Математическая биология и биоинформатика. 2013. Т. 8. № 2. С. 520–528.
URL: http://www.matbio.org/2013/Sibatov_8_520.pdf

========== **БИОИНФОРМАТИКА** ==========

УДК: 519.25:577.21:576.385.5:616-006.6

Статистический анализ радиационно-индуцированной динамики транскриптома раковых клеток по данным ДНК-микроматриц на примере линии HCT116

Сибатов Р.Т.[1], Саенко Ю.В.,[1] Учайкин В.В.[1], Саенко В.В.[1], Морозова Е.В.[1], Шулежко В.В.[1], Кожемякина Е.В.[1], Бызыкчи А.Н.[1], Гусаров Г.Г.[1], Коробко Д.А.[1], Яровикова И.В.[1], Салтыкова К.В.[1], Кожемякин И.И.[2], Журавлев В.М.[1], Журавлев А.В.[1], Айнуллова Н.К.[1]

[1]*Ульяновский государственный университет, Ульяновск, Россия, 432017, Ульяновск, ул. Л.Толстого 42*
[2]*Московский государственный университет имени М. В. Ломоносова, ВМК, 119991, Москва, ГСП-1, Ленинские горы, МГУ, д. 1, стр. 52, 2-й учебный корпус, ВМК*

Аннотация. В работе проведён анализ данных ДНК-микрочипов по радиационно-индуцированной динамике транскриптома раковых клеток линии HCT116 с нормальным и мутантным геном TP53. Транскриптом анализировался через 1, 12 и 24 часа после облучения с использованием микроматрицы Affymetrix серии HGU133A. Получено, что вероятностные характеристики разностей экспрессии существенно зависят от интенсивностей базового уровня, причем для абсолютных разностей эта зависимость нелинейная. Эффект учтён в рамках алгоритма «огибающая шума» на этапах фильтрации, кластеризации и группировки генов. Судить об эффективности применяемых в работе процедур обработки данных ДНК-микрочипов можно по результатам иерархической кластеризации и метода групповых средних. Для генов, прошедших процедуру фильтрации, построены дендрограммы, на основе которых проведено предварительное сравнение динамики ключевых сигнальных путей, связанных с программируемой клеточной смертью и репарацией ДНК.

Ключевые слова: *ДНК-микроматрица, экспрессия генов, иерархическая кластеризация, метод групповых средних.*

ВВЕДЕНИЕ

В работе анализируются данные по радиационно-индуцированной динамике транскриптома раковых клеток на примере линии HCT116. Данные получены с помощью технологии ДНК-микроматриц [1–8]. Как известно, ДНК-микрочип представляет собой нанесённые на подложку в определенном порядке фрагменты одноцепочечных ДНК. Эти синтетические молекулы выступают в роли зондов. Меченные флуоресцентным красителем комплементарные цепи ДНК из исследуемого образца гибридизуются с зондами. Важнейшим шагом в анализе данных ДНК-микрочипов является расчёт уровня экспрессии генов по интенсивности свечения проб на чипе. Как известно, на интенсивность свечения проб влияет множество факторов. Избавление от эффектов технической вариации (связанной с выделением образцов, окрашиванием и гибридизацией) и ошибок измерения подразумевается на этапе предобработки данных ДНК-микроматриц. Предобработка подразумевает три последовательных стадии: фоновую поправку, нормализацию и суммаризацию.

Методы нормализации могут использовать чип-эталон (линейные и нелинейные методы) или всю выборку чипов (циклическая локальная регрессия, метод контрастов, квантильная нормализация и др.) [6–8].

Большой объём информации, получаемый в результате экспериментов с использованием технологии ДНК-чипов, предъявляет серьёзные требования к алгоритмам анализа получаемой информации. Правильный статистический анализ имеет важное значение для успешной интерпретации данных. Необходимо учитывать особенности, такие, например, как высокая плотность ячеек, в которых протекает полимеразная реакция с одновременным синтезом ДНК тысяч генов. Экспрессия только небольшого количества генов будет статистически значимо различаться. До сих пор в области анализа данных, получаемых с помощью микроматриц, существуют многие сложные и нерешенные проблемы. К важным относятся проблемы, связанные с экспериментальным и биологическим шумом, отсутствием стандартизации в статистической оценке ложноположительных и ложноотрицательных результатов и т.п.

В работе использовались данные экспериментов по изучению экспрессии генов в клетках HCT116, облученных в дозе 4 Гр. Транскриптом анализировался через 1, 12 и 24 часа после облучения с использованием микроматрицы Affymetrix серии HGU133A. Обсуждается вопрос о выборе координат профилей экспрессии генов для процедуры кластеризации. Получено, что вероятностные характеристики разностей экспрессии существенно зависят от интенсивностей базового уровня. Процедура фильтрации, выбор метрики подобия, координат профилей экспрессии и метода кластеризации должны учитывать эту зависимость. Метод огибающей шума будет применён на этапах фильтрации, кластеризации и группировки генов. Эффективность метода оценена по результатам кластеризации и метода групповых средних.

МАТЕРИАЛ И МЕТОДЫ ИССЛЕДОВАНИЯ

Данные получены для клеток HCT116 p53+/+ и HCT116 p53–/– через 1, 12 и 24 часа после облучения дозой 4 Гр с использованием микроматрицы Affymetrix (Санта-Клара, Калифорния, США) серии HGU133A (Human Genome U133A), содержащей ампликоны к 22216 генам. РНК выделяли из $3x10^6$ клеток с использованием набора для выделения РНК (RNeasy Mini, Qiagen, США) в соответствии с инструкцией производителя. Если качество биотинилированных кРНК соответствовало расчётному, то тогда проводили гибридизацию с матрицей HGU133A. Матрицу окрашивали стрептовидин-фикоэритрином. Окрашенную матрицу отмывали от несвязавшегося белка и сканировали на сканере Gene Array G2500A (Agilent,Santa Clara, CA, USA).

CEL-файлы данных свечения проб обрабатывались с помощью трех процедур (RMA, GCRMA и MAS5), реализованных на языке R среды Bioconductor [9] в пакете affy [10]. При анализе применялись также пакеты annaffy [11], limma [12], genefilter и annotate [13].

В данной работе применялись бесплатные программы Cluster 3.0 (http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster) для кластеризации данных по экспрессии генов и TreeView (http://rana.lbl.gov/downloads/TreeView/) для визуализации с помощью дендрограмм. Подробное описание CEL-файлов можно найти в [14].

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ И ИХ ОБСУЖДЕНИЕ

После предобработки одним из перечисленных алгоритмов (RMA, GCRMA, MAS5) данные представляются в виде таблицы. В качестве меры δ изменения экспрессии наиболее популярны отношение уровней экспрессии или логарифм этого отношения. Полученные данные должны быть подвержены процедуре фильтрации, направленной на выделение генов, демонстрирующих значимые изменения экспрессии. Такие гены называют *дифференциально экспрессирующими*. Если изменение экспрессии

оценивается, например, путём нахождения логарифма отношения интенсивности к интенсивности контрольного (базового) уровня, то можно считать, что положительные значения соответствуют повышению экспрессии, а отрицательные – её подавлению.

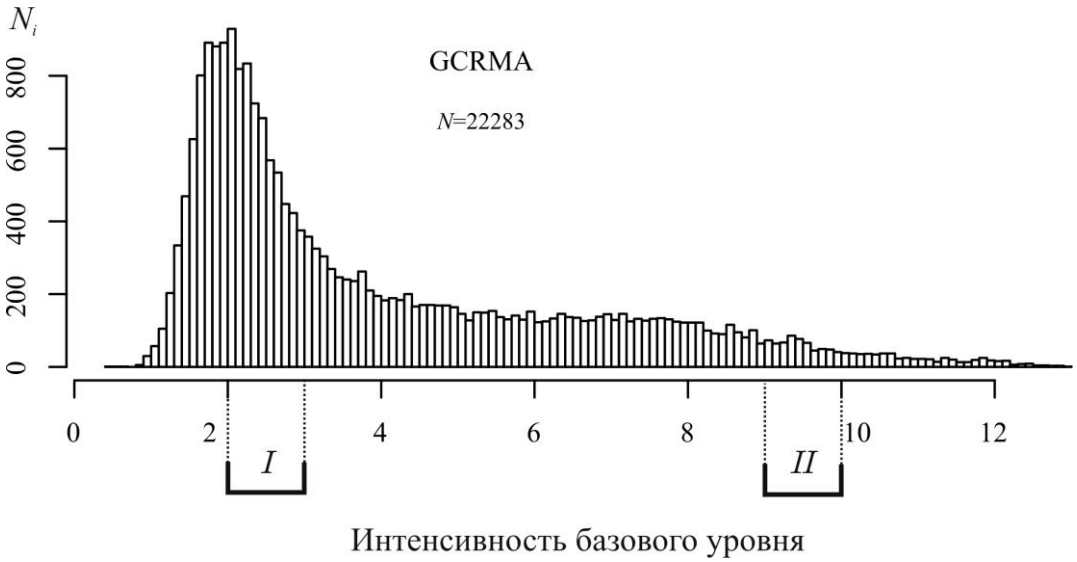

Рис. 1. Гистограмма контрольного уровня интенсивности приведена для случая данных HCT116 p53+/+, предобработанных алгоритмом GCRMA. N_i – число ячеек с интенсивностью, попадающей в нужный интервал разбиения

Для выделения генов, демонстрирующих существенное изменение экспрессии, обычно используют разделяющие уровни с граничными значениями $\delta_{гр} = \pm 1$. Процедура фильтрации может быть распространена и на временные промежутки. Это обобщение подразумевает выделение генов, уровень экспрессии которых превосходил по величине критический уровень в течение нескольких временных участков. Однако такая фильтрация является грубой, поскольку не учитывает зависимость величины шума от уровня интенсивности. На рис. 1 приведена гистограмма базовых интенсивностей для клеточной линии HCT116 p53+/+. Распределение интенсивностей контрольных уровней не является равномерным. Кроме этого, параметры распределений (да и сам их вид) изменений экспрессии для разных интервалов («окон») базовой интенсивности существенно различаются.

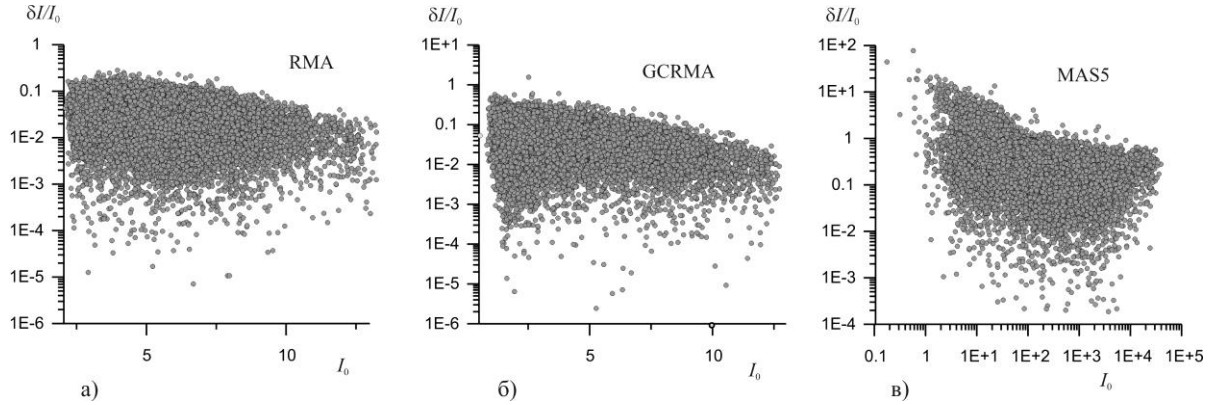

Рис. 2. Диаграмма рассеяния в координатах «базовая интенсивность – относительное изменение» для данных, полученных различными процедурами предобработки ДНК-чипов (RMA, GCRMA, MAS5). I_0 в относительных единицах.

Диаграммы рассеяния на рис. 2 демонстрируют уменьшение разброса относительных изменений с ростом интенсивности базового уровня. Это указывает на

возможную несостоятельность оценки дифференциальной экспрессии генов на основе относительного изменения интенсивности: $\delta = (I - I_0)/I_0$. Аналогичные результаты получены и для некоторых других мер, например, $\delta = \log(I/I_0)$.Среднеквадратичный разброс абсолютной разности растёт с увеличением базовой интенсивности, но рост этот оказывается сублинейным, т.е. медленнее, чем линейная зависимость от I_0. Это приводит к тому, что вероятностные характеристики относительной разности $\delta = (I - I_0)/I_0$ будут зависеть от I_0. Чтобы убедиться в этом, мы разбили интервал значений базовой интенсивности на отрезки («окна»), и для каждого отрезка рассчитали математическое ожидание m_δ, среднеквадратическое отклонение s_δ и эмпирическую плотность распределения (гистограмму) $p(\delta)$ относительных изменений экспрессии. Результаты представлены на рис. 3. Значения m_δ и s_δ зависят от интенсивности контрольного уровня I_0.

Как оказывается, не только первые два момента различаются: гистограммы относительных изменений, построенные для интервалов базовой интенсивности $[1, 2)$, $[5, 6)$, $[10, 11)$, указывают на различие и самих распределений δ. Отметим, что аналогичные замечания справедливы и для уровня экспрессии, вычисляемого как логарифм отношения I к I_0.

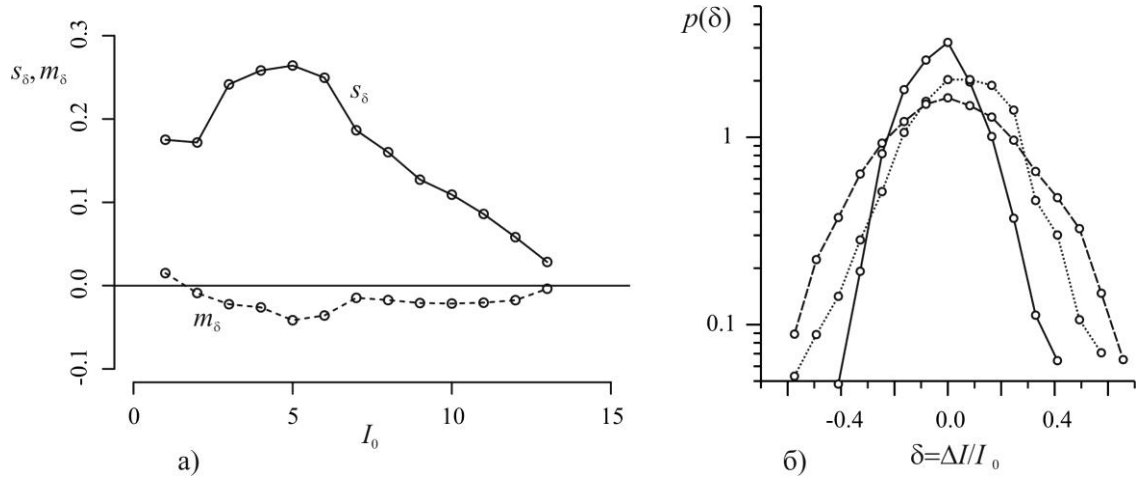

Рис. 3. Математическое ожидание, среднеквадратическое отклонение (а) и эмпирическая плотность распределения (б) относительных изменений экспрессии в заданных «окнах» контрольной интенсивности.

Учёт указанного эффекта должен быть произведён на этапах фильтрации, кластеризации и группировки генов. В данной работе применен алгоритм, основанный на расчёте огибающей шума. Сначала вся совокупность генов подвергалась процедуре фильтрации, направленной на выявление статистически значимых и существенных изменений в экспрессии генов. На рис. 4,а представлена диаграмма «вулкан», представляющая собой диаграмму рассеяния в координатах $-\log_{10} p$ и d, где под d подразумевается величина $\left(I/I_0 - 1 - m_\delta(I_0)\right)/s_\delta(I_0)$. Характеристики $m_\delta(I_0)$ и $s_\delta(I_0)$ рассчитывались следующим образом.

Отрезок, содержащий значения интенсивностей базового уровня I_0, делится на несколько интервалов. Например, для предобработанных методом GCRMA данных отрезок значений контрольного уровня [0, 15] мы разбивали на 15 интервалов. Для каждого интервала рассчитываются характеристики флуктуаций: среднеквадратическое отклонение s_i изменения экспрессии и среднее значение m_i. Далее дискретные зависимости $s_i(I_i)$ и $m_i(I_i)$ аппроксимировались аналитическими функциями $m_\delta(I_0)$ и

$s_\delta(I_0)$ с использованием программы TableCurve 2D. Функции $m_\delta(I_0) \pm s_\delta(I_0)$ представляют собой «огибающие шума», которые использовались вместо уровней ± 1 для отбора генов, демонстрирующих существенное изменение экспрессии.

На рис. 4,а представлена диаграмма «вулкан», визуализирующая процедуру отбора статистически значимых изменений. Дело в том, что на ДНК-чипе несколько ячеек соответствуют одному и тому же гену. С помощью данных этого набора ячеек можно оценить *p*-значение. Кроме статистической значимости подразумевается, что отфильтрованные гены претерпевают существенное изменение экспрессии. На рис. 4,б представлена диаграмма рассеяния, демонстрирующая процедуру выбора генов с существенными изменениями экспрессии. Суперпозиция указанных процедур и приводит к искомому списку дифференциально экспрессирующих генов.

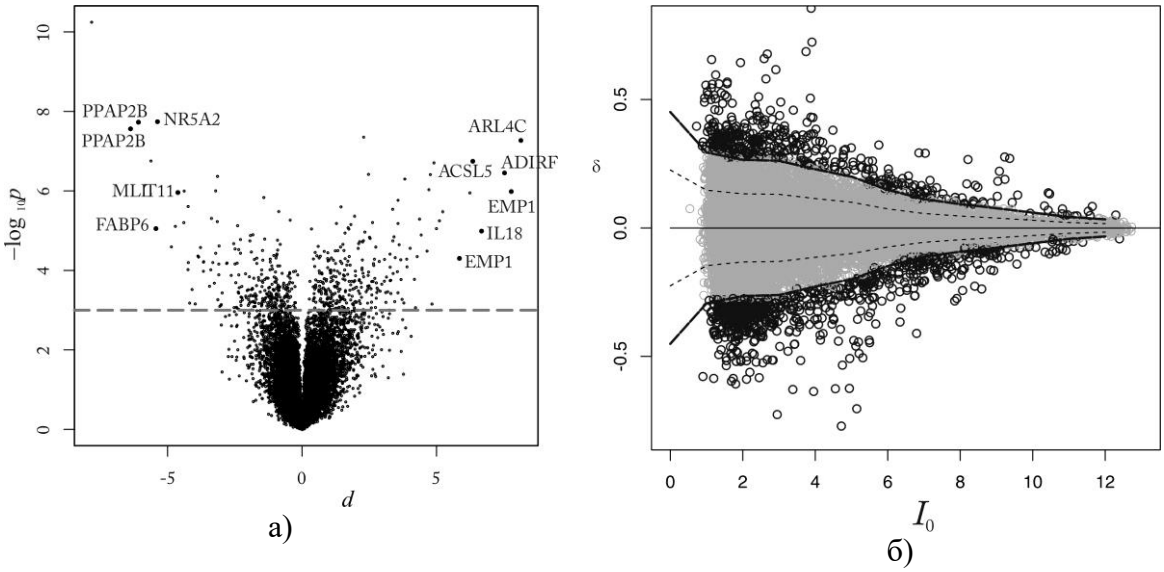

Рис. 4. а) Диаграмма «вулкан», демонстрирующая отбор статистически значимых изменений экспрессии. Подписаны некоторые гены с низкими *p*-значениями. б) Диаграмма рассеяния, демонстрирующая процедуру выбора генов с существенными изменениями экспрессии.

Далее проводилась кластеризация отфильтрованных генов. Кластеризация направлена на группировку генов на основе сходства профилей экспрессии. Цель этой группировки – выявление функционально связанных генов. При этом считается, что функция и соответствующий набор генов заранее неизвестны.

Среди метрик подобия (сходства) [15] чаще применяются коэффициент корреляции Пирсона, абсолютное значение этого коэффициента, нецентрированная корреляция (косинус-расстояние), модуль нецентрированной корреляции (косинус наименьшего угла между векторами), коэффициент ранговой корреляции Спирмена, коэффициент Кендалла τ, евклидово расстояние, манхэттенское расстояние.

Для каждой пары элементов можно вычислить расстояние (сходство) между ними. Если есть N элементов, предназначенных для кластеризации, матрица расстояний будет иметь размерность $N \times N$. На основе этой матрицы может быть построена дендрограмма.

На рисунке 5 представлена данные сравнительного кластерного анализа динамики транскриптома клеточных линий HCT-116p53+/+ (с геном TP53 дикого типа) и HCT-116p53–/– (с мутированным геном TP53) после использования процедуры фильтрации с использованием алгоритма огибающей шума. В качестве метрики подобия выбрана нецентрированная корреляция. Этот выбор связан с тем, что построенная дендрограмма позволяет выделить гены со схожей динамикой экспрессии или различным поведением в клетках с мутированным и нормальным геном TP53. В результате использования алгоритма выявлено 3 кластера генов с дифференциальной экспрессией. 1-й кластер

представлен 2 генами – RAB2A и PHF14 (группа 1). 2-й кластер генами – NME5, SURF2, RRP15, TSFM, RNASEH2B, COPS8, DPH5, ALMS1, FBXO11, MDM1, C5orf54, CCNE2, BRIP1, QRSL1 и ERVMER34-1 (группа 2, рис. 5). 3-й кластер объединяет гены SREBF1, LTBP4, ZFAND5, GREB1, CENTP и GNAT3 (группа 3, рис. 5). Наиболее интересным с биологической точки зрения является 3 кластер. Динамика экспрессии генов этой группы отличается больше всего среди 3-х найденных кластеров. В клетках линии HCT-116p53–/– экспрессия этих генов снижена во всех временных точках, тогда как в клетках линии HCT-116p53+/+ она повышена по сравнению с контрольной группой. На роль генов участвующих в развитии радиорезистентности больше подходят гены ZFAND5 и GREB1. Ген GREB1 кодирует эстроген-зависимый регулятор роста рака груди [16]. Ген ZFAND5, кодирующий "цинковый палец", тип домена AN1, может ингибировать активацию NF-kappa-B через сверхэкспрессию генов RIPK1 и TRAF6. Он также ингибирует экспрессию фактора некроза опухолей TNF, IL-1 и TLR4-индуцированную активацию NF-kappa-B на доз-зависимый манер. Сверхэкспрессия ZFAND5 делает клетки чувствительные к TNF-индуцированному апоптозу. Ген ZFAND5 может быть вовлечен в регуляцию активации NF-kappa-B и апоптоза [17]. Сверх экспрессия генов ZFAND5 и GREB1 делает клетки чувствительные к TNF-индуцированному апоптозу, который опосредуется каспазой 8 и 3, тогда как низкая экспрессия этих генов препятствует индукции программированной клеточной смерти. Из литературных данных известно, что TP53 мутантные раковые опухоли обладают высокой радиорезистентностью и устойчивостью к индукции апоптоза. Причины этого феномена до конца не ясны [18].

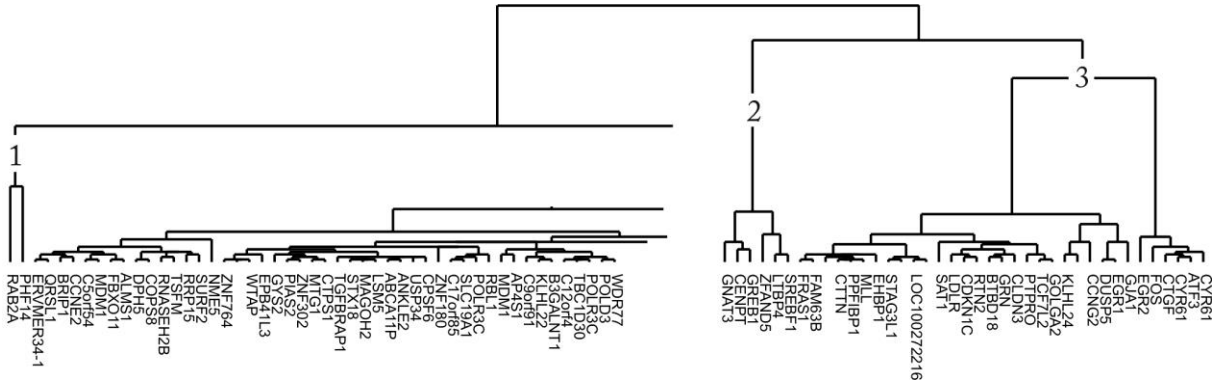

Рис. 5. Часть дендрограммы, построенной на основании профилей экспрессии генов, прошедших процедуру фильтрации. В качестве метрики подобия выбрана нецентрированная корреляция, кластеризация произведена методом средней связи. Учтены профили экспрессии генов через 1ч, 12 ч, 24 ч после облучения в клетках HCT116 p53+/+ и p53-/-.

Отфильтрованные гены могут быть сгруппированы по следующим признакам: молекулярная функция, биологический процесс, белковый класс, сигнальный путь и др. Метод групповых средних подразумевает расчёт динамики среднего значения для групп генов, демонстрирующих статистически значимые изменения экспрессии. Отметим, что группировка до процедуры фильтрации привела бы к существенному вкладу шума, искажающему рассматриваемую динамику.

На рис. 6 показана динамика средних для групп «сцепление клеток», «активация макрофагов», «индукция апоптоза», «репродукция». Ось ординат соответствует эффективной экспрессии $\varepsilon = (1/n)\sum_{i=1}^{n}\left(I_i / I_{0i} - 1 - m_\delta(I_{0i})\right)/s_\delta(I_{0_i})$, рассчитанной с учётом зависимости относительного изменения экспрессии от базового уровня, ось абсцисс – времени в часах. Здесь n – число генов в группе. Значения I_i и I_{0i} относятся к i-му гену группы. Прерывистые линии рассчитаны для клеточной линии с нормальным геном TP53, сплошные – с мутантным. С помощью метода групповых

средних можно судить об адекватности процедур предобработки и фильтрации по реплицированным измерениям. Достаточно близкое поведение эффективной экспрессии некоторых групп свидетельствует об удовлетворительности указанных процедур. Подробный анализ различий в поведении групповых средних для клеток с нормальным и мутантным геном TP53 планируется в отдельной работе.

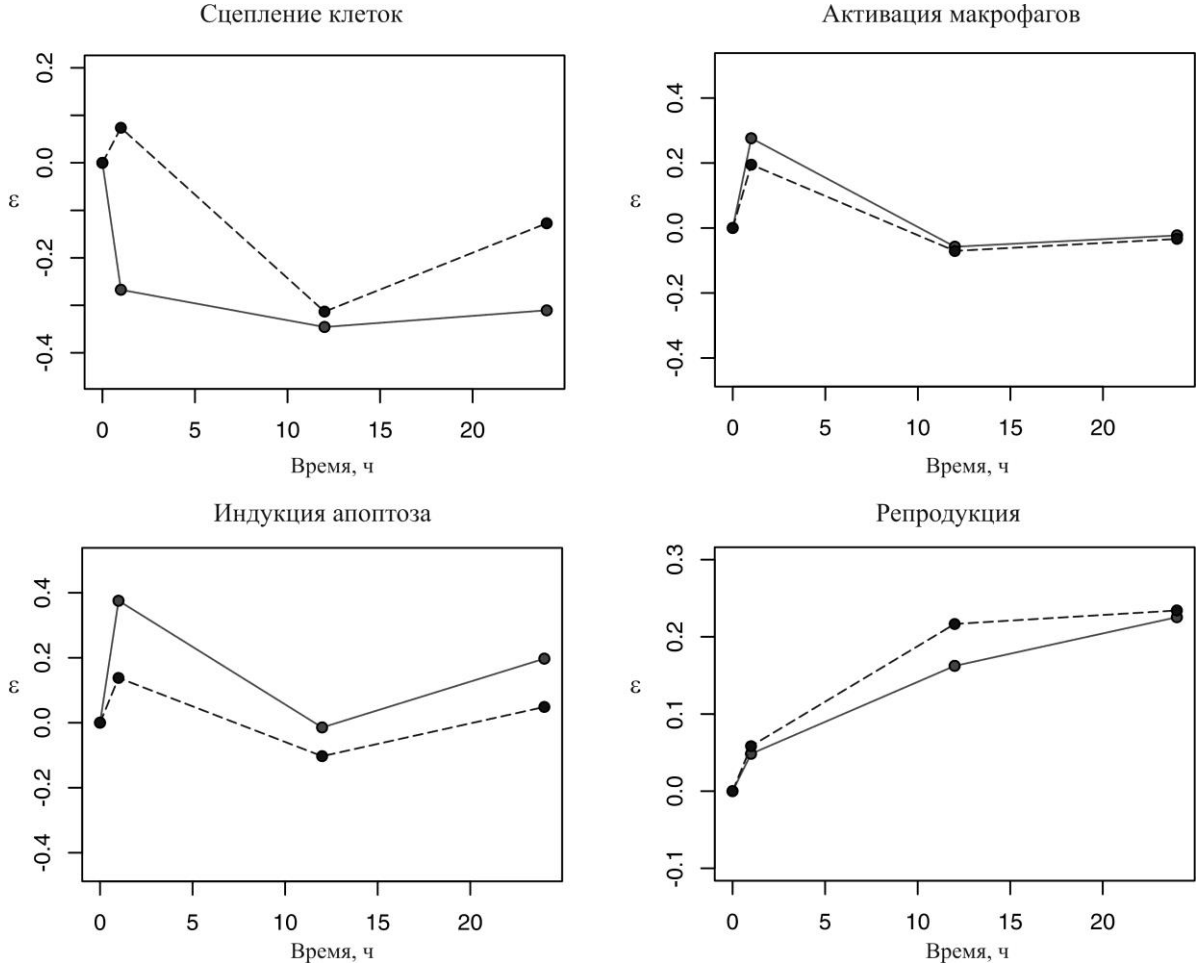

Рис. 6. Динамика групповых средних, рассчитанных на основе дифференциально экспрессирующих генов клеточной линии HCT116 с нормальным (пунктир) и мутантным (сплошные линии) геном TP53.

ЗАКЛЮЧЕНИЕ

В работе проанализированы данные ДНК-микрочипов по радиационно-индуцированной динамике транскриптома раковых клеток линии HCT116 с нормальным и мутантным геном TP53. Получено, что вероятностные характеристики разностей экспрессии существенно зависят от интенсивностей базового уровня, причем для абсолютных разностей эта зависимость нелинейная. В некоторых работах [5, 19] полагается, что средние «флуктуации» экспрессии (в которых скрыты реальные изменения) пропорциональны среднему сигналу ячейки. Это автоматически подразумевается при выборе в качестве оценки изменения экспрессии относительного сдвига $\delta = (I - I_0)/I_0$. В нашей работе показано, что эта зависимость существенно нелинейна. Среднеквадратические отклонения флуктуаций экспрессии убывают с ростом интенсивности базового уровня (и средней интенсивности свечения ячейки) (рис. 3). Учет этого факта произведен с помощью алгоритма «огибающей шума» на этапах фильтрации, кластеризации и группировки генов.

Для генов, прошедших процедуру фильтрации, методами иерархической кластеризации построены дендрограммы. На основе этих дендрограмм планируется

сравнительное изучение динамики изменения экспрессии генов ключевых сигнальных путей связанных с программируемой клеточной смертью и репарацией ДНК в раковых клетках с нормальным и мутантным геном TP53.

После расчёта указанной динамики группы генов могут быть подвержены процедуре кластеризации, позволяющей визуализировать изменения экспрессии уже не отдельных генов, а их групп. Подробный кластерный анализ и исследование корреляций между группами планируется в отдельной работе.

Работа выполнена при поддержке ФЦП «Научные и научно-педагогические кадры инновационной России» грант № 14.B37.21.0558.

СПИСОК ЛИТЕРАТУРЫ

1. Schena M. Microarrays: biotechnology's discovery platform for functional genomics. *Trends in Biotechnology*. 1998. V. 16. № 7. P. 301–306.
2. Young R.A. Biomedical discovery with DNA arrays. *Cell*. 2000. V. 102. Iss. 1. P. 9–15.
3. Butte A. The use and analysis of microarray data. *Nature Reviews Drug Discovery*. 2002. V. 1. № 12. P. 951–960.
4. Slonim D.K. From patterns to pathways: Gene expression data analysis comes of age. *Nature Genetics*. 2002. V. 32. P. 502–508.
5. Stafford P. *Methods in Microarray Normalization*. CRC Press, 2008.
6. Lim W.K., Wang K., Lefebvre C., Califano A. Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. *Bioinformatics*. 2007. V. 23. № 13. P. i282–i288.
7. *Affymetrics Inc. Statistical algorithms description document*. 2013. URL: http://www.affymetrics.com/support/technical/whitepapers.affx (дата обращения: 31.07.2013).
8. Когадеева М.С. *Математическая модель данных микрочипов ДНК и методы оценки её параметров*: дипломная работа. М.: МГУ. 2011. С. 1–65.
9. 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 Biology*. 2004. V. 5. № 10. P. R80:1–R80:16.
10. Gautier L., Cope L., Bolstad B. M., Irizarry R.A. affy-analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*. 2004. V. 20. № 3. P. 307–315.
11. Smith C.A. Browser-based Affymetrix Analysis and Annotation. In: *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. New York: Springer, 2005. P. 313–326.
12. Smyth G.K. Limma: linear models for microarray data. In: *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. New York: Springer, 2005. P. 397–420.
13. Gentleman R., Carey V., Huber W., Hahne F. *Genefilter: methods for filtering genes from microarray experiments*. Version 1.42.0. 2013. URL: http://www.bioconductor.org/packages/2.12/bioc/manuals/genefilter/man/genefilter.pdf (дата обращения: 31.09.2013).
14. URL: http://media.affymetrix.com/support/developer/powertools/changelog/gcos-gcc/cel.html (дата обращения: 30.10.2013).
15. Garrett-Mayer E. Overview of Standard Clustering Approaches for Gene Microarray Data Analysis. In: *DNA Microarrays and Related Genomics Techniques Design, Analysis, and Interpretation of Experiments*. Eds. D.B. Allison, G.P. Page, T.M. Beasley, J.W. Edwards. Taylor & Francis Group, LLC, 2006.

16. Liu M., Wang G., Gomez-Fernandez C.R., Guo S. GREB1 functions as a growth promoter and is modulated by IL6/STAT3 in breast cancer. *PLoS One*. 2012. V. 7. № 10. P. e46410.
17. He G., Sun D., Ou Z., Ding A. The Protein Zfand5 binds and stabilizes mRNAs with AU-rich elements in their 3′-untranslated regions. *Journal of Biological Chemistry*. 2012. V. 287. № 30. P. 24967–24977.
18. Skinner H.D., Sandulache V.C., Ow T.J., Meyn R.E., Yordy J.C., Beadle B.M., Fitzgerald A.L., Giri U., Ang K.K., Myers J.N. *TP53* Disruptive Mutations Lead to Head and Neck Cancer Treatment Failure through Inhibition of Radiation-Induced Senescence. *Clinical Cancer Research*. 2012. V. 18. № 1. P. 290–300.
19. Bolstad B.M., Irizarry R.A., Astrand M., Speed T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*. 2003. V. 19. № 2. P. 185–193.

Материал поступил в редакцию 03.09.2013, опубликован 31.10.2013.