На правах рукописи

Ивлиев Александр Евгеньевич

Анализ генных сетей коэкспрессии для изучения транскриптома опухолей мозга и предсказания функций генов

Специальность 03.01.09

Математическая биология, биоинформатика

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата биологических наук

Москва

2011

Работа выполнена на Факультете биоинженерии и биоинформатики
Московского Государственного Университета имени М.В.Ломоносова

Научный руководитель:
доктор химических наук,
Сергеева Марина Глебовна

Официальные оппоненты:
доктор биологических наук,
Карягина-Жулина Анна Станиславовна,
ГУ НИИ эпидемиологии и микробиологии им. Н.Ф. Гамалеи, Москва

доктор физико-математических наук,
Макеев Всеволод Юрьевич,
Институт общей генетики им. Н.И. Вавилова РАН, Москва

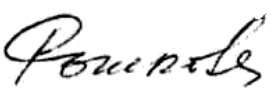
Ведущая организация:
Учреждение Российской академии наук
НИИ биомедицинской химии им. В.Н. Ореховича РАМН, Москва

Защита диссертации состоится 19 декабря 2011 года в 14 часов на заседании
Диссертационного совета Д 002.077.04 при Учреждении Российской академии наук
Институте проблем передачи информации им. А.А. Харкевича РАН
по адресу: 127994, Москва, ГСП-4, Большой Каретный переулок, д. 19, стр. 1.

С диссертацией можно ознакомиться в библиотеке Учреждения Российской академии наук
Институт проблем передачи информации им. А.А. Харкевича РАН.

Автореферат разослан 17 ноября 2011 года.

Ученый секретарь диссертационного совета
доктор биологических наук, профессор
Рожкова Г.И.



Общая характеристика работы

Актуальность темы

Технологические достижения последнего десятилетия сделали возможным исследование живых организмов на уровне генетических последовательностей, экспрессии мРНК и белков в полногеномном масштабе. Важное место в полногеномных исследованиях занимает анализ уровней экспрессии генов. Получаемая при таком анализе информация используется для изучения молекулярных механизмов заболеваний, сравнения типов клеток, поиска функций генов и решения других задач биологии и медицины. Возможность подходить к решению актуальных задач на полногеномном уровне привела к созданию ряда проектов в Европе и США по измерению уровней экспрессии большинства известных генов в тканях человека в норме и при различных заболеваниях. К настоящему времени в открытом доступе имеются массивы данных по многим тысячам разнообразных образцов.

Ключевым инструментом анализа полногеномных данных по экспрессии генов являются генные сети коэкспрессии. Этот метод осуществляет поиск групп (модулей) генов, согласованно экспрессирующихся в эксперименте или наборе клинических образцов. Выделение модулей коэкспрессирующихся генов широко применяется для решения задач двух типов: выявление структуры транскриптомных данных и предсказание функций индивидуальных генов. Первый тип задач, как правило, связан с изучением биологии гетерогенных заболеваний, таких как опухоли. В области изучения рака груди и различных видов лейкемии такие исследования открыли новые возможности для диагностики и разработки подходов химиотерапии. Второй тип задач распространен в фундаментальных исследованиях функции генов и аннотации геномов. В данной работе мы применили генные сети коэкспрессии как инструмент для изучения биологии одного из наиболее гетерогенных групп опухолей – глиальных опухолей мозга (задача первого типа), и предложили новый способ верификации результатов, получаемых в экспрессионных работах по предсказанию функций генов (задачи второго типа).

Актуальность исследования глиальных опухолей мозга (глиом) обусловлена двумя причинами. Во-первых, глиомы относятся к наиболее агрессивным и трудно излечимым видам опухолей. Эффективных методов химиотерапии глиом пока не разработано. Продолжительность жизни пациентов с наиболее распространенным типом глиомы (глиобластомой) составляет в среднем один год. Во-вторых, определение типа глиом в клинической практике основано на гистологических методах, известных своей субъективностью. В связи с этим в клиниках ряда стран активно ведутся работы по изучению биологии глиом на основе транскриптомных данных и поиску мРНК-маркеров для объективной диагностики подтипа глиом. При этом ключевой проблемой является сложность структуры транскриптома глиом: уровни экспрессии ~20 000 генов формируются под действием большого количества разнородных факторов. Это является

препятствием к формированию общего взгляда на молекулярные основы агрессивности и разнообразие экспрессионных классов этих опухолей. Мы предположили, что детальная характеристика структуры транскриптома глиом с помощью генных сетей коэкспрессии позволит сделать новые наблюдения в различных аспектах изучения этих опухолей.

Вторая возможность, которую дают генные сети коэкспрессии, заключается в предсказании функции генов. Поиск функциональной связи генов с клеточными процессами, органеллами, метаболическими и сигнальными путями ведется применительно к широкому спектру живых организмов, включая человека. Ключевой проблемой при этом является верификация экспрессионных предсказаний независимыми методами. В последнее время, благодаря развитию разнообразных (в том числе протеомных) баз данных, появляется возможность верификации предсказаний без проведения направленных экспериментов. Поиск таких подходов может существенно улучшить возможности для верификации. В данной работе мы проверили применимость быстро растущей протеомной базы данных Human Protein Atlas к задаче верификации функциональных предсказаний, сделанных методами генных сетей коэкспрессии.

Для решения этих биологических задач мы также провели методические усовершенствования в нескольких направлениях. Во-первых, большой объем экспрессионных данных, накопленный в электронных базах, требует обеспечения интегрированного доступа к этим базам данных. Во-вторых, по вычислительным причинам анализ генных сетей коэкспрессии трудно реализуем в масштабе всего генома. В связи с этим, на практике исследователи часто используют ограниченные выборки генов, что снижает биологическую ценность анализа. В данной работе мы обратились к решению этих методических проблем.

Цель и задачи исследования

Цель работы – развитие методов анализа экспрессионных данных и их применение для изучения биологии глиом и предсказания функций генов.

<u>В работе были поставлены следующие задачи:</u>

1. Создать доступную через веб-сервер программу, упрощающую процесс поиска и загрузки транскриптомных данных из открытых электронных баз
2. Разработать эвристический метод, позволяющий в короткие сроки проводить анализ коэкспрессии применительно к полному набору генов в геноме (20 000 и более профилей экспрессии)
3. Оценить возможность использования новой крупной протеомной базы данных Human Protein Atlas для верификации функциональных предсказаний, сделанных методом генных сетей коэкспрессии

4. Детально охарактеризовать структуру транскриптома глиальных опухолей мозга методом генных сетей коэкспрессии
5. Применить информацию о структуре транскриптома глиом для развития системы экспрессионной классификации этих опухолей, реконструкции сигнальных путей и поиска потенциальных терапевтических мишеней в глиомах

Научная новизна и практическое значение работы

Впервые детально охарактеризована структура транскриптома глиомы: выделено 20 модулей коэкспрессии, описаны их связи друг с другом и с клиническими характеристиками опухолей. В дополнение к трем известным экспрессионным классам глиомы (мезенхимальному, пролиферативному и пронейральному) показано существование еще одного экспрессионного класса с четкой функциональной интерпретацией – проастроцитарного. Впервые определен список мРНК-маркеров опухолей данного класса: *APOE*, *DAAM2*, *ID4*, *MAP4*, *TJP2* и др. (всего 185 генов). Эти маркеры потенциально могут быть использованы для определения соответствующего класса глиом молекулярными методами (например, ОТ-ПЦР в реальном времени), для которых доступен более высокий уровень стандартизации, чем для принятых в диагностике субъективных гистологических методов.

Предсказано, что в регуляцию одного из ключевых онкогенных сигнальных путей в глиомах, активируемого рецептором эпидермального фактора роста (EGFR), вовлечены белки семейства Sprouty (SPRY1, SPRY2, SPRY4). Этот сигнальный путь известен своей повышенной активностью в наиболее агрессивном типе глиом (глиобластомах). Предсказанный механизм его регуляции важен для понимания биологии этого вида опухолей.

Показано, что существуют статистические закономерности распределения мишеней разрешенных к применению противоопухолевых лекарств в генной сети коэкспрессии в глиоме. По результатам анализа, центральные гены модулей, вовлеченных в патогенез глиом, рекомендованы для дальнейшего изучения в качестве потенциальных новых противоопухолевых мишеней.

На примере изучения эукариотической клеточной органеллы – реснички, показана возможность использования протеомной базы данных Human Protein Atlas для подтверждения экспрессионных предсказаний функций генов. Применение Human Protein Atlas может помочь в задачах предсказания широкого спектра генных функций, которые ассоциированы с неравномерным пространственным распределением соответствующих белков в тканях и клетках человека.

Для 74 генов человека впервые предсказана функциональная связь с клеточной органеллой ресничкой. Согласно результатам анализа данных Human Protein Atlas, около 50% этих экспрессионных предсказаний проходят верификацию на белковом уровне. Идентификация этих

белков, функционально связанных с ресничками, расширяет основу для исследований молекулярных механизмов функционирования этой клеточной органеллы.

Научную новизну и практическую значимость также имеет предложенный в работе эвристический метод, позволяющий многократно ускорить анализ генной коэкспрессии и делающий доступным такой анализ в полногеномном масштабе. Создана программа Microarray Retriever, обеспечивающая интегрированный доступ к существующим экспрессионным базам данных (GEO и ArrayExpress) и упрощающая процесс поиска и загрузки данных.

Апробация работы. Результаты диссертационной работы были представлены на международной конференции Moscow Conference on Computational Biology and Bioinformatics (Москва, 21-24 июля, 2011); на международной конференции 19th International Conference on Intelligent Systems for Molecular Biology & 10th European Conference on Computational Biology (Вена, 17-19 июля, 2011); на международной конференции European Human Genetics Conference 2011 (Амстердам, 28-31 мая, 2011); на XVIII международной научной конференции студентов, аспирантов и молодых учёных «Ломоносов» (Москва, 11-15 апреля, 2011); на I международной научно-практической конференции «Постгеномные методы анализа в биологии, лабораторной и клинической медицине» (Москва, 17-19 ноября, 2010); на XVII международной научной конференции студентов, аспирантов и молодых учёных «Ломоносов» (Москва, 12-15 апреля, 2010); на Всероссийской научной школе для молодежи «Горизонты нанобиотехнологии» (Москва, 12-16 октября, 2009); на XVI международной научной конференции студентов, аспирантов и молодых учёных «Ломоносов» (Москва, 13-18 апреля, 2009).

Публикации. По материалам диссертации опубликовано 13 печатных работ. Из них статей – 3, тезисов устных и стендовых сообщений на конференциях – 10.

Структура и объем работы. Диссертация изложена на 117 страницах, включает 15 таблиц, 20 рисунков, 3 приложения; состоит из введения, обзора литературы, методов, результатов и их обсуждения, выводов и списка литературы, включающего 175 источников.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

I. Решение методических проблем в области работы с экспрессионными данными

Накопление огромных объемов экспрессионных данных в открытых базах сделало возможным решение новых типов задач и обобщение результатов из разных экспериментов. Однако для более эффективного использования этого потенциала требуются методические усовершенствования в области поиска данных и их полногеномного анализа. Такие усовершенствования были проведены в данном разделе работы.

1) Создание программы поиска и загрузки данных. Первым этапом в работе по изучению экспрессии генов на основе ранее опубликованных данных является поиск данных в базе и их

загрузка на локальный компьютер. В связи с этим, актуальна задача обеспечения интегрированного доступа к основным мировым экспрессионным базам данных (GEO и ArrayExpress), который бы позволил осуществлять поиск экспериментов параллельно в обеих базах и загрузку экспериментов группами единовременно.

Для решения этих задач мы создали программу Microarray Retriever (MaRe). Эта программа написана на языке Perl и работает на Веб-сервере с операционной системой UNIX. Программа запускается пользователем через браузер и доступна по ссылке http://www.lgtc.nl/MaRe/. MaRe позволяет осуществлять поиск по наиболее распространенным в области видам запросов: инвентарные номера записей в экспрессионных базах данных, авторы экспериментов, ключевые слова, виды организмов, дата представления данных в базу, характеристики платформы и др. После осуществления поиска пользователь имеет возможность выбрать интересующие эксперименты и загрузить их в виде архива единовременно.

Одновременный доступ к двум базам дает преимущество перед поиском в каждой базе по отдельности. Например, по запросу «ключевое слово: cerebellum» на сайте GEO можно найти 237 экспериментов. По тому же запросу на сайте ArrayExpress можно найти 178 экспериментов, однако сколько из них совпадает между базами не является очевидным. Поиск в двух базах с помощью MaRe находит 237 экспериментов в базе GEO и 28 дополнительных экспериментов в базе ArrayExpress, которые не были найдены в базе GEO. Таким образом, программа автоматизирует процесс сопоставления результатов поиска в двух базах данных.

Дополнительной функцией MaRe является возможность поиска экспрессионных данных через соответствующие публикации в базе PubMed. В этом режиме, программа находит публикации, аннотации которых содержат заданные ключевые слова, и затем – список экспериментов, соответствующих этим публикациям, в экспрессионных базах. Такой поиск помогает найти дополнительные эксперименты, описания которых в экспрессионных базах не содержат заданных ключевых слов.

Таким образом, созданная программа Microarray Retriever дает следующие преимущества по сравнению с использованием сайтов GEO и ArrayExpress:

- интегрированный поиск в двух базах данных;
- поиск экспериментальных данных через ключевые слова в абстрактах PubMed;
- удобный интерфейс для загрузки данных группами.

2) Разработка метода полногеномного анализа коэкспрессии. Построение генной сети коэкспрессии требует измерения корреляций между профилями экспрессии для всех возможных пар генов, входящих в состав сети. Объем оперативной памяти и время, которые необходимы для выполнения этой операции, зависят квадратично от количества анализируемых генов. Например, для построения сети, состоящей из 4 000 генов, широко применяющимся методом WGCNA (от

англ. «Weighted Gene Coexpression Network Analysis») требуется около 60 Мб оперативной памяти и около 5 мин времени, а для включения в сеть всех транскриптов, уровни экспрессии которых измеряются ДНК-микрочипами (~47 000 транскриптов в случае распространенной модели микрочипов Affymetrix U133 Plus 2.0), требуется ~8.5 Гб оперативной памяти и около 4 суток. В связи с этим, при анализе генных сетей коэкспрессии на практике часто ограничиваются сравнительно небольшим количеством генов (4 - 5 тысяч), что снижает биологическую ценность анализа.

Мы разработали эвристический метод, который делает доступным анализ коэкспрессии генов в полногеномном масштабе. Анализ состоит из следующих этапов. (1) На основе случайной выборки генов из генома (нескольких тысяч генов) проводится построение сети коэкспрессии, и в сети выделяются модули. (2) Для каждого модуля вычисляется его характеристический профиль экспрессии путем усреднения профилей входящих в его состав генов. (3) Каждый ген из генома приписывается тому модулю, характеристический профиль экспрессии которого сильнее других коррелирует с индивидуальным профилем экспрессии гена. Таким образом, генный состав модулей, найденных с помощью сети, определяется заново, но уже в полногеномном масштабе. Продолжительность вычислений и объем оперативной памяти на третьем этапе зависят линейно от количества генов.

Оценка качества работы метода проводилась на крупном массиве данных, включающем в себя профили экспрессии 18 000 генов в приблизительно 100 разных тканях человека (массив GSE7307 из базы данных GEO). Во-первых, напрямую проведено сравнение генного состава модулей в сетях, состоящих из 4 000 генов (размер доступный для анализа с помощью метода генных сетей коэкспрессии), с генным составом модулей, определенных на тех же наборах генов (размером 4 000) путем укрупнения модулей-предшественников из сетей меньшего размера с помощью эвристического метода. Метод показал высокую чувствительность (70-80%) и точность (80-90%). Во-вторых, проведена проверка того, как соотносятся функциональные свойства модулей в сетях небольшого размера (2 000 генов) с функциональными свойствами полногеномных аналогов этих модулей, полученных с помощью эвристического метода. Для этого построено 20 сетей коэкспрессии на случайных выборках размером в 2 000 генов. Для каждой исходной сети, определено, какие биологические процессы ассоциированы с модулями (анализ обогащения на основе генных списков из базы Gene Ontology, точный тест Фишера), и рассмотрено, как изменяется средний уровень статистической значимости этих ассоциаций при увеличении размера модулей с помощью эвристического метода (рис. 1A). Оказалось, что статистическая значимость возрастает (рис. 1A), т.е. в модули включаются гены преимущественно тех же функций, которые были ассоциированы с модулями в исходной сети. Эти результаты подтверждают точность работы метода.

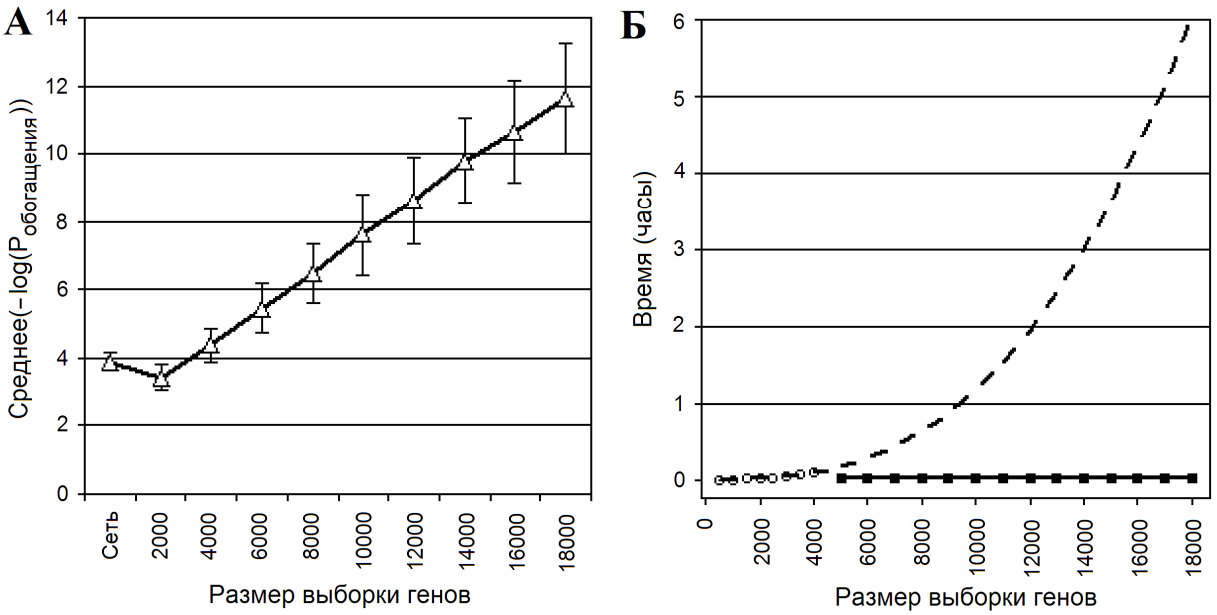


Рис. 1. Результаты тестирования эвристического метода анализа коэкспрессии генов. А. Биологические процессы из базы Gene Ontology, ассоциированные с модулями в исходной сети согласно анализу обогащения, сохраняют свою связь с модулями при перерасчете модулей эвристическим методом. Статистическая значимость обогащения возрастает по мере увеличения размера модулей с помощью эвристического метода. Ось абсцисс - количество анализируемых генов: «сеть» - исходные модули в сети, состоящей из 2000 случайно выбранных генов; «2000» - модули, воспроизведенные с помощью эвристического метода на той же выборке из 2000 генов; «4000» - модули увеличенного размера, полученные эвристическим методом на выборке из 4000 генов (2000 исходных генов + 2000 новых выбранных случайным образом); и т.д. Ось ординат - среднее по модулям и биологическим процессам значение отрицательного десятичного логарифма от Р-значения, оценивающего обогащение модуля генами биологического процесса, с которым был ассоциирован исходный модуль в сети (точный тест Фишера). Б. Зависимость времени вычислений от размера выборки генов. Прозрачные кружки - время работы стандартного метода на выборках генов размером не более 4000. Пунктирная линия - оценка времени работы стандартного метода на более крупных выборках генов. Квадраты, соединенные сплошной линией - время работы эвристического метода.

Чтобы оценить выигрыш во времени, который дает эвристический метод, сначала мы определили время, затрачиваемое стандартным методом для выделения модулей в генных сетях коэкспрессии размером от 500 до 4 000 генов с шагом в 500 генов (рис. 1Б). Время анализа резко возрастало от 2 секунд (сеть размером 500 генов) до ~5 минут (сеть размером 4 000 генов). Путем экстраполяции кривой роста времени, было вычислено, что анализ сети полногеномного масштаба (~18 000 генов) при наличии достаточного ресурса оперативной памяти займет около 6 часов (рис. 1Б). Далее, мы определили модули в полногеномном масштабе (18 000 генов) с помощью эвристического метода, используя в качестве исходной сеть размером 2 000 генов. С учетом времени построения исходной сети, получение полногеномных модулей заняло приблизительно 2 минуты (рис. 1Б). Таким образом, предложенный метод многократно ускоряет анализ и снижает требования к объему оперативной памяти. Это делает доступным анализ коэкспрессии генов в полногеномном масштабе.

II. Проверка возможности верификации экспрессионных предсказаний генных функций с помощью протеомной базы Human Protein Atlas

Ключевой проблемой в работах по предсказанию функций генов методами биоинформатики является верификация этих предсказаний с помощью независимых экспериментальных методов. В тоже время, развитие разнообразных полногеномных ресурсов открытого доступа дает новые возможности для верификации предсказаний компьютерным путем на основе анализа независимых типов данных. Принципиально новым типом крупномасштабных данных являются иммуногистохимические данные в базе Human Protein Atlas. Эта база содержит информацию из развивающегося проекта, цель которого заключается в характеристике локализации белков человека в широком спектре тканей и клеточных типов. База версии 8.0 содержит иммуногистохимические данные для ~11 000 белков, полученные с помощью антител, специфичных к этим белкам, и характеризующие локализацию этих белков в 46 тканях человека. Этот ресурс широко используется для изучения белков. Мы предположили, что он также может служить эффективным инструментом верификации транскриптомных предсказаний. Проверка этого предположения проведена на примере поиска генов с конкретной выбранной функцией.

Для исследований была выбрана клеточная органелла ресничка. Выбор объекта обусловлен: 1) легкостью распознания этой органеллы на иммуногистохимических изображениях (реснички образуют массивные пучки на апикальной поверхности мерцательных клеток в эпителии); 2) актуальностью исследований по поиску генов, вовлеченных в функционирование реснички. В настоящее время показано, что такие заболевания человека как почечный поликистоз, гидроцефалия, первичная цилиарная дискинезия и другие вызваны дисфункцией этой органеллы. Молекулярная биология реснички активно изучается и одним из ее направлений является поиск новых генов, функционально связанных с указанной органеллой. В данном исследовании, в качестве анализируемых тканей были выбраны мозг, дыхательные пути и фаллопиевы трубы, поскольку именно эти ткани содержат наибольшее количество несущих реснички клеток.

12 массивов экспрессионных данных были загружены с помощью программы Microarray Retriever. В каждом массиве данных построена сеть коэкспрессии (4 000 генов) и в ней выделены модули коэкспрессии: от 11 до 46, в зависимости от массива. Генный состав модулей расширен до полногеномного масштаба предложенным выше методом. В 10-ти массивах обнаружен «мерцательный» модуль, обогащенный генами-маркерами ресничек, известными из литературы ($P < 0.001$, точный тест Фишера). Далее, определен консенсусный генный состав этих модулей (статистическая значимость: $FDR < 0.5\%$). Полученный консенсусный модуль состоял из 371 гена. С помощью специализированной базы данных CilDB (обобщает литературу в области изучения протеома ресничек) мы разделили гены в консенсусном модуле на 3 категории по новизне: (I) 237 генов, функциональная связь которых с ресничками уже известна; (II) 60 генов, предсказанных в литературе с низким уровнем достоверности; (III) новые предсказания – 74 гена. Высокое

содержание генов I категории в модуле указывает на повышенную вероятность наличия функциональной связи с ресничками у остальных генов модуля по сравнению со случайной выборкой генов из генома.

На основе иммуногистохимических данных из Human Protein Atlas была изучена локализация белков категории I в ткани фаллопиевых труб и дыхательных путей. База содержала информацию для 136 из 237 белков этой категории. Оказалось, что 76 из них (56%) локализованы специфически в ресничках на субклеточном уровне (на рис. 2А приведен пример иммуногистохимического изображения для белка из категории I с такой локализацией). Еще 34 белка (25%) хотя и локализованы в других субклеточных областях мерцательных клеток, но преимущественно экспрессируются в мерцательных клетках по сравнению с другими клеточными типами. Таким образом, для 81% белков наблюдается специфичная связь с мерцательными клетками. Поскольку эти белки относятся к категории I (функциональная связь с органеллой ресничкой известна), это показывает, что данные из Human Protein Atlas согласуются с уже существующими знаниями об этой органелле.

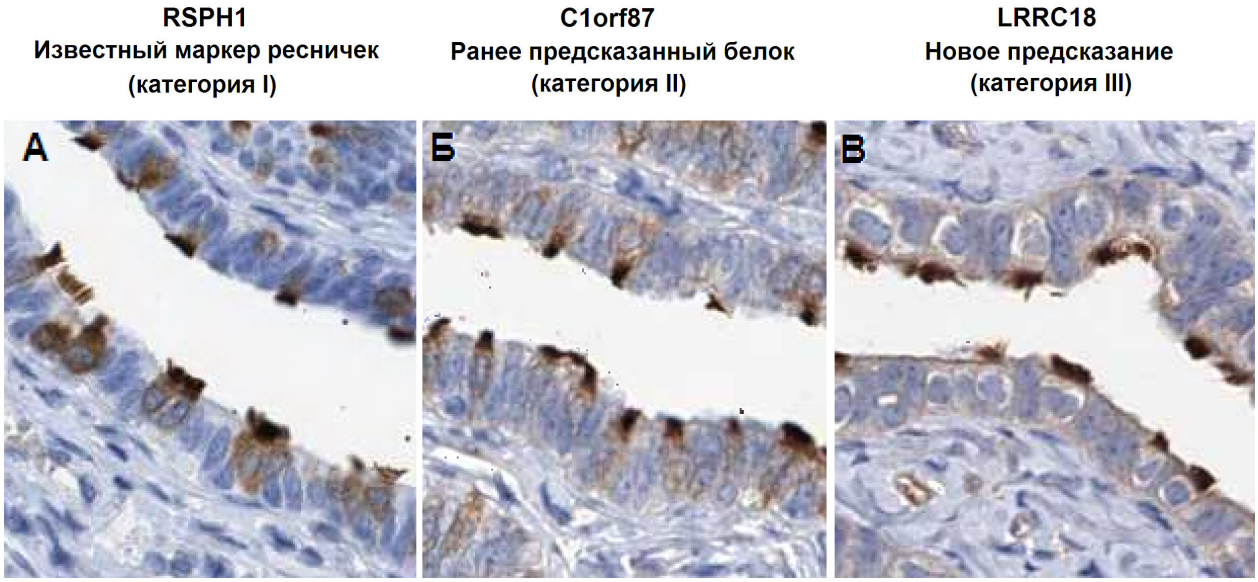


Рис. 2. Найденные белки из разных категорий новизны имеют сходную локализацию в ткани фаллопиевых труб. Мерцательные клетки регулярно перемежаются с эпителиальными клетками без ресничек, что является характерным свойством фаллопиевых труб. Черный цвет соответствует антителам, специфичным к соответствующему белку. На всех изображениях видна преимущественная локализация белков в ресничках мерцательных клеток. Полный набор изображений доступен в базе Human Protein Atlas.

Далее, мы изучили данные для белков из категорий II и III. На рис. 2Б и 2В приведены примеры иммуногистохимических изображений для белков из этих категорий. Всего база содержала информацию для 34 белков из категории II и 48 белков из категории III. 21 белок (62%, категория II) и 25 белков (52%, категория III) оказались локализованы преимущественно в мерцательных клетках (и часть из этих белков – непосредственно в ресничках на субклеточном уровне). Это указывает на функциональную связь указанных белков с изучаемой органеллой.

Таким образом, протеомные данные служат независимым аргументом в пользу правильности соответствующих экспрессионных предсказаний.

Интересно, что некоторые их белков категории III, успешно прошедших верификацию (например, C11orf63 и C1orf129), не имеют никаких аннотированных функций в геномных базах данных и не охарактеризованы в литературе. Для таких белков предсказание функциональной связи с органеллой ресничкой может служить первичной функциональной аннотацией.

Таким образом, протеомная база Human Protein Atlas является эффективным инструментом верификации транскриптомных предсказаний. Применение Human Protein Atlas может помочь в задачах предсказания широкого спектра генных функций, которые ассоциированы с неравномерным пространственным распределением соответствующих белков в тканях и клетках человека.

III. Изучение биологии глиом методами генных сетей коэкспрессии

Глиомы являются трудно излечимым и гетерогенным типом опухолей. С целью изучения молекулярных свойств глиом и поиска мРНК-маркеров для определения подтипов глиом при диагностике, ведутся работы по изучению экспрессии генов в этих опухолях. Однако сложность транскриптомных данных (уровни экспрессии ~20 000 генов, находящиеся под влиянием большого количества биологических факторов) является препятствием к формированию общего взгляда на экспрессию генов в этих опухолях. Мы предположили, что генные сети коэкспрессии позволят структурно охарактеризовать транскриптом глиальных опухолей и сделать новые наблюдения в различных аспектах изучения этих опухолей.

1) Общая характеристика структуры транскриптома глиом. Чтобы детально охарактеризовать структуру транскриптома глиальных опухолей, мы проанализировали 5 наиболее крупных массивов данных (всего 790 пациентов) с помощью генных сетей коэкспрессии. Использованные массивы описаны в таблице 1.

К массиву данных GSE16011 (выборка, содержащая 4 000 генов) применен метод анализа генных сетей коэкспрессии WGCNA. Выбор метода продиктован тем, что WGCNA сглаживает шум в сетях коэкспрессии и более точно определяет границы между модулями, чем стандартные методы. Согласно алгоритму WGCNA, сначала была построена сеть коэффициентов корреляции Пирсона между профилями экспрессии генов. Затем на ее основе построена сглаженная сеть, в которой вес ребра между двумя генами тем выше, чем выше сходство наборов соседей этих генов в исходной сети. Путем иерархической кластеризации сглаженной сети и анализа дерева кластеризации алгоритмом Dynamic Tree Cut выделено 20 модулей коэкспрессии. Далее, генный состав этих модулей расширен до полногеномного уровня с помощью предложенного выше эвристического метода.

Таблица 1. Характеристика использованных массивов данных по экспрессии генов в глиомах

№	Идентификатор массива[1]	Задача	Кол-во образцов	Типы глиом[2]	Платформа[3]
1	GSE16011	Основной анализ	276	ГБ, АА, А, АО, О	U133 plus 2.0
2	Rembrandt_HF	Верификация результатов	159	ГБ, АА, А, АО, О	U133 plus 2.0
3	Rembrandt_O	Верификация результатов	183	ГБ, АА, А, АО, О	U133 plus 2.0
4	GSE4271	Верификация результатов	98	ГБ, АА	U133A
5	GSE4412	Верификация результатов	74	ГБ, АА, АО	U133A

1 - массивы данных GSE16011, GSE4271 и GSE4412 загружены из базы GEO; массивы Rembrandt_HF и Rembrandt_O получены из базы данных Rembrandt (от англ. «Repository of Molecular Brain Neoplasia Data»), которая специализируется на опухолях мозга. Массив Rembrandt_HF (от англ. «Henry Ford») состоит из образцов, полученных в больнице им. Генри Форда, Детройт, США; массив Rembrandt_O (от англ. «other») состоит из образцов, полученных в других больницах в рамках проекта GMDI. 2 - гистологические типы глиомы: «ГБ» - глиобластома, «АА» - анапластическая астроцитома, «А» - астроцитома низких стадий, «АО» - анапластическая олигодендроглиома, «О» - олигодендроглиома низких стадий. 3 - модели олигонуклеотидных микрочипов фирмы Affymetrix.

Мы проверили воспроизводимость модулей коэкспрессии в 4-х независимых массивах данных (табл. 1). Все модули оказались воспроизводимыми в каждом из независимых массивов. Для каждого модуля далее был составлен консенсусный список генов: условием включения гена в консенсусный список была принадлежность гена модулю хотя бы в 3-х из 5-ти массивов данных (FDR < 2%). Чтобы функционально аннотировать консенсусные модули, генный состав модулей был проанализирован с помощью программы DAVID на основе трех источников информации: (1) данные из базы Gene Ontology, (2) сведения о локализации генов на хромосомах и (3) списки генов из предшествующих экспрессионных работ. Модули оказались связанными с широким спектром биологических процессов и структур в глиомах (табл. 2). Поскольку консенсусные модули систематически характеризуют соответствующие процессы и структуры на транскрипционном уровне, они представляют интерес для изучения биологии глиом. Это первая в области изучения глиальных опухолей обширная коллекция экспрессионных генных списков, основанная более, чем на одном массиве данных.

Для того, чтобы охарактеризовать связь уровней экспрессии модулей с клиническими характеристиками опухолей, для каждого модуля мы вычислили его характеристический профиль экспрессии. Далее, для каждого модуля проведено сравнение уровня экспрессии между гистологическими подтипами опухолей (глиобластома, астроцитома, олигодендроглиома) и получена оценка корреляции уровня экспрессии модуля со стадиями заболевания и продолжительностью жизни пациентов. Результаты представлены в таблице 3 и показывают связь транскриптома с клиническими характеристиками болезни.

Таблица 2. Биологическая аннотация модулей

Модуль	Списки генов, которыми обогащен модуль*	Аннотация модуля
Типично раковые процессы		
M11	Митоз (10^{-55})	Пролиферация
M4	Гликолиз (10^{-4})	Ответ на гипоксию
M1	Развитие сосудов (10^{-7})	Капилляры
M6	Синтез внеклеточного матрикса (10^{-16})	Более крупные сосуды
M7	Иммунный ответ (10^{-48})	Иммунный ответ
M3	Регуляция протеин киназ (10^{-9})	Регуляция активности киназ
M2	Противовирусный ответ (10^{-18})	ИФН-зависимые гены
Типы дифференцировки глиомы		
M8	Мезенхимальные маркеры (10^{-29})	Мезенхимальная дифференцировка
M15	Маркеры астроцитов (10^{-26})	Проастроцитарная дифференцировка
M20	Маркеры нейрогенеза (10^{-6})	Пронейральная дифференцировка
Хромосомные аберрации		
M5	Хромосома 19 (10^{-49})	Делеция локуса 19q
M9	Хромосома 1 (10^{-89})	Делеция локуса 1p
M10	Локус 12q13 (10^{-8})	Амплификация локуса 12q13
M16	Хромосома 10 (10^{-19})	Делеция локуса 10q
Нормальные функции мозга		
M13	Маркеры нейронов (10^{-95})	Нормальные нейроны
M19	Маркеры нейронов (10^{-44})	Нейроны и пронейральная дифф.
M14	Образование миелина (10^{-5})	Белое мозговое вещество
Другие		
M12	Ядро (10^{-12})	Ядро
M17	Ядро (10^{-16})	Ядро
M18	Синтез белка (10^{-28})	Синтез белка

* - в скобках приведены P-значения точного теста Фишера

Известно, что различные биологические процессы, например, ангиогенез и ответ клеток на гипоксию, взаимосвязаны в опухолях. Поэтому мы предположили, что профили экспрессии некоторых модулей могут коррелировать, образуя над-модульную структуру. Чтобы охарактеризовать эту структуру, мы кластеризовали модули и пациентов в каждом массиве данных. Результаты были сходными между массивами (пример для массива GSE16011 представлен на рис. 3). Кластеризация выявила две основные группы модулей: A и E (рис. 3). Группа A преимущественно содержала модули, связанные с опухолевой прогрессией. Их повышенный уровень экспрессии наблюдался в опухолях с плохим прогнозом – глиобластомах (табл. 3). Группа E преимущественно содержала модули, экспрессирующиеся в опухолях с благоприятным прогнозом – олигодендроглиомах (табл. 3). Яркая выраженность этих двух групп модулей подчеркивает принципиальный характер различий между глиобластомой и олигодендроглиомой.

Таким образом, выделение 20 воспроизводимых модулей коэкспрессии и подробная характеристика их связи друг с другом, а также с внешними клиническими показателями, дают общий взгляд на структуру транскриптома глиомы. Это важно для дальнейших работ по изучению патогенеза и молекулярного разнообразия глиом с помощью транскриптомных данных. Кроме

того, получен набор из 20 консенсусных экспрессионных генных списков, которые характеризуют широкий спектр биологических процессов в глиоме.

Таблица 3. Связь уровней экспрессии модулей с клиническими характеристиками пациентов

Модуль	Группа	Аннотация	Стадия†	Гистология‡	Время§	COX*
M1	A	Капилляры	0.37	ГБ	↓	1.7×10^{-2}
M2		ИФН-зависимые гены	0.39	ГБ	↓	1.3×10^{-7}
M3		Регуляция акт. киназ	0.44	ГБ	↓	$< 10^{-16}$
M4		Ответ на гипоксию	0.59	ГБ	↓	1.8×10^{-14}
M5		Делеция 19q	0.48	ГБ	↓	6.7×10^{-9}
M6		Кровеносные сосуды	0.52	ГБ	↓	1.8×10^{-6}
M7		Иммунный ответ	0.38	ГБ	↓	2.6×10^{-7}
M8		Мезенхимальная дифф.	0.58	ГБ	↓	$< 10^{-16}$
M9		Делеция 1p	0.52	ГБ	↓	6.6×10^{-16}
M10	B	Амплификация 12q13	0.25	ГБ	↓	3.7×10^{-3}
M11		Пролиферация	0.40	ГБ	↓	2.1×10^{-3}
M12		Ядро	-	-	-	-
M13	C	Нормальные нейроны	-0.28	«Норма»	-	-
M14		Белое вещество	-	«Норма»	-	-
M15	D	Проастроцитарная дифф.	-0.34	A, «Норма»	↑	1.8×10^{-4}
M16	E	Делеция 10q	-0.59	A, O, «Норма»	↑	$< 10^{-16}$
M17		Ядро	-0.42	O	↑	1.7×10^{-10}
M18		Синтез белка	-0.40	O	↑	2.0×10^{-12}
M19		Нейроны и пронейр. дифф.	-0.54	O, «Норма»	↑	2.0×10^{-11}
M20		Пронейральная дифф.	-0.57	O	↑	6.0×10^{-14}

Результаты получены на массиве данных GSE16011. Обоснование объединения модулей в группы приведено на рисунке 3. † – коэффициент корреляции Спирмана между характеристическим профилем экспрессии модуля и стадиями опухолей. ‡ – Гистологический тип глиомы, в котором уровень экспрессии модуля повышен ($p < 0.05$, тест Уилкоксона; «ГБ» - глиобластома, «А» - астроцитома, «О» - олигодендроглиома, «Норма» - образцы опухоли, содержащие примесь нормальных клеток нервной ткани). § – Связь уровней экспрессии модуля с продолжительностью жизни пациентов согласно анализу регрессии Кокса с одной переменной («↓» - модуль активирован у пациентов с плохим прогнозом; «↑» - с благоприятным прогнозом). * - P-значение регрессионного анализа связи модуля с продолжительностью жизни. Символ «-» означает отсутствие статистической значимости.

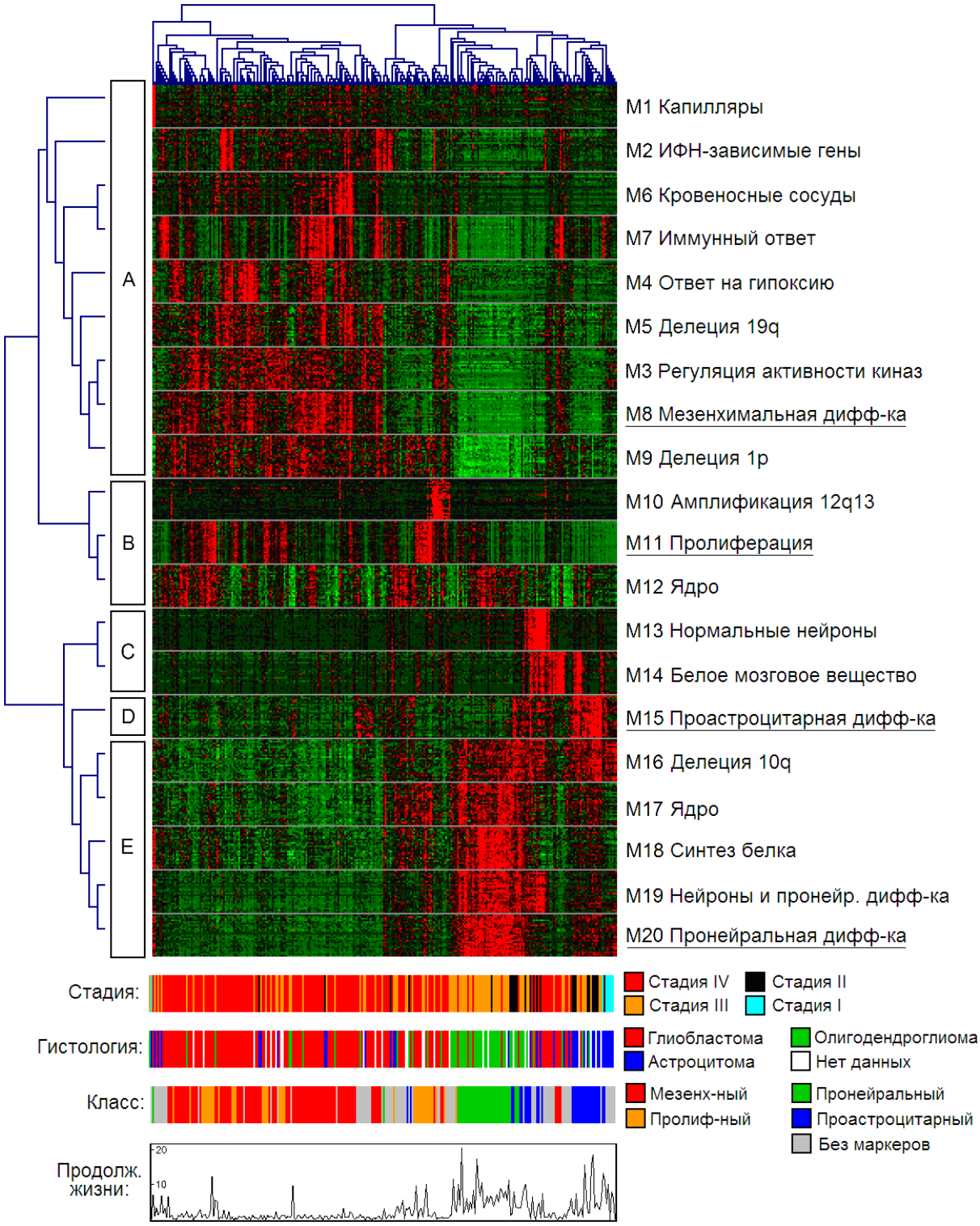


Рис. 3. Структура транскриптома глиомы. Столбцы соответствуют опухолям (276 образцов), строки – генам (приведено по 25 центральных генов из каждого модуля). Высокий уровень экспрессии генов обозначен красным цветом, низкий – зеленым. Модули кластеризованы на основе корреляций между их характеристическими профилями экспрессии; результат кластеризации модулей представлен слева в виде дерева модулей (также отмечены группы модулей A, B, C, D и E). Аннотации модулей приведены справа. Подчеркнуты названия модулей, уровни экспрессии которых использованы далее для разделения опухолей на классы. Опухоли кластеризованы на основе уровней экспрессии генов, представленных на диаграмме. Снизу указаны клинические характеристики (стадия опухоли, гистологический тип опухоли, продолжительность жизни пациента), а также принадлежность опухолей к экспрессионным классам согласно результатам классификации опухолей с помощью центроидов в нашей работе. Диаграмма получена на массиве данных GSE16011.

2) Обнаружение проастроцитарного класса глиом на основе профилей генной экспрессии. Важной проблемой при лечении глиом является субъективность гистологических методов, на которых основано определение типа опухоли при диагностике. Одна и та же опухоль, по мнению двух специалистов, может относиться к разным диагностическим типам. С целью решения этой проблемы, активно ведется разработка системы классификации глиом на основе профилей генной экспрессии. Эта система потенциально может быть использована для объективной диагностики с помощью мРНК-маркеров. Однако трудностью является установление полного списка существующих экспрессионных классов глиомы, что связано со сложностью структуры транскриптома этих опухолей.

В настоящее время установлено существование трех экспрессионных классов глиомы: (1) мезенхимального, (2) пролиферативного и (3) пронейрального. Маркерами соответствующих классов в нашей работе являются модули M8, M11 и M20 (согласно результатам сравнения генных списков этих модулей с ранее опубликованными). Результаты кластеризации образцов (рис. 3) показали, что существует группа опухолей, которые не характеризуются высоким уровнем экспрессии ни одного из этих известных модулей-маркеров, однако образуют отдельный кластер на диаграмме (отличающийся от других кластеров повышенным уровнем экспрессии модуля M15). Это указывает на существование ранее не установленного экспрессионного класса опухолей.

Модуль M15 аннотирован нами как модуль проастроцитарной дифференцировки (табл. 2). В литературе не описаны модули, ассоциированные с такой функцией или имеющие сходный генный состав. Мы оценили свойства опухолей, отличительным свойством которых является активация проастроцитарного модуля M15, в контексте уже известных экспрессионных классов глиомы. Для этого мы кластеризовали образцы на основе профилей экспрессии центральных генов только из 4-х модулей: проастроцитарного (M15), мезенхимального (M8), пролиферативного (M11) и пронейрального (M20). Это позволило выделить 5 классов опухолей: проастроцитарный, мезенхимальный, пролиферативный, пронейральный, а также «класс без маркеров» (в котором все четыре модуля экспрессировались на низком уровне). Наконец, для каждого класса был вычислен характеристический профиль экспрессии (центроид). С помощью центроидов опухоли были классифицированы во всех 5 массивах данных.

Мы сравнили продолжительность жизни пациентов в полученных классах. Мезенхимальный и пролиферативный классы были ассоциированы с низкой продолжительностью жизни, а пронейральный класс – с высокой (рис. 4). Это согласуется с литературными данными. Проастроцитарный класс оказался воспроизводимо ассоциирован с высокой продолжительностью жизни (рис. 4). Наличие у проастроцитарных глиом неслучайных клинических свойств подтверждает правомерность их объединения в группу.

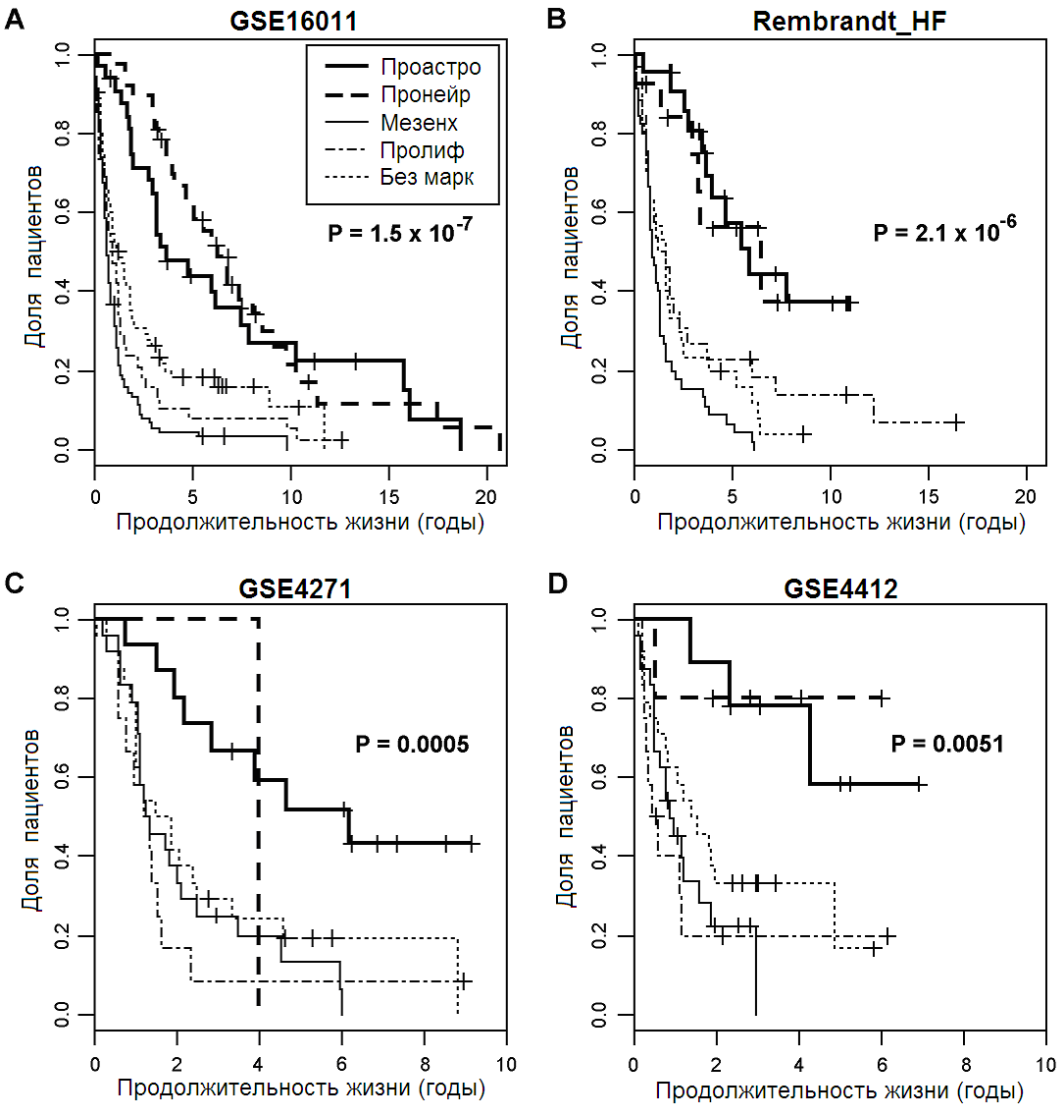


Рис. 4. Проастроцитарный модуль является маркером опухолей с высокой продолжительностью жизни. Представлены диаграммы Каплана-Мейера продолжительности жизни пациентов в 5 молекулярных классах глиомы для следующих независимых массивов данных: **(A)** GSE16011, **(B)** Rembrandt_HF, **(C)** GSE4271, **(D)** GSE4412. Мезенхимальный, пролиферативный, и класс «без маркеров» характеризуются низкой средней продолжительностью жизни; проастроцитарный и пронейральный классы - высокой. P-значения оценивают статистическую значимость различий в продолжительности жизни между проастроцитарным классом и объединением трех классов с неблагоприятным прогнозом (логарифмический ранговый тест). Массив данных Rembrandt_O исключен из анализа в связи с отсутствием данных по продолжительности жизни пациентов.

Таким образом, в дополнение к трем известным экспрессионным классам глиомы нами показано существование еще одного экспрессионного класса с четкой функциональной интерпретацией – проастроцитарного. Мы определили список мРНК-маркеров опухолей данного класса: мРНК генов *APOE*, *DAAM2*, *ID4*, *MAP4*, *TJP2* и др. (всего 185 генов). Найденные проастроцитарные маркеры потенциально могут быть использованы для определения типа глиом молекулярными методами (например, ОТ-ПЦР в реальном времени), для которых доступен более высокий уровень стандартизации, чем для принятых в диагностике субъективных гистологических методов.

3) Предсказание участия белков Sprouty в регуляции онкогенного сигнального пути EGFR в глиомах. Изучение молекулярных механизмов регуляции сигнальных путей, вовлеченных в канцерогенез, является одним из основных направлений современных исследований в онкологии. Понимание таких механизмов важно для разработки методов направленной химиотерапии и изучения механизмов развития лекарственной устойчивости. Известно, что сигнальный путь рецептора эпидермального фактора роста (EGFR) – один из ключевых онкогенных путей в глиомах. Этот путь регулирует способность клеток к пролиферации, миграции и выживанию благодаря передаче сигналов через каскад митоген-активируемых протеин киназ (MAPK). Существенный вклад в активацию пути EGFR в глиомах вносят такие механизмы, как амплификация гена *EGFR* в геноме и повышенное содержание в клетках его мРНК. Хотя сигнальный путь EGFR активно изучается, понимание регуляторных механизмов, контролирующих этот путь, пока не достигнуто.

Для выявления генов потенциально связанных с активностью пути EGFR, мы оценили, как профили экспрессии модулей коррелируют с наличием амплификации гена *EGFR* в геноме и уровнем экспрессии *EGFR* (табл. 4). Оказалось, что модуль M3 активирован в опухолях с амплификациями гена *EGFR* сильнее, чем любой другой из модулей в транскриптоме глиомы ($P = 4.6\times10^{-12}$, тест Уилкоксона). Кроме того, этот модуль превосходил другие по корреляции своего характеристического профиля экспрессии с профилем экспрессии *EGFR* (коэффициент корреляции Пирсона = 0.37, $P = 1.3\times10^{-10}$) (табл. 4). Это указывает на связь модуля M3 с активностью сигнального пути EGFR.

Модуль M3 ранее не был описан в литературе. Анализ генного состава M3 показал, что этот модуль обогащен генами, вовлеченными в регуляцию каскада митоген-активируемых протеин киназ ($P = 1.8\times10^{-4}$). В частности, в состав модуля входили гены семейства Sprouty (*SPRY1*, *SPRY2*, *SPRY4*, *SPRED1*, *SPRED2*). В геноме человека содержится 6 генов этого семейства, и 5 из них принадлежали консенсусному модулю M3 (статистическая значимость обогащения $P = 2.8\times10^{-8}$). Известно, что белки Sprouty регулируют активность сигнального пути EGFR в некоторых нормальных (фибробласты) и опухолевых (меланома, рак груди) типах клеток человека. Однако, данные о вовлеченности белков Sprouty в регуляцию сигнальных путей в глиоме отсутствуют.

Наши результаты дают несколько указаний на вовлеченность генов Sprouty в биологию глиом. Во-первых, повышенный уровень экспрессии этих генов наблюдается у пациентов с низкой продолжительностью жизни. Во-вторых, гены Sprouty коэкспрессируются друг с другом, что указывает на их согласованное функционирование в опухолях. В-третьих, модуль M3, содержащий гены Sprouty, ассоциирован с нарушениями в гене EGFR (табл. 4). Результаты проведенного нами анализа позволяют предположить вовлеченность белков Sprouty в регуляцию ключевого онкогенного сигнального пути EGFR в глиомах.

Эту потенциальную регуляторную связь следует учитывать при изучении ответа клеток глиомы на ингибиторы белка EGFR (гефитиниб, ерлотиниб и др), которые тестируются в настоящее время как противоопухолевые препараты. Полученный результат также указывает на возможность использования аналогичных исследований транскриптомных данных для реконструкции сигнальных путей в опухолях.

Таблица 4. Связь уровней экспрессии модулей в опухолях с наличием амплификации *EGFR* в геноме и уровнем экспрессии этого гена

Модуль	Аннотация	(А) Уровень экспрессии модулей в опухолях с амплификацией гена EGFR			(Б) Корреляция экспрессии модулей с уровнем экспрессии EGFR	
		Высокий/ низкий	P *	Крат-ность †	ККП ‡	P §
M1	Капилляры	↑	6,0E-02	0,99	-0,03	6,4E-01
M2	ИФН-зависимые гены	↑	7,3E-06	1,41	0,24	4,4E-05
M3	Регуляция активности киназ	↑	4,6E-12	1,86	0,38	7,5E-11
M4	Ответ на гипоксию	↑	5,2E-04	1,4	0,03	6,7E-01
M5	Делеция 19q	↑	6,9E-06	1,23	0,16	6,0E-03
M6	Кровеносные сосуды	↑	9,5E-03	1,09	-0,06	3,3E-01
M7	Immune response	-	2,1E-01	1,06	-0,12	5,0E-02
M8	Мезенхимальная дифф.	↑	6,0E-06	1,29	0,07	2,5E-01
M9	Делеция 1p	↑	2,9E-05	1,21	0,16	6,0E-03
M10	Амплификация 12q13	-	5,1E-01	1,39	0,03	6,6E-01
M11	Пролиферация	↑	1,5E-02	1,07	0,14	1,8E-02
M12	Ядро	-	3,5E-01	0,95	0,12	5,4E-02
M13	Нейроны	↓	7,9E-03	0,69	-0,10	9,7E-02
M14	Белое мозговое вещество	-	4,0E-01	1,12	-0,11	7,7E-02
M15	Проастроцитарная дифф.	-	6,9E-01	0,88	-0,10	1,1E-01
M16	Делеция 10q	↓	4,7E-07	0,75	-0,19	1,0E-03
M17	Ядро	↓	1,1E-05	0,78	-0,10	1,1E-01
M18	Синтез белка	↓	3,9E-06	0,8	-0,14	2,0E-02
M19	Пронейральная дифф.	↓	1,4E-07	0,59	-0,22	3,0E-04
M20	Нейрогенез	↓	7,3E-06	0,64	-0,19	1,0E-03

(А) Уровень экспрессии каждого модуля (согласно его характеристическому профилю) сравнивали между опухолями с амплификацией гена *EGFR* и остальными опухолями. Модули с повышенным уровнем экспрессии в опухолях с амплификацией отмечены символом «↑»; модули с пониженным - «↓» ($P < 0.05$, тест Уилкоксона). * - P-значение теста Уилкоксона. † - соотношение уровня экспрессии модуля между двумя группами опухолей. (Б) ‡ - коэффициент корреляции Пирсона между характеристическим профилем экспрессии модуля и профилем экспрессии гена *EGFR*. § - P-значение, оценивающее статистическую значимость коэффициента корреляции. Цветом отмечен модуль, уровень экспрессии которого сильнее других повышен в опухолях с активированным *EGFR*, согласно обоим критериям.

4) Поиск потенциальных терапевтических мишеней в глиомах. Одним из направлений изучения рака является поиск белков, модуляция активности которых подавляет рост опухолей. Такие белки используются в качестве мишеней для разработки методов химиотерапии путем поиска веществ-модуляторов активности этих белков. Верификация возможности использования белка в качестве экспериментальной мишени требует применения широкого спектра направленных молекулярно-биологических и биохимических методов. В то же время, на этапе

первичной идентификации таких белков важную роль играют поисковые полногеномные методы. В связи с этим, мы поставили задачу составить выборку белков, на перспективность которых в качестве экспериментальных мишеней в глиомах указывают результаты анализа структуры транскриптома этих опухолей.

Ранее было предложено осуществлять поиск потенциальных противоопухолевых мишеней среди центральных генов пролиферативного модуля в сетях коэкспрессии (Horvath S *et al* 2006). Чтобы оценить обоснованность этого предположения, мы изучили распределение мишеней лекарств, уже применяющихся для лечения онкологических заболеваний, по модулям коэкспрессии в глиоме. Для этого была использована база данных DrugBank, содержащая информацию о том, с какими белками в организме человека связываются лекарственные вещества. С помощью DrugBank составлен список разрешенных к применению противоопухолевых препаратов (81 лекарство) и соответствующих им мишеней (109 белков). В качестве контрольного использовался список мишеней лекарств, применяющихся для лечения заболеваний неонкологической природы (464 мишени). Далее, изучено, как эти мишени распределены по консенсусным модулям коэкспрессии в глиоме.

Мишенями противоопухолевых лекарств действительно оказался обогащен модуль пролиферации (M11, $P < 0.007$). В него попало 9 мишеней, в то время как количество, ожидаемое в рамках случайной модели, составляет 3 мишени ($P < 2 \times 10^{-3}$). Ассоциация является специфичной, поскольку модуль M11 не обогащен мишенями лекарств из контрольного списка. Мы рассмотрели, как свойства мишеней связаны с их положением внутри пролиферативного модуля. Оказалось, что количество химически различных лекарств, разработанных к мишени, коррелирует с ее близостью к центру пролиферативного модуля (коэффициент Пирсона = 0.88, $P = 0.0034$, рис. 5). Так, каждой из периферических мишеней соответствует лишь по одному-двум лекарствам, в то время как максимальное число лекарств приходится на центральные гены модуля – тимидилатсинтазу и ДНК-топоизомеразу II (шесть и семь препаратов, соответственно) (рис. 5). Известно, что большое количество лекарств, разработанных на одну и ту же мишень, отражает заинтересованность компаний в этой мишени, поскольку она уже показала свою эффективность в клинической практике. Таким образом, результаты анализа подтверждают высказанное ранее предположение о том, что центральные гены пролиферативного модуля возможно использовать в качестве мишеней противоопухолевых лекарств.

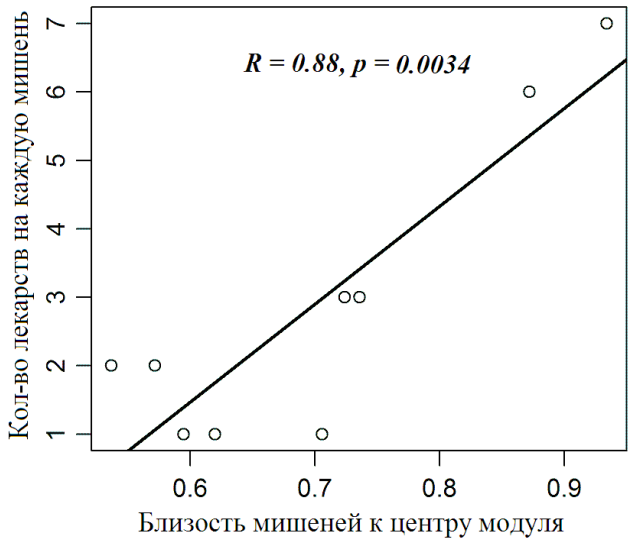


Рис. 5. Близость противоопухолевых мишеней к центру пролиферативного модуля коррелирует с количеством лекарств, действующих через эти мишени. Точки соответствуют мишеням противоопухолевых лекарств, входящим в состав пролиферативного модуля. Близость мишеней к центру модуля вычисляли как коэффициент корреляции Пирсона между профилем экспрессии гена и характеристическим профилем экспрессии пролиферативного модуля. Прямая наилучшего соответствия получена с помощью линейной регрессии. Оценка статистической значимости коэффициента корреляции между величинами, отложенными по осям, проведена с помощью пермутационного теста (случайное перемешивание пар, 100 000 итераций).

Следует отметить, что большинство противоопухолевых препаратов, существующих на фармацевтическом рынке, были разработаны в XX веке в рамках парадигмы поиска цитотоксических препаратов, действующих на активно делящиеся клетки. В настоящее время применяется более широкий спектр подходов при разработке лекарств этого класса, которые привели к появлению препаратов с новыми механизмами действия: иматиниб, сорафениб, сунитиниб (ингибирование сигнальных путей, активируемых рецепторными тирозинкиназами), авастин (подавление ангиогенеза), бортезомиб (модуляция деградации белков) и др. Мишени некоторых из них находятся в модулях коэкспрессии, отличных от пролиферативного: например, PDGFR (сунитиниб) – в модуле кровеносных сосудов M6, VEGF (авастин) – в модуле ответа на гипоксию M4. Поэтому можно ожидать, что гены не только пролиферативного, но также и некоторых других модулей могут являться перспективными противоопухолевыми мишенями.

Чтобы составить выборку белков, представляющих интерес в качестве экспериментальных мишеней в глиоме, мы выбрали модули, удовлетворяющие одновременно двум условиям: ассоциация повышенного уровня экспрессии с плохим прогнозом (табл. 3) и связь модуля с опухолевыми процессами (табл. 2), которые, согласно литературным данным, повышают злокачественность глиом. Согласно этим критериям, выбраны модули M3, M4, M6, M8 и M11 (суммарно 1026 генов). В каждом из модулей гены были ранжированы по близости к центру модуля: среднее по 5 массивам данных значение коэффициента корреляции между профилем экспрессии гена и характеристическим профилем экспрессии модуля. Наиболее близкие к центрам соответствующих модулей гены (например, ANXA5, ACSS3, SPRY2, VAV3, FABP7 в модуле M3;

ADM, VEGFA, GLUT3, ANGPTL4 в модуле M4; и другие) могут быть рекомендованы для изучения независимыми методами в качестве потенциальных противоопухолевых мишеней.

Для того, чтобы выделить в ранжированных списках генов те, которые проще других подвергнуть первичной экспериментальной оценке, мы провели поиск химических соединений, способных связываться с белками, соответствующими этим генам. С помощью базы данных DrugBank составлен список химических соединений, которые хотя и не являются лекарствами при каких-либо заболеваниях, однако находятся в клинических испытаниях либо используются для изучения биологии клетки в фундаментальных исследованиях (822 химических соединений и 282 мишени). Мы определили, какие соединения из этого списка действуют через мишени, принадлежащие модулям M3, M4, M6, M8 и M11. Найдено 138 таких соединений (26 белков). 13 из этих соединений действуют на белки, входящие в число первых 10% белков по близости к центрам соответствующих модулей. Такие химические соединения могут служить потенциальными модуляторами активности соответствующих белков при их дальнейшем экспериментальном изучении.

Выводы

1. Создана программа Microarray Retriever, предоставляющая интегрированный доступ к существующим экспрессионным базам данных (http://www.lgtc.nl/MaRe/).
2. Предложен эвристический метод, делающий доступным поиск модулей коэкспрессии в полногеномном масштабе.
3. При помощи протеомной базы данных Human Protein Atlas верифицированы экспрессионные предсказания функциональной связи с клеточной органеллой ресничкой для 25 генов человека.
4. Детально охарактеризована структура транскриптома глиомы на выборке из 790 больных. Получен набор из 20 воспроизводимых экспрессионных подписей, характеризующих широкий спектр клеточных процессов в глиоме.
5. Показано существование экспрессионного класса глиом, связанного с проастроцитарной дифференцировкой опухолей и благоприятным прогнозом.
6. Предсказано, что в регуляцию одного из ключевых онкогенных сигнальных путей в глиомах, активируемого рецептором эпидермального фактора роста (EGFR), вовлечены белки семейства Sprouty (SPRY1, SPRY2, SPRY4).
7. С использованием базы данных DrugBank, показано существование статистических закономерностей распределения известных в настоящее время противоопухолевых мишеней в сети коэкспрессии генов в глиоме. Предложены новые белки в качестве потенциальных новых противоопухолевых мишеней для дальнейшего изучения.

СПИСОК РАБОТ ОПУБЛИКОВАННЫХ ПО ТЕМЕ ДИССЕРТАЦИИ:

Статьи в научных журналах:

1. Ivliev A.E., ‘t Hoen P.A., Sergeeva M.G. Coexpression network analysis identifies transcriptional modules related to proastrocytic differentiation and Sprouty signaling in glioma. *Cancer Research*. 2010. 70(24), 10060-10070.
2. Ивлиев А.Е., Руднева В.А., Сергеева М.Г. Применимость анализа сетей коэкспрессии генов к поиску мишеней противоопухолевых лекарств. *Молекулярная Биология*. 2010. 44(2), 366–374.
3. Ivliev A.E., 't Hoen P.A., Villerius M.P., den Dunnen J.T., Brandt B.W. Microarray retriever: a web-based tool for searching and large scale retrieval of public microarray data. *Nucleic Acids Research*. 2008. 36(Web Server issue):W327-331.

Тезисы конференций:

1. Ivliev A.E., Sergeeva M.G. Prediction of human cilia-related genes by analysis of open-access transcriptomic and proteomic resources. Moscow Conference on Computational Biology and Bioinformatics, Moscow, Russia, 21-24 July, 2011.
2. Ivliev A.E., ‘t Hoen P.A., Sergeeva M.G. Integrative analysis of gene coexpression modules in glioma based on WGCNA algorithm. 19th International Conference on Intelligent Systems for Molecular Biology & 10th European Conference on Computational Biology, Vienna, Austria, 17-19 July, 2011.
3. Ivliev A.E., ‘t Hoen P.A., Peters D.J., Sergeeva M.G. Integrative analysis of gene coexpression networks identifies novel ciliary proteins in human tissues. European Human Genetics Conference 2011, Amsterdam, the Netherlands, 28-31 May, 2011.
4. Ивлиев А.Е. Анализ транскриптома предсказывает участие белков семейства Sprouty в регуляции онкогенного сигнального пути EGFR в глиомах. XVIII международная научная конференция студентов, аспирантов и молодых учёных «Ломоносов», г. Москва, Россия, 11-15 апреля, 2011.
5. Ивлиев А.Е., Сергеева М.Г. Полногеномный анализ экспрессии генов в изучении биологии глиом. I международная научно-практическая конференция «Постгеномные методы анализа в биологии, лабораторной и клинической медицине», г. Москва, Россия, 17-19 ноября, 2010.
6. Ивлиев А.Е. Выделение прогностических типов глиомы на основе профилей генной экспрессии. XVII международная научная конференция студентов, аспирантов и молодых учёных «Ломоносов», г. Москва, Россия, 12-15 апреля, 2010.
7. Руднева ВА, Ивлиев А.Е. Сравнение мер коэкспрессии по способности обнаруживать функциональную связь между генами. XVII международная научная конференция студентов, аспирантов и молодых учёных «Ломоносов», г. Москва, Россия, 12-15 апреля, 2010.
8. Ивлиев А.Е., Сергеева М.Г. Анализ транскриптома глиомы для разработки подходов к ее химиотерапии. Всероссийская научная школа для молодежи «Горизонты нанобиотехнологии», г. Москва, Россия, 12-16 октября, 2009.
9. Ивлиев А.Е. Изучение функции групп коэкспрессирующихся генов на примере анализа образцов из мозга больных глиобластомой. XVI международная научная конференция студентов, аспирантов и молодых учёных «Ломоносов», г. Москва, Россия, 13-18 апреля, 2009.
10. Руднева В.А., Ивлиев А.Е. Исследование генов с высокой связностью в сети коэкспрессии как потенциальных мишеней лекарственных средств. XVI международная научная конференция студентов, аспирантов и молодых учёных «Ломоносов», г. Москва, Россия, 13-18 апреля, 2009.