SD2: Statistical analysis.

1 Statistical model.

Let B be a nucleotide at a given RefSeq gene position.

It is assumed in this model that a variability exists in both the cancer and normal groups. Our aim was to determine if the variability related to B was the same in both the cancer and normal groups. The numbers p_{1B} and p_{2B} were also compared, where p_{1B} and p_{2B} are the probabilities of a cancer or normal case, respectively, of having the B base on an EST.

This led to the two-sided proportion test, with the following hypotheses :

$$\begin{cases} H_0 : p_{1B} = p_{2B} \\ H_1 : p_{1B} \neq p_{2B}. \end{cases} \tag{1}$$

This test is equivalent to

$$\begin{cases} H_0 : p_{1\bar{B}} = p_{2\bar{B}} \\ H_1 : p_{1\bar{B}} \neq p_{2\bar{B}} \end{cases} \tag{2}$$

where $p_{1\bar{B}}$ and $p_{2\bar{B}}$ denote the probability of having a different base than that of the RefSeq in the cancer or normal case, respectively. For example, if $B = T$, then $\bar{B} = \{A, C, G\}$. $p_{1\bar{B}}$ and $p_{2\bar{B}}$ measure the variabilities in cancer and normal group, respectively.

Moreover,

$$p_{j\bar{T}} = p_{jA} + p_{jC} + p_{jG}, \; j = 1, 2. \tag{3}$$

Note that test (2) cannot be performed since the bases are subject to sequencing/reading errors of the cDNA sequence.

At a given position, it is currently assumed that a sequencing error occurs at a rate of 1 to 5%. This means that the mean number of erroneous readings is 1 to 5 per 100 bases, these bases being replaced by another one on the EST. Therefore variability can't be measured directly. Sequencing errors have also to be introduced in the model.

Let ϵ be the sequencing error probability at a given position, i.e the probability of reading a base that is different from the real one.

It is assumed that:

1. (E_1) The sequencing error probability ϵ does not depend on the cDNA deriving from a normal or cancer case,

2. (E_2) The sequencing error probability ϵ does not depend on the real base B,

3. (E_3) The probability of reading a given base that is different from the real one is $\frac{\epsilon}{3}$.

Let q_{1B} and q_{2B} be the probabilities of the RefSeq base $B \in \{A, T, C, G\}$ to be correctly read for a cancer case and a normal case, respectively.

Let us determine $q_{jT},\ j = 1, 2$:

$$\begin{array}{rcl} q_{jT} & = & P(\text{“ } the\ read\ base\ is\ T \text{ ”}) \\ & = & P(\text{“}the\ real\ base\ is\ T\ and\ there\ is\ no\ sequencing\ error \text{ ”}) \\ & + & P(\text{“}the\ real\ base\ is\ A\ and\ there\ is\ a\ sequencing\ error \\ & & from\ A\ to\ T \text{ ”}) \\ & + & P(\text{“}the\ real\ base\ is\ C\ and\ there\ is\ a\ sequencing\ error \\ & & from\ C\ to\ T \text{ ”}) \\ & + & P(\text{“}the\ real\ base\ is\ G\ and\ there\ is\ a\ sequencing\ error \\ & & from\ G\ to\ T \text{ ”}) \end{array}$$

So under $(E1)$, $(E2)$ and $(E3)$:

$$\begin{array}{rcl} q_{jT} & = & p_{jT}(1-\epsilon) + p_{jA}\frac{\epsilon}{3} + p_{jC}\frac{\epsilon}{3} + p_{jG}\frac{\epsilon}{3} \\ & = & p_{jT}(1-\epsilon) + \frac{\epsilon}{3}(1 - p_{jT}) \\ & = & p_{jT}(1 - 4\frac{\epsilon}{3}) + \frac{\epsilon}{3}. \end{array}$$

In general, for any base B:

$$q_{jB} = p_{jB}(1 - 4\frac{\epsilon}{3}) + \frac{\epsilon}{3}. \tag{4}$$

The two-sided test (1) can also be rewritten as:

$$\begin{cases} H_0 : q_{1B} = q_{2B} \\ H_1 : q_{1B} \neq q_{2B} \end{cases} \tag{5}$$

or

$$\begin{cases} H_0 : q_{1\bar{B}} = q_{2\bar{B}} \\ H_1 : q_{1\bar{B}} \neq q_{2\bar{B}} \end{cases} \tag{6}$$

The value of sequencing error probability ϵ is not needed to perform this test.

Two-sided proportion test.

At a non SNP given position of a gene RefSeq, n_1 cancer ESTs and n_2 normal ESTs are observed. The following contingency table is considered:

	B	$\overline{B}$	Sum
Cancer	$n_1 - k_1$	k_1	n_1
Normal	$n_2 - k_2$	k_2	n_2
Sum	m_1	m_2	n

k_1 and k_2 are computed by counting the number of nucleotides which are different from that of the RefSeq among the number of cancer and normal aligned ESTs, respectively. The proportions $\frac{k_1}{n_1}$ and $\frac{k_2}{n_2}$ refer to the observed percentage of deviations in the cancer and normal tissues.

Let K_j be the random variable which assigns to each case the number of non B nucleotides in the cancer group ($j = 1$) or in the normal group ($j = 2$).

Then k_j is a realization of the random variable K_j.

$\hat{P}$ is denoted as

$$\hat{P} = \frac{K_1 + K_2}{n_1 + n_2}, \tag{7}$$

whose

$$\hat{p} = \frac{k_1 + k_2}{n_1 + n_2} \tag{8}$$

is a realization.

We introduce T as:

$$T = \frac{\frac{K_1}{n_1} - \frac{K_2}{n_2}}{\sqrt{\hat{P}(1 - \hat{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \tag{9}$$

whose

$$t = \frac{\frac{k_1}{n_1} - \frac{k_2}{n_2}}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \tag{10}$$

is a realization.

Under the constraints

$$n > 70, \ \frac{n_i m_j}{n} > 5, \ i = 1, 2, \ j = 1, 2, \tag{11}$$

and when the null-hypothesis H_0 holds

$$T \approx N(0, 1). \tag{12}$$

Finally, let $u(\alpha)$ be the real number as

$$\alpha = P(G > u(\alpha)),$$

where $G \sim N(0, 1)$.

So test (6) is done for each position where (11) holds, and the decision rule is then:

Decision rule :

If

$$|t| > u(\frac{\alpha}{2}), \tag{13}$$

the null hypothesis H_0 is rejected at the confidence level α; otherwise the null hypothesis H_0 is not rejected.

This test allowed us to determine if at a given position, the variabilities in the cancer group and in the normal group are the same or not.

One-sided proportion tests.

As the hypothesis $q_{1\bar{B}} \neq q_{2\bar{B}}$ is equivalent to $q_{1\bar{B}} > q_{2\bar{B}}$ or $q_{1\bar{B}} < q_{2\bar{B}}$, the two following one-sided proportion tests are considered:

1. The first one-sided test

$$\left\{ \begin{array}{l} H_0 : q_{1\bar{B}} \leq q_{2\bar{B}} \\ H_1 : q_{1\bar{B}} > q_{2\bar{B}} \end{array} \right. \tag{14}$$

which is equivalent to the test

$$\left\{\begin{array}{l} H_0 : p_{1\bar{B}} \leq p_{2\bar{B}} \\ H_1 : p_{1\bar{B}} > p_{2\bar{B}}, \end{array}\right. \tag{15}$$

has the same decision rule than the test

$$\left\{\begin{array}{l} H_0 : q_{1\bar{B}} = q_{2\bar{B}} \\ H_1 : q_{1\bar{B}} > q_{2\bar{B}}. \end{array}\right. \tag{16}$$

Decision rule :

If

$$t > u(\alpha), \tag{17}$$

the null hypothesis H_0 is rejected at the confidence level α; otherwise the null hypothesis H_0 is not rejected.

Thus this test allows first to conclude that variabilities are different in both groups when positive, then it measures in this case whether variability is statistically greater in the cancer set.

2. The second one-sided test

$$\left\{\begin{array}{l} H_0 : q_{1\bar{B}} \geq q_{2\bar{B}} \\ H_1 : q_{1\bar{B}} < q_{2\bar{B}} \end{array}\right. \tag{18}$$

which is equivalent to the test

$$\left\{\begin{array}{l} H_0 : p_{1\bar{B}} \geq p_{2\bar{B}} \\ H_1 : p_{1\bar{B}} < p_{2\bar{B}}, \end{array}\right. \tag{19}$$

has the same decision rule than the test

$$\left\{\begin{array}{l} H_0 : q_{1\bar{B}} = q_{2\bar{B}} \\ H_1 : q_{1\bar{B}} < q_{2\bar{B}}. \end{array}\right. \tag{20}$$

Decision rule :

If

$$t < -u(\alpha), \tag{21}$$

the null hypothesis H_0 is rejected at the confidence level α; otherwise the null hypothesis H_0 is not rejected.

Unlike test (14), test (18) verifies the hypothesis that variability is significantly higher in the normal set.

NB: If $\alpha = 5\%$, then $u(\frac{\alpha}{2}) = u(0.025) = 1.96$; if $\alpha = 10\%$, then $u(\frac{\alpha}{2}) = u(0.05) = 1.645$

$p-$**values.**

Introduce the $p-$value notion for tests (6), (14) and (18) .

Let T be the test statistic defined by (9), whose observed realization is t.

1. The probability

$$p = P(|T| > t \mid H_0). \tag{22}$$

is called p-**value** of test (6).

The decision rule of test (6) can also be written as:

Decision rule:

If $p < \alpha$, the null hypothesis H_0 is rejected at the confidence level α; otherwise the null hypothesis H_0 is not rejected.

2. The probability

$$p = P(T > t \mid H_0). \tag{23}$$

is called p-value of test (16), and the decision rule of test (16) can be rewritten as:

Decision rule:

If $p < \alpha$, the null hypothesis H_0 is rejected at the confidence level α; otherwise the null hypothesis H_0 is not rejected.

3. The probability

$$p = P(T < t \mid H_0). \tag{24}$$

is called p-value of test (20), and the decision rule of test (20) can be rewritten as:

Decision rule:

If $p < \alpha$, the null hypothesis H_0 is rejected at the confidence level α; otherwise the null hypothesis H_0 is not rejected.

2 Location Based Estimator: a false positives mean number overestimator.

Definitions : A statistical test is said to be

- positive if H_0 is rejected,
- false positive if H_0 is rejected whereas H_0 is true,
- true positive if H_0 is rejected whereas H_0 is false,
- negative if H_0 is not rejected,
- true negative if H_0 is not rejected whereas H_0 is true,
- false negative if H_0 is not rejected whereas H_0 is false.

For each gene, a statistic test is made at level α for m positions; this way m p-values $\left\{p^k\right\}_{1\leq k\leq m}$ are computed. These p-values are realizations of random variables $\left\{P^k\right\}_{1\leq k\leq m}$.

The following random variables are defined as:

1. $V(\alpha)$, the number of false positives at level α,
2. $S(\alpha)$, the number of true positives at level α,
3. $R(\alpha) = V(\alpha) + S(\alpha)$, the number of positives,
4. $U(\alpha)$, the number of true negatives at level α,

5. $T(\alpha)$, the number of false negatives at level α,

6. $W(\alpha) = U(\alpha) + T(\alpha)$, the number of negatives.

The following contingency table is considered:

	H_0 accepted	H_0 rejected	Sum
H_0 true	$U(\alpha)$	$V(\alpha)$	m_0
H_1 true	$T(\alpha)$	$S(\alpha)$	m_1
Sum	$W(\alpha)$	$R(\alpha)$	m

Note that m_0 and m_1 are unknown constants. On the contrary, $W(\alpha)$ and $R(\alpha)$ are observed variables.

As

$$V(\alpha) \sim B(m_0, \alpha), \tag{25}$$

we have :

$$E[V(\alpha)] = m_0\alpha. \tag{26}$$

An upper-bound of m_0 is given by $2E\left[\sum_{k=1}^{m} P^k\right]$. Indeed :

$$E\left[\sum_{k=1}^{m} P^k\right] = E\left[\sum_{k\ :\ H_0\ true} P^k\right] + \underbrace{E\left[\sum_{k\ :\ H_0\ false} P^k\right]}_{\geq 0}. \tag{27}$$

Whenever H_0 is true, $P^k \sim \mathcal{U}_{[0,1]}$, $1 \leq k \leq m$.

So we have

$$E[P^k] = \frac{1}{2}. \tag{28}$$

Since there are m_0 tests for which H_0 is true, according to (27), we have:

$$E\left[\sum_{k=1}^{m} P^k\right] \geq \frac{1}{2} m_0. \tag{29}$$

With (26) and (29), we obtain:

$$E[V(\alpha)] = m_0\alpha \leq E\left[2\alpha \sum_{k=1}^{m} P^k\right]. \tag{30}$$

$2\alpha \sum_{k=1}^{m} P^k$ is said to be an overestimator of $m_0\alpha$, called **Location Based Estimator** (LBE) [1].

References

[1] C.Dalmasso, P.Broet *Procédures d'estimation du false discovery rate basées sur la distribution des degrés de signification.* Journal de la Société Française de Statistique, tome 146, n°1-2, 2005