СТАТИСТИЧЕСКИЕ ОЦЕНКИ ЭКСПРЕССИИ МОБИЛЬНЫХ ЭЛЕМЕНТОВ В ГЕНОМЕ ЧЕЛОВЕКА НА ОСНОВЕ КЛИНИЧЕСКИХ ДАННЫХ ЭКСПРЕССИОННЫХ МИКРОЧИПОВ

Ю.Л. Орлов, В.М. Ефимов, Н.Г. Орлова

Учреждение Российской академии наук Институт цитологии и генетики Сибирского отделения РАН, Новосибирск, Россия,
e-mail: orlov@bionet.nsc.ru; efimov@bionet.nsc.ru; orlovanina2@mail.ru

Мобильные элементы генома человека способны экспрессироваться в соматических клетках в различных тканях. Прямая экспериментальная оценка уровня их экспрессии затруднена и не имеет большой статистической базы. В то же время огромный массив данных микрочипов, накопленный по клиническим данным экспрессии генов в геноме человека, позволяет исследовать статистические вопросы качества аннотации генов, шумов при измерении экспрессии, а также экспрессии мобильных элементов, включенных в пробы микрочипов. Такие данные по экспрессии генов при раковых заболеваниях накоплены за последние годы на основе платформы Affymetrix в связи с диагностическими и медицинскими задачами в Интернет-репозитариях данных Gene Expression Omnibus (GEO) NCBI и ArrayExpress. Статистические оценки экспрессии мобильных элементов изучались ранее в связи с проблемами качества олигонуклеотидных проб. Были исследованы геномная локализация и качество аннотации целевых последовательностей наборов проб Affymetrix U133 GeneChip и показано, что до 25 % целевых последовательностей перекрываются с мобильными элементами в хромосомных координатах.
В данной работе численно показан статистически значимый эффект изменения экспрессии в раковых тканях для наборов проб, связанных с мобильными элементами. Приведен обзор современных технологий оценки экспрессии мобильных элементов.

Ключевые слова: геном человека, мобильные элементы, экспрессионные микрочипы, статистические оценки, транскрипция.

Введение

Геном человека приблизительно на 45 % состоит из диспергированных повторяющихся элементов (Sela *et al*., 2010). *Alu*-повтор – наиболее распространенный примат-специфичный мобильный элемент, существующий в более чем миллионе копий в геноме, что составляет около 11 % общего размера генома (Zhang *et al*., 2011). *Alu*-повтор представляет семейство *SINE* коротких (до 300 п.н.) диспергированных повторов. Геном человека содержит множество копий других мобильных элементов включая *MIR* (*SINE*), длинные диспергированные повторы (*LINE*, long interspersed nuclear elements), занимающие до 17 % генома, такие, как *LINE-1* (*L1*), *LINE-2* (*L2*), и *CR1* (*L3*) и LTR-повторы. Структура и классификация мобильных элементов в геноме человека детально изучены, существуют базы данных RepBase (http://www.girinst.org/repbase/; Jurka, 2000; Jurka *et al*., 2005) и разметка (аннотация) последовательностей на хромосомах с помощью программы RepeatMasker (Smit *et al*., 1996–2010).

В ряде работ был исследован характер распределения мобильных элементов в геноме человека относительно генов (Sela *et al*., 2007, 2010; Levy *et al*., 2010). Показано, что мобильные элементы имеют тенденцию к кластеризации во внутригенных районах генов, в интронах (Sela *et al*., 2010). Исходя из суммарного размера интронов в геноме доля геномных повторов, находящихся в интронах (около 60 % в зависимости от семейства повторов), значительно превышает ожидаемую.

До 25 % промоторов генов и 4 % экзонов в геноме человека содержат последовательности, происходящие от мобильных элементов (Jordan *et al*., 2003). Известно явление экзонизации – приобретение генами новых экзонов из мобильных элементов, прежде всего *Alu*, в интронах. Показано, что доля экзонов, происходящих от мобильных элементов, в аннотированных генах составляет доли процента (в среднем 0,12 % в зависимости от класса геномных повторов, в том числе 0,2 % для *Alu*) (Sela *et al*., 2010).

Интересно отметить, что несмотря на присутствие мобильных элементов в тысячах копий, избыточность их числа в геноме по отношению к числу генов, они остаются транскрипционно молчащими в нормальных условиях в соматических клетках и не перемещаются в геноме (Hagan, Rudin, 2002). В то же время при повреждающих воздействиях на клетку, в частности при раке, может происходить активация транспозонов различных классов в геноме (Gasior *et al*., 2006; O′Donnell, Burns, 2010; Iramaneerat *et al*., 2011). Недавние исследования (Zhang *et al*., 2011) выявили, что инсерция *Alu*-элементов связана с канцерогенезом. Показано, что гены-супрессоры рака гораздо более насыщены *Alu*-повторами, чем онкогены (Zhang *et al*., 2011). В работе Banaz-Yaşar с соавт. (2010) изучались активация элементов *LINE-1* при раке, влияние ретротранспозиции на факторы системы кроветворения. Один из известных механизмов подавления спонтанной экспрессии мобильных элементов – метилирование ДНК (Daskalos *et al*., 2009; Iramaneerat *et al*., 2011). Нарушение работы этого внутриклеточного механизма, гипометилирование геномных участков, содержащих транспозоны, в частности *HERV-K*, приводят к активации их экспрессии, что может быть использовано в диагностических целях (Iramaneerat *et al*., 2011). Важнейшая задача постгеномной компьютерной генетики – исследование экспрессии мобильных элементов в клетках человека, возможности ее оценки и использования в диагностических целях и в целях тестирования медицинских препаратов (Daskalos *et al*., 2009; Balaj *et al*., 2011).

За последние годы было предложено несколько технологий анализа экспрессии генов как с помощью экзонных микрочипов (Jacox *et al*., 2010), так и с помощью тотального секвенирования (RNA-seq) (Mortazavi *et al*., 2008; Ozsolak, Milos, 2011).

Мы предлагаем оценить экспрессию мобильных элементов в опухолевых тканях статистически, по данным экспрессионных микрочиповых экспериментов, накопленных в базе данных ArrayExpress (www.ebi.ac.uk/arrayexpress). Экспрессионные микрочипы (microarray) представляют собой матрицы для количественного измерения присутствия транскриптов (мРНК) одновременно для большого числа генов в образце клеток или ткани. Микрочипы измеряют аффинность (степень связывания в процессе гибридизации) меченых нуклеотидных последовательностей образца к набору специфичных к заданному гену проб, закрепленных на твердой поверхности микроматрицы. Эксперимент проводится во многих ячейках микроматрицы (десятки тысяч проб) одновременно для заданного технологической платформой множества транскрибирующихся последовательностей генов вплоть до всех аннотированных в геноме генов. Соревнование в сфере технологий производства микрочипов, технологий измерения сигналов гибридизации дало большой толчок научным исследованиям и огромный фактический материал. Следует отметить, что за короткий период в последние 2–3 года на смену микрочипам приходят все более совершенные технологии полного секвенирования транскриптом, имеющие ряд неоспоримых преимуществ, в частности, по способности определения новых вариантов транскриптов гена, по динамической шкале измерения уровня транскрипции, но все же достаточно дорогостоящие (на порядок по сравнению с микрочипами) для широкого использования (Mortazavi *et al*., 2008; Malone, Oliver, 2011). Тем не менее микрочиповая технология позволяет достаточно надежно определять дифференциально экспрессирующиеся гены за счет репликации экспериментов (Malone, Oliver, 2011).

Технология синтеза коротких олигонуклеотидных зондов (25 п.н.) непосредственно на поверхности микрочипа *in situ* с использованием литографических масок была разработана компанией «Аффиметрикс» для изготовления микрочипов GeneChip (Affymetrix, www.affymetrix.com/). Методы измерения уровней экспрессии генов

на основе таких микрочипов получили широкое распространение в медицинских исследованиях (Liu *et al*., 2003; Dai *et al*., 2005; Harbig *et al*., 2005). Олигонуклеотидная матрица GeneChip использует наборы синтезированных *in situ* олигонуклеотидных проб, по 11–20 проб в наборе, каждая размером 25 нуклеотидов, для представления транскриптов генов или их изоформ. Для каждого гена-мишени использованы фрагменты-представители (initial target sequences) длиной 150–450 п.н. для выбора и локализации олигонуклеотидных проб. Сигнал от пробы с совершенным совпадением всех нуклеотидов учитывается после вычитания неспецифического сигнала кросс-гибридизации от пробы с одним центральным несовпадающим нуклеотидом (Affymetrix, 2002, http://www.affymetrix.com/support/).

Здесь будут рассмотрены данные микрочипов Affymetrix GeneChip, относящиеся к клиническим экспериментам на опухолевых тканях, хирургически полученных при лечении, в применении к анализу спонтанной экспрессии мобильных элементов (Orlov *et al*., 2007). Отметим, что дизайн проб этого микрочипа исходно не предназначался для такого анализа, что потребовало разработки специальных статистических оценок.

Непосредственная детекция экспрессии транспозонов, передвижений и встроек мобильных элементов в хромосомы соматических клеток технически ограничена. По методическим и техническим причинам геномные повторы обычно исключаются из дизайна микрочипов (Nellåker *et al*., 2009), в частности из-за избыточности мобильных элементов в геноме и сложности подбора уникальных проб. Таким образом, их потенциальная транскрипционная активность остается недостаточно охарактеризованной, несмотря на многочисленные наблюдения присутствия транскрипции в различных тканях при заболеваниях человека (Frank *et al*., 2008; Karlsson *et al*., 2001). С помощью технологий анализа полноразмерных «кэпированых» транскриптов была показана связь инициации транскрипции с присутствием в 5′-области генов ретротранспозонов (от 6 до 30 %) в геномах мыши и человека (Faulkner *et al*., 2009). Экспрессия генов, содержащих ретротранспозоны в 3′-НТР, уменьшена по сравнению с генами, не содержащими таких транскриптов (Faulkner *et al*., 2009).

Проблема анализа транскрипции с помощью микрочипов в целом связана с рядом технических ограничений и ошибок при создании технологии. Дизайн проб (исходный выбор производителем микрочипов локализации в гене и структуры олигонуклеотидных проб) может не соответствовать целевому транскрипту (гену-мишени) и содержать ряд технических проблем, связанных как с гибридизацией, так и с аннотацией – неверное указание гена-мишени, неоднозначность соответствия один набор проб–один ген. Такой дизайн олигонуклеотидных проб может влиять на регистрацию сигналов гибридизации, нормализацию данных, снижать воспроизводимость экспериментов, вести к противоречивым результатам анализа одних и тех же данных (Gautier *et al*., 2004; Harbig *et al*., 2005; Zhang *et al*., 2005; Okoniewski, Miller, 2006; Orlov *et al*., 2007; Stalteri, Harrison, 2007; Fasold *et al*., 2010).

Мы оценили присутствие мобильных элементов генома человека в целевых последовательностях транскриптов, представленных наборами проб на микрочипе Affymetrix, и возможность статистической оценки транскрипции классов мобильных элементов в соматических клетках.

Методы

Исходно до 2003 г. был разработан микрочип GeneChip U133A, дополненный позднее чипами U133B – U133 plus 2, более полно соответствующими всем известным и проаннотированным на тот момент генам в геноме человека. Уровень экспрессии гена определяется суммой данных всего набора проб (probeset).

Была выполнена независимая аннотация наборов проб микрочипов Affymetrix на основе картирования нуклеотидных последовательностей проб на референсные последовательности генома человека (Gautier *et al*., 2004; Dai *et al*., 2005; Harbig *et al*., 2005; Leong *et al*., 2005). Выявлен ряд несоответствий в аннотации наборов проб; изменения в идентификации генов могут затрагивать до 30–50 % наборов проб (Harbig *et al*., 2005; Okoniewski, Miller, 2006; Fasold *et al*., 2010).

В то же время вопрос картирования проб на целевые последовательности генов, содержащие мобильные элементы в геноме человека, не рассматривался детально. Статистически связь присутствия последовательностей мобильных элементов и систематических изменений в экспрессии генов была показана только в работе Orlov с соавт. (2007).

Мы использовали аннотацию наборов проб Affymetrix U133 GeneChip, выполненную ранее в работе Orlov с соавт. (2007) с целью детального изучения влияния мобильных элементов на изменение экспрессии генов в раковых тканях. Разработан ряд статистических компьютерных методов для количественных оценок изменения активности мобильных элементов через изменение экспрессионного сигнала соответствующих наборов проб. Анализ был сделан на экспрессионных данных GeneChip из больших выборок (базы данных и репозитарии GEO (Gene Expression Omnibus), http://www.ncbi.nlm.nih.gov/geo, и ArrayExpress, http://www.ebi.ac.uk/arrayexpress/) по экспрессии генов в раковых клетках, отличающихся по клиническим и генетическим параметрам, а также по степени агрессивности роста опухоли.

Анализ последовательностей олигонуклеотидных проб на микрочипах

Использовались данные о целевых нуклеотидных последовательностях (мишенях) для наборов проб Affymetrix, микрочипы серий U133A и U133B, загруженные с официального сайта разработчиков платформы NetAffx (http://www.affymetrix.com/analysis/index.affx) (Liu *et al*., 2003). Эти последовательности предназначаются для однозначной детекции транскрибируемых последовательностей в геноме. Для картирования таких целевых последовательностей на референсную последовательность генома человека была использована программа BLAT (http://genome.ucsc.edu/cgi-bin/hgBlat) с порогом отсечения, установленным на 90 %-м уровне сходства. Затем использовалась аннотация геномного браузера UCSC Genome Browser для генов RefSeq, мРНК и сплайсированных вариантов EST на референсные последовательности хромосом генома человека по сборке NCBI Build 36 (hg18). Аннотация наборов проб выполнялась по исходным целевым последовательностям, а не по отдельным 25-мерным нуклеотидных пробам. Данные по наличию геномных повторов были получены с помощью разметки RepBase в геномном браузере UCSC (http://genome.ucsc.edu/cgi-bin/hgTracks, табл. RepeatMasker). Примеры расположения целевой последовательности наборов проб приведены на рис. 1.

Проверка качества целевых последовательностей Affymetrix выполнялась последовательно: сначала были отфильтрованы некартируемые и неоднозначно картируемые последовательности, затем последовательности в неверной ориентации к аннотированным генам. Затем проводилась разметка мобильных элементов из RepBase. Для каждой целевой последовательности, однозначно картированной на геноме, была получена таблица геномных повторов, классифицированных по семействам повторов и типам (*DNA*, LTR, *LINE*, *SINE* включая *MIR* и *Alu*), а также простые тандемные повторы и участки низкой сложности), определены длина и процент длины, занятый геномными повторами данных типов. Суммарная статистика приведена в табл. 1. Как отмечалось ранее, процент «экзонизированных» геномных повторов невелик (Sela *et al*., 2010). Перекрывание целевой последовательности транскрипта гена с геномным повтором не является ошибкой и не указывает на факт экзонизации.

Данные для анализа экспрессии

Было проанализировано распределение значений экспрессии транскриптов, полученных с помощью микрочипов Affymetrix U133A и U133B в 249 образцах первичных опухолей молочной железы (NCBI Gene Expression Omnibus (GEO) http://www.ncbi.nlm.nih.gov/geo/; данные GSE4922). Выборки раковых тканей были разделены на группы, соответствующие гистологическим классам опухоли по степени агрессивности (метастазирования) рака молочной железы. Объем выборок составлял от 40 до 100 образцов (Miller *et al*., 2005). Были использованы также данные экспрессии из нескольких выборок нормальных и раковых тканей мозга (GEO GDS1962), 29 наборов микрочиповых данных Affymetrix, представляющих рак легких (GEO ID: GSE5816; http: //www.ncbi.nlm.nih.gov/geo/) (Shames *et al*., 2006). Все данные

Рис. 1. Примеры перекрывания целевых последовательностей наборов проб с мобильными элементами в геноме человека (визуализация UCSC Genome Browser).

а – пример целевой последовательности Affymetrix 217670_at на хромосоме 11 человека, совпадающей с повторяющимися элементами из RepeatMasker (*SINE*, семейство *Alu*-повторов). Данная целевая последовательность находится в интроне. Возможна «экзонизация» соответствующего повтора в неаннотированных изоформах транскрипта (присутствие сплайсированных мРНК) на данном участке.
б – целевая последовательность набора проб 221200_at не соответствует ни генам, ни мРНК и находится внутри протяженного геномного повтора HERVK3-int (ERVK, LTR).

Таблица 1

Классификация геномных повторов
в целевых последовательностях наборов проб Affymetrix

Группа геномных повторов	Классы повторяющихся элементов по RepBase	Число наборов проб	Доля наборов проб, %
Короткие транспозоны (< 300 п.н.)	*SINE/Alu*, *SINE/MIR*	3200	31,8
– в том числе *Alu*	*Alu*	1807	18,0
Длинные транспозоны (>300 п.н.)	*LINE/CR1*, *LINE/L1*, *L2*	2191	21,8
– в том числе *L1*	*L1*	1394	13,9
LTR	LTR/ERV1/ERVK/ERVL/MaLR	1235	12,3
DNA	*MER1*, *MER2*	1005	10,0
Другие повторяющиеся элементы и сателлитные повторы	RNA, rRNA, Satellite, scRNA, snRNA, srpRNA	52	0,5
Участки низкой сложности, простые повторы	Low_complexity	2373	23,6

прошли контроль качества сигнала экспрессии, нормализацию по алгоритму MAS5 (MAS 5.0 algorithm) (Affymetrix, 2002). Затем значения экспрессии были логарифмиче-ски нормированы для сопоставления экспрессии генов на микрочипах из разных экспериментов. Измерение уровня экспрессии проводилось без выделения отдельных проб.

Статистический анализ

Была использована эмпирическая функция распределения сигнала экспрессии на индивидуальных микрочипах. Эмпирическая функция распределения была построена также для групп наборов проб, классифицированных по степени присутствия повторов в целевых последовательностях. Для сравнения распределений использовалась компьютерная симуляция – были сгенерированы случайные группы наборов того же размера, процедура повторялась с помощью датчика случайных чисел.

Для контроля предсказательной (диагностической) способности наборов проб, содержащих мобильные элементы и не содержащих их, были использованы опубликованные ранее результаты группировки тканей (различных гистологических классов рака молочной железы). Данные группы разделяются по экспрессии около 4000 дифференциально экспрессирующихся генов (Orlov *et al*., 2007).

Для численной оценки использовалась программа SAM (Statistical Analysis of Microarrays) (Tusher *et al*., 2001). Для каждого набора проб данная программа рассчитывает «значимость различия» между двумя выборками данных (группами опухолей) с помощью оценки значения доли ложного предсказания «false discovery rate» (параметр q-value). При фиксированном пороговом уровне q-value программа SAM идентифицирует набор генов (наборов проб), позволяющих достоверно разделить выборки. Зафиксировав значение этого параметра на уровнях 0,05 и 0,015, мы оценили фракцию наборов проб, позволяющих дискриминировать типы опухолей и содержащих при этом мобильные элементы. Дискриминирующая способность проб, содержащих мобильные элементы, была оценена с помощью отношения наблюдаемой и ожидаемой доли проб, содержащих мобильные элементы. Для оценки значимости результатов использовался критерий Манна-Уитни (U-test) и односторонний точный критерий Фишера для таблиц данных.

Результаты

Статистика геномных повторов в наборах проб

В целом до 25 % целевых последовательностей для наборов проб проявляют значимое сходство с мобильными элементами (геномны-

ми повторами), распространенными в геноме человека (Orlov *et al*., 2007). Исходно дизайн микрочипа не предназначался для исследования экспрессии мобильных элементов. Но эти данные могут быть использованы для статистических оценок, так же, как и для фильтрации и калибровки измерения экспрессионного сигнала. Табл. 1 представляет число наборов проб, содержащих полностью или частично, геномные повторы.

Как видно из табл. 1, большое число целевых последовательностей наборов проб на микрочипе проявляют значимое сходство с геномными повторами, представляя, тем не менее, лишь малую часть от более чем 5 млн участков, размеченных RepeatMasker (Smit *et al*., 1996–2010) в геноме человека (таблица UCSC). Напомним, что размеры последовательностей варьируют от 100 до 500 п.н. Доля перекрывания, как правило, невысока, менее 50 % для большинства последовательностей (рис. 2). В то же время несколько тысяч целевых последовательностей перекрываются с геномными повторами более чем на 40 % от своей длины, а около 600 целевых последовательной – более чем на 90 %, что, несомненно, влияет на качество сигнала и может детектировать экспрессию мобильных элементов, а не генов, для которых исходно предназначался дизайн наборов проб. При этом часть последовательностей содержит простые тандемные повторы и участки низкой сложности, занимающие менее 10 % от общей длины целевой последовательности, что не должно оказывать влияние на сигнал экспрессии (рис. 2).

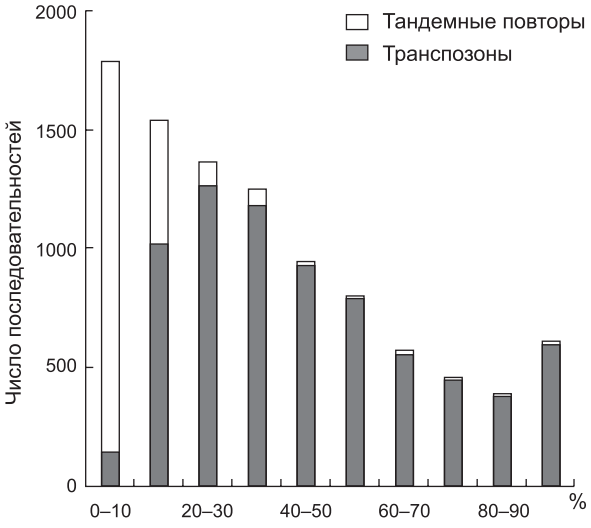

Рис. 2. Распределение числа целевых последовательностей наборов проб Affymetrix U133, пересекающихся с повторяющимися элементами в геноме человека в зависимости от процентной доли геномных повторов в последовательности (ось абсцисс – от 0 до 100 %).

Классификация целевых последовательностей Affymetrix по качеству дизайна и соответствию аннотации генов

Табл. 2 содержит общую статистику различных категорий неверно определенных целевых последовательностей проб Affymetrix U133 на основе геномной сборки Hg18. Около 6 % со-

Таблица 2

Общая классификация проблемных целевых последовательностей микрочипа Affymetrix U133 (чипы A и B)

Группа целевых последовательностей наборов проб	Число	Доля, %
Неоднозначно картируемые на геном человека	1984	4,4
Картированные в обратной ориентации к транскрипту	810	1,8
Перекрывающиеся с геномными повторами	3387	7,6
80–100 % длины последовательности	761	1,7
60–80 %	936	2,1
40–60 %	1690	3,8
Итого не рекомендуется использовать	6181	13,8
Общее число корректных (рекомендуемых к использованию) наборов проб (включая перекрывание менее 40 % длины последовательности)	38511	86,2
Общее число последовательностей микрочипов U133A и B	44692	100

ставляют последовательности, картируемые на геном в различных участках (неоднозначно картируемые), не соответствующие последовательностям генома человека, и последовательности, картируемые в противоположной ориентации к транскрибируемым последовательностям генов. Общая фракция целевых последовательностей наборов проб велика, до 25 %, но большей частью геномные повторы «присутствуют» в таких последовательностях лишь частично, и их можно считать адекватными для измерения экспрессионного сигнала. В целом 86 % наборов проб были рекомендованы к использованию (Orlov *et al*., 2007).

Сравнение средних значений экспрессии наборов проб, содержащих геномные повторы

Мы сравнили средние уровни экспрессии для групп целевых последовательностей, содержащих геномные повторы в зависимости от длины последовательности (с шагом гистограммы 10 %), и для корректных целевых последовательностей (норма). Для каждой группы наборов проб, целевые последовательности которых содержат повторы, мы определили средние значения сигнала гибридизации наборов проб (в логарифмической шкале) и коэффициент вариации (дисперсия, нормированная на среднее значение) на выборках данных опухолей (рис. 3). На рис. 3 показано уменьшение среднего значения сигнала при увеличении доли геномных повторов в целевой последовательности. В то же время коэффициент вариации имеет противоположный тренд и может быть достаточно большим, более 0,1, для последовательностей, почти полностью занятых геномными повторами.

Способность целевых последовательностей, перекрывающихся с геномными повторами, к определению дифференциально экспрессирующихся генов

Было выполнено сравнение способности наборов проб дискриминировать дифференциально экспрессирующиеся гены в выборках образцов опухолей различных типов. Гистологически опухоли молочной железы классов I и III (низко- и высокометастазирующие) различаются, что может быть статистически на микрочипах показано дифференцированной экспрессией нескольких тысяч наборов проб Affymetrix. Используя программное обеспечение SAM

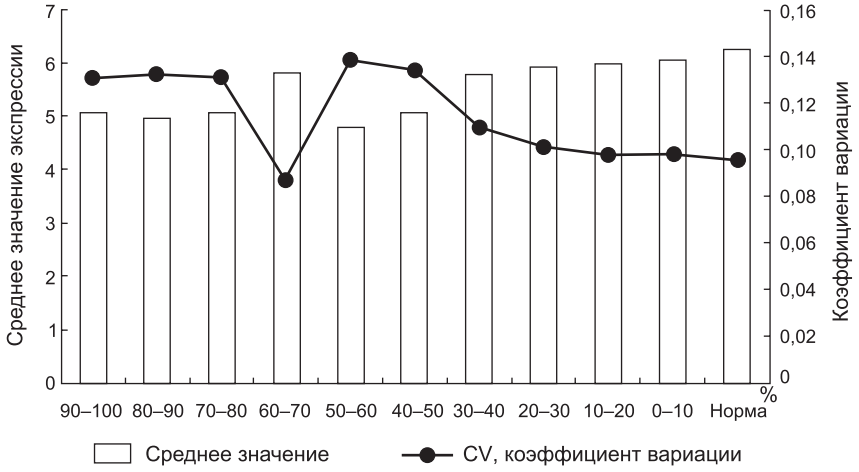

Рис 3. Среднее значение экспрессии и коэффициент вариации наборов проб, перекрывающихся с геномными повторами, по выборке опухолевых тканей.

Ось абсцисс – группа целевых последовательностей, занятых повторами на 90–100 %, 80–90 % и т. д. вплоть до 0 %, корректно определенных последовательностей (норма). По оси ординат слева – среднее значение экспрессии соответствующей группы в логарифмической шкале сигнала гибридизации (колонки гистограммы), справа – коэффициент вариации (линия), безразмерное значение. Видны противоположные тренды – уменьшение среднего значения экспрессии при увеличении доли геномных повторов и увеличение коэффициента вариации (зашумленности сигнала). Данные приведены по выборке образцов опухолей молочной железы (гистологический Grade I).

(Tusher *et al*., 2001), мы отобрали набор из 6144 дифференциально экспрессирующихся генов на микрочипах U133A&B на фиксированном уровне q-value ложного положительного предсказания, не превышающем 1,5 % .

Предполагая, что перекрывание с мобильными элементами приводит к ухудшению качества сигнала из-за неспецифического связывания проб с посторонними транскриптами и не может быть использовано для дискриминации, мы вправе ожидать, что такие наборы проб должны быть недопредставлены во множестве дифференциально экспрессирующихся генов. Было рассчитано число наборов проб, занятых повторами на 10, 20, … , 100 %, найденных в данном множестве дифференциально экспрессирующихся.

Доля дифференциально экспрессирующихся наборов проб, для которых соответствующая целевая последовательность перекрывается с геномными повторами в геноме человека, может быть рассчитана по формуле:

$$r = (R_s / R)/(N_s / N),$$

где N – общее число наборов проб микрочипа; R – число наборов проб, перекрывающихся с геномными повторами, N_s – число наборов проб, дифференциально экспрессирующихся в исследованных видах опухолей по статистическому тесту программы SAM; R_s – число наборов проб, перекрывающихся с геномными повторами и дифференциально экспрессирующихся в тех же видах опухолей по тесту SAM.

Таким образом, r – это отношение наблюдаемой доли наборов проб, связанных с геномными повторами, к ожидаемой доле, ее значение может быть как больше, так и меньше единицы. Было определено, что присутствие простых тандемных повторов и участков низкой сложности не влияет на способность наборов проб дискриминировать опухоли. Простые повторы в целом занимают незначительную часть целевых последовательностей (рис. 2), наборы проб, ассоциированные с простыми повторами, имеют малую вариабельность сигнала (коэффициент вариации по клиническим выборкам данных) и функционально не влияют на дискриминирующие свойства наборов проб при сравнении биологически различных выборок образцов тканей.

В то же время наборы проб, целевые последовательности которых содержат последовательности мобильных элементов, существенно хуже дискриминируют опухоли ($r < 1$). Более того, наблюдается тренд изменения дискриминирующего параметра r в зависимости от доли целевой последовательности, занятой мобильными элементами, особенно для протяженных геномных повторов (LTR и LINE): чем больше геномных повторов присутствует в целевой последовательности набора проб микрочипа, тем меньше отношение r (рис. 4).

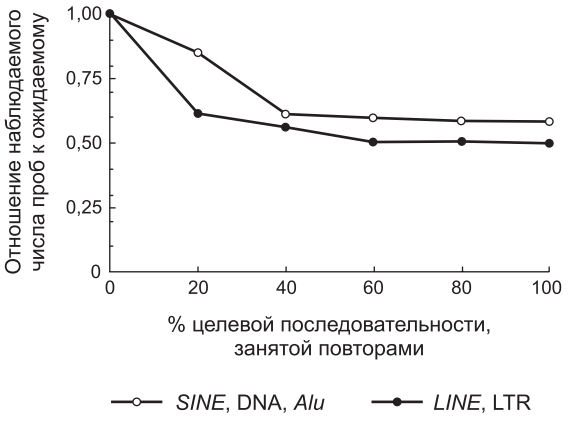

Рис. 4. Оценка отношения r дискриминирующих наборов проб к ожидаемому (при фиксированном q < 1,5 %) как функция процента целевой последовательности, занятой транспозонами.

Статистический анализ данных экспрессии наборов проб, связанных с геномными повторами, на выборках образцов опухолей (молочной железы, тканей мозга) выявил общую воспроизводимую тенденцию: 1) увеличение шума в сигнале экспрессии (коэффициента вариации); 2) уменьшение среднего уровня сигнала экспрессии и 3) увеличение числа ложных корреляций, не связанных с взаимной регуляцией транскрипции. Таким образом, при интерпретации данных экспрессионных микрочипов необходим учет особенностей геномной аннотации мобильных элементов.

Заключение

Оценка активности мобильных элементов в геноме человека может быть выполнена на основе различных технологий включая полное ресеквенирование индивидуальных геномов и

секвенирование индивидуальных транскриптом (Ewing, Kazazian, 2010). В перспективе это позволит более полно описать картину активации мобильных элементов в соматических клетках при повреждающих воздействиях различных типов, описать функционально активные копии и места встраивания мобильных элементов в геноме. Отметим, что новые технологии имеют ряд технических проблем, в частности достаточно высокий уровень ошибок секвенирования, чувствительность к GC-составу и гетерогенности последовательностей (Malone, Oliver, 2011). Это требует разработки специализированных компьютерных методов анализа.

Источники шума в численных данных микрочиповых экспериментов могут быть различны, связаны как с технологическими причинами, так и с неверной интерпретацией (аннотацией) наборов проб (Liu *et al*., 2003; Gautier *et al*., 2004; Harbig *et al*., 2005). Критика качества наборов проб, связанных с присутствием в них участков сходства с геномными повторами, была высказана ранее, однако без статистических оценок применительно к экспрессии мобильных элементов. Здесь показаны статистические оценки влияния повторов на интенсивность экспрессионного сигнала, коэффициент вариации и способность дискриминировать различные биологические классы (число дифференциально экспрессирующихся наборов проб) на больших выборках клинических данных (Orlov *et al*., 2007). Несмотря на ошибки дизайна проб и аннотации последовательностей-мишеней Affymetrix U133A&B, технологически платформа дает воспроизводимые результаты, подтвержденные большим объемом данных. Продолжаются работы по компьютерной аннотации наборов проб и сравнению микрочиповых платформ Affymetrix, основанных на генах (серия U133) и экзонах (Gene 1.0 и Exon 1.0), что позволяет получить новые результаты по экспрессии изоформ транскриптов (На *et al.*, 2009; Risueco *et al*., 2010).

Таким образом, потенциал микрочиповых данных может быть использован гораздо полнее через интеграцию геномной аннотации и клинических данных. Анализ наборов проб проводился для целевых последовательностей-мишеней. Более детальный анализ каждого набора проб на уровне индивидуальных проб только уточняет картину и увеличивает число выявленных неверно аннотированных проб, дающих ненадежный сигнал экспрессии. Дальнейшее изменение аннотации референсного генома, в частности, в связи с новыми проектами ресеквенирования генома человека, может увеличить число неверно аннотированных проб и привести к переоценке данных, накопленных при использовании данного типа микрочипов за последние годы (Fasold *et al*., 2010).

Детекция экспрессии мобильных элементов на данном типе микрочипа не была спланирована первоначально и показана только как результат статистического анализа. Исследование транскрипции в геноме человека с помощью новых технологий секвенирования, в частности RNA-seq, позволяет найти новые транскрипты в геноме, детекция которых невозможна с помощью микрочипов (Faulkner *et al*., 2009). Таким образом, измерение уровней экспрессии мобильных элементов в соматических клетках, в частности в опухолевых тканях, может быть сделано с помощью других подходов. Мы потеряем, к сожалению, огромный массив клинических данных для микрочипов, накопленный в диагностических целях.

Отметим, что гибридизационный сигнал от набора проб с обнаруженным перекрыванием целевой последовательности с каким-либо геномным повтором из RepBase, размеченным с помощью RepeatMasker, не дает информации о транскрипции конкретно этого мобильного элемента или другого гомологичного ему элемента, расположенного на других хромосомах. Таким образом, мы можем сравнивать только классы мобильных элементов и оценивать статистически их влияние на сигнал экспрессии.

Экспонирование клеток к ДНК-повреждающим воздействиям, таким, как лекарства химиотерапии или радиация, может вести к индукции транскрипции *SINE*-элементов, что подтверждает глобальную активацию транспозонов в геноме при стрессовых условиях (Hagan, Rudin, 2002). Есть данные об экспрессии в соматических клетках элементов семейства *L1* (*LINE*) (Gasior *et al*., 2006; Belancio *et al*., 2010; O′Donnell, Burns, 2010; Iramaneerat *et al*., 2011). Механизмы воздействия экспрессии мобильных элементов могут не ограничиваться встройками ДНК и структурными изменениями

генома. Показано, что РНК, транскрибируемая с *Alu*-повторов, может взаимодействовать с РНК полимеразой II и подавлять экспрессию некоторых белок-кодирующих генов (Ponicsan *et al*., 2010). Неравномерное распределение транспозонов в геноме и их активация могут вести к изменению экспрессии генов в раковых тканях и при повреждающих воздействиях, что требует дальнейшего изучения.

Благодарности

Авторы благодарны А.В. Катохину (ИЦиГ СО РАН) и Л. Липовичу (Университет Вэйн, США) за критические замечания. Работа частично поддержана грантами РФФИ 11-04-12167, 11-04-01888-а и ГК 07.514.11.4003.

Литература

Balaj L., Lessard R., Dai L. *et al.* Tumour microvesicles contain retrotransposon elements and amplified oncogene sequences // Nat. Commun. 2011. V. 2. e180.

Banaz-Yaşar F., Steffen G., Hauschild J. *et al. LINE-1* retrotransposition events affect endothelial proliferation and migration // Histochem. Cell Biol. 2010. V. 134. N 6. P. 581–589.

Belancio V.P., Roy-Engel A.M., Pochampally R.R., Deininger P. Somatic expression of *LINE-1* elements in human tissues // Nucl. Acids Res. 2010. V. 38. N 12. P. 3909–3922.

Dai M., Wang P., Boyd A.D. *et al*. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data // Nucl. Acids Res. 2005. V. 33. N 20. e175.

Daskalos A., Nikolaidis G., Xinarianos G. *et al*. Hypomethylation of retrotransposable elements correlates with genomic instability in non-small cell lung cancer // Int. J. Cancer. 2009. V. 124. N 1. P. 81–87.

Ewing A.D., Kazazian H.H. Jr. High-throughput sequencing reveals extensive variation in human-specific *L1* content in individual human genomes // Genome Res. 2010. V. 20. N 9. P. 1262–1270.

Fasold M., Stadler P.F., Binder H. G-stack modulated probe intensities on expression arrays – sequence corrections and signal calibration // BMC Bioinformatics. 2010. V. 11. e207.

Faulkner G.J., Kimura Y., Daub C.O. *et al*. The regulated retrotransposon transcriptome of mammalian cells // Nat. Genet. 2009. V. 41. N 5. P. 563–571.

Frank O., Verbeke C., Schwarz N. *et al*. Variable transcriptional activity of endogenous retroviruses in human breast cancer // J. Virol. 2008. V. 82. P. 1808–1818.

Gasior S.L., Wakeman T.P., Xu B., Deininger P.L. The human *LINE-1* retrotransposon creates DNA double-strand breaks // J. Mol. Biol. 2006. V. 357. P. 1383–1393.

Gautier L., Moller M., Friis-Hansen L., Knudsen S. Alternative mapping of probes to genes for Affymetrix chips // BMC Bioinformatics. 2004. V. 5. e111.

Ha K.Ch., Coulombe-Huntington J., Majewski J. Comparison of Affymetrix Gene Array with the Exon Array shows potential application for detection of transcript isoform variation // BMC Genomics. 2009. V. 10. e519.

Hagan C.R., Rudin C.M. Mobile genetic element activation and genotoxic cancer therapy: potential clinical implications // Am. J. Pharmacogenomics. 2002. V. 2. N 1. P. 25–35.

Harbig J., Sprinkle R., Enkemann S.A. A sequence-based identification of the genes detected by probesets on the Affymetrix U133 plus 2.0 array // Nucl. Acids Res. 2005. V. 33. N 3. P. 31.

Iramaneerat K., Rattanatunyong P., Khemapech N. *et al*. HERV-K hypomethylation in ovarian clear cell carcinoma is associated with a poor prognosis and platinum resistance // Int. J. Gynecol. Cancer. 2011. V. 21. N 1. P. 51–57.

Jacox E., Gotea V., Ovcharenko I., Elnitski L. Tissue-specific and ubiquitous expression patterns from alternative promoters of human genes // PLoS One. 2010. V. 5. N 8. e12274.

Jordan I.K., Rogozin I.B., Glazko G.V., Koonin E.V. Origin of a substantial fraction of human regulatory sequences from transposable elements // Trends Genet. 2003. V. 19. N 2. P. 68–72.

Jurka J. Repbase Update: a database and an electronic journal of repetitive elements // Trends Genet. 2000. V. 9. P. 418–420.

Jurka J., Kapitonov V.V., Pavlicek A. *et al.* Repbase Update, a database of eukaryotic repetitive elements // Cytogenet. Genome Res. 2005. V. 110. P. 462–467.

Karlsson H., Bachmann S., Schroder J. *et al*. Retroviral RNA identified in the cerebrospinal fluids and brains of individuals with schizophrenia // Proc. Natl Acad. Sci. USA. 2001. V. 98. P. 4634–4639.

Leong H.S., Yates T., Wilson C., Miller C.J. ADAPT: a database of affymetrix probesets and transcripts // Bioinformatics. 2005. V. 21. N 10. P. 2552–2553.

Levy A., Schwartz S., Ast G. Large-scale discovery of insertion hotspots and preferential integration sites of human transposed elements // Nucl. Acids Res. 2010. V. 38. N 5. P. 1515–1530.

Liu G., Loraine A.E., Shigeta R. *et al*. NetAffx: Affymetrix probesets and annotations // Nucl. Acids Res. 2003. V. 31. P. 82–86.

Malone J.H., Oliver B. Microarrays, deep sequencing and the true measure of the transcriptome // BMC Biol. 2011. V. 9. e34.

MAS 5.0 algorithm. Affymetrix. Statistical Algorithms Description Document. Santa Clara, CA: Affymetrix, Inc. 2002. (http://www.affymetrix.com/ support/ technical/ whitepapers/sadd-whitepaper.pdf)

Miller L.D., Smeds J., George J. *et al*. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival // Proc. Natl Acad. Sci. USA. 2005. V. 102. N 38. P. 13550–13555.

Mortazavi A., Williams B.A., McCue K. *et al*. Mapping and quantifying mammalian transcriptomes by RNA-Seq // Nat. Methods. 2008. V. 5. N 7. P. 621–628.

Nelleker C., Li F., Uhrzander F. *et al.* Expression profiling of repetitive elements by melting temperature analysis: variation in HERV-W gag expression across human individuals and tissues // BMC Genomics. 2009. V. 10. e532.

O′Donnell K.A., Burns K.H. Mobilizing diversity: transposable element insertions in genetic variation and disease // Mobile DNA. 2010. V. 1. e21.

Okoniewski M.J., Miller C.J. Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations // BMC Bioinformatics. 2006. V. 7. e276.

Orlov Y.L., Zhou J., Lipovich L. *et al.* Quality assessment of the Affymetrix U133A&B probesets by target sequence mapping and expression data analysis // In Silico Biol. 2007. V. 7. N 3. P. 241–260.

Ozsolak F., Milos P.M. RNA sequencing: advances, challenges and opportunities // Nat. Rev. Genet. 2011. V. 12. N 2. P. 87–98.

Ponicsan S.L., Kugel J.F., Goodrich J.A. Genomic gems: SINE RNAs regulate mRNA production // Curr. Opin. Genet. Dev. 2010. V. 20. N 2. P. 149–155.

Risueco A., Fontanillo C., Dinger M.E., De Las Rivas J. GATExplorer: genomic and transcriptomic explorer; mapping expression probes to gene loci, transcripts, exons and ncRNAs // BMC Bioinformatics. 2010. V. 11. e221.

Sela N., Mersch B., Gal-Mark N. *et al*. Comparative analysis of transposed element insertion within human and mouse genomes reveals Alu‘s unique role in shaping the human transcriptome // Genome Biol. 2007. V. 8. N 6. R127.

Sela N., Mersch B., Hotz-Wagenblatt A., Ast G. Characteristics of transposable element exonization within human and mouse // PLoS One. 2010. V. 5. N 6. e10907.

Shames D.S., Girard L., Gao B. *et al*. A genome-wide screen for promoter methylation in lung cancer identifies novel methylation markers for multiple malignancies // PLoS Med. 2006. V. 3. N 12. e486.

Smit A.F.A., Hubley R., Green P. RepeatMasker Open-3.0. 1996-2010 <http://www.repeatmasker.org>.

Stalteri M.A., Harrison A.P. Interpretation of multiple probe sets mapping to the same gene in Affymetrix GeneChips // BMC Bioinformatics. 2007. V. 15. P. 8–13.

Tusher V.G., Tibshirani R., Chu G. Significance analysis of microarrays applied to the ionizing radiation response // Proc. Natl Acad. Sci. USA. 2001. V. 98. P. 5116–5121.

Zhang W., Edwards A., Fan W. *et al. Alu* distribution and mutation types of cancer genes // BMC Genomics. 2011. V. 12. e157.

Zhang J., Finney R.P., Clifford R.J. *et al*. Detecting false expression signals in high-density oligonucleotide arrays by an *in silico* approach // Genomics. 2005. V. 85. P. 297–308.

STATISTICAL ESTIMATES OF TRANSPOSABLE ELEMENT EXPRESSION IN THE HUMAN GENOME BASED ON CLINICAL MICROARRAY DATA ON EXPRESSION

Yu.L. Orlov, V.M. Efimov, N.G. Orlova

Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia,
e-mail: orlov@bionet.nsc.ru; efimov@bionet.nsc.ru; orlovanina2@mail.ru

Summary

Transposable elements (TEs) of the human genome can be expressed in somatic cells in different tissues, but direct experimental information on this item is scarce. Nevertheless, the huge volume of clinical microarray data on TE expression in the human genome is sufficient to investigate statistical problems of gene annotation, noise in expression measurement, and expression of TEs associated with microarray probes. In recent years, data on gene expression in cancer tissues have been collected on the base of the Affymetrix microarray platform in diagnostics and medical studies and submitted to the Internet data repositories Gene Expression Omnibus (GEO) NCBI and ArrayExpress. Statistical estimates of TE expression were studied earlier in connection with probe design quality on expression arrays and genomic location of the probesets. It was shown that up to 25 % of the initial target sequences of Affymetrix U133 GeneChips had overlaps with TEs in the human genome. This work shows statistically significant effects of gene expression changes on probesets associated with TEs. A review of technologies for estimation of TE expression is given.

Key words: human genome, transposable elements, expression microchips, statistical estimates, transcription.