БИООРГАНИЧЕСКАЯ ХИМИЯ, 2013, том 39, № 4, с. 466–476

УДК 577.29

СТРАТЕГИЯ ПОИСКА СОМАТИЧЕСКИХ ИНСЕРЦИЙ РЕТРОЭЛЕМЕНТОВ В ГЕНОМЕ ЧЕЛОВЕКА

© 2013 г. А. А. Курносов*, С. В. Устюгова*, М. В. Погорелый*, А. Ю. Комков*, Д. А. Болотин*, К. В. Ходосевич**, И. З. Мамедов*,#, Ю. Б. Лебедев*

*ФГБУН Институт биоорганической химии им. академиков М.М. Шемякина и Ю.А. Овчинникова РАН, 117997, Москва, ГСП-7, ул. Миклухо-Маклая, 16/10

**German Cancer Research Center, 69120, Heidelberg, Germany

Поступила в редакцию 29.01.2013 г. Принята к печати 14.01.2013 г.

Активность ретроэлементов является одним из значимых факторов, приводящих к генетической изменчивости современного человека. Инсерции ретроэлементов могут являться причиной изменения экспрессии генов, а, следовательно, возникновения функциональных различий между клетками. В последние годы публикуется все больше данных, свидетельствующих о повышенной ретротранспозиционной активности в геномной ДНК некоторых тканей человека и животных. В связи с этим возникает необходимость разработки эффективной системы детекции отдельных соматических ретротранспозиционных событий. В данной работе описывается новый подход к поиску соматических инсерций ретроэлементов в геноме человека. С использованием разработанной стратегии проведен сравнительный анализ уровней соматического мозаицизма в двух тканях одного индивида. Всего было детектировано 3410 соматических инсерций ретроэлементов, принадлежащих к подгруппе AluYa5.

Ключевые слова: ретроэлементы, соматический мозаицизм, секвенирование нового поколения.

DOI: 10.7868/S013234231304012X

ВВЕДЕНИЕ

С момента первого деления зиготы, в клетках начинают накапливаться мутации, приводящие к соматическому мозаицизму организма – присутствию в его тканях генетически разнородных популяций клеток. Причинами таких мутаций могут быть повреждения ДНК свободными радикалами, включение неправильных нуклеотидов в растущую цепь ДНК-полимеразой в ходе репликации, проскальзывание полимеразы на матричной ДНК, ошибки репарации, рекомбинация и другие типы мутационных событий. Основными изменениями в ДНК, из-за которых возникает явление соматического мозаицизма, являются однонуклеотидные замены, различные типы делеций, инсерций и дупликаций, транслокации, анеуплоидия, а также инсерции ретроэлементов.

В первую очередь внимание ученых привлекли случаи соматического мозаицизма, ассоциированные с генетическими заболеваниями. Например, делеции в районе экзона 2 гена дистрофина, приводящие к мышечной дистрофии Дюшена, обычно детектируются именно у пациентов с мозаичным геномом [1], а делеции в гене *NF1* в соматических клетках являются распространенной причиной нейрофиброматоза I типа [2]. С другой стороны, нестабильность генома и возникновение большого числа различающихся генетически популяций клеток может быть следствием заболевания, особенно масштабные геномные перестройки происходят при развитии некоторых типов онкологических заболеваний [3].

В то же время соматический мозаицизм является неотъемлемым свойством тканей здорового организма и далеко не всегда ассоциирован с патологическими состояниями. Наиболее хорошо изучен мозаицизм, возникающий в ходе раннего эмбриогенеза. Исследования эмбрионов человека, находящихся в преимплантационном периоде развития, показали, что до 90% эмбрионов на стадии бластоцисты генетически гетерогенны [4]. Различия в числе копий некоторых участков генома наблюдается у монозиготных близнецов, что является следствием перестроек структуры ДНК на ранних эмбриональных стадиях [5].

Сокращения: NGS – секвенирование нового поколения (англ. next generation sequencing).

Автор для связи (тел./факс: +7 (495) 330-42-88; эл. почта: Imamedov@mx.ibch.ru).

Одним из существенных факторов, обеспечивающих генетическую изменчивость, является активность ретротранспозонов. Ретротранспозоны – класс мобильных элементов, число копий которых увеличивается за счет механизма копирования–вставки, осуществляемого через РНК-интермедиат, с участием обратной транскриптазы [6]. Они составляют около половины генома млекопитающих, в частности, на долю ретроэлементов приходится около 42% генома человека [7]. На протяжении эволюционной истории были активны различные группы ретротранспозонов, в настоящий момент в человеческом геноме сохранили активность мобильные элементы, принадлежащие к некоторым подсемействам L1, Alu-повторов, а также SVA-ретроэлементы. Транспозиционно активные мобильные элементы относятся к молодым с эволюционной точки зрения группам: 60–80 активных ретротранспозонов относятся к подсемейству L1HS семейства L1 [8], большинство транспозиционно активных ретроэлементов семейства Alu, специфического для приматов, относится к подсемействам AluYa5 и AluYb8 [9]. Такие ретротранспозоны являются источником полиморфных инсерций, а также *de novo* интеграций, возникающих в половых клетках [10, 11]. Высокая активность ретроэлементов, приводящая к генетическому мозаицизму, может быть причиной функциональных различий между клетками. Инсерции мобильных элементов могут приводить как к нарушению функций кодирующих участков ДНК, так и к изменению функциональной активности регуляторных последовательностей [12], к появлению новых сайтов связывания транскрипционных факторов [13], сплайс-сайтов [14] и сигналов полиаденилирования [15]. Одним из следствий встраивания ретроэлементов в гены и их регуляторные области являются различные генетические заболевания, а также рак [16].

В последнее время получили развитие исследования, результаты которых свидетельствуют о возникновении в эмбриогенезе соматических инсерций ретроэлементов [17]. Экспериментальные доказательства возможности ретротранспозиционной активности в эмбриональных тканях были получены на крысиных и мышиных моделях с использованием экспрессионного конструкта, несущего кассету – индикатор ретротранспозиции [18]. Менее подробно изучен соматический мозаицизм, возникающий в здоровых тканях взрослого организма. Однако уже опубликован ряд доказательств повышенной активности мобильных элементов в нейрональных тканях, полученных в экспериментах по трансфекции предшественников нервных клеток крысы и человека экспрессионным конструктом, содержащим функциональную копию L1, который несет в 3'-UTR кассету – индикатор ретротранспозиции [19, 20]. Также ретротранспозиции детектировались в нейронах трансгенных мышей, в геном которых встроен функционально активный человеческий L1 ретротранспозон, несущий репортерный ген *EGFP* [19].

Для оценки вклада ретротранспозиционной активности в пластичность генома соматических клеток и ее влияния на экспрессию генов в этих клетках необходимо полногеномное картирование соматических инсерций ретроэлементов. Ранее были разработаны и успешно применялись методы идентификации новых полиморфных интеграций ретроэлементов [9, 21, 22], однако для поиска соматических инсерций, присутствующих в небольшом числе клеток, использование этих методов невозможно, поскольку концентрация таких инсерций в образце ДНК крайне мала. В связи с этим приобретает актуальность задача разработки чувствительного метода, который позволяет детектировать инсерции ретроэлементов, количество копий которых в образце ДНК невелико, по сравнению с количеством копий ретроэлементов, представленных во всех клетках организма.

В данной работе мы предлагаем один из вариантов решения такой задачи. Нами создан новый экспериментальный подход к идентификации соматических инсерций и оценке уровня транспозиционной активности мобильных элементов в различных тканях организма. Этот подход был применен для поиска соматических инсерций ретротранспозонов, относящихся к семейству Alu. Было показано, что в соматических тканях человека возникает большое количество таких ретротранспозиций. Результаты применения разработанного подхода свидетельствуют о его высокой эффективности и специфичности.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Предлагаемый нами подход заключается в сочетании трех следующих друг за другом этапов: получение полногеномных библиотек фрагментов ДНК, фланкирующих инсерции Alu-элементов; проведение крупномасштабного секвенирования созданных библиотек с использованием NGS-платформ (Illumina); применение специализированного алгоритма обработки результатов секвенирования для идентификации соматических инсерций Alu-элементов. Такой подход позволяет выявить отличия в профилях интеграций ретроэлементов между ДНК различных тканей одного организма.

Создание библиотек фрагментов ДНК, фланкирующих Alu-ретроэлементы

Первым этапом разработанного подхода, направленного на идентификацию соматических инсерций ретроэлементов, была подготовка пол-

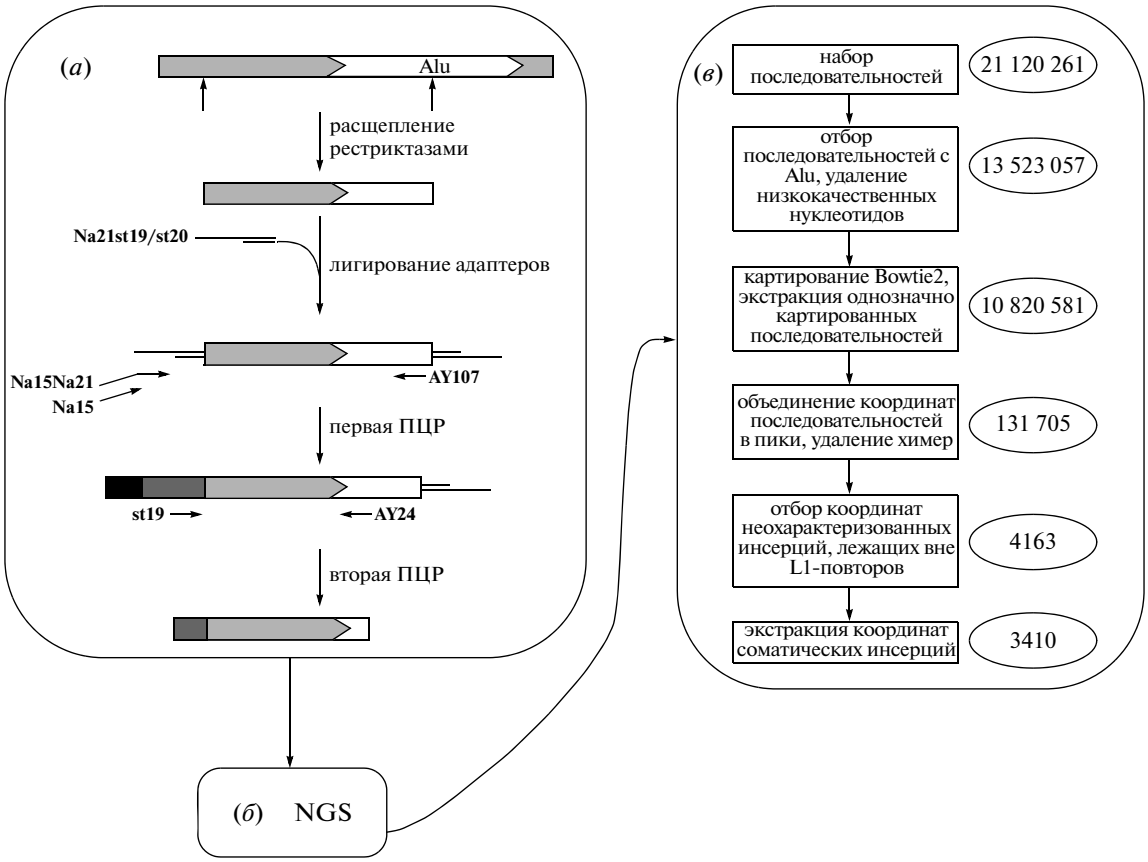

Рис. 1. Стратегия идентификации соматических инсерций ретроэлементов. (*а*) – Схема приготовления библиотеки последовательностей, фланкирующих Alu-ретроэлементы. Нуклеотидные последовательности Alu-ретроэлемента или его фрагмента показаны белой фигурной стрелкой с надписью "Alu" или фрагментами фигурной стрелки; фланкирующие последовательности геномной ДНК показаны светло-серыми прямоугольниками, а последовательности, соответствующие супрессионным адаптерам – темно-серыми и черными прямоугольниками. Вертикальными стрелками в верхней части блока отмечены сайты узнавания рестриктаз. Используемые праймеры показаны горизонтальными или ломаными стрелками с названием праймера; (*б*) – NGS секвенирование; (*в*) – протокол обработки данных для выявления координат потенциально соматических инсерций. В прямоугольных рамках показаны стадии протокола, в овальных рамках – суммарное для всех библиотек число секвенированных последовательностей или координат, полученных на каждой стадии.

ногеномных библиотек фрагментов ДНК, непосредственно прилегающих к участкам интеграции всех копий изучаемых ретроэлементов. Приготовление таких библиотек позволило резко повысить концентрацию молекул ДНК, несущих информацию о координатах инсерций Alu в геноме каждой из клеток анализируемого образца и, следовательно, существенно увеличить информативность результатов последующего секвенирования. Процесс подготовки библиотеки включал стадии фрагментации геномной ДНК эндонуклеазами рестрикции, лигирования супрессионных адаптеров к полученным фрагментам и две стадии супрессионной ПЦР с праймерами, комплементарными последовательности Alu-ретроэлементов и последовательности лигированного олигонуклеотидного адаптера (см. рис. 1*а*).

Поскольку длина молекул ДНК для секвенирования по технологии Illumina должна составлять 100–600 п.о., фрагментацию ДНК на первой стадии проводили частощепящими рестриктазами, сайты узнавания которых имеют длину 4 п.о. Кроме того, для повышения репрезентативности полученных библиотек использовали две рестриктазы – AluI и RsaI.

К полученным рестрикционным фрагментам лигировали олигонуклеотидный адаптер Na21st19/st20 (табл. 1), несущий участок, которому соответствует используемый во втором раунде ПЦР праймер st19 (табл. 1). ПЦР проводили с использованием технологий внешнего праймирования [23] и ПЦР-супрессии [24] (см. рис. 1*а*). Для получения репрезентативной библиотеки в ходе амплификации использовали высокопроцессивную полимеразу, способную с одинаковой эффективностью синтезировать молекулы ДНК разной длины. Первую ПЦР на полученных фрагментах проводили с использованием следующей комби-

Таблица 1. Олигонуклеотиды, использованные в работе

Название	Нуклеотидная последовательность, 5'–3'
AluY-специфические праймеры	
AY107	TCACCGTTTTAGCCGGGA
AY98	TAGCCGGGATGGTCTCGAT
AY24	AGGCGTGAGCCACCGCGC
Супрессионные праймеры и адаптеры	
Na21st19	TGTAGCGTGAAGACGACAGAAAGGGCGTGGTGCGGAGGGCGGT
st20	ACCGCCCTCC
Na15Na21	AGCAGCGAACTCAGTACAACATGTAGCGTGAAGACGACAGAA
Na15	AGCAGCGAACTCAGTACAACA
st19	AGGGCGTGGTGCGGAGGGCGGT
Локус-специфические праймеры	
Ins1 Rev–out	TACCATTTCTACCTTATCATTTCTG
Ins1 Rev	GAGGAAAATTCTTCAGTTCTCAAA
Ins2 Rev–out	GATCTTTCTAATACTTCCTTTGTGG
Ins2 Rev	ACATGAAACTGTTGGAGGCTTG
Ins3 Rev–out	AGCAACTTATGTATGTATGTAGCAG
Ins3 Rev	TAGATGCTTTTCTTTCAAATTTTTA
Ins4 Rev–out	ATGGACATGTGGCCTTTGAG
Ins4 Rev	GCAGCCACACAAGCAAATACT
Ins5 Rev–out	TTCCTTTCCTGTGTCATCTGTC
Ins5 Rev	GTTTAAGCATCTCCATCAGCAC
Ins6 Rev–out	AATTTGTTCTGCCATTGCTTT
Ins6 Rev	CAAACAAATACACAAAACTCACTAAT
Ins7 Rev–out	GAAAATGGATGTTCTAAAAGCAC
Ins7 Rev	CAGATGAACAGCCAACATAGGT
Ins8 Rev–out	GCTTGGTGTGATACAAATCCTG
Ins8 Rev	GAATCAAGGTCCCAAGCCAG
Ins9 Rev–out	GTTCACACAGCACCAGGCAC
Ins9 Rev	CTGTGGTGAGCAGGTGCC
Ins10 Rev–out	CCTGTCTGTCTTCCTTGAGCC
Ins10 Rev	GTATTGAGTTACCGTGCCTCTGA
Ins11 Rev–out	TACAGGAGTTGTAAGGGAATAACA
Ins11 Rev	AGATAATAGAATTAGGGAGAGACTGC
Ins12 Rev–out	CATGCTTGGCTCCAGTGCT
Ins12 Rev	CAGTGGAGGAACATTGGTGTCT
Ins13 Rev–out	GGCACCTTCCACTGTACTCTG
Ins13 Rev	CTCCACAAAGCGACATCTTGA
Ins14 Rev–out	CAACAGACTGTGACTTGAACTGC
Ins14 Rev	CCTGAGAATCAGCCAAGCCT
Ins15 Rev–out	TTGGAAGGTGGATGAGGAGG
Ins15 Rev	GGGATAGATGGAAAGCAGGAA
Ins16 Rev–out	CCTCATTCTAGATCAGTGGAGCC
Ins16 Rev	GCAGGAGATTCAGAATGGAAAC

нации олигонуклеотидов (см. табл. 1): 1) праймер AY107, комплементарный нуклеотидной последовательности ретроэлементов подсемейства AluYa5; 2) 40-звенный праймер Na15Na21, 3'-концевая часть которого соответствует 5'-концу лигированного на предыдущей стадии адаптера Na21st19; 3) 20-звенный праймер Na15, который соответствует 5'-концевой части праймера Na15Na21. Длинные адаптерные последовательности на концах одноцепочечной молекулы ДНК, образующейся на стадии денатурации ПЦР, формируют дуплекс, за счет чего происходит ингибирование ПЦР (ПЦР-супрессия). Реакция амплификации проходит с высокой эффективностью только на молекулах матрицы, в которых содержится сайт отжига специфического праймера. В данной работе использовали специфический праймер AY107, последовательность которого комплементарна диагностическому участку консенсусной последовательности ретроэлементов транспозиционно активного семейства AluYa5. Такой праймер позволяет распознавать ретротранспозоны данной группы среди других Alu-ретроэлементов в ходе амплификации.

Вторая ПЦР с продуктами первой реакции в качестве матрицы необходима для повышения селективности амплификации, праймеры для этой ПЦР заглублены относительно праймеров, используемых в первой реакции. Специфический праймер AY24 комплементарен 5'-концевой последовательности Alu-ретроэлемента, а праймер st19 соответствует 3'-концу адаптера Na21st19.

Продуктом двух раундов селективной амплификации является библиотека последовательностей ДНК, содержащих 5'-концевой фрагмент Alu-ретроэлемента и прилегающую к его 5'-концу геномную последовательность, ограниченную ближайшим сайтом расщепления рестриктаз AluI или RsaI, примененных для фрагментации ДНК. В данном случае использование 5'-концевого фрагмента предпочтительно, поскольку фланкирующая последовательность, прилегающая к 3'-концу Alu-ретроэлемента, может содержать длинный poly(A)-хвост, структура которого не информативна для последующего картирования сайта интеграции.

Секвенирование библиотек фрагментов ДНК, фланкирующих Alu-ретроэлементы

Для секвенирования библиотек использовали технологию секвенирования нового поколения (NGS) Illumina (рис. 1*б*). Преимущество этой технологии перед секвенированием по Сенгеру и большинством других методов NGS заключается в большом числе нуклеотидных последовательностей, которые могут быть прочитаны за один запуск прибора. Поскольку соматические инсерции присутствуют в малом числе клеток, концентрация соответствующих им фланкирующих последовательностей в получаемой библиотеке будет существенно ниже концентрации геномных фланков тех инсерций AluYa5, которые присутствуют во всех клетках организма. Таким образом, детекция соматических инсерций возможна только при большой глубине секвенирования. Современные приборы Illumina позволяют прочитывать до 100 миллионов индивидуальных последовательностей небольшой длины (до 101 п.о.) в ходе одной реакции параллельного секвенирования. Такая длина в большинстве случаев является достаточной для однозначного картирования последовательности, фланкирующей интеграцию ретроэлемента, в референсном геноме человека.

Анализ результатов NGS

Для анализа результатов секвенирования и идентификации соматических инсерций, которые невозможно провести при помощи существующих стандартных пакетов программ для обработки данных NGS, мы создали специализированный протокол обработки данных (см. рис. 1*в*). На первом этапе для дальнейшего анализа были выбраны только те последовательности, в которых на 5'-конце была идентифицирована последовательность Alu-ретроэлемента. С 3'-концов этих последовательностей были удалены фрагменты, прочитанные с низким качеством. Также с 5'-конца секвенированных последовательностей удаляли фрагмент длиной 24 нт., представляющий 5'-концевой участок Alu-ретроэлемента, поскольку такие фрагменты нуклеотидной последовательности соматических и новых полиморфных инсерций отсутствуют в референсном геноме и не могут быть картированы.

Для картирования обработанных подобным образом секвенированных последовательностей была использована программа Bowtie2, которая позволяет быстро проанализировать большое количество данных [25]. В качестве референсного генома человека использовали сборку hg19 (http://www.ncbi.nlm.nih.gov/assembly/2758/). Полученные с помощью Bowtie2 координаты последовательностей, картированных однозначно (т.е. соответствующих единственному участку генома), объединяли в упорядоченные таблицы. Из числа координат были исключены те, которые лежат в L1-элементах, поскольку в большинстве случаев достоверность такого картирования невелика. Это связано с присутствием в геноме большого числа копий L1-элементов, обладающих высокой гомологией нуклеотидных последовательностей. В подавляющем большинстве случаев расположенные рядом (в пределах 1–30 позиций) координаты относились к инсерции одного и того же ретроэлемента. Появление таких групп координат, расположенных рядом и представля-

ющих одну инсерцию, является следствием особенностей алгоритма Bowtie2, проявляющихся при картировании нуклеотидных последовательностей, несущих однонуклеотидные замены, короткие инсерции или делеции, образовавшиеся в ходе пробоподготовки и секвенирования. Координаты, расположенные внутри "окна" размером 30 нт., объединяли в пики, которым присваивали координату, встретившуюся наибольшее число раз (см. рис. 2). Из списка координат удаляли те, для которых в промежутке от –4 до +50 нуклеотидов от координаты сайта интеграции в геноме присутствовал сайт рестрикции фермента, с помощью которого готовился образец ДНК, поскольку, как было показано, большинство таких координат представляло химерные молекулы ДНК. Для выявления в полученном списке координат известных инсерций эти координаты сопоставляли с координатами инсерций Alu-ретроэлементов, присутствующих в референсном геноме hg19, а также в базах данных полиморфных ретроэлементов PRED и dbRIP [26, 27]. Те координаты, которые не принадлежат известным инсерциям Alu-ретроэлементов, могут соответствовать координатам ранее неизвестных полиморфных инсерций, либо координатам *de novo*-инсерций, возникших в половых клетках родителей данного индивида или на ранних этапах эмбриогенеза, либо координатам соматических инсерций. Те из них, которые идентифицировали в списке координат инсерций, принадлежащих только одному из исследованных органов, рассматривали в дальнейшем как координаты потенциально соматических инсерций.

Применение разработанного подхода для поиска соматических инсерций Alu-ретроэлементов

Разработанный подход был применен для полногеномного определения профилей инсерций ретроэлементов семейства AluYa5 в двух различных тканях одного индивида. Были получены библиотеки фланкирующих последовательностей Alu-ретроэлементов, присутствующих в ДНК, выделенной из образцов тканей головного мозга и миокарда. В результате секвенирования библиотек всего с высоким качеством было получено 21 120 261 нуклеотидных последовательностей, из них – 9 622 894 для библиотек, приготовленных из ДНК мозга, и 11 457 367 для библиотек, приготовленных из ДНК миокарда. Всего было однозначно картировано 10 856 703 нуклеотидных последовательности: 2 824 539, 4 463 086, 2 233 097 и 1 299 859 для библиотек "мозг AluI", "миокард AluI", "мозг RsaI" и "миокард RsaI" соответственно (см. табл. 2).

Все координаты, идентифицированные как потенциальные сайты соматических инсерций ретроэлементов, были представлены небольшим

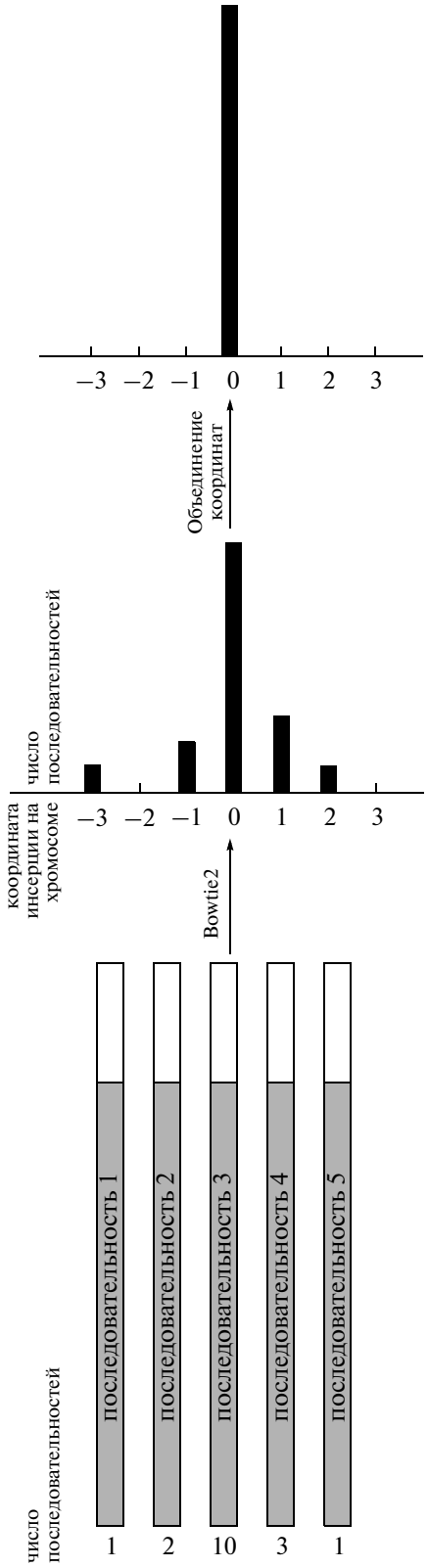

Рис. 2. Пример использования алгоритма слияния координат, возникающих в результате ошибок картирования нуклеотидных последовательностей, фланкирующих одну инсерцию Alu-ретроэлемента.

Таблица 2. Характеристика библиотек последовательностей, фланкирующих инсерции Alu-ретроэлементов

Библиотека	“Мозг AluI”	“Миокард AluI”	“Мозг RsaI”	“Миокард RsaI”
Число высококачественных секвенированных последовательностей	3531061	5575717	2797580	1618699
Число однозначно картированных последовательностей	2824539	4463086	2233097	1299859
Число последовательностей, приходящихся на соматические инсерции (% от картированных)	1239 (0.044)	1506 (0.034)	775 (0.035)	255 (0.020)
Число соматических инсерций	1168	1318	703	233

числом (1–8) секвенированных последовательностей, причем более 95% – одной последовательностью. Доля последовательностей, представляющих соматические инсерции, оказалась выше в ДНК, выделенной из мозга (в среднем по двум библиотекам 0.0395% общего числа секвенированных последовательностей), чем в ДНК, выделенной из миокарда (в среднем по двум библиотекам 0.027%).

Оценка специфичности разработанного подхода

Проведенный анализ данных позволяет оценить специфичность разработанного метода. Подавляющее большинство секвенированных последовательностей (10 222 450) представляет известные Alu-ретроэлементы. Большая доля однозначно картированных последовательностей (84.3%) приходится на фланкирующие последовательности ретроэлементов подсемейства AluYa5, 5.8% представляют ранее неизвестные инсерции, которые отсутствуют в референсном геноме и базах данных полиморфных инсерций. Остальные подсемейства Alu-ретроэлементов составляют в общей сложности 9.9%. Анализ нескольких высокопредставленных в исследуемых библиотеках последовательностей, относящихся к старым семействам Alu-ретроэлементов, показал, что место посадки праймеров в них отсутствует, а точка интеграции не совпадает с началом Alu-ретроэлемента. Вероятно, такие координаты соответствуют новым интеграциям молодых Alu-ретроэлементов в старые повторы. Большинство же координат инсерций ретроэлементов, относящихся к этим семействам, представлено значительно меньшим числом секвенированных последовательностей и, таким образом, составляет совсем незначительный процент от общего числа картированных последовательностей. Таким образом, не менее 90% последовательностей в полученных библиотеках представляют собой те, которые фланкируют инсерции AluYa5-ретроэлементов, что доказывает высокую специфичность использованного метода селективной амплификации (рис. 3).

Оценка эффективности разработанного подхода

Эффективность метода может быть определена, как доля ранее известных AluYa5-ретроэлементов, сайты интеграции которых представлены

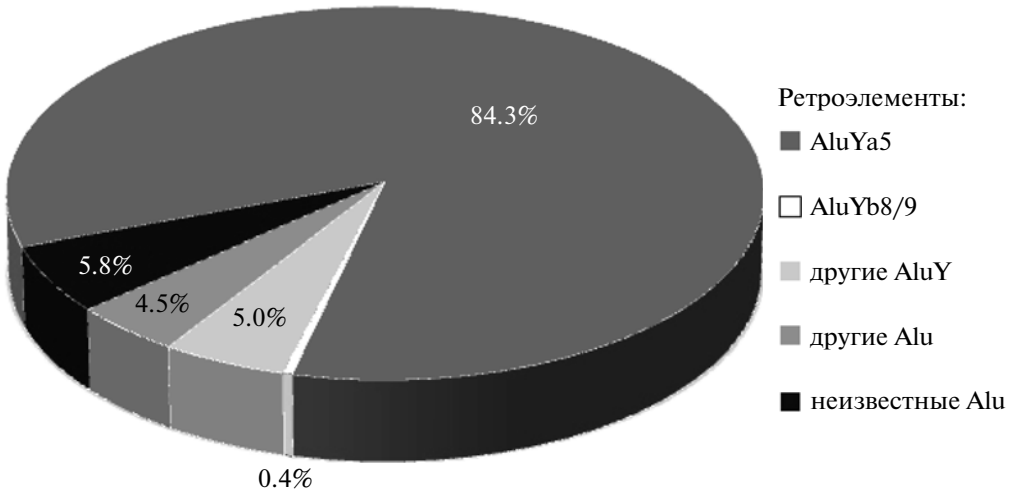

Рис. 3. Диаграмма, отражающая процентное соотношение последовательностей, фланкирующих Alu-ретроэлементы, принадлежащие разным семействам. “Неизвестные Alu-ретроэлементы” – идентифицированные в ходе нашей работы и ранее неохарактеризованные инсерции Alu-ретроэлементов.

Таблица 3. Характеристика выявленных инсерций Alu-ретроэлементов

Ретроэлементы	AluYa5	AluYa8	AluYb8	AluY
Число в референсном геноме hg19	3858	338	2854	120617
Детектировано	2745	40	1325	34392
Детектировано, %	71.1	11.8	46.4	28.5

в конечном наборе координат. В последовательности референсного генома hg19 присутствует 3858 копий AluYa5-ретроэлементов, часть из которых являются полиморфными и могут отсутствовать в геноме конкретного индивида, однако точно определить количество таких инсерций невозможно. В анализируемых библиотеках было детектировано 2745 инсерций AluYa5-ретроэлементов (см. табл. 3). В базах данных полиморфных инсерций ретроэлементов присутствуют 229 инсерций, которые не были найдены в анализируемых библиотеках, следовательно, они могут отсутствовать в геноме исследованного индивида.

Некоторые инсерции, не найденные с помощью нашего подхода, были проанализированы более детально. В 26 случаях из проанализированных 30 в последовательности Alu-ретроэлементов отсутствовали сайты отжига одного или двух праймеров, использованных для амплификации, либо они были сильно дивергированы; в двух случаях сайты, распознаваемые рестриктазами, используемыми при пробоподготовке, располагались далеко от инсерции, и полученные в ходе рестрикции молекулы были слишком длинными для секвенирования по технологии Illumina. В оставшихся двух случаях причину, по которым эти ретроэлементы не были детектированы, установить не удалось, однако, можно предположить, что они являются полиморфными инсерциями, несмотря на то, что отсутствуют в базах данных известных полиморфных инсерций. Таким образом, наш метод позволяет детектировать по меньшей мере 71%, либо, вероятно, даже большую долю AluYa5-ретроэлементов, присутствующих в исследуемом образце геномной ДНК. В табл. 3 также приведены данные о числе представителей некоторых других подсемейств Alu.

Валидация соматических инсерций Alu

Валидацию кандидатных инсерций осуществляли с помощью ПЦР с вложенными праймерами. Поскольку все соматические инсерции были представлены очень небольшим числом последовательностей, вероятно, что каждая из них присутствовала лишь в малом числе клеток, либо даже в одной клетке анализируемой ткани, и амплификация данных инсерций с геномной ДНК в большинстве случаев была невозможна. Поэтому в качестве матрицы использовали продукты первой супрессионной ПЦР, которые содержат достаточно длинный фрагмент Alu-ретроэлементов (см. рис 1). Для валидации инсерций в качестве прямых использовали праймеры AY107 и AY98, специфичные к консенсусной последовательности AluY-ретроэлементов, а в качестве обратных – локус-специфические праймеры Ins Rev-out и Ins Rev, комплементарные последовательности 5'-фланкирующего участка ретроэлемента (рис. 4). Структуру ПЦР-продуктов подтверждали с помощью секвенирования по Сенгеру. Для валидации было выбрано 16 инсерций Alu, из которых было подтверждено 11.

Таким образом, нами была разработана система, позволяющая осуществлять полногеномный поиск соматических инсерций ретроэлементов. Данные о структуре фланкирующих последовательностей Alu-ретроэлементов и составе библиотек позволили сделать вывод о высокой специфичности и эффективности предложенного метода, которые также были подтверждены дополнительными экспериментами по валидации детектированных инсерций. Данный метод может успешно применяться для определения специфического набора соматических инсерций и характера их распределения в тканях человече-

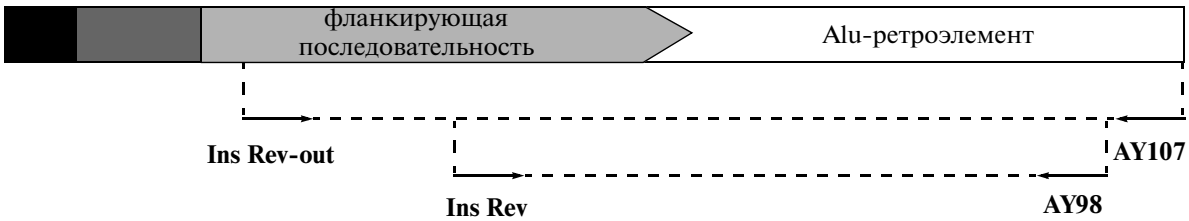

Рис. 4. Схема расположения праймеров для валидации соматических инсерций. Стрелками показаны праймеры, горизонтальными пунктирными отрезками – ПЦР-фрагменты. Для первой реакции использовались праймеры AY107 и Ins Rev-out, для второй – праймеры AY98 и Ins Rev.

ского организма. Особенно перспективными представляются подобные исследования тканей мозга, в которых наблюдается повышенная активность ретроэлементов. Модификации данного метода могут использоваться для детекции инсерций других активных семейств ретротранспозонов, таких как L1 и SVA.

ЭКСПЕРИМЕНТАЛЬНАЯ ЧАСТЬ

Получение образцов геномной ДНК

В работе использованы образцы геномной ДНК, выделенные из тканей одного индивида: мозга (*gyrus dentatus*) и миокарда. Образец ткани массой 100–200 мг гомогенизировали в жидком азоте и инкубировали в 3 мл лизирующего буфера (60 мМ Трис, 100 мМ EDTA, 0.5% SDS) с 0.4 мг РНКазы А (Qiagen, США) при 37°С в течение 1 ч, затем добавляли 1.5 мг протеиназы К (Qiagen, США) и инкубировали в течение еще 10 ч при 50°С. ДНК выделяли фенол-хлороформным методом. К лизату ткани добавляли равный объем фенола (pH 8.0), аккуратно перемешивали в течение 30 мин и центрифугировали в течение 15 мин при 4500 *g*. Водную фазу отбирали, добавляли к ней равный объем смеси фенол–хлороформ (1 : 1), аккуратно перемешивали и центрифугировали в прежнем режиме. Водную фазу отбирали, добавляли к ней хлороформ, перемешивали и центрифугировали. Геномную ДНК из водной фазы осаждали этанолом.

Создание библиотеки последовательностей, фланкирующих Alu-ретроэлементы

Первый этап пробоподготовки – фрагментирование геномной ДНК эндонуклеазами рестрикции AluI или RsaI (Fermentas, Литва). Геномную ДНК (2 мкг) инкубировали при 37°С в течение 12 ч с 20 ед. акт. рестриктазы в буфере Y Tango (Fermentas, Литва). Очистку продуктов рестрикции осуществляли с помощью QIAquick PCR Purification Kit (Qiagen, США).

К рестриктным фрагментам лигировали адаптерные олигонуклеотиды. Предварительно адаптерный олигонуклеотид Na21st19 гибридизовали с олигонуклеотидом-подложкой st20 (структуры олигонуклотидов приведены в табл. 1). Гибридизацию проводили в 4 мкл, смесь содержала олигонуклеотиды (5 мкМ каждый) 10 мМ Трис-HCl, 10 мМ $MgCl_2$. Смесь инкубировали в течение 5 мин при 68°С и охлаждали до 15°С с шагом 0.1°С/с. Лигазная смесь объемом 10 мкл содержала 120 пмоль дуплекса Na21st19/st20 и 6 ед. акт. ДНК-лигазы Т4 в реакционном буфере (Promega, США) и 2 мкг рестрицированной геномной ДНК. Лигирование проводили в течение 14 ч при +16°С.

Амплификацию ДНК для приготовления библиотек проводили в ходе двух последовательных супрессионных ПЦР. Структуры всех олигонуклеотидов приведены в таблице 1. В ходе первой ПЦР реакционная смесь объемом 25 мкл содержала 1/50 от общего количества продуктов лигирования, 0.4 мкМ праймер AY107, 0.16 мкМ праймер Na15, 0.02 мкМ праймер Na15Na21, каждый из dNTP – 0.125 мкМ, 1 ед. акт. полимеразы Encyclo (Evrogen, Россия) в реакционном буфере. Температурный профиль реакции был следующим: предварительная достройка концов в течение 4 мин при 72°С, затем 13 циклов по 20 с при 94°С, 15 с при 65°С, 1 мин при 72°С, затем финальная элонгация в течение 2 мин при 72°С. Для наработки библиотеки проводили 50 идентичных реакций. Продукты ПЦР объединяли, очищали с помощью колонок QIAquick PCR Purification Kit и концентрировали до объема 120 мкл.

В ходе второй ПЦР реакционная смесь объемом 25 мкл содержала 1/1200 объема объединенных продуктов первой реакции, 0.4 мкМ праймер AY24, 0.16 мкМ праймер st19, 0.125 мкМ каждый из dNTP, 1 ед. акт. полимеразы Encyclo в реакционном буфере. Температурный профиль реакции был следующим: 10 циклов по 20 с при 94°С, 15 с при 68°С, 1 мин при 72°С, затем финальная элонгация в течение 2 мин при 72° С. Для наработки библиотеки проводили 30 идентичных реакций. Продукты ПЦР объединяли, очищали с помощью колонок QIAquick PCR Purification Kit и концентрировали до объема 60 мкл.

Концентрацию ДНК в образцах измеряли с помощью прибора Qubit 2.0. Для секвенирования отбирали 500 нг образца, полученного с использованием каждой из рестриктаз, что составляло около 1/5 общего количества ДНК в образце.

Секвенирование фланкирующих последовательностей Alu-ретроэлементов

Нуклеотидные последовательности, фланкирующие Alu-ретроэлементы длиной 101 п.о. были определены с помощью прибора Illumina HiSeq 2000.

Картирование последовательностей и анализ данных секвенирования

Протокол анализа данных включал использование стандартных пакетов программ Bowtie2 и Galaxy [28–30] и компьютерных программ, написанных специально для этой работы на языке программирования Perl. Биоинформатический анализ состоял из следующих этапов:

1. Выявление последовательностей, содержащих Alu-ретроэлементы.

2. Удаление фрагментов последовательностей, структура которых была определена с низкой достоверностью.

3. Удаление с 5'-концов последовательностей фрагментов Alu-ретроэлементов и картирование в референсном геноме человека (hg19) осуществляли с помощью программы Bowtie2. Настройки, отличающиеся от настроек по умолчанию: -p 8 –mp 12.8 -5 24.

4. Экстракция однозначно картированных последовательностей из файлов выдачи данных программы Bowtie2.

5. Построение таблиц координат последовательностей и объединение их в пики.

6. Удаление координат химерных последовательностей из таблиц.

7. Сопоставление полученных координат инсерций с координатами известных Alu, присутствующих в hg19 и базах данных полиморфных ретроэлементов и координатами L1 осуществляли с помощью инструмента Join, входящего в пакет программ Galaxy.

Валидация соматических инсерций

Для валидации выбранных соматических инсерций Alu-ретроэлементов использовали ПЦР с вложенными праймерами. В качестве матрицы для первой реакции были взяты аликвоты продуктов первой супрессионной ПЦР. Реакционная смесь объемом 25 мкл содержала 1 мкл матричной ДНК, 0.4 мкМ праймер AY107, 0.4 мкМ праймер Primer-Rev out, 0.125 мкМ каждый из dNTP, 1 ед. акт. полимеразы HS Taq (Evrogen, Россия) в реакционном буфере. Температурный профиль реакции был следующим: предварительный прогрев в течение 5 мин при 95°C, затем 22 цикла по 20 с при 94°C, 20 с при 63°C, 40 с при 72°C. В качестве матрицы для вложенной ПЦР были взяты продукты первой реакции, разведенные в 100 раз. Реакционная смесь объемом 25 мкл содержала 1 мкл матричной ДНК, 0.4 мкМ праймер AY98, 0.4 мкМ праймер Primer-Rev, 0.125 мкМ каждый из dNTP, 1 ед. акт. полимеразы HS Taq (Evrogen, Россия) в реакционном буфере. Температурный профиль реакции был следующим: предварительный прогрев в течение 5 мин при 95°C, затем 28 циклов по 20 с при 94°C, 20 с при 63°C, 40 с при 72°C. Полученные продукты реакций были секвенированы (Evrogen, Россия) на автоматическом секвенаторе ABI PRISM® 3100 Genetic Analyzer.

БЛАГОДАРНОСТИ

Работа выполнена при поддержке грантами РФФИ №№ 12-04-33065 и 11-04-01159.

СПИСОК ЛИТЕРАТУРЫ

1. *Passos-Bueno M.R., Bakker E., Kneppers A.L., Takata R.I., Rapaport D., den Dunnen J.T., Zatz M., van Ommen G.J.* // Am. J. Hum. Genet. 1992. V. 51. P. 1150–1155.
2. *Petek E., Jenne D.E., Smolle J., Binder B., Lasinger W., Windpassinger C., Wagner K., Kroisel P.M., Kehrer-Sawatzki H.* // J. Med. Genet. 2003. V. 40. P. 520–525.
3. *Lengauer C., Kinzler K.W., Vogelstein B.* // Nature. 1998. V. 396. P. 643–649.
4. *Bielanska M., Tan S.L., Ao A.* // Hum. Reprod. 2002. V. 17. P. 413–419.
5. *Bruder C.E., Piotrowski A., Gijsbers A.A., Andersson R., Erickson S., Diaz de Stahl T., Menzel U., Sandgren J., von Tell D., Poplawski A., Crowley M., Crasto C., Partridge E.C., Tiwari H., Allison D.B., Komorowski J., van Ommen G.J., Boomsma D.I., Pedersen N.L., den Dunnen J.T., Wirdefeldt K., Dumanski J.P.* // Am. J. Hum. Genet. 2008. V. 82. P. 763–771.
6. *Wicker T., Sabot F., Hua-Van A., Bennetzen J.L., Capy P., Chalhoub B., Flavell A., Leroy P., Morgante M., Panaud O., Paux E., SanMiguel P., Schulman A.H.* // Nat. Rev. Genet. 2007. V. 8. P. 973–982.
7. *Bannert N., Kurth R.* // Proc. Natl. Acad. Sci. U.S.A. 2004. V. 101. P. 14 572–14 579.
8. *Feng Q., Moran J.V., Kazazian H.H., Boeke J.D.* // Cell. 1996. V. 87. P. 905–916.
9. *Mamedov I.Z., Arzumanyan E.S., Amosova A.L., Lebedev Y.B., Sverdlov E.D.* // Nucleic Acids Res. 2005. V. 33. P. e16.
10. *Cordaux R., Hedges D.J., Herke S.W., Batzer M.A.* // Gene. 2006. V. 373. P. 134–137.
11. *Mills R.E., Bennett E.A., Iskow R.C., Devine S.E.* // Trends Genet. 2007. V. 23. P. 183–191.
12. *Deininger P.L., Batzer M.A.* // Mol. Genet. Metab. 1999. V. 67. P. 183–193.
13. *Polak P., Domany E.A.* // BMC Genomics. 2006. V. 7. P. 133.
14. *Belancio V.P., Hedges D.J., Deininger P.* // Nucleic Acids Res. 2006. V. 34. P. 1512–1521.
15. *Chen C., Ara T., Gautheret D.* // Mol. Biol. Evol. 2009. V. 26. P. 327–334.
16. *Callinan P.A., Batzer M.A.* // Genome Dyn. 2006. V. 1. P. 104–115.
17. *van den Hurk J.A., Meij I. C., Seleme M.C., Kano H., Nikopoulos K., Hoefsloot L.H., Sistermans E.A., de Wijs I.J., Mukhopadhyay A., Plomp A.S.* // Hum. Mol. Genet. 2007. V. 16. P. 1587–1592.
18. *Kano H., Godoy I., Courtney C., Vetter M.R., Gerton G.L., Ostertag E.M., Kazazian H.H.* // Genes Dev. 2009. V. 23. P. 1303–1312.
19. *Muotri A.R., Chu V.T., Marchetto M.C., Deng W., Moran J.V., Gage F.H.* // Nature. 2005. V. 435. P. 903–910.
20. *Coufal N.G., Garcia-Perez J.L., Peng G.E., Yeo G.W., Mu Y., Lovci M.T., Morell M., O'Shea K.S., Moran J.V., Gage F.H.* // Nature. 2009. V. 460. P. 1127–1131.
21. *Mamedov I., Batrak A., Buzdin A., Arzumanyan E., Lebedev Y., Sverdlov E.D.* // Nucleic Acids Res. 2002. V. 30. P. e71.

22. *Mamedov I., Lebedev Y., Hunsmann G., Khusnutdinova E., Sverdlov E.* // Genomics. 2004. V. 84. P. 597–600.

23. *Garcia-Perez J.L., Marchetto M.C., Muotri A.R., Coufal N.G., Gage F.H., O'Shea K.S., Moran J.V.* // Hum. Mol. Genet. 2007. V. 16. P. 1569–1577.

24. *Diatchenko L., Lau Y.F., Campbell A.P., Chenchik A., Moqadam F., Huang B., Lukyanov S., Lukyanov K., Gurskaya N., Sverdlov E.D., Siebert P.D.* // Proc. Natl. Acad. Sci. U.S.A. 1996. V. 93. P. 6025–6030.

25. *Langmead B., Salzberg S.L.* // Nat. Methods. 2012. V. 9. P. 357–359.

26. *Мамедов И.З., Амосова А.Л., Фисунов Г.Ю., Лебедев Ю.Б.* // Молекулярная биология. 2008. Т. 42. С. 721–727.

27. *Wang J., Song L., Grover D., Azrak S., Batzer M.A., Liang P.* // Hum. Mutat. 2006. V. 27. P. 323–329.

28. *Giardine B., Riemer C., Hardison R.C., Burhans R., Elnitski L., Shah P., Zhang Y., Blankenberg D., Albert I., Taylor J.* // Genome Res. 2005. V. 15. P. 1451–1455.

29. *Blankenberg D., Von Kuster G., Coraor N., Ananda G., Lazarus R., Mangan M., Nekrutenko A., Taylor J.* // Curr. Protoc. Mol. Biol. 2010. V. 89. P. 101–121.

30. *Goecks J., Nekrutenko A., Taylor J., Galaxy Team* // Genome Biol. 2010. V. 11. P. R86.

A Novel Approach to Identification of Somatic Retroelements' Insertions in Human Genome

A. A. Kurnosov*, S. V. Ustyugova*, M. V. Pogorelyy*, A. Yu. Komkov*, D. A. Bolotin*, K. V. Khodosevich, I. Z. Mamedov*,#, Yu. B. Lebedev***

#Phone/fax: +7 (495) 330-42-88; e-mail: Imamedov@mx.ibch.ru

** Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry of the Russian Academy of Sciences Mikluho-Maklaya, 16/10, Moscow, 117997*

*** German Cancer Research Center, 69120, Heidelberg, Germany*

The activity of retroelements is one of the factors leading to genetic variability of the modern humans. Insertions of retroelements may result in alteration of gene expression and functional diversity between cells. In recent years an increasing amount of data indicating an elevated level of retroelements' mobilisation in some human and animal tissues has been reported. Therefore, the development of a system for the detection of somatic retroposition events is required. Here we describe a novel approach to the whole-genome identification of somatic retroelemts' insertions in human genome. The developed approach was applied for the comparisons of somatic mosaicism levels in two tissues of the investigated individual. A total of 3410 insertions of retroelements belonging to AluYa5 subfamily were identified.

Key words: retroelements, somatic mosaicism, next generation, sequencing