

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

Н.С. Сафронова<sup>1,2</sup>, М.П. Пономаренко<sup>1,2</sup>, И.И. Абнизова<sup>3</sup>, Г.В. Орлова<sup>1</sup>, И.В. Чадаева<sup>1</sup>, Ю.Л. Орлов<sup>1,2</sup>

<sup>1</sup> Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук», Новосибирск, Россия

<sup>2</sup> Федеральное государственное автономное образовательное учреждение высшего образования «Новосибирский национальный исследовательский государственный университет», Новосибирск, Россия

<sup>3</sup> 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.

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

N.S. Safronova<sup>1,2</sup>, M.P. Ponomarenko<sup>1,2</sup>,  
I.I. Abnizova<sup>3</sup>, G.V. Orlova<sup>1</sup>, I.V. Chadaeva<sup>1</sup>,  
Y.L. Orlov<sup>1,2</sup>

<sup>1</sup> Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

<sup>2</sup> Novosibirsk State University, Novosibirsk, Russia

<sup>3</sup> 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 flanking monomer repeats define a lower context complexity of sites containing SNPs within the human genome. An effect of local decrease in text complexity in SNP-containing sites is confirmed by analysis of polymorphisms in the rat and mouse genomes. We have found context differences between coding and regulatory sequences. These differences reflect 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 fixation 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 identification of the frequency spectrum of oligonucleotides with fixed lengths, and a comparison of oligonucleotide frequencies in different samples

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

#### КАК ЦИТИРОВАТЬ ЭТУ СТАТЬЮ?

Сафронова Н.С., Пономаренко М.П., 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

#### HOW TO CITE THIS ARTICLE?

Safronova N.S., Ponomarenko M.P., Abnizova I.I., Orlova G.V., Chadaeva I.V., Orlov Y.L. Flanking monomer repeats define lower context complexity of sites containing single nucleotide polymorphisms in the human genome. Vavilovskii Zhurnal Genetiki i Selekcii = Vavilov Journal of Genetics and Breeding. 2015;19(6):668-674. DOI 10.18699/VJ15.092

**И**сследование 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 *НарМар* (International *НарМар* 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 *НарМар* 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; Игнатъева 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; Орлов, Potapov, 2004; Орлов 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; Орлов et al., 2006; Trifonov et al., 2012).

Связь так называемых «горячих точек» мутаций в геномах с окружающим нуклеотидным контекстом была по-

казана 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](http://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 (Тройанская 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 (дубликации коротких последовательностей), необходимых для порождения текста из некоторого базового текста



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

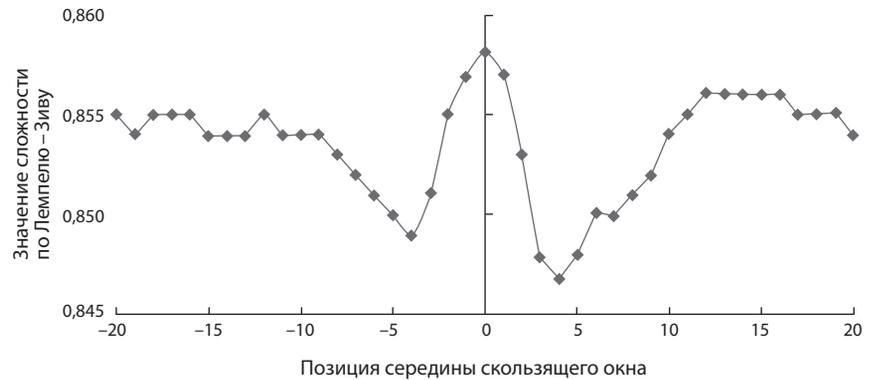
Мера варибельности  $C_{var}$  (мера чередования символов) в скользящем окне также для участков ОНП в геноме человека представлена на рис. 3.

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 политрактатов, в последовательностях, фазированных относительно точки ОНП.

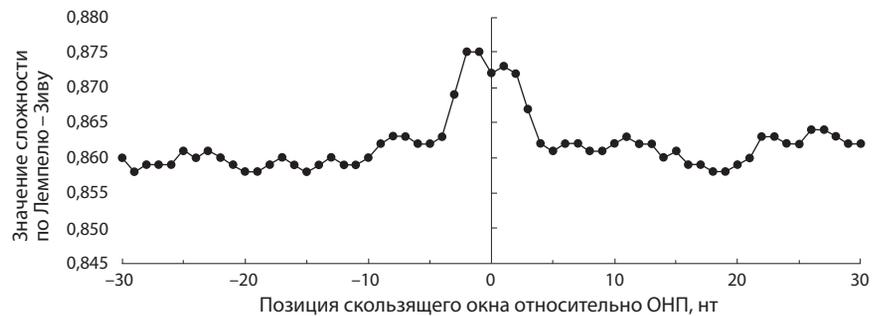
Присутствие сигнала ТАА в горячих точках мутаций было показано еще в ранней работе И.Б. Рогозина и Н.А. Колчанова на ограниченной выборке данных (Rogozin, Kolchanov, 1992). В нашей работе представлен расширенный набор олигонуклеотидов, содержащих короткие повторы нуклеотидов А и Т, непосредственно фланкирующих 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

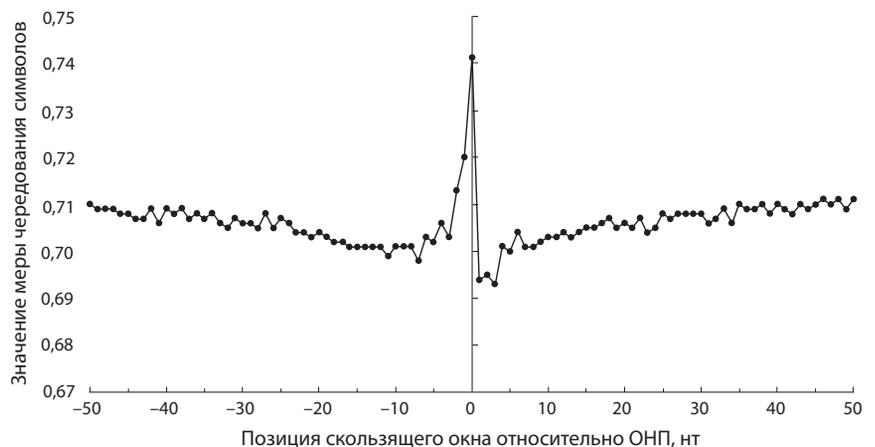


**Рис. 1.** Усредненный профиль сложности по Лемпелю – Зиву в скользящем окне для сайтов ОНП в геноме человека.

Размер окна – 7 нт; замена нуклеотида А на другие три нуклеотида.

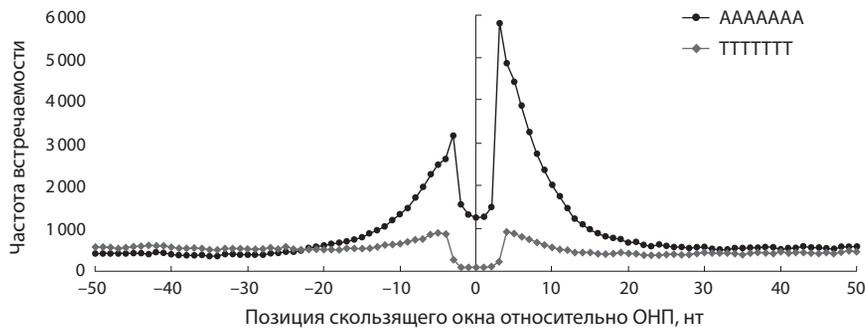


**Рис. 2.** Усредненный профиль изменения сложности по Лемпелю – Зиву в скользящем окне длины 7 нт для сайтов ОНП в геноме крысы.



**Рис. 3.** Профиль изменения частоты смены символа (мера мономеров) для выборки последовательностей ОНП в геноме человека.

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



**Рис. 4.** Профили встречаемости поли-А трактов длины 7 нт в участках, содержащих ОНП в геноме человека.

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 геномной сети. В свою очередь, конвергентное возникновение прямых повторов в белок-кодирующих районах генов связано с кодированием элементов вторичной структуры ( $\alpha$ -спиралей и  $\beta$ -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 Vabenko с коллегами (1999) был обнаружен специфический повторяющийся контекст локального окружения сайта в кор-промоторах генов человека для ТАТА-связывающего белка (ТВР), 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 ТВР 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).

### Благодарности

Компьютерные вычисления проводились на оборудовании ЦКП «Биоинформатика» ИЦиГ СО РАН и ССКЦ СО РАН. Работа была поддержана грантом РФФИ 14-04-01906 (разработка программ геномного анализа), РФФИ-ГФЕН 15-54-53091 (анализ полиморфизмов человека) и бюджетным проектом ИЦиГ СО РАН VI.61.1.2.

### Конфликт интересов

Авторы заявляют об отсутствии конфликта интересов.

### Список литературы

- Игнатъева Е.В., Подколотная О.А., Орлов Ю.Л., Васильев Г.В., Колчанов Н.А. Регуляторная геномика – экспериментально-компьютерные подходы. *Генетика*. 2015;51(4):409-429.
- Кулакова Е.В., Спицина А.М., Орлова Н.Г., Дергилов А.И., Свичкарев А.В., Сафронова Н.С., Черных И.Г., Орлов Ю.Л. Программы анализа геномных данных секвенирования, полученных на основе технологий ChIP-seq, ChIA-PET и Hi-C. Программные системы: теория и приложения. 2015;6(2):129-148.
- Орлов Ю.Л. Анализ регуляторных геномных последовательностей с помощью компьютерных методов оценок сложности генетических текстов: Дис. ... канд. биол. наук. Новосибирск, 2004.
- Орлов Ю.Л., Брагин А.О., Медведева И.В., Гунбин И.В., Деменков П.С., Вишневецкий О.В., Левицкий В.Г., Ощепков В.Г., Подколотный Н.Л., Афонников Д.А., Гроссе И., Колчанов Н.А. ICGenomics: программный комплекс анализа символьных последовательностей геномики. *Вавилонский журнал генетики и селекции*. 2012;16(4/1):732-741.
- Орлов Ю.Л., Левицкий В.Г., Смирнова О.Г., Подколотная О.А., Хлебодарова Т.М., Колчанов Н.А. Статистический анализ последовательностей ДНК, содержащих сайты формирования нуклеосом. *Биофизика*. 2006;51(4):608-614.
- Пономаренко П.М., Савинкова Л.К., Драчкова И.А., Лысова М.В., Аршинова Т.В., Пономаренко М.П., Колчанов Н.А. Пошаговая модель связывания ТВР/ТАТА-боксов позволяет предсказать наследственное заболевание человека по точечному полиморфизму. *Докл. РАН*. 2008;419(6):828-832.
- Поляновский О.Л., Лебедева Е.Н., Деев С.М. ERBB-онкогены – мишени моноклональных антител. *Биохимия*. 2012;77(3):289-311.
- Савинкова Л.К., Пономаренко М.П., Пономаренко П.М., Драчкова И.А., Лысова М.В., Аршинова Т.В., Колчанов Н.А. Полиморфизмы ТАТА-боксов промоторов генов человека и ассоциированные с ними наследственные патологии. *Биохимия*. 2009;74(2):149-163.
- Спицина А.М., Орлов Ю.Л., Подколотная Н.Н., Свичкарев А.В., Дергилов А.И., Чен М., Кучин Н.В., Черных И.Г., Глинский Б.М. Суперкомпьютерный анализ геномных и транскриптомных данных, полученных с помощью технологий высокопроизводительного секвенирования ДНК. Программные системы: теория и приложения. 2015;6:1(23):157-174.
- 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
- 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., McCarroll S.A., Nemes 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., Ghorri 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.
- 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., 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., 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
- 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
- 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. DOI 10.1016/0167-4781(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
- 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., Zoledziwska M., Mulas A., Pistis G., Steri M., Danjou F., Kwong A., Ortega Del Vecchio 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
- 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