DOI: 10.15514/ISPRAS-2023-35(2)-17

Модификация алгоритма выравнивания коротких прочтений для повышения качества пайплайна обработки данных полногеномного секвенирования человека

[1] *Е.П.Гугучкин, ORCID: 0000-0001-7885-9892 <guguchkin@ispras.ru>*
[1] *Е.А.Карпулевич, ORCID: 0000-0002-6771-2163 <karpulevich@ispras.ru>*
[1] *Институт системного программирования РАН,*
109004, Россия, г. Москва, ул. А. Солженицына, д. 25.

Аннотация. Данное исследование подчеркивает важность выравнивания коротких прочтений (ридов) в анализе данных полногеномного секвенирования человека. Процесс выравнивания состоит в определении позиций коротких генетических последовательностей относительно заранее известной референсной последовательности генома человека. Традиционные методы выравнивания используют линейную референсную последовательность, но это может привести к некорректному выравниванию, особенно если в ридах присутствуют генетические варианты. В данной работе была проведена модификация индексного файла референсной последовательности инструмента minimap2. В результате экспериментов было показано, что добавление в индекс инструмента minimap2 информации о часто встречающихся генетических вариантах приводит к повышению количества верно выявленных генетических вариантов, что влияет на качество последующего анализа данных.

Ключевые слова: конвейер обработки данных, секвенирование ДНК, вычислительная биология, методы выравнивания последовательностей, анализ данных NGS, вычислительные методы

Для цитирования: Гугучкин Е.П., Карпулевич Е.А. Модификация алгоритма выравнивания коротких прочтений для повышения качества пайплайна обработки данных полногеномного секвенирования человека. Труды ИСП РАН, том 35, вып. 2, 2023 г., стр. 235–248. DOI: 10.15514/ISPRAS–2023–35(2)–17

Modification of the short read alignment algorithm to improve the quality of the human whole genome sequencing data processing pipeline

[1] *E.P. Guguchkin ORCID: 0000-0001-7885-9892 <guguchkin@ispras.ru>*
[1] *E.A. Karpulevich ORCID: 0000-0002-6771-2163 <karpulevich@ispras.ru>*
[1] *Institute for System Programming of the Russian Academy of Sciences,*
25, Alexander Solzhenitsyn st., Moscow, 109004, Russia.

Abstract. This study emphasizes the importance of aligning short reads in the analysis of human whole-genome sequencing data. The alignment process involves determining the positions of short genetic sequences relative to a known reference genome sequence of the human genome. Traditional alignment methods use a linear reference sequence, but this can lead to incorrect alignment, especially when short reads contain genetic variations. In this work, the index file of the reference sequence was modified using the minimap2 tool. Experimental results showed that adding information about frequently occurring genetic variations to the

minimap2 index increases the number of correctly identified genetic variants, which affects the quality of subsequent data analysis.

Keywords: data processing pipeline, DNA sequencing, Computational biology, Sequence alignment methods, NGS data analysis, Computational methods

For citation: Guguchkin E.P., Karpulevich E.A. Modification of the short read alignment algorithm to improve the quality of the human whole genome sequencing data processing pipeline. *Trudy ISP RAN/Proc. ISP RAS*, vol. 35, issue 2, 2023. pp. 235-248 (in Russian). DOI: 10.15514/ISPRAS-2023-35(2)-17

1. Введение

Геном человека представляет собой последовательность четырех нуклеотидных оснований, обозначаемых как A(аденин), C(цитозин), G(гуанин) и T(тимин). В последовательности генома человека содержится более трех миллиардов нуклеотидов. В геноме содержится генетическая информация, необходимая для поддержания правильной работы живого организма. Исследования генома имеют множество научных и практических применений, включая идентификацию генетических вариантов, связанных с болезнями [1], разработку персонализированных лекарств, исследование эволюционных процессов и биологического разнообразия. Современные технологии в области биологии, такие как редактирование генома и синтетическая биология, появившиеся в результате развития исследований в области генетики, обладают большим потенциалом в развитии медицины, сельского хозяйства и промышленности [2][3][4]. Полногеномное секвенирование (Whole Genome Sequencing) — это технология определения полной последовательности ДНК генома организма [5].

Процесс полногеномного секвенирования происходит в несколько этапов: образец начиная с лабораторного процесса заканчивая вычислительным анализом полученных данных. ДНК извлекается из биологического образца, такого как кровь или ткань, а затем фрагментируется и подготавливается для проведения секвенирования. Затем фрагменты ДНК секвенируют с использованием технологий высокопроизводительного секвенирования. Результатом процесса секвенирования следующего поколения (Next-Generation Sequencing, NGS) [6] является файл, который содержит миллионы коротких подпоследовательностей ДНК, называемых ридами или короткими прочтениями.

Длина рида зависит от конкретной технологии секвенирования, но обычно составляет от 50 до 250 нуклеотидов. Также в зависимости от технологии риды могут быть одиночными (single-end) или парными (paired-end). Одиночные риды — это риды, которые генерируются только с одного конца секвенируемого фрагмента ДНК. Риды с парными концами являются результатом секвенирования обоих концов фрагмента ДНК.

Анализ данных полногеномного секвенирования состоит из следующих основных этапов [7]: контроль качества, выравнивание ридов на референсный геном (известная последовательность нуклеотидов, которая представляет собой образец генома какого-либо вида), и поиск генетических вариантов (отличий генома секвенируемого организма от референса). Риды, полученные в результате полногеномноного секвенирования проходят проверку качества, в результате которой риды с низким качеством могут быть удалены или обрезаны по краям для повышения точности последующего анализа. После этого, риды выравниваются (Рис. 1) на последовательность из референсного генома: каждой последовательности из рида ставится в соответствие некоторая подпоследовательность из референсной последовательности. Для выровненных ридов происходит поиск несоответствий между последовательностью из ридов и референсной последовательностью, таким образом происходит поиск генетических вариантов. Генетические варианты могут по-разному влиять на характеристики человека такие, как физические черты, риск заболевания или реакция на лекарство. Генетические варианты (Рис. 2) бывают нескольких видов: однонуклеотидные полиморфизмы, инсерции(вставки) и делеции. Однонуклеотидный

полиморфизм (SNP) — это тип генетического варианта, в котором происходит замена одного нуклеотидного основания в определенной позиции последовательности ДНК на другое. SNP является наиболее распространенным типом генетических вариантов в геноме человека и встречается примерно один раз на каждые 300 нуклеотидов. Индел (сокращение от инсерция-делеция) — это тип генетического варианта, в котором происходит вставка или удаление одного или нескольких нуклеотидов в последовательности ДНК относительно референсной. Инделы могут иметь размер от одного до нескольких тысяч нуклеотидов и могут встречаться в любом месте генома.

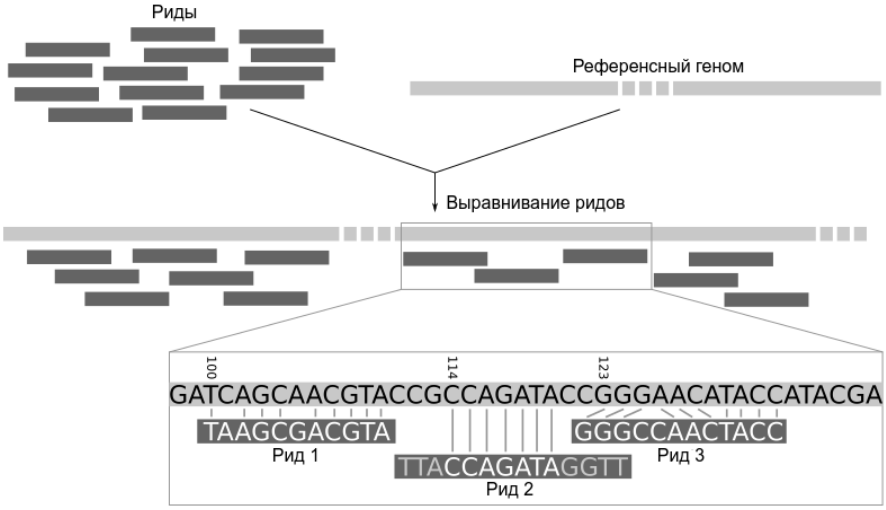

Рис. 1. Выравнивание ридов
Fig. 1 Alignment of short reads

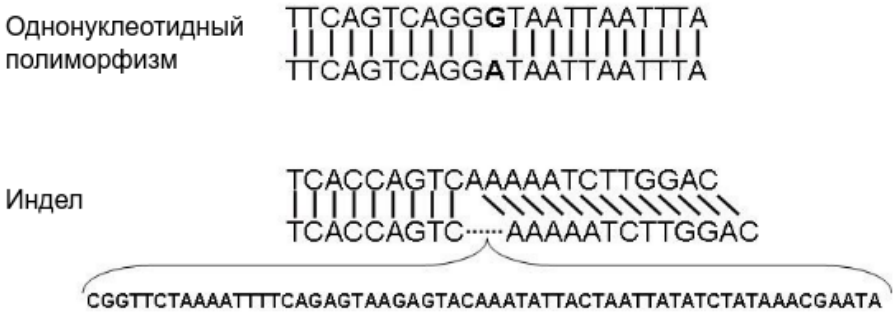

Рис. 2. Однонуклеотидные полиморфизмы и инделы
Fig. 2. SNPs and indels

Генетический вариант может не быть корректно обнаружен, если риды, содержащие этот генетический вариант, не будет выровнен в нужное место на этапе выравнивания. При этом риды, содержащие генетические варианты, если и будут выровнены, то оценка качества выравнивания будет снижена, что также может повлиять на точность нахождения вариантов [8]. В данной работе предложена модификация алгоритма выравнивания ридов для уменьшения количества ложно отрицательных генетических вариантов, с целью снизить негативный эффект, накладываемый при выравнивании ридов с генетическими вариантами.

2. Обзор методов выравнивания ридов на референс

В данном разделе рассмотрены основные методы выравнивания ридов на референсную последовательность. Для точного выравнивания строк существуют такие алгоритмы как алгоритм Смита — Уотермана [9] и алгоритм Нидлмана — Вунша [10]. Различные модификации этих алгоритмов используются в большинстве инструментов выравнивания, однако при выравнивании каждого рида на референсную последовательность длиной несколько миллиардов букв точными алгоритмами выравнивания неэффективно и для ускорения процедуры выравнивания используются дополнительные алгоритмы (на основе методов Seed-and-Extend или Seed-Chain-Align), которые позволяют найти предполагаемые позиции рида в геноме и уже после этого применить алгоритмы точного выравнивания. Помимо алгоритмов выравнивания на линейный референс также существует ряд алгоритмов выравнивания на граф [11], однако они значительно проигрывают по скорости приведенным выше алгоритмам.

2.1 Метод Seed-and-Extend

Основная идея метода Seed-and-Extend заключается в том, чтобы сперва найти короткие точные соответствия (якоря, seeds) между последовательностями ридов и референсной последовательностью, а затем продлить найденные якоря до рида для выполнения выравнивания. Можно выделить следующие этапы данного метода (Рис. 3):

1. **Создание индекса референсной последовательности.** На данном этапе создается индекс референсной последовательности, позволяющий быстро находить определенные регионы последовательности. Примером алгоритма, реализующим создание индекса, является преобразование референса с помощью алгоритма Барроуза — Уилера [12] в суффиксный массив [13].
2. **Вычисление якорей.** Последовательность из рида разбивается на k-меры (подстроки длины k). После этого происходит поиск каждого k-мера в индексе референсной последовательности. В случае если для k-мера из рида существует точное совпадение в референсной последовательности, то такой k-мер становится якорем.
3. **Продление якорей.** Каждый якорь расширяется в обоих направлениях вдоль последовательности из рида и референсной последовательности с использованием алгоритмов динамического программирования (модификации алгоритма Смита — Ватермана), которые ищут лучшее соответствие, начиная с каждой позиции якоря. Расширение проводится до конца последовательности из рида, либо конца референсной последовательности. При оценке выравнивания учитывается такие факторы как несовпадение отдельных элементов последовательностей и индели.
4. **Уточнение выравнивания.** После того, как было выполнено базовое выравнивание, оно может быть дополнительно уточнено для повышения его точности. Уточнение может включать в себя дополнительные итерации выравнивания, а также учет дополнительной информации, например, информации о потенциальных ошибках секвенирования.

5. **Фильтрация.** Данный этап включает в себя фильтрацию выровненных ридов с низким качеством и удаление неоднозначно выровненных ридов.

Метод выравнивания Seed-and-Extend реализован в таких инструментах, как Bowtie2 [14] и BWA-MEM [15].

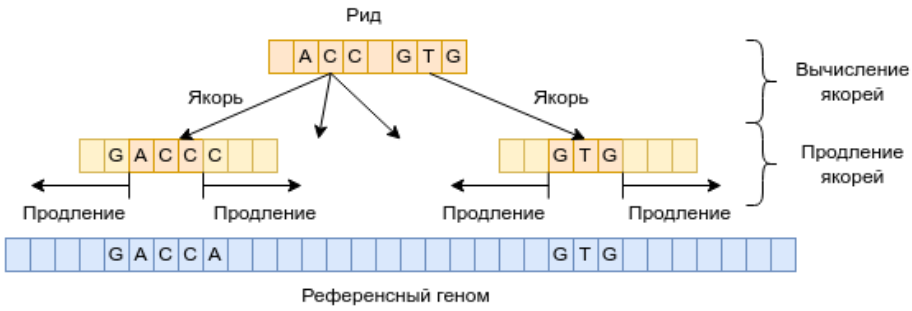

Рис. 3. Выравнивание ридов методом Seed-and-Extend.
Fig. 3. Read alignment using the Seed-and-Extend method.

2.2 Метод Seed-Chain-Align

Несмотря на то, что данный метод имеет определенные сходства с методом seed-and-extend, в нем также имеется ряд отличий, влияющих на качество и производительность. Важным отличием метода Seed-Chain-Align от метода Seed-and-Extend, является использование в процессе выравнивания ридов более коротких k-меров.

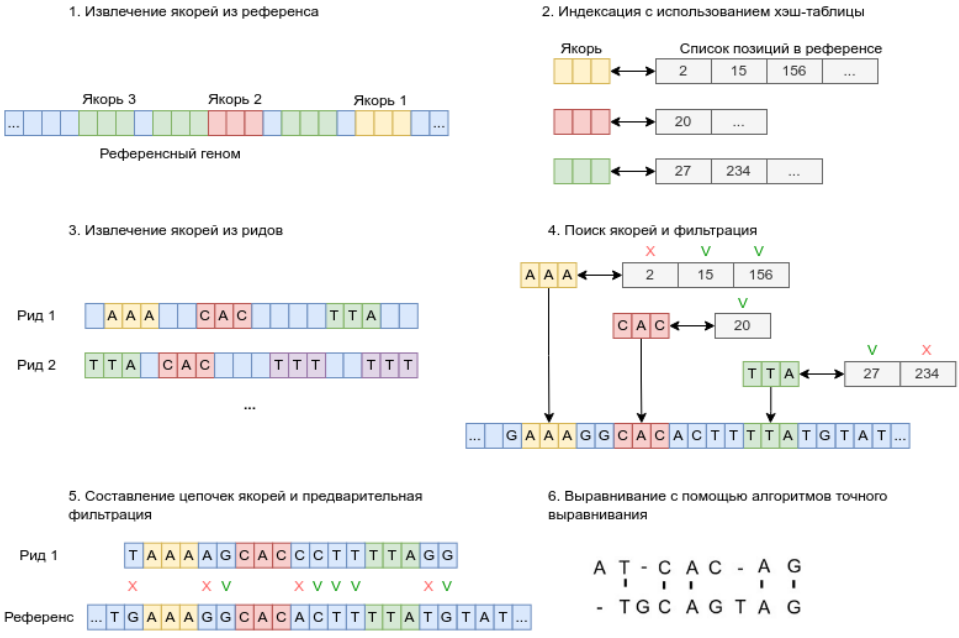

Рис. 4. Выравнивание ридов методом Seed-Chain-Align.
Fig. 4. Read alignment using the Seed-Chain-Align method.

Рассмотрим основные этапы метода Seed-Chain-Align (Рис. 4):

1. **Извлечение якорей из референса.** На данном этапе аналогично методу Seed-and-Extend вычисляются все якоря в референсной последовательности.
2. **Индексация референса.** На данном этапе создается индекс референсной последовательности, вычисленные якоря и их позиции заносятся в хэш-таблицу.
3. **Извлечение якорей из ридов.** На данном этапе вычисляются якоря для ридов.
4. **Поиск якорей и фильтрация.** Выполняется поиск позиций якорей ридов в референсе и фильтрация по различным критериям, например, фильтрация наиболее распространенных якорей
5. **Составление цепочек якорей и предварительная фильтрация.** Для нахождения наиболее точного выравнивания вычисляется оценка оптимальной цепочки якорей. Алгоритм сначала идентифицирует все возможные цепочки совпадающих k-меров между последовательностями из прочтений и референсной последовательности, а затем вычисляет оценку для каждой цепочки на основе количества совпадающих k-меров и размера разрывов между ними. После этого алгоритм выбирает цепочку с наибольшей оценкой в качестве основной.
6. **Выравнивание с помощью алгоритмов точного выравнивания.** На данном этапе проводится глобальное выравнивание между полученными последовательностями из якорей на референсной последовательности и якорей последовательности из рида. Глобальное выравнивание может производится, например, с помощью алгоритма Нидлмана — Вунша.

Метод Seed-Chain-Align реализован в инструменте выравнивания minimap2 [16]. В случае с инструментов minimap2 в качестве якорей используются так называемые минимизаторы. Минимизатор — это короткая подстрока длины k, которая является лексикографически минимальной строкой в окне w (Рис. 5).

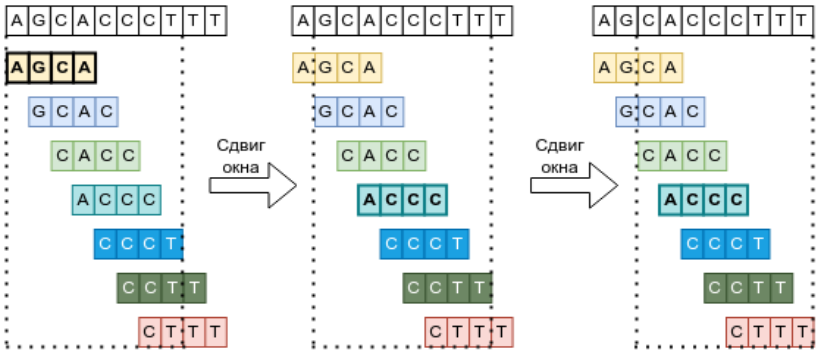

Рис. 5. Поиск минимизаторов длины k=4 в окне длины w=5.
Fig. 5. Finding minimizers of length k=4 in a window of length w=5.

2.3 Выводы

Оба метода (Seed-Chain-Align и Seed-and-Extend) выполняют дополнительные шаги обработки последовательностей перед их точным выравниванием. Данные методы значительно ускоряют выравнивание ридов на линейный референс, однако в случае если в якоре рида который прочитан с определенной позиции в геноме присутствует генетический вариант — данный рид может ошибочно выровняться в другую позицию в геноме или выровняться в правильную позицию, но с меньшим качеством. Если дополнительно учитывать якоря, содержащие генетические варианты, то возможно получится увеличить качество существующего алгоритма выравнивания. Причем индекс в виде хэш-таблицы, в

отличие от индекса в формате суффиксного массива, удобно модифицировать добавлением в таблицу новых позиций якорей с генетическими вариантами.

3. Исследование и построение решения задачи

Выше были рассмотрены основные методы выравнивания ридов и была предложена идея провести модификацию алгоритма выравнивания метода Seed-Chain-Align, реализованного в инструменте выравнивания minimap2.

В данном разделе будут рассмотрены шаги, необходимые для проведения модификации алгоритма выравнивания данных полногеномного секвенирования человека. Помимо самой модификации необходимо провести эксперименты для оценки качества результатов, полученных в результате применения разработанной модификации.

3.1 Модификация алгоритма выравнивания minimap2

Модификация алгоритма выравнивания (Рис. 6) заключается в том что на этапе поиска якорей minimap2 идентифицирует короткие последовательности — минимизаторы в ридах и референсе с добавлением генетических вариантов. В данной работе модификация выполняется добавлением однонуклеотидных полиморфизмов (SNP) в референсную последовательность.На этапе поиска оптимальных цепочек minimap2 учитывает наличие генетических вариантов таких как SNP в референсной последовательности. В частности, minimap2 использует систему подсчета очков, которая учитывает вероятность данного выравнивания на основе наличия генетических вариантов. Если рид содержит генетический вариант, отсутствующий в референсном геноме, minimap2 все еще может выровнять данный рид на референсную последовательность, допуская разрывы или несоответствия в выравнивании. Однако наличие генетического варианта может негативно повлиять на итоговую оценку качества выравнивания, что в дальнейшем приведет к ухудшению поиска генетических вариантов в данном регионе последовательности. Особенно заметным будет эффект, в случае если генетический вариант будет расположен в позиции потенциального якоря, что приведет к ухудшению оценки качества составления цепочки и возможному выбору другой цепочки вместо необходимой.

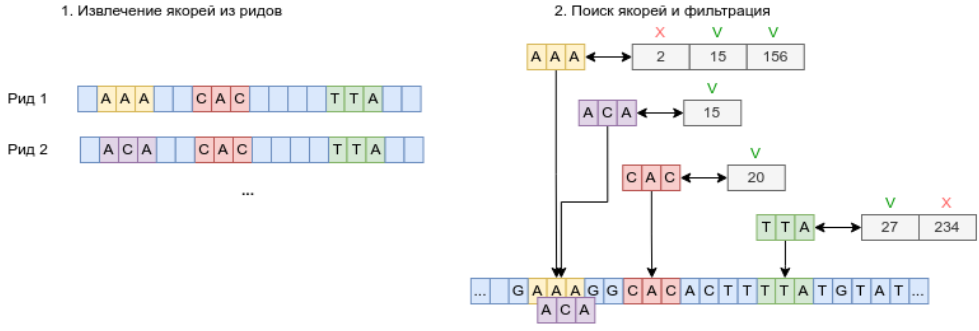

Рис. 6. Поиск якорей в модифицированном индексе
Fig. 6. Finding seeds in a modified index

Добавление часто встречающихся генетических вариантов в референсную последовательность может быть реализовано с помощью дополнительной модификации индекса референсной последовательности, содержащий хэш-таблицу из результатов хэш-функции от минимизаторов и их позиций в референсной последовательности.

В исходном коде инструмента minimap2 содержится функция mm_sketch (файл sketch.c), которая принимает на вход последовательность нуклеотидов и строит все возможные

минимизаторы длины k в окне длины w. Функция возвращает массив, содержащий пары ключ-значение: соответствующие результаты хэш-функции от последовательности и ее позиции в последовательности. Организация хэш-таблицы реализована с использованием библиотеки klib.

Для модификации индекса необходимо подготовить короткие подпоследовательности референсной последовательности, добавить в них генетический вариант (в случае добавления SNP как в данной работе — заменить символ), заново вычислить минимизаторы в модифицированной последовательности, сопоставить позиции в подпоследовательностях и референсной последовательности и, с помощью методов из klib, дополнить существующую хэш-таблицу новыми минимизаторами. В результате, два рида, последовательность одного из которых не содержит генетический вариант в якоре, а второго содержит, будут выравниваться в одинаковую позицию референсной последовательности и будут иметь одинаковую оценку качества выравнивания.

3.2 План экспериментов анализа данных полногеномного секвенирования человека

Для оценки эффекта, получаемого при модификации индекса, необходимо провести ряд экспериментов анализа данных полногеномного секвенирования. Это подразумевает проведение нескольких различных этапов, а именно: выравнивание ридов, поиск дубликатов, вычисление таблицы повторной калибровки, поиск генетических вариантов. Все этапы стоит провести в соответствии с рекомендациями GATK Best Practices [17].

В качестве первого эксперимента планируется взять несколько наборов реальных данных, полученных с помощью технологии секвенирования следующего поколения (NGS). Необходимо взять данные с разным покрытием для одного образца. Под покрытием подразумевается среднее количество ридов для каждого участка геномной последовательности. Разное покрытие позволит судить об эффекте, полученном при модификации индекса.

Соответственно, для каждого набора данных необходимо провести две итерации анализа данных полногеномного секвенирования:

1) Анализ данных полногеномного секвенирования, при котором на этапе выравнивания ридов используется немодифицированный индекс референсной последовательности;
2) Анализ данных полногеномного секвенирования, в котором используется модифицированный индекс с добавлением однонуклеотидных полиморфизмов;

Во втором эксперименте планируется использовать несколько наборов синтетических данных. Синтетические риды — это сгенерированные компьютером последовательности, которые предназначены для имитации характеристик реальных ридов, включая длину считывания, частоту ошибок секвенирования и наличие генетической изменчивости. Аналогично первому эксперименту планируется провести несколько итераций анализа данных полногеномного секвенирования.

4. Описание практической части

4.1 Описание выполненной модификации

Для модификации алгоритма выравнивания была написана функция mm_idx_manipulate, которая вызывается после создания индекса референсной последовательности. Данная функция считывает строки из указанного VCF файла [18] и для каждого генетического варианта(SNP), записанного в отдельной строке, проводятся следующие шаги:

1. В отдельные переменные записываются следующие данные: номер хромосомы, позиция на хромосоме, значение нуклеотидов в референсной последовательности и генетический вариант.
2. Из референсной последовательности (которая также хранится в индексном файле) считывается 2 * (k + w) — 1 символов последовательности, где центральный символ — это генетический вариант.
3. Вызывается функция mm_sketch, которая вычисляет минимизаторы полученной последовательности.
4. Полученные минимизаторы фильтруются, у оставшихся минимизаторов корректируется номера позиции в последовательности.
5. Вызывается написанная функция mm_idx_push, которая добавляет новые минимизаторы в индексную хэш-таблицу.

Полученная модификация имеет линейную вычислительную сложность с незначительными дополнительными затратами памяти порядка O(n) при вычислении каждой генетического варианта. Затраты на добавление элементов в хэш-таблицу можно оценить в среднем O(1) по времени и O(n) по памяти. Время работы алгоритма выравнивания и затраты память при модификации индекса увеличиваются на o(m), где m — длина референсной последовательности. Также, стоит добавить, что вычислять индекс референсной последовательности нужно всего один раз.

4.2 Обработка и обзор данных

Референсный геном

В ходе данной работы использовалась геномная референсная последовательность GRCh38, выпущенная Genome Reference Consortium в 2013 году [19]. Данная референсная последовательность представлена 23-мя хромосомами (от 1-ой до 22-ой, а также X-хромосома) и состоит приблизительно из 3 миллиардов нуклеотидов.

Данные о генетических вариантах в популяции

Данные о генетических вариантах были взяты из проекта 1000 Genomes Project [20]. Данный проект содержит сведения о генетических вариантах для 2504 человек из 26 популяций. Данные представлены в формате VCF, который является стандартным файловым форматом для анализа генетических данных.

Для обработки этих данных использовался инструмент bcftools, который является частью пакета samtools [21]. С его помощью были отфильтрованы генетические варианты, которые встречаются у достаточного числа представителей в популяции (minor allele frequency > 0.05), а также использовался для удаления генетических вариантов тех людей, для которых проводился анализ данных полногеномного секвенирования человека. Помимо этого, bcftools был использован для выделения конкретной популяции из VCF файла.

Данные о ридах

Данные с ридами были взяты из The precisionFDA Truth Challenge — испытании, целью которого было улучшение качества результатов анализа данных полногеномного секвенирования [22]. А также данные, собранные организацией National Institute of Standards and Technology. Датасет состоял из пар файлов с ридами в формате FASTQ [23] для 2 человек: NA12878, NA24631 (сын из набора данных Chinese Trio). Для проведения дальнейших экспериментов были использованы разные наборы данных, отличающихся разным покрытием — средним количеством ридов на всю референсную геномную последовательность. Были использованы данные с покрытием 30X и 60X.

“Уверенные” регионы и эталонные генетические варианты

Из The precisionFDA Truth Challenge были взяты данные об “уверенных” регионах для каждого человека. Эти регионы использовались во время всего пайплайна анализа данных полногеномного секвенирования, а также использовались при валидации с эталонными файлами генетических вариантов (которые были взяты из The precisionFDA Truth Challenge). Стоит отметить, что для уменьшения влияния краевого эффекта на концах “уверенных” регионов, эти регионы были расширены по краям на 1000 нуклеотидов. Расширенные “уверенные” регионы использовались только во время подсчета, во время валидации использовались изначальные версии. Для расширения регионов был написан скрипт interval_expander.py. А для последующей конкатенации расширенных пересекающихся регионов — скрипт interval_concater.py.

Синтетические данные

Для генерации синтетических данным случайным образом было выбрано 10 образцов из проекта 1000 Genomes Project. Были выбраны образцы: HG00472, HG01205, HG01578, HG01680, HG02252, HG02301, HG02536, HG02603, HG03485, HG03520. Аналогично реальным данным создавались наборы парный ридов с длиной каждой последовательности из рида в 100 нуклеотидов. Данные создавались с покрытием 26X. При создании синтетических данных использовались инструменты bcftools и инструмент ART [24]. bcftools использовался для создания геномной последовательности, содержащей генетические варианты конкретного образца, по которым инструмент ART создавал наборы парных ридов.

4.3 Метрика оценки качества

Генетические варианты, полученные в результате анализа, можно разделить на три категории:

- найденные верно генетические варианты
- не найденные генетические варианты
- найденные неверно генетические варианты

Для фильтрации FP, существуют различные инструменты, которые не рассматривались в данной работе. Поскольку для полногеномного секвенирования не существует true negative данных, то было решено использовать метрику recall: RECALL = TP / (TP + FN)

Для сравнения полученных VCF файлов, содержащих генетические варианты, использовался инструмент hap.py [25]. hap.py — инструмент для сравнения диплоидных генотипов на уровне гаплотипов. Вместо того, чтобы сравнивать записи VCF построчно, hap.py генерирует и сравнивает последовательности в суперлокусе. Суперлокус — это небольшой участок генома (размером от 1 до 1000 элементов), который содержит один или несколько генетических вариантов.

4.4 Проведенные эксперименты

Для проверки эффективности модификации алгоритма выравнивания был проведен ряд экспериментов, описанных ниже.

Были выполнены два эксперимента:

1. **Стандартный**: с использованием стандартного индексного файла референсного генома, генерируемый инструментом minimap2
2. **Модифицированный**: с использованием модифицированного индексного файла, в который были добавлены однонуклеотидные полиморфизмы из базы 1000 Genomes Project, превышающие порог MAF>0.05

NA12878 30X

Риды для NA12878 с 30-кратным покрытием генома и с расширением “уверенных” регионов на 100 нуклеотидов для SNP(Табл. 1).

Табл. 1. Результаты полногеномного секвенирования для образца NA12878. Покрытие 30X, SNP.
Table 1. Whole genome sequencing results for sample NA12878. Coverage 30X, SNP.

Эксперимент	TP	FN	FP	Recall	Precision	F1
Стандартный	3206999	47387	15190	0.985439	0.995285	0.990338
Модифицированный	3210000	44386	15681	0.986361	0.995138	0.990730

NA12878 60X
Риды для NA12878 с 60-кратным покрытием генома и с расширением “уверенных” регионов на 100 нуклеотидов для SNP(Табл. 2).

Табл. 2. Результаты полногеномного секвенирования для образца NA12878. Покрытие 60X, SNP.
Table 2. Whole genome sequencing results for sample NA12878. Coverage 60X, SNP.

Эксперимент	TP	FN	FP	Recall	Precision	F1
Стандартный	3218108	36278	16990	0.988853	0.994748	0.991791
Модифицированный	3219938	34448	17577	0.989415	0.994570	0.991986

NA24631 30X
Риды для NA24631 (сын из ChineseTrio) с 30-кратным покрытием генома и с расширением “уверенных регионов” на 100 нуклеотидов для SNP(Табл.3).

Табл. 3. Результаты полногеномного секвенирования для образца NA24631. Покрытие 30X, SNP.
Table 3. Whole genome sequencing results for sample NA24631. Coverage 30X, SNP.

Эксперимент	TP	FN	FP	Recall	Precision	F1
Стандартный	3218619	57012	15758	0.982595	0.995128	0.988822
Модифицированный	3221458	54173	16406	0.983462	0.994933	0.989164

В результате проведенных экспериментов было показано, что при добавлении только однонуклеотидных полиморфизмов достигается наибольшее количество истинно положительных генетических вариантов, однако, в таком случае также возрастает количество ложно положительных генетических вариантов.

Эксперимент с синтетическими данными
Для синтетических данных было проведено по две итерации анализа данных полногеномного секвенирования: анализа данных, в котором на этапе выравнивания использовался стандартный индекс референсной последовательности и модифицированный индекс для SNP(Табл. 4). В ходе данного эксперимента в индекс были добавлены лишь однонуклеотидные полиморфизмы. При этом использовался одинаковый набор полиморфизмов, совпадающий с наборами в экспериментах для NA12878 и NA24631.

Табл. 4. Результаты полногеномного секвенирования синтетических данных. Покрытие 26X, SNP.
Table 4. Results of whole genome sequencing of synthetic data. Coverage 26X, SNP.

Эксперимент	Recall with default index	Recall with modified index
HG00472	0.989396	0.990893
HG01205	0.989052	0.990597

HG01578	0.989461	0.990957
HG01680	0.989332	0.990890
HG02252	0.989665	0.991100
HG02301	0.989426	0.990878
HG02536	0.989446	0.990892
HG02603	0.989307	0.990850
HG03485	0.989463	0.990870
HG03520	0.989239	0.990716

Для случайно выбранных сгенерированных образцов было установлено, что модификация индекса референсной последовательности, также как и в остальных экспериментах, приводит к увеличению итоговой метрики recall.

Заключение

Реализована функция, позволяющая обрабатывать генетические варианты из указанного VCF файла и внедрять их в индекс референсной последовательности, что приводит к улучшению точности выравнивания. Результаты проведенных экспериментов на NA12878 и NA24631 указывают на важность и эффективность алгоритма модификации индекса при решении задачи анализа данных полногеномного секвенирования человека. Использование однонуклеотидных полиморфизмов при добавлении генетических вариантов показало наилучшие результаты в отношении истинно положительных генетических вариантов. Однако, следует отметить, что это также приводит к увеличению количества ложно положительных генетических вариантов. Кроме того, эксперименты с случайно выбранными сгенерированными образцами показали, что модификация индекса референсной последовательности также приводит к повышению метрики recall, что подтверждает улучшение способности обнаружения генетических вариантов с использованием данного алгоритма. Дальнейшие исследования и оптимизации могут помочь сделать алгоритм еще более эффективным для практического применения в анализе геномных данных.

Список литературы / References

[1]. Bagyinszky, E., Youn, Y. C., An, S. S. A., & Kim, S. (2014). The genetics of Alzheimer's disease. Clinical interventions in aging, 535-551.

[2]. Fisher, R. A. (1923). XXI.—On the dominance ratio. Proceedings of the royal society of Edinburgh, 42, 321-341.

[3]. Antonio, K., & Beirlant, J. (2007). Actuarial statistics with generalized linear mixed models. Insurance: Mathematics and Economics, 40(1), 58-76.

[4]. Martin, S. B., & Barclay, D. R. (2019). Determining the dependence of marine pile driving sound levels on strike energy, pile penetration, and propagation effects using a linear mixed model based on damped cylindrical spreading. The Journal of the Acoustical Society of America, 146(1), 109-121.

[5]. Ng, P. C., & Kirkness, E. F. (2010). Whole genome sequencing. Genetic variation: Methods and protocols, 215-226.

[6]. Behjati, S., & Tarpey, P. S. (2013). What is next generation sequencing?. Archives of Disease in Childhood-Education and Practice, 98(6), 236-238.

[7]. Hwang, K. B., Lee, I. H., Li, H., Won, D. G., Hernandez-Ferrer, C., Negron, J. A., & Kong, S. W. (2019). Comparative analysis of whole-genome sequencing pipelines to minimize false negative findings. Scientific reports, 9(1), 3219.

[8]. Ye, H., Meehan, J., Tong, W., & Hong, H. (2015). Alignment of short reads: a crucial step for application of next-generation sequencing data in precision medicine. Pharmaceutics, 7(4), 523-541.

[9]. Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. Journal of molecular biology, 147(1), 195-197.

[10]. Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology, 48(3), 443-453.

[11]. Кондратьева, О. А., & Карпулевич, Е. А. (2022). Модификация метода расчета полигенных рисков с использованием графа вариации. Труды Института системного программирования РАН, 34(2), 191-200.

[12]. Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. bioinformatics, 25(14), 1754-1760.

[13]. Adjeroh, D., Bell, T., & Mukherjee, A. (2008). The Burrows-Wheeler Transform:: Data Compression, Suffix Arrays, and Pattern Matching. Springer Science & Business Media.

[14]. Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods, 9(4), 357-359.

[15]. Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997.

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

[17]. Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., ... & DePristo, M. A. (2013). From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Current protocols in bioinformatics, 43(1), 11-10.

[18]. VCFV4.4 and BCFv2.2 27 Jan 2023 - GitHub Pages. Available at: http://samtools.github.io/hts-specs/VCFv4.4.pdf

[19]. Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H. C., Kitts, P. A., ... & Church, D. M. (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome research, 27(5), 849-864.

[20]. Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., ... & Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. Nature, 526(7571), 75-81.

[21]. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMtools. bioinformatics, 25(16), 2078-2079.

[22]. 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).

[23]. FASTQ format specification (no date) FASTQ Format. Available at: https://maq.sourceforge.net/fastq.shtml (Accessed: 27 July 2023).

[24]. Huang, W., Li, L., Myers, J. R., & Marth, G. T. (2012). ART: a next-generation sequencing read simulator. Bioinformatics, 28(4), 593-594.

[25]. Cleary, J. G., Braithwaite, R., Gaastra, K., Hilbush, B. S., Inglis, S., Irvine, S. A., ... & De La Vega, F. M. (2015). Comparing variant call files for performance benchmarking of next-generation sequencing variant calling pipelines. BioRxiv, 023754.

Информация об авторах / Information about authors

Егор Павлович ГУГУЧКИН является научным сотрудником ИСП РАН. Его научные интересы включают в себя анализ генетических данных и разработку биоинформатических пайплайнов.

Egor Pavlovich GUGUCHKIN is a research fellow at ISP RAS. His research interests include the analysis of genetic data and the development of bioinformatics pipelines.

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