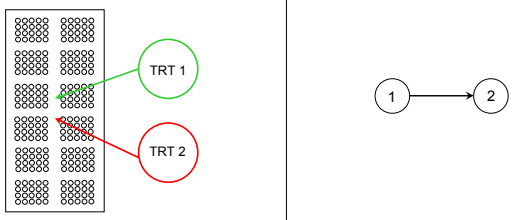
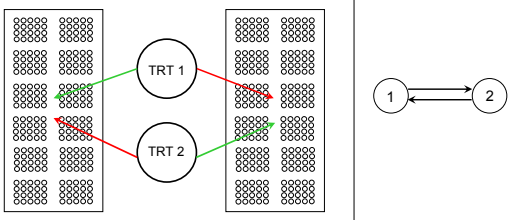
Simple Examples of Analysis for a Single Gene

1/13/2011

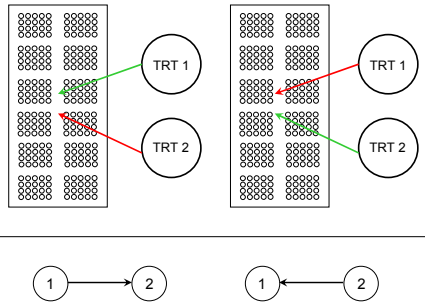
Two-Color Microarray Experimental Design Notation



Microarray Experimental Design Notation

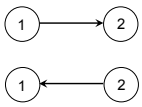


Microarray Experimental Design Notation

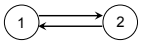


Biological Replicates vs. Technical Replicates

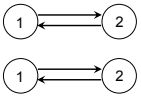
Biological Replication



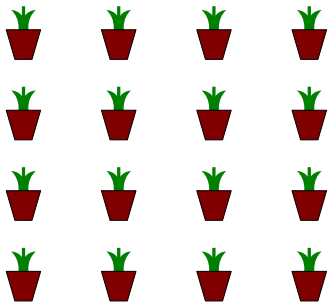
Technical Replication



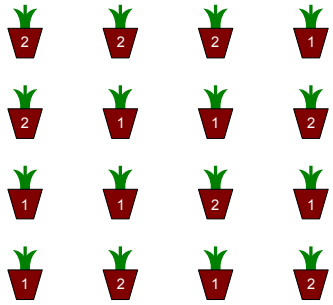
Both Biological and Technical Replication



Example 1: Two-Treatment CRD

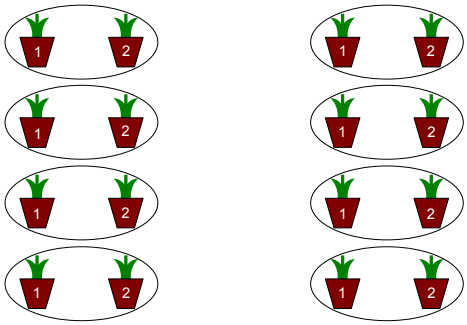


Assign 8 Plants to Each Treatment Completely at Random



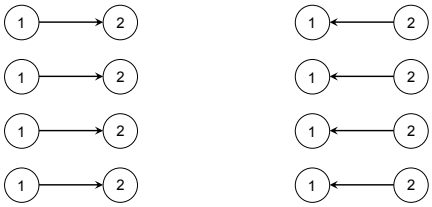
7

Randomly Pair Plants Receiving Different Treatments



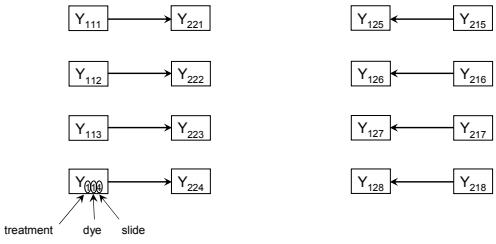
8

Randomly Assign Pairs to Slides Balancing the Two Dye Configurations



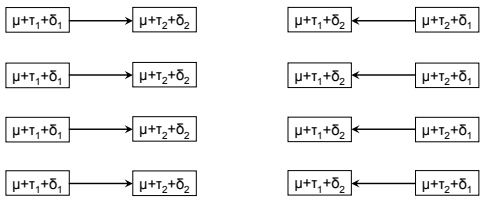
9

Observed Normalized Log Signal Intensities for One Gene



10

Unknown Means Underlying the Observed Normalized Log Signal Intensities (NLSI)



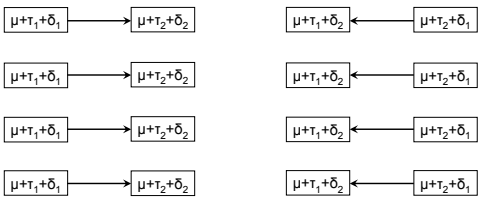
μ is a real-valued parameter common to all observations.

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

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

11

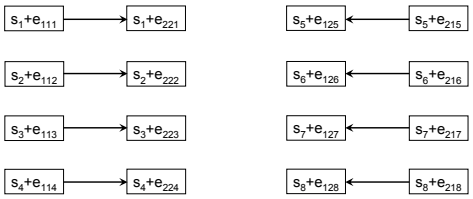
Differential Expression



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

12

Unknown Random Effects Underlying Observed NLSI



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

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

To make our model complete, we need to say more about the random effects.

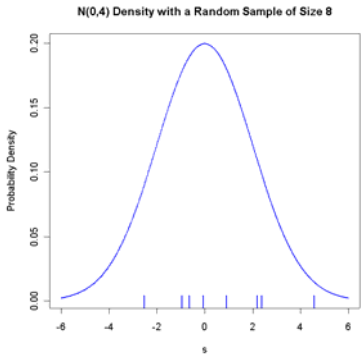
- We will almost always assume that random effects are independent and normally distributed with mean zero and a factor-specific variance.
- $s_1, s_2, \ldots, s_8 \overset{iid}{\sim} N(0,\sigma_s^2)$ and independent of

  $e_{111}, e_{112}, e_{113}, e_{114}, e_{221}, e_{222}, e_{223}, e_{224}, e_{125},$

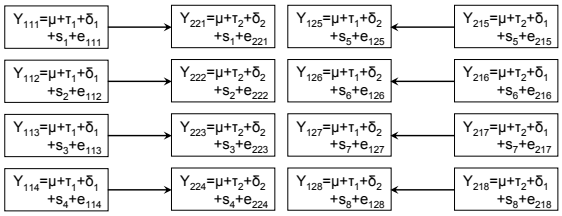
  $e_{126}, e_{127}, e_{128}, e_{215}, e_{216}, e_{217}, e_{218} \overset{iid}{\sim} N(0,\sigma_e^2)$.

  (or just $e_{ijk} \overset{iid}{\sim} N(0,\sigma_e^2)$ to save time and space.)

What does $s_1, s_2, \ldots, s_8 \overset{iid}{\sim} N(0,\sigma_s^2)$ mean?

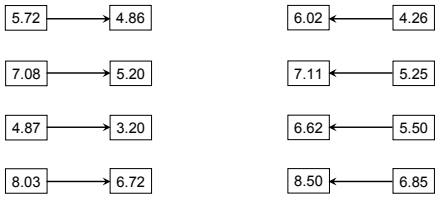


Observed NLSI are Modeled as Means Plus Random Effects



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

Observed Normalized Signal Intensities (NLSI) for One Gene



Given data, our task it to determine whether the gene is differentially expressed and, if so, estimate the magnitude and direction of differential expression.

Analysis of Log Red to Green Ratios

- Rather than working with the normalized log signal intensities, it is often customary to consider the log of the red to green normalized signals from each slide as the basic data for analysis.
- This is equivalent to working with the red – green difference in NLSI from each slide.

$$\log(R/G)=\log(R)-\log(G)$$

Differences for Slides with Treatment 1 Green and Treatment 2 Red

Slide	Difference
$Y_{111}=\mu+\tau_1+\delta_1+s_1+e_{111}$ → $Y_{221}=\mu+\tau_2+\delta_2+s_1+e_{221}$	$Y_{221}-Y_{111}= \tau_2-\tau_1+\delta_2-\delta_1+e_{221}-e_{111}$
$Y_{112}=\mu+\tau_1+\delta_1+s_2+e_{112}$ → $Y_{222}=\mu+\tau_2+\delta_2+s_2+e_{222}$	$Y_{222}-Y_{112}= \tau_2-\tau_1+\delta_2-\delta_1+e_{222}-e_{112}$
$Y_{113}=\mu+\tau_1+\delta_1+s_3+e_{113}$ → $Y_{223}=\mu+\tau_2+\delta_2+s_3+e_{223}$	$Y_{223}-Y_{113}= \tau_2-\tau_1+\delta_2-\delta_1+e_{223}-e_{113}$
$Y_{114}=\mu+\tau_1+\delta_1+s_4+e_{114}$ → $Y_{224}=\mu+\tau_2+\delta_2+s_4+e_{224}$	$Y_{224}-Y_{114}= \tau_2-\tau_1+\delta_2-\delta_1+e_{224}-e_{114}$

Note that according to our original model, these differences are iid $N(\tau_2-\tau_1+\delta_2-\delta_1, 2\sigma_e^2)$.

19

Differences for Slides with Treatment 1 Red and Treatment 2 Green

Difference	Slide
$Y_{125}-Y_{215}= \tau_1-\tau_2+\delta_2-\delta_1+e_{125}-e_{215}$	$Y_{125}=\mu+\tau_1+\delta_2+s_5+e_{125}$ ← $Y_{215}=\mu+\tau_2+\delta_1+s_5+e_{215}$
$Y_{126}-Y_{216}= \tau_1-\tau_2+\delta_2-\delta_1+e_{126}-e_{216}$	$Y_{126}=\mu+\tau_1+\delta_2+s_6+e_{126}$ ← $Y_{216}=\mu+\tau_2+\delta_1+s_6+e_{216}$
$Y_{127}-Y_{217}= \tau_1-\tau_2+\delta_2-\delta_1+e_{127}-e_{217}$	$Y_{127}=\mu+\tau_1+\delta_2+s_7+e_{127}$ ← $Y_{217}=\mu+\tau_2+\delta_1+s_7+e_{217}$
$Y_{128}-Y_{218}= \tau_1-\tau_2+\delta_2-\delta_1+e_{128}-e_{218}$	$Y_{128}=\mu+\tau_1+\delta_2+s_8+e_{128}$ ← $Y_{218}=\mu+\tau_2+\delta_1+s_8+e_{218}$

Note that according to our original model, these differences are iid $N(\tau_1-\tau_2+\delta_2-\delta_1, 2\sigma_e^2)$.

20

If we let d_k denote the difference from slide k, we have

d_1, d_2, d_3, d_4 iid $N(\tau_2-\tau_1+\delta_2-\delta_1, 2\sigma_e^2)$

independent of

d_5, d_6, d_7, d_8 iid $N(\tau_1-\tau_2+\delta_2-\delta_1, 2\sigma_e^2)$.

A standard two-sample t-test can be used to test

$H_0: \tau_2-\tau_1+\delta_2-\delta_1= \tau_1-\tau_2+\delta_2-\delta_1$ which is equivalent to

$H_0: \tau_1= \tau_2$ (null hypothesis of no differential expression).

21

Estimation of the Direction and Magnitude of Differential Expression

- An unbiased estimator of $\tau_1-\tau_2$ is given by

  $\{ \text{mean}(d_5, d_6, d_7, d_8) - \text{mean}(d_1, d_2, d_3, d_4) \} / 2.$

- Because $\tau_1-\tau_2$ is a difference in treatment effects for a measure of log expression level, $\exp(\tau_1-\tau_2)$ can be interpreted as a ratio of expression levels on the original scale.
- $\exp[\{ \text{mean}(d_5, d_6, d_7, d_8) - \text{mean}(d_1, d_2, d_3, d_4) \} / 2]$ can be reported as an estimate of the *fold change* in the expression level for treatment 1 relative to treatment 2.

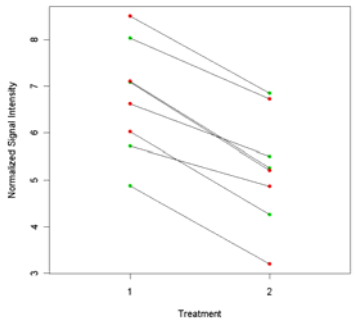
22

Observed Normalized Log Signal Intensities (NLSI) for One Gene

5.72 → 4.86	6.02 ← 4.26
7.08 → 5.20	7.11 ← 5.25
4.87 → 3.20	6.62 ← 5.50
8.03 → 6.72	8.50 ← 6.85

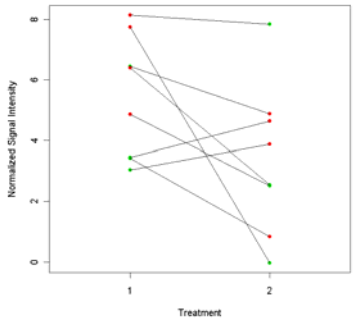
23

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



24

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



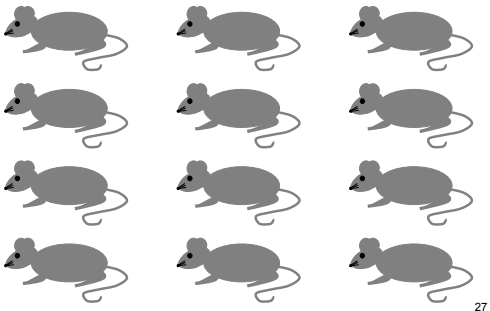
25

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.

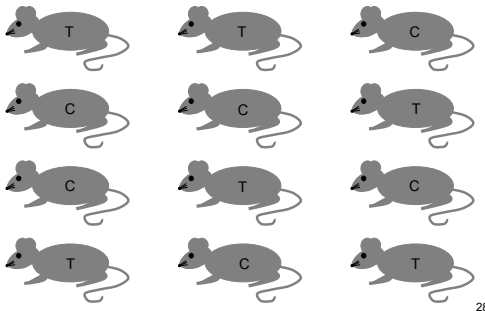
26

Assign 6 mice to each treatment completely at random



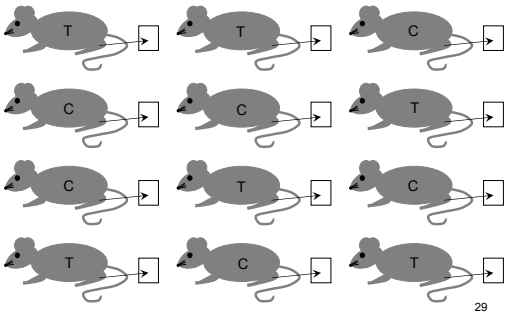
27

Assign 6 mice to each group completely at random



28

Measure Expression in Relevant Muscle Tissue with Affymetrix GeneChips



29

Normalized Log Scale Data

	Experimental Units											
Genes	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

30

Model for One Gene

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

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

μ=real-valued parameter common to all obs.

τ_i=effect due to i^{th} treatment

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

31

Gene 4: Data Analysis

$Y_{11}=8.6$ $Y_{21}=6.2$ $\overline{Y}_{1\cdot}-\overline{Y}_{2\cdot}=\tau_1-\tau_2+\bar{e}_{1\cdot}-\bar{e}_{2\cdot}=2.0$

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

$Y_{13}=9.1$ $Y_{23}=6.6$ $se(\overline{Y}_{1\cdot}-\overline{Y}_{2\cdot})=s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$

$Y_{14}=9.8$ $Y_{24}=6.8$ $=0.7949843\sqrt{\frac{1}{6}+\frac{1}{6}}$

$Y_{15}=7.9$ $Y_{25}=5.5$ $=0.4589844$

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

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

32

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

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

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

$2.0\pm 2.228*0.4589844$

$(0.98, 3.02)$

33

Gene 4: 95% Confidence Interval for Fold Change

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

$(e^{\overline{Y}_{1\cdot}-\overline{Y}_{2\cdot}-t^{(0.975)}_{n_1+n_2-2}se(\overline{Y}_{1\cdot}-\overline{Y}_{2\cdot})}, e^{\overline{Y}_{1\cdot}-\overline{Y}_{2\cdot}+t^{(0.975)}_{n_1+n_2-2}se(\overline{Y}_{1\cdot}-\overline{Y}_{2\cdot})})$

(2.7,20.5)

34

Gene 4: *t*-test

$Y_{11}=8.6$ $Y_{21}=6.2$ $\overline{Y}_{1\cdot}-\overline{Y}_{2\cdot}=\tau_1-\tau_2+\bar{e}_{1\cdot}-\bar{e}_{2\cdot}=2.0$

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

$Y_{13}=9.1$ $Y_{23}=6.6$ $t=\frac{\overline{Y}_{1\cdot}-\overline{Y}_{2\cdot}}{se(\overline{Y}_{1\cdot}-\overline{Y}_{2\cdot})}=\frac{2.0}{0.4589844}=4.3574.$

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

$Y_{15}=7.9$ $Y_{25}=5.5$ Compare to a *t*-distribution

$Y_{16}=7.4$ $Y_{26}=7.7$ with $n_1+n_2-2=10$ d.f. to obtain *p*-value≈0.001427.

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

35