ВИКОНАВЕЦЬ:
СТУДЕНТ ГРУПИ МС-23М-1
ЩЕРБАК РОМАН ОЛЕКСІЙОВИЧ

КЕРІВНИК
КАРНАУХ ЄВГЕН ВОЛОДИМИРОВИЧ

РАНДОМІЗОВАНИЙ КРИТЕРІЙ ПОРІВНЯННЯ СЕРЕДНІХ ДВОХ ГРУП

ПОСТАНОВКА ЗАДАЧІ

Гострий лімфобластний лейкоз (Acute lymphoblastic leukemia, ГЛЛ, ALL) – онкологічне захворювання клітин крови (зокрема, T-лімфоцитів та B-лімфоцитів).

Мета – застосувати методи статистичного аналізу для встановлення істотности різниці між середніми рівнями експресії генів.

Завдання

- Огляд літератури щодо дослідження генної інформації.
- Розгляд основних методів статистичного аналізу генетичної інформації.
- Застосування T_n критерію для виявлення статистичної відмінности між експресією істотних генів.

НАБІР ДАНИХ ALL

Містить інформацію про 128 пацієнтів із Т-лейкемією та В-лейкемією та рівень експресії в них 12 625 генів.

Релевантних для розгляду – 2391 ген.

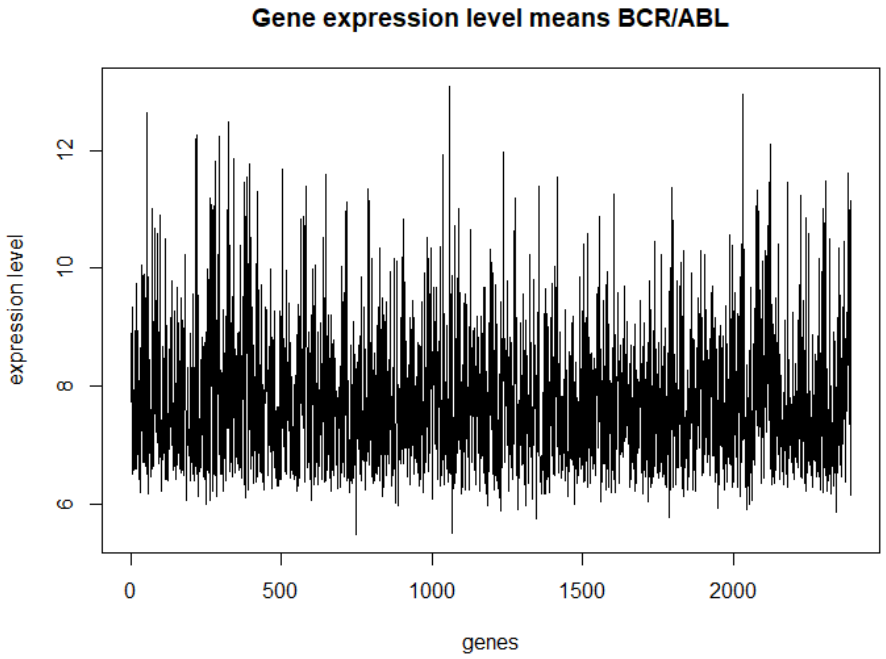
Вилучено гени з низьким рівнем експресії для обох типів хвороби та низькою мінливістю (за IQR).

37 пацієнтів із BCR/ABL, 42 пацієнти з NEG (В-лейкемія).

BCR/ABL – генетична абнормалія 22 хромосоми (філадельфійська хромосома).

NEG – відсутність виявлених генетичних абнормалій.

НАБІР ДАНИХ ALL



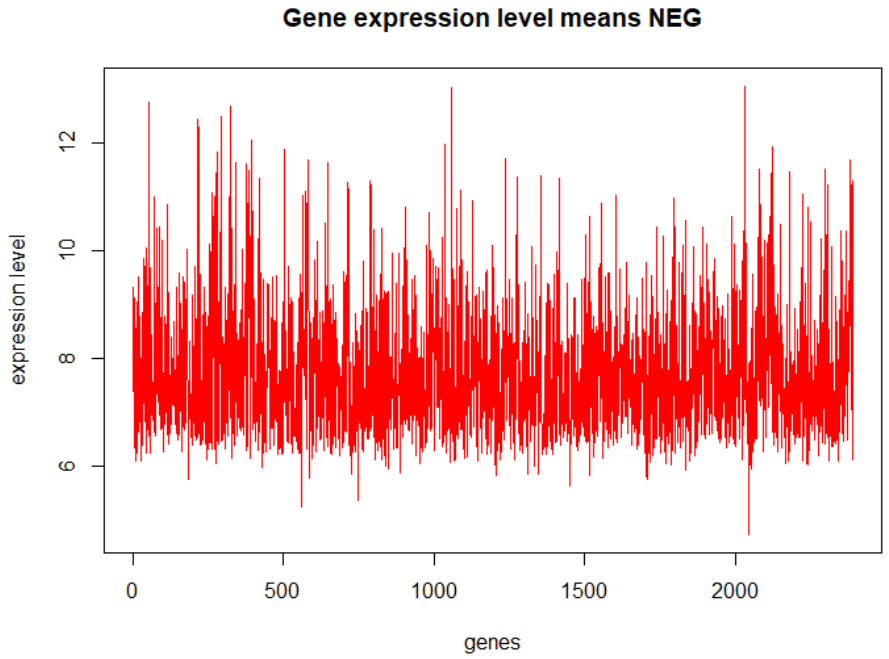


Рис 1. Графік середніх рівнів експресії генів для захворювання типу **BCR/ABL** (чорна лінія) та **NEG** (червона лінія)

СТАТИСТИКА M_n

Стандартно для порівняння середніх застосовують критерій Готелінґа T^2 з фіксованим p.

За $p/n \to \infty$ критерій Готелінґа незастосовний, а за $p/n \to c \in [0,1)$ його потужність зменшується зі збільшенням c

Альтернативні статистики:

1) $M_n = \left(\overline{X}_1 - \overline{X}_2\right)'\left(\overline{X}_1 - \overline{X}_2\right) - \tau tr(S_n)$, де

$$S_n = \frac{1}{n}\left(X_{ij} - \overline{X}_i\right)\left(X_{ij} - \overline{X}_i\right)'$$

$$\tau = \frac{n_1 + n_2}{n_1 n_2}$$

СТАТИСТИКА T_n

2) $T_n =: \frac{\sum_{i\neq j}^{n_1} X'_{1i}X_{1j}}{n_1(n_1-1)} + \frac{\sum_{i\neq j}^{n_2} X'_{2i}X_{2j}}{n_2(n_2-1)} - 2\frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} X'_{1i}X_{2j}}{n_1 n_2}$

Вилучено $\sum_{j=1}^{n_i} X'_{ij}X_{ij}$ для $i = 1$ та 2 з $\left|\left|\overline{X}_1 - \overline{X}_2\right|\right|^2$

$E(T_n) = \left|\left|\mu_1 - \mu_2\right|\right|^2$

Якщо правильна гіпотеза H_1:

$$Var(T_n) = \{\frac{2}{n_1(n_1-1)}tr(\Sigma_1^2) + \frac{2}{n_2(n_2-1)}tr(\Sigma_2^2) + \frac{4}{n_1 n_2}tr(\Sigma_1\Sigma_2)\}\{1 + o(1)\}$$

СТАТИСТИКА Q_n

3) $Q_n = T_n/\widehat{\sigma_{n1}} \to N(0,1)$, $p \to \infty, n \to \infty$, де

$$\sigma_{n1}^2 = \frac{2}{n_1(n_1-1)}\widehat{tr(\Sigma_1^2)} + \frac{2}{n_2(n_2-1)}\widehat{tr(\Sigma_2^2)} + \frac{4}{n_1 n_2}\widehat{tr(\Sigma_1\Sigma_2)}$$

$$\widehat{tr\left(\Sigma_i^2\right)} = \{n_i(n_i-1)\}^{-1} tr\{\sum_{j\neq k}^{n_i}\left(X_{ij} - \overline{X}_{i(j,k)}\right)X_{ij}'\left(X_{ik} - \overline{X}_{i(j,k)}\right)X_{ik}'\}$$

$$\widehat{tr(\Sigma_1\Sigma_2)} = (n_1 n_2)^{-1} tr\{\sum_{l=1}^{n_1}\sum_{k=1}^{n_2}\left(X_{1l} - \overline{X}_{1(l)}\right)X_{1l}'\left(X_{2k} - \overline{X}_{2(k)}\right)X_{2k}'\},$$

$\overline{X}_{i(j,k)}$ – і-е вибіркове середнє після вилучення X_{ij} та X_{ik}, $\overline{X}_{i(l)}$ – і-е вибіркове середнє після вилучення X_{il}.

Якщо $Q_n > \xi_\alpha$, де ξ_α – верхній α квантиль розподілу $N(0,1)$, гіпотезу H_0 відкидають.

ПРОГРАМА

Рівень значущости було встановлено як 0.05.

Містить три модулі:

- main – завантаження та аналіз даних, ділення сум для статистики T_n, обчислення оцінки σ_{n1}^2 та статистики Q_n, обчислення qnorm(1-sign_level/2).
- sum_for_T_stat – обчислення сум для статистики T_n для векторів X1, X2 та для їх комбінації.
- trace_for_Q_stat – обчислення оцінки сліду матриці для статистики Q_n для векторів X1, X2 та для їх комбінації.

У результаті обчислення було отримано значення Q_n – 0.458, тоді як значення ξ_α – 1.959, що не дозволяє відкидати гіпотезу H_0 про рівність середніх.

ВИСНОВКИ

Було розглянуто набір даних про гострий лімфобластний лейкоз, відібрано релевантні гени та пацієнтів із типами захворювання BCR/ABL та NEG

Розглянуто методи статистичного аналізу генетичної інформації, перевірено гіпотезу про рівність середніх за допомогою Q_n-критерію, використовуючи асимптотику для $p/n \to \infty$.

ДЯКУЮ ЗА УВАГУ!