

УДК 575.22

Комбинирование двух технологических платформ для полногеномного секвенирования человека

К.Г. Скрябин¹, Е.Б. Прохорчук^{1*}, А.М. Мазур¹, Е.С. Булыгина¹, С.В. Цыганкова¹, А.В. Недолужко¹, С.М. Расторгуев¹, В.Б. Матвеев², Н.Н. Чеканов³, Д.А. Горанская³, А.Б. Теслюк¹, Н.М. Груздева¹, В.Е. Велихов¹, Д.Г. Заридзе², М.В. Ковальчук¹

¹РНЦ «Курчатовский институт», 123182, Москва, пл. Академика Курчатова, 1

²Научно-исследовательский институт канцерогенеза ГУ РОНЦ им. Н.Н. Блохина РАМН, 115478, Москва, Каширское ш., 24

³Центр «Биоинженерия» РАН, 117312, Москва, просп. 60-летия Октября, 7, корп. 1

*E-mail: Prokhortchouk@biengi.ac.ru

РЕФЕРАТ В настоящее время стремительно развиваются новые технологии секвенирования ДНК, позволяющие быстро и эффективно определять особенности организмов на уровне строения их геномов. В данном исследовании впервые в России было проведено полногеномное секвенирование человека (русского, мужчины) с использованием двух из представленных на современном рынке технологий – циклического лигазного секвенирования SOLiDTM (Applied Biosystems) и технологии секвенирования на молекулярных кластерах с использованием флуоресцентно меченных предшественников на приборе GAII (Illumina). Общее количество накопленных данных о ДНК исследованного генома составило 108.3 млрд нуклеотидных оснований (60.2 млрд с помощью технологии Illumina и 48.1 млрд с помощью технологии SOLiD). Проведенный статистический анализ результатов показал, что данные секвенаторов GAII и SOLiD дают информацию приблизительно о 75 и 96 % генома соответственно. Точность определения коротких полиморфных районов приблизительно одинакова у двух платформ, однако за счет меньшей плотности покрытия платформа SOLiD может предсказывать меньшее количество полиморфизмов. Был установлен оптимальный алгоритм использования новейших методов определения первичной последовательности ДНК при секвенировании индивидуальных геномов человека. Данное исследование является первой российской работой по полногеномному секвенированию человека.

Ключевые слова: геном человека, технологии секвенирования, однонуклеотидные полиморфизмы

Список сокращений: ОП – однонуклеотидный полиморфизм делеции/инсерции.

ВВЕДЕНИЕ

С совершенствованием новых технологий секвенирования ДНК, позволяющих быстро и эффективно определять особенности организмов на уровне строения их геномов, геномика стала одной из самых быстроразвивающихся дисциплин. На сегодняшний день существуют три основные новейшие технологии секвенирования ДНК: технология пиросеквенирования, реализованная в секвенаторах нового поколения (GS FLX, 454 Life Science Inc./Roche), технология циклического лигазного секвенирования (SOLiD, Applied Biosystems) и технология секвенирования на молекулярных кластерах с использованием флуоресцентно меченных предшественников (Illumina). Эти платформы уже продемонстрировали свою состоятельность – за последние два года к уже известному геному человека, последовательность которого была определена рядом ведущих институтов Соединенных Штатов (США), Великобритании и Канады в течение 10 лет при общей стоимости проекта 3 млрд долларов [1], добавились пять новых: геномы выдающихся биологов современности [2, 3], африканца нигерийского происхождения [4, 5], китайца [6] и корейца [7],

не считая геномов других эукариотических и прокариотических организмов [8]. Все эти работы стали возможны именно благодаря новым технологиям, связанным с увеличением производительности секвенирования и его значительным удешевлением. В обозримом будущем количество изученных геномов будет увеличиваться в геометрической прогрессии, так, например, основные мировые научные державы включились в совместный проект под эгидой Европейского союза, США и Китая под названием «1000 геномов» (<http://www.1000genomes.org>). Однако, несмотря на значительный прогресс, достигнутый в развитии технологий «чтения» ДНК, секвенирование крупных геномов, в т.ч. и генома человека, остается нетривиальной задачей. В настоящее время не существует стандартных подходов к анализу этих геномов, а также нет полных и объективных данных по оценке эффективности описанных выше технологий.

В представленной работе впервые в России было проведено полногеномное секвенирование человека (русского, мужчины) с использованием двух современных технологий секвенирования ДНК – циклического лигазного секвениро-

вания технологии SOLiD (Applied Biosystems) и технологии секвенирования на молекулярных кластерах с использованием флуоресцентно меченных предшественников (Illumina). Настоящая работа посвящена оптимизации алгоритмов получения, анализа и представления данных полногеномного секвенирования.

Прежде чем приступить к описанию основных результатов и методик их получения, кратко будут изложены принципы и термины технологии крупномасштабного секвенирования, т.н. «next generation sequencing». Сначала геномная ДНК фрагментируется до размеров 200–1000 п.н. Полученные фрагменты представляют собой основу для создания библиотеки случайных фрагментов (далее по тексту *shotgun*), что достигается путем последовательных ферментативных реакций, пришивки олигонуклеотидных адаптеров с последующей амплификацией в ПЦР. Технологии создания библиотек для платформы GAII и SOLiD описаны на сайтах производителей. Дальнейшие процедуры связаны с получением первичной нуклеотидной последовательности с каждого из двух концов участка ДНК, представленного в библиотеке. Такие последовательности ДНК называют «чтениями» (по аналогии с общепринятым в англоязычной литературе термином «read»). Длина чтений различна для двух платформ и составляет 36 нуклеотидов для GAII и 25 нуклеотидов для SOLiD. Таким образом, каждый фрагмент ДНК в составе библиотеки характеризуется двумя чтениями с длиной и направлением, зависящим от использованной технологической платформы. Затем чтения, полученные в результате секвенирования фрагментов *shotgun* библиотек с двух концов (парноконцевое чтение), располагаются на референсном геноме человека hg18. Этот процесс называется «картирование чтений», в результате которого каждому чтению приписывают координаты его местоположения. Картирование позволяет построить гистограммы покрытия генома, гистограммы расстояния между парными чтениями, найти однонуклеотидные полиморфизмы (ОП) и короткие инсерции/делеции. Более того, расстояния между чтениями и их ориентация служат важной информацией для оценки более существенных структурных перестроек исследуемого генома. Так, если расстояние между картированными чтениями будет существенно превышать физический размер фрагментов ДНК, использованных для создания библиотеки, то это будет означать делецию в анализируемом геноме по сравнению с референсным, произошедшую между этими чтениями. Аналогично, если чтения имеют аномальную ориентацию, противоречащую логике создания библиотек, то это может говорить о возможных инверсиях в рассматриваемом районе. Таким образом, широкомасштабное секвенирование с использованием новых технологий позволяет определять как короткие полиморфные районы, так и указывать на возможные крупные генетические аномалии. Однако последние могут быть точно описаны только в результате *de novo* сборки чтений в протяженный контиги, что выходит за рамки задач представленной работы.

ЭКСПЕРИМЕНТАЛЬНАЯ ЧАСТЬ

Выбор образца ДНК. Выбор человека, геном которого должен был быть секвенирован, происходил на основе результатов анализа этнических групп Российской Федерации

методом основных компонент (далее PCA, от англ. Principal Component Analysis). Одна тысяча триста восемьдесят два человека, представляющих 32 этнические группы, были генотипированы по не менее чем 300 000 аутосомным ОП с помощью высокоплотных микроматриц ДНК. Группа этнических русских состояла из 285 образцов, которые были предоставлены профессором Заридзе Д.Г. Человек, геном которого был выбран для секвенирования, характеризовался основными компонентами, которые располагали его на двумерной карте в координатах первой и второй основной компоненты (PC1-PC2) внутри группы этнических русских. При этом эта область PC1-PC2 пространства не пересекалась ни с одной другой областью других близкородственных этнических групп (статья в процессе подготовки к печати).

Подготовка образцов. Геномная ДНК была выделена из артериальной крови (белых кровяных телец) русского мужчины (пациент N Российского онкологического научного центра им. Н.Н. Блохина РАМН, страдающий раком почки, см. выше). Фрагментирование ДНК проводилось на приборе HydroShear® (Genomic Solutions®, США) до среднего размера фрагментов 500–1000 п.н. Приготовление геномных библиотек и все последующие манипуляции были проведены в соответствии с рекомендациями фирм – производителей оборудования и соответствующих наборов реактивов. Обе геномные библиотеки были пригодны для чтения фрагментов с двух концов. Геномная библиотека, приготовленная для использования на секвенаторах Genome_Analyser_II (Illumina, США) (далее GAII), после стадии лигирования адаптеров была разделена на части: одна часть была заморожена, а другая была использована для проведения ПЦР (в дальнейшем данные этого этапа упоминаются как «амплификация № 1»). После секвенирования этой библиотеки в 9 проточных ячейках была разморожена вторая аликвота, и с ней также была проведена ПЦР. Эти образцы были использованы для секвенирования в 5 проточных ячейках (стадия «амплификация № 2»).

Эта же фрагментированная ДНК была вовлечена в создание геномных библиотек, пригодных для парного чтения на приборах SOLiD v2 (Applied Biosystems, США) (далее по тексту SOLiD). После проведения ПЦР в водно-масляной эмульсии реакционная смесь (ДНК, прикрепленная к магнитным шарикам) была нанесена на проточные ячейки, где и проходила лигазная цепная реакция. В каждом цикле секвенирования фермент (лигаза) пришивал к 5'-концу субстратного комплекса флуоресцентно меченный олигонуклеотид. После идентификации флуоресцентной метки проводилось ее отщепление и регенерация субстратного комплекса, удлинённого на 5 нуклеотидов. Всего было секвенировано 9 проточных ячеек.

Секвенирование. Для расшифровки генетической информации использовались две технологические платформы, разработанные компаниями Illumina и Applied Biosystems.

Первая платформа использует метод детекции флуоресцентных сигналов меченых нуклеотидов, включающихся в процессе синтеза *in situ* в состав поверхностных молекулярных кластеров. Данная технология реализована на секвенаторах GAII (Illumina, США). Длина чтения на этом приборе составляла по 36 нуклеотидов с каждого конца, и всего было использовано 14 проточных ячеек. Вто-

рая платформа базируется на технологии лигазного секвенирования и реализована в приборе SOLiD. Длина чтения на этом приборе составляла по 25 нуклеотидов с каждого конца, и было использовано 9 проточных ячеек.

Генотипирование нефрагментированной геномной ДНК проводилось с использованием технологии Infinium на микроматрицах 610quad (Illumina), согласно рекомендациям производителя. Сканирование микроматрицы было проведено с использованием конфокального сканера iScan. Контроль качества процедуры показал высокую степень соответствия контрольным параметрам (call rate 99.7 %). Всего были достоверно выявлены аллельные варианты 588 702 однонуклеотидных полиморфизмов (далее ОП). Их список представлен на сайте производителя http://www.illumina.com/documents/products/marker_lists/marker_list_human660W_quad.zip.

Анализ данных GAI. Для анализа полученных изображений и их конвертации в последовательности ДНК использовался программный пакет Illumina Genome Analyzer Pipeline версии 1.4.0. Картирование последовательностей на референсный геном (hg18) было осуществлено с помощью программ Eland (входящей в Genome Analyzer Pipeline) и SOAPaligner/soap2 версии 2.20, разработанной в Пекинском Институте геномики (<http://soap.genomics.org.cn/>) (далее SOAP). Полученная библиотека парно-концевых чтений с использованием платформы GAI, доступная для загрузки на персональный компьютер, находится на сайте проекта <http://www.russiangenome.ru>. Данная библиотека позволяет просматривать локализацию чтений и их направление в доступных геномных браузерах, таких как UCSC Genome browser или Ensembl Genome Browser. Для вычисления нуклеотидных несовпадений и коротких инсерций/делеций относительно референсного генома использовался SOAPaligner/soap2.

Анализ данных SOLiD. Подготовка данных SOLiD к картированию велась на программных пакетах, поставляющихся в комплекте с прибором. Картирование последовательностей осуществлялось в оригинальном цветовом пространстве с помощью программного пакета SOLiD System Analysis Pipeline Tool (Corona Lite) версии 4.0r2.0, а также, после конвертирования последовательностей из цветового пространства в формат FASTQ, программой Burrows-Wheeler Aligner (BWA) версии 0.5.1 на кластере в РИЦ «Курчатовский институт». В зависимости от сложности задачи в вычисления было вовлечено от 20 до 250 ядер. Все расчеты и для GAI, и для SOLiD данных, кроме BWA, проводились на выделенном компьютере при вычислительном кластере РИЦ «Курчатовский институт». Аналогично данным с GAI данные SOLiD можно просматривать в геномных браузерах с сайта проекта <http://www.russiangenome.ru>.

Оригинальные методы анализа. Для расчета гистограмм плотности покрытия и расстояний между чтениями в парно-концевых библиотеках, а также для расчета ошибок секвенирования по сравнению с результатами генотипирования на микроматрицах ДНК однонуклеотидных полиморфизмов использовались коды, написанные авторами на языке Perl. Коды могут быть предоставлены по запросу.

Таблица 1. Общая статистика проведенного анализа. Проценты указаны в отношении общего числа чтений отдельно для GAI (оранжевый фоновый цвет) и SOLiD (голубой фоновый цвет)

		GAI (SOAP)	SOLiD (CoronaLite)
Всего нуклеотидов		60 290 962 560	48 151 787 550
Всего чтений		1 674 748 960	1 926 071 502
Всего неоткартированных		17.41 %	32.65 %
Всего одиночных		13.99 %	48.53 %
Уникальных		5.75 %	31.36 %
Множественных		8.24 %	17.17 %
Ошибки при картировании	0	54.85 %	55.67 %
	1	16.95 %	23.13 %
	2	28.20 %	21.20 %
Всего парных		68.60 %	18.82 %
Уникальных		51.16 %	12.20 %
Множественных		17.44 %	6.62 %
Ошибки при картировании	0	74.29 %	28.14 %
	1	16.32 %	21.86 %
	2	9.39 %	50.00 %
		SOAP	CoronaLite
Инсерции	Всего	0.93 %	—
	1 п.н.	0.58 %	—
	2 п.н.	0.19 %	—
	3 п.н.	0.08 %	—
	4 п.н.	0.08 %	—
Делеции	Всего	0.81 %	—
	1 п.н.	0.50 %	—
	2 п.н.	0.17 %	—
	3 п.н.	0.07 %	—
	4 п.н.	0.07 %	—

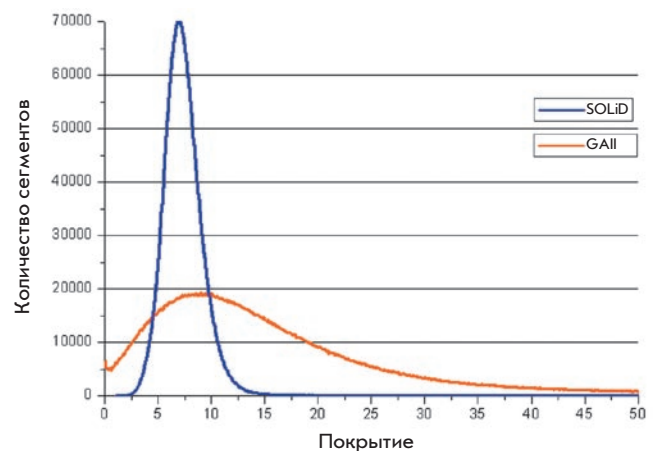
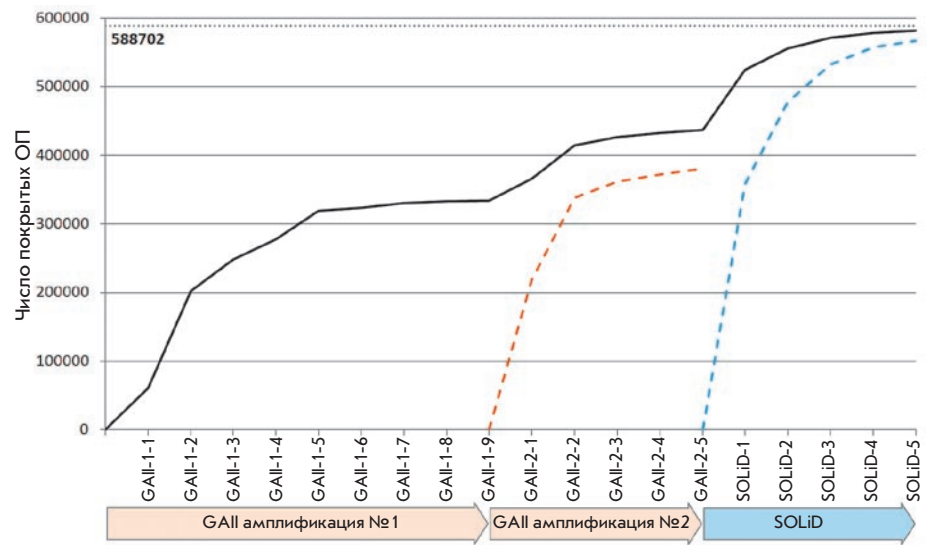


Рис. 1. График плотности покрытия генома mRG-1. Оранжевым и голубым цветами отмечены графики для GAI и SOLiD соответственно

Рис. 2. График наполнения генотипированных на микроматрице ДНК однонуклеотидных полиморфизмов данными по секвенированию генома. Из данных картирования (Eland для GAI, BWA для SOLiD) определялось, какое количество ОП хотя бы раз попало под картированное чтение. Каждый шаг по оси X соответствует получению данных с одной проточной ячейки. Общее количество генотипированных ОП указано горизонтальной асимптотой на уровне 588 702. График наполнения по секвенированию только «амплификации № 2» показан оранжевой пунктирной линией. График наполнения по секвенированию только на SOLiD показан синей пунктирной линией



РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

АНАЛИЗ ДАННЫХ, ПОЛУЧЕННЫХ НА ПЛАТФОРМЕ GAI и SOLiD

Общий объем генетических данных от геномной библиотеки «амплификации № 1» и «амплификации № 2», прошедших внутренние аппаратные контроли (base pairs passed filters), составил 60.2 млрд нуклеотидов или 1 674 748 960 чтений для GAI и 48.1 млрд нуклеотидов или 1 926 071 502 чтений для SOLiD. Картирование чтений на референсный геном человека позволило построить график плотности заполнения генома чтениями (рис. 1). Для его построения геном разбивался на последовательные фрагменты длиной 500 п.о., затем из данных картирования (Eland для GAI и BWA для SOLiD) вычислялось количество чтений, приходящихся на каждый такой фрагмент. Для вычисления плотности покрытия это число умножалось на длину чтения (36 для GAI и 25 для SOLiD) и нормировалось на 500. Оказалось, что для данных, полученных на GAI, такой график имеет форму распределения Максвелла с пиком на 8-кратном покрытии. Следует отметить, что хвост распределения смещается до значений покрытия, равных десяткам тысяч. Такие аномально плотно покрытые последовательности в основном представляли собой центромерные районы. Анализ понуклеотидного покрытия показал, что данные, полученные на платформе GAI, покрывают хотя бы один раз 66.03 % генома или 2 033 881 571 нуклеотид. Общую статистику картирования чтений можно найти в табл. 1. Следует отметить, что количество некартированных на геном чтений в данных SOLiD почти в два раза превышает аналогичное значение для GAI и составляет 32.65 %.

Подсчет количества нуклеотидных несовпадений и коротких инсерций и делеций был проведен только для уникальных выравниваний. Уникальными считаются те выравнивания, которые картируются в референсном геноме лишь один раз, что приписывает им уникальные координаты. Так, из 1.67 млрд чтений около 82.6 % нашло соответствие с минимальными искажениями в референсном

геноме. Оставшиеся 17.4 % были отнесены к классу некартированных чтений. Выборочная проверка случайных 164 чтений из этой категории показала, что ни один из них не может быть картирован на референсный геном с минимальными искажениями (до 2 несовпадений, инсерции/делеции не длиннее 4 нуклеотидов). Всего 13 чтений из 164 были отнесены к другим геномам, в основном к геному обезьян, 2 чтения к геному *Mus musculus* и по одному к геному *Danio rerio* и *E.coli*. Все эти фрагменты имели, однако, не 100 %-ное совпадение с последовательностями ДНК указанных организмов. Основная часть некартированных чтений имела короткое (до 25 нуклеотидов), но полное совпадение с различными участками генома человека.

Проверка совпадений между аллельными вариантами ОП, определенных с помощью секвенирования и генотипирования на микроматрицах ДНК.

С помощью микроматриц ДНК в исследованном геноме были определены аллельные варианты 588 702 ОП с фиксированными координатами. Для того чтобы установить, с какой точностью метод секвенирования может определять ОП, были найдены все чтения, координаты которых включали в себя координаты ОП, выявленных с помощью микроматрицы (далее эти полиморфизмы обозначаются как мОП). Количество ОП, на которые хотя бы раз картировалось чтение на платформах GAI и SOLiD, составило 581 596, или 98.8 % общего количества мОП. Чтения, полученные от платформы GAI, включали 437 056 ОП (74.2 % мОП), а от SOLiD – 566 952 (96.3 %). При секвенировании геномных библиотек, полученных на стадиях «амплификация № 1» и «амплификация № 2» на платформе GAI, удалось получить чтения, пересекающиеся с 333 647 (56.7 %) и 372 483 (55.6 %) мОП соответственно (рис. 2). Для оценки точности предсказания аллельных вариантов с помощью секвенирования были выбраны лишь гомозиготные мОП, число которых, по данным генотипирования, составило 409 760. Также были выбраны только те мОП, на которые картировалось не менее 1, 5 или 10 чтений. В табл. 2 в соответствующих строках «тест на покрытие ≥ 1», «тест на покрытие ≥ 5» и «тест на покрытие ≥ 10» приведено их

Таблица 2. Статистика совпадения предсказания гомозиготных ОП с помощью секвенирования в сравнении с генотипированием

		GAI (Eland)	SOLiD (BWA)	Обе платформы (Eland ИЛИ BWA)
Всего гомозиготных ОП на микроматрице		409 760		
Прохождение тестов	Покрытие (≥ 1)	302 919	394 373	404 564
	Покрытие (≥ 5)	250 353	238 130	349 309
	Покрытие (≥ 10)	194 016	74 902	270 890
		После прохождения всех тестов		
При покрытии ≥ 5	Количество	242 201	218 974	331 873
	Доля	96.74 %	91.96 %	95.01 %
При покрытии ≥ 10	Количество	188 708	71 999	261 537
	Доля	97.26 %	96.12 %	96.55 %

количество. В той же таблице представлены доли мОП, значения аллельных вариантов которых для разных платформ и разной глубины покрытия совпадают с предсказанным генотипированием на микроматрице ДНК. В результате анализа установлено, что с помощью секвенирования удастся определить около 81 % мОП с точностью не менее 95 % (табл. 2, серый столбец «Eland или BWA», при покрытии ≥ 5).

АНАЛИЗ ИНФОРМАЦИИ О ВЗАИМНОМ РАСПОЛОЖЕНИИ ПАРНО-КОНЦЕВЫХ ЧТЕНИЙ

Для оценки количества структурных перестроек, произошедших в исследуемом геноме, была использована информация о взаимном расположении на референсном геноме парно-концевых чтений. На рис. 3 приведен график зависимости количества картированных на hg18 парных чтений от длины референсной последовательности, находящейся между ними. Как видно из рис. 3, форма графика для двух платформ существенно различается, что может отражать принципиальную разницу в протоколах приготовления геномных библиотек, предназначенных для парно-концевого чтения. Следует обратить внимание на форму распределения расстояний между парными чтениями: графики для платформы GAI имеют локальные пики, расположенные в значениях 70 п.о. и 300 п.о., а также высокий пик в районе 700 п.о. Скорее всего, это связано с чтениями, попадающими на повторяющиеся элементы, которые имеют дискретные длины в указанных районах. Такая ситуация уже описывалась выше как проблема ELAND при анализе гистограммы покрытия. График распределения расстояний между чтениями в геномной библиотеке для платформы SOLiD имеет один пик в районе 1000 п.о.

В работе были также проанализированы возможные варианты (3) взаимного расположения и направления парных чтений. Первый вариант – чтения картируются на референсный геном в соответствии с логикой приготовления библиотеки. Для платформы GAI это означает, что два чтения ориентированы навстречу друг другу, если под на-

правлением принять 5'–3' ориентацию. Для платформы SOLiD взаимное расположение чтений на референсном геноме соответствует логике приготовления библиотеки, если они сонаправлены. Второй и третий варианты – это отличия от «нормального» расположения, возможные в том случае, если исследуемый геном имеет существенные перестройки в рассматриваемом районе, что приводит к тому, что одно или оба чтения будут картироваться на референсный геном с одной или с двумя инверсиями. В соответствии с этими определениями все парные чтения были отнесены к трем классам: «нормальные», «с одной инверсией» и «с двумя инверсиями» (табл. 3). Те чтения, которые картируются на разные хромосомы референсного генома, выделены в отдельный класс. Небольшой избыток чтений, картируемых на разные хромосомы, в случае SOLiD объясняется присутствием этапа лигирования тупых концов фрагментов и двуцепочечных олигонуклеотидов, что потенциально создает возможность ковалентного соединения двух фрагментов с разных хромосом. В целом, процент аномально ориентированных чтений приблизительно одинаков у двух платформ.

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ И ВЫВОДЫ

Нами проведено полногеномное секвенирование человека, этнически принадлежащего к русским. Основной особенностью этой работы является факт комбинирования двух технологических платформ – GAI и SOLiD. Сравнительная производительность двух систем, следует отметить, что генерация данных заняла приблизительно одинаковое количество времени, а именно 8 недель. За это время было сделано 14 запусков (по одной проточной ячейке) секвенаторов GAI и 5 запусков (по 2 проточных ячейки) SOLiD. Это позволило получить соответственно 60.2 и 48.1 млрд нуклеотидов. Во время производства работ все основные параметры функционирования технологических платформ находились в пределах, декларируемых производителями как номинальные. Анализ данных был проведен на компьютерном кластере Курчатовского научного центра, куда

данные с секвенаторов передавались по высокоскоростному оптическому каналу, что потребовало создания программного обеспечения по проверке в режиме реального времени сохранности переданных файлов. Анализ первичных данных занял приблизительно 10 недель.

Основное отличие между данными платформ GAI и SOLiD проявляется в том, насколько равномерно покрывают исследуемый геном производимые секвенаторами чтения. По нашим оценкам, чтения GAI покрывают около 75 % генома, а SOLiD – 95 %, и это несмотря на то, что GAI произвел больше нуклеотидов, чем SOLiD. Та же самая тенденция прослеживается и при анализе графика плотности покрытия генома (рис. 1). Значение графика плотности для данных SOLiD в пиковой точке в несколько раз превышает максимальное значение пика для данных GAI. Однако уже при покрытии более 20 данные SOLiD (синий график) практически не обнаруживают таких последовательностей, в то время как данные GAI (оранжевый график) показывают наличие около 10 000 таких плотно покрытых фрагментов. Таким образом, чтения, произведенные платформой SOLiD, покрывают референсный геном существенно более равномерно, чем чтения платформы GAI. Это обстоятельство отражает качество приготовления shotgun библиотек. Тот факт, что для получения равномерного покрытия на платформе GAI необходимо приготовление нескольких библиотек, желательно с различной длиной фрагментов, был отмечен в работах других авторов, в частности при секвенировании генома корейца [7]. Более того, секвенирование даже двух разных ПЦП амплификаций одних и тех же первичных фрагментов, полученных на предварительной стадии создания библиотеки, приводит к тому, что охватываются пусть и сильно пересекающиеся, но все же неидентичные части исследуемого генома. Так, при «амплификации № 1» после 9 запусков прибора GAI график наполнения мОП чтениями выходит на насыщение (рис. 2). После проведения независимой «амплификации № 2» происходит скачок графика наполнения, который, однако, после пяти запусков секвенатора тоже выходит на насыщение. Таким образом, проведение дополнительных запусков с библиотеками «амплификация № 1» и «амплификация № 2» не привело бы к существенному

Таблица 3. Статистика анализа распределения парных чтений. Под инверсией подразумевается изменение направления одного из парно-концевых чтений на референсном геноме по сравнению с расчетным. Под двойной инверсией – изменение направления сразу двух чтений. Данные по GAI и по SOLiD выделены фоновым оранжевым и голубым цветами

		GAI	SOLiD
Парный фрагмент картируется на другую хромосому		3.18 %	4.56 %
Взаимное расположение парных ридов	Нормальное	96.12 %	95.22 %
	Подразумевает инверсию	0.48 %	0.14 %
	Подразумевает двойную инверсию	0.22 %	0.07 %
Размер вставки находится в допустимом диапазоне		93.06 %	95.43 %

увеличению покрытия генома. Добавление к данным GAI данных SOLiD решило проблему частичного покрытия генома. Мы предполагаем, что эта проблема также могла быть решена за счет создания дополнительной библиотеки для GAI, с иным средним размером фрагментов.

Точность секвенирования была оценена за счет сравнения данных аллельных вариантов гомозиготных ОП, определенных с помощью секвенирования и генотипирования на микроматрице ДНК. Оказалось, что при покрытии ОП минимум 10 чтениями ошибка определения значения ОП приблизительно одинакова у двух платформ и колеблется в районе 4–5 %. Однако по причине равномерной, но недостаточной глубины покрытия генома чтениями SOLiD количество таких надежно определенных гомозиготных ОП в несколько раз меньше у SOLiD, чем у GAI.

Полученные результаты позволяют в будущем провести определение всех ОП данного генома, сделать т.н. SNP calling и сравнить аллельные варианты ОП, предсказанные GAI и SOLiD. Анализ данных ОП с фиксированными коор-

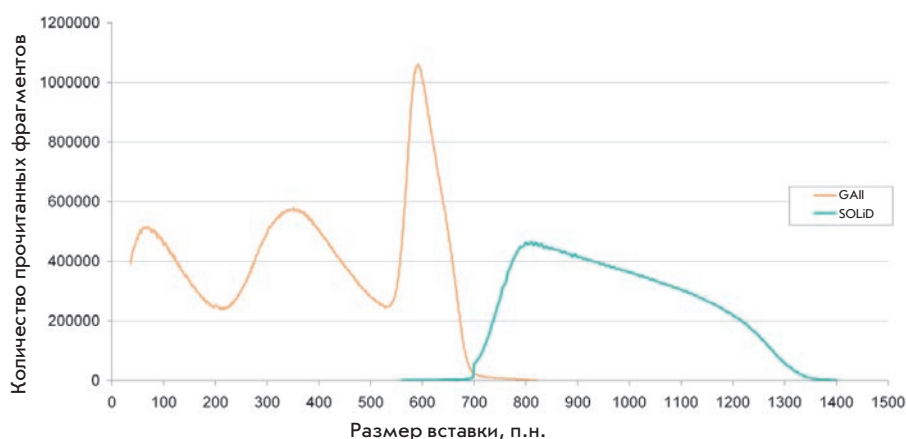


Рис. 3. График распределения расстояний между парными чтениями для библиотек, анализированных на GAI (оранжевая линия) и SOLiD (голубая линия). Чтения картировались на референсный геном человека hg18 (Eland для GAI, Corona Lite для SOLiD), после чего рассчитывалось гипотетическое расстояние между парными чтениями, определенное с помощью референсного генома. Количество фрагментов (по оси Y) было сопоставлено с предсказываемой длиной (по оси X)

динамами, аллельные варианты которых были определены секвенированием и генотипированием на микроматрице ДНК, показывает, что такие результаты во многом должны иметь высокую предсказательную силу и пересекаться не менее чем на 95 %. Более того, безусловный интерес представляет сравнение всех полиморфизмов исследуемого генома с уже известными геномами, в частности Крэйга Вентера и Джэймса Уотсона. Вторым возможным направлением работы является проведение *de novo* сборки протяженных фрагментов из тех чтений, которые не картировались на референсный геном. Потенциально они могут указывать на участки ДНК, непредставленные в референсном геноме hg18, и, тем самым, являться отличительной особенностью исследуемого нами генома.

Наши результаты создают предпосылки для дальнейшего функционального анализа данного генома. В частности, при секвенировании транскриптома знание гетерозиготных ОП в составе экспрессирующихся мРНК в клетках

крови позволит определить те транскрипты, которые имеют сдвиг от биаллельной экспрессии. Такая связь между эпигенетической и генетической составляющей этого генома представляет несомненный интерес для будущих исследований. Однако для этого изучаемый геном должен стать полноценным модельным объектом, что может быть достигнуто за счет иммортализации соматических клеток N и перевода их в линию клеток. Это позволит всем заинтересованным исследователям полноценно использовать информацию, представленную в данной статье. Авторы выражают благодарность Н.В. Равину за внимательное чтение рукописи статьи и важные советы по построению изложения результатов исследования. ●

*Работа была поддержана ФЦП
«Развитие инфраструктуры наноиндустрии
в Российской Федерации
на 2008-2012 годы».*

СПИСОК ЛИТЕРАТУРЫ

1. Lander E., Linton L., Birren B., Nusbaum C., Zody M., et al. // *Nature* 2001. V. 409. P. 860–921.
2. Levy S., Sutton G., Ng P., Feuk L., Halpern A., et al. // *PLoS Biology*. 2007. Sep 4;5(10):e254.
3. Wheeler D., Srinivasan M., Egholm M., Shen Y., Chen L., et al. // *Nature*. 2008. V. 452. P. 872–876.
4. Bentley D., Balasubramanian S., Swerdlow H., Smith G., Milton J., et al. // *Nature*. 2008. V. 456. P. 53–59.
5. McKernan K., Peckham H., Costa G., McLaughlin S., Fu Y., et al. // *Genome Research*. 2009. V. 19. № 9. № 1527–1541.
6. Wang J., Wang W., Li R., Li Y., Tian G. // *Nature*. 2008. V. 456. P. 60 – 65.
7. Ahn S., Kim T., Lee S., Kim D., Ghang H., et al. // *Genome Research*. 2009. V. 19. № 9. P. 1622–1629.
8. <http://www.ncbi.nlm.nih.gov/guide/genomes/>