УДК 577.21:577.214 + 616-006.48

ОПРЕДЕЛЕНИЕ МОЛЕКУЛЯРНЫХ ПОДКЛАССОВ ГЛИОБЛАСТОМ НА ОСНОВЕ АНАЛИЗА ЭКСПРЕССИИ ГЕНОВ

В.В. ДМИТРЕНКО [1], А.В. ЕРШОВ [1], П.И. СТЕЦЮК [2], А.П. ЛИХОВИД [2], Ю.П. ЛАПТИН [2], Д.Р. ШВАРЦ [3], А.А. МЕКЛЕР [3], В.М. КАВСАН [1]

[1] Институт молекулярной биологии и генетики НАН Украины, Киев
E-mail: dmitrenko55@gmail.com
[2] Институт кибернетики им. В.М. Глушкова НАН Украины, Киев
[3] Санкт-Петербургский государственный университет телекоммуникаций им. проф. М.А. Бонч-Бруевича

Две группы глиобластом, отличающихся между собой по уровню экспрессии 416 генов ($P \leq 0,05$), определены с применением математической модели в форме линейного булевого программирования на основе наявных в базе данных Gene Expression Omnibus (GEO) по экспрессии генов в глиобластомах, полученных с помощью анализа микрочипов. Уровень экспрессии 15 генов более чем в два раза выше в первой группе глиобластом (80 образцов) по сравнению со второй группой (144 образца), а 401 гена – более чем в два раза ниже по сравнению со второй группой. Из 15 генов, уровень экспрессии которых преобладает в первой группе глиобластом, 10 кодируют белки, вовлеченные в регуляцию клеточного цикла и пролиферации клеток. Значительную часть 401 генов составляют гены, которые кодируют белки, вовлеченные в функционирование нейральных клеток и принимающие участие в таких процессах, как синаптическая передача, нейрогенез, образование миелиновой оболочки и аксонов. Карта Кохонена, построенная на основе данных 15 генов, уровень экспрессии которых превалирует в первой группе, и 60 (из 401) генов, уровень экспрессии которых выше во второй группе, подтвердила существование двух групп глиобластом со специфическими профилями экспрессии генов. Разделение глиобластом на две группы может отражать два пути развития астроцитарных глиом, один из которых приводит к образованию опухолей с более высоким уровнем экспрессии генов, белковые продукты которых вовлечены в регуляцию клеточного цикла и пролиферации. Вместе с тем существование двух молекулярных вариантов, возможно, является отражением различных стадий развития глиобластом.

Ключевые слова: *сигнатура экспрессии генов, глиобластома, классификация опухолей, «пролиферативный подтип», «пронейральный подтип».*

Введение. Большое количество исследований, направленных на идентификацию и характеристику изменений экспрессии генов в определенном типе опухолей, позволило получить базу данных относительно экспрессии всего генома человека для большого количества биологических образцов. Коллекция такого беспрецедентного количества данных с одной стороны вызвала необходимость развития новых подходов к их анализу, а с другой стороны дала возможность получения более достоверных результатов. Одним из результатов анализа этой обширной базы данных может быть идентификация специфических профилей экспрессии генов в опухолевых клетках, так называемых сигнатур, которые необходимы для понимания процессов возникновения злокачественных новообразований, а в практическом плане – для прогностической оценки и рационализации стратегии лечения пациентов.

Имеются несколько примеров разработки сигнатур для прогностической оценки определенных видов рака, в частности рака молочной железы. Например, в работе ван Вир и соавт. [1] на основе анализа экспрессии генов в опухолях молочной железы с известным клиническим исходом разработана сигнатура из 70 генов, которая успешно испытана на большой группе пациентов для прогностической оценки последствий лечения этого типа злокачественных новообразований [2]. Сигнатура из 76 генов, идентифицированная подобным образом в другой лаборатории [3], также успешно верифицирована на большой группе пациентов с опухолями молочной железы. На основе экспериментальных данных также идентифицирована сигнатура из 16 генов (из 250 генов-кандидатов, протестированных на почти 400 опухолях) [4]. Сейчас существуют три коммерческие сигнатуры для прогностической оценки рака молочной железы: 70-генная MammaPrint (Agendia) [5], сигнатура Oncotype DX (Ge-

nomic Health) из 21 ге-на [6] и сигнатура из двух генов H/I test (Aviara Dx), разработанная в исследованиях Ma и соавт. [7]. Первые две сигнатуры интенсивно используются в клинической практике. Применение сигнатур позволило разработать новую молекулярную классификацию и определение шести подтипов опухолей молочной железы (luminal A, luminal B, HER2-enriched, basal-like, normal breast и claudinlow [8]). Следствием их использования стало лучшее понимание сигнальных путей, которые управляют процессами формирования, поддержания и экспансии опухолей. Теперь больше известно о роли рецепторов эпидермального фактора роста 2 (HER2), эстрогена (ER), инсулиноподобного фактора роста 1 (IGF1R), а также сигнальных путей PI3K/AKT, mTOR, AMPK и ангиогенеза, что способствует развитию новых направленных терапий, которые проходят тестирование в современных клинических испытаниях [8].

Подобных коммерческих сигнатур для прогностической оценки глиальных опухолей пока не существует, хотя есть публикации об идентификации профилей экспрессии генов, которые могут быть использованы для идентификации определенных групп среди глиом, в частности глиобластом. Так, в работе Демут и соавт. [9] с помощью комбинированного использования экспериментальных результатов и ранее опубликованных данных определены специфические профили экспрессии генов миграторных и стационарных клеток глиом. На основе этих данных предложена сигнатура оценки миграции из 22 генов, которая была тестирована на культурах глиальных опухолей. Она позволяет различить миграторные и стационарные клеточные линии, а также может быть использована для оценки инвазивных свойств глиальных опухолей. Другим примером профилирования экспрессии генов в глиальных опухолях является работа Лаи и соавт. [10], в которой авторы на основе данных по экспрессии генов разработали модель для классификации глиом. Она предусматривает две группы глиом (преимущественно олигодендроглиомы и глиобластомы), которые можно разделить на шесть иерархических подтипов с помощью шести генных классификаторов. Наиболее ярким примером молекулярной классификации глиобластом является работа Филлипса и др. [11], где авторы определили три подкласса опухолей, обозначенные как пронейральный (PN), пролиферативный (Prolif) и мезенхимальный (Mes) для распознавания главной особенности списков генов, повышенная экспрессия которых характеризует каждый подкласс, а общий список этих генов (35 генов) авторы предложили как сигнатуру генов, пригодную для отнесения опухоли к одному из подклассов. Пронейральный подтип опухолей имеет более благоприятный прогноз по сравнению с пролиферативным и мезенхимальным, профиль экспрессии генов этого подтипа подобен нормальному головному мозгу и процессу нейрогенеза. Неблагоприятные подтипы глиобластом – пролиферативный и мезенхимальный – характеризуются экспрессией маркеров пролиферации и ангиогенеза. При повторном возникновении глиобластом они имеют тенденцию к переходу в мезенхимальный подтип. Этот переход напоминает собой эпителиально-мезенхимальный переход, ассоциированный с более злокачественным поведением эпителиальных опухолей. Авторы высказывают предположение, что способность изменять подтип является отражением разного состояния дифференциации опухолей, однако не исключают, что некоторые очевидные переходы скорее служат отражением гетерогенности опухолей, чем изменений их характера со временем. Повторное возникновение глиобластомы после стандартной терапии может сопровождаться переходом к более агрессивному мезенхимальному подтипу, и это требует соответствующей коррекции их терапии [11].

Недавно появилось еще несколько публикаций по идентификации сигнатур экспрессии генов глиальных опухолей [12–14]. В подавляющем большинстве работ по определению сигнатуры конечным результатом является отбор авторами нескольких генов для характеристики глиальных опухолей, что недостаточно для их корректной классификации и прогностической оценки. В частности, в работе Колмана и соавт. [12] идентифицирован набор из 9 генов, повышенная экспрессия которых ассоциирована с неблагоприятным прогнозом для пациентов с глиобластомами, а в работе Детайрак и соавт. [13] предложена даже 4-ген-

ная сигнатура, способная независимо оценивать риск неблагоприятного прогноза для высокозлокачественных глиом – глиобластом и анапластичних астроцитом.

Как показал анализ большого списка работ, сигнатуры экспрессии генов для одного и того же типа опухолей, определенные в разных местах, имеют очень незначительное перекрывание списков генов. Так, упомянутые ранее сигнатуры экспрессии генов рака молочной железы из 70 [1] и 76 генов [3] имеют только три общих гена. Авторы объясняют успешность использования различных сигнатур тем, что разные наборы генов, позволяющие успешно классифицировать пациентов, отображают подобные биологические процессы, которые проявляются на уровне экспрессии индивидуальных генов [15]. Списки генов в сигнатурах, предложенных для предсказания риска рецидива у пациентов с опухолями толстой кишки II и III стадий, также имеют минимальное перекрывание [16–22]. Такое незначительное перекрывание вызывает вопрос о возможности клинического применения этих сигнатур, хотя прогностические модели, основанные на нескольких сигнатурах рака толстой кишки, были подтверждены на независимых когортах пациентов [19–21]. Выявление общих биологических закономерностей, лежащих в основе различных по составу сигнатур экспрессии генов, уменьшило высказанное ранее скептическое отношение относительно биологической важности отдельных генов, входящих в состав этих сигнатур.

Приведенное выше касается также и глиальных опухолей. Списки генов сигнатур, определенных в упомянутых работах, очень отличаются. Это может быть обусловлено различными путями образования опухолей вследствие повреждения множественных биологических процессов, и каждая сигнатура охватывает лишь отдельные гены, продукты которых вовлечены в эти процессы. Поэтому определение новых специфических сигнатур экспрессии генов для астроцитарных глиом является актуальным как с точки зрения более полного понимания механизмов возникновения и развития этих гетерогенных по биологическим свойствам новообразований головного мозга, так и возможности клинического использования объединенных сигнатур, определенных в разных работах.

Материалы и методы. *Создание базы данных экспрессии генов в глиальных опухолях на основе результатов анализа микрочипов.* На сайте http://www.ncbi.nlm.nih.gov/geo проводился поиск DataSet-файлов по ключевым словам: glioblastoma, astrocytoma, normal brain. Информация, представленная в DataSet, отражает экспериментальные данные по профилированию экспрессии генов в глиальных опухолях различной степени злокачественности, а также в образцах нормального головного мозга с помощью гибридизационного анализа олигонуклеотидных микрочипов суммарными пробами кДНК глиальных опухолей или нормального головного мезга, которые можно сравнивать статистически в пределах одного файла. Для сравнения данных из различных файлов проводилась нормализация по трем референсным генам. Для этого значение экспрессии всех генов в каждом образце делили на среднее геометрическое значение уровней экспрессии генов *ACTB*, *GAPDH* и *TBP* в этом образце [10]. Для анализа данных микрочипов и их представления в нужном формате использовали скрипты, написанные на языке Perl [24].

Определение количества кластеров в опухолях с помощью метода k-средних. В среде R (свободно доступная программная среда) использовали алгоритм Хартигана и Вонга [25], который разделяет наблюдения на k групп, таких, чтобы сумма квадратов расстояний каждой точки к центру кластера была минимальной. Алгоритм кластеризации: 1) избирали k центроидов (k выбирается произвольно); 2) каждое наблюдение (образец) приписывали к ближайшему центроиду; 3) снова вычисляли центроиды кластеров как среднее всех точек кластера (центроиды – это векторы длины p, где p – количество переменных); 4) приписывали каждую точку данных к ближайшему центроиду. Шаги 3 и 4 повторялись, пока причастность точек к кластерам не становилась постоянной или было достигнуто максимальное число итераций. Массив данных из статьи Филлипса и соавт. [11] без нормализации по генам «домашнего хозяйства» был разбит на части по 600 генов для всех образцов (75 глиобластом), и для каждого из полученных под-

массивов с помощью 26 критериев в среде R с использованием пакета NbClust [26] подсчитано возможное количество кластеров, на которые делятся данные. За максимально возможное число кластеров было принято 15. Вычисления проведены для 22 283 генов, доступных на платформе GPL96. Количество кластеров оценивали по следующим критериям: 1) Калинского и Харабаза [27]; 2) Дуды [28]; 3) Pseudoplot2 [28]; 4) C-критерию Хьюберта и Левина [29]; 5) гамма-критерию Бейкера и Хьюберта [30]; 6) Била [31]; 7) кубическому кластерному критерию Сарле [32]; 8) PtBserial [33–45]; 9) Gplus [35]; 10) Дэвиса и Боулдина [36]; 11) Фрея [37]; 12) Хартигана [25]; 13) tau-критерию [34]; 14) Ратковски [38]; 15) Скотта [39]; 16) Мариотта [40]; 17) Болла [41]; 18) TraceCoW [35]; 19) TraceW [35]; 20) Фридмана [42]; 21) МакКлейна [43]; 22) Рубина [42]; 23) KL-критерию [44]; 24) Сильхуетта [45]; 25) Gap-критерию [46] и 26) Дунна [47].

Кластерный анализ глиобластом с использованием математической модели кластеризации в форме задачи линейного булевого программирования данных. Определение трех возможных подгрупп среди 224 образцов глиобластом проводили с использованием математической модели в форме задачи линейного булевого программирования [48]. Суть вычислений с применением этой модели заключается в том, что все образцы разбиваются на три группы и в каждой из них определяется свой центр – один из образцов, который попал в эту группу. Группа характеризуется суммой всех расстояний (эвклидовых или манхеттеновских) от точек (образцов) группы к центру этой группы. Группы выбираются так, чтобы суммарные расстояния по всем группам были минимальными. Нахождение такого разбиения на группы обеспечивает оптимальное решение задачи линейного булевого программирования.

Нейросетевой анализ и кластеризация образцов с помощью самоорганизующихся карт Кохонена. Для получения критерия оценки данных с целью их классификации использованы самоорганизующиеся карты Кохонена (Self-Organizing Maps, SOM). Отбор генов-кандидатов для сигнатуры экспрессии генов проводился на основе оценки того, насколько хорошо кластеризуются на карте векторы, принадлежащие разным классам, т.е. данные по экспрессии генов в различных группах опухолей [58]. Удовлетворительная кластеризация на SOM свидетельствует о том, что использованные векторы могут успешно классифицироваться нейронными сетями, обучаемыми с учителем (supervised artificial neural networks). Оценка основана на вычислении соотношения топологических расстояний на SOM

$$\alpha_T = \frac{\sum\limits_{i\in M}\sum\limits_{j\in M, j\neq i} Dist(x_i, x_j) + \sum\limits_{i\in L}\sum\limits_{j\in L, j\neq i} Dist(x_i, x_j)}{\sum\limits_{i\in\{L, M\}}\sum\limits_{j\in\{L, M\}j\neq i} Dist(x_i, x_j)},$$

где $Dist(x_i,y_j)$ – топологическое расстояние на карте между двумя нейронами-победителями (НП), M – множество данных, принадлежащих к классу A, L – множество данных, принадлежащих к классу B. Таким образом, сравнивалось расстояние между НП, принадлежащими к одному классу (числитель), и расстояние между НП, принадлежащими к разным классам (знаменатель). Коэффициент α_T является оценкой качества разделения двух классов на кластеры. НП, соответствующие векторам из обеих групп в случае неудачной кластеризации, располагались на карте вперемешку. При этом топологические расстояния между парами НП, принадлежащими к одному классу, и парами НП, принадлежащими к разным классам, были примерно одинаковы, и α_T имело максимальное значение. В случае удачной кластеризации расстояние между нейронами одного класса было меньше, чем между нейронами разных классов, и α_T имел минимальное значение.

Результаты исследований и их обсуждение. DataSet-файлы из базы данных Gene Expression Omnibus (GEO) Datasets (http://www.ncbi.nlm.nih.gov/gds) использовали для создания базы данных экспрессии генов в астроцитарных глиомах всех степеней злокачественности. Эти файлы представляют собой экспериментальные данные по экспрессии генов на основе анализа микрочипов [11, 49–53]. В общем, в базе данных GEO по ключевым словам «normal brain», «astrocytoma» и «glioblastoma» выявлено шесть DataSet-файлов, которые в сумме содержали данные о 71 образце нормального головного мозга, 45 образцах пилоцитарных астроцитом (астроци-

том I степени злокачественности по классификации ВОЗ [23]), 17 образцах диффузных астроцитом (II степени злокачественности), 93 образцах анапластичных астроцитом (III степени злокачественности) и 224 образцов глиобластом (IV степени злокачественности) (табл. 1).

В исследованиях, результаты которых представлены в найденных DataSet-файлах, использовались две платформы микрочипов, поэтому нами с помощью скрипта R отобраны данные только для генов, находящихся на обеих платформах. В общем, была создана таблица данных по экспрессии порядка 20 тысяч генов (20 447 общих проб генов, нанесенных на платформы GPL96 и GPL570) в 453 образцах астроцитарных глиом I–IV степеней злокачественности и нормального головного мозга.

Принимая во внимание возможность того, что опухоли одинаковой степени злокачественности, которые отличаются на молекулярном

Таблица 1. **Характеристика DataSet-файлов из репозитария Gene Expression Omnibus (GEO), использованных для оценки изменений экспрессии генов в глиобластоме**

Файл	Микрочип	Ссылка	Краткое описание экспериментов
GDS1815	GPL96 [HG-U133A] Affymetrix Human Genome U133A Array	Phillips et al., 2006 [11]	Профилирование 77 первичных злокачественных астроцитом и 23 рецидивных опухолей для идентификации изменений экспрессии генов, ассоциированных с периодом выживания пациентов и прогрессией болезни. Идентифицированы новые прогнозные подклассы астроцитом высокой степени злокачественности, имеющие сходство со стадиями нейрогенеза
GDS1975	GPL96 [HG-U133A] Affymetrix Human Genome U133A Array	Freije et al., 2004 [49]	Масштабный анализ экспрессии генов в 85 диффузных глиомах всех гистологических типов. Цель – определение независимого классификатора глиом на базе генной экспрессии
GDS1096	GPL96 [HG-U133A] Affymetrix Human Genome U133A Array	Ge et al., 2005 [50]	Полногеномное экспрессионное профилирование 36 типов нормальных тканей человека. Каждый образец РНК представлял собой пул из нескольких доноров. Идентифицированы 503 тканеспецифических гена. Результаты дают основу для анализа профилей экспрессии генов злокачественных опухолей
GDS1962	GPL570 [HG-U133_ Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array	Sun et al., 2006 [51]	Для уточнения диагноза глиом на молекулярном уровне получены данные по экспрессии генов у пациентов с опухолями головного мозга. 23 образца от пациентов-эпилептиков использовано в качестве неопухолевых образцов. Среди 157 опухолевых образцов было 26 астроцитом, 50 олигодендроглиом и 81 глиобластома
GDS3069	GPL96 [HG-U133A] Affymetrix Human Genome U133A Array	Liu et al., 2007 [52]	Анализ 12 первичных опухолей головного мозга, имеющих некоторые вариации гистологических диагнозов. Полученные результаты вместе с данными по профилированию miРНК на основе ПЦР в реальном времени позволяют понять взаимоотношения между колебаниями уровней miРНК и мРНК
GDS596	GPL96 [HG-U133A] Affymetrix Human Genome U133A Array	Su et al., 2004 [53]	Исследование экспрессии большого количества генов, кодирующих белки человека и мыши, на основе микроарейного анализа панели, содержащей 79 человеческих и 61 мышиных тканей. Получены паттерны экспрессии для нескольких тысяч генов мыши и человека, как уже известных, так и недостаточно охарактеризованных

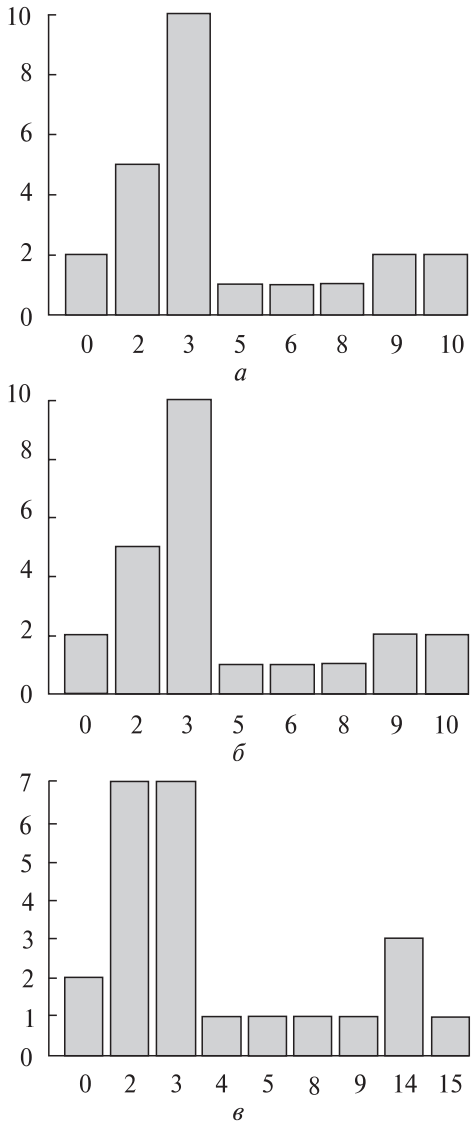

Рис. 1. Определение возможного количества кластеров (по горизонтали) среди глиобластом по данным Филлипса и соавт. [11]: *а* – пробы 1–600 из списка 22 283 проб генов; *б* – 601–1200; *в* – 2401–3000, по вертикали – количество критериев

уровне, проявляют разную чувствительность к терапии, выявление и характеристика молекулярных подтипов опухолей может способствовать развитию эффективных терапевтических подходов для лечения высокозлокачественных глиом. Поэтому на основе данных по экспрессии генов, полученных анализом микрочипов, осуществлен поиск возможных молекулярных вариантов глиобластом. Сначала была проведена оценка количества кластеров, на которые могут распределяться глиобластомы на основе данных по экспрессии генов, с помощью метода *k*-средних, который является одним из наиболее распространенных методов выделения групп в большом объеме данных [54]. Кластеризация с помощью *k*-средних имеет преимущество в том, что может быть использована для более крупных массивов данных, чем позволяет подход иерархического кластерного анализа. К тому же объекты кластеризации не всегда приписаны к одному кластеру – они могут перемещаться в другой кластер, что улучшает конечный результат. Массив данных из статьи Филлипса и соавт. [11] без нормализации по генам «домашнего хозяйства» был разбит на части по 600 генов для всех образцов (75 глиобластом), и для каждого из полученных подмассивов с помощью 26 критериев подсчитано количество кластеров. За максимально возможное число кластеров принято 15 кластеров с целью ограничения объема расчетов. Вычисления выполнены для 22 283 проб генов, имеющихся на микроаррейной платформе GPL96, с помощью которой проводилось профилирование экспрессии генов в глиобластомах. По оценкам большинства критериев глиобластомы с наибольшей вероятностью делятся на два или три кластера. На рис. 1 показаны примеры определения количества кластеров среди глиобластом для различных наборов генов.

Предварительная оценка количества возможных кластеров глиобластом на основе данных по экспрессии 22 283 генов показала, что наиболее вероятной является деление образцов глиобластом на два или три кластера. Принимая во внимание эти данные, провели определение трех возможных подгрупп среди 224 образцов глиобластом с использованием математической модели в форме булевого линейного программирования. В результате вычислений обнаружили три кластера глиобластом, не похожих между собой паттерном экспрессии генов. Построение так называемого «heatmap»-графика с использованием программной среды R позволило наглядно представить результаты кластеризации образцов (рис. 2, см. вклейку в конце номера). В публикациях, описывающих определения различных молекулярных вариантов глиобластом, авторы в большинстве слу-

чаев также подразделяют эти опухоли на три, иногда четыре группы.

Данные по экспрессии генов были нормированы к единице (значение экспрессии гена в каждом образце разделено на максимальную величину экспрессии гена среди всех образцов) и упорядочены по среднему значению экспрессии генов во втором кластере (кластер *2* на рис. 3, см. вклейку в конце номера). Однако как видно из графика, кластеры *1* и *3* очень похожи между собой. После перемещения четырех образцов глиобластом из кластера *3* в кластер *2* (рис. 3, *б*) и объединения кластеров *1* и *3* глиобластомы могут быть распределены на две группы, которые отличаются по профилю экспрессии генов (рис. 3, *в*).

Подобные результаты получены при попытке определения четырех кластеров среди образцов глиобластом. Визуальный анализ «heatmap»-графиков, отражающих результаты выявления четырех кластеров, также показал, что глиобластомы могут быть разделены на две группы – одна размером 80 образцов (кластер *1* на рис. 3, *в*), а другая – 144 образца (кластер *2* на рис. 3, *в*). Для подтверждения распределения глиобластом на две группы проведено определение двух кластеров среди 224 образцов глиобластом с использованием подхода решения задачи булевого линейного программирования. В результате этого анализа глиобластомы разделились на две группы размером 109 и 115 образцов. Сопоставление списков образцов глиобластом в этих двух группах и двух группах размером 80 и 144 глиобластом показало существенное перекрывание – 103 общих образца для групп из 115 и 144 глиобластом и 68 общих образцов для групп из 109 и 80 глиобластом. Таким образом, полученные результаты с высокой вероятностью свидетельствуют о существовании двух молекулярных вариантов глиобластом, отличающихся по профилю экспрессии генов. Сопоставление с клиническими данными, доступными на сайте GEO для образцов глиобластом, которые проанализированы в четырех исследованиях [11, 49, 51, 52] и использованы нами для этой работы, показало, что две группы глиобластом (115 и 109 образцов), определенные нами с помощью подхода булевого линейного программирования, не совпадали с подклассами опухолей (PN, Prolif и Mes), описанными в работе Филлипса и соавт. [11]. Не найдено также корреляции ни с возрастом или полом пациентов, ни с периодом их выживаемости. Сопоставление экспрессии генов в двух группах глиобластом выявило 480 проб генов с более чем двукратной разницей экспрессии с вероятностью $P \leq 0{,}05$. Эти 480 проб представляют 416 генов, из которых уровень экспрессии 15 генов выше в первой группе, а 401 гена – во второй группе глиобластом.

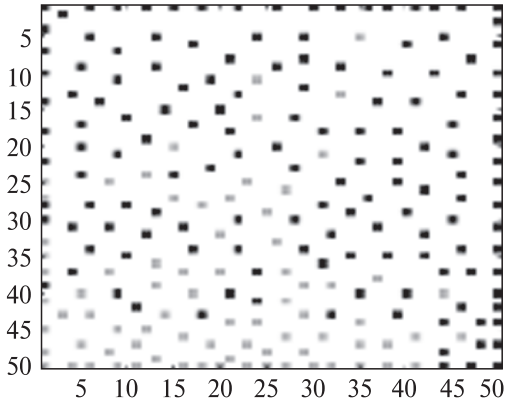

Рис. 4. Распределение двух групп глиобластом на карте Кохонена. Светлыми точками обозначены глиобластомы первой группы (80 образцов), темными – второй группы (144 образца)

Подтверждение распределения глиобластом на две группы проведено также с использованием нейросетевого анализа. Из 401 гена отобраны 60 генов, уровень экспрессии которых с наибольшей вероятностью повышен во второй группе ($P < 1{,}3 \cdot 10^{-10}$), и на основе данных этих 60 генов и 15 генов с повышенным уровнем экспрессии в первой группе построена карта Кохонена. Как видно из рис. 4, две группы глиобластом достаточно хорошо кластеризуются на карте Кохонена.

На основе данных, имеющихся на сайте NCBI (http://www.ncbi.nlm.nih.gov/gene) и в базе данных KEGG (Kyoto Encyclopedia of Genes and Genomes) и REACTOME, проанализированы функции белковых продуктов 416 генов, уровень экспрессия которых отличается между двумя группами глиобластом. Оказалось, что значительный процент из 401 генов, уровень экспрессии которых выше более чем в два раза во второй группе глиобластом, составляют

гены, которые кодируют белки, вовлеченные в функционирование нейральных клеток и принимающие участие в таких процессах, как синаптическая передача – CAMK1G (calcium/calmodulin-dependent protein kinase type 1G), CAMK2B (calcium/calmodulin-dependent protein kinase beta), CPNE6 (copine VI, neuronal), DDN (dendrin), GABRA1 (gamma-aminobutyric acid receptor subunit alpha-1), GABRA5 (gamma-aminobutyric acid receptor subunit alpha-5), GRIN1 (glutamate receptor, ionotropic, N-methyl D-aspartate 1), GRIN2C (glutamate receptor, ionotropic, N-methyl D-aspartate 2C), GRM1 (glutamate receptor, metabotropic 1), SNAP25 (synaptosomal-associated protein, 25kDa), SNCA (synuclein, alpha), SNCG (synuclein, gamma), SYT1 (synaptotagmin I), SYT5 (synaptotagmin V), SYN1 (synapsin I) и SYNGR3 (Synaptogyrin 3), нейрогенез – NRGN (neurogranin), образование миелиновой оболочки – MBP (myelin basic protein) и MOBP (myelin-associated olygodendrocyte basic protein), образование аксонов – NEFL (neurofilament, light polypeptide) и NEFH (neurofilament, heavy polypeptide).

Из 15 генов, уровень экспрессии которых преобладает в первой группе глиобластом, 10 кодируют белки, вовлеченные в регуляцию клеточного цикла и пролиферации клеток (табл. 2, см. http://cytgen.com/articles/4860045s.pdf). Многочисленные публикации демонстрируют участие большинства белков, кодируемых этими 15 генами, в развитии различных видов опухолей. Их продукция в опухолевых клетках обусловливает высокую пролиферативную активность и инвазивный рост опухоли, стимулирует ангиогенез.

Таким образом, в соответствии с профилем экспрессии 416 генов, определенных с использованием подхода булевого линейного программирования, глиобластомы могут быть разделены на две группы, в одной из которых преобладает экспрессия «пролиферативных» генов, а в другой – «пронейральных» генов. Конечно, эти две группы глиобластом не являются однородными и, как видно из рис. 3, *в*, внутри них существуют различные подгруппы со своими специфическими профилями экспрессии генов.

Полученные нами результаты подобны опубликованным в работе Хьюза и соавт. [55], в которой авторы сравнивали молекулярные классификаторы глиобластом, определенные в разных лабораториях. В результате взаимной проверки двух схем классификации глиобластом, которые получены на разных наборах образцов глиобластом и описаны в работах Филлипса и соавт. [11] и Верхака и соавт. [56], было определено, что только два подтипа глиобластом, обозначенные авторами как «пронейральный» (Proneural) и «мезенхимальный» (Mesenchimal), в целом хорошо определяются «пронейральными» и «мезенхимальными» сигнатурами из обеих работ. Другие субтипы глиобластом – «классический» и «нейральный» в работе Верхака и соавт. [56], а также «пролиферативный» в работе Филлипса и соавт. [11] – перераспределяются между другими субтипами при перекрестном применении сигнатур, т.е. не являются однозначными. «Пролиферативная» сигнатура из работы Филлипса и соавт. [11] сочетается с элементами других сигнатур, в частности с «мезенхимальной». Вместе с тем «пролиферативный» субтип глиобластом идентифицирован в работе Фрейдж и соавт. [49], и «пролиферативная» сигнатура оказалась прогностификатором неблагоприятного исхода заболевания, что было подтверждено и для педиатрических глиобластом [57].

Хьюз и соавт. [55] на основе собственных результатов и анализа работ других авторов связывают две основные по их мнению сигнатуры («мезенхимальную» и «пронейральную») с двумя альтернативными механизмами глиомагенеза – развитием первичных и вторичных глиобластом соответственно. Однако, как видно из схемы, приведенной в их работе, и анализа клинических данных, которые приведены в работах по транскрипционному субтипированию глиобластом, в частности в работе Филлипса и соавт. [11], прямой корреляции между «мезенхимальной» сигнатурой и первичными глиобластомами, а «пронейральной» сигнатуры – со вторичными глиобластомами не существует. Возможно, распределение глиобластомы на две группы на основе профилей экспрессии определенных нами 416 генов является отражением двух путей развития астроцитарных глиом, один из которых приводит к образованию более агрессивного типа опухолей с более высоким уровнем экспрессии «пролиферативных» генов по сравнению с другим ти-

пом, характеризующимся более высоким уровнем экспрессии «пронейральных» генов. Вместе с тем существование двух молекулярных вариантов, возможно, является отражением различного состояния развития глиобластом. Определенная в этой работе 416-генная сигнатура тре-бует тщательного анализа с целью оценки ее пригодности для классификации глиальных опухолей и применения в клинической практике.

Работа частично финансировалась Национальной академией наук Украины в рамках совместного конкурса НАН Украины и Российского фонда фундаментальных исследований 2012 г. (проекты № 07-0412 и 12-04-90434-Укр_а) и Государственного фонда фундаментальных исследований Украины (проект № Ф52.4/003).

DETERMINATION OF MOLECULAR GLIOBLASTOMA SUBCLASSES ON THE BASIS OF ANALYSIS OF GENE EXPRESSION

V.V. Dmitrenko, A.V. Iershov, P.I. Stetsyuk, A.P. Lyhovid, Yu.P. Laptin, A.A. Mekler, V.M. Kavsan

Institute of Molecular Biology and Genetics of NAS of Ukraine, Kyiv
V.M. Glushkov Institute of Cybernetics of NAS of Ukraine, Kyiv
The Bonch-Bruevich Saint Petersburg State University of Telecommunications, RF

Two glioblastoma groups, which are distinguished from each other by expression level of 416 genes ($P \le 0,05$), were determined using a mathematical model of linear Boolean programming on the basis of gene expression data, obtained by microarray analysis of the glioblastomas and available in Gene Expression Omnibus (GEO) data base. The expression level of 15 genes was more than two-fold higher in the first group of glioblastoma (80 samples) in comparison with the second group (144 samples) and 401 genes ond – more than two-fold lower as compared to the second group. 10 of 15 genes, which expression level prevailed in the first group, encode the proteins involved in cell cycle regulation and cell proliferation. A significant percentage of 401 genes are the genes that encode proteins involved in the functioning of neural cells and participating in the processes such as synaptic transmission, neurogenesis, the formation of myelin sheath, axon formation. Kohonen map, built on the basis of the data of 15 genes with prevailed expression in the first group and 60 (of 401) genes, whose expression level elevated in the second group, confirmed the existence of two glioblastoma groups with specific gene expression profiles. Distribution of the glioblastomas into two groups may reflect two pathways of astrocytic glioma development, one of which leads to the formation of tumors with higher levels of gene expression, which protein products are involved in cell cycle regulation and proliferation. On the other hand, the existence of two molecular variants may reflect different states of glioblastoma progression.

ВИЗНАЧЕННЯ МОЛЕКУЛЯРНИХ ПІДКЛАСІВ ГЛІОБЛАСТОМ НА ОСНОВІ АНАЛІЗУ ЕКСПРЕСІЇ ГЕНІВ

В.В. Дмитренко, А.В. Єршов, П.І. Стецюк, О.П. Лиховид, Ю.П. Лаптін, О.А. Меклер, В.М. Кавсан

Дві групи гліобластом, що відрізняються між собою за рівнем експресії 416 генів, визначено із застосуванням математичної моделі у формі лінійного булевого програмування на основі даних по експресії генів у гліобластомах, отриманих за допомогою мікроарейного аналізу і наявних у базі даних Gene Expression Omnibus (GEO). Рівень експресії 15 генів у понад два рази вищий в першій групі гліобластом (80 зразків) порівняно з другою групою (144 зразки), а 401 генів – у понад два рази нижчий порівняно з другою групою. 10 з 15 генів, рівень експресії яких переважає у першій групі, кодують білки, залучені до регуляції клітинного циклу та проліферації клітин. Значний відсоток 401 генів складають гени, що кодують білки, залучені до функціонування нейральних клітин і беруть участь у синаптичній передачі, нейрогенезі, утворенні мієлінової оболонки та аксонів. Карта Кохонена, побудована на основі даних 15 генів, рівень експресії яких превалює у першій групі, та 60 (з 401) генів, рівень експресії яких підвищений у другій групі, підтвердила існування двох груп гліобластом із специфічними профілями експресії генів. Розподіл гліобластом на дві групи може відображати два шляхи розвитку астроцитарних гліом, один з яких призводить до утворення пухлин з високим рівнем експресії генів, білкові продукти яких залучені до регуляції клітинного циклу та проліферації. Існування двох молекулярних варіантів, можливо, є відображенням різних стадій розвитку гліобластом.

СПИСОК ЛИТЕРАТУРЫ

1. *Van't Veer L.J., Dai H., van de Vijver M.J. et al.* Gene expression profiling predicts clinical outcome of breast cancer // Nature. – 2002. – **415,** № 6871. – P. 530–536.
2. *Chang H.Y., Nuyten D.S., Sneddon J.B. et al.* Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival // Proc. Nat. Acad. Sci. USA. – 2005. – **102**, № 10. – P. 3738–3743.
3. *Wang Y., Klijn J.G., Zhang Y. et al.* Gene-expression profiles to predict distant metastasis of lymph-node-

negative primary breast cancer // Lancet. – 2005. – **365**, № 9460. – P. 671–679.
4. *Paik S., Shak S., Tang G. et al.* A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer // N. Engl. J. Med. – 2004. – **351**, № 27. – P. 2817–2826.
5. *Tian S., Roepman P., Van't Veer L.J. et al.* Biological functions of the genes in the mammaprint breast cancer profile reflect the hallmarks of cancer // Biomark Insights. – 2010. – **5**. – P. 129–138.
6. *Conlin A.K., Seidman A.D.* Use of the Oncotype DX 21-gene assay to guide adjuvant decision making in early-stage breast cancer // Mol. Diagn. Ther. – 2007. – **11**, № 6. – P. 355–360.
7. *Ma X.J., Hilsenbeck S.G., Wang W. et al.* The HOXB13:IL17BR expression index is a prognostic factor in early-stage breast cancer // J. Clin. Oncol. – 2006. – **24**, № 28. – P. 4611–4619.
8. *Eroles P., Bosch A., Pérez-Fidalgo J.A., Lluch A.* Molecular biology in breast cancer: intrinsic subtypes and signaling pathways // Cancer Treat. Rev. – 2012. – **38**, № 6. – P. 698–707.
9. *Demuth T., Rennert J.L., Hoelzinger D.B. et al.* Glioma cells on the run – the migratory transcriptome of 10 human glioma cell lines // BMC Genom. – 2008. – **9**. – P. 54.
10. *Li A., Walling J., Ahn S. et al.* Unsupervised analysis of transcriptomic profiles reveals six glioma subtypes // Cancer Res. – 2009. – **69**, № 5. – P. 2091–2099.
11. *Phillips H.S., Kharbanda S., Chen R. et al.* Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis // Cancer Cell. – 2006. – **9**, № 3. – P. 157–173.
12. *Colman H., Zhang L., Sulman E.P. et al.* A multigene predictor of outcome in glioblastoma // Neuro-Oncol. – 2010. – **12**, № 1. – P. 49–57.
13. *de Tayrac M., Aubry M., Sankali S. et al.* A 4-gene signature associated with clinical outcome in high-grade gliomas // Clin. Cancer Res. – 2011. – **17**, № 2. – P. 317–327.
14. *Kim Y.W., Koul D., Kim S.H. et al.* Identification of prognostic gene signatures of glioblastoma: a study based on TCGA data analysis // Neuro-Oncol. – 2013. – **15**, № 7. – P. 829–839.
15. *Yu J.X., Sieuwerts A.M., Zhang Y. et al.* Pathway analysis of gene signatures predicting metastasis of node-negative primary breast cancer // BMC Cancer. – 2007. – **7**. – P. 182.
16. *Bandres E., Malumbres R., Cubedo E. et al.* A gene signature of 8 genes could identify the risk of recurrence and progression in Dukes' B colon cancer patients // Oncol. Rep. – 2007. – **17**. – P. 1089–1094.
17. *Barrier A., Boelle P.Y., Roser F. et al.* Stage II colon cancer prognosis prediction by tumor gene expression profiling // J. Clin. Oncol. – 2006. – **24**. – P. 4685–4691.
18. *Eschrich S., Yang I., Bloom G. et al.* Molecular staging for survival prediction of colorectal cancer patients // J. Clin. Oncol. – 2005. – **23**. – P. 3526–3535.
19. *Lin Y.H., Friederichs J., Black M.A. et al.* Multiple gene expression classifiers from different array platforms predict poor prognosis of colorectal cancer // Clin. Cancer Res. – 2007. – **13**. – P. 498–507.
20. *Smith J.J., Deane N.G., Wu F. et al.* Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer // Gastroenterology. – 2010. – **138**, № 3. – P. 958–968.
21. *Wang Y.X., Jatkoe T., Zhang Y. et al.* Gene expression profiles and molecular markers to predict recurrence of dukes' B colon cancer // J. Clin. Oncol. – 2004. – **22**, № 9. – P. 1564–1571.
22. *Shi M., Beauchamp R.D., Zhang B.* A network-based gene expression signature informs prognosis and treatment for colorectal cancer patients // PLoS One. – 2012. – **7**, № 7. – P. e41292.
23. *Louis D.N., Ohgaki H., Wiestler O.D. et al.* The 2007 WHO classification of tumours of the central nervous system // Acta Neuropathol. – 2007. – **114**, № 2. – P. 97–109.
24. *Уолл Л., Кристиансен Т., Орвант Д.* Программирование на Perl = Programming Perl. – М.: Символ-Плюс, 2008. – 1145 с.
25. *Hartigan J., Wong M.* Algorithm AS 136: A k-means clustering algorithm // J. Royal Stat. Soc. Ser. C. – 1979. – **28**, № 1. – P. 100–108.
26. *Charrad M., Ghazzali N., Boiteau V., Niknafs A.* NBclust package: finding the relevant number of clusters in database. http://cran.r-project.org/ web/packages/NbClust/index.html
27. *Caliński T., Harabasz J.* A dendrite method for cluster analysis // Commun. Stat. – 1974. – **3**, № 1. – P. 1–27.
28. *Duda R., Hart P.* Pattern classification and scene analysis. – New York : John Willey & Sons, 1973. – 482 p.
29. *Hubert L., Levin J.* A general statistical framework for assessing categorical clustering in free recall // Psychol. Bull. – 1976. – **83**, № 6. – P. 1072–1080.
30. *Baker F., Hubert L.* Measuring the power of hierarchical cluster analysis // J. Amer. Stat. Ass. – 1975. – **70**, № 349. – P. 31–38.
31. *Beale E.M.L.* Cluster analysis. – London : Sci. Control Syst., 1969. – 18 p.
32. *Sarle W.* Cubic Clustering Criterion SAS technical report A-108. – Cary, NC: SAS Institute, 1983.
33. *Milligan G.* An examination of the effect of six types of error perturbation on fifteen clustering algorithms // Psychometrika. – 1980. – **45**, № 3. – P. 325–342.

34. *Milligan G.* A monte carlo study of thirty internal criterion measures for cluster analysis // Psychometrika. – 1981. – **46**, № 2. – P. 187–199.
35. *Milligan G., Cooper M.* An examination of procedures for determining the number of clusters in a data set // Psychometrika. – 1985. – **50**, № 2. – P. 159–179.
36. *Davies D.L., Bouldin D.W.* A cluster separation measure // IEEE transactions on pattern analysis and machine intelligence. – 1979. – **1**, № 2. – P. 224–227.
37. *Frey T., Van Groenewoud H.* A cluster analysis of the D-squared matrix of white spruce stands in Saskatchewan based on the maximum-minimum principle // J. Ecol. – 1972. – **60**, № 3. – P. 873–886.
38. *Ratkowsky D., Lance G.* A criterion for determining the number of groups in a classification // Austral. Comp. J. – 1978. – **10**. – P. 115–117.
39. *Scott A., Symons M.* Clustering methods based on likelihood ratio criteria // Biometrics. – 1971. – **27**, № 2. – P. 387–397.
40. *Marriott F.* Practical problems in a method of cluster analysis // Biometrics. – 1971. – **27**, № 3. – P. 501–514.
41. *Ball G., Hall D.* ISODATA, a novel method of data analysis and pattern classification / Stanford Res. Inst., 1965. – 61 p.
42. *Friedman H., Rubin J.* On some invariant criteria for grouping data // J. Amer. Stat. Ass. – 1967. – **62**, № 320. – P. 1159–1178.
43. *McClain J., Rao V.* Clustisz : A program to test for the quality of clustering of a set of objects // J. Market. Res. – 1975. – **16**. – P. 545–551.
44. *Krzanowski W., Lai Y.* A criterion for determining the number of groups in a data set using sum-of-squares clustering // Biometrics. – 1988. – **44**, № 1. – P. 23–34.
45. *Kaufman L., Rousseeuw P.* Finding groups in data: an introduction to cluster analysis. – Hoboken : John Willey & Sons, 2009. – 368 p.
46. *Tibshirani R., Walther G., Hastie T.* Estimating the number of clusters in a data set via the gap statistic // J. Royal Stat. Soc. Ser. B. – 2001. – **63**, № 2. – P. 411–423.
47. *Dunn J.* Well-separated clusters and optimal fuzzy partitions // J. Cybernet. – 1974. – **4**, № 1. – P. 95–104.
48. *Стецюк П.І., Березовський О.А., Журбенко М.Г., Кропотов Д.О.* Методи негладкої оптимізації у спеціальних задачах класифікації. – К.: 2009. – 28 с. – (Препринт / Інститут кібернетики ім. В.М. Глушкова).
49. *Freije W.A., Castro-Vargas F.E., Fang Z. et al.* Gene expression profiling of gliomas strongly predicts survival // Cancer Res. – 2004. – **64**, № 18. – P. 6503–6510.
50. *Ge X., Yamamoto S., Tsutsumi S. et al.* Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues // Genom. – 2005. – **86**, № 2. – P. 127–141.
51. *Sun L., Hui A.M., Su Q. Et al.* Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain // Cancer Cell. – 2006. – **9**, № 4. – P. 287–300.
52. *Liu T., Papagiannakopoulos T., Puskar K. et al.* Detection of a microRNA signal in an in vivo expression set of mRNAs // PLoS One. – 2007. – **2**, № 8. – P. e804.
53. *Su A.I., Wiltshire T., Batalov S. et al.* A gene atlas of the mouse and human protein-encoding transcriptomes // Proc. Nat. Acad. Sci. USA. – 2004. – **101**, № 16. – P. 6062–6067.
54. *Steinhaus H.* Sur la division des corps materiels en parties // Bull. Acad. Pol. Sci. – 1956. – **4**. – P. 801–804.
55. *Huse J.T., Phillips H.S., Brennan C.W.* Molecular subclassification of diffuse gliomas: seeing order in the chaos // Glia. – 2011. – **59**, № 8. – P. 1190–1199.
56. *Verhaak R.G., Hoadley K.A., Purdom E. et al.* Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1 // Cancer Cell. – 2010. – **17**, № 1. – P. 98–110.
57. *Paugh B.S., Qu C., Jones C. et al.* Integrated molecular genetic profiling of pediatric high-grade gliomas reveals key differences with the adult disease // J. Clin. Oncol. – 2010. – **28**. – P. 3061–3068.
58. *Меклер А.А., Шварц Д.Р.* Выбор переменных для наиболее качественной классификации объектов // Нейроинформатика-2011: Сб. науч. тр. XIII Всерос. науч.-техн. конф. – М., 2010. – Ч. 1. – С. 136–143.

Поступила 16.07.14