На правах рукописи

КОМИССАРОВ
Алексей Сергеевич

Организация больших тандемных повторов в геноме мыши

03.01.03 – Молекулярная биология

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата биологических наук

Санкт-Петербург
2012

Работа выполнена в Федеральном государственном бюджетном учреждении науки Институт цитологии Российской академии наук

Научный руководитель: доктор биологических наук, профессор
Подгорная Ольга Игоревна
Институт цитологии РАН

Официальные оппоненты: **Родионов Александр Викентьевич**
доктор биологических наук, профессор
Ботанический институт РАН

Тимковский Андрей Леонидович
доктор физ.-мат. наук
Петербургский институт ядерной физики им.Б.П.Константинова

Ведущая организация: Санкт-Петербургский государственный университет

Защита диссертации состоится «20» апреля 2012 г. в 13 часов на заседании Диссертационного совета Д 002.230.01 на базе Института цитологии РАН по адресу: 194064, Санкт-Петербург, Тихорецкий проспект, д.4.

e-mail: cellbio@mail.cytspb.rssi.ru
Сайт: http//www.cytspb.rssi.ru
Факс: 8 (812) 297-35-41

С диссертацией можно ознакомиться в библиотеке Института цитологии РАН.

Автореферат разослан « » марта 2012 г.

Ученый секретарь диссертационного совета,
кандидат биологических наук Е.В. Каминская

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность проблемы. Тандемные повторы формируют значительную часть генома млекопитающих, в том числе генома мыши. В основном они концентрируются в центромерных и перицентромерных районах. Исторически тандемные повторы относили к так называемой «мусорной ДНК», но сейчас становится понятно, что их тандемная организация обеспечивает уникальные структурные и функциональные характеристики. Поле тандемного повтора сформировано многократно повторенной ДНК последовательностью (мономер тандемного повтора), уложенной голова-к-хвосту. Центромеры многих эукариот состоят в основном из тандемных повторов. Ими обогащены также перицентромерные районы. По-видимому, такая организация является критически важной для формирования и поддержания гетерохроматина, для правильной сегрегации хромосом.

Состав тандемных повторов хорошо изучен в геноме человека. Геном содержит широкий спектр тандемных повторов с разной длиной мономеров и разными типами организации мономеров в поля тандемных повторов – от коротких микросателлитов с мономером в несколько нуклеотидных пар до мегасателлитов с длиной мономера, превосходящей несколько тысяч нуклеотидных пар (Ames et al., 2008; Warburton et al., 2008). Микросателлиты и тандемные повторы с варьирующей длиной поля (VNTR – Variable Number Tandem Repeats) являются высокополиморфными и широко используются в качестве генетических маркеров.

У человека центромерный район всех хромосом содержит альфа-сателлитную ДНК, самое крупное семейство тандемных повторов в человеческом геноме. Альфа-сателлитная ДНК сформирована двумя типами полей: содержащими тандемные повторы более высокого порядка (HOR –High Order Repeat) и не содержащими. Считается, что поля с HOR необходимы для обеспечения центромерной функции.

У человека перицентромерные районы состоят из полей альфа-сателлита, окруженного полями «классических» сателлитов, таких как HS1-4 (Human Satellite 1-4). Эти поля также имеют сложную HOR структуру и, возможно, отвечают за пространственную организацию хроматина.

У домовой мыши, *Mus musculus*, центромерные и перицентромерные районы содержат две высококонсервативные тандемно повторенные последовательности. Центромерный минорный сателлит (МиСат), с мономером длиной 120 н.п., локализуется на концах всех телоцентрических хромосом мыши. Эти районы отвечают за формирование кинетохора и присоединение микротрубочек веретена деления (Kipling et al., 1991). Перицентромерный мажорный сателлит (МаСат), сформированный гетеротетрамерным мономером длиной 234

н.п., располагается в областях, прилегающих к области, содержащей МиСат (Guenatri et al., 2004). Показано, что МаСат вовлечен в формирование гетерохроматина и участвует в когезии сестринских хроматид.

В последнее время появились данные о том, что при формировании гетерохроматина важна транскрипция его ДНК. Так, например, для выключения генов в результате эпигенетической модификации типа «гетерохроматин» у *Drosophila melanogaster* включается путь RNAi. В клетках млекопитающих РНК-компонент необходим для ассоциации белка HP1 (Heterochromatic Protein 1) с перицентромерным гетерохроматином. Более того, гиперактивация транскрипции с одной из цепей МаСат необходима для формирования хромоцентров в раннем эмбриональном развитии. Экспрессия МаСат оказалась необходимым механизмом, включающимся во время определённой стадии развития для того, чтобы организовать материнский геном в хромоцентры и включить в общую структуру отцовский геном (Probst et al., 2010). Такие принципиальные открытия, касающиеся роли тандемных повторов, базируются на известной клонированной последовательности МаСат. На настоящий момент для большинства других тандемных повторов невозможно определить их транскрипционный статус, так как до сих пор эти тандемные повторы не описаны и не классифицированы. Недостаток информации о тандемных повторах затрудняет их исследования. В настоящей работе представлен новый подход к анализу тандемных повторов на геномном уровне, который позволил проанализировать и классифицировать ранее не описанные большие тандемные повторы мыши.

Цели и задачи исследования. Цель настоящей работы состояла в анализе и классификации тандемных повторов в геноме мыши.

В процессе работы решались следующие задачи:

1. Выявить все тандемные повторы в геноме мыши.
2. Подобрать параметры программ для отбора тандемных повторов определенного типа, а именно больших тандемных повторов.
3. Определить критерии классификации больших тандемных повторов.
4. Определить характерные особенности каждого из выявленных семейств тандемных повторов, объединить их в суперсемейства и выявить субсемейства.
5. Для некоторых выявленных по критериям *in silico* представителей семейств тандемных повторов сконструировать олигонуклеотидные пробы для проверки положения тандемных повторов *in situ*.

Основные положения, выносимые на защиту.

1. Анализ больших тандемных повторов в полногеномной сборке мыши выявил восемь семейств тандемных повторов, состоящих из 62 подсемейств, из которых только два подсемейства были описаны ранее.
2. Большинство подсемейств являются более GC-богатыми, чем хорошо исследованные МаСат и МиСат.
3. Для многих из новых тандемных повторов показаны HOR-структуры, что позволяет предположить существование хромосомоспецифичных вариантов больших тандемных повторов.
4. Разработан новый подход к конструированию хромосомоспецифичных проб основанных на тандемных повторах.
5. Сконструированы олигонуклеотидные пробы для картирования тандемных повторов *in situ*, основанные на нескольких мономерах.

Научная новизна работы. На примере генома мыши разработан подход к поиску и классификации «больших» тандемных повторов в базе данных любого генома. В геноме мыши впервые показано наличие тандемных повторов, сходных с «большими» классическими тандемными повторами человека. Исправлена асимметрия в распределении тандемных повторов, заключающаяся в том, что для мыши были известны только АТ-богатые тандемные повторы, в то время как большинство геномов эукариот содержит как АТ-, так и GC-богатые тандемные повторы. Выявлены хромосомоспецифичные тандемные повторы мыши, которые могут быть использованы для цитогенетического анализа. Характеристики многих из выявленных повторов предполагают наличие их хромосомоспецифичных вариантов. Предложена гипотеза о существовании хромосомного «штрих-кода», образованного последовательностью разных тандемных повторов.

Теоретическое и практическое значение работы. Теоретическое значение работы состоит в том, что предложена максимально полная и обоснованная классификация больших тандемных повторов. Разработанный подход опробован на геноме мыши, но может быть применен к базам данных любого генома. Практическим следствием работы является выявление хромосомоспецифичных тандемных повторов, которые не были известны до сих пор, что открывает перспективу создания набора проб для использования в цитогенетике.

Материалы диссертации используются в курсах лекций для бакалавров и магистров Биолого-почвенного факультета Санкт-Петербургского государственного университета и могут быть использованы в общих и специальных курсах лекций биологических факультетов других университетов.

Апробация работы. По теме диссертации опубликовано 8 печатных работ, из них 3 статьи. Основные положения представлены и обсуждены на Moscow Conference on Computational Molecular Biology (MCCMB'09), Moscow Conference on Computational Molecular Biology (MCCMB'11), на научных семинарах Отдела клеточных культур Института цитологии РАН; на научном собрании Института цитологии РАН, посвященном юбилею В.И. Воробьева в 2008 г.; на Международной конференции «Хромосома 2009», Новосибирск.

Вклад автора. Автором лично разработан и выполнен полногеномный анализ тандемных повторов в геноме мыши. Также разработан новый подход к конструированию хромосомоспецифичных проб, основанных на тандемных повторах, и сконструированы пробы для представителей трех подсемейств. Экспериментальная проверка сконструированных проб проведена Е.В. Гавриловой, кариотипирование С.Ю. Деминым. Материалы, вошедшие в представленную работу, обсуждались и публиковались совместно с соавторами и научным руководителем.

Объем и структура диссертации. Диссертационная работа состоит из введения, обзора литературы, экспериментальной части, включающей методы и результаты исследования, обсуждения, выводов и списка литературы, содержащего 140 публикаций. Работа изложена на 132 страницах и иллюстрирована 12 рисунками и 6 таблицами.

Работа выполнена при финансовой поддержке гранта Human Genome Project (HUGO DOE USA, 2005), РФФИ (№05-04-49156-а), РФФИ (№05-04-49828-а), РФФИ (№11-04-01700-а), гранта Президиума РАН "Молекулярная и клеточная биология".

МЕТОДЫ

Использованные базы данных. Собранные последовательности ДНК полногеномных сиквенсов мыши (проекты AAHY и CAAA), эталонную, альтернативную и MGSC сборки генома мыши (reference genome build 37.1, alternate genome build 37.1 и MGSC release 3 соответственно) получали с ftp сайта NCBI в FASTA формате. Аннотацию хромосомного бендинга для эталонного генома получали из Genbank. Базу данных аннотированных повторов (Repbase version 15.07) в формате FASTA получали с сайта GIRI. Для создания локальной версии полученных баз данных использовали программу blastdb из программного пакета Blast+ suite.

Использованное программное обеспечение. Выравнивание последовательностей выполняли программами bl2seq и blastn из программного пакета Blast+ suite. Для работы с тандемными повторами были изменены следующие параметры: max target seqs (максимальное

количество последовательностей, для которых будет показан результат выравнивания) и num descriptions (максимальное количество показанных описаний) изменены на 10000, evalue (ожидаемое значение) изменен на 10^{-16}, word size изменен на 10, dust (аргумент для алгоритма DUST) изменен на «no», soft masking параметр (фильтр на простые последовательности) изменен на «false». Для остальных параметров выставлены значения по умолчанию. Для поиска тандемных повторов использовали программу TRF (Tandem Repeat Finder, Benson, 1999) со следующими параметрами: mismatch выставлен равным 5; maximum period size выставлен равным 2000. Для остальных параметров выставлены значения по умолчанию. Для дот-плот анализа использовали программу, написанную на Python.

Анализ тандемных повторов. Для удаления избыточности из результатов работы программы TRF использовали следующие фильтры: (1) удаляли все вложенные тандемные повторы; (2) если два тандемных повтора имели одинаковые координаты, но отличались размером мономера, тандемный повтор с бóльшим размером мономера удаляли. Тандемные повторы с пересекающимися координатами считали независимыми тандемными повторами. Для поиска совпадений с известными повторами использовали поиск программой blastn по базе данных Repbase. Удаляли совпадения, в которых поле тандемных повторов покрыто повторами из Repbase менее чем на 80%, чтобы избавиться от ошибочно-позитивных совпадений (Рис. 1).

Сравнение геномных сборок. Для анализа положения тандемных повторов в собранном геноме использовали три геномных сборки: эталонная сборка генома (the reference genome assembly), альтернативная сборка генома (the alternate or Celera genome assembly) и геномную сборку MGSC. Каждый геном содержит дополнительную Chromosome Unknown (ChrUn), в которую помещены все нелокализованные и некартированные контиги. Для каждой геномной сборки провели поиск тандемных повторов программой TRF.

Дизайн олигонуклеотидов. Конструировали два варианта олигонуклеотидных проб. Первый вариант конструировали на основе нескольких мономеров, выбранных из самого вариабельного участка поля тандемного повтора. Мономеры располагали голова-к-хвосту с суммарной длиной около 150 н.п.. Получившуюся конструкцию фланкировали двумя разными адаптерными последовательностями. Пробы амплифицировали методом ПЦР с праймерами к адаптерным последовательностям и метили биотином. Для второго варианта использовали короткие олигонуклеотиды, основанные на самом консервативном мотиве тандемного поля. Короткие пробы синтезировали с мечеными биотином 3'-/5'-концами.

РЕЗУЛЬТАТЫ

Среди предложенных подходов к секвенированию больших геномов самым популярным методом является *шотган*-секвенирование (shotgun sequencing), которое существует в нескольких вариантах. Ключевое различие между вариантами состоит в том, что *шотган*-секвенирование проводится либо на целом геноме сразу (цельногеномное секвенирование, WGS sequencing, whole genome shotgun sequencing), либо геном сначала разбивают на перекрывающиеся фрагменты, а затем секвенируют каждый отдельный фрагмент (иерархический шотган, hierarchical shotgun). Геном мыши отсеквенирован в 2003 году двумя независимыми группами (Celera и MGSC, Mouse Genome Sequencing Consortium). Celera использовала главным образом цельногеномное секвенирование. MGSC использовал смешанную стратегию, включающую в себя как цельногеномное секвенирование, так и иерархический шотган (Mural et al., 2002; Waterston et al., 2002).

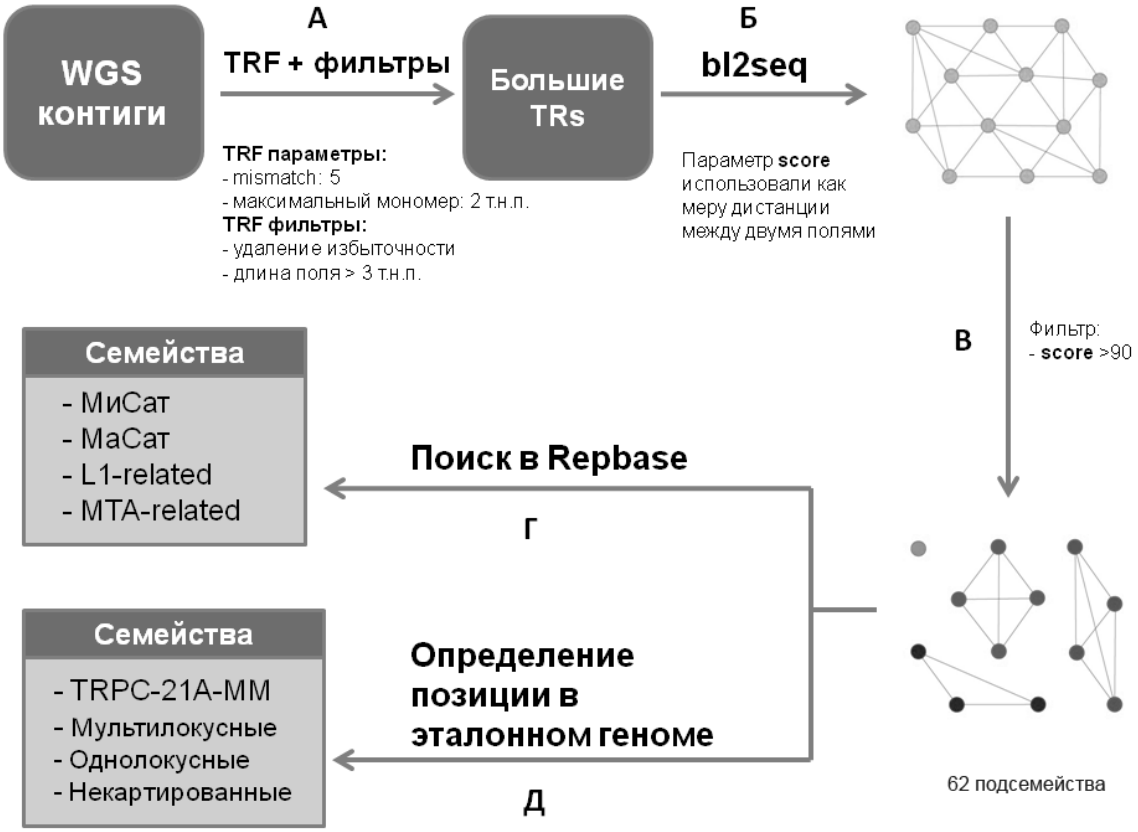

Рисунок 1 – Схема анализа и классификации тандемных повторов. Для каждой программы указаны только те параметры, которые отличаются от параметров по умолчанию, для программы blastn использовали такие же параметры, как для программы bl2seq. Имена семейств больших тандемных повторов указаны согласно таблице 2. Полное описание параметров и методов находится в главе Методы. Этапы анализа отмечены буквами А-Д.

После *шотган*-секвенирования перекрывающиеся короткие геномные последовательности (WGS sequences) собираются *in silico* в контиги (WGSA, WGS assembly). Для того, чтобы расположить контиги из WGSA относительно друг друга, необходимо использовать дополнительные экспериментальные данные, такие как физическая карта, основанная на BAC (BAC-based physical map), библиотека BAC с отсеквенированными концами, физическое картирование с использованием STS (Shot Tag Sites) и другие.
Полученные в результате длинные контиги и скаффолды хромосом могут содержать весьма большие пробелы.

Для большей части фрагментов, обогащенных повторяющимися элементами, особенно тандемными повторами, не удается собрать достаточно длинные контиги и правильно расположить их на хромосомах. Последовательности, которые не удалось собрать и (или) расположить на хромосомах, попадают в «неизвестную» хросомому (ChrUn, Chromosome Unknown). Центромерные и перицентромерные районы хромосом главным образом состоят из повторов, поэтому они практически не собраны, а на каждой хромосоме оставлен участок размером 3 Mb (GPG - Golden Path Gap).

Основным преимуществом использования WGSA для анализа повторов является то, что WGSA содержит как эухроматиновые, так и гетерохроматиновые контиги без пробелов, включая контиги, обогащенные тандемными повторами.

Существует несколько публично доступных вариантов сборки генома мыши: 1) опубликованный в 2003 году геном мыши (MGSCv3); 2) эталонный геном (the reference genome); 3) альтернативная сборка генома, основанная на данных группы Celera (the alternate genome). Для анализа положения найденных тандемных повторов использовали только эталонный геном. Для характеристики тандемных повторов по параметру наличия в ChrUn использовали альтернативную сборку, так как в эталонном геноме ChrUn практически отсутствует.

Поиск и классификация больших тандемных повторов. Одним из стандартных шагов в аннотации секвенированного и собранного генома является поиск повторяющихся элементов. Для поиска тандемных повторов общепринятой практикой является использование программы TRF (Tandem Repeat Finder). Искали тандемные повторы с длиной мономера до 2 т.н.п. в двух WGS сборках (MGSC и Celera), в эталонном собранном геноме мыши и в «неизвестной» хромосоме (Рис. 1, А, Таблица 1). Для классификации использовали только тандемные повторы из WGS сборок, как содержащие наименьшее количество артефактов сборки хромосом. Использовали эталонную геномную сборку для предсказания расположения семейств тандемных повторов на хромосомах и оценки сборки гетерохроматиновых районов.

Интересно, что и в MGSC WGS сборке (2.8%), и в Celera WGS сборке (4.8%) суммарное количество найденных тандемных повторов меньше, чем экспериментально определенное количество одного MaCat (~8%). Celera WGS сборка сильнее обогащена тандемными повторами, что можно объяснить их обильным захватом при использовании чистого цельногеномного секвенирования. Как и ожидалось, ChrUn обогащена тандемными повторами относительно собранных геномов.

Исторически тандемные повторы подразделяют на три класса согласно размеру их мономера: микросателлиты, минисателлиты (VNTR) и сателлитная ДНК (сатДНК). Короткие микро- и минисателлиты относительно хорошо изучены благодаря их участию в нарушении работы генов, что часто приводит к различным заболеваниям. Они высокополиморфны и используются как генетические маркеры. Огромные поля сатДНК изучены главным образом как принимающие участие в образовании центромерного района. К сожалению, точные границы между этими тремя классами размыты и определения классов основаны главным образом на данных, полученных в «прегеномную» эру. Для того, чтобы уйти от классификации, основанной на размере мономера, предложили термин «большие тандемные повторы» (large tandem repeats) для тандемных повторов с размером поля более нескольких т.н.п., для мыши более 3 т.н.п.

Таблица 1. Тандемные повторы в геномных последовательностях мыши. TRs – тандемные повторы (Tandem Repeats).

Сборка	Размер сборки (н.п.)	Кол-во контигов	GC (%)	TRs (все)	доля от сборки (в %)	TRs >3 т.н.п.
MGSC WGS	2 477 633 597	224 713	42,3	834 828	2,8	157
Celera WGS	3 003 109 157	837 963	40,9	1 070 233	4,8	784
Эталонный геном	2 654 895 218	21	38,9	667 513	2,0	211
ChrUn ref	3 350 358	52	37,2	1045	12,2	26

Большие тандемные повторы разделили на 62 группы (подсемейства) согласно сходству их последовательностей (sequence similarity, Рис. 1, Б, В). Сравнили каждую группу с коллекцией известных повторов из Repbase (тандемных и диспергированных, Рис. 1, Г, Д). Из 62 групп нашли только две совпадающие с известными тандемными повторами: MaCat (~76% от найденных полей) и МиСат (~2%). Остальные 60 групп не представлены среди известных тандемных повторов в Repbase и потому были классифицированы согласно их структуре и позиции на эталонном собранном геноме.

Предложили следующую классификацию. Класс больших тандемных повторов разделили на четыре суперсемейства: (1)

центромерные большие тандемные повторы (МиСат, Таблица 2, A); (2) перицентромерные большие тандемные повторы (МаСат, Таблица 2, B); (3) гетерогенные большие тандемные повторы (Таблица 2, C); (4) большие тандемные повторы, родственные диспергированным повторам (TE-related, Таблица 2, D).

Суперсемейство гетерогенных больших тандемных повторов разделили на четыре семейства: (1) TRPC-21A-MM (Таблица 2, C3); (2) мультилокусные большие тандемные повторы, найденные в нескольких локусах в собранном геноме (Таблица 2, C4); (3) однолокусные большие тандемные повторы, найденные только в одном локусе в собранном геноме (Таблица 2, C5); (4) некартированные большие тандемные повторы, не найденные в собранном геноме (Таблица 2, C6).

Суперсемейство TE-related по структурным характеристикам разделили на два семейства: (1) родственные L1 (L1_MM, LINE1) мыши (Таблица 2, D7); (2) родственные MTA (Таблица 2, D8).

Таблица 2. Классификация больших тандемных повторов мыши. Цен – центромерный район; периЦен – перицентромерный; доля от всех тандемных повторов найденных в двух сборках (в %). Групп – количество групп в семействе.

Суперсемейство	N	Семейство	В геноме	Полей	Доля от всех (в %)	Групп
A. Центромерные	1	МиСат	Цен	21	2,2	1
B. Перицентромерные	2	МаСат	периЦен	715	76,0	1
C. Гетерогенные	3	TRPC-21A-MM	периЦен	50	5,3	1
	4	Мультилокусные	Любое	57	6,0	20
	5	Однолокусные	Любое	56	6,0	29
	6	Некартированные	Нет	11	1,2	8
D. TE-related	7	MTA-related	Любое	15	1,6	1
	8	L1-related	Любое	16	1,7	1

Для ранее не описанных тандемных повторов предложили следующую номенклатуру. Название тандемного повтора включает в себя: буквы TR (Tandem Repeat); геномную позицию, если она известна; минимальный размер мономера в нуклеотидных парах; букву для индекса, если существует несколько групп с одинаковым размером мономера и суффикс MM, показывающий что тандемный повтор найден в геноме *Mus musculus*.

Распределение полей тандемных повторов в двух WGS сборках и эталонном геноме в зависимости от GC-обогащенности, длины мономера и вариабельности показаны на Рис. 2. Самая большая группа сформирована МаСат, однако группа не столь гомогенна, как принято считать на основании экспериментальных данных (Vissel , Choo, 1989; Kipling et al., 1994). Лучше всего МаСат представлен в WGS сборке Celera и сильно недопредставлен в эталонном геноме и в MGSC сборке. МиСат располагается под МаСат и формирует отдельную группу. В области с длиной мономера менее 50 н.п. расположены TRPC-21A-MM и другие мультилокусные тандемные повторы. TE-related тандемные повторы формируют отдельную группу, характеризующуюся большим мономером. Однолокусные и некартированные тандемные повторы разбросаны по графику. Вероятно, дополнительные данные, которые будут получены при ресеквенировании генома мыши, уточнят классификацию этих двух семейств.

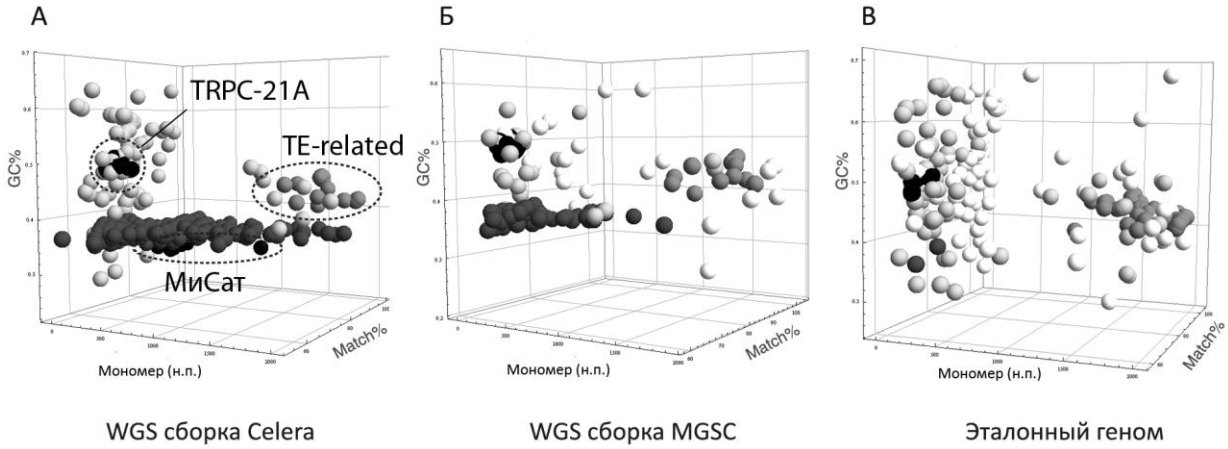

Рисунок 2. Распределение тандемных повторов в зависимости от GC-обогащенности (%) размера мономера и вариабельности. Каждая сфера отображает одно поле. Каждое семейство окрашено согласно Таблице 2: МиСат (темно-серый); МаСат (серый); TRPC-21A (черный); мультилокусные и однолокусные (светло-серый); TE-related (стальной); дополнительно показаны повторы, уникальные для конкретной сборки (белый). Для WGS сборки Celera (А) дополнительно указано положение семейств.

Тандемные повторы на концах хромосом. В наиболее хорошо собранном геноме млекопитающих, геноме человека, только хромосомы 8 и X имеют собранные и картированные последовательности, про которые известно, что они относятся к центромерному району. У мыши ни одна хромосома не заканчивается центромерным МиСат. Хромосомы мыши телоцентрические и экспериментально показано, что на центромерном конце располагаются большие поля МаСат и МиСат, которые невозможно собрать, используя существующие подходы. Поэтому плечо собранных

хромосом обрывается в перицентромерном районе, а на несобранную часть отведено по 3 Mb на каждой хромосоме (GPG, Golden Path Gap).

Для того, чтобы оценить качество сборки концов хромосом, использовали три характеристики: (1) дистанция от GPG до первого гена, для которого известна мРНК; (2) дистанция от GPG до первого большого тандемного повтора; (3) все большие тандемные повторы в пределах 2 млн.н.п., прилегающих к GPG с учетом пробелов. Оказалось, что только две собранные хромосомы из эталонного генома заканчиваются на полях МаСат (хромосомы 9 и 11). Хромосомы 3, 4, 16 и 17 заканчиваются TRPC-21A. На концах хромосом 4 и 17 обнаружили TR-22A и TR-27A, располагающиеся за TR-21A. Нашли также поля TR-22A на концах хромосом 6 и 18. Таким образом, только восемь хромосом содержат на концах большие тандемные повторы, характерные для перицентромерного гетерохроматина; и только две хромосомы содержат МаСат, давно известный как перицентромерный.

МаСат и МиСат. Экспериментальные данные указывали на сравнительную гомогенность двух сатДНК мыши. Для МаСат показана вариабельность менее 5% и для МиСат вариабельность ~5.6%. Оба семейства АТ-богатые (64% и 68%, соответственно) и имеют общие схожие по последовательности фрагменты (Wong, Rattner, 1988).

МиСат нет на хромосомах эталонного генома, тем не менее, поля МиСат с длиной поля до 6 т.н.п. нашли в WGS сборках. Найденные поля характеризуются небольшой вариабельностью мономеров внутри поля и отсутствием выраженных HOR. Нашли поля МиСат с разным размером мономера (112, 120, 223, 232, 1054 н.п.). Разница в размере мономеров может указывать на наличие HOR, однако найденных полей слишком мало для однозначного заключения (Комиссаров и др., 2010).

Известно, что перицентромерный МаСат сформирован гетеротетрамером длиной 234 н.п., состоящим из субмономеров длиной 58-60 н.п. Оказалось, что МаСат – самое большое семейство тандемных повторов в WGS сборках. Экспериментальные данные свидетельствуют об огромных полях МаСат. Однако в базах данных нашли немного полей, превышающих 10 т.н.п. (с максимальной длиной 23 т.н.п.). Отсутствие больших полей может быть объяснено известной сложностью их сборки *in silico*. Большая часть МаСат полей сформирована или мономером длиной 58-59 н.п. (37% полей), или классическим мономером длиной 234 н.п. (30% полей). Очень мало полей МаСат соответствует ранее описанной вариабельности менее 5%. Таким образом, биоинформатический подход не подтвердил расхожее мнение о высокой гомогенности полей МаСат.

Поля с HOR-структурой для МаСат экспериментально не показаны. Однако высокая вариабельность коротких мономеров является предпосылкой и необходимым условием существования HOR. Проверка всех полей МаСат на наличие HOR-структур методом дот-плота (dot-plot)

показала, что около 60% полей MaCat имеют выраженные HOR-структуры, схожие с описанными для альфа-сателлита человека.

Перицентромерный тандемный повтор TRPC-21A-MM. TRPC-21A является вторым после MaCat по количеству полей в WGS сборках. TRPC-21A более GC-богат по сравнению с MaCat и МиСат. Вариабельность мономеров в поле у TRPC-21A такая же, как у MaCat. В четырех случаях, когда он найден в эталонном геноме, он располагается в области, прилегающей к GPG, за исключением хромосомы 7, где он располагается во внутреннем бэнде 7D1. Поля TRPC-21A есть в ChrUn, состоящей главным образом из перицентромерных последовательностей; в некоторых контигах MaCat непосредственно переходит в TRPC-21A. Поэтому, в соответствии с данными in silico, в название тандемного повтора введены буквы PC (pericentromeric). Поля TRPC-21A разделили на 10 групп согласно их сходству с локусами в эталонном геноме. Большинство полей гомологично локусу хромосомы 3, содержащей большое поле TRPC-21A.

Поля TRPC-21A сформированы мономером длиной 21 н.п. и более гомогенны, чем поля MaCat. Метод дот-плот анализа выявляет HOR-структуру у всех полей TRPC-21A.

Характеристики TRPC-21A напоминают «большие классические» сателлиты человека, такие как HS2 и HS3, для которых показана хромосомоспецифичность. Предположили, что на основе TRPC-21A можно создать хромосомоспецифичные пробы, аналогичные пробам для HS2 и HS3.

Мультилокусные и однолокусные тандемные повторы. Суперсемейство гетерогенных тандемных повторов классифицировали на семейства по признаку присутствия (мульти- и однолокусные) или отсутствия полей каждого тандемного повтора в эталонном геноме.

Мультилокусный TR-22A хорошо представлен в WGS сборках; и именно он найден на концах четырех хромосом в эталонном геноме; на трех хромосомах TR-22A прилежит к GPG и на одной расположен в бэнде 7A2. Мультилокусный TR-4A, характеризующийся очень коротким AT-богатым мономером, похож на VNTR. Около половины мультилокусных повторов, в том числе TR-4A, найдены на хромосоме X. Это может быть объяснено более аккуратной сборкой гетерохроматиновых районов хромосомы X по сравнению с другими хромосомами. Однако известно, что половые хромосомы характеризуются уникальными ДНК-повторами (Namekawa, Cooke et al., 1985; Namekawa et al., 2010) и TR-4A может быть одним из них.

На Рис. 2, A группы тандемных повторов из мультилокусного и однолокусного семейств, которые имеют схожий GC-состав, размер мономера и вариабельность, формируют три визуально отличающиеся группы: GC-богатые, AT-богатые и GC-нейтральные. При этом

существенного сходства нуклеотидных последовательностей внутри групп не наблюдается.

Мультилокусный TR-22A находится в центре GC-богатой группы в районе 55-60% GC, а TR-6A, TR-57A, TR-16A и TR-31B плотно прилегают к области, где расположен TR-22A. По крайней мере один однолокусный тандемный повтор также относится к этой группе (TR-31D).

AT-богатая группа располагается в области 40-45% GC и сформирована главным образом однолокусными повторами. GC-нейтральная группа сформирована как мультилокусными, так и однолокусными тандемными повторами.

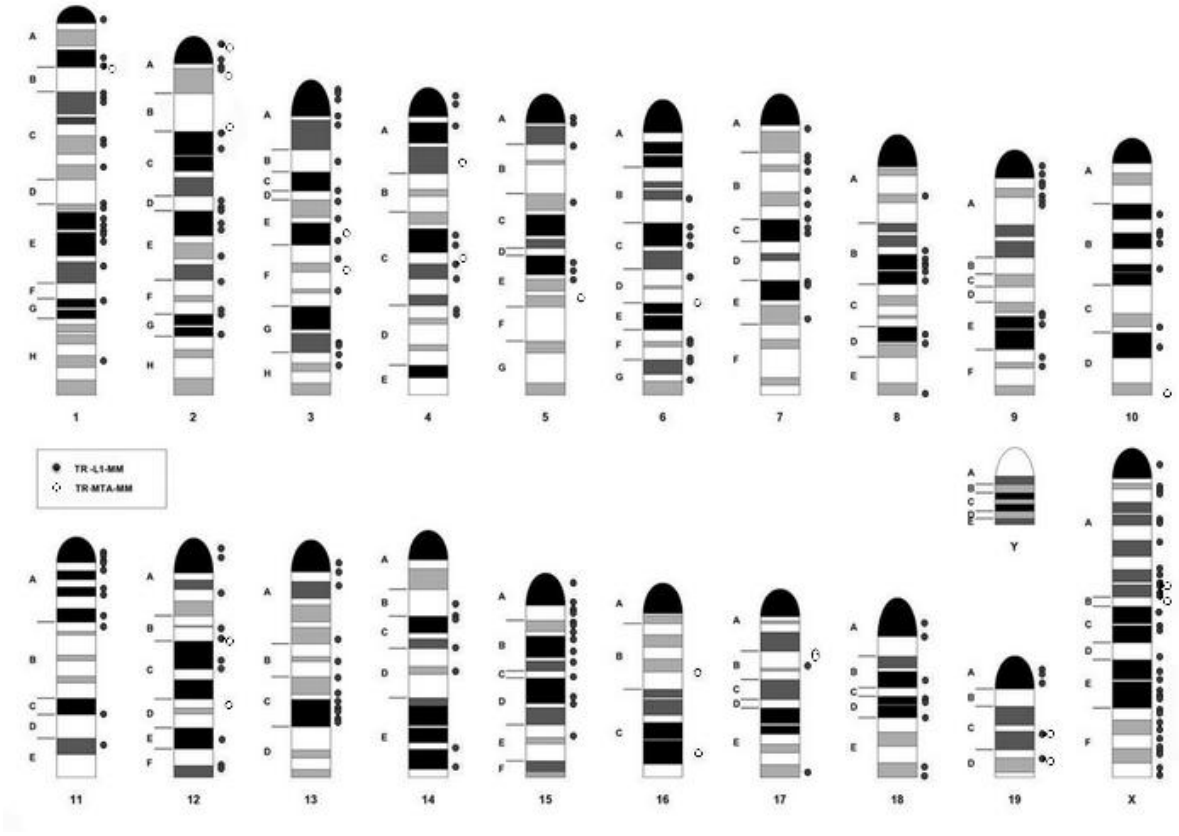

Рисунок 3 – *In silico* картирование TR-MTA и TR-L1 на эталонный геном. Черным кругами отмечена позиция TR-L1-MM, белыми кругами – позиция TR-MTA-MM.

Тандемные повторы, родственные диспергированным повторам. Мономеры двух семейств состоят из фрагментов известных диспергированных повторов. Первое семейство, TR-MTA, сформировано фрагментами MTA (Mammlian apparent LTR-retrotransposons), второе семейство, TR-L1, сформировано фрагментами L1 (LINE1), а именно фрагментом ORF2 и 3'-LTR. Оба семейства картировали на эталонный геном, размер большинства найденных локусов не превышал 5 т.н.п., однако для TR-MTA нашли два локуса размером более 10 т.н.п. Семейство TR-L1 главным образом локализуется в гетерохроматиновых бэндах, с

обогащением хромосомы X, однако собранная хромосома Y не содержит TE-related тандемных повторов (Рис. 3). Экспериментальная проверка локализации этих семейств затруднена тем, что другие нетандемно повторенные ретроэлементы будут давать сигналы при гибридизации *in situ*.

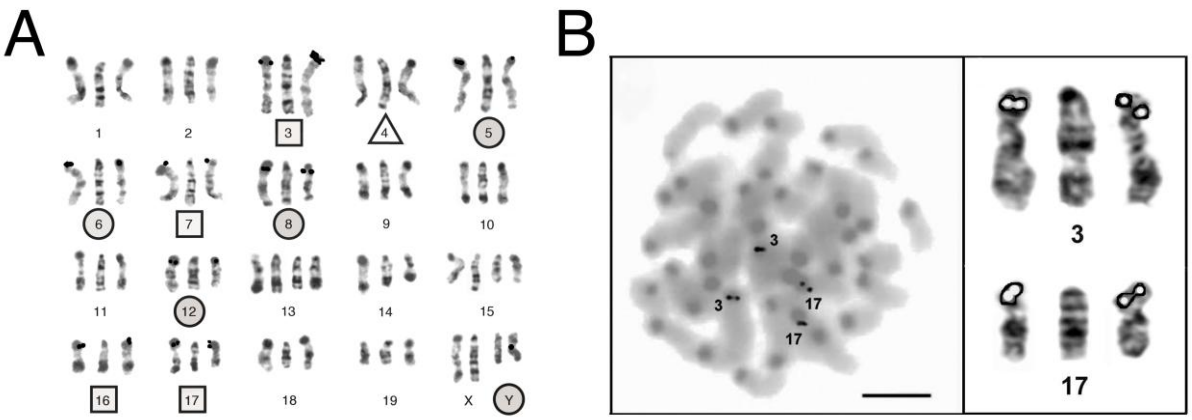

Рисунок 4 – Результаты проверки сконструированных проб методом FISH. А – короткая проба TRPC-21A; кариотипирование метафазной пластинки костного мозга, окраска DAPI (серый), сигнал – черный; оригинальные хромосомы находятся по краям, в центре – хромосома из атласа (Мамаева, 2002). Светло-серые квадраты – предсказанные хромосомы *in silico* и подтвежденные *in vitro,* серые круги – сигнал только *in vitro,* треугольник – положение не подтвердилось. Б – длинная проба TRPC-21A. **a** – метафазная пластинка костного мозга окраска DAPI (серый), сигнал – черный, отмечены номера хромосом. **b** – хромосомы как на панели I, показаны только хромосомы, несущие сигнал (белый с черной обводкой). Масштабный отрезок – 5 мкм (Komissarov et al., 2011).

Картирование больших тандемных повторов. Для того, чтобы проверить методом FISH, соответствуют ли действительности биоинформатические предсказания о позиции вновь найденных повторов (*in silico* versus *in situ*), предложили подход к созданию проб, основанный на использовании уникальных фрагментов полей тандемных повторов, когда нужно повысить хромосомоспецифичность, или на использовании общих для подсемейства фрагментов, когда нужно получить пробу, специфичную ко всем полям тандемных повторов одного подсемейства.

Предположили, исходя из *in silico* картирования, что TRPC-21A (C3, Table 1) должен быть *in situ* локализован в перицентромерном районе. Сконструированная короткая проба дает сигнал на девяти хромосомах: 3, 5, 6, 7, 8, 12, 16, 17 и Y. Во всех случаях, кроме Y, метка находится в перицентромерных районах (Рис. 4, А). Наличие HOR-структур у TRPC-21A дает возможность предположить наличие хромосомоспецифичных

вариантов. Длинную пробу сделали, основываясь на трех мономерах хромосомы 3.

Эта проба гибридизуется с двумя хромосомами (3 и 17) в соответствии с полями, найденными на концах хромосом в собранном геноме (Рис. 4, Б). Для экспериментальной проверки также выбрали мультилокусный TR-22A из-за многочисленности найденных полей как в собранном геноме, так и в ChrUn. Одноцепочечный мономер, меченый с двух концов, гибридизуется на 10 хромосом, 4 из них предсказаны *in silico* (Komissarov et al., 2011).

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Значительная часть генома эукариот не кодирует никаких белков и относится к так называемой «некодирующей ДНК». В настоящее время ее функции практически неизвестны. Главное отличие между геномами заключается в огромных массивах видоспецифичной некодирующей ДНК. Одной из интересных гипотез о роли некодирующей ДНК является предположение о её участии в организации трехмерной структуры ядра, однако эта гипотеза всё еще остается недоказанной (Подгорная и др., 2009).

У эукариот конститутивный гетерохроматин формируется на основе некодирующей ДНК, организованной в виде тандемных повторов. Тандемные повторы составляют до 10% генома большинства высших эукариот, и столь высокое содержание не может быть случайным. Считают, что роль тандемных повторов проявляется на двух уровнях организации: 1) на уровне хромосом и 2) на уровне ядра. На хромосомном уровне участие тандемных повторов в формировании теломер, центромер и интерстициального гетерохроматина может быть весьма важным, так как их наличие часто приводит к формированию трехмерной структуры ДНК, отличной от эухроматиновых участков хромосом. На уровне ядра конститутивный гетерохроматин может принимать участие в организации хромосомных территорий, что необходимо для правильного расположения транскрипционных кластеров в дифференцированных клетках, а также для успешного прохождения клеткой мейоза и митоза (Janssen et al., 2000; Plohl et al., 2008; Mayer et al., 2010).

Гетерохроматин формируется преимущественно в центромерных и перицентромерных участках хромосом. У человека центромерные районы, состоящие из альфа-сателлитных полей, окружены полями «классических» сатДНК или сатДНК с «простой последовательностью». Частным примером таких повторов являются поля HS3, в состав которых входят различные варианты мономера на основе последовательности ДНК: ATTCCA. Поля HS3 присутствуют в различных количествах в перицентромерных районах большинства хромосом человека. Обнаружено несколько типов хромосомоспецифичных вариантов HS3 (Moyzis et al.,

1987). До сих пор аналоги «больших» тандемных повторов не были известны для мыши, но тандемный повтор TRPC-21A демонстрирует все их основные признаки, включая хромосомоспецифичность и локализацию в перицентромерном районе. Обнаружены и другие кандидаты на роль «классических» сатДНК мыши – тандемные повторы TR-22A и TR-27A.

Хромосомоспецифичные тандемные повторы мыши. В WGS идентифицировано 62 подсемейства (группы), предсказанных *in silico* тандемных повторов. Структуры типа HOR найдены для многих тандемных повторов, что предполагает наличие их хромосомоспецифичных вариантов. Пробы для трех представителей семейств сконструированы на основе мономеров. Все хромосомы, с предсказанными тандемными повторами *in silico*, метятся и *in situ*, однако дополнительно метятся и другие хромосомы, по причине того, что гетерохроматиновые районы в эталонном геноме собраны далеко не полностью.

Длинная проба, основанная на варианте TRPC-21A с хромосомы 3, гибридизуется на концах хромосом 3 и 17. До сих пор для мыши не существовало цитогенетических хромосомоспецифичных проб на основе тандемных повторов.

Используя полученные в настоящей работе данные, можно картировать часть новых тандемных повторов, предполагая найти хромосомоспецифичное распределение. Хромосомоспецифичность проб может быть усилена мультимеризацией. Пробы могут быть усовершенствованы вплоть до коммерческого использования.

«Штрих-код» хромосом. Полагаем, что дальнейшие исследования приведут к созданию набора проб, основанных на тандемных повторах для отдельных хромосом. Такой набор будет отражать «штрих-код», придающий хромосоме индивидуальность. Возможно, что гетерохроматиновая подпись, заключенная в «штрих-коде», необходима для аранжировки хромосом в интерфазном ядре в процессе развития и дифференцировки. Гетерохроматиновый «штрих-код», или идентификатор, состоящий из тандемных поворов, потенциально представляет собой основу для гипотетической «Генеральной Программы Развития», предположительно располагающейся в гетерохроматиновых районах (Paris, 2010). Различные наборы тандемных повторов, которые маркируют разные хромосомы, могут обеспечивать механизм правильной ассоциации хромосом и их отдельных районов через РНК интермедиаты.

Набор тандемных повторов в геномах эукариот. Технология поиска и классификации тандемных повторов, описанная в работе, может быть применена к любому секвенированному геному. Международный проект «1000 genomes» непрерывно увеличивает число новых отсеквенированных

геномов различных животных. Однако основной трудностью остается неполнота этих баз данных, поскольку традиционные методы чтения и сборки не справляются с гетерохроматиновыми областями, состоящими из высокоповторяющейся ДНК. Предполагают, что секвенирование следующего поколения (next generation sequencing) может исправить ситуацию.

Возможно, удастся показать, что при анализе тандемных повторов хорошо прочитанных геномов высших эукариот можно выявить общие закономерности, несмотря на видоспецифичность тандемных повторов (сатДНК). Возможно, окажется, что сходные классы со сходными характеристиками необходимы для нормального эукариотического генома. Рабочая гипотеза предполагает, что можно распознать общие структурные особенности для различающихся по первичной последовательности тандемных повторов из разных геномов.

Анализ прочитанной части генома уже дал колоссальный толчок развитию молекулярной биологии, и анализ активно продолжается. Создаются базы данных, содержащие описания отдельных элементов (например, проект ENCODE). Но, к сожалению, тандемные повторы остаются за пределами интересов участников проекта, главным образом из-за отсутствия подходов к их анализу. Определение их функциональной роли как никогда актуально после чтения геномов, но для начала необходимо их выявить и классифицировать. Предложенный в настоящей работе подход доказал свою перспективность для решения этой задачи.

ВЫВОДЫ

1. С использованием предложенного в работе полногеномного анализа тандемных повторов в геноме мыши найдено и охарактеризовано шесть новых семейств больших тандемных повторов. Все они более GC-богаты, чем известные МаСат и МиСат. В результате состав тандемных повторов мыши стал похож на таковой геномов других млекопитающих, содержащих как AT-богатые, так и GC-богатые тандемные повторы.
2. Многие тандемные повторы мыши содержат HOR-структуры, что предполагает наличие хромосомоспецифичных вариантов.
3. Гибридизация с пробами показала, что все хромосомы, несущие тандемные повторы *in silico*, метятся *in situ*, хотя метятся и другие хромосомы из-за недособранности тандемных повторов в эталонном геноме мыши. Длинная проба, основанная на варианте TRPC-21A с хромосомы 3 узнает длинные поля в перицентромерном районе хромосом 3 и 17. До сих пор не существовало хромосомоспецифичных проб на основе тандемных повторов для цитогенетики мыши.
4. На основе вновь охарактеризованных семейств тандемных повторов можно создать набор хромосомоспецифичных проб, маркирующих гетерохроматиновые районы отдельных хромосом.

СПИСОК ПУБЛИКАЦИЙ ПО ТЕМЕ ДИССЕРТАЦИИ

Статьи:

Подгорная О.И., Остромышенский Д.С., Кузнецова И.С., Матвеев И.В., Комиссаров А.С. Парадоксы организации центомера и гетерохроматина // Цитология. – 2009. – Т.51. – № 3. – с.204-211.

Комиссаров А.С., Кузнецова И.С., Подгорная О.И. Центромерные тандемные повторы мыши *in silico* и *in situ* // Генетика. – 2010. – Т. 46. – № 9. – с. 1217–1221.

Komissarov A.S., Gavrilova E.V., Demin S.J., Ishov A.M., Podgornaya O.I. Tandemly repeated DNA families in the mouse genome // BMC Genomics. – 2011. – V.12. – № 1. – p.531.

Тезисы:

Комиссаров А.С., Кузнецова И.С., Подгорная О.И. Состав и тканеспецифичность хромоцентров мыши // XV Всероссийское совещание «Структура и функции клеточного ядра» Материалы конференции. – 2005. – с. 8.

Komissarov A., Podgornaya O. CHRUNTA – tandem repeat search and classification program // Proceedings of the 3-rd Moscow conference on computational molecular biology. – 2007. – p. 155-156.

Komissarov A., Podgornaya O. Similar curved motif surrounds CENPB box in different centromeric satellite DNAs // Proceedings of the international Moscow conference on computational molecular biology. – 2009. – p. 177-178.

Комиссаров А.С. Анализ сателлитной ДНК в геноме мыши // II Конференция молодых ученых Института цитологии РАН. – 2010. – с. 1.

Komissarov A., Gavrilova E., Podgornaya O. Classification of tandemly repeated DNA families in the mouse genome // Moscow Conference on Computational Molecular Biology (MCCMB'11). – 2011. – p. 173-174.

ЦИТИРОВАННАЯ ЛИТЕРАТУРА

Мамаева С.Е. (2002). М.: Научный мир, 236 с. **Ames D, Murphy N, Helentjaris T, Sun N, Chandler V.**. Genetics 2008, **179**:1693-704. **Benson G**. Nucleic acids research 1999, **27**:573-80. **Cooke HJ, Brown WR, Rappold GA**. Nature 1985, **317**:687-92. **Guenatri M, Bailly D, Maison C, Almouzni G.** The Journal of cell biology 2004, **166**:493-505. **Janssen S, Durussel T, Laemmli UK**. Mol Cell 2000, **6(5)**:999-1011. **Kipling D, Ackford HE, Taylor BA, Cooke HJ**. Genomics 1991, **11**:235-41. **Kipling D, Wilson HE, Mitchell AR, Taylor BA, Cooke HJ**. Chromosoma 1994, **103**:46–55. **Mayer C, Leese F, Tollrian R**. BMC genomics 2010, **11**:277. **Moyzis RK, Albright KL, Bartholdi MF, Cram LS, Deaven LL, Hildebrand CE, Joste NE, Longmire JL, Meyne J, Schwarzacher-Robinson T**. Chromosoma 1987, **95**:375-86. **Mural RJ, Adams MD, Myers EW, et al.** Science 2002, **296**:1661-71. **Namekawa SH, Payer B, Huynh KD, Jaenisch R, Lee JT**. Molecular and cellular biology 2010, **30**:3187-205. **Plohl M, Luchetti A, Mestrović N, Mantovani B**. Gene 2008, **409**: 72-82. **Probst AV, Okamoto I, Casanova M, El Marjou F, Le Baccon P, Almouzni G**. Developmental cell 2010, **19**:625-38. **Vissel B, Choo KH**. Genomics 1989, **5**:407-14. **Warburton PE, Hasson D, Guillem F, Lescale C, Jin X, Abrusan G**. BMC genomics 2008, **9**:533. **Waterston RH, Lindblad-Toh K, Birney E, et al.** Nature 2002, **420**:520-62. **Wong AK, Rattner JB**. Nucleic acids research 1988, **16**:11645-61.

Автор выражает искреннюю благодарность сотрудникам Группы некодирующей ДНК Института цитологии РАН, а в особенности научному руководителю О.И. Подгорной.