БОЛЬШИЕ ТАНДЕМНЫЕ ПОВТОРЫ СИРИЙСКОГО ХОМЯЧКА *MESOCRICETUS AURATUS* IN SILICO И IN SITU

Д. Ю. Михеев,[1] ***О. И. Подгорная,***[2] ***Д. И. Остромышенский***[2, *]

[1] *С.-Петербургский научно-исследовательский институт лесного хозяйства,* и
[2] *Институт цитологии РАН, Санкт-Петербург*;
* *электронный адрес*: *necroforus@gmail.com*

Тандемно-повторяющиеся последовательности представляют собой уникальный для эукариот класс ДНК и составляют до нескольких десятков процентов от генома у высших эукариот. Классические методы поиска тандемных повторов, несмотря на более чем полувековую историю исследований, не смогли выявить полный набор тандемных повторов ни для одного позвоночного животного. Ранее была показана возможность поиска тандемных повторов в хорошо собранном геноме домовой мыши с помощью методов биоинформатики. В настоящей работе предпринята попытка поиска тандемных повторов в относительно плохо собранном геноме сирийского хомячка. In silico найдено 19 семейств больших тандемных повторов в геноме этого вида, причем только 1 семейство было ранее клонировано и идентифицировано как тандемный повтор. Для 4 семейств, предсказанных in silico тандемных повторов, с помощью гибридизации in situ показана локализация в области первичной перетяжки хромосом.

Ключевые слова: сателлитная ДНК, тандемные повторы, сирийский хомячок.

Принятые сокращения: MaCat — мажорный сателлит мыши, п. н. — пары нуклеотидов, СатДНК — сателлитная ДНК, ТЕ — диспергированные элементы (transposable element), ТП — большие тандемные повторы, DAPI — 4,6-диамидо-2-фенилиндол, FISH — flourescent in situ hybridization (флуоресцентная гибридизация in situ), WGS — whole genome shotgun (коллекция полногеномных ридов).

Тандемно-повторяющиеся последовательности представляют собой уникальный для эукариотических организмов класс ДНК. Уникальность свойств тандемных повторов — следствие их структурной организации: они состоят из многократно тандемно-повторяющихся («голова-к-хвосту») коротких последовательностей (мономеров). Вследствие особенностей строения и нуклеотидного состава поля тандемных повторов приобретают способность к нетривиальной укладке хроматина (Vogt, 1990). У высших эукариот тандемные повторы составляют до нескольких десятков процента от генома и являются основной частью конститутивного гетерохроматина, расположенного главным образом в центромерных, перицинтромерных и субтеломерных участках хромосом. К настоящему времени вопрос о больших тендемных повторах (больших ТП) как центромерных и перицентромерных последовательностях является практически неизученным. Основная причина этого заключается в том, что существующие методики сборки геномов разработаны главным образом для неповторяющихся последовательностей ДНК и слабо применимы к тандемно-повторяющимся. Полная центромерная последовательность известна только для немногих организмов, например для нескольких одноклеточных, центромер которых имеет небольшой размер, — *Saccharomyces cerevisiae* (Choo, 1997; Pezer et al., 2012) и *Schizosaccharomyces pombe* (Takahashi et al., 1992). Причем у *S. pombe* обнаружены и перицентромерные последовательности, состоящие из тандемно-повторяющихся последовательностей, сходных с генами тРНК.

Тандемные повторы открыты в 1961 г. при центрифугировании тотальной ДНК в равновесном градиенте раствора хлористого цезия благодаря различию в ее плавучей плотности (Kit, 1961). Необходимым условием для обнаружения тандемных повторов при центрифугировании в градиенте является то, что многочисленные повторяющиеся последовательности нуклеотидов должны быть сгруппированы в кластеры. Для их выявления используют также метод, основанный на обработке геномной ДНК эндонуклеазами рестрикции (Manuelidis, 1976).

Исторически обнаруженная фракция ДНК получила название «сателлитной» (СатДНК). После прочтения геномов не было обнаружено ничего такого, что делало бы эту часть генома «сателлитной» (спутниковой): массивы СатДНК в собранном геноме являются продолжением эухроматиновых районов. Для поиска в геномах термин СатДНК неудобен, и мы будем употреблять легко формализуемый термин «большие тандемные повторы» (большие ТП).

С развитием методов секвенирования геномов, в особенности с развитием методов секвенирования нового поколения (высокопроизводительного) для поиска больших ТП, стало возможным использовать биоинформатические методы. Перспективность такого подхода ранее была показана в нашей лаборатории. В геноме домовой мыши найдено около 950 полей больших ТП, которые объединены в 62 подсемейства ТП. Только 2 из них ранее клонированы и идентифицированы как ТП (Komissarov et al., 2011).

Среди представителей мышевидных грызунов (Muridae, Rodentia) наиболее изученными являются геномы

представителей семейства Muridae. К этому семейству принадлежат два вида, широко используемых в лабораторных исследованиях, — *Mus musculus* и *Rattus norvegicus*. На примере хорошо собранного генома *M. musculus* впервые проведена работа по поиску всех больших ТП в геноме с помощью методов биоинформатики (Komissarov et al., 2011). Геномы в целом и ТП в частности другого семейства мышевидных грызунов Cricetidae, к которому принадлежит сирийский хомячок *M. auratus*, исследованы значительно хуже.

Для сирийского хомячка *M. auratus* известно и клонировано 5 больших ТП, расположенных в С-позитивных районах хромосом, но различающихся распределением по хромосомам (Yamada et al., 2006). Одним из этих ТП является центромерный повтор BglII_M11, который обнаружен на всех хромосомах *M. auratus*, а также является весьма консервативным внутри рода *Mesocricetus* (Yamada et al., 2006). Повтор EcoRI_S11, содержащий последовательность, гомологичную LINE-1, обнаружен на большинстве коротких плеч аутосом и на половых хромосомах. Большой ТП EcoRI_M13 располагается на коротких плечах всех аутосом и не найден на половых хромосомах. Остальные два больших ТП являются хромосомоспецифичными: BgIII L24 характерен для Х-хромосомы, а повтор EcoRI L7 — для 2-й хромосомы (Yamada et al., 2006).

Задачей настоящей работы являлась проверка возможности поиска больших ТП в плохо собранном геноме на примере генома сирийского хомячка *M. auratus* и проверка полученных in silico результатов с помощью флуоресцентной гибридизации in situ (FISH).

Материал и методика

Животные. Сирийский хомячок *Mesocricetus auratus* — разводка получена из питомника «Сапфировый хомяк» (Санкт-Петербург).

Поиск больших ТП в геномных основывался на методике, предложенной в работе Комиссарова с сотрудниками (Komissarov et al., 2011). Для поиска ТП использовали программу TRF (Tandem repeat finder; Benson 1999). Параметры работы программы были аналогичными тем, которые использовались в упомянутой работе (Kommisarov et al., 2011): mismatch (вес несовпадения) равен 5; maximum period size (максимальная длина мономера) равен 2000. Остальные значения выставлены по умолчанию. Для обработки результатов работы TRF использовали оригинальную программу, написанную на языке Python (http://github.com/DmitriiOstr/tandem-repeat-family-finder).

Из результатов работы TRF удаляли вложенные повторы, а среди повторов, имеющих одинаковые координаты, оставляли только повтор с наименьшей длиной мономера. Затем оставшиеся повторы фильтровали для удаления SSR (simple sequence repeat, простых коротких повторов), повторов со слишком коротким полем. В качестве кандидатов в большие ТП использовали ТП со следующими параметрами: длина мономера больше 5 bp, длина поля больше 3 kb, содержание GC от 20 до 80 %, энтропия поля больше 1.76.

Для поиска семейств больших ТП поля тандемных повторов сравнивали друг с другом с помощью программы blastn с параметрами evalue 10—16e, dust = «no». Совпадения с параметром score <90 отбрасывали. Затем строили граф, в котором в качестве вершин использовали поля ТП, две вершины связывали ребром в том случае, если между соответствующими этим вершинам полями ТП было найдено совпадение со значением score >90. Для полученного графа искали компоненты связности. Тандемные поля, соответствующие вершинам в каждом компоненте связности графа, считали принадлежащими к одному семейству больших ТП.

Для поиска совпадений с известными повторами использовали программу blastn с параметрами evalue 10—10e, dust = «no». Для поиска совпадений с диспергированными повторами сравнивали поля найденных семейств больших ТП с повторами грызунов из базы данных Repbase версии 19.04. Для удаления ошибочно позитивных совпадений из результатов убирали такие случаи, в которых диспергированные повторы покрывают поле больших ТП менее чем на 80 %. Кроме того, сравнивали поля ТП с известными ТП для *M. auratus*: MAU-BglII_M11 (номер в GenBank AB185080), MAU-BglII S11 (AB185082) и MAU-EcoRI L7 (AB185082).

Для оценки содержания больших ТП в геномах хомяков использовали программу bowtie2 (Langmead, Salzberg, 2012) с параметром чувствительности «local». С помощью этой программы выравнивали на все поля больших ТП, принадлежащих к одному семейству, исходные риды чтения геномов. Для *M. auratus* использовали риды из четырех ранов (SRR396599_2, SRR396609_2, SRR396604_1 и SRR396837_1). В качестве содержания в геноме принимали процентную долю ридов, выровнявшихся на поля больших ТП. Файлы с ранами в формате fastq скачаны из базы данных DDBJ Sequence Read Archive (http://trace.ddbj.nig.ac.jp/dra/index_e.html). Качество чтения последовательностей в каждом файле проверено программой FASTQC и при необходимости с помощью программ из пакета fastx-toolkit удалены последовательности, содержащие адаптеры для секвенирования. У всех ридов в файлах с помощью того же пакета обрезаны первые и последние 10 нуклеотидов. Для каждого семейства подсчитывали среднее значение содержания в каждом ране. Обработку результатов проводили в пакете программ для статистического анализа R.

Номенклатура тандемных повторов. Поскольку устоявшейся номенклатуры больших ТП не существует, для выбора названия новых семейств больших ТП мы использовали следующую схему. Название тандемного повтора включает в себя префикс Maur (первая буква родового названия и три первые буквы видового названия), минимальный размер мономера в п. н. и латинскую букву для индекса с целью различения ТП с одинаковой длиной мономера.

График распределения полей больших ТП. Для создания 3D-графиков распределения полей больших ТП в зависимости от GC (%), размера мономера и вариабельности мономеров внутри поля использовали пакет plot3D из пакета программ для статистического анализа R.

Расчет коротких олигонуклеотидных проб для FISH. Для гибридизации in situ подобрали короткие одноцепочечные олигонуклеотидные зонды к некоторым подсемействам ТП. Подбор последовательностей осуществляли, используя программу на языке Python по следующей методике: 1) из полногеномной сборки брали все поля, принадлежащие к данному подсемейству ТП; 2) для этих полей рассчитывали частоты всех возможных k-mer с длиной слова (k), равной 12; 3) для

Таблица 1

Последовательности олигонуклеотидов, использованных в работе

Вид	Семейство	Последовательность олигонуклеотидов 5′-3′
Mesocricetus auratus	Maur-49A	CTTCATGAAAACTAACGATACAACAC
	Maur-42A	TGCACTATGACATCACAATTATACA
	Maur-62A	GTAGGCAGCTGTAAGACAATGT
	Maur-73A	TCACTTAAATCAATGATGAGGTCTA

создания пробы брали наиболее часто встречающиеся k-mers; 4) для подбора наиболее оптимальных праймеров увеличивали k до 30—60. Контроль за качеством нуклеотидов (например, отсутствием вторичной структуры) осуществляли с помощью программы Primer3. Пробы подбирали для максимально гомогенного поля, чтобы охватить все возможные поля больших ТП. Последовательности зондов к выбранным больших ТП приведены в табл. 1.

Препараты метафазных хромосом готовили прямым способом из костного мозга животных по общепринятой методике (Ford, Hammerton, 1956).

Флуоресцентная гибридизации in situ (олиго-FISH). Для гибридизации использовали синтезированные олигонуклеотиды (Beagle, Россия), модифицированные по 5′- и 3′-концам биотином или флуоресцеином. Препараты метафазных хромосом денатурировали в буфере для денатурации (70%-ный формамид и 2-кратный SSC) в течение 2 мин при 65 °C. Затем препараты дегидратировали, высушивали и наносили гибридизационный буфер (25%-ный формамид и 4-кратный SSC), содержащий 5 мкг/мл модифицированного олигонуклеотида. Гибридизацию проводили во влажной камере при 37 °C в течение ночи. Зонды, меченные биотином, детектировали стрептавидином, конъюгированным с Alexa 488 или Alexa 568 (Invitrogen, США). Затем препараты окрашивали DAPI и заключали в среду VectaShield (Dabco, США). Препараты анализировали на флуоресцентных микроскопах AXIOSKOP-DFC360 (Институт цитологии РАН) и Leica DM 6000 B (С.-Петербургский государственный университет).

Результаты

Сборка генома (WGS) сириийского хомячка *Mesocricetus auratus. B* базе данных геномных сборок (http://www.ncbi.nlm.nih.gov/genome/) присутствует только одна сборка для *M. auratus* (Cricetinae) — APMT (GenBank: APMT00000000). Геном этого вида собран до некартированных скэффолдов, эталонного генома (reference genome) для него нет. Характеристики сборки приведены в табл. 2.

Одной из главных характеристик качества сборки генома является метрика N50. Значение метрики N50 для контигов показывает наибольшую длину контига, такую, что в контигах неменьшей длины содержится не меньше 50 % суммарной длины контигов. Величина метрики N50 и ее применимость для оценки качества сборки зависят от разных факторов, в первую очередь от размера генома, однако для геномов близких размеров значение N50 вполне может использоваться для сравнения качества сборок (Earl et al., 2011). Для использованной в работе геномной сборки значение метрики N50 равно 22.511 (APMT). Для сравнения в таких хорошо собранных геномах, как геном мыши, значение N50 для контигов равно 32 273 079 (для сборки GRCm38.p3) и 26 340 566 (для сборки MGSCv37), а для генома человека — 56 413 054 (сборка GRCh38), т. е. величина N50 для «хороших» геномов больше в 1000 раз. Но мы можем работать и с такими геномами (Остромышенский и др., 2015).

Большие ТП в геноме сириийского хомячка *Mesocricetus auratus.* Геномная сборка *M. auratus* бедна большими ТП. При использовании в качестве фильтра длины поля >3 т. н. п найдено 229 полей ТП, образующих 19 семейств. В базах данных хорошо собранных геномов млекопитающих обычно можно обнаружить по меньшей мере в 2—3 раза больше ТП (табл. 3).

Мажорным большим ТП (СатДНК) сирийского хомяка является ТП Maur-49A (табл. 4). Его значительно больше (4.7 %), чем остальных семейств ТП. Это семейство больших ТП имеет высокую степень сходства с клонированным перицентромерным ТП MAU-BglII_M11, который состоит из повторяющихся единиц длиной 48—49 bp (Yamada et al., 2006). Всего найдено 168 полей семейства Maur-49A, что составляет более 2/3 от общего количества полей больших ТП, найденных в геноме сирийского хомяка. Максимальная длина поля повторов этого семейства составляет около 40 т. п. н., а общая длина полей более 1.7 млн п. н. Семейство Maur-49A включает в себя повторы, богатые AT (содержание GC около 33 %), что в целом характерно для ТП перицентромерной локализации. В полях этого семейства больших ТП присутствуют небольшие вставки последовательностей, гомологичных диспергированному повтору B1. Все свойства ТП Ма-

Таблица 2

Геномная сборка APMT

WGS	Сборка, название	Метод сборки	Платформа секвенирования	Размер сборки, п. н.	Число контигов
APMT	MesAur 1.0	Allpaths v. R44683	Illumina HiSeq	2504925039	237 701

Примечание. WGS — название проекта сборки генома.

Таблица 3

Тандемные повторы в геномной сборке сирийского хомячка

Вид	Сборка	ТП все	ТП >3kb	Количество семейств
Mesocricetus auratus	APMT	978 061	229	19

Примечание. ТП все — общее количество ТП, найденных программой TRF.

ur-49A говорят о том, что это мажорный ТП сирийского хомячка.

Несколько семейств больших ТП с относительно высоким содержанием в геноме показали высокое сходство с диспергированными повторами (TE, transposable element) разных классов — L1 (Maur-85A) и Tcl (Maur-1501A). Высокое их содержание контрастирует с небольшим количеством найденных полей этих больших ТП. Причиной увеличения доли ТЕ могут быть недостатки метода, по которому определяли содержание семейств больших ТП в геноме. Фрагмент ТЕ, на основании которого построен мономер ТП, будет иметь сходство со всеми ТЕ этого типа в геноме, поэтому цифра содержания ТП на основе фрагментов ТЕ, скорее всего, завышена. Кроме того, найдено одно поле повтора Maur-84A, который мы наблюдали и у других животных, в частности у мыши.

Большой ТП 84A содержит фрагмент Zn-finger-домена. Этот домен входит в состав целого ряда ДНК-связывающих белков и распознается в нуклеотидной последовательности их генов. Для большинства семейств нет значительного сходства с повторами ТЕ, представленными в базе данных Repbase. Геном *M. auratus* не аннотирован, его повторы слабо представлены в Repbase, и проследить сходство ТП с другими последовательностями генома пока невозможно. Однако подавляющее большинство больших ТП мыши, описанных после классификации, отличается вполне уникальной организацией (Komissarov et al., 2011).

Для того чтобы проверить, существуют ли какие-либо качественные взаимосвязи между содержанием GC (%), размером мономера и гомогенностью поля ТП у выделенных семейств больших ТП, построили 3-мерный график положения полей в зависимости от этих параметров (рис. 1). Самая большая группа сформирована семейством Maur-49A, причем видно, что группа распадается на два кластера, различающихся по содержанию GC. По сравнению с ТП генома домовой мыши (Komissarov et al.,

Таблица 4

Семейства больших ТП в геноме *Mesocricetus auratus*

Номер	Семейство	Количество полей	Max длина поля, п. н.	Суммарная длина полей, п. н.	Среднее содержание GC, %	Содержание в геноме, %	Комментарий
1	49A	168	40 115	944 816	34	4.6770+/–0.4408	MAU-BglII_M11
2	42A	22	8247	104 294	32	0.4991+/–0.0768	
3	32A	5	8655	30 167	24	0.3663+/–0.1185	
4	73A	5	5806	20 362	35	0.0860+/–0.0064	
5	85A	1	3246	3246	41	0.0789+/–0.0061	L1
6	62A	5	15 575	45 091	41	0.0482+/–0.0029	
7	84A	1	3000	3000	36	0.0374+/–0.0072	Zn-finger
8	163A	1	3058	3058	52	0.0349+/–0.0024	
9	14A	7	11 319	47 717	46	0.0245+/–0.0061	
10	1743A	1	10 886	10 886	41	0.0041+/–0.0004	
11	20A	2	15 967	20 376	37	0.0023+/–0.0002	
12	26A	3	8635	18 570	47	0.0014+/–0.0004	
13	1501A	1	6753	6753	48	0.0008+/–0.0001	Tc1
14	35A	1	4181	4181	57	0.0003+/–0.0000	
15	24A	1	4423	4423	43	0.0003+/–0.0000	
16	126A	1	4098	4098	46	0.0003+/–0.0000	
17	82A	1	4075	4075	40	0.0002+/–0.0001	
18	291A	1	3200	3200	45	0.0002+/–0.0000	
19	30A	2	6772	6772	48	0.0001+/–0.0000	

Примечание. ТП выровнены по убыванию количества в геноме. Количество полей — общее количество полей ТП, найденных в геномной сборке. Max длина поля — длина самого протяженного поля в п. н. Общая длина поля — суммарная длина всех полей в п. н. GC % — среднее GC-содержание во всех полях семейства. Содержание в геноме — количество семейств в ТП в указанных ранах.

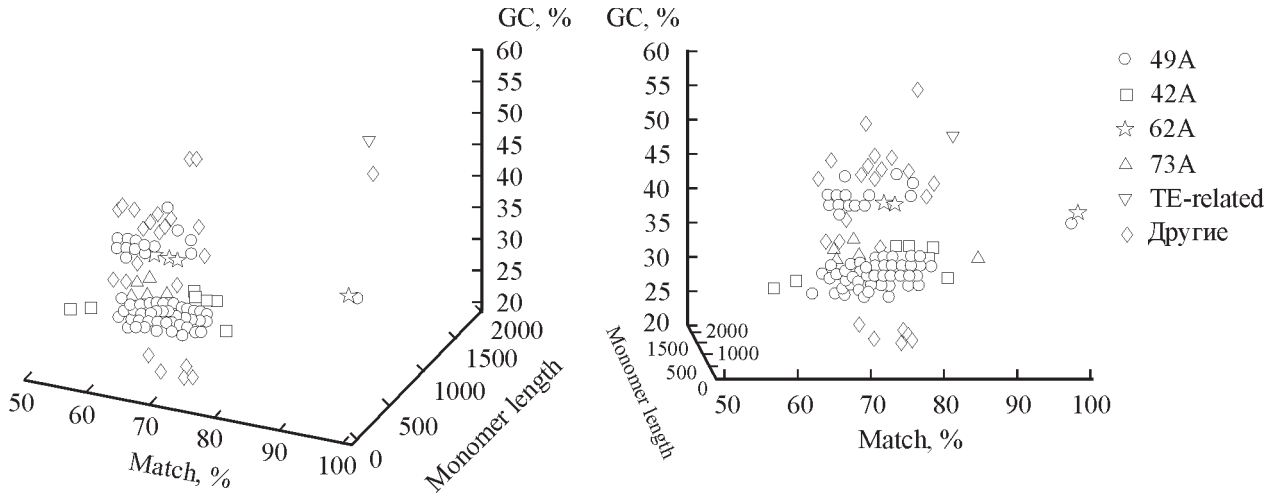

Рис. 1. Распределение больших тандемных повторов, найденных в геноме сирийского хомячка.

Показаны две проекции (*вверху* и *внизу*) визуализации положения каждого найденного поля в зависимости от содержания GC (%), размера мономера (п. н.) и однородности поля (Match, %). Каждое семейство показано определенным значком (4 семейства Maur, TE-related и др.).

2011) наблюдается значительное различие в распределении семейств ТП по осям. В геноме сирийского хомячка мы нашли лишь единичные поля GC, богатых ТП (более 50 %), и ТП с мономером длиной более 500 п. н. Кроме того, серьезное различие состоит в более сильной неоднородности полей ТП по сравнению с ТП домовой мыши. Скорее всего, такие различия обусловлены неполнотой геномной сборки сирийского хомячка. Видно, что некоторые из не поддающихся классификации полей (отмечены *красным цветом*, как остальные на рис. 1) тесно прилегают к группе полей Maur-49A и могут быть переквалифицированы при дальнейшем улучшении сборки.

Экспериментальная проверка тандемных повторов (FISH). Для нескольких семейств больших ТП подобрали олигонуклеотиды для проведения FISH. Критерием для выбора определенных семейств служили

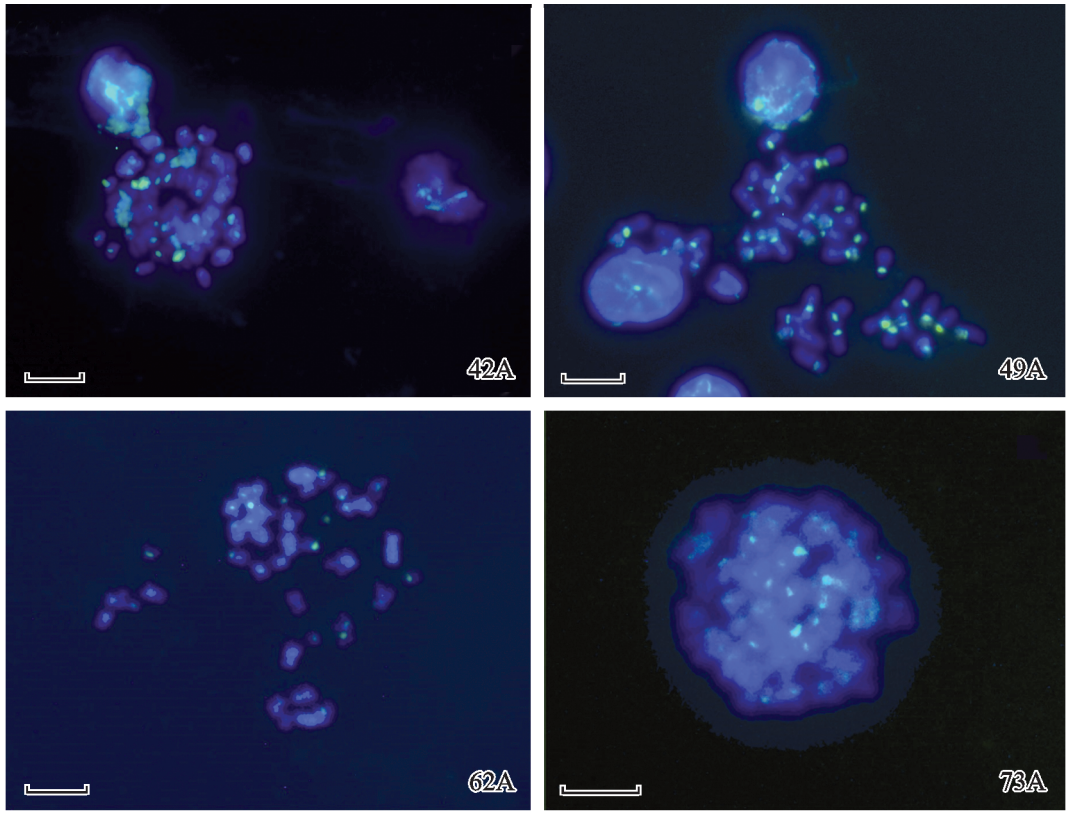

Рис. 2 Флуоресцентная гибридизация in situ с пробами к семействам больших тандемных повторов Maur-49A, Maur-42A, Maur-62A и Maur-73A (*зеленый цвет*) с хромосомами сирийского хомячка *Mesocricetus auratus*.

Хромосомы контрастированы DAPI (*синий цвет*). *Масштабные отрезки* — 10 мкм.

такие факторы, как значительное число полей, относительно высокое содержание ТП в геноме, отсутствие сходства с известными ТЕ и относительно низкое содержание GC Известно, что большинство центромерных и перицентромерных ТП богато АТ (Kusnetsova et al., 2005). Таким критериям в первую очередь соответствуют четыре семейства больших ТП — Maur-49A, Maur-42A, Maur-62A и Maur-73A (отмечены *разными значками* на рис. 1). Последовательности олигонуклеотидов приведены в разделе «Материал и методика».

Разрешение олиго-FISH на метафазных пластинах не дает возможности определить точную локализацию пробы в области первичной перетяжки — центромерную или перицентромерную. Дифференцировать положение пробы можно при использовании метода fiber-FISH. Однако большинство больших ТП описано впервые, и на первом этапе необходимо убедиться в том, что ТП локализованы в гетерохроматиновых областях. Для большинства проб мы будем описывать локализацию пробы как центромерную в широком смысле, т. е. область первичной перетяжки хромосомы, без уточнения позиции.

Все четыре исследованные с помощью FISH семейства больших ТП сирийского хомячка *M. auratus* локализованы в области первичной перетяжки хромосом (рис. 2). Однако характер картины гибридизации, интенсивность и выраженность сигнала различаются для разных семейств. Наиболее четкую и яркую картину гибридизации дает зонд к семейству Maur-49A. Сигнал обнаружен в области первичной перетяжки всех хромосом, на 3—4 парах хромосом сигнал выражен явно сильнее, чем на остальных. Картина гибридизации весьма напоминает гибридизацию мажорной СатДНК мыши (Kusnetsova et al., 2006) и подтверждает данные биоинформатики о том, что именно Maur-49A является мажорным большим ТП этого вида. Сигнал зондов к семействам Maur-42A и Maur-73A обнаружен почти на всех хромосомах также в области первичной перетяжки с более сильным сигналом на 1-й и 2-й парах хромосом. Зонд к семейству Maur-62A дает четкий гибридизационный сигнал в области первичной перетяжки примерно у половины хромосомного набора. Таким образом, все подобранные пробы дают сигнал в области первичной перетяжки, как и предполагалось.

Обсуждение

Даже в хорошо собранном геноме и даже в его некартированной части (Chromosome Unknown) представлены далеко не все поля семейств ТП, т. е. ТП не только не дочитаны, но сильно отфильтровываются уже на начальных этапах сборки. Геном мыши собран преимущественно на основании секвенирования методиками предыдущего поколения. Геном хомячка, использованный в настоящей работе, получен на основании данных секвенирования нового поколения. Это может быть критичным для сборки полей больших ТП, длина мономеров которых превышает длину чтения последовательности при секвенировании (рид). Размер одного рида при секвенировании предыдущего поколения (WGS) составляет несколько сотен п. н., в то время как при секвенировании геномов хомячков на платформе Illumina длина одного рида составляла примерно 100 п. н. Оказалось, что даже в очень плохо собранных геномных сборках можно найти значительное число полей больших ТП, а при использовании более мягких критериев по длине поля ТП их количество оказывается не меньшим, чем для хорошо собранного генома мыши. Так, в полногеномной сборке мыши Celera найдено 784 поля больших ТП длинее 3 тыс. п. н. (Komissarov et al., 2011), в то время как в геномной сборке *M. auratus* при строгом фильтре (длина поля ТП >3 тыс. п. н.) найдено 229 полей ТП, образующих 19 семейств, а при использовании более мягкого фильтра по длине поля ТП (>1 тыс. п. н.) 1029 полей ТП, образующих 65 семейств. Однако при использовании более мягкого фильтра избыточно захватываются диспергированные повторы. Использованная методика подходит не только для хорошо собранных геномов, но и для геномов, степень сборки которых далека от совершенства. Большинство проектов по секвенированию геномов останавливается на стадии первичной сборки до контигов и скэффолдов, что вполне подходит для поиска больших ТП.

Из пяти известных тандемно-повторяющихся элементов *M. auratus* (Yamada et al., 2006) нами в геномной сборки был найден только один BglII_M11, соответствующий семейству Maur-49A. Остальные четыре не найдены. Это может быть обусловлено достаточно большой длиной мономера у этих последовательностей — от 612 п. н. у повтора EcoRI_S11 до 1775 у EcoRI_L7, в то время как в геномной сборке сирийского хомячка мы нашли лишь единичные ТП с длиной мономера больше 500 п. н., что является следствием неполноты геномной сборки.

При помощи метода FISH протестированы 4 олигопробы сирийского хомячка. Maur-49A (BglII_M11) действительно является мажорным ТП и дает сигнал на всех хромосомах. Олигопробы, сконструированные на основании данных in silico, подходят для выявления мажорного ТП. Два больших ТП Maur-42A и Maur-73A также есть на всех хромосомах, но выражены 2—4 более сильных сигнала. Возможно, можно получить хромосомоспецифичные пробы при увеличении избирательности проб за счет длины или использования для расчета олигонуклеотидов хромосом-специфичных вариантов ТП. Большой ТП Maur-63A дает сигнал на половине хромосомного набора, также с выраженными доминантами, т. е. обладает ограниченной хромосомоспецифичностью, и возможно усовершенствование пробы.

Таким образом, в сборке генома сирийского хомячка *M. auratus* найдено 19 семейств больших ТП (>3 тыс. н. п.). Большие ТП четырех семейств картированы на хромосомах сирийского хомячка *M. auratus*, показана их локализация в области первичной перетяжки хромосом. Гибридизация олигопробы Maur-49A подтвердила применимость рассчитанных проб для выявления мажорного ТП этого вида.

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (проект 11-04-01258) и программы президиума РАН «Молекулярная и клеточная биология»

Список литературы

Остромышенский Д. И., *Кузнецова И. С*, *Комиссаров А. С.*, *Картавцева И. В.*, *Подгорная О. И. 2015.* Тандемные повторы геномов мышевидных грызунов в базах данных и их картирование. Цитология. 57 (2) : 102—110. **(***Ostromyshenskii D. L.*, *Kuznetsova I. S.*, *Komissarov A. S.*, *Kartavtseva I. V.*, *Podgornaya O. I. 2015.* Tandem repeats in rodents genome and their mapping. Tsitologiya. 57 (2) : 102—110.**)**

Choo K. H. A. 1997. Centromere DNA dynamics: latent centromeres and neocentromere formation. Amer. J. Hum. Genet. 61 : 1225—1233.

Earl D., *Bradnam K.*, *John J. S.*, *Darling A.*, *Lin D.*, *Fass J. Y. 2011.* Assemblathon 1 : A competitive assessment of de novo short read assembly methods. Gen. res. 21 : 2224—2241.

Ford R., *Hamerton J. L. 1956.* A colchicine hypotonic citrate squash sequence for mammalian chromosoma. Stain Technol. 6 : 247—251.

*Kit S. 1961.*Equilibrium sedimentation in density gradients of DNA preparations from animal tissues. J. Mol. Biol. 3 : 711—716.

Komissarov A. S., *Gavrilova E. V.*, *Demin S. J.*, *Ishov A. M.*, *Podgornaya O. I. 2011.* Tandemly repeated DNA families in the mouse genome. BMC Genomics. 12 : 531.

Kuznetsova I., *Podgornaya O.*, *Ferguson-Smith M. A. 2006.* High-resolution organization of mouse centromeric and pericentromeric DNA. Cytogenet. Gen. Res. 112 : 248—255.

Kuznetsova I. S., *Prusov A. N.*, *Enukashvily N. I.*, *Podgornaya O. I. 2005.* New types of mouse centromeric satellite DNAs. Chromosome Res. 13 : 9—25.

Langmead B., *Salzberg S. 2012.* Fast gapped-read alignment with Bowtie 2. Nature Methods. 4 : 357—359.

Manuelidis L. 1976. Repeating restriction fragments of human DNA. Nucleic Acids Res. 3 : 3063—3076.

Pederson T. 2000. Half a century of «the nuclear matrix». Mol. Biol. Cell. 11 : 799—805.

Pezer Z., *Brajković J.*, *Feliciello I.*, *Ugarković D. 2012.* Satellite DNA-mediated effects on genome regulation. Genome Dynamics. 7 : 153—169.

Takahashi K., *Murakami S.*, *Chikashige Y.*, *Funabiki H.*, *Niwa O.*, *Yanagida M. A. 1992.* Low copy number central sequence with strict symmetry and unusual chromatin structure in fission yeast centromere. Mol. Biol. Cell. 3 : 819—835.

Vogt P. 1990. Potential genetic functions of tandem repeated DNA sequence blocks in the human genome are based on a highly conserved «chromatin folding code». Hum. Genet. 84 : 301—336.

Yamada K., *Kamimura E.*, *Kondo M.*, *Tsuchiya K.*, *Nishida-Umehara C.*, *Matsuda Y. 2006.* New families of site-specific repetitive DNA sequences that comprise constitutive heterochromatin of the Syrian hamster (*Mesocricetus auratus*, Cricetinae, Rodentia). Chromosoma. 115 : 36—49.

Поступила 12 XI 2014

LARGE TANDEM REPEATS OF *MESOCRICETUS AURATUS IN SILICO* AND *IN SITU*

D. Yu. Miheev, *O. I. Podgornaya*, *D. I. Ostromyshenskii*[1]

Institute of Cytology RAS, St. Petersburg; [1] e-mail:necroforus@gmail.com

The class of tandemly repeated sequences exists only in eukaryotic genomes and absent in prokaryotes. The tens percent of eukaryotic genome are built up of the tandem repeats. The whole set of different tandem repeats is not revealed to any of the eukaryote species in spite of the half century history of its investigation by molecular biology methods. Previously we found the set of tandem repeats in the database of well assembled mouse genome with the bioinformatics methods. In the current work we applied the same methods to the poorly assembled hamster *Mesocricetus auratus* genome. 19 tandem repeats families have been found in hamster genome by bioinformatics (*in silico*). Only one of tandem repeats' families found have been cloned previously and exists in the Repbase, the database of all known repetitive fragments. The rest of the families are new and need the experimental verification by FISH (*in situ*). Oligo probes were designed at the base of *in silico* found sequences. Oligo probe for the known tandem repeat gives the same signal as the cloned probe, i. e. probes designed are suitable for oligo-FISH. All four oligo probes tested give signal at the heterochromatic centromeric region as expected, though with different intensities and at different number of chromosomes. The results show the power of the in silico methods for the mostly mysterious genome component, tandem repeats, investigation.

K e y w o r d s: satellite DNA, tandem repeats, *Mesocricetus auratus.*