



Differential Expression – I
jlsolka@gmail.com

BINF636 FALL08 DIFFERENTIAL
EXPRESSION - I (SOLKA)



Acknowledgements

- Tonight's lecture has been adapted from
 - "Advanced Analysis of Gene Expression Microarray Data," Aidong Zhang, World Scientific, 2006.



Example Datasets - I

- Samples were taken from human T (Jurkat) cells grown at 37° (for control samples) and 43° (to explore the influence of heat shock). The expression levels of 1,0416 genes were monitored by cDNA microarrays to identify heat-shock-regulated genes in human T cells. This data set is an example of paired data.
 - Controls are paired with a heat shock exposed sample
 - Log ratios of the two samples are compared
- Shena, M., D. Shalon, R. Heller, A. Chai, P. O. Brown and R. W. Davis (1996) Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. Proc. Natl. Acad. Sci. USA, 93:10614-10619.

Adapted from "Advanced Analysis of Gene Expression Microarray Data," Aidong Zhang, World Scientific, 2006.



Example Datasets - II

- Samples were taken from 14 multiple sclerosis (MS) patients. The expression levels of 4,132 genes were measured by cDNA microarrays for each patient prior to and 24 hours after interferon- β (IFN- β) treatment.
 - Paired data
 - 14 biological replicates
 - Pre and post treatment conditions
- Which genes were differentially expressed in MS treatment?
- Nguyen LT, Ramanathan M, Munschauer F, Brownschidle C, Krantz S, Umhauer M, Miller C, DeNardin E, Jacobs LD. Flow cytometric analysis of in vitro proinflammatory cytokine secretion in peripheral blood from multiple sclerosis patients. *J. Clin. Immunol*, 19 (3): 179-185, 1999.

Adapted from "Advanced Analysis of Gene Expression Microarray Data," Aidong Zhang, World Scientific, 2006.



Example Datasets - III

- Samples were taken from 15 MS patients and 15 age-and sex-matched controls
- Expression profiles for 4,132 humans were monitored by cDNA microarrays.
- unpaired data
- Two groups (MS and Controls)
- Are particular genes differentially expressed between the two groups
- Nguyen LT, Ramanathan M, Munschauer F, Brownschidle C, Krantz S, Umhauer M, Miller C, DeNardin E, Jacobs LD. Flow cytometric analysis of in vitro proinflammatory cytokine secretion in peripheral blood from multiple sclerosis patients. J. Clin. Immunol, 19 (3): 179-185, 1999.

Adapted from "Advanced Analysis of Gene Expression Microarray Data," Aidong Zhang, World Scientific, 2006.



Example Datasets - IV

- This dataset is the union of dataset II and dataset III
- Groups
 - MS, Controls, and IFN- β
- Multi-group data

Adapted from "Advanced Analysis of Gene Expression Microarray Data," Aidong Zhang, World Scientific, 2006.



Example Dataset V

- Consider the following subset of data related to four genes, G1, G2, g3, G4. Their expression levels (log transformed and normalized) in four control tissues C1, C2, C3, C4, and four test tissues, T1, T2, t3, T4 are given below.

	C1	C2	C3	C4	T1	T2	T3	T4
G1	9.011	9.064	9.067	9.008	8.944	9.087	8.963	9.074
G2	10.556	10.373	10.657	10.336	10.101	10.073	10.095	11.273
G3	11.967	12.014	11.757	12.101	11.604	11.782	11.503	11.861
G4	10.211	10.282	10.284	10.087	10.104	9.981	10.131	10.473

- Which genes are differentially expressed?

Adapted from "Advanced Analysis of Gene Expression Microarray Data," Aidong Zhang, World Scientific, 2006.



Some Possible Approaches

- k -fold change
- Parametric tests
- Nonparametric tests
- Methods to deal with multiple hypothesis tests
- One-way ANOVA
- Two-way ANOVA

Adapted from "Advanced Analysis of Gene Expression Microarray Data," Aidong Zhang, World Scientific, 2006.



Statistical Inference

- Population – The entire collection of individuals or objects about which information is desired.
 - In Dataset II the population is the set of all MS patients in the world.
- Sample – A subset of the collection of individuals or objects on which we have obtained some set of information.
 - In Dataset II those MS patients on which we have obtained gene expression measurements.
- Random Variable (RV) – A set of possible outcomes of an experiment along with associated probabilities.
 - We may view our observed set of gene expression values as instantiations of a RV.
- RVs probability density functions (pdfs) are usually parameterized
 - Ex – $X \sim N(\mu, \sigma^2)$
- Since we don't have access to all of the values that we would need to estimate these parameters, we estimated them with values (statistics, calculated on our samples) this is statistical inference.

Adapted from "Advanced Analysis of Gene Expression Microarray Data," Aidong Zhang, World Scientific, 2006.



Hypothesis Testing - I

- We are interested in testing whether gene i is differentially expressed.

$$H_0 : \bar{x}_i = \bar{y}_i$$

$$H_a : \bar{x}_i \neq \bar{y}_i$$

where \bar{x}_i and \bar{y}_i are the mean expression value measured on two groups

- H_0 is rejected in favor of H_a if the observed measurement strongly suggests that H_0 is false



Hypothesis Testing - II

Decision	Truth	
	H_0 is true	H_0 is false
H_0 was rejected	false positive (Type I error) α	true positive (correct decision) $1 - \beta$
H_0 was not rejected	true negative (correct decision) $1 - \alpha$	false negative (Type II error) β

Adapted from "Advanced Analysis of Gene Expression Microarray Data," Aidong Zhang, World Scientific, 2006.



Hypothesis Testing - III

- Def. – The probability of a type I error is usually denoted by α and is commonly referred to as the significance level of a test.
- Def. - The probability of a type II error is usually denoted by β .
- Def. – The power of a test is defined as
 - $1-\beta = 1 - \text{probability of a type II error} = \text{Pr}(\text{rejecting } H_0 | H_1 \text{ true})$



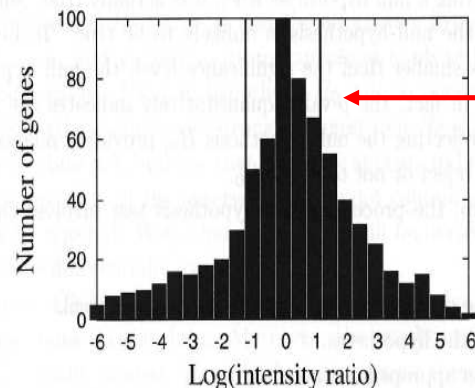
Hypothesis Testing - IV

- Steps in a Hypothesis Test
 1. Define the problem and specify the significance level
 2. Generate the hypothesis
 3. Choose an appropriate statistic
 4. Calculate the statistic value based on the observed data
 5. Calculate the corresponding p-value

Adapted from "Advanced Analysis of Gene Expression Microarray Data," Aidong Zhang, World Scientific, 2006.

k-fold Change - I

- Def. – Fold change is calculated as the average expression over all samples in a condition divided by the average expression over all samples in another condition.

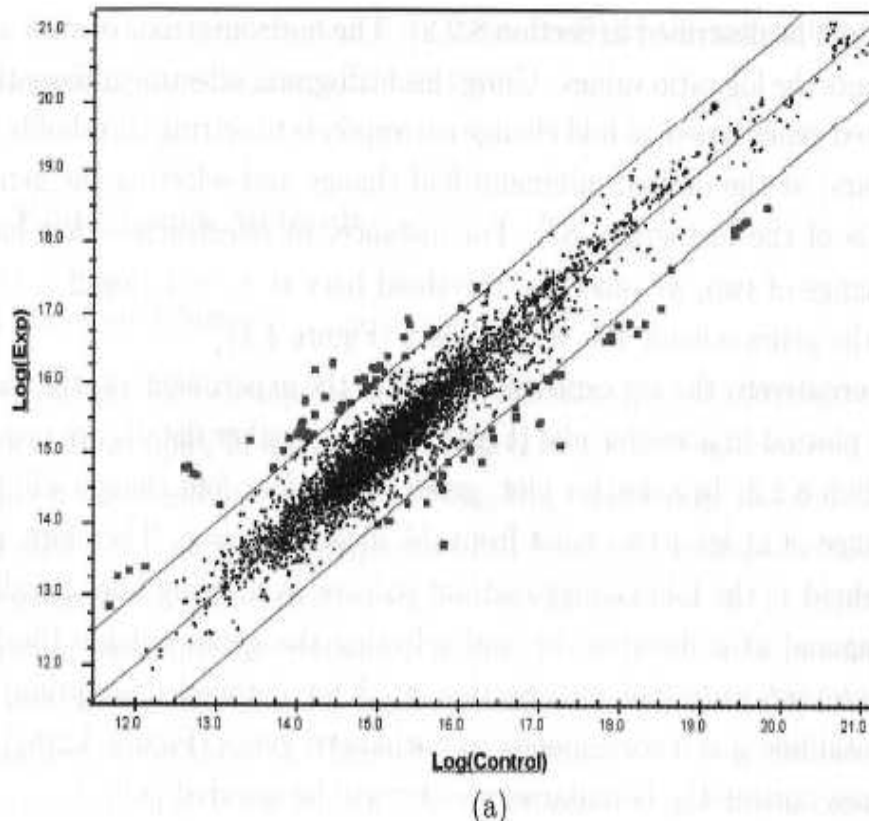


Genes that have at least
a two-fold $\log_2(2)=1$
 $\log_2(2)=-1$

Fig. 4.1 Histogram of log ratios and selection of genes with 2-fold change ($\log_2 2 = 1$).

Adapted from "Advanced Analysis of Gene Expression Microarray Data," Aidong Zhang, World Scientific, 2006.

k-fold Change - II

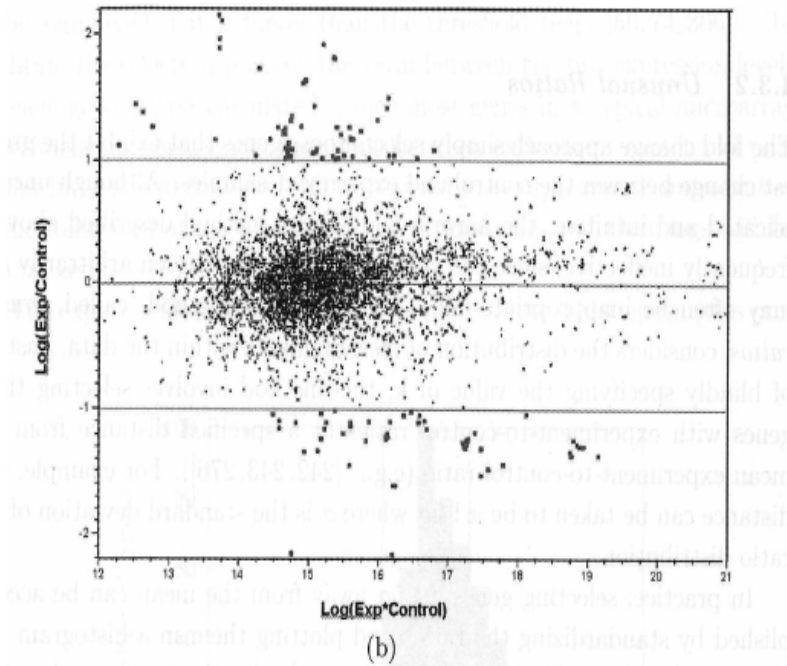


- log expression levels of experiment versus log expression levels of control
- Thresholds are set via the use of parallel lines to the line $y = x$

Adapted from "Advanced Analysis of Gene Expression Microarray Data," Aidong Zhang, World Scientific, 2006.

k-fold Change – III

ratio-intensity plot



y axis is given by

$$\log \frac{Cy3}{Cy5}$$

X axis is given by

$$\log(Cy3 * Cy5)$$

Adapted from "Advanced Analysis of Gene Expression Microarray Data," Aidong Zhang, World Scientific, 2006.

Unusual Ratios - I

- A method to help improve the k-fold approach
- Standardizing transformation

$$z = \frac{x - \mu}{\sigma}$$

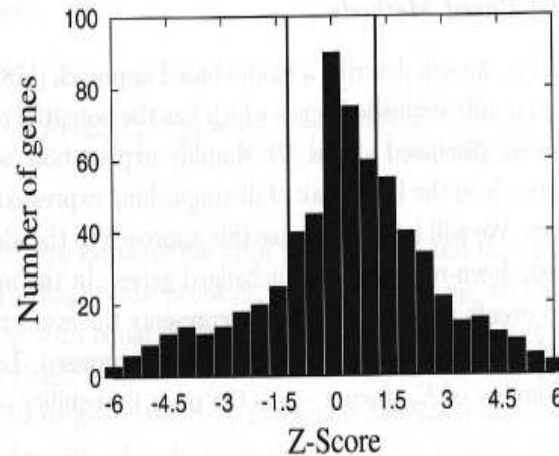
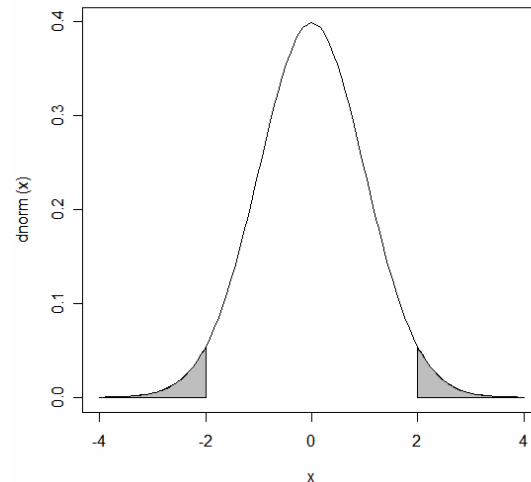


Fig. 4.3 Histogram of standardized log ratios and selection of genes with unusual ratios ($\pm 1.5\sigma$).

Adapted from "Advanced Analysis of Gene Expression Microarray Data," Aidong Zhang, World Scientific, 2006.

Unusual Ratios - II

- Assuming that the gene expression ratios are normally distributed and that the user chooses 2σ for the cutoff threshold then simple statistical theory tells us that we will have 4.56% of the genes.
 - $P(Z < -2) + P(Z > 2) = 0.0228 + 0.0228 = .0456$



Adapted from "Advanced Analysis of Gene Expression Microarray Data," Aidong Zhang, World Scientific, 2006.



Unusual Ratios - III

- So the method always identifies 4.56% of the genes as differentially expressed but what if none of the genes are differentially expressed?
- What if more than this amount are actually differentially expressed?

Adapted from "Advanced Analysis of Gene Expression Microarray Data," Aidong Zhang, World Scientific, 2006.

Model-based Methods - I

- First consider the simpler case of detecting where gene g is expressed or not expressed

E_g = gene g is expressed

\bar{E}_g = gene g is unexpressed

p = prior probability of E_g

$(1 - p)$ = prior probability of \bar{E}_g

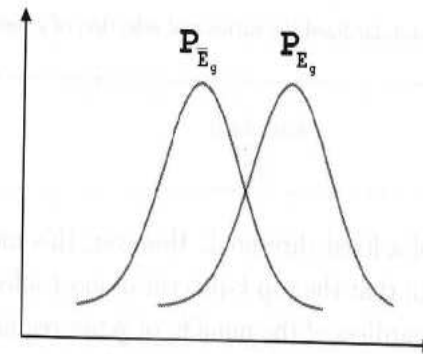


Fig. 4.4 Probability distributions of p_{E_g} and $p_{\bar{E}_g}$. Both distributions are assumed to be normal with equal variance.

Adapted from "Advanced Analysis of Gene Expression Microarray Data," Aidong Zhang, World Scientific, 2006.



Model-based Methods - II

$$\Pr(E_g | Y_g = y) = \frac{p * p_{E_g(y)}}{p * p_{E_g(y)} + (1 - p) * p_{\bar{E}_g(y)}}$$

Use EM algorithm to estimate

p, σ, μ_{E_y} , and $\mu_{\bar{E}_y}$ under the assumption of the individual distributions are normal with equal covariances.

Y_g is the observed ratio for gene g

Adapted from "Advanced Analysis of Gene Expression Microarray Data," Aidong Zhang, World Scientific, 2006.



Model-based Methods - III

- This methodology can be extended to the case of up-regulated, down-regulated, and unchanged

$$p_g(y) = p_1 * p_{Up_g}(y) + p_2 * p_{Down_g}(y) + p_3 * p_{Unchanged_g}(y)$$

- Pros
 - The model-based approach produces an unbiased minimum variance estimator as the sample size increases.
 - Can be used to test hypothesis about models and parameters
- Cons
 - Results become unreliable as sample size decreases
 - Results become unreliable as data deviates from normality

Adapted from "Advanced Analysis of Gene Expression Microarray Data," Aidong Zhang, World Scientific, 2006.



k-fold Summary

- Best suited to data without replication
- Remember our dataset - I

Adapted from "Advanced Analysis of Gene Expression Microarray Data," Aidong Zhang, World Scientific, 2006.



Parametric Tests – Paired t-Test – I (One-Sample Formulation)

- Applicable to datasets where we have paired data such as our dataset II
- If x_1 are the values before treatment and x_2 are the values after treatment then we may form the \log_2 of the ratio

$$\log_2 \left(\frac{x_1}{x_2} \right)$$

- We may use this ratio to formulate the paired t-Test as a one-sample t-Test
- Hypothesis
 - $H_0: \mu = 0$
 - $H_1: \mu \neq 0$

Adapted from "Advanced Analysis of Gene Expression Microarray Data," Aidong Zhang, World Scientific, 2006.

BINF636 FALL08 DIFFERENTIAL

EXPRESSION - I (SOLKA)



Parametric Tests – Paired t-Test – II (One Sample Formulation)

- Eq. 7.10 (One-Sample t Test for the Mean of a Normal Distribution with Unknown Variance (Two-Sided Alternative) To test the hypothesis $H_0: \mu = \mu_0$, versus $H_1: \mu \neq \mu_0$ with a significance level of α , the best test is based on

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

- If $|t| > t_{n-1, 1-\alpha/2}$ then we reject H_0
- If $|t| \leq t_{n-1, \alpha/2}$ then we accept H_0



Parametric Tests – Paired t-Test – III (One Sample Formulation)

- Eq. 7.11 (p-value for the One-Sample t Test for the Mean of a Normal Distribution (Two-Sided Alternative))
- Let

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

$$p = \begin{cases} 2\Pr(t_{n-1} \leq t), & \text{if } t \leq 0 \\ 2[1 - \Pr(t_{n-1} \leq t)], & \text{if } t > 0 \end{cases}$$



Parametric Tests – Paired t-Test – IV (Two Sample Formulation)

- Equation 8.4 (Paired t Test) – Denote the test statistic $\frac{\bar{d}}{s_d / \sqrt{n}}$ by t , where s_d is the sample standard deviation of the observed differences:

$$s_d = \sqrt{\frac{\left[\sum_{i=1}^n d_i^2 - \left(\sum_{i=1}^n d_i \right)^2 / n \right]}{(n-1)}}$$

where n = number of matched pairs

If $t > t_{n-1, 1-\alpha/2}$ or $t < -t_{n-1, 1-\alpha/2}$

Then H_0 is rejected. If

$$-t_{n-1, 1-\alpha/2} \leq t \leq t_{n-1, 1-\alpha/2}$$

then H_0 is accepted.



Parametric Tests – Paired t-Test – IV (Two Sample Formulation)

If $t < 0$, $p = 2x[\text{the area to the left of } t = \frac{\bar{d}}{s_d / \sqrt{n}} \text{ under a } t_{n-1} \text{ distribution}]$

If $t \geq 0$, $p = 2x[\text{the area to the right of } t = \frac{\bar{d}}{s_d / \sqrt{n}} \text{ under a } t_{n-1} \text{ distribution}]$



Unpaired t-Test - I

- This is like our dataset III
- Hypothesis
 - $H_0: \mu_1 = \mu_2$
 - $H_1: \mu_1 - \mu_2 = 0$
- Variants of the test
 - Equal variance
 - Unequal variance

Unpaired t-Test (Equal Variances)

- II

Equation 8.11 (Two-Sample t Test for Independent Samples with Equal Variances) Suppose that we wish to test the hypothesis $H_0: \mu_1 = \mu_2$ versus $H_1: \mu_1 \neq \mu_2$ with a significance level of α for two normally distributed populations, where σ^2 is assumed to be the same for each population. Compute the test statistic:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \text{ where}$$

$$s = \sqrt{\frac{[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]}{(n_1 + n_2 - 2)}}$$

If $t > t_{n_1 + n_2 - 2, 1 - \alpha/2}$ or $t < -t_{n_1 + n_2 - 2, 1 - \alpha/2}$ then H_0 is rejected.

If $-t_{n_1 + n_2 - 2, 1 - \alpha/2} \leq t \leq t_{n_1 + n_2 - 2, 1 - \alpha/2}$ then H_0 is accepted.



Unpaired t-Test (Equal Variances- p-value Computation) - III

Compute the test statistic.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \text{ where}$$
$$s = \sqrt{\frac{[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]}{(n_1 + n_2 - 2)}}$$

If $t \leq 0$, $p = 2 \times$ (area to the left of t under a $t_{n_1 + n_2 - 2}$ distribution)

If $t > 0$, $p = 2 \times$ (area to the right of t under a $t_{n_1 + n_2 - 2}$ distribution)



Unpaired t-Test (Unequal Variances) - IV

- Equation 8.21 Two-Sample t Test for Independent Samples with Unequal Variances (Satterthwaite's Method)

Compute the test statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

Compute the approximate degree of freedom d' , where

$$d' = \frac{\left(s_1^2 / n_1 + s_2^2 / n_2\right)^2}{\left(s_1^2 / n_1\right)^2 / (n_1 - 1) + \left(s_2^2 / n_2\right)^2 / (n_2 - 1)}$$

Round d' down to the nearest integer d''

1. If $t > t_{d'', 1-\alpha/2}$ or $t < -t_{d'', 1-\alpha/2}$ then reject H_0 .
2. If $-t_{d'', 1-\alpha/2} \leq t \leq t_{d'', 1-\alpha/2}$ then accept H_0



Unpaired t-Test (Unequal Variances) - V

- Equation 8.22 Computation of the p-Value for the Two-Sample Test for Independent Samples with Unequal Variances (Satterthwaite Approximation)

Compute the test statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

If $t \leq 0$, $p = 2 * (\text{area to the left of } t \text{ under a } t_{d''} \text{ distribution})$

If $t > 0$, then $p = 2 * (\text{area to the right of } t \text{ under a } t_{d''} \text{ distribution})$

Where d'' is as defined in equation 8.21.



Unpaired t-Test (Testing for Equality of Variances) - VI

- Equation 8.15 F-test for the Equability of two variances Suppose we wish to conduct a test of the hypothesis $H_0: \sigma_1^2 = \sigma_2^2$ versus $H_1: \sigma_1^2$ not equal σ_2^2 with significance level α . Compute the test statistic $F = s_1^2/s_2^2$. If

$$F > F_{n_1-1, n_2-1, 1-\alpha/2} \text{ or } F < F_{n_1-1, n_2-1, 1-\alpha/2}$$

Then H_0 is rejected. If

$$F_{n_1-1, n_2-1, \alpha/2} \leq F \leq F_{n_1-1, n_2-1, 1-\alpha/2}$$

H_0 is accepted.



Unpaired t-Test (Testing for Equality of Variances) - VII

- Equation 8.16 Computation of the p-value for the F test for the Equality of Two Variances Compute the test statistic $F = s_1^2/s_2^2$

If $F \geq 1$, then $p = 2 * \Pr(F_{n_1-1, n_2-1} > F)$

If $F < 1$, then $p = 2 * \Pr(F_{n_1-1, n_2-1} < F)$



Variants of the t-Test - I

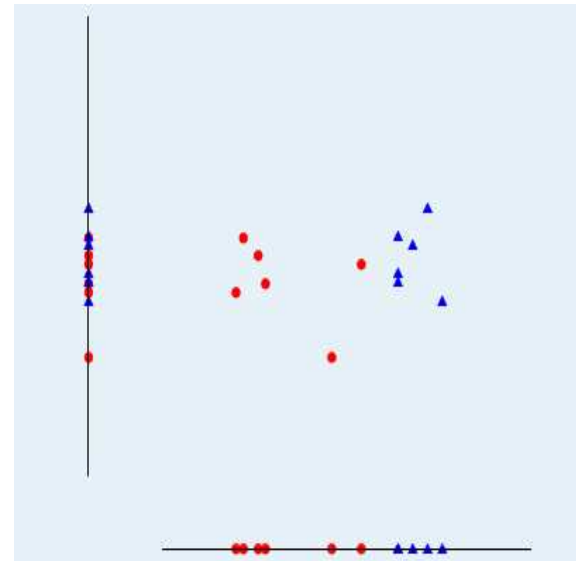
$$P(g) = \frac{\mu_1(g) - \mu_2(g)}{\sigma_1(g) - \sigma_2(g)}$$

- Large values of $|P(g)|$ indicate a strong correlation between gene expression and class distinctions
- Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," Science. 1999 Oct 15;286(5439):531-7.

Adapted from "Advanced Analysis of Gene Expression Microarray Data," Aidong Zhang, World Scientific, 2006.

Variants of the t-Test - II

$$F(g) = \frac{(\mu_1(g) - \mu_2(g))^2}{(\sigma_1^2(g) + \sigma_2^2(g))^2}$$



- This is a straight forward application of Fisher's Linear Discriminant
- Paul Pavlidis, Jason Weston, Jinsong Cai, William Noble Grundy: Gene functional classification from heterogeneous data. Recomb 2001, 249-255.

Adapted from "Advanced Analysis of Gene Expression Microarray Data," Aidong Zhang, World Scientific, 2006.



Nonparametric Tests - I

- The t-test has the tacit assumption that the underlying observations come from a normal distribution
 - The test has been known to have a good degree of robustness to this assumption
- For data that is known to blatantly violate these assumptions it is better to employ nonparametric tests



Nonparametric Tests - II

- Wilcoxon sign-rank test is the nonparametric equivalent of the paired t-Test
- Wilcoxon rank-sum test (aka Mann-Whitney test) is the nonparametric equivalent of the unpaired t-Test
- I will discuss the Mann-Whitney test a little on the next slide
- For a full discussion of these and other tests see
- Bernard Rosner, *Fundamentals of Biostatistics*, Duxbury Press, 2005.



Nonparametric Tests – III (Notional Description of the Mann Whitney Test)

1. Merge all observations from the two classes and rank them in ascending order.
2. Calculate the Wilcoxon statistics by adding all the ranks associated with the observations from the class with a smaller number of observations.
3. Find the p-value associated with the Wilcoxon statistic from the Wilcoxon rand sum distribution table (or use R to calculate the p-value).

Adapted from "Advanced Analysis of Gene Expression Microarray Data," Aidong Zhang, World Scientific, 2006.



Nonparametric Tests – IV (Comparison of Parametric and Nonparametric Tests)

- Nonparametric
 - Pros
 - Less distributional assumptions
 - Less sensitive to outliers
 - Cons
 - Less sensitive to outliers
 - Larger p-values
- Parametric
 - Pros
 - Superior p-values if the data meets the distributional assumptions
 - Cons
 - Distributional assumptions

Adapted from "Advanced Analysis of Gene Expression Microarray Data," Aidong Zhang, World Scientific, 2006.



Bootstrap Analysis – I

- Hypothesis testing steps
 - Form hypothesis
 - Choose test statistic
 - Calculate p-value
- Sometimes closed form methodologies to calculate the p-value are not available
- The bootstrap is a resampling based strategy to mitigate this problem

Adapted from "Advanced Analysis of Gene Expression Microarray Data," Aidong Zhang, World Scientific, 2006.



Bootstrap Analysis - II

- Construct a large number of random data sets via resampling from the original data
 - x_{ij} (measurement of gene i under experimental condition j) is randomly assigned one of the measurements from the dataset
 - This represents the original dataset but disturbs the correlation structure
- Calculate the t statistic for each gene in each bootstrap resampled dataset
- Use the standard t -distribution to find the minimum p -value among the genes
- More formally

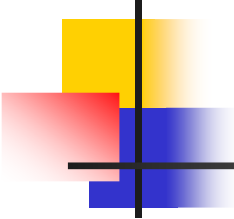
p_i = p -value i -th gene in original dataset

$p_j^{(b)}$ = p -value j -th gene on the b -th bootstrap resample

$\tilde{p}^{(b)} = \min \{ p_j^{(b)} \} =$ minimum p value in the b -th bootstrap resample

$p'_i = \frac{\text{number of random data sets with } \tilde{p}^{(b)} \leq p_i}{\text{total number of random datasets}} =$ adjusted p -value for the i -th gene

Adapted from "Advanced Analysis of Gene Expression Microarray Data," Aidong Zhang, World Scientific, 2006.



Bootstrap Analysis – III (How does this work?)

- If the null hypothesis were true then the original dataset would be very similar to each of the bootstrap samples.
 - Value of test statistic T (and associated p-value) as calculated on the real world data would appear as a typical value in the distribution of T (p-value) as evidenced in the bootstrap sample
- Alternatively
 - Value of T (and the p-value from the real world data) is “significantly abnormal” then we would be confident that the observed data are not formed by chance and we would reject H_0 .
- Pros
 - No distributional assumptions
 - More accurate than standard nonparametric tests
 - Can be used with any statistical measure

Adapted from "Advanced Analysis of Gene Expression Microarray Data," Aidong Zhang, World Scientific, 2006.



Multiple Testing - I

- This is the problem that we face due to the fact that we must run 1000s of hypothesis tests
- Type I Error = Reject H_0 given H_0 is true
- $P(\text{Type I Error}) = \alpha$
- $P(\text{not making a Type I Error}) = (1-\alpha)$
- Suppose we are analyzing N genes
 - $P(\text{globally correct}) = (1-\alpha)^N$
 - $P(\text{at least one error}) = 1 - (1-\alpha)^N$
 - For α small the expected number of Type I errors = αN
 - For N large this can be quite large
 - $N = 10000, \alpha = .05$
 - $N\alpha = 500$

Adapted from "Advanced Analysis of Gene Expression Microarray Data," Aidong Zhang, World Scientific, 2006.



Multiple Testing – II (Approaches to defeat the problem)

- Control global significance level (aka family-wise error rate (FWER))
 - Overly conservative and hence results in too many false negatives
- Control false discovery rate (Type I errors)
 - Assumes independence among the gene expression values
 - Likely to be false
- Permutation based methods
 - More robust to correlation among the genes
- Significance Analysis of Microarray (SAM) data
 - Combines FDR and permutation based methods

Adapted from "Advanced Analysis of Gene Expression Microarray Data," Aidong Zhang, World Scientific, 2006.



Family-Wise Error-Rate

- N = number of statistical tests performed
- p_1, \dots, p_N = p-values observed for the tests
- R = number of tests where H_0 is rejected
- V = number of false positives in these R decisions
- α_s = significance level for a single test
- We reject H_0 in each test if $p_i < \alpha_s$
- α_a = probability of committing at least one false positive among all the tests (i.e. $P(V > 0)$)
- α_a = global error rate aka family-wise error rate (FWER)
- α_s = gene-wise error rate aka per-comparison error rate (PCER)

Adapted from "Advanced Analysis of Gene Expression Microarray Data," Aidong Zhang, World Scientific, 2006.



Sidak Correction and Bonferroni Corrections

Sidak Correction

$$\alpha_a = 1 - (1 - \alpha_s)^N$$

$$\alpha_s = 1 - \sqrt[N]{1 - \alpha_a}$$

Bonferroni Correction

$$\alpha_a = 1 - (1 - \alpha_s)^N = 1 - (1 - N\alpha_s + \dots) = N\alpha_s$$

For α_s small each individual test must be performed at a significance level

of $\frac{\alpha_a}{N}$ in order to achieve an over all significance level of α_a

Adapted from "Advanced Analysis of Gene Expression Microarray Data," Aidong Zhang, World Scientific, 2006.



Holm's Step-wise Correction

- The previously discussed correction methods tend to be overly conservative (no positives at all)
- Holm's step-wise correction procedure
 - (1) Choose the global significance level α_a .
 - (2) Order the genes according to their p -values in the ascending order.
 - (3) Compare the p -value (p_i) of the i -th gene in the ordered list with threshold

$$\tau_i = \frac{\alpha_a}{N - i + 1}.$$

- (4) Report genes $1, \dots, k$ as differentially expressed genes at the chosen α_a significance level, where $k = \max_i \{p_i < \tau_i\}$, where $k = \max_i \{p_i < \tau_i\}$

$$\left(\text{the largest } i \text{ for which } p_i < \frac{\alpha_a}{N - i + 1} \right)$$

Adapted from "Advanced Analysis of Gene Expression Microarray Data," Aidong Zhang, World Scientific, 2006.



False Discovery Rate – I

- All the previous methods may be too stringent in that they attempt to control the probability of committing any Type I errors among the N tests.
- Benjamini and Hockberg decided to control the false discovery rate (FDR)

$$FDR = E \left[\frac{V}{R} \mid R > 0 \right] P[R > 0]$$

- If all the H_0 are true then $FDR = FWER$, in practice this is rarely the case and the more H_0 that are truly false then the smaller the FDR
- Control of FDR is more relaxed than the control of the FWER at the same level of significance.

Adapted from "Advanced Analysis of Gene Expression Microarray Data," Aidong Zhang, World Scientific, 2006.



False Discovery Rate – II

- (1) Chooses the global significance level α_a
- (2) Order the genes according to their p -values in the ascending order.
- (3) Compare the p -value (p_i) of the i -th gene in the ordered list with threshold

$$\tau_i = \frac{i}{N} \alpha_a$$

- (4) Report those genes as differentially expressed if $p_i < \tau_i$

Adapted from "Advanced Analysis of Gene Expression Microarray Data," Aidong Zhang, World Scientific, 2006.

False Discovery Rate – III

(Storey and Tibshirani)

- Storey and Tibshirani noted that an adjustment is only necessary when there are positive findings (i.e. cases where H_0 is rejected)

$$pFDR = E \left[\frac{V}{R} \mid R > 0 \right]$$

R and V are the number of positive findings and the number of false positives respectively.

V is estimated via a permutation procedure

- (1) Construct B permuted data sets by changing the class labels of samples
- (2) Suppose an average of R^* of genes having the p -values smaller than the threshold α_s over the B permuted datasets
- (3) Assume that there are no true positives in any permuted dataset and hence the expected number of false positives is R^*
- (4) Estimate $pFDR$ naturally as

$$pFDR = \frac{R^*}{R}$$

Adapted from "Advanced Analysis of Gene Expression Microarray Data," Aidong Zhang, World Scientific, 2006.



Permutation Correction - I

- Neither FWER or FDR considers the correlation structure among the data
 - A group of three genes may participate in the same pathway

- (1) Permute class labels of the samples
- (2) Compute p -values wrt a statistic such as the t – statistic for all the genes
- (3) Correct these p -values using Holm's stepwise method.
- (4) Repeat (1-3) for a sufficiently large number of times
- (5) Calculate adjusted p -values for each gene

$$\text{p-value for gene } i = \frac{\text{number of permutations for which } p_i^{(b)} \leq p_i}{\text{total number of permutations}}$$

$p_i^{(b)}$ is the corrected p -value by Holm's step-wise method for permutations b , and p_i is the value of the test statistic for the real dataset.

- Westfall P.H., Young S.S. (1993) *Resampling-based Multiple Testing*. Wiley, New York.

Adapted from "Advanced Analysis of Gene Expression Microarray Data," Aidong Zhang, World Scientific, 2006.



Permutation Correction - II

- Pros
 - Takes into account correlation structure
- Cons
 - Slow
- Dudoit, S. et al. (2000). "*Statistical Methods for Identifying Differentially Expressed Genes in Replicated cDNA Microarray Experiments*". Technical Report # 578.
 - Application of the procedure to gene expression data



Sam: Significance Analysis of Microarrays – I

- Summary of where we are so far
 - Bonferroni correct is too stringent
 - W-Y is too stringent
 - FEWR and FDR assume independence
- Significance Analysis of Microarrays (SAM)
 - Assign score to each gene according to its change in gene expression
 - Genes with scores greater than a threshold are considered potentially significant
 - SAM uses permutation of the measurements to estimate the false discovery rate (pFDR)
 - Score threshold for genes is then adjusted iteratively according to the pFDR until a set of significant genes is identified



Sam: Significance Analysis of Microarrays – II

SAM's Score

$$d(i) = \frac{\bar{x}_1(i) - \bar{x}_2(i)}{s(i) + s_0}$$

Pooled Variance Estimate

$$s(i) = \sqrt{\frac{\frac{1}{n_1} + \frac{1}{n_2}}{n_1 + n_2 - 2} \left(\sum_p (x_p(i) - \bar{x}_1(i))^2 + \sum_q (x_q(i) - \bar{x}_2(i))^2 \right)}$$



Sam: Significance Analysis of Microarrays – III

- As compared to the standard t-statistic SAM uses a fudge term s_0 in the denominator
- SAM seeks s_0 so that the dependence of $d(i)$ on $s(i)$ is as small as possible
- One uses a sliding window across all genes and chooses a value of s_0 so that the cov ($d(i)$) is approximately constant

Sam: Significance Analysis of Microarrays – IV

- Given s_0 , $d(i)$ is computed for each gene
- Threshold for significant genes and expectation of false positives is computed as follows
 - Permute the columns of the given data matrix, X , and assign the first n_1 columns to class 1 and the remaining n_2 columns to class 2
 - B permutations will be performed

(1) Sort the $d(i)$ values of the original data in descending order

$$d(1) \geq d(2) \geq \dots \geq d(N)$$

(2) For each permutation B compute

$$d_b(1) \geq d_b(2) \geq \dots \geq d_b(N)$$

(3) Compute the expected order statistic

$$d_E(i) = \sum_{b=1}^B d_b(i) / B$$

(4) Identify those genes whose value is substantially larger than its expected value

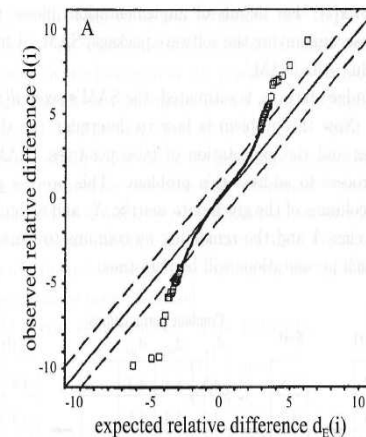


Fig. 4.6 The scatter plot of $d(i)$ vs. $d_E(i)$ to select potential significant genes (Figure is from [289]).



Sam: Significance Analysis of Microarrays – V

- The genes identified via the plot are candidate expressed genes
- To estimate the error rate SAM uses

$$pFDR = \frac{\frac{1}{B} \sum_{b=1}^B C_b}{C}$$

where C_b is the number of potentially significant genes



Methods Not Discussed

- One-way ANOVA
- Two-way ANOVA
- Others
- Thanks for your time
- I will return to talk about clustering and classification with you
- Jeff Solka
 - jlsolka@gmail.com