Introduction to Statistical Design and Analysis of Microarray Experiments

1/13/2009

Microarray Technology

- Microarrays allow researchers to measure the abundance of thousands of mRNA transcripts in multiple biological samples.
- By understanding how transcript abundance changes across experimental conditions, researchers gain clues about gene function and learn how genes work together to carry out biological processes.

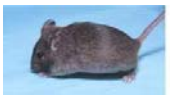
Uses of Microarray Technology

Examples from

Iowa State University

Using Microarrays to Identify Genes Involved in Muscle Development

Wild-Type Mouse



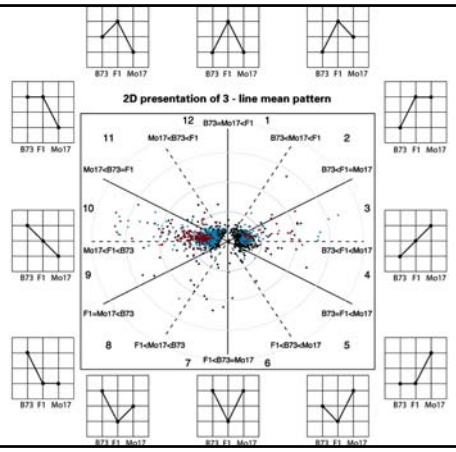
Myostatin Knockout Mouse

Belgian Blue cattle have a mutation in the myostatin gene.

Steelman, C. A., Recknor, J. C., Nettleton, D., Reecy, J.M. (2006). *The FASEB Journal*. 10.1096/fj.05-5125fje.

B73 F1 Mo17 B73 F1 Mo17

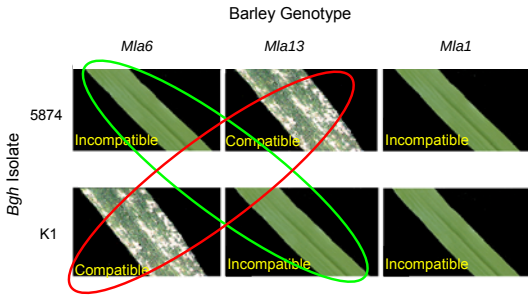
Swanson-Wagner, Jia, DeCook, Borsuk, Nettleton, Schnable (2006) *Proceeding of the National Academy of Science*. **103**, 6805-6810.



FDR
15%
5%
1%

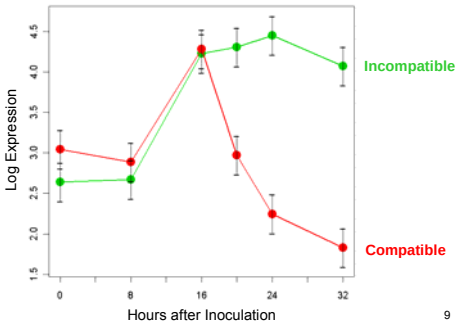
7

Identifying Genes Involved in Pathways That Distinguish Compatible from Incompatible Interactions



Caldo, Nettleton, Wise (2004). *The Plant Cell*. **16**, 2514-2528.

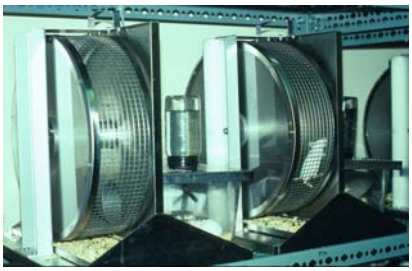
An Example Gene of Interest



9

d1

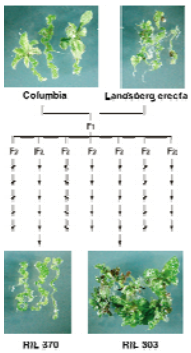
Gene Expression in Brains of Mice Selected for High Voluntary Exercise



Bronikowski, Rhodes, Garland, Prolla, Awad, Gammie (2004) *Evolution* **58**, 2079-2086.

10

Mapping Expression Quantitative Trait Loci (eQTL)



Parental inbred lines

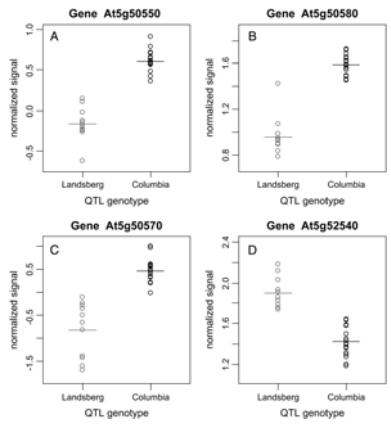
Columbia (C): good shoot producer

Landsberg (L): poor shoot producer

Cross parents to generate recombinant inbred lines (RILs) that vary in shoot production.

Measure gene expression in 30 RILs during shoot production.

DeCook, R., Lall, S., Nettleton, D., Howell, S. H. (2006). *Genetics*. **172** 1155-1164.

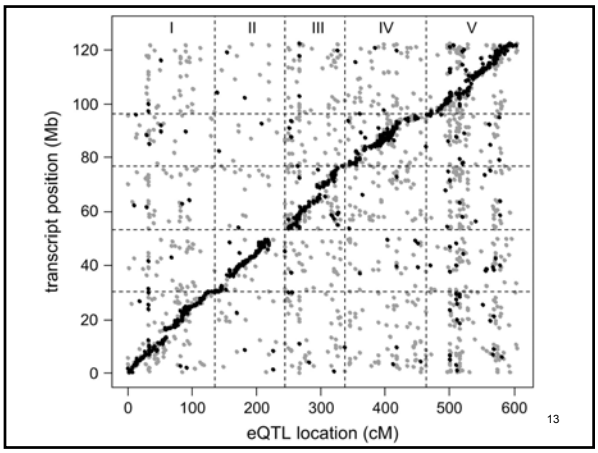


Example genes whose expression is associated with a single genetic position

12

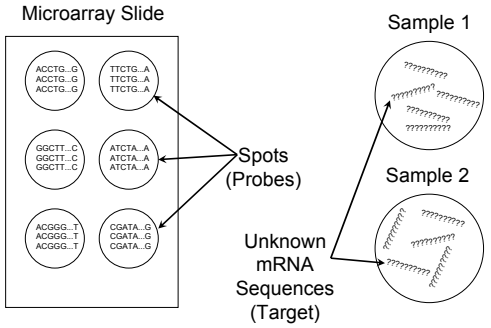
Slide 10

d1 27th generation 15k vs. 5k
dnett, 1/13/2009



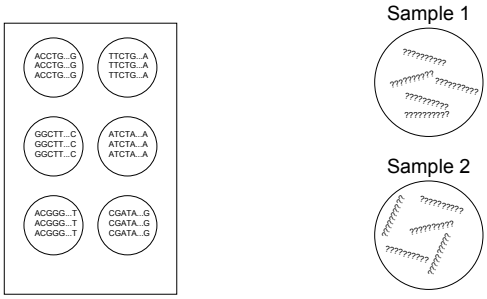
13

Two-Color Microarrays (cartoon version)



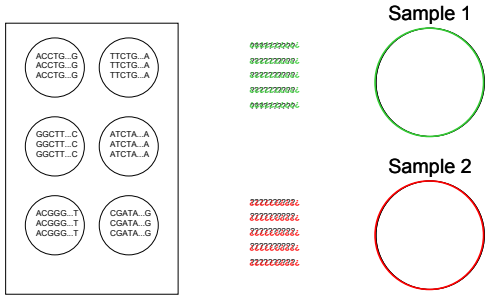
14

Extract mRNA



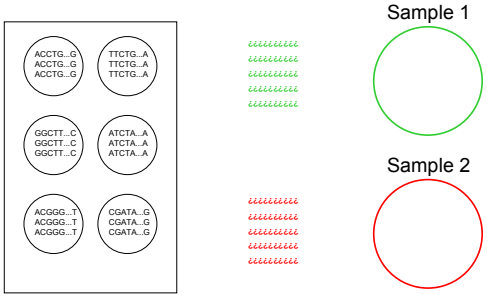
15

Convert to cDNA and Label with Fluorescent Dyes



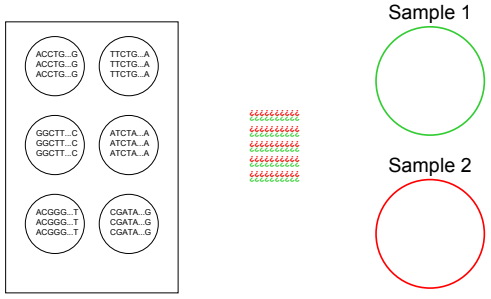
16

Mix Labeled cDNA



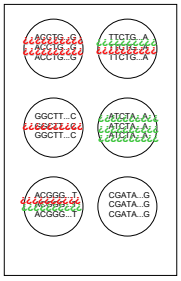
17

Hybridize cDNA to the Slide



18

Excite Dyes with Laser

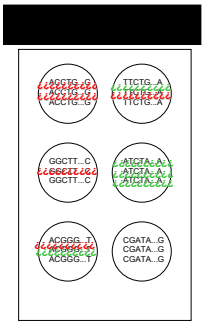


Sample 1

Sample 2

19

Scan

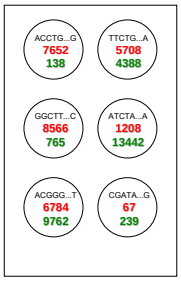


Sample 1

Sample 2

20

Quantify Signals

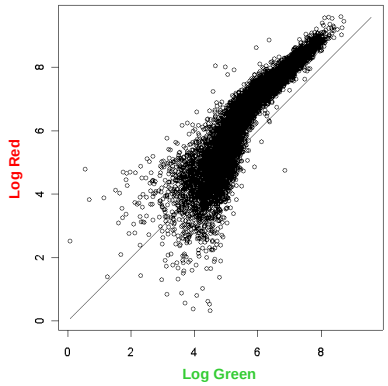


Sample 1

Sample 2

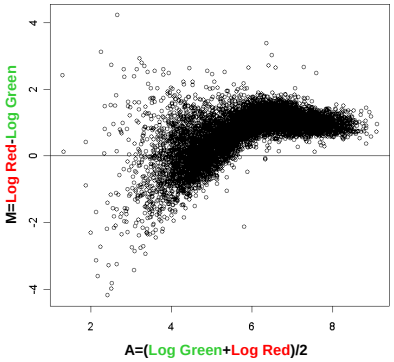
21

Normalization



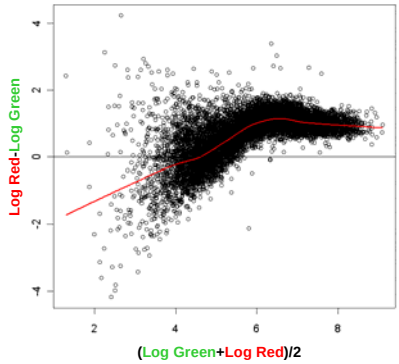
22

M vs. A Plot (45º rotation)



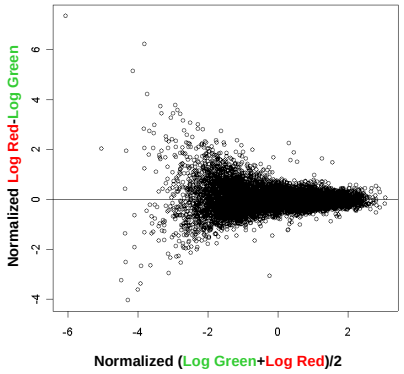
23

Loess Fit



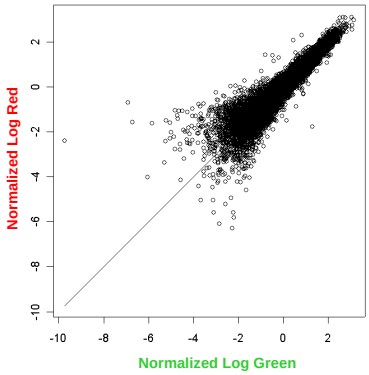
24

M vs. A after Normalization



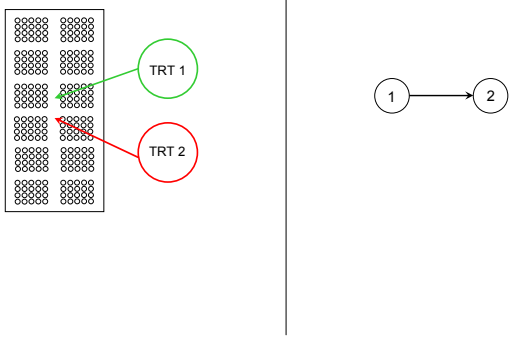
25

Normalized Data



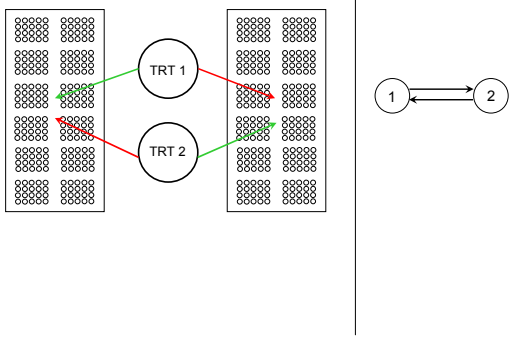
26

Microarray Experimental Design Notation



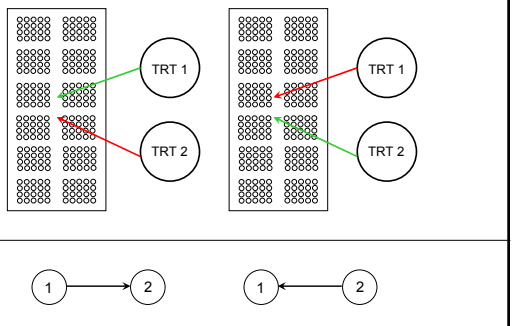
27

Microarray Experimental Design Notation



28

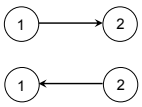
Microarray Experimental Design Notation



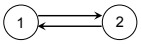
29

Biological Replicates vs. Technical Replicates

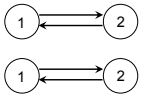
Biological Replication



Technical Replication

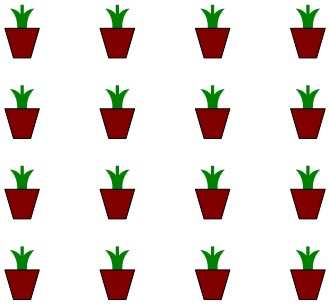


Both Biological and Technical Replication



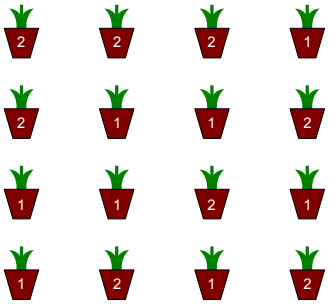
30

Example 1: Two-Treatment CRD



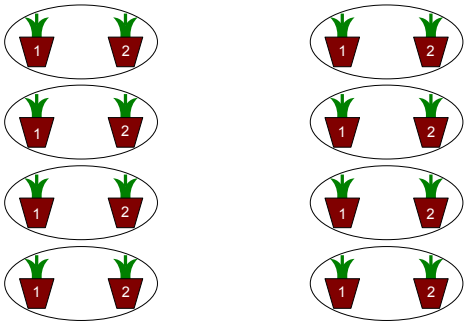
31

Assign 8 Plants to Each Treatment Completely at Random



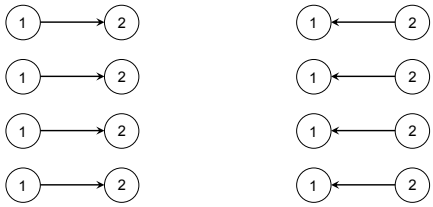
32

Randomly Pair Plants Receiving Different Treatments



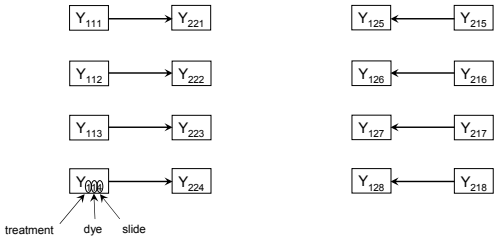
33

Randomly Assign Pairs to Slides Balancing the Two Dye Configurations



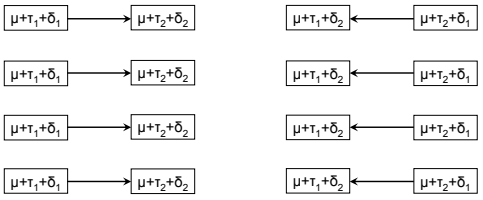
34

Observed Normalized Signal Intensities (NSI) for One Gene



35

Unknown Means Underlying the Observed Normalized Signal Intensities (NSI)



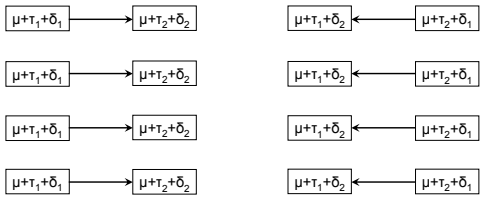
μ represents overall mean of NSI.

τ_1 and τ_2 represent the effects of treatments 1 and 2 on mean NSI.

δ_1 and δ_2 represents the effects of Cy3 and Cy5 dyes on mean NSI.

36

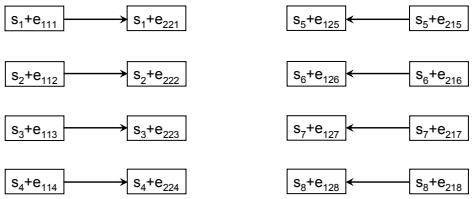
Unknown Means Underlying the Observed Normalized Signal Intensities (NSI)



A gene is differentially expressed if $\tau_1 \neq \tau_2$.

37

Unknown Random Effects Underlying Observed NSI

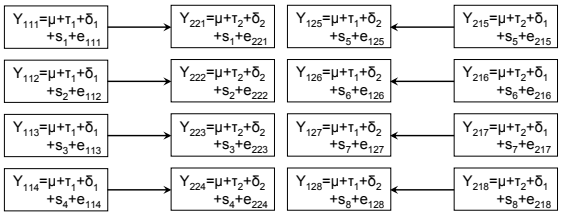


s_1, s_2, s_3, s_4, s_5, s_6, s_7, and s_8 represent slide effects.

$e_{111},\ldots,e_{218}$ represent residual random effects that include any sources of variation unaccounted for by other terms.

38

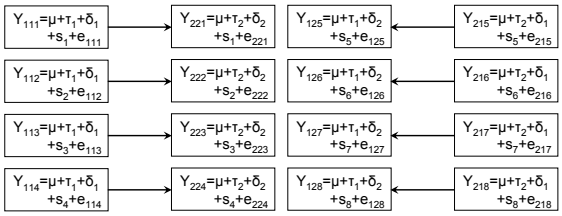
Observed NSI are Means Plus Random Effects



$$Y_{ijk}=\mu+\tau_i+\delta_j+s_k+e_{ijk}$$

39

Observed NSI are Means Plus Random Effects

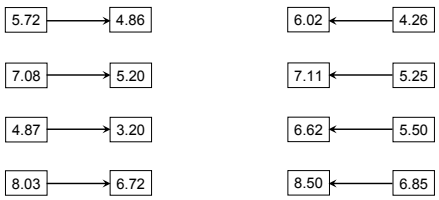


Recall that a gene is differentially expressed if $\tau_1 \neq \tau_2$.

$$\bar{Y}_{1..} - \bar{Y}_{2..} = \tau_1 - \tau_2 + \bar{e}_{1..} - \bar{e}_{2..}$$

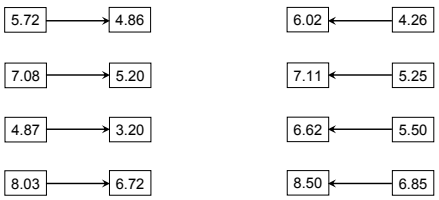
40

Observed Normalized Signal Intensities (NSI) for One Gene



41

Observed Normalized Signal Intensities (NSI) for One Gene

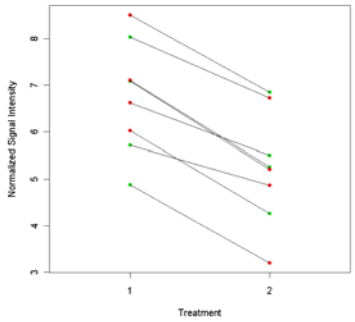


Recall that a gene is differentially expressed if $\tau_1 \neq \tau_2$.

$$\bar{Y}_{1..} - \bar{Y}_{2..} = \tau_1 - \tau_2 + \bar{e}_{1..} - \bar{e}_{2..} = 1.514$$

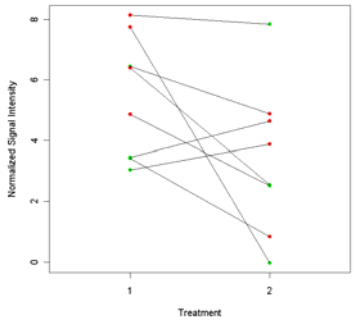
42

P-Value for Testing $\tau_1 = \tau_2$ is < 0.0001 Estimated Fold Change=4.54
95% Confidence Interval for Fold Change 3.23 to 6.38



43

P-Value for Testing $\tau_1 = \tau_2$ is 0.0660 Estimated Fold Change=7.76
95% Confidence Interval for Fold Change 0.83 to 72.49



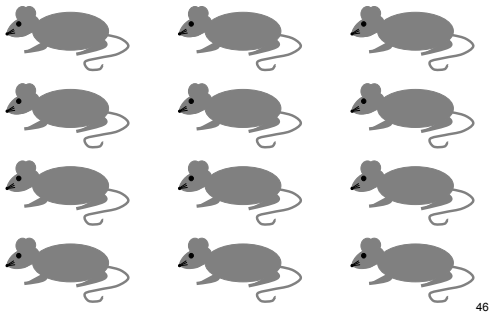
44

Example 2: CRD with Affymetrix Technology

- What genes are involved in muscle hypertrophy?
- Design a treatment that will induce hypertrophy in muscle tissue and an appropriate control treatment.
- Randomly assign experimental units to the two treatments.
- Use microarray technology to measure mRNA transcript abundance in muscle tissue.
- Identify genes whose mRNA levels differs between treatments.

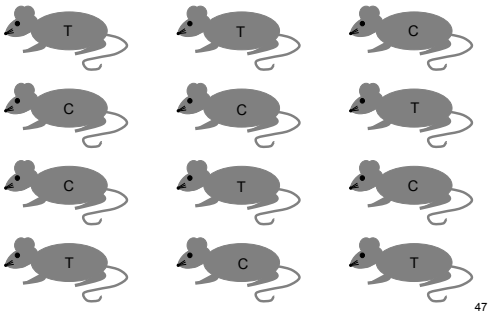
45

Assign 6 mice to each treatment completely at random



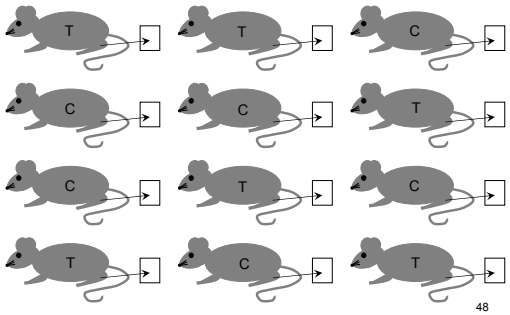
46

Assign 6 mice to each group completely at random



47

Measure Expression in Relevant Muscle Tissue with Affymetrix GeneChips



48

Normalized Data

Genes	Experimental Units											
	T1	T2	T3	T4	T5	T6	C1	C2	C3	C4	C5	C6
1	3.7	4.1	3.9	5.1	5.4	5.0	6.0	5.5	4.0	4.6	4.6	5.3
2	8.2	6.2	7.3	7.6	6.0	6.7	8.1	6.4	5.6	7.6	6.6	8.4
3	6.9	4.1	5.1	3.3	5.4	6.6	6.0	4.9	5.7	9.3	7.4	9.1
4	8.6	8.8	9.1	9.8	7.9	7.4	6.2	6.8	6.6	6.8	5.5	7.7
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
40000	3.5	1.5	2.9	4.5	0.9	0.9	3.0	3.9	3.8	3.1	3.9	1.3

Model for One Gene

$Y_{ij}=\mu+\tau_i+e_{ij}$ $(i=1,2;\ j=1, 2, 3, 4, 5, 6)$

Y_{ij}=normalized signal intensity for the j^{th} experimental unit exposed to the i^{th} treatment

μ=mean normalized signal intensity

τ_i=effect due to i^{th} treatment

e_{ij}=residual effect for the j^{th} experimental unit exposed to i^{th} treatment

Gene 4: Data Analysis

$Y_{11}=8.6$ $Y_{21}=6.2$

$Y_{12}=8.8$ $Y_{22}=6.8$

$Y_{13}=9.1$ $Y_{23}=6.6$

$Y_{14}=9.8$ $Y_{24}=6.8$

$Y_{15}=7.9$ $Y_{25}=5.5$

$Y_{16}=7.4$ $Y_{26}=7.7$

$\bar{Y}_{1\cdot}=8.6$ $\bar{Y}_{2\cdot}=6.6$

$\bar{Y}_{1\cdot}-\bar{Y}_{2\cdot}=\tau_1-\tau_2+\bar{e}_{1\cdot}-\bar{e}_{2\cdot}=2.0$

$$se(\bar{Y}_{1\cdot}-\bar{Y}_{2\cdot}) = s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$= 0.7949843\sqrt{\frac{1}{6}+\frac{1}{6}}$$

$$= 0.4589844$$

Gene 4: 95% Confidence Interval for $\tau_1-\tau_2$

$\bar{Y}_{1\cdot}-\bar{Y}_{2\cdot}=2.0$ $\quad se(\bar{Y}_{1\cdot}-\bar{Y}_{2\cdot})=0.4589844$

$\bar{Y}_{1\cdot}-\bar{Y}_{2\cdot} \pm t^{(0.975)}_{n_1+n_2-2} se(\bar{Y}_{1\cdot}-\bar{Y}_{2\cdot})$

$2.0 \pm 2.228 * 0.4589844$

$(0.98, 3.02)$

Gene 4: 95% Confidence Interval for Fold Change

Estimated Fold Change= $e^{\bar{Y}_{1\cdot}-\bar{Y}_{2\cdot}} = e^{2.0} \approx 7.4$

$$(e^{\bar{Y}_{1\cdot}-\bar{Y}_{2\cdot}-t^{(0.975)}_{n1+n2-2}se(\bar{Y}_{1\cdot}-\bar{Y}_{2\cdot})}, e^{\bar{Y}_{1\cdot}-\bar{Y}_{2\cdot}+t^{(0.975)}_{n1+n2-2}se(\bar{Y}_{1\cdot}-\bar{Y}_{2\cdot})})$$

(2.7,20.5)

Gene 4: *t*-test

$Y_{11}=8.6$ $Y_{21}=6.2$

$Y_{12}=8.8$ $Y_{22}=6.8$

$Y_{13}=9.1$ $Y_{23}=6.6$

$Y_{14}=9.8$ $Y_{24}=6.8$

$Y_{15}=7.9$ $Y_{25}=5.5$

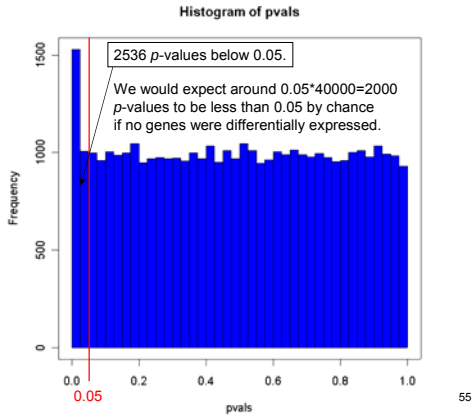
$Y_{16}=7.4$ $Y_{26}=7.7$

$\bar{Y}_{1\cdot}=8.6$ $\bar{Y}_{2\cdot}=6.6$

$\bar{Y}_{1\cdot}-\bar{Y}_{2\cdot}=\tau_1-\tau_2+\bar{e}_{1\cdot}-\bar{e}_{2\cdot}=2.0$

$$t = \frac{\bar{Y}_{1\cdot}-\bar{Y}_{2\cdot}}{se(\bar{Y}_{1\cdot}-\bar{Y}_{2\cdot})} = \frac{2.0}{0.4589844} = 4.3574.$$

Compare to a *t*-distribution with $n_1+n_2-2=10$ d.f. to obtain *p*-value≈0.001427.



55

Which of the 40,000 genes are involved in hypertrophy?

- Create a list of genes so that the estimated proportion of false positive results will be no larger than 5%.
- For this example, we could declare the 163 genes with the smallest *p*-values to be "differentially expressed."
- The estimated False Discovery Rate (FDR) for this list of genes would be slightly less than 5%.
- This would be equivalent to declaring all genes with *p*-value less than 0.0002032437 to be "differentially expressed."
- We will study such computations in great detail later in the course.

56