(19) Евразийское патентное ведомство

(11) 040022 (13) B1

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ЕВРАЗИЙСКОМУ ПАТЕНТУ

(45) Дата публикации и выдачи патента
2022.04.11

(21) Номер заявки
201990935

(22) Дата подачи заявки
2016.10.11

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

(54) СПОСОБ И УСТРОЙСТВО ДЛЯ КОМПАКТНОГО ПРЕДСТАВЛЕНИЯ ДАННЫХ БИОИНФОРМАТИКИ

(43) **2019.11.29**
(86) **PCT/EP2016/074307**
(87) **WO 2018/068829 2018.04.19**

(71)(73) Заявитель и патентовладелец:
ДЖЕНОМСИС СА (CH)

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

(74) Представитель:
Нилова М.И. (RU)

(56) US-A1-2015227686
Anonymous: "SAM", 11 March 2015 (2015-03-11), XP002771304, Retrieved from the Internet: URL:https://web.archive.org/web/20150311045750/http://davetang.org/wiki/tiki-index.php?page=SAM [retrieved on 2017-06-22], section "Using SAM to store various types of alignments", pg. 4
Anonymous: "CRAM format specification (version 3.0)", 8 September 2016 (2016-09-08), XP002771305, Retrieved from the Internet: URL:https://samtools.github.io/hts-specs/CRAMv3.pdf [retrieved on 2017-06-22], pg. 13

(57) Изобретение относится к области биоинформатики. Предложен реализуемый на компьютере способ сжатия данных геномной последовательности, сгенерированных секвенаторами генома. Последовательности нуклеотидов выравнивают по одной или более референсным последовательностям, классифицируют в соответствии со степенями точности совпадения, кодируют в виде множества слоев синтаксических элементов, используя разные модели источников и энтропийные кодеры для каждого слоя, на которые разбиты данные. Дополнительно предложены геномный кодер и геномный декодер (фиг. 21), а также машиночитаемый носитель, реализующие способ согласно изобретению. Изобретение позволяет уменьшить используемый объем памяти и улучшить производительность доступа, предоставляя новые функциональные возможности, которые недоступны с известными из уровня техники способами представления.

040022 B1

040022 B1

Область техники

Изобретение обеспечивает новый способ представления данных секвенирования генома, который уменьшает используемый объем памяти и улучшает производительность доступа, предоставляя новые функциональные возможности, которые недоступны с известными из уровня техники способами представления.

Уровень техники

Надлежащее представление данных секвенирования генома имеет основополагающее значение для обеспечения эффективных приложений для геномного анализа, таких как определение вариантов генома и любой другой анализ, выполняемый с различными целями путем обработки данных и метаданных секвенирования.

Секвенирование генома человека стало доступным благодаря появлению высокопроизводительных технологий секвенирования с низкой стоимостью. Такая возможность открывает новые перспективы в нескольких областях: от диагностики и лечения рака до выявления генетических заболеваний, от наблюдения за патогенами для выявления антител до создания новых вакцин, лекарств и персонализации индивидуального лечения.

Больницы, поставщики услуги геномного анализа, специалисты по биоинформатике и крупные центры хранения биологических данных ищут недорогие, быстрые, надежные и взаимосвязанные решения для обработки геномной информации, которые позволили бы расширить масштабы геномной медицины до мирового масштаба. Поскольку одним из узких мест в процессе секвенирования стало хранение данных, все интенсивнее исследуются способы представления данных секвенирования генома в сжатой форме.

Наиболее часто используемые представления информации о геноме на основе данных секвенирования основаны на форматах FASTQ и SAM. Задача состоит в том, чтобы сжать традиционно используемые форматы файлов (соответственно FASTQ и SAM для невыровненных и выровненных данных). Такие файлы состоят из простых текстовых символов и сжимаются, как упоминалось выше, с использованием подходов общего назначения, таких как схемы LZ (от имен Lempel и Ziv, авторов, опубликовавших первые версии) (хорошо известные zip, gzip и т.д.). С использованием компрессоров общего назначения, таких как gzip, результатом сжатия обычно является один блоб-объект бинарных данных. Информацию в такой монолитной форме очень трудно архивировать, передавать и обрабатывать, особенно когда, как в случае высокопроизводительного секвенирования, объем данных чрезвычайно велик. Формат BAM характеризуется низкой производительностью сжатия из-за того, что он сосредоточен на сжатии неэффективного и избыточного формата SAM, а не на извлечении фактической геномной информации, передаваемой файлами SAM, и из-за применения алгоритмов сжатия текста общего назначения, таких как gzip, вместо того, чтобы использовать специфическую природу каждого источника данных (самих геномных данных).

CRAM - более сложный подход к сжатию геномных данных, который меньше используется, но более эффективен, чем BAM. CRAM обеспечивает более эффективное сжатие для применения дифференциального кодирования по отношению к существующему референсу (он частично использует избыточность источника данных), но ему все еще не хватает таких функций, как инкрементные обновления, поддержка потоковой передачи и выборочный доступ к определенным классам сжатых данных. Эти подходы дают неудовлетворительные коэффициенты сжатия и структуры данных, в которых трудно ориентироваться и манипулировать ими после сжатия. Последующий анализ может быть очень медленным из-за необходимости обработки больших и жестких структур данных даже для выполнения простой операции или для доступа к выбранным областям набора геномных данных. CRAM опирается на концепцию записи CRAM. Каждая запись CRAM кодирует одно картированное или некартированный рид, кодируя все элементы, необходимые для его восстановления.

CRAM имеет следующие недостатки.

1) Для CRAM индексация данных выходит за рамки спецификации (см. раздел 12 спецификации CRAM v 3.0) и реализована в виде отдельного файла. В подходе изобретения, описанного в этом документе, напротив, применяется метод индексации данных, который интегрирован с процессом кодирования, и индексы внедрены в кодированный поток битов.

2) В CRAM все основные блоки данных могут содержать любой тип картированных ридов (идеально совпадающих ридов, ридов только с заменами, ридов с инсерциями или делециями (также называемых "инделами")). Не существует понятия классификации и группировки ридов в классах по результату картирования относительно референсной последовательности.

3) В изобретении нет понятия записи, инкапсулирующей каждый рид, потому что данные, необходимые для восстановления каждого риды, распределены по множеству контейнеров данных, называемых "слоями". Это обеспечивает более эффективный доступ к набору ридов с определенными биологическими характеристиками (например, ридов с заменами, но без "инделов" или идеально картированных ридов) без необходимости декодирования каждого (блока) рида (дидов) для проверки его признаков.

4) В записи CRAM каждый тип данных обозначается определенным флагом. В настоящем изобретении нет понятия флага, обозначающего данные, потому что это по сути определяется "слоем", к которому принадлежат данные. Это подразумевает значительно уменьшенное количество используемых сим-

волов и, как следствие, уменьшение энтропии источника информации, что приводит к более эффективному сжатию. Это связано с тем фактом, что использование разных "слоев" позволяет кодеру повторно использовать один и тот же символ в каждом слое с разными значениями. В CRAM каждый флаг должен всегда иметь одно и то же значение, поскольку отсутствует понятие контекстов, и каждая запись CRAM может содержать данные любого типа.

5) В заменах CRAM инсерции и делеции выражаются в соответствии с различными синтаксисами, тогда как в предлагаемом подходе используется один алфавит и кодирование для замен, инсерций и делеций. Это упрощает процесс кодирования и декодирования и создает модель источника с более низкой энтропией, кодирование которой дает битовые потоки с высокой степенью сжатия.

Изобретение направлено на сжатие геномных последовательностей за счет организации и разбивки данных таким образом, чтобы свести к минимуму избыточную информацию, подлежащую кодированию, и реализовать такие функции, как выборочный доступ и поддержка инкрементных обновлений. Одним из аспектов представленного подхода является определение классов данных и метаданных, которые должны кодироваться отдельно и которые должны быть структурированы в разных слоях. Наиболее важные улучшения этого подхода относительно существующих методов заключаются в следующем:

1) увеличение эффективности сжатия из-за уменьшения энтропии источника информации, созданной путем предоставления эффективной модели для каждого класса данных или метаданных;

2) возможность выполнения выборочного доступа к порциям сжатых данных и метаданных для дальнейшей обработки;

3) возможность инкрементного (без необходимости перекодирования) обновления закодированных данных и метаданных новыми данными и/или метаданными о последовательности и/или новыми результатами анализа.

Краткое описание чертежей

На фиг. 1 - показано, что положение пар картированных ридов кодируется в слое pos как разница с абсолютным положением первого картированного рида;

фиг. 2 - два рида в паре могут происходить из двух цепей ДНК;

фиг. 3 - обратный комплемент рида 2 будет закодирован таким образом, если в качестве референса используется цепь 1;

фиг. 4 - четыре возможных комбинации ридов, составляющих пару ридов и соответствующую кодировку в слое rcomp;

фиг. 5 - схема расчета расстояния спаривания при постоянной длине ридов для трех пар ридов;

фиг. 6 - ошибки спаривания, закодированные в слое pair, позволяют декодеру восстановить правильное спаривание ридов с использованием закодированного MPPPD;

фиг. 7 - кодирование расстояния спаривания, когда рид картируется на другом референсе, чем его партнер по паре. В этом случае к расстоянию спаривания добавляются дополнительные дескрипторы. Один - это сигнальный флаг, второй - идентификатор референса, а затем расстояние спаривания;

фиг. 8 - кодирование N-несовпадений в слое nmis;

фиг. 9 - картированная пара ридов, которая представляет замены относительно референсной последовательности;

фиг. 10 - расчет положений замен как в абсолютных, так и в дифференциальных значениях;

фиг. 11 - расчет символов, кодирующих типы замен, когда коды IUPAC не используются. Символы представляют расстояние - в круговом векторе замен - между молекулой, присутствующей в риде, и молекулой, присутствующей на референсе в этом положении;

фиг. 12 - кодирование замен в слое snpt;

фиг. 13 - расчет кодов замен с использованием кодов неоднозначности IUPAC;

фиг. 14 - кодирование слоя snpt с использованием кодов IUPAC;

фиг. 15 - для ридов класса I используемый вектор замен такой же, как для класса M с добавлением специальных кодов для инсерций символов A, C, G, T, N;

фиг. 16 - некоторые примеры кодирования несовпадений и инделов в случае использования кодов неоднозначности IUPAC. Вектор замен в этом случае намного длиннее, и поэтому возможных рассчитанных символов больше, чем в случае пяти символов;

фиг. 17 - другая модель источника для несовпадений и инделов, где каждый слой содержит положение несовпадений или инсерций одного типа. В этом случае символы для типа несовпадения или индела не кодируются;

фиг. 18 - пример кодирования несовпадений и инделов. Если для рида нет несовпадений или инделов данного типа, в соответствующем слое кодируется значение 0. Символ 0 действует как разделитель и терминатор ридов в каждом слое;

фиг. 19 - модификация в референсной последовательности может преобразовать M-рида в P-рида. Эта операция может уменьшить информационную энтропию структуры данных, особенно в случае высокого перекрывания;

фиг. 20 - геномный кодер 2010 согласно одному варианту реализации этого изобретения;

фиг. 21 - геномный декодер 218 согласно одному варианту реализации этого изобретения.

Краткое описание

Изобретение, в частности, относится к следующим объектам.

1) Реализуемый на компьютере способ сжатия данных геномной последовательности, где 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 различные классы, включающие по меньшей мере

первый класс: если указанные выровненные риды совпадают с указанными одной или более референсными последовательностями без каких-либо ошибок;

второй класс: если указанные выровненные риды совпадают с областью в указанной одной или более референсных последовательностях с числом несовпадений, состоящим из числа положений, в которых секвенатор не смог определить ни одного основания;

третий класс: если указанные выровненные риды совпадают с областью в указанной одной или более референсных последовательностях с числом несовпадений, состоящим из положений, в которых секвенатор не смог определить ни одного основания, и присутствия инсерций, или делеций, или обрезанных нуклеотидов;

четвертый класс: если указанные выровненные риды не находят какого-либо достоверного картирования на указанной одной или более референсных последовательностях в соответствии с указанными ограничениями выравнивания с получением таким образом классов выровненных ридов; и кодирование указанных классифицированных и выровненных ридов в виде множества слоев синтаксических элементов, содержащих дескрипторы, причем для указанного первого класса указанные дескрипторы включают по меньшей мере начальное положение в референсной последовательности, флаг, сигнализирующий о том, что рид должен рассматриваться как обратный комплемент к референсу, расстояние до партнера пары в случае спаренных ридов, значение длины в случае, когда технология секвенирования дает риды переменной длины; для указанного второго класса дескрипторы включают по меньшей мере дескрипторы первого класса и положение несовпадения для каждого несовпадения; для указанного третьего класса дескрипторы включают дескрипторы указанного второго класса и положение несовпадения и тип несовпадения для каждого несовпадения; для указанного четвертого класса дескрипторы включают дескрипторы указанного первого класса и тип несовпадения для каждого несовпадения,

где указанное кодирование указанных классифицированных выровненных ридов в виде множества слоев синтаксических элементов включает выбор указанных синтаксических элементов, содержащих указанные дескрипторы, в соответствии с указанными классами выровненных ридов, причем кодирование указанных классифицированных выровненных ридов в виде множества слоев синтаксических элементов осуществляется с применением конкретного энтропийного кодера 2012-2014.

2) Способ по п.1, характеризующийся тем, что слои синтаксических элементов дополнительно содержат положение варианта относительно референсной последовательности, тип варианта, положение делеции относительно референсной последовательности, положение одного или более символов, отсутствующих в референсной последовательности, но присутствующих в выровненных ридах, тип инсерции в данном положении.

3) Способ по п.1, характеризующийся тем, что указанный энтропийный кодер является контекстно-адаптивным арифметическим кодером.

4) Способ распаковки геномного потока, сжатого способом по п.1, причем указанный способ включает следующие этапы:

синтаксический анализ и декодирование 212-214 сжатого геномного потока в геномные слои синтаксических элементов 215,

декодирование указанных геномных слоев в классы данных 216-217, разворачивание указанных геномных слоев в классифицированные риды последовательностей нуклеотидов,

выборочное декодирование с помощью декодеров 219 классов указанных классифицированных ридов последовательностей нуклеотидов и объединение результатов на одной или более референсных последовательностях с получением несжатых ридов последовательностей нуклеотидов.

5) Геномный кодер 2010 для сжатия данных геномной последовательности 209, причем указанные данные геномной последовательности 209 содержат риды последовательностей нуклеотидов, причем указанный геномный кодер 2010 содержит

модуль выравнивания 201, сконфигурированный для выравнивания указанных ридов с одной или более референсными последовательностями с получением, таким образом, выровненных ридов,

модуль классификации данных 204, сконфигурированный для классификации указанных выровненных ридов в различные классы, включающие по меньшей мере

первый класс: если указанные выровненные риды совпадают с указанными одной или более референсными последовательностями без каких-либо ошибок;

второй класс: если указанные выровненные риды совпадают с областью в указанной одной или более референсных последовательностях с числом несовпадений, состоящим из числа положений, в кото-

рых секвенатор не смог определить ни одного основания;

третий класс: если указанные выровненные риды совпадают с областью в указанной одной или более референсных последовательностях с числом несовпадений, состоящим из положений, в которых секвенатор не смог определить ни одного основания, и присутствия инсерций, или делеций, или обрезанных нуклеотидов;

четвертый класс: если указанные выровненные риды не находят какого-либо достоверного картирования на указанной одной или более референсных последовательностях в соответствии с указанными ограничениями выравнивания с получением, таким образом, классов выровненных ридов;

один или более кодирующих слои модулей 205-207, сконфигурированных для кодирования указанных классифицированных выровненных ридов в виде слоев синтаксических элементов, содержащих дискрипторы, путем выбора указанных синтаксических элементов в соответствии с указанными классами выровненных ридов, где для указанного первого класса указанные дескрипторы включают по меньшей мере начальное положение в референсной последовательности, флаг, сигнализирующий о том, что рид должен рассматриваться как обратный комплемент к референсу, расстояние до партнера пары в случае спаренных ридов, значение длины в случае, когда технология секвенирования дает риды переменной длины; для указанного второго класса дескрипторы включают по меньшей мере дескрипторы первого класса и положение несовпадения для каждого несовпадения; для указанного третьего класса дескрипторы включают дескрипторы указанного второго класса и положение несовпадения и тип несовпадения для каждого несовпадения; для указанного четвертого класса дескрипторы включают дескрипторы указанного первого класса и тип несовпадения для каждого несовпадения,

энтропийный кодер 2012-2014 для энтропийного кодирования указанных слоев синтаксических элементов.

6) Геномный декодер 218 для распаковки геномного потока 211, сжатого геномным кодером по п.6, причем указанный геномный декодер 218 содержит

средства синтаксического анализа и декодирования 210, 212-214, сконфигурированные для синтаксического анализа указанного сжатого геномного потока в геномные слои синтаксических элементов 215,

один или более декодеров слоев 216-217, сконфигурированных для декодирования геномных слоев в классы данных, и дополнительно сконфигурированный для обработки указанных геномных слоев в классифицированные риды последовательностей нуклеотидов 2111,

декодеры классов геномных данных 213, сконфигурированные для выборочного декодирования указанных классифицированных ридов последовательностей нуклеотидов и объединения результата по одной или нескольким референсным последовательностям с получением несжатых ридов последовательностей нуклеотидов.

7) Геномный декодер по п.6, характеризующийся тем, что одна или более референсных последовательностей хранятся в сжатом потоке генома 211.

8) Геномный декодер по п.6, характеризующийся тем, что одна или более референсных последовательностей подаются в декодер по внеполосному механизму.

9) Геномный декодер по п.6, характеризующийся тем, что одна или более референсных последовательностей строятся в указанном декодере.

10) Машиночитаемый носитель, содержащий инструкции, которые при их выполнении приводят к осуществлению по меньшей мере одним процессором способа по любому из пп.1-4.

Признаки независимых пунктов формулы изобретения, приведенные ниже и выше, решают проблему существующих решений предшествующего уровня техники, обеспечивая способ классификации последовательностей генома и способ сжатия с использованием указанной классификации. В одном аспекте предложен способ классификации данных геномной последовательности, генерируемых секвенатором, причем указанные данные геномной последовательности содержат последовательности "оснований" нуклеотидов, причем указанная классификация выполняется в соответствии с референсной последовательностью, причем указанный способ включает следующие этапы:

идентификация последовательностей класса P, содержащих совпадающие (matching) области в референсной последовательности без несовпадений; идентификация последовательностей класса N, содержащих совпадающие области в референсной последовательности, с рядом несовпадений, представленных положениями, в которых секвенатор не смог определить никакое основание; идентификация последовательностей класса M, содержащих совпадающие области в референсной последовательности, с рядом несовпадений, представленных положениями, в которых секвенатор не смог определить никакое основание или определил основание, отличное от основания в референсной последовательности; идентификация последовательностей класса I, содержащих несовпадения класса M плюс присутствие инсерций или делеций;

идентификация последовательностей класса U, содержащих все риды, которые не находят действительного (валидного) картирования в референсной последовательности.

В другом аспекте предложен способ сжатия данных геномной последовательности, генерируемых секвенатором, причем указанные данные геномной последовательности содержат последовательности

нуклеотидов, причем указанный способ включает следующие этапы:

выравнивание указанных ридов по референсной последовательности с получением выровненных ридов;

классификация указанных выровненных ридов в соответствии с множеством (multiplicity) степеней совпадения с референсной последовательностью, тем самым создавая классы выровненных ридов;

кодирование указанных выровненных ридов в виде слоев синтаксических элементов;

причем указанные элементы синтаксиса выбирают в соответствии с указанными классами выровненных ридов.

В другом аспекте предложен способ распаковки сжатого геномного потока, причем указанный способ включает следующие этапы:

синтаксический анализ указанного сжатого геномного потока в геномные слои синтаксических элементов,

разворачивание указанных геномных слоев в классифицированные риды последовательностей нуклеотидов,

выборочное декодирование указанных классифицированных ридов последовательностей нуклеотидов на основании одной или более референсных последовательностей с получением несжатых ридов последовательностей нуклеотидов.

Еще один аспект - геномный кодер 2010 для сжатия данных геномной последовательности 209, причем указанные данные геномной последовательности 209 содержат риды последовательностей нуклеотидов, причем указанный геномный кодер 2010 содержит

модуль выравнивания 201, сконфигурированный для выравнивания указанных ридов с одной или более референсными последовательностями с получением, таким образом, выровненных ридов,

модуль классификации данных 204, сконфигурированный для классификации указанных выровненных ридов в соответствии со степенями точности совпадения совпадения с одной или более референсными последовательностями с получением, таким образом, классов выровненных ридов;

один или более кодирующих слои модулей 205-207, сконфигурированных для кодирования указанных классифицированных выровненных ридов в виде слоев синтаксических элементов путем выбора указанных синтаксических элементов в соответствии с указанными классами выровненных ридов.

В другом аспекте, геномный декодер 218 для распаковки сжатого геномного потока 211, причем указанный геномный декодер 218 содержит:

средства синтаксического анализа (парсинга) 210, 212-214, сконфигурированные для анализа указанного сжатого геномного потока в геномные слои синтаксических элементов 215,

один или более декодеров слоев 216-217, сконфигурированных для декодирования геномных слоев в классифицированные риды последовательностей нуклеотидов 2111,

декодеры классов геномных данных 213, сконфигурированные для выборочного декодирования указанных классифицированных ридов последовательностей нуклеотидов по одной или более референсным последовательностям с получением несжатых ридов последовательностей нуклеотидов.

Подробное описание

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.

Информация секвенирования генома генерируется высокопроизводительными секвенаторами (HTS) в виде последовательностей нуклеотидов ("оснований"), представленных строками букв из определенного словаря. Наименьший словарь представлен пятью символами: {A, C, G, T, N}, представляющими 4 типа нуклеотидов, присутствующих в ДНК, а именно аденин, цитозин, гуанин и тимин. В РНК тимин заменен на урацил (U). N указывает, что секвенатор не смог идентифицировать какое-либо основание, и поэтому реальная природа этого положения не определена. В случае, если секвенатор работает с кодами неоднозначности IUPAC, алфавит, используемый для символов представляет собой (A, C, G, T, U, W, S, M, K, R, Y, B, D, H, V, N или -).

Нуклеотидные последовательности, получаемые секвенаторами, называются "ридами". Последовательность рида может составлять от нескольких десятков до нескольких тысяч нуклеотидов. Некоторые технологии выдают последовательность ридов в парах, где один рид происходит из одной цепи ДНК, а второй - из другой цепи. При секвенировании генома термин "перекрывание" используется для выражения уровня избыточности данных последовательности относительно референсной последовательности. Например, чтобы достичь 30-кратного перекрывания человеческого генома (длиной 3,2 млрд. оснований), секвенатор должен произвести в общей сложности 30×3,2 млрд оснований, чтобы в среднем каждое положение в референсе было "перекрыто" 30 раз.

Во всем этом описании референсная последовательность представляет собой любую последова-

тельность, относительно которой выравниваются/картируются нуклеотидные последовательности, полученные с помощью секвенаторов. Одним из примеров последовательности может быть референсный геном, последовательность, собранная учеными в качестве репрезентативного примера набора генов какого-либо биологического вида. Например, GRC37, геном человека Genome Reference Consortium (сборка 37), получен от тринадцати анонимных добровольцев из Буффало, шт. Нью-Йорк. Однако референсная последовательность может также состоять из синтетической последовательности, сконструированной просто для улучшения сжимаемости ридов с учетом их дальнейшей обработки.

Устройства секвенирования могут вносить ошибки в последовательность ридов, такие как

1) использование неправильного символа (т.е. представляющего другую нуклеиновую кислоту) для представления нуклеиновой кислоты, фактически присутствующей в секвенированном образце; обычно это называется "ошибка замены" (несовпадение);

2) инсерция в одном риде последовательности дополнительных символов, которые не относятся к какой-либо фактически присутствующей нуклеиновой кислоте; обычно это называется "ошибка инсерции";

3) делеция из одного рида последовательности символов, представляющих нуклеиновые кислоты, которые фактически присутствуют в секвенированном образце; обычно это называется "ошибка делеции";

4) рекомбинация одного или более фрагментов в один фрагмент, который не отражает реальность исходной последовательности.

Термин "перекрывание" используется в литературе для количественной оценки степени, в которой референсный геном или его часть могут быть перекрыты доступными ридами последовательности. Перекрывание называется

частичным (менее чем 1-кратным), когда некоторые части референсного генома не картированы ни одним доступным ридом последовательности;

однократным (1×) когда все нуклеотиды референсного генома картированы одним и только одним символом, присутствующим в риде последовательности;

многократным (2×, 3×, N×) когда каждый нуклеотид референсного генома картирован несколько раз.

Настоящее изобретение направлено на разработку формата представления геномной информации, в котором соответствующая информация является эффективно доступной и переносимой, а вес избыточной информации уменьшен.

Основными аспектами раскрытого изобретения являются следующие.

1) Классификация ридов последовательности по разным классам по результатам выравнивания относительно референсных последовательностей с целью обеспечения избирательного доступа к кодированным данным в соответствии с критериями, относящимися к результатам выравнивания и точности совпадения.

2) Разложение последовательности рида данных и метаданных на однородные слои с целью получения отдельных источников информации с уменьшенной информационной энтропией.

3) Возможность моделирования каждого отдельного источника с отдельной моделью источника, адаптированной к каждой статистической характеристике, включая возможность изменения модели источника в каждом классе ридов и слое для каждого доступного блока данных (блоков доступа). Принятие подходящих контекстно-адаптивных вероятностных моделей и соответствующих энтропийных кодеров в соответствии со статистическими свойствами каждой модели источника.

4) Определение соответствий и зависимостей между слоями для обеспечения избирательного доступа к данным без необходимости декодировать все слои, если не вся информация необходима.

5) Кодирование каждого класса данных последовательности и ассоциированных слоев метаданных на основании референсной последовательности, которые могут быть модифицированы таким образом, чтобы уменьшить энтропию классов данных и источников информации о слоях. После первого кодирования на основе референсной последовательности обнаруженные несовпадения могут использоваться для "адаптации/модификации" референсной последовательности с целью дальнейшего уменьшения общей энтропии информации. Этот процесс, который может выполняться итеративно до тех пор, пока уменьшение информационной энтропии является значимым.

Далее каждый из вышеупомянутых аспектов будет описан дополнительно.

Главный заголовок файла.

Классификация ридов последовательности.

В соответствии с результатами выравнивания по одной или более референсным последовательностям сгенерированные секвенаторами прочитанные значения классифицируют 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 последователь-

ности 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 несовпадениями, состоящими из ряда положений, в которых секвенатор не смог определить ни одного основания (или нуклеотида), ИЛИ было определено другое основание, отличное от указанного в референсном геноме. Такой тип несовпадения называется однонуклеотидной вариацией (SNV) или однонуклеотидным полиморфизмом (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 присутствие инсерции или делеции (также называемых инделы). Инсерции представлены последовательностью из одного или более нуклеотидов, отсутствующих в референсе, но присутствующих в последовательности рида. 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".

Оставшиеся некартированные риды относительно референсной последовательности могут быть собраны в одну последовательность с использованием алгоритмов сборки de-novo. После создания вновь собранной референсной последовательности некартированные риды можно дополнительно картировать относительно нее и классифицировать в один из 4 классов P, N, M и I.

Разложение информации, необходимой для представления ридов последовательности, в слои дескрипторов.

После того как классификация рида 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. Структура данных этих синтаксических элементов требует хранения глобальных параметров и метаданных, которые будут использоваться механизмом декодирования. Эти данные структурированы в главном заголовке, описанном в таблице ниже.

Таблица 1. Структура главного заголовка

Элемент	Тип	Описание
Уникальный идентификатор	Байтовый массив	Уникальный идентификатор для закодированного контента
Версия	Байтовый массив	Основная + вспомогательная версия алгоритма кодирования
Размер заголовка	Целое число	Размер в байтах всего закодированного содержимого
Длина ридов	Целое число	Размер рида при постоянной длине рида. Специальное значение (например, 0) зарезервировано для переменной длины рида
Количество референсных последовательностей	Целое число	Количество использованных референсных последовательностей
Счетчики блоков доступа	Байтовый массив (например, целые числа)	Общее количество закодированных блоков доступа на каждую референсную последовательность
Идентификаторы референсных последовательностей	Байтовый массив	Уникальные идентификаторы для референсных последовательностей
Главная индексная таблица *Выравнивание положений первого рида в каждом блоке (блок доступа).* *То есть меньшее положение первого рида референсного генома на каждый блок из 4 классов*	Байтовый массив (например, целые числа)	Это многомерный массив, поддерживающий произвольный доступ к блокам доступа
1 на класс pos (4) на референс		

Сегмент ДНК, относящийся к данной референсной последовательности, может быть полностью выражен следующими параметрами:

начальное положение в 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 в случае, когда технология секвенирования дает переменную длину (len). В случае постоянной длины рида длина рида, ассоциированная с каждым ридом, очевидно, может быть опущена и может быть сохранена в главном заголовке файла;

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);

флаги (296), указывающие специфические характеристики рида последовательности, такие как

шаблон, имеющий множество сегментов в секвенировании;

каждый сегмент правильно выровнен согласно выравнивателю;

некартированный сегмент;

следующий сегмент в шаблоне не картирован;

сигнализация первого или последнего сегмента;
неудача контроля качества;
ПЦР или оптический дубликат;
вторичное выравнивание;
дополнительное выравнивание;
необязательная строка мягко обрезанных нуклеотидов (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.

Таблица 2. Определение слоев для каждого класса данных

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

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.

В следующем разделе подробно описано, как определяются эти дескрипторы.

Слой дескрипторов положения.

В слое положения (pos) только положения картирования первого закодированного рида хранится в заголовке AU как абсолютное положение в референсном геноме. Все остальные дескрипторы положения принимают значение, выражающее разницу относительно предыдущего положения. Такое моделирование источника информации, определяемое последовательностью положений рида, в целом характеризуется пониженной энтропией, особенно для процессов секвенирования, дающих результаты с высоким перекрыванием.

Например, на фиг. 1 показано, как после описания начального положения первого выравнивания в виде "положение "10000" в референсной последовательности положение второго рида, начинающегося в положении 10180, описывается как "180". При высоких значениях перекрывания (> 50×) большинство дескрипторов вектора положения будет иметь очень высокую встречаемость низких значений, таких как 0 и 1, и других маленьких целых чисел. На фиг. 9 показано, как положения трех пар рида описываются в слое pos.

Слой дескриптора обратного комплемента.

Каждый рид пары ридов, полученных с помощью технологий секвенирования, может происходить из любой цепи генома секвенированного органического образца. Однако только одна из двух цепей используется в качестве референсной последовательности.

На фиг. 2 показано, как в паре ридов один рид (рид 1) может происходить из одной нити, а другой (рид 2) - из другой.

Когда в качестве референсной последовательности используется цепь 1, рид 2 может быть закодирован как обратный комплемент соответствующего фрагмента на цепи 1. Это показано на фиг. 3.

В случае сцепленных ридов возможны четыре комбинации пар прямого и обратного комплемента. Это показано на фиг. 4. Слой rcomp кодирует эти четыре возможных комбинации.

Такое же кодирование используется для информации по обратному комплементу для рида, принадлежащего классам P, N, M, I. Чтобы обеспечить расширенный выборочный доступ к данным, информация по обратному комплементу для рида, принадлежащего к этим четырем классам, кодируется в разных слоях, как показано в табл. 2.

Слой дескриптора информации о спаривании.

Дескриптор спаривания хранится в слое "pair". Такой слой хранит дескрипторы, кодирующие информацию, необходимую для восстановления исходных пар ридов, когда используемая технология секвенирования генерирует риды по парам. Хотя на момент раскрытия изобретения подавляющее большинство данных секвенирования генерируется с использованием технологии создания парных ридов, это относится не ко всем технологиям. По этой причине присутствие этого слоя не является необходимым для восстановления всей информации данных секвенирования, если технология секвенирования рассматриваемых геномных данных не генерирует информацию по парным ридам.

Определения.

Партнёр по паре: рид, связанный с другим ридом в паре ридов (например, рид 2 - это пара рида 1 в предыдущем примере).

Расстояние спаривания: количество положений нуклеотидов в референсной последовательности, которые отделяют одно положение в первом риде (якорь спаривания, например, последний нуклеотид первого рида) от одного положения второго рида (например, первый нуклеотид второго рида).

Наиболее вероятное расстояние спаривания (MPPD): наиболее вероятное расстояние спаривания, выраженное в количестве положений нуклеотидов.

Расстояние спаривания в положениях (PPD): PPD - это способ выразить расстояние спаривания в числе ридов, отделяющих один рид от соответствующей пары, присутствующий в слое дескриптора конкретного положения.

Наиболее вероятное расстояние спаривания в положениях (MPPPD): наиболее вероятное число ридов, отделяющих один рид от его пары, присутствующей в слое дескриптора конкретного положения.

Ошибка положений спаривания (PPE): определяется как разница между MPPD или MPPPD и фактическим положением партнёра по паре.

Якорь спаривания: положения последнего нуклеотида первого рида в паре, используемое в качестве референса для вычисления расстояния пары сопряженных элементов, выраженного в числе положении нуклеотидов или числе прочитанных положений.

На фиг. 5 показано, как рассчитывается расстояние спаривания между парами ридов.

Слой дескрипторов пары - это вектор ошибок спаривания, рассчитанный как число ридов, которые необходимо пропустить, чтобы достичь партнёра по паре первого рида пары с учетом заданного расстояния декодирования спаривания.

На фиг. 6 показан пример того, как рассчитываются ошибки спаривания, как в виде абсолютной величины, так и в виде дифференциального вектора (характеризуется меньшей энтропией для высоких значений перекрывания).

Для информации о спаривании ридов, принадлежащих классам N, M, P и I, используются одинаковые дескрипторы. Чтобы реализовать выборочный доступ к различным классам данных, информация о спаривании для ридов, принадлежащих четырем классам, кодируется в другом слое, как изображено на фигуре.

Информация о спаривании в случае ридов, картированных по разным референсам.

В процессе картирования рида последовательности на референсной последовательности нередко бывает, что первый рид в паре картируется на одном референсе (например, хромосоме 1), а второй - на другом референсе (например, хромосоме 4). В этом случае описанная выше информация о спаривании должна быть объединена с дополнительной информацией, относящейся к референсной последовательности, используемой для картирования одного из ридов. Это достигается путем кодирования следующих параметров:

1) зарезервированное значение (флаг), указывающее, что пара картируется на двух разных последовательностях (разные значения указывают, картированы ли рид 1 или рид 2 на последовательности, которая в данный момент не кодирована);

2) уникальный референсный идентификатор, ссылающийся на идентификаторы референса, закодированные в структуре главного заголовка, как описано в табл. 1;

3) третий элемент, содержащий информацию о картировании на референсе, идентифицированном в точке 2, и выраженный как смещение относительно последнего закодированного положения.

На фиг. 7 приведен пример этого сценария.

На фиг. 7, поскольку рид 4 не картируется в кодированной в данный момент референсной последовательности, геномный кодер передает эту информацию, создавая дополнительные дескрипторы в слое pair. В примере, показанном ниже, рид 4 пары 2 картируется на референсе № 4, в то время как закодированный в данный момент референс - № 1. Эта информация кодируется с использованием 3 компонентов:

1) Одно специальное зарезервированное значение кодируется как расстояние спаривания (в этом случае - Oxffffff)

2) Второй дескриптор содержит идентификатор референса, как указано в главном заголовке (в этом случае - 4)

3) Третий элемент содержит информацию о картировании в соответствующем референсе (170).

Дескрипторы несовпадений для ридов класса N.

Класс N включает все риды, в которых на месте определения оснований A, C, G или T присутствуют только несовпадения, представленные "N". Все остальные основания рида идеально соответствуют референсной последовательности.

На фиг. 8 показано, как

положения несовпадений N в риде 1 кодируются следующим образом:

абсолютное положение в риде 1; или

дифференциальное положение относительно предыдущего "N" в том же риде.

Положения несовпадений "N" в риде 2 кодируются следующим образом:

абсолютное положение в риде 2 + длина рида 1 ИЛИ;

дифференциальное положение относительно предыдущего "N".

В слое nmis кодирование каждой пары ридов завершается специальным символом-разделителем.

На фиг. 8 показано, как "N"-несовпадения (где в данном положении картирования "N" присутствует в риде вместо фактического основания в референсной последовательности) кодируется только как положение несовпадения:

1) относительно начала рида; или

2) относительно предыдущего несовпадения (дифференциальное кодирование).

Дескрипторы, кодирующие замены (несовпадения или SNP), инсерции и делеции.

Замена определяется как присутствие в картированном риде нуклеотидного основания, отличного от того, которое присутствует в референсной последовательности в том же положении.

На фиг. 9 показаны примеры замен в картированной паре ридов. Каждая замена кодируется как "положение" (слой snpp) и "тип" (слой snpt). В зависимости от статистической встречаемости замен, инсерций или делеций, могут быть различаться заданные модели источника ассоциированных дескрипторов и сгенерированные символы, закодированные в ассоциированном слое.

Модель источника 1: замены как положения и типы.

Дескрипторы положений замен.

Положения замены вычисляется так же, как значения слоя nmis, т.е.

в риде 1 замены закодированы

как абсолютное положение в риде 1; или

как дифференциальное положение относительно предыдущей замены в том же риде.

В риде 2 замены закодированы

как абсолютное положение в риде 2 + длина рида 1; или

как дифференциальное положение относительно предыдущей замены

На фиг. 10 показано, как замены (где в данном положении картирования символ в риде отличается от символа в референсной последовательности) кодируются как:

1) положение несовпадения

относительно начала рида; или

относительно предыдущего несовпадения (дифференциальное кодирование),

2) тип несовпадения, представленный в виде кода, рассчитанного, как описано на фиг. 10.

В слое snpp кодирование каждой пары ридов завершается специальным символом-разделителем.

Дескрипторы типов замены.

Для класса M (и I, как описано в следующих разделах), несовпадения кодируются индексом (с перемещением справа налево) от фактического символа, присутствующего в референсе, до соответствующего символа замены, присутствующей в риде {A, C, G, T, N, Z}. Например, если выровненный рид показывает C вместо T, который присутствует в том же положении в референсе, индекс несовпадения будет обозначен как "4". Процесс декодирования считывает закодированный синтаксический элемент, нуклеотид в заданном положении на референсе и перемещается слева направо, возвращая декодированный символ. Например, "2", полученное для положения, где в референсе присутствует G, будет декодировано как "N". На фиг. 11 показаны все возможные замены и соответствующие символы кодирования. Очевидно, чтобы минимизировать энтропию дескрипторов, каждому индексу замены могут быть присвоены разные и контекстно-адаптивные вероятностные модели согласно статистическим свойствам каждого типа замены для каждого класса данных.

В случае применения кодов неоднозначности IUPAC механизм замены оказывается в точности таким же, однако вектор замены расширяется следующим образом: S = {A, C, G, T, N, Z, M, R, W, S, Y, K, V, H, D, B}.

На фиг. 12 приведен пример кодирования типов замен в слое snpt.

Некоторые примеры кодирования замен, когда применяются коды неоднозначности IUPAC, пред-

ставлены на фиг. 13. Еще один пример индексов замен представлен на фиг. 14.

Кодирование инсерций делеций.

Для класса I, несовпадения и делеций кодируются с помощью (с перемещением при кодировании справа налево) замены с фактического символа, присутствующего в референсе, на соответствующий символ замены, присутствующий в риде: {A, C, G, T, N, Z}. Например, если выровненный рид показывает C вместо T, присутствующего в том же положении в референсе, индекс несовпадения будет равен "4". В случае, если рид показывает делецию, где в референсе присутствует A, кодированный символ будет "5". Процесс декодирования считывает закодированный синтаксический элемент, нуклеотид в заданном положении на референсе и перемещается слева направо, возвращая декодированный символ. Например, "3", полученное для положения, где в референсе присутствует G, будет декодировано как "Z". Инсерции кодируются как 6, 7, 8, 9,10 соответственно для вставленных A, C, G, T, N.

На фиг. 15 показан пример того, как кодировать замены, инсерции и делеции в паре ридов класса I. Для поддержки всего набора кодов неоднозначности IUPAC вектор замен S = {A, C, G, T, N, Z} должен быть заменен на S = {A, C, G, T, N, Z, M, R, W, S, Y, K, V, H, D, B}, как описано в предыдущем абзаце для несовпадений. В этом случае коды инсерции должны иметь разные значения, а именно 16, 17, 18, 19, 20, если вектор замен имеет 16 элементов. Этот механизм показан на фиг. 16.

Модель источника 2: один слой для каждого типа замен и инделов.

Для некоторых статистических данных может быть разработана модель кодирования для замен и инделов, отличная от описанной в предыдущем разделе, приводящая к источнику с меньшей энтропией. Такая модель кодирования является альтернативой методикам, описанным выше, только для несовпадений, а также для несовпадений и инделов.

В этом случае для каждого возможного символа замены определяется один слой данных (5 без кодов IUPAC, 16 - с кодами IUPAC), плюс один слой для делеций и еще 4 слоя для инсерций. Для простоты объяснения, но не в качестве ограничения применения модели, нижеследующее описание будет сосредоточено на случае, когда коды IUPAC не поддерживаются.

На фиг. 17 показано, как каждый слой содержит положение несовпадений или инсерций одного типа. Если в кодированной паре ридов нет несовпадений или инсерций для этого типа, в соответствующем слое кодируется 0. Чтобы дать возможность декодеру запустить процесс декодирования для слоев, описанных в этом разделе, заголовок каждого блока доступа содержит флаг, сигнализирующий о первом слое, подлежащем декодированию. В примере на фиг. 18 первый декодируемый элемент - это положение 2 в слое C. Когда в паре ридов нет несовпадений или индексов инделов данного типа, к соответствующим слоям добавляется 0. На стороне декодирования, когда указатель декодирования для каждого слоя указывает на значение 0, процесс декодирования переходит к следующей паре ридов.

Кодирование дополнительных сигнальных флагов.

Каждый класс данных, представленный выше (P, M, N, I), может потребовать кодирования дополнительной информации о характере закодированных ридов. Эта информация может быть связана, например, с экспериментом по секвенированию (например, указанием вероятности дублирования одного рида) или может выражать некоторую характеристику картирования рида (например, первого или второго в паре). В контексте этого изобретения эта информация кодируется в отдельном слое для каждого класса данных. Основным преимуществом такого подхода является возможность выборочного доступа к этой информации только в случае необходимости и только в требуемой области референсной последовательности. Другими примерами использования таких флагов являются

спаренный рид;

рид, картированный в правильной паре;

рид или партнер по паре некартированы;

рид или партнер по паре из обратной цепи;

первое/второе в паре;

не первичное выравнивание;

рид не проходит проверку качества платформы/поставщика;

рид представляет собой ПЦР- или оптический дубликат;

дополнительное выравнивание.

Адаптация референсных последовательностей.

Несовпадения, закодированные для классов N, M и I, можно использовать для создания "модифицированных референсов", которые будут использоваться для перекодирования ридов в слое N, M или I (относительно первой референсной последовательности, R0) в качестве p-ридов относительно "адаптированного" генома R1. Например, если мы обозначим через r_in^M i-е рид класса M, содержащее несовпадения относительно референсного генома n, то после "адаптации" мы можем получить

$$r\_in^M=r\_(i(n+1))^P \; c$$

$$A(Refn)=Refn+1,$$

где A - преобразование из референсной последовательности n в референсную последовательность n+1.

На фиг. 19 показано, как риды, содержащие несовпадения (M-риды) относительно референсной по-

следовательности 1 (RS1), могут быть преобразованы в идеально совпадающие риды (P-риды) относительно референсной последовательности 2 (RS2), полученной из RS1 путем модификации несовпадающих положений. Это преобразование может быть выражено как

$$RS2 = A(RS1)$$

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

Модели источника, энтропийные кодеры и режимы кодирования.

Для каждого слоя структуры геномных данных, раскрытой в этом изобретении, могут быть приняты различные алгоритмы кодирования в соответствии с конкретными характеристиками данных или метаданных, переносимых 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".

Каждая "модель", ассоциированная с каким-то режимом кодирования, характеризуется

определением синтаксических элементов, генерируемых каждым источником (например, положения ридов, информация о спаривании ридов, несовпадения относительно референсной последовательности и т.д.);

определением ассоциированной вероятностной модели;

определением ассоциированного энтропийного кодера.

Другие преимущества.

Эта классификация позволяет реализовать эффективные режимы кодирования, использующие меньшую энтропию источника информации, характеризующиеся моделированием последовательностей синтаксических элементов отдельными независимыми источниками данных (например, расстояние, положение и т.д.). Другим преимуществом изобретения является возможность доступа только к подмножеству типов данных, представляющих интерес. Например, одно из наиболее важных приложений в геномике состоит в поиске различий геномного образца относительно референса (SNV) или популяции (SNP). Сегодня такой тип анализа требует обработки ридов полной последовательности, тогда как при принятии представления данных, раскрытого изобретением, несовпадения уже выделены всего в одинтри класса данных (в зависимости от заинтересованности в рассмотрении N-кодов и инделов).

Еще одним преимуществом является возможность выполнения эффективного транскодирования из данных и метаданных, сжатых со ссылкой на специфическую "референсную последовательность", в другую "референсную последовательность", когда публикуется новая "референсная последовательность" или когда выполняется повторное картирование для картированных ранее данных (например, используя другой алгоритм картирования).

На фиг. 20 показано устройство кодирования 207 в соответствии с принципами этого изобретения. Устройство кодирования 207 принимает в качестве входных данных необработанные данные последовательности 209, например созданные устройством секвенирования генома 200. 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 рида 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. 21 показано устройство декодирования 218 в соответствии с принципами этого изобретения. 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. Извлеченные 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 для получения несжатых ридов последовательностей, которые затем могут быть дополнительно сохранены в форматах, известных в данной облас-

ти 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.

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 более процессоров, процессоров цифровых сигналов, микропроцессоров общего назначения, специализированных интегральных схем или других дискретных логических схем.

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

ФОРМУЛА ИЗОБРЕТЕНИЯ

1. Реализуемый на компьютере способ сжатия данных геномной последовательности, где 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 различные классы, включающие по меньшей мере

первый класс: если указанные выровненные риды совпадают с указанными одной или более референсными последовательностями без каких-либо ошибок;

второй класс: если указанные выровненные риды совпадают с областью в указанной одной или более референсных последовательностях с числом несовпадений, состоящим из числа положений, в которых секвенатор не смог определить ни одного основания;

третий класс: если указанные выровненные риды совпадают с областью в указанной одной или более референсных последовательностях с числом несовпадений, состоящим из положений, в которых секвенатор не смог определить ни одного основания, и присутствия инсерций, или делеций, или обрезанных нуклеотидов;

четвертый класс: если указанные выровненные риды не находят какого-либо достоверного картирования на указанной одной или более референсных последовательностях в соответствии с указанными ограничениями выравнивания с получением таким образом классов выровненных ридов; и

кодирование указанных классифицированных и выровненных ридов в виде множества слоев синтаксических элементов, содержащих дескрипторы, причем для указанного первого класса указанные дескрипторы включают по меньшей мере начальное положение в референсной последовательности, флаг, сигнализирующий о том, что рид должен рассматриваться как обратный комплемент к референсу, расстояние до партнера пары в случае спаренных ридов, значение длины в случае, когда технология секвенирования дает риды переменной длины; для указанного второго класса дескрипторы включают по меньшей мере дескрипторы первого класса и положение несовпадения для каждого несовпадения; для указанного третьего класса дескрипторы включают дескрипторы указанного второго класса и положение несовпадения и тип несовпадения для каждого несовпадения; для указанного четвертого класса дескрипторы включают дескрипторы указанного первого класса и тип несовпадения для каждого несовпадения,

где указанное кодирование указанных классифицированных выровненных ридов в виде множества слоев синтаксических элементов включает выбор указанных синтаксических элементов, содержащих указанные дескрипторы, в соответствии с указанными классами выровненных ридов, причем кодирование указанных классифицированных выровненных ридов в виде множества слоев синтаксических элементов осуществляется с применением конкретного энтропийного кодера 2012-2014.

2. Способ по п.1, характеризующийся тем, что слои синтаксических элементов дополнительно содержат положение варианта относительно референсной последовательности, тип варианта, положение делеции относительно референсной последовательности, положение одного или более символов, отсутствующих в референсной последовательности, но присутствующих в выровненных ридах, тип инсерции в данном положении.

3. Способ по п.1, характеризующийся тем, что указанный энтропийный кодер является контекстно-адаптивным арифметическим кодером.

4. Способ распаковки геномного потока, сжатого способом по п.1, причем указанный способ включает следующие этапы:

синтаксический анализ и декодирование 212-214 сжатого геномного потока в геномные слои синтаксических элементов 215,

декодирование указанных геномных слоев в классы данных 216-217,

разворачивание указанных геномных слоев в классифицированные риды последовательностей нук-

леотидов,

выборочное декодирование с помощью декодеров 219 классов указанных классифицированных ридов последовательностей нуклеотидов и объединение результатов на одной или более референсных последовательностях с получением несжатых ридов последовательностей нуклеотидов.

5. Геномный кодер 2010 для сжатия данных геномной последовательности 209, причем указанные данные геномной последовательности 209 содержат риды последовательностей нуклеотидов, причем указанный геномный кодер 2010 содержит

модуль выравнивания 201, сконфигурированный для выравнивания указанных ридов с одной или более референсными последовательностями с получением, таким образом, выровненных ридов,

модуль классификации данных 204, сконфигурированный для классификации указанных выровненных ридов в различные классы, включающие по меньшей мере

первый класс: если указанные выровненные риды совпадают с указанными одной или более референсными последовательностями без каких-либо ошибок;

второй класс: если указанные выровненные риды совпадают с областью в указанной одной или более референсных последовательностях с числом несовпадений, состоящим из числа положений, в которых секвенатор не смог определить ни одного основания;

третий класс: если указанные выровненные риды совпадают с областью в указанной одной или более референсных последовательностях с числом несовпадений, состоящим из положений, в которых секвенатор не смог определить ни одного основания, и присутствия инсерций, или делеций, или обрезанных нуклеотидов;

четвертый класс: если указанные выровненные риды не находят какого-либо достоверного картирования на указанной одной или более референсных последовательностях в соответствии с указанными ограничениями выравнивания с получением, таким образом, классов выровненных ридов;

один или более кодирующих слои модулей 205-207, сконфигурированных для кодирования указанных классифицированных выровненных ридов в виде слоев синтаксических элементов, содержащих дескрипторы, путем выбора указанных синтаксических элементов в соответствии с указанными классами выровненных ридов, где для указанного первого класса указанные дескрипторы включают по меньшей мере начальное положение в референсной последовательности, флаг, сигнализирующий о том, что рид должен рассматриваться как обратный комплемент к референсу, расстояние до партнера пары в случае спаренных ридов, значение длины в случае, когда технология секвенирования дает риды переменной длины; для указанного второго класса дескрипторы включают по меньшей мере дескрипторы первого класса и положение несовпадения для каждого несовпадения; для указанного третьего класса дескрипторы включают дескрипторы указанного второго класса и положение несовпадения и тип несовпадения для каждого несовпадения; для указанного четвертого класса дескрипторы включают дескрипторы указанного первого класса и тип несовпадения для каждого несовпадения,

энтропийный кодер 2012-2014 для энтропийного кодирования указанных слоев синтаксических элементов.

6. Геномный декодер 218 для распаковки геномного потока 211, сжатого геномным кодером по п.6, причем указанный геномный декодер 218 содержит

средства синтаксического анализа и декодирования 210, 212-214, сконфигурированные для синтаксического анализа указанного сжатого геномного потока в геномные слои синтаксических элементов 215,

один или более декодеров слоев 216-217, сконфигурированных для декодирования геномных слоев в классы данных, и дополнительно сконфигурированный для обработки указанных геномных слоев в классифицированные риды последовательностей нуклеотидов 2111,

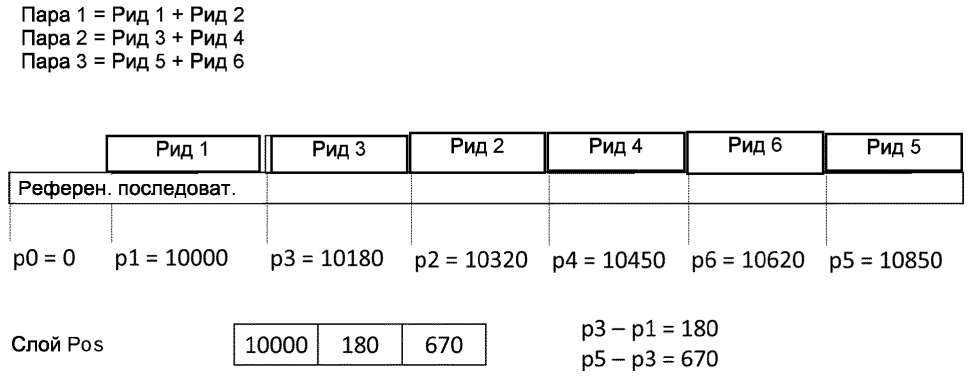
декодеры классов геномных данных 213, сконфигурированные для выборочного декодирования указанных классифицированных ридов последовательностей нуклеотидов и объединения результата по одной или нескольким референсным последовательностям с получением несжатых ридов последовательностей нуклеотидов.

7. Геномный декодер по п.6, характеризующийся тем, что одна или более референсных последовательностей хранятся в сжатом потоке генома 211.

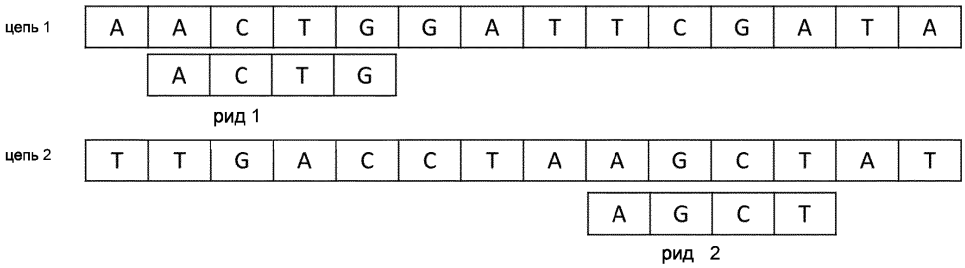
8. Геномный декодер по п.6, характеризующийся тем, что одна или более референсных последовательностей подаются в декодер по внеполосному механизму.

9. Геномный декодер по п.6, характеризующийся тем, что одна или более референсных последовательностей строятся в указанном декодере.

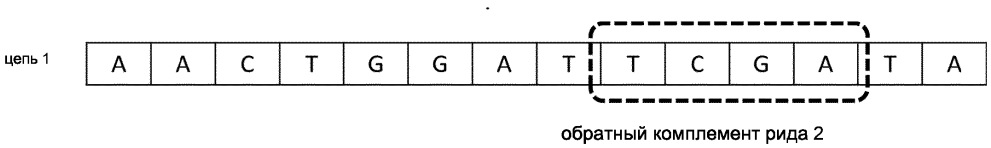
10. Машиночитаемый носитель, содержащий инструкции, которые при их выполнении приводят к осуществлению по меньшей мере одним процессором способа по любому из пп.1-4.



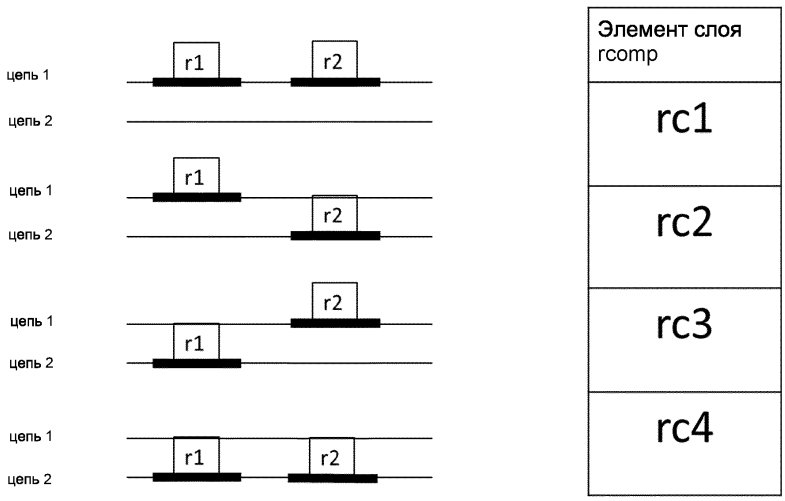
Фиг. 1 (Положение первого рида трех картированных пар прочтений кодируется в слое pos)



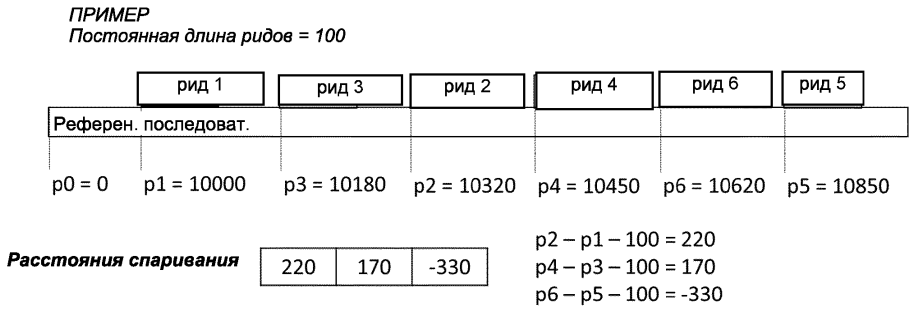
Фиг. 2 (В этой паре ридов рид 1 происходит из цепи 1, а рид 2 - из цепи 2)



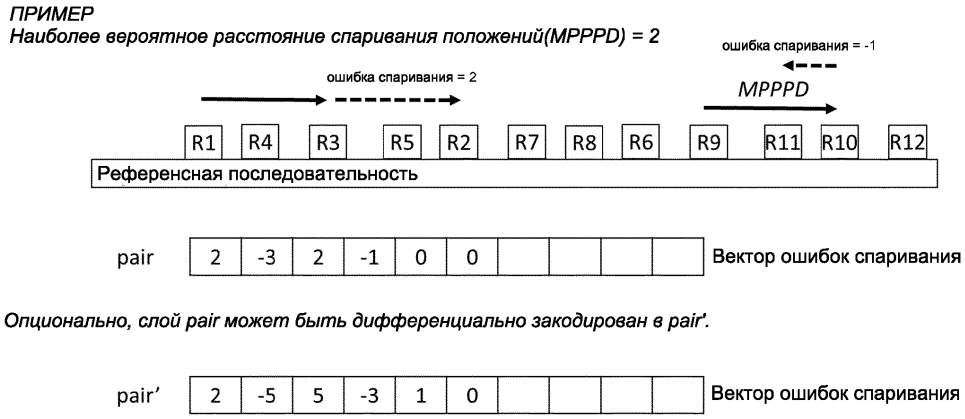
Фиг. 3 (Обратный комплемент прочтения 2 будет закодирован, если в качестве референса используется цепь 1)



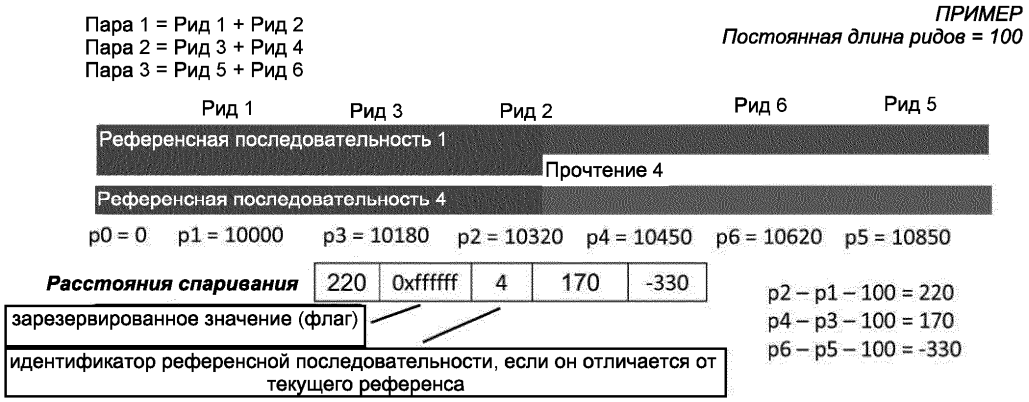
Фиг. 4 (Четыре возможных комбинации ридов, составляющих пару ридов и соответствующее кодирование в слое rcomp)



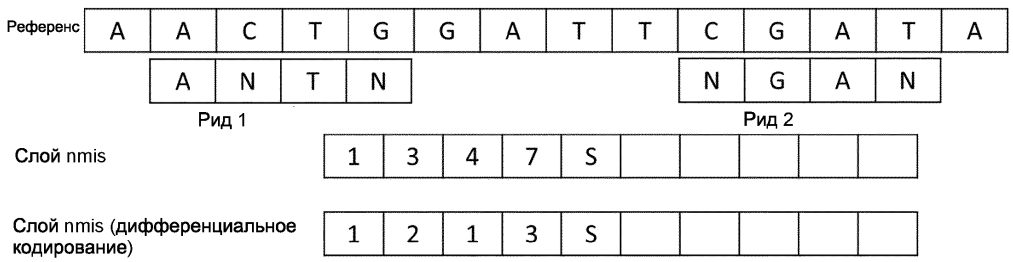
Фиг. 5 (Расчет расстояния спаривания для трех пар ридов)



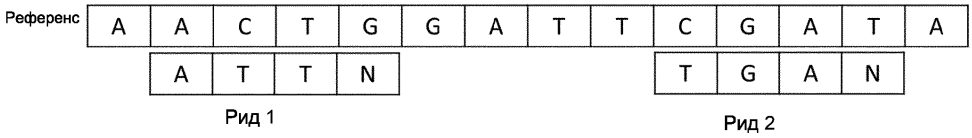
Фиг. 6 (Расчет ошибок спаривания)



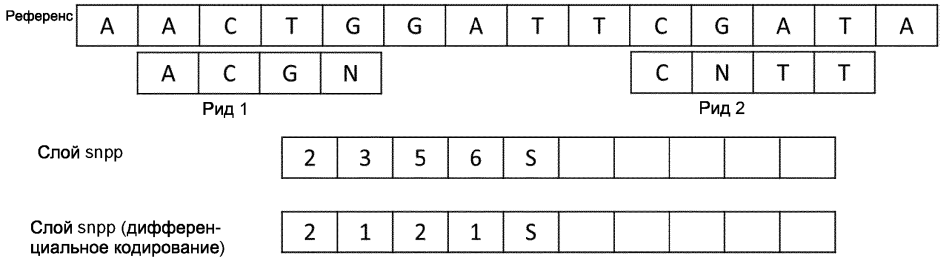
Фиг. 7 (Когда прочтение картируется на другой референс, чем его пара (рид 4), к расстоянию спаривания добавляют дополнительные дескрипторы. Один дескриптор - это сигнальный флаг, второй - идентификатор референса, а затем - расстояние спаривания)



Фиг. 8 (Расчет N-несоответствий в слое nmis)



Фиг. 9 (Замены в картированной паре ридов)



Фиг. 10 (Расчет положений замен в виде абсолютных и дифференциальных значений)

Слой snpt (без кодов IUPAC)
Тип замен рассчитывается как индекс вектора замен, составленного из всех возможных символов. Например:
S = [A, C, G, T, N, Z], **где Z = делеция**
Направление индекса

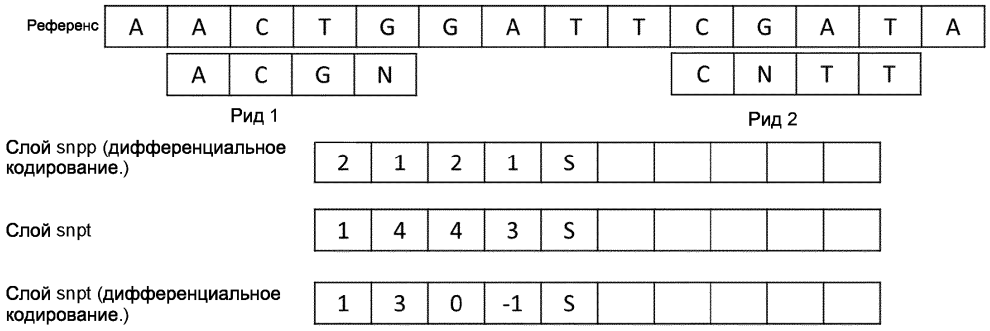
- КОДИРОВАНИЕ справа налево
- ДЕКОДИРОВАНИЕ слева направо

Референс	Прочтение	Закодированный символ
A	C	idx(A,C) = 1
A	G	idx(A,G) = 2
A	T	idx(A,T) = 3
A	N	idx(A,N) = 4
C	A	idx(C,A) = 5
C	G	idx(C,G) = 1
C	T	idx(C,T) = 2
C	N	idx(C,N) = 3
G	A	idx(G,A) = 4
G	C	idx(G,C) = 5
G	T	idx(G,T) = 1
G	N	idx(G,N) = 2
T	A	idx(T,A) = 3
T	C	idx(T,C) = 4
T	G	idx(T,G) = 5
T	N	idx(T,N) = 1

Референс	Рид	Закодированный символ
A	del.	idx(A,Z) = 5
C	del.	idx(C,Z) = 4
G	del.	idx(G,Z) = 3
T	del.	idx(T,Z) = 2

Референс	Рид	Закодированный символ
N	A	idx(N,A) = 2
N	C	idx(N,C) = 3
N	G	idx(N,G) = 4
N	T	idx(N,T) = 5

Фиг. 11 (Расчеты символов, кодирующих замены)



Фиг. 12 (Кодирование замен в слой snpt)

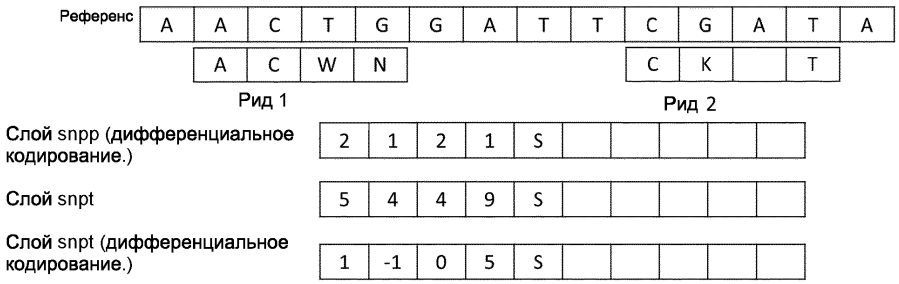
Слой snpt (с кодами IUPAC)
Тип замен рассчитывается как индекс вектора замен, составленного из всех возможных символов. Например:
S = [A, C, G, T, N, Z, M, R, W, S, Y, K, V, H, D, B]
Направление индекса
- КОДИРОВАНИЕ справа налево
- ДЕКОДИРОВАНИЕ слева направо

Референс	Рид	Закодированный символ
N	M	idx(N,M) = 2
N	W	idx(N,W) = 4
N	S	idx(N,S) = 5
N	B	idx(N,B) = 11

Референс	Рид	Закодированный символ
D	M	idx(D,M) = 8
A	Y	idx(A,Y) = 10
A	T	idx(A,T) = 3
A	N	idx(A,N) = 4
C	R	idx(C,R) = 6
C	G	idx(C,G) = 1
C	T	idx(C,T) = 2
C	W	idx(C,W) = 7
G	H	idx(G,H) = 11
G	C	idx(G,C) = 15
G	B	idx(G,B) = 13
G	N	idx(G,N) = 2
T	A	idx(T,A) = 13
T	M	idx(T,M) = 4
T	K	idx(T,K) = 8
T	V	idx(T,V) = 9

Фиг. 13 (Коды замен при использовании кодов неоднозначности IUPAC)

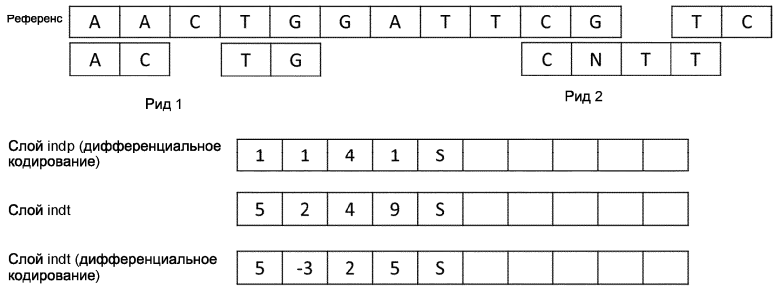
S = [A, C, G, T, N, Z, M, R, W, S, Y, K, V, H, D, B]
Направление
- КОДИРОВАНИЕ справа налево
- ДЕКОДИРОВАНИЕ слева направо



Фиг. 14 (Кодирование слоя snpt при использовании кодов IUPAC)

Слой indt (без кодов IUPAC)
S = [A, C, G, T, N, Z]
Направление
- КОДИРОВАНИЕ справа налево
- ДЕКОДИРОВАНИЕ слева направо

Инсерция	Закодированный символ
A	6
C	7
G	8
T	9
N	10

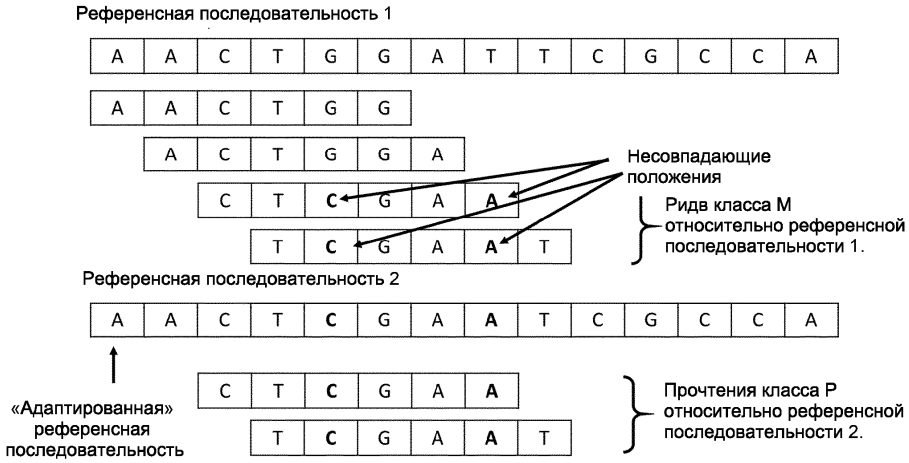


Фиг. 15 (Кодирование замен, инсерций и делеций в паре прочтений класса I)

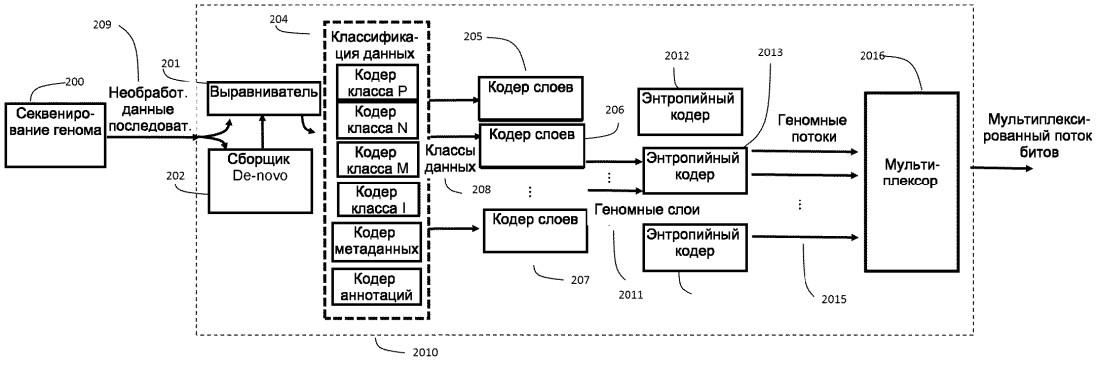
Фиг. 16 (Кодирование несовпадений и инделов при использовании кодов неоднозначности IUPAC)

Фиг. 17 (Каждый слой содержит позицию несовпадения
или инсерции одного типа)

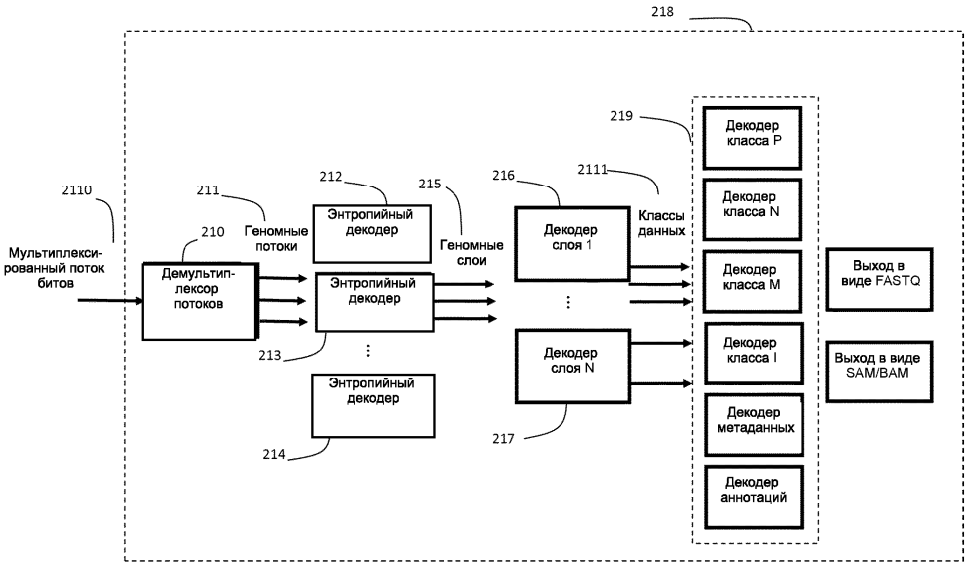
Фиг. 18 (Если для рида нет несовпадений или инделов данного типа, в соответствующем слое кодируется значение 0. Символ 0 действует как разделитель и терминатор ридов в каждом слое)



Фиг. 19 (Модификация в референсной последовательности может преобразовать M-прочтения в P-прочтения)



Фиг. 20 (Геномный кодер)



Фиг. 21 (Геномный декодер)

Евразийская патентная организация, ЕАПВ
Россия, 109012, Москва, Малый Черкасский пер., 2