(19) Евразийское патентное ведомство

(21) 201990933 (13) A1

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ЕВРАЗИЙСКОЙ ЗАЯВКЕ

(43) Дата публикации заявки
2019.11.29

(22) Дата подачи заявки
2016.10.11

(51) Int. Cl. ***G06F 19/28*** (2011.01)
G06F 19/22 (2011.01)

(54) ЭФФЕКТИВНЫЕ СТРУКТУРЫ ДАННЫХ ДЛЯ ПРЕДСТАВЛЕНИЯ ИНФОРМАЦИИ БИОИНФОРМАТИКИ

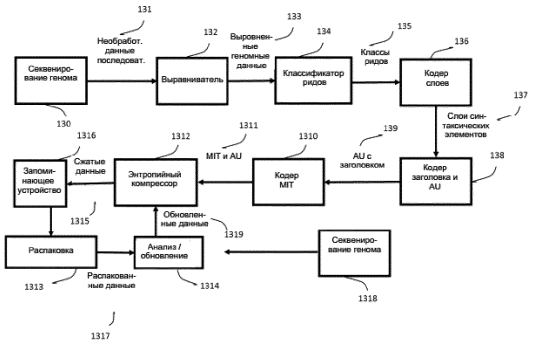
(86) **PCT/EP2016/074297**
(87) **WO 2018/068827 2018.04.19**

(71) Заявитель:
ДЖЕНОМСИС СА (CH)

(72) Изобретатель:
Рензи Даниэле, Зойя Джорджио (CH)

(74) Представитель:
Нилова М.И. (RU)

(57) Способ и устройство для представления данных геномной последовательности, организованных в формате структурированного файла. Соответствующая структура данных содержит представление нуклеотидных последовательностей: в сжатом виде, выровненных по и сопоставленных с одной или более референсными последовательностями и классифицированных в соответствии с различными степенями точности совпадения. Эти классифицированные и выровненные риды закодированы в виде слоев синтаксических элементов, которые, включая информацию заголовка, разбиты на независимые или зависимые блоки доступа.



Общая архитектура.

201990933 A1

201990933 A1

ЭФФЕКТИВНЫЕ СТРУКТУРЫ ДАННЫХ ДЛЯ ПРЕДСТАВЛЕНИЯ ИНФОРМАЦИИ БИОИНФОРМАТИКИ

ОПИСАНИЕ

ОБЛАСТЬ ТЕХНИКИ

[0001] Это изобретение раскрывает слой хранения геномной информации (формат геномных файлов), который определяет структуру геномных данных, включая сбор разнородных данных, связанных с информацией, генерируемой устройствами и приложениями, связанными с секвенированием, обработкой и анализом генома на различных этапах обработки геномных данных (так называемый "жизненный цикл геномной информации").

УРОВЕНЬ ТЕХНИКИ

[0002] Геномная или протеомная информация, генерируемая секвенаторами ДНК, РНК или белков, преобразуется на различных этапах обработки данных для получения неоднородных данных. В решениях предшествующего уровня техники эти данные в настоящее время хранятся в компьютерных файлах, имеющих различные и не связанные структуры. Поэтому эту информацию довольно сложно архивировать, передавать и уточнять.

[0003] Геномные или протеомные последовательности, упоминаемые в данном изобретении, включают, например, помимо прочего, нуклеотидные последовательности, последовательности дезоксирибонуклеиновой кислоты (ДНК), рибонуклеиновой кислоты (РНК) и аминокислотные последовательности. Хотя описание в данном документе является довольно подробным в отношении геномной информации в форме нуклеотидной последовательности, следует понимать, что эти способы и системы для хранения могут быть реализованы также для других геномных или протеомных последовательностей, хотя и с незначительными вариациями, как будет понятно специалисту в данной области.

[0004] Жизненный цикл геномной или протеомной информации от получения данных (секвенирования) до анализа изображен на Фигуре 1, где показаны различные фазы жизненного цикла генома и соответствующие форматы промежуточных файлов. Как показано на Фигуре 1, типичные этапы жизненного цикла геномной информации включают: извлечение ридов последовательности, картирование и выравнивание, детектирование вариантов, аннотацию вариантов и функционально-структурный анализ.

[0005] Извлечение ридов последовательности - это процесс (выполняемый человеком-оператором или машиной-секвенатором) представления фрагментов генетической информации в виде последовательностей символов, представляющих молекулы, составляющие биологический образец. В случае нуклеиновых кислот такие молекулы называются "нуклеотидами". Последовательности символов, полученные в результате извлечения, обычно называют "ридами". Эта информация обычно кодируется в предшествующем уровне техники в виде файлов FASTA, включающих текстовый заголовок и последовательность символов, представляющих секвенированные молекулы.

[0006] Когда биологический образец секвенируют для извлечения из живого организма ДНК, алфавит состоит из символов (A,C,G,T,N).

[0007] Когда биологический образец секвенируют для извлечения из живого организма РНК, алфавит состоит из символов (A,C,G,U,N).

[0008] В случае расширенного набора символов IUPAC, секвенатор также генерирует так называемые "коды неоднозначности", и алфавит, используемый для символов, составляющих риды, состоит из символов (A, C, G, T, U, W, S, M, K, R, Y, B, D, H, V, N или -).

[0009] Когда коды неоднозначности IUPAC не используются, с каждым ридом последлвательности (прочитанной последовательностью) может быть ассоциирована последовательность показателей качества. В таком случае решения предшествующего уровня техники кодируют полученную информацию в виде файла FASTQ.

[0010] Термин "выравнивание последовательностей" относится к процессу упорядочения ридов последовательности путем нахождения областей сходства, которые могут быть следствием функциональных, структурных или эволюционных связей между последовательностями. Когда выравнивание выполняется по ранее существующую нуклеотидную последовательности, называемой "референсной последовательностью", этот процесс называется "картирование". Выравнивание последовательностей также может быть выполнено без ранее существовавшей последовательности (то есть референсного генома), в таких случаях процесс известен в предшествующем уровне техники как выравнивание "de novo". Известные решения хранят эту информацию в файлах SAM, BAM или CRAM. Концепция выравнивания последовательностей для восстановления частичного или полного генома изображена на Фигуре 2.

[0011] Детектирование вариантов (также называемое определением вариантов) - это процесс преобразования выровненных выходных данных секвенаторов (риды, созданные устройствами NGS и выровненные), в сводку уникальных характеристик секвенируемого организма, которые отсутствуют в других ранее существующих последовательностях или могут быть найдены только в нескольких ранее существующих последовательностях. Эти характеристики называются "вариантами", потому что они выражены как различия между геномом исследуемого организма и референсным геномом. В решениях предшествующего уровня техники эта информация хранится в файле определенного формата, который называется файлом VCF.

[0012] Аннотация вариантов - это процесс присвоения функциональной информации геномным вариантам, идентифицированным процессом определения вариантов. Это подразумевает классификацию вариантов по их отношению к кодирующим последовательностям в геноме и по их влиянию на кодирующую последовательность и генный продукт. В предшествующем уровне техники она обычно хранится в файле MAF.

[0013] Процесс анализа цепей ДНК (вариантов, CNV = вариаций числа копий, метилирования и т.д.) для определения их связи с функциями и структурой генов (и

белков) называется функциональным или структурным анализом. В предшествующем уровне техники существует несколько различных решений для хранения этих данных.

[0014] Упрощенная концепция связи между форматами файлов, используемыми в конвейерах обработки генома, изображена на Фигуре 3. В этой диаграмме включение файла не подразумевает существование вложенной файловой структуры, а представляет только тип и объем информации, которая может быть закодирована для каждого формата (т. е. SAM содержит всю информацию в FASTQ, но организован в другой файловой структуре). CRAM содержит ту же геномную информацию, что и SAM/BAM, но обеспечивает большую гибкость в типе сжатия, который можно использовать, поэтому он представлен как расширенный набор SAM/BAM.

[0015] Использование нескольких форматов файлов для хранения геномной информации является крайне неэффективным и дорогостоящим. Наличие разных форматов файлов на разных этапах жизненного цикла геномной информации подразумевает линейный рост используемого пространства хранения, даже если инкрементная информация очень мала по сравнению с исходным объемом данных секвенирования. Это становится нерациональным с точки зрения как объема, так и генерируемых затрат, и ограничивает доступность геномных приложений для более широкой части населения. Другие недостатки известных решений предшествующего уровня техники перечислены ниже.

[0016] 1. Доступ, анализ или добавление аннотаций (метаданных) к необработанным данным, хранящимся в сжатых файлах FastQ или любой их комбинации, требует распаковки и повторного сжатия всего файла с интенсивным использованием вычислительных ресурсов и времени.

[0017] 2. Для поиска и выбоки (retrieve) определенных типов информации, таких как положение картирования рида, положение и тип варианта рида, положение и типы инделов, или любых других метаданных и аннотаций, содержащиеся в выровненных данных, хранящихся в файлах BAM, требуется доступ ко всему объему данных, ассоциированному с каждым ридом. Решения уровня техники не дают возможности выборочного доступа к одному классу метаданных.

[0018] 3. Форматы файлов предшествующего уровня техники требуют, чтобы весь файл был получен конечным пользователем до начала обработки. Например, выравнивание ридов может начаться до того, как процесс секвенирования будет завершен, исходя из соответствующего представления данных. Секвенирование, выравнивание и анализ могут идти и выполняться параллельно.

[0019] 4. Решения предшествующего уровня техники не поддерживают структурирование и не способны различать геномные данные, полученные различными процессами секвенирования, в соответствии с их специфической семантикой генерации (например, секвенирование, осуществленное в разное время жизни одного и того же индивида). Такое же ограничение имеет место для секвенирования, полученного с использованием различных типов биологических образцов одного и того же индивида.

[0020] 5. Шифрование всех или выбранных частей данных не поддерживается решениями предшествующего уровня техники. Например, шифрование выбранных областей ДНК; только последовательности, которые содержат варианты; только химерные последовательности; только некартированные последовательности; специфические метаданные (например, происхождение секвенированной выборки, идентификация секвенированного индивида, тип выборки) невозможно.

[0021] 6. Транскодирование из данных секвенирования, выровненных с заданным референсом (т.е. файлом SAM/BAM), в новый референс требует обработки всего объема данных, даже если новый референс отличается от предыдущего референса только одним положением нуклеотида.

[0022] 7. Передача геномных данных является медленной и неэффективной, поскольку используемые в настоящее время форматы данных организованы в монолитные файлы размером до нескольких сотен гигабайт, которые должны быть полностью перенесены на принимающую сторону для обработки. Это подразумевает, что анализ небольшого сегмента данных требует передачи всего файла со значительными затратами с точки зрения потребляемой ширины пропускания канала и времени ожидания. Зачастую передача в режиме онлайн является невозможной для больших объемов передаваемых данных, и передачу данных осуществляют путем физического

перемещения носителей, таких как жесткие диски или серверы хранения, из одного места в другое.

[0023] 8. Обработка данных является медленной и неэффективной из-за того, что информация не структурирована таким образом, чтобы части различных классов данных и метаданных, необходимые для часто используемых приложений анализа, могли быть извлечены без необходимости доступа к полному объему данных. Этот факт обуславливает то, что обычные конвейеры анализа могут требовать работы в течение нескольких дней или недель, тратя ценные и дорогостоящие ресурсы обработки из-за необходимости на каждом этапе доступа анализировать и фильтровать большие объемы данных, даже когда части данных, относящиеся к конкретной цели анализа, гораздо меньше.

[0024] Существует явная потребность в получении оптимального представления данных и метаданных геномного секвенирования (формат геномного файла) за счет организации и разбивки данных таким образом, чтобы сжатие данных и метаданных было максимизировано, а некоторые функции, такие как выборочный доступ и поддержка инкрементных обновлений, и другие функции обработки данных, полезные на разных этапах жизненного цикла данных генома, были реализованы эффективно.

Основными аспектами раскрытого решения являются:

1. Классификация последовательности ридов по разным классам по результатам выравнивания относительно референсной последовательности для обеспечения избирательного доступа к кодированным данным в соответствии с критериями, связанными с результатами выравнивания. Это подразумевает спецификацию формата файла, который "содержит" элементы структурированных данных в сжатом виде. Такой подход можно рассматривать как противоположность подходов предшествующего уровня техники, например SAM и BAM, в которых данные структурируют в несжатом виде, а затем весь файл сжимают. Первое очевидное преимущество этого подхода заключается в наличии возможности эффективно и естественным образом предоставлять различные формы структурированного

выборочного доступа к элементам данных в сжатом домене, что невозможно или чрезвычайно неудобно в подходах предшествующего уровня техники.

2. Разложение классифицированных ридов на слои однородных метаданных с целью максимально возможного уменьшения информационной энтропии. Разложение геномной информации на конкретные "слои" однородных данных и метаданных представляет собой значительное преимущество, позволяющее определять различные модели источников информации, характеризующихся низкой энтропией. Такие модели не только могут отличаться от слоя к слою, но также могут различаться внутри каждого слоя. Такое структурирование позволяет использовать наиболее подходящее конкретное сжатие для каждого класса данных или метаданных и их части со значительным выигрышем в эффективности кодирования по сравнению с подходами предшествующего уровня техники.

3. Структурирование слоев по блокам доступа, то есть геномной информации, которая может быть декодирована либо независимо с использованием только глобально доступных параметров (например, конфигурации декодера), либо с использованием информации, содержащейся в других блоках доступа. Когда сжатые данные внутри слоев разбиваются на блоки данных, включенные в блоки доступа, могут быть определены различные модели источников информации, характеризующиеся низкой энтропией.

4. Информация структурирована таким образом, что любой соответствующий набор данных, используемый приложениями для геномного анализа, эффективно и выборочно доступен через соответствующие интерфейсы. Эти функции обеспечивают более быстрый доступ к данным и обеспечивают более эффективную обработку. Главная индексная таблица и локальные индексные таблицы обеспечивают выборочный доступ к информации, содержащейся в слоях кодированных (то есть сжатых) данных, без необходимости декодировать весь объем сжатых данных. Кроме того, определен механизм связи (ассоциации) между различными слоями данных, чтобы обеспечить выборочный доступ к любой возможной комбинации подмножеств семантически ассоциированных слоев данных и/или метаданных без необходимости декодировать все слои.

5. Совместное хранение главной индексной таблицы и блоков доступа.

КРАТКОЕ ОПИСАНИЕ ЧЕРТЕЖЕЙ

Фигура 1. Блок-схема типичного жизненного цикла геномной информации.

Фигура 2. Схема, показывающая концепцию выравнивания последовательностей для восстановления частичного или полного генома.

Фигура 3. Концептуальная схема, иллюстрирующая упрощенный вид взаимосвязи между форматами файлов, используемых в конвейерах обработки генома.

Фигура 4. Пары ридов, картированных в референсной последовательности.

Фигура 5. Пример блоков доступа в соответствии с принципами данного раскрытия.

Фигура 6. Пример блоков доступа, включающих заголовок и слои, состоящие из блоков данных.

Фигура 7. Связь между геномными "пакетами данных", "блоками", блоками доступа, слоями и классами потоков ридов.

Фигура 8. Главная индексная таблица с векторами картирования локусов первого рида, содержащихся в каждом блоке доступа.

Фигура 9. Общая структура главного заголовка и частичное представление MIT (главной индексной таблице), показывающие положения картирования первого рида в каждом AU (блоке доступа) pos класса P.

Фигура 10. Хранение второго типа данных в MIT.

Фигура 11. Доступ к блокам доступа, содержащим риды класса P, картированные по референсной последовательности № 2 между положениями 150000 и 250000, осуществляется с использованием значений, содержащихся в векторе T1p.

Фигура 12. Модификация в референсной последовательности может преобразовать M-риды в P-риды.

Фигура 13. Блок-схема, показывающая жизненный цикл геномной информации в соответствии с принципами данного изобретения.

Фигура 14. Экстрактор ридов последовательности согласно принципам данного изобретения.

Фигура 15. Геномный кодер 2010 в соответствии с принципами этого изобретения.

Фигура 16. Геномный декодер 218 в соответствии с принципами этого изобретения.

КРАТКОЕ ОПИСАНИЕ

Признаки п.1 формулы изобретения решают проблему существующих решений предшествующего уровня техники, обеспечивая:

способ хранения представления данных геномной последовательности в формате файла генома, причем указанные данные последовательности генома содержат риды последовательностей нуклеотидов, включающий следующие этапы: выравнивание указанных ридов с одной или более референсными последовательностями, таким образом создавая выровненные риды, классификация указанных выровненных ридов в соответствии с различными степенями точности сопоставления с указанной одной или более референсными последовательностями с получением, таким образом, классов выровненных ридов; кодирование указанных классифицированных выровненных ридов в виде слоев синтаксических элементов, структурирование указанных слоев синтаксических элементов с информацией заголовка с получением, таким образом, последовательных блоков доступа, создание главной таблицы индексов, содержащей один раздел для каждого класса выровненных ридов, включающий положения картирования в референсной последовательности первого рида каждого блока доступа каждого класса данных; совместное хранение указанной главной индексной таблицы и указанных данных блока доступа.

[0025] Благодаря совместному хранению индексных таблиц и указанного представления данных геномной последовательности вместо разных отдельных файлов для каждого типа данных представления данных геномной

последовательности, как указанно в приведенном выше описании жизненного цикла, сразу становятся очевидными многие преимущества, а именно:

- Результаты любой промежуточной стадии обработки данных геномной последовательности могут быть инкрементно добавлены к уже существующим данным без необходимости перекодирования в другой формат файла. Например, информация выравнивания может быть добавлена к необработанным данным без необходимости изменять существующий формат файла. Варианты определения результатов могут быть включены в уже существующие выровненные данные последовательности с инкрементным обновлением.
- Данные геномной последовательности могут быть возвращены в соответствии с их конкретными характеристиками без необходимости доступа ко всему файлу или его областям, которые не соответствуют критериям запроса. Например, могут быть выполнены запросы для выборочного доступа к:
    - Ридам последовательности, которые идеально совпадают с одним или более референсными геномами
    - Ридам последовательности, которые содержат только несовпадения, где присутствует символ "N" вместо символа фактического нуклеотида или аминокислоты
    - Ридам последовательности, которые содержат любой тип несовпадения в виде замены символов относительно одного или более геномов
    - Ридам последовательности, которые содержат несовпадения и инсерции или делеции (инделы)
    - Ридам последовательности, которые содержат несовпадения, инсерции или делеции (инделы) и мягко обрезанные символы относительно одного или более геномов
    - Ридам последовательности, которые не могут быть картированы относительно рассматриваемого референсного генома
    - Всем однонуклеотидным полиморфизмам (SNP), которые присутствуют между указанными порогами глубины
    - Всем ридам химерной последовательности

- Всем последовательностям ридов с оценками качества выше определенного порогового значения
- Всем метаданным, ассоциированным с указанным набором ридов последовательности

Классифицируя выровненные риды в соответствии с достоверностью сопоставления с референсной последовательностью, можно добиться избирательного доступа к закодированным данным в соответствии с критериями, связанными с результатами выравнивания.

За счет кодирования классифицированных выровненных ридов в виде слоев синтаксических элементов кодирование может быть адаптировано в соответствии с конкретными особенностями данных или метаданных, переносимых данным слоем, и его статистическими свойствами.

Путем структурирования слоев синтаксических элементов с информацией заголовка в последовательные блоки доступа кодирование, хранение и передача могут быть адаптированы в соответствии с характером данных. Например, кодирование может быть адаптировано для каждого блока доступа, чтобы использовать наиболее эффективную модель источника для каждого уровня данных с точки зрения минимизации энтропии.

В соответствии с одним раскрытым аспектом, способ извлечения ридов последовательностей нуклеотидов, хранящихся в геномном файле, где указанный геномный файл содержит главную индексную таблицу (MIT) и блокирует данные доступа, сохраненные в соответствии с принципами этого раскрытия, при этом указанный способ содержит следующие этапы: получение команды пользователя, идентифицирующей тип извлекаемых ридов, выборку главной индексной таблицы из указанного геномного файла, выборку блоков доступа, соответствующих указанного извлекаемому типу ридов, восстановление указанных ридов последовательностей нуклеотидов, картирование указанных выбранных блоков доступа по одной или более референсным последовательностям.

[0026] Настоящее изобретение дополнительно раскрывает секвенатор генома, содержащий: модуль секвенирования генома, сконфигурированный для вывода ридов последовательностей нуклеотидов из биологического образца, блок выравнивания, сконфигурированный для выравнивания указанных ридов с одной или более референсными последовательностями тем самым создавая выровненные ридам, модуль классификации, сконфигурированный для классификации указанных выровненных ридов в соответствии с соответствующими степенями точности с указанной одной или более референсными последовательностями, тем самым создавая классы выровненных ридов; модуль кодирования, сконфигурированный для кодирования указанных классифицированных выровненных ридов в виде слоев синтаксических элементов, модуль дальнейшего подразделения, сконфигурированный для структурирования указанных слоев синтаксических элементов с информацией заголовка с получением, таким образом, последовательных блоков доступа, модуль обработки индексной таблицы, сконфигурированный для создания главной индексной таблицы, содержащая один раздел для каждого класса выровненных ридов, включающий в себя положения картирования на одной или более референсных последовательностях первого рида каждого блока доступа каждого класса данных; модуль хранения, сконфигурированный для совместного хранения указанной главной индексной таблицы и указанных данным блоков доступа.

В соответствии с одним раскрытым аспектом, экстрактор для извлечения ридов последовательностей нуклеотидов, хранящихся в геномном файле, в котором указанный геномный файл содержит главную индексную таблицу и данные блоков доступа, сохраненные в соответствии с принципами этого раскрытия, причем указанный экстрактор содержит: средство пользовательского ввода, сконфигурированное для приема команд, идентифицирующих тип ридов, подлежащих извлечению, средство выборки, сконфигурированное для выборки указанной главной индексной таблицы из указанного геномного файла, средство выборки, сконфигурированное для извлечения блоков доступа, соответствующих указанному типу ридов, подлежащих извлечению, средство восстановления, сконфигурированное для восстановления указанных ридов последовательностей нуклеотидов, с

картированием указанных выбранных блоков доступа по одной или более референсным последовательностям.

[0027] В соответствии с одним раскрытым аспектом устройство цифровой обработки запрограммировано для осуществления способа, изложенного в непосредственно предшествующем абзаце. В соответствии с другим раскрытым аспектом устройство цифровой обработки осуществляет доступ к носителю для долговременного хранения информации, хранящему инструкции, выполняемые эти устройством цифровой обработки, для осуществления способа, изложенного в предыдущем абзаце.

[0028] В соответствии с другим раскрытым аспектом этот носитель информации может быть использован цифровым процессором и хранит программное обеспечение для обработки геномных или протеомных данных, представленных в виде строк геномных или протеомных символов, содержащих символы из набора символов биоинформатики, в котором каждое основание или пептид геномных или протеомных данных представлены в формате, описанном в предыдущих абзацах. В некоторых вариантах осуществления программное обеспечение обрабатывает геномные или протеомные данные с использованием преобразований цифровой обработки сигналов.

ПОДРОБНОЕ ОПИСАНИЕ

Классификация ридов последовательности

[0029] В соответствии с результатами выравнивания по одной или более референсными последовательностям сгенерированные секвенаторами риды классифицируются в соответствии с раскрытым изобретением на пять различных "классов".

При выравнивании последовательности ДНК нуклеотидов относительно референсной последовательности возможны пять результатов:

1. Обнаружено, что область в референсной последовательности совпадает с ридом последовательности без каких-либо ошибок (идеальное картирование).

Такая последовательность нуклеотидов будет называться "идеально совпадающий рид" или обозначаться как "класс P".

2. Обнаружено, что область в референсной последовательности совпадает с ридом последовательности с несколькими (множеством) несовпадениями, состоящими из ряда положений, в которых секвенатор не смог определить ни одного основания (или нуклеотида). Такие несовпадения обозначаются буквой "N". Такие последовательности будут обозначаться как "несовпадающие N-риды" или "класс N".
3. Обнаружено, что область в референсной последовательности совпадает с ридом последовательности с несколькими (множеством) несовпадениями, состоящими из ряда положений, в которых секвенатор не смог определить ни одного основания (или нуклеотида), ИЛИ было определено основание, отличное от указанного в референсной последовательности. Такой тип несовпадения называется однонуклеотидная вариация (SNV) или однонуклеотидный полиморфизм (SNP). Такие последовательности будут обозначаться как "несовпадающие M- риды" или "Класс M".
4. Четвертый класс состоит из ридов, представляющих тип несовпадений, который включает в себя несовпадение класса M плюс наличие инсерции (вставки) или делеции – «удаления» (также называемых инделами). Инсерции представлены последовательностью из одного или более нуклеотидов, отсутствующих в референсе, но присутствующих в прочитанной последовательности. В литературе, когда инсертированная последовательность находится на краях последовательности, ее называют "мягко обрезанной" (то есть нуклеотиды не соответствуют референсу, но сохраняются в выровненных ридах в противоположность "жестко обрезанным" нуклеотидам, которые отбрасываются). Делеции - это "дыры" (недостающие нуклеотиды) в выровненном риде относительно референса. Такие последовательности будут называться "несовпадающими I-ридами" или "класс I".
5. Пятый класс включает в себя все риды, которые не находят какого-либо достоверного картирования на референсной последовательности в соответствии с указанными ограничениями выравнивания. Такие последовательности называются некартированными и относятся к "классу U".

Некартированные риды можно собрать в одну последовательность, используя алгоритмы сборки de-novo. После создания новой последовательности некартированные риды могут быть дополнительно картированы по отношению к ней и классифицированы в один из четырех классов P, N, M и I.

Разложение геномной информации на слои.

[0030] После того как классификация риды завершена с определением классов, дальнейшая обработка состоит в определении набора различных синтаксических элементов, представляющих оставшуюся информацию, позволяющую реконструировать последовательность ридов ДНК, когда она представлена в качестве картированной на данной референсной последовательности. Сегмент ДНК, относящийся к данной референсной последовательности, может быть полностью выражен следующими параметрами:

- Начальное положение на референсной последовательности (pos).
- Флаг, сигнализирующий о том, что рид должен рассматриваться как обратный комплемент к референсу (rcomp).
- Расстояние до партнера пары в случае спаренных ридов (pair).
- Значение длины рида (295) в случае технологии секвенирования дает переменную длину ридов. В случае постоянной длины рида длина рида, ассоциированная с каждым ридом, очевидно, может быть опущена и может быть сохранена в главном заголовке файла.
- Дополнительные флажки, описывающие специфические характеристики рида (дубликат рида, первый и второй рид в паре и т.д.).
- Для каждого несовпадения:
    - Положение несовпадения (*nmis* для класса N, *snpp* для класса M и *indp* для класса I)
    - Тип несовпадения (отсутствует в классе N, *snpt* в классе M, *indt* в классе I)
- Необязательная мягко обрезанная нуклеотидная строка, если она присутствует (*indc* в классе I).

Эта классификация создает группы дескрипторов (синтаксических элементов), которые можно использовать для однозначного представления ридов геномной

последовательности. В таблице ниже приведены синтаксические элементы, необходимые для каждого класса выровненных ридов.

	P	N	M	I
pos	X	X	X	X
pair	X	X	X	X
rcomp	X	X	X	X
флаги	X	X	X	X
rlen	X	X	X	X
nmis		X		
snpp			X	
snpt			X	
indp				X
indt				X
indc				X

Таблица 1 – Определение слоев для каждого класса данных.

Риды, принадлежащие к классу P, характеризуются и могут быть полностью восстановлены только по положению, информации об обратном комплементе и смещении между членами пар в случае, если они были получены с помощью технологии секвенирования с получением пар, по некоторым флагам и длине рида.

На Фигуре 4 показано, как риды можно объединить в пары (в соответствии с наиболее распространенной технологией секвенирования, разработанной Illumina Inc.) и картировать на референсной последовательности. Пары ридов, картированные на референсной последовательности, кодируются во множество слоев однородных дескрипторов (то есть положения, расстояния между ридами в одной паре, несовпадения и т.д.).

Слой определяется как вектор дескрипторов, относящихся к одному из множества элементов, необходимых для уникальной идентификации ридов, картированных на референсной последовательности. Ниже приведены примеры слоев, несущих каждый по вектору дескрипторов:

• Слой положений ридов

• Слой обратного комплемента

- Слой информации по спариванию
- Слой положений несовпадения
- Слой типа несовпадения
- Слой инделов
- Слой обрезанных оснований
- Слой длины ридов (присутствует только в случае вариабельной длины ридов)
- Слой флагов BAM

Блоки данных, блоки доступа и слой геномный данных

[0031] Структура данных, также раскрытая этим изобретением, основана на следующих концепциях:

Блок данных определяется как набор элементов вектора дескрипторов одного и того же типа (например, положения, расстояния, флаги обратного комплемента, положение и тип несовпадения), составляющих слой. Один слой обычно состоит из множества блоков данных. Блок данных может быть разбит на пакеты геномных данных, которые состоят из блоков передачи, размер которых обычно определяется в соответствии с требованиями канала связи. Такая функция разбиения желательна для достижения эффективности передачи данных с использованием типичного сетевого протокола связи.

Блок доступа определяется как подмножество геномных данных, которые можно полностью декодировать либо независимо от других блоков доступа, используя только глобально доступные данные (например, конфигурацию декодера), либо используя информацию, содержащуюся в других блоках доступа. Блок доступа состоит из заголовка и результата мультиплексирования блоков данных разных слоев. Несколько пакетов одного типа инкапсулированы в блок, а несколько блоков мультиплексированы в один блок доступа. Эти концепции изображены на Фигуре 5. На Фигуре 6 показан блок доступа, состоящий из заголовка и одного или нескольких слоев блоков данных одинаковой природы. На Фигуре 6 показан пример общей структуры блока доступа, изображенной на Фигуре 5, в которой

- блоки данных слоя 1 содержат информацию, относящуюся к положениям ридов на референсной последовательности;
- блоки данных слоя 2 содержат информацию об обратной комплементарности ридов;
- блоки данных слоя 3 содержат информацию, связанную с информации о спаривании ридов;
- блоки данных слоя 4 содержат информацию о длине ридов.

Слой геномных данных определяется как набор блоков геномных данных, кодирующих данные одного типа (например, блоки положений ридов, идеально совпадающих с референсным геномом, закодированы в одном и том же слое).

Поток геномных данных - это пакетированная версия слоя геномных данных, в которой закодированные геномные данные передаются в качестве полезной нагрузки пакетов геномных данных, включая дополнительные служебные данные в заголовке. На Фигуре 7 приведен пример пакетирования трех слоев геномных данных в три потока геномных данных.

Мультиплекс (множество) геномных данных определяется как последовательность геномных блоков доступа, используемая для передачи геномных данных, связанных с одним или более процессами секвенирования, анализа или обработки генома. На Фигуре 7 представлена схема взаимосвязи между геномным мультиплексом, несущим три потока геномных данных, разложенных в блоки доступа. Блоки доступа инкапсулируют блоки данных, принадлежащие трем потокам и разбитые на геномные пакеты для отправки в сети передачи.

Модели источника, энтропийные кодеры и режимы кодирования.

[0032] Для каждого слоя структуры геномных данных, раскрытых в этом изобретении, могут использоваться разные алгоритмы кодирования в соответствии с конкретными характеристиками данных или метаданных, переносимых данным слоем, и его статистическими свойствами. "Алгоритм кодирования" следует понимать как ассоциацию конкретной "модели источника" дескриптора с конкретным "энтропийным кодером". Конкретная "модель источника" может быть определена и

выбрана для получения наиболее эффективного кодирования данных с точки зрения минимизации энтропии источника. Выбор энтропийного кодера может быть обусловлен соображениями эффективности кодирования и/или особенностями распределения вероятностей и ассоциированными проблемами реализации. Каждый выбор конкретного алгоритма кодирования будет называться "режимом кодирования", применяемым ко всему "слою" или ко всем "блокам данных", содержащимся в блоке доступа. Каждая "модель источника", ассоциированная с режимом кодирования, характеризуется:

- Определением синтаксических элементов, порожденных каждым источником (например, положения ридов, информация о спаривании ридов, несосовпадения по отношению к референсной последовательности и т.д.)
- Определением ассоциированной вероятностной модели.
- Определением ассоциированного энтропийного кодера.

Для каждого слоя данных модель источника, принятая в одном блоке доступа, не зависит от модели источника, используемой другими блоками доступа для того же слоя данных. Это позволяет каждому блоку доступа использовать наиболее эффективную модель источника для каждого слоя данных с точки зрения минимизации энтропии.

ТАБЛИЦЫ

Главная индексная таблица

[0033] Для поддержки выборочного доступа к определенным областям выровненных данных, структура данных, описанная в этом документе, реализует инструмент индексирования, называемый Главной индексной таблицей (MIT). Это многомерный массив, содержащий два класса данных:

1. локусы, по которым конкретные ридом картируются на используемых референсных последовательностях. Значения, содержащиеся в MIT, представляют собой положения картирования первого рида на каждом слое pos, так что

поддерживается непоследовательный доступ к каждому блоку доступа. MIT содержит один раздел для каждого класса данных (P, N, M и I) и для каждой референсной последовательности.

2. указатели на блоки доступа, содержащие данные, необходимые для восстановления блоков ридов, после тех, чьи положения картирования хранятся в векторах положений, указанных в пункте 1. Каждый вектор указателей называется локальной индексной таблицей.

Блоки доступа положений картирования

[0034] На Фигуре 8 показана схема MIT с выделением четырех векторов, содержащих положения картирования на референсной последовательности (возможно, более одной) каждого блока доступа каждого класса данных.

MIT содержится в главном заголовке закодированных данных. На Фигуре 9 показана общая структура главного заголовка и пример вектора MIT для класса P кодированных ридов.

Значения, содержащиеся в MIT, изображенной на Фигуре 9, используются для прямого доступа к интересующей области (и соответствующему блоку доступа) в сжатой области.

Например, со ссылкой на Фигуру 9, если аналитику требуется получить доступ к идеально совпадающим ридам, картированным в области, находящейся между положениями 150 000 и 250 000 на референсе № 2, приложение декодирования сразу будет переходить к вектору положений класса P и второму референсу в MIT и будет искать два значения k1 и k2, так что k1 <150 000 и k2> 250 000. В примере на Фигуре 9 это приведет к положениям 3 и 4 второго блока (второй референс) вектора MIT, относящегося к положению картирования класса P. Затем эти возвращенные значения будут использоваться приложением декодирования для выборки положений соответствующих блоков доступа из слоя pos, как описано в следующем разделе.

Указатели блоков доступа

[0035] Второй тип данных, содержащихся в оставшихся векторах MIT (Фигура 8), состоит из векторов указателей на физическое положение каждого блока доступа в закодированном битовом потоке. Каждый вектор называется локальной индексной таблицей, поскольку его область действия ограничена одним однородным классом кодированной информации.

Для каждого из четырех классов картированных ридов (P, N, M, I) необходимо несколько типов блоков доступа для восстановления закодированных ридов (пар). Конкретные типы блоков доступа, ассоциированных с каждым классом данных, зависят от результата функции сопоставления, примененной к ридам в каждом классе относительно одной или более референсных последовательностей, как описано выше.

В предыдущем примере на Фигуре 9 для доступа к области от 150 000 до 250 000 ридов, выровненных по референсной последовательности № 2, приложение декодирования вернуло положения 3 и 4 из вектора положений класса P в MIT. Эти значения должны использоваться процессом декодирования для доступа к 3-му и 4-му элементам соответствующего вектора блоков доступа (в данном случае второго) MIT. В примере, показанном на Фигуре 11, счетчики общего количества блоков доступа, содержащиеся в главном заголовке, используются для пропуска положений блоков доступа, относящихся к референсу 1 (в примере 4). Индексы, содержащие физические положения запрошенных блоков доступа в кодированном потоке, поэтому рассчитываются как:

Положение запрашиваемого AU = AU референса 1, которые необходимо пропустить + положение, выбранное с использованием MIT

т.е.

Положение первого AU: 4 + 3 = 7

Положение последнего AU: 4 + 4 = 8

Это означает, что интересующая нас область (риды класса P, картированные на референсной последовательности № 2 между положениями 150 000 и 250 000,

содержатся в блоках доступа, на которые указывают указатели, хранящиеся в 7-м и 8-м столбце главной индексной таблицы, строка T1p (Блоки доступа Типа 1 типа p).

На Фигуре 11 показано, как элементы одного вектора MIT (например, Pos класса P) указывают на элементы одной LIT (вектор pos типа 1 в примере на Фигуре 11).

Адаптация референсной последовательности

[0036] Несовпадения, закодированные для классов N, M и I, можно использовать для создания "модифицированного генома", который будет использоваться для перекодирования ридов в слое N, M или I (относительно первого референсного генома, R_0) в качестве p ридов относительно "адаптированного" генома R1. Например, если обозначить $r_{i,n}^{M}$ i-е рид класса M, содержащее несовпадения относительно референсного генома n, то после "адаптации" его можно будет получить $r_{i,n}^{M} = r_{i,(n+1)}^{P}$ с $A(Ref_n)=Ref_{n+1}$ где A - преобразование из референсной последовательности *n* в референсную последовательность *n + 1*.

На Фигуре 12 показано, как риды, содержащие несовпадения (M-риды) относительно референсной последовательности 1 (RS1), можно преобразовать в идеально совпадающие риды (P-риды) относительно референсной последовательности 2 (RS2), полученной из RS1, путем изменения несовпадающих положений. Это преобразование может быть выражено как

RS2 = A(RS1)

Если для выражения преобразования A, которое идет от RS1 к RS2, требуется меньше битов выражения несовпадений, присутствующих в M-ридах, этот способ кодирования приводит к меньшей информационной энтропии и, следовательно, лучшему сжатию.

В некоторых случаях одна или более модификаций в референсном геноме могут уменьшить общую информационную энтропию путем преобразования набора N-, M- или I-ридов в P-риды.

[0037] Архитектура системы в соответствии с принципами этого изобретения теперь описывается согласно Фиг. 13. Исходно одно или более устройств секвенирования

генома 130 и/или приложений генерируют и представляют геномную информацию 131 в формате, который содержит

- Одну или более последовательностей символов, представляющих нуклеиновые кислоты
- Уникальный идентификатор для каждой геномной последовательности
- Дополнительный показатель качества для каждого символа
- Необязательные метаданные
- Одну или более необязательных референсных последовательностей, которые будут использоваться для дальнейшей обработки сгенерированных геномных последовательностей.

[0038] Блок выравнивания ридов 132 принимает необработанные данные последовательности и либо выравнивает их по одним или более доступным референсным последовательностям, либо собирает их в более длинные последовательности, ища перекрывающиеся префиксы и суффиксы, применяя метод, известный как сборка "de-novo".

[0039] Модуль классификации ридов 134 получает выровненные данные геномной последовательности 133 и применяет функцию сопоставления к каждой последовательности относительно:

- одной или более доступных референсных последовательностей или
- внутреннего референса, созданного в процессе выравнивания (в случае сборки de-novo)).

[0040] Модуль кодирования слоев 136 принимает классы ридов 135, созданные модулем 134 классификации, и создает слои синтаксических элементов 137.

[0041] Модуль кодирования заголовка и блоков доступа 138 инкапсулирует слои синтаксических элементов 137 в блоки доступа и добавляет заголовок к каждому блоку доступа.

[0042] Модуль кодирования главной индексной таблицы 1310 создает индекс указателей на полученные блоки доступа 139

[0043] Модуль сжатия 1312 преобразует результат указанного представления в более компактный (сжатый) формат 1315, уменьшея используемый объём памяти;

[0044] Локальное или удаленное запоминающее устройство 1316 сохраняет сжатую информацию 1315.

[0045] Модуль распаковки 1313 распаковывает сжатую информацию 1315 для выборки распакованных данных 1317, эквивалентных геномной информации 131.

[0046] Модуль анализа 1314 дополнительно обрабатывает указанную геномную информацию 1317 инкрементно обновляя содержащиеся там метаданные.

[0047] Одно или более устройств или приложений для секвенирования генома 1318 могут добавлять дополнительную информацию к уже существующим геномным данным путем добавления результатов дальнейшего процесса секвенирования генома без необходимости перекодирования существующей геномной информации; для получения обновленных данных 1319. Выравнивание и сжатие должны применяться к вновь сгенерированным геномным данным до их объединения с уже существующими данными.

[0048] Одним из нескольких преимуществ описанного выше варианта осуществления является то, что устройства и приложения для анализа генома, которым требуется доступ к данным, смогут запрашивать и извлекать необходимую информацию с использованием одной или более индексных таблиц.

[0049] Экстрактор ридов последовательности 140 в соответствии с принципами этого изобретения описан на Фигуре 14.

[0050] Экстрактор 140 использует главную индексную таблицу, описанную в этом изобретении, чтобы иметь произвольный доступ к любым ридам последовательности, хранящимся в формате геномного файла согласно этому раскрытию. Экстрактор 140 содержит средство пользовательского ввода 141 для приема от пользователя команд 142 относительно конкретных данных, подлежащих выборке. Например, пользователь может указать:

a. Область генома с указанием:

i. Начального и конечного абсолютного положения в референсном геноме
ii. Одной полной референсной последовательности (например, хромосомы)

b. Только один конкретный тип закодированных ридов, например:
i. Риды, которые идеально совпадают с одной или более референсными последовательностями
ii. Риды, которые представляют ровно N несовпадений относительно одной или более референсных последовательностей
iii. Риды, которые представляют число несовпадений относительно одной или более референсных последовательностей ниже или выше указанного порога
iv. Риды, которое показывают инсерции и делеции относительно референсной последовательности.

Экстрактор MIT 143 на Фигуре 14 анализирует главный заголовок геномного файла для доступа к содержащейся информации, как показано на Фиг. 9:

c. Уникальный идентификатор
d. Версия используемого синтаксиса
e. Размер главного заголовка в байтах
f. Количество референсных последовательностей, используемых для кодирования ридов последовательности
g. Количество блоков данных, содержащихся в потоке
h. Идентификаторы референсов
i. Главная индексная таблица.

Анализатор MIT и экстрактор AU 145 извлекают запрошенные блоки доступа, используя следующую информацию из главной индексной таблицы:

j. векторы положений по референсному геному первого рида в каждом блоке доступа. На Фигуре 9 показано, как устройство декодирования может считать такое положение и найти, какой блок доступа содержит закодированные риды, картированные в запрашиваемой области.

k. Локальная индексная таблица каждого закодированного слоя. Эти векторы используются для выборки физического положения блоков доступа, определенных на этапах a, которые содержат риды, картированные в геномной области, запрошенной пользователем

l. Локальные индексные таблицы определяются для каждого класса данных, поэтому устройство-экстрактор будет извлекать только те классы, которые ссылаются на риды, запрошенные пользователем. Например, в случае запроса только идеально сопоставленных ридов, устройство-экстрактор получит доступ только к LIT, относящемуся к классу P, как показано на Фигуре 8.

Используя информацию, найденную в выбранных блоках доступа, и одну или более референсных последовательностей, либо закодированных в геномном битовом потоке, либо доступных на устройстве-экстракторе, восстановитель ридов 147 способен восстанавливать исходные риды последовательности.

На Фиг.15 показано устройство кодирования 207 в соответствии с принципами этого изобретения. Схема устройства кодирования дополнительно разъясняет аспекты сжатия, использованные в архитектуре системы согласно Фиг.13, однако создание главной индексной таблицы и блоков доступа в устройстве кодирования по Фиг.15 опущены, что дает сжатый поток без этих метаданных и информации о структуре. Устройство кодирования 207 принимает в качестве входных данных необработанные данные последовательности 209, например, созданные устройством секвенирования генома 200. Устройство секвенирования генома 200 известно в данной области техники, например устройства Illumina HiSeq 2500 или Thermo-Fisher Ion Torrent. Необработанные данные последовательностей 209 поступают в модуль выравнивания 201, который подготавливает последовательности для кодирования путем выравнивания ридов с референсной последовательностью. В качестве альтернативы может использоваться сборщик de-novo 202 для создания референсной последовательности из доступных ридов путем поиска перекрывающихся префиксов или суффиксов, так что из ридов могут быть собраны более длинные сегменты (называемые "контигами"). После обработки сборщиком de-novo 202 риды можно картировать на полученной более длинной последовательности. Выровненные

последовательности затем классифицируются модулем классификации данных 204. Затем классы данных 208 подаются в кодеры слоев 205-207. Геномные слои 2011 затем подаются в арифметические кодеры 2012-2014, которые кодируют слои в соответствии со статистическими свойствами данных или метаданных, содержащихся в этом слое. В результате получают геномный поток 2015.

На Фиг.16 показано соответствующее устройство декодирования 218. Устройство декодирования 218 принимает мультиплексированный геномный битовый поток 2110 из сети или элемента хранения. Мультиплексированный геномный битовый поток 2110 подается в демультиплексор 210 для создания отдельных потоков 211, которые затем подаются в энтропийные декодеры 212-214, для получения геномных слоев 215. Выбранные геномные слои подаются в декодеры слоев 216-217 для дальнейшего декодирования слоев в классы данных. Декодеры классов 219 дополнительно обрабатывают дескрипторы генома и объединяют результаты с получением несжатых ридов последовательностей, которые затем могут быть дополнительно сохранены в форматах, известных в данной области техники, например, в текстовом файле или сжатом zip-файле, или файлах FASTQ или SAM/BAM. Декодеры классов 219 способны восстанавливать исходные геномные последовательности, используя информацию об исходных референсных последовательностях, переносимых одним или более геномными потоками. В случае, если референсные последовательности не транспортируются геномными потоками, они должны быть доступны на стороне декодирования и доступны декодерам классов.

[0051] В одном или более примерах раскрытые в настоящем документе способы изобретения могут быть реализованы в виде аппаратного обеспечения, программного обеспечения, прошивки или любой их комбинации. При реализации в программном обеспечении они могут храниться на компьютерном носителе и реализоваться аппаратным процессором. Аппаратный процессор может содержать один или несколько процессоров, процессоров цифровых сигналов, микропроцессоров общего назначения, специализированных интегральных схем или других дискретных логических схем.

Способы этого раскрытия могут быть реализованы в различных устройствах или приборах, включая мобильные телефоны, настольные компьютеры, серверы, планшеты и тому подобное.

[0052] Многие другие преимущества описаны в следующей формуле изобретения.

ФОРМУЛА ИЗОБРЕТЕНИЯ

1. Способ хранения представления данных геномной последовательности в формате геномного файла, причем указанные данные геномной последовательности содержат риды последовательностей нуклеотидов, включающий следующие этапы:
выравнивание указанных ридов с одной или более референсными последовательностями с получением выровненных ридов,
классификация указанных выровненных ридов в соответствии со степенями точности совпадения с указанной одной или более референсными последовательностямис получением классов выровненных ридов;
кодирование указанных классифицированных выровненных ридов в виде слоев синтаксических элементов,
структурирование указанных слоев синтаксических элементов с информацией заголовка с получением, таким образом, последовательных блоков доступа,
создание главной индексной таблицы, содержащей один раздел для каждого класса выровненных ридов, включающий положения картирования в указанной одной или более референсных последовательностях первого рида каждого блока доступа каждого класса данных;
совместное хранение указанной главной индексной таблицы и данных указанного блока доступа.

2. Способ по п.1, характеризующийся тем, что указанная главная индексная таблица дополнительно содержит вектор указателей на физическое положение каждого последующего блока доступа.

3. Способ по п.1, характеризующийся тем, что указанная главная индексная таблица дополнительно содержит один раздел для каждой референсной последовательности.

4. Способ по п.1, характеризующийся тем, что кодирование указанных классифицированных выровненных ридов в виде слоев синтаксических элементов адаптируют в соответствии с конкретными особенностями данных или метаданных, содержащихся в этом слое.

5. Способ по п.4, характеризующийся тем, что кодирование указанных классифицированных выровненных ридов в виде слоев синтаксических элементов дополнительно адаптируют в соответствии со статистическими свойствами данных или метаданных, содержащихся в этом слое.

6. Способ по п.5, характеризующийся тем, что кодирование указанных классифицированных выровненных ридов в виде слоев синтаксических элементов связывает модель источника дескриптора с конкретным энтропийным кодером.

7. Способ по п.6, характеризующийся тем, что модель источника, применяемая в одном из блоков доступа, не зависит от модели источника, используемой другими блоками доступа для того же слоя данных.

8. Способ извлечения ридов последовательностей нуклеотидов, хранящихся в геномном файле, при котором указанный геномный файл содержит главную индексную таблицу и блоки доступа, сохраненные в соответствии со способом по п.1 формулы изобретения,
причем указанный способ содержит следующие этапы:
получение команд от пользователя, идентифицирующих тип ридов, подлежащих извлечению,
выборку главной индексной таблицы из указанного геномного файла,
выборку блоков доступа, соответствующих указанному типу ридов, подлежащих извлечению,
восстановление указанных ридов последовательностей нуклеотидов
картирование указанных восстановленных блоков доступа на одной или более референсных последовательностях.

9. Способ по п.8, характеризующийся тем, что геномный файл дополнительно содержит одну или более референсных последовательностей.

10. Способ по п.8, характеризующийся тем, что одна или более референсных последовательностей подаются по внеполосному механизму.

11. Геномный секвенатор, содержащий:

модуль секвенирования генома 130, сконфигурированный для вывода ридов последовательностей нуклеотидов 131 из биологического образца,

модуль выравнивания 132, сконфигурированный для выравнивания указанных ридов с одной или более референсными последовательностями с получением, таким образом, выровненных ридов 133,

модуль классификации 134, сконфигурированный для классификации указанных выровненных ридов в соответствии с соответствующими степенями точности с указанной одной или более референсными последовательностями с получением, таким образом, классов выровненных ридов 135;

модуль кодирования 136, сконфигурированный для кодирования указанных классифицированных выровненных ридов в виде слоев синтаксических элементов 137,

модуль дальнейшего подразделения 138, сконфигурированный для структурирования указанных слоев синтаксических элементов с информацией заголовка с получением, таким образом, последовательных блоков доступа 139,

модуль обработки индексной таблицы 1310, сконфигурированный для создания главной индексной таблицы, содержащей один раздел для каждого класса выровненных ридов, содержащий положения картирования в референсной последовательности первого рида каждого блока доступа каждого класса данных;

модуль хранения 1312-1316, сконфигурированный для совместного хранения указанной главной индексной таблицы и указанных данных блока доступа 1311.

12. Секвенатор генома по п.8, характеризующийся тем, что главная индексная таблица дополнительно содержит вектор указателей на физическое положение каждого последующего блока доступа.

13. Секвенатор генома по п.8, характеризующийся тем, что кодирование указанных классифицированных выровненных ридов в виде слоев синтаксических элементов

адаптировано в соответствии с конкретными особенностями данных или метаданных, содержащихся в этом слое.

14. Экстрактор 140 для извлечения ридов последовательностей нуклеотидов, хранящихся в геномном файле, причем указанный геномный файл содержит главную индексную таблицу и данные блоков доступа, сохраненные в соответствии со способом по п.1 формулы изобретения,

причем указанный экстрактор 140 содержит:

средство пользовательского ввода 141, сконфигурированное для приема введенных пользователем параметров 142, идентифицирующих тип ридов, подлежащих извлечению,

средство выборки 143, сконфигурированное для выборки указанной главной индексной таблицы 144 из указанного геномного файла,

средство выборки 145, сконфигурированное для выборки блоков доступа 146, соответствующих указанному типу ридов, подлежащих извлечению,

средство восстановления 147, сконфигурированное для восстановления указанных ридов последовательностей нуклеотидов 148 с картированием указанных извлеченных блоков доступа по одной или более референсным последовательностям.

15. Машиночитаемый носитель, содержащий совокупность инструкций, которые при их выполнении на вычислительном устройстве приводят к выполнению этим вычислительным устройством способа по пп. 1-10.

ФОРМУЛА ИЗОБРЕТЕНИЯ

измененная по ст.34PCT, для рассмотрения на региональной фазе в ЕАПО

1. Реализуемый на компьютере способ хранения представления данных геномной последовательности в формате геномного файла, причем указанные данные геномной последовательности содержат риды последовательностей нуклеотидов, включающий следующие этапы:

- выравнивание указанных ридов с одной или более референсными последовательностями с получением, таким образом, выровненных ридов,
- классификация указанных выровненных ридов в соответствии с тем, имеет ли место идеально картирование по указанной одной или более референсным последовательностям, числом несовпадений с указанными одной или более референсными последовательностями, присутствием замен символов, присутствием вставок, или делеций и мягко обрезанных символов в указанных выровненных ридах относительно указанных одной или более референсных последовательностей, присутствием некартированных ридов, с получением, таким образом, классов выровненных ридов;
- кодирование указанных классифицированных выровненных ридов в виде слоев синтаксических элементов, причем указанные слои синтаксических элементов содержат однородные дескрипторы, которые уникальным образом идентифицируют указанные классифицированные выровненные риды,
- структурирование указанных слоев синтаксических элементов с информацией заголовка с получением, таким образом, последовательных блоков доступа,
- создание главной индексной таблицы, содержащей один раздел для каждого класса выровненных ридов, включающий положения картирования в указанной одной или более референсных последовательностях первого рида каждого блока доступа каждого класса данных;
- совместное хранение указанной главной индексной таблицы и данных указанного блока доступа.

2. Способ по п.1, характеризующийся тем, что указанная главная индексная таблица дополнительно содержит вектор указателей на физическое положение каждого последующего блока доступа.

3. Способ по п.1, характеризующийся тем, что указанная главная индексная таблица дополнительно содержит один раздел для каждой референсной последовательности.

4. Способ по п.1, характеризующийся тем, что кодирование указанных классифицированных выровненных ридов в виде слоев синтаксических элементов адаптируют в соответствии с однородными данными, содержащимися в этом слое.

5. Способ по п.4, характеризующийся тем, что кодирование указанных классифицированных выровненных ридов в виде слоев синтаксических элементов дополнительно адаптируют в соответствии со статистическими свойствами однородных данных, содержащихся в этом слое.

6. Способ по п.5, характеризующийся тем, что кодирование указанных классифицированных выровненных ридов в виде слоев синтаксических элементов связывает модель источника однородных данных с конкретным энтропийным кодером.

7. Способ по п.6, характеризующийся тем, что модель источника, применяемая в одном из блоков доступа, не зависит от модели источника, используемой другими блоками доступа для того же слоя данных.

8. Способ извлечения ридов последовательностей нуклеотидов, хранящихся в геномном файле, при котором указанный геномный файл содержит главную индексную таблицу и блоки доступа, сохраненные в соответствии со способом по п.1 формулы изобретения,
причем указанный способ содержит следующие этапы:

- получение команд от пользователя, идентифицирующих тип ридов, подлежащих извлечению,

- выборку главной индексной таблицы из указанного геномного файла,
- выборку блоков доступа, соответствующих указанному типу ридов, подлежащих извлечению,
- восстановление указанных ридов последовательностей нуклеотидов, картирующих указанные восстановленные блоки доступа на одной или более референсных последовательностях.

9. Способ по п.8, характеризующийся тем, что геномный файл дополнительно содержит одну или более референсных последовательностей.

10. Способ по п.8, характеризующийся тем, что одна или более референсных последовательностей подаются по внеполосному механизму.

11. Геномный секвенатор, содержащий:
- модуль секвенирования генома 130, сконфигурированный для вывода ридов последовательностей нуклеотидов 131 из биологического образца,
- модуль выравнивания 132, сконфигурированный для выравнивания указанных ридов с одной или более референсными последовательностями с получением, таким образом, выровненных ридов 133,
- модуль классификации 134, сконфигурированный для классификации указанных выровненных ридов в соответствии с тем, имеет ли место идеально картирование по указанной одной или более референсным последовательностям, числом несовпадений с указанными одной или более референсными последовательностями, присутствием замен символов, присутствием вставок, или делеций и мягко обрезанных символов в указанных выровненных ридах относительно указанных одной или более референсных последовательностей, присутствием некартированных ридов с указанными одной или более референсными последовательностями с получением, таким образом, классов выровненных ридов 135;
- модуль кодирования 136, сконфигурированный для кодирования указанных классифицированных выровненных ридов в виде слоев синтаксических элементов 137, причем указанные слои синтаксических элементов содержат

однородные дескрипторы, которые уникальным образом идентифицируют указанные классифицированные выровненные риды,

- модуль дальнейшего подразделения 138, сконфигурированный для структурирования указанных слоев синтаксических элементов с информацией заголовка с получением, таким образом, последовательных блоков доступа 139,

- модуль обработки индексной таблицы 1310, сконфигурированный для создания главной индексной таблицы, содержащей один раздел для каждого класса выровненных ридов, содержащий положения картирования в референсной последовательности первого рида каждого блока доступа каждого класса данных;

- модуль хранения 1312-1316, сконфигурированный для совместного хранения указанной главной индексной таблицы и указанных данных блока доступа 1311.

12. Секвенатор генома по п.8, характеризующийся тем, что главная индексная таблица дополнительно содержит вектор указателей на физическое положение каждого последующего блока доступа.

13. Секвенатор генома по п.8, характеризующийся тем, что кодирование указанных классифицированных выровненных ридов в виде слоев синтаксических элементов адаптировано в соответствии с гомогенными данными, содержащимся в этом слое.

14. Экстрактор 140 для извлечения ридов последовательностей нуклеотидов, хранящихся в геномном файле, причем указанный геномный файл содержит главную индексную таблицу и данные блоков доступа, сохраненные в соответствии со способом по п.1 формулы изобретения,

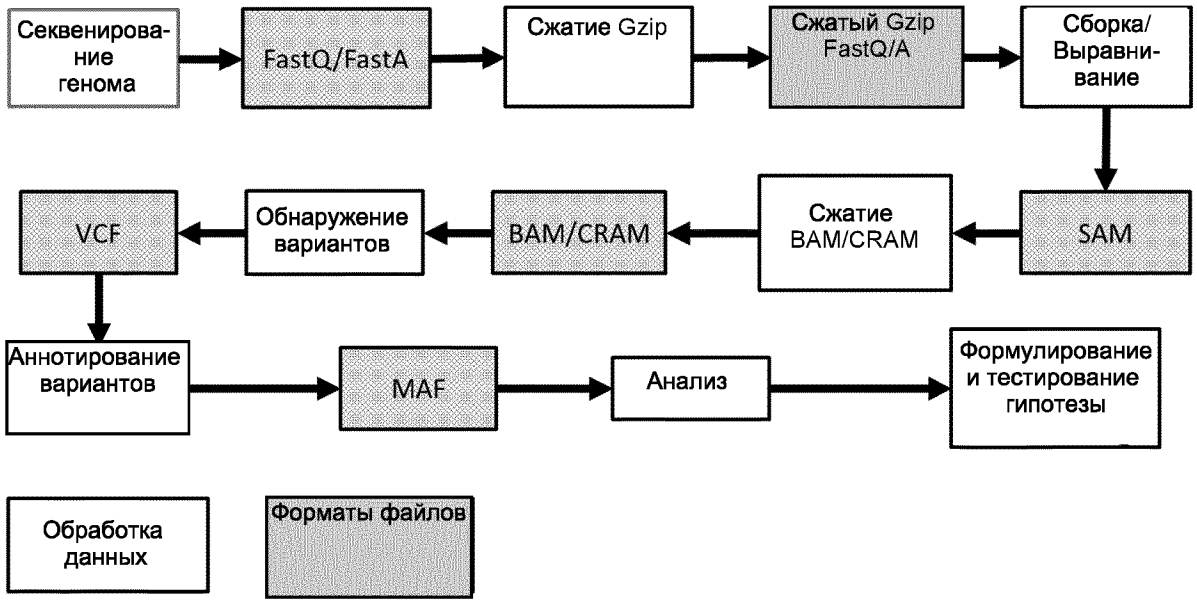
причем указанный экстрактор 140 содержит:

- средство пользовательского ввода 141, сконфигурированное для приема введенных пользователем параметров 142, идентифицирующих тип ридов, подлежащих извлечению,

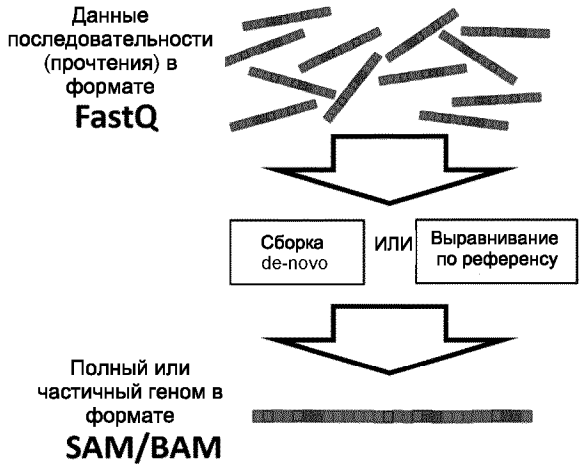
- средство выборки 143, сконфигурированное для выборки указанной главной индексной таблицы 144 из указанного геномного файла,

- средство выборки 145, сконфигурированное для выборки блоков доступа 146, соответствующих указанному типу ридов, подлежащих извлечению,
- средство восстановления 147, сконфигурированное для восстановления указанных ридов последовательностей нуклеотидов 148 с картированием указанных извлеченных блоков доступа по одной или более референсным последовательностям.

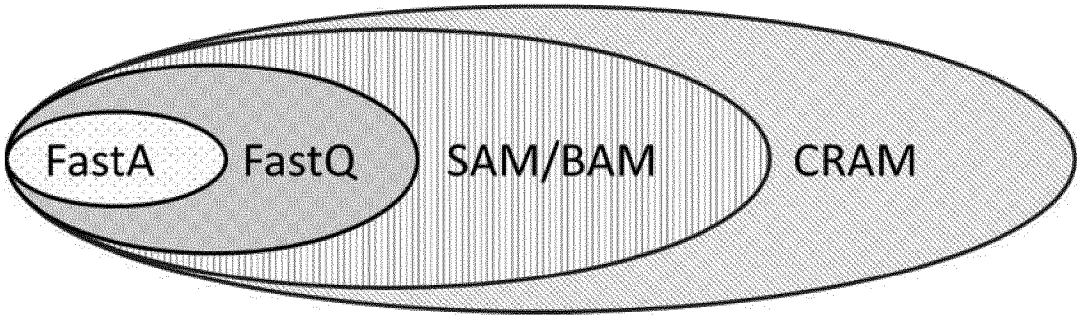
15. Машиночитаемый носитель, содержащий совокупность инструкций, которые при их выполнении на вычислительном устройстве приводят к выполнению этим вычислительным устройством способа по пп. 1-10.



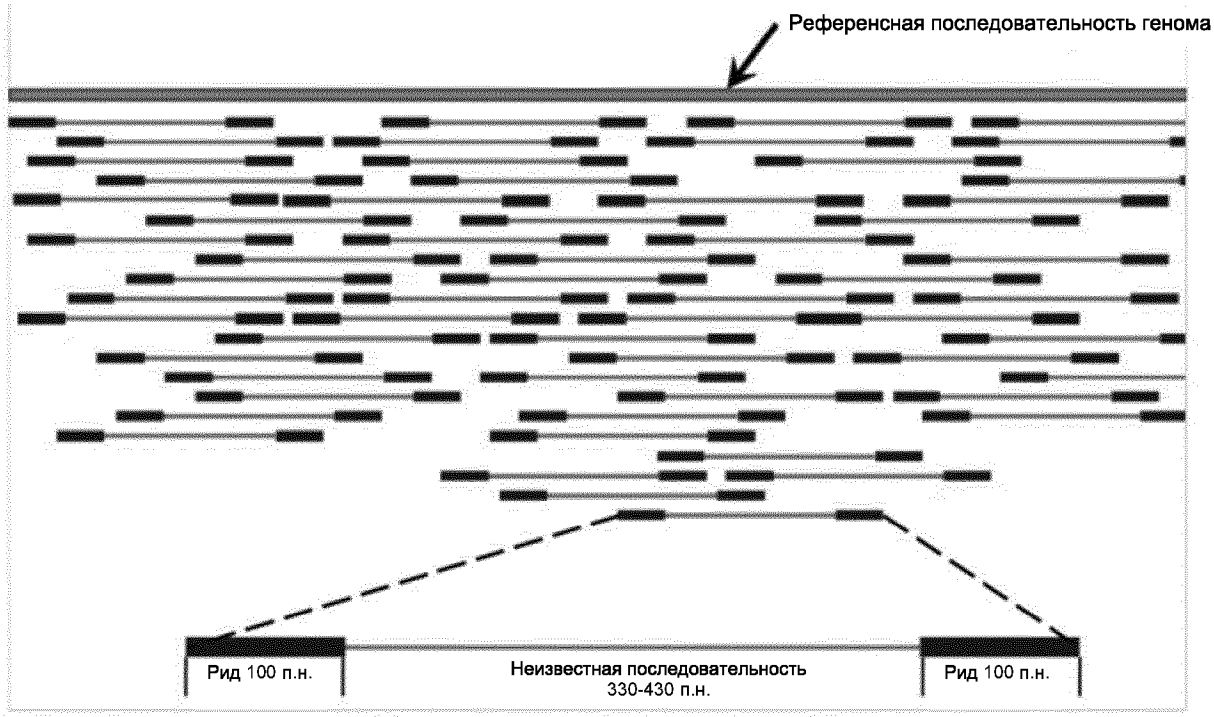
Фигура 1- Жизненный цикл геномных данных от секвенирования до анализа и формулирования гипотезы.



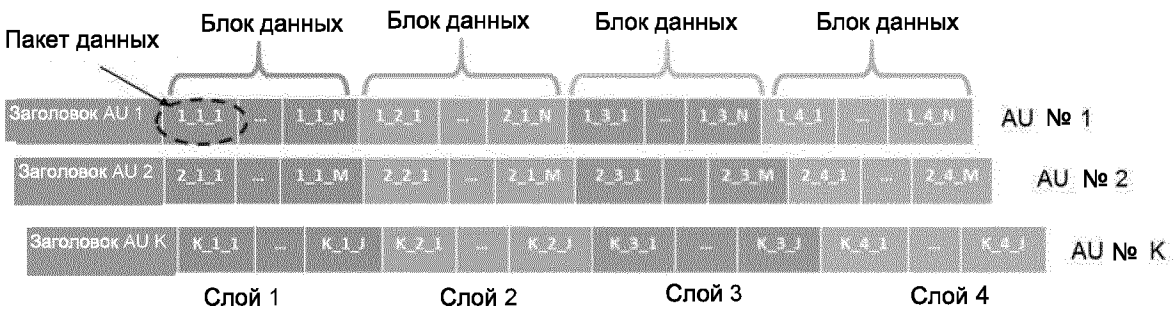
Фигура 2 - Выравнивание прочтений геномной последовательности.



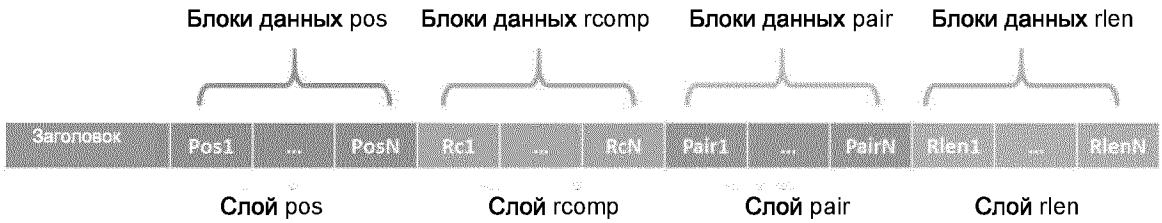
Фигура 3 - Форматы геномных файлов и их связь.



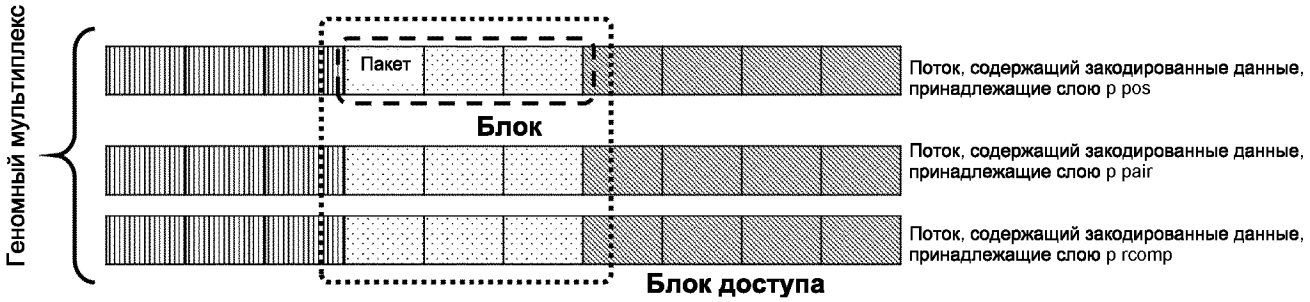
Фигура 4 - Пары ридов, картированные на референсной последовательности (геном).



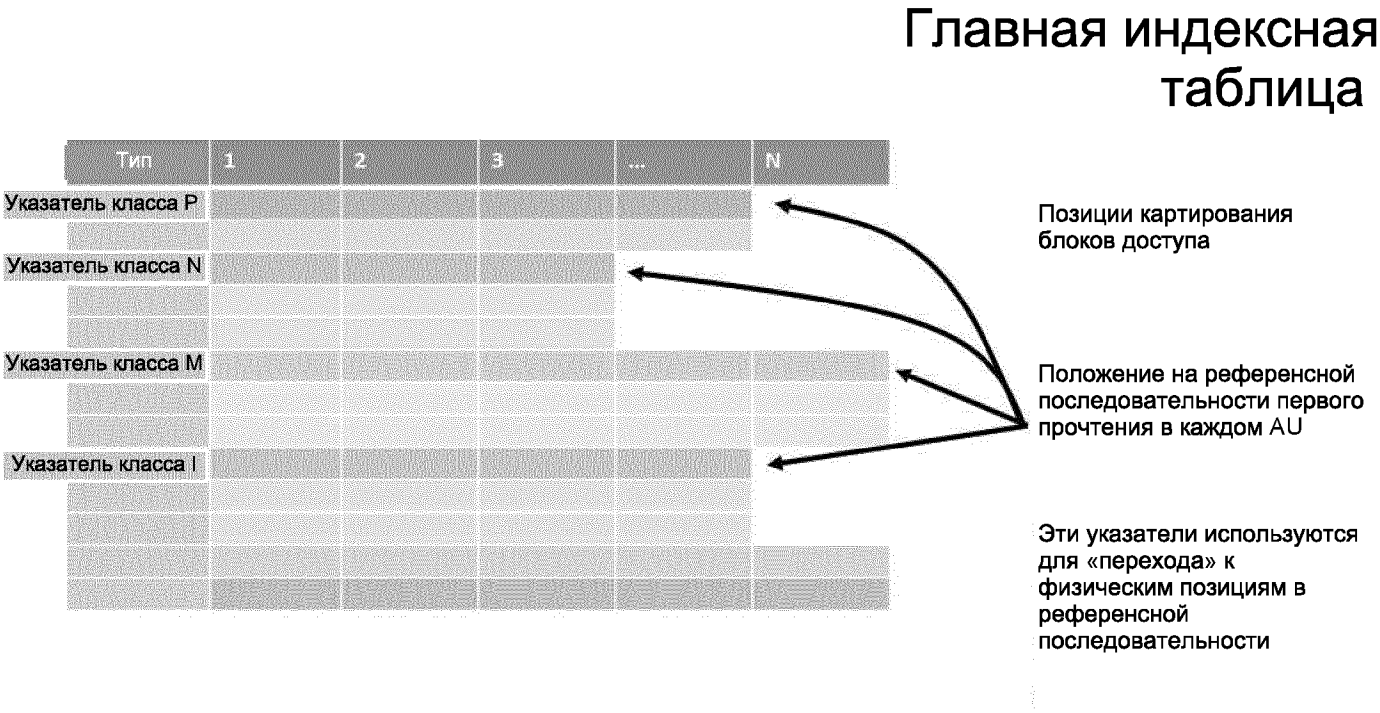
Фигура 5 – Блоки доступа состоят из блоков данных. Каждый блок данных содержит пакеты дескрипторов, принадлежащих к одному типу данных.



Фигура 6 - Пример блока доступа, включающего заголовок и слои, состоящие из блоков данных.

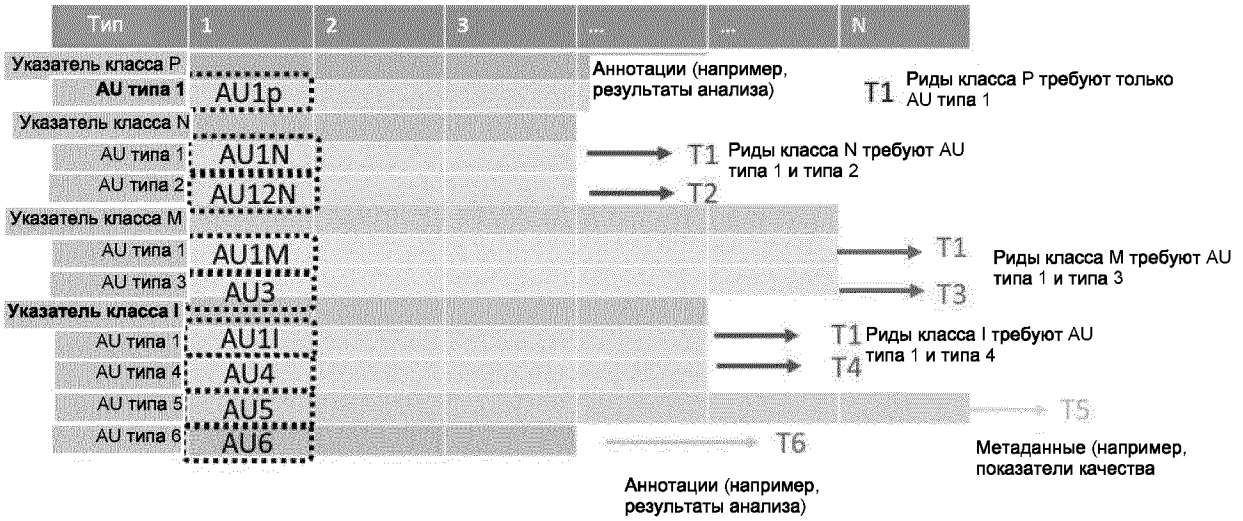


Фигура 7 - Связь между геномными «пакетами данных», «блоками», блоками доступа, слоями и классами потоков ридов



Фигура 8 - Данные первого типа, содержащиеся в MIT, представляют собой векторы локусов картирования первого рида содержащиеся в каждом модуле доступа.

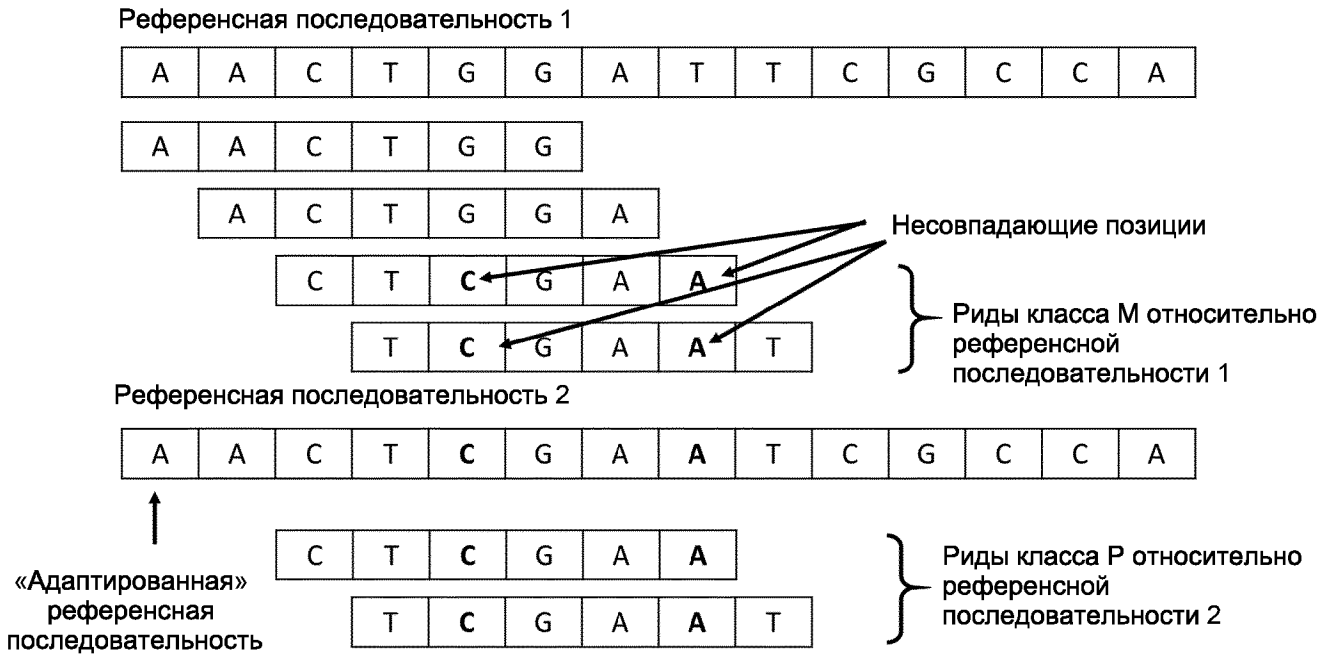
Фигура 9 - Общая структура основного заголовка и частичное представление MIT, показывающие позиции картирования первого рида в каждом AU pos класса P.



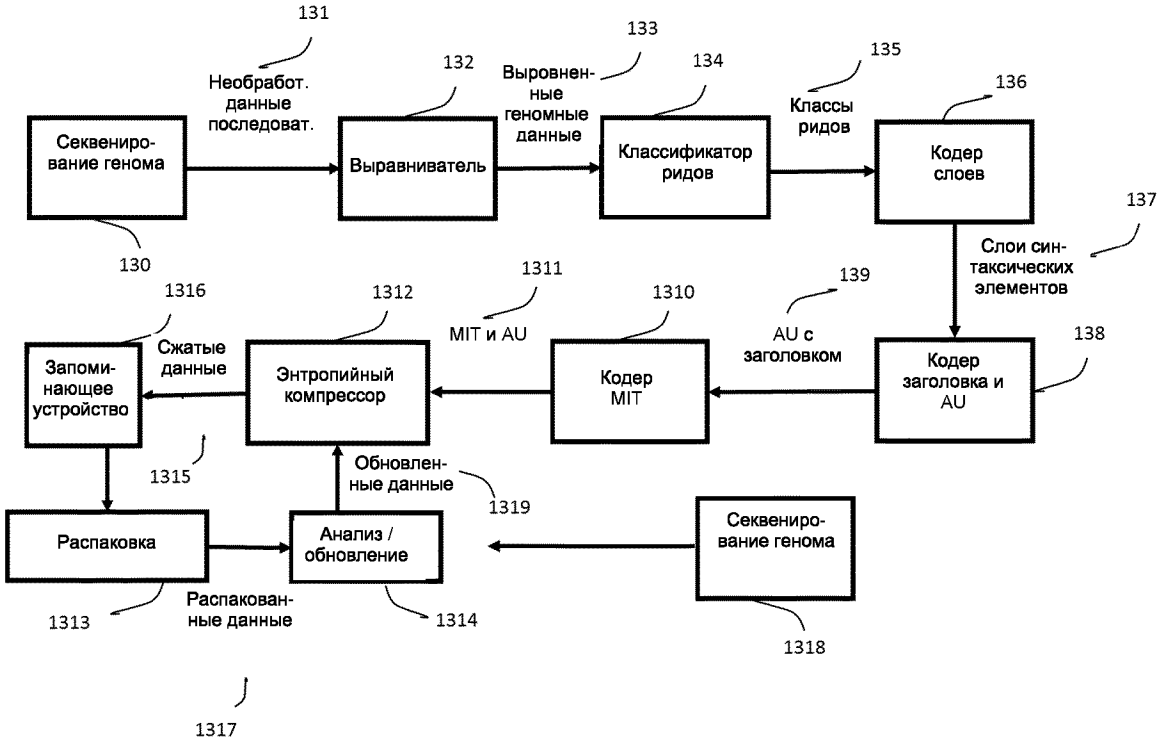
Фигура 10 – Данные второго типа, содержащиеся в MIT, представляют собой векторы указателей на физическое местоположение каждого модуля доступа. Эти векторы называются локальными индексными таблицами.

Полож. класса P	10000	73987	123897	259820	12078	62656	130298	235019	359333
T1p									
полож. класса N	p11N	p12N	p13N	p14N	p21N	p22N	p23N	p24N	p25N
T1N									
T2									
полож. класса M	p11M	p12M	p13M	p14M	p21M	p22M	p23M	p24M	p25M
T1M									
T3									
Полож класса I	p11I	p12I	p13I	p14I	p21I	p22I	p23I	p24I	p25I
T1I									
T4									

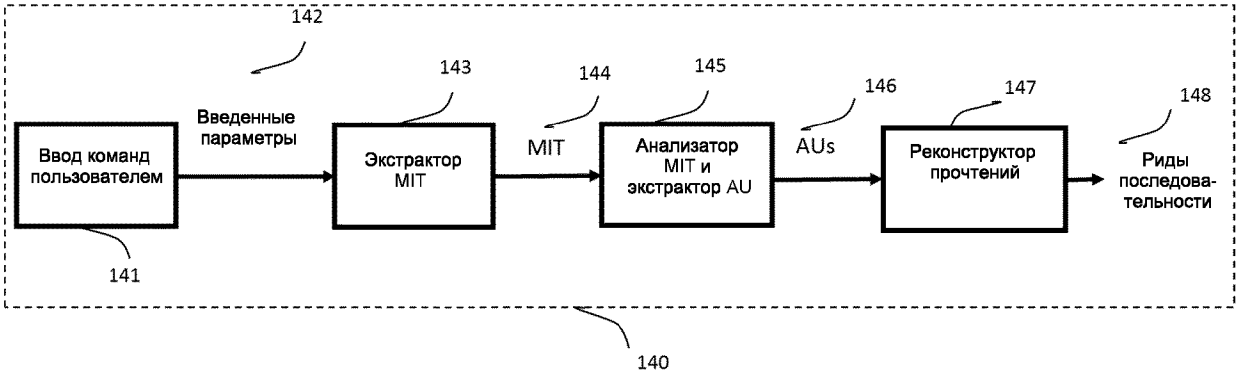
Фигура 11 - Доступ к блокам доступа, содержащим риды класса P, картированные по референсной последовательности № 2 между позициями 150000 и 250000, осуществляется с использованием значений, содержащихся в векторе T1p.



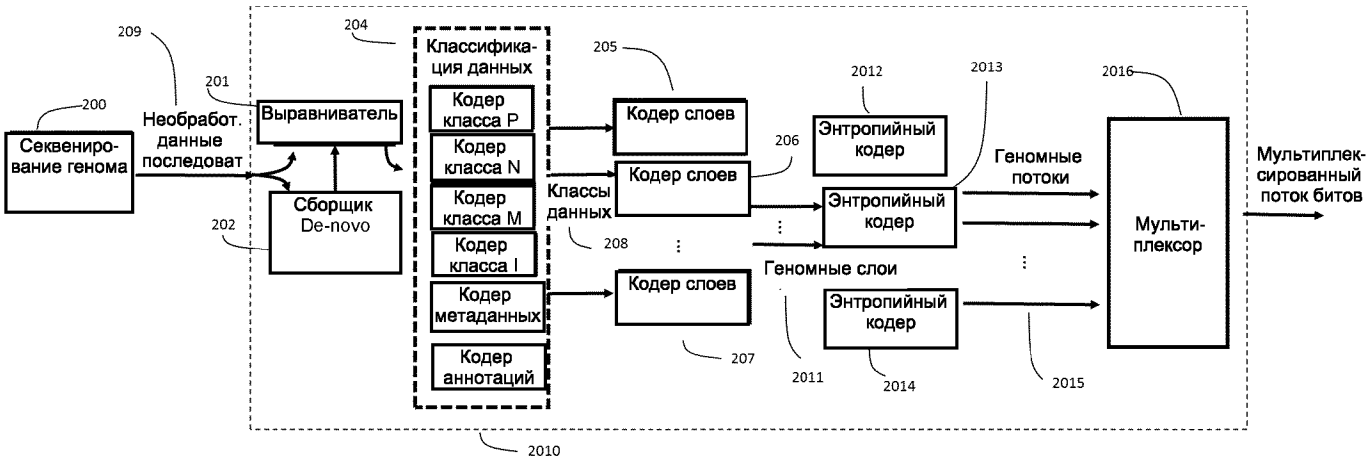
Фигура 12 - Модификация в референсной последовательности может преобразовать M-риды в P-риды.



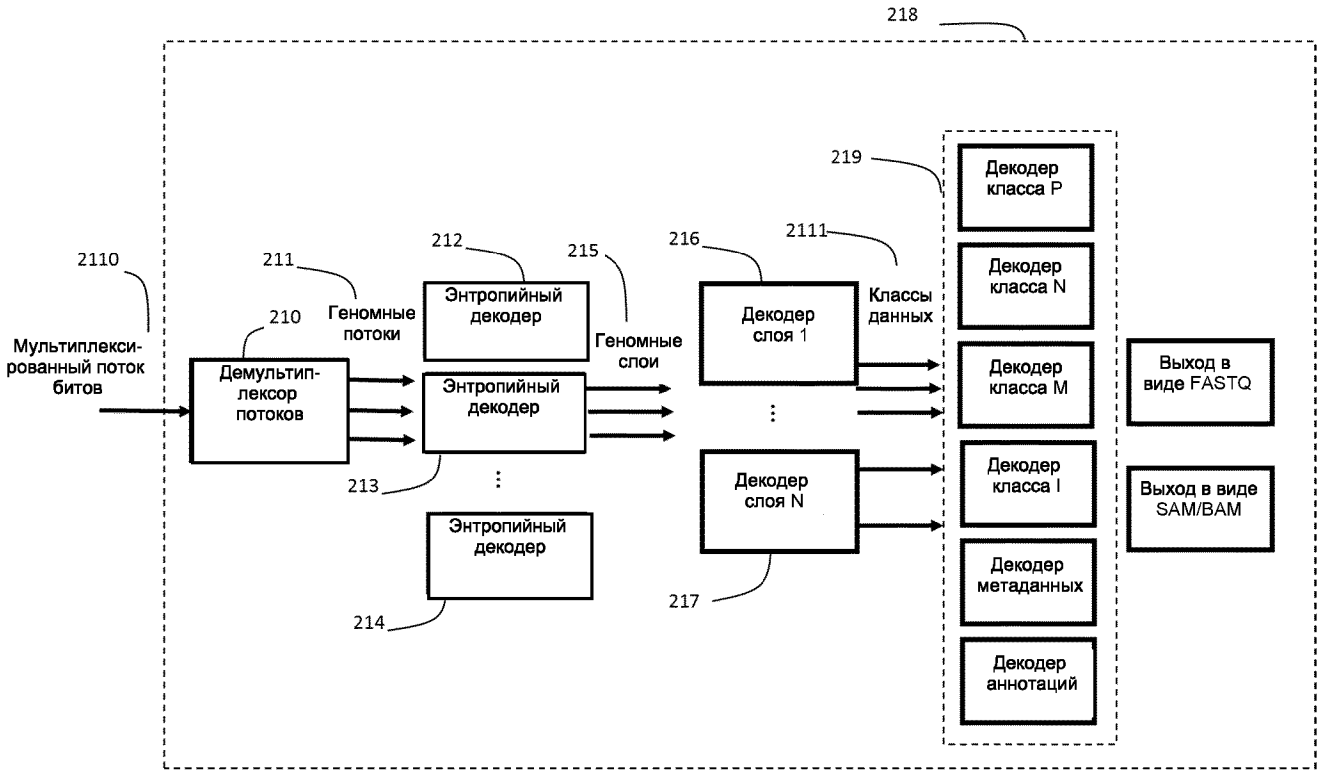
Фигура 13 - Общая архитектура.



Фигура 14 – Экстрактор прочтений



Фигура 15 - Геномный кодер



Фигура 16 - Геномный декодер