Genomics and Polymorphism analysis
original article

Вавиловский журнал генетики и селекции. 2015;19(6):668-674
Doi 10.18699/VJ15.092

Фланкирующие повторы мономеров определяют пониженную контекстную сложность сайтов однонуклеотидных полиморфизмов в геноме человека

Н.С. Сафронова[1, 2], М.П. Пономаренко[1, 2], И.И. Абнизова[3], Г.В. Орлова[1], И.В. Чадаева[1], Ю.Л. Орлов[1, 2]

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 Центр Сенгера, Кембридж, Великобритания

Исследование зависимости частоты возникновения мутаций в геноме человека выполнено на примере набора документированных однонуклеотидных полиморфизмов (ОНП) из проекта «1 000 геномов». Рассмотрены задачи разработки новых компьютерных методов статистического анализа генетических текстов на основе оценок сложности последовательности символов. Показано применение профилей сложности в скользящем окне к анализу сайтов, содержащих однонуклеотидные полиморфизмы в геноме человека. Установлено локальное понижение сложности текста в районе ОНП. На основе анализа профилей сложности в участках, содержащих ОНП, показано, что фланкирующие повторы мономеров определяют пониженную контекстную сложность сайтов однонуклеотидных полиморфизмов в геноме человека. Эффект локального понижения уровня сложности текста последовательностей фланкирующих сайты ОНП подтвержден для данных о полиморфизмах в геномах крысы и мыши. Определены различия в контекстной организации для кодирующих и регуляторных последовательностей, которые отражаются в сложности текста нуклеотидных последовательностей, содержащих ОНП. Изменения в частоте точковых мутаций были ранее показаны для последовательностей, содержащих микросателлиты. С использованием более общего математического аппарата и более полных данных в работе показана насыщенность политрактами и простыми повторяющимися последовательностями локального геномного окружения участков, содержащих ОНП. Определены олигонуклеотиды с повышенной частотой встречаемости в геномном окружении ОНП у человека, показана их связь с политрактами. Присутствие политрактов может свидетельствовать о большей вероятности разрыва двойной цепи ДНК в этой точке, приводящей к повышению частоты замен нуклеотидов. Полученные оценки были определены при помощи разработанного ранее комплекса компьютерных программ, который кроме оценки сложности фазированных выборок позволяет эффективно определять частотный спектр олигонуклеотидов фиксированной длины, производить сравнение частот олигонуклеотидов в выборках большого объема.

Ключевые слова: ОНП; геном; нуклеотидные последовательности; повторы; энтропия; мутации.

Received 17.09.2015
Accepted for publication 27.10.2015

e-mail: taschasafronova@bionet.nsc.ru; orlov@bionet.nsc.ru

Flanking monomer repeats define lower context complexity of sites containing single nucleotide polymorphisms in the human genome

N.S. Safronova[1, 2], M.P. Ponomarenko[1, 2], I.I. Abnizova[3], G.V. Orlova[1], I.V. Chadaeva[1], Y.L. Orlov[1, 2] . Абнизова[3], Г.В. Орлова[1], И.В. Чадаева[1], Ю.Л. Орлов[1, 2]

1 Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
2 Novosibirsk State University, Novosibirsk, Russia
3 Sanger Center, Cambridge, UK

we have investigated a mutation frequency within the human genome for the set of known single nucleotide polymorphisms (SnPs) from the "1000 genomes" project. we have developed and applied novel statistical computational methods to analyze genetic text based on its complexity. A complexity profiling in a sliding window is applied to the sites containing single nucleotide polymorphisms within the human genome. A local decrease in text complexity level in SnP-containing sites has been shown. Analysis of the complexity profiles for SnP-containing sites shows that flan ing monomer repeats define a l wer context complexity of sites containing SnPs within the human genome. An effect of local decrease in text complexity in SnP-containing sites is confi med by analysis of polymorphisms in the rat and mouse genomes. we have found context differences between coding and regulatory sequences. These differences refle t a complexity of SnP-containing loci. The changes in point mutation frequency were shown previously for microsatellite containing sequences. Using enhanced mathematical tools and larger data sets this work shows enrichment of polytracks and simple sequence repeats in local genome surroundings of SnP containing sites. we have found high-frequency oligonucleotides within genomic regions containing SnPs. Such oligonucleotides are related to nucleotide polytracks. The presence of poly-A tracks might be associated with an increased probability of double helix DnA breaks around mutable loci and following fix tion of nucleotide changes. The complexity estimates were computed using a previously developed program tool. This tool allows for both (i) complexity estimation of phased samples, and (ii) rapid and effective identific tion of the frequency spectrum of oligonucleotides with fi ed lengths, and a comparison of oligonucleotide frequencies in different samples.

Key words: SnP; genome; nucleotide sequences; repeats; entropy; mutations.

HOW TO CITE THIS ARTICLE?
Safronova n .S., Ponomarenko M.P., Abnizova i.i., o rlova G.V., c hadaeva i.V., o rlov Y.I . Flanking monomer repeats define lower context complexity of sites containing single nucleotide polymorphisms in the human genome. Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding. 2015;19(6):668-674. Doi 10.18699/VJ15.092

КАК ЦИТИРОВАТЬ ЭТУ СТАТЬЮ?
Сафронова Н.С., Пономаренко МП, Абнизова И.И., Орлова Г.В., Чадаева И.В., Орлов Ю.л. Фланкирующие повторы мономеров определяют пониженную контекстную сложность сайтов однонуклеотидных полиморфизмов в геноме человека. Вавиловский журнал генетики и селекции. 2015;19(6):668-674. Doi 10.18699/VJ15.092

Исследование нуклеотидных полиморфизмов в геноме имеет большое значение для фундаментальной и прикладной медицинской генетики. Изучение нуклеотидных замен изменений, а также предпосылок их возникновения по нуклеотидным последовательностям может позволить ответить на ряд важных молекулярно-биологических вопросов о ходе мутационного процесса, достигнуть успехов в предсказании и лечении генетических заболеваний, связанных с естественной изменчивостью генома. Развиваются международные проекты HapMap (International HapMap 3 Consortium, 2010), «1 000 Genomes» (http://www.1000genomes.org/), идут национальные и региональные исследования генетической изменчивости (Sidore et al., 2015; UK10K Consortium, 2015).

Исследование зависимости частоты возникновения мутаций в геноме человека было выполнено нами на примере набора документированных однонуклеотидных полиморфизмов из проекта «1 000 геномов». Однонуклеотидные полиморфизмы (ОНП, или SNP – Single Nucleotide Polymorphism) – однонуклеотидные различия последовательностей ДНК в геноме или участке генома. Такие полиморфизмы имеют большое значение при изучении различных заболеваний, что требует развития биоинформационных ресурсов анализа (International HapMap 3 Consortium, 2010).

За последние два десятилетия создан большой набор программных продуктов, направленных на изучение свойств и структуры последовательностей ДНК и белков (Babenko et al., 1999; Орлов, 2012; Игнатьева и др., 2015). Одной из важных проблем исследования геномной ДНК является анализ сложности генетических текстов с помощью математических оценок, учитывающих эволюционные ограничения на изменение последовательности (Орлов, 2004; Orlov, Potapov, 2004; Orlov et al., 2006). Исследование сложности текста нуклеотидных и аминокислотных последовательностей как независимой универсальной характеристики имеет свой широкий круг применений: от анализа строения регуляторных районов генов до анализа расположения повторов в полных геномах (Chuzhanova et al., 2002; Orlov et al., 2006; Trifonov et al., 2012).

Связь так называемых «горячих точек» мутаций в геномах с окружающим нуклеотидным контекстом была показана для разных организмов (Rogozin et al., 1991, 2001; Rogozin, Kolchanov, 1992). Изменения в частоте точковых мутаций были ранее показаны для последовательностей, содержащих микросателлиты (Siddle et al., 2011). Ряд исследований, рассматривающих ди- и тринуклеотидные повторы в геноме, показал увеличение скорости мутаций в районе повтора (Vowles, Amos, 2004; Siddle et al., 2011). Опираясь на более общий математический аппарат и более полные данные (Safronova et al., 2015), мы показываем насыщенность политрактами и простыми повторяющимися последовательностями локального геномного окружения участков, содержащих ОНП.

Материалы и методы

В работе использовали базу данных dbSNP и базу документированных однонуклеотидных полиморфизмов из проекта «1 000 геномов». Нуклеотидные последовательности загружали из ресурса UCSC Genome Browser (genome.ucsc.edu). Для численной оценки сложности применяли алгоритмы разработанной ранее компьютерной программы (Orlov, Potapov, 2004): расчет числа операций, необходимых для сжатия текста алгоритмом Лемпеля и Зива (Gusev et al., 1999), расчет комбинаторной (лингвистической) сложности текста (Troyanskaya et al., 2002). Эти алгоритмы были дополнены оценками: 1) энтропии символов (Wootton, Federhen, 1996); 2) присутствия политрактов; 3) частоты чередования символов в последовательности.

Каждое определение сложности текста учитывает определенные аспекты его структурной организации. Энтропия представляет неравномерность олигонуклеотидного состава. Алгоритмы оценки сложности используют представление последовательности символов в виде *l*-граммного дерева (Orlov et al., 2002), которое позволяет рассчитывать число всех возможных 4^l слов длины l в последовательности, определять позиции этих слов и с минимальными затратами компьютерного времени выполнять операции по расчету сложности текста и поиску гомологии.

Операционная сложность, или сложность по методу Лемпеля и Зива – это число операций копирования (дупликаций коротких последовательностей), необходимых для порождения текста из некоторого базового текста

(самой последовательности) в так называемом сложностном разложении (Gusev et al., 1999):

$$C_{LZ}(S) = S[1; i_1]\,S[i_1+1; i_2]\,\ldots\,S[i_{m-1}+1; L], \quad (1)$$

где $S[i_{k-1}+1; i_k]$ – фрагмент исходной строки, генерируемый на k-м шаге из уже сгенерированной последовательности, слева направо, L – длина текста (в нуклеотидах), m – минимальное число фрагментов (число операций копирования), определяющее сложность текста.

Анализ такой сложности учитывает все типы повторов, включая инверсии и симметричные повторы, и дает полную картину всех, а не только локальных, взаимозависимостей в тексте (Orlov, Potapov, 2004; Orlov et al., 2006). Чем короче повторы, тем больше операций копирования и больше сложность текста. Операционная сложность хорошо подходит для поиска повторов в протяженных последовательностях. Она наиболее близка к общему определению сложности, по Колмогорову, как длины кратчайшей программы, необходимой для генерации символьной последовательности.

Комбинаторная, или лингвистическая, сложность определяется как отношение числа встретившихся слов к числу всех возможных слов в последовательности заданной длины (Troyanskaya et al., 2002):

$$C_{ling} = \frac{\sum_1^L N_i}{\sum_1^L N_{max_i}}, \quad (2)$$

где N_i – число встретившихся слов длины i, N_{max_i} – максимальное число возможных слов длины i в последовательности длины L, которое равно минимуму из двух чисел: числа всех возможных слов длины i для используемого алфавита, либо числа слов длины i, которые можно разместить в последовательности длины L.

Чем меньше повторов всех длин, чем больше разнообразие найденных слов, тем больше сложность последовательности. Комбинаторная сложность учитывает существенные особенности текста, позволяет быстро находить простые участки (например короткие повторы) в последовательности.

Сложность частот чередования символов оценивает относительное число смен одного символа на другой в линейном расположении. Чем больше это число, тем более перемешаны символы и тем меньше политрактов – слов, состоящих из одного символа. Таким образом, эта мера позволяет эффективно находить локальную насыщенность политрактов. В то же время не учитывается разнообразие словаря возможных слов. Эта сложность рассчитывается по формуле:

$$C_{var} = Ch/(L-1), \quad (3)$$

где Ch – количество смен текущего символа, L – количество символов в строке (длина последовательности).

Рассмотрим величину C_{var} на примере коротких олигонуклеотидов длины 6: олигонуклеотид AAAAAA в интуитивном понимании проще, чем AAATAA, который в свою очередь проще ATGCGA; таким образом, в рамках введенной меры AAAAAA имеет сложность 0, AAATAA – 0,4 (2/5), ATGCGA – 1 (5/5).

Стандартный метод – оценка энтропии символов (энтропии Шеннона) в тех же обозначениях:

$$CE = -\sum_{i=1}^{K}(n_i/L)\log_K(n_i/L), \quad (4)$$

здесь L – размер последовательности, n_i – число встретившихся символов (нуклеотидов) i-го типа, i = 1, 2, ..., K, K – размер алфавита (K = 4 для ДНК, K = 20 для аминокислот). Поскольку логарифм берется по основанию K, все значения сложности содержатся в интервале (0; 1).

Для анализа коротких повторов в тексте (от 3 до 10 нт) рассмотрены меры сложности олигонуклеотидов. Таким образом, рассматривается сложность элементов (слов, олигонуклеотидов), которые могут быть выделены в исследуемых участках генома. Например, для олигонуклеотида AAATAAA, состоящего из повтора, за исключением одного нуклеотида, T, энтропия составляет 0,296, для ATATATA – 0,493, для TTATGCGC, содержащего уже все четыре нуклеотида, энтропия равна 0,979. Более подробно см. в Дополнительных материалах[1].

Ранее была рассмотрена связь различных мер сложности на геномных последовательностях, было отмечено, что все меры коррелируют между собой, при этом мера Лемпеля–Зива наиболее эффективно выделяет длинные повторы, а мера чередования символов более чувствительна к наличию политрактов (Орлов, 2004; Orlov et al., 2006).

результаты

Нуклеотидные последовательности из генома человека, содержащие аннотированные ОНП, были загружены из UCSC Table Browser (http // genome.ucsc.edu.). Мы использовали последовательности размером 101 нт (1 позиция НП и +/–50 нт окружение). Был выполнен расчет профилей сложности в скользящем окне размеров 10, 7, 5 нт. Рассматривалось отдельно четыре класса ОНП – замена A на другую букву (нуклеотид), C – на другой нуклеотид и т. д. Построенные профили сложности были усреднены.

Замечено, что в районе мутации в скользящем окне профиля происходит резкое изменение сложности (Safronova et al., 2015) (рис. 1).

Как показал расчет стандартного отклонения значений сложности, при большем числе проанализированных последовательностей такое изменение профиля статистически значимо ($p < 0{,}01$). Дальнейший анализ показал, что сложность последовательности в скользящем окне возрастает к точке полиморфизма, а после нее симметрично убывает, в районе ОНП происходит скачок сложности. Было замечено, что у всех наблюдаемых организмов в районе мутации профиль сложности имеет очень схожее поведение: сначала идет убывание, затем следует скачок и снова убывание (яма–пик–яма).

Отмечено локальное симметричное понижение сложности в районе вокруг точки полиморфизма. Дальнейший частотный анализ показал высокую насыщенность области около полиморфизма (+/–10 нт) поли-A и поли-T трактами. Анализ дисперсии значений сложности в точке ОНП и на флангах (5 нт) показал статистическую значимость различия средних значений (рис. 1).

Для модельных организмов – мыши и крысы – были получены схожие результаты, подтвержден эффект, обнаруженный на более полных данных по человеку (рис. 2).

[1] Дополнительные материалы см. в Приложении 1 по адресу: http://www.bionet.nsc.ru/vogis/download/pict-2015-12/appx1.pdf

На рис. 2 показано симметричное изменение профиля сложности с пиком на точке полиморфизма для участков, содержащих ОНП в геноме крысы.

Мера вариабельности C_{var} (мера чередования символов) в скользящем окне также для участков ОНП в геноме человека представлена на рис. 3.

Профиль меры чередования символов имеет тот же вид, что и сложность по Лемпелю – Зиву (число повторяющихся фрагментов), представленная на рис. 1, с локальными минимумами на флангах точки ОНП до 10 нт.

Локальное понижение сложности текста отмечено и с помощью других мер – лингвистической сложности, энтропии и меры чередования нуклеотидов. Минимальные значения сложности соответствуют последовательностям, состоящим почти целиком из одной повторяющейся единицы (тандемного повтора), самое минимальное значение дает политракт (однобуквенный повтор), например $(A)_n$. На рис. 4 приведены профили встречаемости наиболее представленных олигонуклеотидов, в данном случае политрактов, в последовательностях, фазированных относительно точки ОНП.

Присутствие сигнала TAA в горячих точках мутаций было показано еще в ранней работе И.Б. Рогозина и Н.А. Колчанова на ограниченной выборке данных (Rogozin, Kolchanov, 1992). В нашей работе представлен расширенный набор олигонуклеотидов, содержащих короткие повторы нуклеотидов A и T, непосредственно фланкирующих точки ОНП, которые также связаны с повышенным уровнем мутаций.

С использованием набора методов оценок сложности текста были рассчитаны значения сложности для нуклеотидных последовательностей различных типов – белок-кодирующих (экзоны), некодирующих (интроны) и регуляторных (промоторы и энхансеры). Ранее было показано различие мер сложности для кодирующих и некодирующих последовательностей (Orlov et al., 2006). Белок-кодирующие последовательности несут большую нагруженность сигналами различных типов (триплетный код аминокислот, информация о вторичной и пространственной структуре

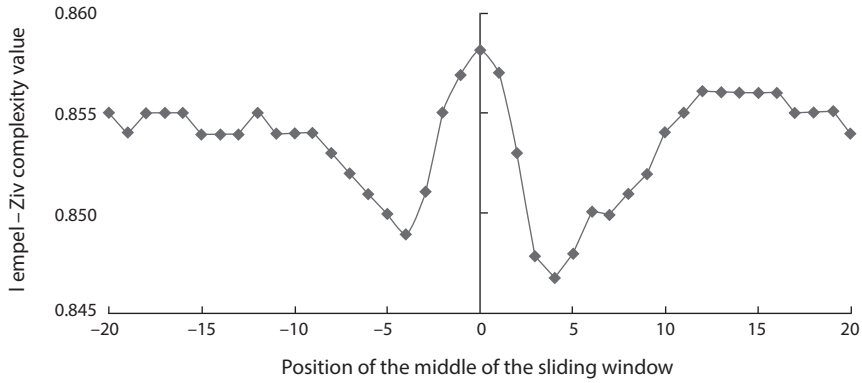

Fig. 1. Profile of l empel – Ziv complexity values in the sliding window for Sn P sites in the human genome.

w indow size 7 nt. r eplacements of nucleotide A in Sn P for three other nucleotides.

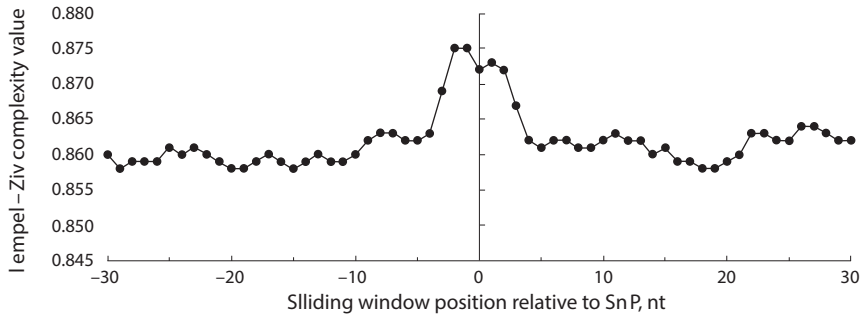

Fig. 2. Profile of l empel – Ziv complexity value changes in the 7-nt sliding window for Sn P sites in the rat genome.

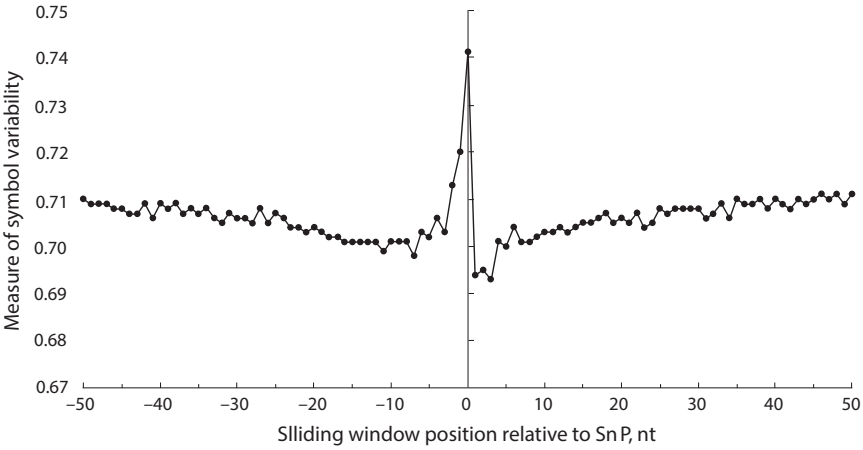

Fig. 3. Profile of the f equency of symbol changes in sequence (monomer variability) for Sn P sites in the human genome.

белка), и, следовательно, большие сложность текста и энтропию символов (Trifonov et al., 2012). Некодирующие последовательности (интроны), напротив, свободны от сигналов структуры белка, соответственно, имеют меньшую сложность. В то же время регуляторные последовательности содержат ряд размытых контекстных сигналов – о сайтах связывания белков, сайте посадки транскрипционного комплекса, но не о кодировании белка. Регуляторные последовательности занимают промежуточное положение между интронами

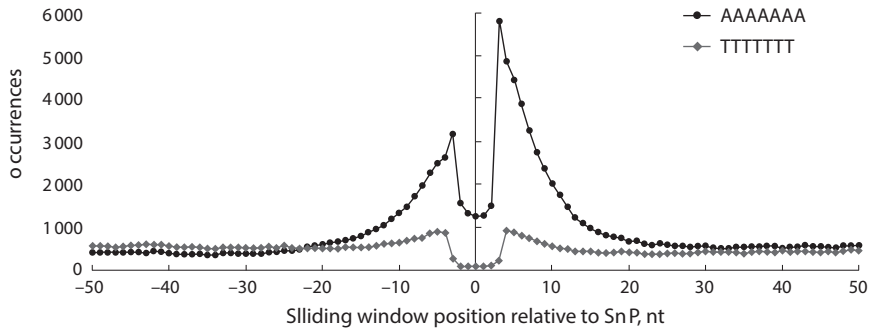

Fig. 4. Profile of 7-bp long pol -A tracts occurrence in Sn P-containing sites in the human genome.

и экзонами. Для более коротких последовательностей результаты различий уже не видны. Поэтому мы использовали построение профилей в скользящем окне для позиций однонуклеотидных полиморфизмов.

Для сайтов, содержащих ОНП, минимальные значения сложности соответствуют последовательностям, состоящим почти целиком из одной повторяющейся единицы (тандемного повтора), самое минимальное значение дает политракт. Таким образом, для участков низкой сложности, фланкирующих ОНП, характерно наличие простых участков (содержащих короткие тандемные повторы, политракты), что подтверждает ранние результаты (Rogozin et al., 1991; Ponomarenko et al., 2002; Medvedeva et al., 2013).

Обсуждение

В связи с возможными молекулярными механизмами патогенеза ОНП в кластерах повторяющейся ДНК представляется интересным, прежде всего, что в обзоре Поляновского с коллегами (2012) к числу самых важных причин устойчивости к моноклональным антителам-ингибиторам онкогенов-мишеней были отнесены наследственные ОНП и соматические мутации в кодирующих районах как этих онкогенов, так и генов для их рецепторов и эффекторов сигнальных путей. Поскольку ранее была открыта статистически значимая корреляция между эффективностью соматического мутагенеза в его горячих точках с количеством содержащих их несовершенных повторов (Rogozin et al., 1991), то степень риска канцерогенеза может быть оценена по числу повторов в анцестральном аллеле гена, в границах которых локализованы ассоциированные с раком ОНП.

Кроме того, большинство инвертированных повторов и комплементарных палиндромов локализованы в белок-кодирующих районах генов, что, как было показано, связано с неравномерностью использования кодонов и оптимизацией вторичной структуры кодируемых ими мРНК по ее устойчивости к ее деградации (Karlin et al., 1989). Это означает, что в числе молекулярных механизмов патогенеза ОНП в границах инвертированных повторов и комплементарных палиндромов в белок-кодирующих районах генов может быть рассогласование координированно экспрессирующихся генов в генной сети. В свою очередь, конвергентное возникновение прямых повторов в белок-кодирующих районах генов связано с кодированием элементов вторичной структуры (α-спиралей и β-нитей) белковых глобул. Это означает, что к числу молекулярных механизмов патогенеза, вызванного ОНП в границах прямых повторов в белок-кодирующих районах генов, могут быть отнесены нарушения пространственных структур белковых глобул.

Наконец, в работе Babenko с коллегами (1999) был обнаружен специфический повторяющийся контекст локального окружения сайта в кор-промоторах генов человека для ТАТА-связывающего белка (TBP), образующего анкерный комплекс для РНК полимеразы II (Ponomarenko et al., 2013a, b). Поэтому в числе молекулярных механизмов патогенеза ОНП в повторяющейся ДНК промоторов генов человека могут быть изменения сродства TBP к этим промоторам и в итоге – изменение экспрессии этих генов (Пономаренко и др., 2008; Савинкова и др., 2009). Таким образом, контекстные оценки сложности текста для геномных участков могут использоваться как характеристики для оценки возникновения полиморфизмов и возможных нарушений регуляции экспрессии генов.

На языке C++ был разработан комплекс программ, который позволяет эффективно определять частотный спектр олигонуклеотидов заданной фиксированной длины, выполнять сравнение частот олигонуклеотидов в различных выборках. С помощью этого программного обеспечения рассчитана контекстная сложность геномных районов, содержащих ОНП человека по базе данных dbSNP. Установленная насыщенность политрактами области точки полиморфизма свидетельствует о более высокой вероятности разрыва двойной спирали ДНК в этой точке, что подтверждает ранее показанные результаты (Siddle et al., 2011; Lenz et al., 2014). Фланги прямых мононуклеотидных повторов подвержены мутациям. Изменения идут в сторону увеличения энтропии, т. е. при возникновении полиморфизмов природа избегает длинных повторов мононуклеотидов и сложность геномного текста в таком участке вновь возрастает. Таким образом, с информационной точки зрения, происходит ослабление контекстных мононуклеотидных сигналов в геноме (Safronova et al., 2015). Пониженная сложность фланкирующих ОНП точки последовательностей наблюдалась также по доступным данным у мыши и крысы, и, следовательно, может иметь общий характер для геномов эукариот (Medvedeva et al., 2013; Lenz et al., 2014).

Интересно отметить, что повышенная частота полиморфизмов в последовательностях, фланкируемых мономерами, может быть связана с открытостью нуклеосомной упаковки в таких участках. Действительно, политракты статистически чаще встречаются в линкерных участках между соседними нуклеосомами, чем в последовательностях ДНК внутри нуклеосомной упаковки (Орлов и

др., 2006; Goh et al., 2010; Trifonov et al., 2012). Анализ контекстной сложности и насыщенности повторами в геномной ДНК содержащих ОНП участков с разной функциональной нагрузкой (регуляторные последовательности, белок-кодирующие последовательности) требует более детального исследования. Данный анализ может быть расширен на протяженные геномные последовательности (Babenko et al., 2015). В дальнейшем планируется интегрировать разработанную программу в комплекс для расчета сложности текста в виде дополнительного программного модуля, провести работу по улучшению и оптимизации пользовательского интерфейса программы и встроить ее в модули анализа геномных последовательностей, заданных генов и участков хромосомных контактов (Орлов и др., 2012; Кулакова и др., 2015; Спицина и др., 2015).

Acknowledgments

Computations were done at the Bioinformatics Shared Access Center of the Institute of the Cytology and Genetics; Siberian Supercomputer Center of SB RAS. This work was supported by the Russian Foundation for Basic Research, project 14-04-01906 (Development of programs for genomic analysis); joint project 15-54-53091 of the Russian Foundation for Basic Research and the National Natural Science Foundation of China (Analysis of polymorphisms in humans); and the Institute of Cytology and Genetics, Budgeted Project VI.61.1.2.

Conflict of interest

The authors declare no conflict of interest.

References

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. DOI 10.1093/bioinformatics/15.7.644

Babenko V.N., Matvienko V.F., Safronova N.S. Implication of transposons distribution on chromatin state and genome architecture in human. J. Biomol. Struct. Dyn. 2015;33(1):10-11. DOI 10.1080/07391102.2015.1032559

Chuzhanova N.A., Krawczak M., Thomas N., Nemytikova L.A., Gusev V.D., Cooper D.N. The evolution of the vertebrate beta-globin gene promoter. Evolution. 2002;56(2):224-232.

Goh W.S., Orlov Y., Li J., Clarke N.D. Blurring of high-resolution data shows that the effect of intrinsic nucleosome occupancy on transcription factor binding is mostly regional, not local. PLoS Comput. Biol. 2010;6(1):e1000649. DOI 10.1371/journal.pcbi.1000649

Gusev V.D., Nemytikova L.A., Chuzhanova N.A. On the complexity measures of genetic sequences. Bioinformatics. 1999;15(12):994-999. DOI 10.1093/bioinformatics/15.12.994

Ignatieva E.V., Podkolodnaya O.A., Orlov Y.L., Vasiliev G.V., Kolchanov N.A. Regulatory genomics: Combined experimental and computational approaches. Genetika = Genetics. 2015;51(4):409-429.

International HapMap 3 Consortium, Altshuler D.M., Gibbs R.A., Peltonen L., Dermitzakis E., Schaffner S.F., Yu F., Peltonen L., Dermitzakis E., Bonnen P.E., Altshuler D.M., Gibbs R.A., de Bakker P.I., Deloukas P., Gabriel S.B., Gwilliam R., Hunt S., Inouye M., Jia X., Palotie A., Parkin M., Whittaker P., Yu F., Chang K., Hawes A., Lewis L.R., Ren Y., Wheeler D., Gibbs R.A., Muzny D.M., Barnes C., Darvishi K., Hurles M., Korn J.M., Kristiansson K., Lee C., McCarrol S.A., Nemesh J., Dermitzakis E., Keinan A., Montgomery S.B., Pollack S., Price A.L., Soranzo N., Bonnen P.E., Gibbs R.A., Gonzaga-Jauregui C., Keinan A., Price A.L., Yu F., Anttila V., Brodeur W., Daly M.J., Leslie S., McVean G., Moutsianas L., Nguyen H., Schaffner S.F., Zhang Q., Ghori M.J., McGinnis R., McLaren W., Pollack S., Price A.L., Schaffner S.F., Takeuchi F., Grossman S.R., Shlyakhter I., Hostetter E.B., Sabeti P.C., Adebamowo C.A., Foster M.W., Gordon D.R., Licinio J., Manca M.C., Marshall P.A., Matsuda I., Ngare D., Wang V.O., Reddy D., Rotimi C.N., Royal C.D., Sharp R.R., Zeng C., Brooks L.D., McEwen J.E. Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467(7311):52-58. DOI 10.1038/nature09298

Karlin S., Ost F., Blaisdell B.T. Patterns in DNA and amino-acid sequences and their statistical significance. Mathematical methods for DNA sequences. Ed. M.S. Waterman. Boca Raton: CRC Press, 1989.

Kulakova E.V., Spitsina A.M., Orlova N.G., Dergilev A.I., Svichkarev A.V., Safronova N.S., Chernykh I.G., Orlov Y.L. Program analysis of genomic sequence data, obtained through technologies ChIP-seq, ChIA-PET and Hi-C. Programmnye sistemy: teoriya i prilozheniya = Program Systems: Theory and Applications. 2015;6(2):129-148.

Lenz C., Haerty W., Golding G.B. Increased substitution rates surrounding low-complexity regions within primate proteins. Genome Biol. Evol. 2014;6(3):655-665. DOI 10.1093/gbe/evu042

Medvedeva S.A., Panchin A.Y., Alexeevski A.V., Spirin S.A., Panchin Y.V. Comparative Analysis of Context-Dependent Mutagenesis Using Human and Mouse Models. BioMed Res. Intern. 2013;2013. Article ID 989410

Orlov Y.L. Analiz regulyatornykh genomnykh posledovatelnostey s pomoshchyu kompyuternykh metodov otsenok slozhnosti geneticheskikh tekstov. Diss. kand. biol. nauk. [Analysis of regulatory genome sequences using computer methods of genetic text complexity. Cand. biol. sci. diss.]. Novosibirsk, 2004.

Orlov Y.L., Bragin A.O., Medvedeva I.V., Podkolodnaia O.A., Khlebodarova T.M., Kolchanov N.A. ICGenomics: Software for analysis of symbol genomics sequences. Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding. 2012;16(4/1):732-741.

Orlov Y.L., Filippov V.P., Potapov V.N., Kolchanov N.A. Construction of stochastic context trees for genetic texts. In Silico Biology. 2002;2(3):257-262.

Orlov Y.L., Levitskii V.G., Smirnova O.G., Gunbin K.V., Demenkov P.S., Vishnevsky O.V., Levitsky V.G., Oshchepkov D.Y., Podkolodnyi N.L., Afonnikov D.A., Grosse I., Kolchanov N.A. Statistical analysis of nucleosome formation sites. Biofizika = Biophisics (Moscow). 2006;51(4):608-614.

Orlov Y.L., Potapov V.N. Complexity: an internet resource for analysis of DNA sequence complexity. Nucl. Acids. Res. 2004;32(Web Server issue):W628-633. DOI 10.1093/nar/gkh466

Orlov Y.L., Te Boekhorst R., Abnizova I.I. Statistical measures of the structure of genomic sequences: entropy, complexity, and position information. J. Bioinform. Comput. Biol. 2006;4:523-536. DOI 10.1142/S0219720006001801

Polanovski O.L., Lebedenko E.N., Deyev S.M. ERBB oncogenes as targets for monoclonal antibodies. Biokhimia = Biochemistry (Moscow). 2012;77(3):289-311.

Ponomarenko J.V., Orlova G.V., Merkulova T.I., Gorshkova E.V., Fokin O.N., Vasiliev G.V., Frolov A.S., Ponomarenko M.P. rSNP_Guide: an integrated database-tools system for studying SNPs and site-directed mutations in transcription factor binding sites. Hum. Mutat. 2002;20(4):239-248. DOI 10.1002/humu.10116

Ponomarenko M., Mironova V., Gunbin K., Savinkova L. Hogness Box. Brenner's Encyclopedia of Genetics. 2nd edn. Eds S. Maloy, K. Hughe. San Diego: Acad. Press, Elsevier Inc. 2013a;3:491-494. DOI 10.1016/B978-0-12-374984-0.00720-8

Ponomarenko M., Savinkova L., Kolchanov N. Initiation Factors. Brenner's Encyclopedia of Genetics, 2nd ed. Eds S. Maloy, K. Hughes. San Diego: Acad. Press, Elsevier Inc. 2013b;4:83-85. DOI 10.1016/B978-0-12-374984-0.00798-1

Ponomarenko P.M., Savinkova L.K., Drachkova I.A., Lysova M.V., Arshinova T.V., Ponomarenko M.P., Kolchanov N.A. A step-by-step model of TBP/TATA box binding allows predicting human hereditary diseases by single nucleotide polymorphism. Doklady RAN = Proceedings of the Russian Academy of Sciences. 2008;419(6):828-832.

Putta P., Orlov Y.L., Podkolodnyy N.L., Mitra C.K. Relatively conserved common short sequences in transcription factor binding sites and miRNA. Vavilov Journal of Genetics and Breeding. 2011;15(4): 750-756.

Rogozin I.B., Kolchanov N.A. Somatic hypermutagenesis in immunoglobulin genes. II. Influence of neighbouring base sequences on mutagenesis. Biochim. Biophys. Acta. 1992;1171(1):11-18. DOI 10.1016/0167-4781(92)90134-L

Rogozin I.B., Pavlov Y.I., Bebenek K., Matsuda T., Kunkel T.A. Somatic mutation hotspots correlate with DNA polymerase eta error spectrum. Nat. Immunol. 2001;2(6):530-536. DOI 10.1038/88732

Rogozin I.B., Solovyov V.V., Kolchanov N.A. Somatic hypermutagenesis in immunoglobulin genes. I. Correlation between somatic mutations and repeats. Somatic mutation properties and clonal selection. Biochim. Biophys. Acta. 1991;1089(2):175-182. DOI10.1016/0167-4781(91)90005-7

Safronova N.S., Babenko V.N., Orlov Y.L. 117 Analysis of SNP containing sites in human genome using text complexity estimates. J. Biomol. Struct. Dyn. 2015;33(1):73-74. DOI 10.1080/07391102. 2015.1032750

Savinkova L.K., Ponomarenko M.P., Ponomarenko P.M., Drachkova I.A., Lysova M.V., Arshinova T.V., Kolchanov N.A. TATA box polymorphisms in human gene promoters and associated hereditary pathologies. Biokhimiya=Biochemistry (Moscow). 2009;74(2): 149-163.

Siddle K.J., Goodship J.A., Keavney B., Santibanez-Koref M.F. Bases adjacent to mononucleotide repeats show an increased single nucleotide polymorphism frequency in the human genome. Bioinformatics. 2011;27(7):895-898. DOI 10.1093/bioinformatics/btr067

Sidore C., Busonero F., Maschio A., Porcu E., Naitza S., Zoledziewska M., Mulas A., Pistis G., Steri M., Danjou F., Kwong A., Ortega Del Vecchyo V.D., Chiang C.W., Bragg-Gresham J., Pitzalis M., Nagaraja R., Tarrier B., Brennan C., Uzzau S., Fuchsberger C., Atzeni R., Reinier F., Berutti R., Huang J., Timpson N.J., Toniolo D., Gasparini P., Malerba G., Dedoussis G., Zeggini E., Soranzo N., Jones C., Lyons R., Angius A., Kang H.M., Novembre J., Sanna S., Schlessinger D., Cucca F., Abecasis G.R. Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. Nat. Genet. 2015; 47(11):1272-1281. DOI 10.1038/ng.3368

Spitsina A.M., Orlov Y.L., Podkolodnaya N.N., Svichkarev A.V., Dergilev A.I., Chen M., Kuchin N.V., Chernykh I.G., Glinskij B.M. Supercomputer analysis of genomics and transcriptomics data revealed by high-throughput DNA sequencing. Programmnye sistemy: teoriya i prilozheniya=Program Systems: Theory and Applications. 2015;6:1(23):157-174.

Trifonov E.N., Volkovich Z., Frenkel Z.M. Multiple levels of meaning in DNA sequences, and one more. Ann. N.Y. Acad Sci. 2012;1267: 35-38. DOI 10.1111/j.1749-6632.2012.06589.x

Troyanskaya O.G., Arbell O., Koren Y., Landau G.M., Bolshoy A. Sequence complexity profiles of prokaryotic genomic sequences: a fast algorithm for calculating linguistic complexity. Bioinformatics. 2002;18(5):679-688. DOI 10.1093/bioinformatics/18.5.679

UK10K Consortium; Walter K., Min J.L., Huang J., Crooks L., Memari Y., McCarthy S., Perry J.R., Xu C., Futema M., Lawson D., Iotchkova V., Schiffels S., Hendricks A.E., Danecek P., Li R., Floyd J., Wain L.V., Barroso I., Humphries S.E., Hurles M.E., Zeggini E., Barrett J.C., Plagnol V., Richards J.B., Greenwood C.M., Timpson N.J., Durbin R., Soranzo N. The UK10K project identifies rare variants in health and disease. Nature. 2015;526:82-90. DOI 10.1038/nature14962

Vowles E.J., Amos W. Evidence for widespread convergent evolution around human microsatellites. PLoS Biol. 2004;2:E199. DOI 10.1371/journal.pbio.0020199

Wootton J.C., Federhen S. Analysis of compositionally biased regions in sequence databases. Methods Enzymol. 1996;266:554-571. DOI 10.1016/S0076-6879(96)66035-2