На правах рукописи

Карпулевич Евгений Андреевич

Построение программного конвейера для выравнивания последовательностей в приложениях биоинформатики

Специальность 2.3.5 –
«Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей»

Автореферат
диссертации на соискание ученой степени
кандидата физико-математических наук

Москва – 2023

Работа выполнена в Федеральном государственном бюджетном учреждении науки Институте системного программирования им. В. П. Иванникова Российской Академии Наук

Научный руководитель: **Турдаков Денис Юрьевич**
кандидат физико-математических наук

Официальные оппоненты: **Макеев Всеволод Юрьевич**,
доктор физико-математических наук, член-корр. РАН, г.н.с. Федерального государственного бюджетного учреждения науки Институт общей генетики им. Н.И.Вавилова Российской академии наук

Алимова Ильсеяр Салимовна, кандидат технических наук, старший преподаватель кафедры программной инженерии Института информационных технологий и интеллектуальных систем Федерального государственного автономного образовательного учреждения высшего образования «Казанский (Приволжский) федеральный университет»

Ведущая организация: Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский Нижегородский государственный университет им. Н.И. Лобачевского»

Защита состоится 07 декабря 2023 г. в 16 часов на заседании диссертационного совета 24.1.120.01 при Федеральном государственном бюджетном учреждении науки Институт системного программирования им. В. П. Иванникова РАН по адресу: 109004, г. Москва, ул. А. Солженицына, дом 25.

С диссертацией можно ознакомиться в библиотеке и на сайте Федерального государственного бюджетного учреждения науки Институт системного программирования им. В. П. Иванникова РАН.

Автореферат разослан “____” _____ 2023 г.

Ученый секретарь
диссертационного совета 24.1.120.01,
кандидат физико-математических наук Зеленов С. В.

Общая характеристика работы

Актуальность проблемы. Данные в ряде прикладных и научных областей могут быть представлены в виде последовательностей. Задача "выравнивания" последовательностей находит применение в таких прикладных областях как сжатие данных, информационный поиск, обработка естественных языков и анализ генетических последовательностей.

Выравнивание последовательностей – это математический метод, используемый для определения сходства или различия между двумя или более последовательностями, обычно строками символов, путем их расположения таким образом, чтобы максимизировать совпадения и минимизировать различия.

Методы решения задачи выравнивания последовательностей стали активно развиваться во второй половине 20-го века. В 1965 году В.И. Левенштейн, сотрудник Института прикладной математики им. М.В. Келдыша, ввел понятие метрики редакционного расстояния. Эта метрика (также известная как расстояние Левенштейна) определяется как минимальное количество односимвольных операций (вставки, удаления, замены), необходимых для превращения одной последовательности символов в другую. В 1970 году Сол Б. Нидлман и Кристиан Д. Вунш представили алгоритм глобального выравнивания последовательностей. Алгоритм Нидлмана-Вунша решает задачу наилучшего (оптимального) выравнивания между двумя последовательностями с использованием их полной длины. В 1981 году Т.Ф. Смит и М.С. Уотерман предложили алгоритм локального выравнивания последовательностей. Алгоритм Смита-Уотермана применяется для идентификации похожих подпоследовательностей в последовательностях.

С развитием возможностей вычислительного анализа научным сообществом были предложены алгоритмы на основе эвристик, алгоритмы с использованием машинного обучения и алгоритмы выравнивания на граф. Так, математик Стивен Альтшул из Национального центра биотехнологической информации США в соавторстве со специалистами из области вычислительной биологии в 1990 году разработал алгоритм и программу BLAST (базовый инструмент поиска локального выравнивания). В ранних 2000-х годах получили развитие методы выравнивания на граф последовательностей. В настоящее время также развиваются алгоритмы выравнивания последовательностей с применением машинного обучения.

Алгоритмы выравнивания последовательностей нашли свое применение в области обработки генетических данных, в частности, данных секвенирования ДНК (ДНК – это молекула, которая может быть представлена в виде последовательности символов из множества {A, C, G, T}). Секвенирование ДНК – экспериментальный метод определения последовательности расположения символов ДНК в биологическом образце организма. На данный момент, наиболее

широко распространенная из существующих технологий секвенирования – секвенирование следующего поколения (NGS). Для получения последовательности ДНК с помощью технологии NGS необходимо произвести несколько шагов: подготовить биологические образцы к секвенированию, получить цифровые данные через обработку подготовленных биологических образцов на специальном приборе (секвенаторе), провести вычислительную обработку выходных данных секвенатора (коротких последовательностей ДНК длиной 50-250 символов). Конечным результатом обработки данных NGS является набор генетических вариантов организма (отличий от референсного генома – заранее известной последовательности ДНК абстрактного организма того же биологического вида). Генетические варианты бывают нескольких типов: однонуклеотидные замены (SNP), вставки и делеции.

Вычислительная обработка данных NGS обычно состоит из нескольких разнородных (с точки зрения требований к вычислительным ресурсам и возможностей распараллеливания) этапов. Одним из ключевых этапов вычислительной обработки данных NGS является выравнивание коротких подпоследовательностей ДНК, полученных от секвенатора, на референсный геном. В инструментах, которые реализуют этап выравнивания генетических последовательностей, могут применяться алгоритмы двух классов: выравнивание на линейный референсный геном и выравнивание на граф, составленный по ДНК нескольких организмов. Первый класс алгоритмов обладает высокой скоростью выравнивания, а второй класс алгоритмов обладает большей точностью. Разработка метода и алгоритмов выравнивания, которые сочетают в себе преимущества обоих подходов, является актуальной задачей.

Количество больших данных NGS (объем данных NGS, полученных из одного биологического образца, составляет от нескольких единиц/десятков до сотен гигабайт) постоянно растет благодаря совершенствованию и удешевлению технологии NGS. Проведение масштабных исследований на тысячах биологических образцов, зачастую с участием нескольких лабораторий, находящихся в разных частях мира, порождает ряд требований к вычислительной обработке данных NGS: автоматизация, масштабируемость, воспроизводимость, контроль качества, поддержка совместной работы и передача накопленного опыта в анализе данных.

Для того чтобы организовать непрерывный цикл разработки, тестирования и эксплуатации масштабируемых биоинформатических программных конвейеров, необходимо использовать современные IT-технологии: системы управления программными конвейерами, облачные вычисления, контроль версий, контейнеризацию, планировщики задач. Требования к оптимизации вычислительных мощностей, снижению затрат на разработку и развитию

программных конвейеров делают актуальной разработку архитектуры воспроизводимых масштабируемых систем анализа данных NGS.

Целью данной работы является разработка алгоритмов и метода выравнивания последовательностей для решения задачи секвенирования ДНК, а также разработка и реализация архитектуры воспроизводимых биоинформатических программных конвейеров обработки данных секвенирования ДНК человека. Разработанная реализация программного конвейера для обработки данных секвенирования ДНК человека должна превосходить существующие реализации по качеству идентификации однонуклеотидных полиморфизмов.

Для достижения поставленной цели необходимо решить следующие **задачи**:

1. Разработать метод выравнивания генетических последовательностей на референсный геном с использованием данных об известных генетических вариантах
2. Разработать алгоритмы в составе метода выравнивания генетических последовательностей и получить аналитические оценки их вычислительной и пространственной сложности
3. Разработать и реализовать архитектуру системы анализа данных NGS на базе программного конвейера, реализующего предложенный метод выравнивания генетических последовательностей, а также экспериментально оценить метрики качества идентификации генетических вариантов на данных NGS

Основные положения, выносимые на защиту:

1. Новый метод выравнивания генетических последовательностей на референсный геном с использованием данных об известных генетических вариантах
2. Алгоритмы в составе метода выравнивания генетических последовательностей и аналитические оценки их вычислительной и пространственной сложности через доказательство соответствующих теорем
3. Архитектура и реализация системы анализа данных NGS на базе программного конвейера для обработки данных секвенирования ДНК человека с использованием модифицированного индекса

Научная новизна. Разработан новый метод выравнивания генетических последовательностей на референсный геном с использованием данных об известных генетических вариантах, который сочетает в себе преимущества методов выравнивания на линейный референсный геном и выравнивания на граф, составленный по ДНК нескольких организмов. Разработаны алгоритмы в составе метода выравнивания генетических последовательностей. Доказаны теоремы о их

вычислительной и пространственной сложности. Оценки, полученные в результате доказательства теорем, показывают, что вычислительная сложность алгоритмов построения создания модифицированного индекса референсной генетической последовательности остается линейной, а вычислительная сложность алгоритмов выравнивания генетических последовательностей на модифицированный индекс не изменяется по сравнению с выравниванием на индекс референсного генома. Теорема об оценке пространственной сложности позволяет оценить количество оперативной памяти необходимой для работы реализации алгоритмов.

Практическая значимость. Разработана и реализована архитектура системы анализа данных NGS на базе программного конвейера, реализующего предложенный метод выравнивания генетических последовательностей. Предложенные метод выравнивания генетических последовательностей на референсный геном с использованием данных об известных генетических вариантах реализован посредством модификации функций существующего инструмента выравнивания генетических последовательностей на референсный геном minimap2. Инструмент для выравнивания последовательностей minimap2 используется в том числе в коммерческих решениях (например, MGI MegaBOLT). Добавление информации об известных генетических вариантах в индекс инструмента minimap2 позволило повысить качество выравнивания ридов, что показано экспериментально. Реализация программного конвейера анализа данных NGS секвенирования ДНК человека с использованием модифицированного инструмента minimap2 позволило снизить количество ложноотрицательных срабатываний на 25% (274 SNP) по сравнению с программным конвейером bgallagher-sentieon, победившем в конкурсе PrecisionFDA Truth Challenge. Описана процедура развертывания разработанного программного конвейера на SLURM кластере в облачной среде Asperitas, проведена оценка функционирования программного конвейера на SLURM-кластере. Результаты работы могут быть использованы в научных исследованиях и промышленных проектах, которые предполагают массовое секвенирование ДНК с помощью технологии NGS.

Апробация работы. Результаты работы докладывались на следующих конференциях:

1. Открытая конференция ИСП РАН, Москва, Россия, декабрь 2021
2. Конференция «MACSPro», Москва, Россия, декабрь 2021
3. Конференция «SIBS» (Сеченовский международный биоинформатический саммит), Москва, Россия, ноябрь 2022
4. Конференция «Ломоносовские чтения» - 2023, Москва, Россия, апрель 2023
5. Конференция «Анализ данных в медицине», Великий Новгород, Москва, июнь 2023

Личный вклад. Все выносимые на защиту результаты получены лично автором.

Публикации. Основные результаты по теме диссертации изложены в трех работах, опубликованных в изданиях, рекомендованных ВАК, кроме того, получено свидетельство о государственной регистрации программы для ЭВМ.

В статье [1] поставлена задача совместно с соавтором, автору принадлежит основная часть: разделы 2-4, реализация инструмента и финальное редактирование текста также выполнены автором.

В статьях [2;3] вместе с соавторами поставлена задача и проводилась редакторская правка, разработка программных конвейеров выполнена автором.

На основе разработанного программного конвейера получено свидетельство о государственной регистрации программы для ЭВМ [4].

Объем и структура работы. Диссертационная работа состоит из введения, четырех глав, заключения и списка литературы, содержащего 96 ссылок. Работа изложена на 123 страницах, содержит 10 рисунков, 17 листингов и 10 таблиц.

Содержание работы

Во **введении** обосновывается актуальность исследований, проводимых в рамках данной диссертационной работы, формулируется цель, ставятся задачи работы, перечисляются основные положения, выносимые на защиту, излагается научная новизна, практическая значимость представляемой работы.

В **первой главе** дается обзор существующих подходов к выравниванию ридов (подпоследовательностей ДНК длиной 50-250 нуклеотидов) на референсный геном. В начале обзора приведены основные понятия и определения биоинформатического домена, в частности, обсуждаются особенности терминологии, приведен краткий обзор технологий секвенирования и вводится ключевое понятие референсного генома. Также описан формат и особенности выходных данных секвенатора и выполнен разбор классического программного конвейера анализа данных секвенирования ДНК из коллекции программных конвейеров Broad Institute “Best Practices Workflows”.

Затем дается краткая сводка по описанным в литературе подходам к выравниванию ридов: методы точного выравнивания последовательностей с помощью алгоритма Нидлмана-Вунша (для глобального выравнивания последовательностей) и алгоритма Смита-Ватермана (для локального выравнивания), двухэтапное выравнивание на линейный референсный геном методом и seed-chain-align, а также выравнивание на пангеномный граф. Пангеномный граф – это структура данных, используемая для представления геномных данных, которая учитывает различия между индивидами (генетические варианты). Далее приводится описание метрик для оценки качества отдельно инструмента выравнивания (метрика MAPQ, mapping quality) и программного конвейера анализа данных секвенирования ДНК в целом (метрики Recall, F-мера, Precision на данных проекта “The Genome in a Bottle”).

Затем рассматриваются проблемы современной биоинформатики с акцентом на масштабируемость анализа генетических данных. Экспоненциальный рост генетических данных, генерируемых с помощью технологии NGS, представляет собой проблему для вычислительной инфраструктуры отдельных лабораторий. Современные решения для хранения данных и распределенных вычислительных систем на базе облачных технологий позволяют решить эту проблему.

Еще одной серьезной проблемой, которую можно решить с помощью использования облачных технологий, является воспроизводимость биоинформатического анализа. Обсуждается важность использования одних и тех же версий инструментов анализа, генетических баз данных в рамках эксперимента для обеспечения воспроизводимости исследований в области биоинформатики. Облачные веб-лаборатории также актуальны в контексте междисциплинарного сотрудничества и необходимости совместной работы исследователей из

различных областей, включая информатику, биологию и статистику, для совместного решения задач биоинформатики.

В обзоре также рассматриваются сложности анализа данных NGS, в том числе, такие вопросы, как особенности выравнивания ридов в задаче секвенирования ДНК. Для эффективного выравнивания последовательностей в таких задачах, как анализ данных секвенирования ДНК, в современных биоинформатических инструментах выравнивания ридов применяется двухэтапный подход seed-chain-align (Рисунок 1), позволяющий ускорить процесс выравнивания большого количества ридов на референсную последовательность.

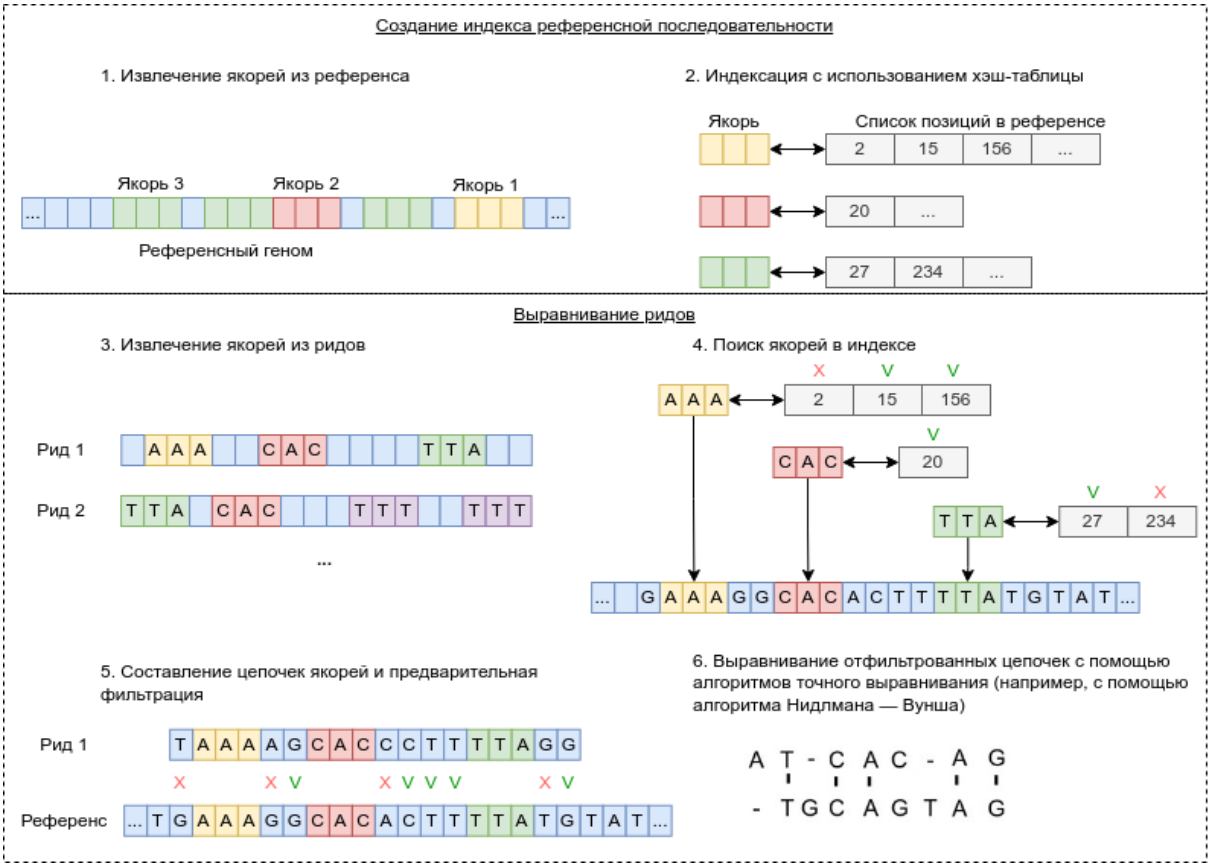

Рисунок 1 – Выравнивание ридов методом Seed-Chain-Align

Описан инструмент minimap2 в котором реализован подход Seed-Chain-Align. Minimap2 — это универсальный инструмент для выравнивания ридов, который использует алгоритмов и методов для эффективного сопоставления коротких и длинных ридов с референсным геномом. Ключевым понятием инструмента minimap2 является минимизатор.

Минимизатор — это короткая подстрока длины k, которая является лексикографически минимальной строкой в окне w (Рисунок 2).

Представлено сравнение алгоритмов выравнивания на линейный референс и на пангеномный граф. В целом, выравнивание на пангеномный граф является более ресурсоемкой задачей, чем выравнивание на линейный референсный геном. Однако это также позволяет учитывать генетическую изменчивость и

обрабатывать регионы с большой вариабельностью более точно, что является важным в генетических исследованиях. Также в первой главе вводится ряд терминов и сокращений.

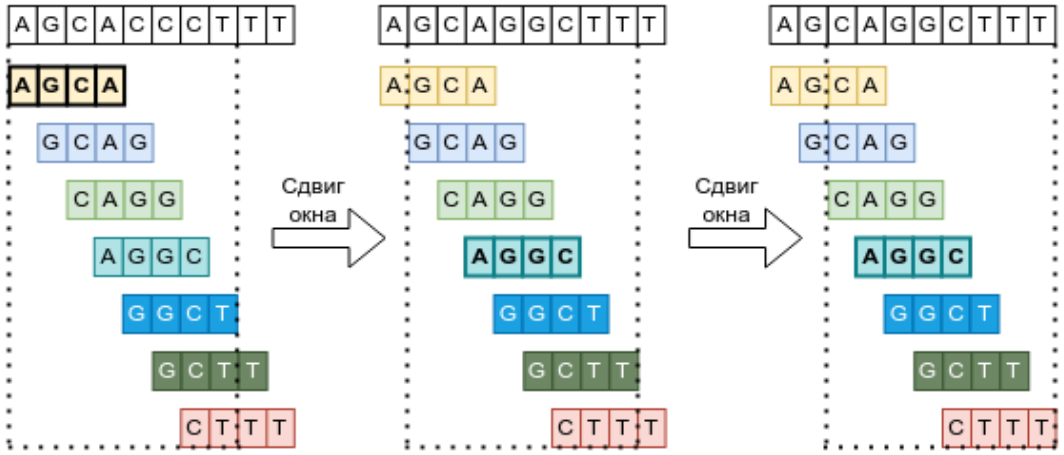

Рисунок 2 – Поиск минимизаторов длины k=4 в окне длины w=5.

Для того чтобы обеспечить возможность обработки больших объемов данных с воспроизводимым вычислительным результатом, необходимо создание программных конвейеров обработки биоинформатических данных. Во **второй главе** предложен метод выравнивания генетических последовательностей на референсный геном с использованием данных об известных генетических вариантах, а также описаны алгоритмы, разработанные автором для использования в составе метода выравнивания генетических последовательностей, и получены аналитические оценки их вычислительной и пространственной сложности. Для эффективного выравнивания последовательностей в таких задачах, как секвенирование ДНК, в современных биоинформатических инструментах выравнивания ридов применяется двухэтапный подход seed-chain-align.

При выполнении выравнивания рида для каждого генетического варианта возможны две ситуации: генетический вариант в секвенируемом организме не попадает в якорь в риде или генетический вариант попадает в якорь в риде.

В случае если генетический вариант не попадает в якорь в риде – позиции якорей определяются верно, то цепочка будет составлена верно и рид будет выровнен в корректную позицию.

В случае если генетический вариант попадает в якорь в риде – позиция якоря может быть определена неверно, что может привести к тому, что рид не будет выровнен или будет выровнен в неверную позицию на геноме из-за ухудшения оценки качества составления цепочки и возможному выбору другой цепочки вместо необходимой.

Выбор якорей можно сделать разными способами. Один из вариантов выбора якорей – все k-меры (последовательность нуклеотидов длины k) референсного генома. Однако при таком способе выбора якорей их количество будет равно длине референсного генома и будет занимать значительное количество памяти и влиять на скорость выравнивания. Один из альтернативных способов выбора якорей — использование минимизаторов.

В качестве удобного для модификаций внутреннего представления индекса может быть использована хэш-таблица позиций минимизаторов. В таком случае существует возможность модификации индекса путем добавления новых минимизаторов. Модификация алгоритма построения индекса (Рисунок 3) для последующего более качественного выравнивания ридов заключается в том, что на этапе поиска якорей после индексации референсного генома дополнительно выполняется добавление минимизаторов и их позиций для известных генетических вариантов.

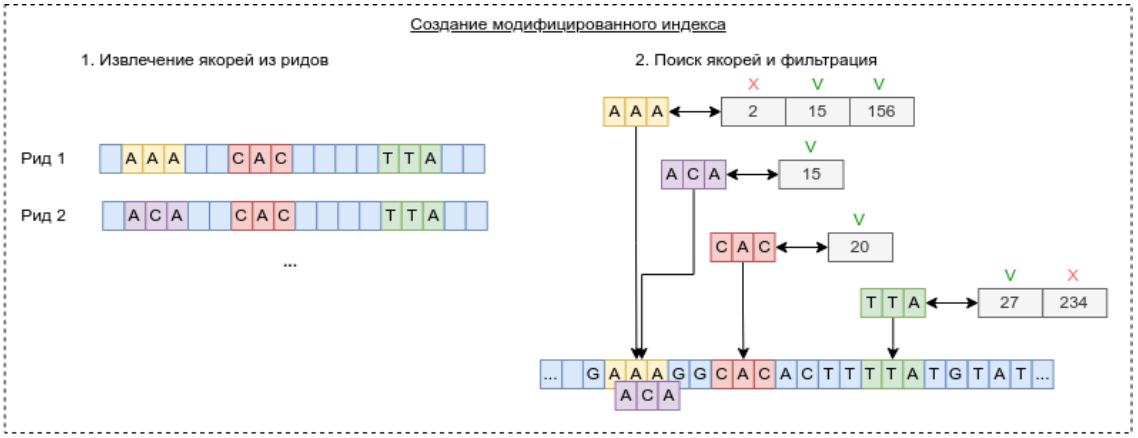

Рисунок 3 – Поиск якорей в модифицированном индексе

Введем следующие обозначения и определения:

Σ = {A; C; G; T} – алфавит нуклеотидов. Для символа (нуклеотида) $a \in \Sigma$ $\bar{a}$ – символ (нуклеотид), комплементарный по Уотсону-Крику.

Строка $s = a_1 a_2 .. a_n$ из символов множества Σ называется последовательностью ДНК. Ее длина $|s| = n$, она обратно комплементарна, т.е. для $s = a_1 a_2 .. a_n$ $\bar{s} = \bar{a}_n \bar{a}_{n-1} .. \bar{a}_1$

k-мером называется последовательность ДНК длины k, таким образом $s_i^k = a_i .. a_{i+k-1}$ – k-мер, начинающийся в i-й позиции, Σ^k – множество всех k-меров.

Для удобства также определим функцию направления π: $\Sigma^* \times \{0,1\} \rightarrow \Sigma^*$ такое, что $\pi(s, 0) = s$, $\pi(\bar{s}, 1) = \bar{s}$. Здесь Σ^* – набор всех последовательностей ДНК.

Файл с генетическими вариантами (VCF_file) содержит N однонуклеотидных полиморфизмов (SNP) для X человек (для каждого человека известны фазированные генотипы для каждого генетического варианта из VCF-файла).

В качестве якорей для создания индекса и последующего выравнивания на него предложено использовать минимизаторы – короткие подстроки длины k, которые являются лексикографически минимальными в окне w.

Комбинацией SNP является набор из одного и более SNP.

Комбинацией SNP в окне длины k является комбинация SNP, позиции которых расположены в одной хромосоме в интервале [min_pos, $min\_pos + k$], где min_pos – минимальная позиция SNP в комбинации.

Допустимой комбинацией SNP в окне длины k является комбинация SNP в окне длины k, которая встречается хотя бы у одного из X человек в VCF_file (в случае если VCF_file содержит информацию о фазировании SNP, то с учетом данной информации).

Ниже представлен метод выравнивания генетических последовательностей на референсный геном с использованием данных об известных генетических вариантах. Метод объединяет в себе преимущества методов выравнивания на линейный референсный геном (скорость) и методов выравнивания на пангеномный граф (качество).

Метод выравнивания генетических последовательностей на референсный геном с использованием данных об известных генетических вариантах состоит в последовательном решении двух задач, первая из которых разбивается на три шага, а вторая на четыре шага, где выполняются вычисления по разработанным алгоритмам, описанным ниже:

1. Создание модифицированного индекса референсной генетической последовательности с добавлением данных об известных генетических вариантах:
    a. Поиск минимизаторов для референсной генетической последовательности s (Листинг 1)
    b. Чтение генетических вариантов из VCF-файла и поиск допустимых комбинаций SNP в окне длины k (Листинг 2)
    c. Модификация участка исходной последовательности референсного генома заменой нуклеотидов в позициях с генетическими вариантами допустимой комбинации SNP и пересчет минимизаторов на модифицированных участках для каждой допустимой комбинации SNP. (Листинг 3)
2. Выравнивание генетических последовательностей на модифицированный индекс, для каждого рида:
    a. Вычисление минимизаторов для рида R
    b. Поиск позиций минимизаторов рида R в модифицированном индексе референсного генома

c. Составление цепочек минимизаторов в референсном геноме в соответствии с их порядком в риде R
d. Точное выравнивание последовательностей с помощью алгоритма Смита-Уотермана и определение позиции наилучшего выравнивания

Листинг 1 – Алгоритм поиска минимизаторов для референсной генетической последовательности *s*

```
Входные данные: параметры для поиска минимизаторов w и k;
последовательность s (|s| >= w+k-1)
Выходные данные:  (w,k)-минимизаторы, и их позиции

Function GetMinimizer (s, k, w) begin
M = ∅      // M - множество без дубликатов
for (i = 1 to |s| - w - k - 1) do
  m = ∞
  for (j = 0 to w - 1) do // найти минимальное значение
    u = si+jk
    v = s⃗i+jk
    if (u != v) then // пропустить если направление не определено
      m = min (m, min (u,v))
  for (j = 0 to w - 1) do // собрать минимизаторы
    u = si+jk
    v = s⃗i+jk
    if (u < v and u = m) then
      M = M U {(m, i+j, 0)}
    if (u > v and v = m) then
      M = M U {(m, i+j, 1)}
return M
```

Листинг 2 – Алгоритм чтения генетических вариантов из VCF-файла и поиск допустимых комбинаций SNP в окне длины *k*

```
Входные данные: VCF_file с SNP, параметры w и k
Выходные данные: (w,k)-минимизаторы, и их позиции

List <SNP> SNP_list = NULL

Function GetMinimizersForSNPs(k, w, VCF_file)) begin
  // считать SNP в однонаправленный список
  for (str in readline(VCF_file)) do
    SNP_list.insert_at_begin(SNP)

  w_beg_ptr = SNP_list //window_start_pointer
  cur_ptr = SNP_list // current_pointer
```

```
  w_end_ptr = SNP_list // window_end_pointer
  gap = k - 1
  // перемещение по списку скользящим окном
  while (w_end_ptr.next) do
    while (w_end_ptr and w_end_ptr.pos > (cur_ptr.pos - gap))do
      w_end_ptr = w_end_ptr.next
    while (w_beg_ptr.pos > (cur_ptr.pos + gap)) do
      w_beg_ptr = w_beg_ptr.next
    return GetMinimizersW(w_beg_ptr, w_end_ptr, cur_ptr, k, w)
    cur_ptr = cur_ptr.next
    w_end_ptr = cur_ptr
  // последняя группа SNP
  while(w_beg_ptr.pos > cur_ptr.pos + gap)do
    w_beg_ptr = w_beg_ptr.next
return GetMinimizersW(w_beg_ptr, w_end_ptr.next, cur_ptr, k, w)
```

Листинг 3 – Алгоритм поиска минимизаторов для допустимых комбинаций SNP в окне длины k

```
Входные данные: параметры w и k, указатели на начало, конец окна и текущий
SNP
Выходные данные:  (w,k)-минимизаторы, и их позиции

Function GetMinimizersW (beg_ptr, end_ptr, cur_ptr) begin
  COMBS = ∅ // Combinations - множество допустимых комбинаций SNP
  Size = size(genotypes)

  for (i = 0 to Size - 1) do
    w_pointer = beg_ptr
    CombinationWithPos = {}
    while (w_pointer != end_ptr) do
      if (w_pointer.genotypes[i] == 0) then
        CombinationWithPos.append((w_pointer.REF, w_pointer.pos))
      if (w_pointer.genotypes[i] == 1) then
        CombinationWithPos.append((w_pointer.ALT, w_pointer.pos))
      w_start_pointer = w_start_pointer.next
    COMBS = COMBS U (CombinationWithPos)
    return AddVariants(Combination, cur_ptr.pos, k, w)
```

В конце главы оценивается вычислительная сложность предложенного алгоритма построения модифицированного индекса, которая задана следующей теоремой:

Теорема 1. Пусть Σ = {A; C; G; T} – алфавит нуклеотидов, строка s = $a_1a_2..a_n$ из символов множества Σ – референсная последовательность длины n. Файл с

генетическими вариантами (VCF-файл) содержит N SNP для X человек. Общая вычислительная сложность алгоритма построения индекса референсной последовательности s модифицированного добавлением N генетических вариантов составляет $O(n)+O(N)$.

Также в конце второй главы приведена оценка потребления памяти предложенным алгоритмом построения модифицированного индекса в следующей теореме:

Теорема 2. Пусть $\Sigma = \{A; C; G; T\}$ – алфавит нуклеотидов, строка $s = a_1a_2..a_n$ из символов множества Σ – референсная последовательность длины n. Файл с генетическими вариантами (VCF-файл) содержит N SNP для X человек. Общее количество потребляемой памяти алгоритма построения индекса референсной последовательности s модифицированного добавлением N генетических вариантов для этапа поиска минимизаторов длины k оценивается как $O(2k^2 \times n + k^5 \times N/3)$ бит.

В качестве модельной реализации алгоритма выравнивания *N_reads* ридов на референсный геном s предложим следующий алгоритм, для каждого рида R необходимо:

1. Вычислить минимизаторы для рида R
2. Найти позиции минимизаторов рида R в индексе референсного генома
3. Составить цепочки минимизаторов в референсном геноме в соответствии с их порядком в риде R
4. Для не более чем *N_chains* цепочек произвести точное выравнивание последовательностей с помощью алгоритма Смита-Уотермана и вернуть позицию наилучшего выравнивания

Для того чтобы выровнять риды на модифицированный референсный геном модифицируем предложенный выше алгоритм следующим образом, для каждого рида R необходимо:

1. Вычислить минимизаторы для рида R
2. Найти позиций минимизаторов рида R в модифицированном индексе референсного генома
3. Составить цепочки минимизаторов в референсном геноме в соответствии с их порядком в риде R
4. Для не более чем *N_chains* цепочек произвести точное выравнивание последовательностей с помощью алгоритма Смита-Уотермана и вернуть позицию наилучшего выравнивания

Вычислительную сложность предложенного алгоритма выравнивания ридов на модифицированный референсный можно оценить с помощью следующей теоремы:

Теорема 3. Пусть Σ = {A; C; G; T} – алфавит нуклеотидов, строка $s = a_1a_2..a_n$ из символов множества Σ – референсная последовательность длины n строка $R = b_1b_2..b_n$ из символов множества Σ – рид длины $|R|=R\_len$. Вычислительная сложность алгоритма выравнивания ридов на модифицированный референсный геном и алгоритм выравнивания ридов на модифицированный референсный геном по сравнению с вычислительной сложностью алгоритма выравнивания ридов на референсный геном остается неизменной.

Третья глава посвящена разработке и реализации архитектуры системы анализа данных NGS на базе программного конвейера, реализующего предложенный в главе 2 метод выравнивания генетических последовательностей.

Создание качественного биоинформатического программного конвейера – это долгосрочный и итеративный процесс, который требует внимания к множеству деталей, начиная от выбора инструментов и алгоритмов и заканчивая оптимизацией производительности.

Зачастую разработка программных конвейеров ведется не систематизированно, без использования систем управления программными конвейерами и версионирования. В таком случае достаточно сложно развивать существующую кодовую базу программного конвейера.

На рисунке 4 представлена архитектура системы анализа данных NGS на базе программного конвейера, которая состоит из следующих частей:

- Реестр контейнеров, содержит docker контейнеры инструментов с фиксированными версиями, что позволяет обеспечить воспроизводимость анализа
- Облачное хранилище данных NGS, содержит данные секвенирования
- Облачное хранилище справочных данных, содержит данные референсных геномов и генетических баз данных различных версий
- Инструмент жизненного цикла ПО, содержит версионированный код программного конвейера (позволяет обеспечить воспроизводимость анализа) и инструменты поддержания жизненного цикла разработки
- Программный конвейер
- Система управления программными конвейерами, это система, которая управляет запуском программного конвейера и логированием вывода биоинформатических инструментов
- Облачная среда осуществляет предоставление вычислительных ресурсов по запросу
- Вычислительный кластер развернутый в облачной среде обеспечивает возможность непрерывного анализа данных с возможностью быстрого подключения новых вычислительных узлов

- Система управления ресурсами кластера позволяет ставить в очередь на выполнение программные конвейеры и отслеживать статус выполнения

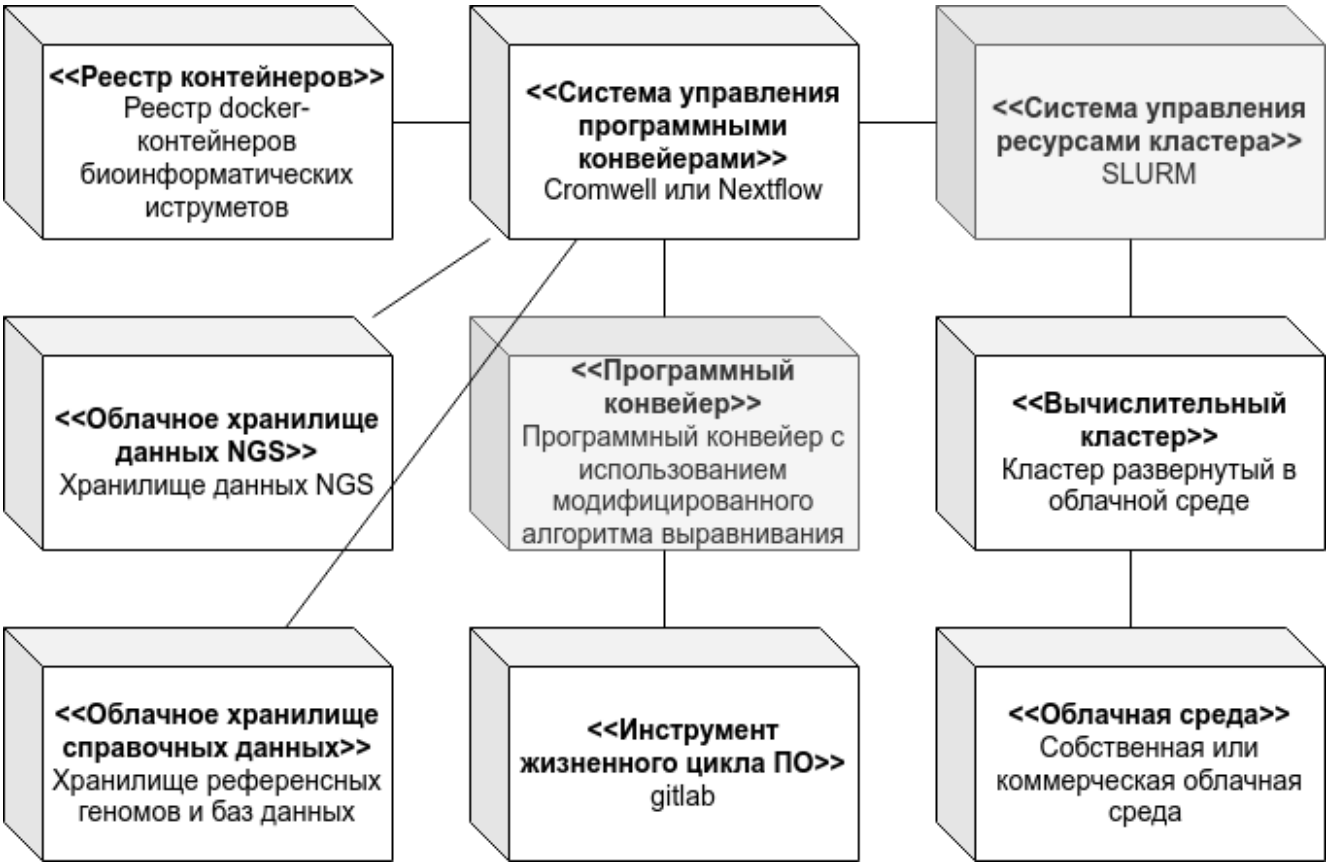

Рисунок 4 – Архитектура системы анализа данных NGS на базе программного конвейера

Целью исследования является создание программного конвейера, принимающего на вход выходные данные секвенатора (файл в формате FASTQ) и выдающего на выходе файл с набором генетических вариантов человека в формате VCF.

Для исследования функционирования программного конвейера разработана методика и инструментарий развертывания программного конвейера на SLURM кластере в облачной среде ИСП РАН Asperitas.

Программный конвейер анализа данных секвенирования ДНК.

Далее приводится описание реализации программного конвейера секвенирования ДНК. Программный конвейер построен согласно методике создания программных конвейеров секвенирования ДНК, разработанной в Институте Броуда (Рисунок 5).

В качестве инструментов для выполнения отдельных шагов программного конвейера согласно методике по построению конвейеров анализа секвенирования ДНК использовались следующие инструменты:

- Выравнивание на референсный геном – minimap2 (версия 2.24)
- Удаление дубликатов – samtools (версия 1.11)

- Перекалибровка качества прочтения нуклеотидов и идентификация генетических вариантов – фреймворк gatk (версия 4.1.7.0, инструменты BaseRecalibrator, HaplotypeCaller и др.)
- Работа с интервалами и определение качества выравнивания – фреймворк picard (версия 2.25.0, инструменты CollectWgsMetrics, ScatterIntervalList, MergeVCF и др.)
- тримминг (за рамками программного конвейера) – cutadapt (версия 4.4)
- сравнение с эталоном и получение итоговых метрик (за рамками программного конвейера) – hap.py (версия 0.3.15)

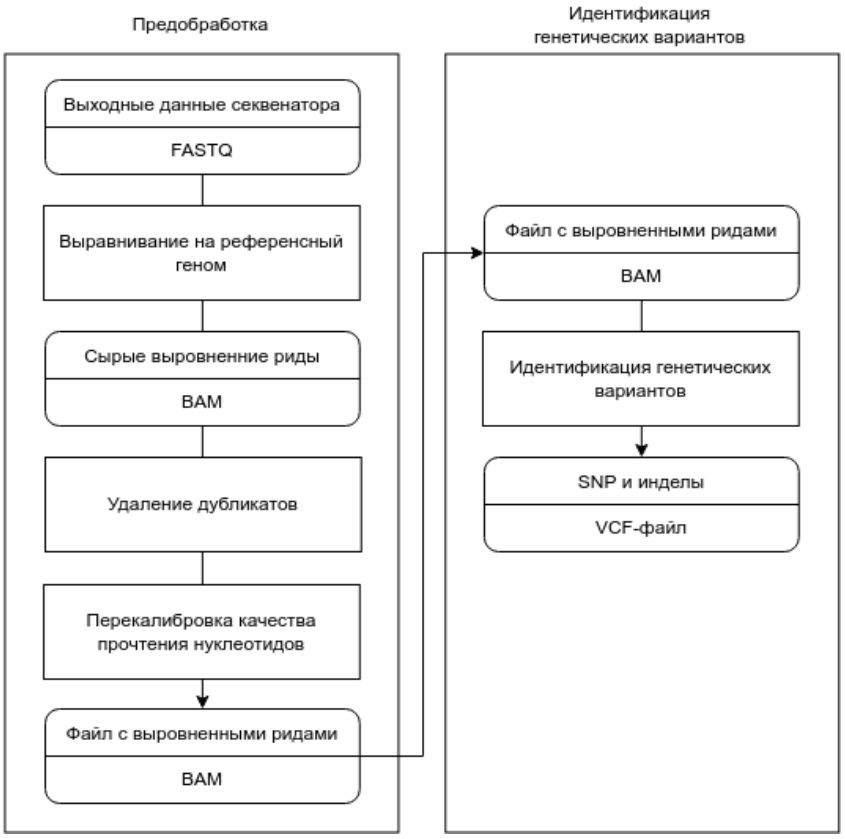

Рисунок 5 – Схема программного конвейера для анализа данных секвенирования ДНК из набора конвейеров “Best Practices Workflows”

Программный конвейер реализован на языке WDL и сконфигурирован для запуска на SLURM-кластере в облачной среде ИСП РАН Asperitas. Все инструменты контейнеризированы и размещены в реестре контейнеров в gitlab.

Программный конвейер принимает в качестве входных данных параметры запуска отдельных инструментов и выходные данные секвенатора (риды). Выходными файлами программного конвейера являются файл генетических вариантов (VCF-файл), файлы с данными анализа качества входных данных и промежуточные файлы, такие как файл выравнивания ридов (Sequence Alignment/Map, SAM-файл). SAM-файл, помимо прочего, для каждого рида

содержит информацию о значении показателя качества выравнивания ридов (Mapping quality, MAPQ, принимает значения от 0 до 60).

Разработка реализации программного конвейера для обработки данных секвенирования ДНК человека и оценка качества идентификации однонуклеотидных полиморфизмов

Метод выравнивания генетических последовательностей на референсный геном реализован в виде библиотеки minimap2_index_modifier которая представляет из себя инструмент minimap2 версии 2.24 с доработанных функционалом построения модифицированного индекса по файлу с генетическими вариантами согласно методу, описанному в главе 2. Для работы с файлом генетических вариантов задействована библиотека htslib версии 1.17.

Оценка качества программного конвейера для обработки данных секвенирования ДНК человека с использованием модифицированного и стандартного индексов проводилась на данных соревнования “The precisionFDA Truth Challenge”. Для сравнения качества идентификации однонуклеотидных полиморфизмов использовалась метрика Recall. Результаты оценки качества идентификации однонуклеотидных полиморфизмов приведены в таблице 1. По метрике Recall программный конвейер с модифицированным индексом превосходит программный конвейер bgallagher-sentieon, который занял первое место в соревновании “The precisionFDA Truth Challenge” (в соревновании использовались данные образца HG002 из проекта "The genome in a bottle").

Таблица 1 – Оценка качества идентификации однонуклеотидных полиморфизмов на покрытии 60X (SNP, образец HG002)

Программный конвейер	**TP**	**FN**	**Recall**
С модифицированным индексом	3054647	943	**0.999691**
bgallagher-sentieon	3054647	998	0.999673

Для дополнительного исследования работы программного конвейера для обработки данных секвенирования ДНК человека с использованием модифицированного индекса были проведены эксперименты по оценке качества разработанного программного конвейера с использованием модифицированного индекса референсного генома по сравнению с программным конвейером, который использует стандартный индекс инструмента minimap2 для разной глубины покрытия (покрытием называется среднее количество раз, которое каждый нуклеотид в геноме считывается в процессе секвенирования).

Для оценки этапа предобработки программного конвейера (Рисунок 5) оценим распределение значений метрики качества выравнивания (Mapping quality,

MAPQ). В качестве метрик оценки конечного результата работы программного конвейера будем использовать такие метрики, как точность (Precision), полнота (Recall) и F-мера. В качестве данных для тестирования возьмем данные секвенирования ДНК биологического образца HG002 из проекта "The genome in a bottle" (в рамках проекта собраны и охарактеризованы эталонные справочные образцы данных секвенирования ДНК человека).

В таблицах 2 и 3 приведены различия в количестве ридов с разными значениями метрики MAPQ после выполнения этапа предобработки программного конвейера при использовании стандартного и модифицированного индекса референсного генома. Значения MAPQ разбиты по группам: 60 (выравнивание ридов без замен и разрывов), 30-60 (выравнивание с высоким качеством), 2-60 (однозначное выравнивание любого качества). Результаты приведены для двух вариантов глубины покрытия: 17X и 60X.

Таблица 2 – Количество ридов с разными значениями MAPQ (HG002 60X)

MAPQ	**Стандартный**	**Модифицированный**	**Разница**
60	949840269	949271649	+568620
30-60	979056628	978898034	+158594
2-60	1007563422	1007539696	+23726

Таблица 3 – Количество ридов с разными значениями MAPQ (HG002 17X)

MAPQ	**Стандартный**	**Модифицированный**	**Разница**
60	250311191	250160776	+150415
30-60	257949672	257906764	+42908
2-60	265409566	265402391	+7175

Анализ распределения значений MAPQ показывает, что в результате модификации индекса увеличивается как количество ридов выровненных без замен и разрывов, так и общее количество качественно выровненных ридов.

Значения метрик для идентификации SNP, полученные с помощью инструмента hap.py при сравнении результатов работы программного конвейера с эталонным VCF-файлом биологического образца HG002 приведены в таблицах 4 и 5.

Разработанный программный конвейер показал снижение ложноотрицательных срабатываний на 25% (274 SNP) на данных с покрытием 60X, а также снижение ложноотрицательных срабатываний на 3% (2945 SNP) на данных с покрытием 17X.

Таблица 4 – Результаты сравнения полученного VCF-файла с эталоном для всего генома на покрытии 60X (SNP, образец HG002)

Тип индекса	**TP**	**FN**	**FP**	**Recall**	**Precision**	**F-score**
Стандартный	3054647	1217	8417	0.999602	0.997251	0.998425
Модифицированный	3054647	943	8382	0.999691	0.997262	0.998475

Таблица 5 – Результаты сравнения полученного VCF-файла с эталоном для всего генома на покрытии 17X (SNP, образец HG002)

Тип индекса	**TP**	**FN**	**FP**	**Recall**	**Precision**	**F-score**
Стандартный	3053446	35828	9786	0.988266	0.996767	0.992499
Модифицированный	3053446	33949	9793	0.988882	0.996767	0.992809

Таким образом показано что разработанный программный конвейер для обработки данных секвенирования ДНК человека с использованием модифицированного индекса превосходит существующие реализации по качеству (по полноте, Recall) идентификации однонуклеотидных полиморфизмов, кроме того разработанный программный конвейер с использованием модифицированного индекса превосходит аналогичный с использованием стандартного индекса по метрике Recall и количеству качественно выровненных ридов, при этом значения остальных метрик (Precision и F–score) не уменьшаются.

Основные результаты третьей главы опубликованы в работе [1].

В **заключении** приведены основные результаты работы, которые заключаются в следующем:

1. Разработан метод выравнивания генетических последовательностей на референсный геном с использованием данных об известных генетических вариантах.
2. Разработаны алгоритмы в составе метода выравнивания генетических последовательностей и аналитические оценки их вычислительной и пространственной сложности через доказательство соответствующих теорем
3. Дана оценка потребления памяти для алгоритма модификации индекса на модельном примере показывает, что для создания модифицированного индекса достаточно мощностей сервера с 500гб оперативной памяти.
4. Разработана архитектура системы анализа данных NGS на базе программного конвейера, реализующего предложенный метод выравнивания генетических последовательностей
5. В рамках архитектуры реализован программный конвейер на основе популярного инструмента выравнивания ридов minimap2. Программный конвейер апробирован на данных проекта "The Genome in a Bottle".

6. Разработанный программный конвейер для обработки данных секвенирования ДНК человека с использованием модифицированного индекса превосходит существующие реализации по полноте (Recall) идентификации однонуклеотидных полиморфизмов, кроме того разработанный программный конвейер с использованием модифицированного индекса превосходит аналогичный с использованием стандартного индекса по качеству (по полноте, Recall) и количеству качественно выровненных ридов, при этом значения остальных метрик (Precision и F–score) не уменьшаются.

Публикации автора по теме диссертации

1. Гугучкин Е.П., Карпулевич Е.А. Модификация алгоритма выравнивания коротких прочтений для повышения качества пайплайна обработки данных полногеномного секвенирования человека. Труды Института системного программирования РАН. 2023;35(2):235-248. https://doi.org/10.15514/ISPRAS-2023-35(2)-17
2. Кондратьева О.А., Карпулевич Е.А. Модификация метода расчета полигенных рисков с использованием графа вариации. Труды Института системного программирования РАН. 2022;34(2):191-200. https://doi.org/10.15514/ISPRAS-2022-34(2)-15
3. Albert, E. A., Kondratieva, O. A., Baranova, E. E., Sagaydak, O. V., Belenikin, M. S., Zobkova, G. Y., Kuznetsova, E. S., Deviatkin, A. A., Zhurov, A. A., Karpulevich, E. A., Volchkov, P. Y., & Vorontsova, M. V. (2023). Transferability of the PRS estimates for height and BMI obtained from the European ethnic groups to the Western Russian populations. In Frontiers in Genetics (Vol. 14). Frontiers Media SA. https://doi.org/10.3389/fgene.2023.1086709
4. Свидетельство о государственной регистрации программы для ЭВМ №2022614027 “Программный комплекс "EVOGEN WEB SYSTEM"”

Карпулевич Евгений Андреевич

Построение программного конвейера для выравнивания последовательностей в приложениях биоинформатики

Автореф. дис. на соискание ученой степени канд. физико-математических наук

Подписано в печать __.__.____. Заказ № _____
Формат 60×90 / 16. Усл. печ. л. 1. Тираж 100 экз.
Типография _____