Evaluation of current methods of testing differential gene expression and beyond

Yi-Ju Li, Ling Zhang, Marcy Speer, Eden Martin

Duke Center for human Genetics

Bioinformatics program in NCSU

Introduction

- Often little replication in microarray experiments
  - what's the appropriate variance estimator to use?
  - How many replicates do we need?

Method we are using

- We are using five t testing methods for T-matrix data from the NCI-60 cancer cell lines data set.
- We selected two cancer groups for comparisons
  - the first is ob-group, including ovarian (OV) vs. breast (BR)
  - the second one is rl-group, including leukaemias (LE) vs. renal carcinoma (RE)
- We try to see the performance of several tests with different variance estimators for these two groups.

T_tests

- test1: uses sample variance of each gene for each disease;
- test2: uses pooled variance across diseases for each gene;
- test3: uses pooled variance across genes for each disease as a common variance for each disease;
- test4: as test 2, but using the common variance obtained from test3;
- test5:permutation test based on test 2.

In Equation,

Let Y_{ijA} and Y_{ijB} be the intensity measurement for gene i ($i=1, \ldots, n$) at the jth cancer cell line for type A and type B cancer, respectively , where $j_A=1,\ldots, r_A$ and $j_B = 1,\ldots, r_b$.We use type A as an example.

Test1

$$\hat{\sigma}_{iA}^2 = \frac{\sum_{j=1}^{rA} (Y_{ijA} - \bar{Y}_i)^2}{r_A - 1}$$

r_A

$$T_i = \frac{\bar{Y}_{iA} - \bar{Y}_{iB}}{\sqrt{\frac{\hat{\sigma}_{iA}^2}{r_A} + \frac{\hat{\sigma}_{iB}^2}{r_B}}}$$

$$df : r_A + r_B - 2$$

Test2

$$\hat{\sigma}_{iP}^2 = \frac{(r_A - 1)\hat{\sigma}_{iA}^2 + (r_B - 1)\hat{\sigma}_{iB}^2}{(r_A + r_B - 2)}$$

$$T_i = \frac{\bar{Y}_{iA} - \bar{Y}_{iB}}{\sqrt{\hat{\sigma}_{ip}^2 \left(\frac{1}{r_A} + \frac{1}{r_B}\right)}}$$

$$df : r_A + r_B - 2$$

Test3

$$\sigma_A^2 = \frac{\sum_{i=1}^{n} \hat{\sigma}_i^2}{n}$$

$$T_i = \frac{\bar{Y}_{iA} - \bar{Y}_{iB}}{\sqrt{\frac{\hat{\sigma}_A^2}{r_A} + \frac{\hat{\sigma}_B^2}{r_B}}}$$

$$df : r_A + r_B - 2$$

Test4

$$\sigma_A^2 = \frac{\sum_{i=1}^{n} \hat{\sigma}_i^2}{n}$$

$$\hat{\sigma}_P^2 = \frac{(r_A - 1)\hat{\sigma}_A^2 + (r_B - 1)\hat{\sigma}_B^2}{(r_A + r_B - 2)}$$

$$T_i = \frac{\bar{Y}_{iA} - \bar{Y}_{iB}}{\sqrt{\hat{\sigma}_p^2 (\frac{1}{r_A} + \frac{1}{r_B})}}$$

$$df : r_A + r_B - 2$$

Permutation

- Compute test2 for observed data.
- Randomly pick (No_of cell lines in first cancer +No_of cell lines in second cancer) samples from each gene
- Compute test 2 for permuted data
- Repeat permutation 1000 times
- compute number of times,N, the test from permuted sample is more extreme than test from observed sample.
- (P-value=N/1000)<0.05

How I did to fulfill this

- Using C++ for tests and Permutation
- Using Splus to do part of calculation

Result1(Number of significant genes for each test)

Number of significant tests	**OV vs. BR**	**LE vs. RE**
test1	105	551
Test2(--)	96	561
Test3(|)	82	526
Test4(+)	75	541
test5	99	560

If each pair of test detects same gene?

- For example:

No of times for sig test1, not in test2

No of times for sig test2, but not in test1

Result2(comparing the ttests results in ob data)

Y*N(ob)	test1	test2	test3	test4	test5
Test1*		15	44	48	51
Test2*(--)	6		36	40	42
Test3*(|)	21	22		7	23
Test4*(+)	18	19	0		37
Test5*	2	2	16	18	

Result3 (comparing the ttests results in rl data)

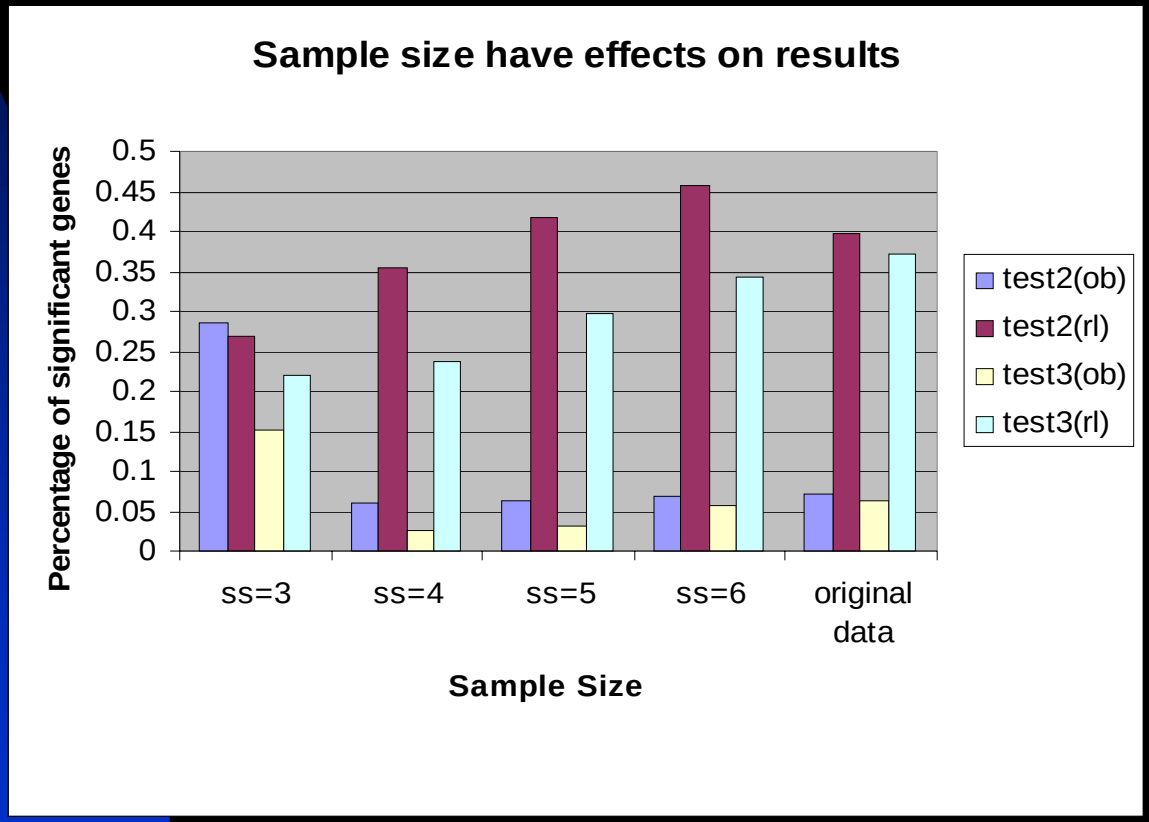
Y*N(rl)	test1	test2	test3	test4	test5
Test1*		14	74	66	52
Test2*(--)	24		74	66	58
Test3*(|)	49	39		0	71
Test4*(+)	56	46	15		79
Test5*	5	1	49	42	

How many replicates should we use?

- Selected different number of cell lines considering them replicates
- Captures variability between cell lines.

Result4

The sample size effect on t-tests



What This Means

- Each test detects many fewer significant genes in ob-group than rl-group
- Number of significant tests is decreasing from test1 to test4.
- The recommended sample size is at least 4 replicates.

Further study

- Use other methods for further study of microarray