https://www.proconference.org/index.php/gec/article/view/gec37-00-006
DOI: 10.30890/2709-1783.2025-37-00-006

УДК 004.9

ANALYSIS OF DIFFERENTIAL GENE EXPRESSION USING THE LIMMA PACKAGE

АНАЛІЗ ДИФЕРЕНЦІЙНОЇ ЕКСПРЕСІЇ ГЕНІВ ІЗ ВИКОРИСТАННЯМ ПАКЕТА LIMMA

Doroshenko I.V. / Дорошенко І.В.
c. p.-m.s., as.prof. / к. ф.-м.н., доц.
ORCID: 0000-0001-8729-1768
Plotnikova O.O. / Плотнікова О.О.
student / студент
Chernivtsi National University, Chernivtsi, Kotsyubynskoho 2, 58012
Чернівецький національний університет, Чернівці, вул.Коцюбинського 2, 58012

Анотація. *В статті проведено аналіз диференційної експресії генів за допомогою пакета limma середовища R на наборі даних, що містить 2000 генів та 62 зразки тканин (22 нормальні та 40 пухлинні). Було застосовано різні методи нормалізації та корекції тестування для підвищення точності аналізу, а також створено теплову карту генів із найбільш вираженими змінами експресії.*

Ключові слова:*. аналіз даних, системна біологія, середовище R.*

Abstract. *The article presents an analysis of differential gene expression using the limma package in the R environment on a dataset containing 2000 genes and 62 tissue samples (22 normal and 40 tumor samples). Various normalization methods and test correction techniques were applied to improve the accuracy of the analysis, and a heatmap was created to visualize genes with the most significant expression changes.*

Keywords: *data analysis, systems biology, R environment.*

Вступ.

Визначення генів, що мають диференційну експресію між пухлинними та нормальними тканинами, є важливим етапом у вивченні молекулярних механізмів раку. Аналіз експресії генів дозволяє не лише виявити потенційні біомаркери захворювання, а й зрозуміти процеси, що лежать в основі розвитку пухлин. У даній роботі проведено аналіз диференційної експресії генів за допомогою пакета *limma* на наборі даних, що містить 2000 генів та 62 зразки тканин (22 нормальні та 40 пухлинні). Було застосовано різні методи нормалізації та корекції тестування для підвищення точності аналізу, а також створено теплову карту генів із найбільш вираженими змінами експресії.

1. Постановка задачі

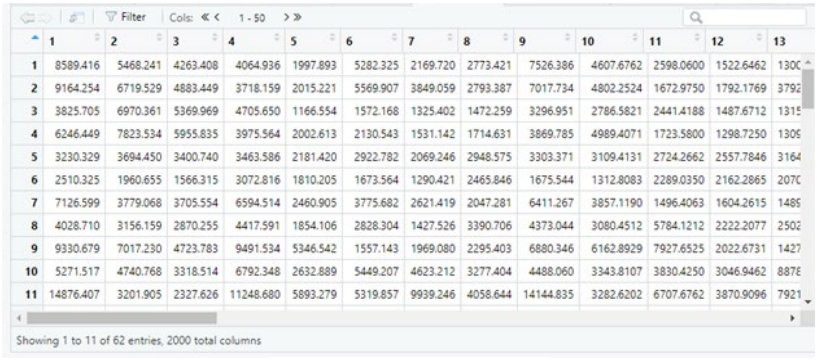
Потрібно провести аналіз диференційної експресії, щоб виявити гени, які мають різний рівень експресії між пацієнтами з раком та здоровими пацієнтами,

використовуючи пакет *limma*. Перевірити відповідність даних розподільчим припущенням. Підсумувати результати аналізу та створіти теплову карту для всіх генів з диференційною експресією.

2. Ознайомлення з даними

Розмірність наших даних - 62 рядки і 2000 стовпців. Це означає, що ми маємо 2000 генів і 62 зразки. З них 22 зразки класу 1, що відповідає нормальним тканинам, і 40 зразків класу 2, що відповідає пухлинним тканинам. (джерело даних: http://microarray.princeton.edu/oncology/affydata/index.html)

Набір даних складається з трьох основних змінних: *X* - матриця, що відображає рівні експресії 2000 генів для 62 зразків. Кожен рядок відповідає пацієнту, кожен стовпчик - гену. *Y* - числовий вектор довжиною 62, що задає тип зразка тканини (пухлина або норма). *gene.names* - вектор, що містить назви 2000 генів для матриці експресії генів X.



	1	2	3	4	5	6	7	8	9	10	11	12	13
1	8589.416	5468.241	4263.408	4064.936	1997.893	5282.325	2169.720	2773.421	7526.386	4607.6762	2598.0600	1522.6462	1300
2	9164.254	6719.529	4883.449	3718.159	2015.221	5569.907	3849.059	2793.387	7017.734	4802.2524	1672.9750	1792.1769	3792
3	3825.705	6970.361	5369.969	4705.650	1166.554	1572.168	1325.402	1472.259	3296.951	2786.5821	2441.4188	1487.6712	1315
4	6246.449	7823.534	5955.835	3975.564	2002.613	2130.543	1531.142	1714.631	3869.785	4989.4071	1723.5800	1298.7250	1309
5	3230.329	3694.450	3400.740	3463.586	2181.420	2922.782	2069.246	2948.575	3303.371	3109.4131	2724.2662	2557.7846	3164
6	2510.325	1960.655	1566.315	3072.816	1810.205	1673.564	1290.421	2465.846	1675.544	1312.8083	2289.0350	2162.2865	2070
7	7126.599	3779.068	3705.554	6594.514	2460.905	3775.682	2621.419	2047.281	6411.267	3857.1190	1496.4063	1604.2615	1485
8	4028.710	3156.159	2870.255	4417.591	1854.106	2828.304	1427.526	3390.706	4373.044	3080.4512	5784.1212	2222.2077	2502
9	9330.679	7017.230	4723.783	9491.534	5346.542	1557.143	1969.080	2295.403	6880.346	6162.8929	7927.6525	2022.6731	1427
10	5271.517	4740.768	3318.514	6792.348	2632.889	5449.207	4623.212	3277.404	4488.060	3343.8107	3830.4250	3046.9462	8878
11	14876.407	3201.905	2327.626	11248.680	5893.279	5319.857	9939.246	4058.644	14144.835	3282.6202	6707.6762	3870.9096	7921

Showing 1 to 11 of 62 entries, 2000 total columns

Рисунок 1 – Дані

3. Попередня обробка даних

На цьому кроці ми спробуємо краще зрозуміти дані. Використовуючи гістограми, Boxplot, Q-Q діаграми та інші графічні тести, ми спостерігатимемо за розподілом наших даних і вирішимо, чи потрібно використовувати будь-яку трансформацію (нормалізацію) для покращення подальшого аналізу.

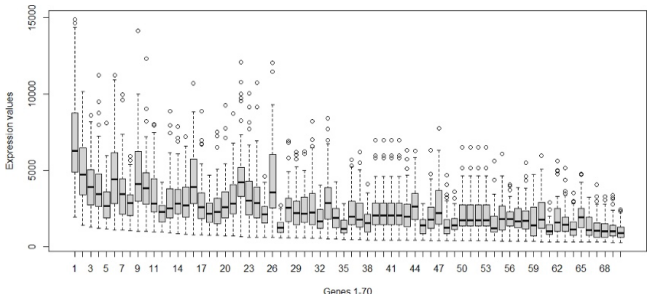


Рисунок 2 – Бокс-діаграма для перших 70 генів

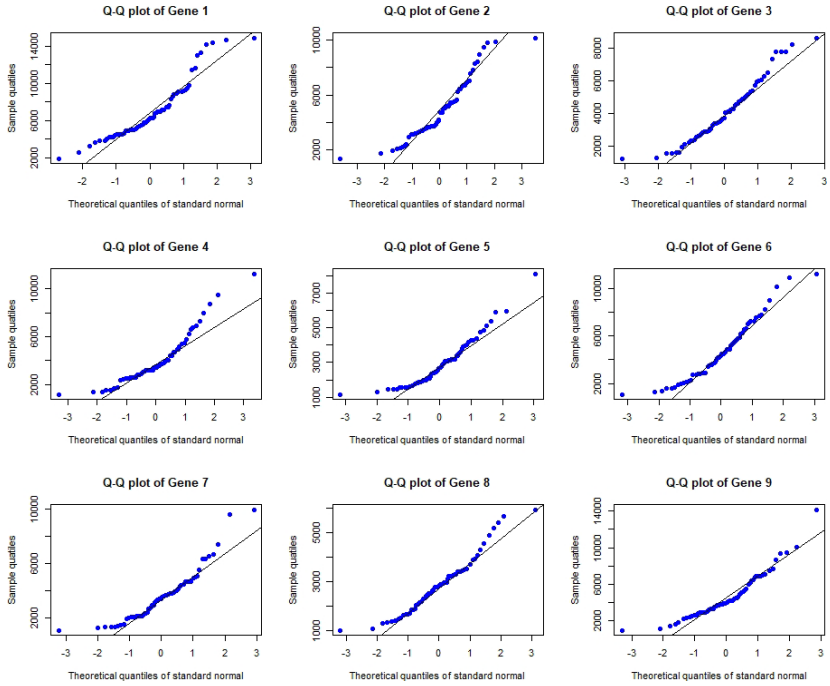


Рисунок 3 – QQ-діаграма

Припущення щодо трансформації та розподілу

1. Перший підхід: getNormMatrix()

У першому підході спробуємо нормалізувати дані за допомогою довідкового пакета *getNormMatrix*(). Результати застосування цієї процедури нормалізації до наших даних не показують значних покращень у розподілі даних.

2. Другий підхід: log2()

У другому підході використовуємо логарифм основи 2 для перетворення даних. І на основі візуалізації виглядає, що другий підхід працює краще, ніж перший.

3. Третій підхід: ZScore

Спробуємо перетворення *Z-рахунку* на даних (*Z* = (*X* - *середнє значення*)/*середнє квадратичне відхилення*). Але знову ж таки, цей процес не був успішним і не перевершив перетворення *log2.*

Спостерігаючи за всіма генами, чітко бачимо різну поведінку перших генів та останніх генів.

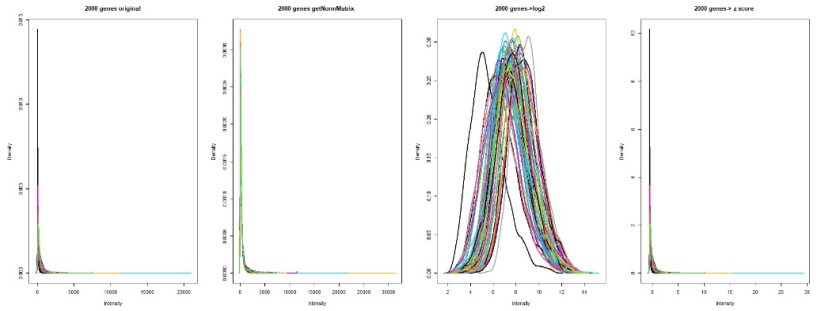


Рисунок 4 - Порівняння графіків щільності трансформації

4. Методи корекції тестування

Отримання списку генів, що експресуються деферентно, означає, що нам потрібно вибрати абсолютний поріг зміни *log2-кратності (стовпчик* log2FoldChange у виводі функцій з пакетів R) та скориговане *p-значення* (стовпчик *padj)*.

Проаналізуємо розподіл -значення за допомогою візуалізації гістограми. Виходячи з розподілу *p*-значень (Рис.6), ми можемо помітити, що для скоригованого *значення p* - ми маємо пік при 0, але також маємо пік до 1.

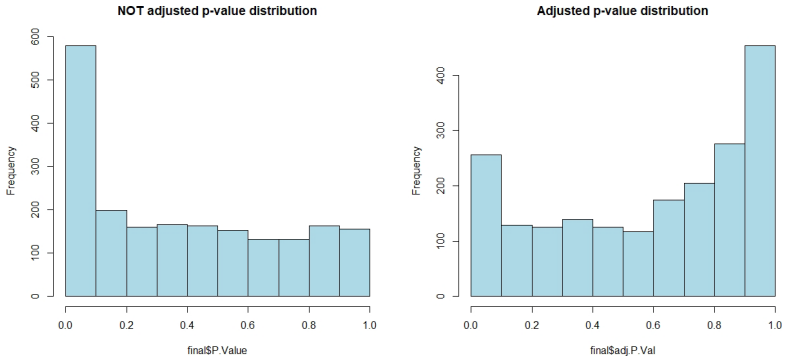


Рисунок 5 – Гістограма *p*-значень

Розглянемо найпоширеніші методи контролю FWER: метод Бонферроні, метод Хольма та міра похибки багаторазового тестування (FDR).

Результати показали, що FDR менш консервативний, і оскільки інші методи дають лише 24, 25 генів, будемо використовувати FDR, вважаючи деферентно експресованими 143 гени із загальної кількості 2000.

У файлі розмітки R ми відняли від оригіналу гени, які були класифіковані FDR як найбільш диференційовано експресовані, і створили підмножину, яка буде використана для подальшого аналізу. На рисунку 8 показано теплову

карту генів до і після того, як створили підмножину генів, відібраних за допомогою FDR.

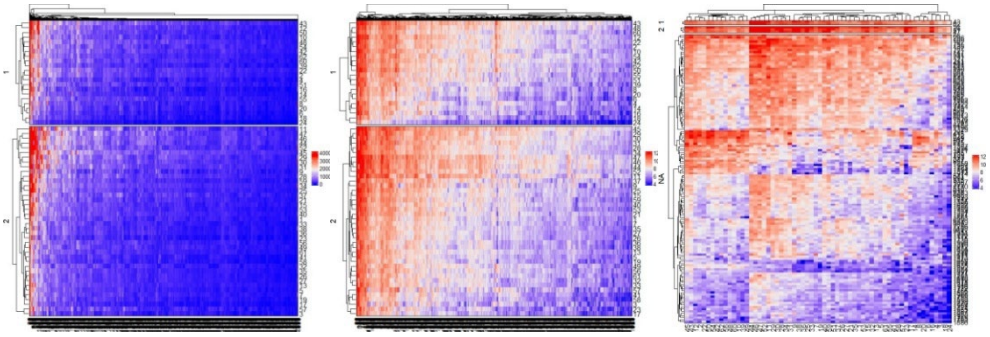


Рисунок 6 – Теплові карти

Висновки.

У даному дослідженні проведено виявлено 143 гени з диференційною експресією, що були відібрані на основі корекції за FDR. Було проведено детальну оцінку методів нормалізації, де найкращі результати продемонстрував підхід log2-трансформації. Використання гістограм *p*-значень та теплових карт дозволило візуалізувати результати та підкреслити значущі зміни у профілі експресії генів. Отримані результати можуть бути використані для подальших досліджень молекулярних механізмів раку та потенційної розробки нових біомаркерів для діагностики захворювання. Всі аналізи та обробка даних виконані за допомогою середовища R.

Література:

[1] . Jae K. Lee. Statistical Bioinformatics: For Biomedical and Life Science Researchers. - Wiley-Blackwell. 2014. – 368 p.

Стаття відправлена: 13.02.2025 р.