

УДК 575.22

Характеристика генома русского мужчины на основе анализа однонуклеотидных полиморфизмов и реконструкции последовательностей ДНК

Н. Н. Чеканов², Е. С. Булыгина¹, А. В. Белецкий², Е. Б. Прохорчук^{1,2##}, К. Г. Скрябин^{1,2#}¹ Российский научный центр «Курчатовский институт», 123182, Москва, пл. Курчатова, д. 1² Центр «Биоинженерия» РАН, 117312, Москва, просп. 60-летия Октября, д. 7, корп. 1

Авторы, внесшие одинаковый вклад в работу (расположены в алфавитном порядке).

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

Поступила в редакцию 03.09.2010 г.

РЕФЕРАТ Ранее нами было проведено ресеквенирование генома клеток крови больного раком почки (результаты депонированы в базу данных Национального центра биотехнологической информации (NCBI) под номером SRA012240). В представленной работе мы определили координаты однонуклеотидных полиморфизмов (ОП) этого генома и сравнили их с координатами полиморфных участков в некоторых опубликованных геномах других людей. Нами выявлено всего 2 921 724 ОП, причем 1 472 679 из них описаны впервые. Анализ ОП позволил определить 63 462 ОП Y-хромосомы и на основании 18 маркеров отнести ее к гаплогруппе R1a1a, доминирующей у русских мужчин. Митохондриальная гаплогруппа определена как весьма распространенная в Европейской части России гаплогруппа U5a. С целью поиска специфичных для исследуемого генома протяженных нуклеотидных последовательностей ДНК проведена реконструкция генетических текстов (более 100 нуклеотидов) *de novo* на основе коротких чтений, полученных ранее с использованием двух технологических платформ секвенирования – Illumina Genome Analyzer II (далее GAI) и Applied Biosystems SOLiD (далее SOLiD). Это позволило выявить специфичные для данного генома последовательности общей длиной 154 т.п.н. (для GAI) и 4.7 т.п.н. (для SOLiD).

КЛЮЧЕВЫЕ СЛОВА геном человека, технологии секвенирования, однонуклеотидные полиморфизмы, биоинформатика.

СПИСОК СОКРАЩЕНИЙ ОП – однонуклеотидный полиморфизм, РКП – реконструированная консенсусная последовательность.

ВВЕДЕНИЕ

Широкое распространение секвенаторов последнего поколения сделало доступным секвенирование геномов отдельных людей. В августе 2010 г. проект «1000 genomes» [1] обнародовал на своем сайте <http://www.1000genomes.org/> предварительные данные о ресеквенировании геномов 2500 человек из разных этнических групп. Ожидается, что статья с результатами этого исследования будет опубликована в ближайшее время. Основной целью проводимых исследований является поиск генетических вариаций с частотой встречаемости более 1% в популяциях человека. Помимо решения фундамен-

тальных задач популяционной генетики, подобные исследования имеют очевидную медицинскую направленность. К примеру, в конце 2009 г. был создан международный консорциум по секвенированию геномов раковых клеток ICGC (International Cancer Genome Consortium) [2]. Россия в нем представлена РНЦ «Курчатовский институт», Центром «Биоинженерия» РАН и РОЦ РАМН им. Н.Н. Блохина, которые участвуют в изучении геномов клеток рака почки. Первое успешное ресеквенирование генома человека в России было осуществлено в конце 2009 г. [3]. Получены наборы коротких последовательностей ДНК (далее чтений) генома пациента N – русского

мужчины, больного раком почки, с использованием двух технологических платформ секвенирования – SOLiD и GAI. Тем самым впервые ресеквенирован геном представителя славянских народов, не представленных в популяционной выборке проекта «1000 genomes». С другой стороны, сделан первый шаг в рамках проекта по секвенированию генома клеток рака почки. В данной работе проведена биоинформатическая обработка результатов ресеквенирования генома пациента N с целью определения ОП. Осуществлена также реконструкция протяженных участков ДНК, специфичных для пациента N.

ЭКСПЕРИМЕНТАЛЬНАЯ ЧАСТЬ

Проявление ОП

Короткие последовательности ДНК, прочитанные на секвенаторе GAI, картировали с использованием программы SOAPaligner/soap2 версии 2.20 [4] с параметрами по умолчанию, за исключением указания размера вставки между парными чтениями. Допустимый размер вставки был установлен в пределах от 100 до 700 нуклеотидов в соответствии с результатами, полученными ранее [3]. В дальнейшем проведено определение ОП с использованием программы SOAPsnr версии 1.02 [5] с параметрами по умолчанию. Короткие последовательности ДНК, прочитанные на секвенаторе SOLiD, картировали с использованием программы Bowtie версии 0.12.5 [6] в цветовом пространстве с допущением двух ошибок на чтение и учетом качества чтения. Допустимый размер вставки был установлен в пределах от 600 до 1400 нуклеотидов также в соответствии с предыдущими данными [3]. Проявление ОП проведено с помощью программного пакета SAM tools версии 0.1.7 [7] с использованием только уникально картированных чтений.

Определение митохондриальных и Y-хромосомных гаплогрупп

Для определения митохондриальной гаплогруппы использовали чтения, полученные на секвенаторе SOLiD и обработанные с помощью программного пакета Corona Lite [3]. Список ОП генома митохондрии с координатами и значениями аллелей взят с сайта <http://www.phyloree.org/> (по состоянию на август 2010 г.). В процессе восхождения по филогенетическому дереву митохондриальных гаплогрупп, взятому с вышеупомянутого сайта, определяли аллель каждого отдельного ОП, а именно: 1) находили аллель по указанным координатам в РКП генома митохондрии, 2) подтверждали правильность этих координат, сравнивая фланкирующие последовательности (не менее 10 п.н. с каждого конца).

Для определения гаплогруппы Y-хромосомы использовали чтения, полученные на платформах GAI и SOLiD и обработанные с помощью программных пакетов Illumina Genome Analyzer Pipeline и Corona Lite соответственно [3]. Список ОП Y-хромосомы взят с сайта <http://isogg.org/> (по состоянию на август 2010 г.), за исключением тех маркеров, что не были включены в базу dbSNP. В процессе восхождения по филогенетическому дереву Y-хромосомных гаплогрупп, также взятому с вышеупомянутого сайта, определяли аллель каждого отдельного ОП следующим образом: 1) в картированных нуклеотидных последовательностях из данных GAI идентифицировали аллель по указанным в dbSNP координатам референсного генома hg18 и подтверждали правильность этих координат по фланкирующим последовательностям (не менее 10 п.н. с каждого конца или не менее 20 п.н. с одного конца); 2) в случае данных SOLiD идентифицировали аллель в РКП Y-хромосомы по фланкирующим последовательностям данного ОП из dbSNP длиной не менее 100 п.н., при наличии не более 50% пропусков в покрытии РКП чтениями. Предковый статус аллелей определяли по описанию ОП в dbSNP.

Реконструкция генетических текстов *de novo*

Сначала выбирали те чтения обеих платформ, которые не были картированы в геноме человека (hg18, за исключением нелокализованных участков). Число таких последовательностей составило 291.57 и 628.86 млн для секвенаторов GAI и SOLiD соответственно. Их использовали в качестве исходных данных для программы-сборщика коротких чтений ABySS версии 1.1.0 [8], которая использует распределенный граф de Bruijn для поиска перекрывающихся между k-мерами (последовательностями длиной k). Было проведено несколько запусков ABySS для оптимизации длины k-мера. Оптимальная длина k-мера, при котором получены контиги максимальной суммарной длины (≥ 200 п.н.), для данных GAI равна 23, а для данных SOLiD – 16.

Затем полученные *de novo* последовательности картировали на референсные геномы человека GRCh37 (hg19), Celera и HuRef с помощью программы NCBI BLAST версии 2.2.23 [9] с использованием алгоритма поиска megablast и включенной фильтрацией повторов (простых и характерных для генома человека). Последовательности, обнаруженные не во всех трех референсных геномах одновременно, повторно картировали на те же референсные геномы человека, а также на геномы приматов с использованием алгоритма поиска discontinuous megablast.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Определение ОП в геноме пациента N

Результаты ресеквенирования генома пациента N, полученные с использованием секвенаторов SOLiD и GAII, представлены в виде набора коротких чтений, которые доступны на сайте Национального центра биотехнологической информации (NCBI) под номером SRA012240. Статистическая обработка данных проведена ранее [3]. Дальнейшей задачей в рамках данной работы стало выявление координат ОП на основе сопоставления всех картированных в этом участке генома чтений. Эта процедура в англоязычной литературе обозначается как «SNP calling». Мы будем использовать в настоящей работе термин «проявление ОП» как наиболее правильный перевод, отражающий суть процесса. Проявление ОП проводили отдельно для данных каждой из использованных технологических платформ GAII и SOLiD. В результате были определены значения аллелей – 1 824 006 и 410 383 ОП соответственно. После перевода данных SOLiD из цветового формата чтений в FASTQ их объединили

с данными GAII и повторили процедуру проявления ОП. Суммарное количество ОП (2 921 724) превышает сумму количеств ОП, полученных при анализе данных двух технологических платформ по отдельности. Это указывает на взаимодополнение двух наборов данных по глубине перекрывания геномных районов. Сравнение координат и значений аллелей ОП проведено для следующих геномов: Крэйга Вентера [10], Джеймса Уотсона [11], Хуанминг Янга [12], а также геномов корейца [13], африканца [14] и европейца (CEU Trio Father NA12891 из проекта «1000 genomes»). Данные представлены в *табл. 1*. Исчерпывающий список координат и значений аллелей ОП приведен на сайте проекта <http://www.russiangenome.ru/>. На *рисунке* суммировано число общих и уникальных ОП, выявленных в геноме пациента N и геномах других индивидуумов. Не обнаружено корреляции между сходством одного или двух одинаковых аллелей ОП (см. *табл. 1*, строки «один одинаковый аллель», «оба аллеля одинаковые») и географическим расстоянием между условным местом происхождения соответствующего индивидуума и Москвой, которая была принята за условное место обитания русских (Вентер и Уотсон представлены как выходцы из Западной Европы). Однако анализ полученных данных методом главных компонент расположил индивидуумов в соответствии с географическими расстояниями между районами их происхождения (данные не приведены). Степень корреляции составила 0.89, при значении p-value, равном 10^{-5} .



Рисунок. Уникальные и общие ОП различных персональных геномов (голубые круги) и генома пациента N (красные круги).

Определение митохондриальной и Y-хромосомной гаплогруппы пациента N

Установленные координаты и значения аллелей ОП позволили определить митохондриальную и Y-хромосомную гаплогруппы генома пациента N. Сначала были отобраны все чтения, полученные с использованием технологической платформы SOLiD и картированные на последовательности референсной митохондриальной ДНК (revised Cambridge Reference Sequence (rCRS), номер в GenBank: NC_012920) [15]. На их основе построена РКП, которая размещена на сайте <http://www.russiangenome.ru/>. Средняя плотность покрытия митохондриального генома составила 291. Из сравнения полученной последовательности с референсной следует, что митохондриальный геном пациента N принадлежит к гаплогруппе U5a (*табл. 2*), одной из наиболее распространенных в Европейской части России.

Y-хромосомная гаплогруппа определена как R1a1a на основании четырех маркеров, выявленных при анализе данных обеих технологических платформ, и 19 маркеров, совпадающих с данными, полученными хотя бы от одной технологиче-

Таблица 1. Сравнение количества ОП, найденных в персональных геномах различных людей и в геноме пациента N

	Вентер	Уотсон	Янг	Кореец	Европеец	Африканец
Всего ОП в геноме	3359375	2060544	3074097	3439107	3049749	3828046
ОП в геноме русского, GAI	1824006					
Общих ОП	510444	365955	518294	570937	532194	479420
Один одинаковый аллель	427096	285913	425024	457469	431977	384934
Оба аллеля одинаковые	81957	79797	92752	113042	99967	89402
ОП в геноме русского, SOLiD	410383					
Общих ОП	179948	141703	187675	204235	192773	178744
Один одинаковый аллель	116376	73735	119837	130518	125589	111031
Оба аллеля одинаковые	27202	57292	30423	34023	33756	32133
ОП в геноме русского, SOLiD+GAI	2921724					
Общих ОП	805127	588131	814751	892529	841279	747617
Один одинаковый аллель	508066	411521	486809	513621	481542	424153
Оба аллеля одинаковые	276881	171052	307802	357562	341765	301925

Примечание. Представлены данные, полученные с использованием двух технологических платформ по отдельности и при их комбинировании.

Таблица 2. Значения аллелей известных полиморфизмов митохондриальной ДНК пациента N, которые характеризуют его принадлежность к гаплогруппе U5a

Гаплогруппа	Координата	Референсный аллель (H2)	Диагностический аллель	SOLiD-аллель
L3	3594	C	C	C
N	10398	A	A	A
N	10400	C	C	C
N	10873	T	T	T
R	12705	C	C	C
UK	12308	A	G	G
U	11467	A	G	G
U5	9477	A	A	A
U5	16270	C	T	T
U5-sub	16399	A	G	G
U5a	14793	A	G	G
U5a	16256	C	T	T

ской платформы (табл. 3). Совпадение аллеля ОП rs2534636 пациента N с предковым аллелем подтверждает гаплогруппу R1a1, поскольку этот полиморфизм считается следствием обратной мутации. В силу того, что Y-хромосома является нерекombинирующей, следует ожидать высокую степень неравновесного сцепления ее генетических маркеров. Таким образом, все 63 462 ОП, определенные в данной работе и отнесенные к Y-хромосоме, потенциально представляют гаплотип, характерный для большинства мужчин Европейской части России, ввиду того, что гаплогруппа R1a1a доминирует

на этой территории. Список всех ОП Y-хромосомы также доступен на сайте проекта.

Реконструкция *de novo* генетических текстов, специфичных для генома пациента N

Возможность получения полной нуклеотидной последовательности персонального генома позволяет найти участки, специфичные для данного человека. Хотя такие данные еще не опубликованы в рамках проекта «1000 genomes», но исследования, проведенные группой Пекинского геномного института под руководством проф. Хуанминг Янга (Huanming Yang),

Таблица 3. Значения аллелей ОП Y-хромосомы пациента N, которые характеризуют его принадлежность к гаплогруппе R1a1a

Гаплогруппа	ОП	GA-аллель	SOLiD-аллель	Предковый аллель
R	rs2032658	Н/Д	G	A
R	rs17307398	T	T	C
R	rs4481791	C	Н/Д	G
R	rs9786261	Н/Д	A	G
R	rs891407	G	G	C
R1	rs17307070	Н/Д	T	G
R1	rs9786232	G	G	T
R1	rs9785959	G	Н/Д	C
R1	rs9786197	Н/Д	C	T
R1	rs7067478	A	Н/Д	G
R1a	rs17222573	Н/Д	G	A
R1a	rs17307677	Н/Д	C	T
R1a	rs17306692	A	Н/Д	C
R1a1	rs17222202	Н/Д	A	T
R1a1	rs17316227	Н/Д	G	A
R1a1	rs2534636	Н/Д	T	T*
R1a1a	rs17222146	Н/Д	T	C
R1a1a	rs17315926	T	T	C
R1a1a	rs17221601	Н/Д	A	T

Примечание. Жирным шрифтом выделены маркеры, выявленные при использовании обеих технологических платформ. *rs2534636 – обратная мутация для гаплогруппы R1a1.

Таблица 4. Результаты статистической обработки реконструированных *de novo* контигов, однозначно отнесенных к одному из трех референсных геномов человека

	Не найдено		Не определена хромосома		Не определена координата на хромосоме		Найдено	
	GA	SOLiD	GA	SOLiD	GA	SOLiD	GA	SOLiD
hg19	292	3	31	6	0	15	154	1
Celera	147	10	47	4	0	3	307	0
HuRef	125	9	69	8	0	0	300	0

показали, что его собственный геном содержит около 7200 уникальных контигов общей длиной приблизительно 5 млн п.н. [16]. Нами проведена реконструкция уникальных генетических текстов *de novo* из генома пациента N. Все собранные контиги длиной более 100 нуклеотидов были разделены на две категории: те, которые после использования в программе BLAST давали однозначный результат поиска (табл. 4), и те, которые не могли быть однозначно интерпретированы и требовали дополнительного анализа (общая статистика приведена в табл. 5). Нуклеотидные последовательности, полученные на секвенаторе SOLiD, не дали существенного результата ни по количе-

ству собранных контигов, ни по их суммарной длине, что, по всей видимости, обусловлено непригодностью последовательностей длиной всего 25 нуклеотидов для реконструкции сложных генетических текстов. Среди собранных контигов из последовательностей, определенных на секвенаторе GAII, наиболее интересны области, которые не имеют гомологий с референсными геномами человека, а также те, которые очень похожи на геномы приматов, но все же имеют небольшие отличия. И если последовательности из первой группы могут быть отнесены к потенциальным ошибкам сборки *de novo* программы ABySS, то последовательности второй группы, очевидно,

Таблица 5. Результаты статистического анализа реконструированных *de novo* контигов, специфичных для генома пациента N

	GA	SOLiD
Однозначно найденные в hg19	146 (44.7)	1 (0.3)
Найденные менее чем в трех референсных геномах человека одновременно	93 (27.4)	3 (0.7)
Не найденные ни в одном геноме человека	72 (21.3)	0 (0)
Найденные в геномах приматов	51 (15.4)	2 (0.5)
Из них с гомологией выше 95%	22 (6)	1 (0.2)
Всего контигов	495 (154)	17 (4.7)

Примечание. В скобках приведены длины контигов в т.п.н.

не могут быть ошибкой сборки и являются характерными для пациента N. Поиск открытых рамок считывания в данных контигах не выявил протяженных (более 30 аминокислот) кодирующих последовательностей. Последовательности всех собранных *de novo* контигов доступны на сайте проекта. Разница в количестве и длине собранных контигов между геномом пациента N и геномом Янга можно объяснить разной плотностью перекрытия геномов при ресеквенировании – 7 и 30 соответственно.

Таким образом, мы приводим характеристику генома пациента N, проведенную в сравнении с опубликованными геномами других людей. Значимость найденных полиморфных и уникальных отличий в 1) формировании генетического разнообразия этносов

и в 2) определении предрасположенности пациента N к различным болезням можно будет оценить только после накопления достаточного количества данных о персональных геномах представителей различных этнических групп и проведения ассоциативных исследований с использованием не только высокоплотных микроматриц ДНК, но и полногеномного секвенирования. ●

Работа поддержана Федеральной целевой программой «Развитие инфраструктуры наноиндустрии в Российской Федерации на 2008–2012 годы». Авторы выражают благодарность М.В. Ковальчуку за всестороннюю помощь и пристальное внимание к работе.

СПИСОК ЛИТЕРАТУРЫ

- Siva N. // Nat. Biotechnol. 2008. V. 26(3). P. 256.
- Hudson T.J., Anderson W., Artz A., et al. // Nature. 2010. V. 464(7291). P. 993–998.
- Скрябин К.Г., Прохорчук Е.Б., Мазур А.М. и др. // Acta Naturae. 2009. Т. 1. № 3. С. 113–119.
- Li R., Yu C., Li Y., et al. // Bioinformatics. 2009. V. 25(15). P. 1966–1967.
- Li R., Li Y., Fang X., et al. // Genome Res. 2009. V. 19(6). P. 1124–1132.
- Langmead B., Trapnell C., Pop M., Salzberg S.L. // Genome Biol. 2009. V. 10(3). P. R25.
- Li H., Handsaker B., Wysoker A., et al. // Bioinformatics. 2009. V. 25(16). P. 2078–2079.
- Simpson J.T., Wong K., Jackman S.D., et al. // Genome Res. 2009. V. 19(6). P. 1117–1123.
- Altschul S.F., Gish W., Miller W., et al. // J. Mol. Biol. 1990. V. 215(3). P. 403–410.
- Levy S., Sutton G., Ng P.C., et al. // PLoS Biol. 2007. V. 5(10). P. e254.
- Wheeler D.A., Srinivasan M., Egholm M., et al. // Nature. 2008. V. 452(7189). P. 872–876.
- Wang J., Wang W., Li R., et al. // Nature. 2008. V. 456(7218). P. 60–65.
- Kim J.I., Ju Y.S., Park H., et al. // Nature. 2009. V. 460(7258). P. 1011–1015.
- Bentley D.R., Balasubramanian S., Swerdlow H.P., et al. // Nature. 2008. V. 456(7218). P. 53–59.
- Andrews R.M., Kubacka I., Chinnery P.F., et al. // Nat. Genet. 1999. V. 23(2). P. 147.
- Li R., Li Y., Zheng H., et al. // Nat. Biotechnol. 2010. V. 28(1). P. 57–63.