***Доклады Международной конференции "Математическая биология и биоинформатика"**. Под ред. В.Д. Лахно. Том 7. Пущино: ИМПБ РАН, 2018. Статья № e103. doi: 10.17537/icmbb18.114*

Проблема доминирования A/T треков в геноме человека

Комаров В.М., Самченко А.А., Кондратьев М.С.

Институт биофизики клетки РАН ФИЦ ПНЦБИ РАН, Пущино, Московская область, Россия

komarov@icb.psn.ru

В работе с использованием методов сравнительной геномики обсуждается проблема неравноправия встречаемости AT- и GC-пар в ДНК структуре генома человека. На фоне общего характера доминирования A/T-треков в структуре целых хромосом обнаружено лимитирующее влияние минисателитных $S(G/C)_n$-треков на формирования специфики GC-состава их экзонных областей. Полученные результаты позволили весь хромосомный набор генома человека систематизировать на III различных типа хромосом, различающихся GC-составом своих экзонных областей и степенью доминирования в них относительно коротких до 35 пар нуклеотидов S-треков смешанной G/C-природы.

Ключевые слова: геном человека, структурный полиморфизм AT и GC пар, нуклеотидные последовательности, доминирование A/T-треков, GC-состав экзонов, систематизация хромосом.

A/T tracts domination problem in the human genome

Komarov V.M., Samchenko A.A., Kondratiev M.S.

Institute of Cell Biophysics of RAS, Pushchino 142290, Moscow region, Russia

Using the comparative genomics technique, the problem of unequal occurrence of the AT and GC pairs in the DNA structure of the human genome is discussed. Contrary to the general predominance of A/T tracks in the whole chromosome structure, the limiting role of S(G/C) tracks in the exon GC content formation was reveal. The results obtained allowed us the entire chromosomal set of the human genome to be systematized on III types, differing in exon GC content and the domination degree of minisatellite S(G/C) tracks.

Key words: *human genome, nucleotide sequences, structural polymorphism of AT- and GC- pairs, domination of A/T tracks, GC-content of exons, systematization of chromosomes.*

Введение

Успехи секвенирования геномов различных видов организмов про и эукариот все больше и больше указывают на нетривиальное своеобразие распределения треков из AT- или GC-пар в структуре ДНК. Хотя функции большинства из найденных повторяющихся или одиночных нуклеотидных последовательностей четко пока не определены некоторые закономерности в частотах появления, например, треков $(A/T)_n$- или $(G/C)_n$-типа в кодирующих и регуляторных областях геномов уже обнаруживаются (см. например [1–4]).

В подобного рода исследованиях распространено представление, что важнейшим фактором структурно-функциональной организации любого генома всегда выступает его GC-состав. Это мнение базируется на известной повышенной термодинамической устойчивости уотсон-криковских GC-пар где в водородном связывании принимают участие три, а не две H-связи по сравнению с AT-парами. В подтверждение правомерности такого подхода используют найденные корреляции между GC-составом некоторых геномов и степенью их устойчивости к УФ-радиации [5], способности к термо-адаптации [6], величиной размера генома [7], длиной кодирующей последовательностей [8], степенью влияния условий окружающей среды [9], мутационной способностью [10], скоростью транскрипции [11].

Однако такой упрощенный подход выглядит, на наш взгляд, весьма ограниченным. Так, в качестве универсального маркера справедливости этого приближения очень часто указывают на широкий диапазон наблюдаемых вариаций GC состава, от 16 % до 75 %, генома микроорганизмов [12, 13], что действительно впечатляет. А вместе с тем, имеющаяся статистика по встречаемости организмов, содержащих в целом более 1700 прокариотических хромосом и плазмид,

***Proceedings of the International Conference "Mathematical Biology and Bioinformatics".** Ed. V.D. Lakhno. Vol. 7. Pushchino: IMPB RAS, 2018. Paper No. e103. doi: 10.17537/icmbb18.114*

говорит [14], что чаще всего встречаются организмы где весьма высок, ~ 65 %, вклад уотсон-криковских AT-, а не GC-пар оснований в составе ДНК.

С другой стороны у геномов эукариот, например высших растений животных и человека, соотношение вкладов AT- и GC-пар тоже оказывается, как ни странно, сильно смещенным в сторону AT-, а не GC-пар (причем с довольно большим превышением [15, 16]). Известны примеры геномов эукариот, где AT-состав оказывается больше 82 % [17, 18], что значительно выше предельного значения GC-состава в 75 %, характерного для прокариот [12, 13].

Наблюдается своеобразие встречаемости и повторяющихся нуклеотидных последовательностей в структуре ДНК. Ранее нами с использованием методов сравнительной геномики впервые было показано, что для 400 представителей архейных и эубактериальных хромосом в 75 % случаев характерным оказывается преобладание oligo(dA)$_n$- и oligo(dT)$_n$-повторов над oligo(dG)$_n$- и oligo(dC)$_n$-нуклеотидными повторами (где $n \geq 5$) [19, 20]. Причем это доминирование сохранялось для хромосом, в составе которых GC-пар было явно больше половины. В структуре геномов эукариот с разным GC-составом, меняющимся от 25 % до 60 %, также впервые нами было выявлено общее доминирование мононуклеотидных и смешанных треков (A/T)$_n$-типа по сравнению с треками из GC-пар [21]. Было дано обоснование наиболее вероятной физической первопричины такого поведения AT-пар в ДНК про- и эукариот. Оно заключается в проявлении скрытого полиморфизма водородного спаривания у одиночных уотсон-криковских AT- и GC- пар. Здесь из-за присутствия в структуре азотистых оснований амино групп пирамидального строения, GC-пара оказывается обладающей четырехкратным вырождением, а AT-пара двукратным вырождением геометрии своего комплементарного H-спаривания. Повышенная неоднозначность исходной геометрии уотсон-криковских GC-пар и инициируют наблюдаемую предпочтительность и надежность «использования» природой AT-пар в структурно-функциональной организации геномной ДНК любого организма [19–21]. Остается необходимость уточнения другого важного момента: существуют ли тогда какие-то свои, дополнительные закономерности встречаемости GC-пар в геномной структуре тех же эукариот, имеющих часто дефицитный характер своего GC-состава?

Цель работы

В данном исследовании, используя методы сравнительной геномики, на примере генома человека с его общим доминированием AT-пар в составе уникальных и повторяющихся нуклеотидных последовательностей, детально разбирается вопрос: каким образом и в каких областях (генных, межгенных, экзонных, интронных) его хромосом происходит значительное перераспределение встречаемости A/T- и G/C-треков и реализуется наблюдаемый геномный GC-состав?

Аннотированные структуры ядерного и митохондриального геномов человека ревизии 36 были нами взяты из базы GenBank [22]. GC-состав ядерного генома ~ 42 %. GC-состав митохондриальной ДНК (мтДНК) ~ 44 %. Все расчеты по обработке геномных данных были выполнены на основе разработанной собственной компьютерной программы, особенности алгоритма которой описаны в [21].

Результаты

Анализ частот встречаемости 4-х типов мононуклеотидных треков (A_n, T_n, G_n, C_n) и 2-х смешанных типов треков W(A/T)$_n$ и S(G/C)$_n$ был выполнен для всех 23 хромосом и мтДНК. Оценивались частоты f появления нуклеотидных последовательностей постепенно увеличивающейся длины n от 1 до n_{max}, где n_{max} – максимальная длина трека, обнаруживаемого в хромосоме. Из-за очень большого диапазона значений частот встречаемости, величины f использовались в логарифмической шкале, $\log f$. Отдельный частотный анализ всех указанных типов последовательностей был выполнен нами и для генных, межгенных, экзонных и интронных областей каждой из хромосом и мтДНК.

Было обнаружено:

1) Как и в целом геноме человека [21], в отдельных хромосомах частоты встречаемости f мононуклеотидных (A_n и T_n) и смешанных W(A/T)$_n$-треков как функция длины трека сохранили общий характер доминирования над встречаемостью G_n, C_n и смешанных S(G/C)$_n$-треков. Несколько неожиданным оказалось весьма близкое подобие хода полученных частотных зависимостей исследуемых треков у разных хромосом.

2) Определенное перераспределение вкладов треков обозначилось при анализе генных, межгенных и интронных хромосомных зон. В качестве иллюстрации на рисунке 1 представлены результаты, полученные для хромосомы № 1. Из-за отмеченного близкого подобия частотных графиков у остальных 22 хромосом мы для экономии места в статье эти данные здесь не приводим.

3) Наиболее важные и интересные результаты дал частотный анализ треков в кодирующих (экзонных) областях. Некоторые из типичных закономерностей представлены на рисунке 2.

Как видно из рисунка 2, необычное перераспределение частотных зависимостей в экзонах хромосом коснулось всех типов треков.

Так, в отличие от предыдущих данных (рис. 1), все четыре типа мононуклеотидных треков A_n, T_n, G_n, C_n с увеличение длины трека повели себя примерно одинаково в частотной зависимости и при размерах треков больше $n = 10$ пар они все симбатно исчезли из данной части хромосом. Доминирование смешанных $W(A/T)_n$-треков, свойственное для целых хромосом здесь тоже исчезло, как и вообще исчезла встречаемость таких треков длины больше $n = 20$ пар оснований.

В то же время смешанные последовательности из G/C-пар повели себя весьма неординарным образом:

– в хромосоме 1 и еще в нескольких хромосомах треки $S(G/C)_n$ оказались доминирующими треками во всем диапазоне длин $n = 1 \div 35$;

– в ряде хромосом, как, например, в хромосоме 4, подобное доминирование обнаружилось лишь у $S(G/C)_n$-треков, начиная с длины $n = 10$;

– у Y-хромосомы практически не оказалось никакого доминирования ни $W(A/T)_n$, ни $S(G/C)_n$ типа треков;

– у мтДНК сохранилось только некоторое подобие доминирования $W(A/T)_n$-треков и то до длины $n \approx 20$;

Для понимания возможных следствий обнаруженного поведения треков в экзонах хромосом уместно напомнить здесь о некоторых известных фактах.

Считается, что генным областям часто свойственен повышенный GC-состав по сравнению со всем геномом. При этом GC-пары концентрируются преимущественно в кодирующих участках хромосом и GC-состав здесь тем выше, чем больше длина экзона.

С другой стороны, не смотря на всю важность гуанин-цитозинового вклада для эффективности протекания процессов генной экспрессии и закрепления в организме последствий естественного отбора именно нуклеотидные последовательности и повторы из GC-пар, а особенно островки CpG (называемые иногда «горячими точками»), чаще всего оказываются подверженными в ДНК явлению точечных спонтанных мутаций. Как ни странно, GC-пары обладают повышенной (иногда в несколько раз) частотой точечных спонтанных мутаций типа транзиций и трансверсий G:C → A:T (G:C → T:A) по сравнению с AT-парами, с их заменами A:T → G:C или A:T → C:G [23, 24]. В ДНК человека, например, обнаружена даже линейная зависимость скорости таких спонтанных мутаций в генных областях с увеличением в них содержания GC-пар [25]. Есть примеры, где повышенная скорость подобных мутаций замечена и для протяженных треков из GC-пар по сравнению с AT-треками [26].

Во введении мы уже отмечали, что возможной физической первопричиной минимизации использования Природой GC-пар в организации структуры генома любого организма является их высокий, 4-х кратный полиморфизм геометрии исходного комплементарного спаривания оснований [21, 27, 28]. Поэтому, выявленная в данной работе специфика нуклеотидного состава экзонов существенно дополняет, на наш взгляд, картину распределения доминирования A/T- или G/C-треков в структуре эукариотических хромосом.

Полученные результаты дают, в частности, возможность весь наблюдаемый 23-мерный хромосомный состав генома человека вместе с его мтДНК подразделить довольно простым образом всего на три типа хромосом по характеру доминирования в их экзонах $(G/C)_n$-треков, рис. 3.

– I тип хромосом – хромосомы 1, 6, 7, 8, 9, 11, 15, 16, 17, 19, 20, 21, 22 с самым высоким GC-составом своих экзонов (51–58 %) и с полным доминированием в них $S(G/C)_n$-треков. Можно предположить, что общим свойством будет повышенный фон частоты точечных спонтанных мутаций кодирующих областей этих хромосом;

– II тип хромосом – хромосомы 2, 3, 4, 5, 10, 12, 13, 14, 18, X с доминированием лишь удлиненных $S(G/C)_n$ треков в диапазоне n от 10 до 35 пар оснований и имеющие пониженный GC-состав экзонных областей (47–51 %). Вполне вероятно, что фон частоты точечных спонтанных мутаций здесь будет уже умеренным;

– III тип хромосом – особый случай. Сюда отнесена Y-хромосома, где нет явного доминирования ни $S(G/C)_n$- ни $W(A/T)_n$ -треков Это уникальное, консервативное образование очень малых размеров, где отсутствует межхромосомный обмен ДНК.

Как отмечено в работе [29] у Y-хромосомы эволюционно минимизирована структура кодирующей области. Наблюдаемая необычно высокая активность этой хромосомы к точечным спонтанным мутациям реализуется за счет особенностей нуклеотидного состава её регуляторных областей.

К этому же типу можно отнести и мтДНК. GC-состав её экзонов ≈ 44 %. В определенном смысле это тоже достаточно автономное структурное образование весьма малых размеров. Здесь нет интронов. Количество белок кодирующих генов минимально. Потенциально активные к спонтанным точечным заменам $S(G/C)_n$-треки сильно минимизированы как по длине, так и по частоте встречаемости в структуре её экзонов (см. рис. 2). Наблюдаемые же в митохондриях повышенные частоты патогенных мутаций по сравнению с другими хромосомам можно объяснить [30] мутациями ядерного генома в кодирующих областях его белков, которые активно участвуют в функционировании митохондрий.

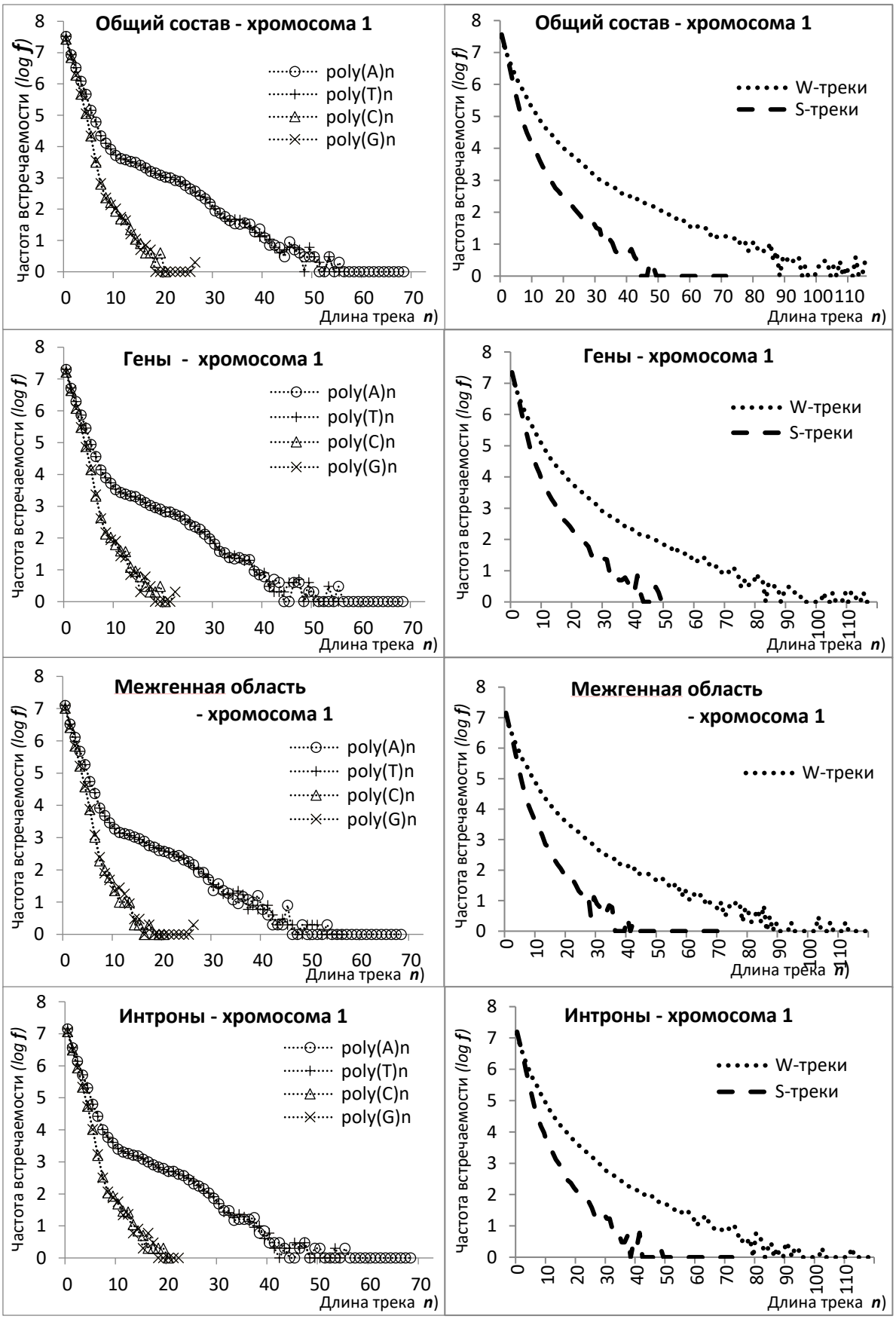

Рис. 1. Зависимость частот встречаемости нуклеотидных треков A_n, T_n, G_n, C_n, $W(A/T)_n$ и $S(G/C)_n$ типа от длины трека n в валовом составе хромосомы № 1 и в составе её отдельных областей – генных, межгенных и интронных.

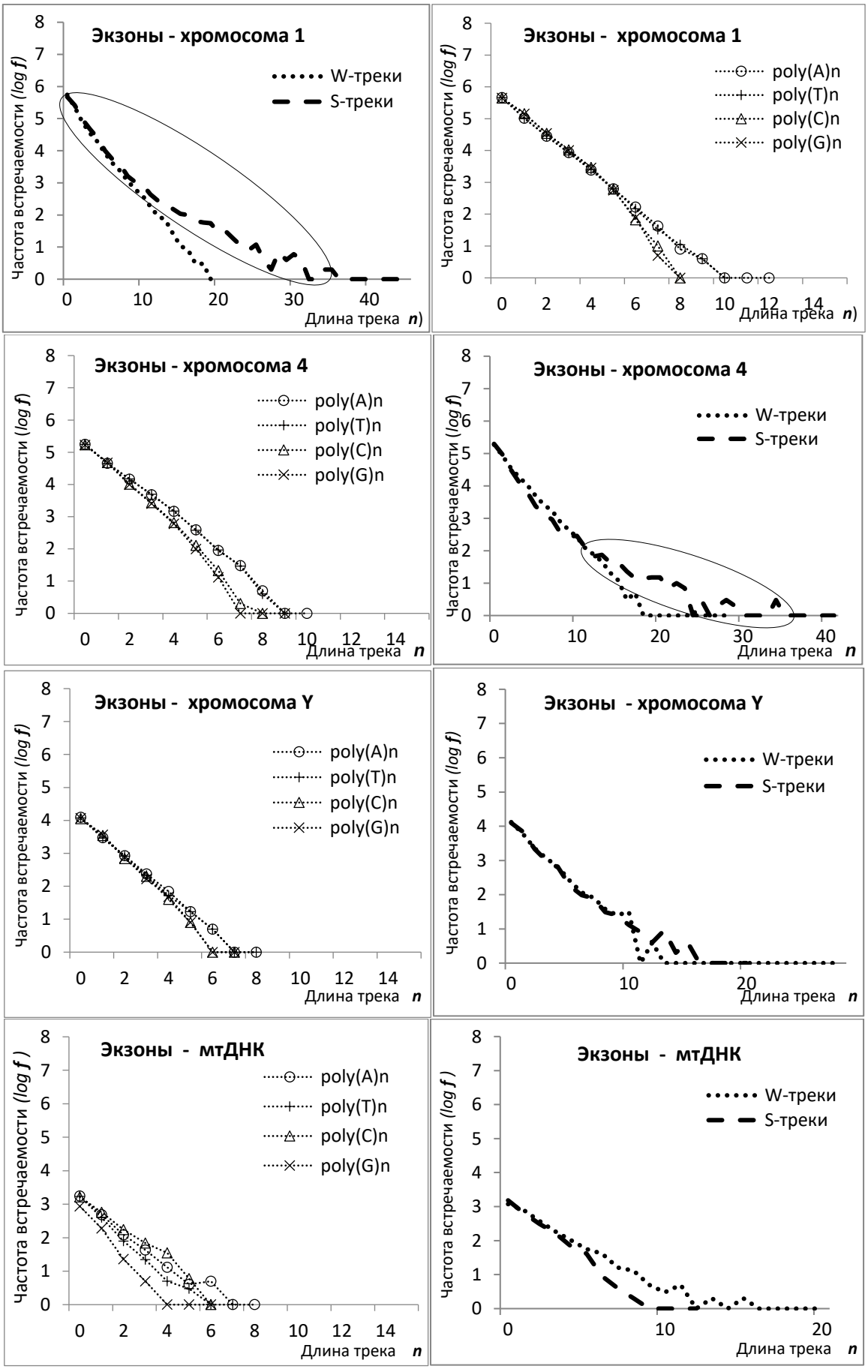

Рис. 2. Особенности поведения частотных зависимостей нуклеотидных треков в функции длины трека *n* в экзонах хромосом 1, 4, Y и мтДНК. Овалом обозначены области доминирования $S(G/C)_n$-треков.

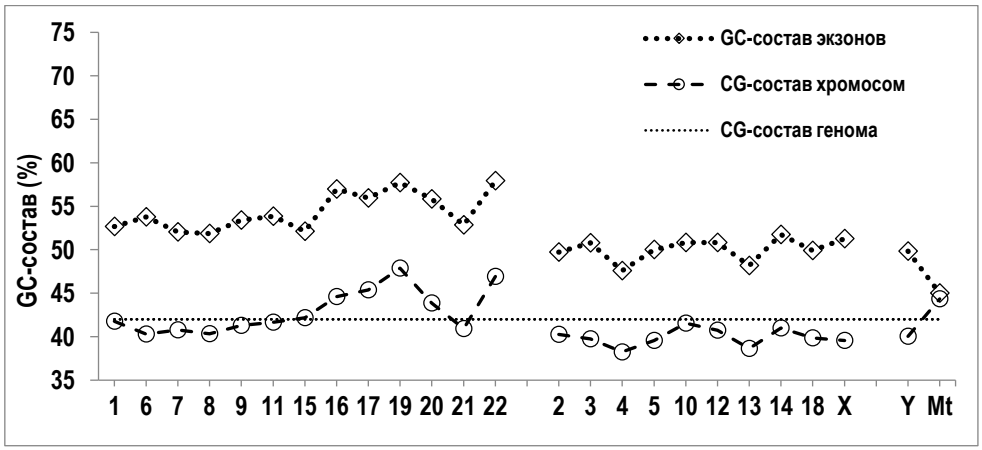

Рис. 3. Разбиение генома человека на три типа хромосом по GC-составу их экзонов. Приведены дополнительные данные по валовому GC-составу каждой из хромосом и всего генома.

Таким образом, представленный здесь материал позволяет утверждать о важном влиянии особенностей скрытого пространственного полиморфизма уотсон-криковских AT- и GC-пар на процесс распределения встречаемости мононуклеотидных и смешанных треков в ДНК генома человека. На фоне общего доминирования A/T-треков в структуре целых хромосом обнаружено лимитирующее влияние последовательностей из GC-пар на специфику нуклеотидной организации экзонных областей. По характеру доминирования в кодирующих участках относительно коротких до 35 пар нуклеотидов S-треков смешанной G/C-природы весь хромосомный набор генома человека может быть систематизирован на III различных типа хромосом, отличающихся GC-составом своих экзонов.

Список литературы

1. Zhou Y., Bizzaro J.W., Marx K.A. Homopolymer tract length dependent enrichments in functional regions of 27 eukaryotes and their novel dependence on the organism DNA (G+ C)% composition. *BMC Genomics.* 2004. V. 5. P. 95–111.
2. Marx K.A., Hess S.T., Blake R.D. Characteristics of the Large (dA).(dT) Homopolymer Tracts in D. discoideum Gene Flanking and Intron Sequences. *J. Biomol. Struct. & Dyn.* 1993. V. 11. P. 057–066.
3. Coenye T., Vandamme P. Characterization of Mononucleotide Repeats in Sequenced Prokaryotic Genomes. *DNA Research.* 2005. V. 12. P. 221–233.
4. Denver D.R., Morris K., Kewalramani A., Harris K.E., Chow A, Estes., S., Lynch M., Thomas W.K. Abundance, Distribution, and Mutation Rates of Homopolymeric Nucleotide Runs in the Genome of Caenorhabditis elegans *J. Mol. Evol.* 2004. V. 58. P. 584–595.
5. Singer C.E., Ames B.N. Sunlight ultraviolet and bacterial DNA base ratios. *Science.* 1970. V. 170. P. 822–825
6. Musto H., Naya H., Zavala A., Romero H., Alvarez-Valin F., Bernardi G. Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. *FEBS Lett.* 2004. V. 573. P. 73–77 .
7. Musto H., Naya H., Zavala A., Romero H., Alvarez-Valin F., Bernardi G. Genomic GC level, optimal growth temperature, and genome. size in prokaryotes. *Biochem. Biophys. Res. Commun*. 2006. V. 347. P. 1–3.
8. Oliver J.L., Marin A. A relationship between GC content and coding sequence length. *J. Mol. Evol*. 1996. V. 43. P. 216–223.
9. Foerstner K.U., von Mering C., Hooper S.D., Bork P. Environments shape the nucleotide composition of genomes *EMBO Rep*. 2005. V. 6. P. 1208–1213.
10. Sueoka N. Directional Mutation Pressure, Mutator Mutations, and Dynamics of Molecular Evolution. *J. Mol. Evol*. 1993. V. 37. P. 137–153.
11. Kudla G., Lipinski L., Caffin F., Helwak A., Zylicz M. High Guanine and Cytosine Content Increases mRNA Levels in Mammalian Cells. *PLoS Biol***.** 2006. V. 4. № 6. P. 933–942.
12. Zhou H.-Q., Ning L.-W., Zhang H.-X., Guo F.-B. Analysis of the Relationship between Genomic GC Content and Patterns of Base Usage, Codon Usage and Amino Acid Usage in Prokaryotes: Similar GC Content Adopts Similar Compositional Frequencies Regardless of the Phylogenetic Lineages. *PLoS ONE*. 2014. V. 9. № 9. doi: 10.1371/journal.pone.0107319.
13. Almpanis A., Swain M., Gatherer D., McEwan N. Correlation between bacterial G+C content, genome size andthe G+C content of associated plasmids and bacteriophages. *Microb. Genom*. 2018. V. 4. № 4. doi: 10.1099/mgen.0.000168.

14. Ussery D.W., Wassenaar T.M., Borini S.B. *Computing for Comparative Microbial Genomics: Bioinformatics for Microbiologists*. Springer-Verlag London, 2009.
15. Watson J.D. *Molecular biology of the gene*, W.A. Benjamin, Inc.: New York, Amstrdam, 1965
16. Karlin S., Mrazek J. Compositional differences within and between eukaryotic genomes *Proc. Natl. Acad. Sci. USA.* 1997. V. 94. P. 10227–10232.
17. Hamilton W.L., Claessens A., Otto T.D., Kekre M., Fairhurst R.M., Rayner J.C., Kwiatkowski D. Extreme mutation bias and high AT content in Plasmodium falciparum. *Nucleic Acids Research*. 2017. V. 45. № 4. P. 1889–1901.
18. Böhme U., et al. Complete avian malaria parasite genomes reveal features associated with lineage specific evolution in birds and mammals. *Genome Res.* 2018. V. 28. № 4. P. 547–556.
19. Киселев С.С., Комаров В.М., Масулис И.С., Озолинь О.Н. Распределение мононуклеотидных повторов в бактериальных хромосомах. A/T-треки преобладают над G/C-треками. *Компьют. исследов. моделир.* 2010. Т. 2. № 2. С. 183–187.
20. Киселев С.С. *Структурообразующие мотивы в геномах прокариот: Закономерности распределения и функциональное значение*: дис…. канд. биол. наук: институт биофизики клетки РАН, Пущино, 2012.
21. Самченко А.А., Киселев С.С., Кабанов А.В., Кондратьев М.С., Комаров В.М. О природе доминирования олигомерных (dA:dT)n треков в структуре геномов эукариот. *Биофизика*, 2016. Т. 6. С. 1045–1058.
22. URL: ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens (дата обращения: 25.09.2018).
23. Lynch M., Sung W., Morris K., Coffey N., Landry C.R., Dopman E.B., Dickinson W.J., Okamoto K., Kulkarni S., Hartl D.L., Thomas W.K. A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc. Natl. Acad. Sci. USA.* 2008. V. 105. № 27. P. 9272–9277.
24. Dillon M.M., Sung W., Lynch M., Cooper V.S. The Rate and Molecular Spectrum of Spontaneous Mutations in the GC-Rich Multichromosome Genome of Burkholderia cenocepacia. *Genetics*. 2015. V. 200. P. 935–946.
25. Schaibley V.M., et al. The influence of genomic context on mutation patterns in the human genome inferred from rare variants. *Genome Res.* 2013. V. 23. P. 1974–1984.
26. Bashir T., Sailer C., Gerber F., Loganathan N., Bhoopalan H., Eichenberger C., Grossniklaus U., Baskar R. Hybridization Alters Spontaneous Mutation Rates in a Parent-of-Origin-Dependent Fashion in Arabidopsis. *Plant Physiol*. 2014. V. 165. P. 424–437.
27. Комаров В.М. Квантово-химическое полуэмпирическое исследование полиморфизма уотсон-криковского спаривания азотистых оснований. *Биофизика.* 1998. Т. 43. № 6. С. 967–974.
28. Kabanov A.V., Komarov V.M. Polymorphism of Hydrogen Bonding in the Short Double Helixes of Oligonucleotides: Semiempirical Quantum-Chemical Study. *Intern. J. Quantum Chem.* 2002. V. 88. № 5. P. 579–587.
29. Самченко А.А., Киселев С.С., Кондратьев М.С., Комаров В.М. Об особенностях механизма точечных спонтанных мутаций в структуре Y-хромосомы человека. *Актуальные вопросы биологической физики и химии.* 2018. Т. 3. № 3. С. 568–573.
30. Мазунин И.О., Володько Н.В., Стариковская Е.Б., Сукерник Р.И. Митохондриальный геном и митохондриальные заболевания человека. *Молекулярная биология.* 2010. Т. 44. № 5. С. 755–772.