(19) Евразийское патентное ведомство

(11) 037106 (13) B1

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ЕВРАЗИЙСКОМУ ПАТЕНТУ

(45) Дата публикации и выдачи патента
2021.02.05

(21) Номер заявки
201990920

(22) Дата подачи заявки
2016.10.11

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

(54) СПОСОБ И СИСТЕМА ДЛЯ ЗАПОМИНАНИЯ БИОИНФОРМАЦИОННЫХ ДАННЫХ И ДОСТУПА К НИМ

(43) **2019.07.31**
(86) **PCT/EP2016/074301**
(87) **WO 2018/068828 2018.04.19**

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

(72) Изобретатель:
Рензи Даниэле, Зоиа Гиоргио (CH)

(74) Представитель:
Пронин В.О. (RU)

(56) The SAM/BAM Format Specification Working Group: "Sequence Alignment/Map Format Specification", http://samtools.github.io/hts-specs/SAMv1.pdf 18 November 2015 (2015-11-18), XP055387586, Retrieved from the Internet: URL: https://web.archive.org/web/20160505133503/http%3A//samtools.github.io/hts-specs/SAMv1.pdf [retrieved on 2017-07-04] pgs. 1, 4; sect. 1, 1.4

Anonymous: "CRAM format specification (version 3.0)", 8 September 2016 (2016-09-08), XP002771305, Retrieved from the Internet: URL: https://samtools.github.io/hts-specs/CRAMv3.pdf [retrieved on 2017-06-22] pgs. 1, 4-6, 7, 14; sect. 1, 5, 7, 10.3

(57) Способ и система для запоминания данных генома и доступа к ним. Результаты секвенирования генома разделяются на единицы доступа исходя из предсказуемости данных, содержащихся в результатах. Единицы доступа классифицируются с разделением по типам, и структурирование обеспечивает возможность избирательного доступа и избирательной обработки данных генома.

037106 B1

037106 B1

Область техники

В данной заявке предлагаются новые способы для эффективного запоминания биоинформационных данных, в частности данных секвенирования генома, доступа к этим данным и их передачи.

Уровень техники

Соответствующая форма представления данных секвенирования генома является основой, позволяющей эффективно обрабатывать, запоминать и передавать данные генома для обеспечения возможности и упрощения приложений анализа - таких как вызов вариантов генома и выполнение полного анализа с постановкой различных целей путем обработки данных секвенирования и метаданных. В настоящее время информация о секвенировании генома генерируется автоматами высокопроизводительного секвенирования (High Throughput Sequencing; (HTS)) в виде последовательностей нуклеотидов (именуемых также "основаниями"), отображаемых строками буквенных символов из определенного словаря.

Эти автоматы секвенирования не считывают геном или гены целиком, но они формируют короткие случайные фрагменты нуклеотидных последовательностей, именуемые считанными фрагментами последовательности.

Каждому нуклеотиду в считанном фрагменте последовательности назначается числовой показатель качества. Этот числовой показатель отображает уровень доверительной вероятности, назначенный автоматом считанному фрагменту специфического нуклеотида на специфической позиции в последовательности нуклеотидов.

Эти исходные данные секвенирования, сгенерированные автоматами NGS (секвенирования следующего поколения), обычно записывают на хранение в файлы FASTQ (см. также фиг. 1).

Словарь минимального объема, отображающий нуклеотидные последовательности, полученные при выполнении процесса секвенирования, состоит из пяти символов: {A, C, G, T, N} отображающих 4 типа нуклеотидов, входящих в состав ДНК (а именно, аденин, цитозин, гуанин и тимин) в сочетании с символом N, указывающим на то, что автомату секвенирования не удалось вызвать то или иное основание с достаточным уровнем доверительной вероятности, что тип основания на такой позиции остается неопределенным в процессе считывания. В рибонуклеиновой кислоте (РНК) тимин заменен урасилом (U). Последовательности нуклеотидов, полученные с помощью автоматов секвенирования, именуют "считанными фрагментами". В случае спаренных считанных фрагментов используют термин "шаблон" для обозначения исходной последовательности, из которой была выделена пара считанных фрагментов. Считанные фрагменты последовательности могут состоять из количества нуклеотидов в диапазоне от нескольких дюжин до нескольких тысяч. В некоторых технологиях считанные фрагменты последовательности считываются в виде пар, где каждый считанный фрагмент может быть получен от одной из двух цепей ДНК.

В области секвенирования генома термин "покрытие" используют для обозначения уровня избыточности данных секвенирования по отношению к референсному геному. Например, для обеспечения покрытия 30× на геноме человека (длиной 3,2 млрд нуклеотидных последовательностей, или оснований) автомат секвенирования должен сформировать в сумме приблизительно 30×3,2 млрд оснований таким образом, что в среднем каждая позиция в эталоне покрывается 30 раз.

Решения в технике известного уровня

Большинство используемых форм представления информации о геноме из данных секвенирования основано на форматах файлов FASTQ и SAM, которые обычно представляют в форме с zip-сжатием для сокращения объема исходного файла. Файлы традиционного формата, соответственно FASTQ и SAM для выровненных и не выровненных данных секвенирования, состоят из символов открытого текста, и поэтому для их сжатия используют универсальные средства - такие как LZ-схемы (хорошо известные zip, gzip и др.), авторами которых являются Lempel и Ziv, опубликовавшие первые версии. Когда используются универсальные компрессоры, такие как gzip, результатом сжатия является обычно одиночный большеобъемный объект двоичных данных. Информационное содержимое таких представленных в монолитной форме результатов очень трудно архивировать, пересылать и обрабатывать, особенно в случае секвенирования с высокой производительностью, когда объемы данных чрезвычайно велики. По окончании секвенирования выполняется поточная обработка информации о геноме, на каждой ступени которой формируются данные, представленные в виде совершенно новой структуры данных (в формате файла), несмотря на то, что, по существу, лишь малая доля полученных данных является новой по отношению к предыдущей ступени.

На фиг. 1 показаны основные ступени типовой поточной обработки информации о геноме с обозначением соответствующей формы представления формата файла.

Общепринятым решениям свойственен ряд недостатков: записанные в архив данные недостаточны в связи с тем, что на каждой ступени поточной обработки информации о геноме используется отличный от других формат файла, что подразумевает множественное копирование данных с сопутствующим быстрым ростом потребности в объеме памяти. Это неэффективно само по себе, поскольку не является логически необходимым, но это неэффективно, нецелесообразно и к тому же не выдерживает нагрузки при увеличении объема данных, генерируемых автоматами высокопроизводительного секвенирования (HTS).

Это является последствиями, зависимыми от доступного объема памяти и от совокупных затрат, и также препятствует предоставлению преимуществ от анализа генома в здравоохранении увеличенной доле населения. Влияние затрат на ИТ, увеличивающихся при экспоненциальном росте объема результатов секвенирования, подлежащих записи в память и анализу, является одним из главных вызовов научному сообществу и тем, с чем вынуждена иметь дело индустрия здравоохранения (см. Скотт Д. Кан "О будущем данных генома", Science 331, 728 (2011) и Д.С. Павлишин, Т. Вайссман и Г. Айона, 2013, "Снова контракты на геном человека", Bioinformatics 29(17): 2199-2202). В это же время был предпринят ряд инициатив с попытками увеличить масштаб распространения секвенирования генома от нескольких выбранных индивидуумов до больших групп населения (См. Джош П. Робертс "Секвенирование для миллиона ветеранов", Nature Biotechnology 31, 470 (2013)).

Пересылка данных генома является медленным и неэффективным процессом, поскольку используемые в настоящее время форматы данных организованы в виде монолитных файлов объемом до нескольких сотен Гбайт, которые необходимо в полном объеме пересылать для обработки на приемный конец. Это означает, что для возможности анализа небольшого сегмента данных требуется пересылка всего файла с существенными затратами, связанными с занятием полосы пропускания и временем ожидания. Часто передача в реальном времени запрещена для больших объемов пересылаемых данных, и транспортировку данных осуществляют путем физического перемещения носителей данных, таких как накопители на жестком диске или серверы хранения данных, из одного места в другое.

Эти ограничения, действующие при использовании подходов известного уровня техники, преодолеваются настоящим изобретением.

Обработка данных является медленной и неэффективной вследствие того факта, что информация не структурирована таким способом, при котором части данных различных категорий и метаданных, требуемые общепринятым приложениям анализа, могут быть считаны без необходимости доступа к данным во всей их совокупности. Данный факт означает, что может потребоваться работа обычных конвейеров поточного анализа в течение нескольких дней или недель с расходованием ценных и дорогостоящих ресурсов обработки по причине необходимости анализировать по синтаксису и фильтровать большие объемы данных даже в том случае, если составные элементы данных, относящиеся к конкретной цели анализа, имеют намного меньший объем.

Эти ограничения препятствуют своевременному получению профессиональными медиками отчетов с результатами анализа генома и оперативному принятию мер по лечению болезней. Настоящее изобретение предоставляет решение в отношении данной потребности.

Имеется еще одно техническое ограничение, которое также преодолевается настоящим изобретением, По существу, целью изобретения является предоставление соответствующей формы представления результатов секвенирования генома и метаданных путем организации и разделения данных таким образом, что сжатие данных и метаданных доводится до максимума и эффективно вводятся в действие несколько функционалов, таких как функционал избирательного доступа и поддержки для инкрементных обновлений и многих других.

Ключевым аспектом изобретения является специфическое определение классов данных и метаданных, которые должны быть представлены в соответствующей исходной модели, по отдельности закодированы (т.е. сжаты) в процессе структурирования на специфических уровнях. Наиболее важными преимуществами настоящего изобретения в сравнении со способами из техники известного уровня являются

повышение эффективности сжатия вследствие снижения энтропии источника информации, обеспечиваемое путем предоставления эффективной модели для каждого класса данных или метаданных;

возможность осуществления избирательного доступа к частям сжатых данных и метаданных с целью любой дальнейшей обработки;

возможность инкрементного обновления (без необходимости повторного кодирования) и добавления закодированных данных и метаданных с данными нового секвенирования и/или метаданными и/или результатами нового анализа;

возможность эффективной обработки данных сразу после их выдачи автоматом секвенирования или инструментами выравнивания без необходимости дожидаться окончания процесса секвенирования или выравнивания.

В настоящей заявке раскрыты способ и система, направленные на решение проблемы эффективного манипулирования, хранения и передачи очень больших объемов данных секвенирования генома путем использования подхода со структурированными единицами доступа.

Настоящая заявка снимает все ограничения подходов из техники известного уровня, связанные с функционалом доступности данных генома, эффективной обработкой поднаборов данных, передачей и потоковой передачей функционала в сочетании с эффективным сжатием.

В настоящее время наиболее распространенным форматом для данных генома является текстовый формат "подстановка выравнивания последовательности" (Sequence Alignment Mapping; SAM) и его двоичный аналог BAM. SAM-файлы представляют собой читаемые человеком текстовые ASCII-файлы, в то время как для BAM принят вариант gzip на блочной основе. BAM-файлы могут быть проиндексированы для обеспечения ограниченной модальности произвольного доступа. Это поддерживается путем созда-

ния отдельного индексного файла. Формат BAM характеризуется низким качеством сжатия по следующим причинам.

1. Он полностью нацелен на сжатие неэффективного и избыточного файла формата SAM вместо извлечения актуальной информации о геноме, пересылаемой посредством файлов SAM, и использования соответствующих моделей для ее сжатия.

2. В нем используется универсальный алгоритм сжатия текста (например, gzip) вместо эксплуатирования специфической природы каждого источника данных (самой информации о геноме).

3. В нем отсутствует какая-либо концепция, связанная с классификацией данных, которая позволила бы реализовать избирательный доступ к специфическим классам данных генома.

Более изощренным подходом к сжатию данных генома получившим меньшее распространение, но более эффективным по сравнению с BAM является CRAM (спецификация CRAM: https://samtools.github.io/hts-specs/CRAMv3.pdf). CRAM обеспечивает более эффективное сжатие благодаря принятию дифференциального кодирования в отношении существующих референсных данных (он частично эксплуатирует избыточность источника данных), однако ему по-прежнему не хватает таких характеристик, как инкрементные обновления, поддержка поточной обработки и избирательный доступ к специфическим классам сжатых данных.

CRAM опирается на концепцию записи CRAM. В каждой записи CRAM кодируется одиночный встроенный подстановкой или невстроенный считанный фрагмент путем кодирования всех элементов, необходимых для его воссоздания.

Основными отличиями раскрытого здесь изобретения от подхода на основе CRAM являются

1. Индексация данных применительно к CRAM выходит за рамки данной спецификации (см. раздел 12 Спецификации CRAM v3.0) и реализуется в виде отдельного файла. В настоящем изобретении индексация данных интегрирована с процессом кодирования и индексы внедряются в закодированный поток данных.

2. В CRAM все блоки основных данных могут содержать выровненные подстановкой считанные фрагменты любого типа (идеально согласующиеся считанные фрагменты, считанные фрагменты только с заменами, считанные фрагменты со вставками-удалениями). В настоящем изобретении отсутствует принцип классификации и объединения в группы по классам в соответствии с результатом выравнивания подстановкой по отношению к референсной последовательности.

3. В описанном изобретении отсутствует концепция записи с инкапсуляцией каждого считанного фрагмента, поскольку данные, требуемые для воссоздания каждого считанного фрагмента, рассредоточены между несколькими контейнерами данных, именуемых "уровнями". Это обеспечивает повышенную эффективность доступа к набору считанных фрагментов со специфическими биологическими характеристиками (например, считанные фрагменты с заменами, но без вставок-удалений или идеально выровненные подстановкой считанные фрагменты) без необходимости декодирования каждого считанного фрагмента (блока считанных фрагментов) для контроля его (их) характеристик.

4. В записи CRAM каждый тип данных обозначается специфическим флажком. В отличие от CRAM в настоящем изобретении отсутствует принцип обозначения данных флажком, поскольку они изнутри определены "уровнем", которому принадлежат данные. Эт означает существенное сокращение количества обязательных для использования символов и вытекающее из этого снижение энтропии источника информации, что приводит к повышению эффективности сжатия. Это связано с тем, что при использовании различных "уровней" кодеру предоставляется возможность повторяющегося использования каждого символа в пределах каждого уровня с различными смысловыми значениями. В CRAM каждый флажок должен постоянно иметь одно и то же смысловое значение, поскольку отсутствует понятие контекстов и каждая запись CRAM может содержать данные любого типа.

5. В CRAM замены, вставки и удаления выражаются по правилам различных синтаксисов, в то время как в настоящем изобретении используется единый алфавит и единое кодирование для замен, вставок и удалений. Благодаря этому упрощается процесс кодирования и декодирования и формируется модель источника с пониженной энтропией, при кодировании которого выдаются битовые потоки, характеризующиеся повышенной эффективностью сжатия.

Алгоритмы сжатия данных генома, используемые в технике известного уровня, можно подразделить на следующие категории:

с преобразованием сигналов

на LZ-основе;

изменение очередности считанных фрагментов;

на основе ассемблирования;

статистическое моделирование.

Двум первым категориям свойственен общий недостаток отсутствия эксплуатирования специфических характеристик источника данных (считанных фрагментов последовательности генома) и обработки данных генома в виде текстовой строки, подлежащей сжатию, без принятия во внимание специфических свойств информации такого вида (например, избыточности среди считанных фрагментов, ссылки на существующую пробу). В двух наиболее современных наборах инструментов для сжатия данных генома,

каковыми являются CRAM и Goby ("Сжатие структурированных данных секвенирования с высокой производительностью", Ф. Кампань, К.К. Дорфф, Н. Чамбве, Д.Т. Робинсон, Д.П. Мезиров, Т.Д. Ву), недостаточно используется арифметическое кодирование, поскольку они в неявном виде моделируют данные в виде независимых данных, равномерно распределяемых методом геометрического распределения. Goby является чуть более изощренным инструментом, поскольку он преобразует все поля в список целых значений, после чего каждый список кодируется независимым образом с привлечением арифметического кодирования без использования контекста. В наиболее эффективном режиме работы Goby способен осуществлять некоторое межсписочное моделирование в пределах списков целых чисел для повышения качества сжатия. Решения из предыдущего уровня техники обеспечивают низкие коэффициенты сжатия и такие структуры данных, которые после сжатия трудно доступны (если вообще доступны) для избирательного доступа и манипулирования. Ступени анализа в нисходящем направлении потока могут в итоге оказаться неэффективными и очень низкоскоростными вследствие необходимости обрабатывать большие по объему и жестко заданные структуры данных даже при выполнении простых операций или при доступе в выбранные области набора данных генома.

Соотношение между форматами файлов, используемыми в конвейерах поточной обработки генома, в упрощенном виде представлено на фиг. 1. На данной диаграмме включение файла не подразумевает наличие гнездовой структуры файлов, но лишь отображает тип и объем информации, которая может быть закодирована для каждого формата (т.е. SAM содержит всю информацию в FASTQ, но с организацией информации в виде файла другой структуры). CRAM содержит такую же информацию по геному, что и SAM/BAM, но обладает более высокой гибкостью в отношении типа сжатия, который может быть использован, поэтому он представлен в виде расширенной версии SAM/BAM.

Использование множества форматов файла для запоминания информации по геному является чрезвычайно неэффективным и дорогостоящим. Использование разных форматов файла на различных ступенях цикла существования информации по геному подразумевает линейное нарастание занимаемой области памяти даже в случае минимума инкрементной информации. Далее перечислены другие недостатки техники известного уровня.

1. Доступ, анализ или добавление аннотаций (метаданных) в отношении исходных данных, записанных в виде сжатых файлов FastQ или любой их комбинаций, требует декомпрессии и рекомпрессии всего файла с большим расходованием вычислительных средств и времени.

2. Для получения специфических поднаборов информации, такой как позиция подстановки считанного фрагмента, тип и позиция варианта считанного фрагмента, типы и позиция вставки-удаления, или любых других метаданных и аннотаций, содержащихся в выровненных данных, записанных на хранение в файлы BAM, требуется доступ ко всему объему данных, соответствующему каждому считанному фрагменту. Решения из техники известного уровня не обеспечивают избирательный доступ к одному классу метаданных.

3. При работе с форматами файлов техники известного уровня требуется получение конечным пользователем всего файла прежде, чем может быть начата обработка. Например, при опоре на соответствующее представление данных можно было бы начать выравнивание считанных фрагментов прежде, чем завершится процесс секвенирования. Секвенирование, выравнивание и анализ могли бы протекать и продвигаться параллельно друг другу.

4. Решения из техники известного уровня не поддерживают структурирование и не способны различать данные генома, полученные с помощью различных процессов секвенирования согласно их специфической семантике генерации (например, при получении результатов секвенирования в разные периоды жизни одного и того же индивидуума). Такое же ограничение действует в отношении секвенирования, осуществляемого по различным типам биологических проб одного и того же индивидуума.

5. Шифрование данных целиком или в выбранной части не поддерживается решениями из техники известного уровня.

Примеры шифрования:

a) выбранные области ДНК,

b) только последовательности, содержащие варианты,

c) только химерные последовательности,

d) только последовательности без подстановок,

e) специфические метаданные (например, источник секвенированной пробы, идентичность секвенированного индивидуума, тип пробы).

6. Для транскодирования данных секвенирования, выровненных по заданным референсным данным (т.е. файла SAM/BAM), относительно новых референсных данных требуется обработка всего объема данных даже в случае, когда новые референсные данные отличаются от предыдущих референсных данных лишь позицией одного нуклеотида.

Поэтому существует потребность в соответствующем уровне запоминания информации по геному (формате файла генома), который обеспечивает эффективное сжатие, поддерживает избирательный доступ в сжатую область, поддерживает инкрементное добавление разнородных метаданных в сжатую область на всех уровнях различных ступеней обработки данных генома.

Настоящее изобретение предоставляет решение в отношении ограничений техники известного уровня путем использования способа, устройств и компьютерных программ, заявленных в сопроводительном наборе пунктов формулы изобретения.

Перечень фигур

На фиг. 1 изображены основные шаги типового конвейера генома и связанные с этим форматы файла;

фиг. 2 иллюстрирует взаимную связь между большинством используемых форматов файла генома;

фиг. 3 - компоновку считанных фрагментов последовательности генома в целом геноме или его части посредством компоновки заново или выравнивания на основе референсной последовательности;

фиг. 4 - вычисление позиций подстановки считанных фрагментов в референсную последовательность;

фиг. 5 - вычисление расстояний образования пар считанных фрагментов;

фиг. 6 - вычисление ошибок образования пар;

фиг. 7 - кодирование расстояния образования пар, когда партнерская пара считанного фрагмента подставлена в другую хромосому;

фиг. 8 показано, как считанные фрагменты последовательности могут поступать из первой или второй ДНК-цепи;

фиг. 9 - взаимосвязь между считанным фрагментом, подставленным в цепь 2, и соответствующим считанным фрагментом обратного замещения в цепи 1;

на фиг. 10 - четыре возможные комбинации считанных фрагментов, составляющих пару считанных фрагментов, и соответствующее кодирование в уровне обратного замещения (rcomp);

фиг. 11 - способ кодирования N рассогласований в уровне nmis;

фиг. 12 - пример замен в выровненной подстановкой паре считанных фрагментов;

фиг. 13 - способ вычисления позиций замен в виде либо абсолютных, либо дифференциальных значений;

фиг. 14 - способ вычисления замен при кодировании символов без использования кодов IUPAC;

фиг. 15 - кодирование типов замен в уровне snpt;

фиг. 16 - способ вычисления замен при кодировании символов с использованием кодов IUPAC;

фиг. 17 - альтернативная модель источника для замен, когда кодируются только позиции, но используется один уровень на тип замены;

фиг. 18 - способ кодирования замен, вставок и удалений в паре считанных фрагментов класса I, когда не используются коды IUPAC;

фиг. 19 - способ кодирования замен, вставок и удалений в паре считанных фрагментов класса I при использовании кодов IUPAC;

фиг. 20 - структура заголовка структуры данных с информацией по геному;

фиг. 21 - обозначение в таблице главного указателя позиций в референсных последовательностях первого считанного фрагмента в каждой единице доступа;

фиг. 22 - пример частичной таблицы главного указателя (MIT) с отображением позиций подстановки первого считанного фрагмента на каждой позиции единицы доступа класса P;

фиг. 23 - способ представления таблицы местного указателя в заголовке уровня в виде вектора указателей к единицам доступа в полезной нагрузке;

фиг. 24 - пример таблицы местного указателя;

фиг. 25 - функциональную взаимосвязь между таблицей главного указателя и таблицей местного указателя;

фиг. 26 - способ составления единиц доступа из блоков данных, принадлежащих нескольким уровням. Уровни составляются блоками, подразделяемыми на пакеты;

фиг. 27 - способ преобразования в форму пакета и инкапсулирования в уплотненную структуру данных генома для геномной единицы доступа типа 1 (содержащей информацию о позициях, образовании пар, обратном замещении и длине считанного фрагмента);

фиг. 28 - составление единиц доступа из заголовков и уплотненных блоков, принадлежащих одному или более уровням однородных данных. Каждый блок может состоять из одного или более пакетов, содержащих действующие дескрипторы информации о геноме;

фиг. 29 - структуру единиц доступа типа 0, которым не требуется ссылка на любую информацию, поступающую от других единиц доступа, предназначенных для доступа или декодирования и доступа;

фиг. 30 - структуру единиц доступа типа 1;

фиг. 31 - структуру единиц доступа типа 2, содержащих данные, соотносимые с единицей доступа типа 1. Имеются позиции N в закодированных считанных фрагментах;

фиг. 32 - структуру единиц доступа типа 3, содержащих данные, соотносимые с единицей доступа типа 1. В закодированных считанных фрагментах указываются позиции и типы рассогласования;

фиг. 33 - структуру единиц доступа типа 4, содержащих данные, соотносимые с единицей доступа типа 1. В закодированных считанных фрагментах указываются позиции и типы рассогласования;

фиг. 34 - первые пять типов единиц доступа;

фиг. 35 - соотношение между единицами доступа типа 1 и требующими декодирования единицами доступа типа 0;

фиг. 36 - соотношение между единицами доступа типа 2 и требующими декодирования единицами доступа типов 0 и 1;

фиг. 37 - соотношение между единицами доступа типа 3 и требующими декодирования единицами доступа типов 0 и 1;

фиг. 38 - соотношение между единицами доступа типа 4 и требующими декодирования единицами доступа типов 0 и 1;

фиг. 39 - единицы доступа, требуемые для декодирования считанных фрагментов последовательности с рассогласованиями, подставленных во второй сегмент референсной последовательности (AU 0-2);

фиг. 40 - способ инкрементного добавления исходных данных геномной последовательности, к которым открыт доступ, к предварительно закодированным данным генома;

фиг. 41 - использование структуры данных на основе единиц доступа для обеспечения пуска анализа данных генома ранее завершения процесса секвенирования;

фиг. 42 - косвенный перенос считанных фрагментов из AU типа 4 в AU типа 3, который может быть вызван новым анализом, выполненным на существующих данных;

фиг. 43 - инкапсулирование заново полученных результатов анализа в новую единицу доступа типа 6 и создание соответствующего индекса в таблице главного указателя;

фиг. 44 - способ транскодирования данных вследствие публикации новой референсной последовательности (генома);

фиг. 45 - перемещение считанных фрагментов, подставленных в новую область генома с повышенным качеством (например, без вставок-удалений) из единицы доступа типа 4 в единицу доступа типа 3;

фиг. 46 показывает, как в случае обнаружения нового места подстановки (например, с меньшим количеством несовпадений) соответствующие считанные фрагменты могут быть перенесены из одной единицы доступа в другую того же типа;

фиг. 47 иллюстрирует возможность применения избирательного шифрования к единицам доступа типа 4 только при наличии в них чувствительной информации, требующей защиты;

фиг. 48 -способ обработки исходных данных последовательности генома или выровненных геномических данных для их инкапсулирования в геномическом уплотненном блоке Для подготовки данных к кодированию могут понадобиться ступени выравнивания, повторного выравнивания, компоновки. Сгенерированные уровни инкапсулируются в единицы доступа и уплотняются посредством мультиплексора генома;

фиг. 49 - извлечение демультиплексором (501) генома содержимого уровней единиц доступа из уплотненной структуры генома; декодер (по одному декодеру на тип единицы доступа (502)) извлекает дескрипторы генома, которые далее декодируются в различные форматы генома, такие как FASTQ и SAM/BAM.

Подробное описание

В настоящем изобретении описываются формат файла мультиплексирования и соответствующие единицы доступа для использования при хранении, транспортировке, доступе и обработке геномической или протеомической информации в форме последовательностей символов, отображающих молекулы.

Например, эти молекулы включают в себя нуклеотиды, аминокислоты и протеин. Одним из наиболее важных элементов информации, представленных в виде последовательности символов, являются данные, сгенерированные устройствами высокопроизводительного секвенирования генома.

Геном любого живого организма обычно представляют в виде цепи символов, отображающей цепочку нуклеиновых кислот (оснований), характеризующую данный организм. Текущее состояние технологии секвенирования генома в технике известного уровня обеспечивает получение лишь фрагментарного представления генома в виде множества (до миллиардов) цепочек нуклеиновых кислот, связанных с метаданными (идентификаторами, уровнем точности и др.). Такие цепочки именуют обычно "считанными фрагментами последовательности" или "считанными фрагментами".

Типовые шаги жизненного цикла информации по геному включают в себя извлечение считанных фрагментов последовательности, подстановки и выравнивание, обнаружение варианта, аннотирование варианта и функциональный и структурный анализ (см. фиг. 1).

Извлечение считанных фрагментов последовательности представляет собой выполняемый либо человеком (оператором), либо машиной процесс представления фрагментов генетической информации в виде последовательностей символов, отображающих молекулы, составляющие биологическую пробу.

В случае нуклеиновых кислот такие молекулы именуют "нуклеотидами". Последовательности символов, полученные в процессе извлечения, обычно именуют "считанными фрагментами". В технике известного уровня данную информацию кодируют обычно в виде файлов FASTA, состоящих из текстового заголовка и последовательности символов, отображающей секвенированные молекулы.

Когда биологическая проба секвенирована для извлечения ДНК живого организма, алфавит состоит из символов (A, C, G, T, N).

Когда биологическая проба секвенирована для извлечения РНК живого организма, алфавит состоит

из символов (A, C, G, U, N).

В случае расширенного набора символов IUPAC машина секвенирования выдает также так называемые "коды неоднозначности", и алфавит, используемый в качестве символов, составляющих считанные фрагменты, включает в себя (A, C, G, T, U, W, S, M, K, R, Y, B, D, H, V, N или …).

Если коды неоднозначности IUPAC не используются, каждому считанному фрагменту последовательности может быть назначена последовательность оценочного показателя качества. В таких случаях техника известного уровня принимает решение о кодировании полученной информации в виде файла FASTQ.

Устройства секвенирования могут вносить в считанные фрагменты последовательности ошибки следующих типов:

1) идентификация ошибочного символа (т.е. представление другой нуклеиновой кислоты) для представления нуклеиновой кислоты, по существу, имеющейся в секвенированной пробе; это именуется обычно "ошибкой замены" (рассогласованием);

2) вставка в считанную выборку последовательности дополнительных символов, которые не соотносятся, по существу, с имеющейся нуклеиновой кислотой; это именуется обычно "ошибкой вставки";

3) удаление из считанной выборки последовательности символов, отображающих нуклеиновые кислоты, по существу, имеющиеся в секвенированной пробе; это именуется обычно "ошибкой удаления";

4) повторное объединение одного или более фрагментов в одиночный фрагмент, не отражающий сущность оригинальной последовательности.

Термин "покрытие" используется в литературе для количественной оценки степени, в которой референсный геном или его часть могут быть охвачены имеющимися считанными фрагментами последовательности. Покрытие по определению бывает

частичным (меньшим 1X), когда некоторые части референсного генома не охватываются никаким имеющимся считанным фрагментом последовательности;

одиночным (1X), когда все нуклеотиды референсного генома охватываются одним и только одним символом, имеющимся в считанных фрагментах последовательности;

множественным (2X, 3X, NX), когда каждый нуклеотид референсного генома охватывается множество раз.

Выравниванием последовательности считается процесс упорядочения считанных фрагментов последовательности путем обнаружения областей подобия, которое может представлять собой сочетание функциональных, структурных или эволюционных соотношений между последовательностями. Если выравнивание выполняют применительно к предварительно заданной последовательности нуклеотидов, именуемой "референсным геномом", процесс именуется "подстановкой". Выравнивание последовательности можно выполнять также при отсутствии предварительно заданной последовательности (т.е. референсного генома), и в таких случаях процесс именуется в технике известного уровня выравниванием "заново". В технике известного уровня данную информацию записывают на хранение в файлы SAM, BAM или CRAM. Концепция выравнивания последовательностей для воссоздания частичного или полного генома представлена на фиг. 3.

Обнаружение варианта (именуемое также вызовом варианта) представляет собой процесс трансляции выровненных выходных данных машин секвенирования генома (считанных фрагментов последовательности, сгенерированных устройствами NGS и выровненными) в резюме уникальных характеристик секвенированного организма, которые не могут быть обнаружены в другой предварительно существующей последовательности или могут быть найдены лишь в некоторых предварительно существующих последовательностях). Эти характеристики именуются "вариантами", поскольку они выражаются в виде различий между геномом исследуемого организма и референсным геномом. В технике известного уровня данную информацию записывают в файл специфического формата, именуемый файлом VCF.

Аннотирование варианта представляет собой процесс назначения функциональной информации вариантам генома, идентифицируемым в ходе процесса вызова варианта. Это подразумевает классификацию вариантов в соответствии с их взаимосвязью с последовательностями кодирования в геноме и согласно их влиянию на последовательность кодирования и на генный продукт. В технике известного уровня данную информацию записывают обычно в файл MAF.

Процесс анализа цепочек ДНК (вариант, CNV = вариация номера копии, метилирование и др.) для определения их взаимосвязи с генами (и протеинами) в части функций и структуры именуется функциональным или структурным анализом. Для запоминания этих данных в технике известного уровня принят ряд решений.

Формат файла генома.

Раскрытое в данном документе изобретение заключается в определении сжатой структуры данных для представления, обработки, манипулирования и передачи данных секвенирования генома с отличием ее от решений в технике известного уровня, по меньшей мере, в следующих аспектах.

Она не опирается на форматы представления информации по геному в технике известного уровня (т.е. FASTQ, SAM).

Она реализует новую оригинальную классификацию данных генома и метаданных согласно их спе-

цифическим характеристикам. Считанные фрагменты последовательности подставляются в референсную последовательность и группируются в различных классах согласно результатам процесса выравнивания. При этом возникают классы данных с пониженной энтропией информации, которые можно кодировать с повышенной эффективностью при использовании различных специфических алгоритмов сжатия.

Это определяет синтаксические элементы и сопутствующий процесс кодирования/декодирования с переносом считанных фрагментов последовательности и информации о выравнивании в форму представления, обладающую повышенной эффективностью при ее обработке для приложений анализа нижестоящих процессов.

Классификация считанных фрагментов согласно результатам их подстановок и кодирования с использованием дескрипторов для записи в уровни (уровень позиции, уровень расстояния до партнера, уровень типа рассогласования и т.д.) предоставляет следующие преимущества:

снижение информационной энтропии при моделировании различных синтаксических элементов посредством специфической модели источника;

повышенная эффективность доступа к данным, ранее организованным в виде групп/уровней, имеющих конкретное смысловое значение для ступеней анализа нижестоящего процесса и допускающих раздельный и независимый доступ;

наличие модульной структуры данных, которую можно обновлять с последовательным приращением шагов путем доступа исключительно к требуемой информации без необходимости декодирования всего информационного содержимого;

информация по геному, выдаваемая машинами секвенирования, имеет высокую степень избыточности по своей внутренней сути вследствие природы самой информации и ввиду необходимости минимизации влияния ошибок, возникающих внутри процесса секвенирования. Это означает, что соответствующая генетическая информация, которую необходимо идентифицировать и проанализировать (вариации по отношению к референсной последовательности) является лишь малой долей полученных данных. Форматы представления данных генома в технике известного уровня не призваны "изолировать" значащую информацию на заданной ступени анализа от остальной части информации с целью оперативного обеспечения ее доступности для приложений анализа;

решение, заложенное в раскрытое здесь изобретение, состоит в представлении данных генома таким образом, что любая относящаяся к делу часть данных легко доступна для приложений анализа без необходимости доступа ко всему объему данных и его декомпрессии с сопутствующим эффективным сокращением избыточности данных за счет эффективного сжатия с целью минимизации требуемого объема памяти и ширины полосы передачи.

Ключевыми элементами изобретения являются

1. Спецификация формата файла, который содержит структурированные и избирательно доступные единицы доступа (AU) элементов данных в сжатом виде. Такой подход можно считать противоположным по сравнению с подходами из техники известного уровня, например, SAM и BAM, в которых данные структурированы в несжатом виде и далее выполняется сжатие всего файла. Первым очевидным преимуществом данного подхода является возможность эффективным и естественным образом обеспечить различные формы структурированного избирательного доступа к элементам данных в сжатой области, что невозможно или крайне затруднительно в подходах из техники известного уровня.

2. Структурирование информации по геному в виде специфических "уровней" однородных данных и метаданных представляет собой существенное преимущество, допускающее возможность определения различных моделей источников информации, характеризуемых низкой энтропией. Такие модели могут различаться между собой не только на разных уровнях, но и внутри каждого уровня, если сжатые данные внутри уровней разделены на блоки данных, включенные в единицы доступа. Данное структурирование обеспечивает возможность использования наиболее подходящего сжатия для каждого класса данных или метаданных и их частей с существенным выигрышем в эффективности кодирования по сравнению с подходами из техники известного уровня.

3. Информацию структурируют в виде единиц доступа (AU) таким образом, что любой подходящий поднабор данных, используемый приложениями геномического анализа, эффективно и избирательно доступен посредством соответствующих интерфейсов. Эти отличительные признаки обеспечивают ускоренный доступ к данным и повышенную эффективность обработки.

4. Определение таблицы главного указателя и таблицы местного указателя, обеспечивающее избирательный доступ к информации, пересылаемой посредством уровней закодированных (т.е. сжатых) данных без необходимости декодировать весь объем сжатых данных.

5. Возможность осуществления повторного выравнивания ранее выровненных и сжатых данных генома, когда требуется новое выравнивание по отношению к заново опубликованным референсным геномам путем выполнения эффективного транскодирования выбранных частей данных в сжатой области. Частый выпуск новых референсных геномов в настоящее время требует потребления ресурсов и расхода времени на процессы транскодирования для повторного выравнивания ранее сжатых и записанных в память данных генома по отношению к заново опубликованному референсному геному, поскольку требуется обработка всего объема данных.

Описанный в данном документе способ направлен на эксплуатирование доступных априорных знаний о данных генома с целью определения алфавита синтаксических элементов с пониженной энтропией. В геномике доступные знания обычно представляются существующей последовательностью генома обычно, но не обязательно, таких же образцов, что и подлежащий текущей обработке. Например, геномы человека различаются для различных индивидуумов только на долю процента С другой стороны, такой малый объем данных содержит информацию, обеспечивающую возможность ранней диагностики, оказания персонализированных медицинских услуг, синтезирования заказных лекарств и др. Данное изобретение направлено на определение формата представления информации по геному, обеспечивающего эффективный доступ к релевантной информации и возможность ее транспортировки и снижение удельного значения избыточной информации. В настоящем изобретении реализуются следующие технические характеристики:

1. Разложение информации по геному на "уровни" однородных метаданных с целью максимально возможного снижения информационной энтропии.

2. Определение таблицы главного указателя и таблиц местных указателей для обеспечения избирательного доступа к уровням закодированной информации без необходимости декодирования всей закодированной информации.

3. Принятие контекстно-адаптивного двоичного арифметического кодирования с целью кодирования элементов синтаксиса, определенных в п.1.

4. Синхронизация уровней для обеспечения избирательного доступа к данным без необходимости декодирования всех уровней (если не требуется).

5. Дифференциальное кодирование в отношении одной или более адаптивных референсных последовательностей, которые могут быть изменены для снижения энтропии. После кодирования на основе первой референсной последовательности записанные рассогласования могут быть использованы для "адаптации/модификации" референсных последовательностей с целью дальнейшего снижения информационной энтропии. Это процесс, который может быть итеративным с повтором в течение всего времени, пока имеет смысл снижение информационной энтропии.

С целью решения упомянутых выше проблем техники известного уровня (в части эффективного доступа к случайным позициям в файле, эффективной передачи и записи в память, эффективного сжатия) в настоящей заявке изменяется очередность данных с формированием из них пакетов, обладающих повышенной степенью однородности и/или семантически значимых с точки зрения простоты обработки. В настоящем изобретении принята также структура данных, основанная на концепции "единицы доступа".

Данные генома структурированы и закодированы в различные единицы доступа. Далее следует описание данных генома, содержащихся в различных единицах доступа.

Классификация данных генома.

Считанные фрагменты последовательности, выданные машинами секвенирования, делятся в раскрытом здесь изобретении на 5 различных "классов" в соответствии с результатами выравнивания по отношению к одной или более референсным последовательностям или геномам.

При выравнивании ДНК-последовательности нуклеотидов по отношению к референсной последовательности возможны результаты пяти видов.

1. Обнаруживается область в референсной последовательности, согласующаяся со считанной выборкой последовательности без единой ошибки (идеальная подстановка). Такая последовательность нуклеотидов именуется "идеально согласующимся считанным фрагментом" или обозначается как "класс Р".

2. Обнаруживается область в референсной последовательности, согласующаяся со считанным фрагментом последовательности с рядом рассогласований, представляющих ряд позиций, на которых машине секвенирования не удалось вызвать никакое основание (или нуклеотид). Такие рассогласования помечаются символом "N". Такие последовательности именуются "считанными фрагментами с рассогласованием N" или "классом N".

3. Обнаруживается область в референсной последовательности, согласующаяся со считанным фрагментом последовательности с рядом рассогласований, представляющих ряд позиций, на которых машине секвенирования не удалось вызвать никакое основание (или нуклеотид) либо было вызвано основание, отличное от того, которое указано в референсной последовательности. Такой тип рассогласования именуется "вариацией одиночного нуклеотида", (SNV) или "полиморфизмом одиночного нуклеотида"(SNP). Последовательность именуется "считанными фрагментами с рассогласованием М" или "классом М".

4. Четвертый класс образуется считанными фрагментами последовательности, представляющими тип рассогласования, включающий в себя те же рассогласования, что и в классе М, с добавлением имеющихся вставок и удалений (именуемых также indels). Вставки представлены последовательностью из одного или более нуклеотидов, отсутствующих в референсной последовательности, но имеющихся в считанной последовательности. В литературе вставная последовательность, расположенная на краях последовательности, именуется "мягко усеченной" (т.е. нуклеотиды не согласуются с референсной последовательностью, однако сохраняются в выровненных считанных фрагментах в противоположность "жестко усеченным"" нуклеотидам, которые сбрасываются). Сохранение или сбрасывание нуклеотидов

обычно осуществляется по решению пользователя, реализуемому в виде конфигурации выравнивающего инструмента. Удаления представляют собой "дыры" (отсутствующие нуклеотиды) в считанном фрагменте, соотнесенном с референсной последовательностью. Такие последовательности именуются "считанными фрагментами с рассогласованием I" или "классом I".

5. Пятый класс включает в себя все считанные фрагменты, для которых не обнаружены никакие действительные подстановки в референсной последовательности согласно указанным ограничениям выравнивания. Такие последовательности именуются последовательностями без подстановок (Unmapped) и относятся к "классу U".

Считанные фрагменты без подстановок могут быть скомпонованы в единую последовательность при использовании алгоритмов компоновки заново. Сразу после создания новой последовательности может быть выполнена подстановка считанных фрагментов, для которых ранее не обнаружены подстановки, по отношению к новой последовательности с последующим включением в один из четырех классов P, N, M и I.

Для структуры данных указанных данных генома требуется запоминание глобальных параметров и метаданных, подлежащих использованию механизмом декодирования. Эти данные структурированы в основном заголовке, описанном в приведенной ниже таблице.

Таблица 1

Структура основного заголовка

Элемент	**Тип**	**Описание**
Уникальный ИД	Массив байтов	Уникальный идентификатор для закодированного содержимого
Версия	Массив байтов	Мажорная + минорная версия алгоритма кодирования
Размер заголовка	Целое число	Размер всего закодированного содержимого в байтах
Длина считанных фрагментов	Целое число	Размер считанных фрагментов в случае постоянной длины считанных фрагментов. Специальное значение (например, 0) резервируют для считанных фрагментов переменной длины
Реф. счет	Целое число	Количество используемых референсных последовательностей
Счетчики единиц доступа	Массив байтов (например, целые числа)	Суммарное количество закодированных единиц доступа на одну референсную последовательность
Реф. идентификаторы	Массив байтов	Уникальные идентификаторы для референсных последовательностей
Таблица главного указателя *Позиции выравнивания первого считанного фрагмента в каждом блоке (единице доступа).* *То есть, позиция с меньшим номером первого считанного фрагмента в референсном геноме на каждый блок 4 классов* *1 на класс pos (4) на референсную последовательность*	Массив байтов (например, целые числа)	Это многомерный массив, поддерживающий произвольный доступ к единицам доступа

Сразу по окончании классификации считанных фрагментов с определением классов дальнейшая обработка состоит в определении набора различных синтаксических элементов, представляющих оставшуюся информацию, что позволяет воссоздавать считанный фрагмент ДНК при представлении его в виде подстановки в заданную референсную последовательность. Сегмент ДНК, соотносимый с заданной референсной последовательностью, может быть полностью выражен посредством указанных ниже элементов;

Начальная позиция в референсной последовательности (pos).

Флажок сигнализации о необходимости рассмотрения считанной выборки в качестве обратного замещения по отношению к референсной последовательности (rcomp).

Расстояние до партнера по паре в случае спаренных считанных фрагментов (pair).

Значение длины считанного фрагмента в случае технологии секвенирования приводит к получению считанных фрагментов переменной длины. В случае постоянной длины считанных фрагментов длину считанного фрагмента, связываемую с каждым из считанных фрагментов, можно по очевидным причинам опускать и можно записывать в основной заголовок файла.

Для каждого рассогласования

позиция рассогласования (nmis для класса N, snpp для класса M и indp для класса I); тип рассогласования (отсутствует в классе N, snpt в классе M, indt в классе I).

Флажки, указывающие на специфические характеристики считанного фрагмента последовательности, такие как

шаблон, имеющий множество сегментов в секвенировании;

каждый сегмент надлежащим уровнем выровнен с помощью выравнивателя;

сегмент без подстановки;

следующий сегмент в шаблоне без подстановки;

сигнализация о первом или последнем сегменте;

сбой управления качеством;

PCR (ПЦР) или оптический дубликат;

вторичное выравнивание;

дополнительное выравнивание.

Выборочно используемая мягко усеченная строка нуклеотидов, если имеется (indc в классе 1).

В рамках данной классификации создаются группы дескрипторов (синтаксических элементов), которые могут быть использованы для однозначного представления считанных фрагментов последовательности генома. В приведенной ниже таблице кратко охарактеризованы синтаксические элементы, необходимые для каждого класса выравниваемых считанных фрагментов.

Таблица 2

Определенные уровни на класс данных

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

Считанные фрагменты, принадлежащие классу P, характеризуются и могут быть идеально воссозданы исключительно по позиции, по информации обратного замещения и смещения между партнерами в случае, когда они были определены согласно технологии секвенирования с образованием смежных пар, по некоторым флажкам и по длине считанного фрагмента. В следующем разделе подробно описаны способы определения этих дескрипторов.

Уровень дескрипторов позиции.

В каждой единице доступа (AU) записана в заголовок AU только позиция подстановки первого закодированного считанного фрагмента в качестве абсолютной позиции в референсном геноме. Все другие позиции выражаются в виде разности по отношению к предыдущей позиции и записываются на хранение на специфическом уровне. Данное моделирование источника информации, определяемого последовательностью позиций считывания, в целом характеризуется пониженной энтропией, в частности, для процессов секвенирования с генерацией результатов с высокой степенью покрытия. Как только записана на сохранение абсолютная позиция первого выравнивания, все позиции других считанных фрагментов выражаются в виде разности (расстояния) по отношению к позиции первого выравнивания.

Например, на фиг. 4 показано, как после кодирования начальной позиции первого выравнивания в

качестве позиции "10000" в референсной последовательности позиция второго считанного фрагмента, начинающаяся на значении 10180, кодируется в виде "180". В случае данных с высокой степенью покрытия (>50×) большинство дескрипторов вектора позиции демонстрируют очень высокую частоту появления низких значений, таких как 0 и 1 и других малых целых чисел. На фиг. 4 показано, как позиции трех пар считанных фрагментов кодируются в уровне pos.

Для позиций считанных фрагментов, принадлежащих классам N, M, P и I, используется одна и та же модель источника. Чтобы обеспечить избирательный доступ к данным в любой комбинации, позиции считанных фрагментов, принадлежащих четырем классам, кодируют в раздельных уровнях, как указано в табл. 1.

Уровень дескрипторов образования пар.

Дескриптор образования пар записывают на хранение в уровень pair. В этом уровне сохраняются дескрипторы, кодирующие информацию, необходимую для воссоздания оригинальных пар считанных фрагментов, если используемая технология секвенирования выдает считанные фрагменты попарно. Несмотря на то, что на дату раскрытия изобретения подавляющее большинство данных секвенирования генерируется с использованием технологии выдачи спаренных считанных фрагментов, это обеспечивается не всеми технологиями. Это причина, по которой наличие данного уровня не является обязательным для воссоздания всей информации с данными секвенирования, если технология секвенирования рассматриваемых данных генома не генерирует информацию в виде спаренных считанных фрагментов.

Определения:

mate pair (партнер по паре): считанный фрагмент, связываемый с другим считанным фрагментом пары считанных фрагментов (например, Read2 является смежной парой Read1 в примере на фиг. 4);

pairing distance (расстояние образования пары): количество позиций нуклеотидов в референсной последовательности, которые отделяют одну позицию в первом считанном фрагменте (анкер спаривания, например, последний нуклеотид первого считанного фрагмента) от одной позиции второго считанного фрагмента (например, первого нуклеотида второго считанного фрагмента);

most probable pairing distance (MPPD) (максимально вероятное расстояние образования пар): это наиболее вероятное расстояние образования пар, выражаемое количеством позиций нуклеотидов;

position pairing distance (PPD) (расстояние образования пар по позициям): PPD является способом выражения расстояния образования пар в виде количества считанных фрагментов, отделяющих один считанный фрагмент от соответствующего ему партнера по паре, находящегося на конкретной позиции в уровне дескриптора;

most probable position pairing distance (MPPPD) (наиболее вероятное расстояние образования пар по позициям): представляет собой наиболее вероятное количество считанных фрагментов, отделяющих один считанный фрагмент от его партнера по паре, находящегося на конкретной позиции уровня дескриптора;

position pairing error (PPE) (ошибка образования пары по позиции): определяется в виде разности между MPPD или MPPPD и фактической позицией партнера по паре;

pairing anchor (анкер образования пары): позиция последнего нуклеотида первого считанного фрагмента в паре, используемая в качестве позиции отсчета для вычисления расстояния партнера по паре в виде количества позиций нуклеотидов или количества позиций считанных фрагментов.

Фиг. 5 илюстрирует вычисление расстояния образования пары среди пар считанных фрагментов. Уровень дескриптора пары представляет собой вектор ошибок образования пары, вычисляемый в виде количества считанных фрагментов, пропускаемых с целью достичь партнера по паре первого считанного фрагмента пары по отношению к определенному расстоянию образования пар при декодировании. На фиг. 6 представлен пример способа вычисления ошибок образования пар в виде как абсолютного значения, так и дифференциального вектора (характеризуется пониженной энтропией для высоких степеней покрытия).

Такие же дескрипторы используют в качестве информации об образовании пар считанных фрагментов, принадлежащих классам N, M, P и I. Чтобы обеспечить избирательный доступ к различным классам данных, информацию об образовании пар считанных фрагментов, принадлежащих четырем классам, кодируют в раздельных уровнях, как показано на фигуре.

Информация об образовании пар в случае считанных фрагментов, подставленных в разные референсные последовательности.

В процессе подстановки считанных фрагментов последовательности в референсную последовательность не редки случаи, когда первый считанный фрагмент в паре подставляют на одну референсную позицию (например, на позицию хромосомы 1), а второй - на другую референсную позицию (например, на позицию хромосомы 4). В этом случае описанную выше информацию об образовании пар необходимо объединять с дополнительной информацией, относящейся к референсной последовательности, используемой для подстановки одного из считанных фрагментов. Это обеспечивается путем кодирования.

1. Резервное значение (флажок), указывающее на то, что пара подставлена в две различающиеся между собой последовательности (различающиеся значения указывают, что считанный фрагмент 1 и считанный фрагмент 2 подставлены в последовательность, которая на текущий момент не закодирована).

2. Уникальный идентификатор референсной последовательности, соотносимый с идентификаторами референсной последовательности, закодированными в структуре основного заголовка, как описано в табл. 1.

3. Третий элемент, содержащий информацию о подстановке в референсную последовательность, идентифицированную в п.2 и выражаемую в виде смещения по отношению к последней закодированной позиции.

На фиг. 7 приведен пример данного сценария. На фиг. 7, поскольку считанный фрагмент 4 не подставлен в закодированную на данный момент последовательность, кодер генома сигнализирует о данном состоянии путем включения дополнительных дескрипторов в уровень пары (pair). В примере, показанном на фиг. 7, считанный фрагмент 4 пары 2 подставляется в референсную последовательность № 4, хотя текущая закодированная референсная последовательность имеет номер 1. Данная информация кодируется с использованием трех компонентов:

1) Одно специальное резервное значение кодируется в виде расстояния при образовании пары (в данном случае Oxffffff).

2) Второй дескриптор предоставляет идентификатор референсной последовательности из списка в основном заголовке (в данном случае 4).

3) Третий элемент содержит информацию о подстановке в соответствующую референсную последовательность (170).

Уровень дескриптора обратного замещения.

Каждый считанный фрагмент пар считанных фрагментов, выдаваемых технологиями секвенирования, может исходить от любых цепочек генома секвенируемой органической пробы. Однако в качестве референсной последовательности используется только одна из двух цепочек. На фиг. 8 показано, как в паре считанных фрагментов один считанный фрагмент (read 1) может поступать от одной цепочки, а другой считанный фрагмент (read 2) может поступать от другой цепочки.

Если цепочка 1 используется в качестве референсной последовательности, то считанный фрагмент 2 может быть закодирован в качестве обратного замещения соответствующего фрагмента в цепочке 1. Это иллюстрируется фиг. 9.

В случае связанных считанных фрагментов возможны четыре комбинации пар с партнерами прямого и обратного замещения. Это иллюстрируется фиг. 10. В уровне rcomp кодируются четыре возможные комбинации.

Такое же кодирование используется для информации обратного замещения считанных фрагментов, принадлежащих классам P, N, M, I. Чтобы обеспечить расширенный избирательный доступ к данным, информацию обратного замещения считанных фрагментов, принадлежащих четырем классам, кодируют в различных уровнях, как показано в табл. 2.

Рассогласования класса N.

Класс N включает в себя все считанные фрагменты, отображающие рассогласования, где N присутствует вместо вызова основания. Все другие основания идеально согласованы в референсной последовательности.

Позиции Ns в считанном фрагменте 1 кодируются в виде

абсолютной позиции в считанном фрагменте 1; или

в качестве дифференциальной позиции по отношению к предыдущей позиции N в этом же считанном фрагменте (к той, которая имеет максимально низкую энтропию).

Позиции Ns в считанном фрагменте 2 кодируются в виде

абсолютной позиции в считанном фрагменте 2 + длины считанного фрагмента 1; или

в качестве дифференциальной позиции по отношению к предыдущей позиции N (к той, которая имеет максимально низкую энтропию).

В уровне nmis кодирование каждой пары считанных фрагментов заканчивают специальным разделительным символом ("сепаратором") "S". Это иллюстрируется фиг. 11.

Кодирование замен (рассогласования или SNP).

Заменой является по определению наличие в подставленном считанном фрагменте нуклеотида, отличающегося от того, который находится на этой же позиции в референсной последовательности (см. фиг. 12). Каждая замена может быть закодирована как "позиция" (уровень snpp) и "тип" (уровень snpt). См. фиг. 13, 14, 16 и 15; или только "позиция", но с использованием одного уровня snpp на тип рассогласования. См. фиг. 17.

Позиции замен.

Позиция замены вычисляется как для значений уровня nmis, а именно

в считанном фрагменте 1 кодируются замены в качестве абсолютной позиции в считанном фрагменте 1; или

в качестве дифференциальной позиции по отношению к предыдущей замене в этом же считанном фрагменте.

В считанном фрагменте 2 кодируются замены

в качестве абсолютной позиции в считанном фрагменте 2 + длины считанного фрагмента 1; или

в качестве дифференциальной позиции по отношению к предыдущей замене. Фиг. 13 иллюстрирует кодирование позиций замены в уровне snpp. Позиции замен могут быть вычислены либо как абсолютные, либо как дифференциальные значения.

В уровне snpp кодирование каждой пары считанных фрагментов заканчивают специальным разделительным символом ("сепаратором").

Дескрипторы типов замен.

Для класса M (и для класса I, как описано в следующих разделах), рассогласования кодируются посредством индекса (перемещающегося слева направо) от фактического символа, находящегося в референсной последовательности, к соответствующему символу замены, находящемуся в считанном фрагменте (A, C, G, T, N, Z). Например, если выровненный считанный фрагмент представляет C вместо T, находящегося на этой же позиции в референсной последовательности, индекс рассогласования помечается как "4" В процессе декодирования считывается закодированный синтаксический элемент, нуклеотид на заданной позиции в референсной последовательности, и перемещается слева направо для выборки декодированного символа. Например, значение "2", принятое для позиции, где в референсной последовательности находится G, декодируется как "N". На фиг. 14 представлены все возможные замены и соответствующие символы кодирования в режиме без использования кодов неоднозначности IUPAC, а на фиг. 15 представлен пример кодирования типов замен в уровне snpt.

В случае использования кодов неоднозначности IUPAC индексы замены изменяются, как показано на фиг. 16.

В случае, когда описанное выше кодирование типов замен привносит высокую информационную энтропию, альтернативный способ кодирования замен состоит в запоминании только позиций рассогласования в отдельных уровнях (по одной на нуклеотид), как показано на фиг. 17.

Кодирование вставок и удалений.

Для класса I рассогласования и удаления кодируются посредством индексов (перемещающихся справа налево) от фактического символа, находящегося в референсной последовательности, к соответствующему символу замены, находящемуся в считанном фрагменте {A, C, G, T, N, Z}. Например, если выровненный считанный фрагмент содержит C вместо T, находящегося на этой же позиции в референсной последовательности, то индеском рассогласования является "4", В случае, когда считанный фрагмент представляет удаление, где в референсной последовательности находится A, закодированным символом является "5". В процессе декодирования считывается закодированный синтаксический элемент, нуклеотид на заданной позиции в референсной последовательности, и перемещается слева направо для выборки декодированного символа. Например, значение "3", принятое для позиции, где в референсной последовательности находится G, декодируется в качестве "Z", что указывает на наличие удаления в считанном фрагменте последовательности.

Вставки кодируются как 6, 7, 8, 9,10 соответственно для вставленных A, C, G, T, N.

В случае принятия кодов неоднозначности IUPAC механизм замены является точно таким же, однако вектор замены расширен следующим образом: = {A, C, G, T, N, Z, M, R, W, S, Y, K, V, H, D, B}.

На фиг. 18 и 19 представлены примеры способа кодирования замен, вставок и удалений в паре считанных фрагментов класса I.

Приводимые далее структуры файла и единицы доступа описываются со ссылкой на элементы кодирования, раскрытые здесь выше. Однако единицы доступа, формат файла и мультиплексирование обеспечивают такое же техническое преимущество также с другими и разными алгоритмами моделирования источника и сжатия данных генома.

Формат файла: избирательный доступ в области данных генома.

Таблица главного указателя.

Чтобы поддержать избирательный доступ в специфические области выровненных данных, описанная в данном документе структура данных реализует инструмент индексации, именуемый таблицей главного указателя (MIT). Это многомерный массив, содержащий обозначения позиций, на которых специфические считанные фрагменты подставляются в используемые референсные последовательности. Значения, содержащиеся в MIT, представляют собой позиции подстановки первого считанного фрагмента в каждом уровне pos таким образом, что поддерживается непоследовательный доступ к каждой единице доступа. MIT содержит одну секцию на каждый класс данных (P, N, M и I) и на каждую референсную последовательность. MIT содержится в основном заголовке закодированных данных. На фиг. 20 представлена обобщенная структура основного заголовка. Фиг. 21 является обобщенным визуальным представлением MIT, а на фиг. 22 изображен пример MIT для класса P закодированных считанных фрагментов.

Значения, содержащиеся в таблице MIT, изображенной на фиг. 22, используются для прямого доступа к представляющему интерес участку (и соответствующей единице доступа) в сжатой области. Например, применительно к фиг. 22, если требуется доступ к участку между позициями 150,000 и 250,000 в референсной последовательности 2, то приложение декодирования выполнит пропуск до второй референсной последовательности в MIT и будет искать два значения k1 и k2, такие, что k1 <150,000 и k2 >250,000. При этом k1 и k2 являются двумя индексами, считанными из MIT. В примере на фиг. 22 полу-

ченным результатом являются позиции 3 и 4 второго вектора MIT. Эти возвращенные значения используются далее приложением декодирования для выборки позиций соответствующих данных из уровня pos таблицы местного указателя, как описано в следующем разделе.

Вместе с указателями на уровень, содержащий данные, принадлежащие четырем классам данных генома, описанным выше, таблица MIT может быть использована в качестве указателя дополнительных метаданных и/или аннотаций, добавляемых к данным генома на протяжении их жизненного цикла.

Таблица местного указателя.

Каждому из описанных выше уровней данных назначается префикс в виде структуры данных, именуемой местным заголовком. Местный заголовок содержит уникальный идентификатор уровня, вектор счетчиков единиц доступа на каждую референсную последовательность, таблицу местного указателя (LIT) и по дополнительному выбору некоторые специфические метаданные уровня. LIT представляет собой вектор указателей на физическое расположение данных, принадлежащих каждой единице доступа в полезной нагрузке уровня. На фиг. 23 изображены в обобщенном виде заголовок уровня и полезная нагрузка, где LIT используется для доступа к специфическим областям закодированных данных в непоследовательном режиме.

В предыдущем примере с целью доступа к области 150,000 ... 250,000 считанных фрагментов, выровненных в референсной последовательности № 2, приложение декодирования осуществило выборку позиций 3 и 4 из MIT. Эти значения должны использоваться процессом декодирования для доступа к третьему и четвертому элементам соответствующей секции LIT. В примере, показанном на фиг. 24, счетчики суммарного количества единиц доступа, содержащиеся в заголовке уровня, используются для пропуска индексов LIT, соответствующих единицам доступа, относящимся к референсной последовательности 1 (в примере - 5). При этом индексы, обозначающие физическое расположение запрашиваемых единиц доступа в закодированном потоке, вычисляются следующим образом: расположение блоков данных, принадлежащих запрашиваемым единицам доступа = блоки данных, принадлежащие пропускаемым единицам доступа референсной последовательности 1+ позиция, выбранная посредством MIT, т.е. позиция первого блока: 5 + 3 = 8. Позиция последнего блока: 5 + 4 = 9.

Блоки данных, выборка которых осуществлена с использованием механизма индексации, именуемого таблицей местного указателя, являются частью запрашиваемых единиц доступа.

На фиг. 26 показано, каким образом блоки данных, выборка которых выполнена с использованием MIT и LIT, составляют одну или более единиц доступа.

Единицы доступа.

Из данных генома, распределенных по классам данных и структурированных в сжатом или несжатом уровнях, организуют различные единицы доступа.

По определению геномными единицами доступа (AU) являются разделы данных генома (в сжатом или несжатом виде), которые воссоздают последовательности нуклеотидов и/или релевантные метаданные и/или последовательности ДНК/РНК (например, виртуальную референсную последовательность) и/или аннотационные данные, выданные машиной секвенирования генома и/или устройством обработки данных генома или приложением анализа.

Единица доступа представляет собой блок данных, который может быть декодирован либо независимо от других единиц доступа при использовании только данных, доступных на общесистемном уровне (например, конфигурационных данных декодера) либо при использовании информации, содержащейся в других единицах доступа.

Единицы доступа содержат информационные сведения, относящиеся к данным генома, в виде информации о позиции (абсолютной или относительной), информации в отношении обратного замещения и, возможно, информации об образовании пар и дополнительных данных. Возможна идентификация единиц доступа нескольких типов. Единицы доступа различают по следующим признакам:

тип, характеризующий природу данных генома и наборы данных, которые переносят эти единицы доступа, и возможный способ доступа к ним;

очередность, характеризующая единый порядок доступа к единицам, относящимся к одному и тому же типу.

Единицы доступа любого типа можно далее классифицировать по различным "категориям".

Далее следует не являющийся исчерпывающим список определений единиц доступа различных типов для данных генома.

1) Единицам доступа типа 0 не требуется ссылка на какую-либо информацию, поступающую от других единиц доступа, предназначенных для доступа или декодирования и доступа (см. фиг. 29). Референсный номер всей информации, переносимый данными или наборами данных, содержащимися в этих единицах доступа, может быть независимым образом считан и обработан устройством декодирования или приложением обработки данных.

2) Единицы доступа типа 1 содержат данные, соотносимые с данными, переносимыми единицами доступа типа 0 (см. фиг. 30) Для считывания или декодирования и обработки данных, содержащихся в единицах доступа типа 1, требуется наличие доступа к одной или более единицам доступа типа 0.

3) Единицы доступа данного типа могут содержать информацию о рассогласовании или разнород-

ности либо о несоответствии по отношению к информации, содержащейся в единице доступа типа 0.

4) Единицы доступа типов 2, 3 и 4 содержат данные, относящиеся к единице доступа типа 1 (см. фиг. 31, 32 и 33). Для считывания или декодирования и обработки данных или наборов данных, содержащихся в единицах доступа типов 2, 3 и 4, требуется информация, переносимая данными или наборами данных, содержащимися в единицах доступа

5) типов 0 и 1. Разница между единицами доступа типов 2, 3 и 4 определяется природой информации, которую они содержат.

6) Единицы доступа типа 5 содержат метаданные (например, оценочные показатели качества) и/или аннотационные данные, соотносимые с данными или наборами данных, содержащимися в единице доступа типа 1. Единицы доступа типа 5 могут быть классифицированы и маркированы в различных уровнях.

7) Единицы доступа типа 6 содержат данные или наборы данных, классифицированные в качестве аннотационных данных. Единицы доступа типа 6 могут быть классифицированы и маркированы в уровнях.

8) Единицы доступа дополнительных типов могут расширять описанные здесь структуры и механизмы. Примером, не подразумевающим ограничений, может служить то, в единицах доступа новых типов могут быть закодированы результаты вызова варианта генома, структурного и функционального анализа. Описанная здесь организация данных в форме единиц доступа не препятствует инкапсулированию в единицы доступа данных любого типа, что является механизмом, полностью прозрачным в отношении природы закодированных данных.

Единицы доступа данного типа могут содержать информацию о рассогласовании или разнородности либо о несоответствии по отношению к информации, содержащейся в единице доступа типа 0.

Фиг. 28 иллюстрирует составление единиц доступа из заголовка и одного или более уровней однородных данных. Каждый уровень может быть образован одним или более блоками. Каждый блок содержит несколько пакетов, и пакеты представляют собой структурированную последовательность дескрипторов, описанных выше, для представления, например, позиций считанных фрагментов, информации об образовании пар, информации об обратном замещении, о позициях и типах рассогласования и др. Каждая единица доступа может иметь в каждом блоке разное количество пакетов, но в пределах единицы доступа все блоки содержат одно и то же количество пакетов.

Каждый пакет данных может быть идентифицирован посредством комбинации из трех идентификаторов XYZ, где

X обозначает единицу доступа, которой принадлежит пакет,

Y обозначает блок, которому принадлежит пакет (т.е. тип инкапсулированных в него данных),

Z обозначает порядковый номер пакета по отношению к другим пакетам в этом же блоке.

На фиг. 28 представлен пример маркировки единиц доступа и пакетов.

На фиг. 34-38 показаны единицы доступа различных типов; общий синтаксис их обозначения является следующим: AU_T_N является единицей доступа типа T, где идентификатор N может косвенно обозначать или не обозначать очередность согласно типу единицы доступа. Идентификаторы используются для однозначного соотнесения единиц доступа одного типа с единицами доступа других типов, требуемыми для полного декодирования переносимых данных генома.

Единицы доступа любого типа могут быть классифицированы и маркированы в различных "категориях" в соответствии с различными процессами секвенирования. Примером, не подразумевающим ограничений, может служить выполнение классификации и маркировки, в случае секвенирования одного и того же организма в разное время (единицы доступа содержат информацию по геному с коннотацией "по времени"), секвенирования органических проб разного вида одних и тех же организмов (например, кожи, крови, волос в качестве проб для людей). Это единицы доступа с "биологической" коннотацией.

Единицы доступа типов 1, 2, 3 и 4 строятся согласно результатам применения функции согласования к фрагментам последовательности генома (именуемым также считанными фрагментами) по отношению к референсной последовательности, закодированной в единицах доступа типа 0, с которыми они соотносятся.

Например, в единицах доступа типа 1 (см. фиг. 30) могут содержаться флажки позиций и обратного замещения тех считанных фрагментов, для которых поучено идеальное согласование (или максимально возможный оценочный показатель качества, соответствующий выбранной функции согласования), когда функция согласования применяется к специфическим областям референсной последовательности, закодированной в единицах доступа типа 0. Вместе с данными, содержащимися в единицах доступа типа 0, такая информация функции согласования является достаточной для полного воссоздания всех считанных фрагментов последовательности генома, представляемых посредством набора данных, переносимого единицами доступа типа 1.

Применительно к классификации данных генома, ранее описанной в данном документе, описанные выше единицы доступа типа 1 будут содержать информацию, относящуюся, к считанным фрагментам класса P (идеальные согласования) последовательности генома.

В случае считанных фрагментов переменной длины и спаренных считанных фрагментов данные,

содержащиеся в единицах доступа типа 1, упомянутых в предыдущем примере, должны быть объединены с данными, представляющими информацию об образовании пар считанных фрагментов и о длине считанных фрагментов для получения возможности полного воссоздания данных генома, включая ассоциативную связь с парами считанных фрагментов. В отношении классификации данных, ранее введенной в данном документе, уровни pair и rlen кодируются в единице доступа типа 1. Функция согласования, примененная в отношении единиц доступа типа 1 для классификации содержимого единиц доступа по типам 2, 3 и 4, может обеспечить получение таких результатов как

каждая последовательность, содержащаяся в единице доступа типа 1, идеально согласуется с последовательностями, содержащимися в единице доступа типа 0, в соответствии с указанной позицией;

каждая последовательность, содержащаяся в единице доступа типа 2, идеально согласуется с последовательностью, содержащейся в единице доступа типа 0, в соответствии с указанной позицией, кроме случаев с наличием символов "N"(отсутствие вызова основания устройством секвенирования) в последовательности в единице доступа типа 2;

каждая последовательность, содержащаяся в единице доступа типа 3, включает в себя варианты в форме заменяемых символов (вариантов) по отношению к последовательности, содержащейся в единице доступа типа 0, в соответствии с указанной позицией;

каждая последовательность, содержащаяся в единице доступа типа 4, включает в себя варианты в форме заменяемых символов (вариантов), вставок и/или удалений по отношению к последовательности, содержащейся в единице доступа типа 0, в соответствии с указанной позицией.

Единицы доступа типа 0 являются упорядоченными (например, пронумерованными), однако не требуется соблюдение этой упорядоченности при передаче и/или записи в память (техническое преимущество: параллельная обработка, образование параллельных потоков, мультиплексирование). Единицы доступа типов 1, 2, 3 и 4 не требуют упорядочения и не требуют соблюдения очередности при передаче и/или записи в память (техническое преимущество: параллельная обработка, образование параллельных потоков).

Количество считанных фрагментов последовательности, содержащихся в каждой единице доступа, является параметром конфигурирования, задаваемым пользователем с помощью пользовательского интерфейса, применяемого при кодировании данных генома согласно настоящему изобретению. Возможна пересылка данного параметра конфигурирования, например, в заголовке соответствующей единицы доступа.

Технические эффекты.

Технический эффект от структурирования информации по геному в виде описываемых здесь единиц доступа состоит в том, что данные генома

1) можно избирательно запрашивать с целью доступа

к специфическим "категориям" данных, например, со специфической коннотацией по времени или биологической коннотацией) без необходимости декомпрессии данных генома в целом и/или наборов данных и/или связанных с ними метаданных;

к специфическим областям генома для всех "категорий", для поднабора "категорий", для одиночной "категории" (с соответствующими метаданными или без них) без необходимости декомпрессии других областей генома;

2) можно обновлять путем постепенного добавления новых данных с предоставлением доступа к ним, когда

выполняется новый анализ надданными или наборами данных генома генерируются новые данные или наборы данных генома путем секвенирования одних и тех же организмов (различных биологических проб, биологической пробы одного и того же типа с дифференциацией характеристик, например, анализа крови с разбросом во времени и др.);

3) можно эффективно транскодировать в новый формат данных в случае необходимости использования новых данных или наборов данных генома в качестве новой референсной последовательности (например, нового референсного генома, переносимого единицей доступа типа 0) обновления спецификации формата кодирования.

В отношении решений из техники известного уровня (таких как SAM/BAM) упомянутые выше технические характеристики адресованы проблемам запрашивания фильтрации данных на уровне приложения, когда выполнена выборка всех данных с декомпрессией из закодированного формата.

Далее приведены примеры сценария приложения, в которых структура единицы доступа становится инструментальной структурой для технологического преимущества.

Избирательный доступ.

В частности, раскрытая здесь структура данных, основывающаяся на единицах доступа различных типов, обеспечивает возможность

извлекать только информацию (данные или наборы данных) считанных фрагментов всего секвенирования всех "категорий" или поднабора (т.е. одного или более уровней) или одиночной "категории" без необходимости декомпрессии также содержимого соответствующих метаданных (ограничение из техники известного уровня: SAM/BAM, которые не способны даже поддерживать различие между различными

категориями или уровнями);

извлекать все считанные фрагменты, выровненные в специфических областях принятой к исполнению референсной последовательности для всех категорий, поднаборов категорий, одиночной категории (с соответствующими метаданными или без них) без необходимости декомпрессии также других областей генома (ограничение из техники известного уровня: SAM/BAM).

На фиг. 39 показано, как для доступа к информации по геному, подставленной во второй сегмент референсной последовательности (AU 0-2) с рассогласованиями, требуется только декодирование единиц доступа AU 0-2 ,1-2 и 3-2. Это пример избирательного доступа, соответствующего как критериям, относящимся к области подстановки (т.е. позиции в референсной последовательности), так и критериям, относящимся к функции согласования, применяемой к считанным фрагментам закодированной последовательности по отношению к референсной последовательности (т.е. с рассогласованиями только в данном примере).

Дополнительное техническое преимущество состоит в том, что намного повышается эффективность запрашивания данных в части доступа к данным и скорости исполнения, поскольку запрашивание основывается на доступе и декодировании только в отношении выбранных "категорий", конкретных областей или геномных последовательностей увеличенной длины и только конкретных уровней для единиц доступа типов 1, 2, 3, 4, которые соответствуют критериям применяемых запросов и любой их комбинации.

Организация единиц доступа типов 1, 2, 3, 4 в виде уровней обеспечивает возможность эффективного извлечения последовательностей нуклеотидов со специфическими вариациями (например, рассогласованиями, удалениями) по отношению к одному или более референсным геномам;

которые не встраиваются подстановкой ни в один из рассматриваемых референсных геномов;

которые идеально встраиваются подстановкой в один или более референсных геномов;

которые встраиваются подстановкой с одним или более уровнями точности.

Инкрементное обновление.

Единицы доступа типов 5 и 6 обеспечивают возможность простой вставки аннотаций без необходимости депакетизировать/декодировать/декомпрессировать весь файл, повышая тем самым эффективность обработки файла, являющуюся ограниченной в подходах из известного уровня техники. При существующих решениях по компрессии для получения доступа к требуемым данным генома может потребоваться доступ к большим объемам сжатых данных и их обработка. Это вызывает неэффективное использование полосы пропускания ЗУПВ и рост потребления мощности, в том числе в аппаратных реализациях. Проблемы потребления мощности и доступа к памяти могут быть решены при использовании подхода на основе описанных здесь единиц доступа.

Механизм индексации данных, описанный применительно к таблице главного указателя (см. фиг. 21), в сочетании с использованием единиц доступа обеспечивает инкрементное обновление закодированного содержимого, как описано ниже.

Вставка дополнительных данных.

Возможно периодическое добавление новой информации по геному к существующим данным генома по нескольким причинам. Например, в следующих случаях

организм секвенирован в разные моменты времени;

несколько различных проб одного и того же индивидуума секвенировано в одно и то же время;

новые данные сгенерированы в ходе процесса секвенирования (потоковой обработки).

В указанных выше ситуациях описанное здесь структурирование данных с использованием единиц доступа и описанная в разделе по формату файла структура данных обеспечивают возможность инкрементного интегрирования заново сгенерированных данных без необходимости перекодирования существующих данных. Процесс инкрементного обновления может быть реализован следующим образом:

1) заново созданные единицы доступа могут быть просто последовательно присоединены в файле к существующим единицам доступа, и

2) индексацию заново созданных данных или наборов данных включают в таблицу главного указателя, описанную в разделе по формату файла в данном документе. Один индекс определяет позицию вновь созданной единицы доступа в существующей референсной последовательности, другие индексы представляют собой указатели вновь созданных единиц доступа в физическом файле для обеспечения прямого и избирательного доступа к ним.

Данный механизм иллюстрируется фиг. 40, где существующие данные, закодированные в трех единицах доступа типа 1 и в четырех единицах доступа на каждый из типов 2-4, обновляются посредством трех единиц доступа на тип с кодированием данных, поступающих, например, из новой последовательности, выполняемой для одного и того же индивидуума.

В специфическом случае использования потоковой обработки данных и наборов данных генома в сжатом виде инкрементное обновление существующего набора данных может быть полезно при анализе данных сразу после их выдачи машиной секвенирования и перед завершением фактического секвенирования. Кодирующее устройство (компрессор) может параллельно компоновать несколько единиц доступа путем образования "кластера" считанных фрагментов последовательности, встраиваемых подстановкой в одну и ту же область выбранной референсной последовательности. Как только количество считан-

ных фрагментов в первой единице доступа доходит до значения выше предварительно сконфигурированного порога/параметра, единица доступа готова к пересылке её приложению анализа. В сочетании с заново закодированной единицей доступа устройство кодирования (компрессор) должно обеспечить, что все единицы доступа, от которых зависит новая единица доступа, уже переданы на приемный конец или пересылаются вместе с новой единицей доступа. Например, для правильного декодирования единицы доступа типа 3 требуется наличие на приемном конце соответствующей единицы доступа типа 0 и типа 1. Посредством описанного механизма приложение вызова варианта на приемной стороне получает возможность инициировать вызов варианта при приеме единицы доступ перед завершением процесса секвенирования на передающей стороне. Схема данного процесса представлена на фиг. 41.

Новый анализ результатов.

В пределах жизненного цикла обработки генома может быть применено несколько итераций анализа генома к одним и тем же данным (например, при вызове другого варианта с использованием другого алгоритма обработки). Использование единиц доступа в определении, приведенном в данном документе, и структуры данных, описанной в разделе данного документа по формату файла, обеспечивает возможность инкрементного обновления существующих сжатых данных с помощью результатов нового анализа.

Например, новый анализ, выполненный на существующих сжатых данных, может привести к созданию новых данных в следующих случаях:

1) новый анализ может изменить существующие результаты, ранее "привязанные" к закодированным данным. Этот случай иллюстрируется фиг. 42 и реализуется путем полного или частичного перемещения содержимого одной единицы доступа из одного типа в другой. В случае, когда требуется создание новых единиц доступа (с учетом предварительно заданного максимально допустимого размера единицы доступа) необходимо создать соответствующие индексы в таблице главного указателя и по мере необходимости выполнить сортировку соответствующего вектора;

2) новые данные создаются в результате нового анализа и должны быть "привязаны" к существующим закодированным данным. В данном случае новые единицы доступа типа 5 могут быть созданы и последовательно присоединены к существующему вектору единиц доступ такого же типа. Этот процесс и сопутствующее обновление таблицы главного указателя изображены на фиг. 43.

Описанные выше случаи, проиллюстрированные фиг. 42 и 43, являются реализуемыми благодаря

1) возможности иметь прямой доступ только к данным с низким качеством подстановок (например, единицам доступа типа 4);

2) возможности повторно подставлять считанные фрагменты в новую область генома путем простого создания новой единицы доступа с возможной принадлежностью к новому типу (например, считанные фрагменты, включенные в единицу доступа типа 4, могут быть повторно подставлены в новую область с меньшим количеством рассогласований (типовое значение 2-3) и включены во вновь созданную единицу доступа);

3) возможности создать единицу доступа типа 6, содержащую только заново полученные результаты анализа и/или связанные с ними аннотации. В данном случае вновь созданным единицам доступа требуется только наличие в содержимом "указателей" на существующие единицы доступа, с которыми они соотносятся.

Транскодирование.

Сжатым данным генома может потребоваться транскодирование, например, в следующих ситуациях:

публикация новых референсных последовательностей;

использование другого алгоритма подстановки (повторной подстановки).

Когда данные генома подставляют в существующий референсный геном общего пользования, для публикации новой версии указанной референсной последовательности или при желании подставить данные с использованием другого алгоритма обработки в настоящее время требуется процесс повторной подстановки (re-mapping). При повторной подстановке сжатых данных с использованием форматов файла из предыдущего уровня техники (например, SAM или CRAM) требуется декомпрессия всех сжатых данных в их "исходную" форму для возможности повторной вставки со ссылкой на заново доступную референсную последовательность или с использованием другого алгоритма подстановки. Это справедливо даже в том случае, когда заново опубликованная референсная последовательность очень незначительно отличается от предыдущей последовательности или другой используемый алгоритм подстановки осуществляет подстановку, которая очень близка (или идентична) предыдущей подстановке.

Преимущество транскодирования структурированных данных генома с использованием описанных здесь единиц доступа состоит в следующем:

1. Для подстановки в отношении нового референсного генома требуется лишь перекодирование (декомпрессия и компрессия) данных из единиц доступа для подстановки в те области генома, где произошли изменения. Дополнительно пользователь может выбирать те сжатые считанные фрагменты, для которых может по той или иной причине потребоваться повторная подстановка даже в том случае, если первоначально не удалось подставить их в измененную область (это может иметь место, когда в пред-

ставлении пользователя предыдущая подстановка является низкокачественной). Данный случай использования иллюстрируется фиг. 44.

2. В случае, когда заново опубликованный референсный геном отличается от предыдущего генома лишь сдвигом целых областей на другие позиции внутри генома, операция транскодирования выполняется особенно просто и эффективно. По существу, для перемещения всех подставленных считанных фрагментов в "область сдвига" достаточно изменить лишь значение абсолютной позиции, содержащееся в заголовке связанных с этим единиц доступа (набора единиц доступа). Каждый заголовок единицы доступа содержит абсолютную позицию, на которую первый считанный фрагмент, содержащийся в единице доступа, подставлен в референсную последовательность, в то время как все другие позиции считанных фрагментов закодированы с дифференциацией относительно первого фрагмента. Поэтому простое обновление значения абсолютной позиции первого считанного фрагмента вызывает соответствующее перемещение всех считанных фрагментов в единице доступа. Данный механизм не может быть реализован на основе подходов из техники известного уровня, таких как CRAM и BAM, поскольку позиции данных генома закодированы в сжатой полезной нагрузке, что означает необходимость полной декомпрессии и повторного сжатия всех наборов данных генома.

3. Если используется другой алгоритм подстановки, можно применять его только к той части сжатых считанных фрагментов, в которой подстановка выполнена с низким качеством. Например, может оказаться целесообразным применение нового алгоритма подстановки только к считанным фрагментам, которые не идеально согласуются с референсным геномом. Применительно к форматам, существующим в настоящее время, невозможно (или возможно лишь частично с некоторыми ограничениями) извлекать считанные фрагменты в соответствии с качеством их подстановки (т.е. наличием рассогласований и их количеством). Если результаты новой подстановки выводятся инструментами новой подстановки, соответствующие считанные фрагменты могут быть транскодированы из одной единицы доступа в другую единицу доступа такого же типа (см. фиг. 46) или из единицы одного типа в единицу другого типа (см. фиг. 45).

Более того, в решениях по сжатию из техники известного уровня для получения возможности доступа к требуемым данным генома может потребоваться доступ к большому объему сжатых данных и его обработка. Это вызывает неэффективное использование полосы пропускания ЗУПВ и повышенное потребление мощности, в том числе в аппаратных реализациях. Проблемы потребления мощности и доступа к памяти могут быть решены при использовании подхода на основе описанных здесь единиц доступа.

Дополнительным преимуществом принятия описанных здесь единиц доступа для геномов является упрощение параллельной обработки и пригодность к аппаратным реализациям. Существующие решения, такие как SAM/BAM и CRAM, концептуально подходят для однопотоковых программных реализаций.

Избирательное шифрование.

Подход на основе единиц доступа, организованным по нескольким типам и уровням, как описано в данном документе, обеспечивает возможность реализации механизмов защиты содержимого, какая невозможна на основе монолитных решений из техники известного уровня.

Специалистам со средним уровнем знаний в данной области известно, что большинство информации по геному, относящейся к генетическому профилю организма, опирается на отличия (варианты) по отношению к известной последовательности (например, референсному геному или популяции геномов). Поэтому индивидуальный генетический профиль, защищаемый от неправомочного доступа, кодируют в единицах доступа типов 3 и 4, как описано в данном документе. Поэтому воплощение на практике регулируемого доступа к максимально чувствительной информации по геному, полученной с помощью процесса секвенирования и анализа, может быть реализовано путем шифрования только полезной нагрузки единиц доступа типов 3 и 4 (см. пример на фиг. 47). При этом обеспечивается существенная экономия в отношении как мощности обработки, так и ширины полосы пропускания, поскольку ресурсы, потребляющие процесс шифрования, применяются только к поднабору данных.

ФОРМУЛА ИЗОБРЕТЕНИЯ

1. Способ сжатия данных генома, картируемых на референсную последовательность, содержащий этап разделения файла данных генома на единицы доступа различных типов, указанные единицы доступа содержат данные генома, классифицируемые согласно подстановке в референсный геном и структурированные в виде уровней однородных данных, причем указанное разделение выполнено так, что

единицы доступа первого типа (280) содержат данные генома, причем указанные данные генома являются частью референсной последовательности, используемой для картирования закодированных данных и данных позиционирования, обозначающих абсолютные координаты местоположения первого считанного фрагмента указанной части референсной последовательности, причем указанные данные генома не соотносятся с единицами доступа любого другого типа (290, 300, 310, 320, 321),

единицы доступа второго типа (290) содержат информацию, относящуюся к данным позиционирования, обозначающим позицию последовательности, идеально вписывающуюся в последовательность, содержащуюся в единице доступа первого типа (280), и информацию об обратном замещении по отно-

шению к информации генома, содержащейся в единице доступа первого типа, и где

указанные данные позиционирования записаны на хранение таким образом, что позиция картирования первого считанного фрагмента запоминается в виде абсолютной позиции, а все другие позиции выражаются в виде разности по отношению к предыдущей позиции и запоминаются на специфическом уровне, и

указанные данные позиционирования и данные обратного замещения структурируются на различных уровнях однородных данных и сжимаются путем применения специфического для соответствующих данных позицинирования и данных обратного замещения алгоритмов сжатия.

2. Способ по п.1, в котором указанные единицы доступа второго типа (290) дополнительно содержат информацию, относящуюся к образованию пар (294) считанных фрагментов генома и/или дополнительную информацию, относящуюся к длине считанных фрагментов (295).

3. Способ по любому из предшествующих пунктов, в котором указанное разделение данных генома на единицы доступа различных типов дополнительно включает в себя единицу доступа дополнительного типа (300), содержащую информацию, относящуюся к позициям рассогласования, где автомату секвенирования не удалось определить какой-либо нуклеотид.

4. Способ по любому из предшествующих пунктов, в котором указанное разделение данных генома на единицы доступа различных типов дополнительно включает в себя единицу доступа дополнительного типа (310), содержащую информацию, относящуюся к позиции рассогласования (311) и к типу рассогласования (312), где указанное рассогласование относится к данным генома, относящимся к единице доступа первого типа (280).

5. Способ по любому из предшествующих пунктов, в котором указанное разделение данных генома на единицы доступа различных типов дополнительно включает в себя единицы доступа добавочного типа (320, 321), содержащие информацию, относящуюся к позиции вставок-удалений и рассогласований (321), к типу вставок-удалений и рассогласований (322) и информацию о мягко усеченных последовательностях нуклеотидов и о последовательностях нуклеотидов с жестким усечением (323).

6. Способ по любому из предшествующих пунктов, в котором разделение дополнительно включает в себя единицы доступа еще одного отличного от других типа, содержащие информацию, относящуюся к метаданным (432) и/или оценочным показателям качества и/или аннотационным данным (431), связанным с единицами доступа.

7. Способ по п.6, в котором разделение дополнительно включает в себя единицы доступа отличного от других добавочного типа, содержащие аннотационные данные.

8. Способ по п.7, котором данные единиц доступа по любому из предшествующих пунктов упорядочены размещением на уровнях, где каждый уровень содержит информацию, относящуюся к отличной от других категории: данные позиционирования, обратное замещение, по дополнительному выбору образование пар, по дополнительному выбору рассогласование и по дополнительному выбору аннотационные данные.

9. Способ по любому из предшествующих пунктов, в котором единицы доступа содержат заголовок и данные полезной нагрузки.

10. Способ по п.1, в котором единица доступа второго типа содержит информацию об образовании пар считанных фрагментов и в котором наличие такой информации помечено в заголовке единицы доступа.

11. Способ по любому из предшествующих пунктов, в котором количество считанных фрагментов, содержащихся в единице доступа первого типа, определяется параметром конфигурирования ввода.

12. Способ по п.11, в котором указанный параметр конфигурирования ввода записан на хранение в заголовок единицы доступа.

13. Способ по любому из предшествующих пунктов, в котором содержимое единиц доступа зашифровано.

14. Устройство для сжатия данных генома, содержащее средства, выполненные с возможностью реализации этапов способа по пп.1-13.

15. Запоминающее устройство для данных генома, содержащее данные генома, разделенные на единицы доступа и сжатые согласно способу по пп.1-13.

16. Считываемый компьютером записывающий носитель с записанной на нем программой, содержащей набор команд для реализации способа по пп.1-13.

17. Способ сжатия данных генома согласно способу по пп.1-13, в котором указанные данные скомпонованы для образования формата файла.

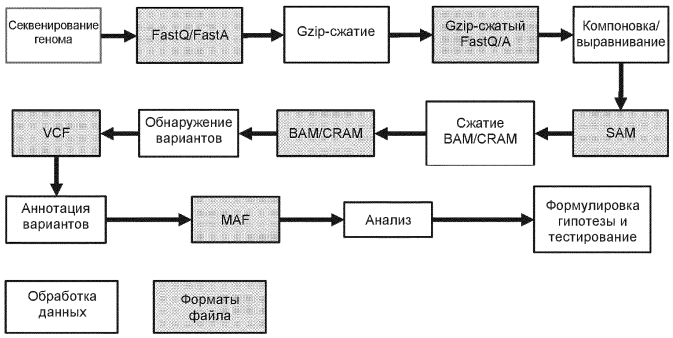
18. Способ транскодирования файла данных генома, сжатого согласно способу по пп.1-13, в котором при восстановлении выравнивания данных по отношению к новому референсному геному изменяют только информацию полезной нагрузки данных доступа.

19. Способ по п.18, в котором структура файла остается неизменной.

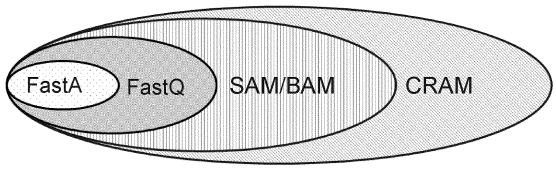
20. Способ по п.19, в котором изменяются только выбранные единицы доступа.

21. Способ по п.20, в котором выбранными единицами доступа являются единицы указанного первого типа (280).

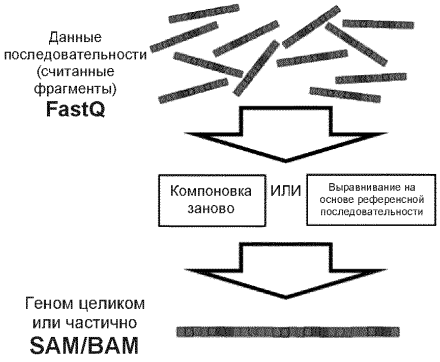
22. Способ по п.21, в котором выбранными единицами доступа являются единицы любого типа (290, 300, 310, 320, 321).



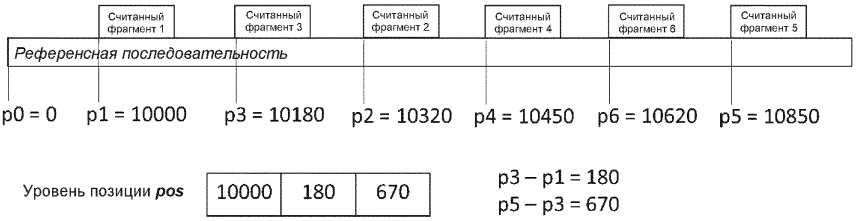
Фиг. 1
Типовой конвейер обработки генома и форматы файла



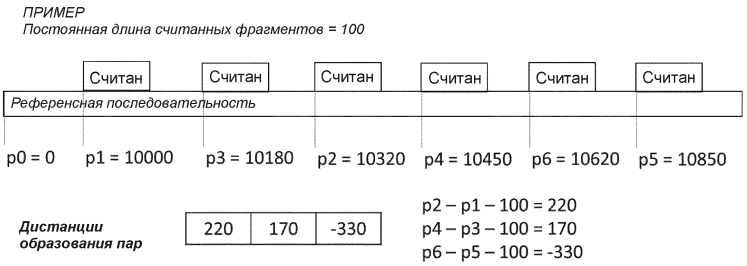
Фиг. 2
Форматы файла генома и их взаимосвязь



Фиг. 3
Выравнивание считанных фрагментов последовательности генома



Фиг. 4
Способ отображения и кодирования позиции первого считанного фрагмента из трех пар считанных фрагментов, подставленных и закодированных на уровне позиции (pos)

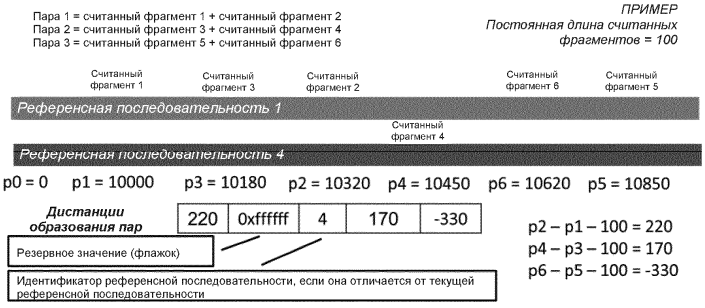


Фиг. 5

Вычисление расстояния образования пар для трех пар считанных фрагментов

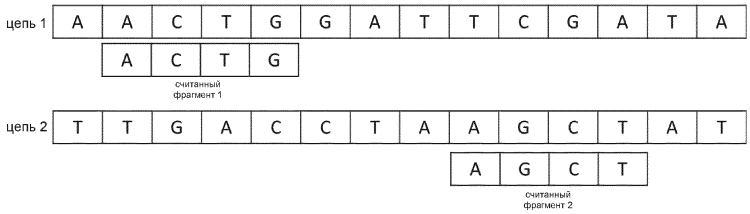
Фиг. 6

Вычисление ошибок образования пар



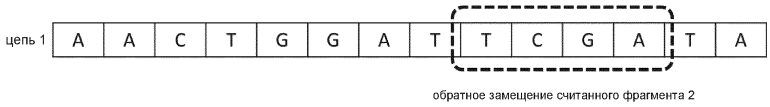
Фиг. 7

Если считанный фрагмент подставляется на другую референсную позицию, нежели второй элемент данной пары (считанный фрагмент 4), то в расстояние образования пары добавляются дополнительные дескрипторы. Одним из дескрипторов является флажок сигнализации, вторым - референсный идентификатор и далее расстояние образования пар



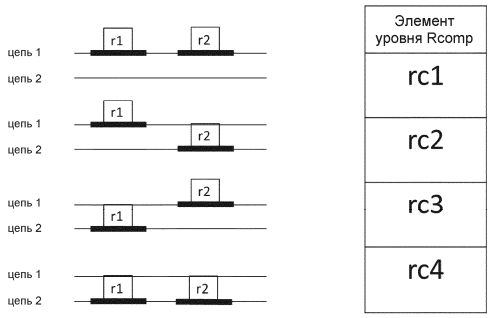
Фиг. 8

В данной паре считанных фрагментов считанный фрагмент 1 поступает от цепи 1 и считанный фрагмент 2 - от цепи 2



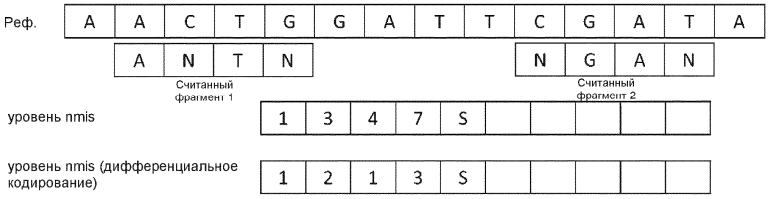
Фиг. 9

Обратное замещение считанного фрагмента 2 кодируется, если в качестве референсной цепи используется цепь 1



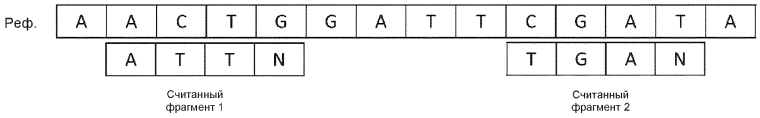
Фиг. 10

Четыре возможные комбинации считанных фрагментов, составляющих пары считанных фрагментов, и соответствующее кодирование на уровне rcomp



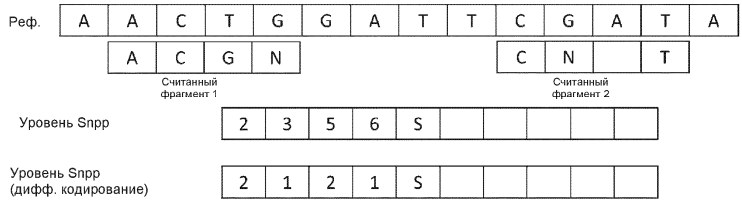
Фиг. 11

Вычисление N рассогласований в уровне nmis



Фиг. 12

Замены в паре считанных фрагментов с подстановкой



Фиг. 13

Вычисление позиций замены в виде абсолютных или дифференциальных выражений

уровень snpt (без кодов IUPAC)

Тип замены вычисляется в качестве индекса вектора замен, составленного всеми возможными символами Например:
S = [A, C, G, T, N, Z] ***где Z = удаление***
Направление индекса

- КОДИРОВАНИЕ выполняется справа налево
- ДЕКОДИРОВАНИЕ слева направо

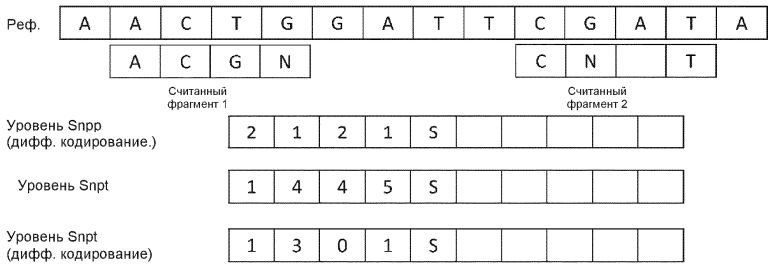
Реф.	Считанный фрагмент	Закодированный символ
A	C	idx(A,C) = 1
A	G	idx(A,G) = 2
A	T	idx(A,T) = 3
A	N	idx(A,N) = 4
C	A	idx(C,A) = 5
C	G	idx(C,G) = 1
C	T	idx(C/T) = 2
C	N	idx(C,N) = 3
G	A	idx(G,A) = 4
G	C	idx(G,C) = 5
G	T	idx(G,T) = 1
G	N	idx(G,N) = 2
T	A	idx(T,A) = 3
T	C	idx(T,C) = 4
T	G	idx(T,G) = 5
T	N	idx(T,N) = 1

Реф.	Считанный фрагмент	Закодированный символ
A	удал.	idx(A,Z) = 5
C	удал.	idx(C,Z) = 4
G	удал.	idx(G,Z) = 3
T	удал.	idx(T,Z) = 2

Реф.	Считанный фрагмент	Закодированный символ
N	A	idx(N,A)=2
N	C	idx(N,C) = 3
N	G	idx(N,G) = 4
N	T	idx(N,T)=5

Фиг. 14

Вычисление замен при кодировании символов без кодов IUPAC



Фиг. 15

Кодирование замен внутри уровня snpt

уровень snpt (без кодов IUPAC)

Тип замены вычисляется в качестве индекса в векторе замен, составляемом всеми возможными символами. Например:

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

Направление индекса

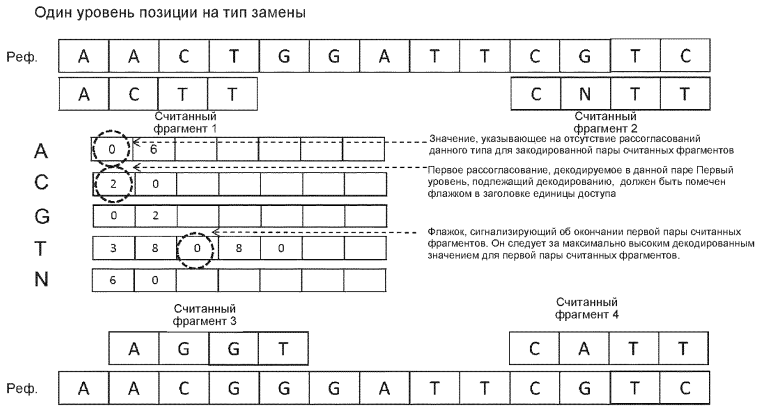
- КОДИРОВАНИЕ выполняется справа налево
- ДЕКОДИРОВАНИЕ слева направо

Реф.	Считанный фрагмент	Закодированный символ
N	M	idx(N,M) = 2
N	W	idx(N,W) = 4
N	S	idx(N,S) = 5
N	B	idx(N,B) = 11

Реф.	Считанный фрагмент	Закодированный символ
D	M	idx(D,M) = 8
A	Y	idx(A,Y) = 10
A	T	idx(A,T) = 3
A	N	idx(A,N) = 4
C	R	idx(C,R) = 6
C	G	idx(C,G) = 1
C	T	idx(C,T) = 2
C	W	idx(C,W) = 7
G	H	idx(G,H)=II
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

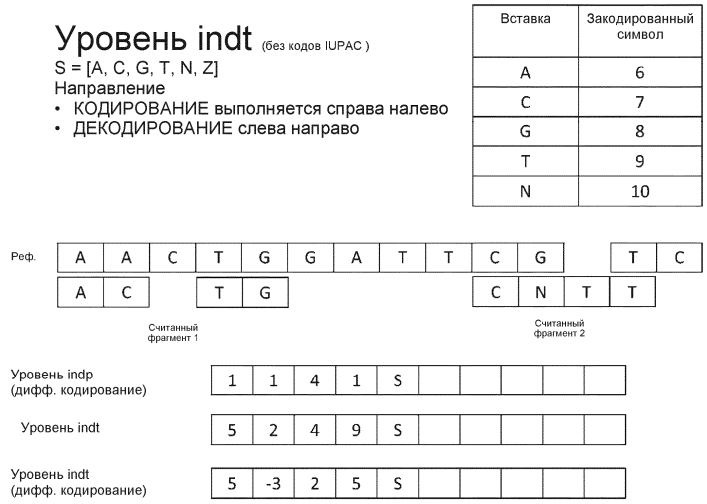
Фиг. 16

Вычисление замен при кодировании символов с использованием кодов IUPAC



Фиг. 17

Альтернативная модель источника для замен, когда кодируются только позиции, но используется один уровень на тип замены



Фиг. 18

Кодирование замен, вставок и удалений в паре считанных фрагментов класса I, когда не используются коды IUPAC

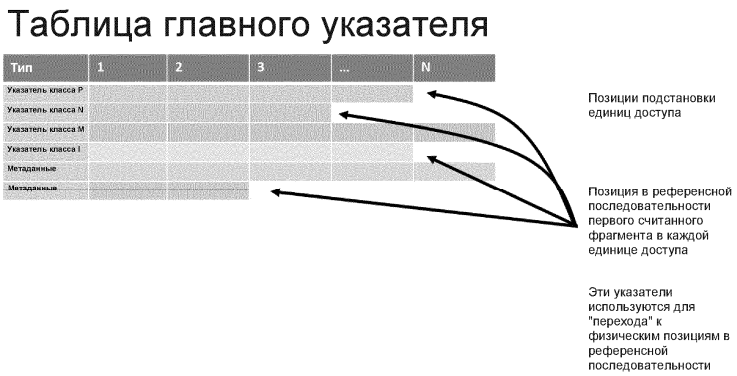
Фиг. 19

Кодирование замен, вставок и удалений в паре считанных фрагментов класса I при использовании кодов IUPAC

Уникальный ИД	Версия	Размер заголовка	Реф. счет	N блоков	Реф. идентификаторы	Таблица главного указателя (M.I.T)

Фиг. 20

Заголовок закодированной информации по геному

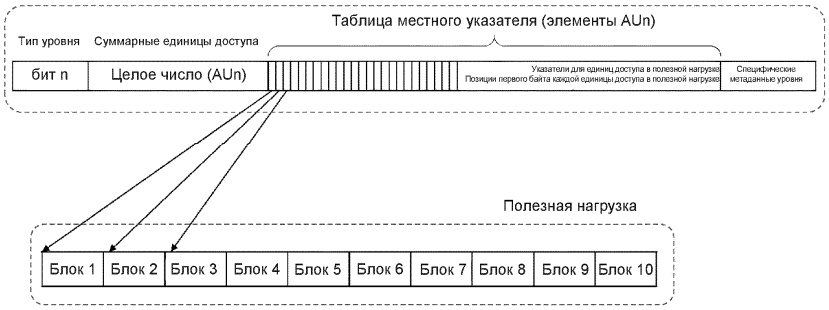


Фиг. 21

В таблице главного указателя обозначены позиции в референсных последовательностях первого считанного фрагмента в каждой единице доступа

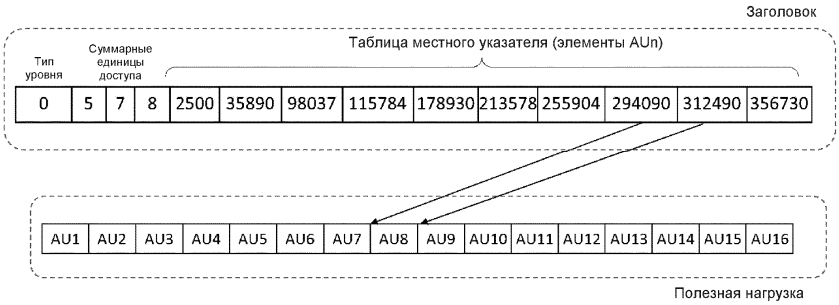
Фиг. 22

Пример частичной таблицы главного указателя (MIT) с отображением позиций подстановки первого считанного фрагмента на каждой позиции единицы доступа класса P



Фиг. 23

Таблица местного указателя в заголовке уровня представляет собой вектор указателей к единицам доступа в полезной нагрузке

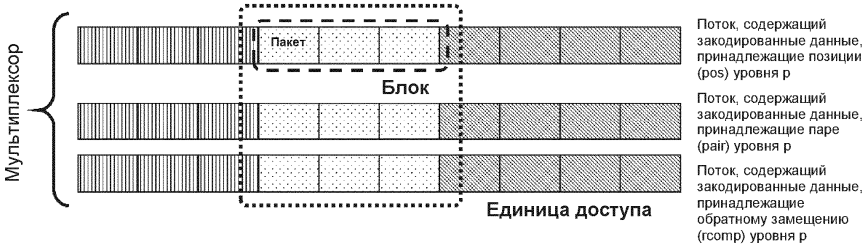


Фиг. 24

Пример таблицы местного указателя (LIT)

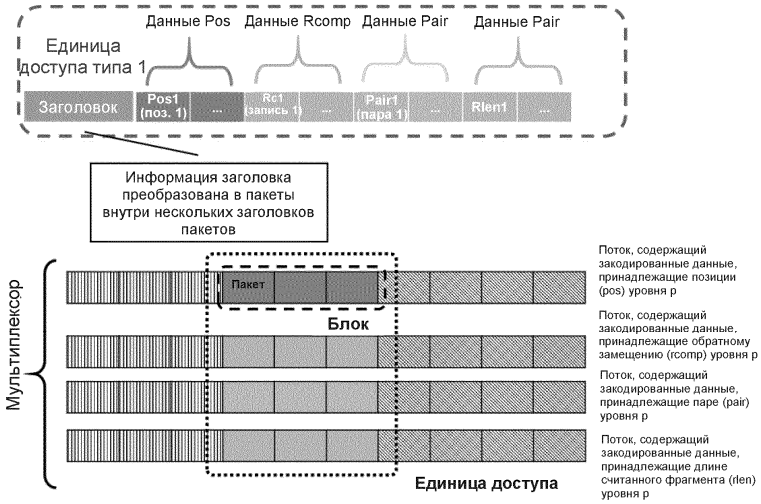
Фиг. 25

Функциональная взаимосвязь между таблицей главного указателя и таблицами местных указателей



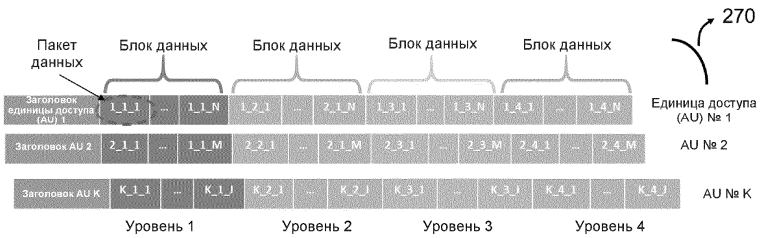
Фиг. 26

Единицы доступа образуются блоками данных, принадлежащими нескольким уровням и индексированным с использованием механизмов MIT и LIT



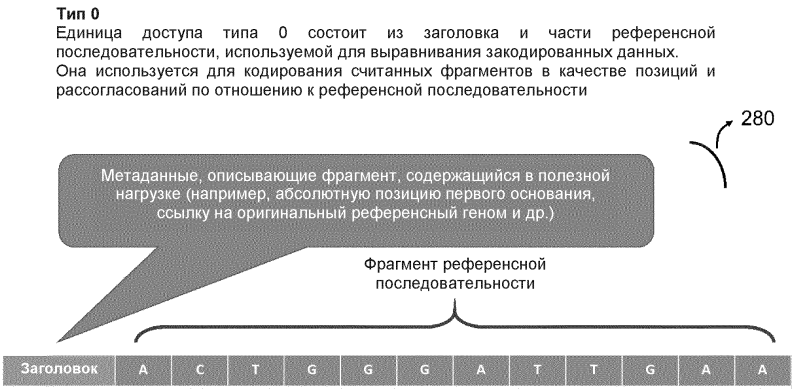
Фиг. 27

Единица доступа типа 1 преобразована в пакет и транспортирована в уплотненном формате данных генома



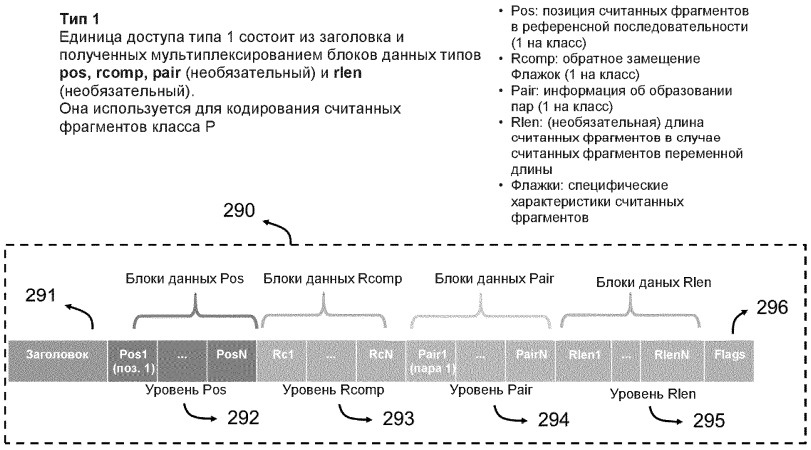
Фиг. 28

Единицы доступа образуются заголовками и уплотненными блоками, принадлежащими одному или более уровням однородных данных. Каждый блок может состоять из одного или более пакетов, содержащих действующие дескрипторы информации о геноме



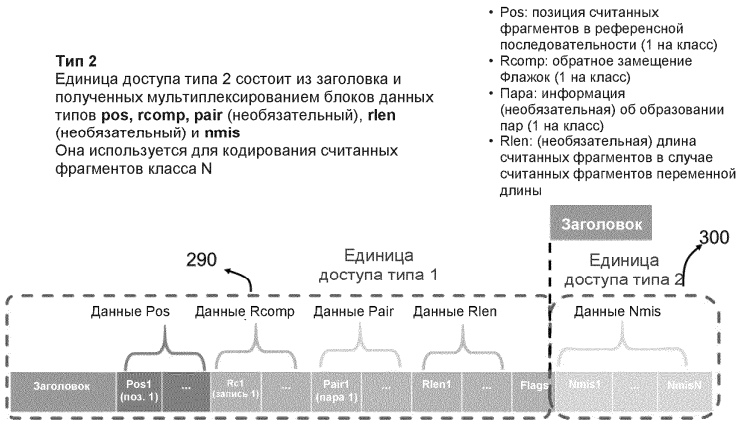
Фиг. 29

Единицам доступа типа D не требуется ссылка на любую информацию, поступающую от других единиц доступа, предназначенных для доступа или декодирования и доступа



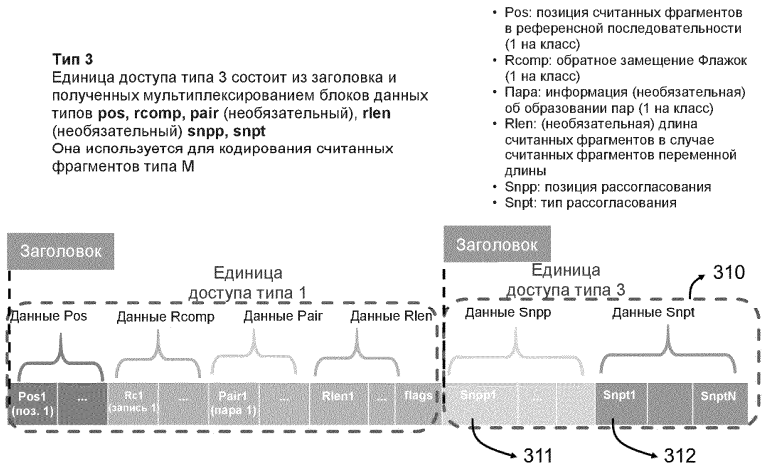
Фиг. 30

Единица доступа типа 1 содержит данные, соотносимые с данными, переносимыми единицами доступа типа 0



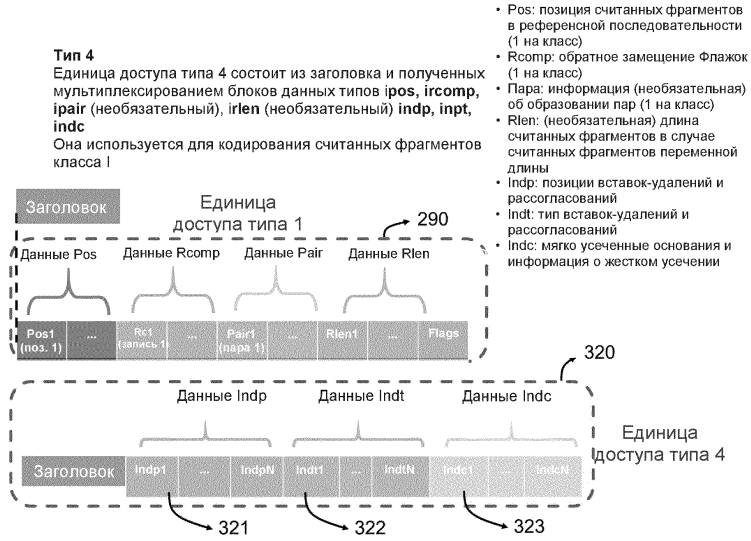
Фиг. 31

Единицы доступа типа 2 содержат данные, соотносимые с единицей доступа типа 1. Имеются позиции N в закодированных считанных фрагментах



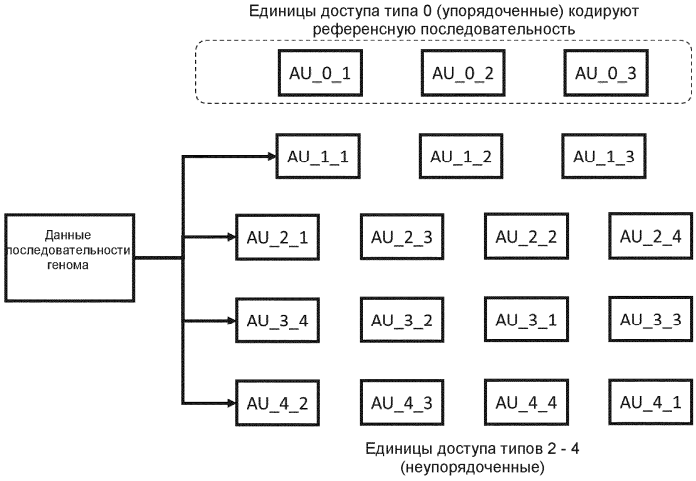
Фиг. 32

Единицы доступа типа 3 содержат данные, соотносимые с единицей доступа типа 1. В закодированных считанных фрагментах указываются позиции и типы рассогласования



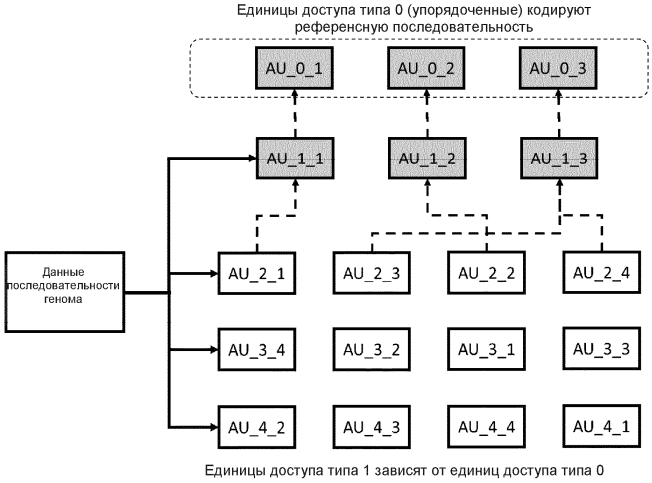
Фиг. 33

Единицы доступа типа 4 содержат данные, соотносимые с единицей доступа типа 1. В закодированных считанных фрагментах указываются позиции и типы рассогласования



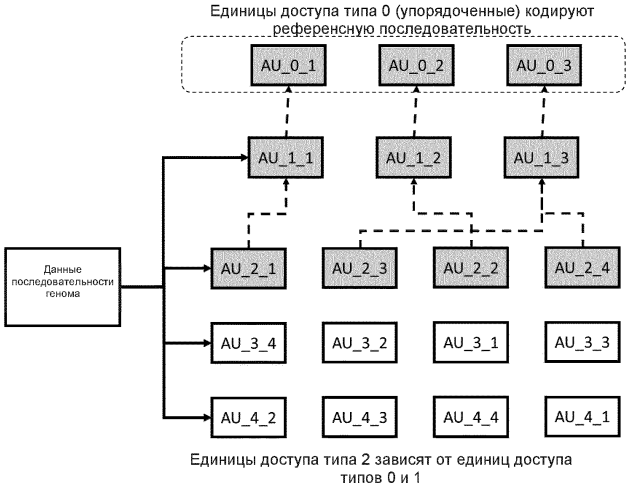
Фиг. 34

Первые пять типов единиц доступа



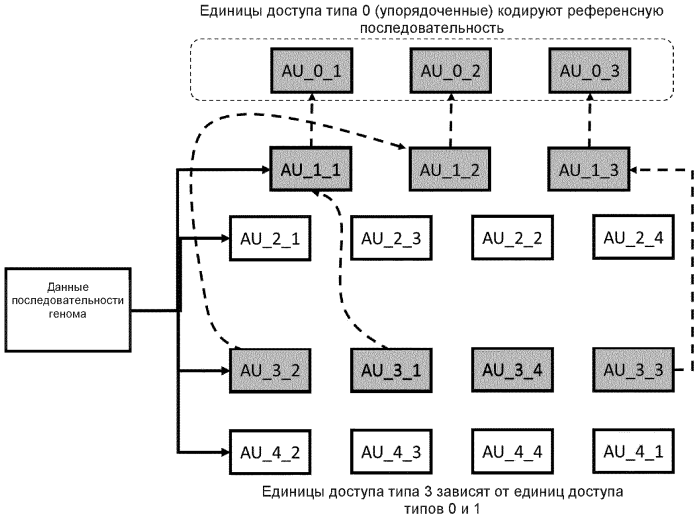
Фиг. 35

Единицы доступа типа 1 соотносятся с требующими декодирования единицами доступа типа 0



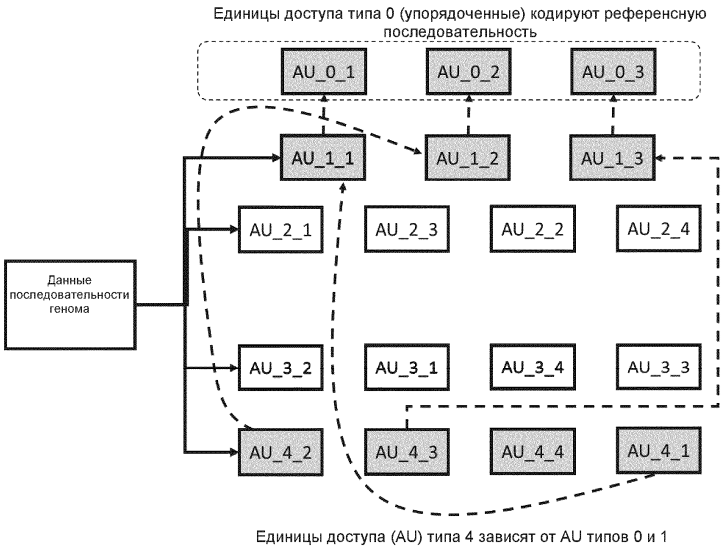
Фиг. 36

Единицы доступа типа 2 соотносятся с требующими декодирования единицами доступа типа 0 и 1



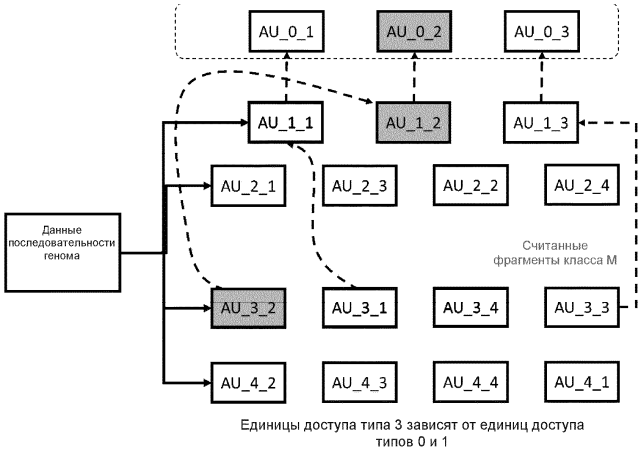
Фиг. 37

Единицы доступа типа 3 соотносятся с требующими декодирования единицами доступа типа 0 и 1



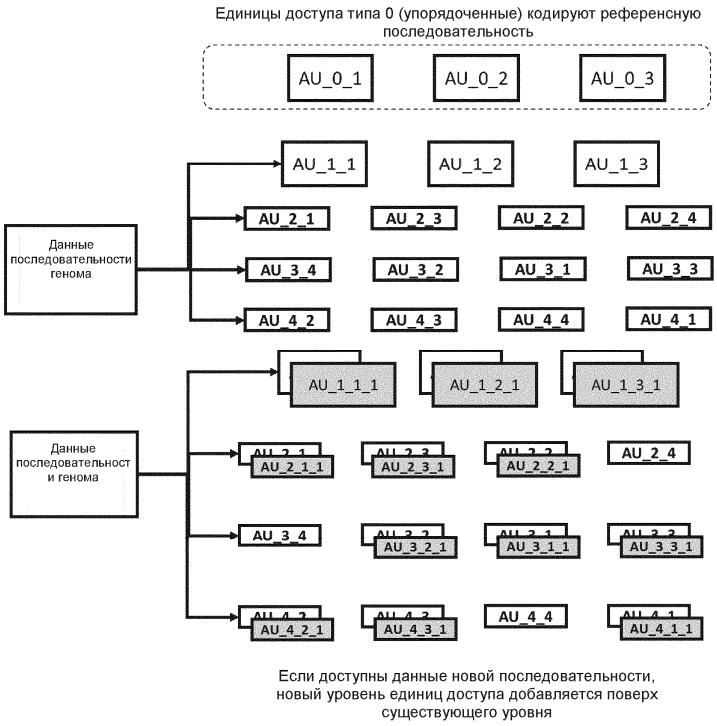
Фиг. 38

Единицы доступа типа 4 соотносятся с единицами доступа типа 1 для кодирования считанных фрагментов класса I



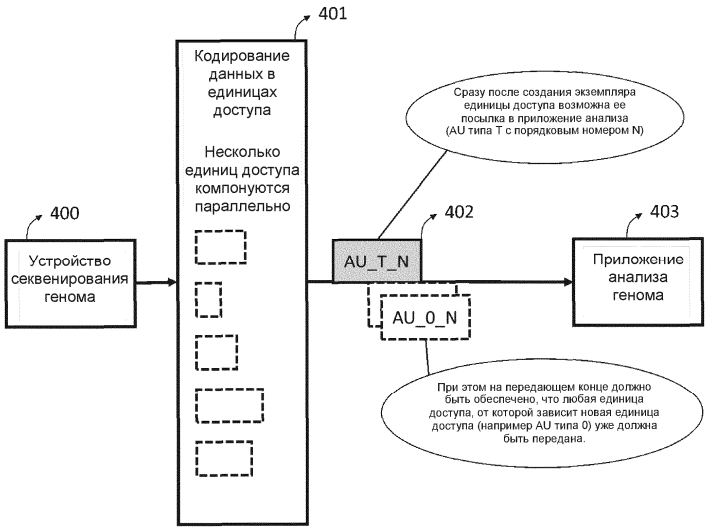
Фиг. 39

Единицы доступа, требуемые для декодирования считанных выборок последовательности с рассогласованиями, подставленными во второй сегмент референсной последовательности (AU 0-2)



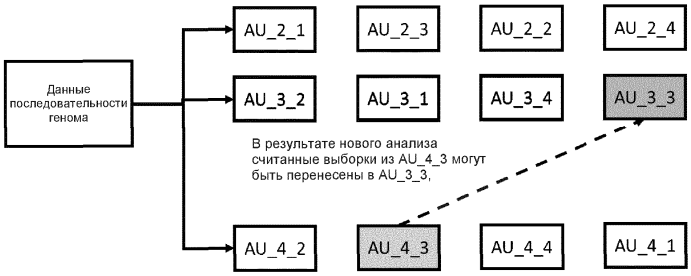
Фиг. 40

Когда доступны новые данные, поверх существующих данных создается новый уровень единиц доступа



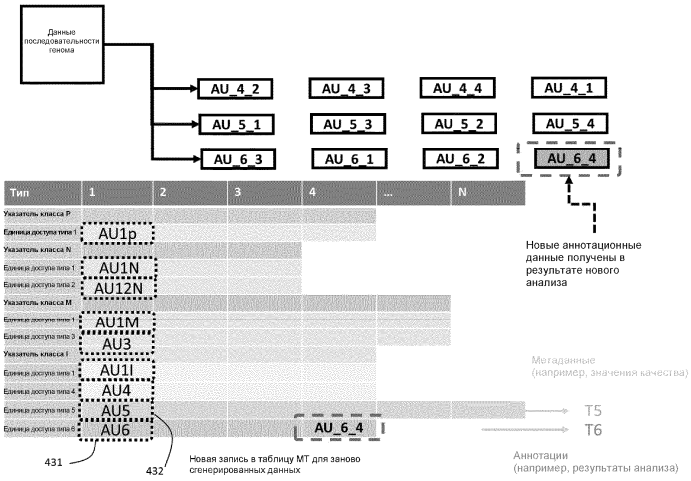
Фиг. 41

Структура данных на основе единиц доступа обеспечивает возможность анализа данных генома перед завершением процесса секвенирования



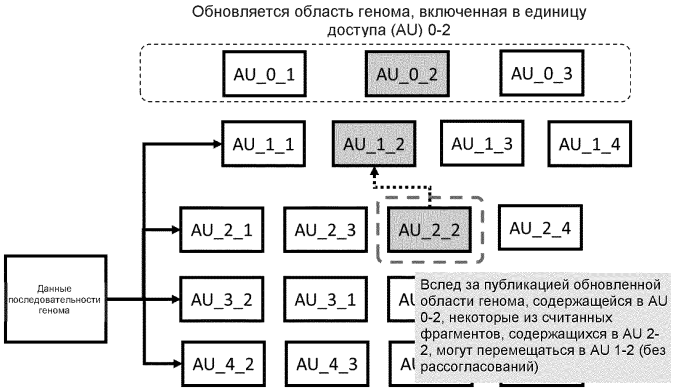
Фиг. 42

Для нового анализа, выполняемого на существующих данных, может понадобиться перемещение считанных фрагментов из единиц доступа типа 4 в единицу доступа типа 3



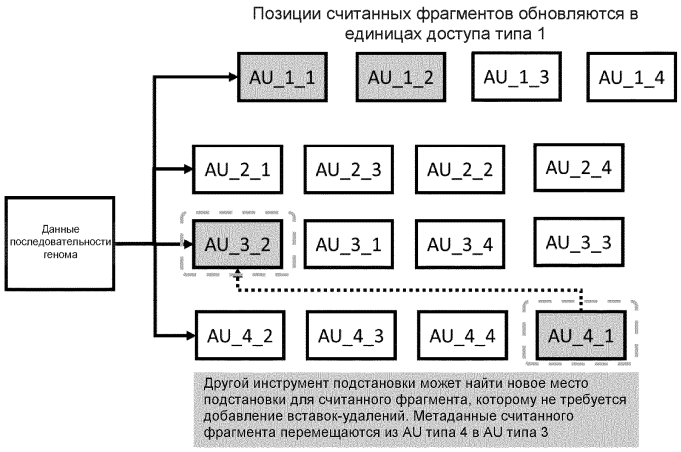
Фиг. 43

Заново полученные результаты анализа инкапсулируются в новую единицу доступа типа 6 и соответствующий индекс создается в таблице главного указателя



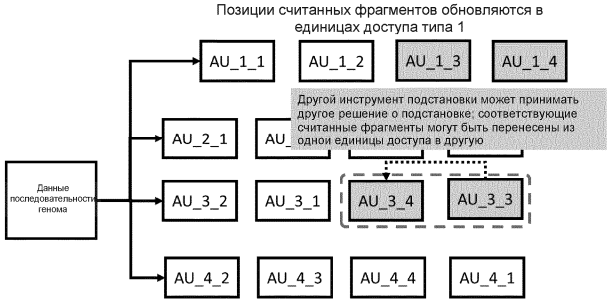
Фиг. 44

Транскодирование вследствие публикации новой референсной последовательности (генома)



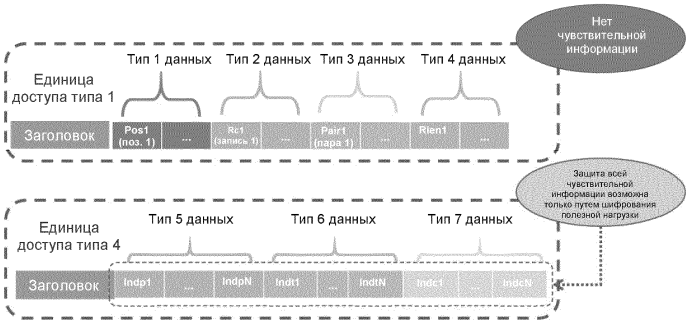
Фиг. 45

Считанные фрагменты, подставленные в новую область генома с повышенным качеством (например, без вставок-удалений), перемещаются из единицы доступа типа 4 в единицу доступа типа 3



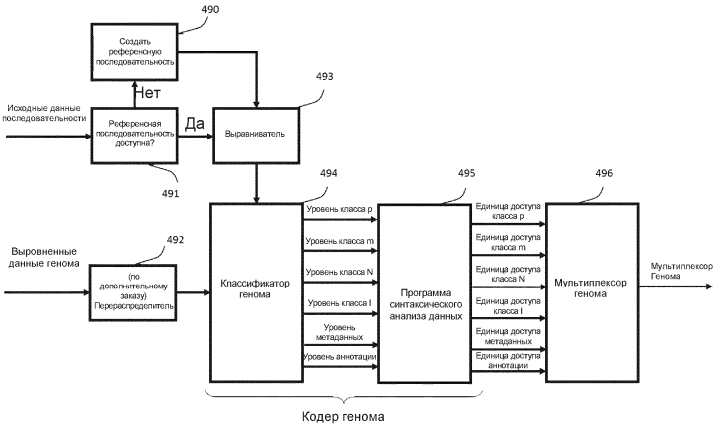
Фиг. 46

В случае обнаружения нового места подстановки (например, с меньшим количеством рассогласований) соответствующие считанные фрагменты могут быть перенесены из одной единицы доступа в другую того же типа



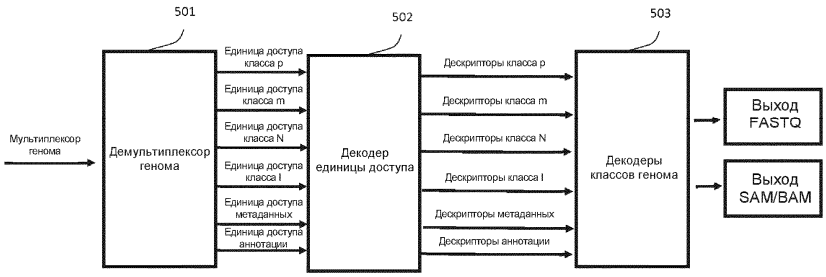
Фиг. 47

Избирательное шифрование может быть применено к единицам доступа типа 4 только при наличии в них чувствительной информации, требующей защиты



Фиг. 48

Кодер генома представляет собой первую ступень классификации данных и вторую ступень синтаксического анализа и кодирования данных. Этому может предшествовать выравнивание на референсной последовательности, сгенерированной внешними или внутренними средствами. Кодер генома может использовать выравниватели и вновь созданные компонующие программы для подготовки данных к кодированию



Фиг. 49

Демультиплексер генома извлекает содержимое уровней единиц доступа из уплотненной структуры генома, декодер (по одному декодеру на тип единицы доступа) извлекает дескрипторы генома, которые далее декодируются в различные форматы генома, такие как FASTQ и SAM/BAM