DOI: 10.15514/ISPRAS-2022-34(2)-15

Модификация метода расчета полигенных рисков с использованием графа вариации

[1,2] *О.А. Кондратьева, ORCID: 0000-0001-6220-5077 <kondratyeva@ispras.ru>*
[1] *Е.А. Карпулевич, ORCID: 0000-0002-6771-2163 <karpulevich@ispras.ru>*

[1] *Институт системного программирования им. В.П. Иванникова РАН, 109004, Россия, г. Москва, ул. А. Солженицына, д. 25*
[2] *Московский государственный университет имени М.В. Ломоносова, 119991, Россия, Москва, Ленинские горы, д. 1*

Аннотация. Представление последовательности ДНК возможно в различном виде. Граф вариации один из самых точных методов, который позволяет работать с нетипичными участками и учитывать все их разнообразие. На основе этой структуры данных и метода полигенной оценки риска была построена система интерпретации ДНК. В результате был получен коэффициент корреляции между путем в графе, отвечающим за конкретную последовательность ДНК, и признаком. Затем мы сравнили его с коэффициентом, полученным аналогичным методом, но использующим представление последовательности с использованием эталонного генома. Такое сравнение помогло оценить эффективность представления в виде графа. После этого был построен модифицированный метод подсчета полигенной оценки на данных выравнивания инструмента vg, который также был сравнен с существующими методами. Модифицированный метод показал улучшение прогноза признака.

Ключевые слова: граф; представление генома; граф вариаций; HISAT2; vg;minimap2; GGP; геномный граф; PRS; полигенная оценка; полигенная оценка риска.

Для цитирования: Кондратьева О.А., Карпулевич Е.А. Модификация метода расчета полигенных рисков с использованием графа вариации. Труды ИСП РАН, том 34, вып. 2, 2022 г., стр. 191-200. DOI: 10.15514/ISPRAS–2022–34(2)–15

Modification of the Method for Calculating Polygenic Risks with Variation Graph

[1,2] *O.A. Kondrateva, ORCID: 0000-0001-6220-5077 <kondratyeva@ispras.ru>*
[1] *E.A. Karpulevich, ORCID: 0000-0002-6771-2163 <karpulevich@ispras.ru>*

[1] *Ivannikov Institute for System Programming of the Russian Academy of Sciences, 25, Alexander Solzhenitsyn st., Moscow, 109004, Russia*
[2] *Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow, 119991, Russia*

Abstract. Representation of the DNA sequence is possible in various ways. The variation graph is one of the most accurate methods that allows you to work with atypical areas and take into account all their diversity. Based on this data structure and the polygenic risk assessment method, a DNA interpretation system was built. As a result, a correlation coefficient was obtained between the path in the column responsible for a specific DNA sequence and the feature. We then compared it with a coefficient obtained by a similar method but using sequence representation using a reference genome. Such a comparison helped to evaluate the effectiveness of the representation in the form of a graph. After that, a modified method for calculating the polygenic score on the alignment data of the vg tool was built, which was also compared with existing methods. The modified method showed an improvement in the prediction of the trait.

Keywords: graph; genome representation; variation graph; HISAT2; vg; minimap2; GGP; genomic graph pipeline; PRS; polygenic score; polygenic risk score.

For citation: Kondrateva O.A., Karpulevich E.A. Modification of the Method for Calculating Polygenic Risks with Variation Graph. Trudy ISP RAN/Proc. ISP RAS, vol. 34, issue 2, 2022, pp. 191-200 (in Russian). DOI: 10.15514/ISPRAS–2022–34(2)–15

1. Введение

У живых организмов их генетическая информация содержится в молекуле дезоксирибонуклеиновой кислоты (ДНК). У всех клеточных организмов, включая человека и растение арабидопсис, геном (совокупность наследственного материала, заключённого в клетке организма) состоит из ДНК. ДНК отличается между разными организмами и определяет наличие различных признаков. Например, у человека это может быть голубой цвет глаз или предрасположенность к какой-нибудь болезни. Такие признаки называются фенотипом организма.

Информация в ДНК хранится в виде кода, состоящего из четырех химических оснований: аденина (A), гуанина (G), цитозина (C) и тимина (T). Секвенирование ДНК – это процесс чтения нуклеотидных оснований в молекуле ДНК. Прибор, с помощью которого проводится секвенирование – секвенатор считывает только небольшие фрагменты ДНК, называемые короткими чтениями (ридами), например, длиной 120 пар оснований. Процесс повторной сборки всей последовательности ДНК из получившихся ридов называется выравниванием. Существует секвенирование парных прочтений, которое позволяет использовать короткие фрагменты и секвенировать и в прямую, и в обратную сторону, что повышает точность дальнейшего выравнивания. Существующие инструменты для осуществления повтороной сборки ДНК в одну последовательность можно разделить на использующие эталонную последовательность (или референс) и графовые.

В случае использования референса в процессе выравнивания каждый короткий прочитанный фрагмент сравнивается с эталонным геномом, чтобы найти наилучшее совпадающее место, где он будет соответствовать наименьшему количеству различий. В случае парных прочтений каждое чтение выравнивается отдельно, а информация по обеим парам объединяется и сообщается в одной строке выравнивания. В высоко вариабельных участках последовательности (человеческие лейкоцитарные антигены в т.ч. главный комплекс гистосовместимости) использование эталонного генома затруднительно и приводит к потере информации. Поэтому вместо референсной последовательности можно использовать графовое представление, которое имеет более сложную структуру и позволяет учитывать вариации включай делеции и вставки.

Графы занимают давнее место в анализе биологических последовательностей, где они часто используются для компактного представления множества возможных последовательностей. Как правило, сами последовательности неявно представляются как обходы в графе. Возможно, самым простым представлением графа является ориентированный граф. В контексте сборки генома графы де Брейна являются популярными представлениями ориентированных графов, в которых каждый узел представляет k-mer (уникальную строку длины k), и каждое направленное ребро представляет собой перекрытие k−1 оснований между суффиксом узла «от» и префиксом узла «к».

Ориентированные графы не полностью выражают концепцию двух цепочек ДНК. То есть они не различают чтение молекулы ДНК в ее прямой и обратной комплементарной ориентации. Чтобы выразить это свойство, ориентированные графы можно обобщить до двунаправленных графов, в которых каждая конечная точка ребра имеет независимую ориентацию, указывающую, должна ли прямая или обратная дополнительная цепочка присоединенного узла быть посещается при входе в узел через эту конечную точку ребра. Инверсии, обратные тандемные дупликации и произвольно сложные перестановки

выражаются в двунаправленном представлении. Такие сложные варианты не могут быть выражены в версии ориентированного графа без создания независимых узлов прямого и обратного обхода и сохранения дополнительной информации для описания этой взаимодополняемости. Версия двунаправленных графов с ребрами дает эквивалентное представление.

2. Существующие решения

2.1 Minimap2

Существует множество инструментов, которые реализуют выравнивание на эталонную последовательность, такие как BLASR [1], BWA-MEM [2], GraphMap [3], Kart [4] и minimap2 [5]. Minimap2 один из новых инструментов и превосходит другие аналогичные инструменты выравнивания для конкретных предметных областей как по скорости, так и по точности [5]. Поэтому он был выбран для сравнительного анализа с другими.

Minimap2 использует один и тот же базовый алгоритм, но разные наборы параметров в зависимости от типов входных данных. Возможности minimap2 основаны на быстром базовом алгоритме выравнивания последовательности [6] и точном алгоритме выравнивания цепочек [7]. Для этих алгоритмов слабым местом производительности является выравнивание длинных последовательностей, что было непрактично медленным 10 лет назад. Алгоритм Судзуки-Касахары (Hajime Suzuki, Masahiro Kasahara) [8] в значительной степени устраняет эту проблему.

2.2 HISAT2

Продолжением развития подхода с использованием эталонного генома, является сначала создание линейного графа референса, а затем добавление мутаций в качестве альтернативных путей по графу (рис. 1). Такой метод был реализован в инструменте HISAT2 [9].

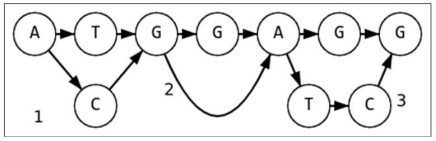

Рис. 1. Представление с использованием графа куска последовательности. Под цифрами указаны различные виды структурных вариаций. 1 - Однонуклеотидный полиморфизм (SNP). 2 - Удаление (делеция). 3 - Вставка

Fig. 1. Representations using the sequence piece graph. Below the numbers are the various types of structural variations. 1 - Single nucleotide polymorphism (SNP). 2 - Deletion (deletion). 3 - Insert

Алгоритм реализует структуру данных на основе графов и использует преобразования Берроуза-Уилера (Michael Burrows, David Wheeler) (BWT) [10]. Вершины представляют собой одно из оснований {A, C, G, T}, а их отношения представлены как ребра. Чтобы обеспечить быстрое нахождение пути в графе, который соответствует конкретному ДНК были сделаны несколько шагов для оптимизации. В отличие от других алгоритмов выравнивания графов, которые используют индексы на основе k-меров, требующих интенсивного использования памяти, таких как vg, HISAT2 использует индекс графа Феррагины-Манзини (Paolo Ferragina, Giovanni Manzini) (ФМ-индекс, GFM) [11].

2.3 Пайплайн геномного графа

В данном инструменте была реализована структура данных генома графа, которая представляет геномные последовательности на ребрах графа. Геномный граф (GGP) строится из популяции геномных последовательностей, так что каждый геном в этой популяции

представлен путем последовательности через граф. Эксперименты, проведенные создателями инструмента [12] по сравнительному анализу, демонстрируют, что использование геномного графа улучшает выравнивание прочтений и поиск вариантов без сопутствующей потери точности.

В настоящее время GGP анализирует образцы по отдельности, и версия, которая выполняет совместный поиск вариантов, находится в стадии разработки. Информация, такая как частоты аллелей каждого варианта и неравновесие по сцеплению между ними, может быть включена в граф, предоставляя дополнительную статистическую информацию для выравнивания считывания и поиска вариантов. В данный момент это реализовано только в графе вариаций.

2.4 Граф вариаций

Граф вариаций [13] представляет собой двунаправленный граф последовательностей ДНК, который представляет генетические вариации в популяции. Эта структура данных обеспечивают сжатое кодирование последовательностей многих геномов. Граф вариаций состоит из:

- вершин, в которых содержится последовательность из {A, C, T, G} и идентификатор;
- ребер, которые соединяют две вершины через любой из их соответствующих концов;
- путей, которые описывают геномы, выровненные последовательностей и аннотации как переходы через вершины, соединенные ребрами.

Чтобы обеспечить сопоставление чтения и другие операции доступа к большим графам последовательностей, инструмент использует краткое представление графа вариаций vg (xg), которое является статическим, но эффективным с точки зрения памяти и времени.

2.5 Сравнение различных представлений

Выравнивание прочтений генома является первым шагом в большинстве рабочих процессов анализа генома. Существует несколько подходов к решению этой задачи, которые можно разделить на две основные категории:

- использование эталонного генома;
- использование графового представления.

Табл. 1. Сравнение инструментов для выравнивания
Table 1. Comparison of alignment tools

Название	Результат	Использование графа	Открытый код
minimap2	sam	Нет	Да
hisat2	sam	Да	Да
GGP	bam	Да	Нет
vg	bam	Да	Да

Хотелось бы отметить, что инструмент GGP, хоть и показал хорошие результаты, к сожалению, не имеет открытого кода. Также для дальнейшей трансформации большинство инструментов биоинформатики принимают и ожидают результатов выравнивания в формате bam, а minimap2 и HISAT2 выдают sam. Это исправляется путем использования инструмента преобразования samtools и возможно, поскольку информация в этих форматах одинаковая, просто bam – это двоичный файл.

3. Задача интерпретирования

По последовательности ДНК возможно определить некоторые признаки организма. Такая операция называется интерпретированием ДНК. Считается, что все признаки организма имеют генетический компонент, однако степень, в которой гены способствуют развитию

фенотипа, варьируется. Существует несколько видов фенотипов - моногенные, хромосомные и полигенные. Моногенные признаки связаны с одним геном. Например, наличие такого заболевания как серповидноклеточная анемия у человека обусловлено мутацией гена HBB. Хромосомные признаки обусловлены изменением количества хромосом. Полигенные признаки же обычно являются результатом сочетания различных генов, каждый из которых увеличивает восприимчивость к этому состоянию. Изучение такого фенотипа всегда представляло сложную задачу. Один из методов интерпретирования называется полигенной оценкой (PRS). Он состоит в том, чтобы на основе данных полногеномного исследования ассоциации (GWAS) и генома организма в формате VCF получить число, которое коррелирует с признаком.

4. Подсчет полигенной оценки

Стандартные оценки полигенного риска обычно строятся на основе взвешенной суммы количества аллелей, то есть:

$$PRS = \sum_i \beta_i x_i$$

Существует несколько подходов для подсчета β_i. Некоторые включенные варианты могут быть ложноположительными, и необработанные оценки размеров эффекта от них могут быть подвержены систематической ошибке отбора [14]. Кроме того, стандартный подход PRS требует тестирования в диапазоне пороговых значений P-value, которые часто выбираются произвольно. Ошибка предсказания, оцениваемая по оптимизированному порогу, также может быть подвержена оптимистическому смещению. Чтобы улучшить предсказание геномного риска, был предложен эмпирические байесовские подходы для восстановления основных размеров эффекта. Этот метод удовлетворительно показал себя при моделировании. Он прост в вычислительном отношении и не требует предположений о распределении размера эффекта. Этот подход заключается в использовании формулу Твиди (Maurice Tweedie) для корректировки оценки β_i:

$$\boldsymbol{\beta_{Twe,i}} = \sqrt{\boldsymbol{\xi}\left(\boldsymbol{z_i} + \frac{\boldsymbol{f'(z_i)}}{\boldsymbol{f(z_i)}}\right)}$$

где ξ – функция преобразования z-статистики в дисперсию объясненной ответственности, описанная в [15], $f(x)$ – ядерная оценка плотности (KDE):

$$f(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

где K является ядром, то есть неотрицательной функцией, а $h > 0$ является сглаживающим параметром, называемым шириной полосы.

Другой подход использует формулу:

$$\beta_{Twe,tdr} = \beta_{Twe}(1 - fdr),$$

где fdr – локальная частота ложных открытий, вероятность нулевого значения с учетом наблюдаемой z-статистики. $(1 - fdr)$, следовательно, является локальной истинной скоростью обнаружения. Этот метод взвешивает каждую оценку величины эффекта по формуле Твиди с вероятностью того, что она не равна нулю. На практике это приведет к дальнейшему уменьшению размеров эффекта до нуля, и часть размеров эффекта станет равной нулю, поскольку локальный fdr равен единице для некоторых маркеров.

Еще один рассматриваемый способ использует:

$$\beta_{tdr} = \beta(1 - fdr),$$

в котором коэффициенты регрессии взвешиваются по локальным истинным показателям обнаружения.

5. Интерпретирование с помощью полигенной оценки

Для анализа генома с помощью оценки полигенного риска для начала требуется выбрать модель для определения вероятности возникновения фенотипа. Поскольку в данном исследовании фенотип представляет собой непрерывную величину, то в таком случае используется линейная регрессия:

$$y = \mu + x\beta + \varepsilon,$$

где μ – коэффициент регрессии.

Линейная регрессия оценивается коэффициентом детерминации R^2:

$$R^2 = \frac{SS_{reg}}{SS_{tot}}$$

где

$$SS_{tot} = SS_{reg} + SS_{res}$$

$$SS_{reg} = \sum_{i=1}^{n} (\hat{y}_i - \acute{y})^2$$

$$SS_{res} = \sum_{i=1}^{n} y_i - \hat{y}_i$$

$\hat{y}_i$ – значение, полученное регрессией, y_i – фактическое значение. Среднее фактических значений:

$$\acute{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

6. Добавление оценки в граф

Геномные графы построены таким образом, что последовательность экземпляра может быть представлена как путь в графе. Чем больше различных вариантов содержит граф, тем точнее будет путь отражать действительную последовательность. Так, при построении графа можно добавить его вершинам оценку, полученную из суммарной статистики (рис. 2), и затем использовать эти данные для подсчета PRS экземпляра.

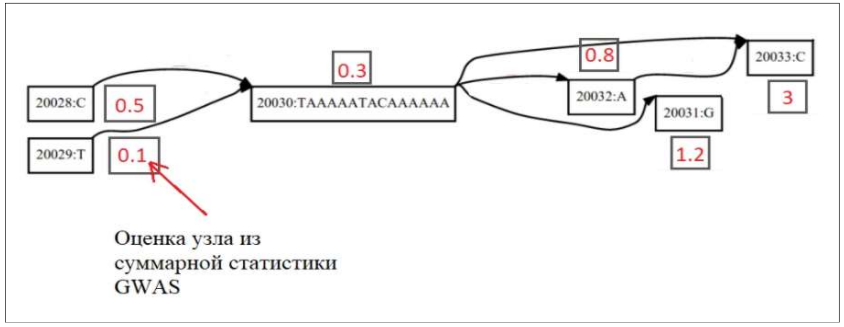

Рис 2. Граф вариаций с оценкой влияния
Fig. 2. Graph of variations with impact assessment

После построения графа возникает следующий этап – выравнивание данных ДНК на него. Эта операция заключается в поиске пути, который бы соответствовал новой последовательности. Благодаря добавленным ранее данным об определенном признаке, в этот момент можно просуммировав весы всех узлов, входящих в найденный путь, получить полногенномную оценку. После этого необходимо построить коэффициент корреляции этой оценки с фенотипом и сравнить результат с исследованием похожего типа.

7. Набор данных

Для оценки эффективности метода были проведены экспериментальные исследования на наборах данных растения Arabidopsis thaliana. Доступность естественно инбредных штаммов позволяет повторять фенотипирование одного и того же адаптированного генотипа в различных контролируемых условиях, что делает Arabidopsis thaliana подходящим для изучения взаимодействия генотип-среда. Существует проект «1001 геном», который был запущен в начале 2008 г. с целью выявления подробной вариации последовательности всего генома как минимум в 1001 образце [16].

Из доступных в каталоге AraGWAS экспериментов было выбрано несколько исследований с доступной суммарной статистикой. Кроме того, был смоделирован собственный GWAS с помощью инструмента plink на 135 экземплярах. Это было сделано, так как все доступные данные об индивидуальных геномах Arabidopsis используются в подсчете. Чтобы проверить, как сильно это влияет, и был построен отдельный GWAS, который был потом протестирован на 44 отдельных экземплярах. В итоге было проведено 2 эксперимента на следующих GWAS.

- M216T665 [17] – признак содержания метаболитов. Профили нецелевых метаболитов ткани листа на основе LC-MS собирали для каждого образца. Характеристика метаболита с отношением массы к заряду 216 и временем удерживания 665 с.
- GWAS, построенный на основе признака в предыдущем пункте.

Собственный GWAS был построен только на 4-й хромосоме, так как исследование M216T665 содержит все значимые варианты только на определенном участке этой хромосомы. Также был построен график (рис. 3), который схож с тем, что показывает каталог для оригинального исследования.

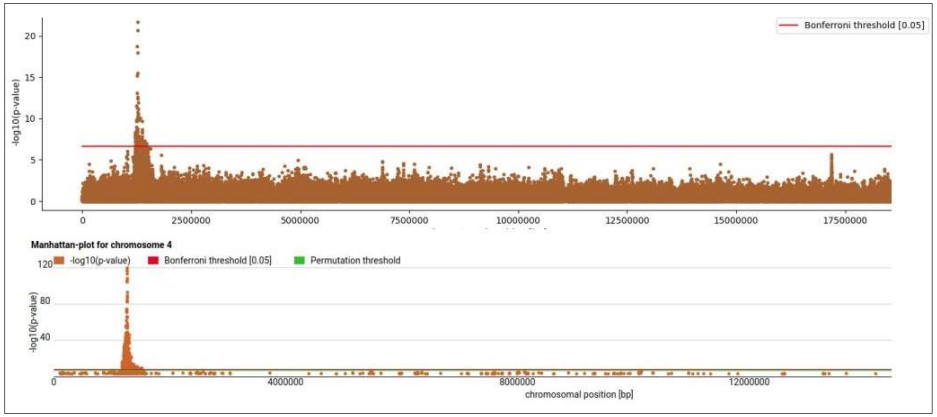

Рис. 3. Манхэттенский график для построенного GWAS на верхнем рисугке и для оригинального GWAS (на нижнем рисунке). Построен для 4-й хромосомы
Fig. 3. Manhattan plot for the constructed GWAS on the top picture and a plot for the original GWAS (on the bottom). Built for chromosome 4

Моделирование данных с секвенаторов было проведено используя данные реального генотипа из проекта 1001 геном. Это исследование содержит данные в формате vcf. С помощью инструмента ART [18] возможно моделирование парных чтений секвенатора Illumina.

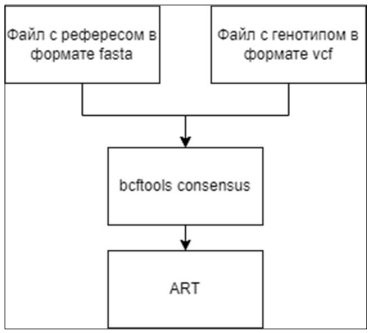

Рис. 4. Моделирование парных чтений секвенатора Illumina
Fig. 4. Modelling of paired readings of the Illumina sequencer

ART выводит данные чтения в формате fastq, но на вход требует последовательность (без чтений) в формате fasta. Для того, чтобы использовать инструмент ART данный генотипа сначала переводятся из vcf в fasta с помощью bcftools consensus [19]. Он позволяет создать согласованную последовательность для экземпляра, где последовательность включает варианты, выбранные для этого индивидуума. Подробно схема работы для моделирования показана на рис. 4.

8. Результаты

Сначала был проведен анализ использования различных коэффициентов для подсчета PRS. Использовался $\beta_{Twe,tdr}$, β_{tdr}, $\beta_{Twe,i}$, где $i = 207$ равен размеру выборки на которой считался PRS. Анализ проводился для GWAS M216T665. Посчитанный R^2 показан в табл. 2.

Табл. 2. Коэффициент детерминации R^2 для различных коэффициентов при подсчете PRS
Table 2. Coefficient of determination R^2 for various coefficients when calculating PRS

Коэффициент	$\beta_{Twe,tdr}$	β_{tdr}	$\beta_{Twe,207}$
minimap2	0.6223	0.5531	0.6956
HISAT2	0.6391	0.5776	0.704
GGP	0.5898	0.5153	0.6751
vg	0.5831	0.5059	0.6663
Среднее значение	0.6086	0.538	0.6853

Наилучший результат показал последний коэффициент, поэтому в дальнейшем именно он будет использоваться для подсчета R^2.

Для тех GWAS, на которых проводились исследования, результаты показаны в табл. 3.

Табл. 3. Коэффициент детерминации R^2 для различных исследований GWAS
Table 3. Coefficient of determination R^2 for various GWAS studies

Исследование	M216T665	Собственный GWAS (158/44)	Среднее значение
minimap2	0.6956	0.3049(0.3298/0.2908)	0.4053
HISAT2	0.704	0.271(0.3096/0.2897)	0.3936
GGP	0.6751	0.289(0.3362/0.2817)	0.3955

vg	0.6663	0.3095(0.3279/0.285)	0.3972
vg + score	0.6692	0.4583(0.5032/0.4884)	0.5298

9. Выводы

Мы разработали модифицированный метод полигенной оценки, который подсчитывается на результатах выравнивания инструмента vg. Он был встроен в существующий инструмент путем добавления опции, которая указывает на необходимость подсчета PRS. Разработанный метод сравнивался с существующими 4 инструментами: HISAT2, minimap2, vg (без модификации) и GGP. Для этого были написаны программы подсчета PRS и R^2.

Для подтверждения результатов было проведено исследования на 3-х наборах GWAS, которые показали одинаковую динамику. Проведение экспериментов было автоматизировано с помощью разработанной программы. Разработанный метод увеличил коэффициент детерминации R^2 для инструмента vg с 0.3972 до 0.5298 в среднем, что показывает улучшение предсказательной способности.

Список литературы / References

[1]. Chaisson M.J., Tesler G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics, vol. 13, 2012, article no. 238, 17 p.

[2]. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997, 2013, 3 p.

[3]. Sović I., Šikić M. et al. Fast and sensitive mapping of nanopore sequencing reads with GraphMap. Nature communications, vol. 7, issue 1, 2017, article no. 11307, 11 p.

[4]. Lin H.-N., Hsu W.-L. Kart: a divide-and-conquer algorithm for NGS read alignment. Bioinformatics, vol. 33, issue 15, 2017, pp. 2281-2287.

[5]. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, vol. 34, issue 18, 2018, pp. 3094-3100.

[6]. Polyanovsky V.O., Roytberg M.A. Tumanyan V.G. Comparative analysis of the quality of a global algorithm and a local algorithm for alignment of two sequences. Algorithms for Molecular Biology, vol. 6, 2011, article no. 25, 12 p.

[7]. Abouelhoda M.I., Ohlebusch E. Chaining algorithms for multiple genome comparison. Journal of Discrete Algorithms, vol. 3, issues 2-4, 2005, pp. 321–341.

[8]. Suzuki H., Kasahara M. Introducing difference recurrence relations for faster semi-global alignment of long sequences. BMC bioinformatics, vol. 19, 2018, article no. 45, 14 p.

[9]. Kim D., Paggi J.M. et al. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature Biotechnology, vol. 37, 2019, pp. 907-915.

[10]. Li H. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics, vol. 25, issue 14, 2009, pp. 1754-1760.

[11]. Simpson J.T., Durbin R. Efficient construction of an assembly string graph using the FM-index. Bioinformatics, vol. 26, issue 12, 2010, pp. i367–i373

[12]. Rakocevic G., Semenyuk V. et al. Fast and accurate genomic analyses using genome graphs. Nature genetics, vol. 51, 2019, pp. 354-362.

[13]. Garrison E., Sirén J. et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. /Nature biotechnology, vol. 36, 2018, pp. 875-879.

[14]. So H.-C. Improving polygenic risk prediction from summary statistics by an empirical Bayes approach, Scientific reports, vol. 7, 2017, article no. 41262, 11 p.

[15]. So H.-C. Uncovering the total heritability explained by all true susceptibility variants in a genome-wide association study. Genetic epidemiology, vol. 35, issue 6, 2011, pp. 447-456.

[16]. Detlef W., R. The 1001 genomes project for Arabidopsis thaliana. Genome biology, vol. 10, 2009, article no. 107, 5p.

[17]. AraGWAS Catalog. 'M216T665' URL: http://aragwas.1001genomes.org/#/study/144.

[18]. Huang W., Li L. et al. ART: a next-generation sequencing read simulator. Bioinformatics, vol. 28, issue 4, 2012, pp. 593-594.

[19]. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics, vol. 27, issue 21, 2011, pp. 2987-2993.

Информация об авторах / Information about authors

Олеся Анатольевна КОНДРАТЬЕВА является студентом магистратуры кафедры системного программирование МГУ, работает в ИСП РАН. Ее научные интересы выключают в себя графовые представления генома, биоинформатику.

Olesia Anatolevna KONDRATEVA is a master's student of the Department of System Programming of the MSU, also works at ISP RAS. Her scientific interests include graph representations of the genome, bioinformatics.

Евгений Андреевич КАРПУЛЕВИЧ является специалистом отдела информационных систем. Сфера научных интересов: применение алгоритмов анализа данных к биомедицинскому домену, разработку систем распределенного хранения и анализа данных.

Evgeny Andreevich KARPULEVICH is a specialist of the Information Systems Department. Research interests: application of data analysis algorithms to the biomedical domain, development of systems for distributed data storage and analysis.