$$

$$X'X = \begin{bmatrix} 8 & 0 & 0 \\ 0 & 8 & -4 \\ 0 & 4 & 8 \end{bmatrix} \qquad \begin{bmatrix} 8 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 8 \end{bmatrix}$$

$$X'X = \begin{bmatrix} 25 & 0 & 0 \\ 0 & 167 & .083 \\ 0 & 083 & .167 \end{bmatrix} \qquad \begin{bmatrix} 25 & 0 & 0 \\ 0 & 25 & 0 \\ 0 & 0 & 125 \end{bmatrix}$$

$\sigma^2(T_A\ T_B) = 4 \times .167 = 0\ 667$ $\qquad 4 \times 125 = 0\ 5$

Spotted cDNA Microarrays



represent genes

Design Factors in a Microarray Experiment:

A_i	arrays	$i = 1, \ldots, K$
D_j	dyes	$j = 1, 2$
G_g	genes	$g = 1, \ldots, N$
V_k	varieties	$k = 1, \ldots K$
S_r	spots	r

Genes

- main effects G

  average expression level
  dye incorporation
  hybridization efficiency

- array by gene A × G

  → variations in spot size

- dye by gene D × G

  → the little green spot

- variety by gene V × G

  → variety specific gene effect

VG_{ig} VG_{jg} = change in expression

Structure of Design Space

- G ⊥ A, D, V
- AG ⊥ DG
- AG, VG partial confound
  DG, VG can be ⊥ or partial confounded
- High order interactions

  A*D*C A*V*G

  D*V*C A*D*V*G

  are confounded with L.O.Ts

Arrays, Dyes and Varieties

- Interaction effects

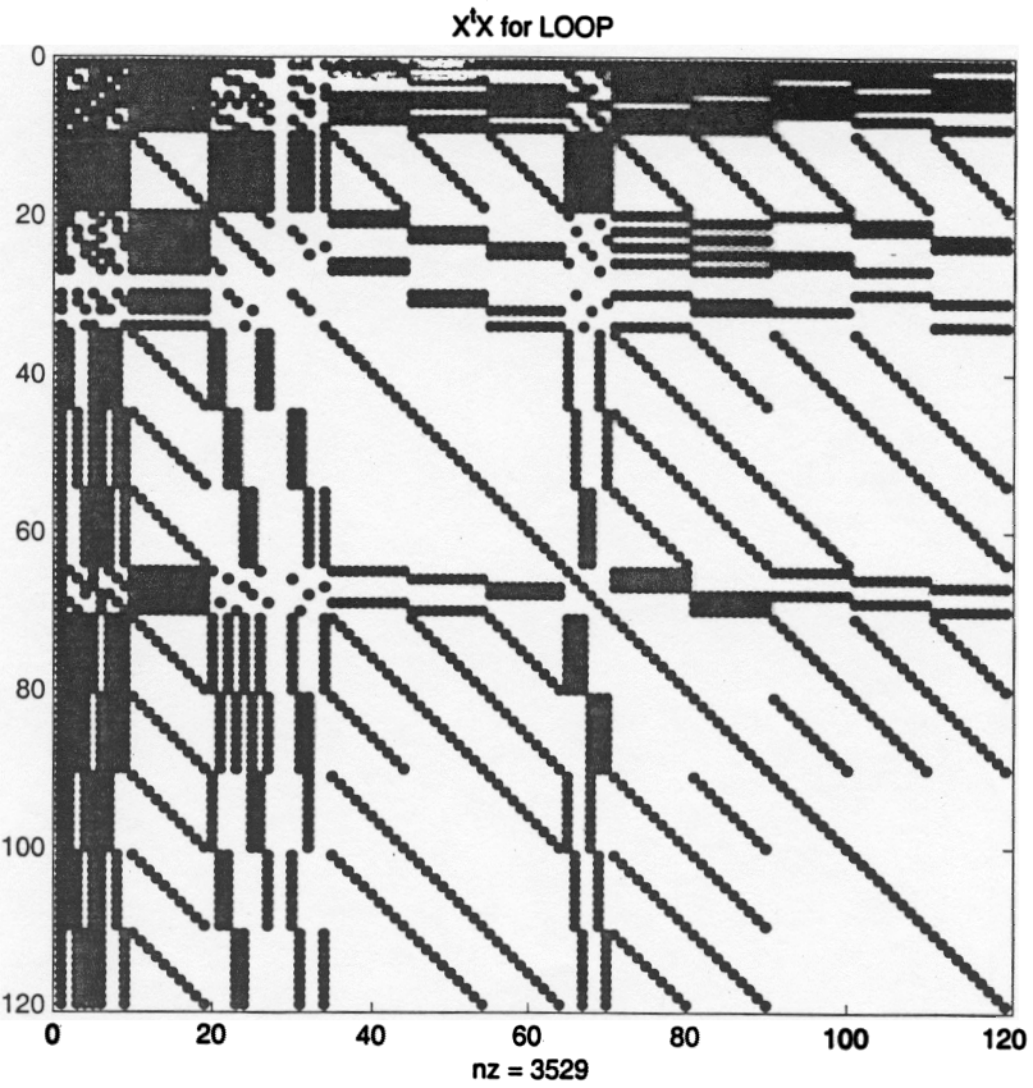
  A*D, A*V, D*V, A*D*

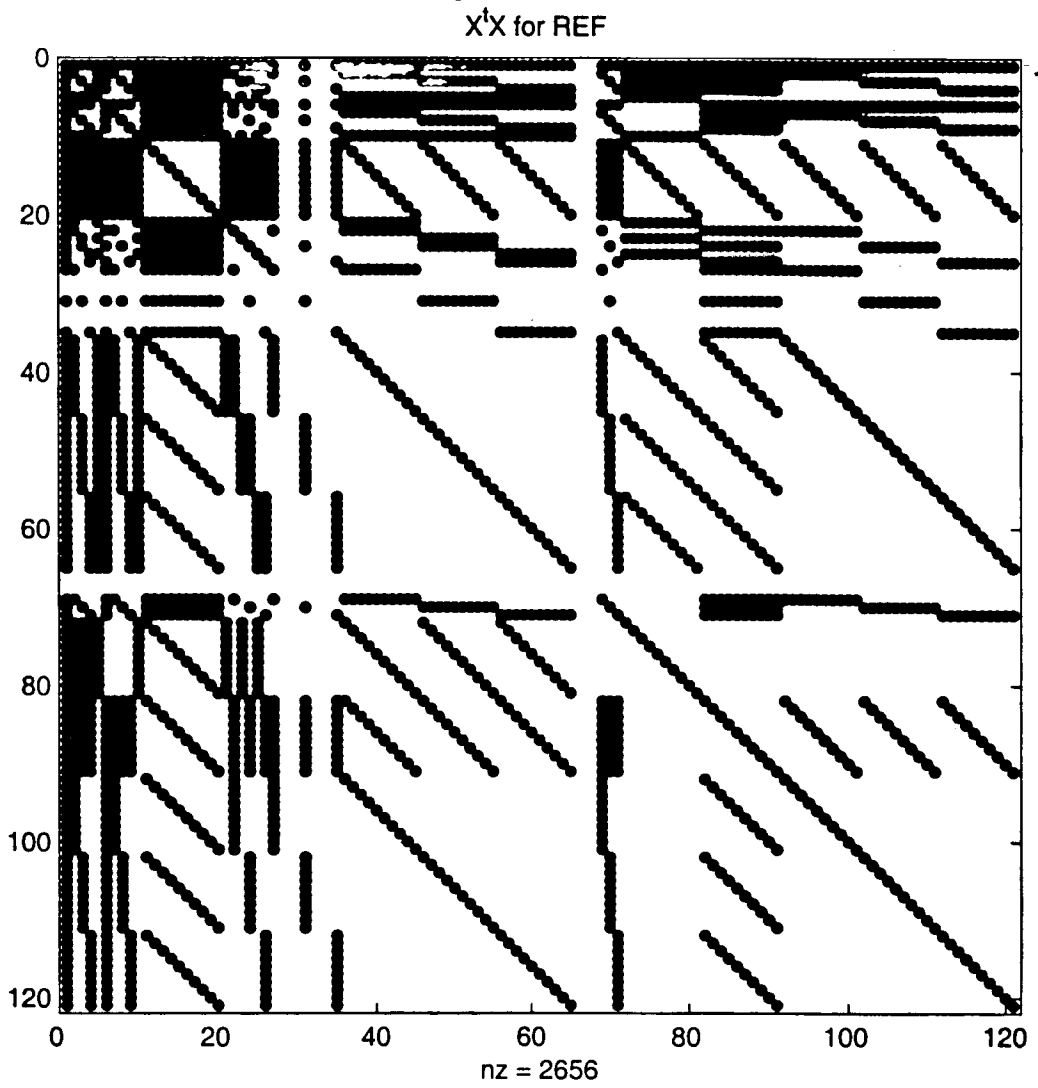
  are confounded with the main effects A, D, V

  A and D are always balanced

- D and V can be balanced

- A and V cannot be balanced if $k > 2$


X^tX for LOOP
0
20
40
60
80
100
120
0
20
40
60
80
100
120
nz = 3529


X^tX for REF
0
20
40
60
80
100
120
0
20
40
60
80
100
120
nz = 2656

A simple ANOVA model includes only the factor main effects and the effects of interest, variety×gene:

$$y_{ijkg} = \mu + A_i + D_j + V_k + G_g + (VG)_{kg} + \epsilon_{ijkg}. \tag{4.1}$$

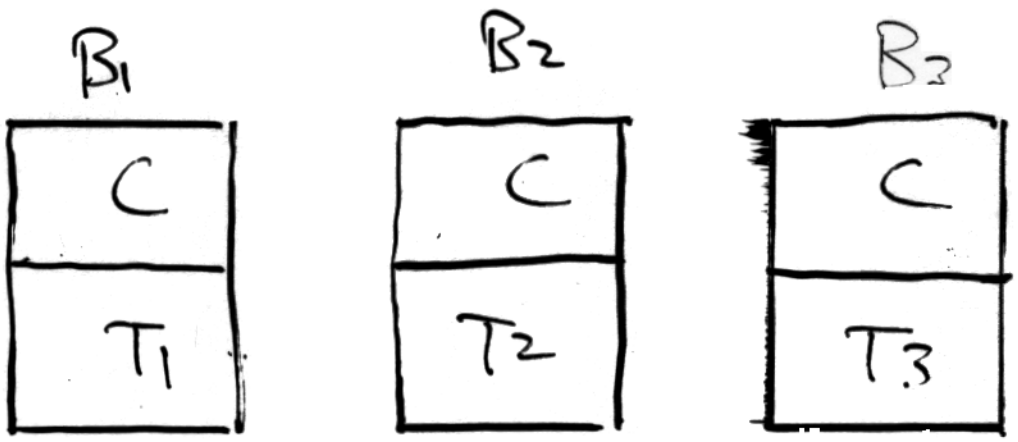
A more plausible model accounts for spot-to-spot variation by including array×gene effects:

$$y_{ijkg} = \mu + A_i + D_j + V_k + G_g + (VG)_{kg} + (AG)_{ig} + \epsilon_{ijkg}. \tag{4.2}$$

Another possibility is to further account for genes interacting with dyes:

$$y_{ijkg} = \mu + A_i + D_j + V_k + G_g + (VG)_{kg} + (AG)_{ig} + (DG)_{jg} + \epsilon_{ijkg}. \tag{4.3}$$

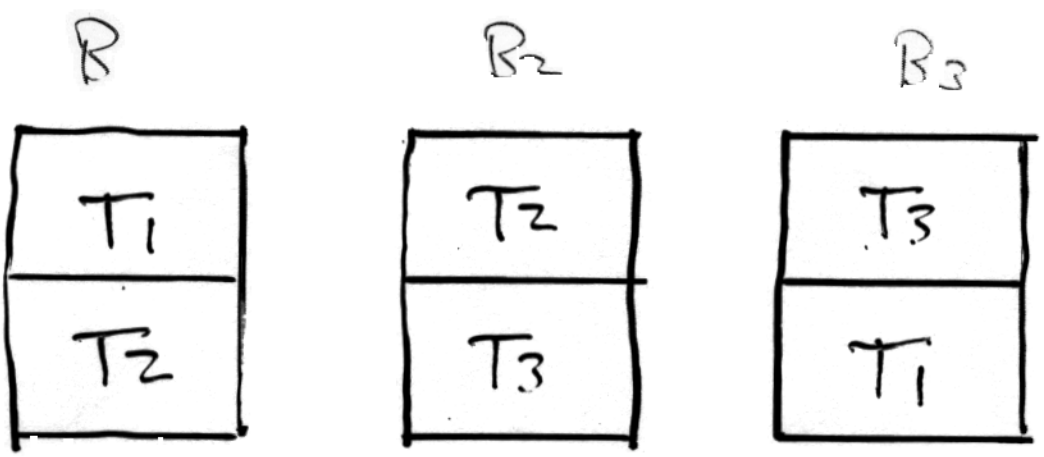
We assume that there is independent, additive error $\epsilon_{ijkg} \sim F$, where F is a distribution with mean 0 and variance σ^2.



Comparison to a common control

$Var(T_A\ T_B) = 4\sigma^2$

error $dF = 6$

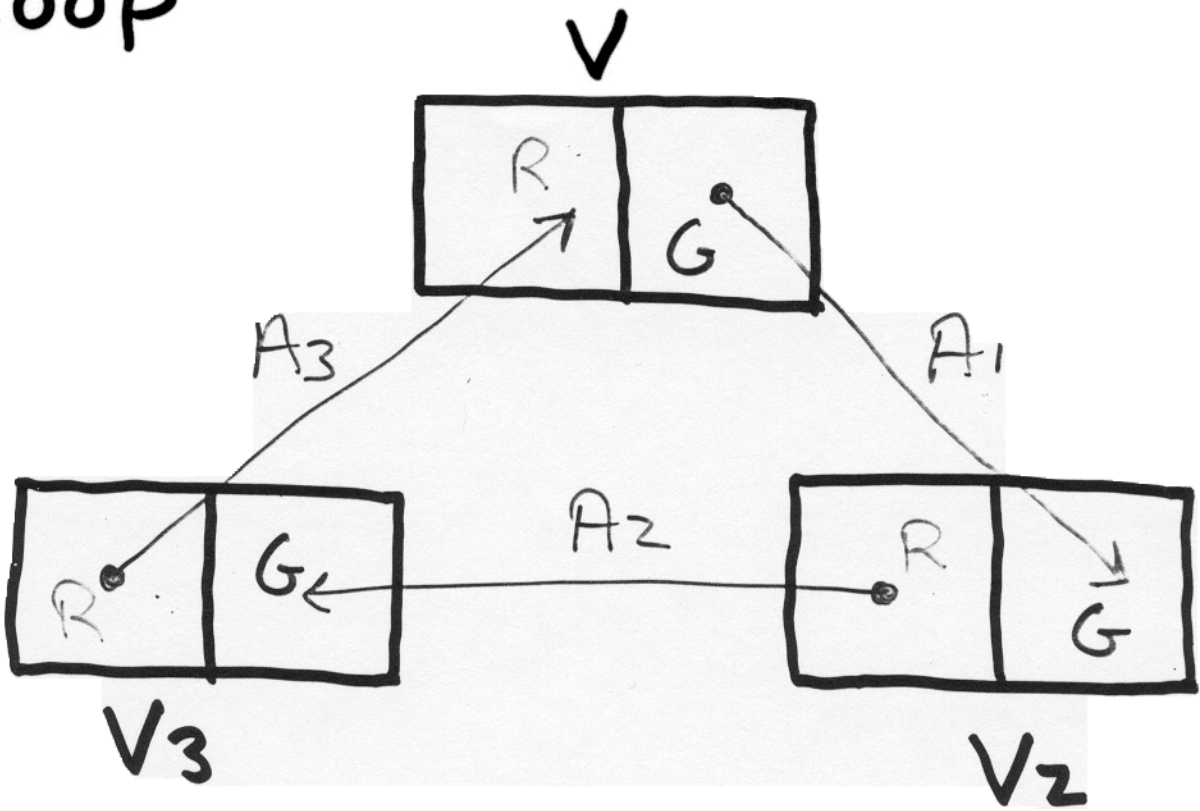


Balanced incomplete blocks

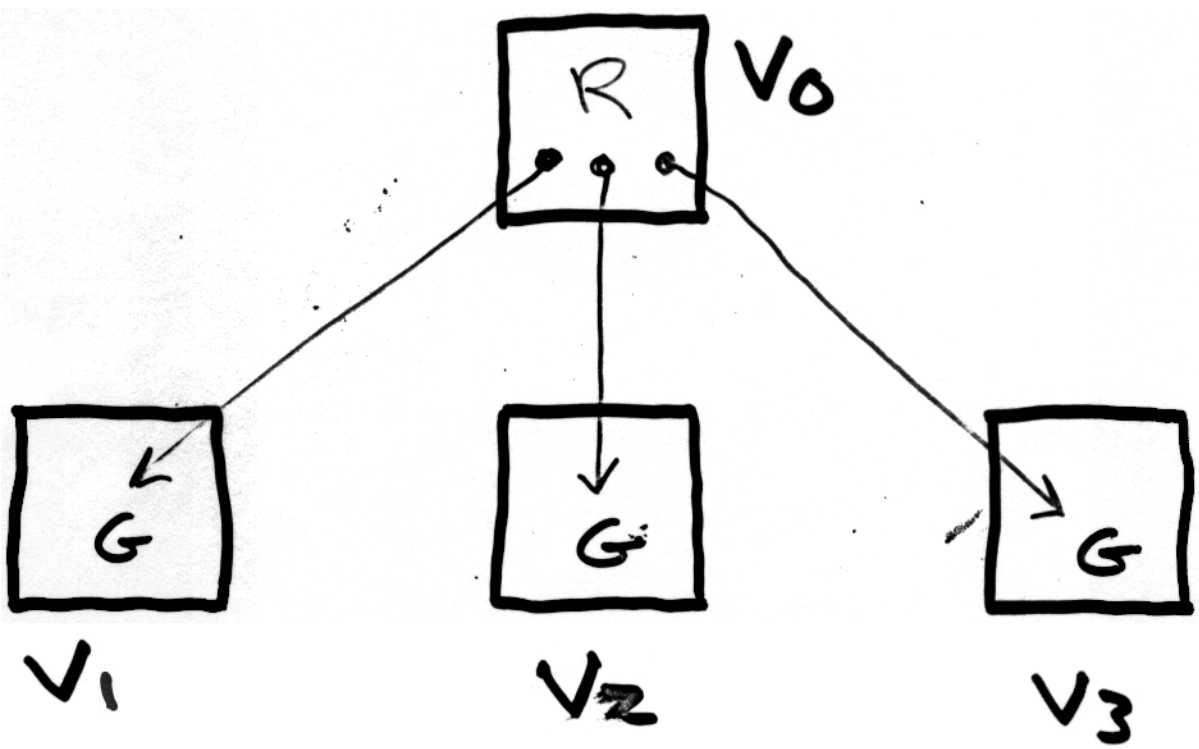
$Var\ (T_A\ T_B) = 4/3\ \sigma$

error $dF =$

Loop



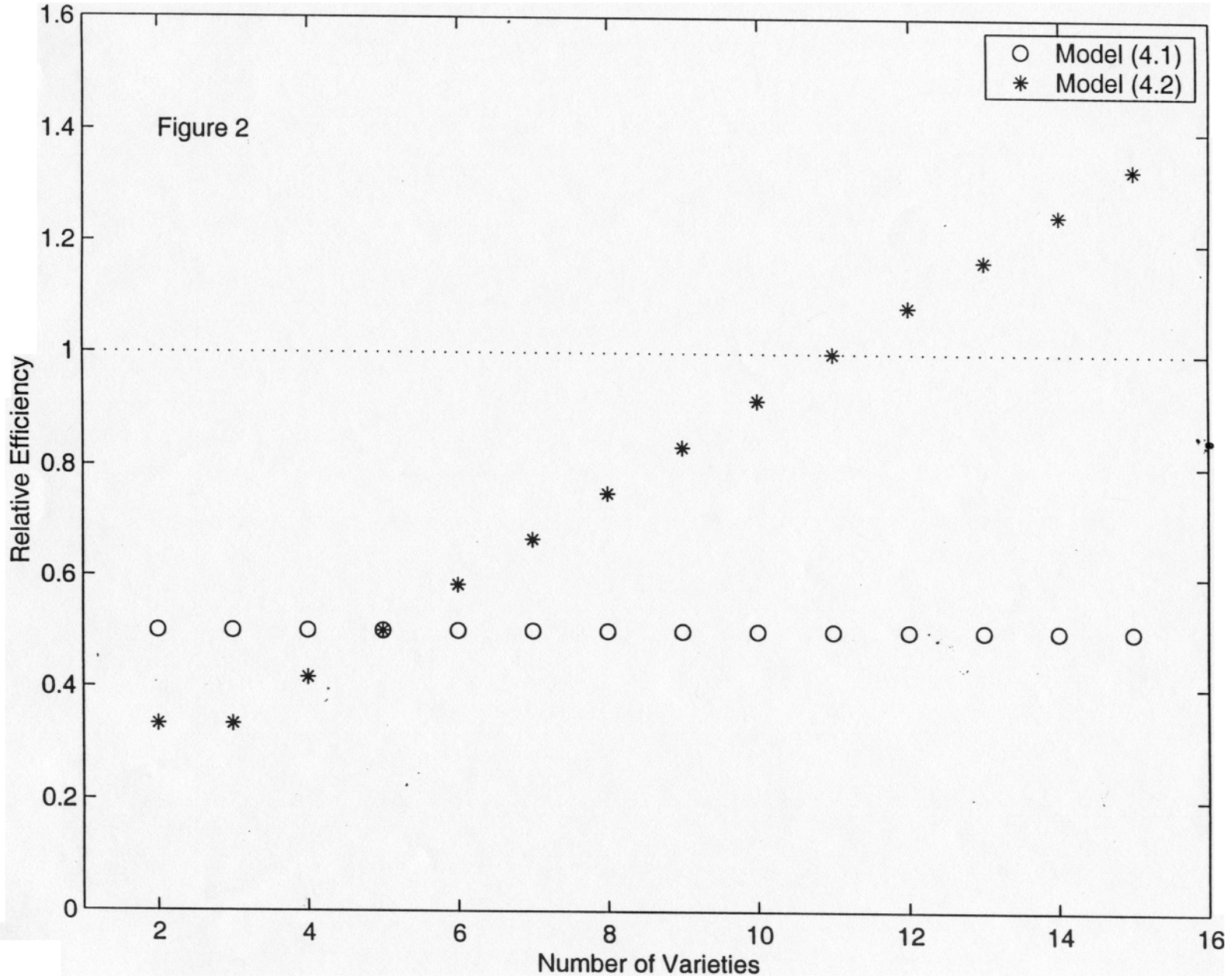
ReFerence



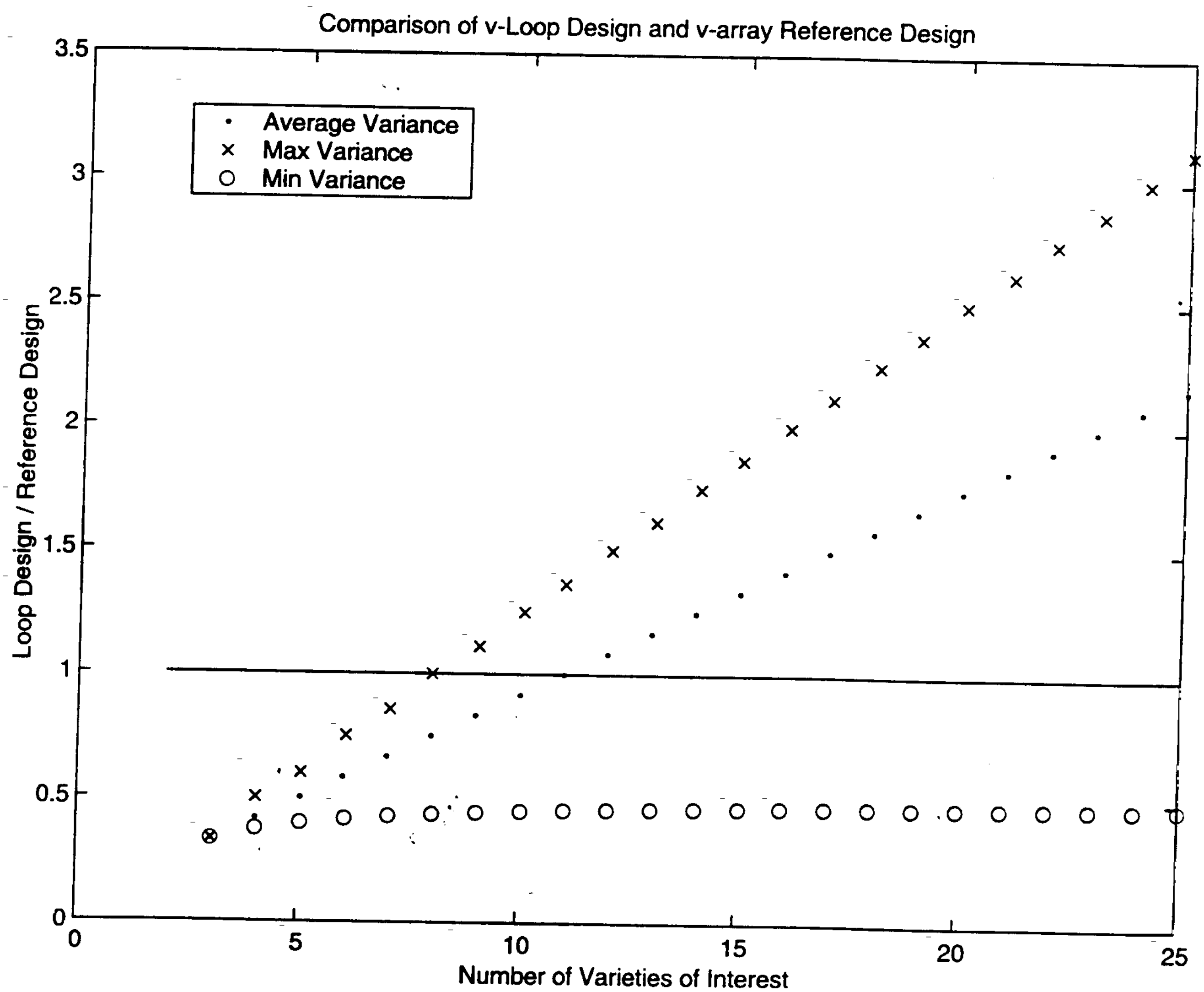
Source	Reference df	Loop df
A,D,T	2K-1	2K-1
G	N-1	N-1
TG	K(N-1)	(K-1)(N-1)
AG	(K-1)(N-1)	(K-1)(N-1)
residual	0	N-1
total	2KN-1	2KN-1

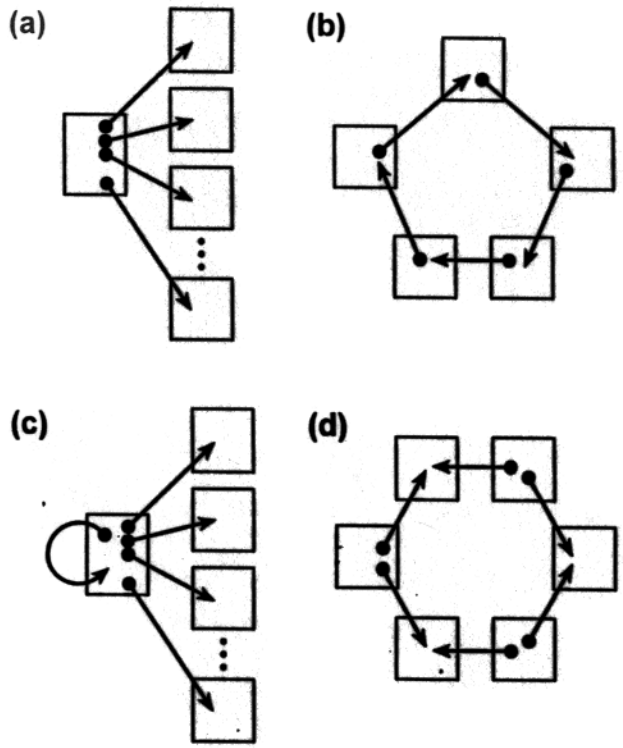
ANOVA table for designs with K arrays and N genes.

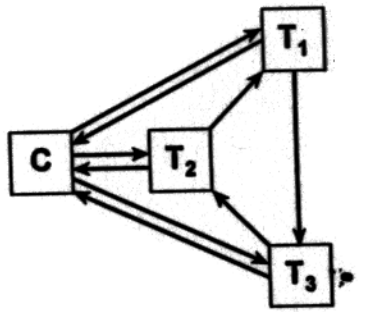
Loop vs Reference Avg. Variance

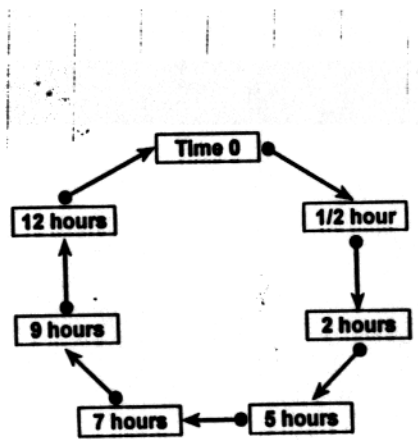


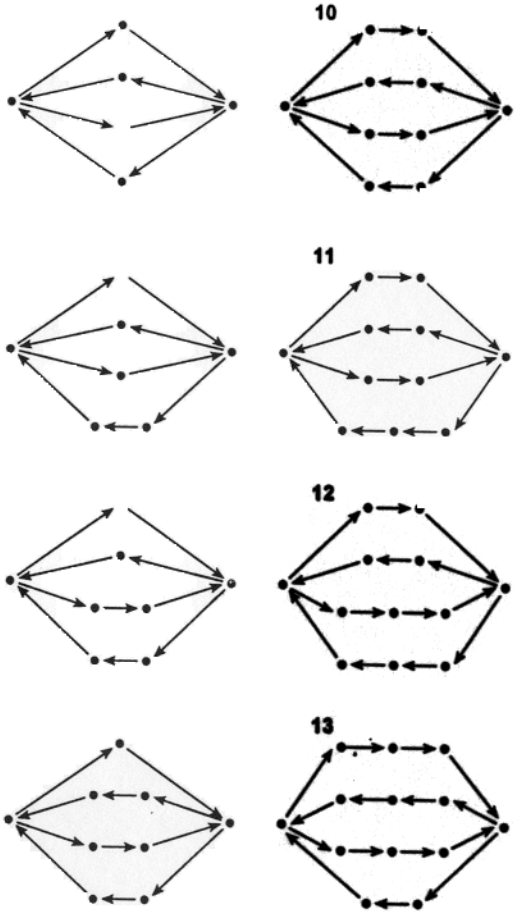
Comparisons within K

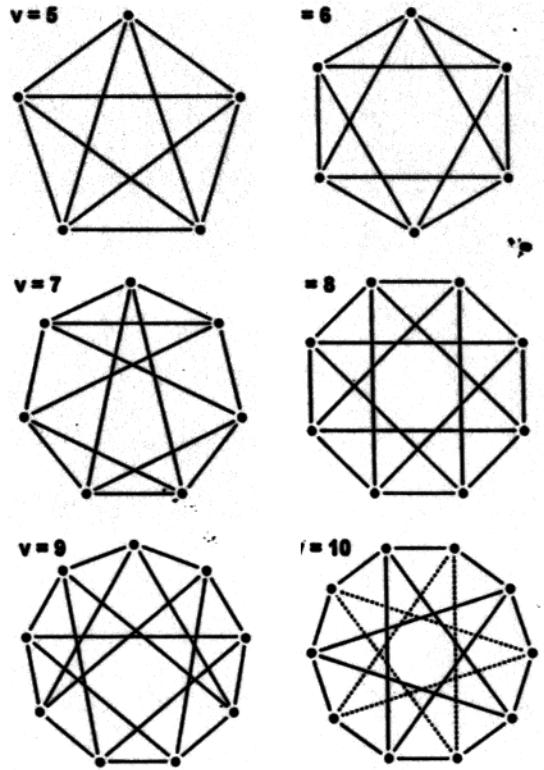



(a)
(b)
(c)
(d)

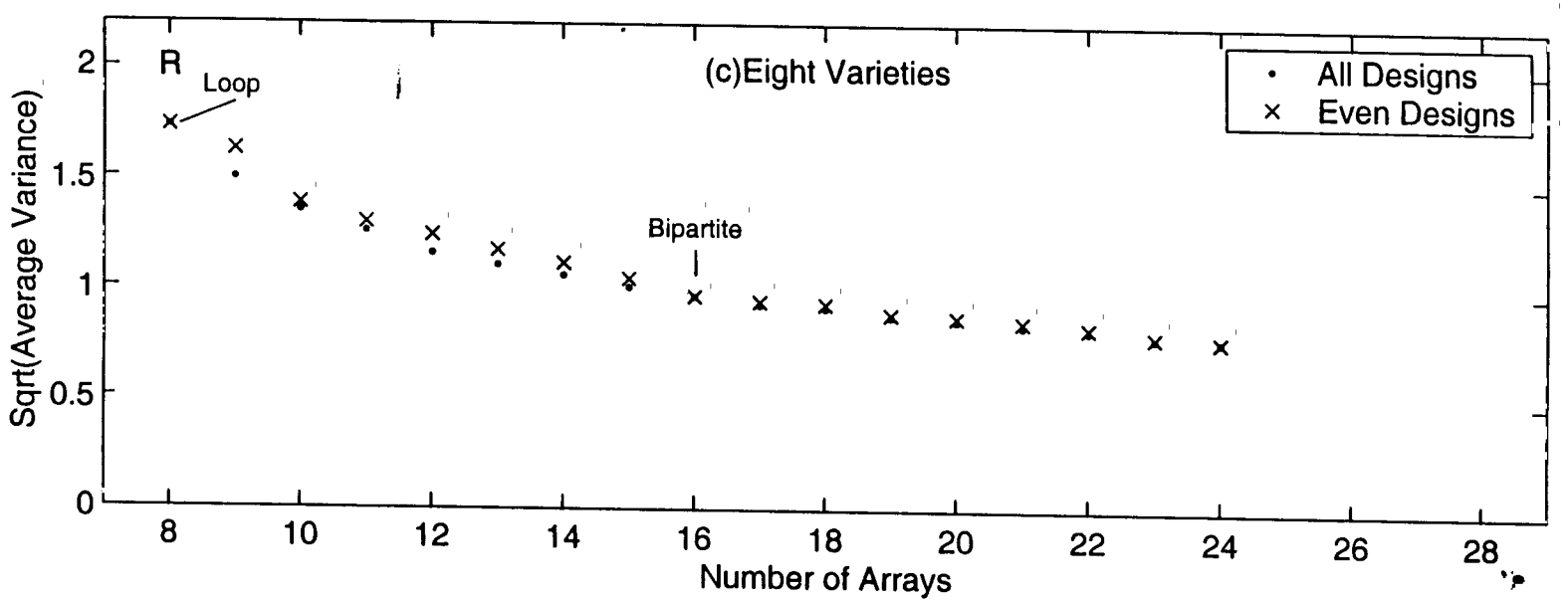

C
T1
T2
T3

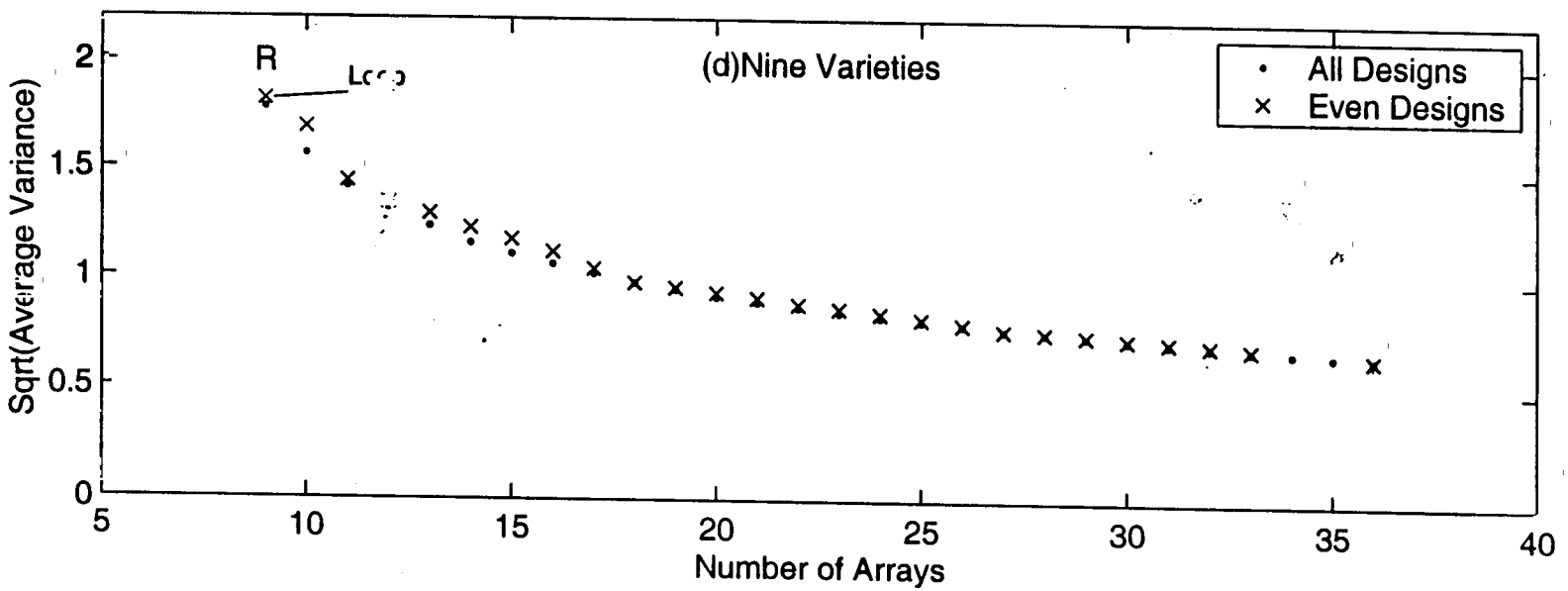

Time 0
1/2 hour
2 hours
5 hours
7 hours
9 hours
12 hours

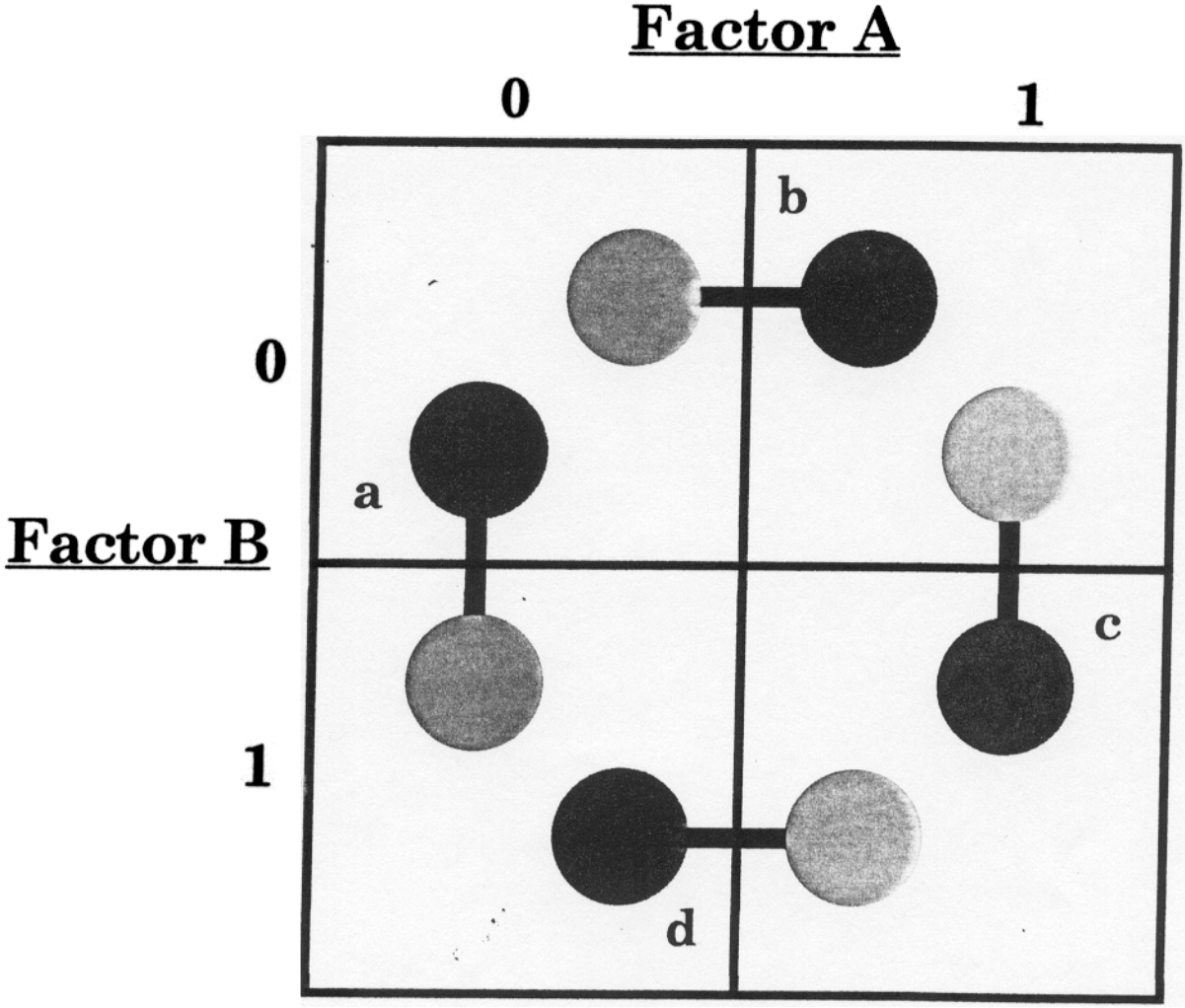

10
11
12
13


v = 5
= 6
v = 7
= 8
v = 9
= 10

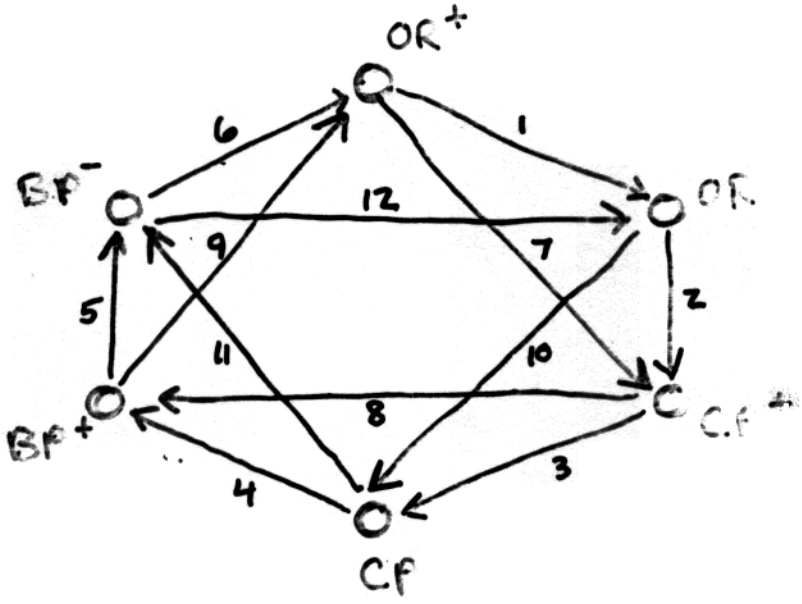
Optimal Designs







Gary Churchill, Ph.D.
The Jackson Laboratory



i	j	k	l
background	genoty	array	cyc
BP	+/+	2	
CP	/+		
OR		2	

pairwise comparisons

OR+ OR $\frac{Y\ Y_7}{Y_6\ Y_9}$ $\frac{Y_2\ Y_{11}}{Y\ Y_{10}}$

contrasts

(CP+ BP OR+) (CP BP OP)

multiple df comparisons

BP- CP OP-

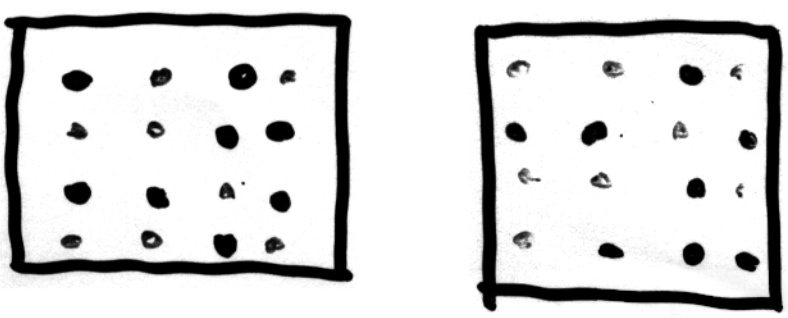
Concluding Remarks

- The key concept that we left out:

  Randomization

- The critical question that no one asked

  IF the design is so bad, why do we get the right answers?



"One characteristic common to all biological material is that it varies."

-Finney 1953

"If I had to replicate my experiments, I could only do half as much"

-Botstein 1999

Synteni Data Analysis

Latin Square Design

Dye	Array 1	Array 2
Red	Liver	Muscle
Green	Muscle	Liver

Analysis of Variance

Source	df	SS	MS
Array	1	92.34	92.34
Dye	1	0.74	0.74
Tissue	1	2.97	2.97
Gene	1285	1885.89	1.47
Gene×Tissue	1285	1357.28	1.06
Residual	2570	242.76	0.09
Corrected Total	5143	3581.99	

Synteni Data Analysis

Reference Sample Design

Dye	Array 1	Array 2
Red	Placenta	Placenta
Green	Liver	Muscle

Analysis of Variance

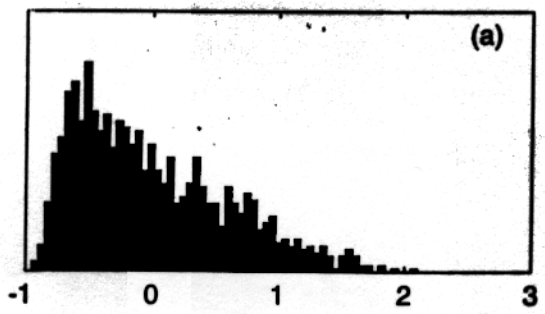
Source	df	SS	MS
Array,Tissue	3	761.97	253.99
Gene	1904	3394.17	1.78
Gene×Tissue	3808	1264.43	0.33
Residual	1904	55.21	0.03
Corrected Total	7619	5475.78	

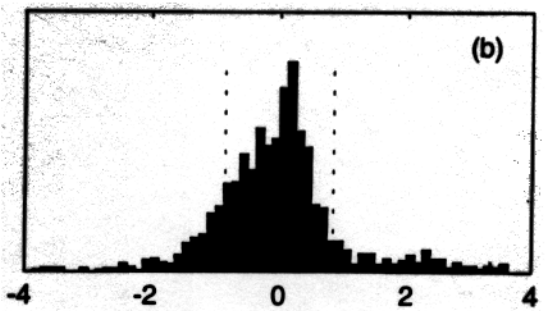
G eFFects

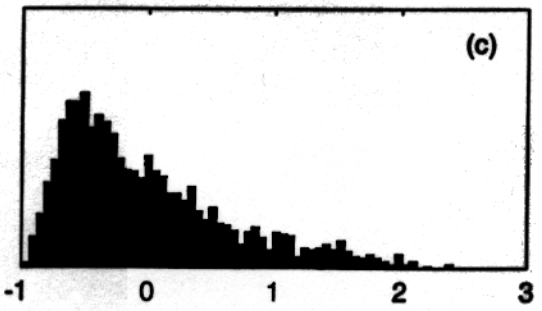
V x G eFFects

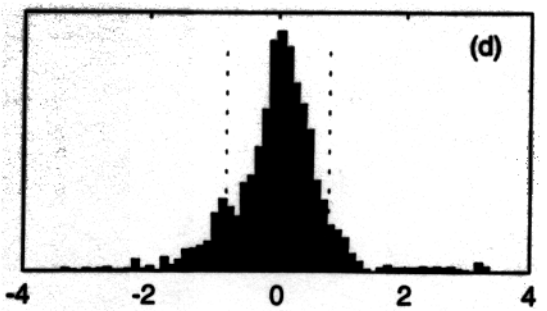
Loop

ReF

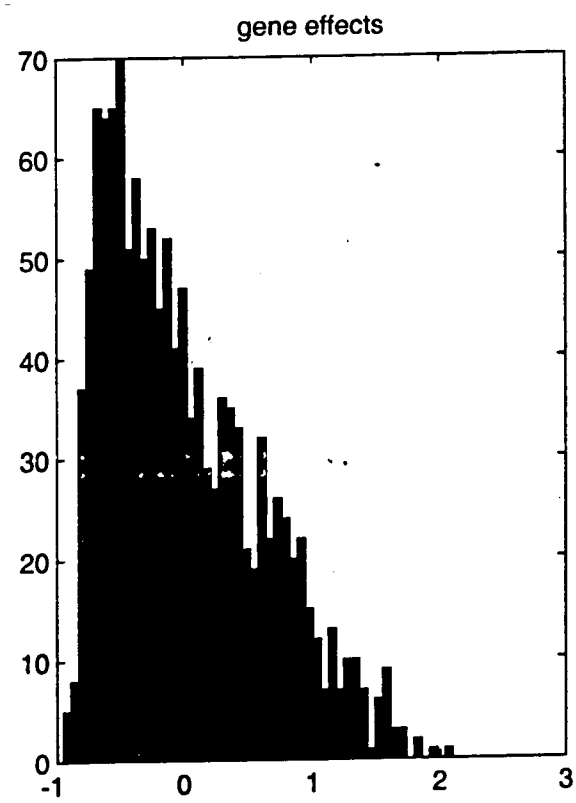


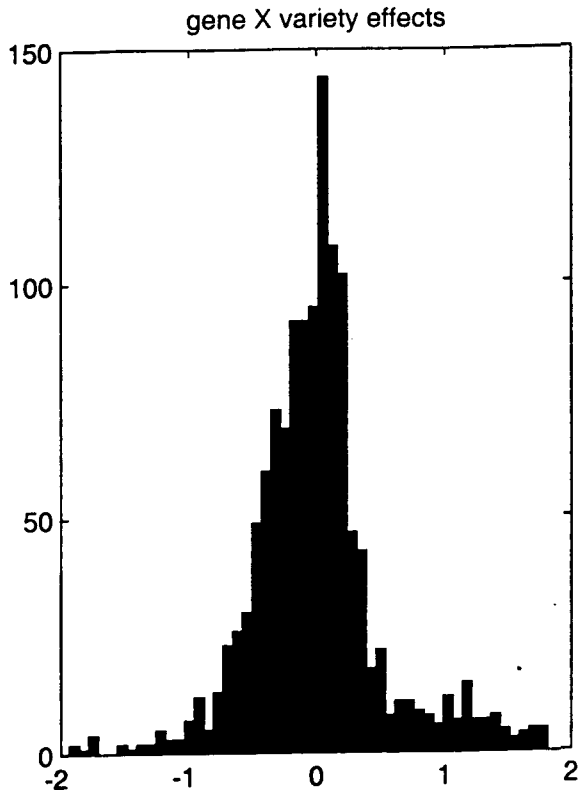


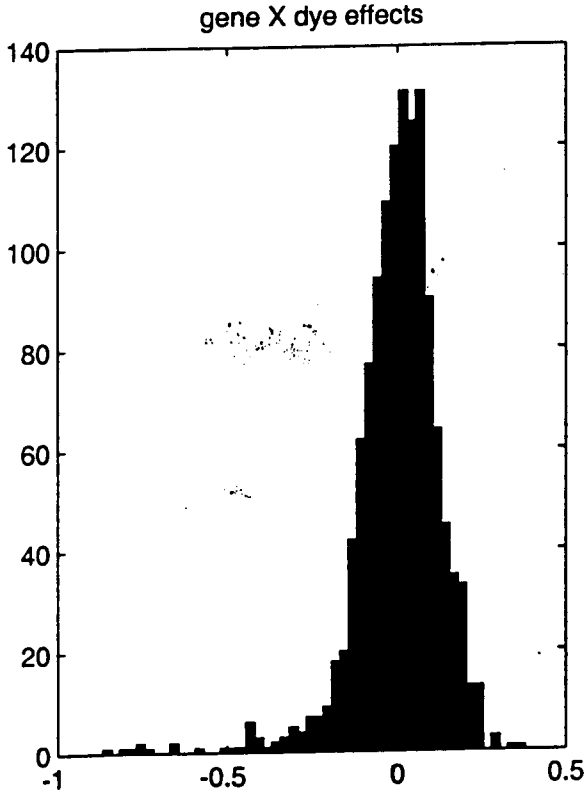


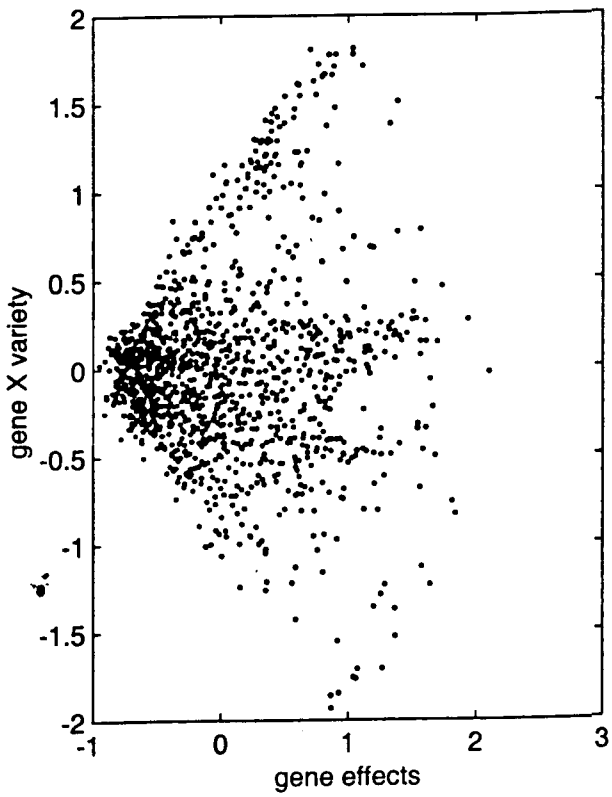


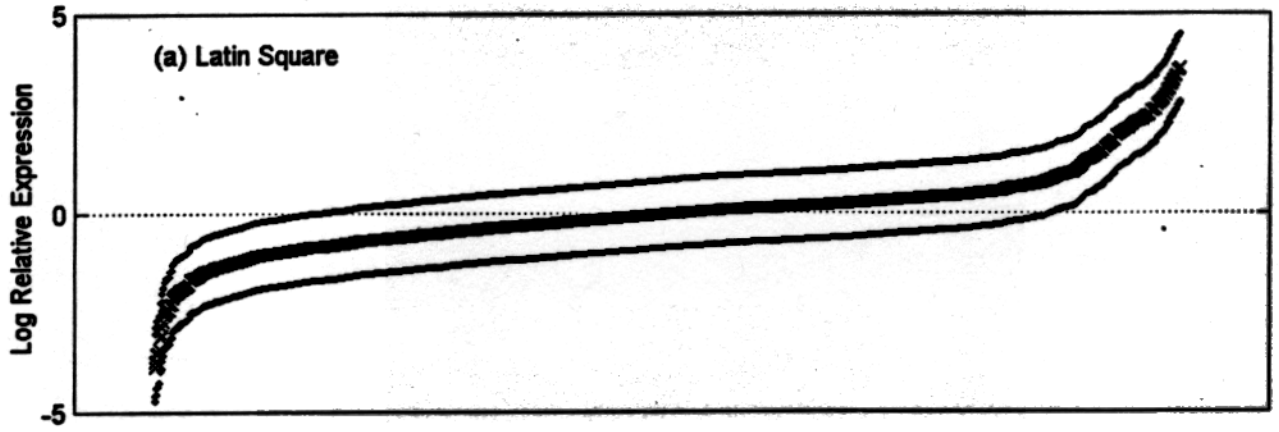
SYNTENI LATIN SQUARE EXPERIMENT

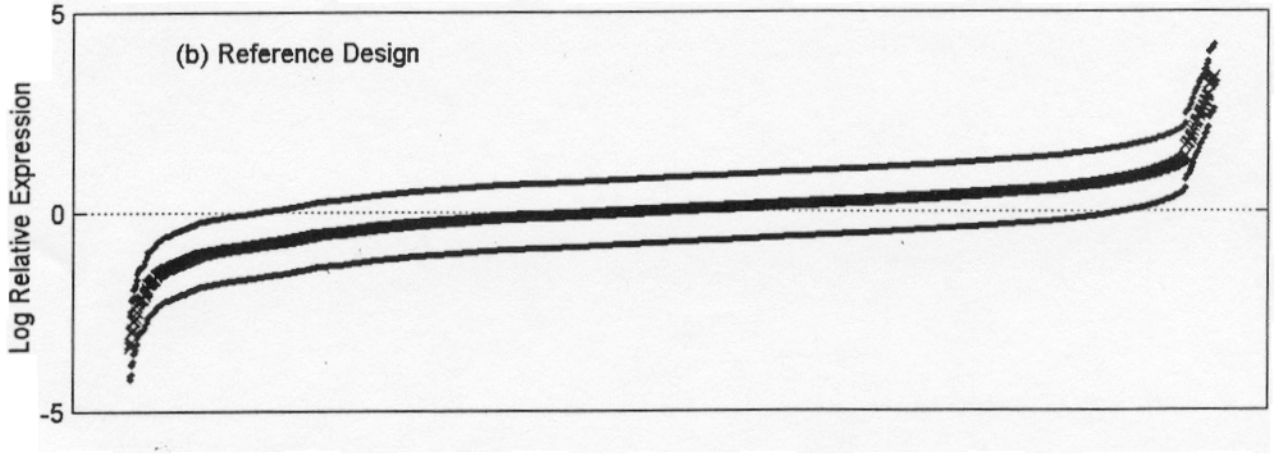


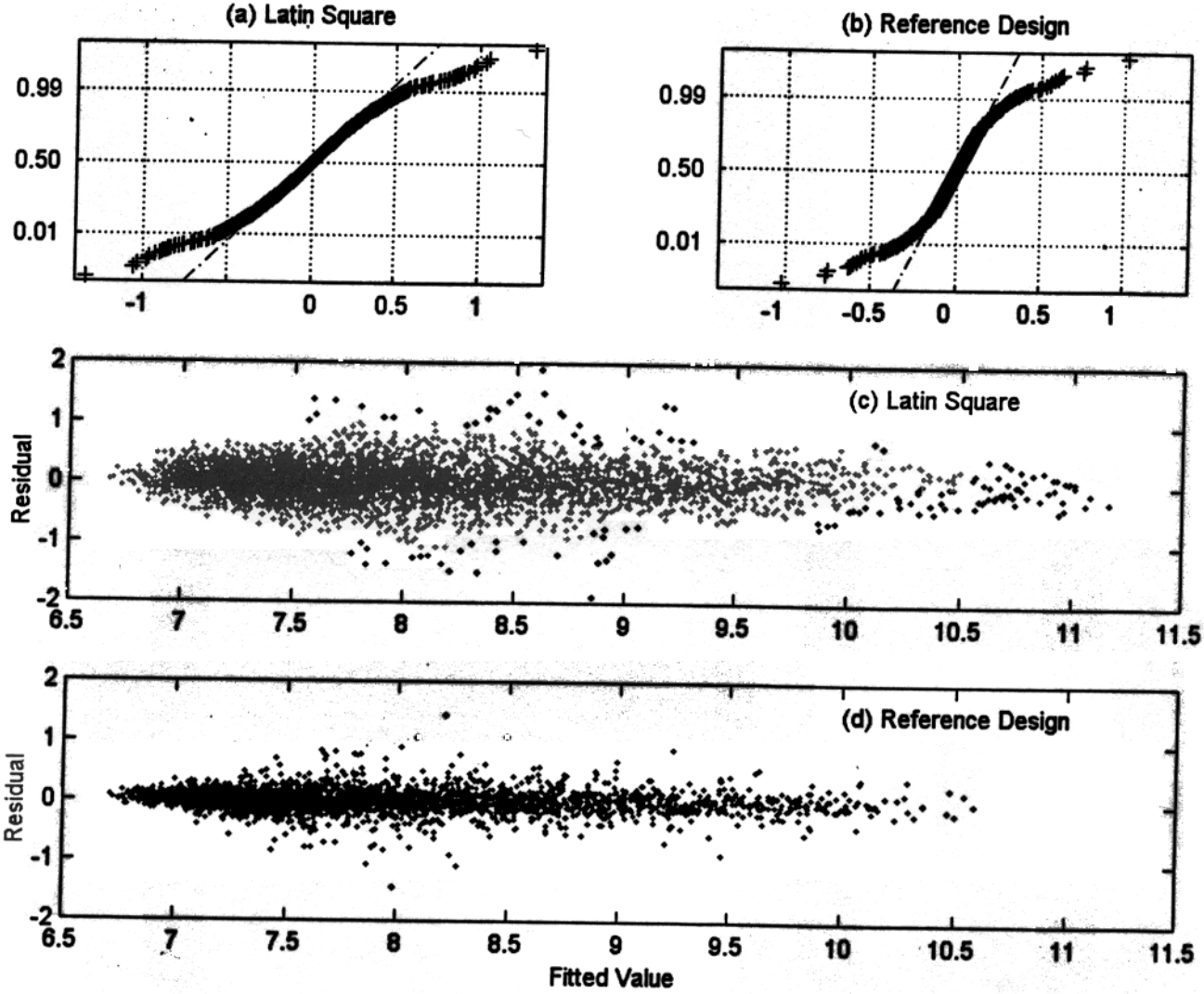


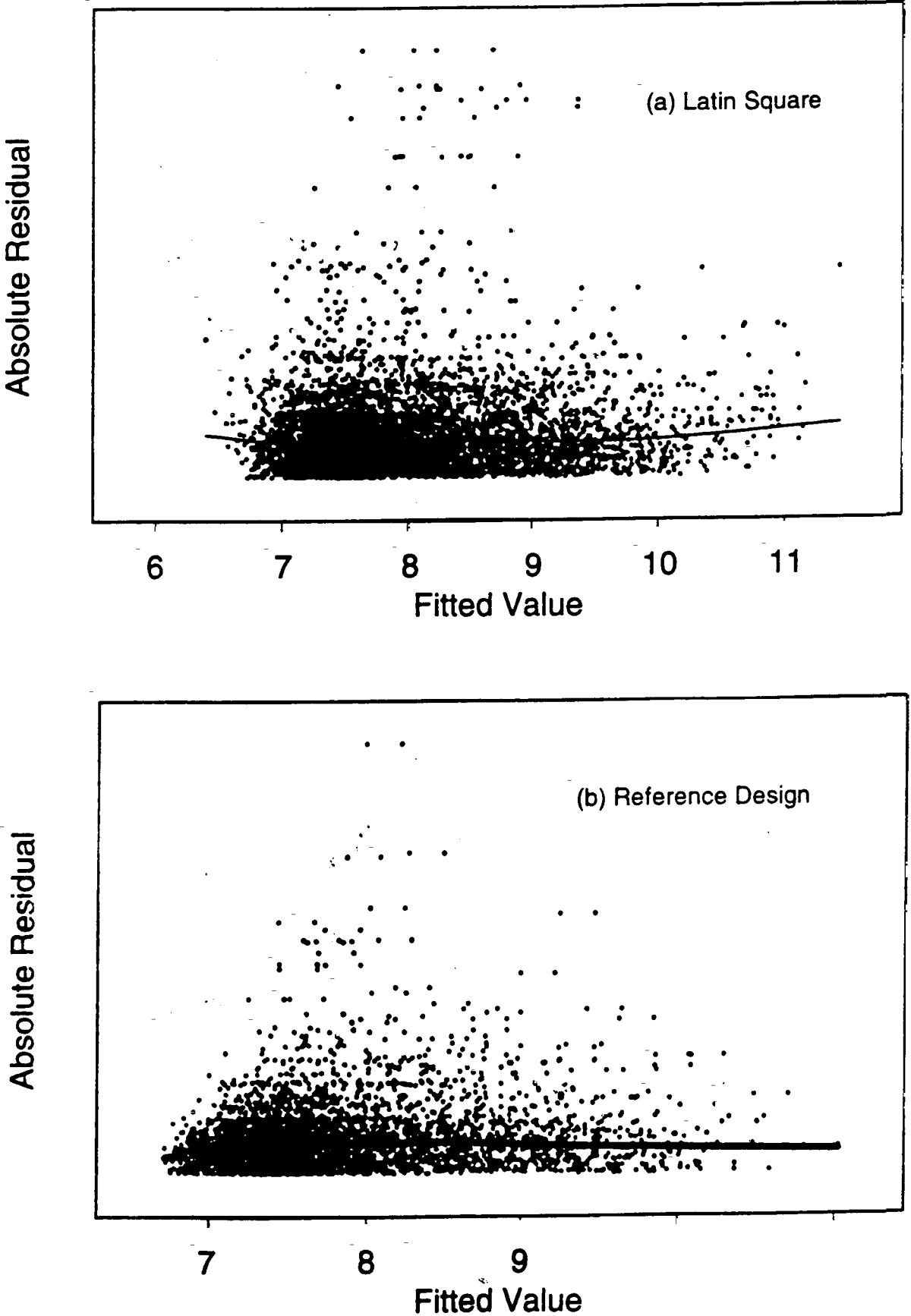





(a) Latin Square
Log Relative Expression
5
0
-5

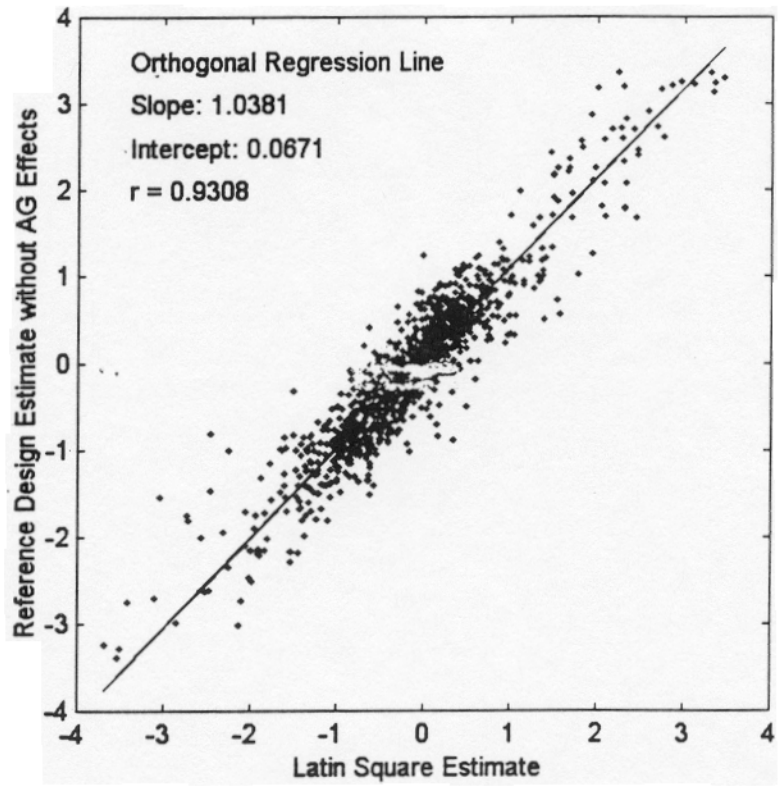

(b) Reference Design
Log Relative Expression
5
0
-5

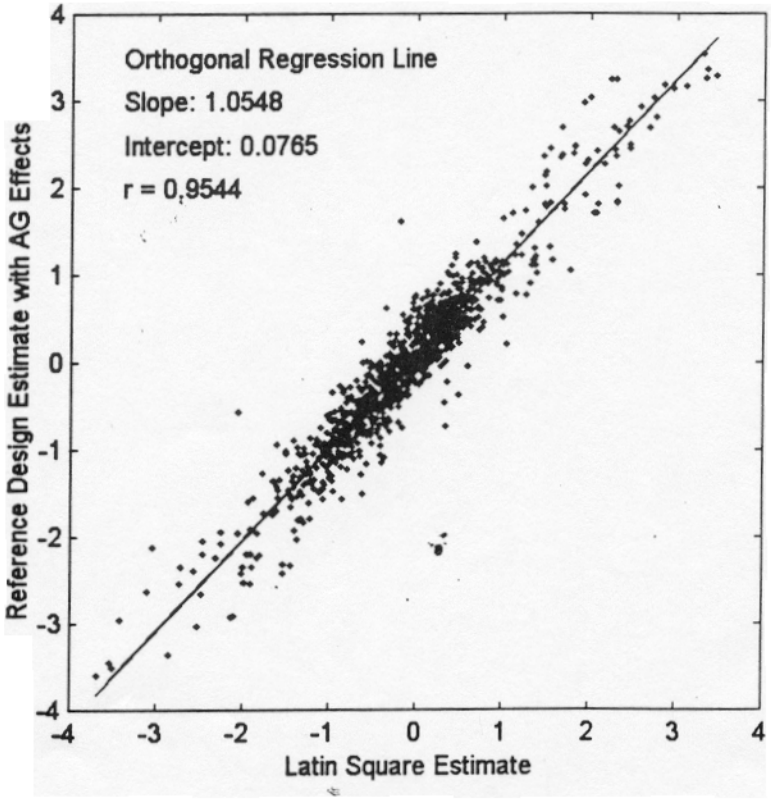

(a) Latin Square
0.99
0.50
0.01
-1
0
0.5
1
(b) Reference Design
0.99
0.50
0.01
-1
-0.5
0
0.5
1
(c) Latin Square
Residual
2
1
0
-1
-2
6.5
7
7.5
8
8.5
9
9.5
10
10.5
11
11.5
(d) Reference Design
Residual
2
1
0
-1
-2
6.5
7
7.5
8
8.5
9
9.5
10
10.5
11
11.5
Fitted Value

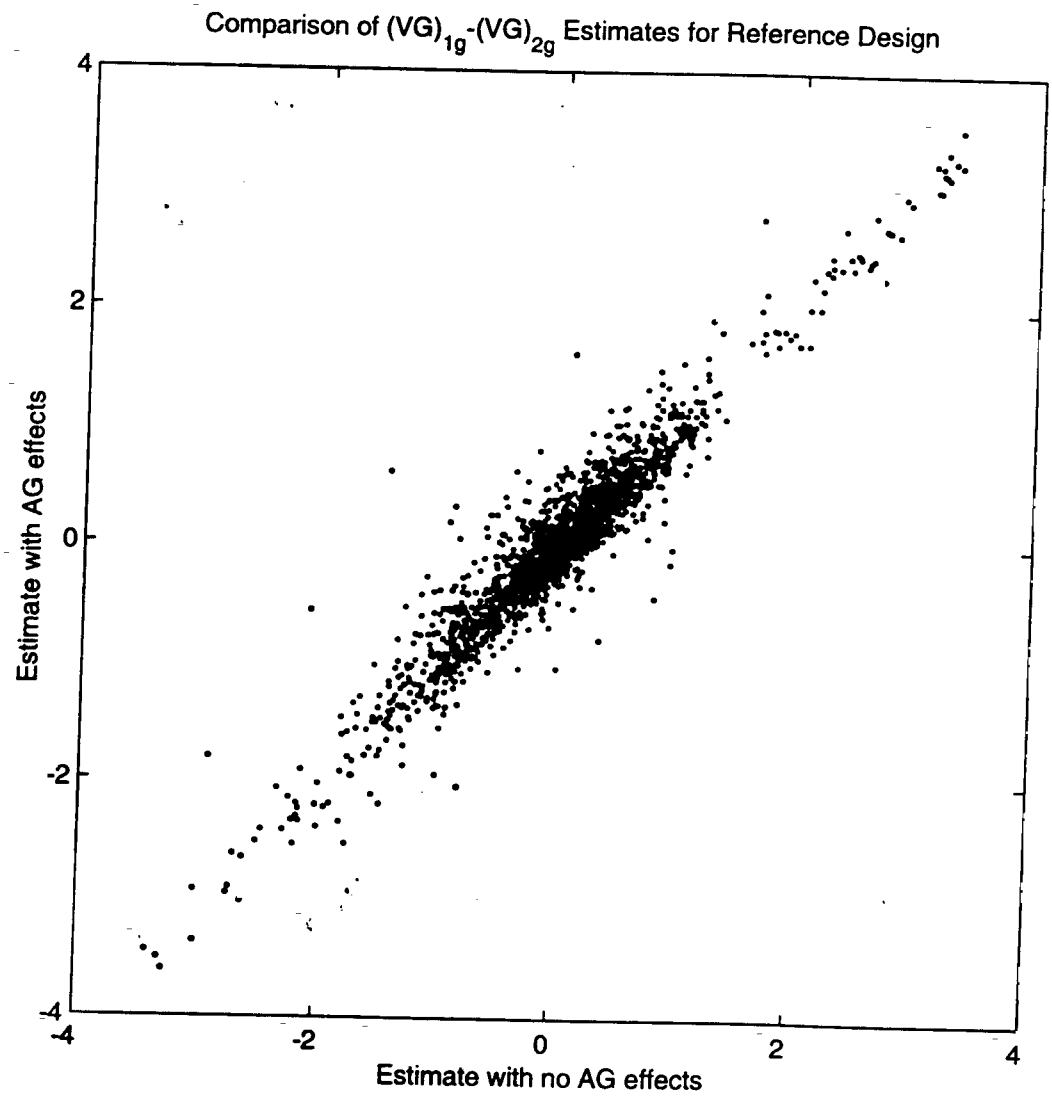

Absolute Residual
(a) Latin Square
6
7
8
9
10
11
Fitted Value
Absolute Residual
(b) Reference Design
7
8
9
Fitted Value

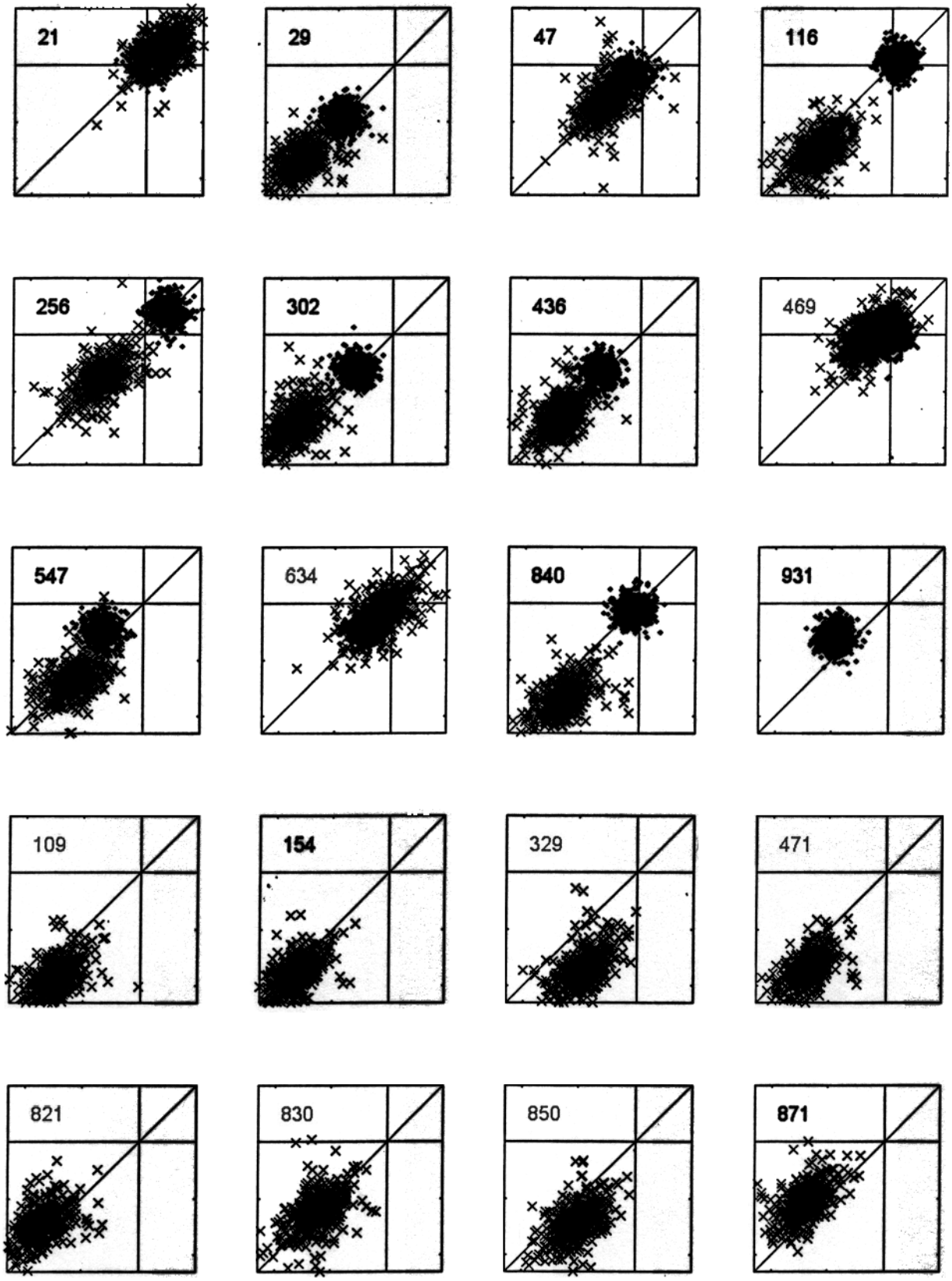
Concordance of the Liver Muscle differences

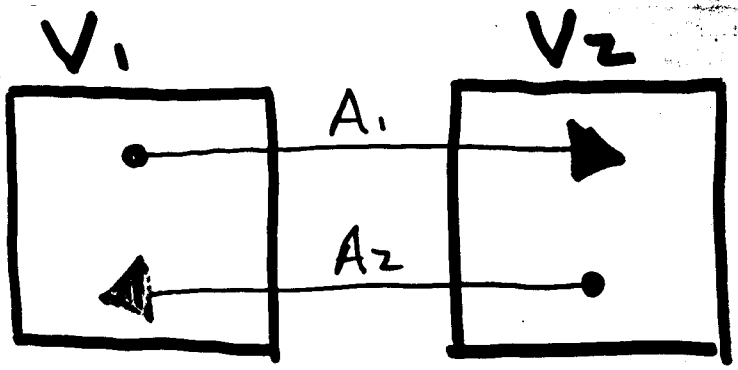
		Reference Design		
		L < M	NC	L > M
Latin Square Design	L < M	55	26	0
	NC	58	792	46
	L > M	0	6	84


Orthogonal Regression Line
Slope: 1.0381
Intercept: 0.0671
r = 0.9308
Reference Design Estimate without AG Effects
Latin Square Estimate
4
3
2
1
0
-1
-2
-3
-4
-4
-3
-2
-1
0
1
2
3
4


Orthogonal Regression Line
Slope: 1.0548
Intercept: 0.0765
r = 0,9544
Reference Design Estimate with AG Effects
Latin Square Estimate
4
3
2
1
0
-1
-2
-3
-4
-4
-3
-2
-1
0
1
2
3
4


Comparison of (VG)1g-(VG)2g Estimates for Reference Design
Estimate with AG effects
Estimate with no AG effects
4
2
0
-2
-4
-4
-2
0
2
4


21
29
47
116
256
302
436
469
547
634
840
931
109
154
329
471
821
830
850
871



The latin square is a **half-fraction** of the theoretical complete **factorial design**. It has a simple aliasing structure.

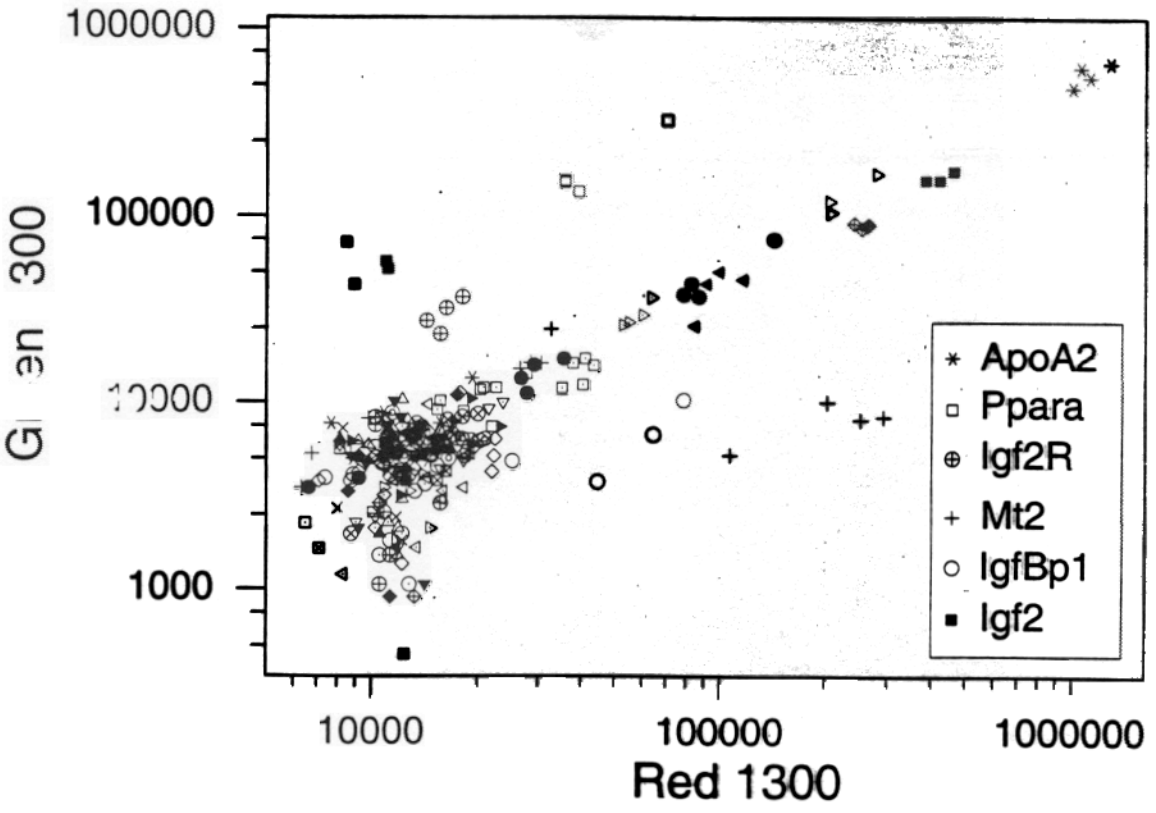
mean	~	ADV
A	~	DV
D	~	AV
V	~	AD
G	~	ADVG
VG	~	ADG
AG	~	DVG
DG	~	AVG

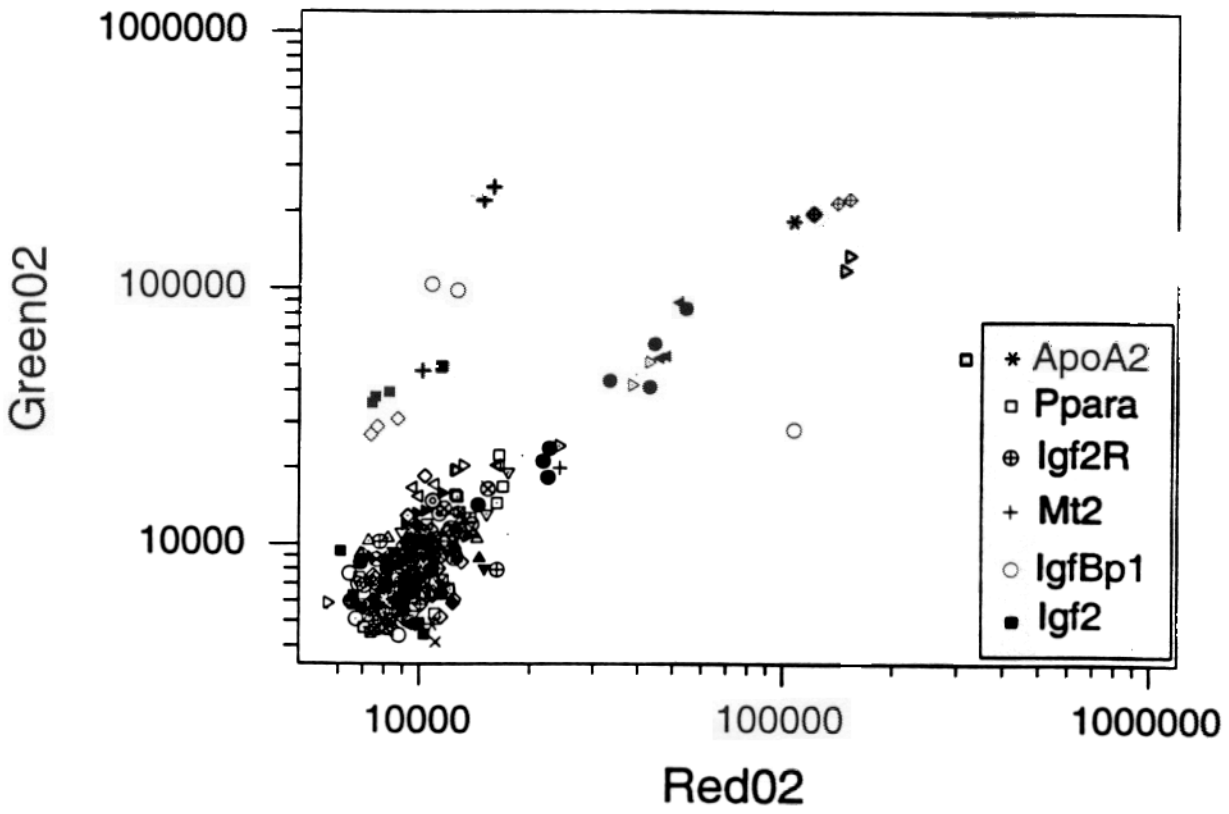
Corning Latin Square Experiment

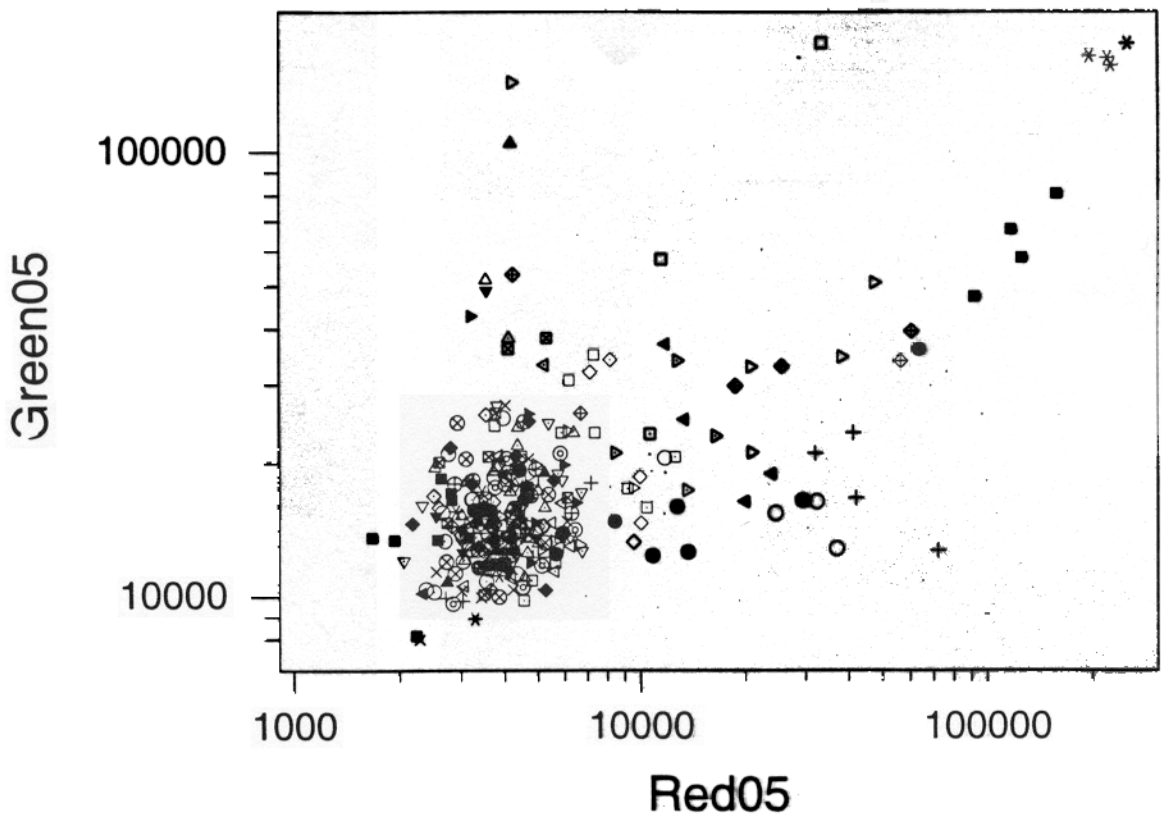
Source	SS	df	MS
Array	3.11	1	3.11
Dye	79.81	1	79.82
Variety	53.06	1	53.06
Gene	954.60	83	11.30
Variety x Gene	70.84	83	0.85
Spot x Gene	165.36	332	0.50
Dye x Gene	30.79	83	0.37
Residual	30.38	759	0.0400
Adj Total	1387.96	1343	

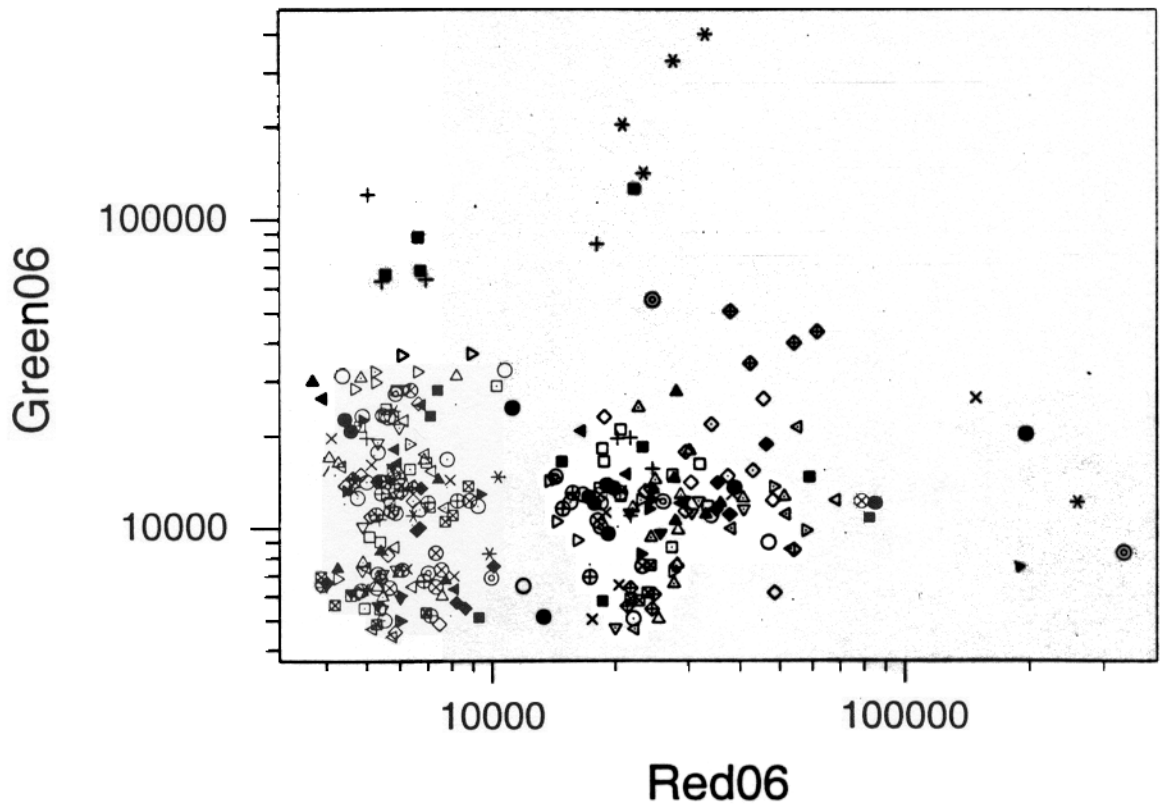
Remove Gene 98/Igf2:

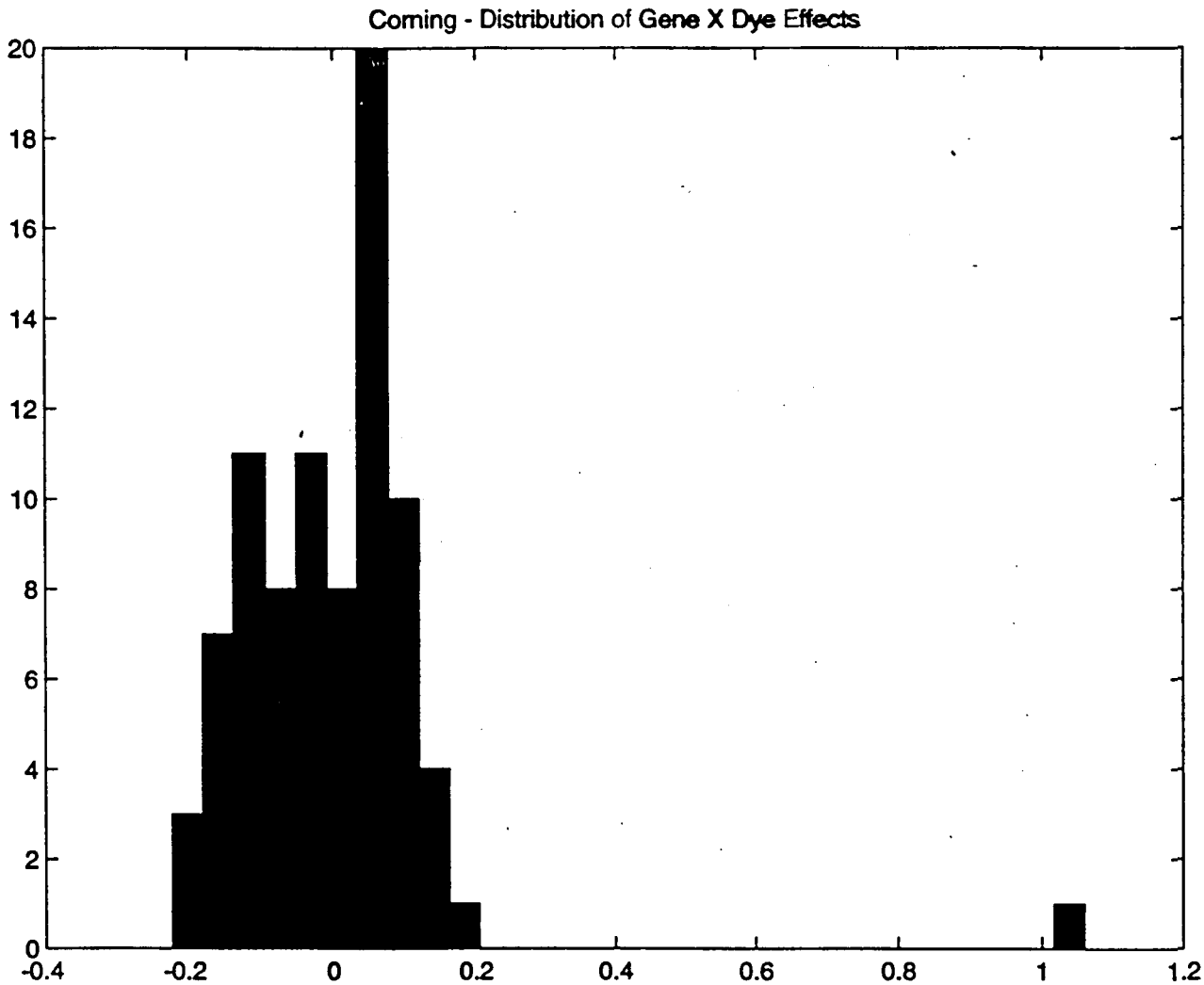
Source	SS	df	MS
Array	3.33	1	3.33
Dye	87.35	1	87.35
Variety	53.99	1	53.99
Gene	949.77	82	11.58
Variety x Gene	69.87	82	0.85
Spot x Gene	164.69	328	0.50
Dye x Gene	12.60	82	0.15
Residual	30.25	750	0.0403
Adj Total	1371.86	1327	

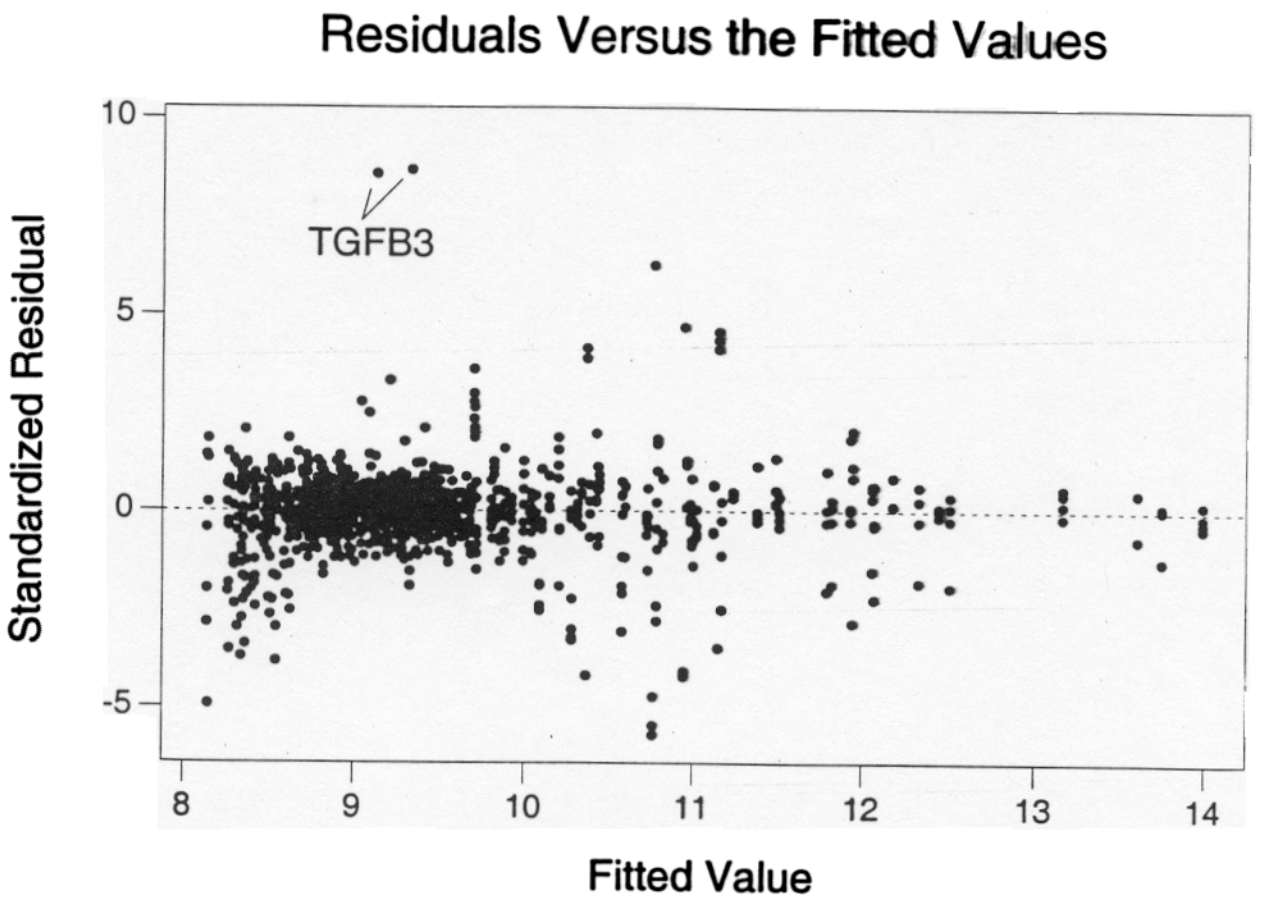

1000000
100000
10000
1000
Gr en 300
10000
100000
1000000
Red 1300
ApoA2
Ppara
Igf2R
Mt2
IgfBp1
Igf2

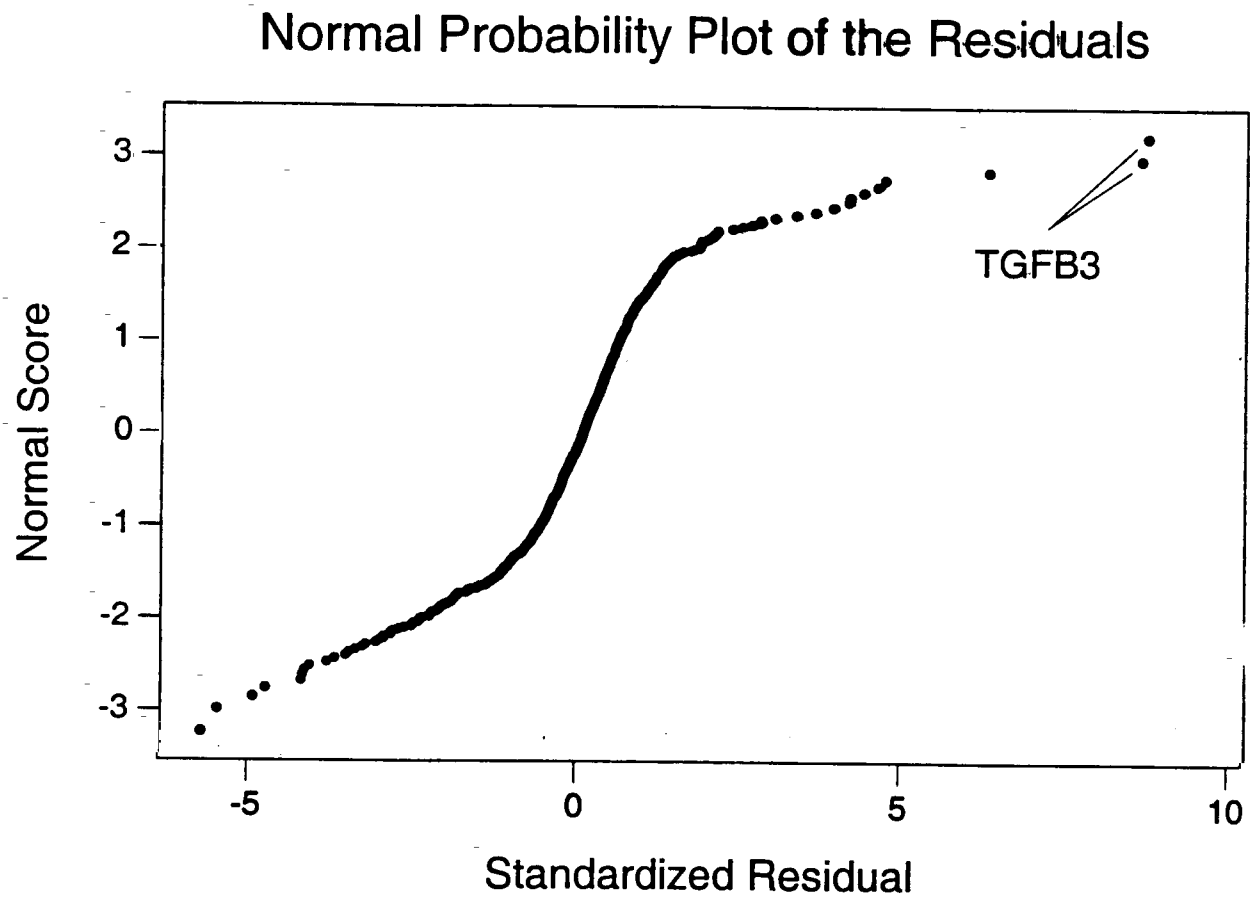

1000000
100000
10000
Green02
10000
100000
1000000
Red02
ApoA2
Ppara
Igf2R
Mt2
IgfBp1
Igf2

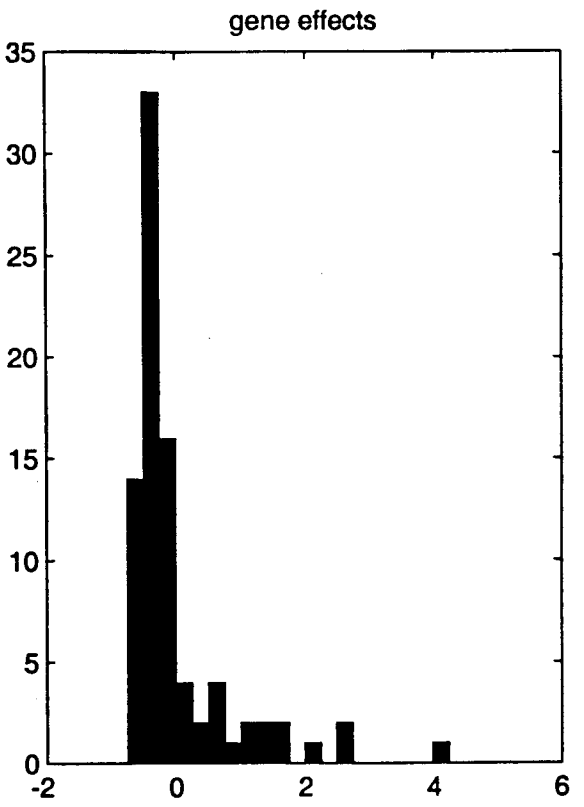

100000
10000
Green05
1000
10000
100000
Red05

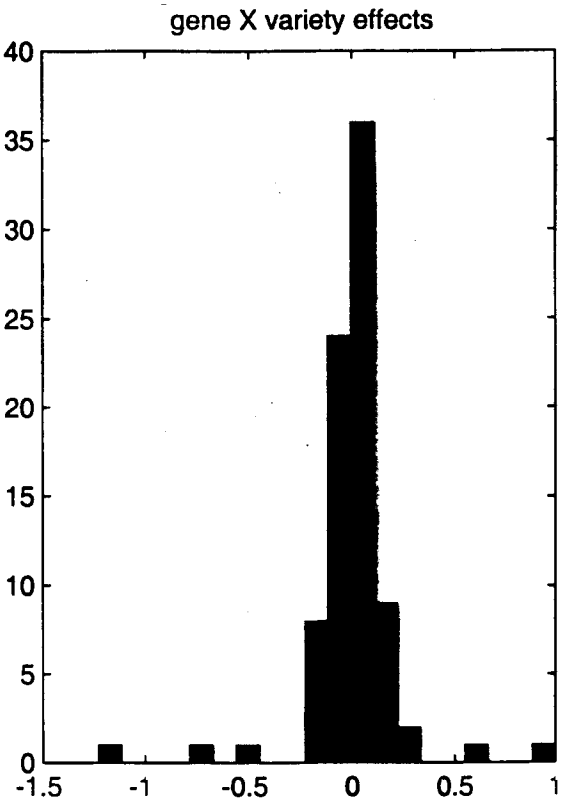

100000
10000
Green06
10000
100000
Red06

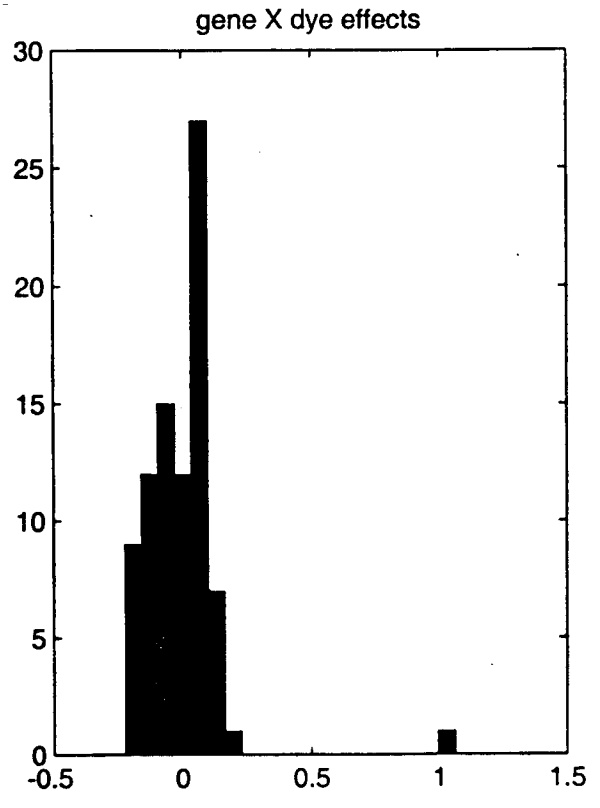

Corning - Distribution of Gene X Dye Effects
20
18
16
14
12
10
8
6
4
2
0
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
1.2

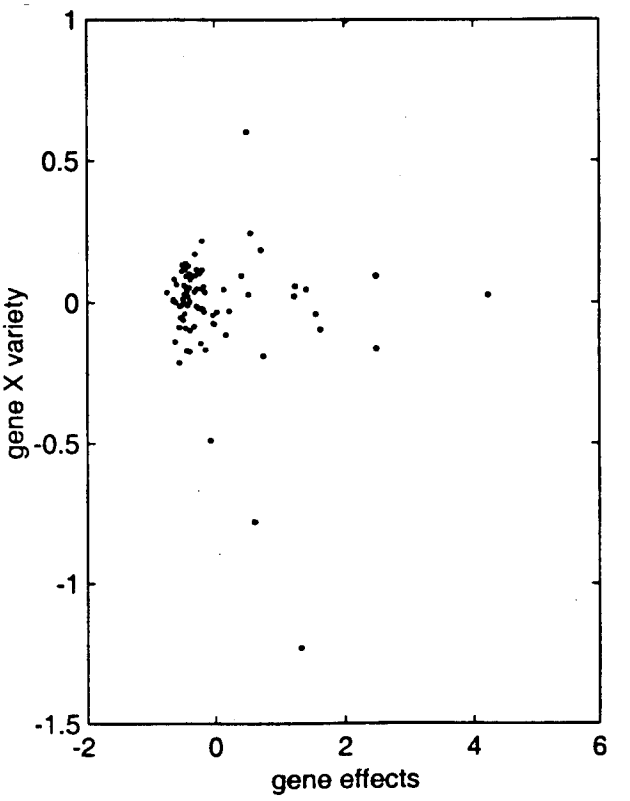

Residuals Versus the Fitted Values
Standardized Residual
Fitted Value
TGFB3
10
5
0
-5
8
9
10
11
12
13
14


Normal Probability Plot of the Residuals
Normal Score
3
2
1
0
-1
-2
-3
TGFB3
-5
0
5
10
Standardized Residual


gene effects
35
30
25
20
15
10
5
0
-2
0
2
4
6


gene X variety effects
40
35
30
25
20
15
10
5
0
-1.5
-1
-0.5
0
0.5
1


gene X dye effects
30
25
20
15
10
5
0
-0.5
0
0.5
1
1.5


gene X variety
1
0.5
0
-0.5
-1
-1.5
-2
0
2
4
6
gene effects

The Question: Is gene g **differentially** expressed across varieties k_1 and k_2?

$$\text{Is } (VG)_{k_1 g} = (VG)_{k_2 g}?$$

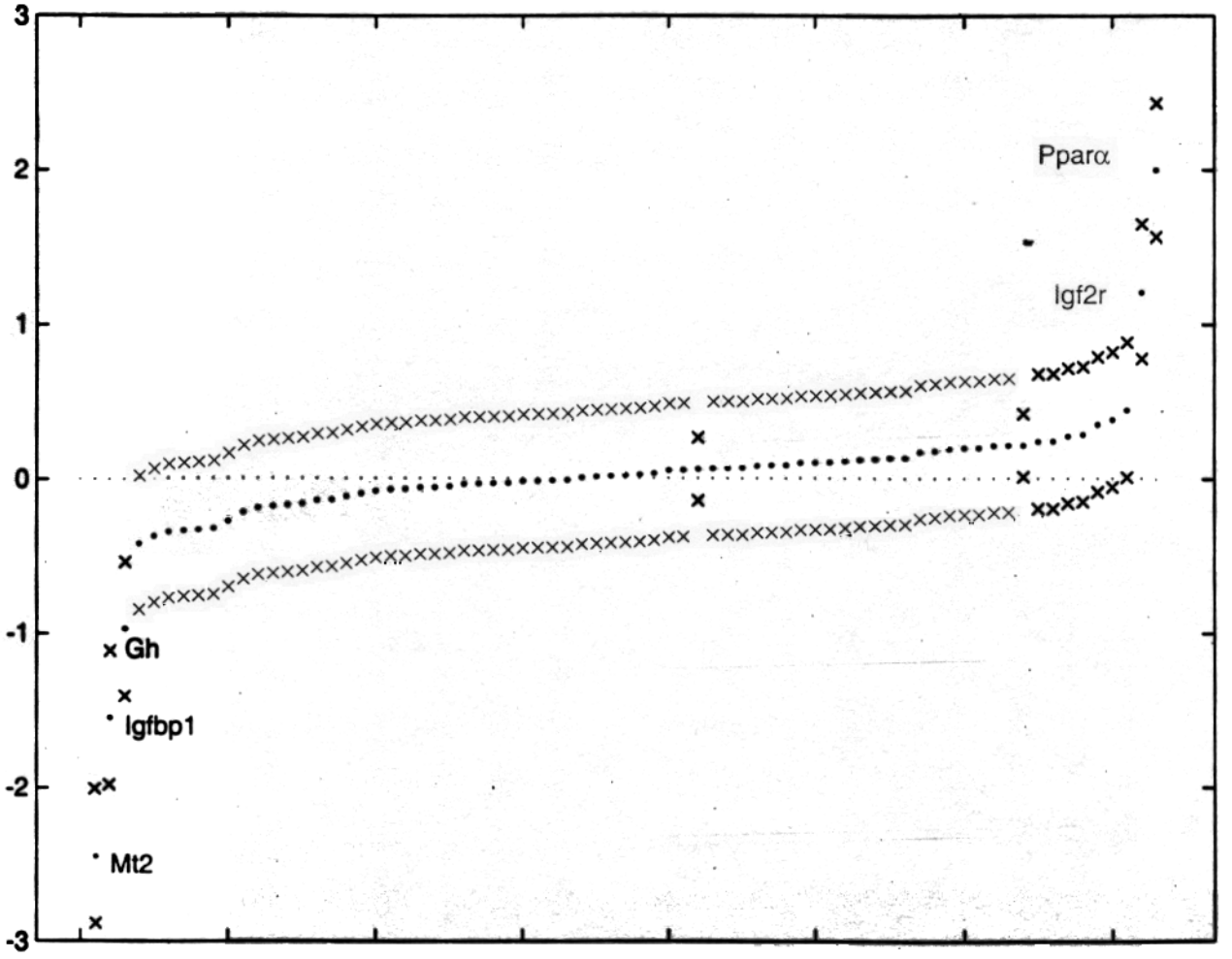
Parametric Bootstrap

Produce bootstrap datasets $\log(y_{ijkgr})^* =$

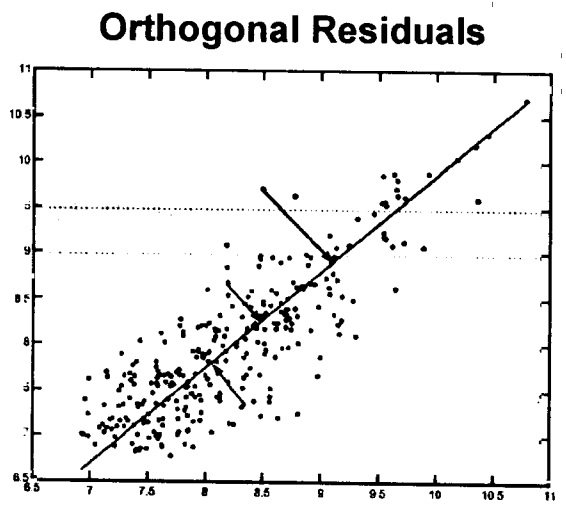
$$\hat{\mu} + \hat{A}_i + \hat{D}_j + \hat{V}_k + \hat{G}_g + (\widehat{VG})_{kg} + \epsilon^*_{ijkgr}.$$

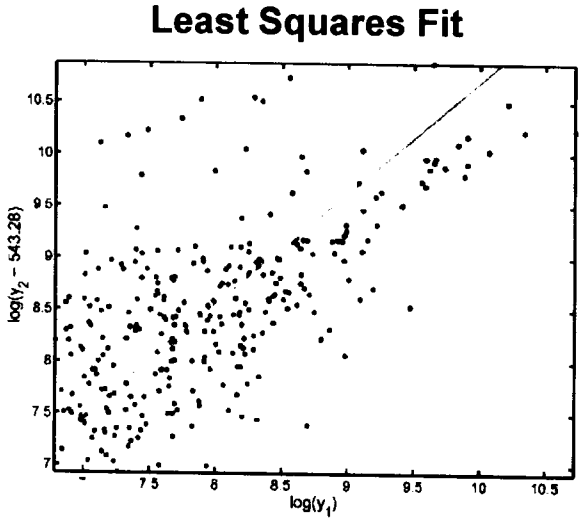
The ϵ^*_{ijkgr} are drawn **independently** from the empirical distribution of **residuals from** the original fit of the model.

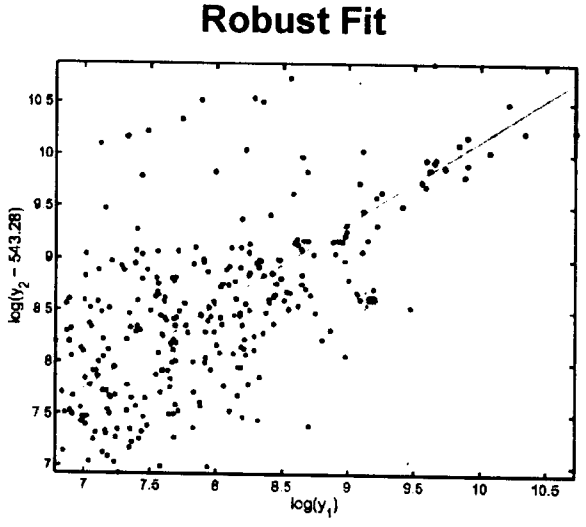
Form 99% confidence intervals using the percentiles of the bootstrap distribution of $(\widehat{VG})^*_{Tg_0}$ $(\widehat{VG})^*_{Cg_0}$ for each gene.


3
2
1
0
-1
-2
-3
Pparα
Igf2r
Gh
Igfbp1
Mt2

Fitting a Regression Line to the Data



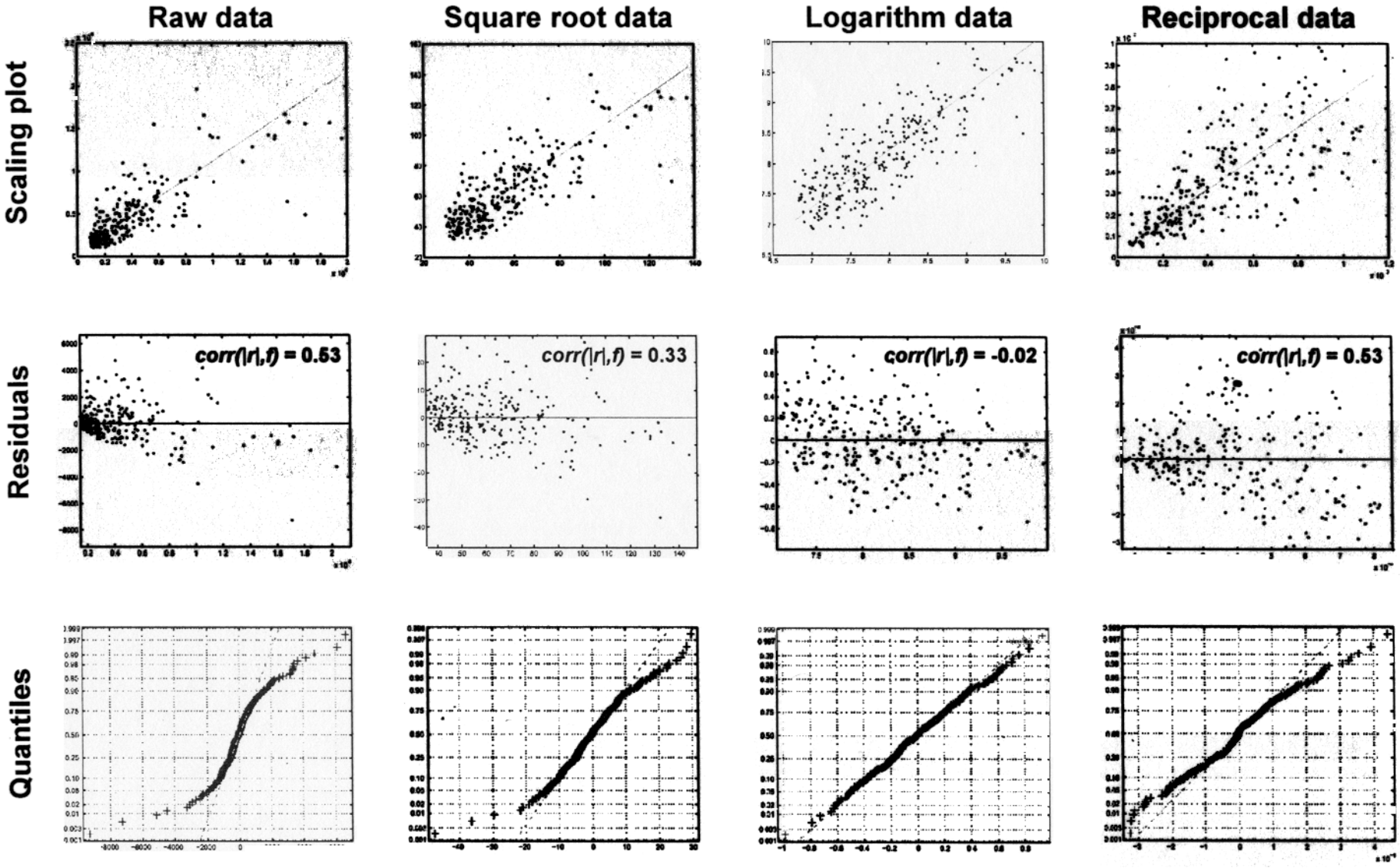




Orthogonal residuals provide a symmetric treatment of the two fluorescent intensities y_1 and y_2.

Robust regression reduces the influence of differentially expressed genes of the Fitted line.

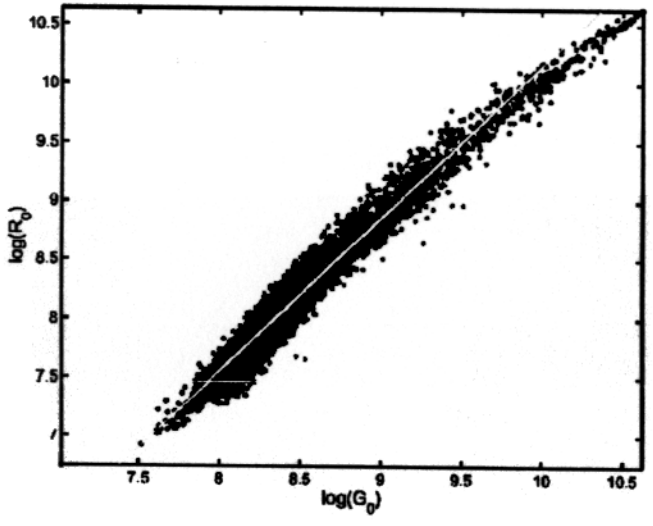
Selection of the Scaling Function:
Placenta-Placenta Self Comparison on GEM Array

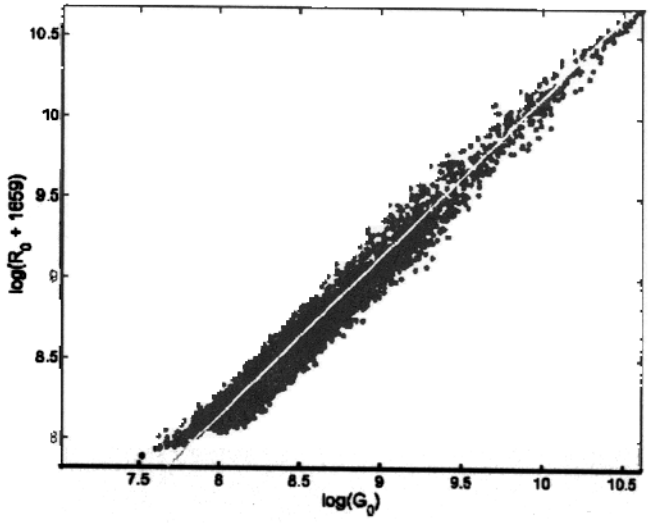


Conclusion: **The residuals from a logarithm scaling transformation are approximately normal and have minimal correlation with fitted values.**

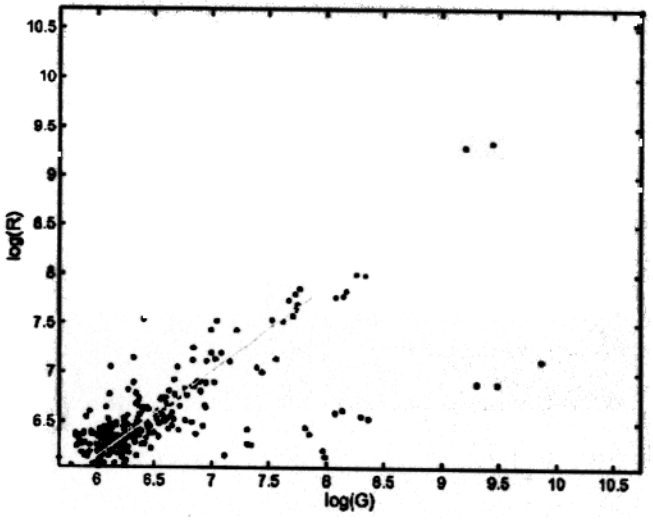
The Shifted Logarithm Scaling Function

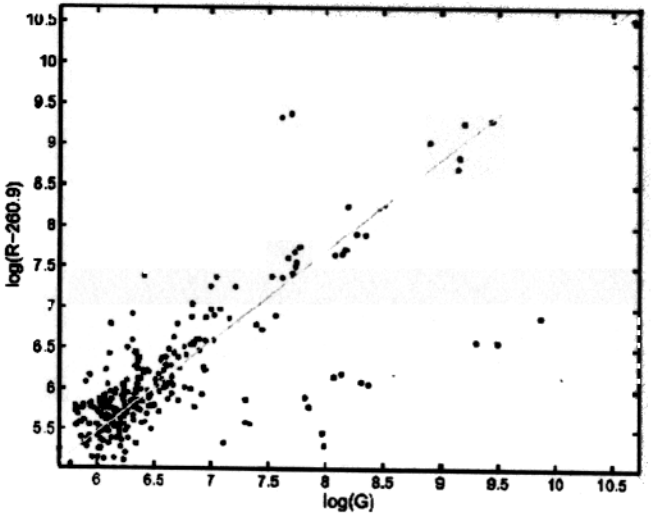
Stanford data





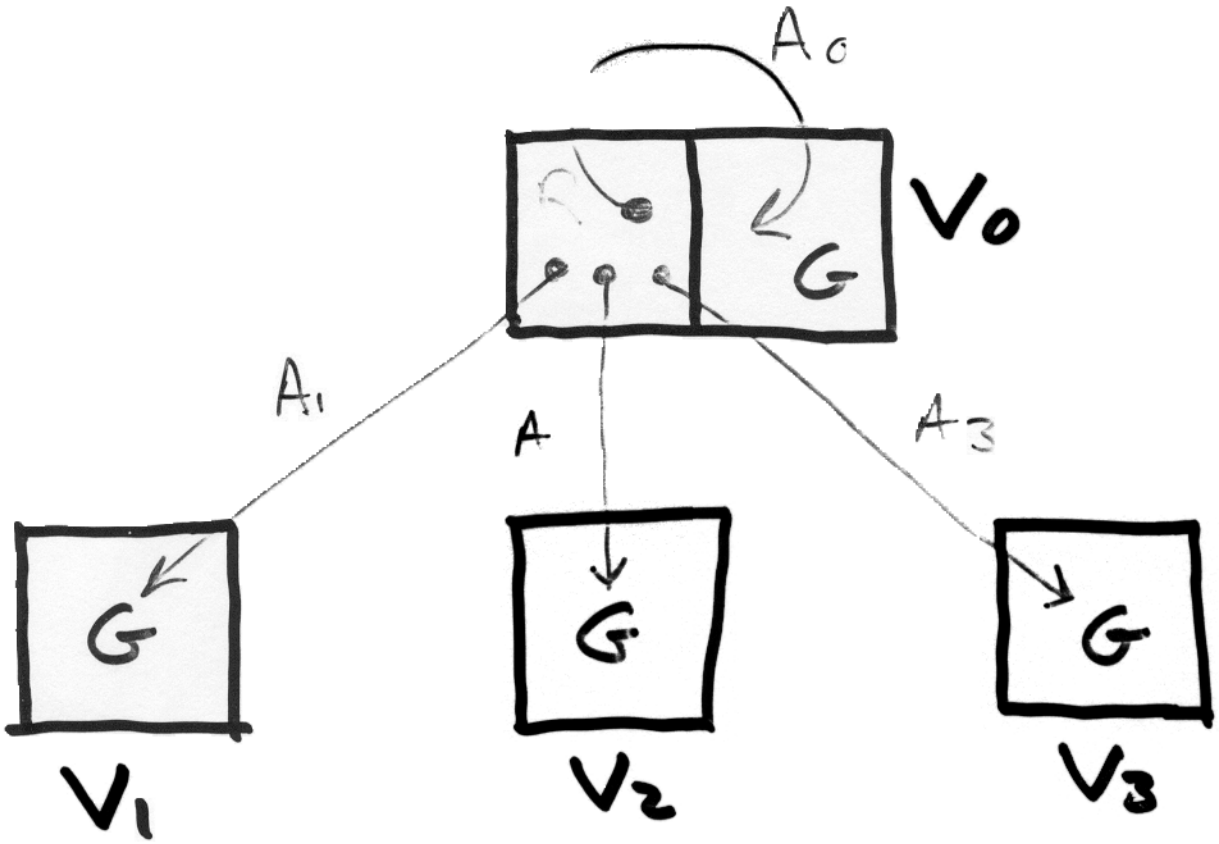
Corning data





Conclusion: **Shifting the raw measurement improves linearity on the logarithm scale.**

Augmented ReFerence Design

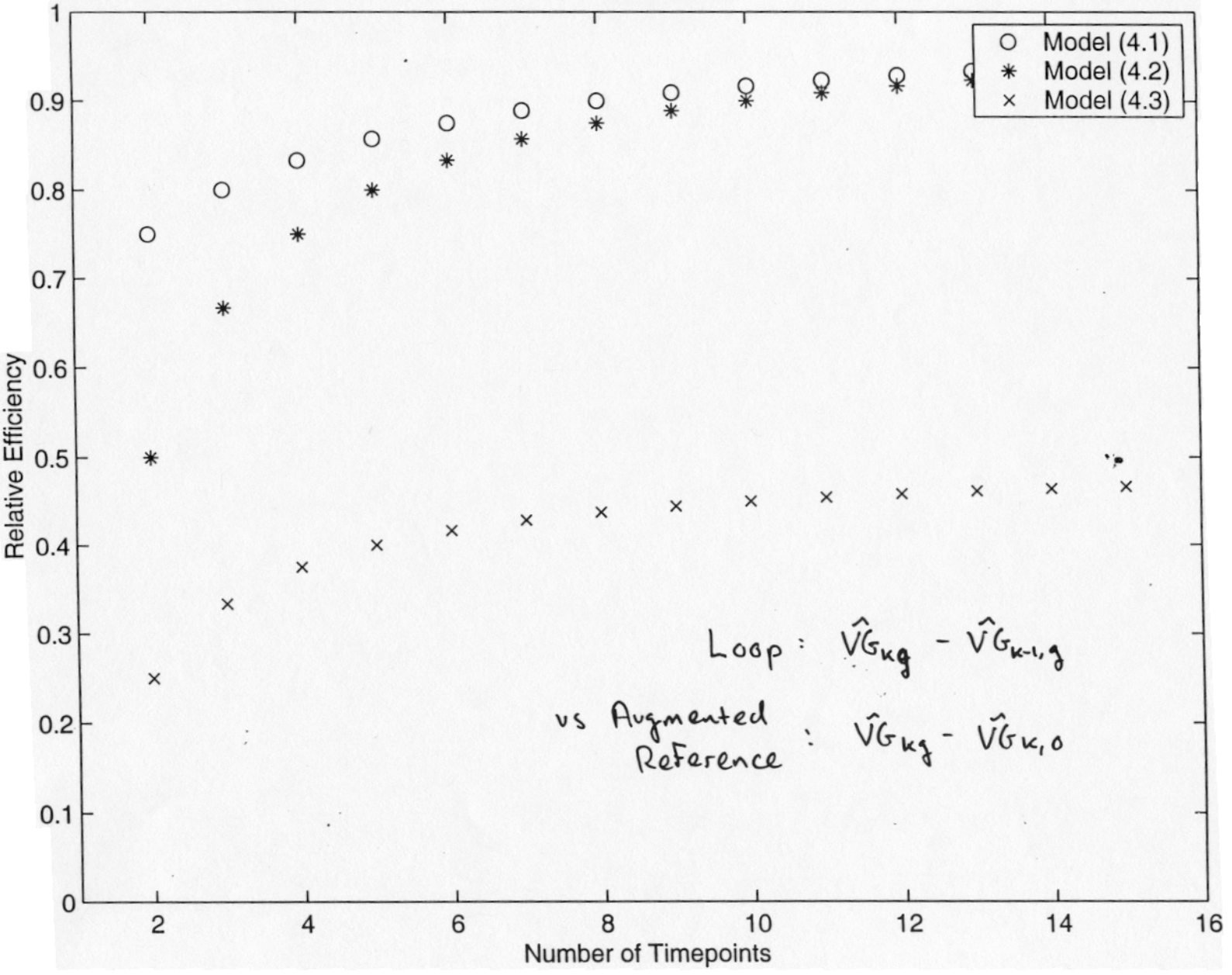


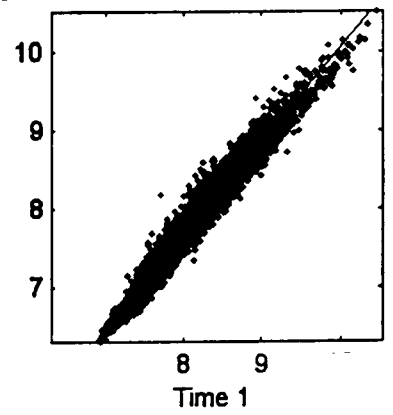
Residual dF

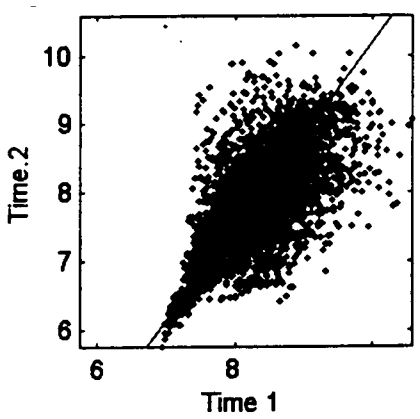
Unbalanced ⇒ partial confounding of VG DG

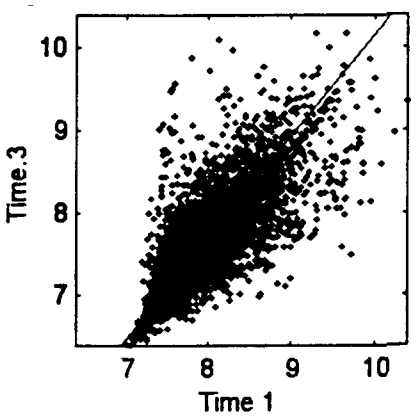
Residuals are all From one array

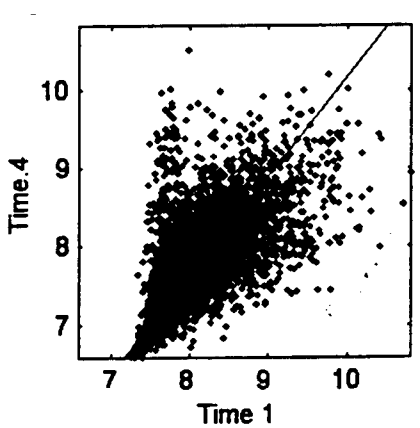
Figure 4

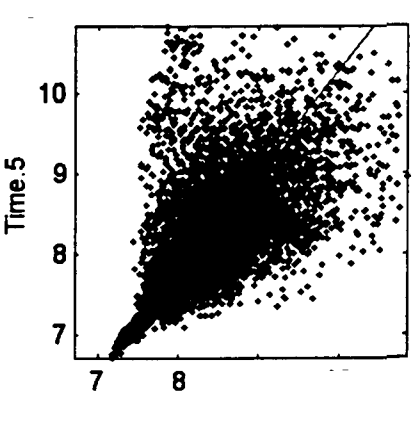

Model (4.1)
Model (4.2)
Model (4.3)
Relative Efficiency
Number of Timepoints
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
2
4
6
8
10
12
14
16
Loop: $\hat{V}G_{kg} - \hat{V}G_{k-1,g}$
vs Augmented Reference: $\hat{V}G_{kg} - \hat{V}G_{k,0}$

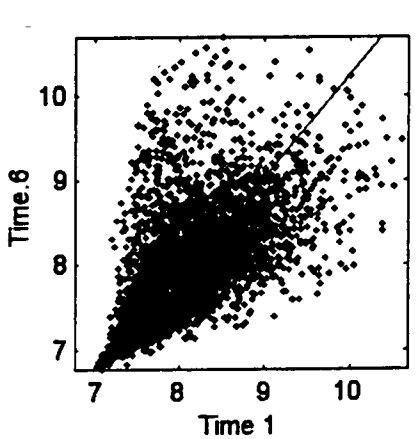

10
9
8
7
8
9
Time 1

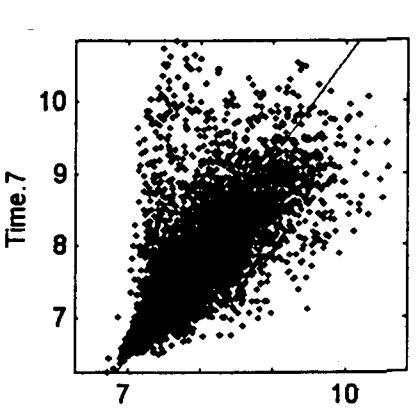

Time.2
10
9
8
7
6
6
8
Time 1


Time.3
10
9
8
7
7
8
9
10
Time 1


Time.4
10
9
8
7
7
8
9
10
Time 1


Time.5
10
9
8
7
7
8


Time.6
10
9
8
7
7
8
9
10
Time 1

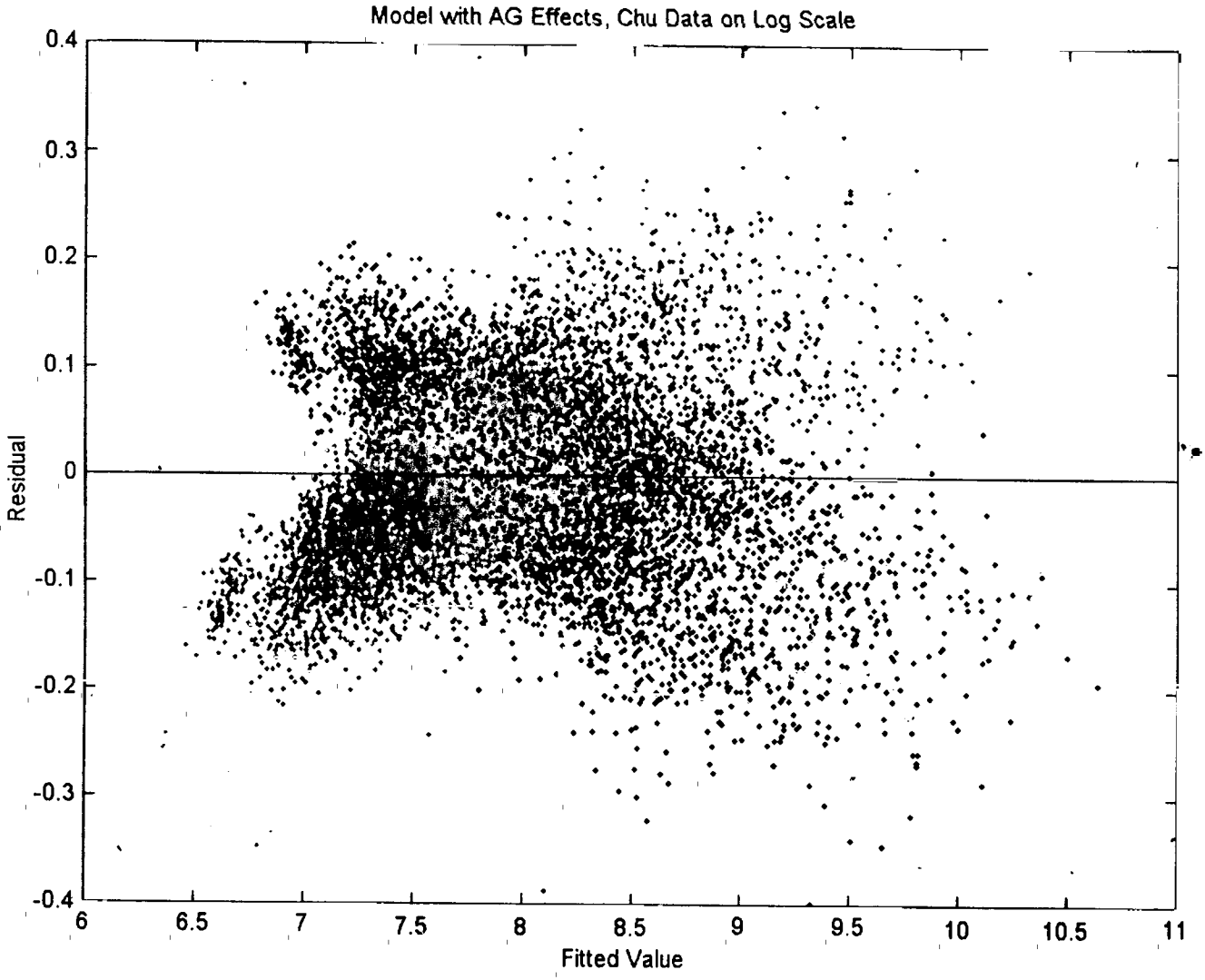

Time.7
10
9
8
7
7
10

Chu, et al Analyses

Source	SS	df	MS
Array,Dye,Variety	4871.16	13	374.70
Gene	18992.28	6117	3.10
→ VG	8250.36	36702	0.22
Residual	2743.69	42819	0.0641
Adj Total	34857.50	85651	

Source	SS	df	MS
Array,Dye,Variety	4871.16	13	374.70
Gene	18992.28	6117	3.10
→ VG,AG	10909.11	73404	0.15
Residual	84.94	6117	0.0139
Adj Total	34857.50	85651	

Source	SS	df	MS
Array,Dye,Variety	4871.16	13	374.70
Gene	18992.28	6117	3.10
→ VG,DG	8596.02	48936	0.18
Residual	2398.03	36702	0.0653
Adj Total	34857.50	85651	

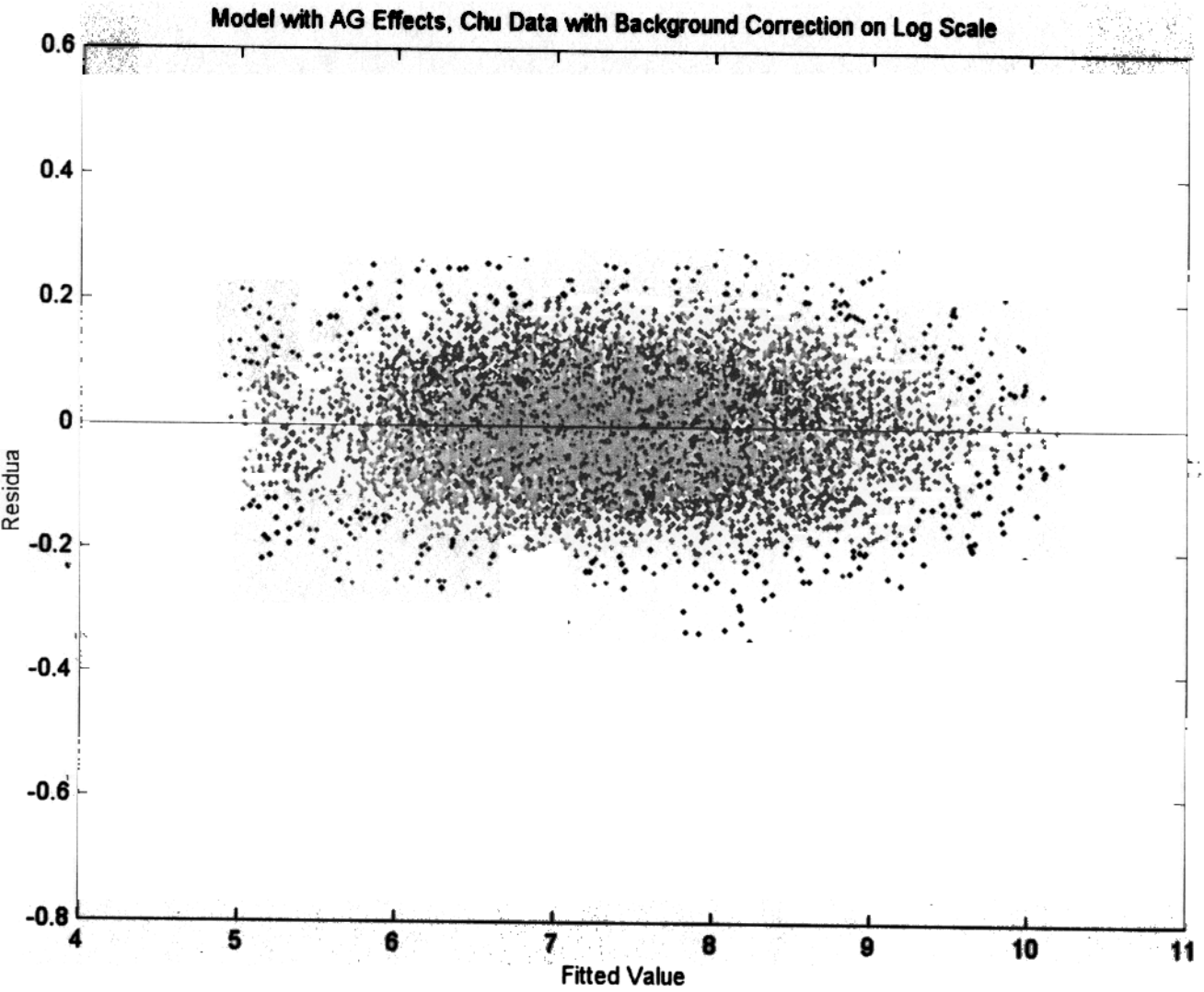

Model with AG Effects, Chu Data on Log Scale
Residual
Fitted Value
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
6
6.5
7
7.5
8
8.5
9
9.5
10
10.5
11

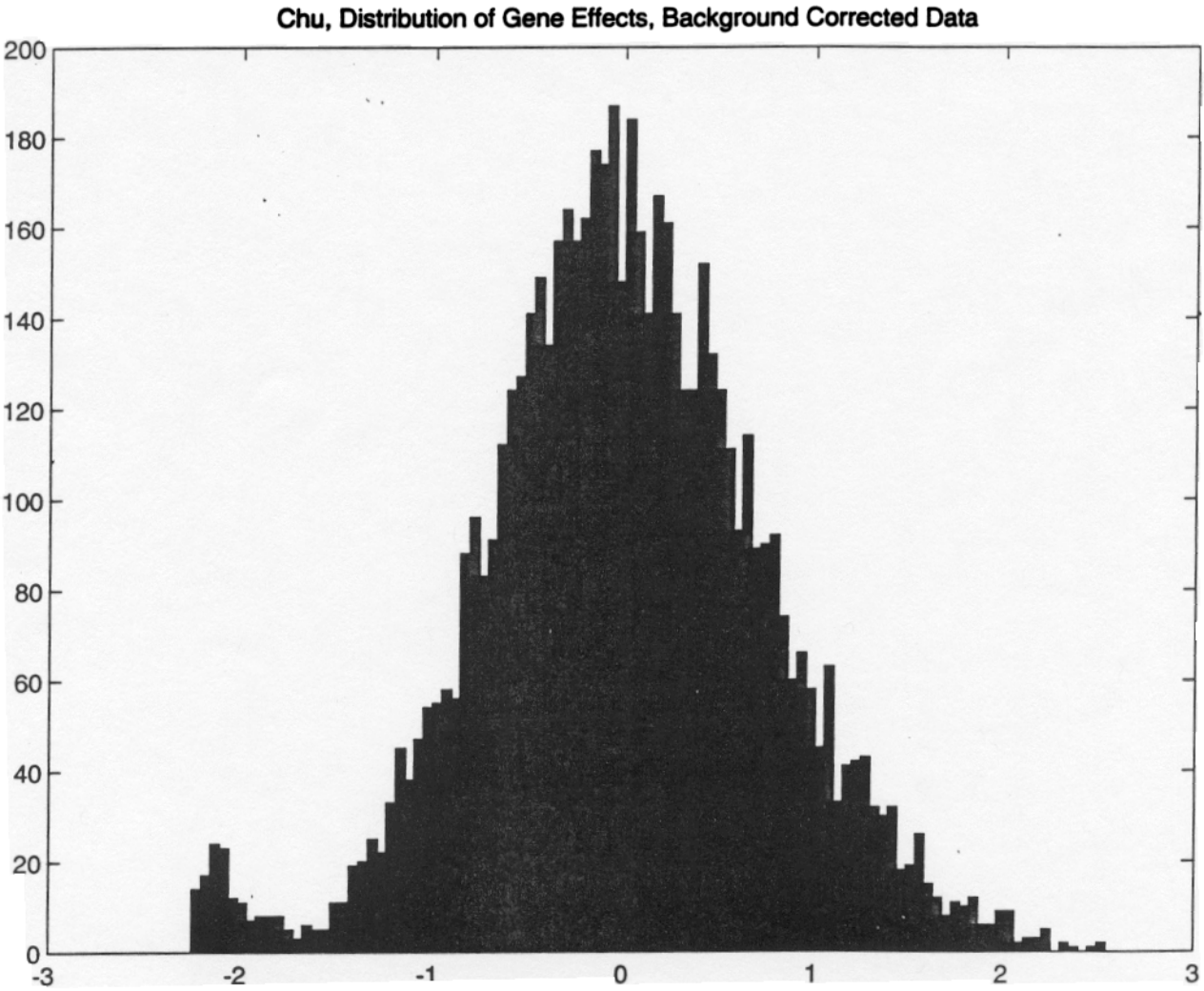
Chu Data, Background Corrected

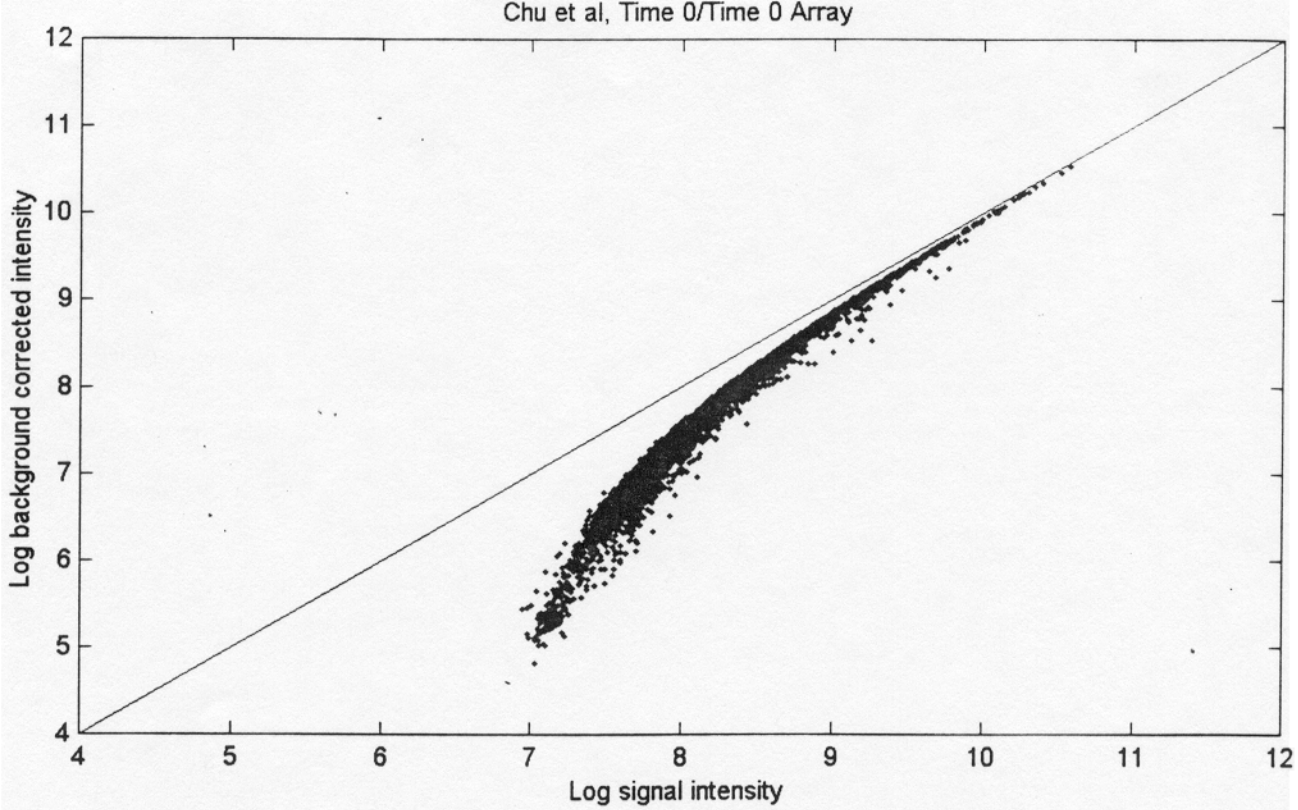
Source	SS	df	MS
Array,Dye,Variety	6896.24	13	530.48
Gene	48329.71	6117	7.90
VG	16681.88	36702	0.45
Residual	6314.46	42819	0.1475
Adj Total	78222.28	85651	

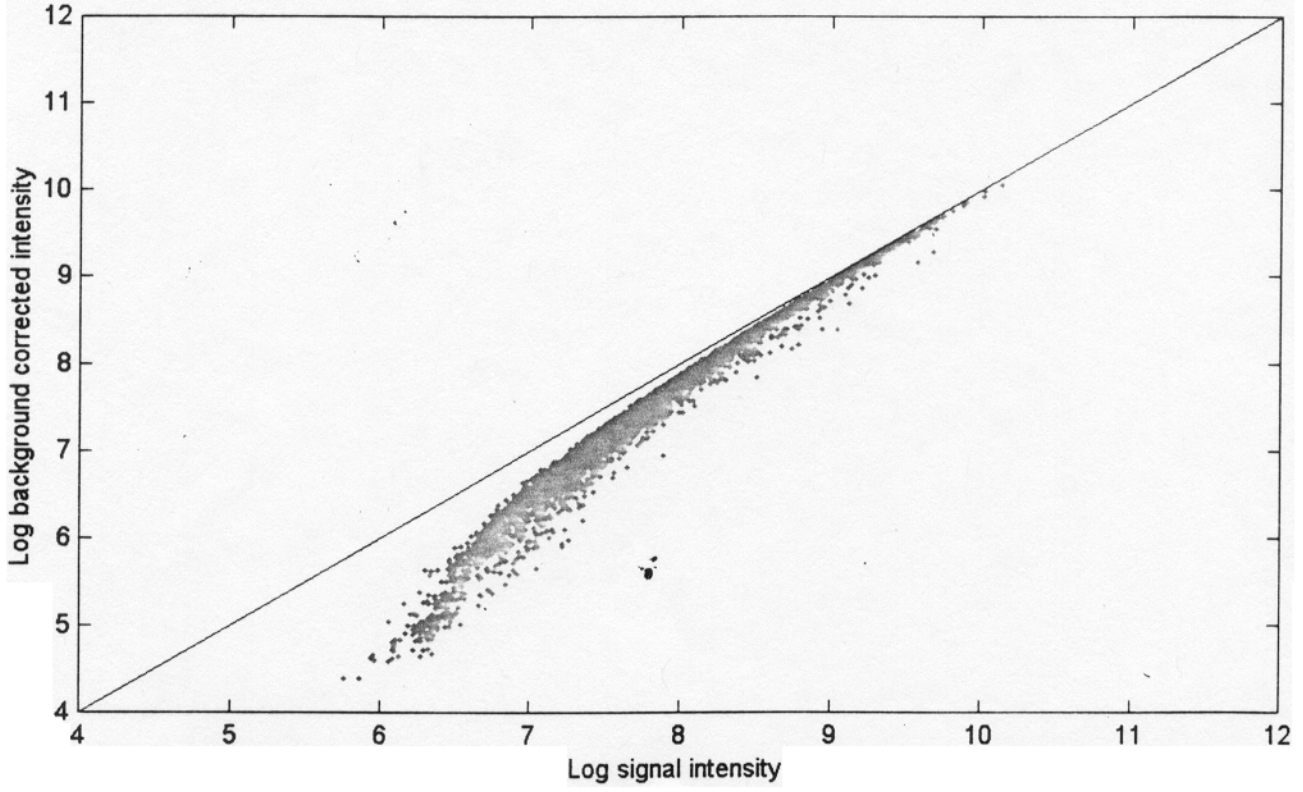
Source	SS	df	MS
Array,Dye,Variety	6896.24	13	530.48
Gene	48329.71	6117	7.90
VG,AG	22907.16	73404	0.31
Residual	89.18	6117	0.0146
Adj Total	78222.28	85651	

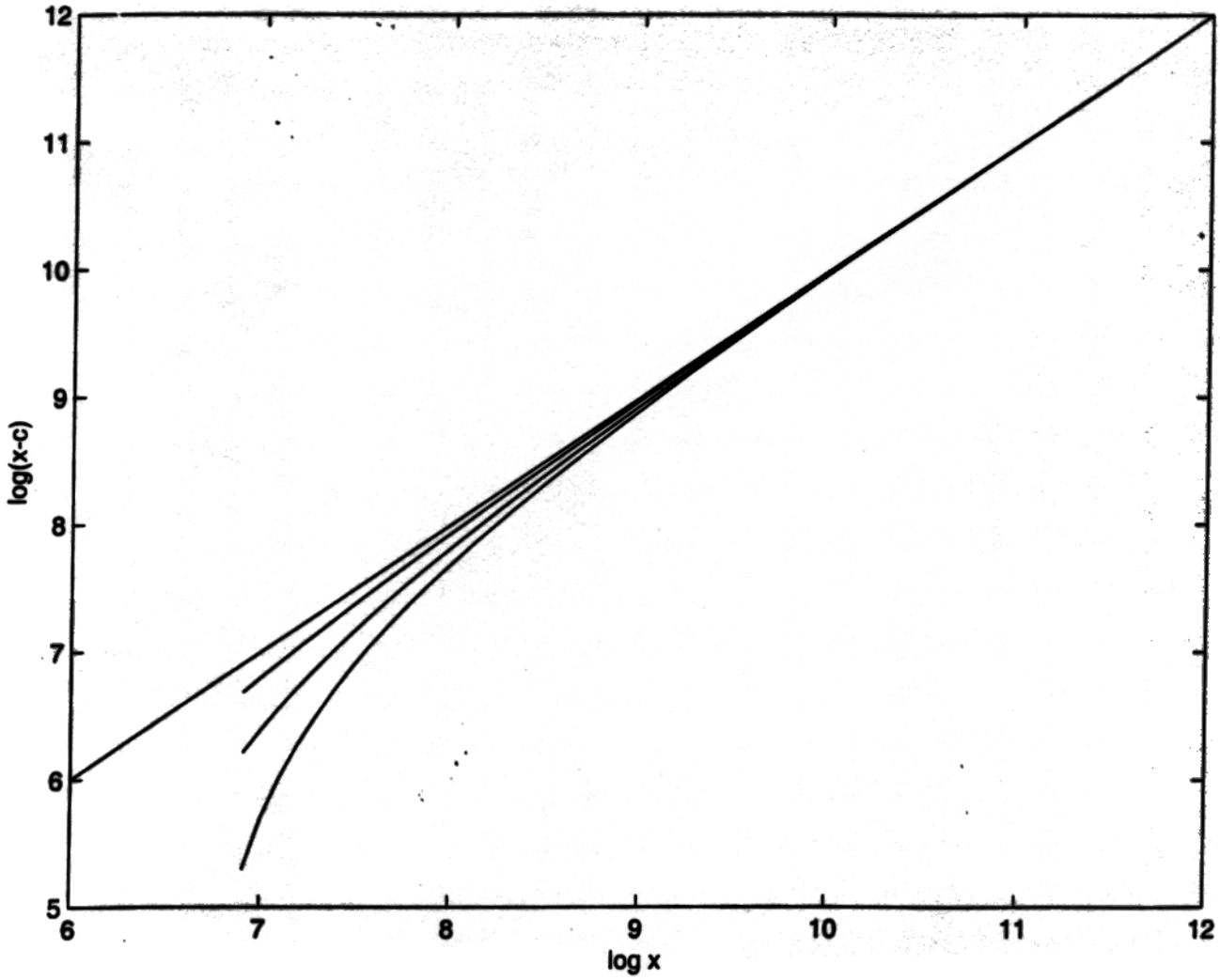
Source	SS	df	MS
Array,Dye,Variety	6896.24	13	530.48
Gene	48329.71	6117	7.90
VG,DG	17215.01	48936	0.35
Residual	5781.33	36702	0.1575
Adj Total	78222.28	85651	


Model with AG Effects, Chu Data with Background Correction on Log Scale
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
Residua
4
5
6
7
8
9
10
11
Fitted Value


Chu, Distribution of Gene Effects, Background Corrected Data
200
180
160
140
120
100
80
60
40
20
0
-3
-2
-1
0
1
2
3


Chu et al, Time 0/Time 0 Array
Log background corrected intensity
Log signal intensity
12
11
10
9
8
7
6
5
4
4
5
6
7
8
9
10
11
12


Log background corrected intensity
Log signal intensity
12
11
10
9
8
7
6
5
4
4
5
6
7
8
9
10
11
12

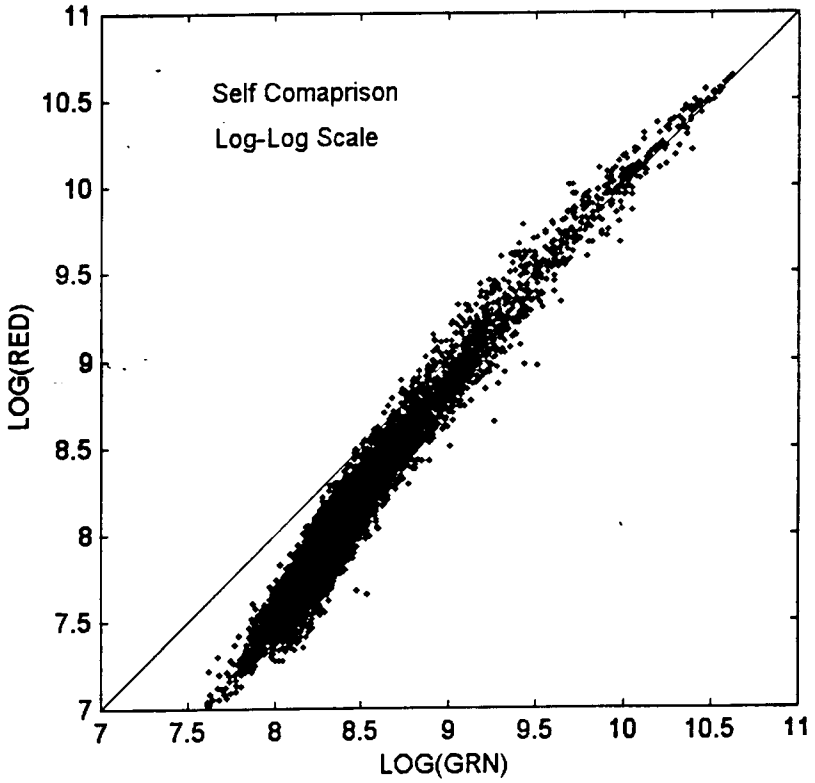

12
11
10
9
8
7
6
5
log(x-c)
6
7
8
9
10
11
12
log x

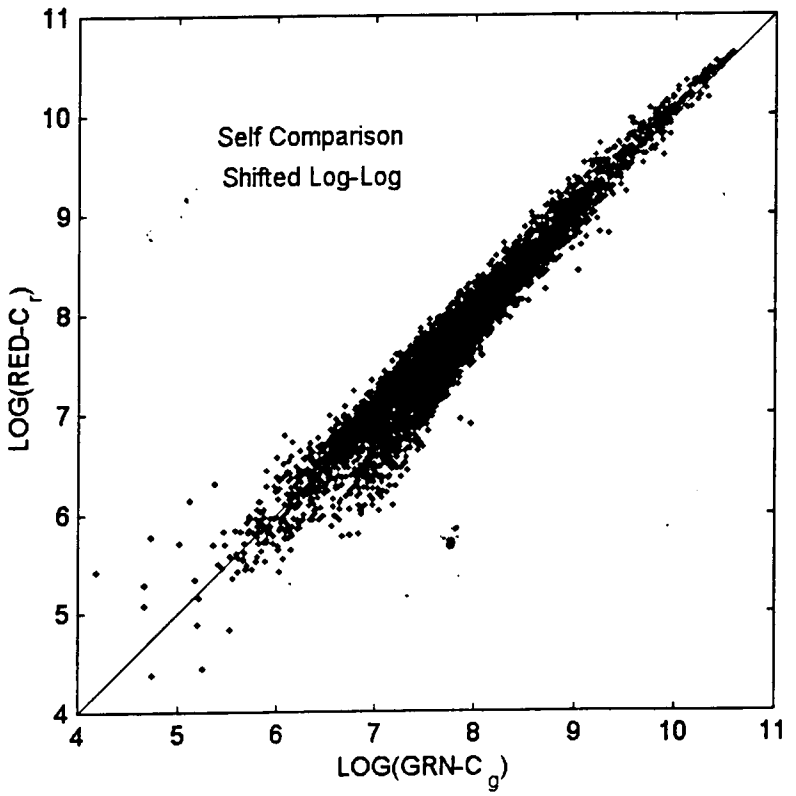
Chu Data with Median Background Correction

Source	SS	df	MS
Array,Dye,Variety	7353.57	13	565.66
Gene	43768.29	5827	7.51
VG	20081.60	34962	0.57
Residual	8855.74	40789	0.2171
Adj Total	80059.21	81591	

Source	SS	df	MS
Array,Dye,Variety	7353.57	13	565.66
Gene	43768.29	5827	7.51
VG,AG	28804.40	69924	0.41
Residual	132.95	5827	0.0228
Adj Total	80059.21	81591	

Source	SS	df	MS
Array,Dye,Variety	7353.57	13	565.66
Gene	43768.29	5827	7.51
VG,DG	20688.91	46616	0.44
Residual	8248.43	34962	0.2359
Adj Total	80059.21	81591	


Self Comaprison
Log-Log Scale
LOG(RED)
LOG(GRN)
11
10.5
10
9.5
9
8.5
8
7.5
7


Self Comparison
Shifted Log-Log
LOG(RED-C r)
LOG(GRN-C g)
11
10
9
8
7
6
5
4

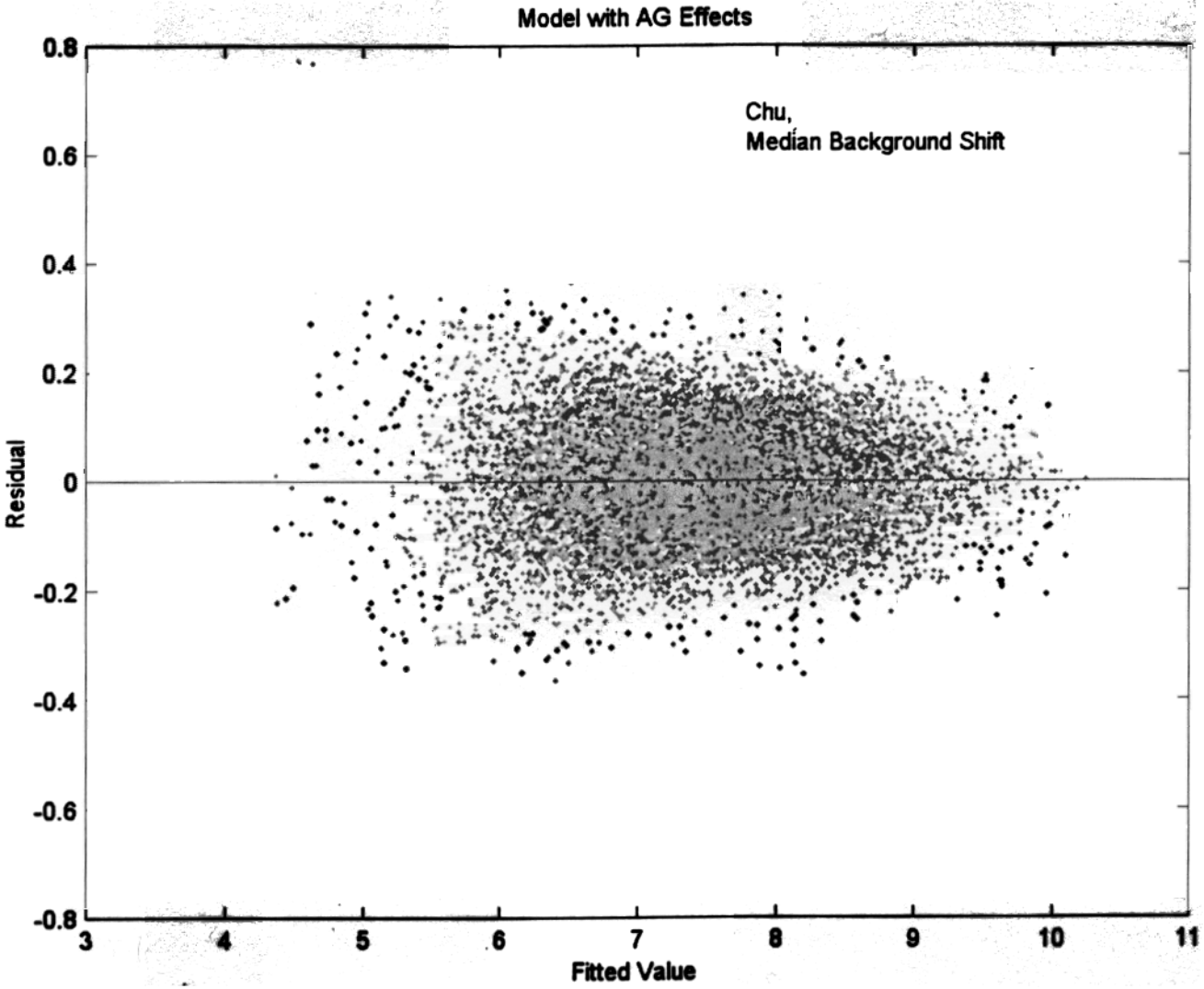
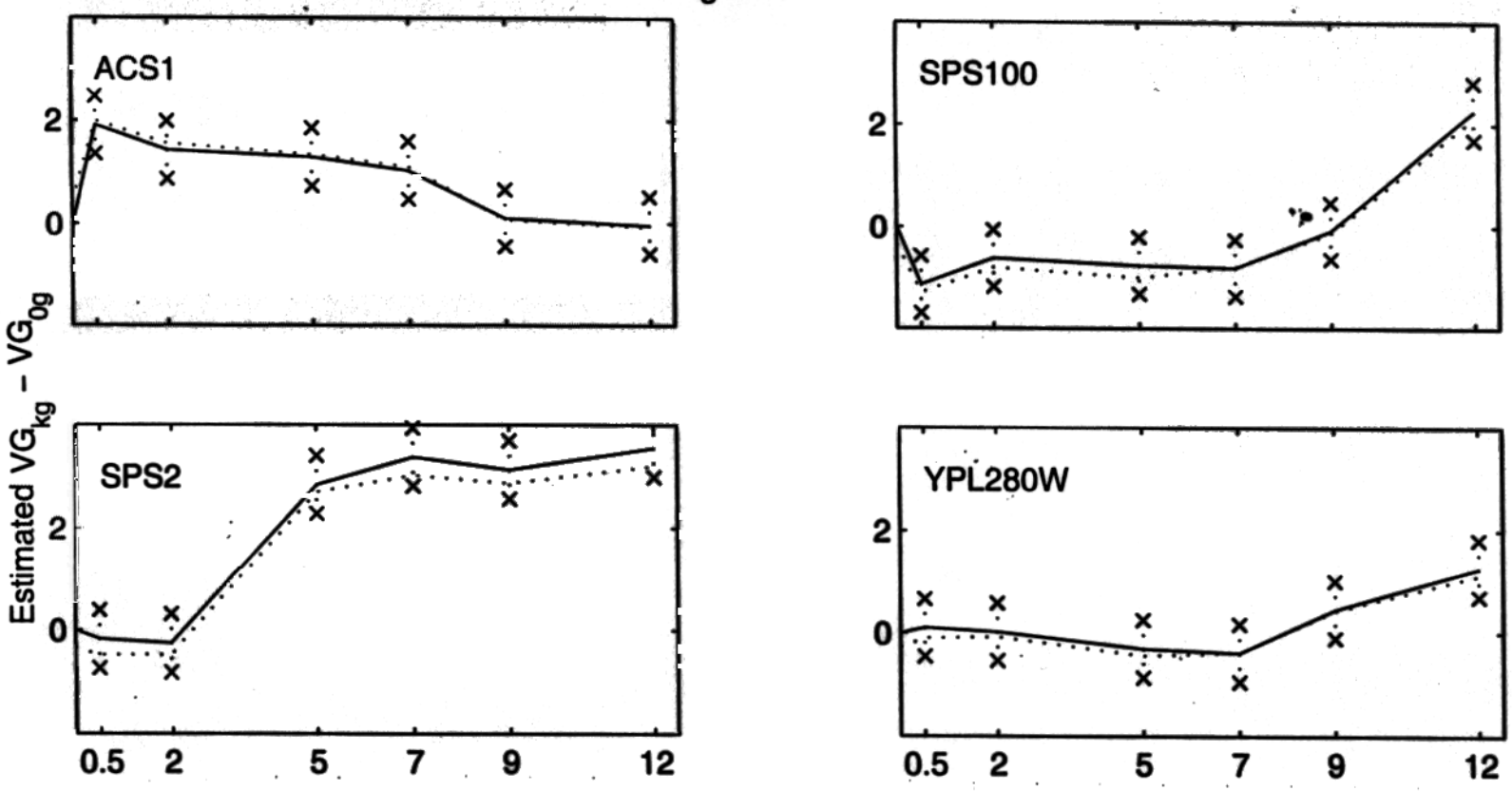
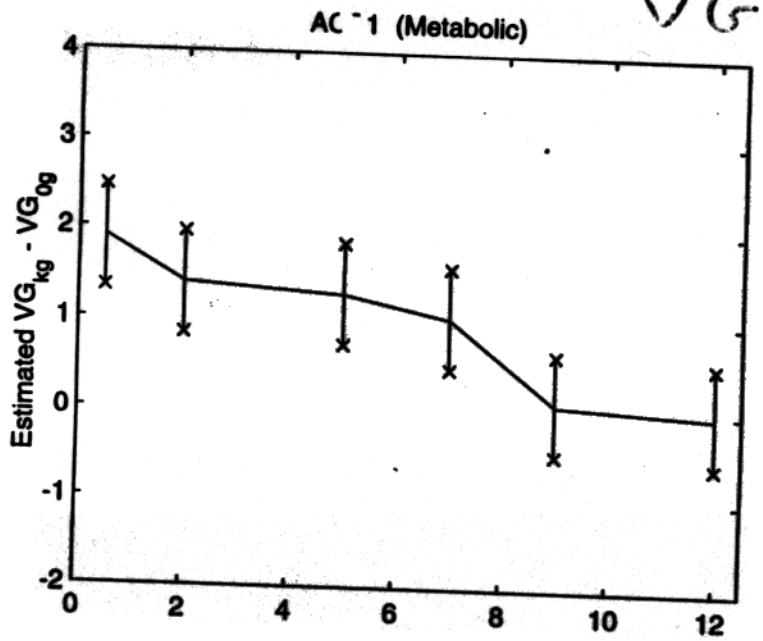
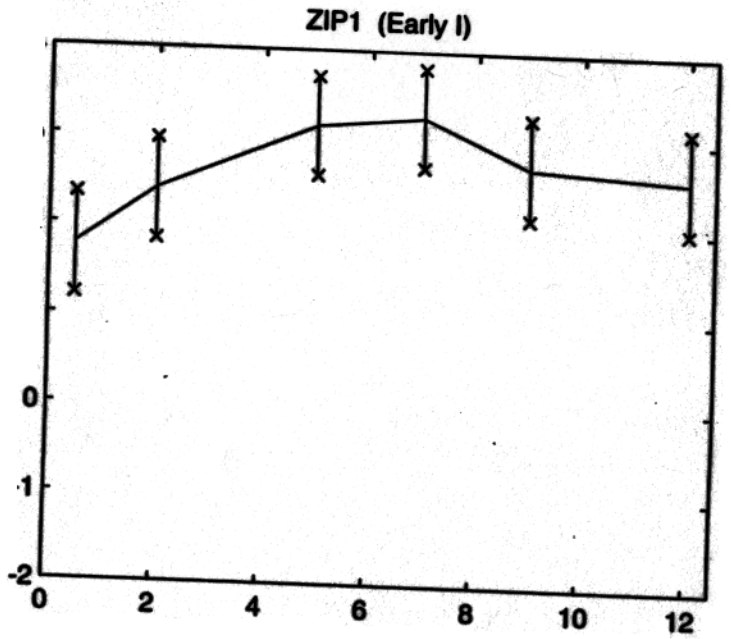
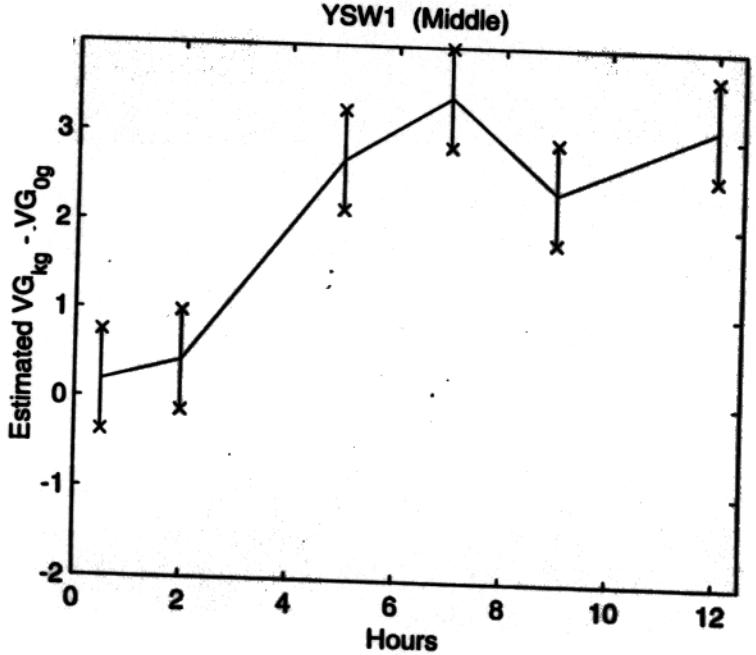

Model with AG Effects
Chu,
Median Background Shift
Residual
Fitted Value
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
3
4
5
6
7
8
9
10
11

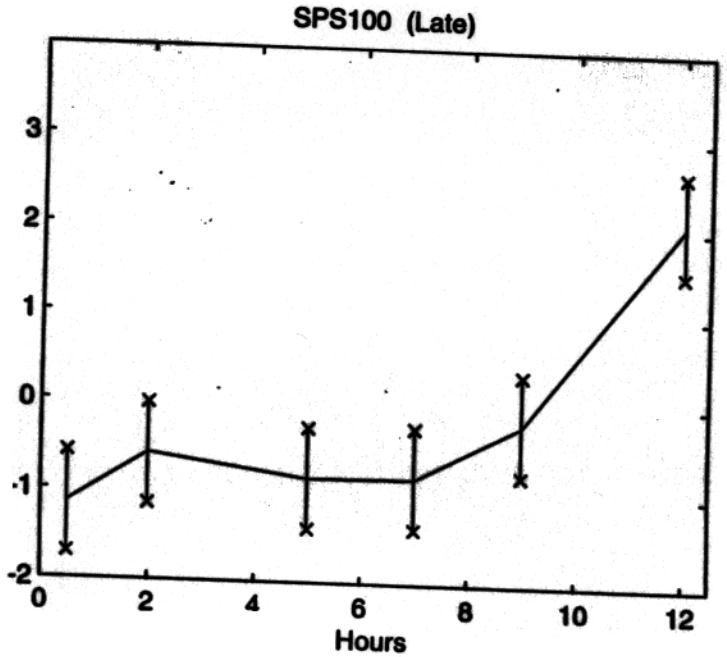
Figure 1



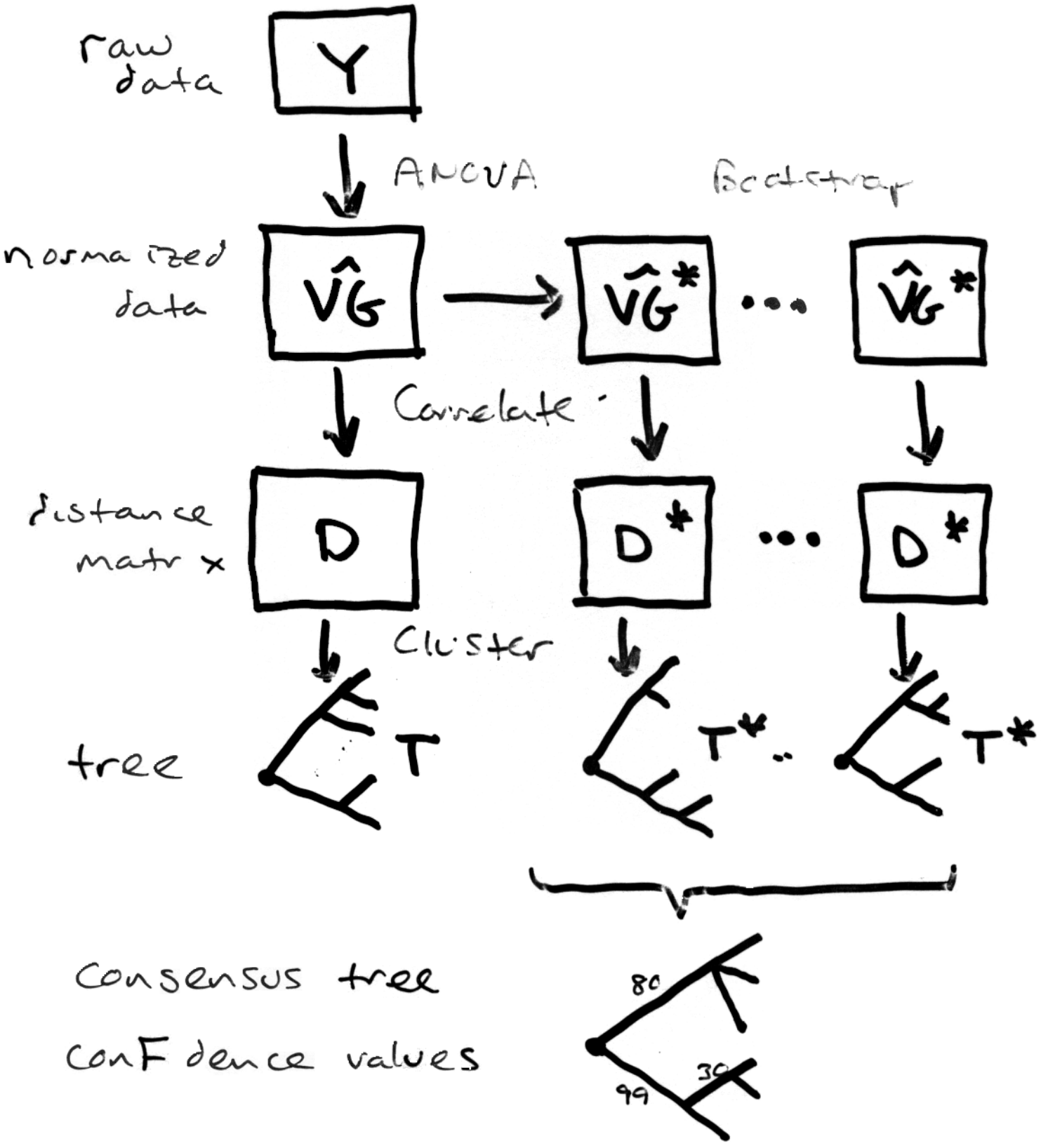

AC 1 (Metabolic)
Estimated VG_kg - VG_0g
4
3
2
1
0
-1
-2
0
2
4
6
8
10
12


ZIP1 (Early I)
1
0
-2
0
2
4
6
8
10
12


YSW1 (Middle)
Estimated VG_kg - VG_0g
3
2
1
0
-1
-2
0
2
4
6
8
10
12
Hours


SPS100 (Late)
3
2
1
0
-1
-2
0
2
4
6
8
10
12
Hours

Bootstrapping Cluster Analysis



Target Profiles
Chu et a

Figure 2

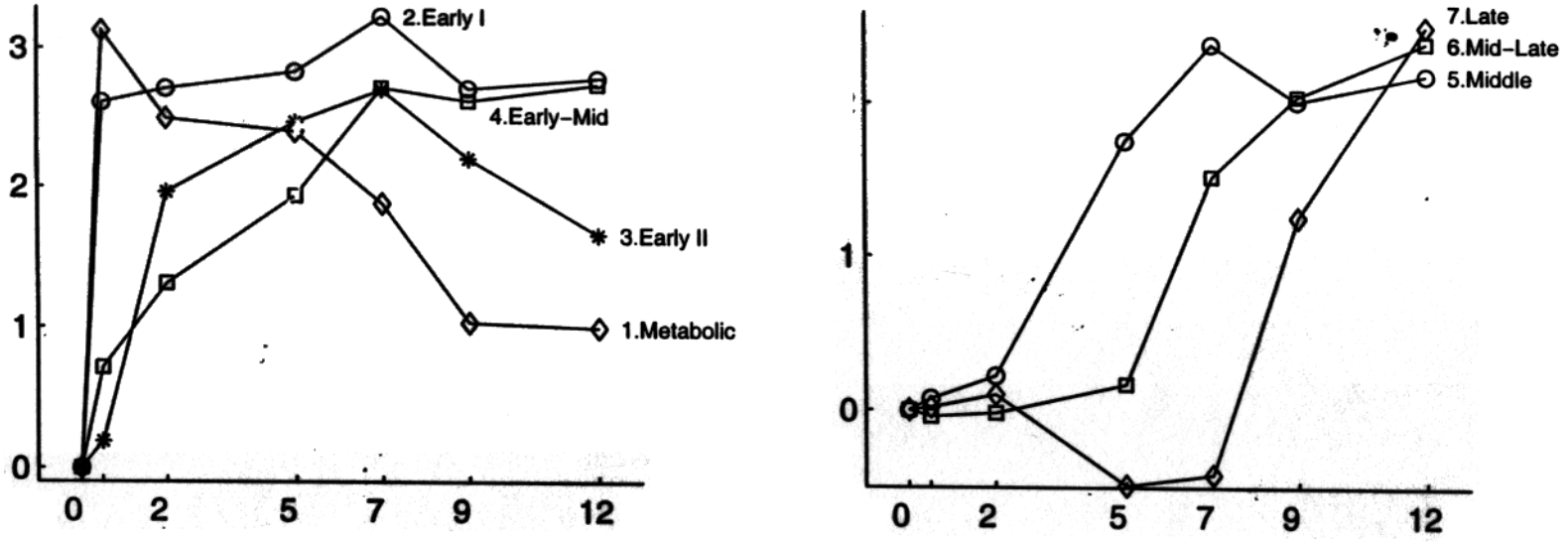
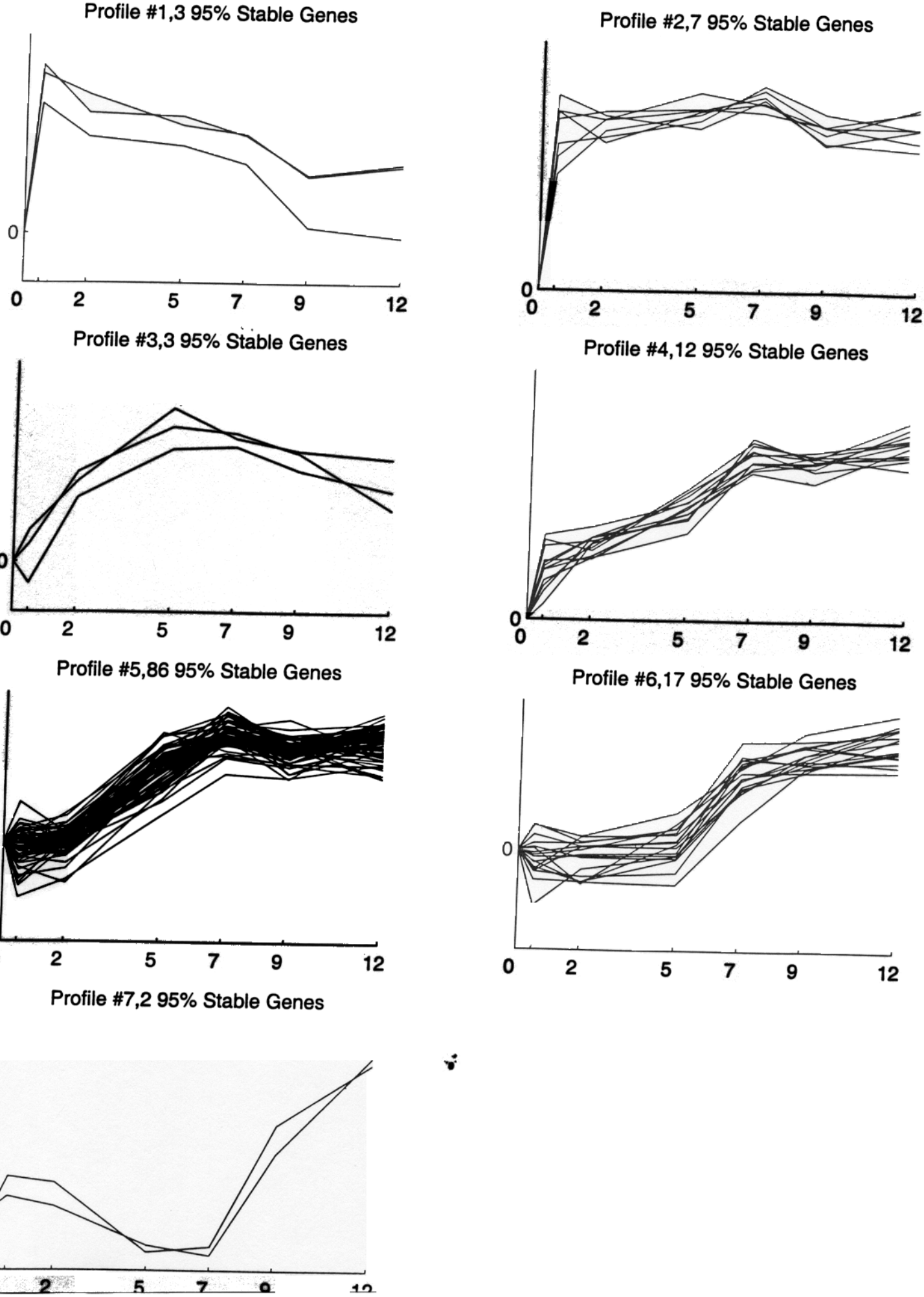


Figure 3


Profile #1,3 95% Stable Genes
0
0 2 5 7 9 12
Profile #2,7 95% Stable Genes
0
0 2 5 7 9 12
Profile #3,3 95% Stable Genes
0
0 2 5 7 9 12
Profile #4,12 95% Stable Genes
0
0 2 5 7 9 12
Profile #5,86 95% Stable Genes
0
0 2 5 7 9 12
Profile #6,17 95% Stable Genes
0
0 2 5 7 9 12
Profile #7,2 95% Stable Genes
0
0 2 5 7 9 12

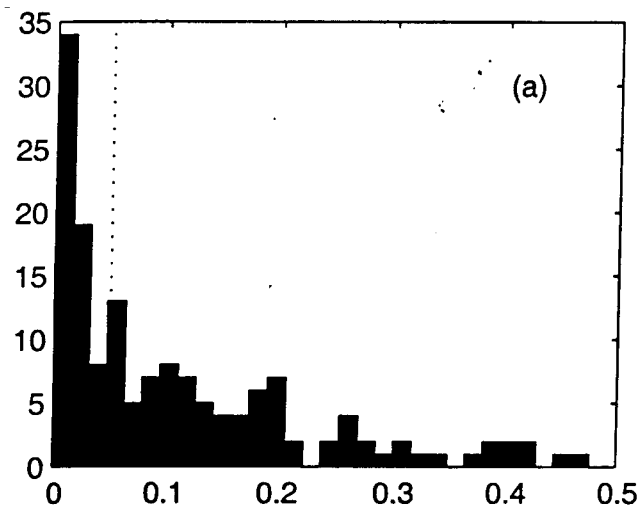
	(a) Chu *et al.*	(b) Nominal	(c) 95% Stable	(d) 80% Stable
Profile 1	52	65	3	8
Profile 2	61	51	7	11
Profile 3	45	74	3	11
Profile 4	95	151	12	27
Profile 5	158	241	86	120
Profile 6	61	145	17	36
Profile 7	5	15	2	6

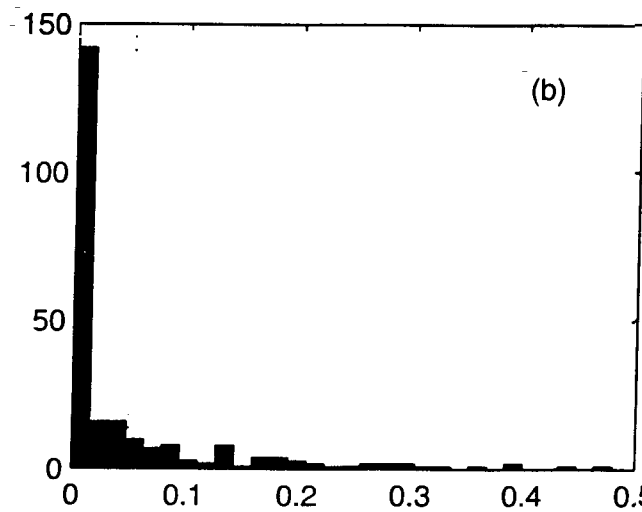
Table 2: Number of genes matching to each profile for (a) Chu *et al.* clustering method, (b) modified clustering method with no reliability measure, (c) modified clustering method requiring 95% stability, (d) modified clustering method requiring 80% stability. Column (d) is included because our choice of 95% for stability is somewhat arbitrary.

Correlations among Chu et al Profiles

	Profile					
	2	3	4	5	6	7
1	.65	.19	.03	-.14	-.39	-.41
2		.73	.77	.60	.40	.11
3			.84	.78	.46	.01
4				.95	.84	.44
5					.83	.36
6						.75

Table 3: Pairwise correlations among the seven profiles.





Summary

Experimental Design Concepts For microarrays

- Confounding and partial confounding effects
- Role of replication is 2-fold
- Relative efficiency for comparing designs

OPEN QUESTIONS

Designs for structured experiments

Designs for large surveys

ANOVA as an analysis too

For microarrays

- Normalization of main effects
- Control of spot effects and D×Gs
- Easy compution / least squares

OPEN QUESTIONS

Scale of measurement

Robustify

- modeling other effects
- Random effects + shrinkage
- gene specific error rates