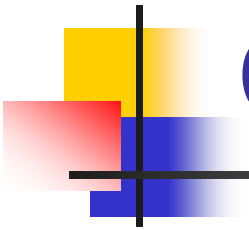
A New Hypothesis to Test Minimal Fold Changes of Gene Expression Levels

Jen-pei Liu, PhD
Division of Biometry, Department of Agronomy
National Taiwan University

at

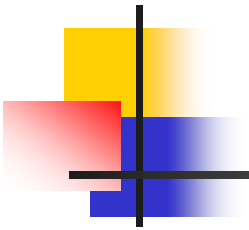
The 5th International Conference on Multiple Comparison Procedures

July 9-July 11, 2007
Vienna, Austria



Outlines

- Introduction
- Current Approaches
- Interval Hypothesis
- Simulation Studies
- Examples
- Summary and Discussion



Introduction

A set of 30 numbers

120, 146, 138, 127, 94, 123, 118, 136,
113, 87, 102, 101, 145, 137, 153, 109,
134, 127, 113, 133, 141, 136, 126, 114,
117, 121, 123, 132, 135

Introduction

- Diagnosis of Diabetes Mellitus
  - Fasting Plasma Glucose > 126 mg/dL
  - 15 subjects (50%)
- Diagnosis of Hypertension
  - Systolic blood pressure > 140 mm Hg
  - 4 subjects (13.3%)

Introduction

- Outcomes of Diabetes Mellitus
  - Acute complications
    - diabetic ketoacidosis, hypoglycemia, etc.
  - Chronic complications
    - macrovascular diseases: coronary artery disease, stroke, peripheral artery diseases, etc.
    - microvascular diseases: retinopathy, neuropathy, nephropathy, etc.

Introduction

- Outcomes of Hypertension
  - Heart: left ventricular hypertrophy, angina, congestive heart failure, etc.
  - Brain: stroke
  - Kidney
  - Peripheral artery diseases
  - Retinopathy

Introduction

- Normal Ranges
  - ALT: 15-35 IU/L
  - AST: 5-40 IU/L
  - LDH: 6.6-8.1 mg/dL
  - BUN: 0.7-1.5 mg/dL
  - Uric Acid: 3.5-7.9 mg/dL (male)
    2.6-6.0 mg/dL (female)

Introduction

- A threshold of 126 mg/dL for FPG or of 140 mm Hg for HP differentiates the occurrence of clinical outcomes.
- FPG or BP are surrogate endpoints of clinical endpoints
- A clinical measurement means nothing without clinical implications or interpretations.

Introduction

- Gene expression levels are surrogates for phenotypes or indicators of clinical outcomes by intervention
- A measurement of gene expression levels means nothing without biological implications or interpretations.
- Thresholds can be determined in expression levels for differentiation of the occurrence of phenotypes and treatment outcomes.

Current Approaches

- Hypothesis of Equality

  Ho: $\mu_{Ti} - \mu_{Ci} = 0$

  vs.

  Ha: $\mu_{Ti} - \mu_{Ci} \neq 0$, i=1,...,G.

Current Approaches

- Notations:
  - Y_{ijk}: normalized log-transformed (base 2) intensity for gene i on array j receiving treatment k, i=1,…,G; j=1,…,n; k=T, C.
  - Sample means and variance

$$\overline{Y}_{ik} = \frac{1}{n_{ik}} \sum_{j=1}^{n_{ik}} Y_{ijk}$$

$$s_{ik}^2 = \frac{1}{n_{ik}} \sum_{j=1}^{n_{ik}} \left(Y_{ijk} - \overline{Y}_{ik} \right)^2$$

Current Approaches

Unpaired t-statistics

$$t_i = \frac{\bar{Y}_{iT} - \bar{Y}_{iC}}{\sqrt{s_{pi}^2 \left(\frac{1}{n_{iT}} + \frac{1}{n_{iC}} \right)}}$$

Issues: Issues:

G is large ≈ tens of thouands

n_{iT} or n_{iC} are small < 10.

Current Approaches

- Adjustment of p-values by multiple comparison procedures
  - Bonferroni method
  - Various methods for false discovery rate
- Significance Analysis of Microarray (SAM)
  - For gene with very low variations in expression levels, the pooled sample may be extremely small and t-statistics may be artificially inflated.
  - Add an empirical constant to the pooled sample variance in order to avoid such scenarios.

Current Approaches

- Fixed Fold Change Rule
  - A gene is differentially expressed if

$$\bar{Y}_{iT} - \bar{Y}_{iC} > C_i$$

or

$$\bar{Y}_{iT} - \bar{Y}_{iC} < -C_i'$$

where $-C_i'$ and C_i are pre-specified minimal requirement for biologically meaningful fold changes.

Current Approaches

- Combined Fold Change rule
  - Affymetrix (2001) and the US Microarray Quality Control Project (MAQC, 2006)
  - Fold change + p-value
  - Observed fold change exceeds the pre-specified meaningful threshold.
  - P-value of the unpaired two-sample t-statistic is less than 0.04.

Current Approaches

- Hypothesis of equality
  - Qualitative
  - Only detect the existence of the difference in mean expression levels
  - Difference in mean expression levels does not imply that genes are differentially expressed for differentiation of the occurrence of phenotypes or treatment outcomes.

Interval Hypothesis

- Interval Hypothesis

$$H_0: -C_i' \le \mu_{iT} - \mu_{iC} \le C_i$$

vs.

$$H_1: \mu_{iT} - \mu_{iC} < -C_i' \text{ or } \mu_{iT} - \mu_{iC} > C_i, \; i=1,...,G$$

- Parameter space
  - $[-C_i', C_i]$: region of no differential expression
  - $(-\infty, -C_i')$: region of under-expression
  - (C_i, ∞): region of over-expression

Interval Hypothesis

- Two One-sided Tests Procedures

$$t_{Ui} = \frac{\bar{Y}_{iT} - \bar{Y}_{iC} - C}{\sqrt{s_{pi}^2\left(\frac{1}{n_{iT}} + \frac{1}{n_{iC}}\right)}} > t_{(\alpha/2, n_{iT}+n_{iC}-2)}$$

or

$$t_{Li} = \frac{\bar{Y}_{iT} - \bar{Y}_{iC} + C}{\sqrt{s_{pi}^2\left(\frac{1}{n_{iT}} + \frac{1}{n_{iC}}\right)}} < -t_{(\alpha/2, n_{iT}+n_{iC}-2)}$$

Interval Hypothesis

- Rejection Region

$$\Re = \left\{ \left(\bar{d}_i, s_{pi} \right) \middle| \; \bar{d}_i > C + t \times m_i \times s_{pi} \;\; \text{or} \;\; \bar{d}_i < -C - t \times m_i \times s_{pi} \right\}$$

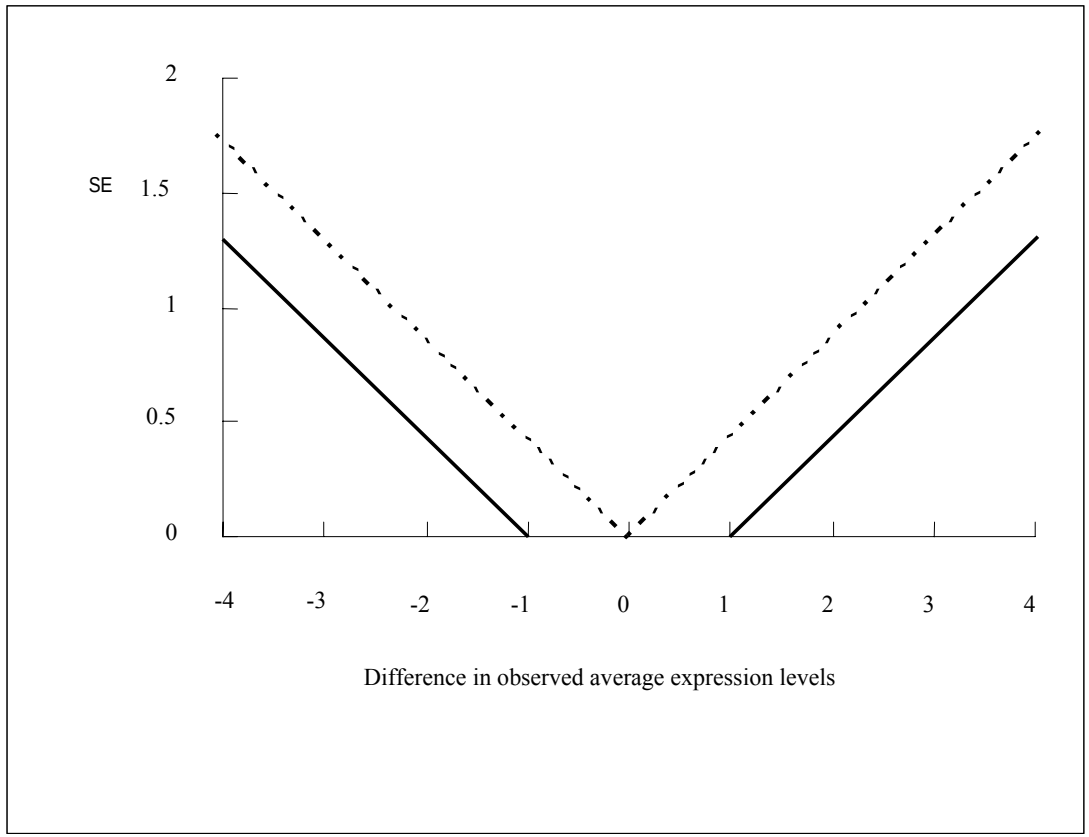
- Combined fold change rule

$$\left| \bar{d}_i \right| > \max \{ C, \, t \times m_i \times s_{pi} \}$$

- Two one-sided test procedure

$$\left| \bar{d}_i \right| > C + t \times m_i \times s_{pi}$$

Interval Hypothesis



Interval Hypothesis Based on Multivariate Permutation Test

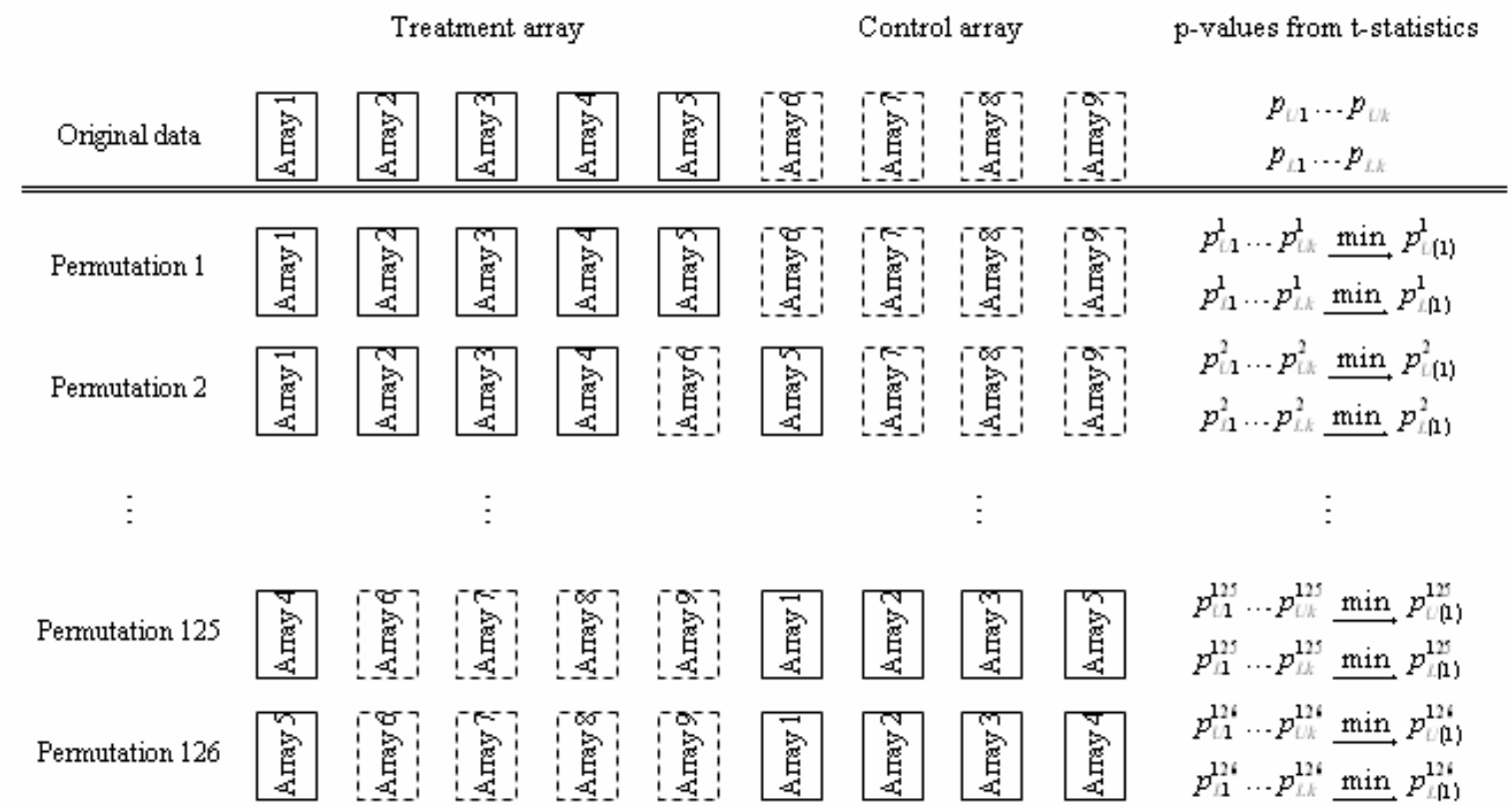


Fig 3.3 A procedure of the multivariate permutation test based on the two one-sided test

Interval Hypothesis Based on Multivariate Permutation Test d

After permutation, we can also get the paired $p^j_{U\,i}, p^j_{L\,i}$ for g genes in an array. After sorting the upper and lower p-values separately, we have to find the smallest p-value $p^j_{U(1)}, p^j_{L(1)}$ of each permutation. Gene i is differentially expressed between the test and control samples at the α significance level if

$$\frac{number\ of\ \ permutations\ where\ p^j_{U(1)} \le p_{Uk}}{\binom{n_{iT}+n_{iC}}{n_{iT}}} \le 0.025$$

$$or\ \frac{number\ of\ \ permutations\ where\ \ p^j_{L(1)} \le p_{Lk}}{\binom{n_{iT}+n_{iC}}{n_{iT}}} \le 0.025$$

Simulation Studies

- Fortran 90 and IMSL STAT/LIBRARY
- Model: Tsai, et al. (2003)
- Number of genes: 500, 1000, 2000
- 10% of genes are truly differentially expressed and 90% of genes are truly unexpressed
- Number of replicates: 3, 4, 5, 8, 10
- Difference:0, $\pm$ 0.5, $\pm$1.0, $\pm$2.0, $\pm$3.0
- Multiplicative error: 0.1 and 0.3
- Additive error: 0.5 and 1.0
- Number replications: 1000

Simulation Studies

- Evaluation criteria
  - Overall type I error rate: proportion of incorrect identification of at least one truly unexpressed genes
  - Average type I error rate: average of the ratios of the number of falsely identified genes to the total number of the truly unexpressed genes
  - Average power: average of the ratios of the number of identified differentially expressed genes to the total number of the truly expressed genes

Simulation Studies

- Except for Bonferroni adjustment, none of methods can control overall type I error rate
- The average type I error rates for t-test and combined fold-change rule are around 4.5% to 6.0%.
- Most of average type I error rates based on the interval hypothesis are below 0.3% for t-test
- The average type I error rates for the multivariate permutation method based on the interval hypothesis range from 0.08% to 3.0%.

Simulation Studies

- Average power increases as the sample size increases or the variation decreases
- Except for Bonferroni procedure, the average powers of the two one-sided tests procedures based on t-test and multivariate permutation method are smaller than all other procedures
- When multiplicative error is small, the average power of multivariate permutation is greater than the t-test

Simulation Studies

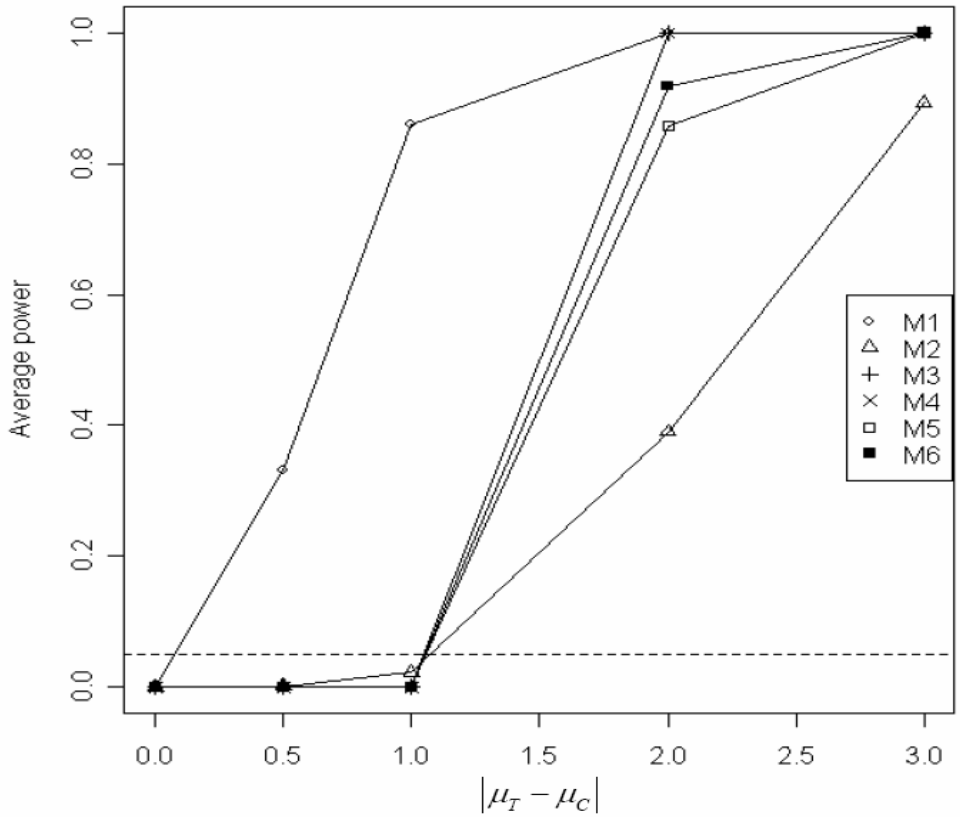
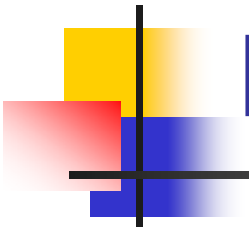
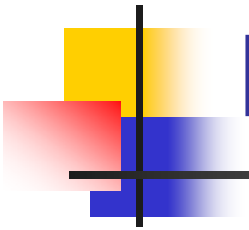


Figure 2 Average power curves of six methods in Model I for G=1000, n=5, ϕ^2=0.1, σ^2=0.5



Example

- Data: Luo et al. (2001)
  - 16 prostate samples
  - 9 samples of benign prostatic hyperplasia
  - 6500 human cDNAs (6112 unique genes)
  - Common reference design
  - Quality score > 0 with 3 replicates each: 5854 genes
  - Log normalized gene expression levels
  - http://research.nhgri.nih.gov/microarray/



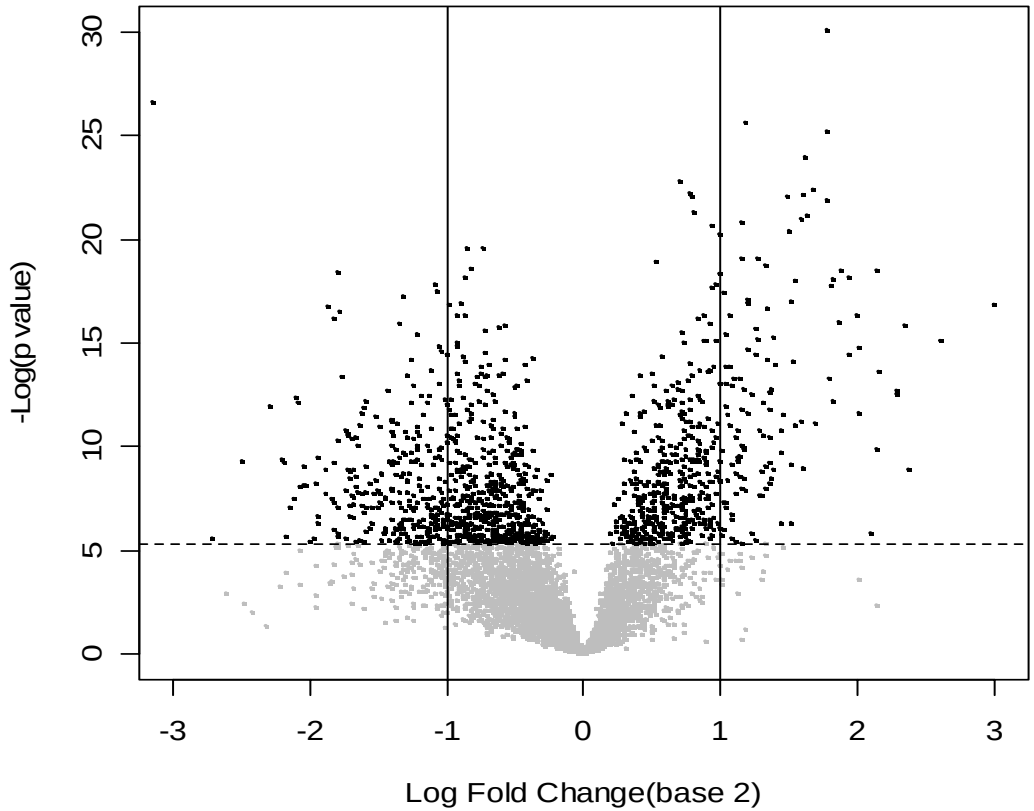
Example

- Numbers of identified genes
  - Unpaired t-test: 1722(29.42%)
  - Bonferroni adjustment: 47 (0.83%)
  - Fixed fold-change rule: 597 (10.20%)
  - Combined fold-change rule: 443 (7.57%)
  - Two one-sided tests procedures
    - T-test: 47 (0.83%)
    - Multivariate permutation: 181 (3.09%)

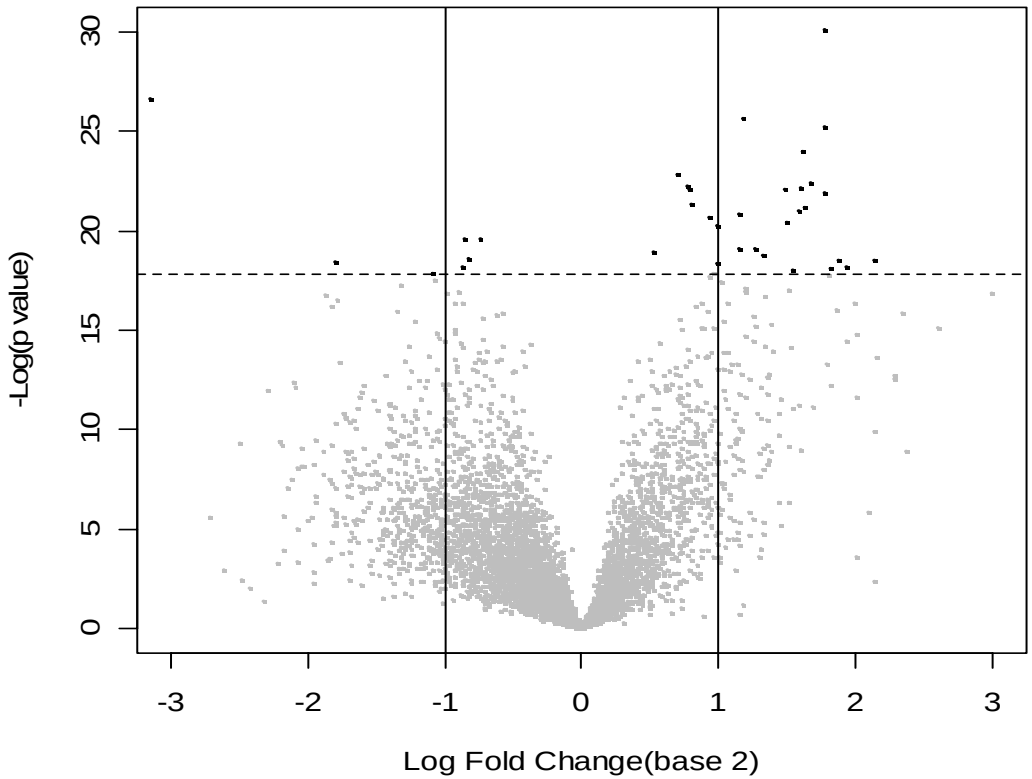
Example

- Unpaired t-test: 1247/1722 genes (72.46%) with fold changes smaller than ±1
- Bonferroni adjustment: 16/47 genes (34.04%) with fold changes smaller than ±1
- Fixed fold-change rule: 122/597 genes (20.44%) with p-values > 0.05
- Combined fold-change rule: clustering around p-value =0.05 and a fold-change around ±1
- Two one-sided tests procedures
  - t-test: minimal fold-change is ±1.5 with p-value < 0.05
  - Multivariate permutation method: 2/181 genes (1.10%) with fold changes smaller than ±1 due to fewer replicates and genes

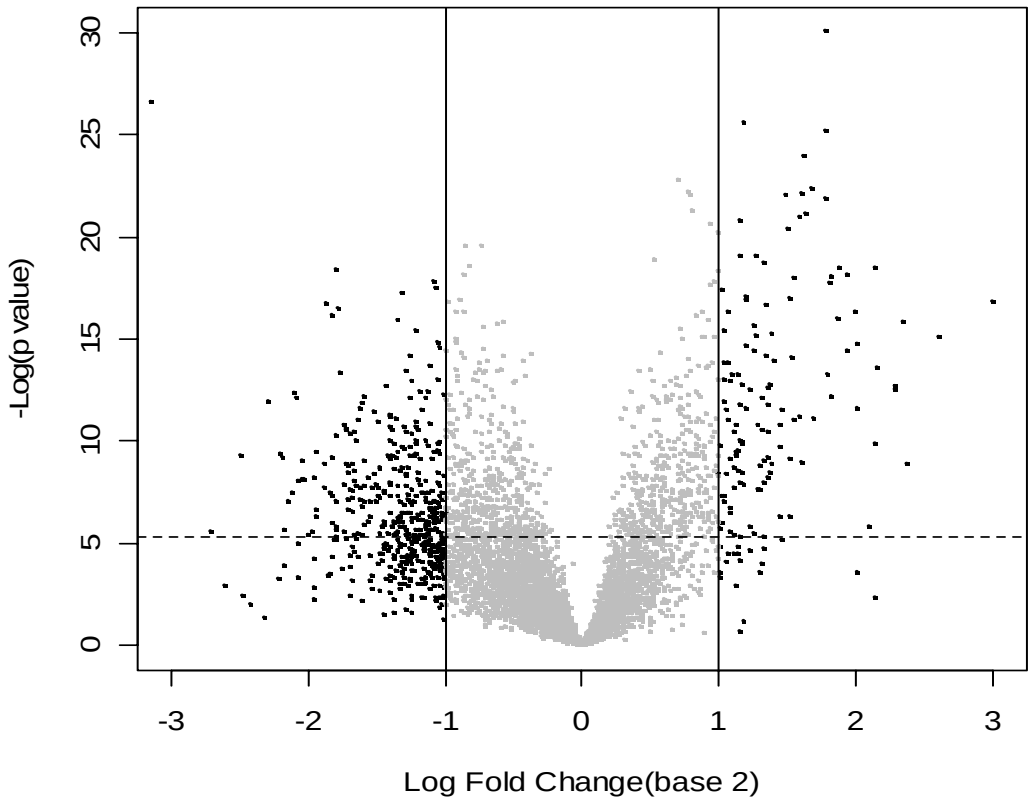
Example (Un-paired t-test)



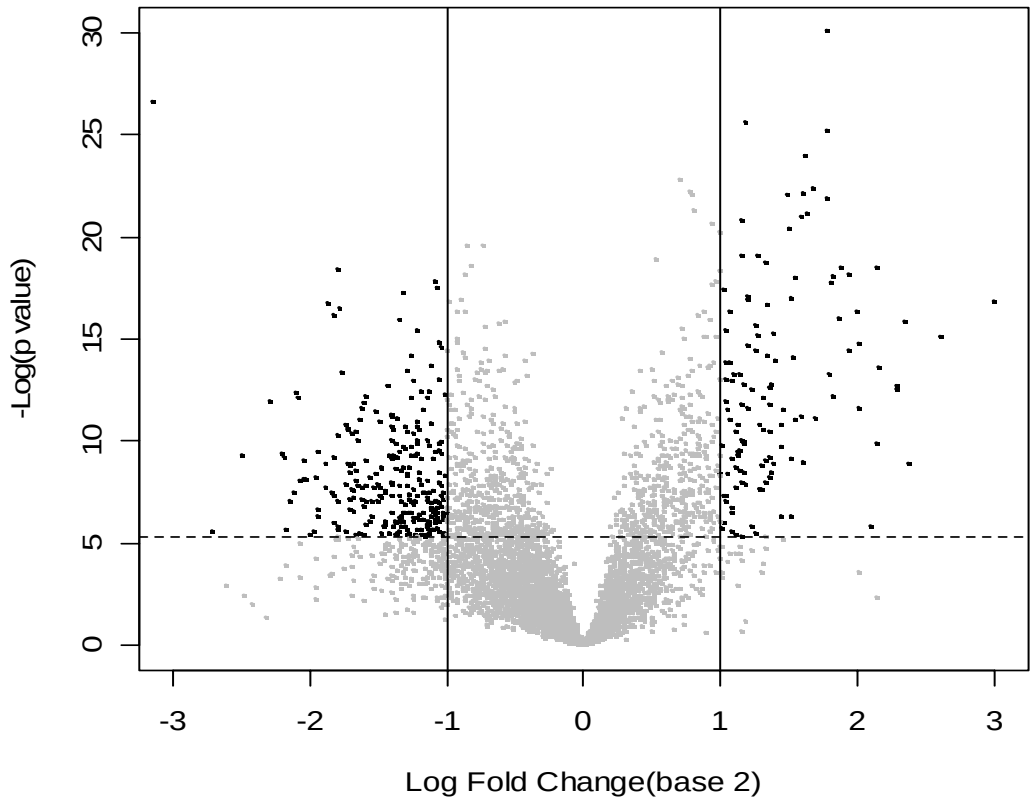
Example (Bonferroni Adjustment)



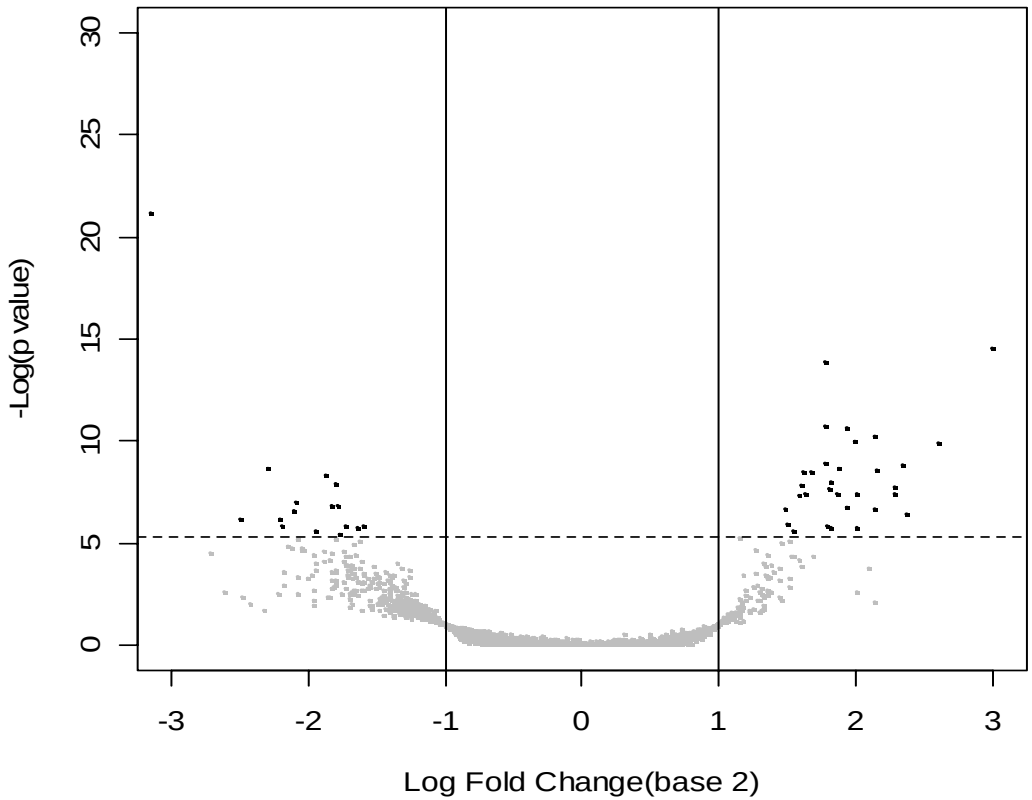
Example (Fixed Fold-Change Rule)



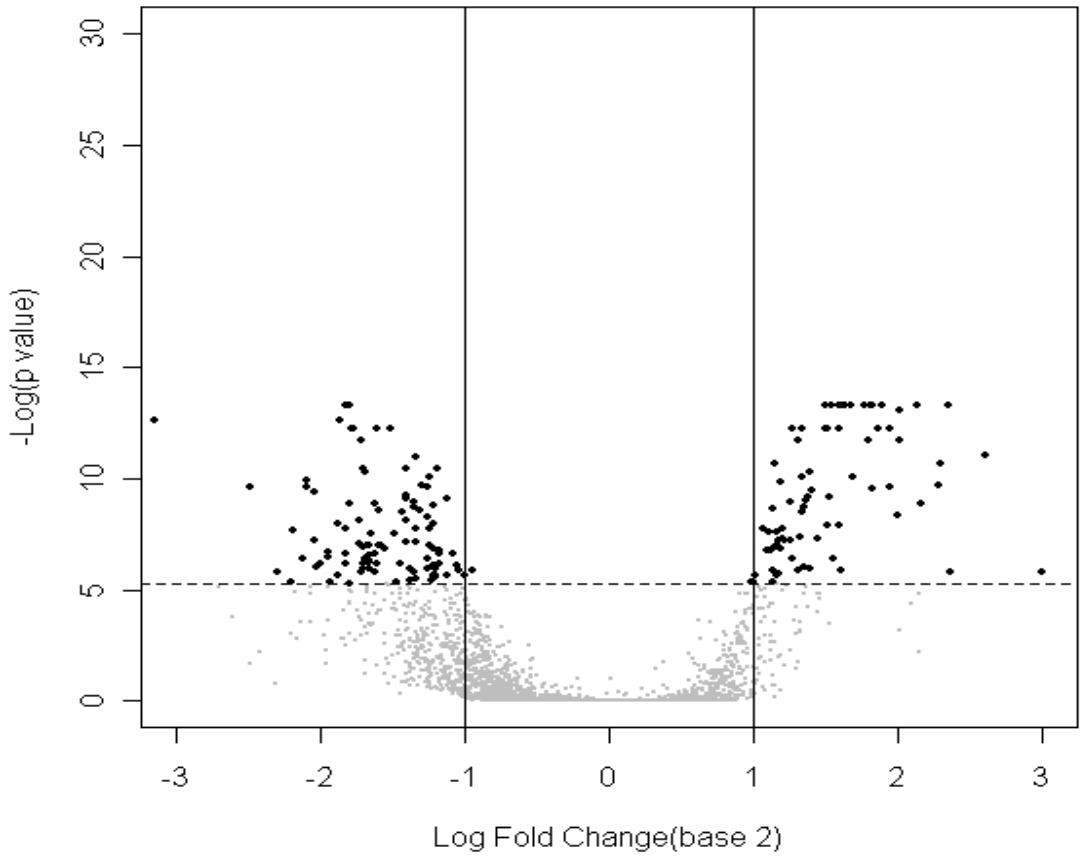
Example (Combined fold-change rule)



Example (interval hypothesis – t test)



Example (interval hypothesis – multivariate permutation)



Summary and Discussion

- Hypothesis of equality is for qualitative determination of existence of difference in mean expression levels: No vs. Yes.
- Hypothesis of equality fails to take into consideration the magnitude of expression levels.

Summary and Discussion

- Expression levels of differentially expressed genes should correlate with differential phenotypes or clinical outcomes intervened by targeted therapy.
- Identification of differentially expressed genes must take the magnitudes of expression levels into account.

Summary and Discussion

- Following the non-inferiority hypothesis in clinical trials, biologically meaningful changes in magnitudes can be used as thresholds for identification of differentially expressed genes.
- Interval hypothesis takes into account both biologically meaningful changes and variability.
- Multivariate permutation can also take interrelationship in expression levels among genes

Thank you very much