(19) Евразийское патентное ведомство

(11) 043338 (13) B1

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ЕВРАЗИЙСКОМУ ПАТЕНТУ

(45) Дата публикации и выдачи патента
2023.05.15

(21) Номер заявки
201991908

(22) Дата подачи заявки
2018.02.14

(51) Int. Cl. ***G06F 7/00*** (2006.01)

(54) СПОСОБ И УСТРОЙСТВО ДЛЯ КОМПАКТНОГО ПРЕДСТАВЛЕНИЯ БИОИНФОРМАЦИОННЫХ ДАННЫХ С ПОМОЩЬЮ НЕСКОЛЬКИХ ГЕНОМНЫХ ДЕСКРИПТОРОВ

(31) **PCT/US2017/017842; PCT/US2017/041591**
(32) **2017.02.14; 2017.07.11**
(33) **US**
(43) **2020.01.21**
(86) **PCT/US2018/018092**
(87) **WO 2018/152143 2018.08.23**

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

(72) Изобретатель:
Алберти Клаудио, Зоиа Гиоргио, Рензи Даниэле (CH), Балуч Мохамед Хосо (US)

(74) Представитель:
Пронин В.О. (RU)

(56) US-A1-20150227686
US-A1-20090055425
US-A1-20130204851
US-A1-20140371109
US-A1-20020159625
US-B2-7698067
US-A1-20160259886

(57) Способ и устройство для сжатия данных геномной последовательности, созданных секвенаторами генома. Прочтения последовательности кодируют путем выравнивания их относительно ранее существующих или построенных референсных последовательностей, причем процесс кодирования состоит из классифицирования прочтений в классы данных с последующим кодированием каждого класса посредством множества блоков дескрипторов. Для каждого класса данных, на которые разбивают данные, и каждого соответствующего блока дескрипторов используют специальные модели источников и энтропийные кодеры.

043338 B1

043338 B1

Перекрестная ссылка на родственные заявки

Настоящая заявка испрашивает приоритет и преимущество следующих заявок на патент: PCT/US2017/017842, поданной 14 февраля 2017 г., и PCT/US2017/041591, поданной 11 июля 2017 г.

Область техники

В соответствии с настоящим изобретением предлагается новый способ представления данных секвенирования генома, который сокращает используемый объем памяти и улучшает характеристики доступа за счет обеспечения новых функциональных возможностей, которые недоступны в способах представления известного уровня техники.

Уровень техники

Надлежащая форма представления данных секвенирования генома является основополагающей для обеспечения эффективных приложений геномного анализа, таких как определение вариантов генома и все другие анализы, выполняемые в различных целях путем обработки данных секвенирования и метаданных.

Секвенирование генома человека стало доступным после появления недорогих технологий секвенирования с высокой пропускной способностью. Такая возможность открывает новые перспективы в нескольких областях, от диагностики и лечения рака до выявления генетических заболеваний, от патогенетического надзора за идентификацией антител до создания новых вакцин, лекарственных препаратов и адаптации персонализированных лечебных процедур.

Больницы, поставщики услуг по анализу геномных данных, специалисты в области биоинформатики и крупные центры хранения биологических данных ищут недорогие, быстрые, надежные и взаимосвязанные решения для обработки геномной информации, которые позволили бы распространить геномную медицину в мировом масштабе. Поскольку одним из узких мест в процессе секвенирования стало хранение данных, все больше внимания уделяется способам представления данных секвенирования генома в сжатой форме.

Наиболее используемые представления геномной информации данных секвенирования основываются на форматах архивирования FASTQ и SAM. Цель состоит в сжатии традиционно используемых форматов файлов (соответственно FASTQ и SAM для невыровненных и выровненных данных). Такие файлы составляют из символов обычного текста и сжимают, как упоминалось выше, с использованием подходов общего назначения, таких как схемы LZ (от Lempel и Ziv, фамилий авторов, опубликовавших первые версии), например, широко известные zip, gzip и т.д. При использовании средств сжатия общего назначения, таких как gzip, результатом сжатия обычно бывает большой двоичный объект. Получаемую в результате информацию в такой монолитной форме довольно трудно архивировать, передавать и обрабатывать, в частности, как в случае высокопроизводительного секвенирования, когда объем данных чрезвычайно велик. Формат BAM отличается низкими характеристиками сжатия, так как он сосредоточен на сжатии неэффективного и избыточного формата SAM, а не на выделении фактической геномной информации, содержащейся в файлах SAM, и адаптирует алгоритмы сжатия текста общего назначения, такие как gzip, вместо использования специфического характера каждого источника данных (самих геномных данных).

Более сложным подходом к сжатию геномных данных, который применяется реже, но более эффективен по сравнению с BAM, является CRAM. CRAM обеспечивает более эффективное сжатие за счет внедрения дифференциального кодирования относительно референса (в частности, использования избыточности источника данных), но ему все же недостает функций, таких как инкрементальные обновления, поддержка потоковой обработки данных и выборочного доступа к конкретным классам сжатых данных. Эти подходы обеспечивают плохие коэффициенты сжатия и структуры данных, которые сложны для перемещения по ним и манипулирования после сжатия. Последующий анализ может быть сильно замедлен ввиду необходимости обработки больших и жестко заданных структур данных даже для выполнения простой операции или получения доступа к выбранным областям геномного набора данных. CRAM опирается на концепцию записи CRAM. Каждая запись CRAM представляет одно картированное или некартированный рид путем кодирования всех элементов, необходимых для его восстановления.

CRAM демонстрирует следующие ограничения и недостатки, которые решаются и преодолеваются изобретением, описанным в данном документе.

1. CRAM не поддерживает индексацию данных и произвольный доступ к подмножествам данных с общими специфическими признаками. Индексация данных выходит за рамки данной спецификации (см. раздел 12 Спецификации CRAM v3.0) и реализуется в виде отдельного файла. Напротив, в подходе согласно изобретению, описанному в данном документе, используется способ индексации данных, который интегрирован с процессом кодирования, и индексы встраиваются в закодированный (т.е. сжатый) двоичный поток.

2. CRAM строят из блоков основных данных, которые могут содержать картированные риды любого типа (идеально совпадающие риды, риды только с заменами, риды с инсерциями или делециями (называемыми также "инделами")). Не существует понимания классификации и группирования ридов в классы в соответствии с результатом картирования относительно референсной последовательности. Т.е. проверять нужно все данные даже в случае поиска ридов с определенными характеристиками. Такое ог-

раничение снимается с помощью данного изобретения за счет классифицирования и разбиения данных на классы перед кодированием.

3. CRAM основан на принципе инкапсуляции каждого рида в "запись CRAM". Это означает, что при поиске ридов, отличающихся определенными биологическими характеристиками (например, ридов с заменами, но без "инделов", или идеально картированных ридов) нужно проверять каждую полную "запись".

В отличие от этого в настоящем изобретении существует понятие классов данных, кодируемых в отдельных информационных блоках, и нет понятия записи, инкапсулирующей каждый рид. Это позволяет эффективнее получать доступ к наборам ридов с определенными биологическими характеристиками (например, к ридам с заменами, но без "инделов", или к идеально картированным ридам) без необходимости декодирования каждого рида (блока ридов) для проверки его характеристик.

4. В записи CRAM каждое поле записи связано с определенным флагом, и каждый флаг должен всегда иметь одно и то же значение ввиду отсутствия понятия контекста, так как каждая запись CRAM может содержать данные какого-либо другого типа. Данный механизм кодирования вносит избыточную информацию и препятствует использованию эффективного контекстного энтропийного кодирования.

Вместо этого в настоящем изобретении не существует понятия флага, обозначающего данные, так как это по сути определено информационным "блоком", которому принадлежат эти данные. Отсюда значительное сокращение символов, которые нужно использовать, и, как следствие, снижение энтропии источника информации, что приводит к более эффективному сжатию. Такое улучшение возможно за счет использования различных "блоков", позволяющих кодеру повторно использовать один и тот же символ в пределах каждого блока в разных значениях в соответствии с контекстом. В CRAM каждый флаг должен всегда иметь одно и то же значение ввиду отсутствия понятия контекстов и каждая запись CRAM может содержать данные любого типа.

5. В CRAM замены, инсерции и делеции представляют с помощью различных дескрипторов, и именно такой выбор увеличивает размер алфавита источника информации и дает повышенную энтропию источника. В подходе описываемого здесь изобретения, напротив, используют один алфавит и одно кодирование для замен, инсерций и делеций. Это упрощается процесс кодирования и декодирования и создает модель источника с пониженной энтропией, кодирование которой обеспечивает потоки двоичных данных, отличающиеся высокими характеристиками сжатия.

Целью настоящего изобретения является сжатие геномных последовательностей путем классифицирования и разделения данных секвенирования так, чтобы избыточная информация, подлежащая кодированию, была сведена к минимуму, а функции, такие как выборочный доступ и поддержка инкрементальных обновлений, действовали непосредственно в сжатой области.

Один из аспектов представленного подхода состоит в определение классов данных и метаданных, структурируемых в различных блоках и кодируемых по отдельности. Более актуальные улучшения такого подхода по сравнению с существующими способами, заключаются в следующем:

1) улучшение характеристик сжатия путем снижения энтропии источника информации за счет обеспечения эффективной модели источника для каждого класса данных или метаданных;

2) возможность осуществления выборочного доступа к частям сжатых данных и метаданных в целях любой дальнейшей обработки непосредственно в сжатой области;

3) возможность инкрементального (т.е. без необходимости в декодировании и повторном кодировании) обновления сжатых данных и метаданных новыми данными секвенирования, и/или метаданными, и/или новыми результатами анализа, связанными с определенными наборами ридов секвенирования.

Краткое описание чертежей

На фиг. 1 показано, как положение картированных пар ридов кодируют в блоке "pos" в виде разницы относительно абсолютного положения первого картированного рида.

На фиг. 2 показано, как два рида в паре могут происходить из двух цепочек ДНК.

На фиг. 3 показано, как кодируют обратный комплемент рида 2, если цепочку 1 используют в качестве референса.

На фиг. 4 показаны четыре возможные комбинации ридов, составляющих пару ридов, и соответствующее кодирование в блоке "rcomp".

На фиг. 5 для трех пар ридов показано, как вычислять расстояние спаривания в случае ридов постоянной длины.

На фиг. 6 показано, как ошибки спаривания, кодированные в блоке "pair", позволяют декодеру восстанавливать правильное спаривание пары с помощью кодированного "MPPPD".

На фиг. 7 показано кодирование расстояния спаривания, когда рид картирован не на тот референс, что его парный рид. В этом случае к расстоянию картирования добавляют дополнительные дескрипторы. Одним из них является флаг сигнализации, вторым - идентификатор референса и еще одним - расстояние спаривания.

На фиг. 8 показано кодирование несовпадение "n-типа" в блоке "nmis".

На фиг. 9 показана картированная пара ридов, в которой присутствуют замены по сравнению с референсной последовательностью.

На фиг. 10 показано, как вычислять положения замен в виде абсолютных или разностных значений.

На фиг. 11 показано, как вычислять символы, кодирующие типы замен, когда используются коды IUPAC. Символы представляют расстояние, в кольцевом векторе замены, между молекулой, присутствующей в риде, и молекулой, присутствующей на референсе в этом положении.

На фиг. 12 показано, как кодировать замены в блоке "snpt".

На фиг. 13 показано, как вычислять коды замены, когда используются коды неоднозначности IUPAC.

На фиг. 14 показано, как кодировать блок "snpt", когда используются коды неоднозначности IUPAC.

На фиг. 15 показано, как для ридов класса I используют вектор замены, тот же самый, что и для класса M, с добавлением специальных кодов для вставок символов A, C, G, T, N.

На фиг. 16 показаны некоторые примеры кодирования несовпадений и инделов в случае кодов неоднозначности IUPAC. В этом случае вектор замены намного длиннее, и поэтому возможных вычисляемых символов больше, чем в случае с пятью символами.

На фиг. 17 показан другая модель источника для несовпадений и инделов, где каждый блок содержит положение несовпадений или инсерций одного типа. В этом случае никаких символов для типа несовпадения или индела не кодируют.

На фиг. 18 показан пример кодирования несовпадений и инделов. Когда в риде нет несовпадений или инделов данного типа, в соответствующем блоке кодируют 0. 0 Действует в качестве разделителя ридов и признака конца в каждом блоке.

На фиг. 19 показано, как изменение в референсной последовательности может преобразовать риды M в риды P. Эта операция может снизить информационную энтропию структуры данных особенно в случае данных с высоким покрытием.

На фиг. 20 показан геномный кодер 2010 в соответствии с одним вариантом осуществления настоящего изобретения.

На фиг. 21 показан геномный декодер 218 в соответствии с одним вариантом осуществления настоящего изобретения.

На фиг. 22 показано, как "внутренний" референс может быть построен путем кластеризации ридов и сборки сегментов, взятых из каждого кластера.

На фиг. 23 показано, как стратегия построения референса заключается в сохранении самых последних ридов После того как к ридам была примерена определенная сортировка (например, в лексикографическом порядке).

На фиг. 24 показано, как рид, принадлежащий классу "некартированных" ридов (класс U), может быть кодировано с помощью шести дескрипторов, хранящихся или содержащихся в соответствующих блоках.

На фиг. 25 показано, как альтернативное кодирование ридов, принадлежащих классу U, где дескриптор pos имеет знак, используют для кодирования положения картирования рида на построенном референсе.

На фиг. 26 показано, как преобразования референса могут быть применены для удаления несовпадений из ридов. В некоторых случаях преобразования референса могут приводить к образованию новых несовпадений или изменению типа несовпадений, обнаруженных при сравнении с референсом до выполнения преобразования.

На фиг. 27 показано, как преобразования референса могут изменить принадлежность ридов к классу при устранении всех или подмножества несовпадений (т.е. рид, принадлежащий классу M до преобразования, назначают классу P После того как преобразование референса было применено).

На фиг. 28 показано, как полукартированные пары ридов (класс HM) могут быть использованы для заполнения неизвестных областей референсной последовательности путем сборки более длинных контигов с помощью некартированных ридов.

На фиг. 29 показано, как кодеры данных класса N, M и I конфигурируют с помощью векторов порогов и формируют отдельные подклассы классов данных N, M и I.

На фиг. 30 показано, как все классы данных могут использовать один и тот же преобразованный референс для повторного кодирования, или для каждого класса N, M и I или любой их комбинации может быть использовано свое преобразование.

На фиг. 31 показана структура заголовка геномного набора данных.

На фиг. 32 показана общая структура таблицы главного индекса (MIT), где каждая строка содержит геномные интервалы нескольких классов данных P, N, M, I, U, HM и, кроме того, указатели на метаданные и аннотации. Столбцы относятся к определенным положениям на референсных последовательностях, связанных с кодированными геномными данными.

На фиг. 33 показан пример одной строки MIT, содержащей геномные интервалы, относящиеся к ридам класса P. Геномные области, относящиеся к различным референсным последовательностям, разделены специальным флагом ("S" в данном примере).

На фиг. 34 показаны общая структура таблицы локального индекса (LIT) и порядок ее использова-

ния для хранения указателей на физическое местоположение кодированной геномной информации в сохраняемых или передаваемых данных.

На фиг. 35 показан пример LIT, используемой для доступа к блокам доступа № 7 и 8 в полезной нагрузке блока.

На фиг. 36 показана функциональная взаимосвязь между несколькими строками MIT и LIT, содержащимися в заголовках геномных блоков.

На фиг. 37 показано, как блок доступа составляют из нескольких блоков геномных данных, переносимых разными геномными потоками, содержащими данные, принадлежащие разным классам. Каждый блок дополнительно содержит пакеты данных, используемые в качестве блоков передачи данных.

На фиг. 38 показано, как блоки доступа составляют из заголовка и мультиплексированных блоков, принадлежащих одному или более блоками однородных данных. Каждый блок может состоять из одного или более пакетов, содержащих фактические дескрипторы геномной информации.

На фиг. 39 показаны множественные выравнивания без сплайсинга. Крайнее левое рид имеет N выравниваний. N является первым значением дескриптора mmap, подлежащим декодированию, и сигнализирует о количестве выравниваний первого рида. Следующие N значений дескриптора mmap декодируют и используют для вычисления P, представляющего собой количество выравниваний второго рида.

На фиг. 40 показано, как дескрипторы pos, pair и mmap используют для кодирования множественных выравниваний без сплайсов. Крайнее левый рид имеет N выравниваний.

На фиг. 41 показаны множественные выравнивания со сплайсами.

На фиг. 42 показано использование дескрипторов pos, pair, mmap и msar для представления множественных выравниваний со сплайсами.

Сущность изобретения

Признаки приведенной ниже формулы изобретения решают проблему решений известного уровня техники путем создания способа кодирования данных геномной последовательности, где данные геномной последовательности содержат риды последовательностей нуклеотидов, причем способ включает в себя этапы

выравнивания ридов на одну или более референсных последовательностей с созданием тем самым выровненных ридов;

классифицирования выровненных последовательностей согласно заданным правилам сопоставления с одной или более референсных последовательностей с созданием тем самым классов выровненных ридов;

кодирования классифицированных выровненных ридов в виде множества блоков дескрипторов, причем кодирование классифицированных выровненных ридов в виде множества блоков дескрипторов включает в себя выбор дескрипторов в соответствии с классами выровненных ридов;

структурирования блоков дескрипторов с помощью информации заголовка с созданием тем самым последовательных блоков доступа.

В соответствии с еще одним аспектом способ кодирования дополнительно включает в себя

дальнейшее классифицирование ридов, которые не удовлетворяют заданным правилам сопоставления, в класс некартированных ридов;

построение набора референсных последовательностей с использованием некоторых некартированных ридов;

выравнивание класса некартированных ридов на набор построенных референсных последовательностей;

кодирование классифицированных выровненных ридов в виде множества блоков дескрипторов;

кодирование набора построенных референсных последовательностей;

структурирование блоков дескрипторов и кодированных референсных последовательностей с помощью информации заголовка с созданием тем самым последовательных блоков доступа.

В соответствии с еще одним аспектов способ кодирования дополнительно включает в себя идентифицирование геномных ридов без несовпадений с референсной последовательностью в качестве первого "класса P".

В соответствии с еще одним аспектом способ кодирования дополнительно включает в себя идентифицирование геномных ридов в качестве второго "класса N", когда несовпадения обнаруживают только в положениях, где секвенатор оказался не в состоянии определить никакого "основания", и количество несовпадений в каждом риде не превышает данного порога.

В соответствии с еще одним аспектом способ кодирования дополнительно включает в себя идентифицирование геномных ридов в качестве третьего "класса M", когда несовпадения обнаруживают в положениях, где секвенатор оказался не в состоянии определить никакого "основания" (называются несовпадениями "n-типа") и/или он определил другое "основание" по сравнению с референсной последовательностью (называются несовпадениями "s-типа"), и количество несовпадений не превышает данных порогов для количества несовпадений "n-типа", "s-типа" и порога, полученного из данной функции (f(n,s)), вычисленной на количестве несовпадений "n-типа" и "s-типа".

В соответствии с еще одним аспектом способ кодирования дополнительно включает в себя иденти-

фицирование геномных ридов в качестве четвертого "класса I", когда они могут иметь несовпадения того же типа, что и несовпадения "класса M", и, дополнительно, по меньшей мере одно несовпадение следующего типа: "инсерция" ("i-типа"), "делеция" ("d-типа"), мягкие отсечения ("c-типа"), и при этом количество несовпадений каждого типа не превышает соответствующего данного порога и порога, обеспечиваемого данной функций (w(n,s,i,d,c)), вычисленной на количестве несовпадений "n-типа", "s-типа", "i-типа", "d-типа " и "c-типа".

В соответствии с еще одним аспектом способ кодирования дополнительно включает в себя идентифицирование геномных ридов в качестве пятого "класса U", который содержит все риды, не отнесенные ни к одному из классов P, N, M, I, определенных ранее.

В соответствии с еще одним аспектом способ кодирования дополнительное отличается тем, что риды геномной последовательности, подлежащие кодированию, являются парными.

В соответствии с еще одним аспектом способ кодирования дополнительно отличается тем, что классифицирование дополнительно включает в себя идентифицирование геномных последовательностей в качестве шестого "класса HM", как содержащего все пары ридов, где один рид принадлежит классу P, N, M или I, а другой рид принадлежит "классу U".

В соответствии с еще одним аспектом способ кодирования дополнительно включает в себя этапы определения того, классифицируются ли два парных рида в один и тот же класс (каждый из P, N, M, I, U) с назначением в таком случае пары в этот же идентифицированный класс.

Определение того, относятся ли два парных рида к разным классам, и если ни одно из них не принадлежит "классу U", то назначение пары ридов классу с самым высоким приоритетом, определяемым в соответствии со следующим приоритетом:

$$P<N<M<I,$$

где "класс P" обладает самым низким приоритетом, а "класс I" - самым высоким приоритетом;

определение того, классифицировано ли только одно из парных ридов как принадлежащее "классу U", и классифицирование пары ридов как принадлежащей последовательностям "класса HM".

В соответствие с еще одним аспектом способ кодирования дополнительно отличается том, что каждый класс ридов N, M, I дополнительно разделяют на два или более подклассов (296, 297, 298) в соответствии с вектором порогов (292, 293, 294), определенным, соответственно, для каждого класса N, M и I посредством количества несовпадений (292) "n-типа", функции f(n,s) (293) и функции w(n,s,i,d,c) (294).

Определение того, классифицированы ли два парных рида в один и тот же подкласс с назначением в таком случае пары в этот же подкласс;

определение того, классифицированы ли два парных рида в подклассы разных классов с назначением в таком случае пары в подкласс, принадлежащий классу более высокого приоритета в соответствии со следующим выражением:

$$N<M<I,$$

где N имеет самый низкий приоритет, а I имеет самый высокий приоритет;

определение того, классифицированы ли два парных рида в один и тот же класс, причем этим классом является N, или M, или I, или в разные подклассы с назначением в таком случае пары в подкласс с самым высоким приоритетом в соответствии о следующими выражениями:

$$N_1<N_2<\ldots<N_k,$$
$$M_1<M_2<\ldots<M_j,$$
$$I_1<I_2<\ldots<I_h,$$

где самый высокий индекс имеет наивысший приоритет.

В соответствии с еще одним аспектом информацию о положении картирования каждого рида кодируют посредством блока дескрипторов pos.

В соответствии с еще одним аспектом информацию о цепочечности (например, из какой цепочки ДНК секвенировали рид) каждого рида кодируют посредством блока дескрипторов rcomp.

В соответствии с еще одним аспектом информацию о парноконцевых ридах кодируют посредством блока дескрипторов pair.

В соответствии с еще одним аспектом дополнительную информацию о выравнивании, такую как, картирован ли рид в правильную пару, не прошло ли оно проверки качества платформы/поставщика, является ли оно дубликатом ПЦР или оптическим дубликатом, является ли это вспомогательным выравниванием, кодируют посредством блока дескрипторов flags.

В соответствии с еще одним аспектом информацию о неизвестных основаниях кодируют посредством блока дескрипторов nmis.

В соответствии с еще одним аспектом информацию о положении замен кодируют посредством блока дескрипторов snpp.

В соответствии с еще одним аспектом информацию о типе замен кодируют посредством блока специальных дескрипторов snpt.

В соответствии с еще одним аспектом информацию о положении несовпадений типа замен, инсерций или делеций кодируют посредством блока дескрипторов indp.

В соответствии с еще одним аспектом информацию о типе несовпадений, таких как замены, инсер-

ций или делеций, кодируют посредством блока дескрипторов indt.

В соответствии с еще одним аспектом информацию об отсеченных основаниях картированного рида кодируют посредством блока дескрипторов indc.

В соответствии с еще одним аспектом информацию о некартированных ридах кодируют посредством блока дескрипторов ureads.

В соответствии с еще одним аспектом информацию о типе референсной последовательности, используемой для кодирования, кодируют посредством блока дескрипторов rtype.

В соответствии с еще одним аспектом информацию о множественных выравниваниях картированных ридов кодируют посредством блока дескрипторов mmap.

В соответствии с еще одним аспектом информацию о сплайсированных выравниваниях и множественных выравниваниях одного и того же рида кодируют посредством блока дескрипторов msar и блока дескрипторов mmap.

В соответствии с еще одним аспектом информацию об оценках выравнивания рида кодируют посредством блока дескрипторов mscore.

В соответствии с еще одним аспектом информацию о группах, которым принадлежат риды, кодируют посредством блока специальных дескрипторов rgroup.

В соответствии с еще одним аспектом способ кодирования дополнительно отличается тем, что блоки дескрипторов содержат таблицу главного индекса, содержащую по одному разделу для каждого класса и подкласса выровненных ридов, причем раздел содержит положения картирования на одну или более референсных последовательностей первого рида каждого блока доступа каждого класса или подкласса данных; при этом таблицу главного индекса и данные блока доступа кодируют вместе.

В соответствии с еще одним аспектом способ кодирования дополнительно отличается тем, что блоки дескрипторов также содержат информацию, относящуюся к типу использованного референса (ранее существующий или построенный) и сегментам ридов, которые не совпадают с референсной последовательностью.

В соответствии с еще одним аспектом способ кодирования дополнительно отличается тем, что референсные последовательности сначала преобразуют в другие референсные последовательности путем применения замен, инсерций, делеций и отсечений, а затем кодируют классифицированные выровненные риды в виде множества блоков дескрипторов, относящихся к преобразованным референсным последовательностям.

В соответствии с еще одним аспектов способ кодирования дополнительно отличается тем, что к референсным последовательностям для всех классов данных применяют одно и то же преобразование.

В соответствии с еще одним аспектов способ кодирования дополнительно отличается тем, что к референсным последовательностям каждого класса данных применяют разные преобразования.

В соответствии с еще одним аспектом способ кодирования дополнительно отличается тем, что преобразования референсных последовательностей кодируют в виде блоков дескрипторов и структурируют с помощью информации заголовка с созданием тем самым последовательных блоков доступа.

В соответствии с еще одним аспектом способ кодирования дополнительно отличается тем, что кодирование классифицированных выровненных ридов и связанных преобразований референсных последовательностей в виде множества блоков дескрипторов включает в себя этап связывания конкретной модели источника и конкретного энтропийного кодера с каждом блоком дескрипторов.

В соответствии с еще одним аспектом способ кодирования дополнительно отличается тем, что энтропийный кодер представляет собой контекстно-адаптивный арифметический кодер, кодер переменной длины или кодер Голомба.

В настоящем изобретении также предложен способ декодирования кодированных геномных данных, включающий в себя следующие этапы:

синтаксический анализ блоков доступа, содержащих кодированные геномные данные, для выделения множества блоков дескрипторов путем использования информации заголовка;

декодирование множества блоков дескрипторов для выделения выровненных ридов в соответствии с заданными правилами сопоставления, определяющими их классификацию относительно одного или более референсных ридов.

В соответствии с еще одним аспектом способ декодирования дополнительно включает в себя декодирование некартированных геномных ридов.

В соответствии с еще одним аспектом способ декодирования дополнительно включает в себя декодирование классифицированных геномных ридов.

В соответствии с еще одним аспектом способ декодирования дополнительно включает в себя декодирование таблицы главного индекса, содержащей по одному разделу для каждого класса ридов и связанные соответствующие положения картирования.

В соответствии с еще одним аспектом способ декодирования дополнительно включает в себя декодирование информации, относящейся к типу используемого референса - ранее существующий, преобразованный или построенный.

В соответствии с еще одним аспектом способ декодирования дополнительно включает в себя деко-

дирование информации, относящейся к одному или более преобразованиям, которые должны быть применены к ранее существующим референсным последовательностям.

В соответствии с еще одним аспектом способ декодирования отличается тем, что геномные риды являются парными.

В соответствии с еще одним аспектом способ декодирования дополнительно включает в себя случай, когда геномные данные энтропийно декодируют.

В настоящем изобретении также предложен геномный кодер (2010) для сжатия данных (209) геномной последовательности, причем данные (209) геномной последовательности содержат риды последовательностей нуклеотидов, при этом геномный кодер (2010) содержит

узел (201) выравнивателя, выполненный с возможностью выравнивания ридов на одну или более референсных последовательностей с созданием тем самым выровненных ридов;

узел (202) генератора конструированной последовательности, выполненный с возможностью создания конструированных референсных последовательностей;

узел (204) классифицирования данных, выполненный с возможностью классифицирования выровненных ридов в соответствии с заданными правилами сопоставления с одной или более ранее существующими референсными последовательностями или построенными референсными последовательностями с созданием тем самым классов выровненных ридов (208);

один или более узлов (205-207) кодирования блоков, выполненных с возможностью кодирования классифицированных выровненных ридов в виде блоков дескрипторов путем выбора дескрипторов в соответствии с классами выровненных ридов;

мультиплексор (2016) для мультиплексирования сжатых геномных данных и метаданных.

В соответствии с еще одним аспектом геномный кодер дополнительно содержит узел (2019) преобразования референсной последовательности, выполненный с возможностью преобразования ранее существующих референсов и классов (208) данных в преобразованные классы (2018) данных.

В соответствии с еще одним аспектом геномный кодер дополнительно отличается тем, что узел (204) классифицирования данных содержит конфигурируемые с помощью векторов порогов кодеры классов данных N, M и I, формирующие подклассы данных классов N, M и I.

В соответствии с еще одним аспектом геномный кодер дополнительно отличается тем, что узел (2019) преобразования референса применяет одно и то же преобразование (300) референса для всех классов и подклассов данных.

В соответствии с еще одним аспектом геномный кодер дополнительно отличается тем, что декодер (2019) преобразования референса применяет разные преобразования (301, 302, 303) к разным классам и подклассам данных.

В соответствии с еще одним аспектом геномный кодер дополнительно содержит функции, пригодные для осуществления всех аспектов ранее упомянутых способов кодирования.

В настоящем изобретении также предложен геномный декодер (218) для разуплотнения сжатого геномного потока (211), причем геномный декодер (218) содержит

демультиплексор (210) для демультиплексирования сжатых геномных данных и метаданных;

средства (212-214) синтаксического анализа, выполненные с возможностью синтаксического анализа сжатого геномного потока с получением геномных блоков дескрипторов (215);

один или более декодеров (216-217) блоков, выполненных с возможностью декодирования геномных блоков в классифицированные риды последовательностей нуклеотидов (2111);

декодеры (219) классов геномных данных, выполненные с возможностью выборочного декодирования классифицированных ридов последовательностей нуклеотидов относительно одной или более референсных последовательностей с получением несжатых ридов последовательностей нуклеотидов.

В соответствии с еще одним аспектом геномный декодер дополнительно содержит декодер (2113) преобразования референса, выполненный с возможностью декодирования дескрипторов (2112) преобразования референса и создания преобразованного референса (2114) для использования декодерами (219) классов геномных данных.

В соответствии с еще одним аспектом геномный декодер дополнительно отличается тем, что одну или более референсных последовательностей хранят в сжатом потоке (211) геномных данных.

В соответствии с еще одним аспектом геномный декодер дополнительно отличается тем, что одну или более референсных последовательностей подают в декодер посредством внеполосного механизма.

В соответствии с еще одним аспектом геномный декодер дополнительно отличается тем, что одну или более референсных последовательностей строят в декодере.

В соответствии с еще одним аспектом геномный декодер дополнительно отличается тем, что одну или более референсных последовательностей преобразуют в декодере с помощью декодера (2113) преобразования референса.

В настоящем изобретении также предложен машиночитаемый носитель информации, содержащий инструкции, исполнение которых вызывает осуществление по меньшей мере одним процессором всех аспектов ранее упомянутых способов кодирования.

В настоящем изобретении также предложен машиночитаемый носитель информации, содержащий

инструкции, исполнение которых вызывает осуществление по меньшей мере одним процессором всех аспектов ранее упомянутых способов декодирования.

В настоящем изобретении также предложена поддержка хранения геномных данных, кодированных в соответствии с выполнением всех аспектов ранее упомянутых способов кодирования.

Подробное описание

В число геномных или протеомных последовательностей, упоминаемых в данном изобретении, входят, например, помимо прочего, нуклеотидные последовательности, последовательности дезоксирибонуклеиновой кислоты (ДНК), рибонуклеиновой кислоты (РНК) и аминокислотные последовательности. Несмотря на то что в данном документе достаточно подробно описана геномная информация в форме нуклеотидной последовательности, ясно, что способы и системы сжатия могут быть реализованы также для других геномных или протеомных последовательностей, хотя и с некоторыми изменениями, понятными специалисту в данной области.

Информация о секвенировании генома формируется приборами высокопроизводительного секвенирования (HTS) в виде последовательностей нуклеотидов (именуемых также "основаниями"), представленных строками буквенных символов из определенного словаря. Минимальный словарь представлен пятью символами: {A, C, G, T, N}, которые обозначают 4 типа нуклеотидов, присутствующих в ДНК, а именно аденин, цитозин, гуанин и тимин. В рибонуклеиновой кислоте (РНК) тимин заменен урацилом (U). N указывает на то, что секвенатор не смог определить никакого основания, из-за чего реальный характер нуклеотида в этом положении не определен. Если в секвенаторе приняты коды неоднозначности IUPAC, то алфавит, используемый для символов, представляет собой (A, C, G, T, U, W, S, M, K, R, Y, B, D, H, V, N или -). Последовательности нуклеотидов, создаваемые секвенаторами, называют "ридами". Риды последовательности могут быть длиной от нескольких десятков до нескольких тысяч нуклеотидов. Некоторые технологии секвенирования создают риды последовательностей в виде "пар", где один из ридов которых происходит из одной цепочки ДНК, а второе - из другой цепочки. В секвенировании генома термин "покрытие" используют для выражения уровня избыточности данных последовательности относительно "референсной последовательности". Например, чтобы достичь покрытия 30× генома человека (длиной 3,2 миллиарда оснований) секвенатор должен создать всего 30×3,2 миллиарда оснований, чтобы в среднем каждое положение в референсе "покрывалась" 30 раз.

Везде в данном описании референсная последовательность - это любая последовательность, на которую выравнивают/картируют последовательности нуклеотидов, создаваемые секвенаторами. Одним примером последовательности может быть "референсный геном" - последовательность, собранная учеными в качестве репрезентативного примера наборов генов вида. Например, GRC37, геном человека Консорциума референсного генома (сборка 37), получен из образцов тринадцати анонимных добровольцев из г. Буффало, штат Нью-Йорк. Однако референсная последовательность может также состоять из синтетической последовательности, задуманной и попросту построенной для улучшения сжимаемости ридов с учетом их дальнейшей обработки. Это более подробно описано в разделе "Дескрипторы класса U и построение "внутренних" референсов для некартированных ридов "класса U" и "класса HM"" и изображено на фиг. 22 и 23.

Секвенаторы могут вносить в риды последовательности ошибки, такие как следующие.

1. Решение о пропуске определения основания ввиду отсутствия уверенности в определении какого-либо конкретного основания. Это называют неизвестным основанием и помечают как "N" (обозначают как несовпадение "n-типа").

2. Использование неверного символа (т.е. представление другой нуклеиновой кислоты) для представления нуклеиновой кислоты, на самом деле присутствующей в секвенированном образце; это обычно называют "ошибкой замены" (обозначают как несовпадение "s-типа").

3. Вставка в один рид последовательности дополнительных символов, которые не имеют отношения ни к одной действительно присутствующей нуклеиновой кислоте; это обычно называют "ошибкой вставки" (обозначают как несовпадение "i-типа").

4. Удаление из одного рида последовательности символов, которые представляют нуклеиновые кислоты, действительно присутствующие в секвенированном образце; это обычно называют "ошибкой удаления" (обозначают как несовпадение "d-типа").

5. Рекомбинация одного или более фрагментов в один фрагмент, который не отражает подлинную сущность исходной последовательности; этот обычно приводит к решению средства выравнивания отсечь основания (обозначают как несовпадение "c-типа").

Термин "покрытие" используют в литературе для количественного определения степени, в которой референсный геном или его часть могут быть охвачены имеющимися ридами последовательности. Покрытие называют

частичным (менее 1X), когда на некоторые части референсного генома не картировано ни одного имеющегося рида последовательности;

однократным (1X), когда на все нуклеотиды референсного генома картирован один и только один символ, представленный в ридах последовательности;

многократным (2X, 3X, NX), когда каждый нуклеотид референсного генома картирован несколько раз.

Цель настоящего изобретения состоит в определении формата представления геномной информации, в котором можно эффективно получать доступ к релевантной информации и транспортировать ее и в котором вес избыточной информации снижен. Раскрытое изобретение имеет следующие элементы новизны.

1. Риды последовательности классифицируют и разделяют на классы данных в соответствии с результатами выравнивания относительно референсных последовательностей. Такие классификация и разделение обеспечивают возможность выборочного доступа к кодированным данным в соответствии с критериями, относящимися к результатам выравнивания и точности совпадения.

2. Классифицированные риды последовательности и связанные метаданные представляют в виде однородных блоков дескрипторов для получения четко различимых источников информации, отличающихся низкой информационной энтропией.

3. Возможность моделирования каждого отдельного источника информации с помощью четко различимой модели источника, адаптированной к статистическим характеристикам каждого класса, и возможность изменения модели источника в пределах каждого класса и в пределах блока дескрипторов для каждого индивидуально доступного блока данных (блока доступа). Внедрение подходящих контекстно-адаптивных вероятностных моделей и соответствующих энтропийных кодеров в соответствии со статистическими свойствами каждой модели источника.

4. Определение соответствий и зависимостей между блоками дескрипторов для обеспечения возможности выборочного доступа к данным секвенирования и связанным метаданным без необходимости декодирования всех блоков дескрипторов, если требуется не вся информация.

5. Кодирование каждого блока класса данных последовательности и связанных метаданных относительно "ранее существующих" (также называемых "внешними") референсных последовательностей или относительно "преобразованных" референсных последовательностей, получаемых путем применения надлежащих преобразований к "ранее существующим" референсным последовательностям, чтобы снизить энтропию источников информации блоков дескрипторов. Дескрипторы представляют риды, разделенные на различные классы данных. После любого кодирования ридов с использованием соответствующих дескрипторов относительно "ранее существующего" референса или "преобразованной" "ранее существующей" референсной последовательности встречаемость различных несовпадений можно использовать для определения подходящих преобразований в референсные последовательности с целью нахождения конечного кодированного представления с низкой энтропией и достижения более высокой эффективности сжатия.

6. Построение одной или более референсных последовательностей (также называемых "внутренними" референсами для отличия их от "ранее существующих", также называемых "внешними" референсными последовательностями), используемых для кодирования класса ридов, которые демонстрируют степень точности совпадения относительно ранее существующих референсных последовательностей, не удовлетворяющую установленным ограничениям. Такие ограничения устанавливают с той целью, чтобы затраты на кодирование представления в сжатой форме класса ридов, выровненных относительно "внутренних" референсных последовательностей, и затраты на представление самих "внутренних" референсных последовательностей были ниже, чем при кодировании невыровненного класса представлений буквально или с использованием "внешних" референсных последовательностей без преобразований или с преобразованиями. Далее каждый из вышеупомянутых аспектов будет дополнительно описан в подробностях.

Классификация ридов последовательности в соответствии с правилами сопоставления.

Риды последовательности, формируемые секвенаторами, в соответствии с настоящим изобретением подразделяют на шесть различных "классов" в зависимости от результатов сопоставления при выравнивании относительно одной или более "ранее существующих" референсных последовательностей.

При выравнивании ДНК-последовательности нуклеотидов относительно референсной последовательности можно выделить следующие случаи.

1. Область в референсной последовательности совпадает с ридами последовательности без единой ошибки (т.е. идеальное картирование). Такую последовательность нуклеотидов называют "идеально совпадающим ридом" или обозначают как "класс P".

2. Область в референсной последовательности совпадает с ридом последовательности, причем тип и количество несовпадений определяются только количеством положений, в которых секвенатору, формировавшему этот рид, не удалось определить никакого основания (или нуклеотида). Несовпадения такого типа обозначают буквой "N", указывающей на неопределенное нуклеотидное основание. В настоящем документе несовпадение этого типа называют несовпадением "n-типа". Такие последовательности принадлежат к ридам "класса N". После того как рид классифицирован как принадлежащий "классу N", полезно ограничить степень неточности совпадения данной верхней границей и разграничить совпадения, считающиеся действительными и не считающиеся действительными. Поэтому риды, назначенные классу N, также ограничивают путем установки порога (MAXN), определяющего максимально допустимое коли-

чество неопределенных оснований (т.е. оснований, определенных как "N"), которое может содержать рид. Такая классификация неявно определяет требуемую минимальную точность совпадения (или максимальную степень несовпадения), общую для всех ридов, принадлежащих классу N, при сравнении с соответствующей референсной последовательностью, что дает полезный критерий для применения выборочного поиска данных к сжатым данным.

3. Область в референсной последовательности совпадает с ридом последовательности, причем типы и количества несовпадений, определяются количеством положений, в которых секвенатору, формирующему рид, не удалось определить никакого нуклеотидного основания, при наличии таковых (т.е. несовпадение "n-типа"), плюс количеством положений, в которых было определено основание, отличное от присутствующего в референсе. Такой тип несовпадения, обозначаемый как "замена", называют также однонуклеотидной вариацией (ОНВ) или однонуклеотидным полиморфизмом (ОНП). В данном документе несовпадение этого типа называют также несовпадением "s-типа". В таком случае рид последовательности называют "ридами с М-несовпадением" и назначают "классу М". Как и в случае "класса N", для всех ридов, принадлежащих "классу М", полезно ограничить степень неточности совпадения данной верхней границей и разграничить совпадения, считающиеся действительными и не считающиеся действительными. Поэтому риды, назначенные классу М, также ограничивают путем установки набора порогов, одного для количества "n" несовпадений "n-типа" (MAXN), при наличии таковых, а другого для количества замен "s" (MAXS). Третье ограничение представляет собой порог, определяемый любой функцией от обоих чисел "n" и "s", f(n,s). Такое третье ограничение позволяет формировать классы с верхней границей неточности совпадения в соответствии с любым значимым критерием выборочного доступа. Например, без ограничений, f(n,s) может быть (n+s)1/2, или (n+s), или линейным или нелинейным выражением, которое устанавливает границу для максимального уровня неточности совпадения, принятого для рида, принадлежащего "классу М". Такая граница представляет собой очень полезный критерий для применения требуемого выборочного поиска данных к сжатым данным при анализе ридов последовательности в различных целях, так как она позволяет устанавливать дополнительную границу для любой возможной комбинации количеств несовпадений "n-типа" и несовпадений (замен) "s-типа" сверх простого порога, применяемого к одному типу или другому типу.

4. Четвертый класс составляют риды последовательности, представляющие по меньшей мере одно несовпадение любого типа из числа "инсерции", "делеции" (иначе называемых инделами) и "отсечения", плюс (при наличии таковых) любые несовпадения типа, принадлежащего классу N или М. Такие последовательности называют "ридами с I-несовпадениями" и назначают "классу I". Инсерции образуются дополнительной последовательностью из одного или более нуклеотидов, отсутствующих в референсе, но присутствующих в последовательности рида. В данном документе несовпадение этого типа называют несовпадением "i-типа". В литературе, когда вставленная последовательность, находится на краях последовательности, ее называют также "мягко отсеченной" (т.е. нуклеотиды не совпадают с референсной последовательностью, но остаются в выровненных ридах, в отличие от "жестко отсеченных" нуклеотидов, которые отбрасывают). В данном документе несовпадение этого типа называют несовпадением "c-типа". Решения оставлять или отбрасывать нуклеотиды принимаются на этапе средства выравнивания, а не раскрытым в данном изобретении средством классифицирования ридов, которое принимает и обрабатывает риды в том виде, в котором они определены секвенатором или на следующей стадии выравнивания. Делеции представляют собой "дыры" (отсутствующие нуклеотиды) в риде в сравнении с референсной последовательностью. В данном документе несовпадения этого типа называют несовпадениями "d-типа". Как и в случае классов "N" и "М" можно и нужно определять передел неточности несовпадения. Определение набора ограничений для "класса I" основывается на тех же принципах, которые используются для "класса М" и указаны в последних строках табл. 1. Помимо порога для каждого типа несовпадения, допустимого для данных класса I, определяют дополнительное ограничение с помощью порога, определяемого посредством любой функции от количеств несовпадений "n", "s", "d", "i" и "c" - w(n,s,d,i,c). Такое дополнительное ограничение позволяет формировать классы с верхней границей неточности совпадения в соответствии с любым значащим критерием выборочного доступа, определяемым пользователем. В качестве примера, но не ограничения w(n,s,d,i,c) может представлять собой (n+s+d+i+c)1/5, или (n+s+d+i+c), или любое линейное или нелинейное выражение, устанавливающее границу для максимального уровня неточности совпадения, который принимают для рида, принадлежащего "классу I". Такая граница представляет собой очень полезный критерий для применения требуемого выборочного поиска данных к сжатым данным при анализе ридов последовательности в различных целях, так как она позволяет устанавливать дополнительную границу для любой возможной комбинации количеств несовпадений, допустимых в ридах "класса I" сверх простого порога, применяемого к каждому типу допустимого несоответствия.

5. Пятый класс включает в себя все риды, для которых не найдено ни одного картирования, считающегося действительным (т.е. не удовлетворяющие набору правил, определяющих верхнюю границу для максимальной неточности совпадения, как указано в табл. 1) для каждого класса данных при сравнении с референсной последовательностью. Такие последовательности называют "некартированными" относительно референсных последовательностей и классифицирую как принадлежащие "классу U".

Классификация пар ридов в соответствии с правилами сопоставления.

Классификация, определенная в предыдущем разделе, относится к одиночным ридами последовательности. В случае технологий секвенирования, формирующих риды парами (т.е. технологий Illumina Inc.), в отношении двух составляющих ридов которых известно, что они разделены неизвестной последовательностью переменной длины, целесообразно рассматривать отнесение всей пары к одному классу данных. Рид, который спарен с другим ридом, называют его "парным ридом".

Если оба спаренных рида принадлежат одному и тому же классу, назначение пары классу очевидно: всею пару назначают тому же классу, что и любую из частей пары (т.е. P, N, M, I, U). Если два рида принадлежат разным классам, но ни одно из них не принадлежит "классу U", то всю пару назначают классу с наивысшим приоритетом, определяемым согласно следующему выражению:

$$P<N<M<I,$$

где "класс P" обладает самым низким приоритетом, а "класс I" - самым высоким приоритетом.

В случае когда одно из ридов принадлежит "классу U", а его парный рид принадлежит любому из классов P, N, M, I, определяют шестой класс как "класс HM", что означает "наполовину картированный".

Определение такого специфического класса ридов мотивировано тем фактом, что его используют в попытке определить гэпы или неизвестные области, существующие в референсных геномах (называемые также малоизвестными или неизвестными областями). Такие области восстанавливают путем картирования пар на краях с использованием парного рида, который может быть картирован на неизвестные области. Затем некартированный рид пары используют для построения так называемых "контигов" неизвестной области, как показано на фиг. 28. Поэтому обеспечение выборочного доступа только к парам ридов такого типа сильно сокращает соответствующую вычислительную нагрузку, позволяя обрабатывать такие данные, происходящие из больших количеств наборов данных, значительно эффективнее, чем с использованием решений известного уровня техники, потребовавших бы полной проверки.

В приведенной ниже таблице обобщены правила сопоставления, применяемые к ридам с целью определения класса данных, которому принадлежит каждый рид. Правила определены в первых пяти столбцах таблицы в форме наличия или отсутствия типов несовпадений (несовпадений типов n, s, d, i и c). В шестом столбце приведены правила в виде максимального порога для каждого типа несовпадения и какой-либо функции (при наличии таковой) f(n,s) и w(n,s,d,i,c) от возможных типов несовпадения.

Таблица 1

Тип несовпадений и набор ограничений, которым должно удовлетворять каждый рид последовательности, относимый к классам данных, определенным в данном описании

Количество и типы несовпадений, обнаруживаемых при сопоставлении рида с референсной последовательностью					Набор ограничений на точность совпадения	Класс назначения
Кол-во Неизвестных оснований («N»)	Кол-во замен	Кол-во делеций	Кол-во инсерций	Кол-во отсеченных оснований		
0	0	0	0	0	0	P
n > 0	0	0	0	0	n ≤ MAXN	N
					n > MAXN	U
n ≥ 0	s > 0	0	0	0	n ≤ MAXN и s ≤ MAXS и f(n,s) ≤ MAXM	M
					n > MAXN или s > MAXS или f(n,s) > MAXM	U

n ≥ 0	s ≥ 0	d ≥ 0*	i ≥ 0*	c ≥ 0*	n ≤ MAXN и s ≤ MAXS и d ≤ MAXD и i ≤ MAXI и c ≤ MAXC и w(n,s,d,i,c) ≤ MAXTOT	I
		***Должно быть по меньшей мере одно несовпадение типа d, i, c (т. е. d > 0, или i > 0, или c > 0)**				
		d ≥ 0	i ≥ 0	c ≥ 0	n > MAXN или s > MAXS или d > MAXD или i > MAXI или	U
					c > MAXC и w(n,s,d,i,c) > MAXTOT	

Правила сопоставления для разделения классов данных ридов N, M и I на подклассы с различными степенями точности совпадения.

Классы данных типа N, M и I, которые определены в предыдущих разделах, могут быть дополнительно разделены на произвольное количество различных подклассов с разными степенями точности совпадения. Такая возможность является важным техническим преимуществом в обеспечении разделения на более мелкие подмножества и, как следствие, намного более эффективного выборочного доступа к каждому классу данных. В качестве примера, но не ограничения для разбиения класса N на к подклассов (подкласс N_1, ..., подкласс N_k) необходимо определить вектор с соответствующими компонентами $MAXN_1$, $MAXN_2$, ..., $MAXN_{(k-1)}$, $MAXN_{(k)}$ при условии, что $MAXN_1<MAXN_2<...<MAXN_{(k-1)}<MAXN$, и назначить каждый рид подклассу самого низкого ранга, который удовлетворяет ограничениям, указанным в табл. 1, при выполнении оценки для каждого элемента вектора. Это показано на фиг. 29, где блок (291) классифицирования данных содержит кодер класса P, N, M, I, U, HM и кодеры для аннотаций и метаданных. Кодер класса N конфигурируют с помощью вектора 292 порогов $MAXN_1$-$MAXN_k$, который формирует к подклассов данных N (296).

В случае классов типа M и I применяют такой же принцип путем определения вектора с теми же свойствами для MAXM и MAXTOT, соответственно, и используют компоненты каждого вектора в качестве порога, чтобы проверять, удовлетворяют ли ограничению функции f(n,s) и w(n,s,d,i,c). Как и в случае подклассов типа N, назначают подклассу самого низкого ранга, удовлетворяющему ограничению. Количество подклассов для класса каждого типа является независимой величиной и допускается любая комбинация подразделений. Это показано на фиг. 29, где кодер 293 класса M и кодер 294 класса I конфигурируют, соответственно, с помощью вектора порогов $MAXM_1$-$MAXM_j$ и $MAXTOT_1$-$MAXTOT_h$. Эти два кодера формируют, соответственно, j подклассов (297) данных M и h подклассов (298) данных I.

Если два рида в паре отнесены к одному и тому же подклассу, то пара принадлежит этому же подклассу.

Если два рида в паре отнесены к подклассам разных классов, то пара принадлежит подклассу класса с более высоким приоритетом в соответствии со следующим выражением:

$$N<M<I,$$

где N имеет самый низкий приоритет, а I имеет самый высокий приоритет.

Если два рида принадлежат разным подклассам одного из классов N, M или I, то пара принадлежит подклассу с самым высоким приоритетом в соответствии со следующим выражениям:

$$N_1<N_2<\ldots<N_k,$$
$$M_1<M_2<\ldots<M_j,$$
$$I_1<I_2<\ldots I_h,$$

где самый высокий индекс имеет наивысший приоритет.

Преобразования "внешних" референсных последовательностей.

Несовпадения, найденные для ридов, отнесенных к классам N, M и I, могут быть использованы для создания "преобразованных" референсов, которые будут применены для более эффективного сжатия представления рида.

Риды, классифицированные как принадлежащие классам N, M или I (относительно "ранее существующей" (т.е. "внешней") референсной последовательности, обозначенной как RS_0), могут быть кодированы относительно "преобразованной" референсной последовательности RS_1 в соответствии с встречаемостью действительных несовпадений относительно "преобразованного" референса. Например, если $рид^{M}_{in}$, принадлежащий классу M (обозначен как i-й рид класса M), содержит несовпадения относительно референсной последовательности RS_n, то после "преобразования" $рид^{M}_{in}=рид^{P}_{i(n+1)}$ может быть получено

с помощью $A(Ref_n)=Ref_{n+1}$, где A является преобразованием из референсной последовательности RS_n в референсную последовательность RS_{n+1}. На фиг. 19 показан пример того, как риды, содержащие несовпадения (принадлежащие классу M) относительно референсной последовательности 1 (RS_1), могут быть преобразованы в риды, идеально совпадающие с референсной последовательностью 2 (RS_2), полученной из RS_1 путем изменения оснований, соответствующих положениям несовпадения. Они остаются классифицированными и их кодируют вместе с другими ридами в одном и том же блоке доступа, но кодирование выполняют с использованием только дескрипторов и значений дескрипторов, необходимых для рида класса P. Это преобразование можно представить в виде следующего выражения:

$$RS_2=A(RS_1)$$

Когда представление преобразования A, которое формирует RS_2, будучи примененным к RS_1, и представление ридов относительно RS_2 соответствуют более низкой энтропии, чем представление ридов класса M относительно RS_1, целесообразно передавать представление преобразования A и соответствующее представление рида относительно RS_2, поскольку достигается более высокое сжатие представления данных.

Кодирование преобразования A для передачи в сжатом потоке двоичных данных требует определения двух дополнительных синтаксических элементов синтаксиса, как указано в приведенной ниже таблице.

Дескрипторы	Семантика	Комментарии
rftp	Положение преобразования референса	Положения отличия между референсом и контигом, используемым для предсказания
rftt	Тип преобразования референса	Тип отличия между референсом и контигом, используемым для предсказания Такой же синтаксис описан для дескриптора snpt, определенного далее в этом документе.

На фиг. 26 приведен пример порядка применения преобразования референса для уменьшения количества несовпадений, подлежащих кодированию, на картированных ридах.

Необходимо отметить последствия применения преобразования к референсу в некоторых случаях:

возможно внесение несовпадений в представления ридов, которые отсутствовали бы при сравнении с референсом до применения преобразования;

возможно изменение типов несоответствий; рид может содержать A вместо G, в то время как все остальные риды содержат C вместо G, но несовпадения остаются в том же положении;

различные классы данных и подмножества данных каждого класса данных могут относиться к одним и тем же "преобразованным" референсным последовательностям или к референсным последовательностям, полученным применением разных преобразований к одной и той же ранее существующей референсной последовательности.

На фиг. 27 также показан пример того, как риды могут менять тип кодирования с одного класса данных на другой с помощью надлежащего набора дескрипторов (например, с использованием дескрипторов класса P для кодирования рида из класса M) после применения преобразования и представления рида с помощью "преобразованного" референса. Это происходит, например, когда преобразование меняет все основания, соответствующие несовпадениям рида, на основания, фактически присутствующие в риде, тем самым фактически преобразуя рид, принадлежащее классу M (если сравнивать с исходной не "преобразованной" референсной последовательностью), в виртуальный рид класса P (если сравнивать с "преобразованным" рефренсом). Определение набора дескрипторов, используемых для каждого класса данных, приведено в следующих разделах.

На фиг. 30 показано, как разные классы данных могут использовать один и тот же "преобразованный" референс $R_1=A_0(R_0)$ (300) для повторного кодирования ридов, или к каждому классу данных могут быть по отдельности применены разные преобразования A_N (301), A_M (302), A_I (303).

Определение информации, необходимой для представления ридов последовательности в виде блоков дескрипторов.

После того как классифицирование ридов завершено определением классов, дальнейшая обработка состоит из определения набора различных дескрипторов, которые представляют остальную информацию, позволяющую восстанавливать последовательность рида, когда ее представляют как картируемую на данную референсную последовательность. Структура данных этих дескрипторов требует хранения глобальных параметров и метаданных, которые будут использованы механизмом декодирования. Эти данные структурированы в заголовке геномного набора данных, описанном в приведенной ниже таблице. Набор данных определяют как совокупность элементов кодирования, необходимых для восстановления геномной информации, относящейся к одному прогону секвенирования генома и всему последующему анализу. Если один и тот же образец для генома секвенируют дважды за два разных прогона, полученные данные будут кодированы в двух разных наборах данных.

Таблица 1а

Структура заголовка геномного набора данных

Элемент	Тип	Описание
Уникальный ИД	Массив байтов	Уникальный идентификатор для кодированного содержимого
Главная торговая марка	Массив байтов	Основная + дополнительная версии алгоритма кодирования
Дополнительная версия	Массив байтов	
Размер заголовка	Целое число	Размер в байтах всего кодированного содержимого
Длина ридов	Целое число	Размер ридов в случае постоянной длины ридов. Специальное значение (например, 0) зарезервировано для ридов переменной длины
Счетчик реф.	Целое число	Количество используемых референсных последовательностей
Счетчики блоков доступа	Массив байтов (например, целые числа)	Общее количество кодированных блоков доступа на референсную последовательность
Идентификаторы референсов	Массив байтов	Уникальные идентификаторы для референсных последовательностей
for (i=0; i<Ref_count; i++) {		
Reference_genome:Ref_ID	Строк:строка	Однозначный ИД в виде строки символов, определяющий референсные последовательности, используемые в данном наборе данных
}		
for (i=0; i<Ref_count; i++) {		
Блоки реф.	Массив байтов	Количество кодированных блоков на каждый референс
}		
Размер метки набора данных	Целое число	Размер следующего элемента

Метка набора данных	Строка	Строка символов, используемая для идентификации набора данных
Тип набора данных	Целое число	Тип данных, кодированных в наборе данных (например, выровненные, невыровненные)
Таблица главного индекса Положения выравнивания первого рида в каждом блоке (блоке доступа). Т. е. младшее положение первого рида на референсном геноме для каждого блока из шести классов. 1 на класс pos (шесть) на референсную последовательность	Массив байтов	Это многомерный массив, поддерживающий произвольный доступ к блокам доступа
Список меток Подмножество главного заголовка, указывающее • количество меток; • для каждой метки: ○ ИД метки; ○ количество референсных последовательностей, имеющих отношение к метке; ○ для каждой референсной последовательности: ▪ идентификатор референса; ▪ количество областей, охватываемых меткой; ▪ для каждой области: • ИД класса; • начальное положение в геномном диапазоне; • конечное положение в геномном диапазоне.	Массив байтов	Это список меток, каждая из которых представлена в виде многомерного массива для поддержки выборочно доступа к конкретным геномным областям или подобластям, или объединениям областей или подобластей
Начальное положение и конечное положение можно заменить «номерами блоков», вместе с ИД референсной последовательности и ИД класса составляющими трехмерный вектор для адресации координат таблицы главного индекса		
Набор параметров	Массив байтов	Параметры кодирования, используемые для конфигурирования процесса кодирования и отправляемые в декодер

Рид последовательности (т.е. сегмент ДНК), относящееся к данной референсной последовательности, может быть выражено с помощью следующих элементов.

Начальное положение на референсной последовательности (pos).

Флаг, сигнализирующий, нужно ли последовательность рассматривать как обратный комплемент относительно референса (rcomp).

Расстояние до парного рида в случае парных ридов (pair).

Значение длины рида в случае технологии секвенирования с получением ридов переменной длины (len). В случае постоянной длины ридов длину ридов, связанную с каждым ридом, очевидно можно опустить и сохранить в главном заголовке файла.

Для каждого несовпадения

положение (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, указывающие определенные характеристики ридов последовательности, такие как

шаблон, имеющий множество сегментов при секвенировании;

каждый сегмент, надлежащим образом выровненный в соответствии со средством выравнивания; некартированный сегмент;

следующий некартированный сегмент в шаблоне;

сигнализация о первом или последнем сегменте;

непрохождение контроля качества;

дубликат ПЦР или оптический дубликат;

вспомогательное выравнивание;

дополнительное выравнивание.

Строка мягко отсеченных нуклеотидов, если присутствуют (indc в классе I).

Флаг, указывающий референс, использованный для выравнивания и сжатия (например, "внутренний" референс для класса U), если применимо (дескриптор rtype).

Для класса U дескриптор indc указывает те части ридов (как правило, края), которые не соответствуют заданному набору ограничений на точность совпадения с "внутренними" референсами.

Дескриптор ureads используют для буквального кодирования ридов, которые невозможно картировать ни на один имеющийся референс, будь то ранее существующая (т.е. "внешняя", такая как фактический референсный геном) или "внутренняя" референсная последовательность.

Данная классификация создает группы дескрипторов (дескрипторы), которые могут быть использованы для однозначного представления ридов геномной последовательности. В приведенной ниже таблице обобщены дескрипторы, необходимые для каждого класса ридов, выровненных относительно "внешних" (т.е. ранее существующих) или "внутренних" (т.е. построенных) референсов.

Таблица 2

Определенные блоки дескрипторов для каждого класса данных

	P	N	M	I	U	HM
pos	X	X	X	X	X	X
pair	X	X	X	X		
rcomp	X	X	X	X	X	X
flags	X	X	X	X	X	X
rlen	X	X	X	X	X	X
nmis		X				
snpp			X		X	
snpt			X		X	
indp				X		X
indt				X		X
indc				X	X	X
ureads					X	X
rtype					X	
rgroup	X	X	X	X	X	X
mmap	X	X	X	X	X	
msar	X	X	X	X	X	
mscore	X	X	X	X	X	

Риды, принадлежащие классу P, отличаются тем, что они могут быть идеально восстановлены с помощью лишь положения, информации об обратном комплементе и смещении между парными ридами В случае когда они были получены с помощью технологии секвенирования, обеспечивающей парноконцевые риды с длинной вставкой, некоторых флагов и длины рида.

В следующем разделе подробно описано определение этих дескрипторов для классов P, N, M и I, а для класса U они описаны в последующем разделе.

Класс HM применим только к парам ридов и представляет собой особый случай, в котором один рид принадлежит классу P, N, M или I, а другое принадлежат классу U.

Дескриптор положения.

В блоке положения (pos) хранится только положение картирования первого картированного рида в виде абсолютного значения на референсной последовательности.

Все остальные дескрипторы положения принимают значение, выражающее разницу относительно предыдущего положения. Такое моделирование источника информации, определяемого последовательностью дескрипторов положения рида, как правило, отличается пониженной энтропией, особенно для процессов секвенирования, формирующих результаты с высоким покрытием.

Например, на фиг. 1 показано, как после описания начального положения первого выравнивания как положения "10 000" на референсной последовательности положение второго рида, начинающегося в положении 10 180, описана как "180". При высоких покрытиях (>50×) большинство дескрипторов вектора положения демонстрируют очень высокую встречаемость низких значений, таких как 0 и 1 и другие малые целые числа. На фиг. 1 показано, как положения трех парных ридов описаны в блоке pos.

Дескриптор обратного комплемента.

Каждый рид пары ридов, создаваемых с помощью технологий секвенирования, может происходить из любой из двух цепочек генома секвенируемого органического образца.

Однако в качестве референсной последовательности используют только одну из двух цепочек. На фиг. 2 показано, как в паре ридов один рид (рид 1) может происходить из одной цепочки, а другой рид (рид 2) может происходить из другой цепочки.

Если цепочку 1 используют в качестве референсной последовательности, то рид 2 может быть кодирован в качестве обратного комплемента соответствующего фрагмента на цепочке 1. Это показано на фиг. 3.

В случае сдвоенных ридов возможны четыре комбинации пар из прямого и обратного парного комплемента пар. Это показано на фиг. 4. Блок rcomp кодирует четыре возможные комбинации.

То же самое кодирование используют для информации об обратном комплементе ридов, принадлежащих классам N, M, P и I. Чтобы обеспечить выборочный доступ к различным классам данных, информацию об обратных комплементах ридов, принадлежащих четырем классам, кодируют в разных блоках, как показано в табл. 2.

Дескриптор информации о спаривании.

Дескриптор спаривания сохраняют в блоке pair. Такой блок хранит дескрипторы, кодирующие информацию, необходимую для восстановления возникающих пар ридов, когда используемая технология секвенирования создает риды парами. Хотя на момент раскрытия изобретения подавляющее большинство данных секвенирования формируют с использованием технологии, создающей парные риды, не все технологии являются таковыми. Именно поэтому наличие данного блока необязательно для восстановления всей информации данных секвенирования, если рассматриваемая технология секвенирования геномных данных не дает информации о парных ридах.

Определения.

Парный рид: рид, связанный с другим ридом в паре ридов (например, в предыдущем примере рид 2 является парным ридом рида 1);

расстояние спаривания: количество положений нуклеотидов референсной последовательности, которые разделяют одно положение в первом риде (якорь спаривания, например, последний нуклеотид первого рида) от одного положения второго рида (например, первого нуклеотида второго рида);

наиболее вероятное расстояние спаривания (MPPD): это наиболее вероятное расстояние спаривания, выраженное в количестве положений нуклеотидов;

расстояние спаривания положений (PPD): PPD представляет собой способ выражения расстояния спаривания в количестве ридов, отделяющих один рид от его соответствующего парного рида, присутствующего в определенном блоке дескрипторов положения;

наиболее вероятное расстояние спаривания положений (MPPPD): это наиболее вероятное количество ридов, отделяющих один рид от его парного рида, присутствующего в определенном блоке дескрипторов положения;

ошибка спаривания положений (PPE): определяется как разница между MPPD или MPPPD и фактическоий положением парного рида;

якорь спаривания: положение последнего нуклеотида первого рида в паре, используемая в качестве точки отсчета для вычисления расстояния до парного рида в количестве положений нуклеотидов или количестве положений рида.

На фиг. 5 показано, как вычисляют расстояние спаривания среди пар ридов.

Блок дескрипторов pair представляет собой вектор ошибок спаривания, вычисленных как количество ридов, которые нужно пропустить, чтобы достичь парного рида первого рида пары относительно определенного расстояния спаривания декодирования.

На фиг. 6 показан пример порядка вычисления ошибок спаривания, как в виде абсолютного значения, так и в виде разностного вектора (отличающегося низкой энтропией при высоких покрытиях).

Те же самые дескрипторы используют для информации о спаривании ридов, принадлежащих классам N, M, P и I. Чтобы обеспечить выборочный доступ к разным классам данных, информацию о спаривании ридов, принадлежащих этим четырем классам, кодируют в разных блоках, как изображено на фиг. 8

(класс N), фиг. 10, 12 и 14 (класс M) и фиг. 15 и 16 (класс I).

Информация о спаривании в случае ридов, картированных на разные референсные последовательности.

В процессе картирования ридов последовательности на референсную последовательность нередко первый риж в паре картируют на одну референсную последовательность (например, хромосому 1), а второй - на другую референсную последовательность (например, хромосому 4). В этом случае описанную выше информацию об образовании пар необходимо объединять с дополнительной информацией, относящейся к референсной последовательности, используемой для картирования одного из ридов. Это достигают путем кодирования

1) зарезервированного значения (флага), указывающего, что пару картируют на две разных последовательности (разные значения указывают, ли рид 1 или рид 2 на последовательность, которая не кодирована в настоящее время);

2) уникального идентификатора референса, относящегося к идентификаторам референса, кодированным в главной структуре заголовка, как описано в табл. 1a;

3) третий элемент содержит информацию о картировании на референс, указанный в пункте 2, и выражен в виде смещения относительно последней кодированного положения.

Пример данного сценария приведен на фиг. 7.

Поскольку на фиг. 7 рид 4 не картирован на текущую кодируемую референсную последовательность, геномный кодер сигнализирует об этой информации, создавая дополнительные дескрипторы в блоке pair. В примере, показанном ниже, рид 4 пары 2 картируют на референс № 4, тогда как в данный момент картируют референс № 1. Эту информацию кодируют с помощью 3 компонентов:

1) одно специальное зарезервированное значение кодируют как расстояние спаривания (в данном случае 0xffffff);

2) второй дескриптор обеспечивает ИД референса, который перечислен в главного заголовке (в данном случае 4);

3) третий элемент содержит информацию о картировании на соответствующий референс (170).

Дескрипторы несовпадения для ридов класса N.

Класс N включает в себя все риды, в которых имеются только несовпадения "n-типа" - вместо основания A, C, G или T в качестве определенного основания обнаружено N. Все другие основания ридов идеально совпадают с референсной последовательностью.

На фиг. 8 показано, как

положение "N" в риде 1 кодируют в виде

абсолютного положения в риде 1, или

виде разностного положения относительно предыдущего "N" в этом же риде;

положение "N" в риде 2 кодируют в виде

абсолютного положения в риде 2 + длина рида 1, или

разностного положения относительно предыдущей N.

В блоке nmis кодирование каждой пары ридов завершают специальным символом "разделителя".

Дескрипторы, кодирующие замены (несовпадения и ОНП), инсерции и делеции.

Замену определяют как наличие в картированном риде другого основания нуклеотида по сравнению с присутствующим в референсной последовательности в этом же положении. На фиг. 9 показаны примеры замен в картированной паре ридов. Каждую замену кодируют в виде "положения" (блок snpp) и "типа" (блок snpt). В зависимости от статистической встречаемости замен, инсерций или делеций можно определять разные модели источника соответствующих дескрипторов и формировать символы, кодируемые в соответствующем блоке.

Модель источника 1: замены в виде положений типов.

Дескрипторы положений замен.

Положение замены вычисляют аналогично значениям блока nmis, а именно следующее.

В риде 1 замены кодируют

как абсолютное положение в риде 1, или

как разностное положение относительно предыдущей замены в этом же риде.

В риде 2 замены кодируют

как абсолютное положение в риде 2 + длину рида 1, или

как разностное положение относительно предыдущей замены.

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

1) положения несовпадения

относительно начала рида, или

относительно предыдущего несовпадения (дифференциальное кодирование);

2) типа несовпадения, представленного в виде кода, вычисленного, как описано на фиг. 10.

В блоке snpp кодирование каждой пары ридов завершают специальным символом "разделителя".

Дескрипторы типов замен.

Для класса M (и для класса I, как описано в следующих разделах) несовпадения кодируют посредством индекса (перемещения справа налево) от фактического символа, присутствующего в референсной последовательности, к соответствующему символу замены, присутствующему в риде (A, C, G, T, N, Z). Например, если в выровненном риде присутствует C вместо T, которая присутствует в этом же положении в референсной последовательности, индекс несовпадения будет обозначен как "4". Процесс декодирования считывает кодированный дескриптор, нуклеотид в данном положении на референсе и перемещается слева направо для извлечения декодированного символа. Например, "2", принятое для положения, где в референсе присутствует G, будет декодировано как "N". На фиг. 11 показаны все возможные ситуации и соответствующие символы кодирования. Чтобы свести к минимуму энтропию дескрипторов, каждому индексу могут быть назначены явно отличающиеся друг от друга и контекстно-адаптивные вероятностные модели в соответствии со статистическими свойствами каждого типа замены для каждого класса данных.

В случае принятия кодов неоднозначности IUPAC результаты механизм замены будут теми же, однако вектор замены расширен 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 блоке snpt. Некоторые примеры кодирования замен при внедрении кодов неоднозначности IUPAC приведены на фиг. 13. Еще один пример индексов замен приведен на фиг. 14.

Кодирование инсерций и делеций.

Для класса I несовпадения и делеций кодируют посредством индексов (перемещения справа налево) от фактического символа, присутствующего в референсной последовательности, к соответствующему символу замены, присутствующему в риде {A, C, G, T, N, Z}. Например, если в выровненном риде присутствует C вместо T в том же положении в референсе, индексом несовпадения будет "4". В случае когда в риде присутствует делеция в положении, где в референсе присутствует "A", кодированным символом будет "5". Процесс декодирования считывает кодированный дескриптор, нуклеотид в данном положении на референсе и переходит слева направо для извлечения декодированного символа. Например, "3", принятое для положения, где в референсе присутствует G, будет декодировано как "Z".

Для вставленных A, C, G, T, N инсерций кодируют как 6, 7, 8, 9, 10, соответственно. На фиг. 15 показан пример, как кодируют замены, инсерции и делеции в парах ридов класса I. Для поддержки полного набора кодов неоднозначности IUPAC вектор замены

S={A, C, G, T, N, Z}

следует заменить 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 данном случае, если вектор замены имеет 16 элементов, коды инсерции должны иметь разные значения, а именно 16, 17, 18, 19, 20. Этот механизм проиллюстрирован на фиг. 16.

Модель источника 2: один блок на тип замены и инделы.

Для некоторых статистик данных может быть разработана отличающаяся от описанной в предыдущем разделе модель кодирования замен и инделов, которая приводит к источнику с более низкой энтропией. Такая модель кодирования является альтернативой методам, описанным выше только для несовпадений и для несовпадений и инделов. В данном случае определяют один блок данных для каждого возможного символа замены (5 без кодов IUPAC, 16 с кодами IUPAC) плюс один блок для делеций и еще 4 блока для инсерций. Для простоты объяснения, но не в качестве ограничения для применения модели, дальнейшее описание будет сконцентрировано на случае, когда коды IUPAC не поддерживаются.

На фиг. 17 показано, как каждый блок содержит положение несовпадений или инсерций одного типа. Если в кодируемой паре ридов нет несовпадений или инсерций данного типа, в соответствующем блоке кодируют 0. Чтоб декодер мог начать процесс декодирования для блоков, описанных в данном разделе, заголовок каждого блока доступа содержит флаг, указывающий посредством сигнализации первый блок, подлежащий кодированию. В примере на фиг. 18 первый элемент, подлежащий декодированию, находится в положении 2 в блоке C. Когда в паре ридов нет несовпадений или инделов данного типа, в соответствующие блоки добавляют 0. Когда на стороне декодирования указатель декодирования для каждого блока указывает на значение 0, процесс декодирования переходит к следующей паре ридов.

Кодирование дополнительных флагов сигнализации.

Для каждого класса данных, введенного выше (P, M, N, I), может потребоваться кодирование дополнительной информации о характере кодируемых ридов. Эта информация может относиться, например, к эксперименту секвенирования (например, указание вероятности дублирования рида) или может выражать некоторые характеристики картирования рида (например, первое или второе в паре). В контексте настоящего изобретения эту информацию кодируют в отдельном блоке для каждого класса данных. Основное преимущество такого подхода состоит в возможности выборочного доступа к этой информации только в случае необходимости и только в требуемой области референсной последовательности. Другие примеры использования таких флагов:

парные риды;
рид, картированный в правильной паре;
некартированный рид или парный рид;
рид или парный рид из обратной цепочки;
первый/второй рид в паре;
не основной рид;
рид не проходит проверки качества платформы/поставщика;
рид представляет собой дубликат ПЦР или оптический дубликат;
дополнительное выравнивание.

Дескрипторы для класса U и построение "внутренних" референсов для некартированных ридов "класса U" и "класса HM"

В случае ридов, принадлежащих классу U или некартированной паре "класса HM", поскольку они не могут быть картированы ни на одну "внешнюю" референсную последовательность, удовлетворяющую заданному набору ограничений на точность совпадения для принадлежности классу P, N, M или I, для сжатого представления ридов, принадлежащих этим классам данных, строят и используют одну или более "внутренних" референсных последовательностей.

К построению подходящих "внутренних" референсов возможны несколько подходов, таких как, для примера, но не ограничения следующие.

Разделение некартированных ридов на кластеры, содержащие риды, у которых имеется общая непрерывная геномная последовательность по меньшей мере минимального размера (сигнатура). Каждый кластер может быть единственным образом идентифицирован по его сигнатуре, как показано на фиг. 22.

Сортировка ридов в любом имеющем смысл порядке (например, в лексикографическом порядке) и использование последних N ридов в качестве "внутреннего" референса для кодирования (N+1)-го. Этот способ показан на фиг. 23.

Выполнение так называемой "сборки заново" на подмножестве ридов класса U для получения возможности выравнивания и кодирования всех или соответствующего подмножества ридов, принадлежащих этому классу в соответствии с заданными ограничениями на точность совпадения или новым набором ограничений.

Если кодируемый рид может быть картирован на "внутренний" референс, удовлетворяющий заданному набору ограничений на точность совпадения, информацию, необходимую для восстановления рида после сжатия, кодируют с помощью дескрипторов, которые могут быть одного из следующих типов.

1. Начальное положение совпадающей части на внутреннем референсе в виде номера рида во внутреннем референсе (блок pos). Это положение можно кодировать в виде абсолютного или разностного значения относительно ранее кодированного рида.

2. Смещение начального положения от начала соответствующего рида во внутреннем референсе (блок pair). Например, в случае постоянной длины рида фактического положения будет pos *length+pair.

3. Возможно присутствующие несовпадения, кодированные в виде положения (блок snpp) и типа (блок snpt).

4. Те части ридов (обычно края, указываемые посредством pair), которые не совпадают с внутренним референсом (или совпадают, но с количеством несовпадений выше заданного порога), кодируют в блоке indc. Для снижения энтропии несовпадений, кодируемых в блоке indc, можно применить операцию заполнения к краям части используемого внутреннего референса, как показано на фиг. 24. Наиболее подходящая стратегия заполнения может быть выбрана кодером в соответствии со статистическими свойствами обрабатываемых геномных данных. Возможные следующие стратегии заполнения:

a) без заполнения;

b) постоянный шаблон заполнения, выбираемый в соответствии с его частотой в текущих кодируемых данных;

c) переменный шаблон заполнения в соответствии со статистическими свойствами текущего контекста, определенными посредством последних N кодированных ридов.

Конкретный тип стратегии заполнения будет сигнализирован специальными значениями в заголовке блока indc.

5. Флаг, который указывает, кодирован ли рид с помощью внутреннего самоформируемого референса, внешнего референса или без референса (блок rtype).

6. Рид, которые кодируют буквально (ureads). Пример такой процедуры кодирования приведен на фиг. 24.

На фиг. 25 показано альтернативное кодирование накартированных ридов на внутренний референс, где дескрипторы pos+pair заменены дескриптором pos со знаком. В данном случае pos выражает расстояние, в единицах положений, на референсной последовательности - положения левого крайнего нуклеотида рида n относительно положения крайнего левого нуклеотида рида n-1.

В случае наличия ридов переменной длины класса U для сохранения длины каждого рида используют дополнительный дескриптор rlen.

Этот подход к кодированию можно расширить для поддержки N начальных положений для каждого

рида, чтобы риды могли быть разбиты на две и более положений референса. Это может быть особенно полезно для кодирования ридов, формируемых с помощью тех технологий секвенирования (например, от Pacific Bioscience), которые создают очень длинные риды (свыше 50K оснований), в которых обычно присутствуют повторяющиеся структуры, образуемые за счет циклов в технологии секвенирования. Такой же подход можно использовать и для кодирования химерных ридов последовательности, определяемых как риды, которые выравнены на две разные положения генома с небольшим перекрытием или без него.

Ясно, что описанный выше подход может быть применен за рамками простого класса U и мог бы быть применен к любому блоку, содержащему дескрипторы, относящиеся к положениям ридов (блоки pos).

Дескриптор оценки выравнивания.

Дескриптор mscore обеспечивает оценку для каждого выравнивания. В контексте настоящего изобретения его используют для представления оценки картирования/выравнивания для каждого рида, формируемого средствами выравнивания ридов геномной последовательности.

Оценку выражают с помощью степени и мантиссы. Количество битов, используемых для представления степени и мантиссы, передают в качестве параметров конфигурации. В качестве примера, но не ограничения в табл. 2а показано, как это определено в стандарте IEEE RFC 754 для 11-битовой степени и 52-битовой мантиссы.

Оценку каждого выравнивания можно представить с помощью

одного бита знака (S);

11 битов для степени (E);

53 битов для мантиссы (M).

Таблица 2а

Оценки выравнивания могут быть выражены
в виде 64-битовых значений двойной точности с плавающей запятой

1	11	52
S	Степ.	Мантисса
63	62 51	0

Основание (radix), предназначенное для вычисления оценок, равно 10, поэтому

$$\text{оценка} = -1^{s} \times 10^{E} \times M$$

Группы ридов.

Во время процесса секвенирования могут быть получены секвенированные риды разных типов. В качестве примера, но не ограничения типы могут относиться к разным секвенированным образцам, разным экспериментам, разным конфигурациям секвенатора.

После секвенирования и выравнивания эту информацию сохраняют в соответствии с настоящим изобретением с помощью специально предназначенного дескриптора под названием rgroup. rgroup - это метка, связанная с каждым кодированным ридом, которая позволяет декодирующему устройству разделять декодированные риды на группы после декодирования.

Дескрипторы для множественных выравниваний

Следующие дескрипторы определяют для поддержки множественных выравниваний. Для случая наличия сплайсированных ридов в настоящем изобретении определен глобальный флаг spliced_reads_flag, который должен быть установлен на 1.

Дескриптор mmap.

Дескриптор mmap используют для сигнализации о том, сколько положений ридов или крайнего левого рида пары были выровнены. Геномная запись, содержащая множественные выравнивания, связана с одним многобайтовым дескриптором mmap.

Первые два байта дескриптора mmap представляют целое число без знака N, которое относится к риду как к единому сегменту (если в кодированном наборе данных нет сплайсов) или, напротив, ко всем сегментам, на которые был сплайсирован рид для нескольких возможных выравниваний (если в наборе данных присутствуют сплайсы).

Значение N указывает, сколько значений дескриптора pos кодированы для шаблона в данной записи. После N следуют одно или более целых чисел без знака M_i, как описано ниже.

Цепочечность множественных выравниваний.

Дескриптор rcomp, описанный в настоящем изобретении, используют для указания цепочечности каждого выравнивания рида с помощью синтаксиса, определенного в настоящем изобретении.

Оценки множественных выравниваний.

В случае множественных выравниваний каждому выравниванию назначают один дескриптор mscore, определенный в настоящем изобретении.

Множественные выравнивания без сплайсов.

Если в блоке доступа нет сплайсов, spliced_reads_flag не установлен.

При парноконцевом секвенировании дескриптор mmap состоит из 16-битового целого числа без

знака N, за которым следуют одно или более 8-битовых целых чисел без знака M_i, причем i принимает значения от 1 до количества полных выравниваний первого (в данном случае крайнего левого) рида. Для каждого выравнивания первого рида, сплайсированного или нет, M_i используют для сигнализации о том, сколько сегментов использованы для выравнивания второго рида (в данном случае без сплайсов оно равно количеству выравниваний) и, кроме того, сколько значений дескриптора pair кодированы для данного выравнивания первого рида.

Значения M_i используют для вычисления $P = \sum_{i=1}^{N} Mi$, которое указывает количество выравниваний второго рида.

Специальное значение M_i (M_i=0) указывает, что i-е выравнивание крайнего левого рида спарено с выравниванием крайнего правого рида, которое уже спарено с k-м выравниванием крайнего левого рида, причем k<i (в таком случае новое выравнивание не обнаружено, что согласуется с приведенным выше уравнением). В качестве примера в самых простых случаях,

если имеются одно выравнивание для крайнего левого рида и два альтернативных выравнивания для крайнего правого рида, N будет равно 1, а M_1 будет равно 2;

если обнаружены два альтернативных выравнивания для крайнего левого рида, но лишь одно для крайнего правого рида, N будет равно 2, M_1 будет равно 1 и M_2 будет равно 0;

когда M_i равно 0, соответствующе значение pair должно быть привязано к существующему выравниванию второго рида;

в противном случае возникнет синтаксическая ошибка, и выравнивание будет считаться нарушенным.

Пример. Если первый рид имеет два положения картирования, а второй рид только одну, N равно 2, M_1 равно 1 и M_2 равно 0, как указано выше. Если за этим следует другое альтернативное вспомогательное картирование для всего шаблона, N будет равно 3, а M_3 будет равно 1.

Фиг. 39 иллюстрирует значение N, P и M_i в случае множественных выравниваний без сплайсов, а на Error! Reference source not found. На фиг. 10 показано, как дескрипторы pos, pair и mmap используют для кодирования информации о множественных выравниваниях.

В отношении фиг. 40 справедливо следующее:

Крайнее правый рид имеет $P = \sum_{i=1}^{N} M_i$ выравниваний.

Некоторые значения M_i могут быть =0, когда i-е выравнивание крайнего левого рида спарено с выравниванием крайнего правого рида, которое уже же спарено с k-м выравниванием крайнего левого рида, причем k<i.

Может присутствовать одно зарезервированное значение дескриптора pair для сигнализации о принадлежности выравниваний другим диапазонам блоков доступа. При наличии такового оно всегда является первым дескриптором pair для текущей записи.

Множественные выравнивания со сплайсами.

При кодировании набора данных со сплайсированными ридами дескриптор msar позволяет представлять длину и цепочечность сплайсов.

После декодирования дескрипторов mmap и msar декодер знает, сколько ридов или пар ридов было кодировано для представления нескольких картирований, и сколько сегментов составляют каждое картирование каждого рида или пары ридов. Это показано на фиг. 41 и 42.

В отношении фиг. 41 справедливо следующее.

Крайний левый рид имеет N_1 выравниваний с N сплайсами ($N1 \leq N$).

N представляет количество сплайсов, присутствующих во всех выравниваниях крайнего левого рида, и его кодируют в качестве первого значения дескриптора mmap.

Крайний правый рид имеет $P = \sum_{i=1}^{N1} Mi$ сплайсов, где M_i представляет собой количество сплайсов крайнего правого рида, который связан в пару с i-м выравниванием крайнего левого рида ($1 \leq i \leq N_1$). Другими словами, P представляет количество сплайсов крайнего правого рида и вычисляется с помощью N значений, следующих за первым значением дескриптора mmap.

N_1 и N_2 представляют количество выравниваний первого и второго рида и вычисляются с помощью N + P значений дескриптора msar.

В отношении фиг. 42 справедливо следующее.

Крайний левый рид имеет N_1 выравниваний с N сплайсами ($N_1 \leq N$). Если N_1=N и N_2=P, сплайсы будут отсутствовать.

Крайний правый рид имеет $P = \sum_{i=1}^{N1} Mi$ сплайсов (t_j $1 \leq j \leq P$) = и N_2 ($N_2 \leq P$) выравниваний.

Количество дескрипторов pair можно вычислить как

$$NP=Max(N_1, P)+M_0,$$

где M_0 - количество M_i со значением 0;

NP должно приращиваться на 1, и в таком случае один специальный дескриптор pair указывает на присутствие выравниваний в других БД.

Оценка выравнивания.

Дескриптор mscore позволяет посредством сигнализации указывать оценку выравнивания. При одноконцевом секвенировании он будет иметь N_1 значений на шаблон; при парноконцевом секвенировании он будет иметь значение для каждого выравнивания всего шаблона (количество различных выравнивание первого рида и, возможно, плюс количество дополнительных выравниваний второго рида, т.е. когда M_i-1>0).

Количество оценок= $MAX(N_1, N_2)+M_0$, где M_0 представляет общее количество M_i=0.

В настоящем изобретении с каждым выравниванием могут быть связаны более одного значения оценки. Количество выравниваний сигнализирует параметр конфигурации as_depth.

Дескрипторы для множественных выравниваний без сплайсов.

Таблица 3

Определение количества дескрипторов, необходимых для представления множественных выравниваний в одной геномной записи в случае множественных выравниваний без сплайсов

Семантика	mmap	mscore	Результат
Рид (или парноконцевой рид) с несколькими картированиями, не сплайсированное	Одиночный рид: N Пара ридов: N, M_i где $1 \le i \le N$	Одиночный рид: N значений Пара ридов: MAX(N, $\sum(M_i)$) значений + $\sum(M_i = 0)$ Введение разделителя позволяет иметь произвольное количество оценок. В противном случае, если он не присутствует, следует использовать 0. Это значения с плавающей запятой, как определено в настоящем изобретении.	Одиночный рид: рид имеет несколько картирований, и его кодируют в виде последовательности из N последовательных сегментов, принадлежащих классу с самым высоким идентификатором. Используют N дескрипторов pos. Пара ридов: пара ридов имеет несколько картирований, и ее кодируют в виде последовательности из: N сегментов для первого рида; P =$\sum(M_i)$ спариваний для выравниваний второго рида. Используют N дескрипторов **pos**. Используют N x P дескрипторов **pair**, причем NP = $\sum_{i=1}^{N} Mi$ + (к-во Mi = 0) + (необязательно) 1 Необязательный дескриптор спаривания используют, когда выравнивания присутствуют на других референсных последовательностях, а не на кодируемой в данный момент. N + 1 дескрипторов mmap

Рид (пара), картированный единственным образом	0	0	

Дескрипторы для множественных выравниваний со сплайсами.

В табл. 4 показано, как определять количество дескрипторов, необходимых для представления множественных выравниваний в одной геномной записи в случае множественных выравниваний без сплайсов.

Таблица 4

Дескрипторы, используемые для представления множественных выравниваний и соответствующих оценок

Семантика	mmap	mscore	Результат
Рид (или парноконцевой рид) с несколькими картированиями, со сплайсами	Одиночный рид: N Пара ридов: N, M_i где $1 \le i \le N$	Одиночный рид: N_1 значений Пара ридов: $Max(N_1, N_2) + \sum(M_i == 0)$ значений N_1 и N_2 с помощью N + P дескрипторов **msar** $P = \sum_{i=1}^{N1} Mi$ Это значения с плавающей запятой, как определено в настоящем изобретении.	Одиночный рид: рид имеет несколько картирований, и его кодируют в виде последовательности из N последовательных сегментов, принадлежащих классу с самым высоким идентификатором. Используют N дескрипторов **pos**. Пара ридов: пара ридов имеет несколько картирований, и ее кодируют в виде последовательности из: N сегментов для первого рида; P =$\sum(M_i)$ спариваний для выравниваний второго рида. Используют N дескрипторов pos.
Рид (пара), картированный единственным образом	0	0	

Множественные выравнивания на разных последовательностях.

Может получиться так, что в процессе выравнивания обнаружатся альтернативные картирования на другую референсную последовательность, а не на ту, на которой расположено основное картирование.

Для пар ридов, которые выровнены единственным образом, дескриптор pair следует использовать для представления абсолютных положений рида, когда имеется, например, химерическое выравнивание с парным ридом на другой хромосоме. Дескриптор pair следует использовать для указания посредством сигнализации референса и положения следующей записи, содержащей дополнительные выравнивания для того же самого шаблона. Последняя запись (например, третья, если альтернативные картирования кодируют в 3 разных БД) должна содержать референс и положение первой записи. В случае когда одно или более выравниваний для крайнего левого рида в паре присутствуют не на той референсной последовательности, которая относится к текущему кодируемому БД, для дескриптора pair использую зарезервированное значение. После зарезервированного значения следуют идентификатор референсной последовательности и положение крайнего левого выравнивания среди всех содержащихся в следующем БД (т.е. первое декодированное значение дескриптора pos для этой записи).

Множественные выравнивания с инсерциями, делециями, некартированными частями.

Когда альтернативное вспомогательное картирование не сохраняет непрерывность референсной области, где выравнивают последовательность, восстановление точного картирования, сформированного средством выравнивания, может оказаться невозможным, так как фактическая последовательность (а значит, дескрипторы, связанные с несовпадениями, такими как замены или инделы) кодирована только

для основного выравнивания. Для представления того, как вспомогательные выравнивания картированы на референсную последовательность В случае когда они содержит инделы и/или мягкие отсечения, следует использовать дескриптор msar. Если msar представлен специальным символом "*" для вспомогательного выравнивания, декодер восстановит вспомогательное выравнивание из основного выравнивания и положения картирования вспомогательного выравнивания.

Дескриптор msar.

Дескриптор msar (Multiple Segments Alignment Record) поддерживает сплайсированные риды и альтернативные вспомогательные выравнивания, которые содержат инделы или мягкие отсечения.

msar предназначен для передачи информации о

длине картированного сегмента;

другой непрерывности картирования (т.е. наличии инсерций, делеций или отсеченных оснований) для вспомогательного выравнивания и/или сплайсированного рида.

В msar используют синтаксис расширенной записи CIGAR, описанный ниже, плюс дополнительный символ, описанный в табл. 5.

Таблица 5

Специальный символ, используемый для дескриптора msar в дополнение к синтаксису, описанному в табл. 6

Символ	Семантика	Описание
*	Вспомогательное выравнивание не содержит инделов или мягких отсечений	Это используют, когда восстановление вспомогательного выравнивания не требует никакой дополнительной информации, кроме положения выравнивания и основного выравнивания

Расширенный синтаксис cigar.

В данном разделе определен расширенный синтаксис CIGAR (E-CIGAR) для строк, которые должны быть связаны с последовательностями и соответствующими несовпадениями, инделами, отсеченными основаниями и информацией о множественных выравниваниях и сплайсированных ридах.

Операции редактирования, описанные в настоящем изобретении, перечислены в табл. 6.

Таблица 6

Синтаксис строки E-CIGAR в формате MPEG-G

Операция	**Семантика**	**Представление E-CIGAR**	**Эквивалент представления CIGAR для SAM**
Приращение указателя на референс *R* и указателя на рид *r* на *n* положений (совпадение)	n совпадающих оснований	n=	nM в более старых версиях (не эквивалентно), = в недавних версиях
Замена нуклеотида в риде основанием C из референса, приращение	Замена символа C (C присутствует в риде и отсутствует в референсе)	C	M в более старых версиях,

указателя на референс *R* и указателя на рид *r* на 1			X в недавних версиях (не эквивалентно)
Приращение указателя на риде *r* на *n* положений (вставка из рида)	n оснований вставляют в риде (не присутствуют в референсе)	n+	nI
Приращение указателя на референс *R* на *n* положений (удаление последовательности S в риде)	n оснований удаляют из рида (но присутствуют в референсе)	n-	nD
Приращение указателя на референс *R* на *n* положений (инсерция в рид). Может иметь место только в начале или конце рида	n мягких отсечений	(n)	nS
Жесткое обрезание. Может иметь место только в начале или конце рида	n жестких отсечений	[n]	nH
Приращение указателя на референс *R* на *n* положений с соблюдением консенсуса сплайса (сплайс в риде)	Ненаправленный сплайс из n оснований	n*	nN
Приращение указателя на референс *R* на *n* положений, соблюдение консенсуса сплайса на прямой цепочке (прямой сплайс в риде)	Прямой сплайс из n оснований	n/	Не существует
Приращение указателя на референс *R* на *n* положений, соблюдение консенсуса сплайса на обратной цепочке (обратный сплайс в риде)	Обратный сплайс из n оснований	n%	Не существует

Модели источника, энтропийные кодеры и режимы кодирования.

Для каждого класса данных, подкласса и соответствующего блока дескрипторов структуры геномных данных в настоящем изобретении могут быть внедрены различные алгоритмы кодирования в соответствии с конкретными особенностями данных или метаданных, содержащихся в каждом блоке, и их статистическими свойствами. "Алгоритм кодирования" следует понимать как взаимосвязь конкретной "модели источника" блока дескрипторов с конкретным "энтропийным кодером". Конкретная "модель источника" может быть определена и выбрана для получения наиболее эффективного кодирования данных в смысле минимизации энтропии источника. При выборе энтропийного кодера можно руководствоваться соображениями эффективности кодирования и/или характеристиками вероятностного распределения и соответствующими проблемами реализации. Каждый выбор специфического "алгоритма кодирования", также называемого "режимом кодирования" может применяться к всему "блоку дескрипторов", связанному с классом или подклассом данных для всего набора данных, или может быть применен свой "режим кодирования" к каждой части дескрипторов, разделенных на блоки доступа. Каждую "модель источника", связанную с режимом кодирования, характеризуют посредством

определения дескрипторов, выдаваемых каждым источником (т.е. набора дескрипторов, используемых для представления класса данных, таких как положение ридов, информация о спаривании ридов,

несовпадения с референсной последовательностью, как определено в табл. 2);

определения связанной вероятностной модели;

определения связанного энтропийного кодера.

Дальнейшие преимущества.

Классифицирование данных последовательности на определенные классы и подклассы данных позволяет реализовывать эффективные режимы кодирования, использующие более низкую энтропию источника информации, отличающиеся моделированием последовательностей дескрипторов одиночными отдельными источниками данных (например, расстояние, положение и т.д.).

Другим преимуществом изобретения является возможность доступа только к подмножеству данных интересующего типа. Например, одной из наиболее важных областей применения в геномике является поиск отличий геномного образца по сравнению с референсом (ОНВ) или популяцией (ОНП). Сегодня анализ такого рода требует обработки полных ридов последовательности, тогда как при внедрении представления данных, раскрытого в настоящем изобретении, несовпадения уже изолированы всего лишь в три класса данных (если представляют интерес, рассматриваются также несовпадения "n-типа" и "i-типа").

Еще одно преимущество состоит в возможности выполнения эффективного транскодирования данных и метаданных, сжатых относительно определенной "внешней" референсной последовательности, в сжатые относительно другой "внешней" референсной последовательности при опубликовании новых референсных последовательностей или при выполнении повторного картирования на уже картированные данные (например, с помощью другого алгоритма картирования) с получением новых выравниваний.

На фиг. 20 показано кодирующее устройство 207 в соответствии с принципами настоящего изобретения. Кодирующее устройство 207 принимает в качестве входа необработанные данные 209 последовательности, например, созданные устройством 200 секвенирования генома. В данной области известны устройства для секвенирования генома, такие как приборы Illumina HiSeq 2500, Thermo-Fisher Ion Torrent. Необработанные данные 209 последовательности подают в узел 201 выравнивателя, который подготавливает последовательности для кодирования путем выравнивания ридов на референсную последовательность 2020. В альтернативном варианте осуществления может быть использован специально предназначенный модуль 202 для формирования референсной последовательности из имеющихся ридов с использованием различных стратегий, как описано в настоящем документе в разделе "Построение внутренних референсов для некартированных ридов "класса U" и "класса HM". После обработки генератором 202 референса риды могут быть картированы на полученную более длинную последовательность. Затем выровненные последовательности классифицируют с помощью модуля 204 классифицирования. После этого к референсу применяют этап преобразования референса, чтобы снизить энтропию данных, сформированных узлом 204 классифицирования данных. Он включает в себя обработку внешнего референса 2020 в узле 2019 преобразования референса, который создает преобразованные классы 2018 данных и дескрипторы 2021 преобразования референса. Затем преобразованные классы 2018 данных подают в кодеры 205-207 блоков вместе с дескрипторами 2021 преобразования референса. После этого геномные блоки 2011 подают в арифметические кодеры 2012-2014, которые кодируют блоки в соответствии со статистическими свойствами данных или метаданных, содержащихся в блоке. В результате получается геномный поток 2015.

На фиг. 21 показано декодирующее устройство 218 в соответствии с принципами настоящего изобретения. Декодирующее устройство 218 принимает мультиплексированный геномный поток 2110 двоичных данных из сети или элемента запоминающего устройства. Мультиплексированный геномный поток 2110 двоичных данных подают в демультиплексор 210 для создания отдельных потоков 211, которые затем подают в энтропийный декодеры 212-214 для получения геномных блоков 215 и дескрипторов 2112 преобразования референса. Выделенные геномные блоки подают в декодеры 216-217 блоков для дальнейшего декодирования в классы данных, а дескрипторы преобразования референса подают в узел 2113 преобразования референса. Декодеры 219 классов выполняют дальнейшую обработку геномных дескрипторов 2111 и преобразованного референса 2114 и объединяют результаты для получения несжатых ридов последовательностей, которые затем также сохраняют в форматах, известных в данной области техники, например, в текстовом файле или сжатом файле формата zip, или в файлах FASTQ или SAM/BAM.

Декодеры 219 классов могут восстанавливать исходные геномные последовательности за счет использования информации об исходных референсных последовательностях, содержащейся в одном или более геномным потоках, и дескрипторов 2112 преобразования референса в кодированном потоке двоичных данных. Если референсные последовательности не транспортируют посредством геномных потоков, они должны быть в наличии на стороне декодирования и доступны для декодеров классов. Обладающие признаками изобретения методы, описанные в настоящем документе, могут быть реализованы в оборудовании, программном обеспечении, встроенном программном обеспечении или любой их комбинации. При реализации в программном обеспечении они могут храниться на машиночитаемом носителе информации и исполняться аппаратным процессорным устройством. Аппаратное процессорное устройство может содержать один или более процессоров, цифровых сигнальных процессоров, микропроцессоров общего назначения, специализированных интегральных схем или других дискретных логических схем.

Методы данного описания могут быть реализованы в самых разных устройствах и приборах, включая мобильные телефоны, настольные компьютеры, серверы, планшеты и подобные устройства.

Формат файла: выборочный доступ к областям геномных данных с помощью таблицы главного индекса.

Для поддержки выборочного доступа к определенным областям выровненных данных структура данных, описанная в настоящем документе, реализует средство индексации, называемое таблицей главного индекса (MIT). Это многомерный массив, содержащий местоположения картирования конкретных ридов на связанные референсные последовательности. Значения, содержащиеся в MIT, представляют собой положения картирования первого рида в каждом блоке pos в целях поддержки непоследовательного доступа к каждому блоку доступа. MIT содержит один раздел на каждый класс данных (P, N, M, I, U и HM) и каждую референсную последовательность. MIT содержится в заголовке геномного набора данных кодированных данных. На фиг. 31 показана структура заголовка геномного набора данных, на фиг. 32 показано общее визуальное представление MIT, а на фиг. 33 показан пример MIT для кодированных ридов класса P.

Изображенные на фиг. 33 значения, содержащиеся в MIT, используют для прямого доступа к интересующей области (и соответствующему БД) в сжатой области. Например, как показано на фиг. 33, если требуется получить доступ к области, содержащейся между положениями 150000 и 250000 на референсе 2, декодирующее приложение должно перейти сразу ко второму референсу в MIT и найти два значения k1 и k2 так, чтобы k1<150000 и k2>250000. Где k1 и k2 являются двумя индексами, считанными из MIT. В примере на фиг. 33 в результате будут получены 3-е и 4-е положения второго вектора MIT. Эти возвращенные значения будут затем использованы декодирующим приложением для извлечения положений соответствующих данных из таблицы локального индекса блока pos, как описано в следующем разделе. Вместе с указателями на блок, содержащий данные, принадлежащие четырем классам геномных данных, описанным выше, MIT можно использовать в качестве индекса дополнительных метаданных и/или аннотаций, добавленных к геномным данным в течение их срока службы.

Таблица локального указателя.

В начале каждого блока геномных данных имеется структура данных, называемая локальным заголовком. Локальный заголовок содержит уникальный идентификатор блока, вектор счетчиков блоков доступа для каждой референсной последовательности, таблицу локального индекса (LIT) и, необязательно, некоторые метаданные, специфичные для блока. LIT представляет собой вектор указателей на физическое положение данных, принадлежащих каждому блоку доступа, в полезной нагрузке блока. На фиг. 34 изображены общий заголовок блока и полезная нагрузка, где LIT используют для доступа к определенным областям кодированных данных непоследовательным образом. В предыдущем примере для доступа к области в диапазоне от 150000 до 250000 ридов, выровненных на референсную последовательность № 2, декодирующее приложение извлекло положения 3 и 4 из MIT. Эти значения должны использоваться процессом декодирования для доступа к 3-му и 4-му элементам соответствующего раздела LIT. В примере, показанном на фиг. 35, счетчики общего количества блоков доступа, содержащиеся в заголовке блока, используют для перехода сразу к индексам LIT, связанным с БД, относящимся к референсу 1 (5 в примере). Поэтому индексы, содержащие физические положения запрашиваемых БД в кодированном потоке, вычисляют следующим образом:

положение блоков данных, принадлежащих запрашиваемым БД = блокам данных, принадлежащим БД референса 1, которые нужно пропустить, + положение, извлеченное с помощью MIT, т.е.

положение первого блока: 5+3=8;

положение последнего блока: 5+4=9.

Блоки данных, извлеченные с помощью механизма индексации, называемого таблицей локального индекса, являются частью запрашиваемых блоков доступа.

На фиг. 36 показано, как блоки, содержащиеся в MIT, соответствуют блокам LIT для каждого класса или подкласса данных.

На фиг. 37 показано, как блоки данных, извлеченные с помощью MIT и LIT, образуют один или более блоков доступа, как определено в следующем разделе.

В варианте осуществления настоящего изобретения LIT может быть встроена в качестве подструктуры MIT. Преимуществом такого подхода является скорость доступа к индексированным данным в случае последовательного синтаксического анализа сжатого файла. Если LIT встроена в MIT в заголовке файла, то в случае выборочного доступа декодирующему устройству нужно будет синтаксически анализировать только малую часть данных, чтобы извлекать запрашиваемую сжатую информацию. Другое преимущество, очевидное для специалиста в данной области, состоит в том, что в случае передачи в потоковом режиме по сети индексированная информация, содержащаяся в MIT и LIT, будет доставляться среди первых блоков данных, позволяя тем самым приемному устройству выполнять операции, такие как сортировка и выборочный доступ, до завершения передачи всех данных.

Блок доступа.

Геномные данные, классифицированные в классы данных и структурированные в сжатые и несжатые блоки, организуют в разные блоки доступа.

Геномные блоки доступа (БД) определяют как разделы геномных данных (в сжатом и несжатом виде), которые восстанавливают нуклеотидные последовательности, и/или соответствующие метаданные, и/или последовательности ДНК/РНК (например, виртуальный референс), и/или данные аннотаций, формируемые секвенатором генома и/или устройством геномной обработки или приложением анализа. Пример блока доступа приведен на фиг. 37.

Блок доступа представляет собой блок данных, который может быть декодирован либо независимо от других блоков доступа с использованием только глобально доступных данных (например, конфигурации декодера), либо с использованием информации, содержащейся в других блоках доступа. Блоки доступа различают по

типу, характеризующему природу геномных данных и наборы данных, которые содержат эти блоки доступа, и возможный способ доступа к ним;

порядку, обеспечивающему уникальный порядок блокам доступа, принадлежащим одному и тому же типу.

Блоки доступа любого типа можно дополнительно классифицировать на различные "категории".

Далее следует неисчерпывающий список определений геномных блоков доступа различных типов.

1) Для получения доступа или декодирования и получения доступа к блокам доступа типа 0 не требуется ссылка ни на какую информацию, поступающую из других блоков доступа. Вся информация в содержащихся в них данных или наборах данных, может быть независимо считана и обработана декодирующим устройством или приложением обработки.

2) Блоки доступа типа 1 содержат данные, которые ссылаются на данные, содержащиеся в блоках доступа типа 0. Для считывания или декодирования и обработки данных, содержащихся в блоках доступа типа 1, требуется наличие доступа к одному или более блокам доступа типа 0. Блок доступа типа 1 кодирует геномные данные, относящиеся к ридам последовательности "класса P".

3) Блоки доступа типа 2 содержат данные, которые ссылаются на данные, содержащиеся в блоках доступа типа 0. Для считывания или декодирования и обработки данных, содержащихся в блоках доступа типа 2, требуется наличие доступа к одному или более блокам доступа типа 0. Блок доступа типа 2 кодирует геномные данные, относящиеся к ридам последовательности "класса N".

4) Блоки доступа типа 3 содержат данные, которые ссылаются на данные, содержащиеся в блоках доступа типа 0. Для считывания или декодирования и обработки данных, содержащихся в блоках доступа типа 3, требуется наличие доступа к одному или более блокам доступа типа 0. Блоки доступа типа 3 кодируют геномные данные, относящиеся к ридам последовательности "класса M".

5) Блоки доступа типа 4 содержат данные, которые ссылаются на данные, содержащиеся в блоках доступа типа 0. Для считывания или декодирования и обработки данных, содержащихся в блоках доступа типа 4, требуется наличие доступа к одному или более блокам доступа типа 0. Блок доступа типа 4 кодирует геномные данные, относящиеся к ридам последовательности "класса I".

6) Блоки доступа типа 5 содержат риды, которые не могут быть картированы ни на одну из доступных референсных последовательностей ("класс U"), и их кодируют с использованием внутренне построенной референсной последовательности. Блоки доступа типа 5 содержат данные, которые ссылаются на данные, содержащиеся в блоках доступа типа 0. Для считывания или декодирования и обработки данных, содержащихся в блоках доступа типа 5, требуется наличие доступа к одному или более блокам доступа типа 0.

7) Блоки доступа типа 6 содержат пары ридов, где один рид может принадлежать любому из четырех классов P, N, M, I, а другое не может быть картировано ни на одну из доступных референсных последовательностей ("класс HM"). Блоки доступа типа 6 содержат данные, которые ссылаются на данные, содержащиеся в блоках доступа типа 0. Для считывания или декодирования и обработки данных, содержащихся в блоках доступа типа 6, требуется наличие доступа к одному или более блокам доступа типа 0.

8) Блоки доступа типа 7 содержат метаданные (например, оценки качества) и/или данные аннотаций, связанные с данными или наборами данных, содержащимися в блоке доступа типа 1. Блоки доступа типа 7 могут быть классифицированы и маркированы в различных блоках.

9) Блоки доступа типа 8 содержат данные или наборы данных, классифицированные как данные аннотаций. Блоки доступа типа 8 могут быть классифицированы и маркированы в блоках.

10) Блоки доступа дополнительных типов могут расширять описанные здесь структуры и механизмы. В качестве примера, но не ограничения результаты определения геномных вариантов, структурного и функционального анализа могут быть кодированы в блоках доступа новых типов. Описанная в настоящем документе организация данных в виде блоков доступа не препятствует инкапсуляции данных любого типа в блоки доступа, что является механизмом, полностью прозрачным в отношении природы закодированных данных.

Блоки доступа типа 0 упорядочены (например, пронумерованы), но их не нужно хранить и/или передавать упорядоченным образом (техническое преимущество: параллельная обработка/параллельная потоковая передача, мультиплексирование).

Блоки доступа типа 1, 2, 3, 4, 5 и 6 не нужно упорядочивать и не нужно хранить и/или передавать упорядоченным образом (техническое преимущество: параллельная обработка/параллельная потоковая

передача).

На фиг. 37 показано, как составляют блоки доступа с помощью заголовка и одного или более блоков однородных данных. Каждый блок может составлен из одного или более блоков. Каждый блок содержит несколько пакетов, а пакеты представляют собой структурированную последовательность дескрипторов, введенных выше для представления, например, положений ридов, информации о спаривании, информации об обратном комплементе, положений и типов несовпадений и т.д.

У каждого блока доступа может быть разное количество пакетов в каждом блоке, но в пределах блока доступа у всех блоков одно и то же количество пакетов.

Каждый пакет данных может быть идентифицирован посредством комбинации из 3 идентификаторов X, Y и Z, где

X указывает блок доступа, которому он принадлежит;

Y указывает блок, которому он принадлежит (т.е. тип данных, которые он инкапсулирует);

Z представляет собой идентификатор, выражающий порядок пакета относительно других пакетов того же блока.

На фиг. 38 показан пример маркировки блоков доступа и пакетов, где AU_T_N - блок доступа типа T с идентификатором N, который может подразумевать или не подразумевать наличие порядка в соответствии с типом блока доступа. Идентификаторы используют для связывания единственным образом блоков доступа одного типа с блоками доступа других типов, требуемыми для полного декодирования содержащихся геномных данных. Блоки доступа любого типа могут быть дополнительно классифицированы и маркированы в различных "категориях" в соответствии с различными процессами секвенирования. В качестве примера, но не ограничения классификация и маркировка могут иметь место при

1) секвенировании одного и того же организма в разное время (блоки доступа содержат геномную информацию по геному с "временной" подоплекой);

2) секвенировании органических образцов различной природы некоторых организмов (например, кожи, крови, волос для человеческих образцов).

Эти блоки доступа с "биологической" подоплекой.

ФОРМУЛА ИЗОБРЕТЕНИЯ

1. Реализуемый на компьютере способ кодирования данных геномной последовательности, где данные геномной последовательности содержат риды последовательностей нуклеотидов, причем способ включает в себя этапы

выравнивания указанных ридов на одну или более первых референсных последовательностей с созданием тем самым выровненных ридов,

классифицирования выровненных ридов согласно заданным правилам сопоставления с указанными одной или более первыми референсными последовательностями на различные классы с созданием тем самым классов выровненных ридов, причем классифицирование включает

идентифицирование в качестве первого класса (класса P) геномных ридов без всяких несовпадений с референсной последовательностью, где в картированном риде нет несовпадений с референсной последовательностью, использованной для картирования;

идентифицирование в качестве второго класса (класса N) геномных ридов, где несовпадения обнаруживают только в положениях, в которых секвенатор оказался не в состоянии определить никакого основания и количество несовпадений в каждом прочтении не превышает данного порога;

идентифицирование в качестве третьего класса (класса M) геномных ридов, где несовпадения обнаруживают в положениях, в которых секвенатор оказался не в состоянии определить никакого основания n-типа и/или он определил другое основание по сравнению с референсной последовательностью s-типа и количество несовпадений не превышает данных порогов для количества несовпадений n-типа, s-типа и порога, полученного из заданной функции (f(n,s));

идентифицирование в качестве четвертого класса (класса I) геномных ридов, которые могут иметь несовпадения того же типа, что и несовпадения третьего класса (класса M), и дополнительно по меньшей мере одно несовпадение следующего типа: инсерция (i-тип), делеция (d-тип), мягкие или жесткие отсечения (c-тип), и при этом количество несовпадений для каждого типа не превышает соответствующего заданного порога и порога, обеспечиваемого заданной функцией (w(n,s,i,d,c)),

кодирования классифицированных выровненных ридов в виде множества блоков дескрипторов, причем кодирование указанных классифицированных выровненных ридов в виде множества блоков дескрипторов включает в себя выбор дескрипторов в соответствии с классами выровненных ридов,

структурирования указанных блоков дескрипторов с помощью информации заголовка с созданием тем самым последовательных блоков доступа,

причем блоки доступа первого класса (класса P) строят с использованием блоков дескрипторов для информации о положении картирования, блоков дескрипторов для информации о том, из какой цепи ДНК секвенировали рид (цепочечность), и флагов для информации об определенных характеристиках рида последовательности, при этом в указанных блоках доступа класса P информацию о спаривании

парноконцевых ридов кодируют с использованием соответствующего блока дескрипторов,

при этом блоки доступа второго класса (класса N) строят с использованием тех же блоков дескрипторов, что для блоков доступа первого класса (класса P), плюс блок дескрипторов для информации о положении неизвестных оснований,

при этом блоки доступа третьего класса (класса M) строят с использованием тех же блоков дескрипторов, что для блоков доступа первого класса (класса P), плюс блоки дескрипторов для информации о положении и типах замен,

при этом блоки доступа четвертого класса (класса I) строят с использованием тех же блоков дескрипторов, что для блоков доступа первого класса (класса P), плюс блоки дескрипторов для информации о положении и типе замен, вставок, делеций и отсеченных оснований,

где способ дополнительно включает в себя

идентифицирование геномных ридов, которые не отнесены ни к одному из классов с первого по четвертый (классов P, N, M, I) в качестве пятого класса (класса U),

построение набора вторых референсных последовательностей с использованием по меньшей мере некоторых из ридов пятого класса,

выравнивание указанных ридов пятого класса на наборе вторых референсных последовательностей,

кодирование ридов пятого класса в виде соответствующих дескрипторов на основании определенных ограничений на точность совпадения в отношении вторых референсных последовательностей,

структурирование указанных соответствующих дескрипторов с помощью информации заголовка с созданием тем самым блоков доступа пятого блока.

2. Способ по п.1, в котором блоки доступа пятого класса строят с использованием одного или более из

блоков дескрипторов для информации о положениях картирования;

блоков дескрипторов для информации о том, из какой цепи ДНК секвенировали рид (цепочечность), и флагов для информации об определенных характеристиках рида последовательности, при этом информацию о спаривании парноконцевых ридов кодируют с использованием соответствующего блока дескрипторов;

блоков дескрипторов для информации о положении и типе замен;

блоков дескрипторов для информации о частях ридов, которые не совпадают со вторыми референсными последовательностями;

блоков дескрипторов для буквального кодирования ридов, которые не могут быть картированы ни на одну референсную последовательность.

3. Способ кодирования по п.1 или 2, в котором

риды геномной последовательности, подлежащие кодированию, являются парными; и

классифицирование дополнительно включает в себя идентифицирование геномных ридов в качестве шестого класса (класса HM), содержащего все пары ридов, где один рид принадлежит классу P, N, M или I, а другой рид принадлежит классу U.

4. Способ кодирования по п.3, дополнительно включающий в себя следующие этапы:

определение того, классифицированы ли два парных рида в один и тот же класс (каждый из классов P, N, M, I, U) с последующим назначением пары в этот же идентифицированный класс;

определение того, относятся ли два парных рида к разным классам, и если ни один из них не принадлежит классу U, то назначение пары ридов классу с самым высоким приоритетом, определяемым в соответствии со следующим приоритетом:

$$P<N<M<I,$$

где класс P обладает самым низким приоритетом, а класс I - самым высоким приоритетом;

определение того, классифицирован ли только один из парных ридов как принадлежащий классу U, и классифицирование пары ридов как принадлежащей последовательностям класса HM.

5. Способ по п.4, в котором каждый класс ридов N, M, I дополнительно разделяют на два или более подклассов (296, 297, 298) в соответствии с вектором порогов (292, 293, 294), определенным соответственно для каждого класса N, M и I посредством количества несовпадений (292) n-типа, функции f(n,s) (293) и функции w(n,s,i,d,c) (294), дополнительно включающий в себя следующие этапы:

определение того, классифицированы ли два парных рида в один и тот же подкласс с назначением в таком случае пары в этот же подкласс;

определение того, классифицированы ли два парных рида в подклассы разных классов с назначением в таком случае пары в подкласс, принадлежащий классу более высокого приоритета в соответствии со следующим выражением:

$$N<M<I,$$

где N имеет самый низкий приоритет, а I имеет самый высокий приоритет;

определение того, классифицированы ли два парных рида в один и тот же класс, причем этим классом является N, или M, или I, или в разные подклассы с назначением в таком случае пары в подкласс с самым высоким приоритетом в соответствии о следующими выражениями:

$$N_1<N_2<\ldots<N_k,$$

$M_1<M_2<\ldots<M_j$,

$I_1<I_2<\ldots<I_h$,

где самый высокий индекс имеет наивысший приоритет.

6. Способ по п.5, в котором

информацию о положении картирования каждого рида кодируют посредством блока дескрипторов pos;

информацию о том, из какой цепочки ДНК секвенировали рид (цепочечности), каждого рида кодируют посредством блока дескрипторов rcomp; и

информацию о парноконцевых ридах кодируют посредством блока дескрипторов pair.

7. Способ по п.6, в котором дополнительную информацию о выравнивании, такую как картирован ли рид в правильную пару, верно ли, что он не прошел проверку качества платформы/поставщика, является ли он дубликатом ПЦР или оптическим дубликатом, является ли это вспомогательным выравниванием, кодируют посредством блока дескрипторов flags.

8. Способ по п.7, в котором информацию о неизвестных основаниях кодируют посредством блока дескрипторов nmis.

9. Способ по п.8, в котором

информацию о положении замен кодируют посредством блока дескрипторов snpp;

информацию о типе замен кодируют посредством блока дескрипторов snpt.

10. Способ по п.9, в котором

информацию о положении несовпадений типа замен, инсерций или делеций кодируют посредством блока дескрипторов indp;

информацию о типе несовпадений, таких как замены, инсерции или делеции, кодируют посредством блока дескрипторов indt;

информацию об отсеченных основаниях картированного рида кодируют посредством блока дескрипторов indc.

11. Способ по п.10, в котором

информацию о некартированных ридах кодируют посредством блока дескрипторов ureads;

информацию о типе референсной последовательности, используемой для кодирования, кодируют посредством блока дескрипторов rtype;

информацию о множественных выравниваниях картированных ридов кодируют посредством блока дескрипторов mmap;

информацию о сплайсированных выравниваниях и множественных выравниваниях одного и того же рида кодируют посредством блока дескрипторов msar и блока дескрипторов mmap;

информацию об оценках выравнивания рида кодируют посредством блока дескрипторов mscore;

информацию о группах, которым принадлежат риды, кодируют посредством блока дескрипторов rgroup.

12. Способ по п.11, в котором блоки доступа класса HM строят с помощью тех же самых блоков дескрипторов, что и блоки доступа класса I для картированных ридов, и с помощью блоков дескриптора ureads для некартированных ридов.

13. Способ по п.12, в котором блоки доступа класса I строят с помощью тех же самых блоков дескрипторов, что и для блоков доступа класса P, плюс блоки дескрипторов indp, indt и indc для информации о положении и типе замен, инсерций, делеций и отсеченных оснований, причем информацию о множественных выравниваниях передают с помощью блоков дескрипторов mmap и msar.

14. Способ по п.13, в котором информацию о сплайсированных выравниваниях передают с помощью расширенной строки cigar, содержащей

символ "=" для указания совпадающих оснований,

символ "+" для указания инсерций,

символ "-" для указания делеций,

символ "/" для указания сплайса на прямой цепочке,

символ "%" для указания сплайса на обратной цепочке,

символ "*" для указания ненаправленного сплайса,

текстовый символ из кодов IUPAC для ДНК для указания замены,

символ "(n)" для указания n мягко отсеченных оснований, где n является целым числом,

символ "[n]" для указания n жестко отсеченных оснований, где n является целым числом.

15. Способ по п.14, в котором

блоки дескрипторов содержат "таблицу главного индекса", содержащую по одному разделу для каждого класса и подкласса выровненных ридов, причем раздел содержит положения картирования на одну или более референсных последовательностей первого рида каждого блока доступа каждого класса или подкласса данных;

указанную таблицу главного индекса и данные блока доступа кодируют вместе.

16. Способ по п.15, в котором

блоки дескрипторов также содержат информацию о типе использованного референса, выбранного

из ранее существующего или построенного, и сегментах рида, которые не картированы на эту референсную последовательность, при этом предпочтительно

референсные последовательности сначала преобразуют в другие референсные последовательности путем применения замен, инсерций, делеций и отсечений, а затем кодируют классифицированные выровненные риды в виде множества блоков дескрипторов, относящихся к преобразованным референсным последовательностям.

17. Способ по п.16, в котором

к референсным последовательностям, используемым для всех классов данных, применяют одно и то же преобразование; или

к референсным последовательностям, используемым для каждого класса данных, применяют разные преобразования, причем преобразования референсных последовательностей кодируют в виде блоков дескрипторов и структурируют с помощью информации заголовка с созданием тем самым последовательных блоков доступа.

18. Способ по п.14, в котором разные преобразования применяют к референсным последовательностями, используемым для каждого класса данных,

причем преобразования референсных последовательностей кодируют в виде блоков дескрипторов и структурируют с помощью информации заголовка с созданием тем самым последовательных блоков доступа,

при этом кодирование классифицированных выровненных ридов и связанных преобразований референсных последовательностей в виде множества блоков дескрипторов включает в себя этап связывания конкретной модели источника и конкретного энтропийного кодера с каждом блоком дескрипторов,

при этом энтропийный кодер представляет собой контекстно-адаптивный арифметический кодер, кодер переменной длины или кодер Голомба.

19. Реализуемый на компьютере способ декодирования кодированных геномных данных, включающий в себя следующие этапы:

синтаксический анализ блоков доступа, содержащих кодированные геномные данные, для выделения множества блоков дескрипторов путем использования информации заголовка; и

декодирование множества блоков дескрипторов для выделения ридов в соответствии с заданными правилами сопоставления, определяющими их классификацию относительно одной или более референсных последовательностей,

причем для блока доступа, принадлежащего первому, второму, третьему или четвертому классу, блоки дескрипторов описывают совпадение указанных ридов с первыми референсными последовательностями в соответствии с заданными правилами сопоставления,

где блок доступа принадлежит первому классу (классу P), если он представляет геномные риды без несовпадений с референсной последовательностью, использованной для картирования;

где блок доступа принадлежит второму классу (классу N), если он представляет геномные риды, для которых несовпадения обнаруживают только в положениях, где секвенатор оказался не в состоянии определить никакого "основания", и количество несовпадений в каждом риде не превышает данного порога;

где блок доступа принадлежит третьему классу (классу M), если он представляет геномные риды, для которых несовпадения обнаруживают в положениях, где секвенатор оказался не в состоянии определить никакого основания n-типа и/или он определил другое основание по сравнению с референсной последовательностью s-типа, и количество несовпадений не превышает заданных порогов для количества несовпадений n-типа, s-типа и порога, полученного из заданной функции (f(n,s));

где блок доступа принадлежит четвертому классу (классу I), если он представляет геномные риды, которые могут иметь несовпадения того же типа, что и несовпадения третьего класса (класса M), и дополнительно по меньшей мере одно несовпадение следующего типа: инсерция (i-тип), делеция (d-тип), мягкие или жесткие отсечения (c-тип), и при этом количество несовпадений для каждого типа не превышает соответствующего заданного порога и порога, обеспечиваемого заданной функцией (w(n,s,i,d,c));

где блоки доступа первого класса (класса P) строят с использованием блоков дескрипторов для информации о положении картирования, блоков дескрипторов для информации о том, из какой цепи ДНК секвенировали рид (цепочечности), и флагов для информации об определенных характеристиках рида последовательности, при этом в указанных блоках доступа класса P информацию о спаривании парноконцевых ридов кодируют с использованием соответствующего блока дескрипторов;

где блоки доступа второго класса (класса N) строят с использованием тех же блоков дескрипторов, что и для блоков доступа первого класса (класса P), плюс блок дескрипторов для информации о положениях неизвестных оснований;

где блоки доступа третьего класса (класса M) строят с использованием тех же блоков дескрипторов, что и для блоков доступа первого класса (класса P), плюс блоки дескрипторов для информации о положении и типах замен;

где блоки доступа четвертого класса (класса I) строят с использованием тех же блоков дескрипторов, что и для блоков доступа первого класса (класса P), плюс блоки дескрипторов для информации о

положении и типе замен, вставок, делеций и отсеченных оснований;

где блок доступа принадлежит пятому классу, если блоки дескрипторов описывают сопоставление указанных ридов со вторыми референсными последовательностями в соответствии с заданными правилами сопоставления.

20. Способ декодирования по п.19, в котором блоки доступа пятого класса строят с использованием одного или более из

блоков дескрипторов для информации о положениях картирования;

блоков дескрипторов для информации о том, из какой цепи ДНК секвенировали рид (цепочечности), и флагов для информации об определенных характеристиках рида последовательности, при этом информацию о спаривании парноконцевых ридов кодируют с использованием соответствующего блока дескрипторов;

блоков дескрипторов для информации о положении и типе замен;

блоков дескрипторов для информации о частях ридов, которые не совпадают со вторыми референсными последовательностями;

блоков дескрипторов для буквального кодирования ридов, которые не могут быть картированы ни на одну референсную последовательность.

21. Способ декодирования по п.20, дополнительно включающий в себя декодирование таблицы главного индекса, содержащей по одному разделу для каждого класса ридов и связанные соответствующие положения картирования.

22. Способ декодирования по п.21, дополнительно включающий в себя

декодирование информации, относящейся к типу используемого референса: ранее существующий, преобразованный или построенный, при этом способ дополнительно включает в себя

декодирование информации, относящейся к одному или более преобразованиям, которые должны быть применены к ранее существующим референсным последовательностям, причем блоки дескрипторов энтропийно декодируют.

23. Способ декодирования по п.22, в котором

риды класса P получают декодированием блоков дескрипторов следующего типа: pos, rcomp, flags и rlen;

риды класса N получают декодированием блоков дескрипторов следующего типа: pos, rcomp, flags, rlen и nmis;

риды класса M получают декодированием блоков дескрипторов следующего типа: pos, rcomp, flags, rlen, snpp и snpt;

риды класса I получают декодированием блоков дескрипторов следующего типа: pos, rcomp, flags, rlen, indp, indt и indc;

риды класса U получают декодированием блоков дескрипторов следующего типа: pos, rcomp, flags, rlen, snpp, snpt, indc, ureads и rtype.

24. Способ декодирования по п.23, в котором парные риды

класса P, N, M и I получают также декодированием блоков дескрипторов типа pair;

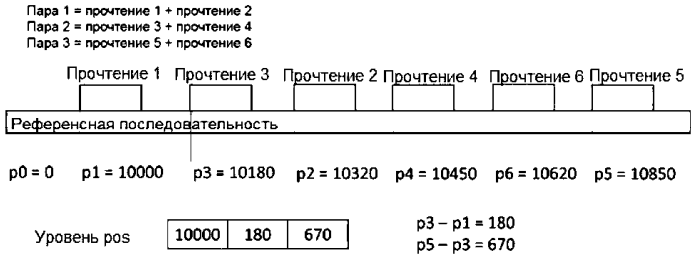
класса HM получают декодированием блоков дескрипторов типа pos, rcomp, flags, rlen, indp, indt, indc и ureads.

25. Геномный кодер (2010) для кодирования данных геномной последовательности, выполненный с возможностью осуществления любого из способов по пп.1-18.

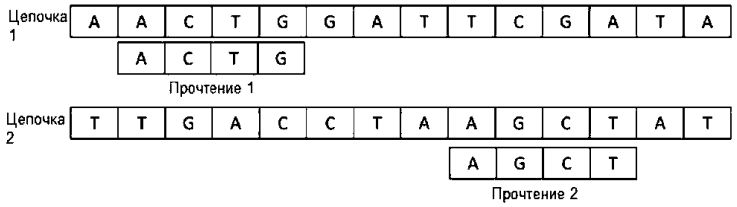
26. Геномный декодер (218) для декодирования геномных данных, выполненный с возможностью осуществления способа по любому из пп.19-24.

27. Машиночитаемый носитель информации, содержащий инструкции, исполнение которых вызывает осуществление по меньшей мере одним процессором способа кодирования по любому из пп.1-18.

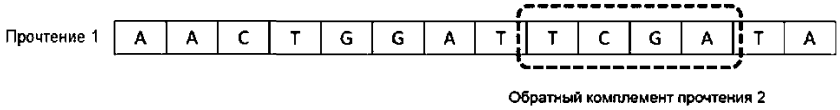
28. Машиночитаемый носитель информации, содержащий инструкции, исполнение которых вызывает осуществление по меньшей мере одним процессором способа декодирования по пп.19-24.



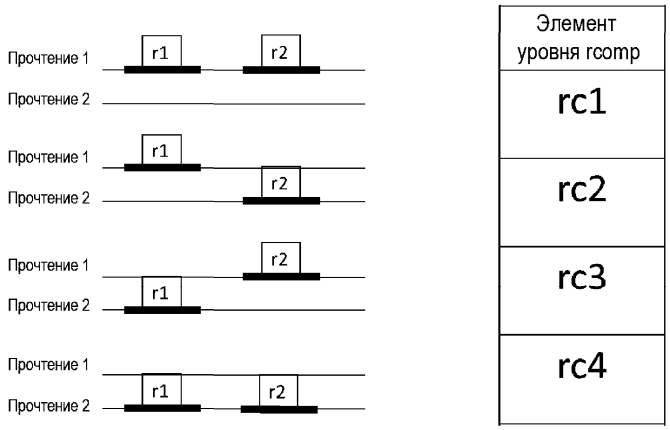
Фиг. 1



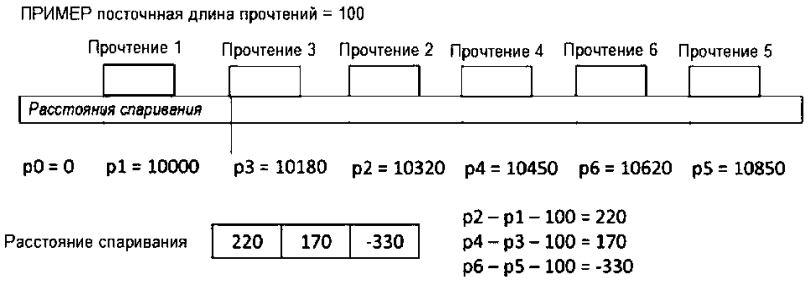
Фиг. 2



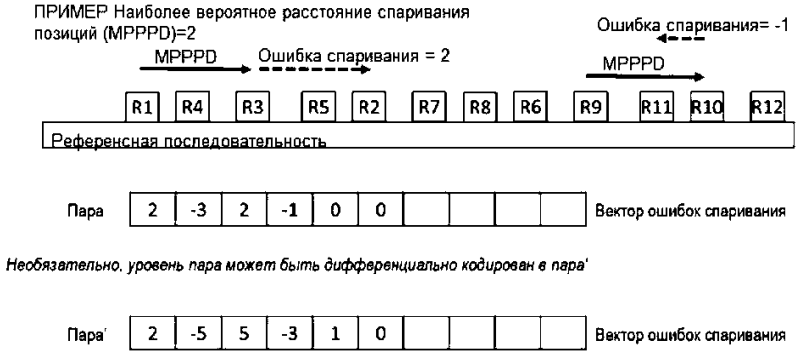
Фиг. 3



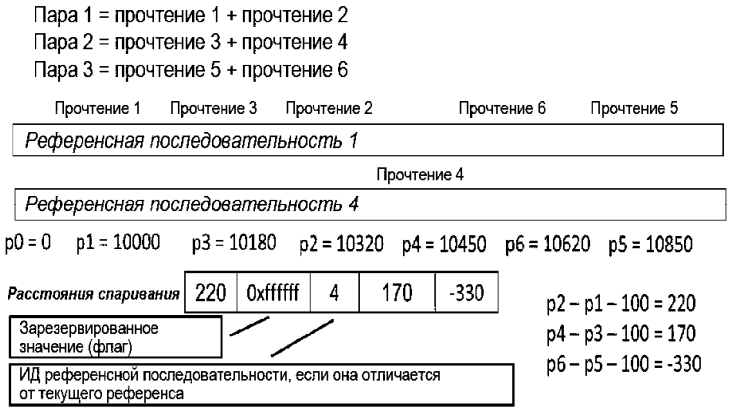
Фиг. 4



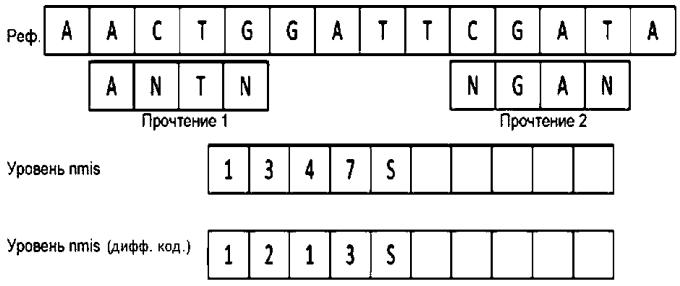
Фиг. 5



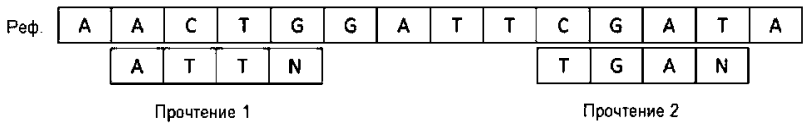
Фиг. 6



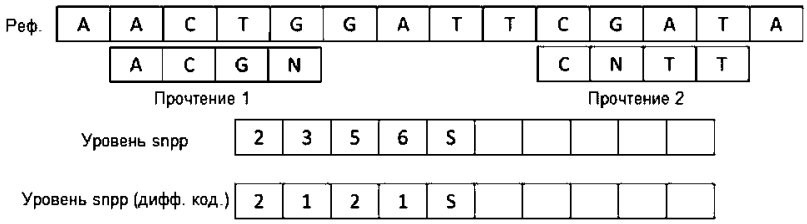
Фиг. 7



Фиг. 8



Фиг. 9



Фиг. 10

Уровень snpt (без кодов IUPAC)

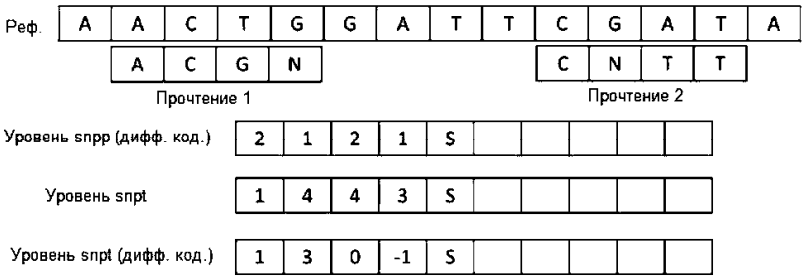
Тип замены вычисляют как индекс вектора замен, составленного из всех возможных символов.
Например:
S = [A, C, G, T, N, Z], где **Z = направление**
Индекс делеции
• КОДИРОВАНИЕ справа налево
• ДЕКОДИРОВАНИЕ слева направо

Реф.	Прочт.	Кодированный символ
A		idx(A,Z) = 5
C		idx(C,Z) = 4
G		idx(G,Z) = 3
T		idx(T,Z) = 2

Реф.	Прочт.	Кодированный 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

Реф.	Прочт.	Кодированный 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

Фиг. 11



Фиг. 12

Уровень snpt (без кодов IUPAC)

Тип замены вычисляют 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

Реф.	Прочт.	Кодированный 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

Реф.	Прочт.	Кодированный 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

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

Уровень indt (без кодов IUPAC)

S = [A, C, G, T, N, Z]

- КОДИРОВАНИЕ справа налево
- ДЕКОДИРОВАНИЕ слева направо

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

Фиг. 15

Уровень indt (с кодами IUPAC)

S = [A, C, G, T, N, Z, M,R,W,S,Y,K,V,H,D,B]

НАПРАВЛЕНИЕ

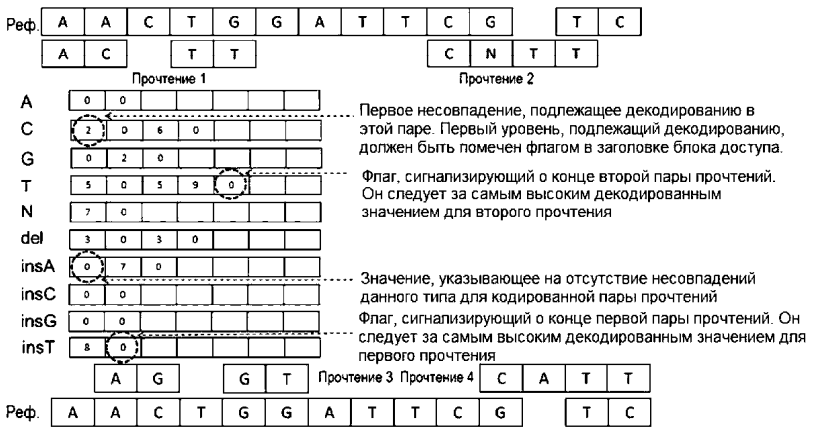
- КОДИРОВАНИЕ справа налево
- ДЕКОДИРОВАНИЕ слева направо

Инсерция	Кодированный символ
A	16
C	17
G	18
T	19
N	20

Фиг. 16

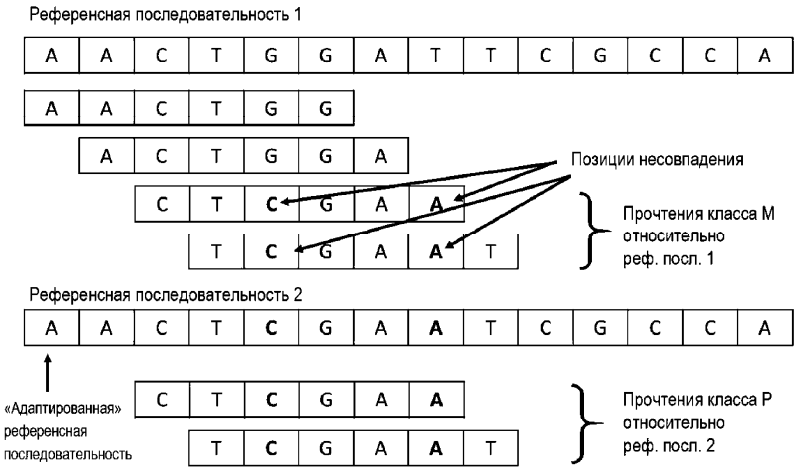
Модель источника 2 (без кодов IUPAC)

Один уровень позиции на тип замены, один на тип делеции и один на тип инсерции

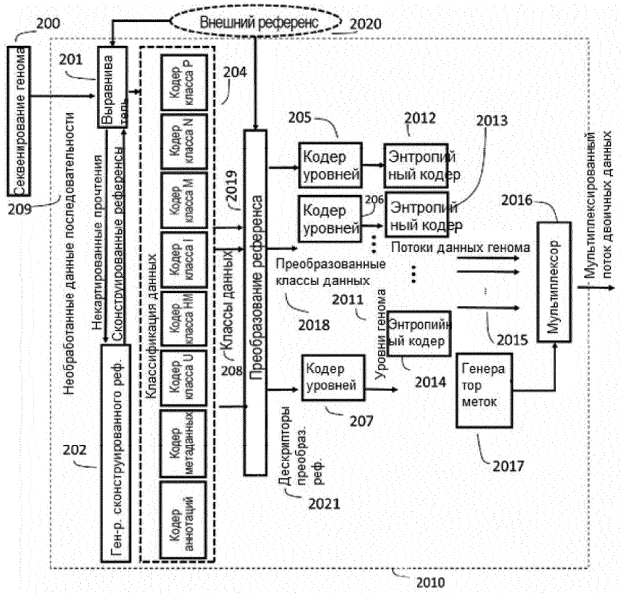


Фиг. 17

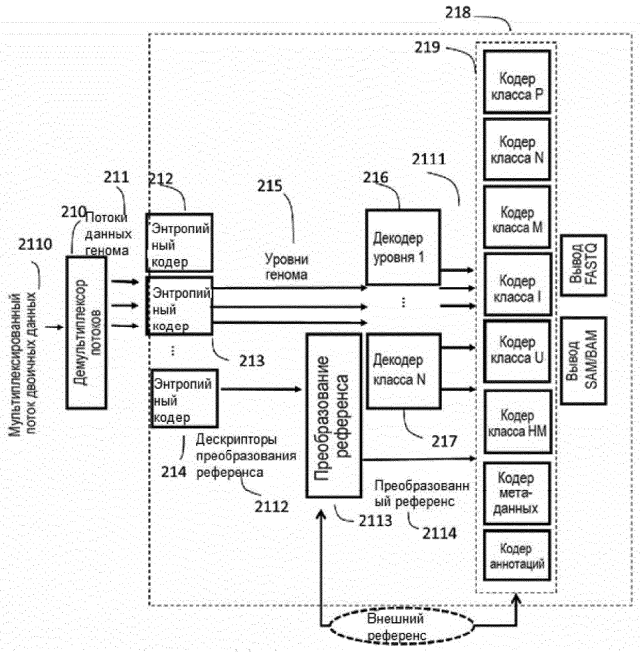
Фиг. 18



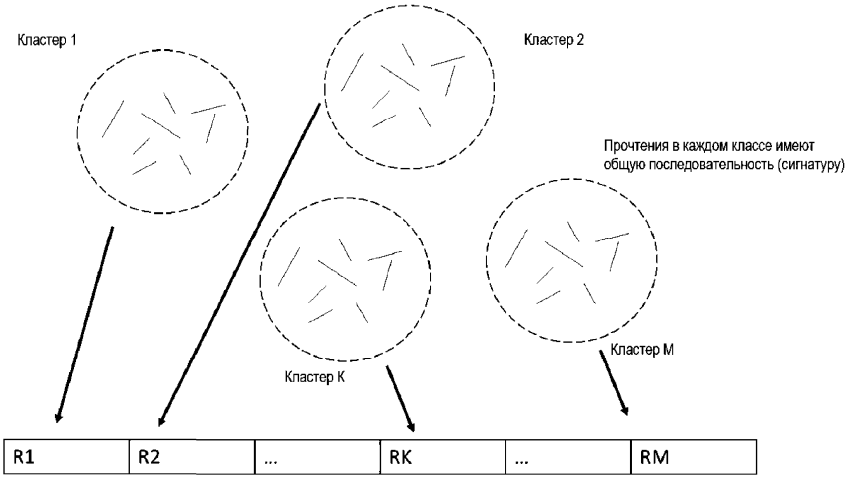
Фиг. 19



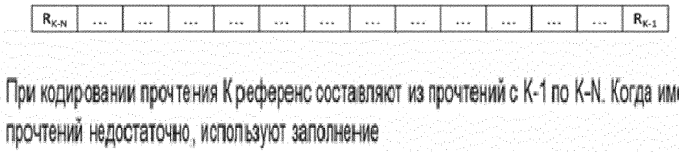
Фиг. 20



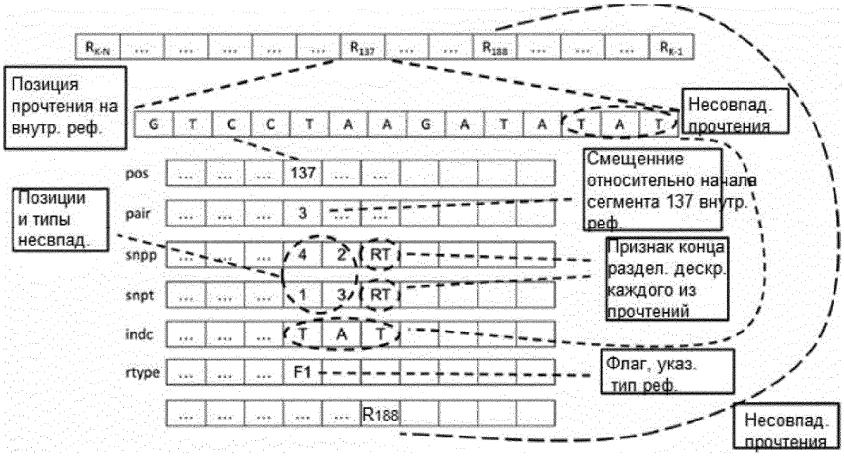
Фиг. 21



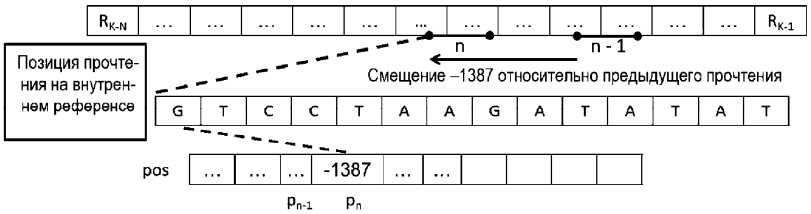
Фиг. 22



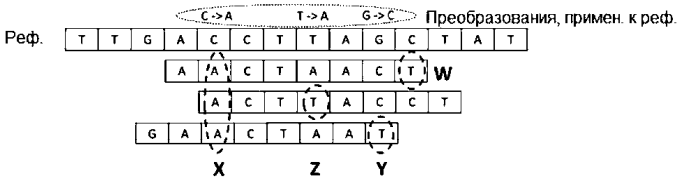
Фиг. 23



Фиг. 24

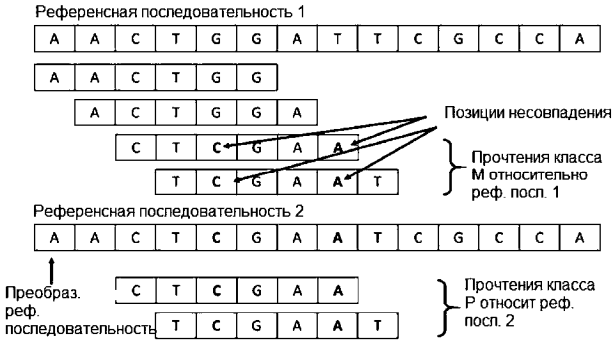


Фиг. 25

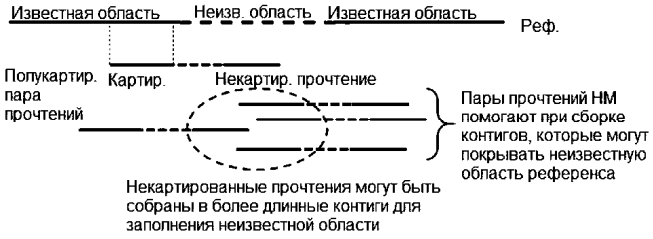


X = несовпадение, которое исчезнет
Y = несовпадение, которое не исчезнет, но код несовпадения изменится
W = несовпадение, которое остается неизменным
Z = заново введенное несовпадение после преобразования

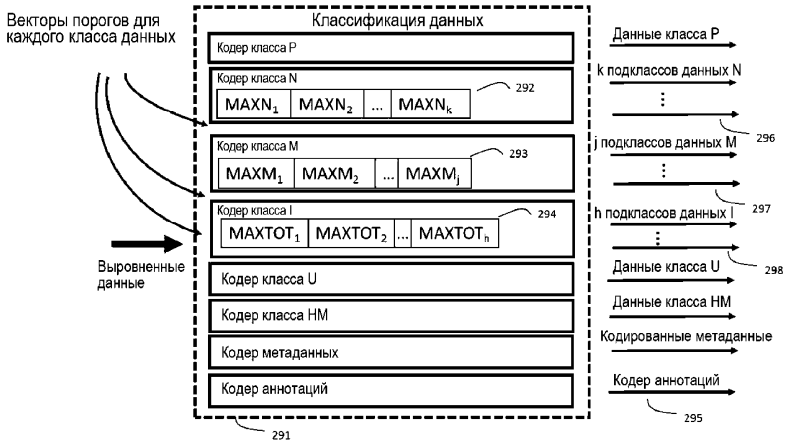
Фиг. 26



Фиг. 27

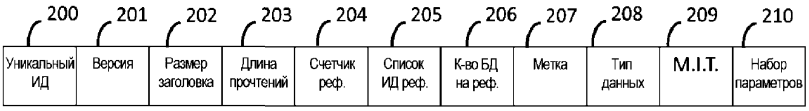


Фиг. 28



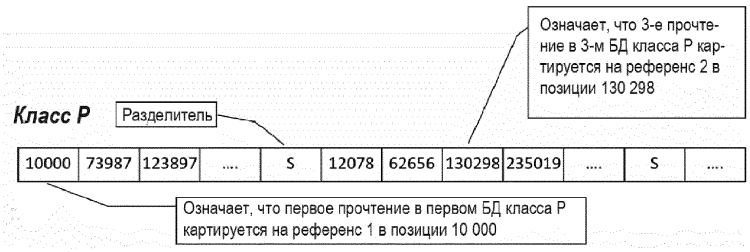
Фиг. 29

Фиг. 30

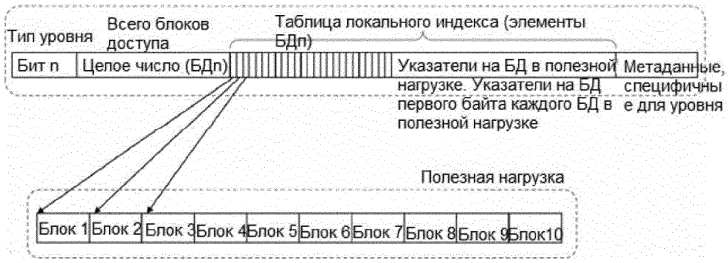


Фиг. 31

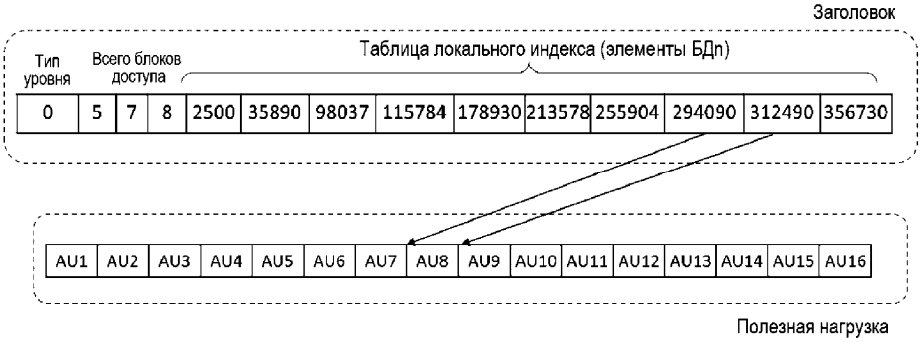
Фиг. 32



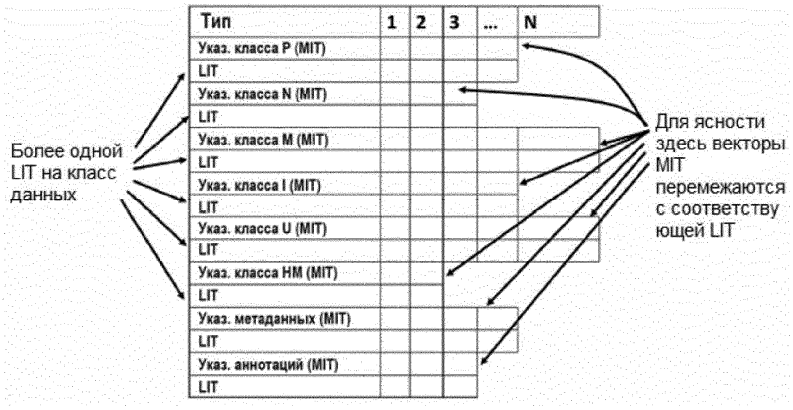
Фиг. 33



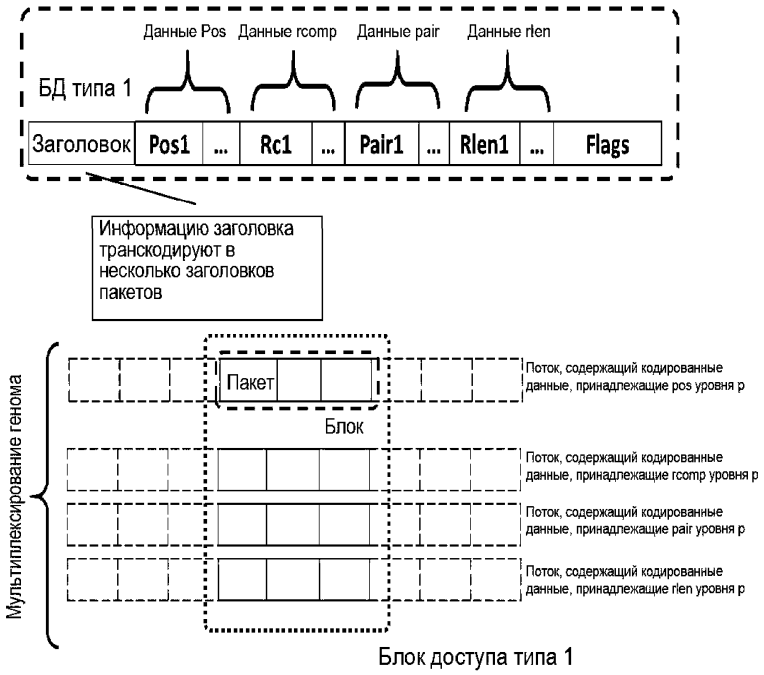
Фиг. 34



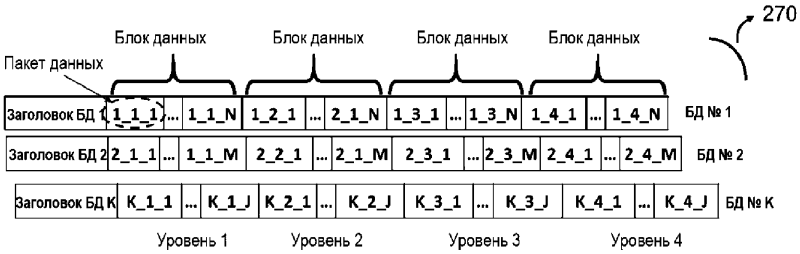
Фиг. 35



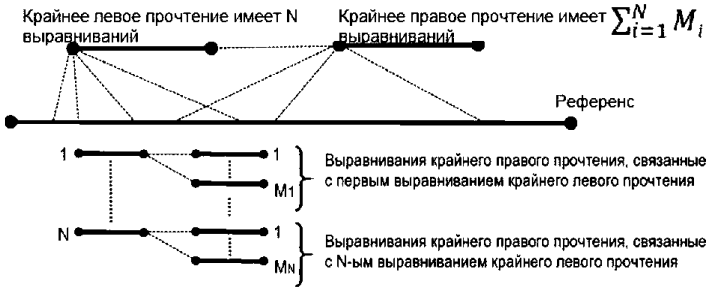
Фиг. 36



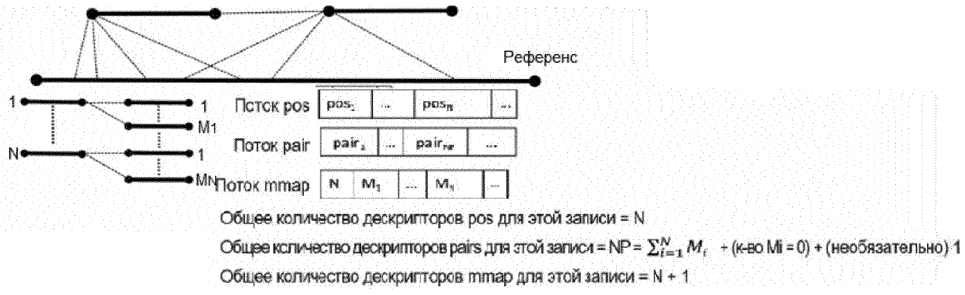
Фиг. 37



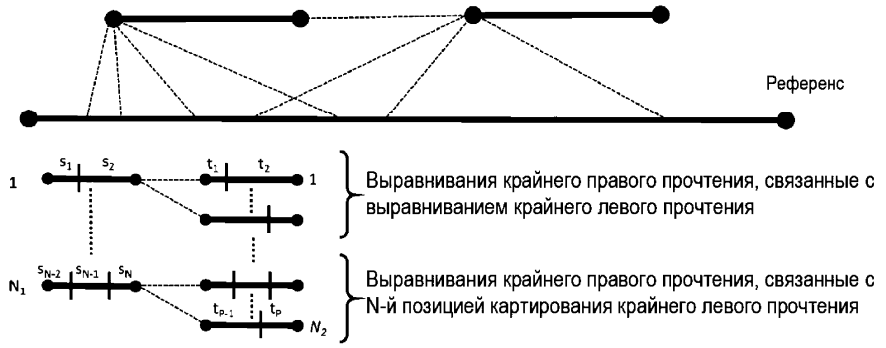
Фиг. 38



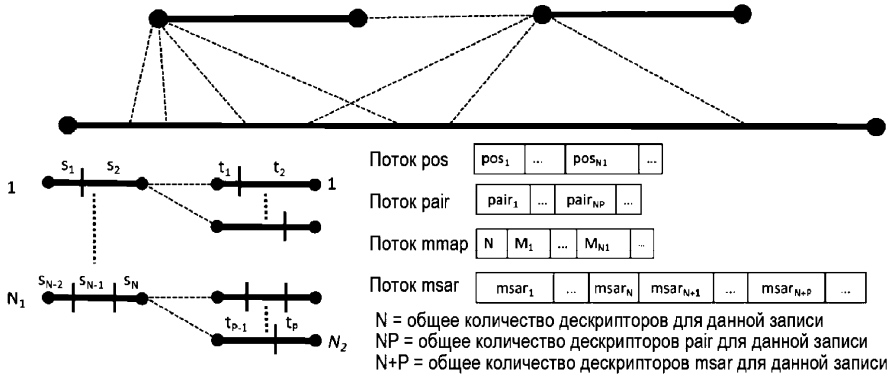
Фиг. 39



Фиг. 40


Референс
1
1
Выравнивания крайнего правого прочтения, связанные с выравниванием крайнего левого прочтения
N1
N2
Выравнивания крайнего правого прочтения, связанные с N-й позицией картирования крайнего левого прочтения

Фиг. 41


1
1
N1
N2
Поток pos
Поток pair
Поток mmap
Поток msar
N = общее количество дескрипторов для данной записи
NP = общее количество дескрипторов pair для данной записи
N+P = общее количество дескрипторов msar для данной записи

Фиг. 42