Optimal significance analysis of microarray data in a class of tests whose null statistic can be constructed

Hironori Fujisawa (The Institute of Statistical Mathematics)

Takayuki Sakaguchi (Yamagata University)

Contents

1. Detection of Differentially Expressed Genes

Gene Expression Value

$$X_i \ : \ i = 1, \ldots, n.$$
$$Y_i \ : \ i = 1, \ldots, m.$$

Null Hypothesis

$$H : \mu_X = \mu_Y$$

Test Statistic

$$T(Z) = \frac{|\bar{X} - \bar{Y}|}{\sqrt{\{(n-1)S_X^2 + (m-1)S_Y\}/(N-2)}}$$

$X = (X_1, \ldots, X_n)'$, $Y = (Y_1, \ldots, Y_m)'$ and $Z = (X', Y')'$

P-value

$$p = \Pr(T(Z) > t \,|\, H) \qquad t = T(z)$$

P-value estimation

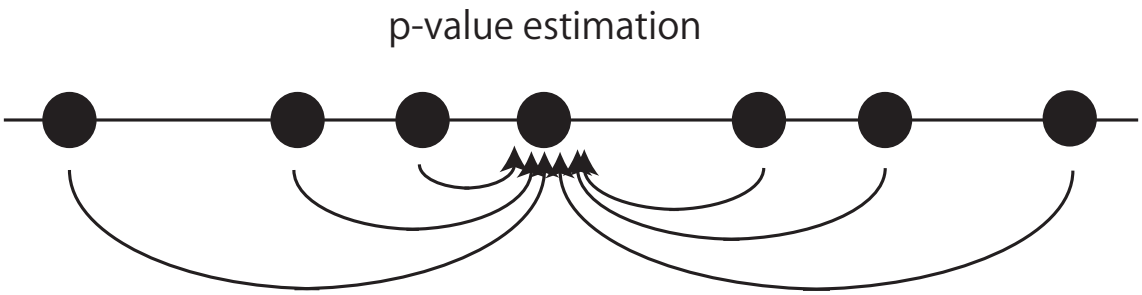
$$\frac{1}{B}\sum_{b=1}^{B} I(T(Z_b^{\#}) > t)$$

$I(\mathcal{A}) = 1$ when $\mathcal{A}$ is true, $I(\mathcal{A}) = 0$ when $\mathcal{A}$ is false.
$Z_b^{\#}$: bth permutation sample drawn from Z

Problem

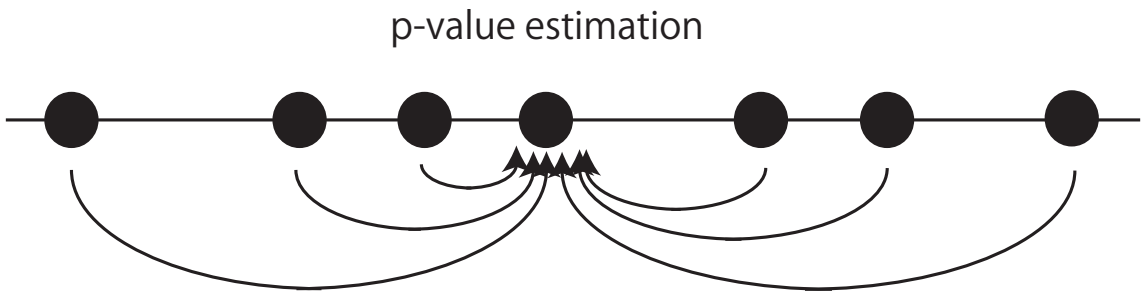
When $n = m = 4$, we have $B = 8!/(4!4!) = 56$. B is too small!
The smallest P-value is about 0.02 (except for 0).

2. Incorporation of Data From Other Genes



P-value Estimation

$$\hat{p}\left(T(Z^{\#}); t_g\right) = \frac{1}{BG} \sum_{b=1}^{B} \sum_{g'=1}^{G} I(T(Z^{\#}_{bg'}) > t_g)$$



Problem

Two types of genes are combined.
The null hypothesis is true on some genes, but not on the other genes.
The P-value should be calculated under the null hypothesis from the definition.

P-value

$$p = \Pr(T(Z) > t \mid H) \qquad t = T(z)$$

3. Devised Testing Procedure

Pan (2003)

$$X_{(1)} = (X_1, \ldots, X_{n_1})', \; X_{(2)} = (X_{n_1+1}, \ldots, X_n)', \; n_2 = n - n_1.$$
$$Y_{(1)} = (Y_1, \ldots, Y_{m_1})', \; Y_{(2)} = (Y_{m_1+1}, \ldots, Y_m)', \; m_2 = m - m_1.$$

Test Statistic

$$T_{\text{Pan}}(Z) = \left| \frac{\bar{X}_{(1)} + \bar{X}_{(2)}}{2} - \frac{\bar{Y}_{(1)} + \bar{Y}_{(2)}}{2} \right| \Bigg/ \sqrt{\frac{1}{4}\left(\frac{S^2_{X(1)}}{n_1} + \frac{S^2_{X(2)}}{n_2} + \frac{S^2_{Y(1)}}{m_1} + \frac{S^2_{Y(2)}}{m_2} \right)}$$

Test Statistic

$$T_{\text{Pan}}(Z) = \left| \frac{\bar{X}_{(1)} + \bar{X}_{(2)}}{2} - \frac{\bar{Y}_{(1)} + \bar{Y}_{(2)}}{2} \right| \Big/ \sqrt{\frac{1}{4}\left(\frac{S^2_{X(1)}}{n_1} + \frac{S^2_{X(2)}}{n_2} + \frac{S^2_{Y(1)}}{m_1} + \frac{S^2_{Y(2)}}{m_2} \right)}$$

Null Statistic

$$\begin{aligned} T^{\text{null}}_{\text{Pan}}(Z) &= T_{\text{Pan}}\left(X_{(1)}, -X_{(2)}, -Y_{(1)}, Y_{(2)}\right) \\ &= \left| \frac{\bar{X}_{(1)} - \bar{X}_{(2)}}{2} + \frac{\bar{Y}_{(1)} - \bar{Y}_{(2)}}{2} \right| \Big/ \sqrt{\frac{1}{4}\left(\frac{S^2_{X(1)}}{n_1} + \frac{S^2_{X(2)}}{n_2} + \frac{S^2_{Y(1)}}{m_1} + \frac{S^2_{Y(2)}}{m_2} \right)} \end{aligned}$$

Device

If the underlying distribution is symmetric around mean parameter, then the distribution of the null statistic does not depend on the mean parameters. This property has no relation on whether the null hypothesis is true or not. The distribution of the test statistic under the null hypothesis is the same as the distribution of the null statistic.

Test Statistic

$$T_{\mathrm{Pan}}(Z) = \left| \frac{\bar{X}_{(1)} + \bar{X}_{(2)}}{2} - \frac{\bar{Y}_{(1)} + \bar{Y}_{(2)}}{2} \right| \Big/ \sqrt{\frac{1}{4}\left(\frac{S^2_{X(1)}}{n_1} + \frac{S^2_{X(2)}}{n_2} + \frac{S^2_{Y(1)}}{m_1} + \frac{S^2_{Y(2)}}{m_2} \right)}$$

Null Statistic

$$\begin{aligned} T^{\mathrm{null}}_{\mathrm{Pan}}(Z) &= T_{\mathrm{Pan}}\left(X_{(1)}, -X_{(2)}, -Y_{(1)}, Y_{(2)}\right) \\ &= \left| \frac{\bar{X}_{(1)} - \bar{X}_{(2)}}{2} + \frac{\bar{Y}_{(1)} - \bar{Y}_{(2)}}{2} \right| \Big/ \sqrt{\frac{1}{4}\left(\frac{S^2_{X(1)}}{n_1} + \frac{S^2_{X(2)}}{n_2} + \frac{S^2_{Y(1)}}{m_1} + \frac{S^2_{Y(2)}}{m_2} \right)} \end{aligned}$$

Relation Between Test Statistic and Null Statistic

$$p_{\mathrm{Pan}} = \Pr(T_{\mathrm{Pan}}(Z) > t \,|\, H) = \Pr(T^{\mathrm{null}}_{\mathrm{Pan}}(Z) > t)$$

We need that the underlying distribution is symmetry around mean parameter.

Realation Between Test Statistic and Null Statistic

$$p_{\text{Pan}} = \Pr(T_{\text{Pan}}(Z) > t \,|\, H) = \Pr(T^{\text{null}}_{\text{Pan}}(Z) > t)$$

We need that the underlying distribution is symmetry around mean parameter.

P-value Estimation

$$\hat{p}\left(T_{\text{Pan}}(Z^*); t_g\right) = \frac{1}{BG} \sum_{b=1}^{B} \sum_{g'=1}^{G} I\left(T^{\text{null}}_{\text{Pan}}(Z^*_{bg'}) > t_g\right)$$

Z^*_{bg}: bth restricted permutation sample on gth gene

$Z^*_{bg} = (X^{\#}_{b'g}, Y^{\#}_{b''g})$

General Viewpoint

Condition that the empirical P-value estimation is possible when the observations on the other genes are incorporated.

There exists a null statistic $h(Z)$ such that

$$(*) \quad p = \Pr(T(Z) > t \,|\, H) = \Pr(h(Z) > t).$$

P-value Estimation

$$\hat{p}\left(T_{\mathrm{Pan}}(Z^*); t_g\right) = \frac{1}{BG} \sum_{b=1}^{B} \sum_{g'=1}^{G} I\left(h(Z^*_{bg'}) > t_g\right)$$

Z^*_{bg}: bth restricted permutation sample on gth gene

$Z^*_{bg} = (X^{\#}_{b'g}, Y^{\#}_{b''g})$

4. Optimal Testing Procedure

General Viewpoint

Condition that the empirical P-value estimation is possible when the observations on the other genes are incorporated.

There exists a null statistic $h(Z)$ such that

$$(*) \quad p = \Pr(T(Z) > t \,|\, H) = \Pr(h(Z) > t).$$

Optimal Testing Procedure

Consider the optimal test (UMP unbiased test) in a class of tests satisfying the above condition.

Summary of Result

Consider a class of tests, $(*A)$, derived from the case where the underlying distribution is symmetric around mean parameter. The UMP unbiased test under normality is similar to that proposed by Pan.

Consider a class of tests, $(*B)$, derived from the case where both underlying distributions for X and Y belong to the same location-family. The UMP unbiased test under normality is different from that proposed by Pan.

The latter test statistic has a one more degree-of-freedom than the former one.
But, the effect is very large for microarray data, because the sample size is small.

Theorem. Consider a class of test statistics expressed as $(*A)$ (derived from the case where the underlying distribution is symmetric around mean parameter). Assume that X and Y are normally distributed with means μ_X and μ_Y and common variance σ^2. The UMP unbiased test for the null hypothesis $H : \mu_X = \mu_Y$ against $K : \mu_X \neq \mu_Y$ is obtained from the test statistic

$$T_{\rm s}(Z) = |Q_{\rm s}|/\sqrt{S^2}$$

where

$$\begin{aligned} Q_{\rm s} &= \left(\frac{\bar{X}_{(1)} + \bar{X}_{(2)}}{2} - \frac{\bar{Y}_{(1)} + \bar{Y}_{(2)}}{2}\right) \Big/ \sqrt{\frac{1}{4}\left(\frac{1}{n_1} + \frac{1}{n_2} + \frac{1}{m_1} + \frac{1}{m_2}\right)} \\ (N-4)S^2 &= (n_1 - 1)S^2_{X(1)} + (n_2 - 1)S^2_{X(2)} + (m_1 - 1)S^2_{Y(1)} + (m_2 - 1)S^2_{Y(2)}. \end{aligned}$$

The threshold c at significance level α is determined from $\Pr(T_{\rm s}(Z) \geq c \,|\, H) = \alpha$. The power is maximized when n_1 and m_1 are the closest integers to $n/2$ and $m/2$, respectively.

Theorem. Consider a class of test statistics expressed as $(*B)$ (derived from the case where both underlying distributions for X and Y belong to the same location-family). Assume that X and Y are normally distributed with means μ_X and μ_Y and common variance σ^2. The UMP unbiased test for the null hypothesis $H : \mu_X = \mu_Y$ against $K : \mu_X \neq \mu_Y$ is obtained from the test statistic

$$T_{\rm p}(Z) = \frac{|Q_{\rm p}|}{\sqrt{\left\{(N-4)S^2 + Q_0^2\right\}/(N-3)}},$$

where

$$\begin{aligned} Q_{\rm p} &= \left\{\frac{n_1 m_1}{n}(\bar{X}_{(1)} - \bar{Y}_{(1)}) + \frac{n_2 m_2}{m}(\bar{X}_{(2)} - \bar{Y}_{(2)})\right\} \Big/ \sqrt{\frac{n_1 m_1}{n} + \frac{n_2 m_2}{m}} \\ Q_0 &= \sqrt{\frac{n_1 m_1 n_2 m_2 / nm}{n_1 m_1/n + n_2 m_2/m}} \left\{(\bar{X}_{(1)} - \bar{Y}_{(1)}) - (\bar{X}_{(2)} - \bar{Y}_{(2)})\right\}. \end{aligned}$$

The threshold c at significance level α is determined from $\Pr(T_{\rm p}(Z) > c \,|\, H) = \alpha$. The power is maximized when n_2 and m_1 are the same closest integer to $1/(1/n + 1/m)$.

5. Example (Golden Spike data)

Choe et al. (2005) presented a control dataset.

This data include 14,010 probe sets with 195,994 probes. They constructed 3,866 probe sets with various *known* fold changes by using spiked-in cRNAs. There were 1,331 probe sets whose fold changes were larger than one. They provided two samples (control and spiked-in samples) with three replicates ($n = m = 3$).

They discussed a lot of combinations of pre-treatments and then recommended some combinations of pre-treatments.

We then analyzed the Golden Spike data with the optimal sample division size where $n_1 = 2$ and $m_1 = 1$.

Performance with Golden Spike data.

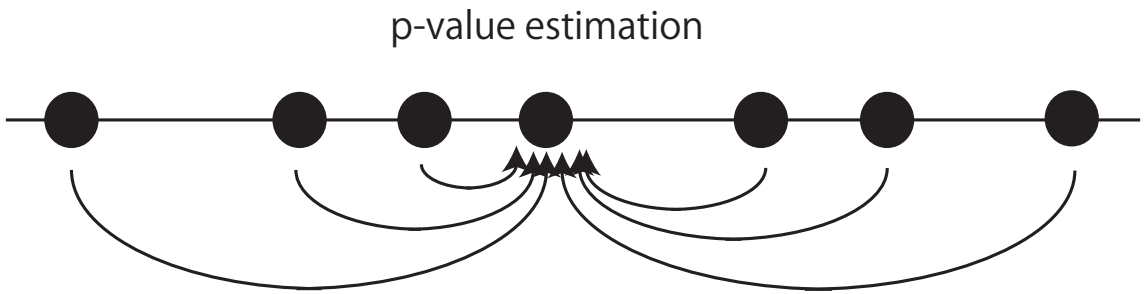
α	10^{-2}	10^{-3}	10^{-4}	10^{-5}	Bon*
	# of detected genes				
T_{s}	340	9	0	0	0
T_{p}	1013	525	161	16	16
	# of truly detected genes				
T_{s}	317	8	0	0	0
T_{p}	890	508	158	16	16

Bon*: Bonferroni correction based on significance level 0.01.
$\alpha = 0.01/3866 \approx 2.59 \times 10^{-6}$.

6. Simulation

6.1. Accuray of P-value Estimation

T_{Pan}: Pan's statistic
T_{s}: Optimal test statistic in a class of $(*A)$
T_{s2}: Modified test statistic of T_{s}
T_{p}: Optimal test statistic in a class of $(*B)$
Perm: Standard permutation method without device



Random Number Genreration

X: Normal with mean zero and variance one.
Y: Normal with mean μ_Y and variance one.
$\mu_Y = 0$ for equally expressed genes
$\mu_Y \sim N(0, 4^2)$ for differentially expressed genes

# of simulations: 50.
# of genes: 1,000.
The proportion of differentially expressed genes: 0.1.
Genes: independent.

P-value estimation

$$\hat{p}\left(T(Z^*); t_g\right) = \frac{1}{BG} \sum_{b=1}^{B} \sum_{g'=1}^{G} I\left(T^{\text{null}}(Z^*_{bg'}) > t_g\right)$$

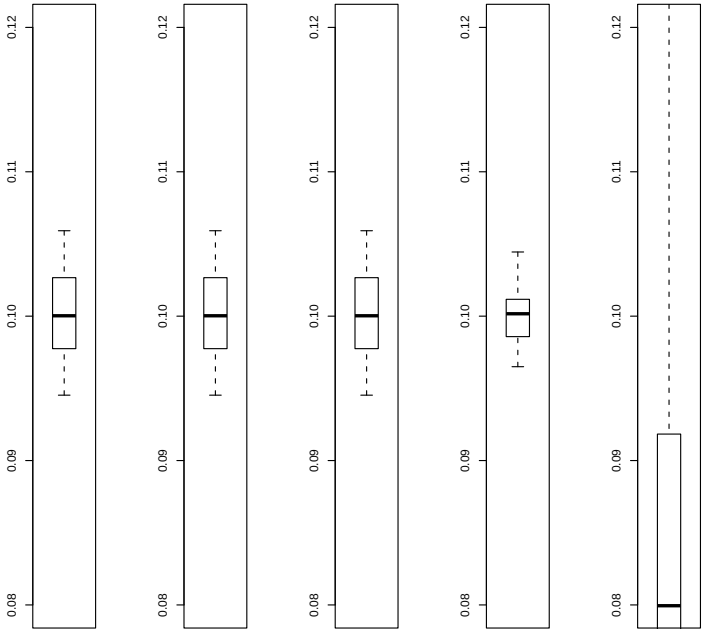
Threshold

$$\Pr\left(T(Z) > t_\alpha \,|\, H\right) = \alpha \quad (= 0.01)$$

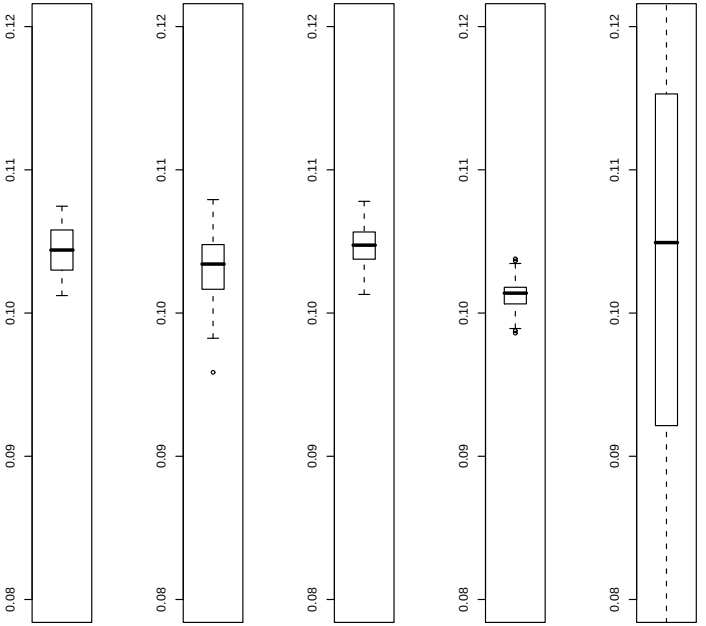
$$\hat{\alpha} = \frac{1}{BG} \sum_{b=1}^{B} \sum_{g'=1}^{G} I\left(T^{\text{null}}(Z^*_{bg'}) > t_\alpha\right)$$

T_{Pan}. T_{s}. T_{s2}. T_{p}. Perm.

(a) $(n, m) = (4, 4)$.



(b) $(n, m) = (10, 5)$.



The following cases were also investigated.

The underlying distribution was skew-normal.
The scale varied according to gene.

Similar behaviors were observed.
The test statistic T_{p} was the best.

6.2. Comparison of Power: T_{Pan}. T_{s}. T_{s2}. T_{p}.

Random Number Generation

X: Normal with mean zero and variance one.
Y: Normal with mean μ_Y and variance σ_Y^2.

The power was estimated by 10,000 simulations.

Power of test when $\sigma_Y^2 = 1$.

	T_{Pan}	T_{s}	T_{s2}	T_{p}
(a) $n = 4,\ m = 4,\ n_1 = 2,\ m_1 = 2$				
		$\alpha = 0.01$		
$\mu_Y = 1$	0.049	0.049	0.049	0.058
$\mu_Y = 3$	0.497	0.497	0.497	0.610
$\mu_Y = 5$	0.920	0.920	0.920	0.974
		$\alpha = 0.001$		
$\mu_Y = 1$	0.007	0.007	0.007	0.008
$\mu_Y = 3$	0.100	0.100	0.100	0.163
$\mu_Y = 5$	0.395	0.395	0.395	0.605

Power of test when $\sigma_Y^2 = 1$.

	T_{Pan}	T_{s}	T_{s2}	T_{p}
(b-i)	$n = 10,\ m = 5,\ n_1 = 5,\ m_1 = 3$			
		$\alpha = 0.01$		
$\mu_Y = 1$	0.133	0.152	0.134	0.156
$\mu_Y = 2$	0.592	0.688	0.602	0.709
$\mu_Y = 3$	0.916	0.973	0.933	0.980
		$\alpha = 0.001$		
$\mu_Y = 1$	0.029	0.032	0.028	0.032
$\mu_Y = 2$	0.233	0.302	0.246	0.325
$\mu_Y = 3$	0.654	0.780	0.650	0.831

Power of test when $\sigma_Y^2 = 1$.

	T_{Pan}	T_{s}	T_{s2}	T_{p}
(b-ii)	$n = 10,\ m = 5,\ n_1 = 7,\ m_1 = 3$			
		$\alpha = 0.01$		
$\mu_Y = 1$	0.108	0.138	0.123	0.158
$\mu_Y = 2$	0.522	0.644	0.580	0.710
$\mu_Y = 3$	0.886	0.966	0.933	0.981
		$\alpha = 0.001$		
$\mu_Y = 1$	0.023	0.029	0.025	0.034
$\mu_Y = 2$	0.182	0.301	0.237	0.329
$\mu_Y = 3$	0.537	0.765	0.638	0.833

Power of test when $\sigma_Y^2 = 2$.

	T_{Pan}	T_{s}	T_{s2}	T_{p}
(a) $n = 4,\ m = 4,\ n_1 = 2,\ m_1 = 2$				
		$\alpha = 0.01$		
$\mu_Y = 1$	0.040	0.040	0.040	0.042
$\mu_Y = 3$	0.342	0.342	0.342	0.424
$\mu_Y = 5$	0.785	0.785	0.785	0.887
		$\alpha = 0.001$		
$\mu_Y = 1$	0.006	0.006	0.006	0.006
$\mu_Y = 3$	0.064	0.064	0.064	0.094
$\mu_Y = 5$	0.253	0.253	0.253	0.403

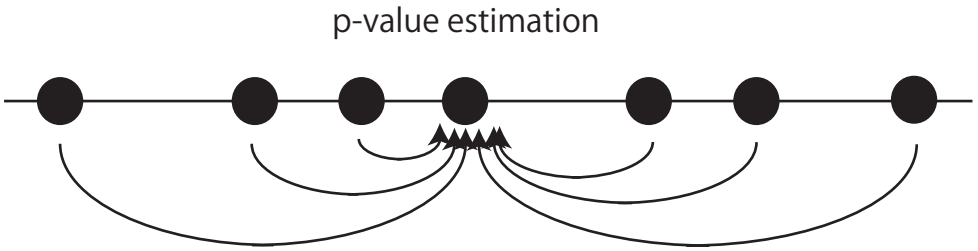
Power of test when $\sigma_Y^2 = 2$.

	T_{Pan}	T_{s}	T_{s2}	T_{p}
(b-i)	$n = 10,\ m = 5,\ n_1 = 5,\ m_1 = 3$			
		$\alpha = 0.01$		
$\mu_Y = 1$	0.094	0.141	0.091	0.142
$\mu_Y = 2$	0.389	0.546	0.390	0.563
$\mu_Y = 3$	0.729	0.906	0.743	0.915
		$\alpha = 0.001$		
$\mu_Y = 1$	0.021	0.034	0.022	0.034
$\mu_Y = 2$	0.141	0.231	0.138	0.257
$\mu_Y = 3$	0.393	0.612	0.368	0.623

Power of test when $\sigma_Y^2 = 2$.

	T_{Pan}	T_{s}	T_{s2}	T_{p}
(b-ii)	$n = 10,\ m = 5,\ n_1 = 7,\ m_1 = 3$			
	$\alpha = 0.01$			
$\mu_Y = 1$	0.089	0.134	0.090	0.146
$\mu_Y = 2$	0.348	0.526	0.371	0.564
$\mu_Y = 3$	0.696	0.892	0.738	0.915
	$\alpha = 0.001$			
$\mu_Y = 1$	0.018	0.032	0.018	0.034
$\mu_Y = 2$	0.106	0.184	0.134	0.230
$\mu_Y = 3$	0.319	0.582	0.387	0.647

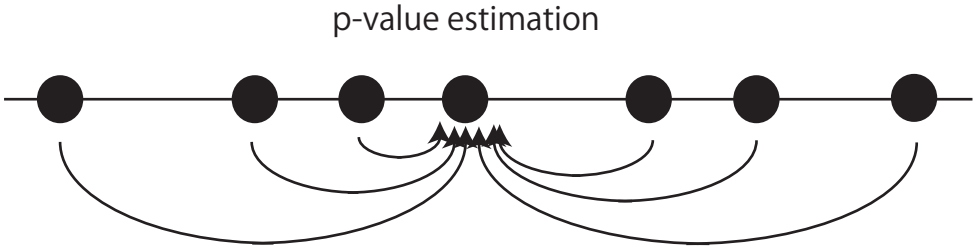
7. Summary



Incorporation of Data From Other Genes
Devised Testing Procedure by Pan
Generalization of Device
Optimal Testing Procedure
Symmetry Case and Location-family Case

Example: Golden Spike Data
Simulation
Accuray of P-value Estimation. Comparison of Power.

THANK YOU



Hironori Fujisawa

The Institute of Statistical Mathematics