УДК: 618.19-006.6-07-092.4:575.001.5:007:004:002.6:681.31
DOI: http://dx.doi.org/10.11603/mie.1996-1960.2016.2.6476

КРОС-ПЛАТФОРМНЕ ОБ'ЄДНАННЯ ДАНИХ МІКРОМАСИВ-ЕКСПЕРИМЕНТІВ ТА ЙОГО ВПЛИВ НА ЗНАЧЕННЯ ГЕННОЇ ЕКСПРЕСІЇ ПРИ АНАЛІЗІ ЗРАЗКІВ РАКОВИХ ПУХЛИН МОЛОЧНОЇ ЗАЛОЗИ ЛЮДИНИ

А. О. Фролова, В. С. Бондаренко[1], М. Ю. Оболенська

Інститут молекулярної біології і генетики НАН України
[1]Київський національний університет імені Тараса Шевченка

У різних галузях біології та медицини накопичилася значна кількість даних широкомасштабних досліджень генної експресії за використання мікромасив-технологій. Тому виникає нагальна потреба в порівнянні, об'єднанні й аналізі цих даних з метою підвищення інформативності та статистичної достовірності результатів аналізу. Однак, процес об'єднання результатів мікромасив-експериментів на базі крос-платформного аналізу ускладнений існуванням багатьох мікроарей-платформ, різних за технологією виготовлення, методикою нанесення проб на чип і різновидністю дизайну проб. Ці особливості кожного з досліджень повинні бути враховані при обробленні та аналізі результатів мікромасив-експериментів, отриманих як на однакових, так і на різних платформах.

Мета роботи: дослідити як різні способи оброблення та об'єднання даних мікромасив-експериментів за використання різних платформ впливають на значення генної експресії.

Зміни в процесі оброблення мікромасив-експериментів впливають на результати даних експресії. Реанотація проб на сучасні версії баз даних геномних/транскриптомних послідовностей дає більш точний результат. Також вибір пробсетів на основі відповідності проб до послідовності конкретного гена (його транскриптів), а саме підходи, показані проектом BrainArray та використаними нами оцінками специфічності та чутливості проб, дають більш зважені та подібні результати, що відображено в аналізі кореляції експресії генів між різними варіантами обробки та кластерному аналізі профілей генної експресії.

Об'єднання даних із різних досліджень допомагає отримати більш прогнозовану похибку при класифікації зразків ракових пухлин молочної залози людини за молекулярними підтипами даних, ніж при використанні даних із окремих досліджень. Цей результат слугує підґрунтям для використання саме такого підходу при дослідженні проблем, що пов'язані з класифікацією будь-яких біологічних даних, особливо якщо заздалегідь інформація про зразки невідома, а також спонукає до подальшого дослідження впливу злиття даних на значення генної експресії.

Ключові слова: генна експресія, мікромасів-технології, ракові пухлини молочної залози людини.

КРОСС-ПЛАТФОРМЕННОЕ ОБЪЕДИНЕНИЕ ДАННЫХ МИКРОМАСИВ-ЭКСПЕРИМЕНТОВ И ЕГО ВЛИЯНИЕ НА ЗНАЧЕНИЕ ГЕННОЙ ЭКСПРЕССИИ ПРИ АНАЛИЗЕ ОБРАЗЦОВ РАКОВЫХ ОПУХОЛЕЙ МОЛОЧНОЙ ЖЕЛЕЗЫ ЧЕЛОВЕКА

А. А. Фролова, В. С. Бондаренко[1], М. Ю. Оболенская

Институт молекулярной биологии и генетики НАН Украины
[1]Киевский национальный университет имени Тараса Шевченко

В различных областях биологии и медицины накопилось значительное количество данных широкомасштабных исследований генной экспрессии с использованием микромасив-технологий. Поэтому возникает насущная необходимость в сравнении, объединении и анализе этих данных с целью повышения информативности и статистической достоверности результатов анализа.

Цель работы: исследовать как различные способы обработки и объединения данных микромасив-экспериментов с использованием различных платформ влияют на значение генной экспрессии.

Объединение данных из разных исследований помогает получить более прогнозируемую погрешность при классификации образцов раковых опухолей молочной железы человека с молекулярными подтипами данных, нежели при использовании данных из отдельных исследований.

Ключевые слова: генная экспрессия, микромасив-технологии, раковые опухоли молочной железы человека.

THE INFLUENCE OF CROSS-PLATFORM MICROARRAY DATA INTEGRATION ON GENE EXPRESSION VALUES IN HUMAN BREAST CANCER SAMPLES

A. O. Frolova, V. S. Bondarenko[1], M. Yu. Obolenska

Institute of Molecular Biology and Genetics of NAS of Ukraine
Taras Shevchenko Kyiv National University

Introduction. Currently advances in different fields of biology and medicine accumulated large amount of high-throughput microarray datasets. It allows comparing, merging and analyzing such data in order to get more informative and statistically significant results. However, the process of cross-platform microarray data integration is complicated by different platform designs, that is the kind of probes used, the hybridization paradigm, the labeling and production methods. Such characteristics of microarray platforms should be considered when analyzing datasets from one study and even more when merging datasets.

Aim. We investigate different variants of data processing and integration pipelines and their influence on gene expression.

Results. The changes in the microarray data processing influence the gene expression values. Probes reannotation on the current version of genome/transcriptome databases produces more accurate results. Additionally, obtaining unique probeset to gene mapping based on probe sequence alignment to the particular gene transcripts (BrainArray project; probes specificity and sensitivity scores) gives more consistent and significant results.

Conlustions. Integrating datasets from different studies average a classification error, which was shown across hundreds of datasets of human breast cancer samples. Such approach allowed to differentiate better between cancer molecular subtypes in comparison with single-study analysis and can be used in other similar studies, especially if the sample specification is unknown.

Key words: gene expression, mikromasiv technology, cancerous tumors of human breast.

Вступ. У різних галузях біології та медицини накопичилася значна кількість даних широкомасштабних досліджень генної експресії за використання мікромасив-технологій, про що свідчить створення багатьох репозиторіїв первинних даних біологічних експериментів із можливістю вільного доступу до них та їхнього використання різними дослідниками. Найбільшою базою даних публічного доступу до результатів мікромасив-експериментів є GEO (Gene Expression Omnimbus), що містить дані з понад 17 тис. різних досліджень та понад 200 тисяч зразків, кількість яких продовжує зростати (рис. 1). GEO має зручний та зрозумілий інтерфейс, базові функції аналізу експериментів, багато посилань на суміжні ресурси та належить до розгалуженої системи баз даних Національних Інститутів Здоров'я США [1].

Також існує європейська версія бази даних із аналогічною функцією - ArrayExpress, що містить порівняну кількість даних, значна частина яких перекривається з GEO, проте на відміну від останньої не містить програмного забезпечення для базового аналізу даних [2].

З накопичення великої кількості даних виникає нагальна потреба в порівнянні, об'єднанні й аналізі даних у кожній галузі досліджень з метою підвищення загальної інформативності даних та їхньої статистичної достовірності. Проте, процес об'єднання результатів мікромасив-експериментів на базі крос-платформного аналізу не є тривіальним, що пов'язано з існуванням мікрарей-платформ, різних за технологією виготовлення, методикою нанесення проб на чип, різновидністю дизайну проб тощо. Ці властивості й особливості кожного з досліджень повинні бути враховані при обробці й аналізі результатів мікромасив-експериментів, отриманих як на однакових, так і на різних платформах.

Мета роботи: дослідити як різні способи оброблення та об'єднання даних мікромасив-експериментів за використання різних платформ впливають на значення генної експресії.

Матеріали та методи дослідження. Проблеми оброблення мікромасив-експериментів. Метод ДНК-мікрочипів або ДНК-мікромасивів (DNA microarray) є сучасним методом молекулярної біології, що заснований на ДНК-ДНК або ДНК-РНК гібридизації між пробою (олігонуклеотидом, закріпленим на певній твердій поверхні) та комплементарною їй мішенню - молекулами РНК або ДНК (рис. 2). Сучасні мікромасив-чипи нараховують сотні тисяч проб олігонуклеотидів та застосовуються в аналізі генної експресії для з'ясування ділянок ампліфікації в геномі, визначення метилування геному, мутацій, сайтів зв'язування транскрипційних факторів та багато іншого [3].

Наразі, переважного застосування набули чипи двох виробників - Affymetrix [4] та Illumina [5], що суттєво відрізняються за дизайном. Чипи Affymetrix виготовлені шляхом in situ синтезу олігонуклеотидів довжиною 25 пар нуклеотидів. Кожен чип може містити до 900 тисяч різних

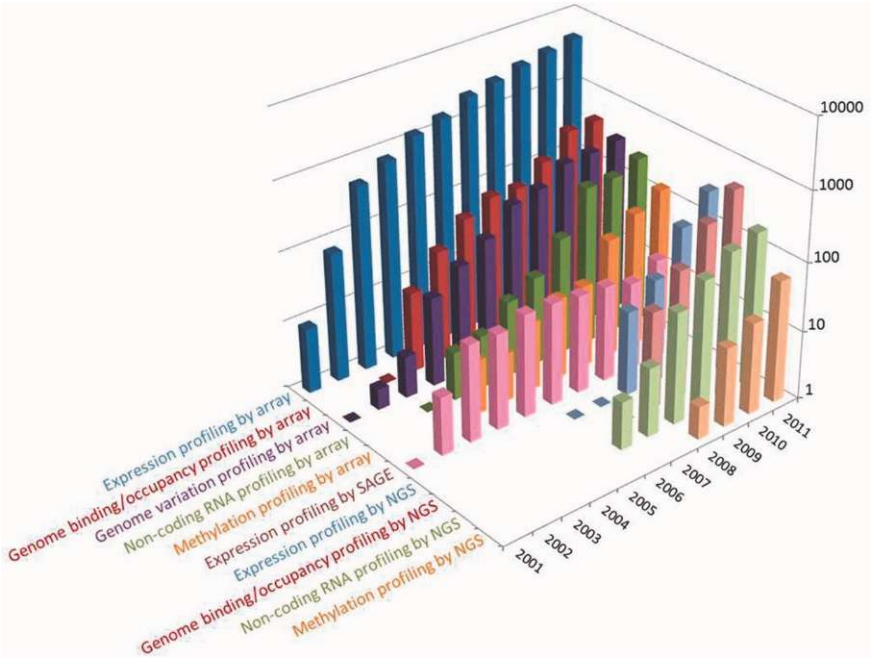

Рис. 1. Кількість різних широкомасштабних геномних досліджень, що з'являються в базі даних GEO кожного року

олігонуклеотидів, де кожнии олігонуклеотид представлений мільйонами копій. Копії одного олігонуклеотиду представлють собою одну пробу. У свою чергу різні проби зібрані у пробсети (probesets). Пробсет - це набір проб, що специфічні до різних ділянок нуклеотидної послідовності певної мішені. Залежно від дизайну конкретної платформи, пробсет містить в середньому від 10 до 20 проб.

На відміну від Affymetrix, платформи Illumina влаштовані інакше. Насамперед, носієм олігонуклеотиду є не тверда пласка поверхня, а кварцеві кульки діаметром 3 мкм. На поверхню кульок нанесені копії олігонуклеотиду довжиною 79 нуклеотидів, з яких послідовність з 50 нуклеотидів на кінці є абсолютно специфічною до мішені, а послідовність з 29 нуклеотидів виконує дві функції: використовуються як лінкер за який олігонуклеотид прикріплюється до поверхні кульки та як штрих-код, послідовність якого означає ген-мішень. Така система кодування використовується для встановлення відповідності між пробою та мішенню. В дизайні платформ Illumina поняття проб та пробсетів ідентичні, чипи можуть містити близько 50 тисяч проб.

У роботі висвітлено проблеми, що виникають при обробленні мікромасив-експериментів.

Проблема застарілої анотації. Дизайн більшості мікромасив-чипів був спроектований у середньому 4-5 років тому, проте якість і повнота баз даних послідовностей усіх (експресованих) генів/транскриптів уже встигла зазнати суттєвих змін, і послідовності старих чипів можуть відповідати іншим генам/транскриптам, ніж тим, що вказані виробником чипів. Наприклад, частота оновлення анотації проб чипів Affymetrix становить раз на рік, а бібліотек анотацій відомого проекту Bioconductor - декілька разів на рік [7]. Тому для отримання найточнішої інформації обов'язковим є встановлення відповідності проб та послідовностей генів/транскриптів у сучасних версіях відповідних баз даних [6].

Проблема визначення взаємно однозначної відповідності пробсетів та їхніх мішеней. Інтеграція даних із чипів, що належать до різних платформ, вимагає визначення однозначної відповідності між

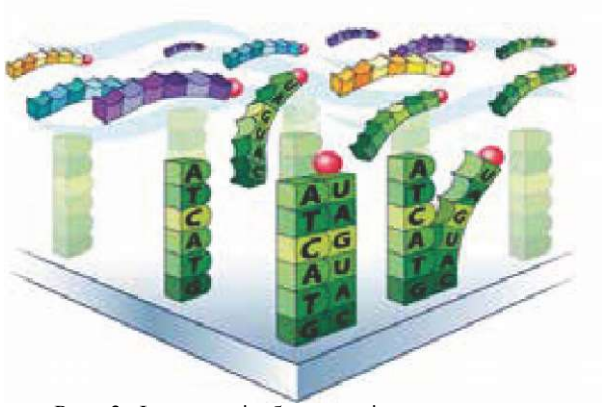

Рис. 2. Ілюстрація будови мікромасив-чипа, на поверхні якого розташовані проби, що є комплементарними до молекул-мішеней. Проби позначені стовпчиками, а мішені - дугами

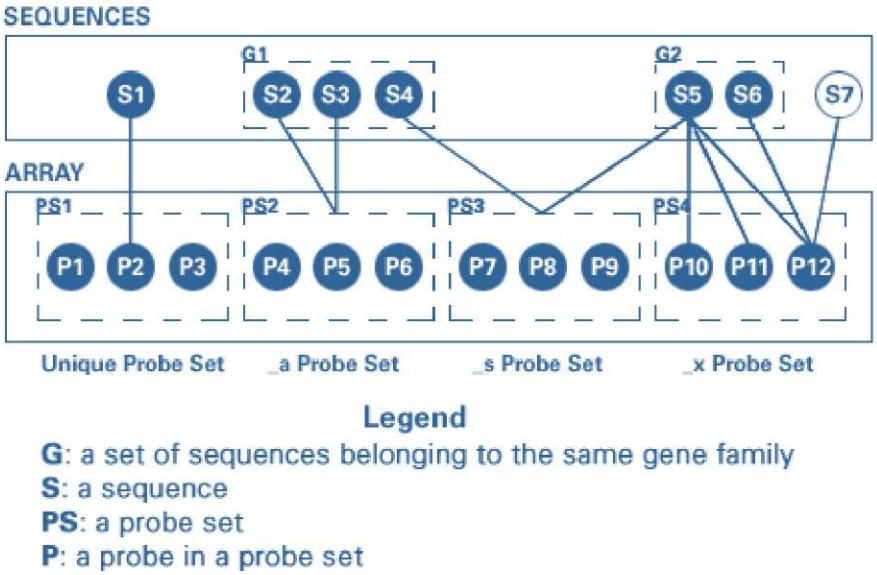

Рис. 3. Дизайн пробсетів чипу Affymetrix: PS1 - унікальний пробсет, що відповідає одному транскрипту мішені; PS2 - пробсет, що відповідає декільком транскриптам; PS3 - пробсет, що відповідає транскриптам різних генів одночасно; PS4 - пробсет, що містить проби, які не відповідають жодним генам

пробсетами та їхніми мішенями (один пробсет - один ген). Нажаль, часто з суто технічних причин, у дизайні платформи наявна певна кількість пробсетів, що впізнають транскрипти кількох різних генів, водночас один ген може бути представлений кількома пробсетами. В різних платформах ця кількість суттєво різниться. Це чудово ілюструється дизайном пробсетів чипів Affymterix (рис. 3).

З метою встановлення однозначної відповідності між пробсетами та генами застосовують декілька методів. Найпростішим варіантом є випадковий вибір одного пробсету серед кількох, що впізнають один і той же ген. Також можна вираховувати середнє значення або суму інтенсивності сигналів проб у цих пробсетах. Більш обґрунтованим методом є розрахунок певного критерію пробсету, на основі якого здійснюється вибір. Такими можуть бути, наприклад, специфічність та чутливість зв'язування пробсету з мішенню [11]. Окремим методом є визначення самим дослідником пробсетів, де проби в кожному відповідали б одному гену/транскрипту. Отже, остаточна анотація пробсетів буде суттєво відрізнятися від запропонованої виробником [10]. Проблема полягає у тому, який із цих методів обрати для отримання найточнішого відображення експресії певного гену.

Результати та їх обговорення. В роботі використано дані мікромасив-експериментів із 3-х різних досліджень раку молочної залози, що отримали з бази даних GEO, які загалом налічують 742 зразка (табл. 1).

Кожен зразок має супровідну інформацію (вік пацієнта, деякі клінічні показники тощо), яку можна знайти за ідентифікатором дослідження в базі даних GEO. В нашій роботі для класифікації профілей генної експресії використовували молекулярні підтипи раку, а саме TNBC (Triple-negative breast cancer), Her2, Luminal A та Luminal B. Ці підтипи визначаються на основі профілей генної експерсії, хоча мають певні клінічні особливості (різний рівень експресії генів рецепторів естрогену, прогестерону та епідермального фактору росту) [8].

Таблиця 1

Використані дані мікромасив-експериментів

Дослідження	Молекулярні підтипи раку молочної залози				Загалом
	TNBC	Her2	Luminal A	Luminal B	
GSE65216	55	39	29	30	153
GSE30682	58	14	145	58	275
GSE58644	50	20	206	38	314
Загалом	163	73	380	126	742

Нами порівняно три варіанти оброблення та поєднання даних мікромасив-експериментів, визначено різні схеми оброблення даних, що відрізняються підходами до вирішення проблеми відповідності пробсетів та їх мішеней.

- **Варіант 1.** Класичний підхід із використанням бібліотек Bioconductor у середовищі програмування R.
- **Варіант 2.** Підхід на основі визначення нових пробсетів (які містять проби, що відповідають лише одному гену) за використання анотацій BrainArray проекту [9].
- **Варіант 3.** Підхід на основі обчислення специфічності та чутливості пробсетів для обрання пробсету з кращою оцінкою.

Варіант 1 передбачає використання анотацій проб із бібліотек Bioconductor та пробсетів, визначених виробником. Однак цей підхід не вирішує проблему взаємно однозначної відповідності пробсет-ген, тож на виході отримуємо декілька пробсетів, що відповідають одному гену. Існує три способи, як цього уникнути: обирати один пробсет із поміж пробсетів, що відповідають одному гену, випадковим чином, сумувати значення пробсетів або ж брати середнє між значеннями пробсетів. У літературі пропонуються всі три підходи, однак, немає відомостей, який із них краще [11].

Варіант 2 передбачає визначення пробсетів, відмінних від тих, що запропоновані виробником, і використовує CDF (chip definition files) файли опису пробсетів за проектом BrainArray [13]. В такі пробсети входять тільки проби, що відповідають одному конкретному гену, тому, відповідно сумарне значення експерсії пробсету буде відрізнятися від того, що отримано у Варіанті 1. Варто зазначити, що BrainArray CDF файли існують тільки для чипів Affymetrix, оскільки будова чипів Illumina зовсім інша - пробсети містять однакові проби (як було зазначено вище). Тож у Варіанті 2 до чипів Illumina застосовано підхід на основі обчислення специфічності та чутливості.

Варіанті 3 передбачає переанотацію усіх проб, тобто нуклеотидне вирівнювання послідовностей проб проти сучасної версії анотації транскриптому людини RefSeq, що здійснювали за допомогою інструменту BLAST [14]. RefSeq ідентифікатори транскриптів були переведені у відповідні генні ідентифікатори Entrez ID за допомогою Biomart [12]. Пробсети, що відповідають транскриптам з ідентифікаторами, що не були конвертовані та представляють собою переважно транскрипти з XM і XN маркуванням, були відфільтровані. Проба вважалась специфічною до певного транскрипту, якщо загальна оцінка вирівнювання (BLAST score) була більше 32 для проб Affymetrix і 63,5 для проб Illumina. Пробсети з менш ніж половиною специфічних проб від вихідної кількості були відфільтровані. Для кожного пробсету цільовий ген/транскрипт був визначений як такий, що специфічно упізнається більшістю проб у пробсеті. Отже, специфічність пробсету нами була вирахувана як сума значень специфічності проб в пробсеті поділена на кількість всіх проб у пробсеті, а чутливість пробсету - кількість транскриптів цільового гена, що специфічно детектуються всіма специфічними пробами поділена на кількість всіх транскриптів цільового гена. Отже, серед низки пробсетів, що відповідають одному гену, обирали такий пробсет, який має кращу оцінку специфічності та чутливості.

Після отримання списку пробсетів у кожному з застосованих варіантів, де кожен пробсет тепер відповідає окремому гену, провели такі операції:

1. Сумування проб у пробсеті (тобто визначення остаточного рівня експресії пробсета) та нормалізація пробсетів. Під нормалізацією мається на увазі корекція фонової інтенсивності сигналу, адже інтенсивність сигналу з кожного пробсету відповідає сумарній кількості зв'язаних із ним мішеней та складається з двох компонентів: дійсного, спричиненого специфічним зв'язуванням мішені з пробами пробсету, та фонового, спричиненого неспецифічним зв'язуванням. Статистичні методи, що застосовують після отримання зображення мікроарей-чипа мають на меті максимально точне розділення двох компонентів сигналу та видалення фонового. Нами було використано найбільш поширений метод RMA (Robust Multi-array Average), який заснований на припущені про те, що фоновий компонент сумарної інтенсивності сигналу пробсета може бути описаний нормальним розподілом, а дійсний - експоненціальним [15].
2. Перевірку якості даних експресії зроблено пакетом arrayQualityMetrics, що включає різноманітні статистичні тести [16].
3. Усунення технічної варіативності в даних експресії (batch-effect removal) - пакетом ComBat [17]. Технічна варіативність, що впливає на значення експресії, може бути зумовлена різними умовами виконання експерименту, яка може призвести до того, що зразки будуть класифі-

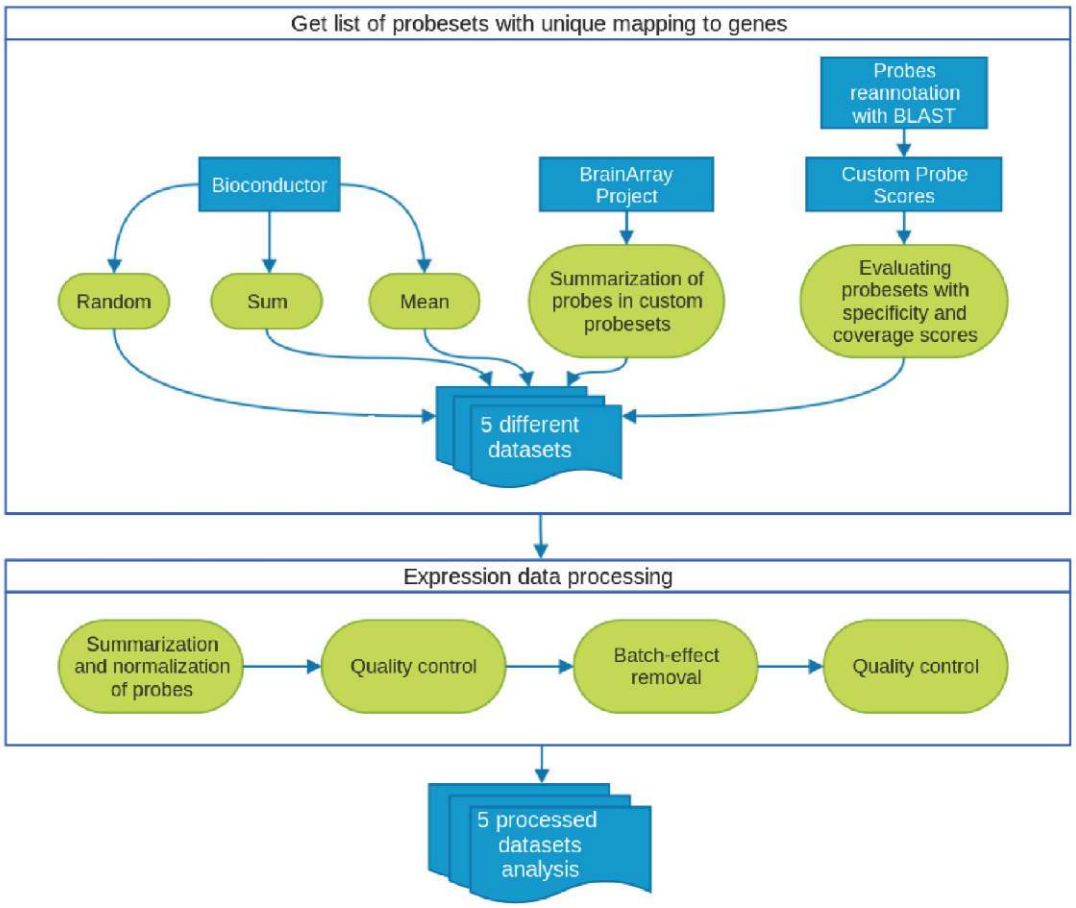

Рис. 4. Схема різних варіантів оброблення даних мікромасив-експериментів. Кінцевий результат 5 масивів даних взятих з однакових досліджень, оброблених у різний спосіб

куватися за часом проведення експерименту, а не за біологічною ознакою.

Отже, отримано 5 різних масивів даних, де 3 - з Варіанту 1, 1 - з Варіанту 2, та 1 - з Варіанту 3 (узагальнена схема зображена на рис. 4). Ще раз варто зазначити, що всі масиви даних отримані з одних і тих же «сирих» даних і відрізняються лише методом обробки.

Кожне з трьох досліджень було зроблено з використанням різної платформи з різною кількістю пробсетів (табл. 2). Платформи Affymetrix та Illumina значно відрізняються за специфічністю проб до транскриптів-мішеней. Медіана розподілу специфічності для Affymetrix платформ становить 0.70 та 0.68, тоді як практично всі пробсети платформи Illumina мають майже абсолютну специфічність. Розподіли ж чутливості пробсетів за формою є в значній мірі подібними між платформами, що свідчить про достатньо високий рівень розпізнання всіх анотованих транскриптів гена. Відсоток пробсетів, картованих на однакові гени для платформи Illumina становить близько 19 %, в платформі HuGene 1.0ST Array - 7 % та в HG U133 Plus 2.0 - близько 25 %.

Після об'єднання даних із усіх трьох платформ отримали близько 12 тисяч пробсетів поміж 3 різних варіантів оброблення даних, де кожен пробсет відповідає окремому гену (табл. 3).

Таблиця 2

Опис платформ мікромасивів

Назва платформи	Варіант 1	Варіант 2	Варіант 3	% генів, представлених кількома пробсетами
Affymetrix HGU133 Plus 2.0	54675	19702 (36 %)	34877(64 %)	25 %
Affymetrix HG 1.0 ST	33297	19700 (59 %)	20252(61%)	7 %
Illumina HumanWG-6 v3.0	48803	-	19257(40 %)	19 %

Таблиця 3

Кількість пробсетів після об'єднання даних з трьох платформ

Варіант 1	Варіант 2	Варіант 3	Перетин між всіма варіантами
16758	13969	12783	11547

Статистичне оцінювання варіантів оброблення. Оскільки кількість кожного з молекулярних підтипів раку молочної залози різниться поміж трьома дослідженнями, для статистичної достовірності правильно було б взяти однакову кількість підтипів із кожного дослідження. Однак, вибір зразків може вплинути на остаточний результат, тому нами згенеровано 100 підмасивів даних для кожного з 5 варіантів оброблення даних, 168 зразків у кожному (14 зразків кожного з 4 підтипів раку, обраних випадковим чином із 3 різних досліджень), як показано на рис. 5. Отже, маємо 100 масивів даних для кожного з 5 випадків, тобто усього 500 наборів даних і можемо статистично достовірно оцінити різні варіанти оброблення даних.

Кореляція експресії генів між різними варіантами оброблення. Проаналізовано як різні варіанти оброблення даних корелюють між собою (рис. 6), порахувавши коефіцієнт кореляції рангу Спірмена, що є непараметричною мірою статистичної залежності між двома змінними (в даному випадку між експресією конкретного гена після різних варіантів оброблення) [18]. Виявилось, що Варіант 2 та 3, тобто обробка на основі BrainArray проекту та визначення оцінки специфічності та чутливості взаємодії проби з мішенню добре корелюють між собою, тобто продукують схожі значення експресії. Так само схожі між собою результати у випадку довільного вибору пробсета та у випадку середнього значення експресії між пробсетами. Однак, сума пробсетів однаково погано корелює з результатами оброблення будь-яким методом. Це може бути пояснено тим, що на відміну від анотації BrainArray, в пробсетах залишаються неспецифічні проби, та сумування таким чином додає небажаний шум в експресію певного гену.

Додатковим доказом впливу методу оброблення на експресію генів є порівняння розподілів інтенсивностей сигналу пробсетів за використання різних анотацій. Так, для платформи HG U133 Plus 2.0 стандартна анотація (Варіант 1) має одномодальний розподіл, у той час як BrainArray

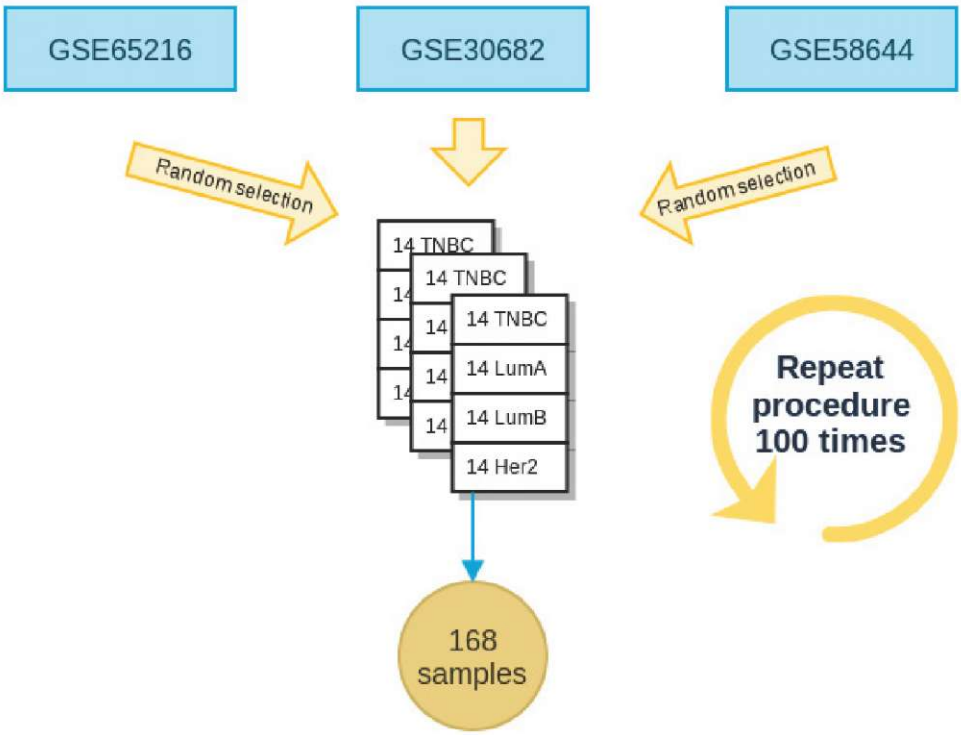

Рис. 5. Процес генерації підмасивів даних, 168 зразків кожен, шляхом довільного вибору зразків із загальної кількості 742

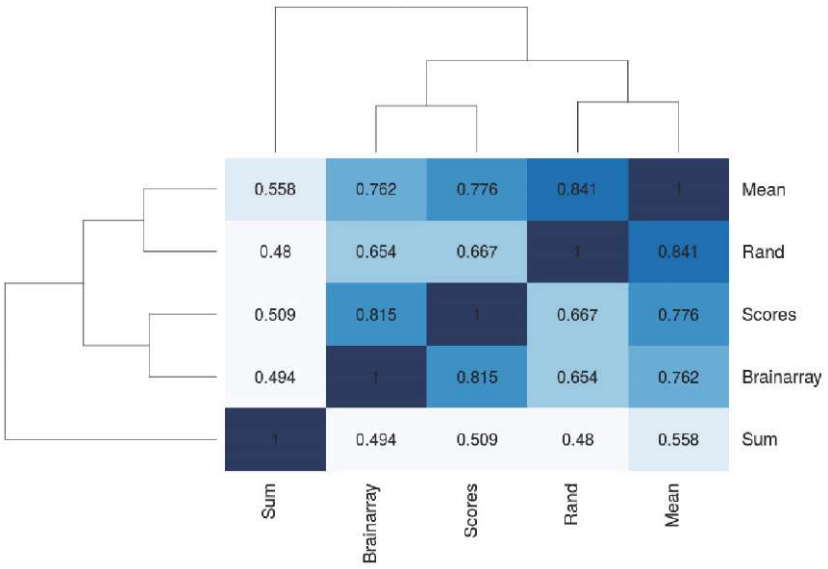

Рис. 6. Кореляція генної експресії після різних варіантів оброблення даних. Вказано середнє значення поміж 100 масивами даних

проект та власна переанотація за допомогою BLAST (Варіант 2 та 3) мають бімодальний розподіл із двома піками (рис. 7). Останнє зазвичай пов'язують з групами генів, що є високоекспресованими та неекспресованими (або низькоекспресованими), тобто маємо більш точну диференціацію між експресіями генів [19, 20].

Кластерний аналіз. Для аналізу того, які масиви даних, отримані після різних варіантів оброблення, дозволяють краще класифікувати зразки за молекулярними типами раку, застосували кластерний аналіз. А саме кластеризацію методом k-середніх (k-means clustering) - впорядкування множини об'єктів у порівняно однорідні групи. В початковий момент роботи алгоритму довільним чином обираються центри кластерів, далі для кожного елемента множини ітеративно обраховується відстань від центрів із приєднанням кожного елемента до кластера з найближчим центром. Для кожного з отриманих кластерів обчислюються нові значення центрів, намагаючись при цьому мінімізувати певну функцію оцінки, після чого повторюється процедура перерозподілу елементів між кластерами. Даний тип аналізу належить до машинного навчання без учителя, тобто не потребує заздалегідь визначених відомостей про приналежність деяких елементів до кластеру, та широко застосовується для аналізу мікромасив-експериментів [21]. Результати розбиття зразків масиву даних на 4 кластери були порівняні з приналежністю цих зразків до одного з 4-х під-

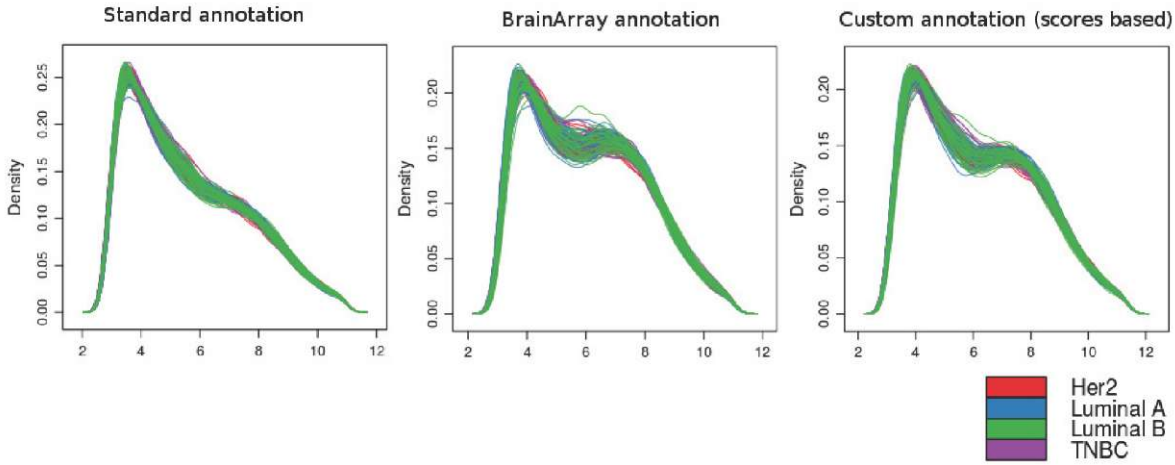

Рис 7. Розподіли інтенсивностей сигналу пробсетів при використанні різних анотацій для даних платформи HG U133 Plus 2.0, що має найбільший відсоток генів, представлених кількома пробсетами

Таблиця 4

Оцінка класифікації об'єднаних зразків за молекулярним підтипом раку методом k-середніх

Похибка класифікації	Варіант 2	Варіант 3	Варіант 1 за середнім значенням	Варіант 1 за випадковим принципом	Варіант 1 за сумою значень
Середнє похибки	0,1883	0,1896	0,1902	0,1912	0,2030
Стандартне відхилення	0,0299	0,0293	0,0299	0,0285	0,0296

типів даних та обчислено рівень похибки класифікації [22]. Процедуру було повторено на кожному зі 100 масивів даних і для кожного з 5-ти варіантів оброблення й обчислено середнє значення похибки класифікації, де 0 - відсутність похибки, а 1 - повністю невірна класифікація. Як видно з результатів у табл. 4, обробка даних за Варіантом 2 забезпечує найкращу класифікацію зразків.

Результат можна показати наочно, використовуючи метод головних компонент для побудови графіку (рис. 8). Цей метод теж широко застосовується для аналізу мікромасив-даних, він дає можливість по m - числу вихідних ознак виділити m головних компонент або узагальнених ознак [23].

Результат кластерного аналізу корелює з порівнянням експресії генів за допомогою рангу Спірмена, тобто в обох випадках сума пробсетів дає результат, відмінний від інший варіантів. Але кластерний аналіз також надає не тільки порівняльний, а й якісний аналіз.

Для того, щоб зрозуміти, як злиття даних із різних досліджень впливає на результат кластеризації, здійснили такий самий кластерний аналіз, але для кожного дослідження окремо. В табл. 5 наведено значення похибок класифікації після обробки даних Варіантом 2, що показав найкращі результати, для об'єднаних даних та для кожного дослідження окремо.

Таблиця 5

Оцінка класифікації за молекулярним підтипом раку методом k-середніх після обробки даних за Варіантом 2

Назва	Похибка класифікації
GSE58644	0,32
GSE30682	0,24
GSE65216	0,10
Об'єднані дані	0,19

Результати, наведені в табл. 5, свідчать про те, що аналізуючи окремі дослідження, можемо отримати як задовільний результат (тобто легко можна відрізнити зразки за різними типами), так і менш задовільний. Однак, дослідник не завжди може оцінити результат, оскільки зазвичай головним завданням як раз і є визначення типів зразків, що заздалегідь невідомо. Тож природно, що важливим стає оцінювання похибки при об'єднанні даних із різних досліджень, яка може бути кращою, ніж у найгіршому з досліджень, але водночас гіршою за похибку в найкращому з досліджень.

Висновки. Результати даних експресії генів у кожному дослідженні й особливо після крос-платформного

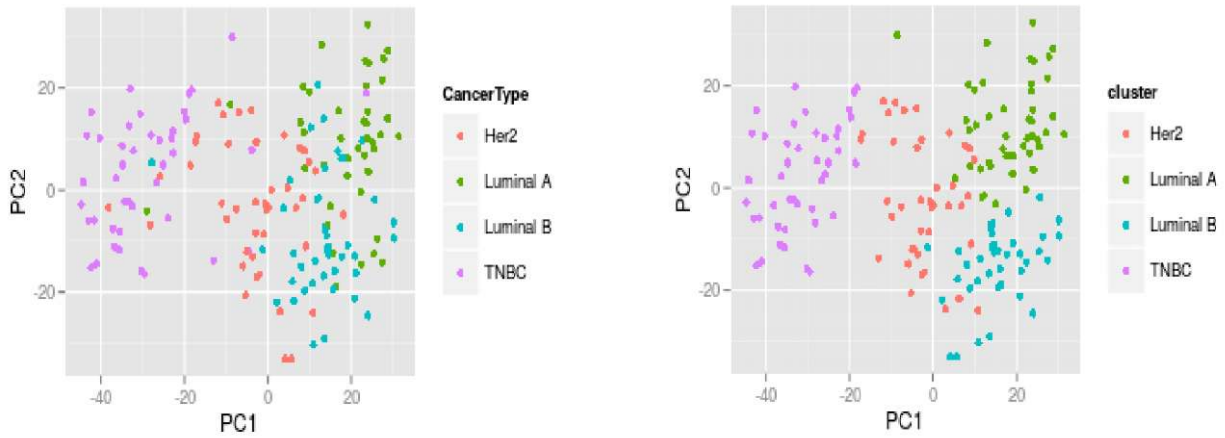

Рис. 8. Графіки головних компонент одного масиву даних із 168 зразків. Ліворуч - вихідна класифікація зразків за молекулярними підтипами раку, праворуч - отримана за допомогою методу k-середніх

об'єднання даних залежать від методу оброблення мікромасив-експериментів. Анотація проб, організація пробсетів та формування взаємно однозначної відповідності ген - пробсет передує іншим етапам обробки та суттєво впливає на кінцевий результат.

У роботі визначено, що реанотація проб на сучасні версії баз даних геномних/транскриптомних послідовностей дає більш точний результат. Також вибір пробсетів на основі відповідності проб до послідовності конкретного гена (його транскриптів), а саме підходи показані проектом BrainArray та використаними нами оцінками специфічності та чутливості проб, дають більш подібні та зважені результати.

Об'єднання даних із різних досліджень допомагає отримати більш прогнозовану похибку при класифікації зразків ракових пухлин молочної залози людини за молекулярними підтипами даних, ніж при використанні даних із окремих досліджень. Цей результат слугує підґрунтям для використання саме такого підходу при дослідженні проблем, що пов'язані з класифікацією будь-яких біологічних даних, особливо якщо заздалегідь інформація про зразки відсутня, а також спонукає до подальшого дослідження впливу злиття даних на значення генної експресії.

Література.

1. Barrett T. et al. NCBI GEO: archive for functional genomics data sets: update // Nucleic acids research 41.D1 (2013): D991-D995.
2. Rustici G. et al. ArrayExpress update - trends in database growth and links to data analysis tools // Nucleic acids research 41.D1 (2013): D987-D990.
3. Stoughton R. B. Applications of DNA microarrays in biology // Annu Rev Biochem; (2005): 74: 53-82.
4. Affymetrix. - Режим доступу: http://www.affymetrix.com/estore/.
5. Illumina. - Режим доступу: http://www.illumina.com/technology/ beadarray-technology.html.
6. Barbosa-Morais N. L. A re-annotation pipeline for Illumina BeadArrays: Improving the interpretation of gene expression data / Barbosa-Morais, N. L., Dunning, M. J., Samarajiwa, S. et al. // Nucleic Acids Res. 38, (2009).
7. BioCon. - Режим доступу: http://www.bioconductor.org/packages/release/ data/annotation/.
8. Schnitt S. J. Classification and prognosis of invasive breast cancer: from morphology to molecular taxonomy // Modern Pathology 23. - (2010): S60-S64.
9. Dai M. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data / Dai M., Wang P., Boyd A. D. et al. // Nucleic Acids Res. 33, 1-9 (2005).
10. Carter S. L. Redefinition of Affymetrix probe sets by sequence overlap with cDNA microarray probes reduces cross-platform inconsistencies in cancer-associated gene expression measurements / Carter S. L., Eklund A. C., Mecham B. H., Kohane I. S., Szallasi Z. // BMC Bioinformatics. - 6, 107 (2005).
11. Li Q. Jetset: selecting the optimal microarray probe set to represent a gene / Li Q., Birkbak N. J., Gyorffy B. et al. // BMC Bioinformatics. - 12, 474 (2011).
12. Durinck S. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt / Durinck S., Spellman P. T., Birney E., Huber W. // Nat Protoc. - 2009, 4(8):1184-1191.
13. BrainAr. - Режим доступу: http://brainarray.mbni.med.umich.edu/ Brainarray/Database/ CustomCDF/genomic_ curated_CDF.asp.
14. Boratyn G. M. BLAST: a more efficient report with usability improvements / Boratyn G. M., Camacho C., Cooper P. S. et al. // Nucleic Acids Res. (2013): 41, 29-33.
15. Irizarry R. Summaries of Affymetrix GeneChip probe level data / Irizarry R. A, Bolstad B. M., Collin F., Cope L. M. et al. // Nucleic Acids. (2003): Res. 31, e15.
16. Kauffmann A. ArrayQualityMetrics-a bioconductor package for quality assessment of microarray data / Kauffmann A., Gentleman R., Huber W. // Bioinformatics. (2009): 25(3), pp. 415-6.
17. Johnson W. E. Adjusting batch effects in microarray expression data using empirical Bayes methods / J. W. Evan, C. Li, A. Rabinovic // Biostatistics 8.1 (2007): 118-127.
18. Boulesteix A.-L. Stability and aggregation of ranked gene lists / Boulesteix A.-L., Slawski M. // Briefings in bioinformatics 10.5 (2009): 556-568.
19. Mokry M. et al. Integrated genome-wide analysis of transcription factor occupancy, RNA polymerase II binding and steady-state RNA levels identify differentially regulated functional gene classes // Nucleic acids research 40.1 (2012): 148-158.
20. Hochreiter S. A new summarization method for Affymetrix probe level data / Hochreiter S., Clevert D., Obermayer K. // Bioinformatics 22.8 (2006): 943-949.
21. Chen T.-Sh. et al. A combined K-means and hierarchical clustering method for improving the clustering efficiency of microarray // Intelligent Signal Processing and Communication Systems, 2005. - ISPACS 2005. Proceedings of 2005 International Symposium on. IEEE, 2005.
22. Hubert L. Comparing partitions / L. Hubert, P. Arabie // Journal of Classification. - 2, 193-218, 1985.
23. Quackenbush J. Computational analysis of microarray data // Nature reviews genetics 2.6 (2001): 418-427.