



Statistical Analysis of Microarray Gene Expression Data

INBRE Microarray Workshop I
Weidong Zhang, Ph.D.

Understand sources of variation

- Treatment
 - Drug
 - Tissue
 - Strain
- Systems/devices
- Random noise

Replication

- Experimental unit
 - Subjects (person, mouse, fish, or a thing etc.) independently receive a treatment condition
- Biological replicates (True replicates)
 - Experimental unit (Most of times)
- Technical replicates (Pseudo-replicates)

Hypothesis testing

- Hypothesis testing

$$H_0: \mu_1 = \mu_2 \dots = \mu_c = \mu$$

$$H_a: \text{Not } H_0$$

- Contrast
 - Any linear combination of means

$$\sum C_i \mu_i$$

With $\sum C_i = 0$

Statistical models (ANOVA)

- One-way ANOVA

$$y_{ij} = T_i + e_{ij}$$

where $i = \text{Treatment}, i = 1, 2, \dots, t$

$j = \text{Samples/Microarrays}, j = 1, 2, \dots, s$

- Two-way ANOVA

$$y_{ijk} = T_i + S_j + TS_{ij} + e_{ijk}$$

where $i = \text{Treatment}, i = 1, 2, \dots, t$

$j = \text{Strain (or Sex etc.)}, j = 1, 2, \dots, s$

$k = \text{Samples/Microarrays}, k = 1, 2, \dots, r$

T test

- T test

$$t = (X_{g1} - X_{g2}) / SE$$

$$X_{g1} = \log_2(\text{Exp}_{g1})$$

$$X_{g2} = \log_2(\text{Exp}_{g2})$$

- Gene specific t test
- Global t test
- Something in between
 - e.g. SAM (Significance Analysis of Microarrays)

F test

- Generalization of t test

$$F_1 = \Delta_g / \hat{\sigma}_g^2,$$

$$F_2 = \Delta_g / \frac{1}{2} (\hat{\sigma}_g^2 + \hat{\sigma}_{\text{pool}}^2),$$

$$F_3 = \Delta_g / \hat{\sigma}_{\text{pool}}^2,$$

$$F_S = \Delta_g / \tilde{\sigma}_g^2.$$

Permutation test

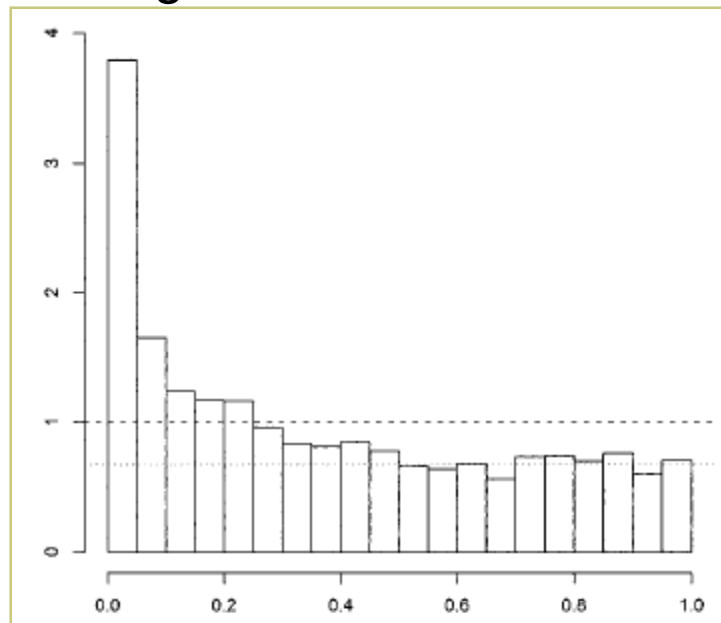
- Issues of statistical test
 - Model assumption for regular statistics
 - No standard form for some derived statistics
- Permutation test
 - Permute sample labels
 - For each shuffled data, calculate the statistics selected
 - Repeat r times
 - Calculate the p value by $\text{num_statistics_greater_than_the_observed}/r$
- Advantage
 - Distribution-free
 - Applied to any statistic
- Disadvantage
 - Small sample size
 - More computation

Multiple testing

- Two errors
 - Type I: probability of reject null when null is true
 - Type II: probability of failing to reject null when alternative is true
- False positive rate
 - The proportion of truly nulls that are called significance
- Controlling family-wise error rate (FWER)
 - Bonferroni
 - Westfall and Young step-down adjustment
 -

False Discovery Rate (FDR)

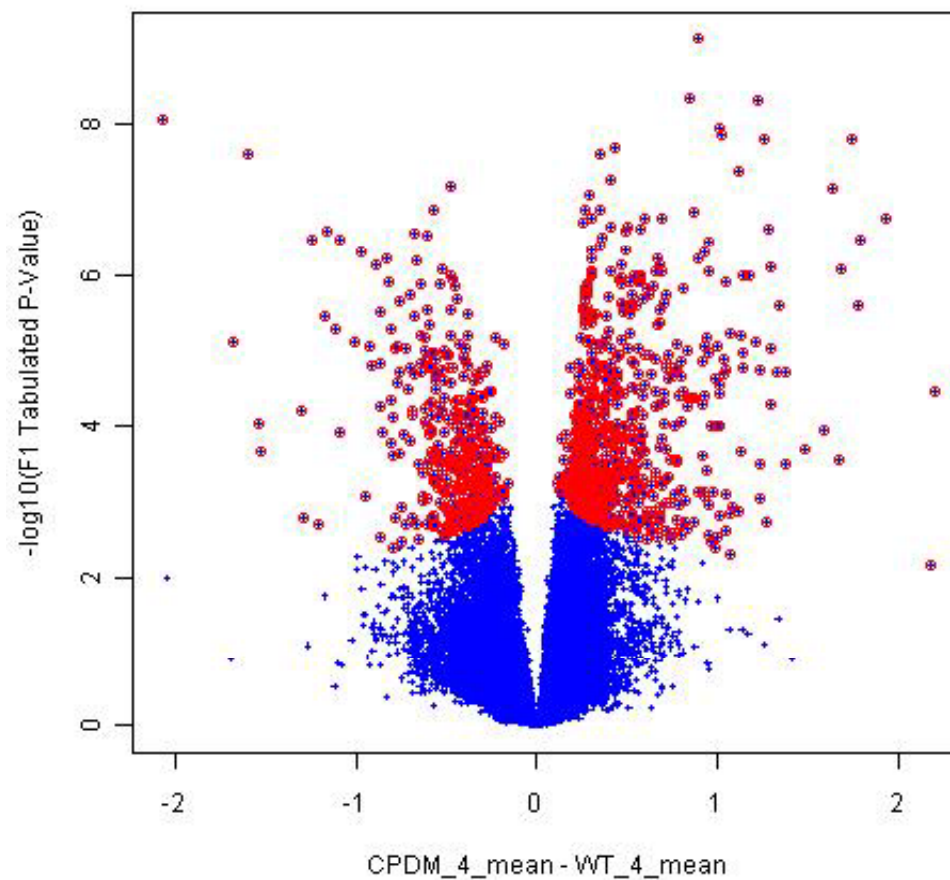
- FDR
 - The proportion of significant genes that are truly null
- Q value
 - Measure of evidence of significance automatically taking into account multiple testing



Storey and Tibshirani, 2003

Volcano plot

Volcano Plot - Contrast 2 - q-value<0.05 - RMA



Summary of the workflow

- Examine your data
- Identify the sources of variation
- Select appropriate models
- Perform statistical testing
- Control FDR and generate gene list
- Interpret results