

DOI: 10.15514/ISPRAS-2025-37(6)-30



Оптимизация выравнивания коротких прочтений с инделями при полногеномном секвенировании

Н.А. Колтунов, ORCID: 0009-0007-8507-8518 <n_koltunov@ispras.ru>

Е.П. Гугучкин, ORCID: 0000-0001-7885-9892 <guguchkin@ispras.ru>

Е.А. Карпулевич, ORCID: 0000-0002-6771-2163 <karpulevich@ispras.ru>

*Институт системного программирования им. В.П. Иванникова РАН,
Россия, 109004, г. Москва, ул. А. Солженицына, д. 25.*

Аннотация. Представлен новый метод выравнивания прочтений для задач полногеномного секвенирования (WGS), ориентированный на повышение точности и практической эффективности этого этапа геномного анализа. В отличие от графовых подходов, предложенный алгоритм интегрирует информацию об известных генетических вариантах напрямую в процесс выравнивания, что позволяет улучшить сопоставление последовательностей с эталонным геномом без строительства сложных графовых структур. Метод продемонстрировал высокую эффективность на реальных данных: наблюдается устойчивый прирост качества выравнивания на участках с высоким уровнем изменений между разными людьми, а также участках, которые сложны для однозначного выравнивания даже при отсутствии изменений в этом месте у конкретного человека. В частности, использование информации о вариантах позволяет точнее выравнивать короткие последовательности (прочтения), содержащие альтернативные аллели, снижая число ошибок в указанных регионах. При этом требуемые вычислительные ресурсы остаются на приемлемом уровне, что делает решение применимым в стандартных WGS-пайплайнах без существенного увеличения нагрузки. Скорость работы алгоритма сопоставима с традиционными решениями, что упрощает его интеграцию в существующие аналитические программные конвейеры. Практическая ценность метода заключается в улучшении точности выравнивания, что напрямую влияет на качество последующего обнаружения вариантов и других анализов. Предлагаемый подход способен служить эффективной альтернативой современным графовым методам выравнивания, обеспечивая сопоставимое повышение качества результатов выравнивания при меньшей сложности реализации. Перспективы дальнейшего развития включают оптимизацию производительности алгоритма, расширение набора учитываемых генетических вариантов и проведение углубленного сравнения с другими инструментами. Эти шаги призваны еще более повысить эффективность и надежность метода, укрепляя его значимость для практического применения в геномике.

Ключевые слова: выравнивание коротких прочтений; инделы; альтернативные контиги; преобразование координат; определение вариантов.

Для цитирования: Колтунов Н.А., Гугучкин Е.П., Карпулевич Е.А. Оптимизация выравнивания коротких прочтений с инделями при полногеномном секвенировании. Труды ИСП РАН, том 37, вып. 6, часть 2, 2025 г., стр. 211–222. DOI: 10.15514/ISPRAS-2025-37(6)–30.

Благодарности: Исследование выполнено при поддержке Российского научного фонда, проект № 25-21-20111.

Optimization of Short Reads Alignment with Indels in Whole-Genome Sequencing

N.A. Koltunov, ORCID: 0009-0007-8507-8518 <n_koltunov@ispras.ru>

E.P. Guguchkin, ORCID: 0000-0001-7885-9892 <guguchkin@ispras.ru>

E.A. Karpulevich, ORCID: 0000-0002-6771-2163 <karpulevich@ispras.ru>

*Ivannikov Institute for System Programming of the Russian Academy of Sciences,
25, Alexander Solzhenitsyn st., Moscow, 109004, Russia.*

Abstract. We present a novel method for aligning reads in whole-genome sequencing (WGS), aimed at improving alignment accuracy and the practical efficiency of this stage of genomic analysis. Unlike graph-based approaches, the proposed algorithm directly integrates knowledge of known genetic variants into the alignment process, enabling more accurate mapping of reads to the reference genome without constructing complex graph structures. The method has demonstrated high effectiveness on real sequencing data: we observed a consistent improvement in read alignment quality in highly variable and difficult-to-map regions of the genome. In particular, using variant information allows more precise alignment of reads that contain alternative alleles, reducing the number of mapping errors in these regions. At the same time, the required computational resources remain at an acceptable level, making this solution applicable in standard WGS pipelines without a significant increase in workload. The alignment speed of the algorithm is comparable to traditional solutions, which facilitates its integration into existing analytical pipelines. The practical value of the method lies in the improved alignment accuracy, which directly affects the quality of downstream variant calling and other analyses. The proposed approach can serve as an effective alternative to current graph-based alignment methods, providing comparable improvements in alignment quality with lower complexity of implementation. Future work will include optimizing the algorithm's performance, expanding the set of genetic variants accounted for, and conducting in-depth comparisons with other tools. These steps are intended to further increase the method's efficiency and reliability, reinforcing its significance for practical use in genomics.

Keywords: short-read alignment; indels; alternate contigs; liftover; variant calling.

For citation: Koltunov N.A., Guguchkin E.P., Karpulevich E.A. Optimization of short reads alignment with indels in whole-genome sequencing. *Trudy ISP RAN/Proc. ISP RAS*, vol. 37, issue 2, part 2, 2025, pp. 211-222 (in Russian). DOI: 10.15514/ISPRAS-2025-37(6)-30.

Acknowledgements. This research was funded by the Russian Science Foundation (RSF) project No 25-21-20111.

1. Введение

Современные технологии секвенирования геномов становятся все дешевле и доступнее, что приводит к переходу от секвенирования отдельных геномов к массовому секвенированию популяций. Ярким примером является проект UK Biobank по секвенированию геномов жителей Великобритании, в рамках которого уже получены сотни тысяч полных геномов вместе с клиническими данными [1]. Накопление данных массового секвенирования породило обширные базы известных генетических вариантов человека, такие как база Human Pangenome Reference Consortium (HPRC) [2], содержащая сотни различных гаплотипов для более полного охвата генетического разнообразия. К ним примыкает и проект 1000 Genomes [3], ставший первой масштабной инициативой по систематическому каталогу распространенных вариантов в разных популяциях и заложивший методологические и ресурсные основы для последующих пангеномных эталонов. Использование информации о ранее известных вариантах генома конкретной популяции открывает возможности для улучшения биоинформатических методов анализа, в частности методов выравнивания прочтений и поиска вариантов.

Полное секвенирование генома (Whole Genome Sequencing, WGS) с применением технологий Next-Generation Sequencing (NGS) генерирует огромные объемы коротких фрагментов ДНК

– коротких прочтений, обычно длиной порядка 150 пар оснований. Выравнивание этих прочтений на референсный геном является одним из самых главных этапов анализа геномных данных. Широко используется двухшаговый подход (seed-and-extend), при котором и референс, и прочтения разбиваются на короткие подпоследовательности (сиды), ищутся совпадающие якорные участки, выстраиваются цепочки якорей, а затем выполняется точное выравнивание алгоритмом, например, Смита–Ватермана или Нидлмана–Вунша. Популярный инструмент выравнивания minimap2 [4] реализует близкую схему (seed-chain-align): после поиска сидов строится цепочка коллинеарных якорей (anchors), и уже вдоль нее выполняется точное выравнивание; при этом референс индексируется по минимайзерам с использованием хеш-структуры, сопоставляющей минимайзеры их позициям в референсе. Благодаря минимайзерам такой индекс позволяет быстро находить соответствия между прочтениями и референсом в масштабах всего генома.

Однако наличие генетических вариантов в секвенируемом геноме затрудняет выравнивание коротких прочтений на линейный референс. Если прочтение содержит индель (вставку или делецию относительно референса) или другой вариант, качество его выравнивания снижается – метрика MAPQ заметно падает. В ряде случаев такие прочтения вовсе не выравниваются в верное положение, что приводит к потерянными (ложно-негативным) вариантам при последующем вызове вариантов. Известно, что более двух третей структурных вариантов (SV) не обнаруживались при анализе данных коротких прочтений именно из-за отсутствия альтернативных аллелей в используемом референсном геноме. Проблемы возникают и в сложных регионах генома (например, повторах), где стандартные алгоритмы выравнивания коротких прочтений работают недостаточно точно [5]. Все это описывает проявление смещения относительно референса, при котором последовательности, отличающиеся от референса, выравниваются хуже.

Одним из путей решения проблемы является переход от линейного референса к пангеномному референсу или графу генома, учитывающему известные варианты. Построение графовой модели референса (genome graph) позволяет представить альтернативные аллели и индели в виде альтернативных путей и узлов графа. Графовые выравниватели (например, VG Giraffe [6]) демонстрируют более высокую точность в переменных регионах генома по сравнению с выравниванием на линейный эталон. Так, консорциум HPRC недавно представил первый черновой пангеномный референс человека, включающий 47 фазированных диплоидных сборок генома, добавив около 119 миллионов пар оснований новых последовательностей и значительно повысив выявление структурных вариантов. Использование этого пангенома при анализе коротких прочтений позволило сократить ошибки обнаружения вариантов и увеличить число выявляемых структурных вариантов по сравнению с традиционным процессом на основе GRCh38. Однако полные графовые решения остаются сложными вычислительно: построение и хранение индекса пангенома требует больших объемов памяти и времени. Это стимулирует поиск компромиссных подходов, сочетающих скорость линейного выравнивания с учетом известных вариантов.

Один из таких подходов – использование расширенного линейного референса с альтернативными контигами, содержащими известные варианты. Еще проект GRCh38 включал альтернативные контиги для сложных регионов, а современные методы развивают эту идею на уровне популяционных данных. Например, в проприетарном инструменте Illumina DRAGEN реализована мультигеномная карта: к основному референсу добавлены дополнительные последовательности – фрагменты гаплотипов из разных популяций [7]. Прочтения выравниваются одновременно на основной референс и на эти альтернативные контиги, связанные с ним через заранее рассчитанные соответствия (лифтовер-цепочки). Далее выравнивания оцениваются группами: если прочтение лучше выровнялось на альтернативный контиг, это выравнивание становится основным, но позиция прочтения все равно переводится на координаты основного референса, и в итоговом SAM/BAM файле

указывается стандартный хромосомный референс. Таким образом, внешне процесс не отличается от классического (все координаты в привычной сборке), а внутренняя точность выравнивания повышается за счет использования информации о популяционных вариантах. Показано, что такой ALT-ориентированный подход заметно улучшает точность выравнивания в сложных областях и снижает число ошибок в вызове SNP и инделов.

В данной работе предложен новый метод улучшения выравнивания коротких прочтений за счет модификации референсной последовательности для инструмента выравнивания `minimap2`. В отличие от полноценных графовых решений, наш подход сохраняет линейную структуру референса, но дополняет его контигами с инделями, отражающими известные варианты. По сути, референс обогащается альтернативными путями для прочтений, содержащих вставки или делеции, характерные для данной популяции. После индексирования такой расширенной базы в `minimap2`, прочтения выравниваются стандартным образом – наиболее подходящие кандидаты для выравнивания будут найдены даже если они содержат варианты, представленные на добавленных контигах. Инструмент `LevioSAM2` [8] затем используется для преобразования координат (лифтовера) полученных выравниваний: все позиции прочтений, выровненных на добавленные контиги, автоматически пересчитываются в координаты оригинального референса. В результате выравнивания и последующий поиск вариантов осуществляются относительно стандартной последовательности, но с повышенной точностью. Предлагаемый подход позволяет уменьшить число ложно не обнаруженных генетических вариантов (особенно инделов) и повысить качество выравнивания без существенного увеличения вычислительной нагрузки. Таким образом, наш метод может рассматриваться как альтернатива графовым методикам при анализе геномных данных человека (и потенциально других видов), сочетая преимущества учета известных вариаций генома с простотой и быстродействием линейного выравнивания.

2. Материалы и методы исследования

2.1 Выравнивание коротких прочтений с помощью `minimap2`

Для выравнивания прочтений применялся инструмент `minimap2` (версия 2.30). Он реализует стандартный алгоритм типа `seed-chain-align`, широко используемый в полногеномных инструментах выравнивания. На этапе индексирования референсный геном разбивается на короткие *k*-меры, из которых отбираются минимайзеры – наиболее лексикографически малые подстроки в скользящем окне фиксированной длины. `Minimap2` собирает такие минимайзеры референса и сохраняет их в хеш-индекс: ключом служит хеш минимайзера, а значением – конкатенированный список позиций данного минимайзера в геноме. За счет минимайзеров размер индекса существенно уменьшается без потери чувствительности.

Далее каждое входное чтение обрабатывается следующим образом. `Minimap2` вычисляет минимайзеры и для прочтения, затем с помощью индекса быстро находит все совпадающие позиции этих минимайзеров на референсе. Эти совпадения служат якорями (*anchors*) для предварительного выравнивания. На основе набора якорных совпадений выполняется поиск максимально длинной цепочки коллинеарных якорей – то есть последовательности совпадений, расположенных на прочтении и в геноме в одном и том же порядке (без нарушения относительного расположения). Это позволяет определить примерное положение сегмента прочтения на геноме, даже если между якорями имеются разрывы из-за инделов или других вариантов. Для найденной цепочки `minimap2` выполняет точное выравнивание прочтения на соответствующий участок референса, используя алгоритм динамического программирования (реализован в библиотеке `KSW2`). В результате формируется окончательное выравнивание с `CIGAR`-строкой, отражающей совпадающие и несовпадающие базы, вставки и делеции. Если для одного прочтения обнаруживается

несколько почти равных по качеству цепочек, инструмент выводит альтернативные (вторичные) выравнивания с пониженным MAPQ. Однако в общем случае minimap2 выбирает одно наилучшее (первичное) выравнивание для каждого прочтения.

Minimap2 изначально разработан для длинных чтений, но включает и режим для коротких прочтений (параметры -x sr и параметры -k27 -w14), который были использованы в наших экспериментах. В этом режиме параметры индексирования и выравнивания адаптированы под длину чтений ~150 пар оснований (“букв” из множества A, C, T, G), что соответствует типичным данным секвенирования платформ Illumina (парные прочтения 2×150 пар оснований). Все прочтения в тестовых наборах данных имели стандартную длину 150 пар оснований и выравнивались на референсный геном человека (сборка GRCh38) с использованием описанного алгоритма. Для оценки нового подхода мы сравнивали качество выравнивания minimap2 в обычном режиме (линейный референс) и с учетом дополнительных последовательностей, как описано ниже.

2.2 Выравнивание коротких прочтений с помощью minimap2

В рамках предлагаемого подхода референсная последовательность была расширена за счет добавления альтернативных контигов, содержащих известные варианты аллели. В качестве источника вариантов использована информация пангеномной базы HPRC (Human Pangenome Reference Consortium), в которой накоплены различные аллели, отсутствующие в линейном референсе GRCh38. Мы сконцентрировались на инсерциях и делециях (инделях) – вариантах типа вставок/делеций, которые особенно затрудняют выравнивание коротких прочтений на линейный эталон. Каждый такой вариант добавлялся в референс в виде отдельного контига: из последовательности основного референса выделялся небольшой фланкирующий участок вокруг позиции варианта, и в него либо вставлялась дополнительная последовательность (для инсерций), либо удалялся фрагмент (для делеций), таким образом формируя альтернативный аллель. Полученный контиг представляет собой альтернативный путь длиной в несколько сотен пар оснований, включающий данный вариант в контексте исходного генома (рис. 1).

Все сгенерированные альтернативные контиги были добавлены в референсный FASTA-файл как отдельные последовательности (аналогично тому, как альтернативные аллели представлены в сборке GRCh38). В названиях контигов указывалась родительская хромосома и идентификатор варианта, что позволяло отличить их от основных хромосом. После этого расширенный референс (включающий стандартные хромосомы и дополнительные контиги) индексировался инструментом minimap2 аналогично обычному геному. Выравнивание прочтений выполнялось одновременно на весь этот комбинированный референс. В результате для прочтений, содержащих известные индели, minimap2 мог находить выравнивания не только на основном эталоне, но и на добавленных альтернативных последовательностях.

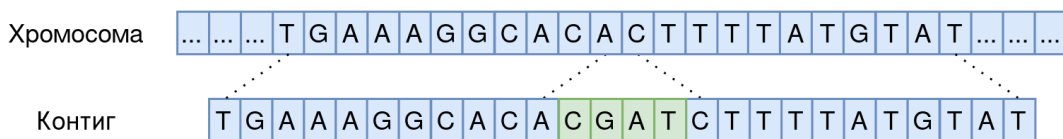


Рис. 1. Пример созданного контига с добавленной инсерцией (A → ACGAT).

Fig. 1. Example of a constructed contig with an added insertion (A → ACGAT).

Чтобы корректно обработать случаи множественного выравнивания, мы реализовали пост-обработку результатов выравнивания перед этапом лифтовера координат. Minimap2 с расширенным референсом нередко выдает два выравнивания для одного прочтения: одно – на основной хромосоме (с пропуском или разрывом в месте варианта), и второе – на альтернативном контиге, содержащем вставку или делецию. Мы сгруппировали такие

выравнивания по идентификатору прочтения и сравнивали качество (баллы выравнивания, MAPQ). Если альтернативный контиг давал более качественное выравнивание (меньше расхождений, выше выровненный процент и MAPQ), то выравнивание на основном референсе считалось субоптимальным и удалялось из файла. И наоборот, если прочтение лучше выровнялось на линейный референс, дополнительное выравнивание на контиг считалось избыточным. Таким образом, для каждого прочтения оставалось не более одного наилучшего выравнивания – фактически мы выбирали между референсным и альтернативным путем для прочтений с вариантом. Данный прием предотвращает снижение суммарной метрики MAPQ из-за наличия двух конкурирующих позиций выравнивания для одного и того же фрагмента.

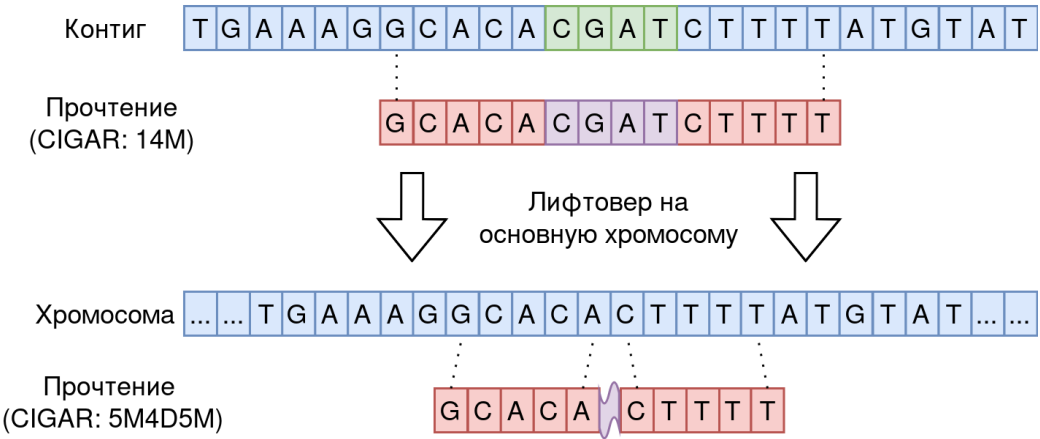


Рис. 2. Пример переноса выровненного прочтения на основную хромосому.

Прочтение содержащее инсерцию A → ACGAT выровнялось на созданный контиг.

Строка CIGAR 14M означает, что всё прочтение, состоящее из 14 п.о. выровнялось без вставок и делеций. После лифтовера прочтения на основную хромосому строка CIGAR стала 5M4D5M, что означает: 5 п.о. без вставок и делеций, 4 п.о. – делеция и снова 5 п.о. без вставок и делеций.

Fig. 2. Example of lifting over an aligned read to the primary chromosome.

The read containing the insertion A → ACGAT aligned to the constructed contig.

The CIGAR string 14M indicates that the entire read of 14 bp aligned with no insertions or deletions.

After liftover of the read to the primary chromosome, the CIGAR string became 5M4D5M, which means: 5 bp matching (no indels), a 4-bp deletion, and then another 5 bp matching (no indels).

После отбора для каждого прочтения единственного наилучшего выравнивания (референсного или альтернативного), следующим шагом было перевести эти выравнивания на координаты основного эталонного генома. Для этого мы воспользовались инструментом LevioSAM2, специально разработанным для высокоточного переноса выравниваний между различными версиями или сборками генома. В нашем случае инструмент LevioSAM2 применял подготовленный нами chain-файл (файл цепочек), описывающий соответствие между последовательностями добавленных альтернативных контигов и участком основной хромосомы, чтобы корректно пересчитать координаты выравниваний. Важно отметить, что инструмент LevioSAM2 не просто смещает координаты: он обновляет всю информацию о выравнивании (название референса, позицию, флаги, CIGAR-строку и т.д.) в соответствии с целевым референсом. В результате выравнивания прочтений, изначально полученные на дополнительных контигах, были преобразованы в эквивалентные выравнивания на линейном референсном геноме (рис 2.).

Хотя LevioSAM2 выполняет перенос координат достаточно точно, в некоторых случаях оставались мелкие неточности выравнивания. Например, если прочтение содержало индель немного отличающийся от представленного на контиге, то после лифтовера его CIGAR-

строка могла сместиться на пару позиций. Для дополнительной корректировки мы применили инструмент локального перевыравнивания ABRA2 (версия 2.24). ABRA2 выполняет локальную пересборку участков с предположительными инделями, что позволяет устранить систематические ошибки выравнивания вокруг инсерций/делетий. В контексте нашего подхода ABRA2 донастраивал выравнивания прочтений в тех местах, где лифтовер мог быть неточным, тем самым выравнивания окончательно приводились в согласованное состояние относительно основного референса.

После этапов лифтовера и перевыравнивания полученные выравнивания проходят стандартную обработку: с помощью samtools выполнялась сортировка по координатам, маркировка дубликатов PCR, и полученный BAM-файл выравнивания поступал в инструмент определения вариантов DeepVariant. Отметим, что за счет обратного переноса координат весь последующий анализ вариаций осуществлялся в системе координат GRCh38, как и в классическом подходе.

3. Результаты

Все эксперименты проводились с использованием эталонной человеческой геномной сборки HG002, взятой из проекта Genome in a Bottle (GIAB) [10], а также парные прочтения из FDA Precision Truth Challenge V2 [11]. Запуски осуществлялись с помощью конвейера nf-core/sarek (v3.4.0). В качестве источника вариантов для создания контигов была использована информация пангеномной базы HPRC, из которой были извлечены индели с необходимыми свойствами. Все запуски сравнивались с эталоном с помощью инструмента hap-py [12]. При эталонном запуске (таблица 1) на этапе выравнивания был применен minimap2 (v2.30 с параметрами -x sr -k 27 -w 14), на этапе вызова вариантов был использован DeepVariant (v1.9.0).

Табл. 1. Эталонный запуск на оригинальной последовательности GRCh38.

Table 1. Reference run on the original GRCh38 sequence.

Type	TOTAL	TP	FN	FP	Recall	Precision	F1_Score
INDEL	525469	522364	3105	1222	0.994091	0.997666	0.995875
SNP	3365127	3348351	16776	4444	0.995015	0.998675	0.996841

После модификации алгоритма выравнивания и создания дополненных геномных референсных последовательностей потребовалось перебрать большое количество комбинаций различных параметров для поиска оптимальных результатов. Первый из важных параметров – размер фланкирующего участка при создании контигов (расстояние влево и вправо от добавленного инделя), значения которого при экспериментах равнялись 100, 150, 200, 300, 500. Брать длину фланкирующего участка меньше 100 не имеет смысла, так как тогда короткое прочтение длиной 150 не сможет на него выровняться, а длина фланкирующего участка больше 500 создает слишком много «шума» на границах этого контига и увеличивает количество ложноположительных вариантов. Также при создании контигов можно учитывать длину самих инделей, так как очень длинные (или очень короткие) могут негативно влиять на результаты. Верхней границей было определено значение 100 (и оно остается таковым при всех запусках, которые будут указаны далее), так как контиги со слишком большими инделями сильно увеличивают ложноположительные варианты, нижней границей взято значение 10, что также уменьшает количество неверных результатов. И еще один из основных перебираемых параметров – это значение AF (Allele Frequency) – это частота альтернативной аллели в выборке, представленная числом от 0 до 1. Она показывает, какая доля всех аллелей в популяции (или наборе образцов) приходится на альтернативную аллель. Для улучшения точности результатов мы можем повышать это значение, чтобы брать только более часто встречающиеся индели.

Были выбраны значения длины инделей от 10 до 100, при этом взялись те индели, значение AF у которых больше 0.1. В дальнейших запусках было исследовано влияние длины фланкирующего участка при создании контигов.

Табл. 2. Результаты запуска на референсе GRCh38 с добавлением контигов. Длина фланкирующего участка 100. AF > 0.1. Индели длины от 10 до 100.

Table 2. Results of the run on the GRCh38 reference with added contigs. Flanking region length: 100 bp. AF > 0.1. Indels 10–100 bp long.

Type	TOTAL	TP	FN	FP	Recall	Precision	F1-Score
INDEL	525469	522515	2954	1228	0.994378	0.997655	0.996014
SNP	3365127	3348470	16657	4412	0.99505	0.998684	0.996864

Табл. 3. Результаты запуска на референсе GRCh38 с добавлением контигов. Длина фланкирующего участка 300. AF > 0.1. Индели длины от 10 до 100.

Table 3. Results of the run on the GRCh38 reference with added contigs. Flanking region length: 300 bp. AF > 0.1. Indels 10–100 bp long.

Type	TOTAL	TP	FN	FP	Recall	Precision	F1-Score
INDEL	525469	522540	2929	1313	0.994426	0.997493	0.995957
SNP	3365127	3348469	16658	4474	0.99505	0.998666	0.996855

Табл. 4. Результаты запуска на референсе GRCh38 с добавлением контигов. Длина фланкирующего участка 500. AF > 0.1. Индели длины от 10 до 100.

Table 4. Results of the run on the GRCh38 reference with added contigs. Flanking region length: 500 bp. AF > 0.1. Indels 10–100 bp long.

Type	TOTAL	TP	FN	FP	Recall	Precision	F1-Score
INDEL	525469	522536	2933	1341	0.994418	0.99744	0.995927
SNP	3365127	3348415	16712	4621	0.995034	0.998622	0.996825

Было получено (табл. 2-4), что на SNP количество верно найденных (TP) вариантов растет при каждом уменьшении длины фланкирующего участка, а количество ложноположительных (FP) уменьшается, что дает рост метрик Recall и Precision, и, как следствие, улучшение метрики F1. На инделях количество верно найденных вариантов немного уменьшается при снижении длины фланкирующего участка, но при это все равно остается выше, чем при эталонном запуске. Количество же ложно-положительных вариантов значительно падает при уменьшении длины фланкирующего участка, что дает хороший рост метрики F1. По результатам данных экспериментов размер окна был зафиксирован на значении 100, как лучший результат.

Далее были рассмотрены варианты границы значения AF >0.05, >0.1 (табл. 2), >0.15.

Табл. 5. Результаты запуска на референсе GRCh38 с добавлением контигов. Длина фланкирующего участка 100. AF > 0.05. Индели длины от 10 до 100.

Table 5. Results of the run on the GRCh38 reference with added contigs. Flanking region length: 100 bp. AF > 0.05. Indels 10–100 bp long.

Type	TOTAL	TP	FN	FP	Recall	Precision	F1-Score
INDEL	525469	522521	2948	1242	0.99439	0.997628	0.996007
SNP	3365127	3348464	16663	4411	0.995048	0.998685	0.996863

Табл. 6. Результаты запуска на референсе GRCh38 с добавлением контигов. Длина фланкирующего участка 100. $AF > 1.5$. Индели длины от 10 до 100.

Table 6. Results of the run on the GRCh38 reference with added contigs. Flanking region length: 100 bp. $AF > 0.15$. Indels 10–100 bp long.

Type	TOTAL	TP	FN	FP	Recall	Precision	F1-Score
INDEL	525469	522496	2973	1215	0.994342	0.99768	0.996008
SNP	3365127	3348471	16656	4411	0.99505	0.998685	0.996864

По результатам (табл. 2, 5, 6) было замечено, что оптимальным по метрике F1 является значение $AF > 0.1$, так как отклонение в большую сторону уменьшает количество верно определенных вариантов на инделях, а отклонение в меньшую сторону увеличивает количество ложно-положительных.

Также в качестве эксперимента (табл. 7) было проверено влияние изменения параметров minimap2 (-k 21 -w 11), однако это снижает количество верно определенных вариантов на снипах ниже эталонного запуска (табл. 1), с которым мы сравнивались, поэтому было оставлено предыдущее значение (-k 27 -w 14).

Табл. 7. Результаты запуска на референсе GRCh38 с добавлением контигов. Длина фланкирующего участка 100. $AF > 0.1$. Индели длины от 10 до 100. Параметры minimap2: -k 21 -w 11.

Table 7. Results of the run on the GRCh38 reference with added contigs. Flanking region length: 100 bp. $AF > 0.1$. Indels 10–100 bp long. Minimap2 parameters: -k 21 -w 11.

Type	TOTAL	TP	FN	FP	Recall	Precision	F1-Score
INDEL	525469	522508	2961	1213	0.994365	0.997684	0.996022
SNP	3365127	3347736	17391	4446	0.994832	0.998674	0.996749

После подбора оптимальных значений AF и длины фланкирующего участка был проведен эксперимент (табл. 8) с удалением нижнего ограничения на длину инделей (>10).

Табл. 8. Результаты запуска на референсе GRCh38 с добавлением контигов. Длина фланкирующего участка 100. $AF > 0.1$. Индели длины до 100.

Table 8. Results of the run on the GRCh38 reference with added contigs. Flanking region length: 100 bp. $AF > 0.1$. Indels less then 100 bp long.

Type	TOTAL	TP	FN	FP	Recall	Precision	F1-Score
INDEL	525469	522707	2762	1334	0.994744	0.997454	0.996097
SNP	3365127	3348441	16686	4486	0.995041	0.998662	0.996849

Данный набор параметров дает наибольшее количество верно определенных вариантов на инделях и, несмотря на рост ложноположительных вариантов, самый лучший по F1-метрике результат. Однако на снипах был получен рост ложноположительных и падение верно определенных результатов, что в сумме дает результат хуже, чем на запуске из табл. 2. После проведения различных экспериментов был сделан вывод, что по F1-метрике лучший результат на инделях дают параметры из таблицы 8, а для снипов – из табл. 2.

4. Заключение

Таким образом, предложенный метод выравнивания демонстрирует высокую эффективность на практике. Он обеспечивает существенный прирост качества выравнивания прочтений по сравнению с традиционными подходами, что особенно заметно в сложных для картирования регионах генома. Достигнутые результаты показывают, что метод может служить эффективной альтернативой графовым подходам к выравниванию прочтений WGS,

обеспечивая сопоставимое улучшение точности без значительного усложнения или увеличения вычислительной нагрузки.

В дальнейшем планируется провести оптимизацию производительности алгоритма для ускорения вычислений и уменьшения потребления ресурсов. Кроме того, мы намерены расширить набор учитываемых в выравнивании генетических вариантов, чтобы еще сильнее повысить полноту и точность покрытия переменных участков генома. Важным шагом станет также сравнительный анализ предложенного решения с современными инструментами выравнивания, такими как VG Giraffe, DRAGEN и другими. Эти шаги позволят всесторонне оценить преимущества метода и подтвердят его конкурентоспособность как надежного и эффективного инструмента для выравнивания прочтений при полногеномном секвенировании.

Определения терминов

Гаплотипы	версия генетической последовательности, унаследованная от одного родителя
Референсная последовательность	стандартная версия генома, используемая как эталон для сравнения
Выравнивание	процесс установления соответствия между позициями коротких последовательностей (прочтений) и референсной последовательности
Генетический вариант (SNP)	замена одного нуклеотида (буквы) в определенной позиции генома
Индель	вставка или удаление участка ДНК (по первым буквам от инсерция/делеция)
Контиг	непрерывный фрагмент последовательности
Лифтовер	процесс переноса выравниваний (текстовых последовательностей) на референсный геном
Лифтовер-цепочки	соответствие между последовательностями добавленных альтернативных контигов и участком основной хромосомы для корректного пересчета координат выравниваний
Минимайзеры	наиболее лексикографически малые подстроки в скользящем окне фиксированной длины

Список литературы / References

[1]. Halldorsson, B. V., Eggertsson, H. P., Moore, K. H., Hauswedell, H., Eiriksson, O., Ulfarsson, M. O., ... & Stefansson, K. (2022). The sequences of 150,119 genomes in the UK Biobank. *Nature*, 607(7920), 732-740.

[2]. Liao, W. W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., ... & Paten, B. (2023). A draft human pangenome reference. *Nature*, 617(7960), 312-324.

[3]. Genomes Project Consortium. A global reference for human genetic variation. *Nature*. 2015;526(7571):68.

[4]. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18), 3094-3100.

[5]. Chaisson, M. J., Sanders, A. D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., ... & Lee, C. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature communications*, 10(1), 1784.

[6]. Sirén, J., Monlong, J., Chang, X., Novak, A. M., Eizenga, J. M., Markello, C., ... & Paten, B. (2021). Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science*, 374(6574), abg8871.

- [7]. Illumina. (n.d.). ALT-aware mapping. Illumina DRAGEN Bio-IT Platform Documentation (v3.7). Available at: https://support.illumina.com/content/dam/illumina-support/help/Illumina_DRAGEN_Bio_IT_Platform_v3_7_1000000141465/Content/SW/Informatics/Dragen/GPipelineAltMap_fDG.html, accessed 12.11.2025.
- [8]. Mun, T., Chen, N. C., & Langmead, B. (2021). LevioSAM: fast lift-over of variant-aware reference alignments. *Bioinformatics*, 37(22), 4243-4245.
- [9]. Mose, L. E., Perou, C. M., & Parker, J. S. (2019). Improved indel detection in DNA and RNA via realignment with ABRA2. *Bioinformatics*, 35(17), 2966-2973.
- [10]. National Institute of Standards and Technology (NIST). (n.d.). Genome in a Bottle. NIST. Available at: <https://www.nist.gov/programs-projects/genome-bottle>, accessed 13.11.2025.
- [11]. Olson, N. D., Wagner, J., McDaniel, J., Stephens, S. H., Westreich, S. T., Prasanna, A. G., ... & Zook, J. M. (2022). PrecisionFDA Truth Challenge V2: Calling variants from short and long reads in difficult-to-map regions. *Cell genomics*, 2(5).
- [12]. Krusche, P., Trigg, L., Boutros, P. C., Mason, C. E., De La Vega, F. M., Moore, B. L., ... & Global Alliance for Genomics and Health Benchmarking Team. (2019). Best practices for benchmarking germline small-variant calls in human genomes. *Nature biotechnology*, 37(5), 555-560.

Информация об авторах / Information about authors

Никита Артемович КОЛТУНОВ – лаборант Федерального государственного бюджетного учреждения науки Институт системного программирования им. В.П. Иванникова Российской академии наук, специалист в области биоинформатики.

Nikita Artemovich KOLTUNOV – Laboratory assistant at the Ivannikov Institute for System Programming of the Russian Academy of Sciences, a specialist in bioinformatics.

Егор Павлович ГУГУЧКИН – аспирант Федерального государственного бюджетного учреждения науки Институт системного программирования им. В.П. Иванникова Российской академии наук, специалист в области биоинформатики.

Egor Pavlovich GUGUCHKIN – postgraduate student at the Ivannikov Institute for System Programming of the Russian Academy of Sciences, a specialist in bioinformatics.

Евгений Андреевич КАРПУЛЕВИЧ – кандидат физико-математических наук, научный сотрудник Института системного программирования им. В.П. Иванникова Российской академии наук, специалист в области биоинформатики.

Evgeny Andreevich KARPULEVICH – Cand. Sci. (Phys.-Math.), Ivannikov Institute for System Programming of the Russian Academy of Sciences.

