

UDC 577.218

Критерії оцінки результатів мікромасив-експерименту з дослідження транскриптому гепатоцитів щура під впливом інтерферону альфа

А. В. Куклін, Б. Т. Токовенко, М. Ю. Оболенська

Інститут молекулярної біології і генетики НАН України
Вул. Академіка Заболотного, 150, Київ, Україна, 03680

Av.kuklin@gmail.com

Транскриптом первинних гепатоцитів щура, культивованих з інтерфероном альфа (ІФНа) впродовж трьох і шести годин, досліджували із залученням олігонуклеотидних мікромасивів. Мета. Проведення поетапного аналізу результатів мікромасив-експерименту і визначення відповідності або невідповідності критеріїв оцінки результатів загальноприйнятим значенням. Методи. Аналіз здійснювали на файлах, отриманих після сканування мікромасивів Affymetrix, із використанням статистичного середовища R, пакетів функцій Bioconductor «affy», «simpleaffy», «affyPLM» і програми BRB Array Tools із застосуванням парного T-тесту. Результати. Усі мікромасиви пройшли контроль якості, нормалізовані і порівнянні між собою. За результатом T-тесту знайдено 28 і 124 диференційно експресованих гени відповідно після трьох і шести годин культивування клітин з ІФН. Висновки. Одержані дані відповідають узвичасним критеріям якості і придатні для подальшого встановлення їхнього біологічного значення. Використані в роботі R-коди можна застосовувати при обробці результатів мікромасив-експериментів.

Ключові слова: мікромасиви ДНК, інтерферон альфа, статистичне середовище R, Bioconductor, нормалізація.

Вступ. Технологія з використанням мікромасивів ДНК є потужним інструментом в арсеналі системної біології, що дозволяє одночасно аналізувати експресію багатьох тисяч генів [1–3]. Фізично мікромасив являє собою носій, на поверхню якого нанесено велику кількість олігонуклеотидів різної послідовності. З-поміж усіх виробників Affymetrix створює мікромасиви з найвищою щільністю зондів, що передбачає можливість аналізу найбільшої кількості генів.

У мікромасивах Affymetrix кожний ген представлено набором зондів (probe set), який складається з 11 типів послідовностей, повністю комплементарних (perfect match) різним ділянкам транскрипту, а також 11 типів послідовностей, які посередині мають один некомплементарний нуклеотид

(miss match). За допомогою останніх встановлюють рівень неспецифічного зв'язування (<http://www.ohsu.edu/xd/research/research-cores/gmsr/about/upload/Rat-Expression.pdf>). При гібридизації кРНК з мікромасивом ДНК відбувається зв'язування комплементарних послідовностей кРНК і ДНК. Оскільки кРНК флуоресцентно мічена, то ділянки взаємодії можна легко ідентифікувати, а за інтенсивністю флуоресценції – визначити кількісну представленість конкретних нуклеїнових послідовностей у досліджуваному зразку.

Методика підготовки зразків РНК до експерименту створена виробниками мікромасивів та є високостандартизованою [4–6]. На відміну від неї комп'ютерна обробка результатів мікромасивів має багато можливостей у виборі інструментів і методів обробки, від чого суттєво залежить кінцевий результат і подальша інтерпретація даних [7–11]. Прог-

рамне забезпечення, що пропонується виробниками мікромасивів (наприклад GCOS у Affymetrix), досить дороге і, як правило, постачається з обладнанням для мікромасивів.

Існує також багато безкоштовних альтернатив, проте найпопулярнішим рішенням є статистичне середовище R і пакети функцій Bioconductor через наявність великої кількості алгоритмів для обробки даних та засобів візуалізації результатів (<http://www.lsw.uni-heidelberg.de/users/christlieb/teaching/UKStaSS10/R-refman.pdf>) [12].

Аналіз первинних даних є окремим завданням, яке найчастіше виконують для замовника спеціалізовані наукові підрозділи. Воно починається з моменту отримання зображення після сканування мікромасивів. Першим і принциповим етапом є контроль якості, в процесі якого сумарно визначають якість РНК, ефективність проведення зворотної транскрипції, транскрипції *in vitro* з утворенням кРНК і її гібридизації з мікромасивами, середню інтенсивність фоновому сигналу, а також артефакти самих мікромасивів. Наступним етапом аналізу результатів експерименту є попередня обробка і нормалізація даних, де коригується фоновий сигнал, числові значення інтенсивності зондів зводять до їхнього логарифму з основою 2, а також підсумовують величини інтенсивностей сигналу різних зондів у наборі зондів певного гена та отримують єдину величину – значення експресії (<http://bioconductor.org/packages/2.12/bioc/manuals/simpleaffy/man/simpleaffy.pdf>). Метою нормалізації є корекція систематичних варіацій, що можуть виникати внаслідок похибок піпетування, різної ефективності ферментативних реакцій при підготовці зразків або різної ефективності гібридизації з мікромасивами. Після проходження контролю якості даних та всіх етапів попередньої обробки і нормалізації вони є придатними для статистичного аналізу і визначення диференційно експресованих генів.

У цій роботі нами поетапно проаналізовано результати мікромасив-експерименту із впливу інтерферону альфа (ІФН α) на транскриптом первинних гепатоцитів щура, культивованих з цитокином упродовж 3 і 6 год, і визначено відповідність або невідповідність критеріїв оцінки результатів експерименту загальноприйнятим значенням. Для обра-

хунку результатів використано статистичне середовище R і пакети функцій Bioconductor та виявлено 28 і 124 диференційно експресованих генів відповідно після 3 і 6 год культивування з ІФН α . Наведені в представленій роботі R-коди можуть бути застосовані іншими дослідниками для аналізу власних експериментів або даних з доступних репозиторіїв (Gene Expression Omnibus, Array Expression) [13].

Матеріали і методи. Первинні гепатоцити щура культивували в середовищі Williams E протягом 3 і 6 год за присутності ІФН α (250 Од/мл) і без нього. По закінченні культивування з клітин виділяли тотальну РНК і гібридували з олігонуклеотидними мікромасивами щура (Affymetrix Rat Genome 2302). Експеримент проводили в трьох біологічних повторях. Результати експерименту отримано у вигляді 12 CEL-файлів (позначені як 1.CEL–12.CEL), з них шість файлів після 3 год інкубації з ІФН α і шість – після 6 год інкубації без нього. Усі процедури проведено за стандартними методиками (виділення клітин, тотальної РНК) згідно з рекомендаціями виробника Affymetrix (<http://www.ucdmc.ucdavis.edu/medmicro/pdfs/desktopgenechipexpressionmanual.pdf>). CEL-файли доступні у базі даних GEO під номером GSE43723.

Контроль якості і попередню обробку здійснювали в статистичному середовищі R з використанням пакетів функцій Bioconductor. Якість перевіряли за допомогою пакету simpleaffy та affyPLM (<http://bioconductor.org/packages/2.12/bioc/manuals/simpleaffy/man/simpleaffy.pdf>) [<http://www.bioconductor.org/packages/2.11/bioc/vignettes/affyPLM/inst/doc/QualityAssess.pdf>] і використовували функції «image», «RLE», «NUSE» у вигляді:

>pset=firPLM(data.raw) – де data.raw – це об'єкт AffyBatch (завантажені дані мікромасивів); >image(pset) – для створення псевдозображення мікромасивів; >RLE(pset) і >NUSE(pset) – для побудови діаграм розподілів відповідно відносного логарифму експресії та нормалізованої стандартної похибки.

Середній фоновий сигнал, якість кРНК, фактор нормалізації та ефективність гібридизації визначали із застосуванням функції qc пакету simpleaffy:

>qc(data.raw); >plot(qc(data.raw)) – для одержання графічного відображення результатів контролю якості.

Попередню обробку результатів і нормалізацію здійснювали, використовуючи функції з пакету *affy* (<http://www.bioconductor.org/packages/release/bioc/manuals/affy/man/affy.pdf>). Нормалізацію проводили у два етапи: на першому – між чотирма групами результатів – із культивування гепатоцитів протягом 3 год з ІФН α і без нього та відповідно впродовж 6 год; на другому – між усіма 12 мікромасивами. При цьому застосовано команди:

```
>data.bgcorrected=bg.correct.rma(data.raw) – для корекції фонового сигналу в усіх мікромасивах. Наступні чотири команди нормалізують окремо кожну з чотирьох експериментальних груп за методом «quantile» (чіпи 1, 3, 5 – ІФН $\alpha$ , 3 год культивування; 2, 4, 6 – 3 год культивування без ІФН $\alpha$ ; 7, 9, 11 – ІФН $\alpha$ , 6 год культивування; 8, 10, 12 – 6 год культивування без ІФН $\alpha$ ):
```

```
>data.normalized.135=normalize(data.bgcorrected[,c(1,3,5)], method=«quantiles»);
```

```
>data.normalized.246=normalize(data.bgcorrected[,c(2,4,6)], method=«quantiles»);
```

```
>data.normalized.7911=normalize(data.bgcorrected[,c(7,9,11)], method=«quantiles»);
```

```
>data.normalized.81012=normalize(data.bgcorrected[,c(8,10,12)], method=«quantiles»).
```

Після нормалізації дані мікромасивів кожної з експериментальних груп поетапно «зливали» в один об'єкт «data.normalized» за допомогою наступних трьох команд:

```
>data.normalized.7911_81012=merge.AffyBatch(data.normalized.7911, data.normalized.81012);
```

```
>data.normalized.135_246=merge.AffyBatch(data.normalized.135, data.normalized.246);
```

```
>data.normalized=merge.AffyBatch(data.normalized.135_246, data.normalized.7911_81012).
```

Фінальну нормалізацію здійснювали між усіма мікромасивам, застосовуючи функцію «*expresso*», де: *data.normalized* – дані, нормалізовані на попередніх етапах; *bgcorrect.method*=«*rma*» – параметр, що визначає метод, за яким коригували фоновий сигнал («*rma*»); *normalize.method*=«*quantile*» – визначає метод, за яким нормалізували дані; *pmcorrect.method*=«*pmonly*» – параметр, за допомогою якого обчислення виконують лише над зондами, повністю комплементарними цільовій кРНК (*perfect match*); *summary.method*=«*medianpolish*» – параметр, що ви-

значає метод, за яким підсумовують дані в наборі зондів до єдиного значення експресії генів.

```
>eset.medianpolish=expresso(data.normalized; bg.correct.method=«none»; normalize.method=«quantiles»; pmcorrect.method=«pmonly»; summary.method=«medianpolish»).
```

При застосуванні цієї команди створено об'єкт «*eset.medianpolish*», що містить нормалізовані, логарифмовані і сумарні значення експресії. Для порівняння даних до і після нормалізації використано команди `>boxplot(data.raw); >boxplot(exprs(eset.medianpolish))`, за якими виводять діаграми розподілу інтенсивностей сигналу. Значення експресії збережено в текстовий файл за допомогою функції `>write.exprs(eset.medianpolish, file=«data.mpolish.txt»)`, що створює текстовий файл «*data.mpolish.txt*» у робочій директорії.

Для статистичної обробки і визначення диференційно експресованих генів ми використали програму BRB Array Tools [14], яку встановлюють як додаток до Microsoft Excel з графічним інтерфейсом. Диференційну експресію виявляли за допомогою парного Т-тесту. Диференційно експресованими генами вважали такі, експресія яких змінювалася щонайменше в два рази відносно контрольної групи зі статистичною вірогідністю $p \leq 0,005$.

Результати і обговорення. Якість мікромасивів можна оцінювати як до, так і після нормалізації. До нормалізації ми перевірили, чи немає на мікромасивах артефактів, тобто плям, подряпин, а також чи рівномірно розподілені сигнали по мікромасивах. Із псевдозображень мікромасивів, отриманих за допомогою функції «*image*» пакету *affyPLM*, видно, що всі мікромасиви в даному експерименті мають рівномірний розподіл сигналів (рис. 1). На рис. 1 також наведено приклад псевдозображення мікромасиву (<http://plmimagegallery.bmbolstad.com/DifferentSites/index.html>), який на своїй поверхні має артефакти у вигляді плям.

Варіації між мікромасивами визначали, використовуючи відносний логарифм експресії (*Relative Log Expression, RLE*) і нормалізовану стандартну похибку (*Normalized Unscaled Standard Error, NUSE*) з функцій пакету *affyPLM*. *NUSE* розраховували для кожного гена на всіх мікромасивах і стандартизували так, щоб медіана стандартної по-

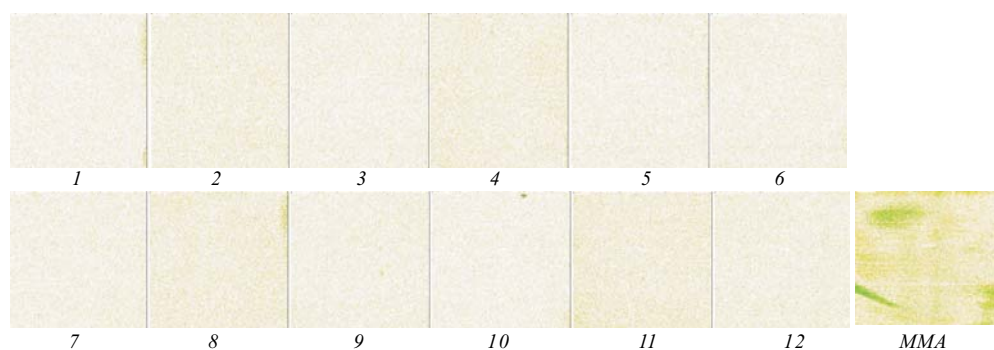


Рис. 1. Псевдозображення проаналізованих мікрмасивів 1–12; MMA – мікрмасив з артефактами у вигляді зон з підвищеною інтенсивністю сигналів (плям) у різних частинах мікрмасиву

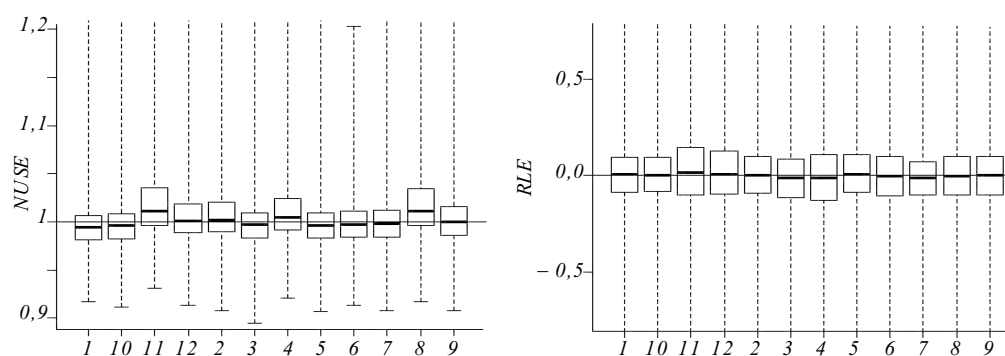


Рис. 2. Діаграми розподілу нормалізованої стандартної похибки NUSE і відносного логарифму експресії RLE; 1–12 – номери мікрмасивів

хибки для кожного гена дорівнювала одиниці. RLE обчислювали також для всіх генів, порівнюючи величини експресії кожного з них на кожному мікрмасиві з медіаною величини експресії для цього гена на всіх мікрмасивах.

Оскільки ми припустили, що більшість генів експресуються на сталому рівні, то значення NUSE і RLE мають наближатися до одиниці і нуля відповідно, а відхилення за даними параметрами в якомусь конкретному мікрмасиві може свідчити про нижчу якість даних, отриманих з нього (рис. 2) (<http://www.bioconductor.org/packages/2.11/bioc/vignettes/affyPLM/inst/doc/QualityAssess.pdf>). У нашому експерименті мікрмасиви 11.CEL і 8.CEL мають трохи збільшені значення NUSE, що може бути пов'язане з вищою відносно решти мікрмасивів загальною інтенсивністю сигналів (рис. 1). Проте такі відхилення за даними показниками не є суттєвими через те, що після нормалізації параметри якості лежать у межах норми.

Пакет *simpleaffy* дозволяє оцінити якість РНК і результати гібридизації за чотирма критеріями: середнім значенням фону; фактором нормалізації; кількістю зондів на мікрмасиві, з якими гібридується кРНК; співвідношенням інтенсивностей сигналів, одержаних внаслідок взаємодії 3'-ділянок

ДНК гліцеральдегідфосфатдегідрогенази (ГАФДГ) і β -актину, до інтенсивностей сигналів у зоні 5'-ділянок цих генів (<http://bioconductor.org/packages/2.12/bioc/manuals/simpleaffy/man/simpleaffy.pdf>).

Значення фонові інтенсивності виражають у безрозмірних величинах і воно зазвичай коливається в межах від 20 до 100. Як видно з рис. 3, дані якого згенеровано функцією *qc* пакету *simpleaffy*, значення інтенсивності фону для всіх експериментів знаходиться в діапазоні 53–66 (рис. 3), що демонструє малий розкид за даним показником.

Щоб нормалізувати інтенсивності сигналів між різними чіпами, частіше розраховують середнє значення середніх інтенсивностей сигналів усіх чіпів, які були в експерименті, і до цього значення зводять середні величини інтенсивності кожного чіпа через спеціально обчислений коефіцієнт – фактор нормалізації. *Affymetrix* рекомендує, щоб значення фактора нормалізації не перевищувало трьох (рис. 3). Результат нормалізації можна також отримати за різницею розподілів інтенсивностей сигналів між мікрмасивами (рис. 4) <http://bioconductor.org/packages/2.12/bioc/manuals/simpleaffy/man/simpleaffy.pdf>.

За величиною співвідношення інтенсивності сигналів з повністю комплементарних (*perfect match*) зондів і контрольних зондів з одним некомплементарним

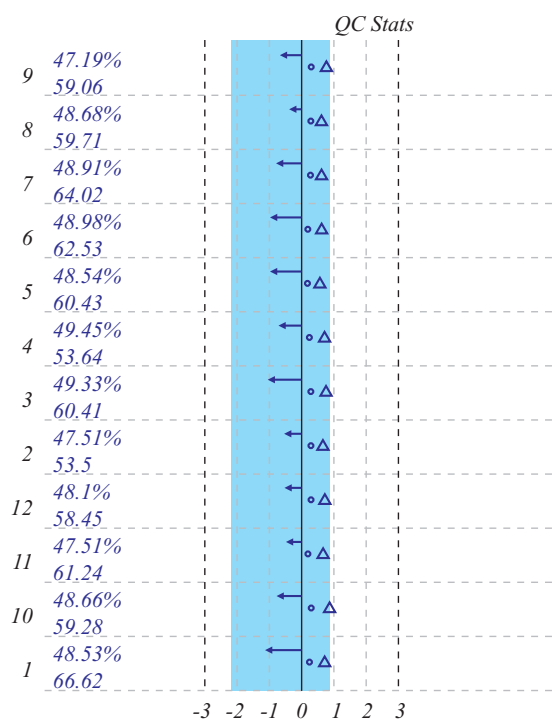


Рис. 3. Діаграма параметрів контролю якості: 0 – співвідношення інтенсивностей сигналу 3'-ділянки до 5'-ділянки гена гліцеральдегідфосфатдегідрогенази (не має перевищувати одиниці); Δ – співвідношення інтенсивностей сигналу 3'-ділянки до 5'-ділянки гена β-актину (не повинне перевищувати трьох); ← – величина фактора нормалізації (має лежати в межах від -3 до 3); цифрові значення показано у стовпчику зліва (кількість «присутніх генів», %), під ними – середнє значення інтенсивності фону. 1–12 – номери мікромасивів

перименті розкид за цим показником був досить малим і коливався від 47,1 до 49,45 % (рис. 3).

Окремим елементом оцінки ефективності гібридизації є визначення співвідношення інтенсивностей сигналів для контрольних зразків внесеної у гібридизаційну суміш кРНК, які походять з локусів BioB, BioC, BioD і Cre *Escherichia coli*. РНК, зчитану з цих локусів, додають до гібридизаційну суміші у зростаючій концентрації: РНК BioB → РНК Cre, а саме – від 1,5 пмоль РНК BioB до 100 пмоль РНК Cre. Гібридизаційний сигнал РНК BioB з відповідним зондом має бути «присутнім» у кожному зразку. У такому разі чутливість методу є оптимальною. У результаті нашого експерименту цей сигнал виявлено на всіх чіпах.

Ще одне співвідношення використовують для оцінки повноти полімеризації при синтезі кДНК і інтактності молекул кРНК, тобто співвідношення інтенсивності сигналів від гібридизації кРНК з 3'- і 5'-ділянками ДНК, які кодують ГАФДГ і β-актин. ГАФДГ має коротший транскрипт, ніж β-актин, тому максимально допустиме співвідношення для нього дорівнює одиниці, а для β-актину – трьома (<http://bioconductor.org/packages/2.12/bioc/manuals/simpleaffy/man/simpleaffy.pdf>).

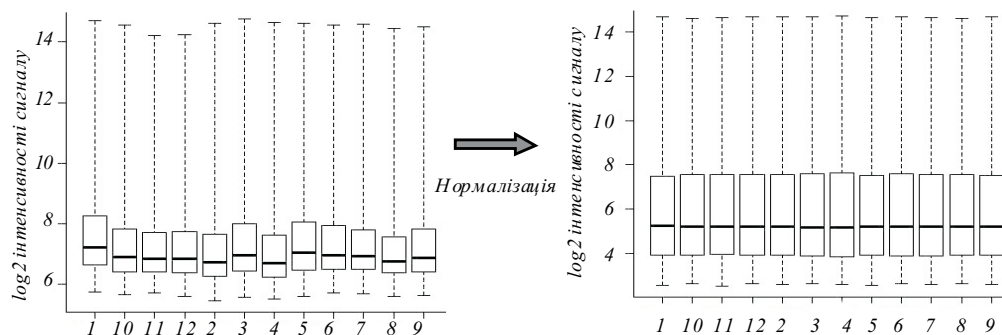


Рис. 4. Розподіл інтенсивностей сигналу до і після нормалізації; 1–12 – номери мікромасивів

тарним нуклеотидом (mismatch) обирають одну з трьох категорій щодо наявності певної кРНК у досліджуваному препараті – «присутня» (P – present), «відсутня» (A – absent) або «присутність достеменно не визначено» (M – marginal). Кількість «присутніх» кРНК в зразку виражають у відсотках по відношенню до загальної кількості зондів і вона не повинна сильно варіювати між різними мікромасивами (<http://bioconductor.org/packages/2.12/bioc/manuals/simpleaffy/man/simpleaffy.pdf>) [15]. У нашому ек-

Усі значення співвідношення інтенсивностей сигналів від 3'- до 5'-ділянок генів ГАФДГ і β-актину не перевищили одиниці, що свідчить про хорошу якість РНК та високу ефективність транскрипції *in vitro* для всіх зразків (рис. 3). Отже, отримані результати засвідчують якість кРНК і проведеної реакції гібридизації. Попередня обробка і нормалізація даних відповідають загальноприйнятим критеріям. Ці етапи передують статистичному аналізу і визначенню диференційно експресованих генів.

Після застосування парного Т-тесту знайдено 32 диференційно експресованих гени після 3 год і 147 генів – після 6 год культивування гепатоцитів з ІФН α , що відповідає заданим критеріям ($p < 0,005$, зміна експресії більш ніж у два рази). Оскільки на мікромасивах Affymetrix одному гену можуть відповідати декілька зондів, то серед вищезазначених генів виявлено відповідно 28 та 124 неповторюваних (унікальних) гени. Аналіз відгуку первинних гепатоцитів на дію ІФН α є предметом окремого дослідження.

A. V. Kuklin, B. T. Tokovenko, M. Yu. Obolenska

Evaluation criteria of rat hepatocytes transcriptome analysis under the influence of interferon alpha by DNA microarray

Institute of Molecular Biology and Genetics, NAS of Ukraine
150, Akademika Zabolotnoho Str., Kyiv, Ukraine, 03680

Summary

The changes induced in transcriptome of rat hepatocytes treated with interferon alpha (IFN α) during three and six hours were analyzed by DNA microarray. **Aim.** To conduct a stepwise analysis of the results of microarray experiment and to determine whether they meet/fail to the conventional requirements. **Methods.** The files obtained after scanning microarrays were subjected to the analysis in statistical environment R by Bioconductor's packages «affy», «simpleaffy», «affyPLM» and BRB Array Tools software for paired T-test. **Results.** All microarrays had quality metrics lying within recommended ranges, passed quality control, were normalized and are comparable with each other. The T-test revealed 28 and 124 differentially expressed genes after three and six hours of cells cultivation with IFN α , respectively. **Conclusions.** The obtained data meet the conventional criteria of quality and are applicable for further evaluation of their biological significance. The R-codes used in this study can be used for the analysis of the microarrays data.

Keywords: DNA microarrays, interferon alpha, statistical environment R, Bioconductor, normalization.

A. В. Куклин, Б. Т. Токовенко, М. Ю. Оболенская

Критерии оценки результатов микромассив-эксперимента при исследовании транскриптома гепатоцитов крысы под влиянием интерферона альфа

Резюме

Транскриптом первичных гепатоцитов крысы, культивированных с интерферона альфа (ИФН α) в течение трех и шести часов, исследовали с привлечением олигонуклеотидных микромассивов. **Цель.** Поэтапный анализ результатов микромассив-эксперимента и определение соответствия или несоответствия критериев оценки результатов общепринятым значениям. **Методы.** Анализ осуществляли на файлах, полученных после сканирования микромассивов, с использованием статистической среды R, пакетов функций Bioconductor «affy», «simpleaf-fy», «affyPLM» и программы BRB Array Tools с применением парного Т-теста. **Результаты.**

ты. Все микромассивы прошли контроль качества, нормализованы и сравнимы между собой. По результатам Т-теста найдены 28 и 124 дифференциально экспрессированных гена соответственно после трех и шести часов культивирования клеток с ИФН α . **Выводы.** Полученные данные отвечают общепринятым критериям качества и пригодны для дальнейшего определения их биологического значения. Используемые в работе R-коды можно применять для обработки результатов микромассив-экспериментов.

Ключевые слова: микромассивы ДНК, интерферон альфа, статистическая среда R, Bioconductor, нормализация.

REFERENCES

- Heller M. J. DNA microarray technology: devices, systems, and applications // Annu. Rev. Biomed. Eng.–2002.–4.–P. 129–153.
- Ideker T., Galitski T., Hood L. A new approach to decoding life: systems biology // Annu. Rev. Genomics Hum. Genet.–2001.–2.–P. 343–372.
- Li X., Quigg R. J., Zhou J., Gu W., Nagesh Rao P., Reed E. F. Clinical utility of microarrays: current status, existing challenges and future outlook // Curr. Genomics.–2008.–9, N 7.–P. 466–474.
- GeneChip^(R) Expression Analysis Technical Manual.–Santa Clara: Affymetrix Inc., 2004.–108 p..
- Slonim D. K., Yanai I. Getting started in gene expression microarray analysis // PLoS Comput. Biol.–2009.–5, N 10.–e1000543.
- Forster T., Roy D., Ghazal P. Experiments using microarray technology: limitations and standard operating procedures // J. Endocrinol.–2003.–178, N 2.–P. 195–204.
- Selvaraj S., Natarajan J. Microarray data analysis and mining tools // Bioinformatics.–2011.–6, N 3.–P. 95–99.
- Irizarry R. A., Bolstad B. M., Collin F., Cope L. M., Hobbs B., Speed T. P. Summaries of Affymetrix GeneChip probe level data // Nucleic Acids Res.–2003.–31, N 4.–e15.
- The analysis of gene expression data: Methods and Software / Eds G. Parmigiani, E. S. Garrett, R. A. Irizarry, S. L. Zeger.–Amsterdam: Springer, 2003.–456 p.
- Amaratunga D., Cabrera J. Exploration and analysis of DNA microarray and protein array data.–New York: Willey-Interscience, 2004.–246 p.
- Quackenbush J. Microarray data normalization and transformation // Nat. Genet.–2002.–32, Suppl.–P. 496–501.
- Gentleman R. C., Carey V. J., Bates D. M., Bolstad B., Dettling M., Dudoit S., Ellis B., Gautier L., Ge Y., Gentry J., Hornik K., Hothorn T., Huber W., Iacus S., Irizarry R., Leisch F., Li C., Maechler M., Rossini A. J., Sawitzki G., Smith C., Smyth G., Tierney L., Yang J. Y., Zhang J. Bioconductor: open software development for computational biology and bioinformatics // Genome Biol.–2004.–5, N 10.–R80.
- Anderle P., Duval M., Draghici S., Kuklin A., Littlejohn T. G., Medrano J. F., Vilanova D., Roberts M. A. Gene expression databases and data mining // Biotechniques.–2003.–Suppl.–P. 36–44.
- Simon R., Lam A., Li M. C., Ngan M., Menezes S., Zhao Y. Analysis of gene expression data using BRB-Array Tools // Cancer Inform.–2007.–3.–P. 11–17.
- Wang Y., Miao Z. H., Pommier Y., Kawasaki E. S., Player A. Characterization of mismatch and high-signal intensity probes associated with Affymetrix genechips // Bioinformatics.–2007.–23, N 16.–P. 2088–2095.

Received 10.02.13