Вавиловский журнал генетики и селекции. 2016;20(6):804-815
DOI 10.18699/VJ16.198

Эволюция CpG-островов путем тандемных дупликаций

В.Н. Бабенко[1, 2] ✉, Ю.Л. Орлов[1, 2], Ж.Т. Исакова[3], Д.А. Антонов[4], М.И. Воевода[1, 5]

[1] Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук», Новосибирск, Россия
[2] Федеральное государственное автономное образовательное учреждение высшего образования «Новосибирский национальный исследовательский государственный университет», Новосибирск, Россия
[3] Научно-исследовательский институт молекулярной биологии и медицины при Национальном центре кардиологии и терапии им. Мирсаида Миррахимова при Министерстве здравоохранения Кыргызской Республики, Бишкек, Кыргызская Республика
[4] Федеральное государственное бюджетное научное учреждение «Научно-исследовательский институт экспериментальной и клинической медицины», Новосибирск, Россия
[5] Федеральное государственное бюджетное научное учреждение «Научно-исследовательский институт терапии и профилактической медицины» Сибирского отделения Российской академии медицинских наук, Новосибирск, Россия

CG-богатые острова (далее CpG-острова; CpG-islands, или CGI) являются важными функциональными элементами генома позвоночных. В частности, они инициируют транскрипцию как промоторы в большинстве (>50 %) генов позвоночных, в ряде случаев двунаправленную, в силу самокомплементарности cg динуклеотидов; формируют глобальный ландшафт метилирования; выступают как «выключатель» транскрипции через метилирование. Вырожденная природа CpG-островов (смещенный cg-состав) предполагает увеличение вероятности тандемных повторов и палиндромов внутри CpG-острова. Данная работа посвящена идентификации тандемных дупликаций полных CpG-островов, т. е. мегамономеров длиной 400–5000 п. н., в геноме человека. Были найдены меж- и внутригенные тандемные дупликации CpG-островов. Найденные межгенные CGI дупликации опосредовались через CG-богатые субцентромерные и теломерные сателлиты, а также SINE элементы. Высокое сходство мономеров тандемов в отдельных случаях предполагает существование давления отбора на структуру таких локусов. Исследование контекста межгенных тандемных CGI повторов указывает на их возможную роль в выравнивании cg-состава в геномном сегменте. Найденные тандемные CGI были транскрипционно активными в широком диапазоне тканей и клеточных линий. Рассмотренный феномен кластерной организации CGI наиболее выражен при рассмотрении хромосомы 19, известной своим избытком сегментных дупликаций и генных экспансий. К уникальным геномным сегментам относится также мегасателлит DXZ4 на q плече хромосомы X, который также попадает в категорию CpG-островов, порожденных тандемными дупликациями.

Ключевые слова: CpG-острова; gc состав; метилирование; эпигенетика, генные дупликации; хромосома 19.

КАК ЦИТИРОВАТЬ ЭТУ СТАТЬЮ:
Бабенко В.Н., Орлов Ю.Л., Исакова Ж.Т., Антонов Д.А., Воевода М.И. Эволюция CpG-островов путем тандемных дупликаций. Вавиловский журнал генетики и селекции. 2016;20(6):804-815. DOI 10.18699/VJ16.198

HOW TO CITE THIS ARTICLE:
Babenko V.N., Orlov Y.L., Isakova Zh.T., Antonov D.A., Voevoda M.I. Evolution of CpG-islands by means of tandem duplications. Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding. 2016;20(6): 804-815. DOI 10.18699/VJ16.198

ORIGINAL ARTICLE
Received 19.05.2015
Accepted for publication 28.12.2015

✉ e-mail: bob@bionet.nsc.ru

Evolution of CpG-islands by means of tandem duplications

V.N. Babenko[1, 2] ✉, Y.L. Orlov[1, 2], Zh.T. Isakova[3], D.A. Antonov[4], M.I. Voevoda[1, 5]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
[3] Institute of Molecular Biology and Medicine at National Center of Cardiology and Therapy named after academician Mirsaid Mirrakhimov under the Ministry of Health of the Kyrgyz Republic, Bishkek, Kyrgyzstan
[4] Institute of Clinical and Experimental Medicine SB RAS, Novosibirsk, Russia
[5] Research Institute of Internal and Preventive Medicine SB RAMS, Novosibirsk, Russia

CG-rich islands (CpG-islands, or CGI) are important functional elements in a genome of vertebrates. In particular, they: a) initiate transcription as promoters in most (> 50 %) genes of vertebrates, in some cases bi-directional, due to self-complement feature of cg dinucleotides; b) form a global methylation landscape; c) act as a transcription "switch" via methylation.
The degenerate nature of CpG-island (elevated CG composition) implies an increase in the probability of tandem repeats and palindromes within CpG-island. This work is devoted to the identification of tandem duplications of complete CpG-islands, i. e. considering mega monomers of size 400–5 000 bp, in the human genome. We found a range of inter- and intragenic tandem duplications of CpG-islands. Intergenic CpGi duplication mediates through CG-rich telomeric satellites, as well as elements of the SINE. One of the most pronounced tandems are located in chromosome 19, known for its abundance of segment duplications and gene expansion. We also underline the unique genomic segment, which is DXZ4 mega satellite, in q arm of chromosome X, also falling into the category of CpG-islands which evolved by tandem duplications rounds.

Key words: CpG-islands; tandem repeats; macrosatellites; human genome; methylation; CpG-island clusters.

Одна из характеристик геномов позвоночных – наличие множества так называемых CpG-островов, характеризующихся избытком CG динуклеотидов (Gardiner-Garden, Frommer, 1987). Формальное определение CpG-островов включает ряд параметров: процентное содержание cg-состава более 60 %; отношение cg_obs/cg_exp > 0.6; длина острова >300 п. н. Упомянутое отношение видоспецифично, в частности у мыши плотность cg динуклеотидов в CGI меньше, чем у человека (Illingworth, Bird, 2009). В геноме человека их число составляет примерно 26 тыс. (www.genome.ucsc.edu).

CpG-острова перекрываются с промоторной частью 50–70 % генов. Такие промоторы называют CpG промоторами (Babenko et al., 1999; Deaton, Bird, 2011). В генах гомеобоксов (HOX, PAX), а также примерно в 5 % других случаев CpG-островов промоторов CGI являются тканеспецифичной мишенью для метилтрансфераз: при метилировании промотора ген репрессируется.

CpG сайты в составе CpG-острова часто гомогенны по метилированию. Поэтому в качестве проб для платформы Illumina methyl 450 в ряде случаев берутся только один-два сайта острова (Sandoval et al., 2011).

Недавно было показано, что для репрессии гена *TTP* в клетках рака печени достаточно метилирования лишь одного CpG динуклеотида, находящегося на границе (берегу – shore) CpG-острова (Sohn et al., 2010). Возможно, такой мощный эффект метилирования одного сайта обусловлен разрушением пространственной структуры хроматина путем ликвидации сайта связывания инсулятора ctcf (Wang et al., 2012; Kang et al., 2015). Не исключена также ситуация, что состояние сайта CpG на границе CpG-острова связано со статусом метилирования острова в силу кооперативного метилирования (Haerter et al., 2014; Grandi et al., 2015).

Остальные 30–40 % островов, не связанных с промоторами генов, находятся во внутригенных районах (20 %) и в межгенных областях (12–15 %) (Deaton, Bird, 2011). Метилирование внутригенного CpG-острова часто усиливает экспрессию гена, предположительно за счет исключения образования спонтанного инициирующего Pol II-комплекса (Deaton, Bird, 2011), а также выраженного отсутствия нуклеосом из-за слабой гибкости ДНК (Vinogradov, 2005; Fenouil et al., 2012). При этом наблюдается градиент корреляции метилирования и экспрессии с максимальным положительным значением от 3′ конца гена.

В настоящее время известны семейства протяженных (более 10 т. п. н.) тандемных cg-богатых межгенных повторов человека, которые были названы мега- или макросателлитами (Chadwick, 2008; Tremblay et al., 2010). Одновременно они являются тандемно повторенными CpG-островами, поэтому мы остановимся на них более подробно. Мегасателлиты имеют специфическую номенклатуру, которая состоит из буквы D (duplication) или RS (repeated sequence) и последующего индекса хромосомы, после чего обычно идет буква Z. Они включают в себя DXZ4 на Xq23-24 (Giacalone et al., 1992; Horakova et al., 2012) и D4Z4 на 4q35/10q24 (Hewitt et al., 1994; Chadwick, 2008; Das, Chadwick, 2016). В базу данных RepeatMasker они не включены, поскольку не являются диспергированными и часто аннотированы как «простые повторы».

Макросателлиты характеризуются двумя общими признаками: высоким cg-составом и значительной (несколько т. п. н.) длиной мономера. Сами мономеры, как правило, не имеют межсателлитной гомологии, а также проявляют индивидуальную вариацию числа копий мономеров (Horakova et al., 2012), что используется как сигнатура популяции по аналогии с микросателлитами (Tremblay et al., 2010).

При наличии ряда межгенных тандемных CGI повторов большинство найденных в работе CGI повторов связаны с рекуррентной тандемной дупликацией генов умеренной длины, каковыми являются, в частности, гены *KRAB-ZNF*, и их промоторов (Lukic et al., 2014). Как правило, число мономеров в таких тандемах немногочисленно, но в некоторых случаях (кластеры некодирующих РНК) может доходить до 30–40.

В контексте генных дупликаций следует отдельно сказать о макросателлите (до недавнего времени генов в них не обнаруживалось) RS447 (4q17), не относящемся к CG-богатым сателлитам, но являющемся классической рекуррентной генной дупликацией. Его мономер содержит ген убиквитин-специфичной пептидазы 17 (11 генов USP17L). Небольшой кластер USP17L (2 гена) расположен также на хромосоме 8p23.1 (chr8:11.974.676-12.006.269). Гомология между мономерами макросателлита RS447 составляет более 99 %, поэтому различить отдельные копии современными методами (кроме клонирования) не представляется возможным (Treangen, Salzberg, 2012). Именно поэтому кластер, хотя он активно экспрессируется в ряде клеток, не аннотирован по состоянию хроматина и экспрессии: NGS секвенирование не может быть однозначно картировано. Это же касается и других мегасателлитов с высокогомологичными мономерами.

Задачей данной работы были поиск и анализ регулярных локальных повторов CpG-островов, представляющих как тандемные, так и диспергированные повторы.

Материалы и методы

Набор из 26 412 CpG-островов взят нами из таблицы cpgIslandExt (www.genome.ucsc.edu; version hg19). Список аннотированных повторов в зоне тандемных регулярных CpG-островов заимствован из базы данных (трека) результатов работы RepeatMasker (Smit, Hubley, 2008).

Оценка достоверности кластеризации *CGI*. Для поиска кластеров CGI геном был разбит на 10 т. п. н. сегменты (243 785). Частота попадания в сегмент острова (λ) равна 26 412/243 785 = 0.1. Оценка достоверно неслучайного числа островов в сегменте выполнялась на основе распределения Пуассона

$$P(X=k) = \exp(-\lambda)\frac{\lambda^{-k}}{k!}.$$

Вычисление вероятности P значения проводилось по формуле

$$P(X>k) = 1 - \sum_{n=0}^{k} P(X=n).$$

При ожидаемом числе CGI в сегменте λ = 0.1 интегральные вероятности (P-значения) следующие:
$P(X > 0)$ = 4.7E–3; $P(X > 1)$ = 1.6E–4; $P(X > 2)$ = 3.8E–6; $P(X > 3)$ = 7.7E–8.

Таким образом, для четырех и более CGI в сегменте вероятность составляет, с учетом поправки на множе-

ственные сравнения, порядка 1E–3, т. е. может считаться достоверным порогом.

Результаты

При рассмотрении распределения длины CGI в исходной выборке выяснилось, что медианная длина островов составляет 1 т. п. н. В этой связи нами выдвинуто предположение, что длинные CGI являются тандемным расширением островов мономеров. В частности, в исходной выборке CGI наблюдалось 16 островов длиннее 10 т. п. н. Они рассмотрены в отдельном разделе в конце статьи.

По остальным CGI был произведен поиск их кластеров по сегментам, согласно описанной методике. Всего выявлено 25 кластеров с общим числом CGI 141. Наибольшее число кластеров находилось на 19, 4, 2, 7 и 10 хромосомах (табл. 1).

Тандемные кластеры CpG-островов на хромосомах 4 и 19, опосредованные диспергированными повторами, относятся к макросателлитам и включают 64 острова. Их характеристика приведена далее. Ген-ориентированные CpG-острова (141 – 64 = 77) рассмотрены ниже.

Большинство кластеров CpG-островов расположено в районах генов гомеобокса (развития). Найденные 25 кластеров с числом CpG-островов более 3 на 10 т. п. н. были проаннотированы вручную (табл. 2). Обнаружено, что большинство кластеров, содержащих в общей сложности 141 CpG-остров, относятся к генам гомеобокса (*HOXa*,*c*) и нервного развития (*PAX2–5*). Число островов, расположенных в районах этих генов, составляет 45. Заметим, что ряд генов гомеобокса находятся в единичных, но протяженных (более 2 т. п. н.) CpG-островах, которые мы не рассматривали по методическим причинам. Например, ген *NKX6-2* на хромосоме 10 расположен внутри CpG-острова длиной более 3 т. п. н., который метилирован в большинстве тканей. О наличии множества CpG-островов в районах генов гомеобокса говорилось и ранее

Table 1. Prevalence of CGI clusters and CGIs on chromosomes, clusters having more than three monomers per 10 kb being taken into consideration

Numbers of	Chromosome													
	2	4	5	6	7	9	10	11	12	14	15	17	19	Total
CGI	13	20	9	4	13	8	12	8	4	4	6	8	32	141
Clusters	3	2	2	1	1	2	3	2	1	1	1	2	4	25

Table 2. Genes associated with CGI clusters

Gene class / location	Gene	Chromosome	Number of CGIs	Position
Homeobox	*ALX4*	chr11	4	44 325 657
Promoter	*Cyp26A1_C1*	chr10	4	94 825 546
Homeobox	*DLX1*	chr2	5	1.73E+08
Homeobox	*EN1*	chr2	4	1.2E+08
Promoter	*FAM89* (и *AS*)	chr11	4	65 337 213
Intragenic	*HMHA1*	chr19	4	1 066 866
Homeobox	*HOXa*	chr7	13	27 146 069
Homeobox	*HOXc12*	chr12	4	54 348 695
Membrane peptide	*HS3ST3B1*	chr17	4	14 200 579
Homeobox	*MSX1*	chr4	4	4 858 389
Myosin (promoter)	*MYO1C*	chr17	4	1 388 686
Homeobox	*NKX2-1*	chr14	4	36 986 362
Homeobox	*NKX2-3*	chr10	4	1.01E+08
Homeobox	*NKX2-5*	chr5	5	1.73E+08
Homeobox	*NRN1*	chr6	4	5 996 185
Homeobox (optical)	*PAX2*	chr10	4	1.02E+08
Homeobox (fetus)	*PAX3*	chr2	4	2.23E+08
Homeobox (CNS)	*PAX5*	chr9	4	37 025 492
Neurospecific gene (*protocadherin)	*PCDHGA1*	chr5	4	1.41E+08
Intragenic	*PTB1*	chr19	4	807 131

Rows with nonhomeobox genes are filled yellow.

Table 3. Locations of CGI tandem segments with > 10 monomers

Chromosome	Location (Mb)	Number of monomers	Gene	Annotated repeats	Segment features	PMID ID
19	36.7	34	ZNF	SST1	Promoter	21078170
4	132.6	17	–	SST1	–	21078170
X	115	45–112	–	Simple	DXZ4	Chadwick, 2008
X	120	12	CT47A	AluS	1 coding exon, 12 genes	16382448
X	3	11	PPP2R3B	AluSx1	Single gene	11173861
Y	2.5	11	PPP2R3B	AluSx1	Single gene	11173861

(Branciamore et al., 2010). В ряде злокачественных опухолей наблюдается гипометилирование данных генов. Такая этиология (гипометилирование *HOX*, *PAX* генов с последующей злокачественной пролиферацией) возникает, в частности, из-за мутаций метилтрансферазы DNMT3A (Qu et al., 2014).

Кроме пяти генов, кодирующих избыточные белки «домашнего хозяйства» и содержащих кластеры CpG-островов в промоторной области, такие как фактор сплайсинга PTB1, миозин MYO1C, гепарансульфат (мембранный белок) HS3ST3B1, транскрипционный фактор FAM89B, ген гистосовместимости *HMHA1*, цитохром P Cyp26A1_C1 (см. табл. 2). Стоит выделить отдельно ген протокадерина, содержащего cg-богатые кассеты экзонов и опосредующего специфическую адгезию нейронов.

Ген протокадерина (*PCDH*). На хромосоме 5 находится локус гена протокадерина (*PCDH*; см. табл. 2). Локус содержит 22 (аннотированных) транскрипта, разделенных на три подсемейства. Подсемейство A содержит 12 генов, подсемейство B – 7 генов и 2 псевдогена, а более отдаленное семейство G – 3 гена. Тандемный массив 22 CGI-содержащих экзонов, составляющих вариабельную часть, завершается константной, общей для всех подсемейств частью, включающей в себя три экзона. Постоянная часть гена кодирует цитоплазматическую часть. В каждом вариабельном экзоне имеется внеклеточный домен, состоящий из пяти кадериновых эктодоменов и трансмембранного района. Предполагается, что данные белки клеточной адгезии играют критическую роль в установлении и функционировании межклеточных связей в мозге.

Интересно, что благодаря CpG-островам каждый экзон вариабельного кластера также содержит промоторную часть для генов семейства. Эволюция этого гена происходила путем дупликации сегмента из двух соседних экзонов. Длина такого мономера превышала 7 т. п. н., длина каждого экзона в нем достигала 2400–2600 п. н. Подробнее механизм генерации транскриптов в генах этого уникального семейства рассмотрен в работах (Ong, Corces, 2014; Guo et al., 2015).

Сателлитные тандемы CGI, опосредованные аннотированными повторами. В отдельную группу были выделены повторы, имеющие в составе мономеров повторы, аннотированные в базе данных RepBase (Smit, Hubley, 2008–2015) (табл. 3).

Table 4. Distribution of SST1 repeats over chromosomes according to RepBase

Chromosome	Number of SST1 monomers	Mean length
chr12	10	710.3
chr16	15	522.7
chr17	27	805.9
chr19	56	1228.5
chr20	49	775.7
chr4	19	1519.9
chr6	3	345.0
chr7	35	757.6
chr9	50	885.9
chrY	321	295.5

Rows for macrosatellites on chromosomes 4 (4SST1) and 9 (19SST11, 19SST12) are filled with yellow.

Повторы, опосредованные SST1. Название умеренного SST1 происходит от рестрикционного энзима Sst1, гидролизующего мономер в одном сайте и приводящего к наблюдению спектра фрагментов, кратных 2.5 т. п. н., при разрезании тандемных кластеров этих повторов. Данный энзим имеет гомологию с аденовирусом (Epstein et al., 1987). Повтор аннотирован в RepBase как центромерный, является G/C богатым (g/c состав 67 %) и имеет наибольшую длину на хромосомах 4 и 19 (табл. 4).

Распределение числа SST1 повторов по хромосомам оценено нами с использованием аннотации RepeatMasker (Smit et al., 2008). Как можно видеть, геном человека содержит 590 SST1 повторов, большинство из которых являются «старыми» и имеют малую длину (см. табл. 4). Множество их следов находится на хромосоме Y.

Два «целевых» кластера макросателлитов SST1 в хромосомах 19 и 4, содержащих в своем мономере CpG-остров, были описаны ранее (Tremblay et al., 2010). Мономеры макросателлитов 4SST1 и 19SST11, 19SST12 на 4 и 19 хромосомах соответственно (Tremblay et al., 2010) являются трехсегментными и включают в себя мотив

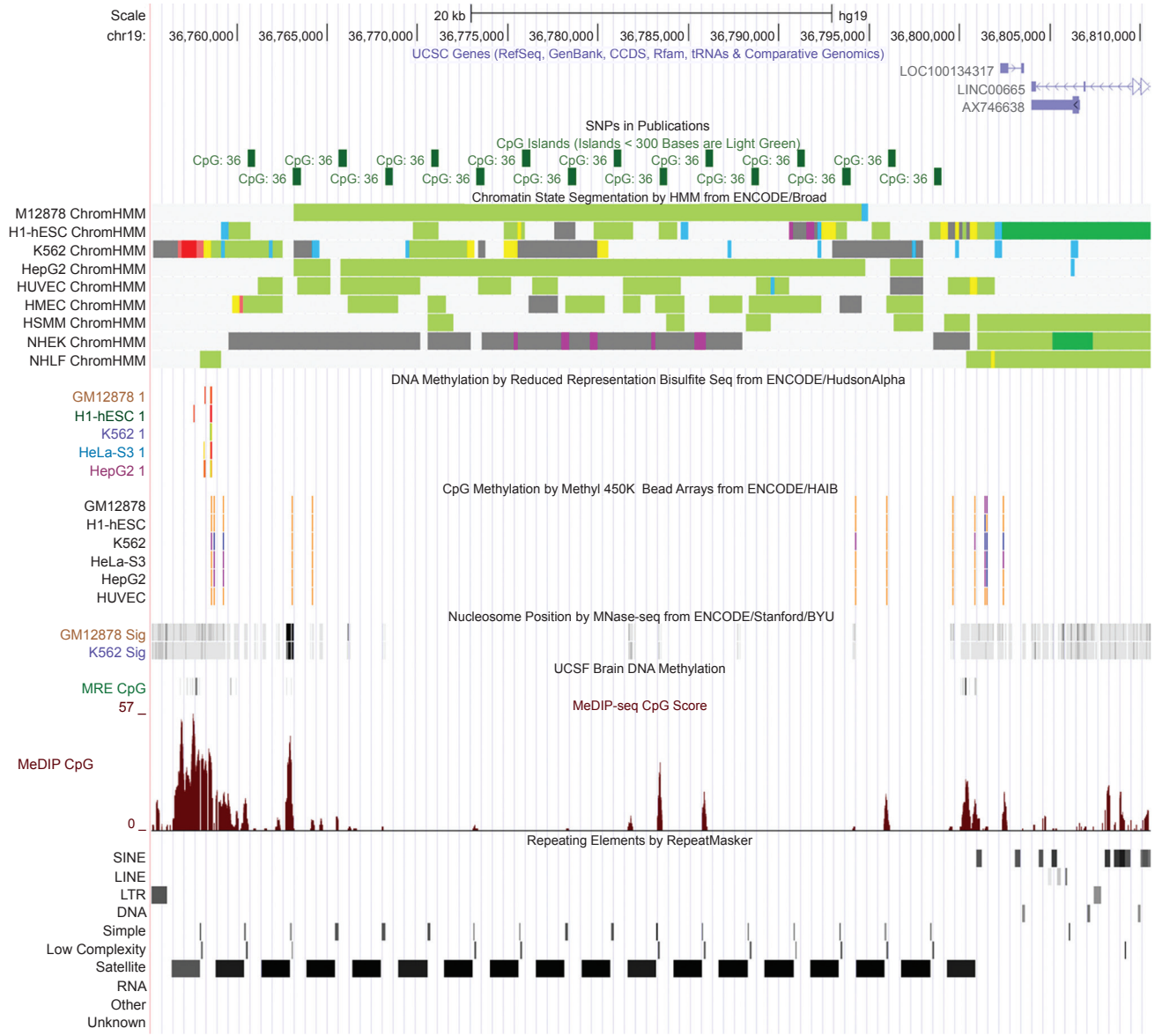

Fig. 1. Macrosatellite 19SST11 comprising monomers SST1_SR_CGI (chr19:36.755.278-36.810.578).

Three segments constituting the monomer are apparent from the RepeatMasker track (below) and green-colored CGI track (above): Satellite (SST1); Simple repeat (SR); CGI.

SST1, простой повтор (трек RepeatMasker) и CpG-остров (зеленый трек CpG-islands): SST1_SR_CGI (рис. 1). Длина данного мономера составляет 2.4 т. п. н.

Выявлено, что при достаточно неравномерной (не тандемной) общей кластеризации SST1 на хромосомах существуют три выраженных высокогомологичных тандема этих повторов: два – на хромосоме 19 и один – на хромосоме 4. От остальных SST1 кластеров они отличаются также тем, что находятся далеко от центромеры (19q12.12, 4q28.4) и содержат в составе мономера CpG-остров. В большинстве своем острова неметилированы, и, возможно, в силу этого три рассмотренных тандема повторов экспрессируются во всех тканях (Tremblay et al., 2010).

Два кластера тандемных высокогомологичных повторов со структурой мономера расположены в районе 19q13.12 (рис. 1 и 3). Мономер повтора состоял из SST1, спейсера (300 п. н.), содержащего в начале несколько простых коротких повторов, и CpG-острова длиной 400 п. н: (SST1, spacer, CGI). Кластеры были ориентированы противонаправленно.

Важно отметить, что и SST1, и CpG-острова ведут себя аналогично по отношению к нуклеосомам, а именно, имеют ярко выраженную «деплецию» нуклеосом (Vinogradov, 2005). При чередовании этих двух повторов (см. рис. 1 и 2) вся область тандемов оказалась свободной от нуклеосом (трек Nucleosome position by MNase-seq).

На основе выравнивания мономеров мы вычислили внутритандемную и межтандемную гомологию для трех рассмотренных тандемов (табл. 5). Из приведенных в таблице данных видна высокая гомология сателлитов, что указывает на возможную эволюционную поддержку

Fig. 2. Macrosatellite 4SST1 comprising 16 tandems of monomer SST1_CGI in the 4q28.3 region with high similarity of the monomers to those of macrosatellites 19SST11 and 19SST12 (see Table 4).

Table 5. Inter- and intra-tandem similarities calculated as the percentage of identity in blast pairwise alignment for three clusters: 19SST11, 19SST12, and 4SST1

Macrosatellite	19SST11	19SST12	4SST1
19SST11	99.8	97.1	93.4
19SST12		98.7	93.4
4SST1			94.6

консервативности этих сателлитов, наблюдающуюся в значительном числе тандемно повторяющихся треков (Warburton et al., 2008).

CpG промоторы 17 генов цинкового пальца (ZNF) между тандемами порождены SST1_CpGi мономерами. Схема района, насыщенная 56 SST1_CGI мономерами и состоящая из двух макросателлитов 19SST11 и 19SST12, а также кластера генов KRAB-ZNF, может быть представлена в виде распределения длин соседних мономеров (рис. 3).

Сверхдлинные CGI. Рассмотрены 16 CpG-островов длиной более 10 т. п. н., выделенные в отдельный класс (см. также (Warburton et al., 2008)). Как видно из аннотации (табл. 6), сверхдлинные острова, за рядом исключений, порождены тандемными повторами и во многих случаях проаннотированы как макросателлиты, поскольку

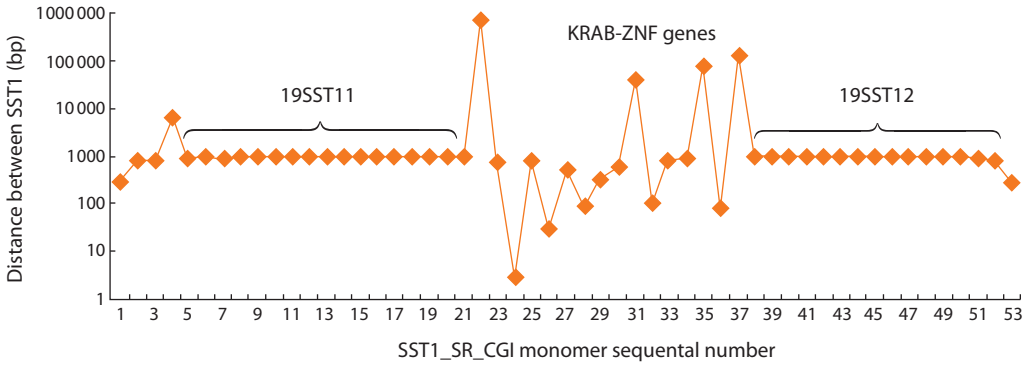

Fig. 3. Distribution of intermonomer distances between SST1 repeats in chromosome 19 macrosatellite doublet 19SST11-19SS12 (Chadwick, 2008).

The intersatellite region comprises 17 KRAB-ZNF genes whose promoters contain SST1_RS_CGI monomers (Table 7).

Table 6. Sixteen ultralong (> 10 kb) CGIs

#	Annotated repeats	Repeat name	Gene	Chromosome	Length (kb)	Hypometh
1	Simple	–	tRNA	chr1	32361	0
2	Simple	–	miRNA	chr1	40058	1
3	Simple	–	ZNF (antisense)	chr10	14472	1
4	–	–	Promoter of EBF3	chr10	10527	1
5	Telomere	LSAU	DUX cluster	chr10	20640	0
6	–	–	piRNA cluster	chr15	12438	0
7	–	–	TBX (AS)	chr17	10206	1
8	–	–	NKX2 homeobox 2	chr20	10782	1
9	–	–	–	chr4	13414	0
10	Telomere	LSAU	DUX cluster (D4Z4)	chr4	27227	0
11	–	–	FOXC1	chr6	11260	1
12	–	–	UNC homeobox (UNCX), mRNA	chr7	13699	1
13	–	–	Promoter of TNRC18	chr7	10753	1
14	–	–	Plectin (PLEC)	chr8	11865	0
15	–	–	Promoter of BCOR	chrX	15813	1
16	DXZ4	–	–	chrX	45712	0

–, no data.

картировать гены в них, в силу повторенности в начальной версии генома, не представлялось технически возможным (Treangen, Salzberg, 2012). Впоследствии гены были найдены, и макросателлиты были переаннотированы в генные кластеры – например, D4Z4 (DUX cluster). Они отличаются от разделенных кластеров CGI, рассмотренных выше, тем, что сливаются в один CpG-остров, хотя и имеют тандемную структуру.

После ручной аннотации (см. табл. 6) мы разделили 16 островов на следующие категории по типам генов:

1) кластеры некодирующих РНК (тРНК, миРНК, пиРНК – tRNA, miRNA, piRNA), найденные на хромосомах 1 (miRNA 40 т. п. н.; tRNA 32 т. п. н., см. табл. 6) и хромосоме 15 (piRNA, 12 т. п. н.). В последнем случае, несмотря на очевидную тандемную структуру, признаки наличия простых повторов отсутствуют, вероятно, вследствие высокой дивергенции;
2) два кластера гомеобокс-генов DUX (хр. 10, 20.6 т. п. н.) и DUX (D4Z4; хр. 4, 27.2 т. п. н.) в теломерах, перемежающиеся теломерными хромосом-специфичными cg-богатыми повторами LSAU в обоих случаях. Заметим, что по этому теломерному макросателлиту (D4Z4) наблюдалась внутрипопуляционная транслокация между сегментами хромосом 4q → 10q (Zhang et al.,

2005; Thijssen et al., 2014), благодаря значительной гомологии мономеров тандемов, как и в случае SST1 повторов (см. табл. 5);

3) гены гомеобокса FOXC1 (хр. 6, 11.2 т. п. н.), NKX2 (хр. 20, 10.7 т. п. н.) и UNCX (хр. 7, 13.7 т. п. н.). В этих случаях CGI образованы уникальными последовательностями, их размеры лишь слегка превышают 10 т. п. н., а гены лежат внутри CGI. Вне эмбриональных клеток эти районы гетерохроматизированы PcG комплексом. Упомянутые выше кластеры DUX также являются генами гомеобоксов, с последующей гетерохроматизацией или метилированием;
4) острова являются сверхдлинными промоторами генов EBF3 (хр. 10, 10.5 т. п. н.), TNRC18 (хр. 7, 10.7 т. п. н.), BCOR (хр. X, 15.8 т. п. н.) и в этом случае имеют структуру без повторов;
5) случаи бинаправленных промоторов генов цинковых пальцев и их антисмысловых партнеров: ZNF (хр. 10, 14.5 т. п. н.), TBX (хр. 17, 10.2 т. п. н.); имеются признаки тандемной структуры CGI;
6) оставшиеся неклассифицированные случаи включают вышеуказанный DXZ4 (хр. X, 45 т. п. н. в референсном геноме) и ген плектина PLEC (хр. 8, 11.8 т. п. н.), у которого протяженный остров находится внутри тела гена (см. табл. 6).

Достаточно часто наблюдалась нетандемная структура сверхдлинных CpG-островов, однако и в этих случаях не исключена гипотеза тандемного механизма возникновения и последующего вырождения: возможно, данные протяженные cg-богатые районы являются следствием древних дупликаций или резко повышенной скорости эволюции для приобретения новых свойств у паралогов.

В каждом случае, особенно для кластеров tRNA, miRNA, piRNA, каждый остров имеет свою историю и происхождение, а также отдельную функцию. Таким образом, несмотря на существующие закономерности тандемной структуры при возникновении сверхдлинных CGI, на этом этапе можно рассматривать данное явление как достаточно уникальный феномен.

В ряде случаев дупликация была, возможно, опосредована сателлитными повторами, как центромерными, так и теломерными, а также SINE Alu. Гены, входящие в сверхдлинные CGI, также являются гомеобоксными, или генами раннего развития. Их расположение в CGI островах обеспечивает возможности ткане-, стадио- и пол-специфичного выключения с помощью метилирования и последующего образования HP1 – гетерохроматина.

В контексте бинаправленных промоторов с антисмысловой lncRNA (см. табл. 6, строки 3 и 7) отметим, что при полногеномном сканировании в большом числе случаев наблюдается CGI промотор (Vavouri, Lehner, 2012) в силу его самокомплементарности, т. е. обогащения cg-состава одновременно в обеих нитях ДНК. Частое наличие инсулятора вблизи CGI промотора (Vavouri, Lehner, 2012) позволяет переключать транскрипцию с одной цепи на другую ткане- и стадиоспецифичным образом (Wood et al., 2013; Anderson, 2014). Следует упомянуть недавнюю статью об антиоксидантом ответе, где наблюдалась массовая экспрессия антисмысловых РНК (Giannakakis et al., 2015).

Обсуждение

Со времени открытия и формального определения CpG-островов (Gardiner-Garden, Frommer, 1987) выяснилось, что эти функциональные элементы играют крайне важную роль в функционировании геномов позвоночных (Deaton, Bird, 2011). В частности, большинство генов «домашнего хозяйства» имеют в промоторной области данный остров, распознаваемый множеством транскрипционных факторов (Sp1, CTCF и др.), и обеспечивают инициацию транскрипции соответствующих генов.

Позже выяснилось, что около 30 % CpG-островов не имеют отношения к промотору и расположены в меж- и внутригенных областях. Эволюционная консервативность данных элементов, показанная в ряде работ (Deaton, Bird, 2011), свидетельствует об их функциональной нагрузке, которая пока до конца не определена. Выдвигались гипотезы, что в целом CpG-острова связаны с глобальным ландшафтом метилирования, т. е. имеют эпигенетическое значение (Deaton, Bird, 2011). Не случайно CpG-острова часто встречаются в районах импринтированных генов (в частности, *PPA2C*, *DNMT1*, *TSHZ3*, *CHST8*, *ZNF225^*, *ZNF229^*, *DMWD*, *ZNF331^*, *LILRB4*, *NLRP2*, *ZIM2^*, *PEG3^*, *MIMT1*, *USP29*, *ZIM3^*, *ZNF264^*, *CHMP2A*, *MZF1^*), причем не только в промоторе, но и в их расширенном контексте (Deaton, Bird, 2011).

Режим эволюции CpG-островов путем дупликации не был рассмотрен ранее. Имеются факты, подтверждающие их адаптационное значение. В частности, на хромосоме 19 произошло достоверно большее число генных и сегментных дупликаций по сравнению с остальными хромосомами (Grimwood et al., 2004). При этом гены, содержащие активный CpG промотор, сохраняли его и при дупликациях. Тем не менее, кроме промоторных CpG, хромосома 19 содержит значительное число 3′UTR и внутригенных CpG-островов, которые в подавляющем числе случаев метилированы, что связано с усилением экспрессии.

Использование протяженных районов сильно смещенного нуклеотидного состава, каковыми являются CpG-острова, в процессе эволюции могло происходить по причинам: а) открытости района, практически полностью свободного от нуклеосом и, как правило, транскрибируемого в широком диапазоне тканей; б) использования cg-богатого состава для дупликации вследствие неравного кроссинговера; в) возможности выключения генов путем метилирования и последующей гетерохроматизации промотора или всего гена.

В нашей работе мы выявили тандемные дупликации CpG-островов в геноме человека. Про cg-богатые тандемные повторы в некодирующих областях сообщалось достаточно давно в контексте полиморфных сателлитов (Schaap et al., 2013). Дело в том, что эти макросателлиты полиморфны по длине, т. е. встречаются в геномах с вариабельным числом в силу повышенной митотической рекомбинации (Schaap et al., 2013). Это свойство позволяет использовать их как популяционно-специфический маркер. Наибольшая вариация cg-богатых тандемов отмечается у африканцев, наименьшая – у азиатов (Schaap et al., 2013). CpG-острова используются в популяционной генетике и имеют следующие названия (Schaap et al., 2013): RS447

Table 7. Genes and CGIs located between macrosatellites 19SST11–19SST12

Cpg_id	CpG island	Location	Gene
		Macrosatellite 19SST11	
9361	CpG: 56	36 869 564	*ZFP14*
9362	CpG: 66	36 909 281	*ZFP82*
9363	CpG: 74	36 912 260	*LOC644189*
9364	CpG: 44	36 980 190	*ZNF566*
9365	CpG: 66	37 018 919	*ZNF260*
9366	CpG: 69	37 063 892	*ZNF529*
9367	CpG: 86	37 095 680	*ZNF382*
9367	CpG: 86	37 095 680	*ZNF529*
9368	CpG: 39	37 157 632	*ZNF461*
9370	CpG: 53	37 263 381	*ZNF850*
9372	CpG: 40	37 288 342	*LOC284408*
9373	CpG: 49	37 328 896	*ZNF790*
9374	CpG: 26	37 340 918	*ZNF345*
9374	CpG: 26	37 340 918	*ZNF790*
9375	CpG: 48	37 406 931	*ZNF568*
9375	CpG: 48	37 406 931	*ZNF829*
9377	CpG: 56	37 568 952	*ZNF420*
		Macrosatellite 19SST12*	
9394	CpG: 45	37 825 101	*HKR1*
9395	CpG: 57	37 861 691	*ZNF527*
9397	CpG: 57	37 957 726	*ZNF569*
9398	CpG: 64	37 959 852	*ZNF570*
9399	CpG: 26	37 997 790	*ZNF793*
9400	CpG: 66	38 039 561	*LOC100507433*
9401	CpG: 56	38 085 148	*ZNF540*
9401	CpG: 56	38 085 148	*ZNF571*
9402	CpG: 119	38 145 826	*ZFP30*
9403	CpG: 47	38 182 793	*ZNF781*
9404	CpG: 50	38 210 107	*ZNF607*
9405	CpG: 26	38 270 79	*ZNF573*

* Immediately after the tandem pair following macrosatellite 19SST12. For continuation of imperfect tandem duplications, see Fig. 3.

(chromosome 4p non CpGi), MSR5p (5p), FLJ40296 (13q), RNU2 (17q) и D4Z4 (4q and 10q), а также X chromosomal DXZ4 и CT47.

Структурные свойства района макросателлитов 19SST11 и 19SST12 и генного кластера между ними. Наибольшими по размеру найденными тандемными кластерами CGI являются макросателлиты 19SST11 и 19SST12 на хромосоме 19 (см. табл. 4). Их длина в общей сложности составляет около 1 млн п. н. Между ними и сразу за ними находятся кластеры генов KRAB-ZNF (табл. 7, см. рис. 3), имеющих в качестве промотора мономер фланкирующих сателлитов. Существует гипотеза о возможности гомологичной неравной рекомбинации на основе двух рассмотренных тандемных инвертированных кластеров 19SST11–19SST12 в районе 19q13.12, которая развернула сегмент 1 млн п. н. между повторами (Tremblay et al., 2010). Высокое сходство тандемов, расположенных в обратной ориентации, делает такую гипотезу правдоподобной. Это также подтверждается локальной мозаичной синтенией с позвоночными, описанной в работе (Grimwood et al., 2004).

Полногеномный анализ SST1 повтора показал, что он транскрипционно активен (в большинстве тканей) на всех хромосомах, кроме облигатно репрессированного сегмента хромосомы Y, где расположены короткие следы повтора (см. табл. 4). Области локусов, включающие в себя SST1,

в большинстве своем не содержат CpG-островов, за исключением рассмотренных трех тандемных кластеров на хромосоме 19 и 4, а также случаев SST1_CpG промоторов генов KRAB-ZNF, расположенных между тандемами на хромосоме 19 (см. рис. 3 и табл. 7).

Тандемные кластеры SST1_CGI хромосомы 19 находятся в районе гетерохроматина B4 (Rao et al., 2014), изобилующего кластерами KRAB-ZNF генов (см. табл. 7). Данные гены имеют отношение к метилированию ретровирусов в эмбриогенезе (Lukic et al., 2014) и реплицируются в ранней G-фазе (Rao et al., 2014). Соответственно, они репрессированы гетерохроматином HP1, который характерен для доменов B4 (Rao et al., 2014). Интродукция кластеров SST1_CGI «открывает» хроматин, освобождая его от нуклеосом. Об этом также свидетельствует базальная транскрипция (см. рис. 1 и 2, зеленые линии в треке HMM from ENCODE для клеточных линий), наблюдаемая в данных районах как на хромосоме 19, так и на хромосоме 4 во всех тканях (Tremblay et al., 2010). Тем не менее в большинстве зрелых тканей эти районы закрыты HP1 хроматином.

Наличие сайтов связывания CTCF в районах кластеров CGI. На рис. 1 и 2 видно также избирательное метилирование определенных CGI (коричневый трек MeDIP CpG) и наличие CTCF сайтов связывания (см. рис. 1, голубые полосы (Das, Chadwick, 2016)). В литературе имеются ссылки на связь тандемных повторов CGI, состояния хроматина, а также ctcf инсуляторов, которые в совокупности модулируют трехмерную конформацию хроматина, в частности на инактивированной X хромосоме (Rao et al., 2014). Избирательное выпетливание, опосредованное ctcf инсуляторами, приводит также к множественному альтернативному сплайсингу CGI-релевантных первых экзонов гена протокадерина PCDHA, рассмотренного выше. Таким образом, данная «структурная аномалия» на хромосоме 19, вполне возможно, связана с пространственной координацией хроматина (Rao et al., 2014; Das, Chadwick, 2016).

Кроме этого, инсуляторный функциональный потенциал, представленный многочисленными сайтами связывания ctcf фактора в CGI тандемах, используется и в случае макросателлита D4Z4 (Zhang et al., 2005; Ottaviani et al., 2010; Darrow, Chadwick, 2014; Das, Chadwick, 2016), а также в генах протокадеринов для осуществления альтернативного сплайсинга (см. табл. 2) (Nichols, Corces, 2015), в кластере тРНК (см. табл. 7) (Darrow, Chadwick, 2014), кластере пиРНК (piwiRNA) (см. табл. 7) (Williams et al., 2015), а также в сцепленных с болезнями кластеров двойных гомеобоксах DUX4 (см. табл. 7) (Das, Chadwick, 2016). Это подчеркивает тесную функционально-регуляторную связь CpG-островов и ctcf инсуляторов как в промоторах (Vavouri, Lehner, 2012; Kang et al., 2015), так и в некодирующих тандемных кластерах, таких как DXZ4 (Ottaviani et al., 2010; Horakova et al., 2012; Wang et al., 2012; Darrow, Chadwick, 2014; Rao et al., 2014). Как правило, все сверхдлинные CGI (см. табл. 7) содержат сайты связывания инсуляторов CTCF.

При оценке внутри- и межтандемной гомологии (см. табл. 3) выяснилось, что CpG-острова имеют крайне высокую гомологию, что говорит об их возможной функциональной значимости. В силу этих же причин тандемы мономеров SST1_CGI могут выступать в качестве рестрикционных макросателлитных маркеров (Tremblay et al., 2010), хотя их использование затруднено наличием по крайней мере трех высокогомологичных тандемных повторов на двух хромосомах, что снижает их специфичность.

DXZ4: феномен сверхдлинного CGI. Макросателлит DXZ4 (45–120 т. п. н., см. табл. 6, строка 16) составляет единый CpG-остров. Мономеры CGI (DXZ4) содержат три структурных элемента, все из которых CG-богатые (Horakova et al., 2012). DXZ4 является консервативным среди приматов и, по всей видимости, кодирует длинную некодирующую РНК, хотя, возможно, происходит просто базальная транскрипция CpG богатого неметилированного сегмента. Его длина сильно варьирует среди популяций по числу тандемов: от 12 до 99, среднее число мономеров 54 (Horakova et al., 2012). Таким образом, DXZ4 – также популяционный маркер.

Тандемный массив DXZ4 (Chadwick, 2008; Horakova et al., 2012; Rao et al., 2014) разделяет инактивированную хромосому X на два пространственных супердомена: (0–115 млн п. н.) (DXZ4) и (115–153.3 млн п. н.) (Rao et al., 2014). Инициация эффекта «ламповых щеток» при инактивировании X хромосомы происходит эухроматизацией DXZ4 путем деметилирования, в отличие от гиперметилированного состояния на активной X хромосоме. На инактивированной X хромосоме с сегментом DXZ4 связывается ряд CTCF факторов, позволяя обеспечить до 27 больших суперпетель размером от 7 до 74 млн п. н. (Rao et al., 2014). Организация суперпетель идет подавляющим образом с помощью ctcf факторов: его содержат 23 из 24 районов основания петель (Rao et al., 2014). Четыре района основания суперпетель содержат длинные некодирующие РНК (lncRNA), а именно XIST, loc550643, DXZ4 и FIRRE. Интересно, что два первых являются cg-богатыми и гиперметилированы в активной X хромосоме, что предотвращает связывание ctcf факторов (Wang et al., 2012).

Феномен инактивации X хромосомы, а также результаты исследования взаимодействия метилирования и сайтов связывания ctcf факторов (Wang et al., 2012; Kang et al., 2015) выявляют антагонистическое взаимодействие ctcf фактора и метилирования в cg-богатых районах, каковыми являются CGI.

Заключение

Сосредоточившись в основном на аннотации районов и выявлении общих характеристик, для их поиска мы использовали метод кластеризации CGI, что является грубым приближением при поиске тандемных дупликаций. Тем не менее в работе по выявлению длинных тандемных повторов с помощью специализированной программы Tandem Repeats Finder – TRF (Warburton et al., 2008) содержатся все основные макросателлиты/тандемные дупликации, которые мы нашли с учетом CGI-релевантности, за исключением несовершенных, как piРНК кластер.

Мы представили относительно немногочисленные, но выраженные случаи множественной дупликации CpG-островов, а также сверхдлинных CGI (см. табл. 2 и 6). Выяснилось, что основные аннотированные повторы,

опосредующие консервативную дупликацию, относятся к типам SINE (AluS на половых хромосомах) и центромерным повторам SST1 на хромосомах 4 и 19. Кроме этого, существуют неаннотированные «простые» cg-богатые повторы, которые создали уникальный якорный CGI остров DXZ4 на хромосоме X, участвующий в ее инактивации, а также D4Z4 (см. табл. 2 и 6).

В работе показано, что большинство плотных кластеров CGI, как раздельных (см. табл. 2), так и слитных (см. табл. 6), порождено тандемными дупликациями. Эти кластеры связаны с генами раннего развития, включая протяженный кластер piPHK (piwiRNA) (Williams et al., 2015), а также KRAB-ZNF гены, расположенные на хромосоме 19 в районе B4 хроматина.

Acknowledgments

This work was supported by the Russian Science Foundation, project 14-24-00123. The authors are grateful to the SB RAS supercomputing center (http://www2.sscc.ru) for access to computational resources.

Conflict of interest

The authors declare no conflict of interest.

References

Anderson S.K. Probabilistic bidirectional promoter switches: noncoding RNA takes control. Mol. Ther. Nucl. Acids. 2014;3:e191.

Babenko V.N., Kosarev P.S., Vishnevsky O.V., Levitsky V.G., Basin V.V., Frolov A.S. Investigating extended regulatory regions of genomic DNA sequences. Bioinformatics. 1999;15(7-8):644-653.

Branciamore S., Chen Z.X., Riggs A.D., Rodin S.N. CpG island clusters and pro-epigenetic selection for CpGs in protein-coding exons of HOX and other transcription factors. Proc. Natl. Acad. Sci. USA. 2010;107(35):15485-15490.

Chadwick B.P. DXZ4 chromatin adopts an opposing conformation to that of the surrounding chromosome and acquires a novel inactive X-specific role involving CTCF and antisense transcripts. Gen. Res. 2008;18:1259-1269. DOI 10.1101/gr.075713.107.

Darrow E.M., Chadwick B.P. A novel tRNA variable number tandem repeat at human chromosome 1q23.3 is implicated as a boundary element based on conservation of a CTCF motif in mouse. Nucl. Acids Res. 2014;42(10):6421-6435.

Das S., Chadwick B.P. Influence of repressive histone and DNA methylation upon D4Z4 transcription in non-myogenic cells. PLoS One. 2016;11(7):e0160022.DOI 10.1371/journal.pone.0160022.

Deaton A.M., Bird A. CpG islands and the regulation of transcription. Gen. Dev. 2011;25(10):1010-1022.

Epstein N.D., Karlsson S., O'Brien S., Modi W., Moulton A., Nienhuis A.W. A new moderately repetitive DNA sequence family of novel organization. Nucl. Acids Res. 1987;15:2327-2341.

Fenouil R., Cauchy P., Koch F., Descostes N., Cabeza J.Z., Innocenti C., Ferrier P., Spicuglia S., Gut M., Gut I., Andrau J.C. CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. Genome Res. 2012;22(12):2399-2408.

Gardiner-Garden M., Frommer M. CpG islands in vertebrate genomes. J. Mol. Biol. 1987;196:261-282.

Giacalone J., Friedes J., Francke U. A novel GC-rich human macrosatellite VNTR in Xq24 is differentially methylated on active and inactive X chromosomes. Nat. Genet. 1992;1(2):137-43.

Giannakakis A., Zhang J., Jenjaroenpun P., Nama S., Zainolabidin N., Aau M.Y., Yarmishyn A.A., Vaz C., Ivshina A.V., Grinchuk O.V., Voorhoeve M., Vardy L.A., Sampath P., Kuznetsov V.A., Kurochkin I.V., Guccione E. Contrasting expression patterns of coding and noncoding parts of the human genome upon oxidative stress. Sci. Rep. 2015;5:9737. DOI 10.1038/srep09737.

Grandi F.C., Rosser J.M., Newkirk S.J., Yin J., Jiang X., Xing Z., Whitmore L., Bashir S., Ivics Z., Izsvák Z., Ye P., Yu Y.E., An W. Retrotransposition creates sloping shores: a graded influence of hypomethylated CpG islands on flanking CpG sites. Genome Res. 2015; 25(8):1135-1146.

Grimwood J., Gordon L.A., Olsen A., Terry A., Schmutz J., Lamerdin J., … Stubbs L., Rokhsar D.S., Myers R.M., Rubin E.M., Lucas S.M. The DNA sequence and biology of human chromosome 19. Nature. 2004;428(6982):529-535.

Guo Y., Xu Q., Canzio D., Shou J., Li J., Gorkin D.U., Jung I., Wu H., Zhai Y., Tang Y., Lu Y., Wu Y., Jia Z., Li W., Zhang M.Q., Ren B., Krainer A.R., Maniatis T., Wu Q. CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function. Cell. 2015; 162(4):900-910. DOI 10.1016/j.cell.2015.07.038.

Haerter J.O., Lövkvist C., Dodd I.B., Sneppen K. Collaboration between CpG sites is needed for stable somatic inheritance of DNA methylation states. Nucl. Acids Res. 2014;42(4):2235-2244.

Hewitt J.E., Lyle R., Clark L.N., Valleley E.M., Wright T.J., Wijmenga C., van Deutekom J.C.T., Francis F., Sharpe P.T., Hofker M., Frants R.R., Williamson R. Analysis of the tandem repeat locus D4Z4 associated with facioscapulohumeral muscular dystrophy. Hum. Mol. Genet. 1994;3(8):1287-1295. DOI 10.1093/hmg/3.8.1287.

Horakova A.H., Moseley S.C., McLaughlin C.R., Tremblay D.C., Chadwick B.P. The macrosatellite DXZ4 mediates CTCF-dependent long-range intrachromosomal interactions on the human inactive X chromosome. Hum. Mol. Genet. 2012;21(20):4367-4377.

Illingworth R.S., Bird A.P. CpG islands – 'a rough guide'. FEBS Lett. 2009;583(11):1713-20. DOI 10.1016/j.febslet.2009.04.012.

Kang J.Y., Song S.H., Yun J., Jeon M.S., Kim H.P., Han S.W., Kim T.Y. Disruption of CTCF/cohesin-mediated high-order chromatin structures by DNA methylation downregulates PTGS2 expression. Oncogene. 2015;34:5677-5684. DOI 10.1038/onc.2015.17.

Lukic S., Nicolas J.C., Levine A.J. The diversity of zinc-finger genes on human chromosome 19 provides an evolutionary mechanism for defense against inherited endogenous retroviruses. Cell Death Differ. 2014;21(3):381-387. DOI 10.1038/cdd.2013.150.

Nichols M.H., Corces V.G. A CTCF code for 3D genome architecture. Cell. 2015;162(4):703-705. DOI 10.1016/j.cell.2015.07.053.

Ong C.T., Corces V.G. CTCF: an architectural protein bridging genome topology and function. Nat. Rev. Genet. 2014;15(4):234-246.

Ottaviani A., Schluth-Bolard C., Gilson E., Magdinier F. D4Z4 as a prototype of CTCF and lamins-dependent insulator in human cells. Nucleus. 2010;1(1):30-36.

Qu Y., Lennartsson A., Gaidzik V.I., Deneberg S., Karimi M., Bengtzén S., Höglund M., Bullinger L., Döhner K., Lehmann S. Differential methylation in CN-AML preferentially targets non-CGI regions and is dictated by DNMT3A mutational status and associated with predominant hypomethylation of HOX genes. Epigenetics. 2014; 9(8):1108-1119.

Rao S.S., Huntley M.H., Durand N.C., Stamenova E.K., Bochkov I.D., Robinson J.T., Sanborn A.L., Machol I., Omer A.D., Lander E.S., Aiden E.L. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159(7):1665-1680.

Sandoval J., Heyn H., Moran S., Serra-Musach J., Pujana M.A., Bibikova M., Esteller M. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. Epigenetics. 2011; 6(6):692-702.

Schaap M., Lemmers R.J., Maassen R., van der Vliet P.J., Hoogerheide L.F., van Dijk H.K., Baştürk N., de Knijff P., van der Maarel S.M. Genome-wide analysis of macrosatellite repeat copy number variation in worldwide populations: evidence for differences and commonalities in size distributions and size restrictions. BMC Genomics. 2013;4(14):143. DOI 10.1186/1471-2164-14-143.

Smit A.F.A., Hubley R. RepeatModeler Open-1.0. 2008–2015. http://www.repeatmasker.org.

Sohn B.H., Park I.Y., Lee J.J., Yang S.J., Jang Y.J., Park K.C., Kim D.J., Lee D.C., Sohn H.A., Kim T.W., Yoo H.S., Choi J.Y., Bae Y.S., Yeom Y.I. Functional switching of transforming growth factor-beta1 signaling in liver cancer via epigenetic modulation of a single CpG site in tristetraprolin promoter. Gastroenterol. 2010;138:1898-1908.

Thijssen P.E., Balog J., Yao Z., Pham T.P., Tawil R., Tapscott S.J., van der Maarel S.M. DUX4 promotes transcription of FRG2 by directly activating its promoter in facioscapulohumeral muscular dystrophy. Skeletal Muscle. 2014;4:19. DOI 10.1186/2044-5040-4-19.

Treangen T.J., Salzberg S.L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat. Rev. Genet. 2012;13:36-46.

Tremblay D.C., Alexander G., Jr., Moseley S., Chadwick B.P. Expression, tandem repeat copy number variation and stability of four macrosatellite arrays in the human genome. BMC Genomics. 2010;15(11):632. DOI 10.1186/1471-2164-11-632.

Vavouri T., Lehner B. Human genes with CpG island promoters have a distinct transcription-associated chromatin organization. Genome Biol. 2012;13(11):R110.

Vinogradov A.E. Noncoding DNA, isochores and gene expression: nucleosome formation potential. Nucl. Acids Res. 2005;33(2):559-563.

Wang H., Maurano M.T., Qu H., Varley K.E., Gertz J., Pauli F., Lee K., Canfield T., Weaver M., Sandstrom R., Thurman R.E., Kaul R., Myers R.M., Stamatoyannopoulos J.A. Widespread plasticity in CTCF occupancy linked to DNA methylation. Gen. Res. 2012;22(9):1680-1688.

Warburton P.E., Hasson D., Guillem F., Lescale C., Jin X., Abrusan G. Analysis of the largest tandemly repeated DNA families in the human genome. BMC Genomics. 2008;9:533. DOI 10.1186/1471-2164-9-533.

Williams Z., Morozov P., Mihailovic A., Lin C., Puvvula P.K., Juranek S., Rosenwaks Z., Tuschl T. Discovery and characterization of piRNAs in the human fetal ovary. Cell Rep. 2015;13(4):854-863.

Wood E.J., Chin-Inmanu K., Jia H., Lipovich L. Sense-antisense gene pairs: sequence, transcription, and structure are not conserved between human and mouse. Front. Genet. 2013;4:183.

Zhang Y.Z., Sun S.C., Wu H.C., Fan Q.S., Song Y.J., Yu W., Jeanpierre M., Urtizberea J.A. Polymorphism of the D4Z4 locus associated with facioscapulohumeral muscular dystrophy 1A in Shanghai population. Zhonghua Yi Xue Yi Chuan Xue Za Zhi (=Chinese J. Med. Genetics). 2005;22(4):380-382.