РОССИЙСКАЯ ФЕДЕРАЦИЯ

(19) **RU** (11) **2 804 029**(13) **C9**

(51) МПК
G16B 50/30 (2019.01)
G16B 50/00 (2019.01)

ФЕДЕРАЛЬНАЯ СЛУЖБА
ПО ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ

(12) СКОРРЕКТИРОВАННОЕ ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ПАТЕНТУ

Примечание: библиография отражает состояние при переиздании

(52) СПК
G16B 50/30 (2023.05); G16B 50/00 (2023.05)

(21)(22) Заявка: **2021134292, 11.01.2017**

(24) Дата начала отсчета срока действия патента:
11.01.2017

Приоритет(ы):
(30) Конвенционный приоритет:
11.01.2016 US 62/277,445
Номер и дата приоритета первоначальной заявки, из которой данная заявка выделена:
2018120941 11.01.2016

(43) Дата публикации заявки: **01.04.2022** Бюл. № **10**

(45) Опубликовано: **26.09.2023**

(15) Информация о коррекции:
Версия коррекции №1 (W1 C2)

(48) Коррекция опубликована:
19.12.2023 Бюл. № 35

Адрес для переписки:
190900, Санкт-Петербург, BOX 1125, Нилова Мария Иннокентьевна

(72) Автор(ы):
ВАН РОЙН, Питер (US),
МАКМИЛЛЕН, Роберт Дж. (US),
РЮЛЕ, Майкл (US),
МЕХЬО, Рами (US)

(73) Патентообладатель(и):
ЭДИКО ГЕНОМ, КОРП. (US)

(56) Список документов, цитированных в отчете о поиске: **US 20130275486 A1, 17.10.2013. WO 2014113736 A1, 24.07.2014 . MILLER N.A. et al. "A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases." Genome medicine, 2015, 7(1): 1-16. RU 2015144109 A, 21.04.2017.**

(54) ГЕНОМНАЯ ИНФРАСТРУКТУРА ДЛЯ ЛОКАЛЬНОЙ И ОБЛАЧНОЙ ОБРАБОТКИ И АНАЛИЗА ДНК И РНК

(57) Реферат:

Группа изобретений относится к биоинформатике, более конкретно к системам, устройствам и способам для реализации биоинформатических протоколов для осуществления одной или более функций для анализа геномных данных на интегральной схеме. Способ, энергонезависимый машиночитаемый носитель, система и платформа для анализа геномных данных предусматривают выполнение одним или более компьютерами операций, включающих: получение посредством программного интерфейса приложения (API) первых данных, идентифицирующих конкретный конвейер обработки геномных данных, вводимых посредством графического пользовательского интерфейса; конфигурирование с использованием API схем программируемого логического устройства для исполнения конвейера обработки геномных данных на основе указанных первых данных; получение вторых данных, представляющих набор геномных данных или набор данных, полученных из геномных данных; использование одним или более компьютерами указанного конвейера обработки геномных данных, сконфигурированного в схемах программируемого логического устройства, на основе первых данных, для обработки полученных вторых данных; получение

RU 2804029 C9

RU 2804029 C9

результирующих данных и предоставление на их основе выходных данных. Группа изобретений обеспечивает ускорение обработки данных при большей процентной точности. 4 н. и 18 з.п. ф-лы, 21 ил., 2 табл.

RU 2804029 C9

RU 2804029 C9

RUSSIAN FEDERATION

FEDERAL SERVICE
FOR INTELLECTUAL PROPERTY

(19) **RU** (11) **2 804 029**(13) **C9**

(51) Int. Cl.
G16B 50/30 (2019.01)
G16B 50/00 (2019.01)

(12) ABSTRACT OF INVENTION

Note: Bibliography reflects the latest situation

(52) CPC
G16B 50/30 (2023.05); G16B 50/00 (2023.05)

(21)(22) Application: **2021134292, 11.01.2017**

(24) Effective date for property rights:
11.01.2017

Priority:
(30) Convention priority:
11.01.2016 US 62/277,445
Number and date of priority of the initial application, from which the given application is allocated:
2018120941 11.01.2016

(43) Application published: **01.04.2022** Bull. № **10**

(45) Date of publication: **26.09.2023**

(15) Correction information:
Corrected version no1 (W1 C2)

(48) Corrigendum issued on:
19.12.2023 Bull. № 35

Mail address:
190900, Sankt-Peterburg, BOX 1125, Nilova Mariya Innokentevna

(72) Inventor(s):
VAN ROJN, Piter (US),
MAKMILLEN, Robert Dzh. (US),
RYULE, Majkl (US),
MEKHO, Rami (US)

(73) Proprietor(s):
EDIKO GENOM, KORP. (US)

RU 2804029 C9

RU 2804029 C9

(54) GENOMIC INFRASTRUCTURE FOR LOCAL AND CLOUD PROCESSING AND ANALYSIS OF DNA AND RNA

(57) Abstract:

FIELD: bioinformatics.

SUBSTANCE: systems, devices and methods for implementing bioinformatic protocols for performing one or more functions for analysing genome data in an integrated circuit. A method, a non-volatile computer-readable medium, a system, and a platform for analysing genome data involves one or more computers performing operations including: receiving, via an application programming interface (API), the first data identifying a particular genome data processing pipeline entered through a graphical user interface; configuring, using the Programmable Logic Device API circuits, to execute the genome data processing pipeline based on the said first data; obtaining the second data representing a genome data set or a data set derived from the genome data; one or more computers using the said genomic data processing pipeline configured in the programmable logic circuits based on the first data to process the received second data; obtaining the resulting data and providing output data based on them.

EFFECT: faster data processing with greater percentage accuracy.

22 cl, 21 dwg, 2 tbl

ПЕРЕКРЕСТНАЯ ССЫЛКА НЕ РОДСТВЕННУЮ ЗАЯВКУ

[0001] Настоящая заявка испрашивает приоритет на основании предварительной заявки на патент США № 62/277,445, поданной 11 января 2016 г., содержание которой полностью включено в настоящий текст посредством ссылки.

ОБЛАСТЬ ТЕХНИКИ

[0002] Описанный в настоящем документе объект относится к биоинформатике, и, более конкретно, к системам, устройствам и способам для реализации биоинформатических протоколов, например, для осуществления одной или более функций для анализа геномных данных на интегральной схеме, например, на аппаратной процессорной платформе.

УРОВЕНЬ ТЕХНИКИ

[0003] Перед исследователями и практикующими специалистами в области здравоохранения стоить задача повышения безопасности, качества и эффективности медицинских услуг для каждого пациента. Персонализированная медицина направлена на решение этих задач на индивидуальном уровне. Например “геномика” и/или “биоинформатика” представляют собой области, задачей которых является обеспечение безопасности, качества и эффективности профилактических и терапевтических средств и способов лечения на персонифицированном, индивидуальном уровне. Соответственно, применением методик геномики и/или биоинформатики позволяет определить генетические характеристики индивидуума, например, состав ее или его генов, а эту информацию можно использовать для разработки лекарственных средств и/или профилактических схем, включая способы лекарственного лечения, персонализированные для этого индивидуума, что позволяет адаптировать медицинские услуги под потребности каждого отдельного лица.

[0004] Стремление предоставлять персонализированные медицинские услуги трансформирует систему здравоохранения. Эта трансформация системы здравоохранения по-видимому основана на прорывных инновациях на пересечении медицинской науки и информационных технологий, примером которых являются области геномики и биоинформатики. Соответственно, геномика и биоинформтика составляют фундамент, на котором будет построено это будущее. С того момента как в 2000г с общими затратами более 1 млрд. долларов США был впервые секвенирован полный геном человека, наука существенно шагнула вперед. Сегодня мы стоим на пороге секвенирования с высоким разрешением при стоимости менее 1000 долларов США за генома, что делает выход за пределы исследовательской лаборатории в широкую область медицинской помощи реальным с эколномической точки зрения. Соответственно, геномные данные могут внести важнейший вклад в диагностический скрининг, разработку профилактических и/или терапевтических лекарственных средств и/или лечение болезней.

[0005] Более конкретно, геномики и биоинформатика представляют собой области, применяющие информационные технологии и информатику в области молекулярной биологии. В частности, биоинформатические методы можно применять для обработки и анализа разнообразных геномных данных, таких как данные о геноме индивидуального субъекта, для получения качественной и количественной информации об этих данных, которую специалисты в разных областях смогут затем применять для разработки методов профилактики и терапии для предотвращения или по меньшей мере облегчения болезненных состояний, и, таким образом, улучшения безопасности, качества и эффективности медицинской помощи на уровне индивидуальных субъектов.

[0006] Благодаря тому, что биоинформатика направлена на обеспечения

персонализированных медицинских услуг, она стимулирует развитие индивидуализированного здравоохранения, которое является проактивным, а не реактивным, что дает пациенту возможность принимать более активное участие в повышении качества своей жизни. Во-первых, можно установить федеральное управление для поддержки исследований, направленное на эти индивидуальные аспекты заболеваний и предотвращения заболеваний; основной задачей может быть подбор диагностики и превентивных мероприятий, подходящий для уникальных генетических параметров каждого индивидуального субъекта. Дополнительно, можно создать “сеть сетей” для сбора данных, относящихся к здравоохранению, чтобы помочь исследователям в определении паттернов и идентификации генетических “картин” существующих заболеваний.

[0007] Преимущество применения биоинформатических технологий в этих ситуациях заключается в том, что они позволяют осуществлять качественный и/или количественный анализ молекулярно-биологических данных на более широких множествах образцов с значительно более высокой скоростью и зачастую большей точностью, что способствует быстрому развитию системы персонализированного здравоохранения.

[0008] Соответственно, в различных вариантах молекулярные данные для обработки на платформах, работающих на основе биоинформатических методов, обычно связаны с геномными данными, такими как данные о дезоксирибонуклеиновых кислотах (ДНК) и/или рибонуклеиновых кислотах (РНК). Например, хорошо известный способ генерации данных о ДНК и/или РНК включает ДНК/РНК-секвенирование. ДНК/РНК-секвенирование может осуществляться вручную, например, в лаборатории, или может осуществляться при помощи автоматического секвенатора, например, в профильном центре секвенирования, с целью определения генетического состава образца генетического материала субъекта, например, ДНК и/или РНК. Генетическую информацию о субъекте можно затем использовать для сравнения с эталоном, таким как референсная последовательность, гаплотип или теоретический гаплотип, чтобы таким образом определить отличия от него. Такую информацию об отличиях можно затем использовать для определения или предсказания возникновения болезненного состояния у субъекта.

[0009] Например, ручное или автоматизированное секвенирование ДНК/РНК можно применять для определения последовательности нуклеотидных оснований в образце ДНК/РНК, таком как образец, взятый у субъекта. Применение различных биоинформатических процедур позволяет затем объединить эти последовательности вместе и получить геномную последовательность субъекта. Затем эту последовательность можно сравнить с референсной геномной последовательностью, чтобы определить отличия геномное последовательности субъекта от эталона. Этот процесс включает определение отличий (вариантов) в последовательности образца и представляет основную проблему биоинформатических методик.

[0010] Например, основной проблемой в секвенировании ДНК является построение полноразмерных геномных последовательностей, например, хромосомных последовательностей, из образца генетического материала, которые можно было бы сравнить с референсной геномной последовательностью для определения отличий в полноразмерной геномной последовательности из образца. В частности, методы, применяемые в протоколах секвенирования, не выдают полноразмерные последовательности хромосом, содержащихся в образце ДНК.

[0011] Вместо этого получают фрагменты последовательностей, которые обычно

имеют длину 100-1000 нуклеотидов, без каких-либо указаний на то, какой области генома они соответствуют. Таким образом, чтобы получить полноразмерные геномные хромосомные конструкты, эти фрагменты последовательностей ДНК необходимо картировать, выровнять, соединить и/или сравнить с референсной геномной последовательностью. Такие процессы позволяют определить отличия геномных последовательностей из образцов от референсных геномных последовательностей.

[0012] Однако геном человека состоит из приблизительно 3,1 миллиардов пар оснований, а каждый фрагмент последовательности обычно имеет длину лишь от 100 до 500 нуклеотидов, поэтому время и ресурсы, уходящие на построение полноразмерных геномных последовательностей и определения отличий в них, весьма велики, что частно обуславливает необходимость применения нескольких различных вычислительных ресурсов, использующих несколько различных алгоритмов, в течение длительного времени.

[0013] В одном из конкретных примеров от тысяч до миллионов фрагментов последовательностей ДНК генерируют, выравнивают и объединяют для получения геномной последовательности, которая приближается по длине к хромосоме. Один из этапов этого процесса может включать сравнение фрагментов ДНК с референсной последовательностью для определения области генома, с которой она выравнивается (т.е. которой она соответствует).

[0014] В построении последовательностей, имеющих длину хромосомы, и определении вариантов (т.е. отличий от референсной последовательности) последовательности образца участвует несколько таких этапов. Соответственно, для реализации этих этапов разработано большое количество разнообразных методов. Например, существуют широко применяемое программное обеспечение, реализующее один или ряд этапов биоинформатической системы. Однако общей чертой таких биоинформатических способов и систем на основе программного обеспечения является то, что они требуют больших трудозатрат, их выполнение на обычных процессорах занимает длительное время, и они имеют тенденцию к ошибкам.

[0015] Соответственно, была бы полезна биоинформатическая система, которая могла бы осуществлять алгоритмы, реализуемые в таком программном обеспечении с меньшими трудозатратами и/или способом, требующих менее интенсивной обработки, при большей процентной точности. Однако даже несмотря на то, что мы приблизились к стоимости анализа "Геном за 1000$", хранение и обеспечение коллективного доступа к этим необработанным цифровым данным, значительно опережают стоимость их получения. Это узкое место анализа является основным препятствием на пути от этих необработанных данных, количество которых постоянно растет, к реальным медицинским знаниям, которые мы стремимся из них получить.

[0016] Соответственно, в настоящем документе представлены системы, устройства и способы для применения протоколов геномики и/или биоинформатики, например, для реализации одной или более функций для анализа геномных данных, например, при помощи программных реализаций и/или на интегральной схеме, например, на аппаратной платформе для обработки. Например, как описано ниже в настоящем документе, в различных вариантах реализации комбинации программно-реализуемых и/или решений аппаратного ускорения, таких как решения, включающие интегральные схемы и программное обеспечение для взаимодействия с ними, можно применять для решения задач, связанных с биоинформатикой, при этом интегральные схемы могут состоять из одной или более аппаратно соединенных (смонтированных) цифровых логических схем, которые могут быть связаны множеством физических электрических

соединений, которые могут быть организованы в виде набора модулей обработки, причем каждый модуль обработки может быть сконфигурирован для выполнения одного или более этапов биоинформатического протокола генетического анализа. Преимущество такой организации заключается в том, что эти задачи, связанные с биоинформатикой, можно осуществлять быстрее, чем при применении только программного обеспечения, как это обычно бывает при решении таких задач. Тем не менее, применение такой методики аппаратного ускорения в настоящее время обычно не применяется в области геномики и/или биоинформатики.

КРАТКОЕ ОПИСАНИЕ

[0017] Настоящее изобретение относится к выполнению таких задач, как задачи, являющиеся частью биоинформатических протоколов. В различных вариантах выполняется множество задач, а в некоторых случаях эти задачи выполняют таким образом, что они образуют конвейер, причем каждая задача и/или ее почти полное выполнение действует как строительный блок для каждой последующей задачи до достижения желаемого конечного результата. Соответственно, в различных вариантах реализации настоящее изобретение относится к реализации одного или более способов на одном или большем числе устройств, где указанное устройство оптимизировано для реализации этих способов. В некоторых вариантах реализации указанные один или более способов и/или одно или более устройство скомпонованы в одну или более систем.

[0018] Например, в некоторых аспектах настоящее изобретение относится к системам, устройствам и способам для реализации протоколов геномики и/или биоинформатики, таким как, в различных вариантах, для осуществления одной или более функций для получения и/или анализа генетических данных с применением инновационного программного обеспечения и/или на интегральной схеме, например, реализованных в комбинации программного обеспечения и/или аппаратной платформы для обработки. Например, в одном аспекте предложена геномная и/или биоинформатическая система. Эта система может включать выполнение различных функций для получения биоаналитических данных и/или анализа, которые оптимизированы для более быстрого осуществления и/или повышения точности. Способы осуществления этих функций могут быть реализованы в программных или аппаратных решениях. Соответственно, в некоторых вариантах представлены способы, включающие получение и/или сбор данных и/или анализ, которые могут включать реализацию один или более алгоритмов, причем указанные алгоритм(ы) оптимизированы в соответствии с, например, например, программным обеспечением, аппаратным обеспечением, или комбинацией аппаратного и программного обеспечения, в которых они реализуются. В частности, в тех случаях, когда алгоритм реализуется в программном решении, алгоритм и/или составляющие его процессы могут быть оптимизированы для более быстрого выполнения и/или выполнения с повышенной точностью при реализации с применением конкретного средства. Аналогично, в тех случаях, когда алгоритм реализуется в аппаратном решении, аппаратное обеспечение разрабатывается специально для осуществления этих функций и/или составляющих их процессов с оптимизацией для более быстрого выполнения и/или выполнения с повышенной точностью при реализации с применением конкретного средства. Далее, в тех случаях, когда функции включают комбинацию программных и/или аппаратных решений, эти функции и составляющие их процессы разработаны и сконфигурированы для оптимальной совместной работы, что позволяет достичь скорости, недоступной ранее, при той же или повышенной точности.

[0019] Соответственно, в одном аспекте настоящего изобретения предложены системы, устройства и способы для реализации биоинформатических протоколов,

например, для выполнения одной или более функций для генерации и/или анализа генетических данных, например, с применением одного или более разработанных и/или оптимизированных алгоритмов и/или на одной или более оптимизированных интегральных схемах, например, на одной или более аппаратных платформ обработки. Соответственно, в одном варианте предложены способы для реализации одного или более алгоритмов для осуществления одного или более этапов для генерации и/или анализа геномных данных в протоколе геномики и/или биоинформатики. В другом случае предложены способы реализации функций одного или более алгоритмов для осуществления одного или более этапов для анализа геномных данных в биоинформатическом протоколе, причем эти функции по меньшей мере частично реализованы на интегральной схеме, такой как схема, состоящая из одной или более аппаратно соединенных цифровых логических схем. В таком варианте аппаратно соединенные цифровые логические схемы могут быть связаны, например, посредством множества физических электрических соединений, и могут быть организованы для работы в виде одного или более модулей обработки. В различных вариантах предложено множество аппаратно соединенных цифровых логических схем, причем аппаратно соединенные цифровые логические схемы сконфигурированы в виде набора модулей обработки, причем каждый модуль обработки может выполнять один или более этапов биоинформатических протоколов генетического анализа, как в случае конвейера биоинформатической обработки.

[0020] Более конкретно, в одном варианте предложена система для получения данных о генетических последовательностей, например, включающая устройства и способы для секвенирования нуклеиновых кислоты, и/или для реализации конвейера анализа последовательности по таким данным генетического секвенирования. Система может включать одно или более электронных источников данных, таких как источники, связанные с устройством для секвенирования ДНК/РНК, таким как описанное в настоящем документе, памятью и/или интегральной схемой. Например, один вариант реализации включает электронный источник данных, причем указанный электронный источник данных может быть сконфигурирован для генерации и/или обеспечения одного или более цифровых сигналов, таких как цифровые сигналы, представляющие один или более ридов генетических данных, где, например, каждый рид геномных данных включает последовательность нуклеотидов. Далее, память может быть сконфигурирована для хранения одной или более референсных генетических последовательностей, и дополнительно может быть сконфигурирована для хранения индекса, такого как индекс одной или более референсных генетических последовательностей, и/или аннотированные данные о границах сплайсинга.

[0021] Далее, предложены устройство и/или способ для получения данных о генетических последовательностях. Например, предложен подход к анализу ДНК/РНК, например, для генетической диагностики и/или секвенирования, включающий одну или более реакций из гибридизации, детектирования, и/или секвенирования нуклеиновых кислот. В различных вариантах, подход может включать устройства для гибридизации и/или детектирования и/или процедуры для выполнения одного из следующих этапов. В частности, для целей генетического анализы образец РНК тдт ДНК субъекта можно выделить и иммобилизовать, например, напрямую и/или опосредованной, на субстрате, таком как субстрат, содержащий химически чувствительный одномерный (1-D) и/или двумерный (2D) реакционный слой, например, графеновый реакционный слой, и/или трехмерный (3D) реакционный слой, а зонд известной или детектируемой генетической последовательности, например, маркера заболевания, можно проливать по поверхности,

или наоборот. В различных вариантах, один или более из образцов РНК или ДНК и/или зонда могут содержать метку.

[0022] В других вариантах, как в случае, когда субстрат включает 1D- или 2D, например, графеновый, реакционный слой, и/или другой химически чувствительный реакционный слой, метку или зонд, такая химическая или радиоактивная метка не обязательно является необходимой и/или присутствует. В любом из вариантов, если присутствует маркер заболевания, будет происходить событие связывания, например, гибридизация, а поскольку события связывания можно детектировать, например, путем мечения аналита или зонда и/или за счет соответствующей организации реакционного слоя, описанной в настоящем документе, присутствие маркера заболевания также можно детектировать. В случае, когда маркер заболевания не присутствует, реакция будет отсутствовать, и, соответственно, маркер не будет детектироваться. Безусловно, в некоторых случаях, показательным событием может быть отсутствие события связывания. Соответственно, система может быть выполнена таким образом, что событие гибридизации может детектироваться, либо не детектироваться, что указывает на присутствие или отсутствие маркера заболевания во взятом у субъекта образце.

[0023] Аналогичным образом, для секвенирования ДНК и/или РНК сначала неизвестную последовательность нуклеиновой кислоты, которую необходимо идентифицировать, например, однонитевую последовательность ДНК или РНК субъекта, выделяют, амплифицируют и иммобилизируют на подложке, которая, как описано в настоящем документе, может включать 1D-, 2D-, например, графеновый, слой, 3D, или другой реакционный слой, нанесенный на нее. Затем известную нуклеиновую кислоту, например, нуклеотидное основание, которое может быть помечено идентифицируемой меткой, приводят в контакт с неизвестной последовательностью нуклеиновой кислоты в присутствии полимеразы. Как отмечалось выше, в случае, когда происходит реакционное событие происходит вблизи соответствующим образом выполненного (сконфигурированного) реакционного слоя, например, содержащего графен реакционного слоя, реагент с меткой можно не использовать.

[0024] Соответственно, когда происходит гибридизация, нуклеиновая кислота связывается с соответствующим комплементарным основанием в неизвестной последовательности, например, в секвенируемой ДНК или РНК из образца, и иммобилизуется на поверхности подложки, например, вблизи реакционного слоя. Событие связывания можно детектировать, например, оптическими средствами, электрическими средствами, и/или посредством подходяще детектируемой реакции, происходящей на реакционном слое. Затем эти этапы повторяют до тех пор, пока ДНК или РНК из образца не будет полностью секвенирована. Обычно эти этапы выполняют на секвенаторе следующего поколения известным в технике образом, или они могут быть выполнены в соответствии с устройствами и способами, описанными в настоящем документе, что позволяет провести и/или обработать от тысяч до миллионов реакций секвенирования одновременно, а цифровые данные, полученные в результате этих реакций можно анализировать с применением инновационных устройств и способов для секвенирования, раскрытых в настоящем документе, например, в мультиплексном конвейере для биоинформатической обработки.

[0025] Например, в одном аспекте, например, связанном с инновационными устройствами для секвенирования, представленными в настоящем документе, сконфигурированная соответствующим образом платформа может быть выполнена в виде полевого транзистора (FET), содержащего химический реакционный слой, такого как транзистор для применения в реакции гибридизации и/или секвенирования. В

частности, такой полевой транзистор (FET) может быть выполнен на первичной структуре, такой как подложка, например, кремниевая подложка. В различных вариантах первичная структура может включать одну или более дополнительных структур, например, в пакетной конфигурации, например, в форме слоя изолирующего материала. Например, материал изолятора может быть расположен поверх первичной структуры кремниевой подложки, и может представлять собой неорганический материал, такой как диоксид кремния, или азотистый кремний, или органический материал, такой как полимид, бензоциклобутен или аналогичный материал.

[0026] Первичная структура и/или слой изолятора могут включать дополнительную структуру, содержащую один или более истоков носителей заряда и/или стоков носителей заряда, например, отделенные друг от друга некоторым пространством, и интегрированную в первичную структуру, и/или слой изолирующего материала, и/или может лежать в одной плоскости с верхней и/или нижней поверхностью изолятора с образованием верхнего и/или нижнего затвора. В различных вариантах структуры, например, структура кремниевой подложки, могут дополнительно включать, либо могут быть связаны с интегральной схемой, такой как процессор, например, микропроцессор, для обработки сгенерированных данных, таких как данные с сенсора, например, данные, полученные в результате реакции секвенирования, например, вблизи области затвора. Соответственно, множество структур может быть выполнено в форме, или включать, интегральную схему, и/или может быть представлено в форме ASIC (интегральной схемы специального назначения), структурированной ASIC или FPGA (программируемой пользователем вентильной матрицы, ППВМ).

[0027] В частности, эти структуры могут быть выполнены в форме комплементарной структуры металл-оксид-полупроводник (КМОП), которая, в свою очередь, может быть выполнена в форме химически чувствительного сенсора на полевых транзисторах, содержащего один или более истоков носителей заряда, стоков носителей заряда и/или реакционную область, такую как затвор, причем он сам может включать конфигурацию микро- или наноканала, камеры и/или лунки, причем такой сенсор может быть адаптирован для взаимодействия с процессором. Например, полевой транзистор может включать КМОП-конфигурацию, содержащую или связанную с интегральной схемой, которая выполнена на кремниевой подложке, которая дополнительно включает слой изолятора, причем указанный слой изолятора включает исток носителей зарядов и сток носителей зарядов, интегрированные в слой изолятора, причем исток и сток могут быть выполнены из металла, такого как инкрустированная медь. В различных вариантах КМОП и родственные структуры могут включать поверхность, например, верхнюю поверхность, которая может включать канал и/или камеру, с образованием реакционной лунки, причем поверхность реакционной лунки может быть простираться за пределы истока носителей заряда к стоку носителей заряда и может быть приспособлена для приема различных реагентов, применяемых при проведении биохимических реакций, таких как реакции гибридизации и/или секвенирования ДНК или РНК.

[0028] В некоторых вариантах поверхность, и/или канал, и/или камера могут включать одномерный транзисторный материал, двумерный транзисторный материал, трехмерный транзисторный материал и/или нечто подобное. Различные варианты могут включать одномерный (1D) транзисторный материал, где 1D-материал может состоять из углеродной нанотрубки или полупроводникового нанопровода, которые в различных вариантах могут быть выполнены в форме пласта или канала, и/или в различных вариантах могут включать нанопору, хотя во многих вариантах нанопора отсутствует или не является необходимой. Некоторые варианты реализации могут содержать

двумерный (2D) транзисторный материал, где 2D-материал может включать графеновый слой, силицен, дисульфид молибдена, черный фосфор и/или конфигурация. В различных вариантах поверхность и/или канал могут включать диэлектрический слой. Дополнительно, в различных вариантах, реакционный слой, например, оксидный слой, может быть расположен на поверхности и/или в канале и/или камере, например, может быть наслоен или иным образом размещен на 1D, 2D, например, графеновой 3D-слое (слоях). Такой оксидный слой может представлять собой оксид алюминия или оксид кремния, например, диоксид кремния. В различных вариантах, пассивирующий слой может быть расположен на поверхности, и/или канале, и/или в камере, например, он может быть наслоен или нанесен иным образом на 1D-, 2D-, например, графеновый, или 3D-слой (слои) и/или на связанном реакционном слое на поверхности и/или канале и/или камере.

[0029] В конкретных вариантах первичная и/или вторичная и/или третичная структуры могут быть изготовлены или иным образом выполнены таким образом, что они включают структуру лунки или камеры в и/или на поверхности, например, с образованием реакционной области. Например, структура с лунками или камерами может быть расположена на части поверхности, например, внешней поверхности первичной, и/или вторичной и/или третичной структур. В некоторых вариантах содержащая лунки структура может быть выполнена в виде микро- или нанокамеры и может быть выполнена сверху, или может включать по меньшей мере часть одномерного, двумерного, например, графенового, и/или трехмерного материала, и/или дополнительно включать реакционный, например, оксидный и/или пассивирующий слои. В различных вариантах, структура, содержащая камеру и/или лунку, может образовывать отверстие, например, отверстие, обеспечивающее доступ внутрь камеры, например, дающее возможность прямого контакта с одномерной, например, углеродной нанотрубкой или нанопроводом, двумерной, например, графеновой, или трехмерной поверхностью, и/или каналом, и/или камерой. В конкретных вариантах камера и/или лунка могут иметь размер, позволяющий определить ее как микро- или нанокамеру.

[0030] Соответственно, один из дополнительных аспектов настоящего изобретения представляет собой биосенсор, например, биосенсор для осуществления реакции секвенирования нуклеиновых кислоты. Биосенсор включает КМОП-структуру, которая может быть выполнена в виде химически чувствительного сенсора на основе полевых транзисторов и может включать металл-содержащий исток и сток, например, исток и/или сток, который дополнительно включает последовательность, такую как реакционная зона, которая включает последовательность с одномерным или двумерным слоем, например, графеновым слоем, или 3D-последовательность, которая расположена от истока до стока. В частности, реакционная область может содержать, либо может быть выполнена в виде структуры лунки или камеры, которая может быть расположена на части внешней поверхности лунки с 1D- или 2D-слоем. В таком варианте структура лунки может быть выполнена таким образом, чтобы ограничивать отверстие, которое обеспечивает возможность прямого контакта с нанотрубкой, нанопроводом и/или графеновой поверхностью лунки или камеры. В различных вариантах оксидный и/или пассивационный слой может быть расположен в или на поверхностях камеры. Соответственно, в некоторых вариантах может быть предложен химически чувствительный транзистор, такой как полевой транзистор (FET), включающий одну или более нано- или микролунок для реакций секвенирования.

[0031] В некоторых вариантах реализации химически чувствительный полевой транзистор может включать множество лунок и может быть выполнен в виде матрицы,

например, матрицы сенсоров. Такая матрица или матрицы могут быть выполнены с возможностью детектирования присутствия и/или изменения концентрации различных типов аналитов в широком диапазоне химических и/или биологических процессов, включая реакции гибридизации ДНК и/или РНК и/или секвенирования ДНК или РНК. Например, описанные в настоящем документе устройства и/или системы, включающие такие устройства, могут применяться в способе анализа биологических или химических материалов, например, для секвенирования и/или анализа полного генома, генотипирования, микроматричного анализа, анализа панелей, экзомного анализа, анализа микробиома и/или клинического анализа, такого как анализа раковых заболеваний, NIPT-анализ (неинвазовная перинатальная диагностика) и/или анализа карциносаркомы матки и т.п.

[0032] Соответственно, в одном из конкретных вариантов реализации можно применять матрицу на грефеновых полевых транзисторах (gFET) для осуществления методик секвенирования ДНК и/или РНК и обработки этих данных, например в конвейере генетического анализа, описанного в настоящей заявке. Например, матрица на полевых транзисторах по технологии, например, на графеновых полевых транзисторах (gFET), может быть сконфигурирована таким образом, что она будет включать реакционные лунки с реакционным слоем, с возможностью детектирования изменений в концентрации ионов водорода (pH), изменений концентраций других аналитов, и/или событий связывания, связанных с химическими процессами, такими как процессы, связанные с синтезом ДНК или РНК, например, в управляемой реакционной камере или лунке сенсора на основе gFET. Такой химически чувствительный полевой транзистор может включать или может быть адоптирован для соединения с одной или большим числом интегральных схем и/или адаптирован для повешения чувствительности измерения и/или точности сенсора и/или связанной матрицы (матриц), например, за счет включения одной или более последовательностей в реакционной камере или реакционной лунке, где по меньшей мере на одну последовательность нанесен слой одномерного (1D) и/или двумерного (2D) и/или трехмерного (3D) материала, диэлектрический реакционный, пассивирующий слой, и/или тому подобное.

[0033] Соответственно, один из аспектов настоящего изобретения может включать одну или более интегральных схем, которые могут состоять из одного или более наборов аппаратно соединенных цифровых логических схем, как в случае, когда набор аппаратно соединенных цифровых логических схем имеет межсоединения, например, например, посредством множества физических электрических межсоединений, и может быть адаптирована для осуществления и/или детектирования реакций гибридизации и/или секвенирования ДНК или РНК, например, первичной обработке, и/или может быть дополнительно адаптирована для реализации их результатов, например, как в одном или более этапах вторичной и/или третичной обработки. В таких вариантах интегральная схема может включать ввод, какой как реализованный посредством одного или более из множества физических электрических межсоединений, что обеспечивает возможность соединения с электронным источником генерации данным, таким как секвенирование КМОП FET согласно настоящему изобретению и/или секвенатор следующего поколения (Next Gen Sequencer), сконфигурированные для генерации таких данных, например, в форме множества секвенированных сегментов, например, ридов, геномных данных. В конкретных вариантах указанные одна или более интегральных схем могут включать набор аппаратно соединенных цифровых логических схем, которые сконфигурированы для реализации конвейера вторичной и/или третичной обработки сгенерированных ридов геномных данных, например, посредством одного или более связанных

межсоединений.

[0034] В этом случае фиксированные цифровые логические схемы и/или объединяемые межсоединения могут быть сконфигурированы с возможностью получать один или более ридов геномных данных, например, из электронного источника данных. В конкретных вариантах одна или более аппаратно соединенных цифровых логических схем могут быть организованы в виде набора модулей для обработки, например, в виде набора модулей для обработки, где каждый модуль для обработки составлен из поднабора аппаратно соединенных цифровых логических схем и сконфигурирован для выполнения одного или более этапов конвейера секвенирования и/или анализа, например, множества ридов геномных данных. В таких вариантах каждый поднабор аппаратно соединенных цифровых логических схем может, в некоторых вариантах, иметь аппаратно соединенную конфигурацию, позволяющую осуществлять один или более этапов в конвейере секвенирования и/или анализа. Однако, как отмечалось выше, один или более этапов в конвейере секвенирования и/или анализа могут быть реализованы в программном обеспечении, как, например, в случае, когда программное обеспечение и/или аппаратное обеспечение адаптируют для оптимизированного взаимодействия друг с другом в процессе работы.

[0035] Соответственно, в различных вариантах, предложено множество аппаратно соединенных (смонтированных) цифровых логических схем, причем указанные аппаратно соединенные цифровые логические схемы организованы в форме набора модулей обработки, причем один или более модулей обработки включают один или более модулей секвенирования и/или модулей картирования (картировщиков) и/или модулей выравнивания и/или модулей сортировки (сортировщиков) и/или модулей определения вариантов и/или модулей третичной обработки, описанных в настоящем документе. Например, в различных вариантах реализации указанные один или более модулей обработки могут включать модуль картирования (картировщик), причем картировщик может быть электрически соединен и дополнительно может быть сконфигурирован для взаимодействия с памятью, на устройстве, либо может быть связан с ним иным образом, например, посредством подходящим образом скорфигурированного межсоединения с обеспечением доступа к индексу, содержащему одну или более референсных генетических последовательностей, один или более ридов данных, сгенерированных в результате секвенирования, и/или индексу границ сплайсинга (например, в случае секвенирования РНК), при этом картировщик применяется для выполнения одной или более операций картирования.

[0036] В частности, подходящим образом сконфигурированные модуль или модули обработки могут включать, либо могут быть адаптированы для применения в качестве модуля картирования, для выполнения одной или более операций картирования, например, включая доступ к индексу одной или более референсных генетических последовательностей из памяти, например, посредством одного или более из множества физических электронных межсоединений, например, таким образом, чтобы картировать множество ридов на один или более на один или более сегментов одной или более референсных генетических последовательностей. Дополнительно, в различных вариантах реализации указанные один или более модулей обработки могут включать модуль выравнивания, который может быть связан (wired) с памятью и может быть сконфигурирован для оценки одной или более референсных генетических последовательностей из памяти, например, посредством одного или более из множества физических электронных межсоединений, например, с возможностью выравнивания множество ридов с одним или большим числом сегментов одной или большего

количества референсных генетических последовательностей.

[0037] Далее, в различных вариантах реализации указанные один или более модулей обработки могут включать модуль сортировки, причем указанный модуль сортировки может быть связан (wired) и может быть сконфигурирован для оценки одного или более выровненных ридов из памяти, например, посредством одного или более из множества физических электронных межсоединений, например, с возможностью сортировки каждого выровненного рида, например, в соответствии с одним или большим числом положений этого рида в одном или большем числе референсных геномных последовательностей. В таких вариантах указанные один или более из множества физических электрических межсоединений могут включать вывод интегральной схемы, например, для передачи результирующих данных из модуля картирования и/или модуля выравнивания и/или модуля сортировки. Далее, в конкретных вариантах реализации, как указано выше, один или более модулей обработки могут быть сконфигурированы для взаимодействия с различным программно реализованными функциями, например, посредством одного или более межсоединений, например, множества физических электронных межсоединений, для осуществления одно или более этапов в конвейере анализа, включая реализацию одного или более протоколов секвенирования РНК и/или ДНК и/или протокола определения варианта.

[0038] В различных вариантах, множества указанные одна или более электронных схем могут включать главный контроллер с получением соединенной (wired) конфигурации для каждого поднабора аппаратно соединенных цифровых логических схем, например, для выполнения одной или более функций картирования, выравнивания и/или сортировка, причем указанные функции могут быть реализованы в виде одного или более этапов конвейера анализа последовательности и/или могут включать выполнение одного или более аспектов секвенирования и/или функции определения вариантов. Далее, в различных вариантах реализации указанные одно или более интегральных схем, раскрытых в настоящем документе, могут быть выполнены в виде программируемой пользователем вентильной матрицы (FPGA), содержащей аппаратно соединенные цифровые логические схемы, как в случае, когда соединенная конфигурация может быть получена после изготовления интегральной схемы, и, соответственно могут быть энергонезависимыми. В других различных вариантах реализации интегральной схемы специального назначения (ASIC), содержащей аппаратно соединенные цифровые логические схемы. В других различных вариантах реализации интегрированная схема может быть сконфигурирована в виде структурированной интегральной схемы специального назначения (Structured ASIC), содержащей аппаратно соединенные цифровые логические схемы.

[0039] В некоторых вариантах указанные одна или более интегральных схем, например, выполненных по технологии КМОП на полевых транзисторах для секвенирования, и/или биосенсор, и/или одно или более подсоединяемых запоминающих устройств, могут быть расположены на расширительной плате, такой как плата расширения PCI, например, в различных вариантах реализации интегральная схема (схемы) согласно настоящему изобретению может представлять собой чип (интегральную схему) с картой PCIe. В различных вариантах, интегральная схема и/или чип могут быть компонентами секвенатора, например, автоматизированного секвенатора, в котором используется сенсоры на полевых транзисторах, и/или секвенатора нового поколения (NGS), и/или, в других вариантах реализации может быть обеспечена возможность доступа к интегральной схеме и/или расширительное плате могут через интернет, например, через облако. Далее, в некоторых вариантах,

запоминающее устройство может представлять собой энергозависимое ОЗУ (RAM) или динамическое ОЗУ (DRAM).

[0040] Соответственно, в одном аспекте предложено устройство для выполнения одного или более этапов конвейера анализа последовательности, такого как анализа генетических данных, причем генетические данные включают одно или более из референсной генетической последовательности (последовательностей), индекса одной или более генетической последовательности (последовательностей), индекса одной или более референсной генетической последовательности (последовательностей), индекса одной или более границ сплайсинга, например, аннотированного индекса или таблицы границ сплайсинга, и/или множества ридов, например, генетических данных, например, ДНК или РНК. В различных вариантах устройство может содержать интегральную схему, причем указанная интегральная схема может включать одну или более, например, набор, аппаратно соединенных цифровых логических схем, причем указанный набор аппаратно соединенных цифровых логических схем может быть соединен межсоединениями, например, множеством физических электрических межсоединений. В некоторых вариантах указанные один или более из множества электрических межсоединений могут включать ввод, такой как вводя для получения множества ридов геномных данных, например, с устройства для секвенирования, как описано в настоящем документе. Дополнительно, набор аппаратно соединенных цифровых логических схем может быть дополнительно выполнен в соединенной конфигурации с доступом к индексу одной или более референсных генетических последовательностей и/или аннотированным границам сплайсинга, посредством одного из множества физических электрических межсоединений, и с возможностью картирования множества ридов ДНК и/или РНК на один или более сегментов одной или более референсных генетических последовательностей, например, в соответствии с индексом или индексами.

[0041] В различных вариантах реализации индекс может включать одну или более хеш-таблиц, таких как первичная и/или вторичная и/или таблица границ сплайсинга. Например, может присутствовать первичная хеш-таблица, причем в таком варианте набор аппаратно соединенных цифровых логических схем может быть сконфигурирован для выполнения одной из следующих операций: извлечения одной или более затравок генетических данных из множества ридов генетических данных; выполнения первичной функции хеширования, например, с одной или более затравками генетических данных, к генерацией справочного адреса для каждой из одной или более затравок, и оценки первичной хеш-таблицы с применением справочного адреса с получением положения в одном или большем числе референсных геномных последовательностей для каждой из одной или более затравок генетических данных. В различных вариантах указанные одна или более затравок генетических данных могут содержать фиксированное число нуклеотидов.

[0042] Далее, в различных вариантах реализации индекс может включать вторичную хеш-таблицу, например, как в случае, когда набор аппаратно соединенных цифровых логических схем сконфигурирован для по меньшей мере одного из следующего: удлинения по меньшей мере одной из одной или более затравок за счет дополнительных соседних нуклеотидов с получением по меньшей мере одной удлиненной затравки генетических данных; выполнения функции хеширования, например, вторичной функции хеширования в отношении по меньшей мере одной удлиненной затравки генетических данных с генерацией второго справочного адреса для по меньшей мере одной удлиненной затравки, и оценки вторичной хеш-таблицы, например, с применением второго справочного адреса, с получением положения в одной или более референсных

генетических последовательностей для каждой из одной или более удлиненных затравок генетических данных. В различных вариантах вторичная функция хеширования может быть выполнена набором аппаратно соединенных цифровых логических схем, например, как в случае, когда первичная хеш-таблица возвещает запись об удлинении, которая является инструкцией для аппаратно соединенных цифровых логических схем удлинить одну или более затравок дополнительными соседними нуклеотидами. В некоторых вариантах запись об удлинении может указывать число дополнительны соседних нуклеотидов, которыми следует удлинить указанную по меньшей мере одну затравку, и/или способ, которым следует удлинить затравку, например, одинаковым образом на равное число “x” нуклеотидов по каждому из концов затравки.

[0043] Далее, как известно, ДНК кодирует гены. Однако для экспрессии гена необходима трансляция и транскрипция генетического кода в белки. В частности, ген может транскрибироваться в ядре клетки ферментами РНК-полимеразами в транскрипт - матричную (информационную) РНК (мРНК) или другие типы РНК (например, транспортную РНК). Непосредственно РНК-транскрипт представляет собой однонитевую копию гена, за тем исключением, что основание тимин (T), присутствующее в ДНК, заменено на урацил (U). Однако сразу после образования этой копии ее последовательность включает копии как различных экзонов, так и интронов, причем различные копии интронов обычно вырезаются в процессе сплайсинга, например, сплайсосомами, в результате чего остаются только копии экзонов, которые должны быть соединены друг с другом по “стыкам (точкам, границам) сплайсинга” (которые после этого не всегда можно легко выявить), с образованием областей кодонов. Затем сплайсированная мРНК, содержащая области кодонов, транспортируется из ядра клетки на рибосому, которая расшифровывает ее в белок, при этом каждая группа из трех нуклеотидов РНК образует кодон, который кодирует одну аминокислоту. В ходе процесса расшифровки происходит связывание аминокислот в цепочки, а в результате связывания и гликозилирования образуются белки, из которых состоят клетки, ткани и органы организма. Таким образом гены в ДНК выполняют функцию исходных инструкций для производства белков.

[0044] Соответственно, поскольку ДНК включает как кодирующие области, например, экзоны, так и некодирующие области, например, интроны, картирование и/или выравнивание и/или сортировка РНК относительно исходного генетического предшественника в геномной ДНК могут быть затруднены. В частности, каждый ген существует на одной нити двунитевой двойной спирали ДНК, часто в виде ряда экзонов (кодирующих сегментов), разделенных интронами (некодирующими сегментами). Некоторые гены содержат единственный экзон, но большинство содержит несколько экзонов (разделенных интронами), а некоторые содержат сотни экзонов или тысячи экзонов. Длина экзонов обычно составляет несколько сотен нуклеотидов, но они могут быть и меньшей длины, до одного нуклеотида, или большей длины до десятков, сотен или тысяч нуклеотидов. Длина интронов обычно составляет тысячи нуклеотидов, А у некоторых превосходит миллион нуклеотидов. Соответственно, при картировании, выравнивании, и/или сортировке по РНК, например, сплайсированной мРНК, части этой сплайсированной мРНК могут происходить из различных областей ДНК, которые могут быть отделены друг от друга одним, или двумя, или даже миллионом или более нуклеотидов. Это обуславливает высокую сложность обработки РНК.

[0045] Однако один из аспектов настоящего изобретения позволяет преодолеть эти сложности благодаря способам, раскрытым в настоящем документе, и, соответственно, обеспечивает возможность быстрого и точного полнотранскриптомного

секвенирования, картирования, выравнивания и/или сортировки РНК. Более конкретно, в вариантах, включающих обработку РНК, указанный выше индекс может включать одну или более таблиц, например, хеш-таблицу или другой индекс, включающих или связанных с таблицей, которая позволяет легко обращаться к различным известными или определенным границам сплайсинга, применяемым в биологических системах при транскрибировании РНК с ДНК, как подробно описано ниже. Соответственно, в таких вариантах РНК-компетентный модуль картирования/модуль сортировки может быть сконфигурирован для обработки таких границ сплайсинга и сопоставления ридов последовательности РНК с сегментами транскрибированной и сплайсированной РНК, как в случае, когда рид пересекает одну или более границ сплайсинга; что, в отношении ДНК-ориентированного референсного генома означает, что первая часть рида происходит из первого экзона и должна картироваться на первый экзон, а вторая часть рида происходит из второго экзона и должна картироваться на него, а так далее. Соответственно, индекс может включать или быть связанным с одной или более таблиц границ сплайсинга, а набор аппаратно соединенных цифровых логических схем может быть сконфигурирован для выполнения одной из следующих операций: применения указанных данных о границах сплайсинга для определения и/или извлечения одной или более затравок генетических данных, например, РНК, из множества ридов генетических данных на основе РНК; выполнения функции, например, функции хеширования, например, для указанных одной или более затравок генетических РНК-данных, с генерацией справочного адреса для каждой из указанных одной или более затравок; и оценки хеш-таблицы с применением справочного адреса с получением положения в одном или большем числе референсных геномных последовательностей для каждой из указанных одной или более затравок генетических РНК-данных.

[0046] Дополнительно, в одном аспекте предложено устройство для выполнения одного или более этапов конвейера анализа последовательности по данным генетической последовательности, например, ДНК или РНК, где указанные данные генетической последовательности включают одну генетическую последовательность или множество генетических последовательностей, которые могут включать как экзоны, так и интроны, индекс одной или более референсных генетических последовательностей и/или индекс аннотированных границ сплайсинга, и множество ридов геномных. В различных вариантах, устройство может включать интегральную схему, причем указанная интегральная схема может включать одну или более, например, набор, аппаратно соединенных цифровых логических схем, причем указанный набор аппаратно соединенных цифровых логических схем может быть соединен межсоединениями, например, множеством физических электрических межсоединений. В некоторых вариантах указанные одно или более из множества физических электрических межсоединений могут включать ввод, такой как ввод для получения множества ридов геномных данных, причем риды могут быть предварительно обработаны как описано в настоящем документе, например, могут быть картированы. Дополнительно, набор аппаратно соединенных цифровых логических схем может быть дополнительно соединен с обеспечением доступа к одной или более референсных генетических последовательностей, посредством одного из множества физических электрических межсоединений, для получения информации о положении, например, от картировщика, указывающей один или более сегментов одной или более референсных последовательностей, и для выравнивания множества ридов с указанными одним или более сегментами одной или более референсных генетических последовательностей.

[0047] Соответственно, в различных вариантах, набор аппаратно соединенных

цифровых логических схем в соединенной (wired), сконфигурирован для выравнивания множества ридов генетических данных ДНК РНК с одним или большим числом сегментов одной или более референсных генетических последовательностей и дополнительно включают модуль волновой обработки данных, который может быть образован соединенной конфигурацией набора аппаратно соединенных цифровых логических схем. В некоторых вариантах реализации указанный модуль волновой обработки данных может быть сконфигурирован для обработки массива элементов матрицы выравнивания, такой как матрица, задаваемая поднабором набора аппаратно соединенных цифровых логических схем. Например, в некоторых вариантах матрица выравнивания может задавать первую ось, например, представляющую один из множества ридов, и вторую ось, например, представляющую один или более сегментов одной или более референсных генетических последовательностей. В таком варианте указанный модуль волновой обработки данных может быть сконфигурирован для генерации профиля фронта волны лунок, которые расположен по массиву элементов от первой оси до второй оси, и также может быть сконфигурирован для генерации балла, например, балла для каждого элемента в волновом профиле элементов, где указанный балл может представлять степень совпадения указанного одного из множества ридов и одного или более сегментов одной или более референсных генетических последовательностей.

[0048] В таком варианте указанный модуль волновой обработки данных может быть дополнительно сконфигурирован для перемещения профиля фронта волны элементов по матрице выравнивания таким образом, что наивысший балл может быть центрирован на профиле фронта волны лунок. Дополнительно, в различных вариантах реализации указанный модуль волновой обработки данных может быть дополнительно сконфигурирован для обратного отслеживания одного или более, например, всех положений в профиле волны элементов с присвоенными баллами по предшествующим положениям в матрице выравнивания; отслеживания одного или более, например, всех отслеженных траекторий до схождения, и генерации строки CIGAR на основании следа схождения.

[0049] В некоторых вариантах реализации соединенная конфигурация набора аппаратно соединенных цифровых логических схем для выравнивания множества ридов с одним или большим числом сегментов одной или более референсных генетических последовательностей может включать соединенную конфигурацию для применения алгоритма Барроуза-Уилера, как описано выше, например, для картирования перед выравниванием, и/или для реализации алгоритма оценки Смита-Уотермана и/или Нидлмана-Вунша. В таком варианте алгоритм оценки Смита-Уотермана и/или Нидлмана-Вунша может быть сконфигурирован для реализации параметра оценки, чувствительного к оценкам качества оснований. Далее, в некоторых вариантах реализации алгоритм оценки Нидлмана-Вунша может представлять собой аффинный алгоритм Нидлмана-Вунша.

[0050] В конкретных вариантах реализации устройство может включать интегральную схему, причем указанная интегральная схема может включать одну или более, например, набор аппаратно соединенных цифровых логических схем, причем указанный набор аппаратно соединенных цифровых логических схем может быть соединен межсоединениями, например, множеством физических электрических межсоединений. В некоторых из этих вариантов указанные одно или более из множества физических электрических межсоединений могут включать ввод, такой как ввод для получения множества ридов геномных данных, причем риды могут быть предварительно

обработаны как описано в настоящем документе, например, могут быть картированы и/или выровнены. Дополнительно, набор аппаратно соединенных цифровых логических схем может дополнительно находиться в соединенной конфигурации, которая обеспечивает доступ к одной или большему числу генетических последовательностей, посредством одного из множества физических электрических межсоединений, для получения информации о положении, например, от модуля картирования и/или модуля выравнивания, которая указывает на один или более сегментов одной или более референсных последовательностей, и для сортировки множества ридов с одним или большим числом сегментов одной или более референсных генетических последовательностей.

[0051] Соответственно, в одном аспекте может быть предложен способ секвенирования генетического материала, например, для получения электронных генетических данных. В конкретных вариантах указанный способ включает применение секвенатора следующего поколения для секвенирования геномной ДНК и/или соответствующей РНК, как описано в общих чертах в настоящем документе и как известно в данной области. В других вариантах указанный способ включает применение секвенатора следующего поколения, модифицированного, как описано в настоящем документе, для секвенирования геномной ДНК и/или соответствующей РНК. В других вариантах указанный способ включает применение секвенатора на полевых транзисторах и/или на основе технологии КМОП, например, секвенатора на чипе, как подробно описано в настоящем документе, для секвенирования геномной ДНК и/или соответствующей РНК. В различных вариантах генетический материал после получения может быть преобразован в электронную форму, например, в цифровую форму, которую можно передавать потоком или иным образом переносить в один или более модулей конвейера, описанных в настоящем документе.

[0052] Дополнительно, после получения электронных, например, аналоговых или цифровых генетических данных, таких как данные секвенирования, другой аспект настоящего изобретения относится к исполнению последовательных операций конвейера анализа последовательности с такими данными секвенирования генетического материала. Эти генетические данные могут включать одну или более референсных генетических последовательности, один или более индекс йодной или более референсных генетических последовательностей и/или список одной или более аннотированных границ сплайсинга (например, в случае секвенирования РНК), относящихся к ним, и/или множество ридов геномных данных (например, ДНК и/или РНК). Способ может включать одно или более различных итераций приема, оценки, картирования, выравнивания, и/или сортировки указанных данных генетической последовательности. Например, в некоторых вариантах реализации способ может включать прием, на входе интегральной схемы от источника электронных данных, одного или более из множества ридов геномных данных, причем каждый геномных данных может включать последовательность нуклеотидов. В таком варианте интегральная схема может быть образована набором аппаратно соединенных цифровых логических схем, например, соединенных множеством физических электрических межсоединений, причем указанные электрические межсоединения могут включать одно или более из множества физических электрических межсоединений, включающих ввод.

[0053] Способ может дополнительно включать оценку, интегральной схемой на одной или более из множества физических электрических межсоединений от запоминающего устройства, индекса одной или более референсных генетических последовательностей и/или, в случае секвенирования РНК, аннотированных границ

сплайсинга. В частности, если в модуль картирования поступают аннотированные границы сплайсинга, они могут быть использованы для повышения чувствительности картирования. В таком варианте список аннотированных границ можно загрузить в запоминающие устройство таким образом, чтобы модуль картирования имел к нему доступ, что облегчает картирование генетического материала РНК. В предпочтительном варианте аннотированные границы могут быть представлены в форме таблицы, например, например, хеш-таблицы или индекса, который может быть связан с ней таким образом, чтобы модуль каритирования легко могу получить к ним доступ. Соответственно, способ может включать картирование, первым поднабором аппаратно соединенных цифровых логических схем интегральной схемы, множества генетических ридов, например, ридов ДНК или РНК, по одному или более сегментом одной или более референсных генетических последовательностей. Дополнительно, способ может включать оценку, интегральной схема на одном или более из множества физических электрических межсоединений от запоминающего устройства, указанных одного или более картированных ридов риды и/или референсных генетических последовательностей; и выравнивание, вторым поднабором аппаратно соединенных цифровых логических схем интегральной схемы, множества ридов, например, картированных ридов, с одним или большим числом сегментов одной или более референсных генетических последовательностей.

[0054] В различных вариантах реализации способ может дополнительно включать оценку, интегральной схемой на одном или более из множества физических электрических межсоединений от памяти, выровненного множества ридов ридов. В таком варианте способ может включать сортировку, третьим поднабором аппаратно соединенных цифровых логических схем интегральной схемы, выровненного множества ридов в соответствии с их положениями в одной или большем числе референсных геномных последовательностей. В некоторых вариантах способ может дополнительно включать вывод, например, на одном или более из множества физических электрических межсоединений интегральной схемы, результирующих данных картирования, и/или выравнивания и/или сортировки, как в случае, когда результирующие данные включают положения картированного, и/или выровненного, и/или отсортированного множества ридов.

[0055] Далее, после генерации и/или обработки генетических данных, например, в одном или большем числе протоколов вторичной обработки, например, путем картирования, выравнивания и/или сортировки, например, с получением одного или более файлов определения вариантов, например, для определения того, как указанные данные генетической последовательности субъекта отличаются от одной или более референсных последовательностей, дополнительный аспект настоящего изобретения может быть связан с выполнением одной или более аналитических функций с сгенерированными и/или обработанными генетическими данным, например, для дальнейшей третичной обработки. Например, система может быть сконфигурирована для дальнейшей обработки сгенерированных данных и/или данных после вторичной обработки, например, путем обработки в одном или большем числе конвейеров третичной обработки, например, одном или более из геномного конвейера, эпигеномного конвейера, метагеномного конвейера, конвейера совместного генотипирования, конвейера MuTect2 или другого конвейера третичной обработки. В частности, в различных вариантах, может быть предложен дополнительный уровень обработки, например, для диагностики заболеваний, терапевтического лечения и/или профилактического лечения, такого как с применением процедур NIPT, NICU, Cancer,

LDT, AgBio (сельскохозяйственный и биологический анализ) и других подобных процедур для диагностики, профилактики и/или лечения заболеваний, с применением данных, сгенерированных одним или более из представленных первичных, и/или вторичных, и/или третичных конвейеров. Соответственно, устройства и способы, раскрытые в настоящем документе, можно применять для генерирования данных о генетических последовательностях, причем указанные данные могут затем применяться для генерации одного или более файлов определения вариантов и/или других связанных данных, которые затем могут быть переданы для обработки в конвейерах третичной обработки в соответствии с устройствами и способами, раскрытыми в настоящем документе, например, для частной и/или общей диагностики заболеваний, а также для профилактического и/или терапевтического лечения и/или, для приложений, связанных с развитием.

[0056] Соответственно, в различных вариантах, варианты реализации различных аспектов настоящего изобретения могут включать, но не ограничиваются следующими: устройства, системы и способы, включающие один или более признаков, подробно описанных в настоящем документе, а также изделия, которые содержат материальный машиночитаемый носитель, выполненный с возможностью заставлять одну или более машин (например, компьютеров, и т.д.) выполнять операции, описанные в настоящем документе. Аналогичным образом, также описаны компьютерные системы и/или сети, которые могут включать один или более процессоров и/или одно или более запоминающих устройств, связанных с указанным одним или более процессорами напрямую или удаленно. Соответственно, реализуемые при помощи компьютера способы, соответствующие одному или более вариантов реализации настоящего изобретения, могут быть реализованы посредством одного или более процессора данных, расположенного в одной вычислительной системе или во множестве вычислительных систем, например, в одном или более кластерах компьютеров. Такие множественные системы вычислений могут быть связаны с и могут обмениваться данными и/или командами или другими инструкциями и т.п. посредством одного или более соединений, включая следующие, но не ограничиваясь ими: соединение через сеть (например, интернет, беспроводную глобальную сеть, локальную сеть, глобальную сеть, проводную сеть и т.п.), или посредством прямого соединения между оной или более вычислительными системами и т.п. Память, которая может включать машиночитаемый носитель для хранения информации, может содержать, кодировать, хранить и т.п. одну или более программ, которые заставляют процессор выполнять одну или более операций, описанных в настоящем документе.

[0057] Варианты реализации описанного в настоящем документе объекта подробно раскрыты ниже на прилагающихся к тексту графических материалах и в описании ниже. Другие признаки и преимущества описанного в настоящем документе объекта станут понятны из описания и графических материалов, а также из формулы изобретения. В то время как некоторые признаки описанного здесь объекта в иллюстративных целях раскрыты применительно к корпоративной программной системе или другим бизнес-решениям или архитектуре, очевидно, что такие признаки не являются ограничивающими. Предполагается, что признаки, которые следуют за этим раскрытием, определяют объем защищенного объекта.

ОПИСАНИЕ ГРАФИЧЕСКИХ МАТЕРИАЛОВ

[0058] Сопровождающие настоящий текст графические материалы, которые включены в настоящее описание и составляют его часть, иллюстрируют некоторые аспекты раскрытого здесь объекта и, вместе с описанием, служат для объяснения некоторых

принципов, связанных с раскрытыми вариантами реализации. В этих графических материалах:

[0059] На ФИГ. 1 изображен РНК-рид, иллюстрирующий пересечение между одной или более границами (точками) и затравки, покрывающие границы сплайсинга рида.

[0060] На ФИГ. 2 изображен пример РНК-рида, иллюстрирующий, что короткие (L оснований) затравки можно сконфигурировать таким образом, чтобы они лучше вписывались в короткие экзоны и учитывали выступающие части экзонов или сегменты экзонов, вырезаемые при редактировании, такие как однонуклеотидные полиморфизмы (SNP).

[0061] На ФИГ. 3 изображены примеры референсных интервалов в пределах диапазона поиска успешно картированных затравок из K-оснований, которые могут быть использованы для запроса в фиксированной хеш-таблице затравок, например, с использованием затравок из L оснований.

[0062] На ФИГ. 4 показано сравнений частей ридов слева и справа от положения сшивки.

[0063] На ФИГ. 5 изображен абстрактный прямоугольник выравнивания с конкатенированной последовательностью запроса по вертикальной оси и конкатенированной референсной последовательностью по горизонтальной оси.

[0064] На ФИГ. 6 показано устройство в соответствии с одним из вариантов реализации настоящего изобретения.

[0065] ФИГ. 7 показано другое устройство в соответствии с альтернативным вариантом реализации настоящего изобретения.

[0066] На ФИГ. 8 изображена блок-диаграмма геномной инфраструктуры для локальных и/или облачных обработки и анализа методами геномики.

[0067] На ФИГ. 9 изображена блок-диаграмма локальных и/или облачных вычислений согласно ФИГ. 8 для геномной инфраструктуры для локальных и/или облачных обработки и анализа методами геномики.

[0068] ФИГ. 10 изображена блок-диаграмма согласно ФИГ. 9, на которой более подробно показана реализация вычислений для геномной инфраструктуры для локальных и/или облачных обработки и анализа методами геномики.

[0069] ФИГ. 11 изображена блок-диаграмма согласно ФИГ. 8, на которой более подробно показана осуществляемая 3ей стороной аналитика для геномной инфраструктуры для локальных и/или облачных обработки и анализа методами геномики.

[0070] На ФИГ. 12 изображена блок-диаграмма, иллюстрирующая конфигурацию гибридных облачных вычислений.

[0071] ФИГ. 13 изображена блок-диаграмма согласно ФИГ. 12, более подробно иллюстрирующая конфигурацию гибридных облачных вычислений.

[0072] ФИГ. 14 изображена блок-диаграмма согласно ФИГ. 13, более подробно иллюстрирующая конфигурацию гибридных облачных вычислений.

[0073] ФИГ. 15 изображена блок-диаграмма, иллюстрирующая конвейер первичного, вторичного и/или третичного анализа согласно настоящему раскрытию.

[0074] На ФИГ. 16 изображена блок-схема для конвейера анализа, описанного в настоящем документе.

[0075] ФИГ. 17 показан пример дизайна и сборки интегральной схемы.

[0076] ФИГ. 18 представляет собой блок-диаграмму аппаратной архитектуры процессора в соответствии с другим вариантом реализации раскрытого объекта.

[0077] ФИГ. 19 представляет собой блок-диаграмму аппаратной архитектуры

процессора в соответствии с другим вариантом реализации раскрытого объекта.

[0078] На ФИГ. 20 показан конвейер анализа генетической последовательности.

[0079] На ФИГ. 21 показаны этапы обработки с использованием аппаратной платформы для анализа генетической последовательности.

[0080] Где это уместно, одинаковые ссылочные номера обозначают одинаковые структуры, признаки или элементы.

ПОДРОБНОЕ ОПИСАНИЕ

[0081] Для решения этих и, возможно, других проблем, связанных с доступными в настоящее время решениями, способы, системы, изделия и т.п., соответствующие одному или нескольким вариантам осуществления настоящего изобретения могут, среди прочих возможных преимуществ, обеспечивать устройство анализа последовательности для выполнения конвейера анализа последовательностей на основе данных генетических последовательностей.

[0082] Далее представлено подробное описание различных вариантов реализации платформы секвенирования, поточного анализа последовательностей, а также системы для выполнения одного или нескольких протоколов третичной обработки.

[0083] В самом общем виде, тело состоит из клеток, эти клетки образуют ткани, ткани образуют органы, органы образуют системы, а эти системы функционируют вместе для обеспечения функционирования организма и поддержания жизни индивида. Таким образом, указанные клетки тела являются строительными кирпичиками жизни. Более конкретно, каждая клетка имеет ядро, а внутри указанного ядра каждой клетки располагаются хромосомы. Хромосомы образованы дезоксирибонуклеиновыми кислотами, которые имеют организованную, но свернутую структуру двойной спирали. Сама по себе ДНК состоит из двух противоположно направленных, но комплементарных цепей нуклеотидов, которые образуют гены, которые кодируют белки, придающие клетке ее структуру и влияющие на функционирование тканей и органов тела, а также регулирующие их функции. В общем, белки выполняют большую часть работы клеток по поддержанию нормальных процессов и функций организма.

[0084] С учетом большого количества компонентов тела и сложности, связанной с тем, как они взаимодействуют друг с другом для поддержания различных процессов и функций организма, существует множество способов того, как указанный организм может функционировать неправильно на любом из этих различных уровнях организации. Например, в одном таком варианте, нарушение может заключаться в том, как конкретный ген кодирует заданный белок, и это нарушение зависит от указанного белка, при этом характер этого нарушения может приводить к началу процесса перехода в состояние смерти.

[0085] Соответственно, в диагностике, предотвращении и/или лечении таких патологических состояний, определение генетических характеристик субъекта может быть исключительно полезным. Например, генетические характеристики субъекта, если они известны, например его или ее состав генов, могут быть использованы для целей диагностики и/или для определения, есть ли у субъекта какое-либо патологическое состояние или потенциал для его развития, и, поэтому, может быть использовано для целей профилактики. Также, знание о геноме субъекта может быть полезным при определении потенциальных терапевтических способов воздействия, таких как лекарственные средства, которые могут или не могут быть использованы в профилактических или терапевтических схемах лечения без причинения вреда пациенту. В различных вариантах, знание о геноме субъекта также может применяться при определении эффективности лекарственного средства, и/или вызывающие проблемы

побочные эффекты использования такого лекарственного средства могут быть предсказаны и/или идентифицированы. Потенциально, указанное знание о геноме субъекта может быть использовано для создания синтетического «дизайнерского» лекарственного средства, когда такое лекарственно средство создано индивидуально и оптимизировано в соответствии со специфическими генетическими характеристиками субъекта. В частности, в качестве одного из примеров, сконструированный белок или последовательность нуклеотидов могут быть произведены в соответствии с уникальными генетическими характеристиками субъекта так, чтобы выключать или включать транскрипцию генов, которые производят белки либо в избыточном, либо в недостаточном количестве, и, таким образом, устранять патологические состояния.

[0086] Следовательно, в некоторых примерах, целью обработки данных биоинформатики является определение индивидуальных геномов людей, при этом такие определения могут быть использованы в протоколах детектирования генов, а также для целей профилактических и/или терапевтических способах воздействия, для улучшения качества жизни каждого конкретного субъекта и человеческого рода в целом. Кроме того, знание о геноме субъекта может быть использовано, например, при открытии лекарственных средств и/или в клинических исследованиях для лучшего более тщательного предсказания, какое лекарственное средство вероятно будет работать в субъекте, если вообще будет, и/или какое лекарственное средство вероятно может вызвать пагубные побочные эффекты, посредством анализа индивидуального генома и/или профиля белка, получаемого из него, и его сравнения с предсказанным биологическим ответом на введение такого лекарственного средства.

[0087] Такая геномика и обработка данных биоинформатики обычно включает три хорошо разработанные, но обычно используемые раздельно фазы обработки информации. Первая фаза включает в себя определение последовательности ДНК/РНК, когда получают ДНК/РНК субъекта и подвергают ее различным исследованиям, получая таким образом генетический код субъекта, преобразованный в машиночитаемый цифровой код, например, файл FASTQ. Вторая фаза включает использование полученного цифрового генетического кода субъекта для определения генетических характеристик субъекта, например, определение последовательностей нуклеотидов генов субъекта и/или файла с указанием вариантов, например, того, как геном субъекта отличается от одного или нескольких геномов сравнения. Третья фаза включает проведение одного или нескольких анализов генетических характеристик субъекта для получения на их основе терапевтически полезной информации. В порядке их следования, эти фазы могут быть названы первичная, вторичная и третичная обработка данных, соответственно.

[0088] Предварительно, например, в Фазе I или при первичной обработке данных, генетический материал должен быть предварительно обработан, например, посредством определения последовательности нуклеотидов с получением данных о последовательности генома, которые могут быть использованы в дальнейшем. Определение последовательности нуклеиновых кислот, таких как дезоксирибонуклеиновая кислота (ДНК) и рибонуклеиновая кислота (РНК), является фундаментальной частью биологического исследования. Такое определение является необходимым для большого числа задач и очень часто используется в научных исследованиях, а также в медицинских разработках. Например, области геномики и биоинформатики связаны посредством применения информационных технологий и компьютерных наук в области генетики и/или молекулярной биологии. В частности, методы биоинформатики, такие как описанные в настоящей заявке, могут применяться

для генерирования, обработки и анализа различных геномных данных, например, полученных у человека, для определения качественной и количественной информации об этих данных, которые затем могут быть использованы различными специалистами в разработке индивидуальных и/или глобальных диагностических, профилактических и/или терапевтических методов для выявления, профилактики и/или, по крайней мере, улучшения патологического состояния и, следовательно, повышения безопасности, качества и эффективности медицинской помощи человеку и/или обществу.

[0089] Обычно метод анализа ДНК/РНК, такой как генетическая диагностика, включает гибридизацию и детекцию нуклеиновых кислот. Например, различные типичные подходы гибридизации и детекции включают следующие этапы. Для генетического анализа образец РНК или ДНК субъекта, подлежащего анализу, может быть выделен и иммобилизован на подложке, зонд с известной генетической последовательностью, например маркер заболевания, может быть помечен и нанесен на всю подложку. Если присутствует маркер заболевания, происходит связывание, например, гибридизация, и, поскольку зонд был помечен, событие гибридизации может либо обнаруживаться, либо не обнаруживаться, что указывает на наличие или отсутствие маркера заболевания в образце субъекта. В качестве альтернативы, как указано выше, когда реакция гибридизации происходит рядом с реакционным слоем, например, сконфигурированным для обнаружения реагента и/или побочного продукта реакции, использующегося в подходящем устройстве с полевым транзистором (FET), нет необходимости в применении меченого зонда.

[0090] Обычно для определения последовательности нуклеотидов сначала неизвестную последовательность нуклеиновой кислоты, например, одноцепочечную последовательность ДНК и/или РНК субъекта, которую необходимо идентифицировать, выделяют, амплифицируют и иммобилизуют на подложке. Затем нуклеиновую кислоту с известной последовательностью, меченную идентифицируемой меткой, вводят в контакт с нуклеиновой кислотой с неизвестной последовательностью в присутствии полимеразы. Когда происходит гибридизация, меченая нуклеиновая кислота связывается с ее комплементарным основанием в неизвестной последовательности, иммобилизованной на поверхности подложки. Затем может быть обнаружено событие связывания, например, оптическим или электрическим способами. Затем эти шаги повторяют до тех пор, пока весь образец ДНК не будет полностью секвенирован.

[0091] Как правило, эти этапы выполняют вручную или с использованием автоматического секвенатора, такого как секвенатор следующего поколения (Next Gen Sequencer, NGS), в котором одновременно могут быть получены от тысяч до миллионов последовательностей в процессе секвенирования следующего поколения. При этом, как описано в настоящей заявке, используют систему прямого секвенирования ДНК и/или РНК без использования меток, такую как система на компьютерном чипе, таком как чип с дополнительным металлооксидным полупроводником (CMOS), где различные компоненты или весь сенсорный аппарат секвенатора может быть воплощен внутри указанного полупроводникового чипа или иным образом связан с этим чипом. Такая система, как предоставлено в данной заявке, обеспечивает беспроблемную интеграцию первичной, вторичной и/или третичной обработки информации, например, с использованием одного и того же набора полупроводниковых микросхем.

[0092] Более конкретно, типичная процедура секвенирования, независимо от типа используемого устройства секвенирования, включает получение биологического образца от субъекта, например посредством венепункции, из волос и т.д., и обработку образца для выделения из него генетического компонента. После выделения, где генетическим

образцом является ДНК, указанная ДНК может быть денатурирована и ее цепи разделены. Поскольку РНК уже является одноцепочечной, этот этап может не потребоваться при обработке РНК. Затем выделенную ДНК и/или РНК или ее части могут быть амплифицированы, например, с помощью полимеразной цепной реакции (ПЦР), чтобы создать библиотеку реплицированных цепей, которые теперь готовы для секвенирования и считывания, например, с помощью автоматического секвенатора, который сконфигурирован для считывания реплицированных цепей, например, путем синтеза, и, таким образом, определения последовательностей нуклеотидов, составляющих ДНК и/или РНК. Кроме того, в различных вариантах, таких как создание библиотеки реплицированных и амплифицированных цепей, может оказаться полезным обеспечение избыточного перекрытия при предварительной обработке данной части ДНК и/или РНК. Для выполнения такого избыточного перекрытия, например, с использованием ПЦР, могут потребоваться увеличенные ресурсы и время для подготовки образца, и, следовательно, он будет более дорогостоящим, но это часто повышает вероятность того, что конечный результат будет более точным.

[0093] После создания библиотеки реплицированных цепей ДНК/РНК их можно ввести в автоматический секвенатор, например, NGS, который затем может прочитывать указанные цепи, например, путем синтеза, для определения их последовательности нуклеотидов. Например, реплицированная одноцепочечная ДНК или РНК может быть прикреплена к стеклянному шарику и введена в испытательный сосуд, например, в матрицу. Все необходимые компоненты для репликации его комплементарной цепи, включая меченые нуклеотиды, также последовательно добавляют в указанный сосуд. Например, все «A», «C», «G» и «T», которые могут быть помечены, добавляют либо по одному, либо все вместе, если помечены, чтобы увидеть, какой из нуклеотидов связывается в положении «один» одноцепочечной ДНК или РНК.

[0094] После каждой стадии присоединения в маркированной модели на матрицу подают свет, например, лазер. Если указанная композиция флуоресцирует, то создается изображение, указывающее, какой нуклеотид связан в соответствующем местоположении. В немеченой модели событие связывания может быть обнаружено, например, по изменению сопротивления в затворе, например канала с раствором, вблизи реакционного слоя, где расположен стеклянный шарик, содержащий реплицированную одноцепочечную ДНК или РНК. Более конкретно, когда нуклеотиды добавляют по одному за один раз, если происходит событие связывания, то будет наблюдаться его индикативная флуоресценция или изменение сопротивления. Если событие связывания не происходит, тестируемый сосуд можно промыть и процедуру повторять до тех пор, пока соответствующий один из четырех нуклеотидов не свяжется с его комплементом в соответствующем местоположении и не будет наблюдаться его индикативное изменение условий. Когда все четыре нуклеотида добавляются одновременно, каждый может быть помечен различными флуоресцентными метками, и может быть определен нуклеотид, который связывается с его комплементом в соответствующем положении, например по цвету его флуоресценции. Это значительно ускоряет процесс синтеза.

[0095] После того, как произошло событие связывания, указанный комплекс промывают и этапы синтеза повторяют для положения «два». Например, меченый или иным образом маркированный нуклеотид «A» может быть добавлен к реакционной смеси, чтобы определить, является ли комплемент в положении «один» в связанной молекуле-матрице последовательности комплементом для «A», и, если это так, меченый реагент «A» будет связываться с последовательностью матрицы, имеющей этот

комплемент, и, следовательно, будет флуоресцировать, после чего все образцы будут промыты, чтобы удалить любые избыточные количества нуклеотидные реагенты. Там, где произошло связывание, связанный нуклеотид не смывается. Этот процесс будет повторяться для всех нуклеотидов для всех положений до тех пор, пока не будут секвенированы все избыточно перекрытые сегменты нуклеиновой кислоты, например риды, и не будут собраны данные. В качестве альтернативы, когда все четыре нуклеотида добавляются одновременно, каждый из которых помечен различным флуоресцентным индикатором, только один нуклеотид будет связываться с его комплементом в указанном положении, а другие будут вымываться, так что после промывания сосуда указанный сосуд может освещаться лазером, и может быть определено какой нуклеотид связался с его комплементом, например, по цвету его флуоресценции. Однако, когда используется датчик полевого транзистора CMOS, как описано ниже, событие связывания может быть обнаружено по изменению проводимости, которое происходит вблизи сконфигурированного подходящим образом затвора или другой области реакции.

[0096] В частности, из-за необходимости использования оптически детектируемых, например, флуоресцентных, меток в осуществляемых реакциях секвенирования, требуемое оборудование для выполнения такого высокопроизводительного секвенирования может быть громоздким, дорогостоящим, трудоемким в использовании и непереносным. По этой причине в настоящей заявке предлагается новый подход к прямому определению последовательности ДНК и/или РНК без меток. Например, хотя в различных вариантах осуществления предоставлены улучшенные способы для выполнения обработки NGS, в других вариантах осуществления предлагаются улучшенные способы и устройства для секвенирования нуклеиновой кислоты и/или обработки, не обязательно включающие NGS. Например, в частности, в настоящей заявке предлагается способ детектирования, основанный на использовании различных электронных аналитических устройств. Такие методы прямого электронного детектирования имеют ряд преимуществ перед обычной платформой NGS.

[0097] Более конкретно, датчик и/или устройство обнаружения, раскрытое в настоящей заявке, может быть встроено в саму подложку, такую как используется в устройстве «биосистема на чипе», такую как дополнительное устройство на основе металлооксидного полупроводника «CMOS». В частности, при использовании устройства CMOS в генетической детекции сигнал, представляющий событие гибридизации, например, гибридизации и/или секвенирования нуклеиновой кислоты, может быть непосредственно получен и обработан на самом микрочипе. В таком случае автоматическое распознавание возможно достичь в режиме реального времени и с меньшими затратами, чем это возможно достичь в настоящее время с использованием обычной обработки с использованием NGS. Кроме того, стандартные устройства с подложкой CMOS могут использоваться для такого электронного обнаружения, что делает такой процесс простым, недорогим, быстрым и портативным.

[0098] Например, для того чтобы секвенирование следующего поколения стало широко использоваться в качестве диагностического средства в отрасли здравоохранения, необходимо, чтобы секвенирующее оборудование массово производилось с высокой степенью качества, мобильности и экономичности. Одним из способов достижения этого является перевод процесса определения последовательности ДНК/РНК в формат, который полностью использует производственную базу, созданную для компьютерных микросхем, таких как изготовление микросхем на основе комплементарных металл-оксидных полупроводников (CMOS), что является вершиной современного крупномасштабного,

высококачественного недорогого производства с использованием высоких технологий. Для достижения этого в идеале весь сенсорный аппарат секвенатора может быть размещен в стандартном полупроводниковом чипе, изготовленном, например, в тех же самых мощностях, специализирующихся на производстве интегральных схем, которые используются для производства микросхем логики и памяти.

[0099] Соответственно, в другом аспекте раскрытия в настоящем документе представлен полевой транзистор (FET), который может быть изготовлен или иным образом связан с чипом CMOS, который сконфигурирован для использования при выполнении одной или нескольких реакций гибридизации последовательностей ДНК/РНК. Такой полевой транзистор FET может включать в себя затвор, участок канала, соединяющий электроды истока и стока, и изолирующий барьер, который может быть выполнен для отделения затвора от канала. Оптимальная работа такого полевого транзистора зависит от управления проводимостью канала и, таким образом, от контроля тока на стоке, например, посредством напряжения, которое может быть приложено между указанным затвором и указанным электродом истока.

[00100] Для высокоскоростных приложений и в целях повышения чувствительности датчика, FET, представленные в данном документе, могут работать таким образом, чтобы быстро реагировать на изменения напряжения на затворе (VGS). Однако это требует использования коротких затворов и быстро движущихся в этом канале носителей. Ввиду этого современные датчики FET, например, для использования в реакциях гибридизации нуклеиновых кислот и/или реакций секвенирования, сконфигурированы таким образом, чтобы иметь каналы очень тонкие по вертикали и/или по горизонтали, чтобы обеспечить высокую скорость переноса носителей, а также повышенную чувствительность и точность датчиков, что дает существующим датчикам особые преимущества при проведении реакций секвенирования нуклеиновых кислот. Следовательно, устройства, системы и способы их использования, представленные в настоящей заявке, являются идеальными для осуществления анализа геномной информации и в таких приложениях, как секвенирование нуклеиновых кислот и/или генетическая диагностика.

[00101] Следовательно, одним аспектом настоящего изобретения является химически чувствительный транзистор, такой как полевой транзистор (FET), который предназначен для анализа биологических или химических материалов, который решает многие из текущих проблем, связанных с секвенированием нуклеиновых кислот и генетической диагностикой. Такие полевые транзисторы могут быть изготовлены на первичной структуре, такой как пластина, например кремниевая пластина. В различных вариантах, указанная первичная структура может включать в себя одну или несколько дополнительных структур, например, в пакетной конфигурации, такую как слой изоляционного материала. Например, изолирующий материал может быть нанесен поверх первичной структуры и может быть неорганическим материалом, таким как оксид кремния, например, диоксид кремния или нитрид кремния, или органическим материалом, таким как полиимид, BCB или другой подобный материал.

[00102] Первичная и вторичная структуры, например, включающие изолирующий слой, могут включать в себя дополнительную структуру, содержащую один или более из проводящего истока и/или проводящего стока, отделенных друг от друга пространством и встроенных в первичную структуру и/или материал изолятора и/или расположенных в плоскости верхней поверхности изолятора. В различных вариантах указанные структуры могут дополнительно включать в себя процессор или могут быть иным образом связаны с процессором, например, для обработки сгенерированных

данных, таких как данные, полученные с датчиков. Соответственно, структуры могут быть сконфигурированы в виде интегральной схемы или могут иным образом включать в себя интегральную схему, такую как описанная в настоящей заявке, и/или может быть схемой ASIC, структурированной схемой ASIC или схемой FPGA.

[00103] В частности, указанные структуры могут быть выполнены в виде комплементарного металлооксидного полупроводника (CMOS), который, в свою очередь, может быть выполнен в виде химически чувствительного полевого транзистора, содержащего одно или более из следующего: проводящий исток, проводящий сток, канал или лунку и/или процессор. Например, полевой транзистор может включать в себя структуру CMOS, имеющую интегральную схему, которая изготовлена на кремниевой пластине, которая дополнительно включает в себя слой изолятора, причем этот слой изолятора включает в себя проводящий электрод истока и проводящий электрод стока, например, встроенные в него, где электроды истока и стока могут быть выполнены из металла, например, электрода истока из дамасской меди и электрода истока из дамасской меди. В различных вариантах осуществления указанные структуры могут включать в себя поверхность, например верхнюю поверхность, причем эта поверхность может включать в себя канал, например так, что поверхность и/или канал могут быть выполнены проходящими от истока к стоку, формируя тем самым зону реакции.

[00104] В определенных случаях указанная поверхность и/или канал может включать в себя материал одномерного транзистора, материал двумерного транзистора, материал трехмерного транзистора и/или тому подобное. В различных вариантах может быть включен материал одномерного (1D) транзистора, который может состоять из углеродной нанотрубки или полупроводниковой нанопроволоки. В других случаях указанная камера и/или канал состоит из материала одномерного транзистора, содержащего одну или несколько углеродных нанотрубок и/или полупроводниковых нанопроводов, таких как лист полупроводниковых нанопроводов.

[00105] В частности, может быть включен двумерный (2D) транзисторный материал, например, 2D-материал, который может иметь толщину один или два атома и может быть распределен в плоскости. В таких случаях 2D материал может включать или иным образом состоять из таких элементарных 2D материалов, как графен, графин (аллотропная модификация углерода, состоящая из решетки бензольных колец, связанных ацетиленовыми связями), борофен (аллотропная модификация бора), германен (аллотропная модификация германия), герман (еще одна аллотропная модификация германия), силикен (аллотропная модификация кремния), станен (аллотропная модификация олова), фосфорен (аллотропная модификация фосфора, иногда называемая черным фосфором) или одноатомные слои металлов, таких как палладий или родий; дихалькогениды переходных металлов (которые содержат один атом переходного металла на каждые два атома халькогена), такие как дисульфид молибдена (MoS_2, иногда называемый молибденитом), диселенид вольфрама (WSe_2), дисульфид вольфрама (WS_2) или другие; MXены (карбиды и/или нитриды переходных металлов, как правило, формулы M_n+1X_n, где M представляет собой переходный металл и X представляет собой углерод и/или азот), такие как Ti_2C, V_2C, Nb_2C, Ti_3C_2, Ti_3CN, Nb_4C_3 или Ta_4C_3 (более того, MXены могут заканчиваться O, OH или F для получения полупроводников с небольшой шириной запрещенной зоны); или металлоорганические соединения, такие как Ni HITP (Ni_3(2,3,6,7,10,11-гексаиминотрифенилен)$_2$ или супракристаллы 2D (супракристаллами называются надатомные периодические структуры, где атомы,

которые обычно находятся в узлах структуры, замещены их симметричными комплексами). Следует отметить, что дихалькогениды переходных металлов могут содержать один атом любого переходного металла (Sc, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Y, Zr, Nb, Mo, Tc, Ru, Rh, Pd, Ag, Cd, Hf, Ta, W, Re, Os, Ir, Pt, Au, Hg, Rt, Db, Sg, Bh, Mt, Ds или Rg) в паре с двумя атомами любого из халькогенидов (S, Se или Te). В частности, 2D материал может включать один или несколько слоев графена, силикена, дисульфида молибдена, черного фосфора и/или дихалькогенидов металлов. В различных вариантах, трехмерный (3D) материал может быть нанесен на поверхность и/или канал может включать в себя диэлектрический слой.

[00106] Кроме того, в различных вариантах, на поверхности и/или канале может быть расположен реакционный слой, например слой оксида, нанесенный в виде слоя или иным образом нанесенный на 1D слой, 2D слой, например, слой графена, или 3D слой. Такой слой оксида может представлять собой слой оксида алюминия или оксида кремния, такого как диоксид кремния. В различных вариантах на поверхности и/или канале и/или на соответствующий реакционный слой на поверхности и/или канале может быть расположен пассивирующий слой, нанесенный в виде слоя или иным образом нанесенный на 1D слой, 2D слой, например, слой графена, или 3D слой.

[00107] В некоторых вариантах, первичные и/или вторичные структуры могут быть изготовлены или иным образом сконфигурированы так, чтобы включать в себя структуру в виде камеры или структуру в виде лунки в указанной поверхности и/или на ней. Например, структура в виде лунки может быть расположена на участке поверхности, например внешней поверхности, первичной и/или вторичной структуры. В некоторых случаях структура в виде лунки может быть сформирована поверх, по меньшей мере, части материала 1D слоя, 2D слоя, например, графена, и/или 3D слоя, или может включать в себя эти части, и/или может дополнительно включать реакционные слои, например, слои оксида, и/или пассивирующие слои. В различных вариантах, структура в виде камеры и/или луки может образовывать отверстие, например отверстие, которое обеспечивает доступ к внутренней части камеры, например, обеспечивает прямой контакт с 1D слоем, например, углеродной нанотрубкой или нанопроволокой, 2D слоем, например, графеном, поверхностью и/или каналом.

[00108] Соответственно, в различных вариантах осуществления настоящее изобретение относится к биосенсору. Указанный биосенсор включает в себя структуру CMOS, которая может включать металлсодержащий впускной конец, например, из дамасской меди, а также металлосодержащий выпускной конец, например, из дамасской меди, 1D или 2D слой, например, слой графена, поверхность или канал, идущий от электрода истока к электроду стока, и структуру в виде лунки или камеры, которая может быть расположена на участке внешней поверхности слоистой 1D, 2D или 3D структуры в виде лунки. В таком случае структура в виде лунки может быть сконфигурирована так, чтобы образовывать отверстие, которое обеспечивает прямой контакт с нанотрубкой, нанопроволокой и/или графеновой лункой или поверхностью камеры. В различных вариантах на поверхностях камеры может быть расположен слой оксида и/или пассивирующий слой. Следовательно, в некоторых случаях может обеспечиваться химически чувствительный транзистор, такой как полевой транзистор (FET), включающий одну или несколько нано- или микролунок.

[00109] В некоторых вариантах осуществления химически чувствительный полевой транзистор может включать в себя множество лунок и может быть сконфигурирован в виде матрицы, например матрицы датчиков. По существу, указанная система может включать в себя массив лунок, включающий в себя один или более, например, множество

датчиков, где каждый из датчиков включает в себя химически чувствительный полевой транзистор, имеющий исток, сток и реакционную поверхность или канал, проходящий от истока к стоку. Такая матрица или такие матрицы могут быть использованы, например, для обнаружения присутствия и/или изменения концентрации различных типов анализируемых веществ в широком спектре химических и/или биологических процессов, включая реакции гибридизации ДНК/РНК и/или реакции секвенирования. Например, описанные настоящей заявке устройства и/или системы, включающие в себя такие устройства, могут использоваться в способе диагностики заболевания и/или анализа биологических или химических материалов, таких как анализ всего генома, анализ типирования генома, анализ на микроматрицах, панельный анализ, экзомный анализ, микробиомный анализ и/или клинический анализ, такой как анализ рака, анализ NIPT и/или анализ UCS.

[00110] В конкретном варианте осуществления FET может представлять собой FET матрицу графена (gFET), описанную в настоящей заявке, и может использоваться для опрощения методик секвенирования ДНК/РНК и/или гибридизации, основанных на мониторинге изменений концентрации ионов водорода (pH), изменений концентраций других анализируемых веществ и/или событий связывания, связанных с химическими процессами, связанными с синтезом ДНК/РНК, например, в управляемой реакционной камере или в лунке датчика на основе gFET. Например, химически чувствительный полевой транзистор может быть сконфигурирован как биосенсор CMOS и/или может быть адаптирован для увеличения чувствительности измерения и/или точности датчика и/или соответствующей матрицы (матриц), например, путем включения одного или более поверхностей или лунок, имеющих поверхность, на которую нанесен 1D и/или 2D и/или 3D материал, диэлектрический или реакционный слой, пассивирующий слой и/или подобный им слой. Например, в конкретном варианте осуществления предоставляется химически чувствительный графеновый полевой транзистор (gFET), такой как gFET, имеющий структуру CMOS, где указанный датчик gFET, например, биосенсор, может включать в себя слой оксида и/или пассивирующий слой, такой как слой, расположенный на поверхности лунки или камеры, для повышения чувствительности измерения и/или точность датчика и/или соответствующих матриц. Слой оксида, если он присутствует, может состоять из оксида алюминия, оксида кремния, диоксида кремния и подобных им соединений.

[00111] Система может дополнительно включать в себя один или более текучих компонентов, необходимых для осуществления реакции, компонента схемы, необходимого для запуска процессов реакции, и/или вычислительного компонента для управления и/или обработки. Например, жидкий компонент может быть включен, когда жидкий компонент сконфигурирован для управления одним или более потоками реагентов над матрицей и/или одной или несколькими ее камерами. В частности, в различных вариантах осуществления указанная система включает в себя множество мест проведения реакции, таких как поверхности или лунки, которые, в свою очередь, включают в себя множество датчиков и/или множество каналов и дополнительно включают в себя один или более источников жидкости, содержащих жидкость, содержащую множество реагентов и/или анализируемых веществ для доставки на одну или более поверхностей и/или лунок для проведения в них одной или более реакций. В некоторых случаях также может быть включен механизм создания одного или более электрических и/или магнитных полей.

[00112] Система может дополнительно включать в себя компонент схемы, который может включать в себя схему выборки и хранения, декодер адреса, схему смещения и/

или, по меньшей мере, один аналого-цифровой преобразователь. Например, схема выборки и хранения может быть сконфигурирована для хранения аналогового значения напряжения, которое должно быть приложено к выбранному столбцу и/или строке матрицы устройства согласно настоящему изобретению, например, во время интервала считывания. Кроме того, декодер адреса может быть сконфигурирован для создания сигналов выбора столбца и/или строки для какого-либо столбца и/или строки матрицы, для получения доступа к датчику с заданным адресом в указанной матрице. Схема смещения может быть связана с одной или более поверхностями и/или камерами матрицы и может включать в себя компонент смещения, который может быть приспособлен для приложения напряжения считывания и/или напряжения смещения к выбранным химически чувствительным полевым транзисторам матрицы, например, к электроду затвора указанного транзистора. Аналого-цифровой преобразователь может быть сконфигурирован для преобразования аналогового значения в цифровое значение.

[00113] Также может быть включен вычислительный компонент, например, когда указанный вычислительный компонент включает в себя один или более процессоров, таких как процессор сигналов; модуль вызова базы, сконфигурированный для определения одного или более оснований одного или более ридов секвенированной нуклеиновой кислоты; модуль картирования, сконфигурированный для генерации одного или более начальных значений из одного или нескольких ридов секвенированных данных и для выполнения функции картирования одного или более начальных значений и/или ридов; модуль выравнивания, сконфигурированный для выполнения функции выравнивания на одном или более отображаемых ридах; модуль сортировки, сконфигурированный для выполнения функции сортировки одного или более сопоставленных и/или выровненных ридов; и/или модуль вызова варианта, сконфигурированный для выполнения функции вызова варианта на одном или более сопоставленных, выровненных и/или отсортированных ридах. В конкретных вариантах, базовый вызывающий элемент базового вызывающего модуля может быть сконфигурирован для коррекции множества сигналов, таких как потеря фазы и сигнала, для нормализации по ключу и/или для генерации множества исправленных базовых вызовов для каждого потока в каждом датчике для получения множества ридов последовательностей. В различных вариантах осуществления указанное устройство и/или система может включать в себя, по меньшей мере, один электрод сравнения.

[00114] В частности, указанная система может быть сконфигурирована для осуществления реакции секвенирования. В таком случае устройство секвенирования FET может включать в себя матрицу датчиков, имеющих один или более химически чувствительных полевых транзисторов, связанных с ними. Такие транзисторы могут включать каскадный транзистор, имеющий один или более электродов истока, электродов стока и/или электродов затвора. В таком случае электрод истока указанного транзистора может быть прямо или косвенно связан с электродом стока указанного химически чувствительного полевого транзистора. В некоторых случаях может быть включен одно- или двухмерный канал, который может проходить от электрода истока до электрода стока, например, когда материал 1D-канала может представлять собой углеродную нанотрубку или нанопроводу, а материал двумерного канала может состоять из графена, силикена, фосфорена, дисульфида молибдена и дихалькогенида металла. Указанное устройство может быть дополнительно сконфигурировано так, чтобы включать в себя множество линий столбцов и строк, соединенных с датчиками в матрице датчиков. В таком случае каждый столбец во множестве столбцов может быть прямо или косвенно связан с электродами стока указанных транзисторов,

например, каскадных транзисторов, соответствующего множества пикселей в матрице, и аналогично каждая строка во множестве строк может быть прямо или косвенно связана с электродами истока указанных транзисторов, например, каскадных транзисторами, соответствующего множества датчиков в матрице.

[00115] В некоторых случаях может быть использовано множество электродов истока и стока, имеющих множество реакционных поверхностей, например, элементов канала, расположенных между ними, когда каждый элемент канала включает в себя одно-, двух- или трехмерный материал. В таком случае множество первых и/или вторых проводящих слоев может быть соединено с первым и вторым электродами истока/стока химически чувствительных полевых транзисторов в соответствующих столбцах и строках в матрице. Кроме того, может быть предусмотрена схема управления, которая может быть соединена с множеством столбцов и строк, например, для считывания информации с конкретного датчика, подключенного к конкретному столбцу и/или конкретной строке. Схема также может включать в себя компонент смещения, который может быть сконфигурирован для приложения напряжения считывания к выбранной строке и/или для приложения напряжения смещения к электроду затвора транзистора, такого как полевой транзистор и/или каскадный транзистор конкретного датчика. В конкретном варианте осуществления схема смещения может быть соединена с одной или несколькими камерами матрицы и может быть выполнена с возможностью применения смещения считывания к выбранным химически чувствительным полевым транзисторам через проводящий столбец и/или строку. В частности, схема смещения может быть сконфигурирована для подачи напряжения считывания на конкретную строку и/или для подачи напряжения смещения на электрод затвора транзистора, например, каскадного транзистора, во время интервала считывания.

[00116] К матрице для считывания заряда, связанного с одной или несколькими конфигурациями затворов выбранного химически чувствительного полевого транзистора может быть добавлена и присоединена схема считывания. Схема считывания также может быть сконфигурирована для считывания конкретного датчика на основе уровня выборочного напряжения в конкретной строке и/или строке столбца. В таком случае схема считывания может включать в себя одну или несколько схем предварительной зарядки, например, для предварительной зарядки конкретного столбца до уровня напряжения предварительной зарядки передо интервала считывания; и схему выборки, например, для выбора уровня напряжения на электроде стока выбранного транзистора, например, каскадного транзистора, во время интервала считывания. Схема выборки также может быть включена и содержать схему выборки и удержания, сконфигурированную для хранения аналогового значения напряжения на конкретном столбце в течение интервала считывания, и может дополнительно включать в себя аналого-цифровой преобразователь для преобразования аналогового значения в цифровое значение.

[00117] В другом аспекте, указанные интегральные схемы 1D, 2D или 3D FET, например gFET, датчики и/или массивы согласно настоящему изобретению, могут быть изготовлены, например, с использованием любого подходящего метода комплементарной обработки металл-оксидного полупроводника (CMOS), известного из уровня техники. В определенных случаях такая технология обработки CMOS может быть сконфигурирована для увеличения чувствительности измерения и/или точности датчика и/или матрицы, и в то же время для получения датчика со значительно меньшими размерами и области датчика камеры с плотным gFET. В частности, усовершенствованные технологии изготовления, описанные в настоящей заявке, с

использованием 1D, 2D, 3D и/или оксида в качестве реакционного слоя, обеспечивают быстрое получение данных от маленьких датчиков в больших и плотных массивах датчиков. В конкретных вариантах осуществления, когда используется ионоселективная проницаемая мембрана, мембранный слой может включать полимер, такой как перфторсульфоновый материал, перфторкарбоновый материал, PEEK, PBI, Nafion и/или PTFE. В некоторых вариантах осуществления ионоселективная проницаемая мембрана может включать неорганический материал, такой как оксид или стекло. Один или несколько различных слоев, например реакционных, пассивирующих и/или проницаемых мембранных слоев, могут быть изготовлены или иным образом нанесены методом центрифугирования, анодирования, PVD и/или методом «золь-гель».

[00118] Соответственно, описанное в настоящей заявке устройство полевого транзистора CMOS может быть использовано для секвенирования образца нуклеиновой кислоты, когда образец нуклеиновой кислоты служит в качестве матрицы для синтеза и секвенирования ДНК/РНК, которые могут быть связаны или могут располагаться близко к поверхности зоны реакции, например, поверхности, покрытой графеном. После иммобилизации последовательность матрицы может быть секвенирована и/или проанализирована посредством выполнения одного или нескольких из следующих шагов. Например, могут быть добавлены праймер и/или полимераза, например, ДНК- и/или РНК-полимераза, и/или один или несколько субстратов, например дезоксинуклеотидтрифосфаты дАТФ, дГТФ, дЦТФ и дТТФ, последовательно в реакционную камеру, например после начала реакции гибридизации, чтобы вызвать реакцию удлинения. Как только соответствующий субстрат гибридизуется с ее комплементом в последовательности шаблона, происходит сопутствующее изменение индивидуального электрического характеристического напряжения, например, напряжения истока-стока (Vsd), измеренного в результате нового локального эффекта затвора. Если добавлен реакционный слой, такой как оксидный слой, нанесенный на 1D, 2-D или 3-D поверхность, чувствительность, с которой происходит регистрация события связывания, может быть усилена, например, когда реакционный слой сконфигурирован для получения и/или мониторинга изменений концентрации ионов водорода (pH), изменений концентрации других анализируемых соединений.

[00119] Следовательно, для каждой реакции удлинения с соответствующим, например, комплементарным субстратом, будет происходить изменение характеристического напряжения и/или концентрации pH. Например, как описано в настоящей заявке, полевое устройство для секвенирования нуклеиновой кислоты и/или обнаружения гена может быть расположено в камере для образца или лунке проточной кюветы, и раствор для образца, например, содержащий полимеразу и один или несколько субстратов, например нуклеиновые кислоты, могут быть введены в камеру с раствором образца, например, через один или несколько жидкостных компонентов системы. В различных вариантах осуществления электрод сравнения может быть расположен выше по потоку, ниже по потоку или в жидкостном контакте с полевым устройством, и/или электроды истока и/или стока могут сами по себе служить в качестве электродов, использующихся для обнаружения гибридизации, и напряжение для затвора может подаваться когда это нужно.

[00120] В частности, в типичной реакции удлинения, описанной выше, синтезируются полинуклеотиды, если добавленный субстрат является комплементарным последовательности основания целевой матрицы ДНК/РНК и/или матрицы праймера. Если добавленный субстрат не комплементарен следующей доступной последовательности основания в шаблоне, гибридизация и удлинение не происходят.

Поскольку нуклеиновые кислоты, такие как ДНК и РНК, имеют отрицательный заряд в водных растворах, гибридизация, приводящая к удлинению, может последовательно обнаруживаться за счет изменения плотности заряда на поверхности реакции и/или в реакционной камере. Такое обнаружение может быть улучшено за счет возможности обнаружения увеличения концентрации ионов, например, путем обнаружения изменения pH. Поскольку субстраты добавляются последовательно, можно легко определить, какой нуклеотид связан с матрицей, тем самым облегчая реакцию удлинения. Соответственно, в результате удлинения отрицательный заряд на поверхности затвора с графеновым слоем, поверхности изолирующей пленки и/или поверхности боковой стенки реакционной камеры будет увеличен. Такое увеличение затем может быть обнаружено, например, за счет изменения напряжения затвора истока и/или концентрации ионов, как подробно описано в настоящей заявке. Устанавливая добавление того субстрата, который привел к изменению сигнала или pH в напряжении затвор-исток, можно определить и/или проанализировать последовательность оснований целевой нуклеиновой кислоты.

[00121] В частности, независимо от используемого устройства для секвенирования, такого как устройство секвенирования на основе NGS и/или FET, как описано в настоящей заявке, этот итеративный процесс синтеза продолжается до тех пор, пока вся матричная цепь ДНК/РНК не будет реплицирована в указанном сосуде. Обычно типичная длина последовательности, реплицируемой таким образом, составляет от около 100 до около 500 пар оснований, например от 150 до 400 пар оснований, в том числе от около 200 до около 350 пар оснований, например от около 250 пар оснований до около 300 пар оснований, в зависимости от используемого протокола секвенирования. Кроме того, длина нуклеотидов этих сегментов матрицы может быть заранее определена, например, сконструирована, чтобы соответствовать любому конкретному механизму секвенирования и/или протоколу, в соответствии с которым он выполняется.

[00122] Конечным результатом является считывание или прочитывание, которое состоит из реплицированного сегмента ДНК/РНК, например, длиной от около 100 до около 1000 нуклеотидов или более, который либо был мечен таким образом, что каждый нуклеотид в последовательности, например, рид, известен на основании своей метки, или определен и известен по изменению характеристики затвора, такому как изменение напряжения и/или pH. Следовательно, поскольку человеческий геном состоит из примерно 3,2 миллиарда пар оснований, а различные известные протоколы секвенирования обычно приводят к получению меченых реплицированных последовательностей, например, ридов, длиной от примерно 100 или 101 оснований до примерно 250, или примерно 300, или примерно 400 оснований, общее количество полученных сегментов, которые необходимо упорядочить, и, следовательно, общее количество генерируемых ридов, может составлять от примерно 10000000 до примерно 40000000, например, от примерно 15000000 до примерно 30000000, в зависимости от того, насколько длинны реплицированные меченые последовательности. Следовательно, секвенатор может обычно генерировать около 30000000 ридов, например, когда длина считывания составляет 100 нуклеотидов в длину, чтобы охватить указанный геном один раз. Однако, как указано в настоящей заявке, из-за сжатой природы формата настоящей микросхемы для определения последовательности, представленной в настоящей заявке, могут быть достигнуты гораздо более существенные длины считывания, такие как 800 оснований, 1000 оснований, 2500 оснований, 5000 оснований, вплоть до 10000 оснований.

[00123] Кроме того, как указано выше, в таких процедурах может оказаться полезным

избыточный набор ридов ДНК/РНК, больший примерно в 5 раз, или примерно в 10 раз, или примерно в 20 раз, или примерно в 25 раз, или примерно в 30 раз, или примерно в 40 раз, или примерно в 50 раз, или примерно в 100 раз, или примерно в 200 раз, или примерно в 250 раз, или примерно в 500 раз, или примерно в 1000 раз, или примерно в 5000 раз, или даже примерно в 10000 раз или более, и, таким образом, объем первичной обработки, который необходимо выполнить, и время, затрачиваемое на это, могут быть весьма значительным. Например, с 40-кратным избытков информации, при использовании которого различные синтезированные риды сконструированы так, чтобы в некоторой степени перекрываться, может потребоваться синтез до 1,2 миллиарда ридов. Как правило, подавляющее большинство, если не все эти меченые последовательности, могут генерироваться параллельно. Конечным результатом является то, что исходный биологический генетический материал обрабатывается, например, с помощью протокола секвенирования, такого как те, которые кратко изложены в настоящей заявке, и генерируется цифровое представление этих данных, при этом указанное цифровое представление данных может подвергаться протоколу первичной обработки.

[00124] В частности, генетический материал субъекта может быть реплицирован и его последовательность определена таким образом, что генерируется измеряемый электрический, химический, радиоактивный и/или оптический сигнал, который затем преобразуется, например, секвенатором и/или устройством обработки, связанным с ним, в цифровое представление генетического кода субъекта. Более конкретно, первичная обработка может включать в себя преобразование изображений, таких как записанные вспышки света или другие данные электрических или химических сигналов, в данные файла FASTQ. Соответственно, эта информация сохраняется в виде файла FASTQ, который затем может быть отправлен для дальнейшей, например, вторичной обработки. Типичный файл FASTQ включает в себя большой набор ридов, представляющих нуклеотидные последовательности с цифровой кодировкой, в которых каждое предсказанное основание в последовательности было определено и дано значение вероятности того, что указанное основание в указанной позиции неверно.

[00125] Во многих случаях может оказаться полезной дополнительная обработка указанных последовательностей с цифровым кодированием, полученных из секвенатора и/или протокола секвенирования, например, путем осуществления вторичной обработки данных, представленных в цифровом виде. Эта вторичная обработка, например, может быть необходима для того, чтобы собрать полный геномный профиль индивида, например, когда определяются все генетические факторы индивида, где определяется положение в последовательности каждого нуклеотида каждой хромосомы так, чтобы состав всего генома человека был идентифицирован. При такой обработке геном индивида может быть собран, например, путем сравнения с референсным геном, таким как стандарт, например, один или несколько геномов, полученных в рамках проекта генома человека, чтобы определить, чем генетические характеристики индивида отличаются от референсного генома(ов). Этот процесс широко известен как определение вариантов. Поскольку разница между ДНК/РНК различных лиц составляет от 1 до 1000 пар оснований, такой процесс определения вариантов может быть очень трудоемким и времязатратным.

[00126] Соответственно, в типичном протоколе вторичной обработки генетическое строение субъекта собирают посредством сравнения с референсным геномом. Это сравнение включает реконструкцию генома индивида из миллионов и миллионов коротких последовательностей рида и/или сравнение всего ДНК и/или РНК индивида

с примером модели последовательности ДНК и/или РНК. В типичном протоколе вторичной обработки из секвенатора принимают файл FASTQ, содержащий необработанные секвенированные данные рида. Например, в некоторых случаях может иметься до 30000000 или более ридов, охватывающих геном субъекта, исходя из того, что отсутствует избыточная выборка образцов, например, когда каждый рид имеет длину приблизительно 100 нуклеотидов. Следовательно, в таком примере, чтобы сравнить геном ДНК/РНК субъекта со стандартным референсным геномом, необходимо определить, где каждый из этих ридов картируется на референсный геном, например, как каждый из них выравнивается относительно другого и/или как каждый рид может быть также отсортирован по порядку хромосом, чтобы определить, в какой позиции находится каждый рид, и какой хромосоме он принадлежит. Одна или более из этих функций могут предшествовать выполнению функции определения вариантов на полноразмерной последовательности. После того, как определено, какой части генома принадлежит каждый рид, можно определить полноразмерную генетическую последовательность, а затем можно оценить различия между генетическим кодом субъекта и генетическим референсным кодом.

[00127] Так как длина человеческого генома составляет более 3 миллиардов пар оснований, были разработаны эффективные автоматизированные протоколы и оборудование для секвенирования для выполнения секвенирования таких геномов ДНК/РНК в течение периодов времени, которые могут быть полезными с клинической точки зрения. Таким инновации в автоматизированном секвенировании обеспечили возможность секвенирования всего генома за считанные часы или дни в зависимости от количества секвенируемых геномов, с учетом избыточной выборки образцов, и количества обрабатывающих ресурсов, вовлеченных в работу. Следовательно, учитывая этот прогресс в сгенерировании, большое количество секвенируемых данных может быть сгенерировано за относительно короткий период времени. Результат такого прогресса, однако, развитие ограничивающего фактора («бутылочного горлышка») на стадии вторичной обработки. В попытках помочь преодолеть эти ограничивающие факторы были разработаны различные алгоритмы на основе программного обеспечения, такие как описаны в настоящем документе, для способствования ускорению процесса сборки секвенированных ДНК и/или РНК субъекта, например посредством процесса основанной на референсе сборки.

[00128] Например, основанная на референсе сборка в типичном протоколе сборки вторичной обработки включает в себя сравнение секвенированной геномной ДНК и/или РНК субъекта с секвенированной геномной ДНК и/или РНК одного или более стандартов, например, известных референсных последовательностей. В качестве помощи для ускорения этого процесса разработаны различные алгоритмы. Обычно данные алгоритмы включают в себя некоторый вариант одного или более из: картирования, выравнивания и/или сортировки миллионов ридов, полученных из цифровых файлов, например файлов FASTQ, которые переданы секвенатором, для определения, к какому месту каждой хромосомы относится или как другим образом на ней расположен каждый конкретный рид. Часто общей особенностью функционирования этих различных алгоритмов является использование ими индекса и/или массива для ускорения их функции обработки.

[00129] Например, что касается картирования, большое количество секвенированных ридов, например, все, могут быть обработаны для определения возможных местоположений в референсном геноме, на который могли бы быть выровнены эти риды. Один из методов, который может быть использован в этих целях, заключается

в прямом сравнении рида с референсным геномом, чтобы найти все позиции совпадения. Другой метод состоит в использовании массива префиксов или суффиксов или построении дерева префиксов или суффиксов с целью картирования ридов на различные позиции в референсном геноме ДНК/РНК. Типичным алгоритмом, полезным при выполнении такой функции, является преобразование Барроуза-Уилера, которое используют для картирования ридов на референс с помощью формулы сжатия, которая сжимает повторяющиеся последовательности данных.

[00130] Еще один метод заключается в использовании хэш-таблицы, например, когда выбранное подмножество ридов, k-мер выбранной длины «k», например, затравку, помещают в хэш-таблицу в качестве ключей, а референсную последовательность разбивают на части, равные по длине k-меру, и эти части и их местоположения вставляют с помощью алгоритма в хэш-таблицу в те места таблицы, на которые они отображаются в соответствии с функцией хэширования. Типичным алгоритмом для выполнения этой функции является «BLAST», Инструмент для поиска базового локального выравнивания (Basic Local Alignment Search Tool). Такие программы на основе хэш-таблицы сравнивают исследуемые нуклеотидные или белковые последовательности с одной или более баз данных стандартных референсных последовательностей и вычисляют статистическую значимость совпадений. Подобным образом можно определить вероятное местоположение любого данного рида относительно референсного генома. Эти алгоритмы полезны, поскольку они требуют меньше памяти, преобразований и, следовательно, требуют меньше вычислительных ресурсов и времени при выполнении своих функций, чем было бы в ином случае, например, если бы геном субъекта собирали путем прямого сравнения, например без использования этих алгоритмов.

[00131] Кроме того, может быть выполнена функция выравнивания для определения всех возможных местоположений картирования данного рида на геном, например в тех случаях, когда рид можно картировать на множество позиций в геноме, которые в действительности являются местоположением, из которого он был фактически получен, например путем секвенирования с этого места с помощью исходного протокола секвенирования. Эту функцию можно выполнить на ряде ридов генома и можно получить строку упорядоченных нуклеотидных оснований, представляющую частично или полностью геномную последовательность ДНК и/или РНК субъекта. Наряду с упорядоченной генетической последовательностью каждой нуклеотидной позиции можно присвоить оценку, представляющую для любой данной нуклеотидной позиции вероятность того, что нуклеотид, например, «A», «C», «G», «T» (или «U»), предполагаемый в этой позиции, действительно является нуклеотидом, который принадлежит этой назначенной позиции. Типичными алгоритмами для выполнения функция выравнивания являются алгоритмы Нидлмана-Вунша и Смита-Ватермана. В любом случае эти алгоритмы выполняют выравнивания последовательностей между строкой исследуемой геномной последовательности ДНК и/или РНК субъекта и строкой референсной геномной последовательности, тем самым вместо сравнения полногеномных последовательностей друг с другом сравнивают выбранные сегменты возможных длин.

[00132] После того как ридам назначены позиции, например, относительно референсного генома, что может включать в себя определение принадлежности рида конкретной хромосоме и/или его смещения от начала этой хромосомы, риды можно отсортировать по позиции. Это может позволить в последующих анализах использовать преимущества избыточной выборки, описанные в настоящем документе. Все риды, которые перекрывают данную позицию в геноме, будут рядом друг с другом после

сортировки и могут быть организованы в скопление (pileup) и без труда исследованы, чтобы определить, согласуются ли большинство из них с референсным значением или нет. Если нет, вариант можно отметить флагом.

[00133] Хотя данные и другие подобные им алгоритмы решают ограничивающие факторы, присущие вторичной обработке, по своему, более быстрое время работы и большая точность все еще являются необходимыми. В частности, хотя в выработке необработанных данных, например в вырабатываемых данных последовательности ДНК/РНК, есть развитие, развитие информационных технологий не стоит на месте, что приводит к ограничивающим факторам при анализе данных. Указанные ограничивающие факторы в некторой степени уменьшены благодаря разработке различных алгоритмов, например таких, как описаны выше, которые помогают ускорить данный анализ, однако все-равно существует потребность в новых технологиях для выполнения выработки и получения данных, вычисления, хранения и/или анализа таких данных, в особенности это относится к анализу геномной последовательности, например к этапу вторичной обработки.

[00134] Например, с применением стандартных технологий нового поколения секвенирование человеческого генома может занять несколько часов, вплоть до суток, а при использовании стандартных протоколов для выполнения вторичной обработки таких полученных секвенированных геномных данных обработка секвенированных данных может занять до трех (3) суток или даже до недели или более для выработки релевантной с клинической точки зрения информации о геномной последовательности индивида. С применением различных оптимизированных устройств, алгоритмов, способов и/или систем время, затраченное на обработку, от первичной до вторичной, может быть снижено и составлять лишь от 27 до 48 часов. Однако для достижения таких быстрых результатов обычно требуется по существу одновременная и параллельная обработка всех выработанных ридов, например 30 миллионов ридов по 100 нуклеотидов каждый. Такая параллельная обработка требует дорогих обрабатывающих мощностей, включая массивные ресурсы ЦПУ, и все еще занимает относительно много времени.

[00135] Кроме того, в различных примерах необходима повышенная точность рещультатов. Такая повышенная точность может быть достигнута за счет обеспечения некоторой избыточной выборки секвенируемого генома. Например, как описано выше, может быть необходимо обрабатывать ДНК субъекта таким образом, что в любом данном местоположении в последовательности нуклеотидов имелась избыточная выборка этой области. Как указано выше, может требоваться избыточная выборка в любой данной области генома, включая в 10X, 15X, 20X, 25X, 30X, 40X, 50X, 100X, 250X, или даже 500X или 1000X раз или более. Однако, когда имеется избыточная выборка генома, например в 40X раз, количество ридов, которые необходимо обработать, составляет приблизительно 30 миллионов x 40 (в зависимости от длины ридов), что приводит к тому, что необходимо обработать 1,2 миллиарда ридов, когда имеется избыточная выборка генома в 40X раз. Следовательно, хотя такая избыточная выборка обычно приводит к большей точности, это происходит ценой большего времени и требует более масштабных обрабатывающих ресурсов, так как каждый участок генома покрыт от 1 до 40 раз. Более того, в некоторых случаях применения в онкологии, в которых медицинский специалист пытается отличить мутировавший геном раковых клеток в потоке крови от генома здоровых клеток, может использоваться избыточная выборка в 500X, 1000X, 5000X или даже в 10000X раз.

[00136] Настоящее изобретение, таким образом, направлено на такие новые

технологии, которые могут быть реализованы в одном геномном и/или биоинформационном протоколе, например в конвейере, или в их последовательности, для выполнения генетического получения и/или анализа, например первичной и/или вторичной обработки, полученных геномных последовательных данных или их части. Последовательные данные могут быть получены непосредственно от автоматизированных систем секвенирования высокой производительности, например посредством автоматизированного секвенатора 454 «Секвенирования посредством синтеза» («Sequencing by Synthesis») фирмы ROCHE, автоматизированных секвенаторов HiSeq x Ten или Solexia фирмы ILLUMINA, секвенаторов «Секвенирование посредством лигирования и обнаружения олигонуклеотида» («Sequencing by Oligonucleotide Ligation and Detection» (SOLiD)) или Ионное полупроводниковое секвенирование (Ion Torrent) фирмы LIFE TECHNOLOGIES и/или секвенатора «Флуоресцентное секвенирование одной молекулы» («Single Molecule Fluorescent Sequencing») фирмы HELICOS GENETIC ANALYSIS SYSTEMS, или тому подобного, например, посредством прямого соединения с секвенирующим обрабатывающим блоком, или секвенированные данные могут быть получены напрямую, например при секвенировании на конфигурации чипа, такой как графеновый слоистый датчик на полевых транзисторах, содержащий секвенирующий чип на комплементарных металло-оксидных полупроводниках (CMOS), как описано в настоящем документе. Такие секвенированные данные также могут быть получены удаленно, например из базы данных, например через Интернет, или из другого удаленного мета, доступного посредством беспроводных протоколов связи, например по WiFi, Bluetooth или тому подобному.

[00137] В соответствии с определенными аспектами эти технологии генетического получения и/или анализа могут использовать усовершенствованные алгоритмы, которые могут быть реализованы программным обеспечением, которое выполняется с менее интенсивной обработкой, и/или с меньшими временными затратами, и/или более высоким процентом точности. Например, в некоторых вариантах реализации обеспечены усовершенствованные устройства и способы производства информации генетической последовательности, например как в протоколах первичной обработки, как раскрыто в настоящем документе, и/или усовершенствованные алгоритмы выполнения над ней вторичной обработки, как раскрыто в настоящем документе. В различных конкретных вариантах реализации усовершенствованные устройства, системы, способы их использования и применяемые алгоритмы направлены на более эффективное и/или более точное выполнение одной или более из функций секвенирования, картирования, выравнивания и/или сортировки, например для выработки и/или анализа цифрового представления данных последовательности ДНК/РНК, полученных от платформы секвенирования, например, в формате файла FASTQ, полученного из автоматизированного секвенатора и/или секвенатора на чипе, такого как один из описанных выше.

[00138] Кроме того, в некоторых вариантах реализации предложены усовершенствованные алгоритмы, направленные на более эффективное и/или более точное выполнение одной или более из функций локального повторного выравнивания, маркировки дубликатов, перекалибровки качественной оценки оснований, определения вариантов, сжатия и/или распаковки. Кроме того, как более подробно описано ниже в настоящем документе, согласно определенным аспектам эти технологии генетического производства и/или анализа могут использоваться по одному или более алгоритмам, таким как усовершенствованные алгоритмы, которые могут быть реализованы с помощью аппаратного обеспечения, которые выполняются с менее интенсивной

обработкой, и/или с меньшими временными затратами, и/или более высоким процентом точности, чем различные программные реализации для выполнения того же самого.

[00139] В конкретных вариантах реализации предложена платформа технологий для секвенирования ДНК/РНК для получения данных генетической последовательности и/или для выполнения генетических анализов, причем платформа может включать в себя выполнение одной или более из функций: секвенирования, картирования, выравнивания, сортировки, локального повторного выравнивания, маркировки дубликатов, перекалибровки качественной оценки основания, определения вариантов, сжатия и/или распаковки, и/или может также включать протоколы третичной обработки, как описано в настоящем документе. В определенных случаях реализация одной или более из этих функций платформы предназначена для генерации и/или выполнения одного или более из определения и/или реконструкции консенсусной геномной последовательности субъекта, сравнения геномной последовательности субъекта с референсной последовательностью, например, референсной или модельной генетической последовательностью, определения того, каким образом геномная ДНК и/или РНК субъекта отличается от референсной, например, определения вариантов, и/или для выполнения третичного анализа на геномной последовательности субъекта, например, для полного анализа генома, например, для анализа вариации по всему геному и/или для анализа типирования генома, функционального анализа генов, функционального анализа белков, например, анализа связывания белков, численного и/или сборного анализа геномов и/или транскриптомов, анализа микропанели, анализа панелей, анализа экзом, микробиомного анализа и/или клинического анализа, например анализа рака, анализа с неинвазивным пренатальным тестированием (NIPT) и/или анализа UCS, а также для различных анализов диагностической, и/или профилактической, и/или терапевтической оценки.

[00140] В частности, после того, как генетические данные сформированы и/или обработаны, например, в одном или более протоколах первичной и/или вторичной обработки, например, картированы, выровнены и/или отсортированы, например, для создания одного или более файлов определения вариантов, например, для определения того, как данные генетической последовательности субъекта отличаются от одной или более референсных последовательностей, согласно другому аспекту настоящее изобретение может относиться к выполнению одной или более других аналитических функций над сформированными и/или обработанными генетическими данными, например, для дальнейшей обработки, такой как третичная обработка. Например, система может быть выполнена с возможностью выполнения дальнейшей обработки сгенерированных и/или обрабатываемых во вторую очередь данных, например посредстом их прогона через один или более конвейер для третичной обработки, такой как один или более из конвейера генома, конвейера эпигенома, конвейера метагенома, совместного генотипирования, конвейера MuTect2 или другого конвейера для третичной обработки, например посредством устройств и способов, раскрытых в настоящем документе. Например, в различных примерах может быть обеспечен дополнительный уровень обработки, например для диагностики заболевания, терапевтического лечения и/или профилактического предупреждения, включая, например, неинвазивное пренатальное тестирование (NIPT), реанимацию и интенсивную терапию новорожденных (NICU), рак, проводимые в лаборатории исследования (LDT), агробиологию (AgBio) и другие виды диагностики, профилактики и/или способов лечения таких заболеваний, в которых применяются данные, сгенерированные одним или более из указанных первичных, вторичных и/или третичных конвейеров. Следовательно, устройства и

способы, описанные в настоящем документе, могут быть использованы для формирования данных генетических последовательностей, которые затем могут быть использованы для формирования одного или более файлов определения вариантов и/или другой связанной информации, которая может быть в дальнейшем подвергнута обработке другими конвейерами третичной обработки в соответствии с устройствами и способами, описанными в настоящем документе, например, для диагностики конкретных и/или общих заболеваний, а также для профилактических и/или терапевтических мер и/или методов воздействия на развитие.

[00141] Кроме того, в различных вариантах реализации биоинформационный режим обработки, раскрытый в настоящем документе, может применяться для создания одной или более масок, геномной референсной маски, стандартной маски, маски заболевания и/или маски итерационной обратной связи, которые могут быть добавлены в картировщик и/или выравниватель, например вместе с референсом, причем набор масок выполнен с возможностью идентификации конкретной интересующей области или объекта. Например, в одном варианте реализации способы и устройства, раскрытые в настоящем документе, могут применяться для создания геномной референсной маски, например путем создания набора масок, который может быть загружен в картировщик и/или выравниватель вместе с референсом, причем набор масок выполнен с возможностью идентификации особо важных и/или релевантных областей, например для специалиста и/или субъекта, и/или для идентификации областей с повышенной чувствительностью к ошибкам. В различных вариантах реализации набор масок может обеспечивать интеллектуальное направление картировщика и/или выравнивателя, например, на каких областях генома сфокусироваться для повышения качества. Таким образом, маски могут быть созданы слоистым образом для обеспечения изменяемых уровней или итераций направления на основе различных конкретных случаев применения. Соответственно, каждая маска может идентифицировать интересующие области и обеспечивать минимальный целевой показатель качества для области. Кроме того, стандартная маска может применяться для обеспечения направления, например на идентифицированные, например типичные, «особо важные» области генома. Такие области могут включать известные кодированные области, контролируемые области и т.д., а также области, в отношении которых хорошо известно, что они производят ошибки. Кроме того, маска заболевания, или специфичная для случая применения маска, может применяться в наборе масок, который идентифицирует особо важные области, например, области, которым требуются высокие уровни точности на основании известных маркеров, например, рак. Помимо этого, может применяться маскировка итерационной обратной связи, например путем добавления новой, подходящей к данному случаю маски, которая может быть специально спроектирована посредством использования обратной связи от системы третичного анализа (наподобие Cypher Genomics), которая идентифицирует представляющие интерес обласли на основе наблюдаемых ошибок или несоответствий.

[00142] Как указано выше, согласно одному аспекту одна или более из этих функций платформы, например, функций картирования, выравнивания, сортировки, повторного выравнивания, маркировки дубликатов, перекалибровки качественной оценки основания, определения вариантов, одного или более модулей третичной обработки, сжатия и/или распаковки, выполнены с возможностью реализации в программном обеспечении. В другом варианте реализации одна или более из этих функций платформы, например, функций картирования, выравнивания, сортировки, локального повторного выравнивания, маркировки дубликатов, перекалибровки качественной оценки

основания, распаковки, определения вариантов, третичной обработки, сжатия и/или распаковки, выполнены с возможностью реализации в аппаратном обеспечении.

[00143] Соответственно, в определенных случаях в настоящем документе предложены способы, где способы включают в себя выполнение алгоритма, такого как алгоритм для реализации одной или более функций генетического анализа, таких как картирование, выравнивание, сортировка, повторное выравнивание, маркировка дубликатов, перекалибровка качественной оценки основания, определение вариантов, сжатие и/или распаковка, где алгоритм оптимизирован в соответствии со способом, которым он должен быть реализован. В частности, когда алгоритм должен быть реализован в программном решении, алгоритм и/или обслуживающие его процессы оптимизированы таким образом, чтобы они работали быстрее и/или с более высокой точностью при выполнении этой средой. Аналогичным образом, когда функции алгоритма должны быть реализованы в аппаратном решении, аппаратное обеспечение разработано для выполнения этих функций и/или обслуживающих их процессов оптимальным образом, чтобы работать быстрее и/или с более высокой точностью при выполнении этой средой. Эти способы, например, могут быть использованы, например, в процедуре итеративного определения вариантов.

[00144] Поэтому согласно одному аспекту в настоящем документе предложены системы, устройства и способы для реализации протоколов биоинформатики, например, для выполнения одной или более функций анализа генетических данных, таких как геномные данные, например, посредством одного или более оптимизированных алгоритмов и/или на одной или более оптимизированных интегральных схемах, например на одной или более аппаратных платформах обработки. Таким образом, в одном случае предложены системы и способы для реализации одного или более алгоритмов для выполнения одного или более этапов анализа геномных данных в протоколах биоинформатики, например, когда этапы могут включать в себя выполнение одного или более из: картирования, выравнивания, сортировки, локального повторного выравнивания, маркировки дубликатов, перекалибровки качественной оценки основания, определения вариантов, сжатия и/или распаковки. В другом случае предложены системы и способы для реализации функций одного или более алгоритмов для выполнения одного или более этапов анализа геномных данных в протоколе биоинформатики, как указано в настоящем документе, причем функции реализуются на аппаратном ускорителе, который может быть соединен или не соединен с одним или более процессорами общего назначения и/или суперкомпьютерами.

[00145] Точнее говоря, в некоторых случаях предложены способы для выполнения вторичной аналитики над данными, имеющими отношение к генетическому составу субъекта. В одном случае аналитика, подлежащая выполнению, может включать в себя основанную на референсе реконструкцию генома субъекта. Например, основанное на референсе картирование включает в себя использование референсного генома, который может быть сформирован в результате секвенирования генома одного или множества индивидов, или он может быть объединением принадлежащих различным людям ДНК, которые объединены таким образом, чтобы создать прототипный стандартный референсный геном, с которым можно сравнить ДНК любого индивида, например, для определения и реконструкции генетической последовательности индивида и/или для определения разницы между их генетическим строением и этим стандартным референсом, например, для определения вариантов.

[00146] В частности, причина выполнения вторичного анализа над секвенированной ДНК субъекта состоит в том, чтобы определить, как ДНК субъекта отличается от ДНК

референса. Более конкретно, чтобы определить одно, множество или все отличия нуклеотидной последовательности субъекта от нуклеотидной последовательности референса. Например, отличия между генетическими последовательностями любых двух случайно выбранных людей встречаются 1 раз на 1000 пар оснований, что с учетом свыше 3 миллиардов пар оснований в полном геноме составляет вариацию из до 3000000 отличающихся пар оснований на человека. Определение этих отличий может быть полезным, например, в протоколе третичного анализа, например, для прогнозирования возможности возникновения болезненного состояния, например, вследствие генетического нарушения, и/или вероятности успеха профилактического или терапевтического воздействия, например, на основе того, каким ожидается взаимодействие профилактики или терапии с ДНК субъекта или формируемыми при этом белками. В различных случаях может оказаться полезным выполнение реконструкции генома субъекта как впервые, так и на основе референса, чтобы подкрепить результаты одной результатами другой, и чтобы улучшить точность протокола определения вариантов, если требуется.

[00147] В различных примерах, как указано выше, при выполнении протокола первичной обработки может быть полезным обеспечить избыточную выборку для одной или более областей генома субъекта. Эти области могут быть выбраны на основе известных областей с повышенной вариативностью, области с предполагаемой повышенной вариативностью, например на основании состояния субъекта и/или всего генома в целом. В базовой форме, как указано выше, на основании типа выполняемых протоколов секвенирования, секвенирование производит считывания, например риды, которые являются цифровым представлением кода генетической последовательности субъекта. Эти длины ридов обычно сформированы на основании типа используемого секвенирующего оборудования. Например, автоматизированный секвенатор 454 фирмы ROCHE обычно производит риды длиной от 100 или 150 пар оснований приблизительно до 1000 пар оснований; для некоторых технологий фирмы ILLUMINA длина ридов обычно составляет приблизительно от 100 или 101 приблизительно до 150 пар оснований и 250 пар оснований для других технологий; для технологии SOLiD фирмы LIFE TECHNOLOGIES длина ридов обычно составляет приблизительно от 50 приблизительно до 60 пар оснований, а для технологии Ion Torrent от 35 до 450 пар оснований; а для HELICOS GENETIC ANALYSIS SYSTEMS длина ридов может варьироваться, но обычно может составлять менее 1000 нуклеотидов.

[00148] Однако так как обработка образца днк, требуемая для получения ридов заданной длины специфического размера является трудозатратной, а также затратной с точки зрения химии, и так как секвенирование само по себе часто зависит от работы секвенирующего оборудования, существует возможность появления ошибок в процессе секвенирования, тем самым внося нарушения в ту часть секвенированного генома, в которой возникла ошибка. Такие ошибки могут вызывать проблемы, особенно, когда целью реконструкции генома субъекта является определение он или по меньшей мере часть генома отличается от стандарта или референсной модели. Например, ошибка устройства или химии, приводящая к замене одного нуклеотида, например в риде, другим, даст ложную индикацию вариации, которой на самом деле нет. Это может привести к некоректному определению оснований и также может привести к ложной индикиции состояния заболевания или тому подобного. Соответственно, вследствие возможности ошибки оборудования, химии и/или даже человека при выполнении протокола секвенирования во многих случаях является желательным обеспечение избыточности в системе анализа, например посредством обеспечения избыточности

образцов по всему геному. В частности, при обеспечении автоматизированным секвенатором файла FASTQ, определяющего последовательность ридов, содержащих нуклеотиды в заданных позициях, вместе с вероятностью того, что соответствие определения заданного нуклеотида определенной позиции на самом деле является ложным, например определение основания, зачастую необходимо использовать способы, такие как избыточная выборка образцов, для обеспечения того, что определение оснований, выполненные в процессе секвенирования, могут быть обнаружены и скорректированы.

[00149] Следовательно, при выполнении способов, описанных в настоящем документе, в некоторых случаях, протокол первичной обработки выполняют таким образом, чтобы получать секвенированный геном, в котором для части или для всего генома обеспечена избыточная выборка образцов прблизительно в 10X, 15X, 20X, 25X, 30X, 40X, 50X раз или более. Соответственно, когда длина ридов составляет приблизительно 50-60 пар оснований, такая избыточная выборка образцов может приводить к наличию приблизительно от 2 приблизительно до 2,5 миллиардов ридов, или когда длина ридов составляет приблизительно 100 или 101 пару ридов, избыточная выборка образцов может приводить к наличию приблизительно от 1 приблизительно до 1,2 миллиардов ридов, а когда длина ридов составляет приблизительно 1000 пар оснований, секвенатором могут быть сгенерированы приблизительно от 50 приблизительно до 100 миллионов ридов, например когда избыточная выборка образцов составляет приблизительно 40X раз. В частности, в таком примере вследствие избыточной выборки образцов в 40X раз, ожидается, что в любом заданной месте генома будет иметься 40 ридов, покрывающих любое положение, несмотря на то, что заданная позиция может находиться в начале одного рида, середине другого и конце еще одного, однако ожидается, что оно будет покрыто приблизительно 40 раз.

[00150] Таким образом, такая избыточная выборка образцов обеспечивает области секвенированного генома, которые покрыты множеством ридов, например, дубликатами, например, до 40 ридов, например, когда избыточная выборка образцов составляет приблизительно 40X раз. Указанные по меньшей мере частичные дубликаты являются полезными при определении, является ли любое заданная вариация в любом конкретном риде на самом деле действительной геномной вариацией или лишь артефактом оборудования или химии. Таким образом, избыточная выборка образцов может применяться для повышения точности при реконструкции генома субъекта, особенно в случаях, когда геном субъекта необходимо сравнить с референсным геномом для определения тех случаев, когда генетическая последовательность субъекта отличается от референсной генетической последовательности. Таким образом, как описано более подробно ниже в настоящем документе, можно подтвердить, что любая данная вариация между реконструированной последовательностью и моделью на самом деле вызвано наличием действительного варианта, а не ошибкой первичной обработки образца ДНК, программным обеспечением для выравнивания ридов и т.д.

[00151] Например, при построении генетической последовательности секввенированной ДНК субъекта, необходимо определить, где должен быть какой нуклеотид в растущей строке нуклеотидов. Для определения того, где должен быть какой нуклеотид, различные риды могут быть организованы и может быть получено скопление ридов, покрывающих дублирущиеся положения. Это позволяет выполнить сравнение всех ридов, покрывающих одни и те же положения, для более точного определения, имеется ли действительный вариант в какой-либо данной позиции, или имеет ли место ошибка в каком-либо из ридов в исследуемой позиции в скоплении.

Например, если имеется только один или два рида из 40, в которых конкретный нуклеотид находится в позиции X, а все 38 или 39 других ридов сходятся в том, что в данной позиции находится другой нуклеотид, то два выпадающих рида могут быть исключены как ошибочные, по меньшей мере в указанном конкретном положении.

[00152] В частности, когда имеется множество ридов, сгенерированных для какого-либо места генома субъекта, то наиболее вероятно, что для любой данной позиции нуклеотидов будет множество пересечений или скоплений. Эти скопления покрывают любое конкретное положение и могут быть полезными для определения корректной последовательности генома субъекта с большей точностью. Например, как указано выше, в результате секвенирования получают риды и, в различных примерах, полученные риды подвержены избыточной выборке образцов, так что в различных позициях различные конкретные риды будут перекрываться. Это перекрывание полезно для определения действительного генома образца, например с высокой вероятностью того, что он будет корректным.

[00153] Задача, таким образом, может заключаться в инкрементальном сканировании референсного генома множество раз, как более подробно описано ниже в настоящем документе, для более точной реконструкции генома субъекта, и при необходимости определении, как геном субъекта отличается от другого генома, например генома модели, использование скоплений может идентифицировать ошибки с большей точностью, например ошибки химических реагентов, оборудования или рида, и отличать их от действительных вариантов. В частности, когда субъект имеет действительную вариацию в позиции X, большинство ридов в скоплении должны подтвердить, т.е. содержать, указанную вариацию. Затем могут быть проведены процедуры статистического анализа, например такие, как раскрыты в настоящем документе, для определения действительной генетической последовательности субъекта со всеми ее вариантами относительно референсного генома.

[00154] Например, когда генетическую последовательность субъекта необходимо перестроить относительно использования референсного генома, после генерации ридов, например скопления ридов, следующими этапами могут быть картирование, выравнивание и/или сортировка ридов относительно одного или более референсного генома (например, чем больше взятых в качестве примера референсных геномов доступно в качестве моделей, тем лучше должен быть анализ) и, таким образом, перестроение генома субъекта, в результате этого наборы ридов оказываются картированными и/или выравненными относительно референсного генома (референсных геномов) во всех возможных позициях вдоль цепи, в которой имеется совпадение, и в каждой такой позиции им присвоена оценка вероятности, отражающая вероятность того, что они в действительности относятся к данной позиции.

[00155] Соответственно, в различных примерах, после генерации ридов, их позиции картируют, например, могут быть определены потенциальные места референсного генома, на которые риды могут быть картированы, а их последовательный порядок выравнен, может быть определена действительная последовательность генома субъекта, например вследствие выполнения функции сортировки выравненных данных. Кроме того, после того, как фактический геном образца известен и сравнен с референсным геномом, между этим двумя геномами можно определить вариации, может быть определен и вызван список всех вариаций/отклонений между референсным геномом и геномом образца. Такие вариации между двумя генетическими последовательностями могут быть обусловлены рядом причин.

[00156] Например, возможен однонуклеотидный полиморфизм (ОНП, SNP), например

там, где одно основание в генетической последовательности субъекта было заменено на другое; возможны более обширные замены множества нуклеотидов; возможны инсерция или делеция, например когда одно или множество оснований добавлены в генетическую последовательность субъекта или удалены из нее; и/или возможен структурный вариант, например такой, который вызван скрещиванием ножек двух хромосом, и/или возможно просто смещение, приводящее к сдвигу в последовательности. В различных примерах можно сформировать файл определения вариантов, содержащий все вариации генетической последовательности субъекта относительно референсных последовательностей. В частности, в различных вариантах реализации способы по настоящему изобретению могут включать в себя формирование файла определения вариантов (VCF), идентифицирующего один или более, например, все, генетические варианты у индивида, ДНК которого секвенировали, например, в соответствии с одним или более референсных геномов. Файл VCF в базовом виде представляет собой список положений вариаций и их тип: например, хромосома 3, в позиции X, «A» заменен «T» и т.д.

[00157] Однако, как указано выше, чтобы сформировать такой файл, геном субъекта необходимо секвенировать и снова построить, прежде чем определять его варианты. Однако существуют несколько проблем, которые могут возникнуть при попытке формирования такой сборки. Как указано выше, возможны проблемы с химией, секвенатором и/или человеческими ошибками, которые происходят в процессе секвенирования. Кроме того, возможны генетические артефакты, которые делают такую реконструкцию проблематичной. Например, проблемой при выполнении таких сборок является то, что иногда имеются огромные части генома, которые повторяют сами себя, например, длинные секции генома, которые включают в себя одни и те же строки нуклеотидов. Следовательно, так как любая генетическая последовательность уникальна не везде, возможны трудности с определением того, где в геноме в действительности картируется и выравнивается идентифицированный рид.

[00158] Например, в зависимости от применяемого протокола секвенирования, могут быть получены более короткие или более длинные риды. Более длинные риды являются полезными тем, что чем длиннее рид, тем меньше вероятность того, что он появится во множестве мест генома. Необходимость оценки меньшего количества возможных положений также может ускорить работу системы. Однако, чем длиннее риды, тем больше проблем они могут вызывать, так как больше вероятность того, что они содержат реальные или ложные вариации, например, обусловленные ОНП, инделом (инсерцией или делецией) (InDel), ошибкой оборудования или тому подобным, в результате чего будут отсутствовать совпадения между ридом и референсным геномом. С другой стороны, более короткие риды являются полезными, так как чем короче рид, тем меньше вероятность того, что он покроет позицию, которая кодирует вариант. Проблема с более короткими ридами, однако, заключается в том, что более короткий рид более вероятно появится в множестве позиций генома, вследствие чего требуется дополнительное время и ресурсы для обработки для того, чтобы определить, какое из всех возможных позиций наиболее вероятно является действительным положением выравнивания. Идеал, которого можно достичь, например практикуя способы, описанные в настоящем документе, заключается в том, что может быть получен файл определения вариантов, в котором сформирован список секвенированного генома (исследуемого генома), который показывает, где находятся все варианты пар оснований, гарантируя, что каждый определенный вариант является действительным вариантом, а не просто химической ошибкой, ошибкой оборудования или вызванной человеческим

фактором ошибкой рида.

[00159] Таким образом, для вариации существуют две основных возможности. Во-первых, существует действительная вариация в данном конкретном исследуемом месте, например, когда геном человека в конкретном месте действительно отличается от референса, например, имеется естественная вариация, обусловленная ОНП (заменой одного основания), инсерцией или делецией (длиной в один или более нуклеотидов), и/или имеется структурный вариант, например, когда материал ДНК из одной хромосомы перекрещивается на другую хромосому или ножку, или когда определенная область дважды встречается в ДНК. В качестве альтернативы, вариация может быть вызвана наличием проблемы в данных рида из-за ошибки химии или оборудования, секвенатора или выравнивателя, или иной человеческой ошибки. Соответственно, способы, описанные в настоящем документе, могут быть использованы таким образом, чтобы компенсировать эти типы ошибок и, в частности, чтобы отличать ошибки в вариации, обусловленные химией, оборудованием или человеком, от реальных вариаций в секвенированном геноме. Точнее говоря, способы, устройства и системы для их реализации, описанные в настоящем документе, разработаны таким образом, чтобы четко различать эти два различных типа вариаций и, следовательно, лучше обеспечивать точность любых сформированных файлов вариантов, чтобы правильно выявлять истинные варианты.

[00160] Кроме того, в различных вариантах реализации после того, как реконструирован геном субъекта и/или сформирован файл VCF, такие данные могут быть затем подвергнуты третичной обработке с целью их интерпретации, например, для определения того, что эти данные означают с точки зрения выявления заболеваний, от которых может страдать или которым может подвергнуться этот человек, и/или для определения терапии или изменения стиля жизни, которыми, возможно, пожелает воспользоваться данный субъект, чтобы устранить и/или предотвратить болезненное состояние. Например, генетическая последовательность субъекта и/или его файл определения вариантов могут быть проанализированы для определения релевантных с клинической точки зрения генетических маркеров, которые указывают на наличие или возможность болезненного состояния и/или эффективность, с которой может воздействовать на субъекта рекомендуемый терапевтический или профилактически режим. Затем эти данные могут быть использованы для обеспечения субъекту одного или более терапевтических или профилактических режимов для того, чтобы улучшить качество жизни субъекта, например, вылечить и/или предотвратить болезненное состояние.

[00161] В частности, медицинские научные технологии развиваются в совокупности с развитием информационных технологий, которое повысило наши возможности в отношении хранения и анализа медицинских данных. Следовательно, после того, как определены одна или более генетических вариаций индивида, такая информация файла определения вариантов может быть использована для подготовки полезной с медицинской точки зрения информации, которая, в свою очередь, может быть использована для определения, например, с использованием известных моделей статистического анализа, относящихся к здоровью данных и/или полезной с медицинской точки зрения информации, например, в диагностических целях, например, для диагностирования заболевания или его возможности, клинической интерпретации (например, поиска маркеров, которые представляют вариант заболевания), того, следует ли включить субъект в различные клинических испытания или исключить из них, и для других таких целей. Поскольку существует конечное число болезненных состояний,

которые вызываются генетическими нарушениями, при третичной обработке варианты определенного типа, например, известные тем, что они связаны с возникновением болезненных состояний, могут быть уточнены, например, путем определения того, включены ли один или более генетических маркеров болезни в файл определения вариантов субъекта.

[00162] Поэтому в различных случаях способы, описанные в настоящем документе, могут включать в себя анализ, например, сканирование, VCF и/или сформированной последовательности на предмет известных связанных с заболеваниями вариантов последовательности, например, присутствующих по этой причине в базе данных геномных маркеров, чтобы выявить наличие генетического маркера в VCF и/или сформированной последовательности, и при наличии такового проверять присутствие или возможность генетически обусловленного болезненного состояния. Так как существует огромное количество известных генетических вариаций и огромное количество индивидов, страдающих от заболеваний, вызываемых такими вариациями, в некоторых вариантах реализации способы, описанные в настоящем документе, могут охватывать формирование одной или более баз данных, связывающих секвенированные данные полного генома и/или связанного с ним файла определения вариантов, например, от одного или множества индивидов, с болезненным состоянием, и/или поиск в сформированных базах данных с целью определения того, имеет ли конкретный субъект генетический состав, который предрасполагает его к наличию такого болезненного состояния. Такой поиск может включать в себя сравнение одного полного генома с одним или более другими, или фрагмента генома, такого как фрагмент, содержащий только вариации, с одним или более фрагментами одного или более других геномов, например, в базе данных референсных геномов или их фрагментов.

[00163] Кроме того, следует понимать, что генетические последовательности, вовлеченные в данные способы, могут представлять собой ДНК, одноцепочечную ДНК (оцДНК), РНК, мРНК, рРНК, тРНК и тому подобное. Следовательно, хотя в настоящем раскрытии приведены различные упоминания различных способов и устройств анализа ДНК генома, в различных примерах системы, устройства и способы, раскрытые в настоящем документе, в равной степени подходят для выполнения их соответствующих функций, например, анализа, над всеми типами генетического материала, включая ДНК, оцДНК, РНК, мРНК, рРНК, тРНК и тому подобное. Кроме того, в различных примерах способы по настоящему изобретению могут включать анализ сформированной генетической последовательности, например ДНК, оцДНК, РНК, мРНК, рРНК, тРНК и тому подобного, субъекта и определение на ее основании вариаций белков, которые наиболее вероятно вызваны генетической последовательностью, и/или определение и/или прогнозирование на ее основании возможного болезненного состояния, например вызванного ошибкой в экспрессии белка. Следует отметить, что полученная генетическая последовательность может представлять интрон или экзон, например, генетическая последовательность может относиться лишь к кодированной части ДНК, например, где получен экзом, и с использованием лишь известных технологий обработки могут быть секвенированы только кодированные области, или некодированные области, что может приводить к более быстрому секвенированию и/или более быстрой обработке, хотя и с учетом более сложной процедуры подготовки образца.

[00164] В настоящее время такие этапы и анализы, описанные в настоящем документе, обычно выполняют на различных отдельных или несвязанных этапах, зачастую использующих различное аналитическое оборудование в различных местоположениях. Соответственно, согласно различным аспектам, способы и системы по настоящему

изобретению реализуются одним устройством и/или в одном местоположении, например в совокупности с автоматизированным секвенатором или другим устройством, выполненным с возможностью формирования данных генетической последовательности. В различных примерах множество устройств может использоваться в одном местоположении или во множестве удаленных местоположений, и в некоторых случаях способы могут использовать два или более обрабатывающих блоков, развернутых в двух или более местоположений.

[00165] Например, согласно различным аспектам может быть предусмотрен конвейер, который включает в себя выполнение одной или более аналитических функций, как описано в настоящем документе, над геномной генетической последовательностью одного или более индивидов, например, на данных, полученных в цифровом файловом формате, например, FASTQ, из автоматизированного секвенатора. Типичный конвейер, подлежащий выполнению, может включать в себя один или более секвернированных генетических материалов, например, часть или весь геном, одного или более субъектов, причем генетический материал может содержать ДНК, оцДНК, РНК, рРНК, тРНК и тому подобное, и/или, в некоторых случаях, генетический материал может представлять кодируемые или некодируемые области, такие как экзомы, эписомы ДНК. Конвейер может включать в себя одно или более из выполнения операции определения оснований и/или исправления ошибки, например, в оцифрованных генетических данных, и/или может включать в себя одно или более из выполнения функции картирования, выравнивания и/или сортировки над генетическими данными. В определенных случаях конвейер может включать в себя выполнение одного или более из повторного выравнивания, удаления дубликатов, перекалибровки оценки качества основания, редукции и/или сжатия и/или распаковки над оцифрованными генетическими данными. В определенных случаях конвейер может включать в себя выполнение операции определения вариантов над генетическими данными.

[00166] Таким образом, в различных случаях конвейер по данному изобретению может содержать один или более модулей, где модули выполнены с возможностью осуществления одной или более функций, таких как операция определения оснований и/или исправления ошибок, и/или функция картирования, выравнивания, и/или сортировки генетических данных, например, секвенированных генетических данных. И в различных случаях конвейер может содержать один или более модулей, которые выполнены с возможностью осуществления одного или более из локального повторного выравнивания, удаления дубликатов, перекалибровки качественной оценки основания, определения вариантов, редукции и/или распаковки над генетическими данными. Многие из этих модулей могут выполняться либо программным обеспечением, либо аппаратным обеспечением, или удаленно, например, посредством программного обеспечения или аппаратного обеспечения, например, в облаке или на удаленном сервере и/или банке серверов.

[00167] Кроме того, многие из этих этапов и/или модулей конвейера являются необязательными и/или могут быть расположены в любом логическом порядке и/или полностью опущены. Например, программное обеспечение и/или аппаратное обеспечение, описанные в настоящем документе, могут включать в себя или не включать определение оснований или алгоритм исправления последовательности, например, когда могут быть опасения, что такие функции могут привести к статистической систематической ошибке. Следовательно, система будет включать в себя или не включать в себя функцию определения оснований и/или исправления последовательности соответственно, в зависимости от требуемого уровня точности и/или эффективности.

И, как указано выше, одна или более функций конвейера могут быть использованы при формировании геномной последовательности субъекта, например, посредством реконструкции генома на основе референса. Кроме того, в определенных случаях выходными данными из конвейера является файл определения вариантов, указывающий частично или полностью варианты в геноме или его части.

[00168] Соответственно, как указано выше, выходные данные выполнения протоколов секвенирования, например, как одних или более из указанных выше, обычно являются цифровым представлением генетического материала субъекта, например в файловом формате FASTQ. Однако, также может использоваться авторад (autorad), выраженный в цифровом виде. В частности, выходные данные от протокола секвенирования могут содержать множество ридов, где каждый рид содержит последовательность, например строку, нуклеотидов, где позиция каждого нуклеотида была определено, и качественную оценку, представляющую вероятность того, что нуклеотид определен неверно. Однако качество этих выходных данных может быть повышено посредством различных протоколов предварительной обработки для достижения более высокого качества оценок, причем один или более из указанных протоколов могут быть применены в способах, описанных в настоящем документе.

[00169] Например, в некоторых случаях, необработанные данные файла FASTQ могут быть обработаны для очистки первоначальных определений оснований, полученных из секвенатора/ридера, например на этапе первичной обработки, например перед вторичной обработкой, описанной выше в настоящем документе. В частности, секвенатор/ридер обычно анализирует данные секвенирования, например флуоресцентные данные, указывающие на то, какой нуклеотид находится в какой позиции, и преобразует данные изображения в определение основания с качественной оценкой, например где качественная оценка основана на сравнительной яркости флуоресценции в каждой позиции. Может применяться специализированный алгоритм, например на этапе первичной обработки, для корректного анализа этих различий во флуоресценции для более точного выполнения определения основания. Как указано выше, этот этап может быть включен в конвейер этапов и может быть реализован посредством программного обеспечения и/или аппаратного обеспечения, однако, в данном случае он будет частью платформы первичной обработки.

[00170] Дополнительный этап обработки может включать функцию корректировки ошибок, которая может включать попытку учесть от миллиона до миллиарда ридов в файле FASTQ и скорректировать некоторую часть механических ошибок секвенирования с помощью информации, относящейся к определению основания и качественной оценке, доступных перед каким-либо последующим этапом обработки, например, функцией картирования, выравнивания и/или сортировки, и т.д. Например, риды в файле FASTQ могут быть проанализированы для определения, имеются ли какие-либо подпоследовательности в каких-либо ридах, которые появляются в других ридах, которые вследствие дублирующего покрытия могут повысить уверенность в том, что подпоследовательности в ридах могут быть корректными. Это может быть реализовано посредством построения хэш-таблицы, содержащей все возможные k-меры выбранной длины, k, из каждого рида, и хранения каждой из них его частоты, а также того, какие основания следуют непосредственно за ним и с какой вероятностью. Затем с использованием хэш-таблицы каждый рид может быть постоянно сканирован. Так как каждый k-мер в конкретном риде может быть найден в хэш-таблице, может быть выполнена оценка, вероятно ли, что основание, следующее непосредственно за тем k-мером, является корректным или нет. Если вероятность этого мала, то оно может быть

заменено тем, которое наиболее вероятно следует за ним, исходя из таблицы. Последовательные k-меры для этого рида затем будут содержать корректное основание в качестве значения в той позиции и процесс повторят. Это может быть высокоэффективным при корректировке ошибок, так как избыточная выборка образцов обеспечивает сбор точной статистики для прогнозирования того, что последует за каждым k-мером. Однако, как указано выше, такие корректировки могут привнести статистическое смещение в систему, например вследствие ошибочных корректировок в данных, так что эти процедуры могут быть при необходимости пропущены.

[00171] Соответственно, согласно аспектам настоящего изобретения, в различных примерах способы, устройства и/или системы по настоящему изобретению могут включать получения данных рида, которые либо были, либо не были заранее обработаны, например вследствие из получения непосредственно из файла FASTQ автоматизированного секвенатора, и подвергание полученных данных одной или более функции из картирования, выравнивания и/или сортировки. Выполнение таких функций может быть полезным, например, как указано выше, в различных примерах, вследствие того, что сгенерированные данные обычно сформированы различными автоматизированными секвенаторами, например, риды, имеют длину существенно меньше, чем вся анализируемая геномная последовательность, и так как человеческий геном обычно имеет множество повторяющихся участков и известно, что он имеет различные повторяющиеся паттерны, то может иметься множество мест, в которых любая данная последовательность ридов может соответствовать сегменту человеческого генома. Следовательно, учитывая все возможности, данный рид может совпадать с последовательностью генома, например, вследствие различных повторяющихся последовательностей в геноме и т.д., необработанные данные рида могут не явно указывать, какая из возможностей в действительности является корректным положением, из которого он был получен. Таким образом, для каждого рида потребуется определить, куда в геноме рид в действительности картируется. Кроме того, также может быть полезным определить последовательное выравнивание ридов для определения действительной идентичности последовательности субъекта, и/или также может быть полезным определить положение в хромосоме для каждой части последовательности.

[00172] В различных примерах способы по настоящему изобретению могут быть направлены на картирование, выравнивание и/или сортировку необработанных данных рида файла FASTQ для нахождения всех мест, в которых данный рид вероятно может быть выравнен, и/или определение действительной идентичности последовательности субъекта, и/или определение положения в хромосоме для каждой части последовательности. Например, картирование может быть применено для картирования сформированных ридов на референсный геном и, таким образом, поиска положения, в котором появляется каждый рид, для того, чтобы хорошо совпадать с геномом, например, поиска всех мест, в которых может быть хорошая оценка для выравнивания какого-либо данного рида относительно референсного генома. Таким образом, картирование может включать учет одного или более, например, всех, из необработанных или предварительно обработанных ридов из файла FASTQ и сравнение указанных ридов с одним или более референсных геномов, а также определение, где указанный рид может совпасть с референсным геномом (геномами). В своей базовой форме картирование включает поиск положения (положений) в референсном геноме, где совпадает один или более из ридов FASTQ, полученный из секвенатора.

[00173] Аналогичным образом, выравнивание может быть применяться для оценки возможных положений ридов индивида относительно окна референсного генома для

определения, где и как последовательность ридов наилучшем образом выравнена с геномом. Однако выполнение выравнивания может быть трудным вследствие замен, инсерций, делеций, структурных вариантов и тому подобного, что может мешать риду выравниваться точно. Таким образом, существуют несколько различных путей получить выравнивание, но для этого может потребоваться внести изменения в рид, причем каждое изменение, которое необходимо внести для получения надлежащего выравнивания, приводит к снижению оценки достоверности. Например, любой данный рид может иметь замены, инсерции и/или делеции при сравнении с референсным геномом, и эти вариации необходимо учитывать при выполнении выравнивания.

[00174] Соответственно, вместе со спрогнозированным выравниванием также может быть дана оценка вероятности того, что спрогнозированное выравнивание является верным. Указанная оценка указывает на наилучшее выравнивание для любого данного рида среди множества положений, в которых рид может быть выравнен. Например, оценку выравнивания основана на том, насколько хорошо данный рид совпадает с потенциальным положением картирования, и может включать растяжение, уплотнение и изменение частей рида для получения наилучшего выравнивания.

[00175] Оценка будет отражать все изменения рида для соответствия референсу. Например, для создания выравнивания между ридом и референсом может потребоваться инсерция одного или более промежутков в рид, причем инсерция каждого промежутка представляет собой делецию в риде по отношению к референсу. Аналогичным образом, может потребоваться выполнить делецию в риде, причем каждая делеция представляет собой инсерцию в риде по отношению к референсу. Кроме того, может потребоваться изменить различные основания, например, вследствие одной или более замен. Каждое из этих изменений выполняют для того, чтобы обеспечить более точное выравнивание рида (ридов) относительно референса, но каждое изменение выполняется ценой оценки качества, которые являются измерением того, как хорошо весь рид совпадает с некоторой областью референса. Достоверность таких качественных оценок затем определяют посредством наблюдения за всеми положениями, на которые рид может быть картирован и сравнения оценок в каждом положении, а также выбора местоположения с наивысшей оценкой. В частности, когда имеется множество позиций с высокой качественной оценкой, то достоверность является низкой, но когда разница между первой и второй лучшими оценками велика, то достоверность является высокой. В конце все предложенные риды и оценки достоверности оценивают и выбирают наилучшее совпадение.

[00176] После того как ридам назначены позиции относительно референсного генома, что заключается в определении того, какой хромосоме принадлежит рид, и его смещения от начала этой хромосомы, их можно отсортировать, например по позиции. Это позволяет в последующих анализах использовать преимущества различных протоколов с избыточной выборкой, описанных в настоящем документе. Все риды, которые перекрывают данную позицию в геноме, могут быть рядом друг с другом после сортировки и могут быть организованы в скопление и без труда исследованы, чтобы определить, согласуются ли большинство из них с референсным значением, или нет. Если нет, как указано выше, вариант можно пометить.

[00177] Как указано выше, файл FASTQ, полученный из секвенатора, состоит из множества, например, от миллионов до миллиарда или более, ридов, состоящих из коротких строк данных последовательности нуклеотидов, представляющих часть или весь геном индивида. Картирование, как правило, включает в себя нанесение ридов на все местоположения в референсном геноме, где имеется совпадение. Например, в

зависимости от размера рида, могут быть одно или множество местоположений, где этот рид по существу совпадает с соответствующей последовательностью в референсном геноме. Соответственно, картирование и/или другие функции, описанные в настоящем документе, могут быть выполнены с возможностью определения того, какое из всех возможных местоположений, где одно или более ридов могут совпадать с референсным геномом, действительно является истинным местом, куда они картируются.

[00178] Можно сравнить каждый рид с каждой позицией в референсном геноме из 3,2 миллиардов для определения того, где риды, в случае совпадения, совпадают с референсным геномом. Это может быть выполнено, например, когда длина рида достигает приблизительно 100000 нуклеотидов, приблизительно 200000 нуклеотидов, приблизительно 400000 нуклеотидов, приблизительно 500000 нуклеотидов и даже приблизительно 1000000 или более нуклеотидов. Однако, когда риды имеют существенно меньшую длину, например когда имеется 50 миллионов ридов или более, например, 1 миллиард ридов, этот процесс может занять очень много времени и потребовать большого количества вычислительных ресурсов. Соответственно, существуют несколько способов, например как описаны в настоящем документе, которые были разработаны для выравнивания ридов FASTQ с референсным геномом значительно более быстрым образом. Например, как описано выше, один или более алгоритмов могут применяться для картирования одного или более ридов, сформированных секвенатором, например, в файле FASTQ, и соотнесения их с референсным геномом для определения, где в референсном геноме рассматриваемый рид потенциально может быть картирован.

[00179] Например, в различных способах формируют индекс референса, чтобы можно было искать риды или части ридов в индексе, извлекая указатели местоположений в референсе для картирования ридов на референс. Такой индекс референса можно построить в различных формах и обращаться к нему различным образом. В некоторых способах индекс может содержат дерево префиксов и/или суффиксов. В других различных способах индекс может включать преобразование референса Барроуза-Уилера. В других способах индекс может содержать одну или более хэш-таблиц, а хэш-функция может быть выполнена на одной или более частях ридов с целью картирования ридов на референс. В различных примерах один или более из этих алгоритмов может быть выполнен последовательно или одновременно для точного определения, где один или более, например, значительная часть ридов или каждый рид, из ридов верно совпадает с референсным геномом.

[00180] Каждый из этих алгоритмов может иметь достоинства и/или недостатки. Например, дерево суффиксов и/или префиксов и/или преобразование Барроуса-Уиллера можно применять к данным последовательности так, что индекс референсного генома составляется и/или запрашивается как древовидная структура данных, причем начиная с одноосновной или короткой подпоследовательности рида, подпоследовательность постепенно расширяют в пределах рида, на каждом шаге обращаясь к индексу, отслеживая путь в древовидной структуре данных до тех пор, пока подпоследовательность не станет достаточно уникальной, например, достигнута оптимальная длин а пути, и/или достигнут листовой узел древовидной структуры, где листовой или последний достигнутый узел дерева показывает одну или более позиции в референсном геноме, с которой начинается рид. Таким образом, эти алгоритмы, как правило, не имеют фиксированной длины для последовательностей ридов, которые могут быть картированы путем поиска по индексу. Хеш-функция же часто задействует единицу сравнения фиксированной длины, которая может быть длиной всего рида, но часто в несколько раз больше длины некоторой части рида, которая называется

затравкой. Такие затравки могут быть короче или длиннее, но в отличие от деревьев префиксов и/или суффиксов и/или преобразований Барроуза-Уилера, задействованные в хеш-функции затравки ридов имеют заранее выбранную, фиксированную длину.

[00181] Дерево префиксов и/или суффиксов представляет собой структуру данных, которая выстраивается из референсного генома таким образом, что каждое ребро (link), соединяющие родительский узел с дочерним, помечено или ассоциировано с нуклеотидом или последовательностью нуклеотидов, а каждый путь от корневого элемента через различные ребра и узлы соответствует некоторой подпоследовательности референсного генома. Узел, достигнутый таким путем, неявно ассоциирован с референсной последовательностью, соответствующей пути от корневого элемента. Начиная от корневого элемента, дерево префиксов соответствует продолжению последовательности референсного генома вперед, в то время как подпоследовательности дерева суффиксов соответствует продолжению референсного генома в обратном направлении. Оба вида деревьев, префиксов и суффиксов, можно использовать в гибридных алгоритмах префиксов/суффиксов, так что подпоследовательности могут выстраиваться в обоих направлениях. Деревья префиксов и суффиксов могут также содержать дополнительные ребра, такие как переходы от узла, ассоциированного с одной референсной подпоследовательностью, к другому узлу, ассоциированному с более короткой подпоследовательностью.

[00182] Например, запросы к древовидной структуре данных, служащей в качестве индекса референсного генома, можно выполнять, определяя путь в дереве, соответствующий подпоследовательности рида, который нужно картировать. Путь выстраивается по мере добавления к последовательности нуклеотидов, каждый из которых определяет следующий узел при обходе дерева, до тех пор, пока не получится уникальная последовательность. Эту уникальную последовательность также можно назвать затравкой, которая может представлять ветвь и/или корень древовидной структуры последовательности. Альтернативно, спуск по древовидной структуре можно прекратить до того, как совокупная последовательность будет полностью уникальной, так что полученная затравка может соответствовать нескольким местам референсного генома. В частности, дерево можно строить для каждой начальной позиции референсного генома, в этом случае полученные риды можно сравнивать с ветвями и/или корневыми элементами дерева, так что можно обходить дерево согласно этим последовательностям, чтобы определять, какому месту референсного генома соответствует рид. Более конкретно, риды файла FASTQ можно сравнивать с ветвями и корневыми элементами референсного дерева, что позволяет при соответствии немедленно определить их положение в референсном геноме. Например, образец рида можно использовать при обходе дерева до тех пор, пока набранная последовательность не будет настолько уникальной, чтобы определить, что данный рид действительно выровнен с определенной позицией референсного генома. В частности, можно спускаться по дереву до листового узла.

[00183] Недостатком, однако, такого дерева префиксов/суффиксов является то, что необходимо множество раз осуществлять доступ к огромной структуре данных, по мере того как дерево обходят для того, чтобы картировать риды к на референсный геном. С другой стороны, преимуществом хеш-таблицы, как подробно описано далее в этой заявке, является то, что как только таблица получена, достаточно, как правило, одного обращения к ней, чтобы определить где, и есть ли такое место, может быть соответствие между ридом и референсным геномом. В общем случае потребуется множество, например 5, 10, 15, 20, 25, 50, 100, 1,000, или более, и т.д., обращений к дереву

префиксов/суффиксов, чтобы определить, имеется ли соответствие. Далее, из-за двойной спиральной структуры ДНК, может также потребоваться построить и исследовать обратное комплементарное дерево, поскольку может потребоваться найти обратный комплемент к референсному геному. Учитывая вышесказанное, дерево данных можно описывать как структуру, построенную на основании референсного генома, которую затем сравнивают с секвенированными из ДНК субъекта ридами, однако, следует понимать, что дерево данных может быть изначально как из референсной последовательности, так и из ридов образца, или того и другого, после чего проводится сравнение одного с другим, как описано выше.

[00184] Альтернативно, или в дополнение к применению дерева префиксов или суффиксов, к данным можно применить преобразование Барроуза-Уилера. Например, преобразование Барроуза-Уилера можно использовать для хранения древовидной структуры данных, абстрактно эквивалентной дереву префиксов/суффиксов, в компактном формате, например в пространстве, выделенном для хранения референсного генома. В различных вариантах, хранимые данные представляют собой не древовидную структуру, а именно, данные референсной последовательности хранятся в линейном списке, который можно выстраивать в различном порядке с тем, чтобы преобразовать его весьма специфичным образом, так что сопутствующий алгоритм позволяет искать в последовательности участки, соответствующие ридам образца, таким образом эффективно обходя «дерево». Преимущество преобразования Барроуза-Уилера, по сравнению с деревом префиксов или дерево суффиксов, в том, что оно, как правило, требует существенно меньше памяти для хранения данных, а по сравнению с хеш-таблицами в том, что оно поддерживает переменную длину затравки, и следовательно, в нем можно определить уникальную последовательность, и найти соответствие. Однако так же, как и в случае дерева префиксов/суффиксов, то, сколько нуклеотидов данной последоваельности потребовалось для определения уникальности, или для картирования на сравнительно небольшое число референсных позиций, определяет длину затравки. В то же время для хеш-таблицы все затравки имеют одну, предопределенную длину. Однако недостатком преобразования Барроуза-Уилера является необходимость в общем случае несколько, например два или больше раза обращаться к структуре данных, как и во время последовательного спуска по дереву.

[00185] Альтернативно, или в дополнение к применению одного из или обоих способов, дерева префиксов/суффиксов и/или преобразования Барроуза-Уилера, к референсному геному и данным последовательности субъекта с целью узнать, где одно картируется на другое, еще один такой способ задействует создание индекса хеш-таблицы и/или вычисление хеш-функции. Индекс хеш-таблицы может представлять из себя большую ссылочную структуру, которая выстравивается из последовательностей референсного генома, который могут сравнивать с одной или более частью рида, чтобы определить возможное соответствие одного другому. Точно так же, индекс хеш-таблицы можно выстраивать по частям рида, которые затем можно сравнивать с одной или более последовательностью референсного генома, что в свою очередь позволяет определить возможное соответствие.

[00186] Более конкретно, в любом из описанных в этой заявке алгоритмов картирования, подходящих для реализации этапов способов, описанных здесь, один или все три алгоритма картирования, иди другие известные из уровня техники, могут применяться в программном или аппаратном обеспечении для того, чтобы картировать идну или более последовательностей образца секвенированного ДНК на один или более референсный геном. Как очень подробно описано далее, все эти действия можно

проводить как программно, так и аппаратным способом, например с помощью интегральных цепей, таких как чип, возможно как часть печатной платы. Например, исполнение одного или нескольких этих алгоритмов может быть встроено в чип, например с ППВМ (FPGA) (программируемая пользователем вентильная матрица), ИССН (ASIC) (интегральная схема специального назначения), или со Структурированной ИССН (интегральная схема специального назначения), и может быть оптимизировано для большей производительности посредством имплементации .

[00187] Дополнительно, одна или более, например, два или все три эти функции картирования могут формировать модуль, такой как модуль картирования, который может образовывать часть системы, например, конвейер, который используется в процессе для определения полной фактической последовательности генома, или ее части, субъекта. Выходные данные, полученные после применения функции картирования, могут представлять собой список вероятностей того, что один или более, наример, каждый рид картируется на один или более референсный геном. Например, выходными данными для каждого картированного рида может быть список возможных мест референсного генома, которым соответствует прошедший процедуру картирования рид. В различных вариантах осуществления можно искать точное совпадение с референсным геномом если не всего рида, то хотя бы части, например затравки. Соответственно, в различных вариантах, не обязательно иметь точное соответствие всех частей всех ридов всем частям референсного генома.

[00188] Далее, одна или все эти функции могут быть запрограммированы так, чтобы получать точное и/или приблизительное соответствие и/или иметь возможность редактировать результаты. Следовательно, все эти процессы также можно сконфигурировать на поиск неполных совпадений, если есть такая потребность, например в соответствии с заранее выбранным возможным расхождением, например совпадение на 80%, 85%, 95%, 99%, или более. Однако, как подробно описано ниже, поиск неполного совпадения может быть существенно более затратным по времени и вычислительным ресурсам, поскольку может потребовать множество правок, например, таких как например однонуклеотидный полиморфизм (SNP) или вставка или удаление одного или более нуклеотидных оснований, например, может потребоваться 1 или 2 или 3 или 5 и более правок до достижения приемлемого соответствия. Такие правки вероятно будут широко применяться при имплементации протоколов хеширования, или построении деревьев префиксов и/или суффиксов и/или преобразовании Барроуза-Уилера.

[00189] Что касается хеш-таблиц, их можно получать разными способами. В одном варианте, хеш-таблица строится разбиением референсного генома на сегменты стандартной длины, например затравки длиной от примерно 16 до примерно 30 нуклеотидов или более, или такие как от примерно 18 до примерно 28 нуклеотидов, приводя их к формату таблицы по которой можно вести поиск, после чего составляется индекс всех референсных сегментов из которых секвенировано ДНК, например, 1 или более рид, или его часть, могут сравниваться для определения соответствия. Более конкретно, индекс хеш-таблицы можно генерировать разбиением референсного генома на сегменты нуклеотидных последовательностей известной, одинаковой длины, например, затравки, и сохраняя их в случайной последовательности в отдельные ячейки таблицы. Это можно делать для части или целого референсного генома, чтобы построить фактический референсный индекс, который можно использовать, чтобы сравнивать части референсного генома с частями одного или более ридов, например из файла

FASTQ, с целью определить соответствие..

[00190] Этот способ можно затем повторять приблизительно тем же образом для части, например, для большинства ридов в файле FASTQ, чтобы генерировать затравки подходящей, например, выбранной длины. Например, риды из файла FASTQ можно использовать для создания затравок заранее определенной длины, которые можно преобразовать в бинарный вид и подать на вход хеш-функции, и запросить в индексе хеш-таблицы, где для бинарной формы затравки можно найти соответствие с бинарным представлением сегмента референсного генома, что дает представление о месте генома, в котором имеется соответствие с затравками образца.

[00191] Например, если рид имеет в длину приблизительно 100 нуклеотидных оснований, типичная затравка может быть примерно в два или три раза меньше, например примерно от 27 до примерно 30 нуклеотидных оснований в длину. Следовательно, в такой реализации, требуется несколько, приблизительно 3 или 4 затравки, в зависимости от длины рида и/или длин затравок, чтобы полностью покрыть рид. Каждую затравку можно преобразовать в бинарный вид и/или затем подать на вход хеш-таблицы, после чего может быть получен разультат, говорящий о соответствующем месте референсного генома. В таких вариантах реализации, нет необходимости сравнивать весь рид с каждой возможной позицией в референсном геном, достаточно будет сравнить части ридов, например одну или нескольких затравок для каждого рида, с индексом, содержащим эквивалентные по размеру порции референсного генома. Следовательно, в различных вариантах, хеш-таблица может быть сформирована таким образом, что достаточно будет одного обращения к памяти для определения места в референсном геноме, с которым имеется совпадения у затравки, и, следовательно, рида образца. Однако в некоторых вариантах, может появиться необходимость применить хеш-функцию и поиск к одной или более перекрывающимся секциям затравок из одного рида. В этих вариантах реализации, затравки можно выбирать таким образом, чтобы по крайней мере часть их последовательностей перекрывались. Это может помочь избежать машинных и/или человеческих ошибок, или различий в геноме субъекта и референсном геноме, и содействовать нахождению точного совпадения.

[00192] В некоторых вариантах, построение хеш-таблицы, так же как и одно или несколько различных сравнений проводятся посредством хеш-функции. Хеш-функция отчасти является скремблером данных. Она принимает на вход данные и отдает то, что на первый взгляд выглядит как случайная последовательность. В этой реализации, скремблер хеш-функции разбивает референсный геном на сегменты заранее выбранной и случайным образом размещает их в хэш-таблице. Альтернативно, пространство для хранения данных может быть сегментировано и/или для целей хранения могут присваиваться разные веса. Более, конкретно, хеш-функция - это функция, которая принимает любые входные данные и на выходе дает число, например как выходные данные бинарных шаблонов, причем выходые значения могут быть какими угодно, за исключением того, что при одном и том же входном значении, выходное значение также всегда будет одним и тем же. Следовательно, если две последовательности, подаваемые на вход хеш-функции, практически одинаковы, из-за того, что они в точности не совпадают, на выходе будут получаны два совершенно разных значения.

[00193] Далее, поскольку генетический материал может быть составлен из четырех нуклеотидных оснований, например, “A”, “C”, “G”, и “T” (или “U” в случае РНК), отдельные нуклеотиды последовательности, например референсные сегменты и/или риды, или их части, которые будут поданы на вход хеш-функции, могут быть

оцифрованы и представлены в бинарном формате, например в таком, в котором каждое из 4 нуклеотидных оснований представлено в виде двубитного цифрового кода, например “A” = 00, “C” = 01, “G” = 11, и “T”/”U” = 10. В некоторых вариантах именно это бинарное значение “затравки” случайным образом помещается в хэш-таблицу в известном месте, имеющем значение, равное его бинарному представлению. Хеш-функция, таким образом, служит для разбиения референсного генома на бинарные представления референсных затравок и вставляет каждую бинарную затравку в случайное место, например, ячейку хеш-таблицы, на основании численного значения. Наряду с цифровым бинарным кодом, например, кодом доступа, каждая ячейка может содержать фактические точки входа, где сегмент был взят из реального референсного генома, например, референсную позицию. Референсная позиция, таким образом, может представлять собой число, показывающее позицию исходной референсной затравки в геноме. То же самое может быть проделано для перекрывающихся позиций, которые размещены в таблице в случайном порядке, но в известных местах, например посредством хеш-функции. Способом, например как этот, можно получать индекс хеш-таблицы, причем индекс включает цифровой бинарный код для части или для всего разнообразия сегментов одного или более референсных геномов, на которые может ссылаться одна или более последовательностей генетического материала, например, один или более рид, или их части, от одного или более субъектов.

[00194] Когда хэш-таблицу и/или функцию реализуют как модуль, например как один из модулей конвейера, на программном (так что ширина в битах равна 2x числу нуклеотидных оснований в затравке как описано выше) и/или аппаратном обеспечении, по ссылке выше, хеш-таблицу можно строить так, чтобы бинарное представление затравок референсного рида имеет любую желаемую ширину в битах. Поскольку затравки могут быть длинными или короткими, бинарные представления могут быть больше или меньше, но обычно длину затравки нужно выбирать таким образом, чтобы она была достаточно длинной, чтобы обеспечить уникальность, но не слишком длинной, когда становится слишком сложно искать совпадения между затравками референсного генома и затравками рида образца из-за наличия ошибок или большого числа вариантов. Например, как указано выше, человеческий геном состоит из до примерно 3.1 миллиарда пар, а типичный рид может быть порядка 100 нуклеотидных оснований в длину. Следовательно, длина эффективной затравки может быть примерно от 16 или примерно 18 нуклеотидных основания и менее до примерно 28 или 30 нуклеотидных оснований или более. Например, в некоторых вариантах длина рида может быть сегментом длиной 20 нуклеотидных оснований. В других вариантах, длина рида может быть сегментом длиной 28 нуклеотидных оснований.

[00195] Следовательно, если длина затравки составляет сегмент в 20 нуклеотидных оснований, то в цифровом представлении данные будут иметь длину 40 бит, например, 40-битовое цифровое представление затравки. Например, если выбирают два бита для представления нуклеотида, например так, что A = 00, C = 01, G = 10, и T = 11, затравка длиной 20 нуклеотидных оснований x 2 бита на нуклеотидное основание = 40-битный (5-байтовый) вектор, например, число. Если длина затравки составляет 28 нуклеотидных оснований в длину, цифровое, например, бинарное представление будет представлять собой 56-битовый вектор. Следовательно, если длина затравки составляет приблизительно 28 нуклеотидных оснований, потребуется 56 бит, чтобы получить цифровое представление 28-основной затравки. Более конкретно, если бинарные представления затравок референсного генома случайным образом распределены по хеш-таблице и имеют длину 56 бит, еще 56 бит можно задействовать для представления

затравок ридов, которые предстоит сравнивать с затравками референсного генома. Эти 56 бит можно подать на вход полиномиальному преобразованию, на выходе которого получается также 56 бит с соответствием 1:1. Без увеличения или уменьшения количества бит на выходе, это преобразование рандомизирует место хранения соседних входных значений, так что различные затравки будут равномерно распределены по всем возможным местам хранения. Это также помогает минимизировать количество коллизий между значениями, которые хеш-функция приводит к одинаковым местам хранения. В частности, в типичной реализации хеш-таблицы, описанной в этой заявке, только часть 56 бит служит для обращения к таблице, остальные же биты хранятся для дальнейшего подтверждения соответствия. Если бы не применялась хеш-функция, огромное количество паттернов имело бы одинаковые адресные биты и различные хранимые биты, и должны были бы храниться в одних и тех же местах таблицы.

[00196] Более конкретно, имеется сходство между тем, как заполняется таблица, например, посредством программного и/или аппаратного обеспечения затравки референсного геном случайным образом помещается в хеш-таблицу, и тем, как происходит обращение к таблице при хешировании затравок ридов, а именно, обращение к хеш-тиблице в этих случаях происходят одинаково. Следовательно, затравки референсного генома и затравки ридов, имеющие одинаковый бинарный код, окажутся в одном и том же месте таблицы, например, области памяти с одним адресом, поскольку доступ к ней осуществляется одинаково, например, для одинаковых входных векторов (паттерн). Это самый быстрый способ поиска совпадений в паттерах. Каждое обращение к таблице занимает примерно константное время. Это можно противопоставить методу Барроуза-Уилера, который требует несколько проверок (их количество может меняться в зависимости от того, сколько бит требуется, чтобы найти уникальную последовательность) на запрос, чтобы найти одно совпадение, или метод бинарного поиска, который требует $\log_2(N)$ проверок, где N - число затравок в таблице.

[00197] Далее, хотя хеш-функция и может разбить референсный геном на сегменты затравок любой длины, например, 28 пар нуклеотидных оснований, и затем может перевести эти затравки в цифровое, например, бинарное 56-битное представление, не обязательно нужно иметь доступ ко всем 56 битам одновременно одним и тем же образом. Например, хеш-функция может быть реализована так, что адрес каждого сида определяется числом менее 56 бит, например числа от примерно 20 до примерно 45 бит, от примерно 25 бит до примерно 40 бит, от примерно 28 бит до примерно 35 бит, включая от примерно 28 бит до примерно 30 бит могут быть как исходный ключ или адрес доступа к хеш-таблице.

[00198] Например, в некоторых вариантах от примерно 26 до примерно 29 бит можно использовать в качестве первичного ключа доступа к хеш-таблице, без учета 27 до 30 оставшихся бит, которые можно использовать как средство двойной проверки первого ключа, например, если и первый и второй ключи оказываются в одной ячейке хеш-таблицы, тогда относительно понятно, что они принадлежат указанной локации. Так, для экономии места и уменьшения потребляемой памяти и/или времени обработки хеш-модуля, например когда хеш-таблица и/или хеш-функция реализована на аппаратном обеспечении, для представления первичного ключа доступа можно оставлять от примерно 26 до примерно 29 бит из исходных 56, представляющих оцифрованную затравку определенного секвенированного рида, а оставшиеся от 27 до 30 можно использовать для двойного контроля.

[00199] Более конкретно, в различных вариантах, от примерно 26 до примерно 29 бит из 56 бит бинарного представления референсной затравки можно использовать для

получения первичного адреса, который определяется 26 до 29 и может иметь рандомизированное место в хеш-таблице, которое в свою очередь может быть заполнено местоположением, которому изначально принадлежала референсная затравка, а также оставшимися 27 до 30 битов затравки, так что точное совпадение может быть подтверждено. Затравки запроса, представляющие риды генома субъекта, преобразованные в бинарную форму, можно хешировать той же самой функцией так, что они также будут представлены 29 битами, содержащими первичный ключ доступа. Если 29 бит, представляющие референсную затравку в точности совпадают с 29 битами, представляющими затравку запроса, они оба будут направлены в одно и то же место в хеш-таблице. При полном совпадении с референсной затравкой, ожидается, что в том месте будет найдена запись, содержащая те же оставшиеся 27 до 30 бит. В этом варианте реализации, 29 зарезервированных адресных бит референсной последовательности используют для обращения к таблице, чтобы определить позицию в референсной последовательности, к которой выровнен рид запроса, из которого была выделена затравка для запроса.

[00200] Однако, что касается оставшихся 27 до 30 бит, эти биты могут представлять вторичный ключ доступа, который также моет быть импортирован в хеш-таблицу для подтверждения соответствий найденных по 26-29-битному первичному ключу. Поскольку хеш-таблица представляет собой идеальное 1:1 преобразование последовательности 28 нуклеотидных оснований/56 бит, и только 26-29 этих битов используется для определения адреса, эти 26-29 бит были сначала проверены и таким образам определили правильный адрес при первом проходе. Эти данные, таким образом, не нуждаются в подтверждении. Однако оставшиеся примерно 27-примерно 30 бит вторичного ключа нужно проверить. Соответственно, оставшиеся примерно 27-30 бит затравки запроса вставляют в таблицу как средство завершения поиска совпадения. Такая реализация может быть короче, чем сортировка полных 56-битных ключей, и таким образом, позволяет экономить место и снижает требуемое количество памяти и время обработки модуля.

[00201] Таким образом, хеш-таблица может быть сконфигурирована как индекс, где последовательности одного или более референсных геномов, разбитые на последовательности предопределенной длины, например 28 нуклеотидных оснований, собраны в таблицу случайным образом, а последовательность одного или более ридов или их «затравочных порций», полученных секвенированием ДНК или РНК субьекта, можно прогонять через индекс хеш-таблицы, например в соответствии с хеш-функцией, чтобы определить положение затравок в таблице, позволяя получить одну или более позицию, например, место в референсном геноме, когда затравка образца соответствует месту референсного генома. При использовании прямого перебора «в лоб» для сканирования референсного генома с целью обнаружить места, совпадающие с затравками, пришлось бы перебрать 3 миллиона позиций. Однако, при хешировании, определение места соответствия каждой затравки требует приблизительно константного времени. Часто, место совпадения можно подтвердить за одно обращение к таблице. В случае, когда несколько затравок подходят к одному месту в таблице, можно сделать несколько дополнительных обращений, чтобы уточнить результат. Следовательно, хотя потенциально риду длиной 100 нуклеотидных оснований может соответствовать 30М позиций в референсном геноме, с помощью хеш-таблицы и хеш-функции можно быстро определить, входит ли рид в референсный геном. Соответственно, при использовании индекса хеш-таблицы не нужны в определении положения, с которым выравнивается рид.

[00202] Как было указано выше, хромосомы имеют структуру двойной спирали, которая состоит из двух противонаправленных комплементарных цепей последовательностей нуклеиновой кислоты, которые связаны друг с другом с образованием двойной спирали. Например, при образовании структуры двойной спирали, эти комплементарные пары оснований связываются друг с другом в соответствии со следующей формулой: “A” связывается с “T”, а “G” связывается с “C”. Соответственно, это дает две равные и противонаправленные цепочки последовательностей нуклеотидов, которые комплементарны друг другу. Более конкретно, основания последовательности нуклеотидов одной нити зеркально отображаются комплементаными им основаниями на другой нити, что дает две комплементарные нити. Однако транскрипция ДНК происходит только в одном направлении, начинаясь с одного конца ДНК и продвигаясь к другому. Соответственно, получается, что для одной нити ДНК транскрипция происходит в одном направлении, а для комплементарной ей нити, транскрипция происходит в другом направлении. Соответственно, получается, что две нити последовательностей ДНК обратно комплементарны друг другу, что означает, что если порядок последовательности одной нити ДНК сравнить с другой, то мы увидим две нити, к которых буквы, обозначающие нуклеотиды в одной цепочки заменены на комплементарные им в другой например, буквы “A” на “T”, а “G” на “C” и наоборот, а их порядок обращен.

[00203] Двойная спиральная структура ДНК обуславливает то, что на подготовительном этапе перед секвенированием ДНК, разделяют, т.е. денатурируют, разделяют на две отельные нити, а затем лизируют на более мелкие фрагменты заранее определенной длины, например, по 100-300 оснований в длину, которые затем секвенируют. Можно разделить нити перед секвенированием, таким образом, чтобы секвенировалась только одна нить, но обычно нити ДНК не разделяют и секвенируют обе нити ДНК. Соответственно, в таком случае примерно половина ридов в файле FASTQ будут обратно комплементарными.

[00204] Безусловно, обе нити референсного генома, например, комплемент и обратный комплемент, могут быть обработаны и хешированы как описано выше, однако это увеличило бы размер хеш-таблицы в два раза, и в два раза удлинило бы работу функции хеширования, например, это могло бы потребовать в два раза больше вычислений для сравнения последовательностей как комплемента, таки и обратного комплемента двух геномных последовательностей. Соответственно, для экономии места в памяти, снижения потребности в вычислительных ресурсах и/или снижения времени обработки в различных вариантах нужно хранить в качестве референса в хеш-таблице только одну нить модельной ДНК.

[00205] Однако, поскольку в соответствии с обычными протоколами секвенирования, когда две нити ДНК субъекта не отделяются друг от друга, любой рид, сгенерированный из секвенированной ДНК, может относиться к любой из нитей: комплементу или обратному комплементу, соответственно, может быть затруднительно определить, какая нить обрабатывается: комплемент или обратный комплемент. Конкретнее, в различных вариантах, поскольку для генерации хеш-таблицы нужно использовать только одну нить референсного генома, половина ридов, генерируемых протоколом, секвенирования может не совпадать к данной конкретной нитью, например, комплементарной или обратно комплементарной, модели референсного генома, поскольку половину времени обрабатываемый рид является обратно комплементарным относительно хешированных сегментов референсного генома. Соответственно, только риды, сгенерированные с одной нити ДНК, будут иметь совпадение с индексированными

последовательностями референсного генома, а риды, сгенерированные с другой нити, теоретически, будут их обратными комплементами и не будут иметь совпадений ни с какими областями референсного генома. Далее, дополнительную сложность может представлять то, что для любого данного рида, который является обратным комплементом сохраненного референсного генома, этот рид все же может дать ошибочное совпадение с частью референсного генома, например, по чистой случайности. Ввиду вышесказанного для эффективного осуществления картирования в различных вариантах необходимо не только определить, с каким местом в референсном геноме совпадает рид, но и определить, является ли рид обратно комплементарным. Соответственно, хеш-таблица и/или модуль функций следует сконструировать таким образом, чтобы минимизировать эти затруднения и/или типы ошибок, возникающих в результате этих затруднений.

[00206] Например, как указано выше, в одном варианте в хеш-таблицу можно поместить и комплемент, и обратный комплемент для референсного генома, таким образом, для каждого рида или его обратного комплемента секвенированной ДНК субъекта можно установить совпадение в соответствующей нити референсной геномной ДНК. В таком варианте для любой данной затравки в риде, затравка теоретически должна иметь совпадение с одной из нитей, комплементом или обратным комплементом референса, если не принимать во внимание ошибки и отклонения. Однако хранение обеих нитей референсного генома в хеш-индекса может потребовать в два раза больше места для хранения (например, может потребовать вместо 32 гигабайт 64 гигабайта), а также может требовать в два раза больше ресурсов для обработки. Далее, такое решение не решает проблему палиндромов, которые могут иметь совпадения в обоих направлениях, например, комплементарной и обратно комплементарной цепей.

[00207] Соответственно, хотя индекс хеш-таблицы может быть сконструирован таким образом, чтобы включать обе нити референсной геномной последовательности, в различных вариантах, хеш-таблица может быть сконструирована таким образом, чтобы включать только одну нить модельного генома в качестве референса. Это может быть полезно, поскольку хранение хеш-таблицы в памяти потребует половины ресурсов для хранения и/или обработки, по сравнению с теми, которые потребовались бы для хранения и обработки обеих нитей, и, соответственно, поиск также потребует меньше времени. Однако хранение только одной нити генома в качестве референса может вызвать затруднения, поскольку, как указано выше, когда секвенируемая ДНК субъекта является двунитевой, обычно не известно, с какой конкретно нити сгенерирован каждый конкретный рид. Соответственно, в таком варианте хеш-таблица должна быть сгенерирована с учетом того факта, что картируемый рид может быть из любой нити и, соответственно, может представлять собой комплемент или обратный комплемент сохраненных сегментов референсного генома.

[00208] Соответственно, в различных вариантах, как в случае, когда только один вариант ориентации затравок из референса помещают в хеш-таблицу, при выполнении функции хеширования с затравками, сгенерированными по ридам из файла FASTQ, можно сначала провести поиск затравки в ее текущей ориентации и/или можно затем преобразовать в обратный комплемент и провести поиск с обратным комплементом. Это может потребовать провести два поиска в хеш-индексе, т.е. в два раза больше, но либо затравка, либо ее обратный компонент должны совпасть с комплементарным сегментом в референсном геноме, при условии отсутствия ошибок и вариаций, и это должно сократить общую потребность в вычислительных ресурсах, например, за счет меньшего использования памяти, а также сокращения времени, например, когда

сравнивается не такое большое число последовательностей.

[00209] Более конкретно, как в случае, когда затравка в одной конкретной ориентации состоит из 28 нуклеотидов, которые например, представлены в цифровом виде в 56 битах двоичного формата, как описано выше, затравка может быть преобразована в обратный комплемент, и обратный комплемент также можно представить в цифровом виде в 56 битах двоичного формата. Двоичный формат для каждого представления последовательности затравки и ее комплемента дает число, например, целое число, имеющее значение, представленное этим числом. Эти два значения, например, два целых числа (номера), можно сравнить и число с более высоким или более низким значением, например, более высоким или более низким абсолютным значением, можно выбрать в качестве канонического варианта ориентации, и он будет храниться в хеш-таблице и/или обрабатываться хеш-функцией. Например, в некоторых вариантах для обработки хеш-функцией можно отбирать число с более высоким значением.

[00210] Для конструирования затравок можно применять другой способ, в котором каждая затравка состоит из нечетного числа оснований. В качестве канонической ориентации можно выбрать нити, в которых основание в середине представляет собой “A” или “G”, но не “T” или “C”, или наоборот. Затем можно применить хеш-функцию к затравкам, отвечающим требованиями канонической ориентации. Таким образом, для сравнения с целью определения, которая имеет более высокое значение, нужно сравнить только два бита, представляющие срединное основание, и только эти 2 бита этой последовательности можно использовать для поиска. Соответственно, нужно учитывать только биты, представляющие два срединных основания. Обычно это хорошо работает, поскольку затравка имеет нечетную длину и, соответственно, ее срединное основание всегда преобразуется в обратный комплемент. Тем не менее, хотя это может работать для затравок с нечетной длиной, хеширование затравок с более высоким, или более низким значением, описанное выше, должно работать для затравок любой длины, даже если такой способ может потребовать обработки, например, поиска большего количества битов данных.

[00211] Эти способы можно осуществлять для любого числа затравок, например, всех затравок референса, и/или любого числа затравок, например, всех, полученных из всех или из части ридов из файла FASTQ. Приблизительно половину времени двоичное представление затравок данной ориентации, например, комплемента, будет иметь более высокое значение, и приблизительно половину времени более высокое значение будет у двоичного представления затравок противоположной ориентации, например, обратного комплемента. Но при рассмотрении бинарных чисел, в хеш-таблицу будут всегда попадать те, значение которых больше. Например, можно сравнивать двоичные целые числа для каждого рида и его комплемента, и первую встреченную 1 выбирать в качестве нити для хранения в хеш-таблице и/обработки функцией хеширования. Если обе нити содержат первую 1 в одном и том же положении, то тогда выбирают нить со второй следующей 1 и т.д. Безусловно, можно также выбирать рид с более низким значением, в этом случае выбирают нить с первым по порядку первым 0 или первым 0 имеющим более высокий номер по порядку. Указатель, например, флаг, также можно включить в хеш-таблица, где флаг указывает, какую ориентации, комплемент или обратный комплемент, хранится, представляет сохраненная и/или хешированная нить, например, флаг 1RC, для обратного комплемента.

[00212] Более конкретно, при реализации хеш-функции и оценке хеш-таблицы, затравки из геномной референсной ДНК и затравки из ридов данных о последовательности подвергают одним и тем же операциям, например, переводят в

двоичную форму и сравнивают с их обратным комплементом, при этом целые числа с более высокими, или более низкими, значениями выбирают в качестве канонических ориентаций применяют к ним хеш-функцию и передают в хеш-таблицу для поиска и проверки совпадения друг с другом. Однако, поскольку, одну и ту же операцию проводят по существу одинаковым образом с референсными последовательностями и последовательностями ридов, будут получаться одинаковые записи, если две последовательности, референсной затравки и затравки субъекта, начинаются с одной и той же последовательности, даже если одна из них является обратным комплементом, они будут отправляться в одну и ту же ячейку в хеш-таблице.

[00213] Соответственно, если некоторую затравку в референсной последовательности, имеющую данную последовательность в конкретной ориентации, преобразовать в двоичную форму и хешировать, а затем затравку, полученную из рида образца, имеющую ту же последовательность, но в обратной ориентации, например, обратный комплемент, и ее подвергают описанным выше протоколам, в силу раскрытых выше способов, когда определяют двоичное значение и выполняют функцию хеширования, поиск будет направлен на тот же адрес в хеш-таблице, как если бы хеш-функцию выполняли сначала с комплиментарной затравкой. Соответственно, при таком подходе не важно, какую ориентацию имеет обрабатываемая затравка, поскольку она всегда будет направляться на один и тот же адрес.

[00214] Соответственно, при таком подходе, раскрытые способы могут хешировать и, соответственно, определять положение затравки в таблице вне зависимости от ее ориентации, и благодаря флагу в записи, также будет известно, если какая-либо затравка является обратным комплементом. Например, также будет известно, если затравка будет перевернута относительно исходной ориентации в референсе, а также будет известно, если затравка будет перевернут относительно исходной ориентации в риде субъекта. Соответственно, если в обоих случаях было принято одно и то же решение, рид и референс имеют одинаковую ориентацию. Тем не менее, если одна сторона перевернута, а другая нет, можно заключить, что рид каптируется в обратно комплементарной ориентации относительно референса. Соответственно, при помощи хеш-таблицы можно определить, какому месту в геноме соответствует данный рид или его часть, например, затравка, и является ли она обратно комплементарной. Далее, следует понимать, что хотя сказанное выше относится к генерации хеш-таблицы по референсному геному и выполнению различных вспомогательных процессов хеш-функции с затравками, сгенерированными из ридов, например, из файла FASTQ, система также может быть структурирована так, что индекс хеш-таблицы генерируется из затравок, полученных из ридов секвенированной ДНК субъекта, а различные вспомогательные процессы хеш-функции, описанные в настоящем документе, осуществляются с затравками, сгенерированными из референсного генома.

[00215] Как указано выше, преимущество применения хеш-таблицы и/или хеш-функции заключается в том, что при применении затравок можно определить соответствие большинства ридов секвенированный ДНК с референсным геномом, часто за один поиск в хеш-таблице, и в различных вариантах, не все затравки, полученные из рида, нужно хешировать и/или искать. Затравки могут быть любой подходящей длины, например, могут быть относительно короткими, например, иметь длину 16 нуклеотидов или меньше, например, примерно 20 нуклеотидов, например, примерно 24 нуклеотидов, например, примерно 28 нуклеотидов, например, примерно 30 или примерно 40, или примерно 50, или 75, или примерно 100 нуклеотидов, или даже до 250, или 500, или 750, или даже 999, или даже примерно 1000 нуклеотидов в длину; или могут быть

относительно длинными, например, больше примерно 1000 нуклеотидов, или больше примерно 10000, или больше примерно 100000, или больше 1000000, или больше нуклеотидов в длину. Тем не менее, как описано выше, у применения затравок есть некоторые недостатки, такие как связанные с хеш-таблицей, в частности, в отношении выбора ридов подходящей длины.

[00216] Например, в функции картирования можно применять затравку любой подходящей длины, но у применения затравок относительно большой длины и относительно маленькой длины есть преимущества и недостктки. Например, чем меньше длина затравки, тем меньше вероятность того, что он будет включать ошибку или вариант, которые могут препятствовать нахождения совпадения в хеш-таблице. Однако, чем меньше длина сита, тем менее он уникален, и тем больше вероятность совпадения между затравкой референсного генома и затравками, полученными из ридов секвенированной ДНК субъекта. Далее, чем короче затравка, тем больше поисков нужно осуществить хеш-функцией, что занимает больше времени и требует повышенной вычислительной мощности.

[00217] С другой стороны, чем больше длина затравки, тем выше ее уникальность, и тем меньше вероятность существования нескольких совпадающих положений между затравками референса и запроса. Далее, чем длиннее затравка, тем меньше затравок нужно на каждый рид, тем меньше поисков и, соответственно, требуется меньше времени и меньше вычислений. Однако, чем длиннее затравка, тем больше вероятность того, что затравки, полученные из секвенированной ДНК, содержат ошибку, такую как ошибка секвенирования, и/или могут включать отклонение от последовательности, которые препятствуют определению совпадения. Кроме того, для более длинным ридам может быть дополнительно присущ недостаток, заключающийся в том, что они с большей вероятность попадают на конец рида и/или конец хромосомы. Соответственно, если длина затравки составляет всего 20 - 100 нуклеотидов, может быть несколько совпадений в хеш-таблице, но если длина затравки составляет 1000 или более нуклеотидов, будет значительно меньше совпадений, но при этом совпадений может не быть совсем.

[00218] Существует насколько методов минимизации этих недостатков. Один способ заключается в том, чтобы обеспечить достаточно избыточную выборку, генерируемую на этапах обработки ДНК перед секвенированием. Например, если известно, что обычно присутствует по меньшей мере одно отклонение (вариант) на каждую 1000 пар оснований, длина затравки может быть выбрана таким образом, чтобы максимизировать совпадения, в то же время минимизируя несовпадения из-за возможных ошибок и/или вариантов. Дополнительно, применение избыточной выборки, например, на этапах подготовки к секвенированию и/или секвенирования, можно использовать в качестве дополнительного способа минимизации различных проблем, связанных с применением затравок, например, в хеш-функции.

[00219] Как указывалось выше, избыточность выборки порождает наслоения. Наслоения представляют собой группы ридов, которые картируются с перекрыванием на в целом одно и то же место генома. Для большинства ридов из образца такие наслоения могут не быть обязательными, как в случае, когда риды, и/или сгенерированные из них затравки, не включают вариант и/или не картируются на множественные положения в хеш-таблице (например, не дублируются полностью в геноме). Однако для тех ридов и/или затравок, которые могут включать вариант, и/или ошибку, и/или другое несовпадение между затравкой и/или ридом и референсным геномом, образование наслоений для каждой данной области генома может быть

полезно. Например, даже несмотря на то, что нужны только точные попадания между затравкой, сгенерированной из рида образца генома, для картирования рида образца не референсный геном, тем не менее, может иметь место техническая ошибка или истинный вариант в последовательности ДНК образца, который может препятствовать определению точного совпадения между ридом и референсом, что часто делает формирование перекрывающихся наслоений на этапе подготовки к секвенированию или на этапе секвенирования полезным.

[00220] Например, для тех случаев, в которых затравки образца на самом деле содержат вариант или ошибку, формирование наслоений ридов может быть полезно для того, чтобы отличить фактическое отклонение от технических и/или химических ошибок. В таком варианте наслоение можно применять, чтобы определить, является ли кажущееся отклонение реальным отклонением. Например, если 95% ридов в наслоении содержат “C” в некотором положении, то правильное решение определяют остальные, даже если референсный геном содержит “T” в этом положении. В таком варианте несовпадение может быть связано с SNP, например, заменой “T” на “C “ в этом положении генома, в тех случаях, когда генетический код субъекта действительно отличается от референса. В таком варианте глубину наслаивания можно применять для сравнения перекрывающихся частей наслоения в положении отклонения, и на основании доли ридов в наслоении, которые имеют это отклонение, можно определить, действительно ли это отклонения связано с реальным отклонением в последовательности образца. Соответственно, фактическую последовательность ридов, которая наилучшим образом укладывается в геномную последовательность, можно частично определить на основании информации, отражаемой глубиной наслоений. Однако применение наслоений имеет недостаток, заключающийся в том, что оно требует больше времени на обработку всех избыточных ридов и/или генерируемых ими затравок.

[00221] Другой способ минимизации проблем, присущих коротким или длинным ридам, заключается в применении вторичной хеш-таблицы параллельно или совместно с первой, например, первичной хеш-таблицей. Например, вторую хеш-таблицу и/или хеш-функцию можно применять для затравок, которые не имеют ни одного попадания в первичнший хеш-таблице, или для затравок с множественными попаданиями в первичной хеш-таблице. Например, сравнение одной затравки в другой может дать несколько возможных результатов. В одном случае, может иметь место отсутствие попаданий например, отсутствие совпадений между двумя последовательностями, что указывает на возможную ошибку или отклонение, например, затравки рида от субъекта в сравнении с затравкой, полученной из референсного генома. Либо может быть найдено одно или множество совпадений. Однако обнаружение большого количество совпадений может создать проблему.

[00222] Например, в отношении первичной хеш-таблицы, если каждая затравка в хешируемой референсной последовательности появляется лишь небольшое число раз, например, один, два или три раза и т.д., то может не быть необходимости во вторичной хеш-таблице и/или функции хеширования. Однако, если одна или более затравок появляются большее число раз, например, 5, 10, 15, 20, 25, 50, 100, 1,000 или более раз, это может быть проблематично. Например, известны области генома человека, для которых с математической значимостью было определено, что они повторяются множество раз. Соответственно, картирование любой затравке на одно из этих положений по существу может автоматически картироваться на множество этих положений, как в случае, когда затравка содержит нуклеотиды перекрывающихся последовательностей. В таком варианте определить, с каким из всех возможных

вариантов выравнивается эта затравка, может быть затруднительно. Однако, поскольку эти повторяющиеся области известны и/или когда они станут известны, любая затравка любую затравку, обычно картирующуюся на одну или более из этих областей, можно выделить и распределить во вторичную хеш-таблицу для обработки первой или вторичной функцией хеширования, что, соответственно, позволяет не тратить зря время и вычислительные мощности на попытки использования первичной функции хеширования для определения того, что с большой вероятностью определить нельзя.

[00223] Более конкретно, сравнения затравок референсной геномной последовательности с затравками, сгенерированными из ридов генома субъекта, может дать любое число положений совпадений от 1 до сотен или даже тысяч. Тем не менее настоящая система может быть сконфигурирована для работы с некоторым числом повторных совпадений, в том числе без необходимости в дополнительных этапах обработки, как в случае, когда число совпадений меньше примерно 50, или меньше примерно 40, или меньше примерно 30, например, меньше примерно 25 или меньше примерно 20, например, меньше примерно 16 совпадений или меньше примерно 10 или примерно 5 совпадений. Однако, в случае, если число возвращаемых совпадений реальных попаданий превосходит это значение, систему можно сконфигурировать для реализации вторичной функции хеширования, например, с применением вторичной хеш-таблицы.

[00224] Соответственно, вместо того, чтобы помещать такие затравки, о которых известно, что они имеют повышенную вероятность избыточности, в первичную хеш-таблицу, такие затравки можно помещать во вторичную хеш-таблицу, или вторичную область в первой хеш-таблице. Дополнительно, в некоторых случаях, запись, которая не несет никакой информации о множественности потенциального размещения на карте для этой затравки, но передает команду на доступ к вторичной хеш-таблице, например, запись о расширении, может быть помещена в первичную хеш-таблицу. Например, запись о расширении может представлять собой инструкцию, например, инструкцию увеличить длину первичной, например, неуникальной или повторяющейся затравки, до более длинной, дающей более уникальную затравку, например, путем добавления к ней одного или более дополнительных соседних оснований, например, на конце (концах) затравки, чтобы превратить ее в более длинную последовательность затравки, которую затем можно хешировать и искать, например, во вторичной таблице.

[00225] Запись может быть выполнена таким образом, что она информирует или иным образом инструктирует относительно увеличения известной избыточную затравки на заданное количество, а также может содержать инструкции относительно того, где и/или как увеличивать затравку. Например, поскольку хеш-таблицу обычно рассчитывают заранее, например, исходно конструируют из затравок, сгенерированных по референсному геному (геномам), до создания таблице может быть известно, какая из затравок, сгенерированных по референсному геному, вероятно встретиться несколько раз (если такие будут). Соответственно, в различных вариантах, модно заранее определить, какие затравки может быть нужно переместить во вторичную хеш-таблицу. Например, при конструировании индекса хеш-таблицы, характеристики последовательностей референсных заявок, помещаемых в хеш-таблицу в виде индекса, известны, поэтому для каждой затравки можно определить, есть ли вероятность того, что она породит множество попаданий, например, от 10 до 10000 попаданий.

[00226] Более конкретно, в различных вариантах, может быть реализован алгоритм для определения предсказанных совпадений данной затравки, поученной из референсных ридов и/или ридов субъекта. Если определяется, что для какой-либо конкретной затравки

есть вероятность возвращения множества совпадений, может быть сгенерирован флаг, например, запись, например, в ячейке хеш-таблицы, указывающий на то, что эта конкретная заявка характеризуется высокой частотой попаданий. В таком варианте запись может также содержать инструкции на пропускание хеширования этой затравки и затравок, подобных ей, поскольку непрактично осуществлять определенное число, например, 20-10000 или более оценок такой затравки, необходимое для точного определения мета ее картирования. В таком варианте первичная хеш-функция может быть неспособна точно определить, какое положение их всех возможных положений, с которыми может совпасть эта затравка, является тем самым, с которым рид на самом деле выравнивается, и соответственно, для практических целей, поскольку такую затравку невозможно точно картировать на этом этапе, первичная хеш-функция вряд ли вернет полезный результат, такой как результат, точно указывающий, какому месту генома на самом деле соответствует эта затравка.

[00227] В таком варианте алгоритм функции хеширования может быть сконфигурирован для вычисления действий, необходимых для того, чтобы сделать избыточную затравку более уникальной. Например, вторичная хеш-функция может определять, насколько необходимо удлинить затравку, и в каком порядке, и в каком положении, чтобы гарантировать, что затравка перестанет быть избыточной, а приобретет степень уникальности, подходящую для хеширования. Соответственно, запись также может включать инструкции по удлинения избыточной затравки (т.е. затравки, представленной в избыточном количестве), например, на два, на четыре, на шесть и т.д., но одном или на обоих концах затравки для получения заранее определенного уровня уникальности. Таким образом. Таким образом, затравки которые сначала казались идентичными, могут определяться как неидентичные.

[00228] Например, в некоторых случаях, типичная запись может содержать инструкции на то, что дублирующуюся затравку следует удлинить на до X нечетных или четных оснований, но, в некоторых случаях, удлинить на четное число оснований, например от примерно 2 до 4, до примерно от 8 до 16, до примерно или примерно 64 или более оснований, например, на одинаковое число оснований с каждой стороны. Например, в тех случаях удлинения на до 64 оснований, запись может содержать инструкции на добавление 32 оснований с каждой стороны затравки. Число оснований, на которое удлиняют затравку, можно менять, и оно может представлять собой любое подходящее число в зависимости от конструкции системы. В некоторых вариантах можно применять вторичную функцию хеширования для определения, не сколько оснований нужно удлинить затравку, чтобы получить более разумное число совпадений. Соответственно, удлинение может быть точной относительной уникальности, как в случае, когда имеется только 1, 2, 3, или даже до 16, или 25, или 50 положений совпадения для этого паттерна. В различных вариантах, одинаковое удлинение обоих концов затравки может быть полезно для того, чтобы избежать проблемы с обратными ридами, но в различных вариантах затравку можно удлинять на одно или большее, неодинаковое число оснований с обеих сторон.

[00229] Более конкретно, как в одном примере, если затравка содержит 28 оснований, и запись об удлинении, такая запись об удлинении, расположенная в ячейке первичной хеш-таблицы, заставляет хеш-функцию удлинить затравку, например, на 64 основания, то эта запись может также содержать дополнительные инструкции относительно того как следует удлинять затравку, например, путем добавления 32 оснований с каждой стороны затравки. Тем не менее, удлинение может осуществляться в любом удобном положении рида и может быть симметричным или асимметричным. В некоторых

вариантах запись может содержать инструкции для функции хеширования на симметричное удлинение затравки, поскольку в некоторых случаях такое симметричное удлинение может быть более эффективным, как, например, в случае с обратными комплементами, обсуждаемом в настоящем документе. В таком варианте при удлинении к каждой из противоположных сторон затравки будет добавлено одинаковое число оснований. Тем не менее, в других вариантах удлинение может осуществляться путем добавления четного или нечетного числа оснований несимметричным образом, и, соответственно, операция удлинения не обязательно будет удлинять затравку на одинаковое число оснований с каждой стороны. Обычно первичная хеш-таблица сконфигурирована таким образом, что она не полностью заполнена. Например, желательно задействовать не более 80% или 90% ее емкости. Это необходимо для высокой эффективности скорости поиска. В случае, когда имеет место большое количество коллизий при хешировании затравок в одно и то же место при создании таблицы, механизм сортировки создаст цепочку ссылок на другие положения, и благодаря этому механизм поиска сможет найти положение, присвоенное перегруженной затравке. Чем плотнее таблица, тем выше число коллизий и тем более длинная цепочка необходима для нахождения фактического совпадения.

[00230] В различных вариантах, как в случае, когда исходная, избыточная затравка, имеет длину 28 оснований, запись передает инструкцию удлинить ее, например, на с от 18 до 32 до 64 оснований, например, на каждой из противоположных сторон затравки, цифровое представление затравки, цифровое представление затравки может составлять примерно 64 оснований × 2 бита на основание = 128 бит. Соответственно, в зависимости от выполнения модуля картирования, этот размер может быть слишком большим для обработки в первичной хеш-таблице. Соответственно, в некоторых вариантах для того, чтобы справится с потребностью в такой экстенсивной обработке, в некоторых вариантах реализации модуль вторичного хеширования может быть сконфигурирован для хранения информации, связанной с затравками большего размера. Поскольку число затравок, требующих удлинения, представляет собой некоторую долю общего числа затравок, вторичная хеш-таблица может быть меньше, чем первичная хеш-таблица. Однако в других вариантах, например, для уменьшения потребности в вычислениях модуля, например, для экономии битов, известные избыточные части последовательности, например, первичной последовательности, можно заменить заранее выбранными переменными, например, заранее определенной длиной последовательности. В таком варианте, поскольку избыточная последовательность уже известна и идентифицирована, нет необходимости в том, чтобы полностью представлять ее в цифровом виде. Вместо этого, в различных вариантах, все, что нужно сделать - это заменить известную, избыточную последовательность известной вариабельной последовательностью, а все, что на практике нужно искать - это удлиненные части, например, хвосты, которые были добавлен с одной из сторон последовательности, поскольку они являются единственными неизбыточными и новыми частями исходной последовательностями. Соответственно, в некоторых вариантах первичную последовательность можно заменить более коротким кодом-идентификатором (таким как 24-битный прокси вместо 56-битного представления), а затем к прокси можно добавить основания удлинения, например 36-битное расширение (например, с получением 60 битов), которое затем можно поместить в запись расширения первичной таблицы. Таким образом, сложности, связанные со слишком короткими и/или слишком длинными ридами, можно минимизировать и получить преимущество, заключающееся в том, что нужно поддерживать только один или небольшое число

ридов в хеш-таблице.

[00231] Как указано выше, описанную выше хеш-функцию можно реализовать в программном обеспечении и/или аппаратном обеспечении. Преимущество реализации модуля хеширования в аппаратной части заключается в возможности ускорения процессов и, соответственно их значительно более быстрого осуществления. Например, в тех случаях, когда программное обеспечение может включать различные инструкции на выполнение одной или более из различных функций, реализация таких инструкций часто требует хранения и/или вызова, и/или интерпретации данных и инструкций, например, перед исполнением. Однако, как отмечено выше и описано более подробно в настоящем документе, чип можно аппаратно смонтировать для выполнения этих функций без необходимости в вызове, интерпретации и/или осуществлении одной или более из последовательности инструкций. Вместо этого чип может быть выполнен (смонтирован) с возможностью прямого выполнения таких функций. Соответственно, в различных аспектах настоящее изобретение относится к специальному смонтированному устройству, которое может быть сконфигурировано так, что часть или весь описанный выше модуль хеширования может быть реализован одной или более сетевыми схемами, такими как интегральные схемы, смонтированные на чипе, например, FPGA, ASIC или структурированный ASIC.

[00232] Например, в различных вариантах, индекс хеш-таблицы может быть создан и хеш-функция может выполняться на чипе, а в других вариантах индекс хеш-таблицы генерируется вне чипа, например, программным обеспечением, запускаемым на основном процессоре, но после генерации его загружают на чип и используют так, например, в процессе работы модуля хеширования. В некоторых вариантах чип может иметь любое подходящее число гигабайт, такое как 8 гигабайт, такое как 16 гигабайт, такое как 32 гигабайт, такое как 64 гигабайт, например, примерно 128 гигабайт. В различных вариантах, чип может быть сконфигурирован так, чтобы различные процессы модуля хеширования выполнялись с использованием только части или всех ресурсов памяти. Например, в тех случаях, когда можно построить обычный референсный геном, большую часть памяти можно отдать под хранение хеш-индекса референса, и/или для хранения ридов, и/или для того, чтобы зарезервировать место для других используемых функциональных модулей, как в случае, когда 16 гигабайт выделено под хранение ридов, 8 гигабайт может быть выделено под хранение хеш-индекса, а еще 8 гигабайт может быть выделено под другие функции обработки. В другом примере, если 32 гигабайт выделено под хранение ридов, 26 гигабайт можно выделить под хранение первичной хеш-таблицы, 2.5 гигабайт можно выделить под хранение вторичной таблицы, а 1.5 гигабайт можно выделить для хранения референсного генома.

[00233] В некоторых вариантах реализации вторичная хеш-таблица может быть сконструирована таким образом, что ее цифровое представление будет больше, чем первичная хэш-таблица. Например, в различных вариантах первичная хэш-таблица может быть сконфигурирована для хранения хэш-записей по 8 байт каждая, по 8 записей на контейнер хеша, что суммарно дает 64 байта на контейнер, а вторичная хеш-таблица может быть сконфигурирована для хранения 16 хеш-записей, что суммарно дает 128 байт на контейнер. Для каждой хеш-записи, содержащей избыточные биты хеша, те же биты хеш-ключа, возвращаются положения возможных совпадений в референсном геноме. Соответственно, для первичной хеш-таблицы, может определяться до 8 положений. Для вторичной хеш-таблицы может определяться до 16 положений.

[00234] В не зависимости от того, выполнена реализация аппаратно или программно, может быть полезно структурировать хеш-таблицу так, чтобы избежать коллизий.

Например, может присутствовать множество затравок, которые из-за различных артефактов системы, будут стремиться встроиться в хеш-таблицу в том же месте, вне зависимости от того, имеет ли место совпадение или нет. Часто, коллизий можно частично избежать за счет структурирования хеш-таблицы. Соответственно, в различных вариантах хеш-таблица может быть структурирована так, чтобы исключить коллизии, и соответственно, может быть сконфигурирована таким образом, чтобы включать один или более виртуальных хеш-контейнеров.

[00235] В различных вариантах, хеш-таблица может быть структуру с представлением в 8-байтном, 16-байтном, 32-байтном, 64-байтном, 128-байтном формате и т.п. Но в различных примерах реализации может быть полезно представить хеш-таблицу в 64-байтном формате. Это может быть полезно, например, в тех случаях, когда хеш-функция будет использовать доступ к памяти, такой как DRAM (динамическое ОЗУ), например, в стандартном форм-факторе DIMM или SODIMM, как в случае, когда минимальный размер пакета обычно составляет 64 байта. В таком варианте дизайн процессора для доступа к данному запоминающему устройству будет таким, что число байтов, необходимое для образования контейнера в хеш-таблице, также равно 64, и, соответственно, можно реализовать максимальную эффективность. Тем не менее, если таблица структурирована в 32-байтном формате, этого будет недостаточно, поскольку примерно половина байтов, доставляемых в пакете, будет содержать информацию, которая не нужна процессору. Это снизит степень эффективности доставки байтов вполовину. И наоборот, если число байтов, используемое для образования контейнера в хеш-таблице, в несколько раз превышает минимальный размер пакета, например, равно 128, эффективность не будет страдать, поскольку процессору действительно нужно вся информация, возвращаемая при одном обращении. Соответственно, в случаях, когда оптимальный размер пакета доступа к памяти, имеет определенный размер, например, 64 байта, хеш-таблица может быть сконструирована таким образом, чтобы оптимизировать использование размера пакета, как в случае, когда количество байтов, выделенных для представления сегментов в хеш-таблице, обрабатываются функцией картирования, например, 64 байта, совпадает с размером пакета памяти. Соответственно, когда пропускная способность памяти является ограничивающим фактором, хеш-таблицу можно структурировать для оптимального использования этих ограничений.

[00236] Далее, отмечается также, что хотя запись можно уместить в 8 байт, хеш-функцию можно сконструировать таким образом, чтобы не было так, что 8 байт из таблицы считываются для обработки одной записи, поскольку это может быть неэффективно. Вместо этого можно прочитывать все 8 записей в контейнере, можно читать одновременно, или их часть. Это может быть полезно для оптимизации скорости обработки системы, поскольку, с учетом описанной выше архитектуры, на обработку всех 8 записей уйдет столько же времени, сколько на обработку всего лишь 1 записи. Соответственно, в некоторых вариантах модуль картирования может включать хеш-таблицу, которая сама может включать несколько подразделов, например, виртуальных разделов или контейнеров, где каждый контейнер может иметь 1 или более слотов, например, 8 слотов, таким образом, что в него можно поместить одну или более записей, чтобы контролировать коллизии. Тем не менее в некоторых условиях указанные один или более контейнеров, могут полностью заполняться записями, поэтому может быть предложено средство для сортировки дополнительных записей в другие контейнеры и записи информации в исходный контейнер, указывающей на то, что механизм поиска в хеш-таблице должен искать дальше для обнаружения совпадения.

[00237] Соответственно, в некоторых случаях также может быть полезно применять один или более дополнительных способов, например, для контроля коллизий, и один из таких способов может включать одно или более из линейного зондирования и/или формирования хеш-цепочки. Например, если не известно, что точно ищут в хеш-таблице или ее порции, например, в одном контейнере хеш-таблицы, и конкретный контейнер полон то функция поиска в хеше может быть сконфигурирована таким образом, что если один контейнер, и при поиске в нем не обнаруживается нужная запись, то функция может быть направлена в следующий контейнер, например, +1 контейнер, и этот контейнер может быть в свою очередь проверен. Таким образом, при поиске конкретной записи можно проверить все контейнеры. Соответственно, такой поиск можно осуществлять, просматривая один контейнер за другим, до тех пор пока не станет понятно, что запись не найдется, как в случае, когда обнаруживается пустой слот в по меньшей мере одном из контейнеров. В частности, если каждый контейнер заполняется последовательно и поиск по каждому контейнеру проводиться по очереди в соответствии с порядком заполнения, то если находится пустой слот, как в случае, последовательной проверки контейнеров в поиске конкретной записи, пустой слот может указывать на то, что такой записи не существует, поскольку если бы она существовала, она должна была бы быть помещена в этот пустой слот, если не в предыдущие контейнеры.

[00238] Более конкретно, в тех случаях, когда 64 байта выделяются для хранения в хеш-контейнере, где хранится 8 записей, после получения вызванного контейнера процессор картирования может работать со всеми 8 записями одновременно для определения, какие из них совпадают, а какие нет. Например, при осуществлении поиска, такого как поиск затравки из рида, полученного из секвенированной ДНК из образца, относительно затравки, сгенерированной из референсного генома, цифровое представление затравки можно сравнивать с затравками во всех, например, 8, записях для поиска совпадения. Это может привести к нескольким результатам. Может быть найдено прямое совпадение. Затравка образца может быть помещена в хеш-таблица и, в некоторых случаях, совпадение не будет найдено, например, из-за того, что она не является полностью идентичной ни одной соответствующей затравке в референсе, например, из-за аппаратной ошибки или ошибки секвенирования в отношении затравки или рида, из которого она была сгенерирована, или из-за того, что генетическая последовательность этого лица отличается от референсного генома. Или затравка может быть помещена в хеш-таблицу и может быть возвращено множество совпадений, как в случаях, когда затравка образца совпадает с 2, 3, 5, 10, 15, 20 или большим числом мест в таблице. В таком варианте может быть возвращено несколько записей, и все они показывают на различные и разные места в референсном геноме, с которыми совпадает данная конкретная затравка, записи для этих совпадений могут находиться либо в том же контейнере, либо можно произвести зондирование множества контейнеров, в результате чего будут возвращены все значимые результаты, например, совпадения.

[00239] В некоторых вариантах, как в случае, когда, возможно, место становится ограниченным фактором в хеш-таблице, например, в контейнерах хеш-таблицы контейнеры, можно реализовать дополнительный механизм для разрешения коллизий и/или экономии места. Например, когда место ограничено, например, когда необходимо хранить более 8 записей в контейнере, или если когда это желательно, можно выполнить создания цепочки хеширования. Создание хеш-цепочки может включать, например, замену записи, содержащей конкретное положение в геномной последовательности, записью с указателем цепочки, которая вместо того, чтобы указывать на положение в геноме, указывает на некоторый другой адрес, например, второй контейнер в текущей

хеш-таблице, например, первичной хеш-таблице или вторичной хеш-таблице. Преимуществом этого способа перед методом линейного зондирования является то, что он дает механизму поиска в хеше прямой доступ к контейнеру, содержащему нежную запись, вместо последовательной проверки контейнеров по порядку.

[00240] Такой способ может быть полезен для определенной архитектуры системы. Например, хешируемые первичные затравки, как в случае первичного поиска, размещаются в определенном положении в таблице, например, в их исходном положении (позиции), а при сцеплении эти затравки помещают в положение, которое может отличаться от их исходного контейнера. Соответственно, как указано выше, первую часть затравки в цифровом представлении, например, от примерно 26 до примерно 29 бит, можно хешировать и можно искать на первом этапе. На втором этапе можно ввести в хеш-таблицу оставшиеся от примерно 27 до примерно 30 бит, например, как в цепочке хеширования, в качестве средства проверки первого прогона. Соответственно, для любой затравки ее исходный адрес можно хешировать на первом этапе, а биты вторичного адреса можно использовать на втором, проверочном этапе. Соответственно, первая порция затравок может быть помещена в первичное положение записи, а вторая часть может быть помещена в таблицу во вторичном положении записи. И, как указано выше, в различных вариантах, эти два разных положения записи могут быть расположены раздельно, как в формате сцепленной записи. Соответственно, в любом конечном контейнере сцепления сцепленный формат записи может позиционно разделять элементы/записи, которые предназначены для доступа и зондирования локальных первичных контейнеров, и записи, предназначенные для сцепления.

[00241] Такие цепочки хеширования можно продолжать до различных значений длины. Преимущество такого сцепления заключается в том, что в случае, когда один или более контейнеров включают одну или более, например, 2, 3, 4, 5, 6 или боле пустых слотов для записей, эти пустые слоты для записей можно применять для хранения данных цепочек хеширования. Соответственно, в некоторых вариантах сцепление хеша может включать начало с пустого слота в одном контейнере и сцепление этого слота с другим слотом в другом контейнере, где эти два контейнера могут быть расположены в удаленных положениях в хеш-таблице. Также можно принимать меры, чтобы избежать путаницы между записями, размещаемыми в удаленном контейнере в качестве части цепочки хеширования, и «нативными» записями, которые хешируются непосредственно в тот же контейнер. Как обычно, остальные от примерно 27 до примерно 30 бит вторичного ключа доступа хранятся в записях, размещенных на удалении в сцепленном контейнере, но из-за того, что этот сцепленный контейнер находится на расстоянии от исходного хеш-контейнера, подтверждение этих примерно от 27 до 30 будет не достаточно для того, чтобы гарантировать, что совпадающая хеш-запись соответствует исходной затравке, попавшей в этот контейнер в результате сцепления, в отличие от другой затравки, которая достигает того же контейнера в результате прямого доступа. (например, проверка приблизительно от 27 до 30 может быть полной верификацией, если примерно от 26 до 29 бит, используемых для адресации хеш-таблицы, косвенно проверяются по близости к исходному хеш-контейнеру, к которому осуществляют доступ).

[00242] Для того, чтобы избежать возвращения неправильной хеш-записи без необходимости хранить все хеш-ключи в записях, можно использовать позиционную систему в сцепленном контейнере. Соответственно, сцепленный контейнер должен содержать запись в формате непрерывной цепочки, которая содержит дополнительный указатель цепочки для продолжения цепочки контейнеров в случае необходимости; эта

запись продолжения цепочки должна находиться в слоте контейнера после всех “нативных” записей, соответствующих прямому доступу к хешу, и перед всеми удаленными записями, относящимися к цепочки. В ходе обработки запроса перед переходом по какому-либо указателю цепочки, следует игнорировать любые записи, расположенные после записи продолжения цепочки, а после перехода по какому-либо указателю цепочки должна игнорироваться любая запись, возникающая после записи продолжения цепочки.

[00243] Например, в случаях, когда контейнеры заполнены приблизительно на 75%-85%, можно просканировать 8 контейнеров и при этом найти лишь 15-25 слотов, подходящих для использования, в то время как хеш, сцепляющий эти слоты, можно найти в 2, или 3, или 4 контейнерах. В таком варианте имеет значение число этапов зондирования или сцепления, необходимое для хранения хеш-записи, поскольку оно влияет на скорость системы. Во время работы, если для нахождения записи необходимо зондирование, может быть необходимо осуществлять многократный доступ для поиска в хеше, например, 64-байтного рида в контейнере, что замедляет систему. Сцепление хеша помогает минимизировать среднее количество доступов, которые приходится выполнять, поскольку в сцепленный контейнер, который может быть выбран из широкой области, обычно можно помещать больший избыток хеш-записей, чем в контейнер зондирования, которые должны строго следовать по порядку. Соответственно, данное число избыточных хеш-записей обычно можно поместить в более короткую последовательность сцепленных контейнеров, чем в обязательно следующие по порядку контейнеры зондирования, что также ограничивает количество операций доступа, необходимых для локализации этих избыточных записей в запросе. Тем не менее зондирование сохраняет ценность для меньших количества избыточных хеш-записей, поскольку зондирование не требует жертвовать контейнерным слотам в указателе цепочки.

[00244] Например, после того, как было определено, где находятся все возможные совпадения затравок с референсным геномом, необходимо определить, какое из всех этих возможных местоположений, где возможно совпадение данного рида, действительно является правильной позицией, с которой он выровнен. Соответственно, после картирования может быть множество позиций, где один или более ридов предположительно совпадают с референсным геномом. Следовательно, может существовать множество затравок, которые предположительно указывают в точности одно и то же, например, они могут совпадать в точности с одной и той же позицией на референсе, если учитывать позицию затравки в риде.

[00245] Поэтому для каждого данного рида необходимо определить фактическое выравнивание. Это определение можно осуществить несколькими различными способами. В одном случае можно оценить все риды для определения их правильного выравнивания относительно референсного генома на основе позиций, указанных всеми затравками из рида, которые вернули информацию о позиции во время хэшированного поиска. Однако в различных вариантах перед выполнением выравнивания можно выполнить функцию фильтрации затравочной цепочки на одной или более затравок.

[00246] Например, в некоторых вариантах затравки, ассоциированные с данным ридом, которые предположительно картируются на на одно и то же общее место в референсном геноме, могут быть агрегированы в одну цепочку, которая ссылается на ту же общую область. Все затравки, ассоциированные с одним ридом, могут быть сгруппированы в одну или более затравочных цепочек, так, чтобы каждая затравка входила только в одну цепочку. Именно такие цепочки затем приводят к выравниванию

рида с каждой указанной позицией в референсном геноме. В частности, в различных вариантах все затравки, которые имеют одни и те же подтверждающие данные, указывающие на то, что они принадлежат одним и тем же общим местоположениям в референсе, могут быть собраны вместе для формирования одной или более цепочек. Соответственно, затравки, которые группируются вместе, или по меньшей мере создают впечатление, что они окажутся рядом друг с другом в референсном геноме, например, в пределах определенной полосы, будут сгруппированы в цепочку затравок, а затравки за пределами указанной полосы будут превращены в другую цепочку затравок.

[00247] После того как эти различные затравки агрегированы в одну или более различных затравочных цепочек, можно определить, какая из цепочек действительно представляет правильную цепочку, подлежащую выравниванию. Это можно осуществить, по меньшей мере отчасти, с помощью алгоритма фильтрации, который представляет собой эвристический алгоритм, разработанный для устранения слабых затравочных цепочек, с большой вероятностью не являющихся верными. Обычно более длинные затравочные цепочки, применительно к длине, охваченной ридом, с большей вероятностью являются правильными; также затравочные цепочки, куда входит большее число затравок, с большей вероятностью являются правильными. Согласно одному примеру может быть использован эвристический алгоритм, согласно которому относительно сильная «вышестоящая» затравочная цепочка, например, длинная или содержащая много затравок, отфильтровывает относительно слабую «нижестоящую» затравочную цепочку, например, короткую или содержащую малое число затравок.

[00248] В одном варианте реализации длина нижестоящей цепочки определяет пороговую длину, например, двукратную, так, чтобы ее могла отфильтровать вышестоящая цепочка, имеющая по меньшей мере указанную пороговую длину. В другом варианте число затравок в нижестоящей цепочке определяет пороговое число затравок, например, пятикратно большее число затравок, так, чтобы ее могла отфильтровать вышестоящая цепочка, содержащая по меньшей мере указанное пороговое число затравок. В другом варианте длина нижестоящей цепочки определяет пороговое число затравок, например, результат вычитания из двукратно большего числа затравок длины затравок, так, чтобы ее могла отфильтровать вышестоящая цепочка из по меньшей мере порогового числа затравок. В некоторых вариантах, например, если требуется выравнивание химерных ридов, только вышестоящие затравочные цепочки, по существу перекрывающие нижестоящие затравочные цепочки в пределах рида, могут их отфильтровать.

[00249] Указанный процесс отсеивает затравки с низкой вероятностью идентификации области референсного генома, где может быть обнаружено высококачественное выравнивание рида. Это, соответственно, может быть полезно, поскольку уменьшает число выравниваний, которые необходимо выполнить для каждого рида, повышая таким образом скорость обработки и обеспечивая экономию времени. Соответственно, указанный процесс может быть использован, отчасти, в качестве характеристики настройки, с помощью которой, если требуется более высокая скорость, например, в высокоскоростном режиме, осуществляют более подробную фильтрацию затравочных цепочек, а если требуется бóльшая общая точность, например, в режиме повышенной точности, осуществляют менее подробную фильтрацию затравочных цепочек, например, оценивают все затравочные цепочки.

[00250] В различных вариантах реализации может быть проведено редактирование затравок, например, до этапа фильтрации затравочных цепочек. Например, для каждого рида, если все затравки указанного рида обрабатывают функцией картирования и ни

одна из них не возвращает попаданий, может иметь место высокая вероятность одной или более ошибок в риде, например, ошибки секвенатора. В таком случае функция редактирования, например, процесс редактирования с одиночными изменениями, например, процесс редактирования ОНП, может быть выполнена на каждой затравке, например, на затравках, не вернувших совпадений. Например, в позиции X функция редактирования с одиночными изменениями может дать указание заменить означенный нуклеотид на один из 3 других нуклеотидов; определяют, привела ли указанная замена, например, ОНП-замена, к попаданию, например, к совпадению. Указанное редактирование с одиночными изменениями может быть выполнено таким же образом на каждой позиции в затравке и/или на каждой затравке рида, например, с заменой каждого альтернативного основания для каждой позиции в затравке. Кроме того, при введении одиночного изменения в одну затравку эффекты, оказываемые этим изменением на все другие перекрывающие затравки, могут быть определены на основании указанного одиночного изменения.

[00251] Такое редактирование могут также быть выполнено для инсерций, например, когда один из четырех нуклеотидов добавляют в заданную позицию для инсерции, X, и определяют, возникает ли попадание в результате введения указанной замены. Указанное редактирование может быть выполнено для всех четырех нуклеотидов и/или для всех позиций (X, X+1, X+2, X+3 и т.п.) в затравке, и/или для всех затравок в ридах. Такое редактирование может также быть выполнено для делеций, например, когда в затравке делетирован один из четырех нуклеотидов в заданной позиции, X, и определяют, возникает ли попадание в результате введения указанной делеции. Указанное редактирование может быть затем повторено для всех позиций X+1, X+2, X+3 и т.п. Такое редактирование, однако, может значительно увеличивать объемы дополнительной обработки и затраченного времени, например, требуя значительного числа дополнительных поисков, например, 2, или 3, или 4, или 5, или 10, или 50, или 100, или 200 и т.п. Тем не менее, такие дополнительные объемы обработки и затрачиваемого времени могут быть полезны, если при помощи такого редактирования может быть определено действительное попадание, например, совпадение полученное там, где ранее совпадение отсутствовало. В таком случае, как правило, затем может быть определено, что возникла ошибка, а затем - что ошибка была исправлена, с сохранением таким образом рида.

[00252] Также может быть использован дополнительный эвристический алгоритм для определения целесообразности выполнения функции редактирования, выполняющий расчет для определения вероятности получения попадания при выполнении такого редактирования. При достижении определенного порога вероятности, такого как 85% правдоподобие, может быть выполнено такое редактирование затравочных цепочек. Например, указанная система может генерировать различные статистические показатели для затравочных цепочек, такие как вычисление количества присутствующих высокочастотных попаданий и/или количества затравочных цепочек, которые содержат высокочастотные попадания, и таким образом определять, будет ли с некоторой вероятностью редактирование затравочных цепочек иметь значение для определения совпадений. Например, в том случае, когда определено наличие значительной доли высокочастотных попаданий, редактирование затравочных цепочек может быть пропущено, поскольку, вероятно, оно не позволит получить ряд последовательностей, достаточно уникальных, чтобы получить попадание в результате разумного числа поисков в хэш-таблицах, такого как 100 или менее, 50 или менее, 40 или менее, 30 или менее, 20 или менее, или 10 или менее. Такие статистические показатели могут быть

проанализированы и затем может быть принято решение о выполнении или не выполнении редактирования затравок. Например, если статистические показатели показывают, что для любого рида половина позиций показывает отсутствие совпадений, а остальные показывают высокочастотные совпадения, редактирование затравок может быть целесообразным, поскольку при отсутствии возвращения совпадений возможно наличие ошибки, однако при возвращении большого числа высокочастотных совпадений выполнение редактирования затравок может просто не иметь смысла.

[00253] Результатом выполнения одной или более из указанных функций картирования, фильтрации и/или редактирования является список ридов, который включает для каждого рида список всех возможных местоположений, в которых рид может совпасть с референсным геномом. Следовательно, функцию картирования можно выполнить так, чтобы быстро определить, где риды из файла FASTQ, полученного из секвенатора, картируются на референсный геном, например, туда, куда в полном геноме картируются различные риды. Однако при наличии ошибки в любом из ридов или генетической вариации можно не получить точного совпадения с референсом; и/или могут быть несколько мест, с которыми, предположительно, совпадают один или более ридов. Соответственно, необходимо определить, где различные риды действительно выровнены относительно генома в целом.

[00254] Соответственно, после картирования, и/или фильтрации, и/или редактирования определены позиции местоположений для большого количества ридов, причем для некоторых индивидуальных ридов определено множество позиций местоположений, и после этого требуется установить, какие из всех возможных местоположений в действительности являются истинными или наиболее вероятными местоположениями, с которым выравниваются различные риды. Такое выравнивание может быть выполнено с помощью одного или более алгоритмов, таких как алгоритм динамического программирования, который ищет совпадения для картированных ридов с референсным геномом и выполняет функцию выравнивания на нем.

[00255] Пример функции выравнивания сравнивает один или более, например, все риды с референсом, например, путем взаимного наложения в графическом режиме, например, в таблице, такой как виртуальный массив или матрица, где последовательность одного из референсных геномов или картированных ридов помещают на одно измерение или ось, например, на горизонтальную ось, а другую помещают на противоположные измерения или ось, например, на вертикальную ось. Затем воображаемый фронт волны оценки пропускают по массиву для определения выравнивания ридов с референсным геномом, например, путем вычисления оценок выравнивания для каждой ячейки в матрице.

[00256] Фронт волны оценки представляет одну или более, например, все ячейки матрицы, или часть указанных ячеек, которые могут быть оценены независимо и/или одновременно в соответствии с правилами динамического программирования, применимыми к алгоритму выравнивания, такому как алгоритмы Смита-Ватермана, и/или Нидлмана-Вунша, и/или родственные алгоритмы. Например, если за начало матрицы (соответствующее началу рида и/или началу окна референса воображаемого фронта волны оценки) принять левый верхний угол, сначала может оцениваться только верхняя левая ячейка с координатами (0,0) в матрице, например, фронт волны из 1 ячейки; далее могут оцениваться две ячейки в направлении вправо и вниз с координатами (0,1) и (1,0), например, фронт волны из 2 ячеек; далее могут оцениваться три ячейки с координатами (0,2), (1,1) и (2,0), например, фронт волны из 3 ячеек. Указанные примеры фронта волны могут затем расширяться по диагонали вдоль прямых линий, от нижнего

левого угла до верхнего правого, и пошаговое движение фронта волны через матрицу происходит по диагонали из верхнего левого угла к правому нижнему углу. Вычисление оценок выравнивания может быть выполнено последовательно или в другом порядке, например, путем вычисления всех оценок в верхнем ряду в направлении слева направо, а затем всех оценок в следующем ряду в направлении слева направо, и т.п. При таком подходе прокатывающийся по диагонали диагональный фронт волны представляет собой оптимальную последовательность пакетов оценок, вычисляемых одновременно или параллельно в серии этапов фронта волны.

[00257] Например, в одном варианте реализации окно референсного генома, содержащее сегмент, на который был картирован рид, помещают на горизонтальную ось, а рид располагают на вертикальной оси. Подобным образом генерируют массив или матрицу, например, виртуальную матрицу, что позволяет сравнить нуклеотид в каждой позиции в риде с нуклеотидом в каждой позиции в окне референса. По мере прохождения фронта волны по массиву рассматривают все потенциальные способы выравнивания рида с окном референса, в том числе выясняют, требуются ли изменения в одной последовательности, чтобы обеспечить совпадение рида с референсной последовательностью, например, путем изменения одного или более нуклеотидов рида на другие нуклеотиды, или инсерции одного или более новых нуклеотидов в одну последовательность, или делеции одного или более нуклеотидов из одной последовательности.

[00258] Генерируют оценку выравнивания, отражающую степень изменений, которые требуется осуществить для достижения точного выравнивания, причем эта оценка и/или другие связанные данные могут быть сохранены в заданных ячейках массива. Каждая ячейка массива соответствует вероятности того, что нуклеотид в ее позиции на оси ридов выровнен с нуклеотидом в ее позиции на оси референса, а оценка, сформированная для каждой ячейки, представляет частичное выравнивание, заканчивающееся позициями ячейки в риде и окне референса. Наивысшая оценка, сгенерированная в любой клетке, представляет лучшее общее выравнивание рида с окном референса. В различных вариантах выравнивание может быть глобальным, когда весь рид должен быть выровнен с некоторой частью окна референса, например, с применением алгоритма Нидлмана-Вунша или аналогичного алгоритма; или, в других случаях, выравнивание может быть локальным, когда только часть рида может быть выровнена с частью окна референса, например, с применением алгоритма Смита-Ватермана или аналогичного алгоритма.

[00259] Окно референса может иметь любой подходящий размер. Например, поскольку длина типичного рида может составлять от приблизительно 100 до приблизительно 1000 нуклеотидов, длина окна референса, соответственно, в некоторых случаях может составлять от приблизительно 100 до 1000 нуклеотидов или более. Однако в некоторых случаях длина ридов может быть больше, и/или длина окна референса может быть больше, например, может составлять приблизительно 10000, 25000, 50000, 75000, 100000, 200000 нуклеотидов или более. Может быть благоприятным некоторое удлинение окна референса по сравнению с ридом, например, включение 32, или 64, или 128, или 200, или даже 500 дополнительных нуклеотидов в окно референса за пределами сегмента референсного генома, на который был картирован рид, например, чтобы было возможно полностью оценить инсерции и/или делеции возле концов рида. Например, если только часть рида был картирована на сегмент референса, может применяться дополнительное удлинение для окна референса, соответствующего некартированным частям рида, или удлинение с некоторой кратностью, например, на

10%, или 15%, или 20%, или 25%, или даже 50%, или более, чтобы обеспечить полное выравнивание некартированных частей объема рида с окном референса. Однако в некоторых случаях может быть выбрана длина окна референса меньше длины ридов, например, когда длинная часть рида не картирована на референс, например, более или менее 1000 нуклеотидов на одном конце рида, например, для фокусирования на выравнивании картированной части.

[00260] Фронт волны выравнивания может иметь неограниченную длину, или может быть ограничен любой подходящей фиксированной длиной, или иметь вариабельную длину. Например, могут оцениваться все ячейки вдоль полной диагональной линии каждого шага фронта волны, проходящей полностью от одной оси до другой оси. Как вариант, ограниченная длина, например, шириной 64 ячейки, может оцениваться на каждом шаге фронта волны, например, путем прокладывания в матрице диагональной полосы оцениваемых ячеек шириной 64 ячейки, при этом ячейки за пределами указанной полосы остаются без оценки. В некоторых случаях может быть необязательным вычисление оценок далеко от полосы вокруг истинного пути выравнивания, и за счет вычисления оценок только в ограниченной полосе пропускания с использованием фронта волны оценки фиксированной длины, согласно описанию в настоящем документе, может быть сэкономлен существенный объем работы.

[00261] Соответственно, в различных вариантах функция выравнивания может быть выполнена, например, на данных, полученных из модуля картирования. Соответственно, в различных вариантах функция выравнивания может образовывать модуль, например, модуль выравнивания, который может образовывать часть системы, например, конвейер, который используют, например, наряду с модулем картирования, в процессе определения фактической полной геномной последовательности индивидуума или ее части. Например, выходные данные, возвращаемые после выполнения функции картирования, например, из модуля картирования, например, список возможных вариантов картирования одного или более, или всех ридов на одну или более позиций в одном или более референсных геномах, могут быть использованы функцией выравнивания для определения фактического выравнивания последовательности секвенированной ДНК субъекта.

[00262] Такая функция выравнивания может быть полезна во всех случаях, поскольку, согласно описанию выше, часто по разнообразным причинам секвенированные риды не обязательно точно совпадают с референсным геномом. Например, в одном или более ридов может присутствовать ОНП (однонуклеотидный полиморфизм), например, замена одного нуклеотида на другой в единственной позиции; может присутствовать «индел», инсерция или делеция одного или более оснований на протяжении одной или более последовательностей рида, не присутствующие в референсном геноме; и/или может присутствовать ошибка секвенирования (например, ошибки приготовления образцов, и/или ридов секвенатора, и/или выходных данных секвенатора и т.п.), вызывающая одну или более из указанных очевидных вариаций. Соответственно, когда рид отличается от референса, например, наличием ОНП или индела, это может быть обусловлено отличием референса от истинной последовательности ДНК в образце, или отличием рида от истинной последовательности ДНК в образце. Проблема состоит в том, чтобы выполнить корректное выравнивание ридов с референсным геномом, с учетом того факта, что, по всей вероятности, между указанными двумя последовательностями будет множество разных отличий.

[00263] Соответственно, в различных вариантах, входными данными функции выравнивания, например, из функции картирования, такой как дерево префиксов/суффиксов, или преобразование Барроуза-Уилера, или хэш-таблица, и/или хэш-функция,

может быть список возможных вариантов совпадений одного или более ридов с одной или более позиций в одной или более референсных последовательностей. Например, любой данный рид может совпадать с любым количеством позиций в референсном геноме, например, в 1 местоположении или в 16, или в 32, или в 64, или в 100, или в 500, или в 1000 или более местоположениях, где данный рид картируется на геном. Однако любой отдельный рид был получен, например, секвенирован, только из одной определенной части генома. Соответственно, чтобы найти истинное местоположение, из которого происходит данный конкретный рид, можно выполнить функцию выравнивания, например, выравнивание с гэпами Смита-Ватермана, выравнивание Нидлмана-Вунша и т.п., для определения фактического места происхождения в геноме одного или более ридов, например, путем сравнения всех возможных местоположений, где имеет место совпадение, и определения того, какой из всех возможных вариантов является наиболее вероятным местоположением, из которого был секвенирован рид, исходя из наивысшей оценки выравнивания местоположений.

[00264] Как было указано, для выполнения такой функции выравнивания обычно используют алгоритм. Например, для выравнивания двух или более последовательностей друг относительно друга можно использовать алгоритм выравнивания Смита-Ватермана и/или Нидлмана-Вунша. В этом случае они могут быть использованы так, чтобы для любой данной позиции, где рид картируется на референсный геном, определить вероятности того, что картирование действительно выполнено в позиции, откуда происходит рид. Как правило, эти алгоритмы выполнены с возможностью осуществления программным обеспечением, однако в различных вариантах реализации, таких как представленные в настоящем документе, один или более из этих алгоритмов может быть выполнен с возможностью осуществления в аппаратном обеспечении, согласно более подробному описанию ниже в настоящем документе.

[00265] В частности, функцию выравнивания используют по меньшей мере частично для выравнивания одного или более, например, всех ридов на референсный геном, несмотря на наличие одной или более несовпадающих частей, например, ОНП, инсерций, делеций, структурных артефактов и т.п., для определения того, где указанные риды, вероятно, правильно впишутся в геном. Например, один или более ридов сравнивают с референсным геномом, и определяют наилучшее возможное совпадение рида с геномом, учитывая при этом замены, и/или инделы, и/или структурные варианты. Однако для того, чтобы лучше определить, какая из модифицированных версий рида лучше всего вписывается в референсный геном, необходимо учитывать предполагаемые изменения, и, соответственно, можно также выполнить функцию оценки.

[00266] Например, можно выполнить функцию оценки, например, как часть общей функции выравнивания, при этом при выполнении модулем выравнивания своей функции и введении одного или более изменений в последовательность, сравниваемую с другой последовательностью, например, например, чтобы достичь более хорошего или наилучшего соответствия между ними, для каждого изменения, вносимого для достижения более хорошего выравнивания, из начальной оценки, например, либо из идеальной оценки, либо из нулевой начальной оценки вычитают некоторое число, так, чтобы при выполнении выравнивания определять также оценку этого выравнивания, например, когда обнаруживают совпадения, оценку увеличивают, а при каждом внесенном изменении накладывают штраф; таким образом, может быть определено лучшее возможное соответствие для возможных выравниваний, например, путем выявления из всех возможных модифицированных ридов того рида, соответствие которого геному имеет наивысшую оценку. Соответственно, в различных вариантах

функция выравнивания может быть выполнена с возможностью определения лучшей комбинаций изменений, которые нужно внести в рид(-ы) для достижения выравнивания с наивысшей оценкой, которое может быть, соответственно, определено как правильное или наиболее вероятное выравнивание.

[00267] Соответственно, ввиду вышеизложенного, по меньшей мере две цели могут быть достигнуты за счет выполнения функции выравнивания. Одна из целей представлена отчетом о наилучшем выравнивании, включающим в себя позицию в референсном геноме и описание изменений, которые необходимы для того, чтобы рид совпал с референсным сегментом в указанной позиции, а другая цель представлена оценкой качества выравнивания. Например, в различных вариантах выходные данные из модуля выравнивания могут представлять собой отчет «Compact Idiosyncratic Gapped Alignment Report», например, строку CIGAR, при этом указанная выходная строка CIGAR представляет собой отчет, подробно описывающий все изменения, внесенные в риды для достижения их наиболее соответствующего выравнивания, например, подробные инструкции по выравниванию, указывающие, каким образом происходит действительное выравнивание исследуемой последовательности с референсом. Вывод такой строки CIGAR может быть полезным на последующих стадиях обработки для более хорошего определения того, что для данной геномной нуклеотидной последовательности субъекта прогнозированные вариации в сравнении с референсным геномом действительно являются истинными вариациями, а не просто обусловлены ошибкой машины, программного обеспечения или человека.

[00268] Как было указано выше, в различных вариантах реализации выравнивание, как правило, выполняют последовательно, причем алгоритм принимает данные последовательности рида, например, из модуля картирования, принадлежащие риду, и одно или более возможных местоположений, где этот рид потенциально может быть картирован на один или более референсных геномов, а также принимает данные геномной последовательности, например, из одной или более памятей, относящиеся к одной или более позиций в одном или более референсных геномах, на которые может быть картирован рид. В частности, в различных вариантах реализации модуль картирования обрабатывает риды, например, из файла FASTQ, и картирует каждый из них на одну или более позиций в референсном геноме, на которые они, возможно, выровнены. Затем выравниватель берет указанные прогнозированные позиции и использует их для выравнивания ридов на референсный геном, например, путем построения виртуального массива, с помощью которого риды можно сравнивать с референсным геномом.

[00269] При выполнении этой функции выравниватель оценивает каждую картированную позицию для каждого отдельного рида и, в частности, оценивает те риды которые картированы на множество возможных местоположений в референсном геноме, и для каждой позиции оценивает возможность того, что она является правильной. Затем он сравнивает лучшие оценки, например, две лучшие оценки, и принимает решение о том, где действительно выравнивается конкретный рид. Например, при сравнении первой и второй лучших оценок выравнивания выравниватель проверяет разницу между оценками, и если разница между ними большая, то оценка достоверности того, что позиция с большей оценкой является правильной, будет высокой. Однако если разница между ними маленькая, например, нулевая, то оценка достоверности выбора из двух позиций одной из них в качестве правильной позиции, из которой получен рид, низкая; и, возможно, будет полезна дополнительная обработка, чтобы четко определить истинное местоположение в референсном геноме, из которого получен рид.

Соответственно, выравниватель, в частности, ищет наибольшую разницу между первой и второй лучшими оценками достоверности для принятия решения о том, что данный рид картируется на данное местоположение в референсном геноме. В идеале оценка лучшего возможного варианта выравнивания значительно выше оценки второго лучшего выравнивания для данной последовательности.

[00270] Существует множество различных способов реализации метода оценки выравнивания, например, можно оценивать каждую ячейку массива или подмножество ячеек, например, в соответствии со способами, описанными в настоящем документе. Как правило, каждое совпадение при выравнивании, соответствующее шагу по диагонали в матрице выравнивания, вносит вклад в положительную оценку, например, +1, если соответствующие нуклеотиды рида и референса; и в отрицательную оценку, например, -4, если два нуклеотида не совпадают. Далее, каждая делеция в референсе, соответствующая шагу в горизонтальном направлении в матрице выравнивания, вносит вклад в отрицательную оценку, например, -7, и каждая инсерция в референсе, соответствующая шагу в вертикальном направлении в матрице выравнивания, вносит вклад в отрицательную оценку, например, -7.

[00271] В различных вариантах параметры оценки для совпадений нуклеотидов, несовпадений, инсерций и делеций нуклеотидов могут принимать любые различные положительные или отрицательные, или нулевые значения. В различных вариантах указанные параметры оценки могут быть модифицированы на основании доступной информации. Например, в некоторых вариантах аффинная функция накладывает штраф на гэпы при выравнивании (инсерции или делеции) в зависимости от длины гэпа, например, -7 для первого делетированного (или, соответственно, инсертированного) нуклеотида, но только -1 для каждого дополнительного делетированного (или, соответственно, инсертированного) нуклеотида в непрерывной последовательности. В различных вариантах реализации аффинные штрафы на гэп могут осуществляться путем разделения штрафов на гэп (за инсерцию или делецию) на два компонента, например, на штраф на открытие гэпа, например, -6, применяемого к первому шагу в гэпе; и штраф на продление гэпа, например, -1, применяемого к каждому или последующим шагам в гэпе. Аффинные штрафы на гэп могут обеспечивать более точное выравнивание, например, позволяя выравниваниям, включающим длинные инсерции или делеции, получать адекватно высокие оценки. Далее, стоимость каждого перемещения по горизонтали может быть одинаковой или разной, например, стоимость шага может быть одинаковой, и/или, при наличии гэпов, такие гэпы могут иметь более высокую или более низкую стоимость, таким образом, стоимость перемещений выравнивателя по горизонтали может быть меньше, чем стоимость гэпов. Соответственно, в различных вариантах реализации может быть реализована аффинная оценка гэпов, однако это может требовать больших расходов на программное обеспечение и/или аппаратное обеспечение, поскольку она, как правило, требует множества, например, 3 оценок, для каждой подлежащей оцениванию ячейки, и, соответственно, согласно различным вариантам реализации аффинную оценку гэпов не реализуют.

[00272] В различных вариантах параметры оценки могут также быть чувствительны к «оценкам качества оснований», соответствующим нуклеотидам в риде. Некоторые данные ридов секвенированной ДНК, в таких форматах, как FASTQ, могут включать оценку качества оснований, ассоциированную с каждым нуклеотидом, указывающую на расчетную вероятность некорректности нуклеотида, например, из-за ошибки секвенирования. В некоторых данных ридов оценки качества оснований могут указывать

на вероятность наличия инсерционной и/или делеционной ошибки секвенирования в каждой позиции или в смежной с ней позиции, или дополнительные оценки качества могут обеспечивать эту информацию отдельно. Более точное выравнивание, соответственно, может быть достигнуто путем использования параметров оценки, включающих любые или все из оценок совпадения нуклеотидов, оценок несовпадения нуклеотидов, штрафов на гэп (инсерцию и/или делецию), штрафов на открытие гэпа и/или продление гэпа, варьирующих в зависимости от оценки качества оснований, ассоциированной с текущим нуклеотидом или текущей позицией рида. Например, бонусы и/или штрафы при оценке могут быть снижены, если оценка качества оснований указывает на высокую вероятность наличия ошибки секвенирования или другой ошибки. Чувствительная в качеству оснований оценка может быть реализована, например, с применением фиксированной или конфигурируемой таблицы подстановки, к которой обращаются с применением оценки качества оснований, возвращающей соответствующие параметры оценки.

[00273] В случае аппаратной реализации в интегральной схеме, такой как FPGA, ASIC или структурированной ASIC, фронт волны оценки может быть реализован в виде линейного массива ячеек оценки, например, 16 ячеек, или 32 ячеек, или 64 ячеек, или 128 ячеек или т.п. Каждая из ячеек оценки может быть построена из цифровых логических элементов в монтажной конфигурации для вычисления оценок выравнивания. Соответственно, для каждого этапа фронта волны, например, каждого тактового цикла, или некоторой другой фиксированной или вариабельной единицы времени, каждая из ячеек оценки, или часть ячеек, вычисляет оценку или оценки, требуемые для новой ячейки в виртуальной матрице выравнивания. Теоретически считается, что различные ячейки оценки находятся в различных позициях матрицы выравнивания, соответствующих фронту волны оценки согласно описанию в настоящем документе, например, вдоль прямой линии, проходящей из нижней левой части в верхнюю правую часть матрицы. Как хорошо известно в области разработки цифровых логических устройств, физические ячейки оценки и составляющая их цифровая логика не обязательно должны быть физически расположены подобным образом на интегрированной схеме.

[00274] Соответственно, по мере того, как фронт волны шаг за шагом прокатывается по виртуальной матрице выравнивания, воображаемые позиции ячеек оценки соответствующим образом обновляют каждую ячейку, например, умозрительно «перемещаются» на шаг вправо или, например, на шаг вниз в матрице выравнивания. Все ячейки оценки совершают одинаковое относительное воображаемое перемещение, сохраняя порядок диагонального фронта волны. Каждый раз, когда фронт волны перемещается в новое положение, например, за счет шага в вертикальном направлении вниз или шага в горизонтальном направлении вправо в матрице, ячейки оценки прибывают в новые воображаемые позиции и вычисляют оценки выравнивания для ячеек виртуальной матрицы выравнивания, в которые они вошли.

[00275] В такой реализации соседние ячейки оценки в линейном массиве соединены для обмена исследуемыми (принадлежащими риду) нуклеотидами, референсными нуклеотидами и ранее вычисленными оценками выравнивания. Нуклеотиды окна референса могут последовательно подаваться на один конец фронта волны, например, в верхнюю правую ячейку оценки в линейном массиве, и могут последовательно сдвигаться оттуда вниз вдоль фронта волны, так, что в любой данный момент времени сегмент референсных нуклеотидов, равный по длине количеству ячеек оценки, присутствует в этих ячейках, по одному из следующих один за другим нуклеотидов в каждой следующей одна за другой ячейке оценки.

[00276] Соответственно, каждый раз, когда фронт волны перемещается на шаг в горизонтальном направлении, следующий референсный нуклеотид подается в верхнюю правую ячейку, а другие референсные нуклеотиды сдвигаются вниз и влево по фронту волны. Указанный сдвиг референсных нуклеотидов может быть реальным отражением воображаемого перемещения фронта волны ячеек оценки вправо в матрице выравнивания. Соответственно, нуклеотиды рида могут последовательно подаваться на противоположный конец фронта волны, например, в нижнюю левую ячейку оценки в линейном массиве, и могут последовательно сдвигаться оттуда вверх вдоль фронта волны, чтобы в любой данный момент времени сегмент исследуемых нуклеотидов, равный по длине количеству ячеек оценки, присутствовал в этих ячейках, по одному из следующих один за другим нуклеотидов в каждой следующей одна за другой ячейке оценки.

[00277] Сходным образом, каждый раз, когда фронт волны перемещается на шаг в вертикальном направлении, следующий исследуемый нуклеотид подают в нижнюю левую ячейку, а другие исследуемые нуклеотиды сдвигаются вверх вправо по фронту волны. Указанный сдвиг исследуемых нуклеотидов является реальным отражением воображаемого перемещения фронта волны ячеек оценки вниз в матрице выравнивания. Соответственно, подавая команду на сдвиг референсных нуклеотидов, можно перемещать фронт волны на шаг в горизонтальном направлении, а подавая команду на сдвиг исследуемых нуклеотидов, можно перемещать фронт волны на шаг в вертикальном направлении. Соответственно, для получения в целом диагонального перемещения фронта волны, например, чтобы следовать типичному выравниванию исследуемой и референсной последовательностей без инсерций или делеций, можно попеременно подавать команды на перемещение фронта волны на шаг в вертикальном и горизонтальном направлениях.

[00278] Соответственно, соседние ячейки оценки в линейном массиве могут быть соединены для обмена ранее вычисленными оценками выравнивания. В различных алгоритмах оценки выравнивания, например, Смита-Ватермана или Нидлмана-Вунша, или подобных вариантах оценка или оценки выравнивания в каждой ячейке виртуальной матрицы выравнивания могут быть вычислены на основании ранее вычисленных оценок в других ячейках матрицы, например, трех ячейках, расположенных непосредственно слева от текущей ячейки, выше текущей ячейки и вверх влево по диагонали от текущей ячейки. Когда ячейка оценки вычисляет новую оценку или оценки для другой позиции матрицы, в она должна извлечь такие ранее вычисленные оценки, соответствующие таким другим позициям матрицы. Указанные ранее вычисленные оценки могут быть получены из хранилища ранее вычисленных оценок внутри этой же ячейки и/или из хранилища ранее вычисленных оценок в одной или двух соседних ячейках оценки в линейном массиве. Это обусловлено тем, что три вносящие вклад в оценку позиции в виртуальной матрице выравнивания (непосредственно слева, сверху и сверху слева по диагонали) могли быть оценены либо текущей ячейкой оценки, либо одной из соседних с ней ячеек оценки в линейном массиве.

[00279] Например, ячейка непосредственно слева в матрице могла быть оценена текущей ячейкой оценки, если самый последний шаг фронта волны был произведен в горизонтальном направлении (вправо), или могла быть оценена соседней ячейкой снизу слева в линейном массиве, если самый последний шаг фронта волны был произведен в вертикальном направлении (вниз). Аналогичным образом, ячейка непосредственно сверху в матрице могла быть оценена текущей ячейкой оценки, если самый последний шаг фронта волны был произведен в вертикальном направлении (вниз), или могла быть

оценена соседней ячейкой сверху справа в линейном массиве, если самый последний шаг фронта волны был произведен в горизонтальном направлении (вправо). Аналогичным образом, ячейка сверху слева по диагонали в матрице могла быть оценена текущей ячейкой оценки, если два самых последних шага фронта волны были произведены в разных направлениях, например, вниз, а затем вправо; или вправо, а затем вниз; или могла быть оценена соседней ячейкой сверху справа в линейном массиве, если два самых последних шага фронта волны были оба в горизонтальном направлении (вправо), или могла быть оценена соседней ячейкой снизу слева в линейном массиве, если два самых последних шага фронта волны были оба в вертикальном направлении (вниз).

[00280] Соответственно, с учетом информации о направлениях последних одного или двух шагов фронта волны, ячейка оценки может выбрать надлежащие ранее вычисленные оценки, получив доступ к ним внутри себя и/или в соседних ячейках оценки, задействуя соединение между соседними ячейками. В одном варианте для ячеек оценки, располагающихся на двух концах фронта волны, на наружных входах для оценок могут быть жестко смонтированы недопустимые, или нулевые, или имеющие минимальное значение оценки, чтобы они не влияли на новые вычисления оценок в этих крайних ячейках.

[00281] Благодаря реализуемому таким образом фронту волны в линейном массиве ячеек оценки, с таким соединением для сдвига референсных и исследуемых нуклеотидов по массиву в противоположных направлениях, для умозрительного перемещения фронта волны пошагово в вертикальном и горизонтальном направлении, и соединении для доступа к оценкам, ранее вычисленным соседними ячейками, чтобы вычислять оценку или оценки выравнивания в новых позициях ячеек виртуальной матрицы, в которые входит фронт волны, можно, соответственно, оценивать полосу ячеек в виртуальной матрице, ширину фронта волны, например, путем подачи команд на последовательное пошаговое перемещение фронта волны, чтобы он прокатился по матрице. Чтобы выровнять новый рид и окно референса, соответственно, фронт волны может начинаться изнутри матрицы оценки, или, предпочтительно, может постепенно входить в матрицу оценки снаружи, начиная, например, слева, или сверху, или по диагонали слева и сверху с верхнего левого угла матрицы.

[00282] Например, фронт волны может начинаться с его верхней левой ячейки оценки, расположенной сразу слева от верхней левой ячейки виртуальной матрицы, и затем фронт волны может вкатываться вправо в матрицу с помощью серии горизонтальных шагов, оценивая горизонтальную полосу ячеек в верхней левой области матрицы. Когда фронт волны достигает прогнозированного соотношения выравнивания между референсной и исследуемой последовательностью, или когда обнаруживается совпадение на основе возрастания оценок выравнивания, фронт волны может начать прокатываться по диагонали вниз вправо, за счет попеременных шагов в вертикальном и горизонтальном направлении, оценивая диагональную полосу клеток посередине матрицы. Когда нижняя левая ячейка оценки фронта волны достигает нижней границы матрицы выравнивания, фронт волны может начать снова прокатываться вправо за счет последовательных шагов в горизонтальном направлении, оценивая горизонтальную полосу ячеек в нижней правой области матрицы, до тех пор, пока некоторые или все ячейки фронта волны не выйдут за границы матрицы выравнивания.

[00283] В одном варианте повышенной эффективности фронта волны выравнивания можно добиться путем разделения его ячеек оценки между двумя последовательными операциями выравнивания. Следующую матрицу выравнивания устанавливают заранее,

так как когда верхняя правая часть фронта волны выходит из нижней правой области текущей матрицы выравнивания, она может сразу же или после пересечения минимального зазора, такого как одна ячейка или три ячейки, входить в верхнюю правую область следующей матрицы выравнивания. Таким образом горизонтальный выход фронта волны из одной матрицы выравнивания может представлять собой то же самое действие, что и горизонтальный вход фронта волны в следующую матрицу выравнивания. Осуществление указанного способа может включать поступление референсных и исследуемых оснований следующего выравнивания в ячейки оценки, переходящих в следующую матрицу выравнивания, и может уменьшать среднюю продолжительность времени, необходимого для каждого выравнивания, на время выполнения ряда шагов фронта волны, практически равного числу ячеек выравнивания во фронте волны, например, 64 или 63 шагов, или 61 шага, которые могут занимать, например, 64 или 63 тактовых цикла, или 61 тактовый цикл.

[00284] Число ячеек оценки при реализации фронта волны выравнивания может быть выбрано таким образом, чтобы уравновешивать различные факторы, в том числе точность выравнивания, максимальную длину инсерции и делеции, площадь, стоимость и потребление энергии цифровыми логическими устройствами, тактовую частоту логики выравнивателя и производительность общей интегральной схемы. Длинный фронт волны желателен для хорошей точности выравнивания, в частности, поскольку фронт волны из N ячеек способен выполнять выравнивание поверх инделов длиной приблизительно N нуклеотидов, или немного короче. Однако при использовании более длинного фронта волны требуется больше логических устройств, что ведет к потреблению большего количества энергии. Далее, более длинный фронт волны может повышать сложность трассировки и увеличивать задержки на интегральной схеме, что приводит к более низким максимальным тактовым частотам, уменьшающим чистую производительность выравнивателя. Более того, если интегральная схема отличается ограниченным размером или потреблением энергии, применение более длинного фронта волны может требовать реализации меньшего количества логических устройств на любых участках ИС, например, воспроизведения меньшего количества полных фронтов волны, или других компонентов логики выравнивателя или картирующего устройства, что снижает чистую производительность ИС. В одном конкретном варианте реализации 64 ячейки оценки во фронте волны могут обеспечивать приемлемый баланс указанных факторов.

[00285] Соответственно, при ширине фронта волны X ячеек, например, 64 ячейки оценки, ширина оцениваемой полосы в матрице выравнивания, аналогичным образом, будет составлять 64 ячейки (при измерении по диагонали). Ячейки матрицы за пределами указанной полосы не обязательно обрабатывать, а также вычислять их оценки, при условии, что путь оптимального (с наилучшей оценкой) выравнивания в матрице остается в пределах оцениваемой полосы. Для относительно небольшой матрицы, соответственно, используемой для выравнивания относительно коротких ридов, например, ридом длиной 100 нуклеотидов или 250 нуклеотидов это может быть безопасным допущением, например, если фронт волны покатывается точно по диагонали вдоль прогнозированной выровненной позиции рида.

[00286] Однако в некоторых случаях, например, в случае большой матрицы выравнивания, используемой для выравнивания длинных ридов, например, длиной 1000, или 10000, или 100000 нуклеотидов, может присутствовать существенный риск накопления инделов, вызывающего отклонение истинного выравнивания от точной диагонали, суммарно достаточно сильное для ухода из оцениваемой полосы. В таких

случаях может быть полезно наведение фронта волны таким образом, чтобы максимальный набор оценок располагался возле центра фронта волны. Таким образом, по мере прокатывания фронта волны, если наивысшие оценки начинают двигаться в ту или иную сторону, например, слева направо, фронт волны сдвигается, отслеживая указанное движение. Например, если наивысшие оценки наблюдаются в ячейках оценки, расположенных существенно выше и правее центра фронта волны, указанный фронт волны может быть наведен на некоторое расстояние точно направо посредством поступательных шагов в горизонтальном направлении, до возвращения наивысших оценок в область около центра фронта волны.

[00287] Соответственно, автоматический наводящий механизм может быть реализован в логике контроля фронта волны для определения целевой позиции для наведения в пределах длины фронта волны, на основании текущих и прошлых оценок, наблюдаемых в ячейках фронта волны оценки, и для наведения фронта волны на указанную мишень, если она находится за пределами центра. Более конкретно, позиция с максимальной оценкой в последней оцененной позиции фронта волны может быть использована в качестве мишени для наведения. В некоторых случаях этот способ эффективен. Однако в некоторых случаях позиция с максимальной оценкой может быть неудовлетворительной мишенью для наведения. Например, при некоторых комбинациях параметров оценки выравнивания, при возникновении длинного индела, когда, соответственно, оценки начинают снижаться, вдоль фронта волны может формироваться паттерн с двумя пиками более высоких оценок и впадиной более низких оценок между ними, и по мере продолжения индела указанные два пика продолжают расходиться.

[00288] Поскольку невозможно легко определить, представляет ли собой прогрессирующее явление инсерцию или делецию, важно отслеживать фронт волны по диагонали до повторного начала успешного совпадения, либо на некотором расстоянии в направлении направо в случае делеции, либо на некотором расстоянии в направлении вниз в случае инсерции. Однако при формировании двух распространяющихся пиков оценки один из них, вероятно, будет несколько выше другого, и может притянуть автоматическое наведение в этом направлении, приводя к потере выравнивания фронтом волны в том случае, если фактический индел располагался в другом направлении. Более устойчивый способ, соответственно, может заключаться в вычитании дельта-значения из максимальной наблюдаемой оценки фронта волны для определения пороговой оценки, идентификации двух крайних ячеек оценки, по меньшей мере равных указанной пороговой оценке, и использовании средней точки между указанными крайними ячейками в качестве мишени для наведения. Указанный способ обеспечивает тенденцию направления по диагонали посередине паттерна оценки с двумя пиками. Могут, однако, легко применяться и другие критерии наведения, которые служат для поддержания более высоких оценок возле центра фронта волны. При задержке реакции между получением оценки от фронта волны ячеек оценки и принятием соответствующего решения о наведении может быть целесообразно использование гистерезиса для компенсации решений о наведении, принимаемых в промежуточном периоде, чтобы избежать осциллирующих паттернов автоматического наведения фронта волны.

[00289] Одна или более таких процедур выравнивания может быть выполнена с применением любого подходящего алгоритма выравнивания, такого как алгоритм выравнивания Нидлмана-Вунша и/или алгоритм выравнивания Смита-Ватермана, который мог быть модифицирован с учетом функциональных характеристик, описанных в настоящем документе. В общем случае, оба указанных алгоритма и сходные с ними алгоритмы в целом функционируют, в некоторых случаях, аналогичным образом.

Например, как было указано выше, указанные алгоритмы выравнивания, как правило, строят виртуальный массив аналогичным образом, например, в различных вариантах, горизонтальная верхняя граница может сконфигурирована для представления геномной референсной последовательности, которая может быть распределена по верхнему ряду массива в соответствии с составом пар оснований. Сходным образом, вертикальная граница может быть сконфигурирована для представления секвенированных и картированных исследуемых последовательностей, которые были расположены в порядке в направлении сверху вниз в первом столбце, таким образом, чтобы порядок последовательности их нуклеотидов в целом совпадал с последовательностью нуклеотидов референса, на который они картированы. Промежуточные ячейки могут затем быть заполнены оценками в соответствии с вероятностью того, что релевантное основание исследуемой последовательности в заданной позиции находится в указанном местоположении применительно к референсу. При выполнении указанной функции полоса захвата может перемещаться по диагонали по матрице, заполняя промежуточные ячейки оценками, и может быть определена вероятность нахождения каждого основания исследуемой последовательности в указанной позиции.

[00290] Применительно к функции выравнивания Нидлмана-Вунша, которая генерирует оптимальные глобальные (или полуглобальные) выравнивания, выравнивающие всю последовательность рида с некоторым сегментом референсного генома, наведение фронта волны может быть сконфигурировано таким образом, чтобы он в общем прокатывался полностью от самого верхнего края матрицы выравнивания до нижнего края. После того как фронт волны прокатился, выбирают максимальную оценку на нижнем крае матрицы выравнивания (соответствующем концу рида), и выравнивание отслеживают в обратном направлении до ячейки на верхнем краю матрицы (соответствующем началу рида). В различных случаях, описанных в настоящем документе, риды могут быть любой длины, могут быть любого размера, и не обязательно требуется всесторонние параметры рида для определения того, как осуществляется выравнивание, например, в различных вариантах, длина рида может соответствовать длине хромосомы. В таком случае, однако, размер памяти и длина хромосомы могут представлять собой ограничивающий фактор.

[00291] Применительно к алгоритму Смита-Ватермана, который генерирует оптимальные локальные выравнивания, выравнивающие всю последовательность или часть последовательности рида с некоторым сегментом референсного генома, указанный алгоритм может быть сконфигурирован для нахождения наилучшей возможной оценки на основании полного или частичного выравнивания рида. Соответственно, в различных вариантах оцениваемая фронтом волны полоса может не доходить до верхнего и/или нижнего краев матрицы выравнивания, например, если очень длинный рид содержит только затравки в середине картирования на референсный геном, однако, как правило, фронт волны все же может выполнять оценку с верхнего до нижнего края матрицы. Локальное выравнивание, как правило, достигается посредством двух регулировок. Во-первых, запрещено падение оценок выравнивания ниже нуля (или некоторого другого минимального уровня), и если вычисленная иным образом оценка ячейки принимает отрицательное значение, его заменяют на нулевую оценку, соответствующую началу нового выравнивания. Во-вторых, максимальную оценку выравнивания, полученную в любой ячейке в матрице, не обязательно расположенную вдоль нижнего края, используют в качестве окончания выравнивания. Выравнивание отслеживают в обратном порядке от этой максимальной оценки вверх и влево по матрице до нулевой оценки, которую используют в качестве стартовой позиции локального выравнивания,

даже если она не располагается в верхнем ряду матрицы.

[00292] Принимая во внимание вышеизложенное, имеется несколько разных возможных путей через виртуальный массив. В различных вариантах реализации фронт волны начинается с верхнего левого угла виртуального массива и движется вниз к идентификаторам максимальной оценки. Например, результаты всех возможных выравниваний могут быть собраны, обработаны, коррелированы и оценены для определения максимальной оценки. Когда достигнут конец границы или конец массива, и/или произведено давшее наивысшую оценку вычисление для всех обработанных ячеек (например, идентифицирована общая наивысшая оценка), может быть выполнено обратное отслеживание, чтобы найти путь, приведший к указанной наивысшей оценке.

[00293] Например, может быть идентифицирован путь, который приводит к прогнозированной максимальной оценке, и после идентификации может быть выполнена проверка, чтобы определить, каким образом была получена указанная максимальная оценка, например, путем обратного перемещения вдоль указателей выравнивания с наилучшей оценкой, с прослеживанием пути, который привел к достижению идентифицированной максимальной оценки, например, вычисленной с помощью ячеек оценки фронта волны. Указанная обратная реконструкция или обратное отслеживание включает начало движения с определенной максимальной оценкой, и возвращение назад через предыдущие ячейки, с прокладыванием пути через ячейки с оценками, которые привели к достижению максимальной оценки, вверх до края таблицы и обратно к начальной границе, например, к началу массива, или к нулевой оценке в случае локального выравнивания.

[00294] Во время обратного отслеживания, после достижения конкретной ячейки в матрице выравнивания следующий шаг обратного отслеживания совершается в соседнюю ячейку непосредственно слева или сверху, или по диагонали вверх и налево, которая внесла вклад в наилучшую оценку, выбранную для конструирования оценки в текущей ячейке. Указанным образом может быть определена эволюция максимальной оценки, с выяснением таким образом, как была достигнута максимальная оценка. Обратное отслеживание может быть завершено в углу, или на крае, или на границе, или может быть завершено на нулевой оценке, например, в верхнем левом углу массива. Соответственно, именно с помощью такого обратного отслеживания идентифицируют правильное выравнивание с получением таким образом строки вывода CIGAR, например, 3M, 2D, 8M, 4I, 16M и т.п., показывающую, каким образом образец геномной последовательности, происходящий от индивидуума, или его часть совпадает или иным образом выравнивается с геномной последовательностью референсной ДНК.

[00295] Соответственно, после того как было определено, где картирован каждый рид, и дополнительно определено, где выровнен каждый рид, например, каждому релевантному риду была присвоена позиция и оценка качества, отражающая вероятность того, что указанная позиция является правильным выравниванием, так что последовательность нуклеотидов ДНК субъекта известна, может быть верифицирован порядок различных ридов и/или геномная последовательность нуклеиновой кислоты субъекта, например, путем выполнения функции обратного отслеживания, передвигающейся в обратном направлении вверх по массиву, с определением идентичности каждой нуклеиновой кислоты в правильном порядке в образце геномной последовательности. Следовательно, согласно некоторым аспектам настоящее изобретение относится к функции обратного отслеживания, например, представляющей собой часть модуля выравнивания, выполняющего как выравнивание, так и функцию обратного отслеживания, например, модуля, который может быть частью конвейера

модулей, такого как конвейер, который предназначен для приема необработанных данных рида последовательности, например, в виде геномного образца индивидуума, и картирования и/или выравнивания этих данных, которые могут быть затем классифицированы.

[00296] Для облегчения операции обратного отслеживания полезно сохранять вектор оценки для каждой оцененной ячейки в матрице выравнивания, кодирующий решение по выбору оценки. В случае классической оценки Смита-Ватермана и/или Нидлмана-Вунша с линейными штрафами на гэп вектор оценки может кодировать четыре возможности, которые могут необязательно храниться в виде 2-битового целого числа от 0 до 3, например: 0 = новое выравнивание (выбрана нулевая оценка); 1 = вертикальное выравнивание (выбрана оценка из ячейки сверху, модифицирована штрафом на гэп в последовательности); 2 = горизонтальное выравнивание (выбрана оценка из ячейки слева, модифицирована штрафом на гэп в последовательности); 3 = диагональное выравнивание (выбрана оценка из ячейки сверху и слева, модифицирована оценкой совпадения или несовпадения нуклеотида). Необязательно, можно сохранять вычисленную оценку или оценки для каждой оцененной ячейки матрицы (помимо максимальной достигнутой оценки выравнивания, которую обычно сохраняют), однако это, как правило, не является необходимым для обратного отслеживания и может требовать затрат больших объемов памяти. Выполнение обратного отслеживания, в таком случае, превращается в следование за векторами оценки; при достижении при обратном отслеживании определенной ячейки в матрице следующий шаг обратного отслеживания определяет сохраненный вектор оценки для указанной ячейки, например,: 0 = окончание обратного отслеживания; 1 = обратное отслеживание в направлении наверх; 2 = обратное отслеживание в направлении налево; 3 = обратное отслеживание по диагонали наверх и налево.

[00297] Такие векторы оценки могут храниться в двумерной таблице, скомпонованной в соответствии с размерами матрицы выравнивания, где заполнены только элементы, соответствующие ячейкам, оцениваемым с помощью фронта волны. Как вариант, для экономии памяти, упрощения регистрации векторов оценки по мере их генерации и упрощения подгонки матриц выравнивания разных размеров, векторы оценки могут храниться в таблице, размер каждой строки которой подходит для хранения векторов оценки из одного фронта волны ячеек оценки, например, 128 битов для хранения 64 2-битовых векторов оценки фронта волны, состоящего из 64 ячеек, а количество строк равно максимальному числу шагов фронта волны в операции выравнивания.

[00298] Дополнительно, в этом варианте можно регистрировать направления различных шагов фронта волны, например, сохраняя дополнительный, например, 129-й, бит в каждой строке таблицы, с кодированием, например, с помощью 0 вертикального шага фронта волны, предшествующий указанной позиции фронта волны, и с помощью 1 - горизонтального шага фронта волны, предшествующий указанной позиции фронта волны. Указанный дополнительный бит можно использовать при обратном отслеживании, чтобы следить за тем, каким позициям виртуальной матрицы оценки соответствуют векторы оценки в каждой строке таблицы, чтобы после каждого последовательного шага обратного отслеживания можно было извлекать правильный вектор оценки. Если шаг обратного отслеживания является вертикальным или горизонтальным, следующий вектор оценки следует извлекать из предыдущей строки таблицы, однако если шаг обратного отслеживания является диагональным, следующий вектор оценки следует извлекать из двух предыдущих строк, так как фронт волны должен был выполнить два шага для перехода от оценки любой ячейки к оценке ячейки,

расположенной по диагонали справа внизу от нее.

[00299] В случае аффинной оценки гэпов информация о векторе оценки может быть расширена, например, до 4 битов на оцениваемую ячейку. Помимо, например, 2-битового индикатора направления выбора оценки могут быть добавлены два 1-битовых флага - флаг вертикального продления и флаг горизонтального продления. В соответствии с методами расширения аффинной оценки гэпов для алгоритмов Смита-Ватермана или Нидлмана-Вунша, или аналогичных алгоритмов выравнивания, для каждой ячейки, в дополнение к первичной оценке выравнивания, представляющей выравнивание с лучшей оценкой, заканчивающееся в указанной ячейке, должна быть сгенерирована «вертикальная оценка», соответствующая максимальной оценке выравнивания, достигающей указанной ячейки на последнем шаге в вертикальном направлении, а также «горизонтальную оценку», соответствующую максимальной оценке выравнивания, достигающей указанной ячейки на последнем шаге в горизонтальном направлении; при этом при вычислении любой из указанных трех оценок шаг в вертикальном направлении в ячейку может быть вычислен с использованием большего из двух значений: либо первичной оценки из ячейки сверху за вычетом штрафа на открытие гэпа, либо вертикальной оценки из ячейки сверху за вычетом штрафа на продление гэпа; а шаг в горизонтальном направлении в ячейку может быть вычислен с использованием большего из двух значений: либо первичной оценки из ячейки слева за вычетом штрафа на открытие гэпа, либо горизонтальной оценки из ячейки слева за вычетом штрафа на продление гэпа. В тех случаях, когда выбрана вертикальная оценка за вычетом штрафа на продление гэпа, должен быть установлен флаг вертикального продления в векторе оценки, например, установлено значение «1», а в ином случае он должен быть снят, например, установлено значение «0». В тех случаях, когда выбрана горизонтальная оценка за вычетом штрафа на продление гэпа, должен быть установлен флаг горизонтального продления в векторе оценки, например, установлено значение «1», а в ином случае он должен быть снят, например, установлено значение «0». При обратном отслеживании в случае аффинной оценки гэпов каждый раз, когда при обратном отслеживании происходит шаг в вертикальном направлении вверх от заданной ячейки, если для указанной ячейки установлен флаг вертикального продления в векторе оценки, следующий шаг обратного отслеживания также должен быть вертикальным, независимо от вектора оценки для ячейки сверху. Сходным образом, каждый раз, когда при обратном отслеживании происходит шаг в горизонтальном направлении налево от заданной ячейки, если для указанной ячейки установлен флаг горизонтального продления в векторе оценки, следующий шаг обратного отслеживания также должен быть горизонтальным, независимо от вектора оценки для ячейки слева.

[00300] Соответственно, такая таблица векторов оценки, например, 129 битов на строку для 64 ячеек при использовании линейной оценки гэпов, или 257 битов на строку для 64 ячеек при использовании аффинной оценки гэпов, с некоторым числом (NR) строк, подходит для обеспечения обратного отслеживания после завершения оценки выравнивания, если фронт волны оценки выполняет NR шагов или менее. Например, при выравнивании ридов из 300 нуклеотидов число необходимых шагов фронта волны может быть всегда меньше 1024, таким образом, размер таблицы может составлять 257×1024 битов, или приблизительно 32 килобайта, что во многих случаях может представлять собой разумный объем локальной памяти в составе ИС. Однако в том случае, если требуется выравнивание очень длинных ридов, например, из 100000 нуклеотидов, требования к памяти для векторов оценки могут быть достаточно значительными, например, 8 мегабайт, что может быть очень затратным для включения

в качестве локальной памяти в состав ИС. Для такого обеспечения информация о векторах оценки может регистрироваться в память большой емкости вне ИС, например, DRAM, однако в этом случае требования к полосе пропускания, например, 257 битов на тактовый цикл на модуль выравнивания, могут быть чрезмерно большими, что может сдерживать работу и резко снижать производительность выравнивателя.

[00301] Соответственно, желательно иметь способ отработки векторов оценки до завершения выравнивания, позволяющий ограничить требования к их хранению, например, способ выполнения инкрементных обратных отслеживаний, с генерацией инкрементных частичных строк CIGAR, например, из начальных частей истории векторов оценки выравнивания, чтобы потом такие начальные части векторов оценки могли быть удалены. Трудность состоит в том, что обратное отслеживание, в принципе, должно начинаться с конца выравнивания в ячейке с максимальной оценкой, неизвестной до завершения оценки выравнивания, поэтому любое обратное отслеживание, начатое до завершения выравнивания, может начинаться с неверно выбранной ячейки, располагающейся не на потенциальном итоговом пути оптимального выравнивания.

[00302] Соответственно, предложен способ для выполнения инкрементного обратного отслеживания на основе частичной информации о выравнивании, например, содержащей частичную информацию о векторах оценки для уже оцененных ячеек матрицы выравнивания. Исходя из границы выполненного к текущему моменту выравнивания, например, конкретной оцененной позиции фронта волны, начинают обратное отслеживание из всех позиций ячеек на указанной границе. Такое обратное отслеживание из всех ячеек на границе может быть выполнено последовательно или, предпочтительным образом, в частности, в случае аппаратной реализации, все обратные отслеживания выполняют совместно. Нет необходимости в выделении нотаций выравнивания, например, строк CIGAR, указанных нескольких обратных отслеживаний; необходимо только определить, через какие позиции матрицы выравнивания они проходят при обратном отслеживании. При реализации одновременного обратного отслеживания от границы оценивания можно использовать ряд 1-битовых регистров, соответствующих числу ячеек выравнивания, все из которых установлены в исходное значение, например, «1», которые показывают, проходит ли какое-либо из обратных отслеживаний через соответствующую позицию. Для каждого шага одновременного обратного отслеживания могут быть исследованы векторы оценки, соответствующие всем текущим значениям «1» в указанных регистрах, например, из одной строки таблицы векторов оценки, для определения следующего шага обратного отслеживания, соответствующего каждому значению «1» в указанных регистрах, ведущего в следующую позицию для каждого значения «1» в указанных регистрах, для следующего шага одновременного обратного отслеживания.

[00303] Важно отметить достаточно большую вероятность слияния нескольких значений «1» в регистрах в общие позиции, соответствующие нескольким одновременным обратным отслеживаниям, сливающимся на общих путях обратного отслеживания. После того как два или более одновременных отслеживаний сливаются, они остаются слитыми неограниченное время, поскольку с этого момента используют информацию вектора оценки из одной и той же ячейки. Согласно наблюдениям, из эмпирических соображений и по теоретическим причинам, с высокой вероятностью все одновременные обратные отслеживания сливаются в сингулярный путь обратного отслеживания за относительно небольшое число шагов обратного отслеживания, например, число, с небольшой кратностью, такой как 8, превышающее число ячеек оценки во фронте волны. Например, в случае фронта волны из 64 ячеек, с высокой

вероятностью все обратные отслеживания с заданной границы фронта волны сольются в один путь обратного отслеживания не более чем за 512 шагов обратного отслеживания. Как вариант, также возможно и нередко происходит завершение всех обратных отслеживаний в пределах указанного числа, например, 512, шагов обратного отслеживания.

[00304] Соответственно, несколько одновременных обратных отслеживаний может быть выполнено с границы оценивания, например, например, с оцененной позиции фронта волны, достаточно далекой, чтобы все они либо завершились, либо слились в один путь обратного отслеживания, например, за 512 шагов обратного отслеживания или менее. Если все они сливаются в сингулярный путь обратного отслеживания, то выполнение инкрементного обратного отслеживания на основании частичной информации о выравнивании возможно с местоположения в матрице оценки, где они сливаются, или на любом расстоянии в обратном направлении вдоль сингулярного пути обратного отслеживания. Дополнительно начинают обратное отслеживание с точки слияния, или на любом расстоянии в обратном направлении, с использованием обычных способов одиночного обратного отслеживания, в том числе регистрации соответствующей нотации выравнивания, например, частичной строки CIGAR. Указанное инкрементное обратное отслеживание и, например, частичная строка CIGAR должны быть частью любого возможного окончательного обратного отслеживания, и, например, полной строки CIGAR, получаемых после завершения выравнивания, за исключением случаев, когда такое окончательное обратное отслеживание заканчивается до достижения границы оценивания, где началось одновременное обратное отслеживание, поскольку, если оно достигает границы оценивания, оно должно следовать по одному из путей одновременного обратного отслеживания и сливаться с сингулярным путем обратного отслеживания, к этому моменту инкрементно выделенным.

[00305] Соответственно, все векторы оценки для областей матрицы, соответствующих инкрементно выделенному обратному отслеживанию, например, во всех строках таблицы для позиций фронта волны, предшествующих началу выделенного сингулярного обратного отслеживания, могут быть безопасно отброшены. Если при осуществлении окончательного обратного отслеживания из ячейки с максимальной оценки оно заканчивается до достижения границы оценивания (или, как вариант, если оно заканчивается до достижения начала выделенного сингулярного обратного отслеживания), нотация инкрементного выравнивания, например, частичная строка CIGAR, может быть отброшена. Если окончательное обратное отслеживание продолжается до начала выделения сингулярного обратного отслеживания, его нотация выравнивания, например, строка CIGAR, может затем быть присоединена к нотации инкрементного выравнивания, например, частичной строке CIGAR.

[00306] Кроме того, при очень длинном выравнивании процесс выполнения одновременного обратного отслеживания с границы оценивания, например, оцененной позиции фронта волны, до тех пор, пока все обратные отслеживания не завершатся или не сольются, с последующим сингулярным обратным отслеживанием с выделением нотации выравнивания, может быть повторен несколько раз, начиная с различных последовательных границ оценивания. Нотация инкрементного выравнивания, например, частичная строка CIGAR, из каждого последовательного инкрементного обратного отслеживания может затем быть присоединена к аккумулированным предыдущим нотациям выравнивания, если новое одновременное обратное отслеживание или сингулярное обратное отслеживание не заканчивается рано; в этом случае аккумулированные предыдущие нотации выравнивания могут быть отброшены.

Потенциальное окончательное обратное отслеживание сходным образом присоединяет собственную нотацию выравнивания к самым последним аккумулированным нотациям выравнивания для полного описания обратного отслеживания, например, строки CIGAR.

[00307] Соответственно, указанным образом объем памяти для хранения векторов оценки может быть ограничен, при условии, что одновременные обратные отслеживания всегда сливаются между собой за ограниченное число шагов, например, 512 шагов. В редких случаях, когда одновременные обратные отслеживания не сливаются и не заканчиваются за ограниченное число шагов, могут быть предприняты различные исключительные действия, в том числе отказ от текущего выравнивания как от ошибочного или его повторение с бóльшим ограничением или без ограничения, вероятно, другим или традиционным способом, таким как сохранение всех векторов оценки для полного выравнивания, например, во внешней DRAM. В одном варианте может быть целесообразно отказаться от такого выравнивания как от ошибочного ввиду его чрезвычайной редкости, и еще реже такое ошибочное выравнивание представляет собой выравнивание с наилучшей оценкой для включения в отчет о выравнивании.

[00308] В необязательном варианте хранение вектора оценки может быть разделено, физически или логически, на ряд отдельных блоков, например, по 512 строк каждый, и конечная строка в каждом блоке может быть использована в качестве границы оценивания для начала одновременного обратного отслеживания. Необязательно может требоваться окончание или слияние одновременного обратного отслеживания в пределах одного блока, например, 512 шагов. Необязательно, если одновременные обратные отслеживания сливаются за меньшее число шагов, слитое обратное отслеживание может, тем не менее, продолжаться по всему блоку, до начала выделения сингулярного обратного отслеживания в предыдущем блоке. Соответственно, после того, как векторы оценки полностью записаны в блок N и начинают записываться в блок N+1, одновременное обратное отслеживание может начинаться в блоке N, с последующим сингулярным обратным отслеживанием и выделением нотации выравнивания в блоке N-1. Если все скорости одновременного обратного отслеживания, сингулярного обратного отслеживания и оценки выравнивания аналогичны или идентичны, и все они могут быть выполняться одновременно, например, в параллельном аппаратном обеспечении в ИС, то сингулярное обратное отслеживание в блоке N-1 может происходить одновременно с заполнением векторами оценки блока N+2, а когда нужно будет приступать к заполнению блока N+3, блок N-1 может быть освобожден и возвращен в работу.

[00309] Соответственно, при такой реализации могут применяться как минимум 4 блока векторов оценки, которые могут быть использованы циклично. Соответственно, общий объем для хранения векторов оценки для модуля выравнивателя может включать 4 блока по 257×512 битов каждый, например, или приблизительно 64 килобайта. В одном варианте, если текущая максимальная оценка выравнивания соответствует более раннему блоку, чем текущая позиция фронта волны, указанный блок и предыдущий блок может быть сохранен, а не возвращен в работу, чтобы окончательное обратное отслеживание могло начинаться с указанной позиции, если она остается позицией с максимальной оценкой; сохранение дополнительных 2 блоков указанным образом сдвигает минимум, например, до 6 блоков. В другом варианте для обеспечения перекрывающихся выравниваний, при фронте волны оценки, постепенно переходящем от одной матрицы выравнивания к следующей согласно описанию выше, могут

использоваться дополнительные блоки, например, 1 или 2 дополнительных блоков, например, в общей сложности 8 блоков, например, приблизительно 128 килобайтов. Соответственно, если такое ограниченное число число блоков, например, 4 блока или 8 блоков, используют циклично, возможны выравнивание и обратное отслеживание ридов произвольной длины, например, 100000 нуклеотидов, или всей хромосомы, без использования внешней памяти для векторов оценки.

[00310] Согласно описанию выше, определенные области ДНК представлены генами, которые кодируют белки или функциональную РНК. Каждый ген существует на одной нити двунитевой двойной спирали ДНК, часто в виде ряда экзонов (кодирующих сегментов), разделенных интронами (некодирующими сегментами). Некоторые гены содержат единственный экзон, но большинство содержит несколько экзонов (разделенных интронами), а некоторые содержат сотни экзонов или тысячи экзонов. Длина экзонов обычно составляет несколько сотен нуклеотидов, но они могут быть и меньшей длины, до одного нуклеотида, или большей длины, до десятков, сотен или тысяч нуклеотидов. Длина интронов обычно составляет тысячи нуклеотидов, а у некоторых превосходит миллион нуклеотидов.

[00311] Ген может быть транскрибирован РНК-полимеразными ферментами в матричную РНК (мРНК) или другие типы РНК. Непосредственно полученный РНК-транскрипт представляет собой одноцепочечную копию гена, за исключением того, что тиминовые (T) основания ДНК транскрибированы в урациловые (U) основания РНК. Однако копии интронов непосредственно после продуцирования указанной копии, как правило, подвергаются сплайсингу сплайсингсомами, оставляя копии экзонов сцепленными на «стыках сплайсинга» (которые после этого не являются прямо очевидными). Сплайсинг РНК не всегда происходит одинаково. Иногда сплайсируются один или более экзонов, а иногда стыки сплайсинга не попадают на наиболее распространенные границы интронов/экзонов. Соответственно, один ген может продуцировать несколько разных транскрибированных РНК-сегментов в результате этого процесса, иногда называемого альтернативным сплайсингом.

[00312] Сплайсированная мРНК транспортируется (у эукариот) из клеточного ядра в рибосому, которая декодирует ее в белок, при этом каждая группа из трех нуклеотидов РНК (кодон) кодирует одну аминокислоту. Указанным образом, гены в ДНК служат в качестве оригиналов инструкций для получения белков.

[00313] В основном сплайсинг РНК происходит на стабильных границах экзонов/интронов, которые характеризуются типичным составом последовательностей, в частности, возле концов интронов. В частности, первые два и последние два основания интрона, называемые мотивом интрона, имеют одну из всего 3 последовательностей «канонических» мотивов интрона в подавляющем большинстве случаев (примерно 99,9%). Наиболее распространенным каноническим мотивом интрона является «GT/AG», что означает, что первые два основания интрона представлены «G» и «T», а последние два интрона представлены «A» и «G». Мотив GT/AG встречается примерно в 98,8% случаев. Другие канонические мотивы интронов представлены GC/AG, который встречается примерно в 1,0% случаев, и AT/AC, который встречается примерно в 0,1% случаев. Указанные канонические мотивы и показатели их распространенности достаточно стабильны у разных видов, однако не обязательно универсальны.

[00314] Не все гены транскрибируются, а транскрипция транскрибируемых генов может происходить с разной скоростью. Многие факторы могут влиять на то, будет ли заданный ген транскрибироваться в РНК, и насколько часто. Некоторые из указанных факторов наследуются, некоторые варьируют в зависимости от специализации

клеток в разных тканях, а некоторые варьируют с течение времени под воздействием условий окружающей среды или заболеваний. Соответственно, две клетки с совершенно одинаковой ДНК могут продуцировать достаточно разные типы и количества белков и функциональной РНК. Из-за этого секвенирование (чтение) РНК, присутствующей в одной или более клеток, дает информацию, отличающуюся от информации при секвенировании ДНК. Более полную картину состояния и активности клеток обеспечивает комбинация секвенирования ДНК и секвенирования РНК.

[00315] Секвенирование полного транскриптома РНК обычно проводят, сначала выбирая целевую РНК, например, кодирующую белок РНК, а затем используя ферменты с активностью обратной транскриптазы для обратного преобразования РНК-сегментов в цепи комплементарной ДНК (кДНК). Указанная ДНК может быть амплифицирована с применением полимеразной цепной реакции (ПЦР) и/или фрагментированы с получением требуемого распределения длин последовательностей. Затем фрагменты ДНК секвенируют на секвенаторе ДНК, таком как секвенатор нового поколения, использующий «метод дробовика».

[00316] Полученные риды ДНК представляют собой либо обратно-комплементарные, либо прямые копии исходных цепей РНК, за исключением того, что «U», аналогично вышеописанному, заменяют на «T». При получении некоторых библиотек и выполнении некоторых протоколов секвенирования ориентация секвенированных цепей ДНК относительно оригинальной РНК может быть сохранена или помечена; но в обычных протоколах приблизительно 50% секвенированной ДНК обратно-комплементарна оригинальной РНК, без прямого указания на ориентацию (хотя имеются косвенные указания).

[00317] Риды ДНК из протоколов РНК-секвенирования (RNA-seq) отличаются от получаемых при полногеномном или полноэкзомном секвенировании ДНК другими способами. Во-первых, кроме загрязняющих примесей, секвенируется только транскрибированная РНК, соответственно, некодирующая ДНК и неактивные гены, как правило, не появляются. Во-вторых, количество секвенированных ридов, соответствующих различным генам, связано с показателями биологической транскрипции указанных генов. В-третьих, в результате сплайсинга интронов, в ридах РНК-секвенирования наблюдается тенденция к пропуску интронных (некодирующих) сегментов в составе генов.

[00318] Риды РНК-секвенирования обычно обрабатывают способом, достаточно выраженно отличающимся от обработки ридов ДНК. Хотя оба типа ридов, как правило, картируют на референсный геном и выравнивают с ним, техники картирования и выравнивания ДНК и РНК различаются (см. следующий раздел). После картирования и выравнивания риды обычно сортируют по позициям картирования на референсе, в случае как ДНК, так и РНК. Двойная маркировка, которая может быть необязательно использована при обработке ДНК, обычно не используется для данных РНК-секвенирования.

[00319] После этого риды ДНК обычно обрабатывают с помощью определителя вариантов для идентификации различия между ДНК образца и референсным геномом. Риды РНК-секвенирования обычно не используют для определения вариантов, хотя иногда это делается. Чаще выровненные и сортированные риды РНК анализируют, чтобы определить, какие гены были экспрессированы и в каких относительных количествах, или какие из различных транскриптов в результате альтернативного сплайсинга были продуцированы и в каких относительных количествах. Указанный анализ обычно включает подсчет количества ридов, которые выравниваются с

различными генами, экзонами и т.п., и может также включать сборку транскриптов (на основе референсов или de novo) чтобы на основании относительно коротких ридов РНК-секвенирования сделать вывод о вероятном сплайсинге более длинных транскриптов РНК из ДНК.

[00320] Анализ экспрессии генов, экзонов или транскриптов часто расширяют до анализа дифференциальной экспрессии, при котором данные РНК секвенирования нескольких образцов, часто из двух или более разных классов (субпопуляций или фенотипов), сравнивают, чтобы определить степень различий экспрессии указанных генов, экзонов или транскриптов в разных классах. Указанный анализ может включать вычисление правдоподобия «нулевой гипотезы», заключающейся в том, что соответствующие уровни экспрессии одинаковы в разных классах, а также расчет «кратности изменений» экспрессии в разных образцах, например, 8-кратного или 10-кратного изменения, или изменений с большей кратностью.

[00321] Во многих вариантах применения секвенирования ДНК или РНК на ранней стадии обработки проводят картирование ридов на референсный геном и выравнивание с ним. Обычно используют ДНК-ориентированный референсный геном при секвенировании как ДНК, так и РНК, содержащий «T», а не «U», в частности, с учетом того, что РНК-секвенирование обычно включает обратную транскрипцию в кДНК перед секвенированием. В случае РНК-секвенирования, как и при полноэкзомном секвенировании в этом отношении, референсный геном теоретически может быть ограничен известными кодирующими областями или областями вблизи кодирующей ДНК. Тем не менее, обычной практикой является картирование и выравнивание целого референсного генома для исследуемого организма.

[00322] Самое большое отличие, необходимое для подходящего для РНК картирующего устройства/ выравнивателя заключается в способности обрабатывать стыки сплайсинга. Поскольку риды РНК-секвенирования соответствуют сегментам транскрибированной и сплайсированной РНК, обычно рид пересекает один или более стыков сплайсинга. Применительно к ДНК-ориентированному референсному геному это означает, что первая часть рида происходит из первого экзона и должна быть картирована на него, а вторая часть рида должна быть картирована на второй экзон, и т.д. Например, в риде длиной 100 оснований первые 40 оснований могут быть взяты из экзона на хромосоме 3 со смещением 2 345 000, а остальные 60 оснований могут быть взяты из другого экзона, расположенного на расстоянии 100000 оснований, начинающегося на хромосоме со смещением 2 445 040. Выравнивание для такого рида может быть представлено позицией картирования Chr3:2345000 и строкой выравнивания CIGAR «40M100000N60M», в которой «40M» и «60M» представляют части, выровненные относительно соответствующих экзонов, а «100000N» представляет интрон длиной 100000 оснований, и указанные 100000 оснований референса пропускаются при выравнивании рида. (Абстрактно говоря, указанную строку CIGAR можно рассматривать как эквивалентную «40M100000D60M», где «100000D» представляет делецию размером 100000 оснований из референса, однако предполагаемые сплайсированные интроны принято обозначать как «N» в отличие от делеций предполагаемых мутаций или ошибок секвенирования, обозначаемых как «D».)

[00323] Практическое различие между событиями CIGAR «N» (интрон) и «D» (делеция) связано с их типичной длиной. Делеции редко бывают длиннее 50 оснований и, таким образом, успешно обнаруживаются и точно позиционируются при использовании алгоритма Смита-Ватермана или аналогичных алгоритмов для выравнивания последовательностей. Длина интронов часто составляет несколько тысяч оснований

или даже миллион оснований или более, и использование выравнивателей типа Смита-Ватермана для детекции таких длинных гэпов выравнивания непрактично. Следовательно, первоначальное исследование стыков сплайсинга в большей мере относится к «картированию», чем к «выравниванию».

[00324] Проблема картирования состоит в том, что каждый рид может быть разделен на сегменты экзона на неизвестных границах, и различные сегменты экзона, вероятно, будут картированы на далеко отстоящие друг от друга геномные местоположения, которые потребуется определять индивидуально. Методики картирования сегментов экзонов на соответствующие им местоположения в референсе могут быть аналогичны методикам картирования целого рида на один референсный сегмент, однако картирование со сплайсингом (первое из перечисленных) является более сложным, поскольку каждый экзон может быть значимо короче целого рида, и, соответственно, содержит значительно меньше информации для направления картирующего устройства. Так, длина одиночного экзона может составлять всего одно (1) основание, такое как «G», и без дополнительной информации будет нецелесообразным пытаться определить, куда должно быть картировано указанное одиночное основание при потенциальном диапазоне длин интронов, включающем миллионы оснований.

[00325] Помимо исследования картирования двух последовательных сегментов экзонов в риде, должен быть точно позиционирован стык сплайсинга между ними, по меньшей мере, для некоторых вариантов применения. Хотя может быть ясно, что первые примерно 40 оснований и последние примерно 60 оснований рида длиной 100 оснований картируются на местоположения, расположенные точно на расстоянии в 100000 оснований в хромосоме 3, часто значительно менее ясно, какое именно количество оснований рида картируется на каждое из указанных двух местоположений, или где именно стык сплайсинга, граница между двумя сегментами экзонов, попадает в рид. Корректная строка CIGAR может выглядеть, например, не только как «40M100000N60M», но и как «39M100000N61M» или «42M100000N58M». Точное позиционирование стыков сплайсинга скорее представляет собой операцию «выравнивания», а не «картирования».

[00326] Подходящее для РНК картирующее устройство (картировщик) также может успешно делать выводы, с какой из двух цепей ДНК была транскрибирована последовательность рида. В типичных ненаправленных протоколах РНК-секвенирования заданный рид может выравниваться либо в прямом, либо в обратно-комплементарном относительно референса виде (со стыками сплайсинга или без них). В протоколах РНК-секвенирования спаренных концов, как правило, два сопряженных рида ориентированы «FR» (прямо/обратно), так что ранее картированный сопряженный рид в референсном геноме ориентирован прямо, а другой сопряженный рид обратно-комплементарен. Однако в типичных протоколах ненаправленного РНК-секвенирования указанные ориентации картирования не определяют, какая цепь ДНК несла ген, с которого была транскрибирована РНК для указанного рида, отчасти из-за того, что при амплификации кДНК с помощью ПЦР получают обе ориентации.

[00327] Наконец, РНК-ориентированный картировщик может благоприятным образом улучшать вводную базу данных «аннотированных» известных границ сплайсинга. Все общие гены человека были подробно изучены, например, и границы сплайсинга более общих и менее общих РНК-транскриптов аннотированы в геномных базах данных. Данная информация не является на 100% всеобъемлющей; любые отдельные образцы наиболее вероятно будут иметь некоторый «новый» сплайсинг, не записанный в базах данных. Однако аннотированные границы сплайсинга могут служить

полезными указателями для повышения точности картирования и выравнивания секвенированной РНК. После картирования/выравнивания РНК-секвенированных ридов с аннотированными границами сплайсинга, или без них, усовершенствованный способ заключается в обнаружении набора границ сплайсинга, наблюдаемых в выровненных ридах, с некоторыми критериями, такими как минимальное число выравниваний, покрывающее границу сплайсинга, а также использовании данного набора эмпирически обнаруженных границ сплайсинга в качестве аннотированных границ сплайсинга для второго прогона картирования/выравнивания РНК. Это может повысить чувствительность посредством использования границ сплайсинга, обнаруженных в некоторых ридах для направления картирования других ридов.

[00328] Картирование начальной затравки для РНК-секвенированных ридов происходит аналогично ридам ДНК. Первичная длина затравки выбирается, в идеале, немного длиннее, чем логарифм по основанию 4 размера референсного генома для того, чтобы обеспечить картирование затравок приблизительно однозначно, например K=18 или K=21 для всего референсного генома человека. Хэш-таблица построена, заполнена некоторыми или всеми затравками из референсного генома, хэш-запись в хэш-таблицы для каждой заполненной затравки указывает ее положение и ориентацию в референсе. Хэш-таблица, загруженная в память, доступна аппаратному обеспечению модуля картировщика, например, модулям DRAM на плате FPGA, соединенным разводкой с контактами ППВМ, реализуя аппаратное обеспечение модуля картировщика.

[00329] Модуль картировщика принимает риды РНК, поступающие от секвенатора РНК или ДНК (зачастую обратно транскрибированные в кДНК перед секвенированием). Из каждого рида картировщик извлекает затравки длиной K, в идеале, плавающее окно множества перекрывающихся затравок с K оснований, выбранное по некоторой модели, например, начиная в каждом положении основания или начиная в каждом четном положении. Картировщик осуществляет доступ к хэш-таблице в памяти для каждой затравки, получая по меньшей мере ноль, одно или более положений в референсном геноме, с которыми затравка совпадает. Что касается картирования ДНК, затравки могут быть динамически расширены при осуществлении повторяющегося доступа к хэш-таблице со все более длинными затравками, когда необходимо уменьшить большое количество совпадающих положений до приемлемо малого количества, например до 16 или менее совпадений. Совпадения затравок объединяются в цепочки затравок, содержащие затравки, совпадающие с одинаковой ориентацией (прямой или с обратным соответствием по отношению к референсу) вдоль одинаковых диагоналей выравнивания.

[00330] Для РНК-секвенированных ридов предпочтительным может быть дополнительный этап, выполняемый движком картировщика, для детализации картирования начальной затравки посредством картирования привязанной короткой затравки. Например, как показано на ФИГ. 1, риды РНК зачастую пересекают одну или более границ сплайсинга и затравку, пересекающую границу сплайсинга, обычно не удается картировать, так как ее реальное изображение в референсе разделено между двумя местоположениями. Когда рид содержит более короткий экзон, чем начальная длина K затравки, или рид перекрывает более длинный экзон меньшим количеством оснований, чем K, то при картировании затравки может не удаться определить положение в соответствующем референсе для указанного экзона. Даже когда полный или частичный экзон немного длиннее, чем K оснований, но короче, чем полный рид, он может быть уязвим к появлению ошибки картирования затравки, когда он содержит по меньшей мере одно редактирование (отличи) по отношению к референсу, например,

одиночный нуклеотидный полиморфизм (ОНП, SNP) или инсерцию или делецию (индел) вследствие мутации в образце по отношению к референсу или вследствие ошибки секвенирования. Например, на ФИГ. 1 показан пример ошибки при картировании всех сегментов экзона с длинными (с K оснований) затравками. По этой причине для хорошей чувствительности картирования затравок предпочтительно обращаться к более коротким затравкам, которые могут поместиться в короткие экзоны или выходящие за пределы коротких ридов части экзонов, или между редактированиями.

[00331] Может быть в некоторой степени нецелесообразно обращаться к хэш-таблице всего генома за затравками, которые сильно короче, чем минимальная длина, относящаяся к логарифмам по основанию 4 размера референсного генома, так как более короткие затравки будут иметь тенденцию к совпадению с большим количеством местоположений. Например, при референсе по всему геному человека размером приблизительно 3,1 миллиард оснований логарифм по основанию 4 равен приблизительно 15,8, а минимальная приемлемая длина затравки, к которой обращаются, может составлять K=16 или 18, возможно когда K=21 является желаемой настройкой; не приемлемо обращаться к затравкам с K=11 оснований, так как каждая модель с 11 основаниями будет совпадать в среднем более чем с 700 местоположениями референса.

[00332] Однако, после картирования начальных затравок, например, при затравках с K=21 основаниями, можно детализировать картирование затравок посредством привязанных затравок меньшей длины, например с L=11 основаниями. Для картирования привязанных затравок хэш-таблицу с привязанными затравками (которая может быть той же, что и первичная хэш-таблица, или отдельной хэш-таблицей) заполняют затравками с L оснований из референса, которые связаны с конкретными областями референса, например ячейки некоторого размера, например, 2^{16}=65 536 оснований. Каждая области референса или ячейке присвоен уникальный идентификатор, например, ее начальное положение в референсном геноме, разделенное на размер ячейки. Затравки с L оснований в каждой ячейке референса занесены в хэш-таблицу привязанных затравок с использованием хэш-ключа, образованного из L оснований затравки и идентификатора ячейки.

[00333] Модуль картировщика может обращаться к хэш-таблице привязанных затравок для каждой данной затравки с L оснований в каждой данной ячейке с использованием хэш-ключа обращения, образованного таким же образом из L оснований затравки и идентификатора ячейки. Посредством такого обращения будут обнаружены только совпадения затравки с L оснований с конкретной ячейкой референса. Так как ячейка значительно меньше, чем весь референсный геном, короткая затравка с L оснований имеет достаточно информации, чтобы зачастую иметь лишь одно совпадение. Например, логарифм по основанию 4 ячейки размером 65 536 равен 8, так что L=11 (или 10, 12 и т.д.) является приемлемой длиной привязанной затравки для заполнения и обращения. Как показано на ФИГ. 2, короткие (с L оснований) затравки более легко поместить в короткий экзон, выступающие части короткого экзона или сегменты экзона, вырезанные редактированием, такие как ОНП.

[00334] Ключ к работе картирования привязанных затравок состоит в том, что обращения модуля картировщика к хэш-таблице привязанных затравок руководствуются картированием начальных затравок. Начальные совпадения, например, при затравках с K=21 основаниями, могут не обеспечить успешное картирование всех сегментов экзона рида, но наиболее вероятно они картируют по меньшей мере один сегмент экзона каждого рида или его парного сопряженного на конце рида. Имея по меньшей мере

одно совпадение с K оснований в пределах по меньшей мере одного сегмента экзона в риде РНК или его сопряжении, любые другие сегменты экзона в риде, которые не были успешно картированы посредством затравок с K основаниями, наиболее вероятно совпадут относительно близко в референсном геноме.

[00335] Например, приблизительно 99% интронов человека короче, чем 65 536 оснований, так что если один сегмент экзона картирован при затравке с K оснований в данном положении референса, то другие некартированные сегменты экзона наиболее вероятно совпадут в пределах той же ячейке референса с 65 536 оснований, или в пределах соседней ячейки. Как показано на ФИГ. 3, может быть определен диапазон поиска, например, размер ячейки, ½ или ¼ размера ячейки, или два размера ячейки, и можно обратиться к одной или более ячейкам референса в пределах диапазона поиска успешно картированных затравок с K основаниями в хэш-таблице привязанных затравок с использованием затравок с L основаниями. Таким образом, совпадения затравок с K основаниями служат как точки привязки для локальных поисков с более короткими затравками с L оснований. Велика вероятность найти дополнительные совпадения с ранее некартированными сегментами экзона рида. Таким образом чувствительность картирования затравок повышена для ридов РНК.

[00336] В дополнение, существуют различные способы, как модуль картировщика может использовать картирование привязанных коротких затравок. В одном варианте реализации после того, как картировщик обращается к начальной затравке с K оснований в хэш-таблице и объединяет совпадения в цепочки затравок, картировщик затем извлекает затравки с L основаниями из рида и обращается с ними к расположенным рядом ячейкам референса (в пределах выбранного диапазона поиска для текущих цепочек затравок) для поиска дополнительных совпадений для более коротких затравок с L оснований, которые модуль картировщика затем объединяет в дополнительные цепочки затравок или добавляет к существующим цепочкам затравок с такими же диагоналями выравнивания. В таком варианте реализации предпочтительно, чтобы хэш-таблица привязанных затравок была той же, что и первичная хэш-таблица, или, для отдельных первичной хэш-таблицы и хэш-таблицы привязанных затравок, чтобы они находились в памяти, к которой осуществляется доступ, одновременно. В обоих случаях, для того, чтобы хэш-таблица (хэш-таблицы) с затравками с K основаниями и с L основаниями помещались в память, может быть использовано примерно в два раза больше памяти, например, 64 Гб DRAM вместо 32 Гб DRAM, или, в качестве альтернативы, могут быть заполнены примерно половина затравок каждой длины, например, плотность заполнения затравок составляет 50% вместо 100%. Для ограничения количества требуемых обращений к хэш-таблице привязанных затравок в качестве точек привязки могут использоваться только наиболее перспективные цепочки начальных затравок и/или затравки с L основаниями могут быть извлечены из рида только из некоторых областей, например, областей, где затравки с K основаниями не были картированы успешно.

[00337] В другом варианте реализации картирование и/или выравнивание набора ридов может быть подведено к завершению при первом прогоне с использованием лишь затравок с K основаниями. Результаты картирования/выравнивания каждого рида могут быть затем изучены, например, посредством программного обеспечения вне модуля картировщика, для определения, каким ридам требуется детальное картирование с использованием коротких привязанных затравок. Одним индикатором, запускающим детализацию посредством привязанных затравок, является то, что выравнивания первого прогона являются обрезанными, в особенности с обрезанием

около или более, чем длина L короткой затравки. Другим индикатором, запускающим детализацию с помощью привязанных затравок, может являться существенное количество несовпадений, наблюдаемых в выравниваниях первого прогона. Другим индикатором, запускающим детализацию с помощью привязанных затравок, может являться то, что оба парных сопряженных на концах ридов не картировались успешно, картировались далеко друг от друга или с неожиданной относительной ориентацией. Преимущественно, если один рид выбран для детализации с применением короткой затравки, то его парный сопряженный на конце рид также будет выбран. Преимущественно, только часть выравниваний первого прогона может потребовать детализации с применением коротких затравок, например 15% или 30%.

[00338] Для каждого рида в подмножестве, к которому применяют детализацию с применением коротких затравок, может быть выбрана одна или более ячеек референса для поиска, например, ячейки, перекрывающие радиус поиска вокруг результатов выравнивания первого прогона для рида и/или его сопряженного рида (см. ФИГ. 3). Затем в отношении подмножества ридов, выбранных для детализации, может быть выполнен второй прогон картирования/выравнивания. Во втором прогоне могут обращаться с затравками с L основаниями из каждого рида к хэш-таблице привязанных затравок, связанной с одной или более выбранных ячеек референса для каждого рида. Обычно по меньшей мере для некоторых ридов, выбранных для второго прогона, в результате получают улучшенное картирование/выравнивание, например с более высокими оценками выравнивания; результаты второго прогона могут быть оставлены в тех случаях, когда они улучшены, или, в других случаях, могут быть оставлены результаты первого прогона. При необходимости, первичная хэш-таблица могут быть загружена в доступную движку память перед первым прогоном картирования и хэш-таблица привязанных затравок может быть загружена перед вторым прогоном картирования, устраняя необходимость размещения обеих хэш-таблиц (или одной объединенной хэш-таблицы) в памяти одновременно, хотя и обе из них могут быть загружены одновременно, или снижения плотности заполнения референсных затравок для размещения обеих одновременно.

[00339] В некоторых вариантах реализации ячейки референса имеют конфигурируемый размер, радиус поиска является конфигурируемым и длина начальной затравки (K) и длина привязанной затравки (L) являются конфигурируемыми. В других вариантах реализации размер ячейки референса является степенью двойки. Приведенные в качестве примера предпочтительные настройки для обработки полного транскриптома секвенирования РНК человека представляют собой K=21, L=11, размер ячейки референса 2^{16}=65536, а радиус поиска 2^{14}=16384.

[00340] Если аннотированные границы сплайсинга переданы в модуль картировщика, они могут быть использованы для повышения чувствительности картирования. Список аннотированных границ загружают в память, доступ к которой осуществляется движком картировщика. Преимущественно, аннотированные границы могут быть форматированы в таблицу, легко доступную для модуля картировщика, например, таблицу с записью для каждой, например, с 1024 основаниями, ячейки референса, которая или содержит информацию об интроне по меньшей мере с одной конечной точкой в указанной ячейке, или указывает на список (в пространстве, размещенном после начальной таблицы) множества дескрипторов интронов. Каждый дескриптор интрона указывает на референсное положение обеих конечных точек соответствующего интрона и также может переносить дополнительную информацию, например, на какой нити ДНК находится ген интрона, мотив интрона и показатель того, как часто появляется граница

сплайсинга.

[00341] После картирования затравок (начальных затравок и/или коротких привязанных затравок) и образования цепочек затравок, осуществляют доступ к таблице аннотированных границ в рядах, соответствующих областям референса, охваченным каждой цепочкой затравок, или областям вблизи концов длинных цепочек затравок. Получают список интронов по меньшей мере с одной расположенной вблизи конечной точкой и сравнивают по меньшей мере с одной цепочкой затравок, для которой был осуществлен доступ. Каждый интрон отбрасывают, если он не является возможной или вероятной границей сплайсинга из цепочки затравок. В частности, посредством сравнения местоположения конечной точки интрона в референсе с конечной точкой цепочки затравок в референсе и в риде вычисляют действительное местоположение границы сплайсинга в риде. Если указанное действительное местоположение находится за пределами границ рида или в значительной степени перекрывает цепочку затравок (например, более, чем maxSpliceOlap = 16 в пределах конечной точки цепочки затравок) или находится слишком далеко вне пределов цепочки затравок в риде (например, более, чем на maxSpliceGap = 150 оснований вне цепочки затравок), то аннотированную границу отбрасывают как наименее вероятно являющуюся релевантной.

[00342] Каждый оставшийся дескриптор интрона рассматривают как возможную границу сплайсинга с одного конца соответствующей цепочки затравок. Эту информацию используют двумя способами. Во-первых, противоположный конец интрона в референсе берут как наиболее вероятное местоположение, на которое смежная с часть рида должна картироваться, даже если это местоположение было обнаружено посредством картирования затравок. В действительности, наиболее вероятная диагональ выравнивания на противоположном конце интрона вычисляется в точности посредством сложения длины интрона с или вычитания длины интрона из (в зависимости от ориентации) диагонали выравнивания на соответствующем конце текущей цепочки затравок. Если указанное референсное местоположение диагональ выравнивания не согласуются с существующей цепочкой затравок, то формируют новую (псевдо) цепочку затравок, начиная с референсного местоположения на противоположном конце интрона, и начиная в риде в соответствующем положении, предполагаемом посредством вычисленной диагонали выравнивания. Таким образом, вероятные местоположения картирования сегментов экзона рида обнаруживают без картирования затравок в их пределах, посредством выведения их местоположений в интронах на основании существующих цепочек затравок.

[00343] Во-вторых, аннотированную информацию об интроне используют для установления известной связи между двумя цепочками затравок, которые представляют смежные сегменты экзона в риде. Информацию о связи записывают в один или оба дескрипторов цепочки затравок, идентифицирующих другую цепочку, с которой она связана, посредством аннотированной границы сплайсинга. Кроме того, точное положение границы сплайсинга известно (принимая, что аннотированная граница сплайсинга является верной), вычислено при помощи вычитания аннотированных конечных точек интрона и диагоналей выравнивания границы сплайсинга. Указанное точное позиционирование границы сплайсинга также записывают в один или оба дескрипторов цепочки затравок.

[00344] Если обнаружено, что множество аннотированных границ сплайсинга связанны с одной цепочкой затравок, информацию о связи и положении границы сплайсинга могут записать различными способами. Например, каждую связь между двумя цепочками может быть необходимо записать только в одну из двух цепочек так,

чтобы не было конфликта, если, например, она всегда записана на «целевом» конце связи. Один дескриптор цепочки затравок может иметь место для хранения множества связей или иметь динамическое пространство для информации о связи. Кроме того, копии существующих дескрипторов цепочки затравок могут быть выполнены для содержания дополнительной информации о связи.

[00345] Просмотр аннотированной границы сплайсинга преимущественно может повторяться. Начиная с покрытия одной цепочки затравок, например, первой 1/3 рида, может быть обнаружена аннотированная граница сплайсинга, связанная с не обнаруженным ранее референсным местоположением, из чего формируют новую цепочку затравок. Доступ к таблице аннотированных границ могут повторно осуществлять для новой сформированной цепочки затравок, возможно обнаруживая, что после второй 1/3 рида имеется другая известная граница с другим необнаруженным референсным местоположением. Преимущественно, записи в таблице аннотированных соединений могут указывать на расстояние (проходящее в том же направлении, что и аннотированная граница) до достижения ближайшей другой аннотированной границы в пределах транскриптор одного и того же гена, или в целом. Когда указанное расстояние, измеренное после вычисленного местоположения границы сплайсинга в риде, проходит за пределы конца рида, не требуется снова осуществлять доступ к таблице аннотированных границ, так как ничего не будет найдено.

[00346] В пределах модуля картировщика совпадения затравок с одинаковой ориентацией (прямой или с обратным соответствием по отношению к референсу) и одинаковыми диагоналями выравнивания объединены в цепочки затравок с намерением, что одиночная операция выравнивания без промежутка или с промежутком может затем изучить и оценить выравнивание между ридом и референсом для каждой цепочки затравок. Диагональ выравнивания может быть представлена как путь диагональ-ориентированного выравнивания, покрытый совпадающей затравкой, в прямоугольнике выравнивания, образованном в пределах последовательности рида на одной оси и последовательности рида на другой оси; одно представление как целое может быть вычислено для прямых выравниваний путем вычитания положения затравки в риде из ее положения в референсе, и для выравниваний с обратным соответствием посредством сложения положения затравки в риде с ее положением в референсе.

[00347] Когда рид точно совпадает с сегментом референса, например положения от 0 до 100 в риде совпадают с положениями от 1 200 000 до 1 200 100 в референсе, все затравки обычно совпадают по одной диагонали, например 1 200 000 - 0 = 1 200 100 - 100 = 1 200 000; конкретная затравка с 21 основанием от оснований с 30 по 50 в риде совпадет с основаниями от 1 200 030 до 1 200 050 в референсе, также на той же диагонали 1 200 030 - 30 = 1 200 000. Совпадения затравок с одной и той же ориентацией и диагональю обычно включены в одну цепочку затравок, но также затравки на немного отличающихся диагоналях выравнивания могут быть включены в ту же цепочку затравок, например затравки, диагонали которых отличаются не более, чем на 20 или не более, чем на 50, или по более сложным правилам. Обеспечивая такие допуски для разницы диагоналей полезно, так как риды иногда содержат инделы (инсерции или делеции) относительно референса и выравнивание с промежутком, например выравнивание Смита-Ватермана в модуле выравнивателя может решить и оценить такие инделы для одной цепочки затравок, пока инделы не слишком велики, например не более 50 оснований инсерций или делеций.

[00348] Но РНК-секвенированные риды часто пересекают границы сплайсинга, в которых шаг от одного основания рида до следующего основания рида переходит

через весь интрон в референсе, который может быть тысячи оснований в длину, или даже более миллиона оснований в длину. В таких случаях затравки с одной стороны границы сплайсинга в риде будут картироваться на референс со значительно отличающимися диагоналями выравнивания от диагоналей на другой стороне границы сплайсинга; целочисленная разность диагоналей равна длине пропущенного интрона, вероятно тысячи или более, чем миллион. Такие затравки могут не быть приняты в ту же цепочку затравок, так как выравниватель с промежутком не может напрямую разрешить такой длинный промежуток в референсе.

[00349] Таким образом, для картирования РНК, в отличии от ДНК, следует ожидать, что верное выравнивание данного рида может содержать множество цепочек затравок, причем каждая цепочка затравок соответствует различным сегментам экзона в риде. Каждое вероятное выравнивание, таким образом, может содержать последовательность нескольких цепочек затравок. Следующий этап в модуле картировщика определяет такие вероятные последовательности цепочек затравок, известных в настоящем документе как скаффолды.

[00350] Каждый скаффолд, как последовательность одной или более цепочек затравок, имеет физическую интерпретацию как кусочное выравнивание следующих друг за другом сегментов экзона рида с соответствующими сегментами экзона в референсном генома. Таким образом, каждая цепочка затравок в одном скаффолде обычно должна покрывать только часть рида, указанные части, проходящие от начала рида в направлении конца рида вдоль последовательности цепочек затравок; и соответствующие референсные сегменты цепочек затравок должны проходить в фиксированном направлении через референс с попадающимися промежутками, соответствующими ожидаемым длинам интронов. Каждый скаффолд будет прогнан через модуль выравнивателя для разрешения точных выравниваний и их оценки, а также выравнивания наиболее вероятного варианта. Но получение списка скаффолдов из необработанного списка цепочек затравок является затруднительным.

[00351] На практике картирование затравок для одного РНК-картированного рида может быть получено из малого числа цепочек затравок до десятков или более чем сотен цепочек затравок. Имея более сотни цепочек затравок, число потенциальных последовательностей цепочек затравок является чрезвычайно большим. Таким образом, проблемой является и получение приемлемо короткого списка скаффолдов для обработки в модуле выравнивателя, и определение указанного списка скаффолдов из данного списка цепочек затравок за приемлемое количество времени так, чтобы не замедлять модуль картировщика. Для того, чтобы выполнять это эффективным образом, представлен рекурсивный способ.

[00352] Во-первых, очень полезно сортировать цепочки затравок в порядке покрываемых ими положений в риде, например в возрастающем порядке начальных положений цепочек затравок в риде. Цепочки затравок могут быть свободно построены в таком порядке посредством обращения с затравками к хэш-таблице с начала рида до конца и формируя из них цепочки затравок в том же порядке. Но если это не имеет место или если порядок нарушен последующими изменениями списка цепочек затравок при картировании с применением привязанных затравок или поиском аннотированных границ сплайсинга, то цепочки затравок должны быть отсортированы перед формированием скаффолдов, например, посредством «быстрой сортировки» или другого алгоритма сортировки.

[00353] Далее устанавливают правила, по которым одна цепочка затравок (B) может следовать непосредственно за другой цепочкой затравок (A) в одном и том же

скаффолде, устанавливая связь цепочек затравок от A к B. В правилах имеется значительная гибкость, которая может хорошо работать, на правила должны допускать вероятные связи цепочек затравок в скаффолде с верным выравниванием, в то же время исключая настолько много маловероятных связей цепочек затравок, насколько это возможно. Существует хорошо работающий набор правил с различными названными параметрами и хорошими значениями по умолчанию.

[00354] Критерии для цепочки затравок B для следования за цепочкой затравок A в скаффолде:

A и B имеют одинаковую ориентацию

(Промежуток между A и B в риде) =: gap ≤ maxSpliceGap = 150

(Перекрывание между A и B в риде) =: olap ≤ maxSpliceOlap = 16

(Промежуток между началом A и началом B в риде) =: nhead ≥ olap + (olapAdj = 4)

(Промежуток между концом A и концом B в риде) =: tail ≥ olap + (olapAdj = 4)

(Референсный промежуток A/B минус промежуток рида A/B) =: intronLen ≥ minIntronLen = 20

(Референсный промежуток A/B минус промежуток рида A/B) =: intronLen ≤ maxIntronLen = 1000000

[00355] Когда используют аннотированные границы сплайсинга и аннотированная связь была записана между цепочками затравок A и B, то они всегда могут следовать друг за другом.

[00356] Далее приведен рекурсивный алгоритм для формирования множества скаффолдов:

Сортируют N цепочек затравок по начальному положению в риде, при необходимости

Зацикливают c0 = 0 до N-1

Пропускают c0, если уже использовано внутри скаффолда

Инициализируют last = 0, scaf[0] = c0, start = c0 + 1, stack[0] = 1, stack[1] = 0, stackPos = 0

Зацикливают пока last ≥ 0

Зацикливают c = start до N - 1

Если цепочка c может следовать за цепочкой scaf[last]:

scaf[++last] = c

stack[last] = 0 если stack[last] = c

В противном случае если last > stackPos и цепочка c может следовать за цепочкой scaf [last-1]:

stack[last] = c

Выходной скаффолд scaf[0 .. last]

Устанавливают stackPos = maximum в (0 .. last) с stack[stackPos] > 0

Устанавливают start = scaf[stackPos] + 1

Устанавливают last = stackPos - 1

[00357] Значения терминов и переменных в алгоритме:

“chain”: индекс 0..N-1 цепочки затравок

scaf[] = формируемый скаффолд, каждый слот получает цепочку 0..N-1

c0 = первая цепочка в скаффолде (слот 0)

last = конечный слот # (на данный момент) в скаффолде

start = первая цепочка в поисковом цикле

stack[i] = альтернативная цепочка с наивысшим номером для scaf[i], или 0, если отсутствует. Это представляет конечную точку поиска альтернатив для scaf[i] после рекурсивного резервирования.

stackPos = целевой слот скаффолда, предназначенный для замены посредством рекурсивного резервирования. Наблюдают, что когда происходит резервирование, stack[stackPos] является интегрированным и не будет обновлен до его очистки.

[00358] Рекурсивный поиск осуществляют в физической логической части модуля картировщика. Может быть доступно время для исполнения указанного алгоритма без значительного замедления модуля с использованием способов распараллеливания за счет аппаратных средств. В частности, группу цепочек затравок для данного рида могут буферизировать для относящейся к скаффолду логической части для обработки позже в цепочке конвейера обработки, параллельно с логической обработкой следующего рида, относящейся к картированию затравок и формированию цепочек.

[00359] Иногда на практике рекурсия может развиваться, так что необходимо ее ограничивать. Полезный способ ограничения рекурсии, а также ограничения числа получаемых скаффолдов, является фильтрование меньших скаффолдов при их получении. Представлен полезный способ фильтрации скаффолдов. Для каждого скаффолда вычисляют его охват рида как количество оснований рида, покрытое одной или более цепочками затравок в скаффолде. Большее покрытие скаффолдов наиболее вероятно представляет верное выравнивание. В частности, если обнаружено максимальное покрытие рида среди всех образованных скаффолдов (на данный момент), то скаффолды с большой дельтой покрытия за пределами максимального покрытия менее вероятно представляют верное выравнивание.

[00360] Кроме того, для каждого скаффолда вычисляют его охват референсного генома, расстояние между наиболее внешними основаниями в референсе первой и последней цепочки затравок в скаффолде. Скаффолды с очень большим охватом референса менее вероятно представляют верное выравнивание. Комбинируя два указанных результата измерения следующим образом является мощным фильтром скаффолдов: filter_metric = (max_coverage - coverage) + floor(25 * ($\log_2$(ref_span + 2^{13}) - 13). Постоянные 25 и 13 должны быть конфигурируемыми параметрами: rna-filt-ratio = 25, и rna-span-log-min = 13. Отфильтровывают скаффолды с множеством цепочек, где указанный показатель превышает конфигурируемый порог, например rna-max-covg-gap = 150. Порог, равный 200, значительно ослабляет фильтр, а 100 значительно уплотняет.

[00361] Фильтр может быть применен к завершенным или незавершенным скаффолдам, полученным из списка цепочек затравок для данного рида, посредством отслеживания или вычисления максимального покрытия среди всех скаффолдов, а также сканирования списка скаффолдов, и отбрасывания тех, для которых filter_metric > rna-max-covg-gap.

[00362] Такой фильтр может также быть применен в качестве рекурсивного отсечения в середине скопления скаффолдов. При добавлении каждой новой цепочки затравок в скаффолд аннотируют обновленное покрытие референса, а также потенциальное покрытие, вычисляемое посредством вычитания промежутков покрытия в пределах и перед указанной цепочкой затравок из длины рида. С использованием указанного частичного охвата и потенциального покрытия, если показатель запустит фильтрацию, то любой более длинный скаффолд, использующий текущий потенциальный скаффолд в качестве префикса, наиболее вероятно будет отфильтрован, так как охват референса будет только повышаться, а потенциальное покрытие будет только уменьшаться. Таким образом, вся рекурсия, имеющая текущий потенциальный скаффолд в качестве префикса, может быть пропущена. Отсечение рекурсии при помощи фильтра скаффолдов таким образом может значительно уменьшить длину рекурсии для формирования набора

скаффолдов из длинного списка цепочек затравок.

[00363] Также может быть оптимизировано осуществление рекурсивного поиска скаффолдов. Рекурсивные циклы повторно сканируют части списка цепочек затравок и скорость алгоритма таким образом значительно лучше, когда лист цепочек затравок короче. Но в действительности не обязательно выполнять рекурсивный алгоритм над всем списком цепочек затравок, когда некоторые цепочки затравок не могут образовывать скаффолд с другими цепочками затравок. Один способ оптимизации заключается в обнаружении «изолированных» цепочек затравок, которые расположены дальше, чем maxIntronLen (например, 1000000 оснований) от любой другой цепочки затравок в референсе. Изолированные цепочки затравок могут автоматически быть выделены как скаффолды с одной цепочкой и удалены из списка цепочек затравок перед дальнейшим формированием скаффолдов, тем самым укорачивая список цепочек затравок, сканируемого в процессе рекурсии. Аналогичным образом могут быть обнаружены хорошо отделенные подмножества цепочек затравок, например, подмножества с каждой хромосомой или подмножества, отделенные более, чем maxIntronLen в референсе, и рекурсивное формирование скаффолдов может быть выполнено отдельно над каждым таким подмножеством, обеспечивая в результате сниженное общее время исполнения.

[00364] Первый этап модуля выравнивателя для каждого скаффолда заключается в точном расположении каждой границы сплайсинга, границы между двумя сегментами экзона (представленные двумя соответствующими цепочкам затравок в скаффолде). Это называют «сшиванием» сегментов экзона вместе или сшивание границ сплайсинга. Более точное сшивание все еще необходимо после картирования затравок и формирования скаффолдов, так как две последовательные цепочки затравок сами по себе могут не давать ясного представления, где между ними пролегает верная граница. Например, последовательные цепочки затравок могут быть разделены некоторым расстоянием в риде, если затравки не удалось картировать сразу на одной или обеих сторонах границы сплайсинга; или последовательные цепочки затравок могут перекрывать друг друга в риде, особенно если последовательность рида на конце одного экзона совпадает с последовательностью в начале следующего экзона. Даже если последовательные цепочки затравок примыкают без промежутка или перекрываются, это не гарантирует, что граница между ними пролегает в верном положении границы сплайсинга.

[00365] Сшивание границ сплайсинга таким образом в первую очередь представляет собой анализ для выбора лучшего положения для сшивания между последовательными сегментами экзона в риде, соответствующего наиболее вероятному положению границы сплайсинга. Для определения этого полезными являются два фактора. Первый фактор заключается в сравнении последовательности рида с левой и правой последовательность референса в двух местоположениях картирования сегментов экзона в референсном геноме. Данное положение сшивания подразумевает. что основания рида, левые по отношению ко шву, картируются на левую область референса, а основания рида, правые по отношению ко шву, картируются на правую область референса. При перемещении потенциального положения сшивания слева направо в риде основания рида меняют свое картирование при их пересечении с правой области референса на левую.

[00366] Как показано на ФИГ. 4, верное положение границы сплайсинга наиболее вероятно будет иметь хорошее совпадение между левой частью рида и левой областью референса, а также между правой частью рида и правой областью референса. Общее количество несовпадений (или ОНП) могут быть подсчитаны на обеих сторонах

потенциального положения сшивания посредством сравнения с соответствующей областью референса; и положения сшивания с меньшим числом ОНП наиболее вероятно будут верными. Сравнение частей рида, левых и правых для каждого положения сшивания, показано на фиг. 4.

[00367] Подсчет ОНП эффективно моделируют посредством сканирования положений сшивания через посредством окна возможных положений в риде, например перекрывая каждые из двух цепочек затравок не более чем на некоторое расстояние, например, на 48 оснований. Указанное сканирование запускают в аппаратном модуле выравнивателя, например, на скорости в одно положение за цикл. Каждый раз, когда сканирование проходит один шаг, например слева направо, только одно основание рида меняет свое картирование с правой области референса на левую область референса. Таким образом, сравнение левых последовательностей либо приобретает одно ОНП, или остается неизменным, а сравнение правых последовательностей либо теряет одно ОНП, либо остается неизменным; и таким образом прямое количество ОНП изменяется на -1, 0 или +1. Такое постепенное изменение количества ОНП для каждого шага может быть вычислено посредством сравнения одного основания рида (того, которое было пересечено при шаге положения сшивания) с двумя основаниями референса. Если такое постепенное изменение количества ОНП суммируется при выполнении шагов слева направо, то текущая сумма может быть взята в качестве относительной оценки, где минимальная оценка является наилучшей. Эквивалентным образом, каждому совпадающему основанию может быть присвоена положительная оценка совпадения, а каждому несовпадающему основания негативный штраф несовпадения; и сумма постепенных изменений оценки должна быть максимальной для наилучшего положения сшивания.

[00368] Другим фактором, предполагаемым каждым положением сшивания, является мотив интрона. Мотив интрона определен как первые два основания и последние два основания пропущенного сегмента референса, или интрона. Эквивалентным образом, мотив для любого потенциального положения сшивания образован из двух первых оснований после левой области референса и последних двух оснований перед правой областью референса, как показано на фиг. 4. Этот предполагаемый мотив интрона зависит от положения сшивания и обычно меняется при сканировании положения сшивания в окне возможных положений сшивания. Некоторые «каноничные» мотивы интрона возникают значительно боле часто, чем другие, при сплайсинге естественной РНК. Положение сшивания, которое соответствует каноничному мотиву интрона, наиболее вероятно будет верным положением границы сплайсинга, особенно, если он является одним из наиболее частых каноничных мотивов.

[00369] Таблица трех каноничных мотивов интронов в человеческой РНК показана в ТАБЛИЦЕ I. Для каждого мотива также показано его обратное соответствие, так как в большинстве протоколов секвенирования РНК риды могут картироваться как прямо, так и с обратным соответствием относительно транскрибированной генной нити, так что хотя возникает только «прямой» каноничный мотив в оригинальной считываемой генной нити, его обратное соответствие может возникнуть в РНК-секвенированных ридах. Для каждого каноничного мотива и для оставшихся классовых или неканонических мотивов показан средняя частота в сплайсинге человеческой РНК вместе со штрафом оценки образца, который может быть использован, например, с совпадениями оснований с оценкой +1 и несовпадениями оснований с оценкой -4.

ТАБЛИЦА I

Мотив интрона	Обратное соответствие	Приблизительная частота	Штраф оценки

GT/AG	CT/AC	98,73%	0
GC/AG	CT/GC	1,03%	10
AT/AC	GT/AT	0,11%	15
250 неканоничных мотивов		0,13%	25

[00370] Модуль сшивания сплайсинга соответственно сканирует потенциальные положения сшивания в окне возможных местоположений сшивания, например слева направо, суммируя постепенное изменение оценки вследствие изменения основаниями того, на какую область они картируются, а также вычитания в каждом потенциальном положении сшивания штрафа мотива интрона в соответствии с мотивом интрона, наблюдаемым непосредственно после левой области референса и непосредственно перед правой областью референса, и выбирает для сшивания положение с максимальной оценкой.

[00371] Кроме того, некоторые специальные результаты могут быть рассмотрены и оценены. Сшивание на левом краю или правом краю окна возможных положений сшивания может рассматриваться как ошибка сшивания и наиболее вероятно возникнет, когда одна из двух областей референса не является верным положением картирования для сегмента экзона рида, например, когда следовали аннотированной границе сплайсинга, но оказывается верным для данного рида. Сшиванию по левому или правому краю преимущественно могут присвоить бонусную оценку, например, 25 по той же приведенной в качестве примера шкале оценок, так что для успешного сшивания должно появиться значительное доказательство верной границы сплайсинга.

[00372] Кроме того, если аннотированная граница сплайсинга идентифицирована как связывающая две сшиваемых цепочки затравок, аннотированная граница находится в известном положении в окне возможных положений сшивания. В качестве опции, операция сшивания может быть пропущена, просто принимая известное положение аннотированной границы. В качестве другой опции, операция сшивания может быть выполнена, но известному положению аннотированной границы может быть присвоена бонусная оценка и/или автоматически присвоен наилучший из доступных штрафов мотива интрона, или нулевой штраф. В качестве другой опции, вместо штрафа мотива интрона известному положению аннотированной границы может быть присвоена бонусная оценка или штраф, связанные с наблюдаемой общностью или редкостью той границы сплайсинга, которая аннотирована в базе данных аннотирования. Если известное положение аннотированной границы сплайсинга выбрано для сшивания, то граница сшивания может быть отмечена как согласующаяся с аннотированной границей, так что этот факт может быть сообщен, если указанная граница сплайсинга возникнет в выходном выравнивании рида.

[00373] Выбранные положения сшивания могут быть аннотированы в скаффолды различными способами. В предпочтительном варианте реализации составляющие цепочки затравок скаффолда редактируют для начала и конца, непосредственно соседних к выбранным положениям сшивания.

Кроме того, предпочтительно, чтобы модуль выравнивателя в некоторых обстоятельствах выполнял большие редактирования скаффолдов на основании результатов сшивания. Если сшивание не удается, то скаффолд могут усечь или разделить на два скаффолда в точке ошибки. Кроме того, может быть выполнена попытка сшивания между несмежными цепочками затравок в скаффолде, например, с пропуском одной цепочки затравок. Например, для скаффолда, содержащего цепочки затравок 1, 2, 3 и 4, сшивание границы сплайсинга должно естественным образом выполняться между парами цепочек (1,2), (2,3) и (3,4); но, в дополнение, может быть выполнена

попытка сшивания между парами цепочек (1,3) и (2,4). Если сшивание от 1 до 3 имеет лучшую оценку, чем сшивание от 1 до 2 с последующем сшиванием от 2 до 3, то цепочку 2 затравок могут отбросить из скаффолда.

[00375] Определив точные положения границы сплайсинга в рассматриваемом скаффолде посредством сшивания, могут быть определены соответствующие завершенные выравнивания и оценки выравнивания для каждого скаффолда посредством использования модуля беспромежуточного выравнивателя или выравнивателя с промежутками (например, Смита-Ватермана). Для скаффолда только с одной цепочкой затравок это не сильно отличается от выравнивания для ридов ДНК и могут быть использованы такие же аппаратные модули и способы. Для скаффолда с множеством цепочек затравок необходимы некоторые дополнительные способы для получения завершенного, возможно со сплайсингом (содержащего интронные операции), выравнивания.

[00376] Одним способом, посредством которого могут быть определены завершенные выравнивания со сплайсингом, является отдельное выравнивание каждого сегмента экзона рида, соответствующего каждой цепочке затравок в скаффолде, с соответствующем ему сегментом референса, посредством беспромежуточного выравнивателя и/или выравнивателя с промежутками. Такой подход имеет недостатки, когда требуются локальные (т.е., возможно, обрезанные) выравнивания. Если выравнивания отдельных сегментов экзона получают без обрезания, то их нелегко собрать вместе в завершенное выравнивание со сплайсингом. Если выравнивания отдельных сегментов экзона получают без обрезания, то их оценки выравнивания могут изучить для определения того, будет ли обрезать наилучшее общее локальное выравнивание один или более сегментов экзона, но подходящее ограничение в произвольных местоположениях в пределах сегментов экзона определить нелегко. Таким образом, может быть дорого получать обрезанный и необрезанные версии выравнивания каждого сегмента экзона для решения указанных трудностей.

[00377] Один способ определение завершенных выравниваний со сплайсингом для скаффолда с множеством цепочек включает объединение сегментов экзона вместе перед выравниванием. В каждый модуль выравнивателя - беспромежуточный или с промежутком - можно подавать две последовательности нуклеотидов для выравнивания, рассматриваемую последовательность (рида) и референсную последовательность. Объединенная рассматриваемая последовательность может представлять собой просто целый рид, представляющий собой объединение его сегментов экзона, при необходимости с обрезанием начала или конца рида, если первый или последний сегмент экзона не проходит в начало или конец рида.

[00378] Объединенная референсная последовательность получена посредством извлечения сегмента референсного генома, представляющего собой картированное изображение каждой цепочки затравок сегмента экзона, и объединения указанных референсных сегментов вместе. Следует отметить, что для данного сегмента экзона (цепочки затравок) его референсный сегмент может отличаться по длине от его сегмента рида в случае, когда самые левые затравки в цепочке затравок попадают в несколько отличающуюся диагональ выравнивания, чем самые правые затравки; например, затравки в цепочке затравок предполагают наличие индела. В таком случае следует использовать выравниватель с промежутком.

[00379] Кроме того, для выравнивания с промежутком первый и последний сегменты экзона референсной последовательности могут проходить наружу, например, проходить через первый сегмент экзона с 50 предшествующих оснований референса и последний

сегмент экзона и 50 последующих оснований референса для обеспечения пространства для делеций в пределах первого и последнего сегментов экзона.

[00380] При помощи объединения сегментов экзона скаффолда из рида и референса, можно обеспечить подачу единой объединенной рассматриваемой последовательности и единой объединенной референсной последовательности в модуль выравнивателя, который затем может функционировать практически тем же образом для выравниваний РНК со сплайсингом, как и для выравнивания (РНК или ДНК) без сплайсинга. Однако некоторые другие модификацию будут обеспечивать преимущество. Во-первых, для определения подходящего обрезания локальных выравниваний в любом положении в пределах любого сегмента экзона некоторые штрафы оценок (или бонусы) могут быть применены к каждой границе сплайсинга, которую пересекает выравнивание. В одном варианте реализации, штраф оценки для каждой границы сплайсинга относится к ее мотиву интрона и статусу аннотированной границы сплайсинга, и может представлять собой тот же штраф оценки, использованный при сшивании границ сплайсинга.

[00381] Соответственно, неаннотированная граница сплайсинга с редким или неканоничным мотивом интрона может иметь большой штраф связанной оценки и наиболее вероятно, что один или более из всех сегментов экзона будут обрезаны из выравнивания сплайсинга для исключения такой маловероятной границы сплайсинга до тех пор, пока не произойдет достаточное количество совпадений последовательности на обеих сторонах границы сплайсинга для того, чтобы служить в качестве убедительного доказательства того, что граница сплайсинга действительно присутствует, посредством преодоления штрафа ее оценки. В предпочтительном варианте реализации каждая из объединенных опрашиваемой и референсной последовательностей имеет ложные основание, вставленное между последовательными сегментами экзона, и подходящий штраф оценки для каждой границы сплайсинга прикреплен к соответствующему ему ложному основанию. Это позволяет включить штрафы оценки границы сплайсинга без специализированной логической части и обеспечивает пространство для возможного обрезания выравнивания с любой стороны ложного основания границы сплайсинга.

[00382] Кроме того, в различных примерах для выравнивания с промежутком, например, Смита-Ватермана, может быть преимуществом вынуждать части выравнивания проходить через заранее определенные границы сплайсинга. Другими словами, ни один путь выравнивания не должен переходить от одного рассматриваемого сегмента экзона к следующему без одновременного перехода от соответствующего сегмента экзона референса к следующему. Одной причиной указанного ограничения является то, что только надлежащим образом синхронизированная граница сплайсинга будет иметь надлежащую оценку на основании мотива интрона, определенного во время сшивания границы сплайсинга.

[00383] Другой причиной является то, что объединенная референсная последовательность была образована с использованием точных границ сегмента экзона референса, соответствующих выбранному положению сшивания каждой границы сплайсинга, так что отсутствуют дополнительные основания референса для выравнивателя с промежутком для свободной регулировки границ сплайсинга. Кроме того, для того, чтобы избежать выравниваний, которые сложно интерпретировать (например, такие как строки CIGAR с операциями ‘I’ или ‘D’, смежными с ‘N’ операций), желательно потребовать, чтобы по меньшей мере одно рассматриваемое основание и основание референса перед каждой включенной границей сплайсинга были выровнены «по диагонали» (рассматриваемое основание выровнено с основанием референса как

операция ‘M’ CIGAR), и одно рассматриваемое основание и основание референса после каждой включенной границей сплайсинга были выровнены «по диагонали».

[00384] Для усиления указанных ограничений объединенные референсная и рассматриваемая последовательности разделены на зоны, которым присвоены идентификаторы или ID зоны, например, целочисленные значения. В одном варианте реализации один присвоен каждому ложному основанию между сегментами экзона, другой ID зоны присвоен последнему основанию каждого сегмента экзона, предшествующему границе сплайсинга (но не последнему основанию объединенной последовательности), и другой ID зоны присвоен всем оставшимся основаниям каждого сегмента экзона.

[00385] Например, для скаффолда с тремя сегментами экзона (цепочками затравок), каждый длиной 20 оснований, может быть 4 ID зоны: зона 1 для оснований 1-19 первого сегмента экзона, зона 2 для основания 20 первого сегмента экзона, зона 3 для ложного основания между первым и вторым сегментами экзона, зона 4 для оснований 1-19 второго сегмента экзона, зона 5 для основания 20 второго сегмента экзона, зона 6 для ложного основания между вторым и третьим сегментами экзона и зона 7 для оснований 1-20 третьего сегмента экзона. Одинаковое картирование зон применяют к объединенной рассматриваемой последовательности и к объединенной референсной последовательности, учитывая, что соответствующие рассматриваемые зоны и референсные зоны с множеством оснований с одинаковыми ID зоны могут иметь различные длины вследствие инделов в цепочках затравок. Затем, в беспромежуточном выравнивателе, ячейка оценки модифицирована так, чтобы допускать только верные оценки выравнивания на пересечении между идентичными ID зоны, т.е. когда рассматриваемый ID зоны совпадает с ID зоны референса.

[00386] Соответственно, в различных примерах, настоящее изобретение направлено на устройства и способы для использования при картировании и выравнивании ДНК и/или РНК. Таким образом, в конкретных примерах обеспечена жестко смонтированная цифровая логическая схема, например, интегральная схема, причем интегральная схема включает конфигурацию, например, жестко смонтированную и/или предварительно сконфигурированную конфигурацию, которая выполнена с возможностью реализации одного или более этапов операции картирования и/или выравнивания ДНК и/или РНК. В частности, раскрытые в настоящем документе устройства могут быть выполнены с возможностью реализации различных типов анализа над РНК, например, типы анализа РНК, реализуемые одним или более жестко смонтированных модулей обработки или их подмножества.

[00387] Например, в некоторых вариантах реализации обеспечены устройство и/или система для реализации конвейера анализа последовательности ДНК и/или РНК над данными последовательности ДНК и/или РНК, например, над ридом полученных из РНК геномных данных. В таком примере система может содержать одно или более из: памяти, например, для хранения одной или более референсных последовательностей ДНК и/или РНК, например, полученных из РНК геномных референсных данных, индекса одной или более референсных последовательностей ДНУ и/или РНК и множества ридов геномных данных, например, когда каждая из референсных последовательностей ДНК и/или РНК и множество ридов данных последовательности содержат последовательность нуклеотидов; а также интегральную схему, раскрытую в настоящем документе. В частности, интегральная схема может быть образована из множества жестко смонтированных цифровых логических схем, которые могут быть соединены друг с другом посредством множества физических электрических

межсоединений. В таком примере одно или более из множества физических электрических межсоединений могут содержать интерфейс памяти для доступа интегральной схемы к памяти. Кроме того, жестко смонтированные цифровые логические схемы могут быть выполнены в виде набора модулей обработки, такого где один или более из модулей обработки сформированы подмножеством жестко смонтированных цифровых логических схем и выполнены с возможностью реализации по меньшей мере одного этапа в конвейере анализа геномных последовательностей ДНК и/или РНК на множестве ридов секвенированных данных. Следует отметить, что в различных примерах рид полученных из РНК геномных данных может указывать на рид, который был получен непосредственно секвенированием образца РНК или посредством секвенирования некоторого дальнейшего продукта, полученного из образца РНК, например, обратно транскрибированную кДНК и тому подобное, и могут называться в настоящем документе как «рид РНК» или «рид длинных РНК», который содержит общие сведения об источнике, из которого получены данные РНК.

[00388] В частности, множество модулей обработки может содержать модуль картирования ДНК и/или РНК, модуль выравнивания, модуль сортировки и/или модуль определения вариантов, который может содержать модуль скрытой марковской модели (HMM) и/или модуль Смита-Ватермана (SW). Например, в первой конфигурации жестко смонтированная цифровая логическая схема, раскрытая в настоящем документе, может быть выполнена с возможностью осуществлять доступ в памяти, посредством интерфейса памяти, по меньшей мере к некоторым из последовательностей нуклеотидов ДНЕ и/или РНК в выбранном риде из множества ридов и к индексу одной или более референсных последовательностей ДНК и/или РНК, и для картирования выбранного рида ДНК и/или РНК на один или более сегментов указанных одной или более генетических референсных последовательностей на основе индекса для получения картированного рида. В конкретных случаях, например, в отношении картирования РНК, модуль картирования РНК может быть выполнен с возможностью реализации одного или более из картирования с применением привязанных затравок, поиска аннотированной границы сплайсинга и/или формирования скаффолдов из цепочек затравок и/или тому подобного по отношению к этапам картирования РНК.

[00389] Аналогичным образом, во второй конфигурации жестко смонтированных цифровых логических схем может быть обеспечен модуль выравнивания, выполненный с возможностью доступа к одной или более референсных последовательностей ДНК и/или РНК из памяти посредством интерфейса памяти для выравнивания картированных ридов ДНК и/или РНК, например, из модуля картирования, на одно или более положений в одном или более сегментов одной или более референсных последовательностей ДНК и/или РНК для получения выровненного рида. В конкретных случаях, например, в отношении выравнивания РНК, модуль выравнивания РНК может быть выполнен с возможностью реализации одного или более из сшивания границ сплайсинга и/или выравнивания рида со сплайсингом, и/или тому подобного по отношению к этапам выравнивания РНК.

[00390] Соответственно, в различных примерах может быть обеспечена жестко смонтированная цифровая логическая схема, причем цифровая логическая схема, или их подмножество, содержит модуль картирования и/или выравнивания, которые могут быть выполнены с возможностью содержать множество сконфигурированных, например, предварительно сконфигурированных, модулей обработки для реализации одного или более этапов конвейера анализа РНК, например так, что один или более этапов могут включать картирование с применением привязанных коротких затравок,

поиск аннотированной границы сплайсинга, формирование скаффолдов цепочек затравок, сшивание границ сплайсинга, выравнивание рида со сплайсингом и/или один или более других связанных этапов с реализацией операций картирования и/или выравнивания, например в конвейере генетического анализа.

[00391] Кроме того, в некоторых примерах обеспечен модуль определения оснований (VC), сконфигурированный, например, в третьей конфигурации жестко смонтированных цифровых логических схем, с возможностью доступа к выровненному риду ДНК и/или РНК и по меньшей мере одной из референсных последовательностей и выполнения одного или более из приведенных далее этапов. Например, модуль определения оснований может быть выполнен с возможностью сравнения последовательности нуклеотидов в выровненном риде ДНК и/или РНК с последовательностью нуклеотидов по меньшей мере одной генетической референсной последовательности так, чтобы определять одну или более разницу между последовательностью нуклеотидов в выровненном риде ДНК/РНК и последовательностью нуклеотидов ДНК/РНК по меньшей мере в одной генетической референсной последовательности, а также с возможностью генерации одного или более определения вариантов, представляющих указанную одну или более разностей. Кроме того, в отношении раскрытой интегральной схемы, одно или более из множества физических электрических межсоединений может также включать в себя выход из интегральной схемы для обмена результирующими данными из модуля картирования, и/или модуля выравнивания, и/или модуля определения вариантов.

[00392] В частности, интегральная схема по настоящему изобретения может содержать одно или более множеств жестко смонтированных цифровых логических схем и/или их подмножеств, например, включая первое, второе, третье и более множеств сконфигурированных, например, предварительно сконфигурированных, жестко смонтированных цифровых логических схем, которые сконфигурированы как один или более модулей обработки для реализации одного или более отдельных этапов в конвейере анализа последовательностей ДНК и/или РНК. Например, жестко смонтированная цифровая логическая схема может содержать первое подмножество цифровых логических схем, которые выполнены как модуль обработки так, чтобы принимать рид данных ДНК и/или РНК посредством одного или более физических электрических межсоединений. Кроме того, может быть обеспечено второе подмножество жестко смонтированных цифровых логических схем, которое выполнено как модуль обработки для выделения части рида ДНК или РНК для формирования затравки, например, когда затравка представляет подмножество последовательности нуклеотидов ДНК или РНК, представленной ридом, например, для реализации одного или более или картирования с применением привязанных коротких затравок. Может быть включено одно или более дополнительных подмножеств цифровых логических схем, например, модуль обработки для поиска аннотированной границы сплайсинга и/или для выполнения картирования цепочек затравок. Кроме того, могут быть включены подмножества цифровых логических схем, сконфигурированные как модуль обработки, например, для выполнения одной или более функций выравнивания над данными РНК, включая подмножество цифровых логических схем для выполнения операции сшивания границ сплайсинга и/или выравнивания рида со сплайсингом.

[00393] Соответственно, интегральная схема по настоящему изобретению может содержать одну или более цифровых логических схем или их подмножество для выполнения одного или более этапов при выравнивании с применением привязанных коротких затравок. Как более подробно раскрыто в настоящем документе, картирование

с применением коротких затравок может быть выполнено для усовершенствования чувствительности базы данных, например, на основе хэш-таблицы, картирования затравок, например посредством использования совпадений более длинных затравок в качестве точек привязки для направления локализованных поисков с более короткими затравками. Это является полезным для того, чтобы картирование на основе хэш-таблиц хорошо работало для ридов РНК, но также это полезно для улучшения чувствительности в отношении картирования ДНК. В частности, картирование начальных затравок может использовать одну или более затравок с K основаниями, полученных из рида геномных данных ДНК и/или РНК для обращения к первому индексу, например, индексу на основе хэш-таблицы, референсного генома ДНК и/или РНК. В таком примере последующее картирование с применением привязанных коротких затравок может быть выполнено, например, с использованием затравок с L основаниями (L<K) для обращения ко второму индексу на основе хэш-таблицы множества ячеек референса, например, когда каждая из множества ячеек референса может быть завершенным или незавершенным подмножеством референсного генома. Это также может быть полезным, когда используется затравка с K и/или с L основаниями для отдельного обращения к первому и/или второму индексу, например, к индексу на основе хэш-таблицы, так, чтобы таким образом выявлять каждую одну или более привязанных ячеек, выбранных из множества ячеек референса.

[00394] В таком примере может быть обеспечена одна или более структур данных, например, одиночная или множественная структура данных, причем структура(ы) данных может включать первый индекс, например, индекс на основе хэш-таблицы, и может также включать второй индекс, например, второй индекс на основе хэш-таблицы. Кроме того, когда обеспечена одна или более привязанных ячеек, выявление одной или более привязанных ячеек может быть включено в качестве этапа в способах. Следовательно, такое выявление может подразумевать включение идентификатора каждой привязанной ячейки в хэш-ключ, например, хэш-ключ, используемый для обращения к первому и/или второму индексам на основе хэш-таблицы. В некоторых примерах одна или более привязанных ячеек, например, для обращений с затравками с L основаниями, могут быть выбраны, например, в модуле картирования, на основании совпадений, обнаруженных посредством обращений с затравками с K основаниями. Соответственно, модуль картирования может быть выполнен с возможностью реализации операции картирования с применением привязанных коротких затравок перед или после вывода местоположений совпадения, например, когда одна или более привязанных ячеек для обращений с затравками с L основаниями выбраны в или вне модуля картирования на основании вывода местоположений совпадения посредством модуля картирования. В некоторых случаях модуль картирования может выполнять картирование с применением привязанных коротких затравок, когда при вторичной процедуре картирования могут проходить через по меньшей мере подмножество входных ридов.

[00395] Кроме того, в отношении выполнения выравнивания, одна или более привязанных ячеек для обращений с затравками с K или с L основаниями могут быть выбраны в или вне модуля картирования и могут быть основаны на выходных данных выравнивания на выходе модуля выравнивания, либо в программном обеспечении, либо в аппаратном обеспечении, например, когда модуль выравнивания принимает местоположения совпадения из модуля картирования. В таком примере может быть выбрано подмножество входных ридов, с включением или исключением ридов с достаточно обрезанными выравниваниями. В некоторых примерах может быть выбрано

подмножество входных ридов, с включением или исключением ридов с выравниваниями с достаточно низкой оценкой. И в различных других примерах входные риды могут быть ридами со спаренными концами, и может быть выбрано подмножество входных ридов, с включением или исключением пар ридов с отсутствием выравниваний в надлежащим образом спаренных конфигурациях.

[00396] Кроме того, модуль картирования для выполнения анализа последовательности РНК может включать один или более модулей обработки, которые выполнены с возможностью реализации поиска одной или более границ сплайсинга. Поиск границ сплайсинга может применяться для улучшения чувствительности картирования специфичных для РНК ридов. Например, после картирования частей рида РНК с использованием референсного индекса РНК, например, который может представлять или не представлять собой индекс на основе хэш-таблицы, может быть сгенерирована индексу на основе хэш-таблицы опрошена «база данных», или любая другая подходящая форма структуры данных, например, в той же памяти, что и референс и/или его индекс. В особенности, «база данных» может быть сгенерирована на основании известных и/или определенных границ сплайсинга РНК для исследуемых видов и к ней может осуществляться доступ на основании картированных положений. Следует отметить, что каждая известная граница сплайсинга может представлять по возможности длинный сегмент «интрона» (до 1Mbp или длиннее) в референсе, например, где выравнивания рида обычно будут «перепрыгивать» от одной конечной точки интрона к другой.

[00397] Соответственно, доступ к базе данных могут осуществлять таким образом для извлечения известных границ сплайсинга, которые имеют одну конечную точку в каждом из сегментов референса, или вблизи него, вдоль уже картированных частей рида, но которые могут иметь другую часть рида, которая проходит за пределы расположенной рядом конечной точки интрона. В таких примерах ориентировочно может быть принято, что указанная другая часть рида затем продолжит совпадение с референсом после перескакивания через интрон, даже если предыдущие попытки картирования могли не обнаружить какое-либо совпадение в указанной области. Более позднее выравнивание со сплайсингом и оценивание затем может измерить, насколько хорошо рид в действительности совпадает с выравниванием по указанной границе сплайсинга. Этот способ, таким образом, может улучшить способность интегральной схемы обнаруживать наиболее вероятное выравнивание со сплайсингом рида, несмотря на препятствия, например, короткие экзоны, выступающие части короткого экзона и/или редактирования (ОНП и т.д.), блокирующие совпадение в сегменте экзона средней длины.

[00398] Следовательно, может быть обеспечена память, выполненная с возможностью соединения с интегральной схемой, причем память содержит индекс референсного генома и список аннотированных границ сплайсинга в пределах референсного генома. Модуль картирования, таким образом, может картировать первую часть рида полученных из РНК геномных данных на местоположение совпадения в референсном геноме, например, посредством осуществления доступа к индексу референса с использованием первой части рида. Модуль картирования затем может осуществлять доступ к списку аннотированных границ сплайсинга и извлекать дескриптор интрона, например, когда дескриптор интрона включает первую конечную точку и вторую конечную точку в референсном геноме, например, когда первая конечная точка находится в пределах ограниченного расстояния от местоположения совпадения в референсном геноме.

[00399] Модуль картирования может затем картировать ворую часть рида полученных из РНК геномных данных на предполагаемое местоположение в референсном геноме, например, когда предполагаемое местоположение в референсном геноме может быть смежным со второй конечной точкой в референсном геноме дескриптора интрона. Картировщик затем выдает картированные местоположения рида полученных из РНК геномных данных, например, когда картированные полученные из РНК геномные данные включают по меньшей мере местоположение совпадения в референсном геноме и предполагаемое местоположение в референсном геноме. Соответственно, в различных примерах может быть обеспечен список аннотированных границ сплайсинга, причем список может быть сформулирован в виде таблицы, содержащей запись для одной или более, например, каждой, из множества ячеек референса, таким образом формируя разделение референсного генома. К списку аннотированных границ сплайсинга затем могут осуществлять доступ таким образом, который включает определение по меньшей мере одной ячейки референса в пределах ограниченного расстояния от местоположения совпадения в референсном геноме и осуществление доступа к записям таблицы, которые соответствуют по меньшей мере одной ячейке референса.

[00400] В таком примере действительное местоположение границы сплайсинга в риде может быть определено, например, посредством использования первой части рида, например, совпадающей с местоположением в референсном геноме, посредством использования первой конечной точки и второй конечной точки дескриптора интрона в референсном геноме. Кроме того, в некоторых примерах ограниченное расстояние может быть определено по меньшей мере для того, чтобы было предусмотрено, что действительное местоположение границы сплайсинга в риде не находится вне границ рида. Соответственно, может быть определена первая цепочка затравок, например с использованием по меньшей мере первой части рида полученных из РНК геномных данных и местоположения совпадения в референсном геноме. Затем может быть определена вторая цепочка затравок, например, посредством использования по меньшей мере второй части рида полученных из РНК геномных данных и предполагаемого местоположения в референсном геноме. Затем может быть установлена связь между первой цепочкой затравок и второй цепочкой затравок, и выходные данные могут быть картированы на местоположения рида полученных из РНК геномных данных, которые включают описания первой цепочки затравок, второй цепочки затравок и связи между ними.

[00401] Кроме того, модуль картирования для выполнения анализа последовательности РНК может включать один или более модулей обработки, которые выполнены с возможностью реализации операции формирования скаффолдов цепочек затравок. Следует отметить, однако, что хотя в настоящем документе упоминаются цепочки «затравок», такие цепочки «затравок» не ограничены контекстом картирования затравок на основе хэш-таблицы, эта концепция может быть распространена на любое картирование части рида на сегмент референсного генома. Таким операции формирования скаффолдов полезны для перевода списка цепочек затравок в список скаффолдов, например, когда каждый скаффолд представляет собой последовательность одной или более цепочек затравок, которая представляет вероятное выравнивание со сплайсингом рида.

[00402] В различных примерах последовательные части рида могут картироваться на последовательные сегменты одной хромосомы референса, например, в согласующихся ориентации и порядке. Соответственно, способ формирования множества скаффолдов может быть избирательным, так как по существу для списка цепочек затравок

количество объединенных последовательностей и количество цепочек затравок может быть очень велико. Таким образом, список скаффолдов может быть сгенерирован так, чтобы быть достаточно обширным для того, чтобы включать верное выравнивание со сплайсингом рида с высокой степенью достоверности, например, без генерации слишком большого количества ложных скаффолдов. В таком примере каждый скаффолд затем может быть оценен, например посредством выравнивания со сплайсингом.

[00403] Соответственно, модуль картирования по настоящему изобретению могут быть выполнен с возможностью определения списка цепочек затравок, так что каждая цепочка затравок списка цепочек затравок представляет совпадение между соответствующей частью рида полученных из РНК геномных данных и соответствующего сегмента референсного генома. Посредством изучения списка цепочек затравок картировщик может затем выработать список скаффолдов, например, в котором каждый скаффолд может содержать одну или более отдельных цепочек затравок из списка цепочек затравок, и/или скаффолд предполагает соответствующую последовательность части рида одной или более соответствующей части рида полученных из РНК геномных данных. В последовательность части рида может развиваться в едином направлении в риде полученных из РНК геномных данных. Аналогичным образом, скаффолд может также предполагать соответствующую последовательность сегмента референса одного или более соответствующих сегментов референсного генома, так что развитие последовательности сегмента референса происходит в едином направлении в референсном геноме.

[00404] Кроме того, получение списка скаффолдов может дополнительно включать сортировку списка цепочек затравок в возрастающем или убывающем порядке соответствующих сегментов референсного генома, например, когда список скаффолдов получают в соответствии с набором правил, который определяет, когда одно цепочка затравок может следовать за другой цепочкой затравок в скаффолде. Такой набор правил включает минимально и/или максимально допустимый промежуток и/или минимально и/или максимально допустимое перекрывание между последовательными частями рида в последовательности частей рида. В некоторых примерах последовательность частей рида и последовательность сегментов референса могут предполагать, что последовательность длины интрона вычисленных диагоналей выравнивания смещается из каждой части рида и соответствующего сегмента референса на следующую часть рида. Кроме того, набор правил может включать минимально допустимую длину интрона и максимально допустимую длину интрона для последовательности длины интрона.

[00405] Кроме того, в различных примерах выполнение списка скаффолдов может предполагать выполнение начальной части скаффолда, имеющей частичную последовательность из одной или более отдельных цепочек затравок, и может дополнительно включать выполнение последовательность двух отдельных скаффолдов в списке скаффолдов, так что один или более из по меньшей мере двух отдельных скаффолдов может представлять собой расширения начальной части скаффолда до более длинных скаффолдов. Такое выполнение списка скаффолдов может включать отфильтровывание меньших скаффолдов, например, как меньших в соответствии с вычисленным показателем фильтрации, например, показатель фильтрации, вычисленный с использованием разности между охватом каждого скаффолда рида полученных из РНК геномных данных и/или максимального покрытия рида полученных из РНК геномных данных, например, максимального покрытия, вычисленного по всему списку скаффолдов. В некоторых примерах показатель фильтрации может быть вычислен с

использованием охвата каждого скаффолда в референсном геноме.

[00406] Как указано выше, в различных примерах может быть обеспечена жестко смонтированная цифровая логическая схема, причем цифровая логическая схема, или подмножество логических схем, содержит модуль выравнивания, который может быть выполнен с возможностью содержать множество модулей обработки для реализации одного или более этапов выравнивания в конвейере анализа РНК, например так, что один или более этапов могут включать сшивание границ сплайсинга и/или выравнивание рида со сплайсингом. В частности, модуль выравнивания для выполнения анализа последовательности РНК может включать один или более модулей обработки, которые выполнены с возможностью реализации одной или более операций сшивания границ сплайсинга и/или одного или более выравниваний рида со сплайсингом.

[00407] В частности, в различных примерах может быть сгенерирована и/или другим образом обеспечена пара частичных картирований, например, цепочек затравок, например, двух последовательных цепочек затравок в скаффолде, в риде РНК. В различных примерах частично картированные цепочки затравок могут представлять два сегмента экзона вероятного картирования со сплайсингом, которое пропускает возможный интрон в референсе. В таком примере модуль обработки модуля выравнивания может быть выполнен с возможностью реализации операции сшивания границ сплайсинга, которая предназначена для точного, например, аккуратного, более или менее точного определения наиболее вероятного положения в риде, где был пропущен интрон. Результат представляет собой позицию сшивания между двумя основаниями рида таким образом, что основания слева от точки сшивания выравниваются с первым сегментом экзона, а основания справа от точки сшивания выравниваются со вторым сегментом экзона. Это может быть выполнено посредством 1) тестирования множества возможных положений сшивания и 2) оценки результатов тестирования. Такая оценка может быть основана на множестве различных критериев, например, на количестве наблюдаемых несовпадений оснований и/или отсутствии или присутствии и типе каноничных мотивов интрона, наблюдаемых на двух концах предполагаемого протяжения интрона в референсе, например, в соответствии с данным положением сшивания.

[00408] Операция сшивания сплайсов может быть сконфигурирована как процедура предварительной обработки для выравниваний со сплайсингом, которые в обратном случае будут ресурсоемкими и/или дорогими в выполнении, если операцию необходимо выполнить таким образом, который рассматривает все положения сшивания. Соответственно, модуль выравнивания может принимать картирование со сплайсингом для рида полученных из РНК геномных данных, причем картирование со сплайсингом включает по меньшей мере первую часть и вторую часть рида полученных из РНК геномных данных и по меньшей мере первый сегмент и второй сегмент референсного генома.

[00409] Кроме того, модуль выравнивания может быть выполнен с возможностью реализации операции сшивания сплайсов таким образом, чтобы определить наилучшее положение сшивания в пределах рида полученных из РНК геномных данных. Такое наилучшее положение сшивания может быть реализовано посредством оптимизации множества факторов сшивания, которые относятся к каждому рассматриваемому положению сшивания. Факторы сшивания могут включать степень совпадения между первой частью рида, например, с согласованием по длине с концом рассматриваемого положения сшивания, и первым сегментом референсного генома с идентичным согласованием по длине. Факторы сшивания могут также включать степень совпадения

второй частью рида с согласованием по длине с началом рассматриваемого положения сшивания, и вторым сегментом референсного генома с идентичным согласованием по длине. В данном случае факторы сшивания могут также содержать вероятность мотива интрона, соответствующего рассматриваемому положению сшивания, например так, что мотив интрона включает по меньшей мере два референсных основания, смежных с согласованным по длине первым сегментом референса, и по меньшей мере два референсных основания, смежных с согласованным по длине вторым сегментом референса.

[00410] В различных примерах факторы сшивания могут быть объединены в оценку и рассматриваемое положение сшивания с наилучшей в числовом отношении оценкой может быть определено как наилучшее положение сшивания. В некоторых примерах переход может быть сделан от первого рассматриваемого положения сшивания ко второму рассматриваемому положению сшивания через по меньшей мере один являющийся помехой нуклеотид в риде полученных из РНК геномных данных. В таком примере оценка второго рассматриваемого положения сшивания может быть вычислена частично посредством регулировки оценки для первого рассматриваемого положения сшивания для учета любой разницы между тем, как хорошо по меньшей мере один являющийся помехой нуклеотид совпадает с первым сегментом референсного генома и как хорошо по меньшей мере один являющийся помехой нуклеотид совпадает со вторым сегментом референсного генома. В некоторых примерах наилучшее положение сшивания может быть передано в модуль беспромежуточного выравнивания, который может быть выполнен с возможностью определения наилучшего беспромежуточного выравнивания рида полученных из РНК геномных данных с объединением полученные из РНК двух согласованных по длине сегментов референсного генома. Наилучшее положение сшивания затем может быть передано в модуль выравнивания с промежутком, который затем может определять наилучшее беспромежуточное выравнивание или выравнивание с промежутком рида полученных из РНК геномных данных с объединением полученные из РНК двух согласованных по длине сегментов референсного генома.

[00411] Модуль выравнивания может также включать модуль, выполненный с возможностью реализации выравнивания рида со сплайсингом. Например, при выполнении выравнивания рида со сплайсингом может быть сгенерирована последовательность множества частичных картирований, например, цепочек затравок, например, последовательных цепочек затравок в скаффолде, в риде РНК. Множество частичных картирований может представлять множество сегментов экзона вероятных картирований со сплайсингом, которые могут пропустить один или более возможных интронов в референсной последовательности (например, подвергнувшись сшиванию границ сплайсинга для каждого интрона). В таких примерах операция выравнивания со сплайсингом может быть выполнена посредством выравнивания рида (ридов) с объединением множества сегментов экзона в референсе.

[00412] В частности, модуль выравнивания может быть обеспечен и выполнен с возможностью приема картирования со сплайсингом для одного или более ридов полученных из РНК геномных данных. Картирование со сплайсингом может включать определение последовательности множества прилегающих частей между ридами полученных из РНК геномных данных и одной или более соответствующими последовательностями, имеющими один или более, например, множество, сегментов референсного генома. Модуль выравнивания затем может реализовывать операцию выравнивания со сплайсингом над указанными последовательностями для определения

наилучшего выравнивания со сплайсингом рида полученных из РНК геномных данных с последовательностью множества сегментов референсного генома. Например, операция выравнивания со сплайсингом может содержать объединение множества сегментов референсного генома в объединенную референсную последовательность, в которой каждый из множества сегментов референсного генома соединен в местах соединений объединения. Кроме того, последовательность рида, имеющая по меньшей мере множество прилегающих частей рида полученных из РНК геномных данных, может быть сгенерирована и соединена в таких местах соединения объединения.

[00413] Наилучшее выравнивание последовательности рида может быть вычислено в отношении объединенной референсной последовательности, например, когда наилучшее выравнивание последовательности ограничено таким образом, что места соединения объединения в последовательности рида выровнены с соответствующими местами соединения объединения в объединенной референсной последовательности. Выравнивание последовательности может быть отредактировано в выравнивание со сплайсингом, причем указанное редактирование может включать вставку дескрипторов интронов в положения выравнивания, соответствующие местам соединения объединения, например так, что дескрипторы интронов кодируют длину интрона, которая может быть равна соответствующим расстояниям между сегментами референсного генома. Модуль выравнивания затем может выдавать выравнивание со сплайсингом.

[00414] В таком примере наилучшее выравнивание последовательности может быть определено как выравнивание с величиной наилучшей в числовом отношении оценки среди вычисленных оценок для всех вероятных выравниваний, причем оценку каждого вероятного выравнивания вычисляют так, чтобы включать штрафы несовпадения для каждого нуклеотида последовательности рида, который не совпал с выровненным нуклеотидом объединенной референсной последовательности. Оценка вероятного выравнивания также может быть вычислена таким образом, чтобы включать один или более других штрафов, таким как штраф индела для каждой инсерции или делеции в вероятном выравнивании и/или штраф сплайсинга для каждого места соединения объединения, включенного в вероятное выравнивание, например так, что штраф сплайсинга определяют по меньшей мере частично в соответствии с мотивом интрона, связанным с соединением сегментов референсного генома. В различных примерах объединенная референсная последовательность может быть выполнена так, чтобы включать ложные основания, например, в местах соединения объединения, причем ложные основания могут быть использованы для переноса значений соответствующих штрафов сплайсинга.

[00415] В данных примерах вычисление наилучшего выравнивания последовательности может включать динамическое программирование для вычисления оценок ячеек для двумерной матрицы ячеек оценок, указанные два измерения соответствуют последовательности рида и объединенной референсной последовательности. В таких примерах каждой из множества прилегающих частей рида может быть присвоен ID зоны; и, в дополнение, каждому из множества сегментов референсного генома может быть присвоен ID зоны. Такой ID зоны может быть сконфигурирован так, чтобы быть аналогичным каждой соответствующей части рида и сегмента референса, например так, что когда каждая ячейка оценки в матрице ячеек оценки имеет ID зоны рида ячейки, который аналогичен ID зоны соответствующей части рида. Кроме того, ID зоны референса ячейки может быть сконфигурирован так, чтобы быть аналогичным ID зоны соответствующего сегмента референса, и наилучшее выравнивание последовательности может быть ограничено так, чтобы проходить

только через ячейки оценки, ID зоны рида ячейки которых аналогичны их ID зоны референса ячейки.

[00416] На ФИГ. 5 показан абстрактный прямоугольник выравнивания, в котором объединенная рассматриваемая последовательность проходит по вертикальной оси, а объединенная референсная последовательность проходит по горизонтальной оси. Ложные основания каждой объединенной последовательности затенены (зоны 3 и 6). На прямоугольник выравнивания наложена сетка для того, чтобы показать границы между зонами на каждой оси. Вспомогательные прямоугольники с совпадающими ID зон являются действительными областями выравнивания, а другие (затененные) вспомогательные прямоугольники являются недействительными областями выравнивания. Показан пример действительного выравнивания, которое проходит от конца до конца в рассматриваемой последовательности и содержит инсерцию (вертикальный сегмент) во втором сегменте экзона (зона 4) и делецию (горизонтальный сегмент) в третьем сегменте экзона (зона 7). Действительное выравнивание проходит диагонально через границы сплайсинга (зоны 3 и 6).

[00417] Беспромежуточное выравнивание или выравнивание с промежутком, с использованием объединенный рассматриваемой и референсной последовательности обеспечивает получение корректных оценок выравнивания, но след выравнивания (например, строка CIGAR) требует редактирования, так как он еще не включает интронные (‘N’) операции. Например, выравнивание со сплайсингом рида из 100 оснований без инделов может происходить из выравнивания со “101M” CIGAR, что означает, что 101 основание выровнено по диагонали без инделов. В указанной CIGAR необходимы две регулировки. Во-первых, ложное основание между сегментами экзона учитывается в CIGAR, но не должно. Во-вторых, интронная операция, например, длинной 895 оснований, должна быть вставлена в положении ложного основания. Верная CIGAR может быть, например, «40M895N60M».

[00418] Если принять, что скаффолд, содержащий цепочки затравок, определяющие конечные точки сегментов экзона, сшит, то посредством прямой арифметики можно обнаружить положение каждой границы сплайсинга в CIGAR, удалить «1M» для ложного основания и заместить его интронной операцией надлежащей длины. Для локальных выравниваний этот процесс может учитывать возможность того, что один или боле из всех сегментов экзона были вырезаны из выравнивания. Тем же арифметическим процессом можно вычислять корректные начальное и конечное положения выравнивания в референсном геноме.

[00419] После получения оценок выравнивания, начального и конечного положения и строк CIGAR для каждого выровненного скаффолда, обработка для выбора и выдачи наилучшего выравнивания, возможно, со сплайсингом, является идентичным обработки ДНК. Парные концевые вероятные выравнивания изучают, чтобы найти надлежащим образом расположенные и ориентированные пары выравнивания. Вероятным парам выравнивания, включая не надлежащим образом образованные вероятные пары, присваивают штрафы оценки за то, что они не имеют пары или имеют маловероятные эмпирические длины вставки; оценки пар сформированы посредством объединения (например, добавления) оценок выравнивания их каждого сопряжения и штрафа на образование пары; и пару выравниваний с наилучшей оценкой выбирают и обеспечивают выдачу из модуля выравнивания.

[00420] Очевидная длина вставки, обычно измеренная как охват в референсе, покрываемом выравниваниями двух сопряженных ридов, может оказаться очень длинной из-за интронов в одном или обоих сопряженных ридов или ненаблюдаемых

интронов в промежутке между сопряженными ридами. (Физические инсерции потенциально намного короче и имеют длину секвенированной молекулы РНК или кДНК, в которых интроны удалены посредством сплайсинга.) Таким образом, значительно более длинные очевидные длины вставок должны рассматриваться как образующие пару надлежащим образом и им должна быть присвоены нулевые или малые штрафы на формирование пар; это может быть выполнено в соответствии с распределением известных длин интронов в видах, от которых получен образец, и/или распределением наблюдаемых очевидных вставок в обрабатываемых РНК-секвенированных данных.

[00421] В одном варианте реализации для каждого обрабатываемого рида обеспечивают выдачу оценки выравнивания, начального положения и кодированной строки CIGAR из модуля выравнивания. Кроме того, в другом варианте реализации для каждой границы сплайсинга в выравнивании обеспечивают выдачу мотива интрона и статуса аннотирования. Качество или достоверность картирования, например, параметр «MAPQ» по шкале phred также может быть оценен и выдан. В предпочтительном варианте реализации MAPQ предварительно оценивают в пропорции к разности между наилучшей оценкой пары и второй наилучшей оценкой пары с различным выравниванием для текущего рида. Дополнительные вероятные выравнивания, или вторичные выравнивания, могут также быть выданы для каждого рида, например, ограниченное число других вероятных вариантов, оценка которых находится в пределах определенного и сконфигурированного порогового значения разности оценок.

[00422] Следует понимать, что касается вышеизложенного, то хотя в некоторых случаях была описана функция картирования, например, как картировщик, и/или в некоторых случаях была описана функция выравнивания, например, как выравниватель, эти различные функции могут выполняться последовательно одной и той же архитектурой, которую в данной области техники обычно называют выравнивателем. Соответственно, в различных случаях функция картирования и функция выравнивания, как описано в настоящем документе, могут выполняться общей архитектурой, которую можно понимать как выравниватель, особенно в тех случаях, когда для выполнения функции выравнивания сначала необходимо выполнить функцию картирования.

[00423] Выходными данными из модуля выравнивания является файл SAM (текст) или BAM (например, двоичная версия файла SAM) вместе с оценкой качества картирования (MAPQ), которая отражает достоверность того, что спрогнозированное и выровненное местоположение рида относительно референса действительно то самое, откуда получен рид. Соответственно, после того, как определено, где каждый рид картирован, и также определено, где каждый рид выровнен, например, каждому соответствующему риду даны положение и оценка качества, отражающая вероятность того, что это положение является правильным выравниванием, так что нуклеотидная последовательность для ДНК субъекта известна, как и то, как ДНК субъекта отличается от референса (например, определена строка CIGAR), различные риды, представляющие геномную последовательность нуклеиновых кислот субъекта могут быть отсортированы по местоположению в хромосоме так, чтобы можно было определить точное местоположение рида на хромосомах. Поэтому в соответствии с некоторыми аспектами настоящее изобретение относится к функции сортировки, например, которая может быть выполнена модулем сортировки, который может быть частью конвейера модулей, например, конвейера, предназначенного для приема необработанных данных рида последовательности, например, в виде геномного образца индивида, и картирования

и/или выравнивания этих данных, которые могут быть затем отсортированы.

[00424] В частности, после того, как ридам присвоены положения, например, относительно референсного генома, что может включать в себя определение того, какой хромосоме принадлежит рид, и/или его смещения от начала этой хромосомы, риды могут быть отсортированы по положению. Сортировка может быть полезна, например, при последующих анализах, так как при помощи нее все риды, которые перекрывают данное положение в геноме, могут быть сформированы в скопление так, чтобы находится друг возле друга, например, после обработки модулем сортировки, в результате чего можно легко определить, согласуются ли большинство ридов с референсным значением, или нет. Таким образом, если большинство ридов не согласуются с референсным значением, определение варианта может быть помечено. Следовательно, сортировка может включать в себя одну или более сортировок ридов, которые выровнены относительно одной и той же позиции, например, одной и той же позиции хромосомы, для создания скопления, чтобы все риды, которые покрывают одно и то же местоположение, были физически сгруппированы вместе; и может также включать в себя анализ ридов в скоплении для определения того, где риды могут указывать фактический вариант в геноме по сравнению с референсным геномом, причем этот вариант можно отличить, например, с помощью согласования скопления, от ошибки, такой как ошибка считывания машиной или ошибка в способах секвенирования, которая может проявляться малой частью ридов.

[00425] После того, как данные получены, имеются один или более других модулей, с помощью которых можно очистить эти данные. Например, один модуль, который может быть включен, например, в конвейер анализа последовательностей, такой как для определения геномной последовательности индивида, может быть модулем локального повторного выравнивания. Например, часто трудно определить инсерции и делеции, которые возникают в конце рида. Причина в том, что алгоритм Смита-Ватермана или аналогичный процесс выравнивания испытывает недостаток контекста за пределами индела, чтобы можно было выполнить оценку для обнаружения его присутствия. Поэтому фактический индел может быть указан в отчете как один или более ОНП. В таком случае точность спрогнозированного местоположения для любого данного рида может быть улучшена за счет выполнения локального повторного выравнивания на картированных, и/или выровненных, и/или сортированных данных рида.

[00426] В таких случаях могут быть использованы скопления для помощи в выяснении надлежащего выравнивания, например, когда рассматриваемое положение находится в конце любого данного рида, это же положение, вероятно, будет в средине некоторого другого рида в данном скоплении. Соответственно, при выполнении локального повторного выравнивания могут быть проанализированы различные риды в скоплении, чтобы определить, указывают ли некоторые риды в скоплении на наличие инсерции или делеции в данном положении, где другой рид не содержит индела или, скорее, имеет замену в этом положении, тогда можно вставить индел, например, в референс, где он не присутствует, и можно повторно выровнять риды в локальном скоплении, которое перекрывает эту область, чтобы посмотреть, будет ли при этом достигнута более хорошая совокупная оценка, чем в том случае, когда там не было инсерции и/или делеции. Соответственно, если улучшение имеется, весь набор ридов в скоплении может быть пересмотрен, и если оценка всего набора улучшилась, то становится понятно, что в этой позиции действительно был индел. Подобным образом можно компенсировать отсутствие достаточного контекста для более точного выравнивания рида в конце

хромосомы для любого отдельного рида. Поэтому при выполнении локального повторного выравнивания исследуют одно или более скоплений, где могут находиться один или более инделов, и определяют, можно ли улучшить общую оценку выравнивания путем добавления индела в любое данное положение.

[00427] Другой модуль, который может быть включен, например, в конвейер анализа последовательностей, такой как для определения геномной последовательности индивида, может быть модулем маркировки дубликатов. Например, функция маркировки дубликатов может быть реализована для компенсации ошибок химии, которые могут возникать во время фазы секвенирования. Например, как описано выше, во время некоторых процедур секвенирования последовательности нуклеиновых кислот прикрепляют к бусинам и наращивают на них с помощью меченных нуклеотидных оснований. В идеале получится по одному риду на бусину. Однако иногда к одной бусине прикрепляется множество ридов, что приводит к избыточному количеству копий прикрепленных ридов. Это явление известно как дупликация ридов.

[00428] Такая дупликация ридов может сбивать статистику и создавать статистическое отклонение, так как вместо того, чтобы иметь эквивалентное представление всех ридов, различные риды были дублированы, например, вследствие того, что дублирующие шаблонные последовательности, прикрепленные более, чем к одной бусине, представлены в избытке. Соответственно, это может быть определено, так как любой рид, который выравнивается с точно таким же положением и имеет точно такую же длину, скорее всего является дубликатом. После того, как это определено системой, только один рид необходимо подвергнуть дальнейшей обработке, а другие могут быть помечены как дубликаты и, следовательно, могут быть отброшены или проигнорированы. Типичной ситуацией, в которой это происходит, является то, когда с самого начала отсутствует достаточное количество генетического материала для обработки и система совершает попытки компенсировать это в избыточной степени.

[00429] Другой модуль, который может быть включен, например, в конвейер анализа последовательностей, такой как для определения геномной последовательности индивида, может быть перекалибровщиком оценки качества оснований. Например, каждое основание каждого рида имеет оценку по шкале Phred, которая указывает на вероятность того, что определенное основание в указанном положении является неверным. Например, оценка по шкале Phred для любого основания частично обусловлена природой основания, которая ему предшествует, и профиль ошибки будет отличаться в зависимости от того, какое основание предшествует рассматриваемому основанию. Кроме того, велика вероятность ошибка, возникающей на концах рида, например такой, когда на концах рида химия начинает терять свои качества. Перекалибровка оценки качества оснований является ковариантом анализа, который может возвращаться и измерять эмпирические качества оценки качества основания как функции всех тех вещей, вследствие которых она изменяется.

[00430] В различных примерах это включает два прохода, на первом собирают все фактические, эмпирически измеренные данные и статистику частоты наблюдаемых ошибок как функции всех переменных, а второй проход включает фактическую перекалибровку оценок посредством проведения всех ридов через фильтр, модифицирующий оценки качества для каждого отдельного основания как функцию переменных на основании того, что было в действительности измерено эмпирическим путем в множестве данных. Это компенсирует все различия в данных вследствие различных переменных и очищает все данные и оценки. Назначение всех указанных очисток заключается в обеспечении достижения наилучшего определения вариантов.

Множество определителей вариантов основывают свои решения частично на сообщенном качестве каждого нуклеотида, которые скапливаются в каждом положении в геноме. Если оценки качестве не точны, в результате может быть легко определено неверное основание.

[00431] Другой модуль, который может быть включен, например, в конвейер анализа последовательностей, такой как для определения геномной последовательности индивида, может быть модулем сжатия, который выполняет функцию сжатия. Как указано выше, может быть полезным на некотором этапе взять сгенерированные и обработанные данные и передать их в удаленное местоположение, например, облако, и поэтому данные может быть необходимо сжать на некотором этапе обработки, так чтобы после их сжатия они могли быть переданы и/или другим образом загружены, например, в облако или на группу серверов и т.д., например, для реализации модуля определения оснований. Полученные результаты затем могут быть распакованы и/или сохранены в памяти, в базе данных в облаке, например, базе данных здоровья в электронном виде и/или исследовательской базе данных и тому подобном, которая, в свою очередь, может быть доступной для третичной обработки, и т.д.

[00432] В частности, после того, как генетические данные сформированы и/или обработаны, например, в одном или более протоколах первичной и/или вторичной обработки, например, картированы, выровнены и/или отсортированы, например, для создания одного или более файлов определения вариантов, например, для определения того, как данные генетической последовательности субъекта отличаются от одной или более референсных последовательностей, согласно другому аспекту настоящее изобретение может относиться к выполнению одной или более других аналитических функций над сформированными и/или обработанными генетическими данными, например, для дальнейшей обработки, такой как третичная обработка. Например, система может быть выполнена с возможностью выполнения дальнейшей обработки сгенерированных и/или обрабатываемых во вторую очередь данных, например посредством их прогона через один или более конвейер для третичной обработки, такой как один или более из конвейера генома, конвейера эпигенома, конвейера метагенома, совместного генотипирования, конвейера MuTect2 или другого конвейера для третичной обработки, например, посредством устройств и способов, раскрытых в настоящем документе. Например, в различных примерах может быть обеспечен дополнительный уровень обработки, например для диагностики заболевания, терапевтического лечения и/или профилактического предупреждения, включая, например, неинвазивное пренатальное тестирование (NIPT), реанимацию и интенсивную терапию новорожденных (NICU), рак, проводимые в лаборатории исследования (LDT), агробиологию (AgBio) и другие виды диагностики, профилактики и/или способов лечения таких заболеваний, в которых применяются данные, сгенерированные одним или более из указанных первичных, вторичных и/или третичных конвейеров. Следовательно, устройства и способы, описанные в настоящем документе, могут быть использованы для формирования данных генетических последовательностей, которые затем могут быть использованы для формирования одного или более файлов определения вариантов и/или другой связанной информации, которая может быть в дальнейшем подвергнута обработке другими конвейерами третичной обработки в соответствии с устройствами и способами, описанными в настоящем документе, например, для диагностики конкретных и/или общих заболеваний, а также для профилактических и/или терапевтических мер и/или методов воздействия на развитие.

[00433] Соответственно, как указано выше в настоящем документе, согласно

различным аспектам, настоящее изобретение относится к системам, устройствам и способам реализации геномных и/или биоинформационных протоколов, таких как, в различных примерах, для выполнения одной или более функций для анализа генетических данных на интегральной схеме, например, на реализованной на платформе аппаратной обработки. Например, согласно одному аспекту, обеспечена биоинформационная система, которая может включать реализацию различных биоинформационных функций, которые были оптимизированы для более быстрой реализации и/или для реализации с повышенной точностью в аппаратном исполнении. Соответственно, в различных примерах способы и системы, раскрытые в настоящем документе, могут включать в себя реализацию одного или более алгоритмов для исполнения указанных функций, причем указанные алгоритмы могут быть реализованы в виде аппаратного решения, например, так что алгоритмы оптимизированы таким образом, чтобы осуществляться посредством интегральной схемы, образованной на одной или более жестко смонтированных цифровых логических схем. В таком примере жестко смонтированные цифровые логические схемы могут быть взаимно соединены, например, посредством одного или множества физических электрических межсоединений, и могут быть выполнены с возможностью функционирования как один или более модулей обработки. В различных примерах обеспечено множество жестко смонтированных цифровых логических схем, которые сконфигурированы в виде набора модулей обработки, причем каждый модуль обработки выполнен с возможностью реализации одного или более этапов в биоинформационном протоколе генетического анализа.

[00434] В частности, согласно одному аспекту предложена система для осуществления конвейера анализа последовательностей, например на данных генетической последовательности. Система может включать в себя один или более электронных источников данных, память и интегральную схему. Например, в одном варианте реализации имеется электронный источник данных, который может быть выполнен с возможностью обеспечения одного или более цифровых сигналов, например, цифровых сигналов, представляющих один или более ридов генетических данных, например, где каждый рид геномных данных включает последовательность нуклеотидов. Кроме того, память может быть выполнена с возможностью хранения одной или более генетических референсных последовательностей и может быть также выполнена с возможностью хранения индекса, такого как индекс одной или более генетических референсных последовательностей.

[00435] Также в различных примерах одно или более из множества физических электрических межсоединений могут содержать вход, например, в интегральную схему, и могут также быть соединены с источником электронных данных для обеспечения возможности приема одного или более ридов геномных данных. В различных вариантах реализации жестко смонтированные цифровые логические схемы могут быть выполнены в виде набора модулей обработки, такого где каждый модуль обработки может быть сформирован подмножеством жестко смонтированных цифровых логических схем и может быть выполнен с возможностью реализации одного или более этапов в конвейере анализа последовательностей, например, на оцифрованных генетических данных, например на множестве ридов геномных данных. В таких примерах каждое подмножество жестко смонтированных цифровых логических схем может быть сконфигурировано с соединением при помощи разводки, так чтобы выполнять один или более этапов в конвейере анализа последовательности, например, где один или более этапов могут включать в себя выполнение одной или более из операции

определения оснований и/или исправления ошибки, например, в оцифрованных генетических данных, и/или может включать в себя одно или более из выполнения функции картирования, выравнивания и/или сортировки над генетическими данными. В определенных случаях конвейер может включать в себя выполнение одного или более из повторного выравнивания, удаления дубликатов, перекалибровки оценки качества основания, редукции и/или сжатия и/или распаковки на оцифрованных генетических данных. В определенных случаях конвейер может включать в себя выполнение операции определения вариантов над генетическими данными.

[00436] Таким образом, в различных вариантах реализации, системах, аппаратах и методах для реализации протоколов геномики и/или биоинформатики, о которых здесь говорится, может быть задействовано использование процессов, которые обычно могут быть выполнены с помощью программного обеспечения, и встраивание указанных функций в интегральную схему, такую как микросхема (чип) 100, например, в качестве составной части системной платы 105, такой, в которой указанные функции оптимизированы, чтобы повысить ее производительность на микросхеме. В связи с этим, в одном из вариантов реализации, как можно видеть на фиг. 6 и 7, предусмотрена микросхема 100, причем микросхема 100 спроектирована так, чтобы эффективно выполнять указанные функции программного конвейера. В разных конкретных вариантах реализации микросхема 100 может представлять собой программируемую пользователем вентильную матрицу (FGPA), интегральную схему специального назначения (ASIC), или структурную интегральную схему специального назначения (sASIC), или тому подобное.

[00437] Так, например, выполнение одного или более алгоритмов может быть встроено в микросхему, такую как FGPA или ASIC или структурную микросхему ASIC, и может быть оптимизировано для более эффективной их работы благодаря реализации аппаратными средствами. Соответственно, в одном из вариантов реализации предусмотрена микросхема FGPA, где такая микросхема может являться конфигурируемой, например, программная часть микросхемы может быть изменена так, чтобы быть сделать ее более адаптируемой к конкретным потребностям пользователя в части выполнения различных геномных функций, подробно описанных в настоящем документе. В этом случае пользователь может изменить или модифицировать задействованные алгоритмы в зависимости от ключевых параметров, обозначаемых для системы в целом, например, для придания дополнительной функциональности или замены того, что было перед этим на микросхеме, например, для перенастраивания конфигурации микросхемы для того, чтобы задействовать другой алгоритм.

[00438] Далее, в другом варианте реализации предусмотрена микросхема FGPA или структурная микросхема ASIC, где такая микросхема является конфигурируемой как полностью, так и частично, например, некоторая ее программная часть может быть изменена таким образом, чтобы сделать ее более адаптируемой к конкретным потребностям пользователя в части выполнения различных геномных функций, подробно описанных в настоящем документе. В соответствии с другим вариантом реализации, предусмотрена ASIC, например, преобразованная в ASIC-микросхему FGPA или sASIC, функциональность которой может быть заблокирована в микросхеме. В таком случае различные параметры, как, например, различные параметры, относящиеся к конкретной функции одного или более алгоритмов, изложенных в настоящем документе, могут быть выбираемы пользователем, например, управляющие тем, как следует функционировать различным модулям, и при этом конкретный механизм

функционирования таких модулей заблокирован.

[00439] В различных вариантах реализации, как показано на фиг. 6 и 7, микросхема 100 может являться частью системной платы, такой как часть платы расширения 104, например, платы межсоединения периферийных компонентов (PCI), включая плату PCIe, которая в различных вариантах реализации может быть объединена, например, сопряжена с возможностью передачи сигнала, например, подключена электрически, с автоматизированным устройством секвенирования таким образом, чтобы функционировать в качестве неотъемлемой части секвенатора, так чтобы файлы данных, например, файлы в формате FASTQ, генерируемые секвенатором, передавались прямо на микросхему для вторичной обработки генома, как непосредственно следующей за генерацией файла в формате FASTQ и/или первичной обработкой, например, непосредственно после того, как выполнена функция секвенирования.

[00440] Таким образом, в некоторых случаях предусмотрена плата PCI 104, где такая плата PCI может включать в себя микросхему с шиной PCIe 105, в которой плата 102 и/или микросхема 100 может включать в себя один или более модулей управления конфигурацией, таких как контроллер конфигурации (Cent-Com); устройство прямого доступа к памяти (например, драйвер); программный интерфейс приложения (API); интерфейс уровня клиента (CLI); библиотека; блок памяти, например, оперативной памяти (RAM) или динамической оперативной памяти (DRAM); и/или межсоединение на уровне микросхемы, такое как DDR3. Так, например, в различных случаях, в состав может входить модуль управления конфигурацией, где такой модуль управления конфигурацией управляется, например, файлом параметров. В таком случае модуль управления конфигурацией может быть адаптирован таким образом, чтобы настраивать конфигурацию различных модулей программного конвейера. В различных случаях он может быть редактируемым, позволяя пользователю определять какие модули программного конвейера будут использоваться, например, все или подгруппа модулей в количестве меньшем, чем все, например, для некоторого набора данных, такого как некоторый набор файлов в формате FASTQ.

[00441] Например, в различных вариантах реализации функционирование программного конвейера является конфигурируемым в значительной степени, так что один или более модулей, например, встроенных в микросхему, могут быть запущены или не запущены, если потребуется. Далее, конфигурация каждого используемого модуля также может быть настроена для запуска в соответствии с одним или более предварительно выбранных параметров, которые могут контролироваться пользователем, например, относящихся к тому, как данный модуль будет выполняться и функционировать. Таким образом, может присутствовать два разных набора файлов конфигурации, таких что, например, один управляет базовыми операциями системы в целом и может быть скрыт от пользователя, а другой управляется пользователем и, таким образом, позволяет выбирать различные параметры, которые будут запускаться одной или более подсистемами, например, модулем, микросхемы 100 и/или платы PCI 104.

[00442] Таким образом, различные модули, из числа описанных выше, могут быть соединены на постоянной основе с микросхемой или являться внешними по отношению к микросхеме, но находиться с последней в сопряженном состоянии, например, на плате PCI 104, или быть удаленными по отношению к микросхеме, например, на другой плате PCI или даже другом сервере, например, на сервере, доступном через облако 30. Например, в некоторых реализациях, один или более из числа описанных выше модулей, могут быть соединены на постоянной основе с микросхемой 100, а сама микросхема

установлена на печатной плате 104 автономного устройства 300, или сопряжены с секвенатором, и при этом пользователь настраивает конфигурацию и запускает систему самостоятельно в соответствии с собственными предварительно выбранными параметрами. В альтернативном варианте, как указано в настоящем документе, один или более из числа описанных выше модулей могут находиться в некоторой системе, доступной через облако 30, в котором управление функционированием программного конвейера и/или соответствующих модулей может включать в себя регистрацию пользователя на сервере, например, на удаленном сервере, и передачу данных в прямом и обратном направлении, и, таким образом, выбор модулей, которые следует запускать для набора данных. В некоторых случаях один или более модулей могут выполняться удаленно, например, через доступный в облаке сервер.

[00443] В некоторых случаях при настройке конфигурации системы, микросхема, такая как микросхема 100 на плате расширения 104, такой как плата PCI, может входить в состав сервера 300, так что различные приложения системы запускаются с этого сервера. В некоторых случаях сервер 300 может содержать подключаемый терминал с предоставляемым пользователю оконным интерфейсом, так что пользователь может выбирать модули для запуска и параметры, по которым они запускаются, например, выбирая один из блоков в меню блоков. В других случаях, однако, файл параметров может представлять собой текстовый файл, подробно описывающий категории модулей под именами файлов, которые пользователь может впоследствии редактировать для того, чтобы выбирать какие модули будут запускаться и в соответствии с какими параметрами. Например, в различных вариантах реализации каждая микросхема может содержать все модули или выбираемый набор модулей, например, один или более из следующих: модуль поиска оснований, корректировки ошибок, картрирования, выравнивания, сортировки, локального перевыравнивания, маркировки дубликатов, повторной калибровки, определения вариантов, модуль сжатия и/или модуль распаковки, из которых пользователь может выбирать, какие модули будут запускаться, когда и в каких пределах они будут запускаться, без изменения функционирования лежащих в основе алгоритмов, по которым отдельные модули работают.

[00444] Кроме того, а различных случаях, в состав могут входить устройство прямого доступа к памяти (DMA) в микросхеме и драйвер DMA, причем драйвер DMA содержит код, запускаемый в ядре. Таким образом, драйвер DMA может лежать в основании целой операционной системы. Например, виртуальное пользовательское пространство может располагаться на уровень выше пространства литеральной адресации, в котором ядро запускает код. Такое программное обеспечение операционной системы, следовательно, работает между этими уровнями, управляя преобразованием данных из виртуального пространства в физическое. В частности, ядро представляет собой самый нижний уровень кода, предоставляющего платформе доступ к PCI 104, например PCIe, шине 105, с которой сопряжена микросхема 100. Таким образом, поскольку в различных вариантах реализации конфигурация микросхемы 100 может быть настроена как в плате расширения 104 с шиной расширения PCIe 105, и такая плата 104 может быть сопряжена с различными аппаратными средствами некоторого устройства, например, секвенатора, и драйвер DMA может функционировать таким образом, чтобы осуществлять обмен информацией с аппаратными средствами секвенатора, и его конфигурация в дальнейшем может быть настроена для запуска на уровне ядра на ЦП 100, чтобы также осуществлять обмен информацией с устройством DMA в микросхеме 100, и/или работы в виртуальном пользовательском пространстве, так чтобы получать команды от пользователя.

[00445] Для того, чтобы обеспечить такой обмен сообщениями на микросхеме и/или между микросхемой и одной или более плат, каждый отдельный конфигурируемый параметр модуля может быть приписан к некоторому адресу в регистре. В таком случае, плата может иметь собственное адресное пространство, причем такое адресное пространство может отличаться от адресного пространства для одного или более блоков памяти, например, 64 Гигабайт памяти, и/или каждый модуль может дополнительно содержать регистры и связанную с ним локальную память - каждую с собственным адресным пространством. Таким образом, драйвер знает, где все находится, все адреса, и знает, как осуществлять обмен информацией между микросхемой 100, платой PCI 104, и/или аппаратными средствами сервера. Далее, зная расположение всех адресов и осуществляя обмен информацией с API, драйвер может считывать созданный пользователем файл параметров и отыскивать тот параметр, в котором файл фактически находится в системе главного компьютера, и будет считывать и интерпретировать значение, записанное в данном файле, и доставлять это значение в надлежащий регистр в надлежащее место на микросхеме. Таким образом, драйвер может управлять доставкой команд, связанных с выбранным параметром, например, таких, которые имеют отношение к выбранной пользователем конфигурации, и доставлять такие данные на микросхему через устройство DMA для конфигурации ее рабочих функций.

[00446] В частности, после того, как генетические данные сгенерированы и/или обработаны, например, с использованием одного из первичных и/или вторичных протоколов, например, картрированы, выровнены, и/или отсортированы, например, для создания одного или более файлов определения вариантов, например, для определения того, как данные генетической последовательности исследуемого объекта отличаются от одной или более контрольных последовательностей, и следующий аспект изобретения может быть связан с выполнением одной или более аналитических функций в отношении сгенерированных и/или обработанных генетических данных, таких как дальнейшая, например, третичная, обработка. Например, конфигурация системы, представленной на фиг. 8-11, может быть настроена для последующей обработки сгенерированных данных и/или данных, полученных в результате вторичной обработки, например, для ее запуска через один или более программных конвейеров третичной обработки 700, например, через один или более программных конвейеров генома, программных конвейеров эпигенома, программных конвейеров метагенома, совместного генотипирования, программных конвейеров MuTect2 или других программных конвейеров третичной обработки, например, с помощью устройств и методов, описанных в настоящем документе. Например, в различных случаях может обеспечиваться дополнительный уровень обработки 122, например, для диагностики заболеваний, терапевтического лечения и/или профилактических действий, например, включающий в себя NIPT, NICU, Cancer, LDT, AgBio и другие подобные методы диагностики заболеваний, профилактики, и/или лечения, в которых задействованы данные, создаваемые одним или более таких первичных и/или вторичных и/или третичных программных конвейеров.

[00447] Таким образом, описанные в настоящем документе устройства и методы могут использоваться для генерации данных генетической последовательности, которые могут в дальнейшем использоваться для создания одного или более файлов определения вариантов и/или других, связанных с этим данных, которые в дальнейшем могут являться объектом исследования для других запускаемых программных конвейеров третичной обработки в соответствии с описанными в настоящем документе устройствами и

методами, например, как для диагностики конкретного и/или системного заболевания, так и для профилактики и/или терапевтического лечения и/или методик, находящихся в стадии разработки.

[00448] Далее, в различных случаях, в состав может входить API, где конфигурация такого API настроена таким образом, чтобы в нем содержался список обращений к функциям, которые может осуществлять пользователь, так чтобы конфигурировать и управлять системой. Например, API может быть определен в файле заголовка, который описывает функциональные возможности и задает способ, которым вызывается какая-либо функция, например, передаваемые параметры, входные и выходные данные, что входит, что выходит, и что возвращается. Например, в различных вариантах реализации, один или более элементов программного конвейера могут быть конфигурируемыми, например командами, которые вводит пользователь и/или одним или более сторонними приложениями. Такие команды могут посылаться на микросхему через API, который осуществляет обмен информацией с драйвером, указывая драйверу какие составляющие, например, модули, следует активировать, когда и в каком порядке, с учетом предварительно выбранных параметров конфигурации.

[00449] Как указано выше, драйвер DMA запускается на уровне ядра и имеет собственный простой API очень низкого уровня, который обеспечивает доступ к аппаратным средствам и функциям для того, чтобы получить доступ к применяемым регистрам и модулям. Над таким уровнем построен виртуальный уровень служебных функций, которой образует конструктивные блоки, используемые для всего многообразия функций, отправляющих файлы в ядро и получающих результаты, и, впоследствии, выполняет функции более высокого уровня. Над этим уровнем находится дополнительный уровень, использующий указанные служебные функции, который представляет собой уровень API, с которым взаимодействует пользователь, и который функционирует, в первую очередь, для конфигурации, загрузки файлов и выгрузки результатов. Такая конфигурация может включать в себя обмен информацией с регистрами, а также выполнение вызовов функций.

[00450] Например, как описано выше в настоящем документе, один из вызовов функций может быть осуществлен для создания таблицы расстановки с помощью алгоритма перемешивания. Причем, поскольку в некоторых вариантах реализации эта функция может быть создана на основе одного контрольного генома, причем только однажды для каждого контрольного генома, и может возникать необходимость создавать используемые в модуле картрирования таблицы расстановки на основе контрольной последовательности, то существует вызов функции, который выполняет такую функцию, и который будет получать имя файла, связанного с расположением контрольного файла и будет создавать после этого один или более файлов данных, содержащих таблицу расстановки и контрольную последовательность. Другой вызов функции может служить для загрузки таблицы расстановки, созданной с помощью алгоритма перемешивания, и ее передачи в память на микросхеме 100, и/или размещения в подходящем месте, определяемом аппаратными средствами. Конечно, возникнет необходимость загрузки контрольной последовательности на микросхему 100, что потребуется также для работы функции выравнивания. Модуль управления конфигурацией может выполнять такую функцию, например, с помощью загрузки всего необходимого для того, чтобы модули микросхемы 100 выполняли свои функции в памяти на микросхеме или прикрепленной к микросхеме 100.

[00451] Дополнительно, конфигурация API может быть настроена так, чтобы позволить микросхеме 100 взаимодействовать с системной платой секвенатора, в случае

если она входит в его состав, так чтобы получать файлы последовательностей в формате FASTQ прямо из секвенатора, например, непосредственно после того, как они сгенерированы, и затем передавать эту информацию в модуль управления конфигурацией, который затем направляет эту информацию в подходящие хранилища данных в аппаратных средствах 100, которые делают эту информацию доступной для соответствующих модулей в аппаратных средствах, так что они могут выполнять предписанные им функции, как то осуществлять поиск оснований, картрировать, выравнивать, сортировать и т.д. образец ДНК по отношению к контрольному геному.

[00452] Более того, в состав может входить интерфейс уровня клиента (CLI), где такой CLI может позволять пользователю вызывать одну или более из указанных функций напрямую. В различных вариантах реализации, CLI может представлять собой программное приложение, адаптированное для конфигурации использования аппаратных средств. Таким образом, CLI может представлять собой программу, которая принимает команды, например, аргументы, и делает доступной все функциональные возможности с помощью простого вызова приложения. Как указано выше, CLI может быть реализована на основе командной строки или GUI (графического пользовательского интерфейса). Строковые команды могут работать на уровне ниже GUI, в то время как GUI имеет в своем составе оконный диспетчер файлов с функцией выбора мышью блоков, на которых схематически изображены используемые модули и параметры их использования. Например, если поступили соответствующие команды, CLI в процессе своей работы будет определять местоположение контрольной последовательности, определять, существует ли необходимость создания таблицы расстановки и/или указателя, и, если они уже созданы, определять место их хранения, и направлять выгрузку созданных таблицы расстановки и/или указателя, и т.п. Команды такого типа могут возникать по усмотрению пользователя при работе с GUI, в случае, если пользователь выберет для работы описываемую микросхему.

[00453] Кроме того, в состав может входить библиотека, где такая библиотека может содержать предшествующие редактируемые файлы конфигурации, например, файлы, созданные для выбираемого пользователем типового функционирования аппаратных средств, например, для анализа части или целого генома, например, для анализа происхождения, диагностики заболеваний, поиска новых лекарственных средств, профилирования белков, и т.д. Предустановленные параметры такого типа, например, для выполнения подобных анализов, могут храниться в библиотеке. Например, если описанная в настоящем документе платформа задействована, например, для онкологических исследований, предустановленные параметры могут компоноваться не так, как в случае, если платформа направлена на генеалогические исследования.

[00454] Более конкретно, в онкологии точность может являться важным фактором, и, следовательно, параметры системы могут устанавливаться таким образом, чтобы обеспечить повышенную точность, приводя при этом к возможному падению скорости. Однако скорость может являться ключевым определяющим фактором для других прикладных задач геномики и, следовательно, параметры системы могут устанавливаться таким образом, чтобы сделать скорость максимальной, что, однако, может быть в ущерб точности. Соответственно, в различных вариантах реализации, настройки параметров, часто используемые для выполнения различных задач, могут быть предустановлены в библиотеке, чтобы обеспечить простоту использования. Такие настройки параметров могут также содержать информацию о программных приложениях, которые необходимо задействовать при запуске системы. Например, такая библиотека может содержать код, запускающий API, и, впоследствии, файлы-

примеры, командные файлы и любую вспомогательную информацию, необходимую для запуска системы. Таким образом, конфигурация библиотеки может быть настроена для компиляции программного обеспечения с целью запуска как API, так и различных выполняемых программ.

[00455] В различных случаях, PCI 104 и/или микросхема 100 может также включать в себя блок памяти, например, оперативной памяти (RAM) или динамической оперативной памяти (DRAM) с, например, интерфейсом DDR3, и такая память может использоваться для обеспечения работоспособности различных модулей, описанных в настоящем документе, например, модуля картрирования, выравнивателя, и/или сортировщика. Например, DRAM может использоваться там, где могут храниться контрольная последовательность, таблица расстановки и/или указатель таблицы расстановки, и/или считываемые фрагменты. Далее, как показано на фиг. 9., память может использоваться для обеспечения работоспособности различных других модулей, например, 114, описанных в настоящем документе, таких как, например, модуль устранения повторов, модуль местного перевыравнивания, модуль повторной калибровки качества оснований, модуль поиска вариантов, модули сжатия и/или распаковки. Например, DRAM может использоваться там, где могут храниться отсортированные считываемые фрагменты, аннотированные считываемые фрагменты, сжатые считываемые фрагменты и/или генетические варианты. Далее, конфигурация памяти может настраиваться таким образом, чтобы иметь в своем составе отдельный интерфейс для каждого из различных модулей памяти, задействованных выравнивателем и/или любым другим модулем, так что при этом каждый модуль памяти может содержать файловый уровень и логический уровень. Как указано выше, поскольку может быть задействовано несколько блоков памяти и/или несколько модулей, в состав может входить межсоединение на уровне микросхемы, чтобы обеспечивать обмен информацией через микросхему 100.

[00456] Таким образом, в различных случаях, аппаратура, описываемая в настоящем изобретении, может включать в себя микросхему 100, где такая микросхема включает в себя интегральную схему, образованную группой соединенных на постоянной основе цифровых логических схем, которые могут быть соединены между собой одним или более электрических межсоединений. В различных вариантах реализации одно или более физических электрических межсоединений включает в себя входной элемент, с помощью которого интегральная схема может соединяться с источником электронных данных для получения данных. Далее, в различных вариантах реализации, соединенные на постоянной основе цифровые логические схемы могут быть организованы в набор устройств обработки, например, где каждое устройство обработки может быть образовано подгруппой соединенных на постоянной основе цифровых логических схем, конфигурация которых настроена для выполнения одного или более этапов цифрового конвейера анализа последовательностей. Более конкретно, каждая подгруппа соединенных на постоянной основе цифровых логических схем находится в монтажной конфигурации для выполнения одного или более этапов цифрового конвейера анализа последовательностей.

[00457] В различных случаях, группа устройств обработки может включать в себя один или более модулей картрирования 112, модулей выравнивания 113, и/или модулей сортировки 114a, например, таким образом, что один или более этих модулей находятся в монтажной конфигурации. Например, в состав может входить модуль картрирования, находящийся в монтажной конфигурации, и такой модуль картрирования может получать доступ к указателю, например, одной или более контрольных генетических

последовательностей, например, из памяти, например, через одну или более совокупностей физических электронных межсоединений, например, для картрирования совокупности считываемых фрагментов одного или более сегментов одной или более контрольных генетических последовательностей. Далее, в различных случаях, в состав может входить модуль выравнивания, и такой модуль выравнивания может получать доступ к одной или более контрольных генетических последовательностей, например, из памяти, например, через одну или более совокупностей физических электронных межсоединений, например, для выравнивания совокупности считываемых фрагментов одного или более сегментов одной или более контрольных генетических последовательностей. Более того, в различных случаях, в состав может входить модуль сортировки, и такой модуль сортировки может получать доступ к одной или более выровненных последовательностей, например, из памяти, например, через одну или более совокупностей физических электронных межсоединений, например, для сортировки совокупности считываемых фрагментов одной или более хромосом одной или более контрольных генетических последовательностей. Таким же образом, в различных случаях, микросхема может содержать один или более из числа модулей местного перевыравнивания, маркировки дубликатов, повторной калибровки качества оснований и/или определения вариантов, например, в монтажной конфигурации таким же образом, как и в случае описанных выше модулей, для выполнения соответствующих функций.

[00458] Как указано выше, в различных случаях конфигурация одной или более интегральных схемсогласно настоящему изобретению, может быть настроена как для одной или более микросхем, например, FGPA, ASIC и/или структурной микросхемы ASIC. Например, интегральная схема, по своему типу, является набором электронных схем, расположенных на небольшой пластине или «чипе» из полупроводникового материала, например, кремния. Интегральные схемы обычно включают в себя схемные элементы, которые могут быть нераздельно с ними связаны и соединены между собой электрически. Прототипичная цифровая интегральная схема включает в себя разнообразные элементы, такие как, например, один или более логических вентилей, триггеров, концентраторов и других различных элементов, конфигурация которых настроена или настраиваема для функционирования схемы в качестве микропроцессора или другого микроконтроллера, например, для бинарной обработки сигналов «нулей» и «единиц», например, при выполнении одной или нескольких операций согласно настоящему изобретению.

[00459] Более конкретно, один или более программируемых с использованием маски логических вентилей могут быть сконфигурированы или запрограммированы для выполнения логических операций, таких как булевы функции одного или более входных логических переменных, так чтобы получать одиночную выходную логическую переменную. Такие логические вентили могут быть сконфигурированы с использованием одного или более диодов или транзисторов таким образом, чтобы вентили работали как электронные переключатели. В различных случаях, логические вентили могут включаться каскадом, подобно тому, как могут компоноваться булевы функции, что, таким образом, делает возможным построение физической модели любой булевой логики и, следовательно, любых алгоритмов и любых математических моделей, которые могут быть описаны булевой логикой, таких как описанные в настоящем документе, и могут быть реализованы логическими вентилями интегральных схем настоящего изобретения. В различных вариантах реализации набор вентилей может находиться на полупроводниковой пластине, формируя, таким образом, вентильную матрицу, например, схему вентильной матрицы.

[00460] В различных случаях, интегральная схема может также включать в себя один или более триггеров. Триггер может представлять собой схему или, как минимум, часть схемы, сконфигурированный в качестве клапана. Обычно триггер имеет два стабильных состояния и может переходить из одного в другое, например, с помощью сигнала, подаваемого на один или более контрольных входов, и, следовательно, триггер будет иметь один или два выхода. Триггеры используются для хранения информации о состоянии и, следовательно, могут применяться в качестве базовых элементов хранения, например в последовательных логических операциях. Интегральная схема может также включать в себя концентратор. Концентратор может быть сконфигурирован для выбора одного из нескольких входных сигналов, таких как цифровые (или аналоговые) входные сигналы, и, далее, может быть сконфигурирован для ретрансляции выбранного входного сигнала на выход. Таким образом, концентратор может использоваться для увеличения количества данных, пересылаемых через сеть в течение определенного отрезка времени при заданной пропускной способности.

[00461] В некоторых случаях, упоминаемых в настоящем документе, типичная интегральная схема может включать в себя любое количество, от одного до миллионов, таких элементов, сконфигурированных для выполнения операций, таких как те, которые описаны в настоящем документе, где различные элементы занимают лишь несколько квадратных миллиметров. Небольшие размеры таких схем обеспечивают высокую скорость, низкое рассеивание мощности и сниженную стоимость производства.

[00462] Такие интегральные схемы могут быть изготовлены с использованием широкого диапазона различных технологий, однако, обычно производятся в виде монолитной интегральной схемы. Например, типичная интегральная схема, например, полупроводниковая, может изготавливаться при помощи послойного процесса, например, такого процесса, который включает в себя три главных этапа, таких как перенос изображения, размещение и травление. В различных случаях, один или более из этих этапов могут дополняться этапами дальнейшей обработки, такими как легирование, очистка и тому подобными. Например, типичная процедура изготовления может предполагать наличие полупроводниковой пластины, например, монокристаллической кремниевой пластины, используемой в качестве подложки на которой интегральная схема собирается или, например, печатается. Затем используется фотолитография для печати на полупроводниковой пластине, например, для разметки разных областей подложки, которые затем могут быть легированы и/или на которых будут печататься проводящие дорожки, например, металлические, например алюминиевые.

[00463] В типичном случае интегральная схема состоит из одной или более совокупностей перекрывающихся слоев, таких, что структура каждого из слоев задается с помощью фотолитографии. Некоторые слои могут образовывать диффузионные слои с разметкой тех мест, в которых различные легирующие примеси диффундировали в подложку, а на других слоях могут быть заданы те места, в которые могут быть имплантированы дополнительные ионы. Дополнительные слои могут быть заданы как проводящие (например, поликремниевые, металлические или тому подобные) или как обеспечивающие соединение между проводящими слоями. Например, транзистор может быть сформирован в одном из мест, где вентильный слой (поликремниевый или металлический) пересекается с диффузионным слоем, и, в различных случаях, изогнутые полоски могут использоваться для формирования встроенных резисторов. Типичные интегральные схемы могут включать в себя: ASIC, FGPA, и/или структурную ASIC.

[00464] Интегральные схемы часто изготавливаются в качестве схем общего

назначения. Однако, в различных случаях, например, некоторых из описанных в настоящем документе, интегральная схема может быть специализированной, такой как интегральная схема специального назначения, или “ASIC.” Обычно ASIC относится к классу “стандартных ASIC,” то есть интегральных схем, специализированных под конкретное, не универсальное, применение. Обычно ASIC может содержать большое количество логических вентилей, например, в некоторых случаях, более 100 миллионов вентилей, которые могут быть сконфигурированы для выполнения большого числа разных операций, например, сконфигурированы для работы в качестве микропроцессоров и/или блоков памяти, включая ROM, RAM, EEPROM, флэш-память, и других больших конструктивных блоков, таких как те, которые предназначены для выполнения операций описываемых в настоящем документе. Уникальность ASIC состоит в том, что, поскольку она является микросхемой, разработанной для выполнения определенного набора прикладных задач, она может изготавливаться таким способом, чтобы стать специализированной, например, с использованием проектного протокола вентильной матрицы.

[00465] Так, например, вентильная матрица или нескоммутированная логическая матрица (ULA) могут использоваться при проектировании и производстве интегральных схем специального назначения (ASIC). В таком случае ASIC может производиться из предварительно изготовленной микросхемы, содержащей активные устройства, такие как вентили, например, вентили NAND, которые могут быть не соединены между собой на начальном этапе, но соединяться в последствии, например, в соответствии с проектным протоколом вентильной матрицы, например, с помощью дополнительного металлического слоя, например на производственном предприятии. Таким образом, что касается производства ASIC, вентильная матрица может быть предварительно изготовлена на плате кремниевой микросхемы, которая в результате не имеет конкретной функции, но содержит один или более транзисторов, стандартных логических вентилей NAND или NOR и может содержать в дальнейшем другие активные устройства, расположенные в заданных местах на полупроводниковой пластине, которая в этом случае может называться «базовый кристалл». Таким образом, создание схемы, выполняющей определенные специализированные функции, может завершаться добавлением последнего поверхностного слоя или слоев металлических межсоединений с микросхемами на базовом кристалле на последних стадиях производственного процесса, и соединением элементов, которые обеспечивают требуемую специализированность микросхемы, например, в соответствии с проектным протоколом.

[00466] Более конкретно, проектный протокол вентильной матрицы подразумевает использование метода производства, при котором различные диффузионные слои, например, с транзисторами и другими активными элементами, такими как описанные выше, готовы к использованию и построены на полупроводниковых пластинах общего назначения, и до выполнения этапа металлизации находятся в таком состоянии, что различные из элементов остаются несоединенными. В таком случае, микросхема может быть специализирована в дальнейшем, на некоторой стадии, в соответствии с различными связанными с ее специальным назначением параметрами, например, на стадии физического конструирования, когда определяются межсоединения готового устройства. Так, например, базовые кристаллы вентильной матрицы обычно изготовлены заранее и хранятся на складе в большом количестве в ожидании стадии специализации. Схему, предназначенную для конкретного применения, необходимо собирать на вентильной матрице таким способом, чтобы схема имела достаточное количество вентилей, монтажных соединений и контактов входа-выхода (I/O) для

выполнения требуемых функций.

[00467] Поскольку требования могут варьироваться, пластины вентильных матриц часто выпускаются стандартными сериями, как более габаритных изделий, имеющих, например, больше ресурсов, но являющихся, соответственно, более дорогостоящими, так и сравнительно небольших изделий, имеющих ограниченный выбор ресурсов, но являющихся менее дорогостоящими. Выбор подходящего стандарта полупроводниковой пластины должен быть основан на количестве ресурсов, требуемых для выполнения отобранных функций. Количество ресурсов, которые необходимо будет задействовать, можно сравнительно легко определить, например, подсчетом необходимого количества контактов ввода-вывода, однако, количество проводящих дорожек разводки может варьироваться в некоторых пределах, и, следовательно, должно выбираться с осторожностью. Однако, поскольку базовый кристалл обычно изготовлен заранее, проектирование и изготовление, в соответствии с требованиями конкретного проектного протокола, могут быть выполнены за меньшее время по сравнению со стандартным проектом частичной или полной специализации (FPGA). Также такой подход с использованием вентильных матриц снижает расходы на изготовление шаблонов, так как необходимо изготавливать меньшее количество шаблонов. Дополнительно к этому, время и стоимость выполнения инструментального тестирования также снижены, так как одни и те же контактирующие приспособления могут использоваться для всех изготавливаемых вентильных матриц с одним размером кристалла.

[00468] В таком случае, одна изготавливаемая стандартная микросхема, например, ASIC, может состоять из любого числа от двух до девяти, или десяти, или двенадцати и более слоев, таких, что при этом один или более, например, всех, идущих один за другим металлических слоев располагаются перпендикулярно находящимся под ними. Такие методы изготовления удобны, потому что позволяют получить микросхему сравнительно специализированной конструкции за относительно небольшое время сборки, поскольку завершающий процесс металлизации может быть выполнен быстро. Однако, такие микросхемы вентильных матриц, например, ASIC, часто не оптимальны, так как при размещении данной конструкции на «имеющейся в наличии» полупроводниковой пластине предоставленные ресурсы обычно не используются на 100%. Другим недостатком ASIC являются затраты на единовременное проектирование (NRE), которые могут составлять до миллионов долларов. Тем не менее затраты в пересчете на единицу продукции могут быть сравнительно низкими для ASIC.

[00469] Альтернативой стандартной ASIC при изготовлении специализированных микросхем является программируемая пользователем вентильная матрица или “FPGA”. В FPGA используется перезаписываемые программируемые логические блоки и межсоединения, которые позволяют спроектировать или, как минимум, частично перепроектировать FPGA для использования в разных прикладных задачах или в одной прикладной задаче, но решаемой большим количеством разных способов в разные моменты времени. Более конкретно, программируемая пользователем вентильная матрица является интегральной схемой, которая спроектирована так, чтобы ее конфигурация была настраиваемой один или много раз, например, заказчиком или проектировщиком, например, после изготовления.

[00470] Обычно FPGA имеют большое количество ресурсов в виде логических вентилей и/или блоков памяти, например, RAM, блоков, которые могут быть сконфигурированы для реализации сложных цифровых вычислений. Например, FPGA содержат как программируемые логические компоненты, именуемые «логические блоки», так и большое количество, например, иерархическую структуру, межсоединений

с перенастраиваемой конфигурацией, которые позволяют указанным блокам быть в "смонтированном" состоянии. Более конкретно, FGPA могут содержать большое количество изменяемых логических вентилей, которые могут быть смонтированы в разнообразных конфигурациях, образуя, таким образом, логические блоки, которые могут быть сконфигурированы для выполнения широкого диапазона сложных комбинационных функций, таких как те, что имеют отношение к операциям, подробно изложенным в настоящем документе. В различных случаях, логические блоки FPGA могут быть сконфигурированы так, чтобы включать в себя элементы памяти, например, такие простые, как триггеры, или более полноценные блоки памяти как ROM или RAM. Поскольку в FPGA используются очень быстрые I/O и двунаправленные шины данных, проверка подлинности отсчетов временных интервалов действительных данных и времени удержания может быть, в некоторых случаях, затруднена. Таким образом, в некоторых случаях, при надлежащем поуровневом планировании размещения элементов может допускаться выделение ресурсов на FPGA для того, чтобы учесть ограничения, связанные с контролем времени. FPGA, следовательно, может использоваться для реализации любой логической функции, которую может выполнять ASIC. Однако, возможности обновления и корректировки функциональности после доставки, частичной перенастройки конфигурации части структуры, и низкие затраты на единовременное проектирование по отношению к ASIC (несмотря на более высокую стоимость в пересчете на единицу продукции), определяют ее преимущества для многих прикладных задач.

[00471] В некоторых случаях, при изготовлении типичной FPGA применяется подход крупномодульной архитектуры таким образом, чтобы объединить логические блоки и межсоединения традиционной FPGA со встроенными микропроцессорами и относящимися к ним периферическими устройствами для создания полноценной «системы на программируемой микросхеме». В некоторых случаях, FPGA, описываемые в настоящем изобретении, могут быть перепрограммированы "во время работы" что, в соответствии с методами, изложенными в настоящем документе, может делать возможным выполнение вычислений с перенастраиваемой конфигурацией или создание систем с перенастраиваемой конфигурацией, например, ЦП, который может перенастроить собственную конфигурацию, чтобы стать совместимым с операциями, описанными в настоящем документе. В некоторых случаях могут использоваться микропроцессоры с конфигурируемым программным обеспечением для формирования матрицы ядер процессоров и FPGA-подобных программируемых ядер, которые могут находиться на одной микросхеме.

[00472] Обыкновенная архитектура FPGA может включать в себя матрицу конфигурируемых логических блоков, контактные площадки ввода-вывода, и/или один или более трассировочных каналов. Обычно логический блок может включать в себя одну или множество логических ячеек, и при этом типичная ячейка может содержать таблицу перекодировки (lookup table, LUT) с четырьмя входами, сумматор с тремя входами (full adder, FA), и/или триггер, и нечто подобное, функционирующее для формирования выходного сигнала. В различных случаях, выход может быть или синхронным или асинхронным. Прикладная схема может быть реализована в FPGA, и количество логических блоков, I/O, и проводящих дорожек разводки, которое необходимо включить в схему и которое может варьироваться, может быть определено в соответствии с проектной документацией. Необходимо отметить, что поскольку неиспользуемые дорожки могут увеличивать стоимость и снижать производительность интегральной схемы, не принося никакой выгоды, количество проводящих дорожек

разводки должно быть достаточным, то есть, подобранным для обеспечения разводки таблиц перекодировки (LUT) и I/O, но не быть в избытке. Далее, так как сигналы синхронизации обычно проходят по специально выделенной маршрутизируемой сети (например, глобальному буферу), такие сигналы могут управляться отдельно.

[00473] FPGA, описываемый в настоящем документе, может также содержать элементы, обеспечивающие функциональность более высокого уровня, реализованную на схеме, такие как один или более умножителей, блоки плат цифровой обработки сигналов (DSP) общего типа, встроенные процессоры, высокоскоростные логические элементы I/O, и/или встроенные блоки памяти. Включение в состав этих функций общего типа, встроенных в полупроводниковую пластину, уменьшает требуемую площадь и обеспечивает повышенную скорость выполнения таких функций. Необходимо отметить, что описываемые FPGA могут использоваться для проверки работоспособности систем, в том числе, на стадиях до и после завершения изготовления и стадии разработки встроенной программы, например, для проверки работоспособности итоговой конструкции перед производством «в использование» микросхем, таких как стандартная ASIC или структурная ASIC, которые могут представлять собой итоговый продукт.

[00474] При производстве типовой интегральной схемы, такой как FPGA или т.п., имеющей заданную функциональность, описанную в настоящем документе, может выполняться один или более из числа следующих этапов, в любой логической последовательности. На первом этапе могут быть использованы язык описания аппаратных средств (HDL) или схемотехническое проектирование. Далее, система автоматизированного проектирования, например, CAD, может быть задействована для создания технологического списка соединений. Список соединений затем может быть вписан непосредственно в архитектуру FPGA, например, с помощью процесса «размещения и разводки», с использованием соответствующего программного обеспечения для размещения компонентов и трассировки соединений. После завершения проектирования и проверки работоспособности сгенерированный файл в двоичной форме может использоваться для настройки/перенастройки конфигурации FPGA.

[00475] При реализации типичного проектного протокола, может выполняться моделирование работоспособности конструкции на различных этапах проектирования. Сначала описание логики регистровых передач, RTL, такое как VHDL или Verilog, может моделироваться с помощью создания испытательных стендов для моделирования системы и наблюдением за результатами. В некоторых случаях, устройство синтезирования может преобразовывать предполагаемый дизайн в список соединений, и после того, как устройство синтезирования закончит преобразование предполагаемого дизайна в список соединений, список соединений может быть переведен в форму описания уровня вентилей. На этой стадии моделирование может быть выполнено, например, еще раз, для того, чтобы убедиться, что синтез прошел без ошибок. После этого проект переходит в стадию размещения в FPGA, на которой могут добавляться задержки на прохождение сигнала, и моделирование может быть выполнено, например, еще раз, с добавлением этих данных в список соединений, например, перед итоговой проверкой работоспособности и дальнейшим изготовлением, например, одной или более ASIC или структурных ASIC микросхем.

[00476] Таким образом, структурная ASIC является комбинацией ASIC и FPGA, то есть, находится между FPGA и ASIC. Традиционная «стандартная ASIC», описанная выше, обычно является дорогостоящей, и требует значительных временных затрат на разработку. Так, например, при разработке стандартной ASIC может потребоваться изготовление большого количества фотолитографических шаблонов, для каждой

конструкции стандартной ASIC в отдельности. Однако, после этих первоначальных вложений в разработку типичные затраты на производство становятся очень низкими, а рабочие параметры, имеющие отношение к мощности, частоте и производительности, могут быть легко оптимизированы.

[00477] В отличие от стандартных ASIC, разработка типичных FPGA и/или CLPD, содержащих программируемые логические блоки, относительно не дорогая и быстрая, главным образом, потому что существующие устройства являются электронно программируемыми, и фотолитографические шаблоны не требуются. Однако, что касается рабочих параметров, имеющих отношение к мощности, частоте и производительности, такие устройства значительно уступают стандартным ASIC, и их стоимость в пересчете на единицу продукции может быть очень высокой, особенно для устройств с высокой производительностью.

[00478] Структурная ASIC, с другой стороны, является компромиссом между двумя упомянутыми выше устройствами. В отличие от вентильнаых матриц, структурные ASIC, как правило, содержат готовые конфигурируемые блоки памяти и/или аналоговые блоки. Таким образом, стоимость их разработки значительно ниже по сравнению со стандартными ASIC, так как необходимо изготавливать только несколько фотолитографических шаблонов для каждой конструкции структурной ASIC в отдельности, например, для конфигурируемых металлических слоев. И, несмотря на то, что стоимость в пересчете на единицу продукции значительно выше, чем для стандартных ASIC, она, тем не менее, существенно ниже, чем для FPGA. Что касается мощности и частоты, структурная ASIC занимает промежуточное положение между стандартными ASIC и FPGA, однако, по производительности она близка к самым крупным FPGA. Таким образом, во многих случаях, структурная ASIC может являться технологическим решением, которое уменьшает первоначальные затраты и время на разработку новой специализированной интегральной схемы.

[00479] Что касается проектирования и изготовления структурной ASIC, до того, как начнется серийная разработка структурных ASIC, «базовый кристалл» может быть разработан, например, с использованием методологии стандартной ASIC. Как указано выше, базовый кристалл может включать в себя типовые слои интегральных схем, такие как один или более транзисторов, блоков или ячеек памяти, ячеек входа/выхода, схем фазовой автоподстройки или других тактовых генераторов, или тому подобных. При необходимости базовый кристалл может содержать триггеры, релейные элементы с фиксацией воздействия, и/или мульти-транзисторные комбинационные вентили. Базовый кристалл может содержать некоторое количество монтажных соединений между компонентами, но необязательно полный их набор, реализующий полноценную логическую схему, поскольку соединения могут быть добавлены впоследствии. Отметим, что базовый кристалл может теоретически быть сконструирован таким образом, чтобы включать в себя стандартные ASIC, потенциально содержащие большие сложные модули, и рабочие параметры (мощность, частота, производительность) базового кристалла могут быть оптимальными, такими же, как и для стандартных ASIC. Фотолитографические шаблоны могут изготавливаться для содержимого базового кристалла, при этом количество шаблонов будет примерно таким же или меньше, как в случае стандартной ASIC. Таким образом, базовый кристалл включает в себя набор цифровых логических схем, которые могут быть соединены или еще не соединены на постоянной основе для определенного функционирования устройства.

[00480] После того, как базовый кристал сконструирован, может быть изготовлена серия из одного или более структурных ASIC, например, построена на одном и том же

базовом кристалле. Обычно при конструировании нескольких структурных ASIC используются ресурсы одного базового кристалла для амортизации стоимости базового кристалла при выполнении нескольких проектов. Каждая отдельная конструкция структурной ASIC может быть реализована набором новых монтажных соединений между компонентами (транзисторами и т.п.) в базовом кристалле, которые обеспечат построение вентилей более высокого уровня, триггеров, релейных элементов с фиксацией воздействия, блоков памяти, крупных логических модулей. Соответственно, определенные таким образом монтажные соединения могут быть реализованы при помощи меньшего числа дополнительных «конфигурируемых» металлических слоев 904A и 904B построенных на базовом кристалле, например, соединением контактных площадок или через базовый кристалл, например, с помощью монтажных соединений на конфигурируемых металлических слоях. Такие дополнительные металлические слои называют слоями “ конфигурируемыми”, потому что их можно специализировать под любую проектную конструкцию ASIC; однако, они фиксируются в процессе изготовления и не могут быть перемонтированы физически, а только с использованием реализованной в устройстве логической схемы. Количество конфигурируемых металлических слоев может быть любым.

[00481] Большинство из предполагаемых логических схем может быть, таким образом, реализовано с использованием базового кристалла и надлежащих монтажных соединений металлических слоев, поскольку базовый кристалл содержит достаточное количество логических ресурсов (транзисторов, блоков памяти и т.п.) для формирования всех необходимых элементов логической схемы. Количество конфигурируемых металлических слоев может меняться от одного проекта структурной ASIC к другому, но обычно может составлять приблизительно от 1 до 5 конфигурируемых слоев. Небольшое количество фотолитографических шаблонов может быть изготовлено в зависимости от числа конфигурируемых металлических слоев, и во время изготовления устройства полный набор шаблонов (шаблоны базового кристалла и шаблоны конфигурируемых металлических слоев) может использоваться для изготовления полупроводниковых пластин готового кристалла структурной ASIC. В других случаях, полупроводниковые пластины базовых кристаллов могли бы быть изготовлены заранее серийно, а металлические слои добавлены на готовые полупроводниковые пластины специализированных конструкций структурных ASIC на поздних стадиях изготовления.

[00482] Предпочтительным является проектирование базового кристалла структурной ASIC в один этап, например, первым проектировщиком, тогда как логические схемы структурных ASIC, построенных на этом базовом кристалле, могут проектироваться на втором этапе, например, разными проектировщиками, использующими результаты работы проектировщика структурных ASIC. В частности, различные исполнители обычно могут быть ответственны за «клиентскую часть» логической схемы, специализированной под требуемую функциональность интегральной схемы, например, RTL (логику регистровых передач), разработку кода, моделирование работоспособности, эмуляцию, регрессионное тестирование, отладку и тому подобное; тогда как проектировщик структурной ASIC может быть ответственным за проектирование «скрытой от пользователя части», в том числе синтезирование, размещение и прокладку межсоединений, анализ статических рисков сбоя, встраивание тестовой логики, и/или передачу в производство. Дополнительно может быть задействовано, например, микроэлектронное производство, для изготовления фотолитографических шаблонов и пластин, и/или тестирования и/или упаковки устройств. В различных случаях, проектировщик структурной ASIC может также проектировать специализированные

базовые кристаллы для определенного класса прикладных задач, например, содержащих логические ресурсы, тип и количество которых специализированы для таких прикладных задач.

[00483] Таким образом, благодаря наличию готовых металлических слоев (что уменьшает время производства) и определенным заранее характеристикам того, что находится на полупроводниковой пластине, например, базовый кристалл (что уменьшает время проектирования), циклы проектирования и полного изготовления могут быть завершены за меньшее время в случае структурной ASIC по сравнению с производством обычной ASIC. Например, при проектировании стандартной ASIC или FPGA, например, матрицы вентилей, пользователю зачастую может быть необходимым проектировать структуры питания, тактовые и тестовые структуры самостоятельно. Однако в случае структурной ASIC они могут быть готовы к использованию, что может позволить сократить время производства и расходы по сравнению со стандартными ASIC и матрицами вентилей.

[00484] В частности, задача проектирования структурных ASIC состоит в размещении схемы в фиксированной расстановке заданных ячеек. Более конкретно, сравнительная архитектура структурной ASIC может обычно включать в себя два главных уровня, таких как уровень структурных элементов и уровень матрицы структурных элементов. Такие структурные элементы могут включать в себя и комбинационные и последовательные функциональные блоки, которые могут функционировать или как логические элементы или элементы хранения. Дополнительно, что касается матриц структурных элементов, однородные или неоднородные стили матриц могут использоваться, например, в фиксированной расстановке структурных элементов.

[00485] Вследствие этого при проектировании структурной ASIC логические шаблонные слои могут быть определены заранее. В таком случае конструктивные изменения и специализация могут выполняться, например, с помощью создания специализированных металлических слоев, которые создают специализированные соединения между определенными заранее логическими элементами более низкого уровня. Также инструменты проектирования структурных ASIC могут быть значительно ниже по стоимости и проще (быстрее) в использовании, чем инструменты для проектирования стандартных ячеек, поскольку отсутствует необходимость выполнять все те функции, которые выполняются с помощью инструментов проектирования стандартных ячеек. Более конкретно, при проектировании могут использоваться существующие CAD для стандартных ячеек. В некоторых случаях, однако, могут использоваться CAD, специально разработанные для структурных ASIC. Могут также использоваться специализированные монтажные инструменты. Далее, как описывается в настоящем документе, разработаны новые и улучшенные алгоритмы, использующие модульный принцип организации структурных ASIC, с улучшенным контролем времени. Дополнительно, методы, описываемые в настоящем документе, могут быть задействованы для совершенствования процессов количественного определения и анализа, как отмечалось выше.

[00486] Таким образом, структурная ASIC может являться промежуточным технологическим звеном между конструкциями программируемой пользователем вентильной матрицы и стандартной ASIC. Более конкретно, поскольку требуется только небольшое количество специализированных слоев микросхемы, затраты на единовременное проектирование (NRE) могут быть значительно ниже, чем для микросхем со «стандартными ячейками» или «полностью специализированных» микросхем, когда требуется изготавливать полный набор шаблонов для каждой

конструкции. Таким образом, структурная ASIC обеспечивает высокую производительность (такую же, как в случае типичной ASIC) при низких затратах на NRE (такие же, как и в случае FPGA). Таким образом, процесс изготовления структурной ASIC обеспечивает быстрый вывод на рынок продукта, который сравнительно легко разработать, и обладающего сниженной стоимостью.

[00487] В некоторых случаях, однако, предпочтительно использовать FPGA, в котором межсоединения и логические блоки являются программируемыми после изготовления. Это обеспечивает большую адаптивность конструкции и легкость отладки при прототипировании. Однако, возможности построения больших схем, по размеру и скорости, на FPGA зачастую ограничены, что в некоторых случаях связано с присущими ей сложностью программируемой прокладки межсоединений и/или значительным объемом пространства, занимаемым различными включенными в нее программными элементами. С другой стороны, ASIC имеют ряд недостатков, например, дорогостоящий процесс разработки, связанный, в частности, с тем, что для каждой новой конструкции обычно требуется полный набор новых шаблонов. Структурная ASIC, следовательно, может быть оптимальным решением при выборе между двумя упомянутыми выше конструкциями. Она может иметь в своей основе структуру, такую же, как и FPGA, но быть программируемой по шаблону, как ASIC, вместо программируемости пользователем, посредством настроек конфигурации слоев в сквозных отверстиях между металлическими слоями. Например, один или более, например, каждый бит конфигурации статической оперативной памяти (SRAM) может быть заменен включением или не включением в конфигурацию сквозного или межгруппового металлического контакта.

[00488] Например, что касается архитектуры структурной ASIC, типичная архитектура часто может быть мелкоструктурной, среднеструктурной и/или иерархической. Мелкоструктурная архитектура может включать в себя большое число соединений на входе и выходе структурного элемента, в то время как структуры более крупного размера имеют меньше соединений со структурным элементом, и могут ухудшать поддерживаемую им фунциональность. Каждая отдельная конструкция будет иметь свой набор преимуществ в зависимости от ее структурных особенностей. Более конкретно, мелкоструктурная архитектура может включать в себя структурные элементы, содержащие разрозненные несоединенные компоненты, такие как транзисторы, резисторы и другие элементы управления, которые могут быть соединены впоследствии. Среднеструктурная архитектура элементов может включать в себя как типовую логику, так и вентили, MUX, LUT и/или элементы хранения, такие как триггеры. В противоположность этому, иерархическая архитектура может включать в себя мини-структурные элементы, которые, например, включают в себя вентили, MUX и LUT, но обычно не включают в себя типичные элементы хранения, такие как триггеры. В других случаях мини-элемент может быть объединен с регистрами или триггерами.

[00489] Что касается реализации структурной ASIC, различные этапы производства могут включать в себя один или более этапов проектирования логики регистровых передач (RTL); логический синтез для переноса RTL на структурные элементы; встраивание тестовых функций для улучшенной тестируемости и обнаружения неисправностей; монтаж для переноса каждого структурного элемента на матричный элемент и размещения каждого из элементов в фиксированном порядке; физический синтез для улучшенного отсчета временных интервалов на схеме и оптимизированного размещения каждого из элементов; синтез тактовой синхронизации для распределения сети тактовой синхронизации и минимизации искажений и задержек тактовой

синхронизации; прокладка межсоединений или, по-другому, установка монтажных соединений между различными элементами. В различных случаях такие этапы могут выполняться в различном логическом порядке, таким образом, чтобы сделать процесс проектирования, например относящийся к логическому синтезу, менее сложным, а также собрать более полную готовую библиотеку структурной ASIC, которая содержит расширенное представление о том, что конкретно может быть реализовано в соответствии с проектом.

[00490] Более того, обычной практикой при разработке ядер процессоров является лицензирование проектной конструкции процессора при его встраивании различными пользователями в их собственные полупроводниковые устройства. Список таких встроенных ядер может включать в себя процессоры общего назначения, такие как ARM, PowerPC, Krait и т.д., и также более специализированные процессоры, такие как графические процессоры (GPU) или векторные процессоры. Ядра встроенных процессоров могут быть большими, сложными конвейерными логическими модулями, работающими на высоких рабочих частотах, например, от 1 или 2 ГГц до примерно 3 или 6 ГГц или выше. Для достижения таких высоких частот работы ядер процессоров и связанной с ним буферной памяти требуются точные физическое размещение элементов и прокладка межсоединений, и, как результат, технология встроенных процессоров может зачастую сопровождаться «аппаратным макросом» (например, для определения точного размещения и прокладки межсоединений субкомпонентов), предназначенным для конкретного процесса изготовления.

[00491] Однако такое встроенное ядро процессора может быть недостаточно оптимальным при реализации структурной ASIC с использованием конфигурируемых металлических слоев. Аппаратные макросы обычно не применяются для конфигурируемых структурных ASIC, и даже если бы встроенный процессор был реализован, насколько это возможно, в соответствии с его аппаратным макросом в конфигурируемых металлических слоях, он бы, скорее всего, был ограниченным по частоте (например, 30% or 50% от номинальной рабочей частоты), и, скорее всего, потреблял очень большие объемы доступных ресурсов базового кристалла. Относительно малая площадь структурной ASIC по сравнению со стандартной ячейкой могла бы привести к тому, что встроенный процессор занял бы существенно большую физически доступную площадь на кремниевом кристалле, и, в комбинации с пониженной рабочей частотой, производительность в пересчете на единицу площади (или стоимость) могла быть ниже, чем при конструировании стандартной ячейки с тем же встроенным ядром.

[00492] Тем не менее, внедрение одного или более ядер процессоров в базовый кристалл структурной ASIC является целесообразным, например, с использованием методологии конструирования стандартной ячейки, описанной в настоящем документе, включающей в себя использование аппаратного макроса. Это позволило бы сохранить рабочую частоту, полную производительность и нормальную занимаемую площадь. Ядро процессора и/или монтажные соединения входа и выхода буферной памяти могли бы быть соединены с другими ресурсами базового кристалла или, что предпочтительно, вмонтированы в разводку конфигурируемого металлического слоя, что позволило бы соединить встроенные ядра с любыми инфраструктурными и логическими модулями, реализованными в каждой конкретной конструкции структурной ASIC. Таким образом, встроенные ядра процессоров становятся ресурсом базового кристалла, доступным для реализации в дальнейшем многих различных проектов структурных ASIC с использованием этого базового кристалла.

[00493] Встроенные в структурную ASIC ядра процессоров могут быть соединены с логической инфраструктурой, так что программное обеспечение (встроенное программное обеспечение), запускаемое в ядрах, может получать доступ и использовать различные блоки памяти и другие ресурсы на микросхеме и вне микросхемы, и осуществлять обмен информацией с одним или всеми другими логическими модулями на микросхеме через блоки памяти и/или напрямую. Таким образом, ядра процессоров могут работать параллельно с другими логическими модулями и/или взаимодействовать с другими логическими модулями для совершения совместной работы, например, когда ядра процессоров посылают запрос на выполнение задач другими модулями или другие модули посылают запрос на выполнение задач ядрами процессоров, или и то и другое.

[00494] Когда биоинформационные (Bio-IT) ускоряющие модули (например, картрирования, выравнивания, сортировки, маркировки дубликатов, повторной калибровки качества оснований, местного перевыравнивания, сжатия, распаковки и т.д., описанные в настоящем документе) реализованы на FPGA и/или структурной ASIC со встроенными ядрами процессоров, итоговая система на микросхеме (SOC) обладает рядом важных преимуществ, особенно в сочетании со скоростью и адаптивностью. Наибольшая скорость может достигаться с помощью аппаратных средств ускоряющих модулей, и наибольшая адаптивность может достигаться благодаря полной программируемости ядер процессоров. При помощи перепрограммирования ядер процессоров исполняемые алгоритмы Bio-IT могут быть легко модифицированы, и эти алгоритмы могут работать на порядок быстрее традиционных ЦП, поскольку интенсивные вычислительные операции могут быть возложены на ускорители аппаратных средств. Обмен информацией и организация памяти могут быть оптимизированы для совместной работы процессора и ускорителя. Дополнительное ускорение программных алгоритмов может достигаться с помощью дополнительных аппаратных модулей сконструированных для предварительной и послеоперационной обработки данных, используемых ядрами процессоров, например, для преобразования данных перекрывания считываемыми фрагментами участка контрольного генома в структуру данных наложения, для представления их в ядрах процессоров. В некоторых архитектурах набор команд может быть расширен для того, чтобы задействовать ресурсы аппаратных средств; в Bio-IT SOC могут быть определены новые команды процессора для доступа к функциям аппаратного ускорения Bio-IT.

[00495] Как представлено ниже в таблице II, структурная ASIC, следовательно, имеет несколько определенных заранее, перед изготовлением, преимуществ, например, перед ASIC или FPGA. Например, различные компоненты могут быть «почти» соединены, например, в разнообразных готовых конфигурациях, и многочисленные центральные и локальные тактовые генераторы могут быть изготовлены заранее. Это означает, следовательно, что проблемы с целостностью сигнала и отсчетом временных интервалов должны по существу решаться. Дополнительно необходимо изготавливать только несколько металлических слоев. Далее, в отличие от стандартных FPGA, структурная ASIC должна иметь мощности, производительность, и параметры энергопотребления близкие к стандартной ASIC. Это позволяет ускорить и облегчить процесс разработки, а также снизить затраты на единовременное проектирование (NRE) по сравнению со стандартными ASIC, и существенно уменьшить время от разработки устройства до его выпуска. Более того, должна отсутствовать необходимость решения проблемы искажений.

ТАБЛИЦА II

	FPGA	Структурная ASIC	Стандартная ASIC

Площадь кристалла	Очень большая	Маленькая	Очень маленькая
Использование ресурсов мощности	Высокое	Низкое	Очень низкое
Рабочая частота	Низкая	Высокая	Высокая
Логическая емкость	Средняя	Средняя	Высокая
Стоимость разработки	Очень низкая	Низкая	Высокая
Стоимость единицы продукции	Очень высокая	Низкая	Очень низкая

[00496] Структурная ASIC, следовательно, обладает несколькими обеспечивающими ее преимущества свойствами, в том числе, одним или более из следующих: низкие затраты на единовременное проектирование, упрощенные требования к выполнению опытно-конструкторских работ, сниженные затраты на производство шаблонов по сравнению со стандартной ASIC, с дополнительными преимуществами в виде высокой производительности, низкого потребления энергии, меньшего числа изготавливаемых слоев, меньшей сложности в части конфигурации готовых для размещения схемных элементов предварительно изготовленных блоков ячеек, что приводит к уменьшению времени на изготовление. Однако структурная ASIC обладает некоторыми недостатками, например, нехватка адекватных инструментов проектирования, которые могут являться дорогостоящими и которые получают с помощью изменения традиционных инструментов ASIC. Далее, такие новые архитектуры по-прежнему нуждаются в формальной оценке и сравнительном анализе. И существуют проблемы выбора оптимальной таблицы перекодировки между LUT с 3, 4, и 5 входами и/или оптимального размера распределенной памяти RAM.

[00497] Таким образом, принимая во внимание вышесказанное, ASIC, FPGA и структурная ASIC обладают как своими преимуществами, так и недостатками. Например, стандартная ASIC является непростой для проектирования, требует больше времени на разработку, имеет высокие затраты на NRE. Однако ASIC может поддерживать большие конструкции, сложные конструкции, обладает высокой производительностью при низком энергопотреблении, что может, следовательно, приводить к низкой стоимости в пересчете на единицу продукции (при больших объемах). С другой стороны, FPGA является простой для проектирования, требует меньше времени на разработку, имеет низкие затраты на NRE. Однако FPGA могут иметь ограниченный размер конструкции и/или сложность, могут обладать ограниченной производительностью и высоким энергопотреблением, что может приводить к высокой стоимости в пересчете на единицу продукции. Во многих случаях структурная ASIC может быть спроектирована для максимального увеличения указанных преимуществ и уменьшения указанных недостатков. Например, по большому счету, отношение для числа вентилей на единицу площади может примерно выражаться как 100:33:1 для стандартной ASIC, структурной ASIC и FPGA; для производительности (основанной на тактовой частоте) - как 100:75:15; для мощности - как 1:3:12.

[00498] Как указано выше, в различных случаях микросхема 100 настоящего изобретения может быть сконфигурирована как плата расширения, например, если указанная микросхема включает в себя шину PCIe и расположена так, чтобы обмениваться информацией с одним или более блоков памяти, например, окружена блоками памяти, например, в значительной степени окружена блоками памяти, например, полностью окружена блоками памяти. В различных вариантах реализации, микросхема может представлять собой FPGA с высокой плотностью упаковки и/или быстродействующую FPGA, которая различных случаях, может быть преобразована в ASIC или структурную ASIC. В различных случаях, микросхема может быть структурной ASIC, которая может быть преобразована в ASIC. В некоторых случаях,

микросхема может представлять собой ASIC.

[00499] Как указано выше, модули, описанные в настоящем документе, могут быть реализованы в аппаратных средствах микросхемы, например, соединены с ней на постоянной основе, и в таких случаях реализация может приводить к тому, что их функционирование может происходить с более высокой скоростью по сравнению с программной реализацией таким образом, что, например, минимальное количество команд вызывается, считывается и/или исполняется. Таким образом, при такой уникальной реализации в аппаратных средствах, модули, описанные в настоящем изобретении, могут функционировать прямо в соответствии с их рабочими параметрами, например, без необходимости вызывать, считывать и/или исполнять команды. Дополнительно, требования к памяти и времени обработки могут быть упрощены, например, когда обмен информацией внутри микросхемы происходит через файлы, а не через доступ к памяти. Конечно, в некоторых случаях, микросхема и/или плата могут быть увеличены в размерах, например, для большего количества встроенной памяти, так чтобы расширить возможности параллельной обработки, что может приводить к еще большему увеличению скорости обработки. Например, в некоторых вариантах реализации, микросхема, описываемая в настоящем изобретении, может включать в себя встроенную DRAM для того, чтобы микросхема не зависела от внешней памяти, что, следовательно, привело бы к дополнительному увеличению скорости обработки, например, при работе алгоритма Барроуза-Уилера вместо таблицы расстановки и функции перемешивания, которые могут в различных случаях, зависеть от внешней, например, центральной памяти. В таких случаях работа всего программного конвейера может быть завершена через 6 минут или меньше, например, от ее начала до ее конца.

[00500] Как указано выше и показано на фиг.8, существуют различные точки, в которых данный модуль может быть расположен на аппаратных средствах, или расположен удаленно, например, на сервере с доступом через облако. В случае, если данный модуль расположен на микросхеме, например, соединен с микросхемой на постоянной основе, его функция может выполняться аппаратными средствами, однако, если потребуется, модуль может быть расположен удаленно, и в таком случае платформа может включать в себя необходимый набор инструментов для отправки соответствующих данных на сервер, например, сервер с доступом через облако, так что конкретная фунциональность модуля может привлекаться для дальнейшей обработки данных в соответствии с выбранным пользователем протоколом. Таким образом, часть платформы может включать в себя интерфейс удаленного доступа через интернет для выполнения одной или более задач, имеющих отношение к функционированию одного или более модулей, описанных в настоящем документе. Например, если модули картрирования 112, выравнивания 113, и/или сортировки 114a это все модули, находящиеся на микросхеме, то, в различных случаях, один или более модулей локального перевыравнивания 114d, маркировки дубликатов 114b, повторной калибровки качества оснований 114c, и/или поиска вариантов 115 могут находиться на облаке.

[00501] В частности, после того, как генетические данные сгенерированы и/или обработаны, например, с использованием одного из первичных и/или вторичных протоколов, например, картрированы, выровнены, и/или отсортированы, например, для создания одного или более файлов определения вариантов, например, для определения того, как данные генетической последовательности исследуемого объекта отличаются от одной или более контрольных последовательностей, следующий аспект изобретения может быть связан с выполнением одной или более аналитических функций

в отношении сгенерированных и/или обработанных генетических данных, таких как дальнейшая, например, третичная, обработка. Например, конфигурация системы, представленной на фиг. 8-11, может быть настроена для последующей обработки сгенерированных данных и/или данных, полученных в результате вторичной обработки, например, для ее запуска через один или более программных конвейеров третичной обработки 700, например, через один или более программных конвейеров генома, программных конвейеров эпигенома, программных конвейеров метагенома, совместного генотипирования, программных конвейеров MuTect2 или других программных конвейеров третичной обработки, например, с помощью устройств и методов, описанных в настоящем документе. Например, в различных случаях может обеспечиваться дополнительный уровень обработки 122, например, для диагностики заболеваний, терапевтического лечения и/или профилактических действий, например, включающий в себя NIPT, NICU, Cancer, LDT, AgBio и другие подобные методы диагностики заболеваний, профилактики, и/или лечения, в которых задействованы данные, получаемые в результате работы одного или более таких первичных и/или вторичных и/или третичных программных конвейеров. Таким образом, описанные в настоящем документе устройства и методы могут использоваться для генерации данных генетической последовательности, которые могут в дальнейшем использоваться для генерации одного или более файлов определения вариантов и/или других, связанных с этим данных, которые в дальнейшем могут являться объектом исследования для других запускаемых программных конвейеров третичной обработки в соответствии с описанными в настоящем документе устройствами и методами, например, как для диагностики конкретного и/или системного заболевания, так и для профилактики и/или терапевтического лечения и/или методик, находящихся в стадии разработки.

[00502] Как описано выше, система 1 представленная в настоящем документе, может включать в себя генерирование, например, с помощью технологического решения, при котором секвенатор находится на микросхеме, описанного в настоящем документе, или, в других случаях, получение данных генетической последовательности, и может включать в себя выполнение одного или более протоколов вторичной обработки, например, одного или более из следующих: картрирование, выравнивание, сортировка сгенерированных данных генетической последовательности, например, для получения одного или более файлов поиска вариантов, например, для того чтобы, определить как данные генетической последовательности одного объекта исследования отличаются от одной или более последовательностей контрольных геномов. Следующий признак изобретения может быть связан с выполнением одной или более аналитических функций в отношении полученных и/или обработанных генетических данных, таких как дальнейшая, например, третичная, обработка, которая может выполняться на той же микросхеме или совместно с той же микросхемой или набором микросхем, на которых находится упомянутый выше секвенатор.

[00503] В первом случае, например, что касается генерации, получения и/или передачи данных генетической последовательности, как показано на фиг. 8, такие данные могут быть получены локально или удаленно и/или в результате могут быть затем обработаны, например, с использованием локальных вычислительных ресурсов 100, или могут быть переданы в удаленное местоположение, например, к удаленным вычислительным ресурсам 300, для последующей обработки. Например, сгенерированные данные генетической последовательности могут обрабатываться локально и непосредственно, например, если один набор микросхем и/или одно устройство имеет функциональные возможности секвенирования и вторичной обработки. Сгенерированные данные

генетической последовательности также могут обрабатываться локально, но не непосредственно, если функциональные возможности секвенирования и вторичной обработки находятся на различных аппаратах, которые находятся в одном помещении, но разделены в пространстве и при этом соединены с возможностью передачи сигнала, например через локальную сеть 100. В других случаях, данные генетической последовательности могут обрабатываться удаленно, например, с помощью NGS, и результаты могут передаваться через облачную сеть 50 в удаленное местоположение, например, географически не совпадающее с секвенатором.

[00504] В частности, как показано на фиг. 8-11, в различных вариантах реализации, может обеспечиваться наличие секвенатора нуклеотидов непосредственнно в рабочем пространстве, например, секвенатора на микросхеме или NGS, где указанный секвенатор объединен с локальным вычислительным ресурсом прямо или косвенно, например соединением через локальную сеть 10. Локальный вычислительный ресурс 100 может включать в себя или, в других случаях, быть объединен с одним или более механизмами генерации данных 110 и/или получения данных 120. Такие механизмы могут представлять собой любые механизмы, сконфигурированные для генерации и/или, в других случаях, получения данных, например, аналоговых, цифровых или электромагнитных данных, имеющих отношение к одной или более генетических последовательностей исследуемого объекта или группы объектов.

[00505] Например, такой механизм генерации данных 110 может представлять собой устройство первичной обработки, например, секвенатор, например, NGS, секвенатор на микросхеме или другой подобный механизм генерации информации о генетической последовательности. Далее, такой механизм получения данных 120 может представлять собой любой механизм, сконфигурированный для приема данных, например, сгенерированной информации о генетической последовательности, и/или, при совместном использовании с механизмом генерации данных 110 и/или вычислительных ресурсов 150, способный задействовать один или более протоколов вторичной обработки, таких как программный конвейер вторичной обработки, сконфигурированный для исполнения протоколов картрирования, выравнивания, сортировки, и/или поиска вариантов в отношении сгенерированных или полученных данных последовательности, как поисано в настоящем документе. В различных случаях, аппараты генерации данных 110 и/или получения данных 120 могут соединяться, например, через локальную сеть 10, например, для локального хранения 200, или могут соединяться через облачную сеть 30, например, для передачи и/или приема данных, таких как цифровые данные, имеющие отношение к первичной и/или вторичной обработке информации о генетической последовательности, например, в удаленное местоположение 30 или из удаленного местоположения 30, например для удаленных обработки 300 и/или хранения 400. В различных вариантах реализации, один или более таких компонентов могут быть сопряжены с возможностью передачи сигнала при помощи гибридной сети, как описано в настоящем документе.

[00506] Локальный вычислительный ресурс 100 может также включать в себя или, в других случаях, быть объединен с компилятором 130 и/или процессором 150, такими, что компилятор 130 сконфигурирован для компиляции сгенерированных и/или полученных данных и/или связанных с ними данных, а процессор 150 сконфигурирован для обработки сгенерированных и/или полученных и/или скомпилированных данных и/или управления системой 1 и ее компонентами, как описано в настоящем документе. Далее, локальные вычислительные ресурсы 100 могут включать в себя устройство сжатия данных 160, сконфигурированное для сжатия данных, таких как сгенерированные

и/или полученные данные и/или данные, являющиеся результатом первичной и/или вторичной обработки, которые могут быть сжаты, например, перед отправкой через локальную сеть 10 и/или облачную сеть 30 и/или гибридную облачную сеть 50.

[00507] В конкретных примерах, как можно видеть на фиг. 8-11, система 1 может быть выполнена с возможностью выполнения дальнейшей обработки сгенерированных и/или обрабатываемых во вторую очередь данных, например посредством локальных 100 и/или удаленных вычислительных ресурсов, например посредстом их прогона через один или более конвейер для третичной обработки, такой как один или более из конвейера генома, конвейера эпигенома, конвейера метагенома, совместного генотипирования, конвейера MuTect2 или другого конвейера для третичной обработки. Такие данные затем могут быть сжаты и/или локально 200 локально и/или могут быть переданы для удаленного хранения.

[00508] В дополнительных примерах система 1 может содержать дополнительный уровень обрабатывающих модулей, например выполненных с возможностью выполнения дополнительной обработки, например с возможностью диагностики, обнаружения заболевания и/или терапевтического свойства, и/или их профилактики. Например, в различных примерах могут быть обеспечены дополнительные уровни обработки, например для диагностики заболевания, терапевтического лечения и/или профилактического предупреждения, включая, например, неинвазивное пренатальное тестирование (NIPT), реанимацию и интенсивную терапию новорожденных (NICU), рак, проводимые в лаборатории исследования (LDT), агробиологию (AgBio) и другие виды диагностики, профилактики и/или способов лечения таких заболеваний, в которых применяются данные, сгенерированные одним или более из указанных первичных, вторичных и/или третичных конвейеров.

[00509] Соответственно, в настоящем документе представлена система 1 для получения и применения глобальной гибридной облачной сети 50. Например, согласно настоящему документу, облако 30 в первую очередь используется для хранения, например в удаленном месте 400 хранения. В таком примере вычисление данных выполняют локально 100 посредством локальных вычислительных ресурсов 150, и в случае, когда требования к хранилищу являются существенными, осуществляется доступ в облако 30 для хранения данных, сгенерированных локальными вычислительными ресурсами 150, например с применением удаленных ресурсов 400 хранения. Следовательно, управление сгенерированными данными обычно любо полностью осуществляют локально 100, либо полностью осуществляют удаленно 300 в облаке 30.

[00510] В частности, в общем варианте осуществления платформы для биоинформационного анализа функции вычисления 150 и/или хранения 200 выполняют локально, а когда требования к хранилищу превосходят общую емкость хранилища, или когда есть необходимость обеспечить доступ к сохраенным данным для других удаленных пользователей, такие данные могут быть переданы через сеть Интернет 30 в облако для удаленного хранения 400 в нем. В таком примере, когда вычислительные ресурсы 150, которые требуются для выполнения вычислительных функций, являются минимальными, а требования к зранилищу являются существенными, вычислительная функция 150 может выполняться локально 100, а функция 400 хранения может выполняться удаленно, при этом полностью обработанные данные передают в обе стороны между функцией 150 обработки, например торлько для локальной обработки, и функцией 400 хранения, например для удаленного хранения 400 обработанных данных.

[00511] Например, можно привести пример в отношении функции секвенирования, например посредством типичного секвенатора нового поколения (NGS), где

вычислительные ресурсы 150 выполнены с возможностью реализации функций, необходимых для секвенирования генетического материала, для получения генетических секвенированных данных, например ридов, которые вырабатывают на месте 100. Указанные риды, после их генерации, например посредством локального NGS, затем могут быть переданы, например по облачной сети 30, например для хранения 400 в удаленном местоположении 300 таким образом, что при необходимости их можно повторно ихзвлечь из облака 30, например для дальнейшей обработки, к примеру, для выполнения одной или более из функций вторичной и/или третичной обработки, которые выполняют в местоположении, удаленном от средств 400 хранения, т.е. локально. В таком примере локальные ресурсы 150 хранения предназначены лишь в качестве кэша хранилища, в котором расположены данные, при ожидании передачи из облака 30, или в него, например для формирования удаленных средств 400 хранения.

[00512] Подобным образом, когда вычислительная функция является существенной, например если ей для обработки данных требуется одно или более ядер 300 удаленных кластеров компьютеров, и где требования к хранилищу для хранения обработанных данных 200 сравнительно малы по сравнению с вычислительными ремурсами 300, требуемыми для обработки данных, подлежащие обработке данные могут быть отправлены, например по облаку 30, для их обработки удаленными вычислительными ресурсами, которые могут включать одно или более ядер или один или более кластеров вычислительных ресурсов, например одни или более супер вычислительных ресурсов. В таком примере после того, как данные были обработаны компьютерным ядром 300 на облачной основе, обработанные данные могут затем быть переданы по облачной сети 30 для локального 200 зранения и быть доступны для использования локальными вычислительными ресурсами 150, например для локального анализа и/или локальной диагностики.

[00513] В качестве примера можно привести типичную функцию вторичной обработки, например когда осуществляют доступ к предобработанным секвенированным данным, т.е. риду, хранимым локально 200, например посредством локальных вычислительных ресурсов 100, и передают по облачной сети Интернет 30 на удаленные вычислительные средства 300 для дальнейшей обработки в них, например при выполнении функции вторичной обработки, для получения рузельтирующих обработанных данных, которые затем могут быть отправлены обратно на локальные средства 100 для хранения 200 в них. Это может быть случаем, когда локальный специалист генерирует секвенированные данные рида с использованием локальных ресурсов 100 для генерации данных, например автоматического секвенатора, а затем отправляет данные по сети 30 на удаленные вычислительные средства 300, которые затем выполняют одну или более функций над указанными данными, такие как преобразование Барроуза-Уилера или функция выравнивания Нидлмана-Вунша и/или Смита-Ватермана над указанными данными последовательности, для генерации результирующих данных, которые могут быть переданы по сети Интернет 30 на локальные вычислительные ресурсы 100 для их изучения по одному или более локально администрируемым протоколам обработки и/или для их локального 200 хранения.

[00514] Однако необходима безшовная интеграция между взаимодействием между локальной 100 и удаленной 300 компьютерной обработкой, а также между локальным 200 и удаленным 400 хранилищем, например в гибридной системе на основе облака 50, представленной в настоящем документе. В таком примере система может быть выполнена таким образом, что локальные 100 и удаленные 300 вычислительные ресурсы выполнены с возможностью работы совместно бесшовным образом, так что

подлежащие обработке при помощи них данные могут быть назначены в режиме реального времени локальным 200 или удаленным 300 вычислительным ресурсам без существенных ухудшений вследствие скорости передачи и/или существенных ухудшений в эффективности работы. Это может иметь место, например, когда программное обеспечение и/или аппаратное обеспечение, предназначенные для развертывания или другим образом запуска на вычислительных ресурсах, выполнены так, чтобы соответствовать друг другу и/или являются одинаковыми или функционально схожими, например аппаратное обеспечение и/или программное обеспечение выполнено одним и тем же образом, чтобы выполнять одни и те же алгоритмы одинаковым образом над сгенерированными и/или полученными данными.

[00515] Например, как можно видеть в отношении фиг. 8-9, локальные вычислительные ресурсы 100 могут быть выполнены с возможностью генерации данных и, следовательно, могут содержать механизм 110 геренации данных, например для генерации и/или анализа первичных данных, например для получения файла последовательности в формате FASTQ. Указанный механизм 110 генерации данных может представлять собой локальный компьютер, как описано в настоящем документе, имеющий процессор, который может быть выполнен с возможностью исполнения одного или более программных приложений и/или может быть жестко смонтирован для выполнения одного или более алгоритмов, например в монтажной конфигурации, над сгенерированными и/или полученными данными. Например механизм 110 геренации данных может быть выполнен с возможностью генерации данных, таким как данных 111 секвенирования, которые могут представлять собой измеренные данные 111a, например данные, которые можно обнаружить в виде изменения напряжения, ионной концентрации, электромагнитного излучения и тому подобного; и/или механизм 110 геренации данных может быть выполнен с возможностью генерации и/или обработки сигнала, например данных аналогового или цифрового сигнала, такие как данные, представляющие одно или более значений нуклеотидов в последовательности или цепи связанных нуклеотидов. В таком примере механизм 110 геренации данных, например секвенатор 111, может также быть выполнен с возможностью предварительной обработки сгенерированных данных для выполнения одной или более операций 111c определения основания, например над данными для получения идентификационных данных последовательности, например файл в формате FASTQ.

[00516] Следует отметить, что в данном примере данные 111, сгенерированные таким образом, могут быть сгенерированы локально, например посредством ресурсов 150 для генерации и/или вычисления данных, например секвенаторе на микросхеме; или они могут быть получены удаленно, например посредством удаленных вычислительных и/или генерирующих ресурсов, таких как удаленный NGS 300, но переданы по облаку 30/50 на локальные вычислительные ресурсы 100, например для вторичной обработки 150 и/или хранения посредством них в локальных ресурсах 200 зранения, например во время ожидания дальнейшей локальной обработки 150. В таком примере, когда генерирующие данные ресурсы 300 являются удаленными от локальных обрабатывающих 100 ресурсов и/или ресурсов хранения 200, соответствующие ресурсы могут быть выполнены таким образом, что протоколы удаленного и/или локального хранения, удаленной и локальной обработки и/или обмена информацией, используемые каждыми ресурсами, могут быть выполнены с возможностью плавной и/или бесшовной интеграции друг с другом, например посредством выполнения одного и того же, схожего и/или эквивалентного программного обеспечения и/или вследствие обладания одинаковыми, схожими и/или эквивалентными конфигурациями аппаратного

обеспечения, и/или вследствие применения одних и тех же протоколов обмена информацией и/или передачи, которые, в некоторых допустимых пределах, могут быть реализованы при производстве или позднее.

[00517] В частности, указанные функции могут быть реализованы в жестко смонтированной конфигурации, например в которой функция секвенирования и функция вторичной обработки выполняются в одной или связанных микросхемах или наборах микросхем, например в таких, в которых секвенатор и вторичный процессор напрямую соединены на микросхеме, как описано в настоящем документе, или могут быть реализованы посредством программного обеспечения, оптимизированного для обеспечения связи двух удаленных устройств друг с другом бесшовным образом. Также может применяться комбинация оптимизированных вариантов реализации аппаратного обеспечения и программного обеспечения для выполнения указанных функций, как описано в настоящем документе. В различных вариантах реализации ресурсы для генерации данных, такие как секвенатор 111, выполненный в виде программного обеспечения и/или аппаратного обеспечения или в виде их комбинации, также могут содержать первоначальный уровень процессоров 500, таких как диспетчер, различные аналитические устройства, сравнивающие устройства, регистрирующие устройства, выпускающие устройства и тому подобное для помощи генератору 111 данных, например секвенатору, при преобразовании биологической информации в первичные данные ридов, например в файлы в формате 111d FASTQ.

[00518] Аналогичным образом это может выполняться и в отношении выполнения других функций, которые могут быть развернуты на локальных 100 и/или удаленных 300 вычислительных ресурсах. Например локальные вычислительные ресурсы 100 могут содержать программное обеспечение и/или аппаратное обеспечение, выполненное с возможностью реализации одной или более функций 112-115 обработки вторичного уровня 700 над данными, сгенерированными удаленно и/или локально, такими как данные генетической последовательности, таким образом, что обработка и ее результаты могут быть бесшовным образом разделены друг с другом и/или сохранены в них. В частности, локальная вычислительная функция 100 и/или удаленная вычислительная функция 300 могут быть выполнены с возможностью генерации и/или приема первичных данных, таких как данные генетической последовательности, например в виде файла в формате FASTQ или тому подобного, и исполнения одного или более протоколов 600 вторичной обработки над сгенерированными и/или полученными данными, причем указанные протоколы могут быть реализованы в виде программного обеспечения, аппаратного обеспечения или их комбинации. Например ресурсы 110 для генерации и/или вычисления данных могут быть выполнены с вохможностью реализации одной или более картирующих операций 112, операций 113 выравнивания или других связанных функций 114 над полученными или сгенерированными данными.

[00519] Более конкретно, ресурсы 110 для генерации данных могут содержать картирующий модуль 112, как описано в настоящем документе, или могут быть другим образом запрограммированы для исполнения алгоритма картирования над данными генетической последовательности, например для выполнения преобразования Барроуза-Уилера и/или других алгоритмов построения хэш-таблицы и/или исполнения хэш-функции 112a над указанными данными, например для картирования семян с помощью хэш-функции для генерации картированных данных последовательности. Ресурсы 110 для генерации данных могут также содержать выравнивающий модуль 113, как описано в настоящем документе, или могут быть другим образом запрограммированы для исполнения алгоритма выравнивания над данными генетической последовательности,

например над картированными секвенированными данными, например для выполнения выравнивания Смита-Ватермана с промежутками и/или без промежутков, и/или алгоритма Нидлмана-Вунша или других алгоритмов 113a оценивания над указанными данными для генерации выровненных данных последовательности. Ресурсы 110 для генерации данных могут также быть выполнены содержащими один или более других модулей 114, выполненных с возможностью выполнения одной или более других функций обработки над данными генетической последовательности, например над картированными и/или выровненными данными последовательности, и, следовательно, могут содержать подходящим образом сконфигурированный модуль 114 или могут быть другим образом запрограммированы для исполнения одной или более других функций обработки, таких как сортировка 114a, удаление 114b дубликатов, рекалибровка 114c, местное повторное выравнивание 114d, дублирующая маркировка 114f, функция (функции) Рекалибровки 114g оценки качества оснований и/или функция сжатия (например для получения файла в формате BAM, Reduced BAM и/или сжатый и/или распакованный CRAM) 114e, в соответствии со способами, раскрытыми в настоящем документе, причем указанные функции обработки могут быть выполнены в виде одного или более конвейеров системы 1. Аналогичным образом, система 1 может содержать модуль 115, выполненный с возможностью обработки данных, например секвенированных, картированных, выровненных и/или сортированных данных, так чтобы получать файл 116 определения вариантов, например с функциональными возможностями обработки, основанными на аппаратнои обеспечении и/или программном обеспечении. Более конкретно, система 1 может содержать модуль 115 определения вариантов для исполнения одной или более функций определения вариантов, например функции 115a скрытой марковской модели (CMM, HMM) и/или программного обеспечения геномного анализа (GATK), например в монтажной конфигурации и/или посредством одного или более приложений программного обеспечения, локально или удаленно, и/или преобразователь 115b для тех же целей.

[00520] В некоторых вариантах реализации, показанных на фиг. 8 и 10, система 1 может содержать локальную вычислительную функцию 100, которая может быть выполнена с возможностью применения компьютерных обрабатывающих ресурсов 150 для выполнения одной или более дополнительных функций компьютерной обработки над данными, сгенерированными системным генератором 110 или полученными посредством механизма 120 получения системы (как описано ниже), например посредством передачи в них, например третьими лицами 121, к примеру, посредством облака 30 или гибридной облачной сети 50. Например, анализатор 121 третьих лиц может развертывать удаленные вычислительные ресурсы 300 для генерации релевантных данных с необходимостью дальнейшей обработки, таких как данные генетической последовательности или тому подобное, причем указанные данные могут быть переданы системе 1 по сети 30/50 для дальнейшей обработки. Это может быть полезным, например, когда удаленные вычислительные ресурсы 300 представляют собой NGS, выполненный с возможностью приема первичных биологических данных и преобразования их в цифровую форму, например в форму одного или более файлов в формате FASTQ, содержащих риды данных генетической последовательности, и когда требуется дальнейшая обработка, например для определения, как сгенерированная последовательность объекта отличается от одной или более контрольных последовательностей, как описано в настоящем документе, и/или требуется подвергнуть ее результаты дальнейшей, например третичной, обработке.

[00521] В таком примере система 1 может быть выполнена с возможностью

обеспечения доступа одних или более лиц, например пользователей в виде первых, вторых и/или третьих лиц, к связанных локальным обрабатывающим ресурсам 100 и/или подходящим образом конфигурированным удаленным обрабатывающим ресурсам 300, связанным с ними, таким образом, чтобы обеспечить пользователю возможность выполнения одной или более количественной и/или качественной функции 152 обработки над сгенерированными и/или полученными данными. Например, в одной конфигурации, система 1 может содержать, например в дополнение к конвейерам для первичной 600 и/или вторичной 600 обработки, третий уровень обрабатывающих модулей 700, которые могут быть выполнены с возможностью реализации одной или более функций обработки над сгенерированными и/или полученными первичными и/или вторичными обработанными данными.

[00522] В частности, в одном из вариантов реализации, система 1 может быть выполнена с возможностью генерации и/или приема обработанных данных 111 генетической последовательности, которые были картированы удаленно или локально для генерации файла 116 определения вариантов, который затем может быть подвергнут дальнейшей обработке, например в системе 1, например в ответ на аналитический запрос 121 вторых и/или третьих лиц. Более конкретно, система 1 может быть выполнена с возможностью приема запросов на обработку от третьих лиц 121, а также с возможностью выполнения такой запрошенной третичной обработки 700 над сгенерированными и/или полученными данными. В частности, система 1 может быть выполнена с возможностью выработки и/или получения данных 111 генетической последовательности, с возможностью получения указанных данных генетической последовательности и выполнения их картирования 112, выравнивания 113 и/или сортировки 114a для получения одного или более файлов 116 определения вариантов (VCFs), и дополнительно система 1 может быть выполнена с возможностью реализации функции 700 третичной обработки над данными, например в отношении одного или более файлов VCF. Система 1 может быть выполнена с возможностью реализации любой формы третичной обработки 700 над сгенерированными и/или полученными данными, например подвергая их конвейерных функции 700 обработки, например для генерации данных 122a генома, данных 122b эпигенома, данных 122c метагенома и тому подобного, включая совместное генотипирование 122d, конвейеры анализа GATK 122e и/или MuTect2 122f. Кроме того, система 1 может быть выполнена с возможностью реализации дополнительного уровня обработки над сгенерированными и/или обработанными данными, включая, например, одно или более из неинвазивного пренатального тестирования (NIPT) 123a, реанимации и интенсивной терапии новых пациентов (N/P ICU) 123b, относящихся к раку видов диагностики и/или терапевтических методов 123c, проводимых в лаборатории исследований (LDT) 123d, применений 123e, относящихся к агробиологии (Ag Bio), или другой такой функции обработки, относящейся 123f к области здравоохранения.

[00523] Следовательно, в различных вариантах реализации, где первичный пользователь может осуществлять доступ к системе 1 или ее компонентам напрямую, и/или конфигурировать их напрямую, например посредством прямого доступа к ним, например через локальные вычислительные ресурсы 100, как описано в настоящем документе, система 1 также может быть выполнена с возможностью доступа к ней вторичными лицами, например подключенными к системе 1 через локальную сеть или внутреннее соединение 10 для конфигурации и запуска системы 1 в локальной среде. В дополнение, в некоторых вариантах реализации, показанных на фиг. 2B, система также может быть выполнена с возможностью доступа к ней и/или ее конфигурации третьими

лицами 121, например по связанной гибридной облачной сети 50, соединяющей третьих лиц 121 с системой 1, например через программный интерфейс приложения (API), доступный через один или более компонентов графического пользовательского интерфейса (ГПИ, GUI). Такой ГПИ может быть выполнен с возможностью обеспечения третьим лицам доступа к системе 1 и, с использованием API, с возможностью конфигурации различных компонентов системы, модулей, связанных конвейеров и других связанных функциональных элементов для генерации и/или обработки данных для запуска только тех компонентов системы, которые необходимы и/или полезны третьим лицам и/или запуск которых необходим или требуется.

[00524] Соответственно, в различных примерах система 1, представленная в настоящем документе, может быть выполнена с возможностью ее конфигурации перыичным, вторичным или третичным пользователем системы. В таком примере система 1 может быть выполнена с возможностью ее конфигурации пользователем и, таким образом, выполнения ее компонентов таким образом, чтобы обеспечить возможность исполнения одних, всех или выбранных аналитических ресурсов системы, например 152, над данными, сгенерированными, полученными системой, или другим образом переданными в нее, например первичным, вторичным или третичным пользователем, так что система 1 запускает только те части системы, которые необходимы или полезны для исполнения аналитических функций, запрошенных пользователем, для получения необходимых результатов. Например для этих и других целей API может содержаться в системе 1, причем API содержит или другим образом операционно связан с графическим пользовательским интерфейсом (ГПИ), включая операционное меню и/или соответствующий список вариантов функционирования системы, из которых пользователь может выбрать и/или другим образом конфигурировать и управлять системой и ее компонентами, как это требуется.

[00525] В таком примере меню и/или варианты функционирования системы ГПИ могут управлять одной или более выбираемыми пользователем операциями 600 первого уровня, включая: секвенирование 111, картирование 112, выравнивание 113, сортировку 114a, определение 115 вариантов и/или другие связанные функции 114 в соответствии с приведенной в настоящем документе информацией, например в отношении функций первичной и/или вторичной обработки, описанных в настоящем документе. Кроме того, при необходимости меню и/или варианты функционирования системы в ГПИ могут управлять одной или более операциями 700 второго уровня, включая: конвейер 122a генома, конвейер 122b эпигенома, конвейер 122c метагенома, конвейер 122d совместного генотипирования, конвейеры анализа GATK 122e и/или MuTect2 122f. Кроме того, при необходимости меню и варианты функционирования системы в ГПИ могут управлять одной или более выбираемыми пользователем операциями 800 третьего уровня, включая: неинвазивное пренатальное тестирование (NIPT) 123a, N/P ICU 123b, относящиеся к раку виды диагностики и/или терапевтические методы 123c, проводимые в лаборатории исследования (LDT) 123d, применения 123e, относящиеся к агробиологии (Ag Bio), или другие такие функции обработки, относящиеся 123f к области здравоохранения.

[00526] Соответственно, меню и варианты функционирования системы могут включать одну или более из функций первичной, вторичной и/или третичной обработки для обеспечения конфигурации системы и/или ее компонентов для реализации одного или более конвейеров для анализа данных, выбранных и сконфигурированных пользователем. В таком примере локальные вычислительные ресурсы 100 могут быть выполнены так, чтобы соответстовать удаленным вычислительным ресурсам и/или

зеркально отображать их, и/или аналогичным образом локальные ресурсы 200 хранения могут быть выполнены так, чтобы соответствовать удаленным ресурсам 400 хранения и/или зеркально отображать их, так чтобы различным компоненты системы могут запускаться и/или сгенерированные ими данные могут храниться как локально, так и удаленно бесшовным образом по выбору пользователя системы 1. Кроме того, в некоторых вариантах реализации к системе 1 может быть обеспечен доступ третьих лиц для запуска собственных протоколов 121a анализа над сгенерированными и/или обработанными данными, например посредством прогона через интерфейс искуственного интеллекта, выполненного с возможностью поиска корреляций между ними.

[00527] Система 1 может быть выполнена с возможностью реализации любой формы третичной обработки над сгенерированными и/или полученными данными. Следовательно, в различных вариантах реализации первичный, вторичный или третичный пользователь может осуществлять доступ и/или конфигурировать любой уровень системы 1 и различных ее компонентов напрямую, например посредством прямого доступа к вычислительным ресурсам 100, опосредованно, например через локальное сетевое соединение 10, или по связанной гибридной облачной сети 50, соединяющей указанное лицо с системой 1, например посредством соответствующим образом сконфигурированного API, имеющего соответствующие разрешения. В таком примере компоненты системы могут быть представлены как меню, такое как меню ГПИ с возможностью выбора, в котором пользователь может выбрать из всех различных вариантов обработки и хранения, которые необходимо исполнить над данными, представленными пользователем. Кроме того, в некоторых примерах пользователь может хагружать собственные системные протоколы для принятия и исполнения их системой для обработки различных данных образом, разработанным и выбранным пользователем. В таком примере ГПИ и связанный API обеспечат пользователю доступ к системе 1 и, с использованием API, возможность добавления и конфигурации различных компонентов системы, модулей, связанных конвейеров и других связанных функциональных элементов для генерации и/или обработки данных для запуска только тех компонентов системы, которые необходимы и/или полезны указанному лицу и/или запуск которых необходим или требуется.

[00528] Хотя приведенная выше в отношении фиг. 8 и 9 информация направлена на генерацию 110 данных, например на локальную генерацию 100 данных, использующую локальные вычислительные ресурсы 150; как описано выше в отношении фиг. 9, один или более из указанных выше отдельных модулей и их соответствующие функции и/или связанные ресурсы могут быть выполнены с возможностью удаленной реализации, например посредством удаленных вычислительных ресурсов 300, и также могут быть выполнены с возможностью их передачи в систему 1, например по протоколу бесшовной передачи по Интернет-соединению 30/50 на облачной основе, например посредством сконфигурированного подходящим образом механизма 120 получения данных.

[00529] Соответственно, в таком примере локальные вычислительные ресурсы 100 могут содержать механизм 120 получения данных, например выполненный с возможностью передачи и/или приема таких полученных данных и/или связанной информации. Например система 1 может содержать механизм 120 получения данных, выполненный таким образом, чтобы обеспечивать непрерывную обработку и/или хранение данных бесшовным и надежным образом, например по облачной или гибридной сети 30/50, в которой функции обработки распределены локально 100 и/или удаленно 300, и, аналогичным образом, в которой один или более результатов такой

обработки могут храниться локально 200 и/или удаленно 400, так что система бесшовным образом распределяет, каким из локальных и удаленных ресурсов должна быть направлена выбранная задача для обработки и/или хранения, вне зависимости от того, где физически расположены указанные ресурсы. Такая распределенная обработка, передача и получение могут включать одно или более из секвенирования 111, картирования 112, выравнивания 113, сортировки 114a, дублирующей маркировки 114c, удаления дубликатов, рекалибровки 114d, местного повторного выравнивание 114e, функции (функций) Рекалибровки 114f оценки качества оснований и/или функции сжатия 114g, а также функции 116 определения вариантов, как описано в настоящем документе. Когда обработанные данные хранятся локально 200 или удаленно 400 в любом состоянии, они в процессе могут быть выполнены доступными для локальных 100 или удаленных 300 обрабатывающих ресурсов, например для дальнейшей обработки перед повторной передачей и/или повторного хранения.

[00530] В частности, система 1 может быть выполнена с возможностью выработки и/или получения данных 111 генетической последовательности, может быть выполнена с возможностью обработки указанных данных 111 генетической последовательности локально 150 или передачи указанных данных по соответствующим образом конфигурированной облачной 30 или гибридной облачной 50 сети, например на удаленные обрабатывающие средства для удаленной обработки 300. Кроме того, после обработки данных система 1 может быть выполнена с возможностью хранения обработанных удаленно 400 или передачи их обратно для локального хранения 200. Соответственно, система 1 может быть выполнена с возможностью локальной или удаленной генерации и/или обработки данных, например когда этапы генерации и/или обработки могут формировать первый уровень функций 600 первичной и/или вторичной обработки, который может включать одно или более из: секвенирования 111, картирования 112, выравнивания 113 и/или сортировки 114a для получения одного или более файлов 116 определения вариантов (VCFs). Аналогично, система 1 может быть выполнена с возможностью локальной или удаленной генерации и/или обработки данных, например когда этапы генерации и/или обработки могут формировать второй уровень функций 700 третичной обработки, который может включать одно или более из генерации и/или получения данных согласно конвейеру 122a генома, программному конвейеру 122b эпигенома, конвейеру 122c метагенома, конвейеру 122d совместного генотипирования, конвейеру анализа GATK 122e и/или MuTect2 122f. Дополнительно, система 1 может быть выполнена с возможностью локальной или удаленной генерации и/или обработки данных, например когда этапы генерации и/или обработки могут формировать третий уровень функций обработки, как показано на фиг. 11, который может включать одно или более из генерации и/или получения данных, относящихся к и включающих: неинвазивное пренатальное тестирование (NIPT) 123a, N/P ICU 123b, относящиеся к раку виды диагностики и/или терапевтические методы 123c, различные проводимые в лаборатории исследования (LDT) 123d, применения 123e, относящиеся к агробиологии (Ag Bio), или другие такие функции обработки, относящиеся 123f к области здравоохранения.

[00531] В некоторых вариантах реализации, показанных на фиг. 8, 9 и 10, система 1 может также быть выполнена с возможностью обеспечения доступа одних или более лиц к системе и к передаче информации на связанные локальные обрабатывающие 100 и/или удаленные обрабатывающие ресурсы, или от них, а также к хранению информации локально 200 или удаленно 400 таким образом, который позволяет пользователю выбрать, какую информацию обработать и/или сохранить в системе 1. В таком примере

пользователь может не только решать, какие функции первичной, вторичной и/или третичной ообработки выполнять над сгенерированными и/или полученными данными, но также как указанные ресурсы могут быть реализованы и/или где результаты такой обработки будут храниться. Например, в одной конфигурации, пользователь может выбрать, будут ли данные сгенерированы локально или удаленно, или комбинированно, будут ли они подвергнуты вторичной обработке, и если да, то какими модулями вторичной обработки они будут обработаны, и/или какие ресурсы исполняют какие из указанных процессов, и далее может определить, будут ли сгенерированные или полученные данные далее подвергнуты третичной обработке, и если да, то какими модулями и/или какими уровнями третичной обработки 700 они будут обработаны, и/или какие ресурсы исполняют какую из указанных обработок, и, аналогичным образом, где результаты указанных обработок будут храниться для каждого этапа операций.

[00532] В частности, в одном варианте реализации пользователь может конфигурировать систему 1 таким образом, что генерация данных 111 генетической последовательности происходит удаленно, например посредством NGS, а вторичная обработка 600 данных происходит локально. В таком примере пользователь затем может определять, какие из функций вторичной обработки выполняются локально 100, например посредством выбора функций обработки, таких как картирование 112, выравнивание 113, сортировка 111 и/или получение файла VCF 116, в меню доступных вариантов обработки. Пользователь может затем выбирать, подвергнуть ли локально обработанные данные третичной обработке, и если да, то какие модули привести в действие для дальнейшей обработки данных, и будет ли такая третичная обработка происходить локально 100 или удаленно 300. Аналогичным образом, пользователь может выбрать различные варианты для различных уровней вариантов третичной обработки, а также где какие-либо сгенерированные и/или полученные данные должны храниться, локально 200 или удаленно 400, на любом заданном этапе или в любой заданный момент времени в процессе работы.

[00533] Более конкретно, первичный пользователь может конфигурировать систему для приема запросов на обработку от третьих лиц, причем третьи лица могут конфигурировать систему для выполнения запрошенной первичной, вторичной и/или третичной обработки над сгенерированными и/или полученными данными. В частности, пользователь или вторые или третье лицы могут конфигурировать систему 1 для выработки и/или получения данных генетической последовательности, локально 100 или удаленно 200, могут конфигурировать систему 1 для получения указанных данных генетической последовательности и выполнения их картирования, выравнивания и/или сортировки, локально или удаленно, для получения одного или более файлов определения вариантов (VCFs), и дополнительно могут конфигурировать систему 1 для выполнения функции третичной обработки над данными, например в отношении одного или более файлов VCF, локально или удаленно. Более конкретно, пользователь или другие лица могут конфигурировать систему 1 для выполнения любой формы третичной обработки над сгенерированными и/или полученными данными, а также определять где указанная обработка будет происходить в системе. Следовательно, в различных вариантах реализации пользователь в виде первых, вторых или третьих лиц 121 может осуществлять доступ и/или конфигурировать систему 1 и ее различные компоненты напрямую, например посредством прямого доступа к локальной вычислительной функции 100, через локальное сетевое соединение 10 или по связанной гибридной облачной сети 50, соединяющей указанное лицо 121 с системой 1, например через программный интерфейс приложения (API), доступный через один или более

компонентов графического пользовательского интерфейса (ГПИ). В таком примере пользователь в виде третьих лиц может осуществлять доступ к системе 1 и использовать API для конфигурации различных компонентов системы, модулей, связанных конвейеров и других связанных функциональных элементов для генерации и/или обработки данных для запуска только тех компонентов системы, которые необходимы и/или полезны третьим лицам и/или запуск которых необходим или требуется, и также назначать, какие вычислительные ресурсы будут обеспечивать запрошенную обработку и где будут храниться результирующие данные.

[00534] Соответственно, в различных примерах система 1 может быть выполнена с возможностью ее конфигурации первичным, вторичным или третичным пользователем системы, который может конфигурировать систему 1 для настройки ее компонентов таким образом, чтобы обеспечить возможность исполнения одних, всех или выбранных аналитических ресурсов системы над данными, которые непосредственно генерирует, обеспечивает их генерацию системой 1 или обеспечивает их переданчу в систему 1, например по связанной с ней сети, например посредством механизма 120 получения данных. Таким образом система 1 может быть выполнена с возможностью запуска только тех частей системы, которые необходимы или полезны для выполнения аналитической функции, желаемой и/или запрошенной лицами, которые осуществляют запрос. Например для этих и других целей может быть обеспечен API, который содержит операционное меню и/или соответствующий список системных функциональных вариантов ГПИ, из которых пользователь может выбирать для конфигурирования и управления системой, как это требуется. Кроме того, в частных вариантах реализации система 1 может быть доступна третьим лицам, например государственным регулирующим структурам, например Федеральному управлению по лекарственным средствам (FDA) 70b, или обеспечивать третьим лицам возможность сопоставлять, компилировать и/или осуществлять доступ к базе данных генетической информации, извлеченной или другим образом полученной и/или компилированной системой 1 для формирования базы данных 70a электронных медицинских записей (EMR) и/или для обеспечения государственного доступа к системе и/или надзора за системой, например для FDA для Оценки разработки лекарственных средств. Система 1 может также накапливать, компилировать и/или аннотировать данные 70c и/или обеспечивать доступ других высокоуровневых пользователей к ним.

[00535] Соответственно, в различных вариантах реализации, как можно видеть на фиг. 13, обеспечено гибридное облако 50, выполненное с возможностью соединения локальных вычислительных ресурсов 100 и/или ресурсов 200 хранения с удаленными вычислительными ресурсами 300 и/или ресурсами 400 хранения, например когда локальные и удаленные ресурсы отделены друг от друга дистальным образом, в пространстве, географически и тому подобное. В таком примере локальные и дистальные ресурсы могут быть выполнены с возможностью обмена информациией друг с другом так, чтобы разделять информацию, например цифровые данные, бесшовным образом друг с другом. В частности, локальные ресурсы могут быть выполнены с возможностью реализации одного или более типов обработки данных, например перед передачей по гибридной сети 50, а удаленные ресурсы могут быть выполнены с возможностью реализации одного или более типов дальнейшей обработки данных.

[00536] Например, в одной конкретной конфигурации система 1 может быть выполнена таким образом, что функция 152 генерации и/или анализа сконфигурирована для ее выполнения локально 100 посредством локальных вычислительных ресурсов, например для выполнения функции первичной и/или вторичной обработки для генерации

и/или обработки данных генетической последовательности, как описано в настоящем документе. Дополнительно, в различных вариантах реализации локальные ресурсы могут быть выполнены с возможностью реализации одной или более функций третичной обработки над данными, такой как одно или более из анализа генома, экзома и/или эпигенома, или анализа рака, микробиома и/или другие виды анализа с обработкой ДНК/РНК. Кроме того, когда предполагается передача таких данных, например на удаленные вычислительные ресурсы 300 и/или ресурсы 400 хранения, данные могут быть преобразованы, например подходящим образом сконфигурированным преобразователем 151, который может быть выполнен с возможностью индексации, преобразования, сжатия и/или шафрования данных, например перед передачей по гибридной сети 50.

[00537] В частных примерах, например когда сгенерированные и обработанные данные передают на удаленные вычислительные ресурсы 300 для дальнейшей обработки, такая обработка может иметь глобальную природу и может включать прием данных из множества локальных вычислительных ресурсов 100, сопоставление таких множеств данных, аннотирование данных и их сравнение, например для интерпретации данных, определения их тенденций, анализ тенденций различных биомаркеров и способствование развитию методов диагностики, терапевтических методов и/или методов профилактики. Соответственно, в различных примерах удаленные вычислительные ресурсы 300 могут быть выполнены в виде ядра обработки данных, например где данные из множества различных источников могут передаваться, обрабатываться и/или храниться, например посредством доступа к ним локальных вычислительных ресурсов 100. Более конкретно, удаленное ядро 300 обработки может быть выполнено с возможностью приема данных из множества ресурсов 100, обработки данных и распределения обработанных данных обратно на множество различных локальных ресурсов 100 для обеспечения возможности взаимодействия между исследователями и/или ремурсами 100. Такое взаимодействие может включать различные протоколы разделения данных и дополнительно может включать подготовку данных, подлежащих передаче, например обеспечивая пользователю системы 1 возможность выбора между различными протоколами безопасности и/или настройками приватности для управления тем, как данные будут подготовлены к передаче.

[00538] В одном конкретном примере, как показано на фиг. 11, обеспечены локальные вычислительные ресурсы 100 и/или ресурсы 200 хранения, например в местоположении пользователя. Вычислительные ресурсы 100 и ресурсы 200 хранения могут быть соединены с ресурсами 121 для генерации данных, такими как NGS и/или секвенатор на микросхеме, как описано в настоящем документе, например по прямому или внутреннему соединению 10, причем секвенатор 121 выполнен с возможностью генерации данных генетической последовательности, например файла в формате FASTQ. Например, секвенатор 121 может являться частью и/или может быть расположен в том же устройстве, как и секвенатор вычислительных ресурсов 100 и/или блока хранения 200 для того, чтобы иметь прямое соединение с ним с возможностью обмена информацией и/или управления, или секвенатор 121 и вычислительные ресурсы 100 и/или ресурсы 200 хранения могут быть частью отдельных друг от друга устройств, но могут быть расположены на одном объекте и, таким образом, могут быть соединены посредством проводного или внутреннего 10 соединения. В некоторых примерах секвенатор 121 может быть расположен на другом объекте, отдельном от объекта с вычислительными ресурсами 100 и/или ресурсами 200 хранения, и, таким образом, может быть соединен по сети Интернет 30 или гибридному облачному соединению 50.

[00539] В таких примерах данные генетической последовательности могут быть обработаны 100 и сохранены локально 200 перед их передачей посредством подходящим образом сконфигурированного преобразователя 151, или сгенерированные данные последовательности могут быть переданы непосредственно на один или более преобразователей 151 и/или анализаторов 152, например по подходящим образом сконфигурированному локальному соединению 10, внутреннему 30 или гибридному облачному соединению 50, как описано выше, например перед выполнением обработки локально. В частности, как и ресурсы 121 для генерации данных, преобразователь 151 и/или анализатор 152 могут являться частью и/или могут быть расположены в том же устройстве, как и те же элементы вычислительных ресурсов 100 и/или блока хранения 200 для того, чтобы иметь прямое соединение с ним с возможностью обмена информацией и/или управления, или преобразователь 151 и/или анализатор 152 и вычислительные ресурсы 100 и/или ресурсы 200 хранения могут быть частью отдельных друг от друга устройств, но могут быть расположены на одном объекте и, таким образом, могут быть соединены посредством проводного или внутреннего 10 соединения. В некоторых примерах преобразователь 151 и/или анализатор 152 могут быть расположены на другом объекте, отдельном от объекта с вычислительными ресурсами 100 и/или ресурсами 200 хранения, и, таким образом, могут быть соединены по сети Интернет 30 или гибридному облачному соединению 50.

[00540] В таких примерах, как показано на фиг. 13, преобразователь 151 может быть выполнен с возможностью подготовки данных к передаче перед проведением анализа или после проведения анализа, например посредством подходящим образом конфигурированных вычислительных ресурсов 100 и/или анализатора 152. Например, анализатор 152 может выполнять функцию вторичной и/или третичной обработки над даннымит, как описано в настоящем документе, например для анализа сгенерированных данных последовательности в отношении определения их геномных и/или экзомных характеристик 152a, их эпигеномных особенностей 152b, каких-либо различных интересующих маркеров ДНК и/или РНК и/или индикаторов рака 152c, и их взаимосвязи с одним или более микробиомов 152d, а также одним или более вторичным и/или третичным процессами, как описано в настоящем документе. Как было указано, сгенерированные и/или обработанные данные могут быть преобразованы, например посредством подходящим образом конфигурированного преобразователя 151, например перед передачей через систему 1 от одного ее компонента к другому, например по прямому, локальному 10, Интернет 30 или гибридному облачному 50 соединению. Такое преобразование может включать одно или более из преобразования 151d, такого, при котором данные преобразуют из одной формы в другую; обеспечение понимания 151c, включая кодирование, декодирование и/или извлечение данных другим образом из недоступной пониманию формы, и преобразование их в доступную для понимания форму, или из одной доступной для понимания формы в другую; индексация 151b, например включая компилирование и/или сопоставление сгенерированных данных от одних или более ресурсов и обеспечение возможности их поиска и/или установления местоположения, например посредством сгенерированных индексов; и/или шифрование 151a, например, создание блокируемого и разблокируемого защищенного паролем набора данных, например перед передачей по сети Интернет 30 и/или по гибридному облаку 50.

[00541] Следовательно, в этих и/или других таких примерах гибридное облако 50 может быть выполнено с возможностью обеспечения возможности бесшовной и защищенной передачи данных по всем компонентам системы, например, когда

гибридное облако 50 выполнено с возможностью разрешения различным пользователям системы конфигурировать ее составляющие части и/или саму систему для удовлетворения потребностей пользователя в отношении исследовательских, диагностических, терапевтических и/или профилактических открытий и/или разработок. В частности, гибридное облако 50 и/или различные компоненты системы 1 могут быть выполнены с возможностью функционального соединения с совместимыми и/или соответствующими интерфейсами API, которые выполнены с возможностью обеспечения пользователю возможности удаленного конфигурирования различных компонентов системы 1 для развертывания требуемых ресурсов нужным образом, причем локальным, удаленным или комбинированным способом, например на основе потребностей системы и особенностей выполняемых видов анализа, обеспечивая при этом обмен информацией в защищенной среде с возможностью шифрования. Другой приведенный в качестве примера вариант реализации гибридной облачной системы, представленной в настоящем документе, показан на фиг. 12.

[00542] Например, как показано на фиг. 10-12 и, в частности, на фиг. 15, система 1 может быть многоуровневой и/или мультиплексированной платформой для биоаналитической обработки, которая содержит уровни блоков обработки, каждый из которых имеет один или более конвейеров для обработки, которые могут быть развернуты систематически, одновременно и/или последовательно для обработки генетической информации, начиная со стадии 400/500 ее первичной обработки для получения данных генетической последовательности, например в виде одного или более файла в формате FASTQ; к стадии 600 вторичной обработки для получения одного или более файла определения вариантов; и далее к выбору одного или более файлов определения вариантов или других связанных обработанных данных и выполнению одной или более других операций над ними, например для выполнения с ними одной или более из диагностических, профилактических и/или терапевтических процедур, например в ответ на запрос 121 третьих лиц и/или в ответ на данные, отправленные третьими лицами 121. Такая дальнейшая обработка может включать различные конвейерные протоколы 700, например выполненные с возможностью исполнения аналитических функций над определенными данными генетической вариации одного или более субъектов, включая аналитику генома, эпигенома, метагенома и/или генотипирования, например на одном уровне, и/или различные типы диагностики заболеваний и/или исследовательские протоколы 800, которые могут включать одно или более из NIPT, NICU, раковых, LDT, биологических, агробиологических применений и тому подобное. В частности, система 1 может быть также выполнена с возможностью приема и/или передачи различных данных 900, относящихся к процедурам и процессам, описанным в настоящем документе, например, имеющим отношение к данным электронных медицинских записей (EMR), данным испытаний и/или структурирования Федерального управления по лекарственным средствам (США), относящимся к аннотации данным и тому подобному. Такие данные могут быть полезными ввиду предоставления пользователю возможности создания и/или осуществления доступа к сформированным медицинским, диагностическим, терапевтическим и/или профилактическим методам, разработанным посредством использования системы 1 и/или доступным с ее помощью.

[00543] Следовательно, одна или более, например все, из указанных функций таким образом могут быть выполнены локально, например на месте 10, в облаке 30 или посредством управляемого доступа через гибридное облако 50. В таком примере создана среда разработки, которая обеспечивает пользователю возможность управления

функциональными особенностями системы для соответствия его или ее индивидуальным нуждам и/или для обеспечения доступа к ней для других лиц, которым требуются те же или подобные результаты. Следовательно, различные компоненты, процессы, процедуры, средства, уровни и иерархии системы могут быть выполнены с возможностью конфигурирования через интерфейс ГПИ, который позволяет пользователю выбирать, какие компоненты системы использовать, над какими данными, в какое время и в каком порядке в соответствии с установленными пользователем требованиями и протоколами для генерации соответствующих данных и соединений между данными, которые могут быть безопасно переданы по всей системе, локально или удаленно. Как было указано, эти компоненты могут быть выполнены с возможностью бесшовного обмена данными между собой независимо от местоположения и/или вида соединения, например посредством возможности конфигурирования, чтобы выполнять одни и те же или подобные процессы одинаковым или аналогичным образом, например путем использования соответствующих интерфейсов API, рассредоточенных по всей системе, применение которых позволяет различным пользователям конфигурировать различные компоненты для исполнения различных процедур аналогичным образом.

[00544] Например, API может быть определен в заголовочном файле в отношении процессов, которые должны исполняться каждым конкретным компонентом системы 1, причем заголовок описывает функциональные возможности и определяет, как вызывать функции, например параметры, которые передаются, принимаемые входные и передаваемые выходные данные, и способ, каким это происходит, что поступает и каким образом, что выдается и каким образом, что возвращается и каким образом. Например, в различных вариантах реализации один или более компонентов и/или их элементов, которые могут формировать один или более конвейеров одного или более уровней системы, могут быть выполнены с возможностью их конфигурирования, например с помощью инструкций, вводимых пользователем и/или одним или более приложений вторых и/или третьих лиц. Эти инструкции могут передаваться в систему через соответствующие API, которые обмениваются информацией с одним или более различных драйверов системы, указывая драйверу (драйверам), какие части системы, например какие модули и/или какие процессы в них, нужно активировать, когда и в каком порядке, с учетом предварительно выбранной конфигурации параметров, которая может быть определена с помощью интерфейса, который может быть выбран пользователем, например ГПИ.

[00545] Как описано выше, один или более драйверов ПДП системы 1 могут быть выполнены с возможностью работы соответствующим образом, например на уровне ядра каждого компонента и системы 1 в целом. В таком примере одно или более из предусмотренных ядер может иметь свой собственный базовый API очень низкого уровня, который обеспечивает доступ к аппаратному обеспечению и функциям различных компонентов системы 1, так что обеспечен доступ к соответствующим реестрам и модулям для конфигурирования и управления процессами и тем, как они исполняются в системе 1. В частности, сверху этого слоя может быть сформирован виртуальный слой служебных функций для формирования строительных блоков, которые используются для множества функций, которые отправляют файлы ниже в ядро (ядра) и получают обратно результаты, кодируют, шифруют и/или передают соответствующие данные и далее выполняют над ними функции более высокого уровня. Поверх этого слоя может быть сформирован дополнительный слой, использующий указанные служебные функции, которые могут быть уровня API, с которым может

взаимодействовать пользователь, причем этот слой может быть выполнен с возможностью функционирования в основном для конфигурирования системы 1 в целом или ее составляющих частей, скачивая файлы и загружая результаты, при этом файлы и/или результаты могут быть переданы по всей системе, локально или глобально.

[00546] Такая конфигурация может включать обмен информацией с реестрами, а также выполнение обращений к функциям. Например, как описано в настоящем документе выше, одно или более обращений к функциям, необходимых и/или полезных для выполнения этапов, например последовательно, для выполнения картирования, выравнивания, сортировки и/или определения вариантов, или другой вторичной и/или третичной функции, как описано в настоящем документе, могут быть осуществлены в соответствии с операциями аппаратного обеспечения и/или соответствующими алгоритмами для генерации необходимых процессов и выполнения требуемых этапов.

[00547] В частности, так как в определенных вариантах реализации одна или более из указанных операций могут быть основаны на одной или более структурах, возможно потребуется построить различные структуры, необходимые для осуществления указанных операций. В связи с этим потребуется обращение к функции, которая выполняет данные действия, причем указанное обращение к функции приведет к построению требуемой структуры для выполнения операции, и поэтому указанное обращение примет имя файла, где хранятся файлы параметров структуры, и затем сгенерирует один или более файлов данных, которые содержат и/или конфигурируют требуемую структуру. Другое обращение к функции может быть предназначено для загрузки структуры, которая была сгенерирована посредством соответствующего алгоритма, и передачи ее в память на микросхеме и/или в систему 1 и/или помещения ее в нужное место, где ее нахождение предполагается аппаратным обеспечением. Конечно, потребуется скачивание различных данных в микросхему и/или осуществление передачи в системный генератор другим образом, а также выполнение различных других выбранных функций системы 1, и эти функции может выполнять диспетчер конфигураций, например посрежством загрузки всего, что необходимо для того, чтобы модули конвейеров уровней платформ микросхемы и/или системы в целом выполняли свои функции, в память, находящуюся на микросхеме и/или системе или прикрепленную или иным образом связанную с микросхемой и/или системой.

[00548] Кроме того, как показано на фиг. 16, API может быть выполнен так, чтобы обеспечивать одной или более микросхемам системы 1 возможность взаимодействия с печатной платой секвенатора 121, вычислительного ресурса 100/300, преобразователя 151, анализатора 152, интерпретатора 310, коллаборатора 320 или другого компонента системы, когда они включены в нее, для приема файла в формате FASTQ и/или других файлов сгенерированной и/или обработанной генетической последовательности непосредственно от секвенатора или другого обрабатывающего компонента, например немедленно после их генерации и/или обработки, и затем передачи этой информации в диспетчер конфигураций, который затем направляет данную информацию в соответствующие банки памяти в аппаратном и/или программном обеспечении, которые делают эту информацию доступной соответствующим модулям аппаратного обеспечения, программного обеспечения и/или системе в целом так, чтобы они могли выполнять назначенные им функции над этой информацией для определения оснований, картирования, выравнивания, сортировки и т.д. образца ДНК/РНК относительно контрольного генома и/или исполнения на ней связанных операций вторичной и/или третичной обработки.

[00549] Соответственно, в различных вариантах реализации может быть включен

интерфейс уровня клиентов (CLI), которй может обеспечивать пользователям возможность непосредственного обращения к одной или более из этих функций. В различных вариантах реализации CLI может быть приложением программного обеспечения, например имеющим ГПИ, которое выполнено с возможностью конфигурирования доступности и/или использования системных приложений аппаратного обеспечения и/или различного другого программного обеспечения. Таким образом, CLI может быть программой, которая принимает инструкции, например аргументы, и делает функциональные возможности доступными просто путем обращения к прикладной программе. Как указано выше, CLI может быть основан на командной строке или ГПИ (графическом пользовательском интерфейсе). Уровень команд для командной строки ниже уровня ГПИ, причем ГПИ содержит диспетчер файлов на базе окон, имеющий функциональные блоки, выполненные с возможностью выбора щелчком мыши, которые изображают, какие модули, какие конвейеры, какие уровни каких платформ будут использованы, а также параметры их использования. Например, во время работы, если предписано, CLI будет находить ссылку, будет определять, нужно ли сгенерировать хэш-таблицу и/или индекс, или, если они уже сгенерированы, устанавливать, где они хранятся, и управлять загрузкой сгенерированной хэш-таблицы и/или индекса и т. д. Инструкции этих видов могут появляться в виде пользовательских вариантов на ГПИ, которые пользователь может выбирать для выполнения связанными микросхемой (микросхемами) и/или системой 1.

[00550] Кроме того, может быть включена библиотека, которая может содержать уже существующие редактируемые файлы конфигурации, например файлы, ориентированные на типичные выбираемые пользователем функциональные возможности аппаратного обеспечения и/или связанного программного обеспечения, например относящиеся к анализу части или всего генома и/или белка, например для различных видов анализа, таких как анализ персональных медицинских историй и родословной, или диагностика заболеваний, или открытие новых лекарственных средств, терапевтика и/или один или более других аналитических методов и т.д. Параметры этих типов могут быть предварительно установлены, например для выполнения таких видов анализа, и могут быть сохранены в библиотеке. Например, если описанная в настоящем документе платформа используется, например, для исследований в области NIPT, NICU, рака, LDT, AgBio и связанных исследований на собирательном уровне, настоящие параметры могут быть сконфигурированы иным образом по сравнению с тем, если бы платформа была направлена просто на проведение геномного и/или генеалогического исследования, например на индивидуальном уровне. См., например, фиг. 11.

[00551] Более конкретно, в случае определенной диагностики индивида точность может быть важным фактором и, таким образом, параметры системы могут быть установлены так, чтобы обеспечивать повышенную точность, хотя и в обмен на возможное снижение скорости. Однако в других относящихся к геному случаев применения скорость может быть основным определяющим фактором, и, таким образом, параметры системы могут быть установлены для обеспечения максимальной скорости, однако это может привести к некоторому снижению точности. Соответственно, в различных вариантах реализации часто используемые настройки параметров для выполнения разных задач могут быть предварительно установлены в библиотеку для облегчения их использования. Такие настройки параметров могут также включать в себя необходимые приложения программного обеспечения и/или конфигурации аппаратного обеспечения, используемые при работе системы 1. Например, библиотека может содержать код, который исполняет API, и может также содержать файлы образов,

сценарии и любую другую вспомогательную информацию, необходимую для работы системы 1. Следовательно, библиотека может быть выполнена с возможностью компиляции программного обеспечения для исполнения API, а также различных исполняемых объектов.

[00552] Кроме того, как показано на фиг. 12-14, система может быть выполнена таким образом, что один или более системных компонентов могут выполняться удаленно, например когда компонент системы выполнен с возможностью исполнения одной или более функций сравнения над данными, например функции 310 интерпретации и/или функции 320 совместной работы. Например, когда к данным применяют протокол интерпретации, протокол 312 интерпретации может быть выполнен с возможностью анализа и заключения выводов о данных и/или определения различных взаимосвязей в отношении них, могут также выполняться один или более других аналитических протоколов, включающих аннотирование 311 данных, выполнение диагностики 313 над данными и/или анализ данных для определения присутствия или отсутствия одного или боле биомаркеров 314. Кроме того, при выполнении протокола совместной работы система 1 может быть выполнена с возможностью обеспечения электронного форума, где можно делиться 321 данными, причем протокол разделения данных может включать настройки безопасности 324 и/или приватности 322, которые могут быть выбраны пользователем и обеспечивают возможность шифрования данных и/или защищиты их паролем, чтобы можно было скрыть идентификаторы и источники данных от пользователя системы 1. В конкретных примерах система 1 может быть выполнена с возможностью обеспечения анализатору 121 третьих лиц возможности выполнения виртуального моделирования над данными. Кроме того, после генерации данные, интерпретированные и/или подвергнутые одному или более совместно осуществляемым видам анализа, могут быть сохранены удаленно 400 или локально 200, чтобы сделать их доступными для удаленных 300 или локальных 100 вычислительных ресурсов, например для дальнейшей обработки и/или анализа.

[00553] Соответственно, с учетом приведенного в настоящем документе раскрытия, согласно одному аспекту предложено устройство для выполнения одной или более из множества операций при выполнении операции анализа последовательности генома. В некоторых примерах устройство может представлять собой вычислительное решение в виде «рабочего места», имеющее набор микросхем, который связан с платой PCIe, которая сама по себе может быть вставлена в вычислительное устройство, чтобы тем самым быть связанной с одним или более внутренних центральных процессоров (ЦП), графических процессоров (ГП) и/или связанными запоминающими устройствами. В частности, вычислительное устройство, обрабатывающие блоки, связанные запоминающие устройства и/или связанная(ые) плата(ы) PCIe, имеющие один или более наборы микросхем ППВМ/ИССН, как описано в настоящем документе, могут быть связаны с возможностью обмена информацией друг с другом и могут быть расположены внутри корпуса, например в виде коробочной версии, которая типична в данной области техники и может быть выполнена с возможностью использования в качестве рабочего места и/или она может быть обеспечена и/или выполнена с возможностью использования в серверной стойке. В других вариантах реализации наборы микросхем и/или связанная (ые) соединительная(ые) плата(ы) могут быть связаны внутри устройства секвенирования нового поколения с образованием с ним одного блока.

[00554] Однако, в различных примерах одна или более интегриральных схем, раскрытых в настоящем документе, могут быть обеспечивать и быть выполненными с возможностью доступа к ним посредством интерфейса на облачной основе, например

как показано на фиг. 12-15. В некоторых примерах коробочная версия может быть выполнена с возможностью удаленного доступа к ней, например когда конфигурация в виде коробочной версии выполнена портативной по отношению к облаку. Однако в других случаях одна или более из интегральных схем, раскрытых в настоящем документе, могут быть частью серверной стойки, например когда доступная на сервере система выполнена с возможностью удаленного доступа к ней, например посредством облака 50.

[00555] Например, в одном варианте реализации сервер, имеющий одно или более, например множество, из ядер ЦП и/или ГП и связанных запоминающих устройств, может быть связан с одной или более ППВМ/ИССН, раскрытых в настоящем документе. В частности, в одном варианте осуществления может быть обеспечена коробочная версия/сервер, содержащая(ий) от 18 до 24 или более ядер ЦП, имеющая(ий) твердотельные накопители (SSD), ОЗУ 128×8, и одну или более BioIT-систем ППВМ/ИССН, к которой(ому) обеспечена возможность удаленного доступа. В одном примере одна или более ППВМ могут быть выполнены с возможностью изменения их конфигурации, например частичного изменения их конфигурации, между одним или более различных этапов конвейера для геномного анализа. В других примерах серверная система может содержать до 36 ядер ЦП/ГП и приблизительно 972 Гб RAM, которая может быть связана приблизительно с 8 ППВМ, выполненных с возможностью конфигурации как описано в настоящем документе.

[00556] Более конкретно, предложенные ППВМ могут быть выполнены с возможностью специализации на выполнении одной или более вычислительноемких операций в конвейере BioIT, например когда предусмотрена одна ППВМ, которая специально предназначена для выполнения операции картирования, и предусмотрена другая ППВМ, которая выполнена с возможностью реализации операции выравнивания, хотя, в некоторых примерах, может быть предусмотрена одна ППВМ, которая выполнена с возможностью по меньшей мере частичного изменения ее конфигурации между выполнением обоих операций, картирования и выравнивания. В число других операций в конвейере, которые могут осуществляться специализированными FPGA, могут входить выполнение операции HMM, операции местного повторного выравнивания, например операции Смита-Ватермана, и/или различных других операций определения вариантов. Аналогичным образом различные операции конвейера могут быть выполнены с возможностью исполнения одного или более из связанных ЦП/ГП системы. Такие операции могут быть одной или более вычислительноемкими операциями конвейера, например для выполнения сортировки, удаления дубликатов и других операций определения вариантов. Следовательно, системы может быть выполнена с возможностью осуществления комбинации операций, частично с помощью ЦП и частично с помощью аппаратного обеспечения, такого как ППВМ/ИССН системы.

[00557] Соответственно, в различных вариантах осуществления системы могут быть предусмотрены примеры различных ЦП и жестко смонтированных интегральных схем для выполнения специализированных функций анализа генома при помощи конвейера, предложенные в настоящем документе. Например, могут быть предусмотрены различные экземпляры ППВМ для выполнения специализированных операций геномного анализа, например экземпляр FPGA для выполнения картирования, другой экземпляр для выполнения выравнивания, еще один для выполнения местного повторного выравнивания и/или других операций Смита-Ватермана, другой для выполнения операций HMM и тому подобное. Аналогичным образом могут быть предусмотрены различные экземпляры ЦП/ГП для выполнения специализированных

операций геномного анализа, например экземпляры ЦП/ГП для выполнения сортировки, удаления дубликатов, сжатия, различных операций определения вариантов и тому подобного. В таких примерах могут быть обеспечены одно или более связанных запоминающих устройств, например между различными вычислительными этапами конвейера, для приема полученных данных, полученных в результате вычисления, компиляции и обработки по всей системе, например между различными экземплярами ЦП и/или ППВМ. Кроме того, следует отметить, что размер различных экземпляров ЦПУ и/или ППВМ может меняться в зависимости от вычислительных потребностей системы и может изменяться в диапазоне от малого до среднего, крупного и очень крупного, аналогичным образом может меняться количество экземпляров ЦП/ГП и ППВМ/ИССН.

[00558] Следовательно, система может также содержать диспетчер рабочих потоков, который выполнен с возможностью планирования и направления перемещения данных по всей системе и от одного экземпляра к другому или из одной памяти в другую. В некоторых примерах запоминающее устройство может представлять собой специализированное запоминающее устройство, специфичное для конкретных экземплярах, а в других примерах запоминающее устройство может быть выполнено с возможностью быть эластичными и, таким образом, способными переключаться с одного экземпляра на другой, например как эластичное блочное запоминающее устройство хранения. В других примерах запоминающее устройство может быть неспецифичным для экземпляров и, следовательно, быть выполненным с возможностью соединения с возможностью обмена информацией с множеством экземпляров, например для эластичного хранения файлов.

[00559] Кроме того, диспетчер рабочих потоков может являться специализированным экземпляром, например ядром ЦП/ГП, которое специально предназначено для и/или выполнено с возможностью определения, какие задания необходимо выполнять, и когда и какие ресурсы будут использованы при выполнении указанных заданий, а также постановки в очередь заданий и направления их от ресурсов к ресурсам. Диспетчер рабочих потоков может содержать или может быть иным образом выполнен как оценщик загрузки и/или может образовать эластичный узел управления, представляющий собой специализированный экземпляр, который может исполняться процессором, например ядром ЦП с четырьмя ядрами, или может исполняться без множества ядер. В различных примерах диспетчер рабочих потоков может иметь базу данных, соединенную с ним, которая может быть выполнена с возможностью управления всеми заданиями, которые должны быть обработаны, обрабатываются или были обработаны. Следовательно, диспетчере может быть выполнен с возможностью обнаружения и управления потоками данных по всей системе, определения того, как назначить системные ресурсы, и когда привести в действие больше ресурсов.

[00560] Как указано выше, в определенных примерах может быть предусмотрено решение для рабочего места, где система содержит множество серверов с X ядрами ЦП, которые подают данные в ППВМ размера Z, где X, Y и Z являются числами, которые могут меняться в зависимости от потребностей обработки системы, но их следует выбирать и/или иным образом конфигурировать, чтобы они были оптимальными. Например, типичные конфигурации системы не оптимизированы для выполнения BioIT-операций системы, описанной в настоящем документе. В частности, определенные конфигурации системы не оптимизированы таким образом, чтобы максимально увеличивать поток данных от различных экземпляров ЦП/ГП в различные интегральные схемы, такие как ППВМ, системы. Более конкретно, архитектура системы

может быть выполнена таким образом, чтобы аппаратное обеспечение ЦП/ППВМ работало оптимально эффективным образом для поддержания платформ обоих экземпляров постоянно занятыми в течение работы.

[00561] Следовательно, хотя в целом хорошо, когда ППВМ имеет большие возможности, способность обработки большого количества данных может не быть эффективной, если в систему подают не достаточное количество данных, которые необходимо обработать. Например, в различных примерах ядра четырехъядерного ЦП могут быть выполнены с возможностью подавать данные в ППВМ среднего размера, например 2,5X. Однако когда ЦП не генерирует достаточно работы для занятости и/или полного использования ППВМ, конфигурация не будет идеальной. В таких конфигурациях экземпляр(ы) ЦП не вырабатывают достаточно работы для поддержания занятости доступных ППВМ. Таким образом, в настоящем документе предложена архитектура способ ее осуществления, которые выполнены так, что система работает таким образом, что программное обеспечение/аппаратное обеспечение ЦП/ППВМ работают эффективно для обеспечения того, что имеющиеся ЦП оптимально подают данные в доступные ППВМ таким образом, чтобы обеспечивать постоянную занятость обеих имеющихся платформ во время работы. Соответственно, обеспечение доступа к такой системе с помощью облака обеспечит, чтобы подаваемое в нее множество данных помещались в очередь диспетчером рабочих потоков и направлялись в определенные ресурсы ЦП/ППВМ, которые выполнены с возможностью и имеют конфигурацию для приема и обработки данных оптимальным эффективным образом, например когда ЦП обрабатывает менее вычислительноемкие данные, а ППВМ выполняет вычислительноемкие задачи, и запоминающие устройства обеспечивают хранение данных между раздличными этапами процедуры и/или между различными типами экземпляров и экземплярами, тем самым устраняя задержку между экземплярами. В частности, во время картирования и выравнивания ЦП/ГП используются очень мало, поскольку из-за интенсивного характера вычислений эти задачи выполнены с возможностью их осуществления аппаратными реализациями. Аналогичным образом во время определения вариантов указанные задачи могут быть разделены таким образом, чтобы они были примерно одинаково распределены между экземплярами ЦП/ППВМ в виде их задач, например когда операции HMM и Смита-Ватермана могут быть выполнены аппаратным обеспечением, а различные другие операции могут быть выполнены программным обеспечением, исполняемым на одном или более экземплярах ЦП/ГП.

[00562] Соответственно, параметры архитектуры, указанные в настоящем документе, необязательно ограничены однажды установленной архитектурой, наоборот, система выполнена с возможностью обладания большей гибкостью в организации ее реализаций и опирается на диспетчер рабочих потоков для определения, какие экземпляры активны, когда, как и как долго, и указывает, какие вычисления выполняются на каких экземплярах. В частности, архитектуры на облачной основе, описанные в настоящем документе, например, показанные на приложенных чертежах, показывают, что различные известные недостатки предыдущих вариантов архитектуры могут привести к проявлениям неэффективности, которые можно преодолеть за счет гибкого предоставления большему количеству ядер ЦП/ГП доступа к различным экземплярам аппаратного обеспечения, например ППВМ, которые организованы более целенаправленным образом, чтобы быть в состоянии специально назначать надлежащий экземпляр для выполнения предназначенных функций для оптимизации за счет реализации в таком формате, так что большая часть доступных экземпляров ЦП может

быть постоянно занята, вырабатывая результирующие данные, которые могут быть оптимальным образом поданы в доступные экземпляры ППВМ для обеспечения постоянной занятости выбранных экземпляров ППВМ. Таким образом, желательно обеспечить структурированную архитектуру, которая максимально эффективна и постоянно занята. Следует отметить, что конфигурации, в которых слишком мало ЦП снабжают слишком много ППВМ, так что одна или более ППВМ используются не в полной мере, являются неэффективными и их следует избегать.

[00563] В одном варианте осуществления архитектура может быть выполнена с возможностью виртуального включения нескольких различных слоев, таких как первый уровень, имеющий первое количество X ядер ЦП, например, от 4 до приблизительно 30 ядер ЦП, и второй уровень, имеющий от 1 до 12 экземпляров ППВМ, причем размер может меняться в диапазоне от малого до среднего или большого и т.д. Может также быть включен третий уровень ядер ЦП и/или четвертый уровень дополнительных ППВМ и т.д. Следовательно, так как в облаке имеется множество доступных экземпляров, например экземпляров, которые просто содержат ЦП или ГП и/или ППВМ, или их комбинации, например на одном или более уровнях, описанных в настоящем документе. Соответственно, подобным образом архитектура может быть организована так, чтобы наиболее интенсивные специальные вычислительные функции выполнялись экземплярами аппаратного обеспечения, а те функции, которые могут быть исполнены с помощью ЦП, направлялись на подходящий ЦП на подходящем уровне в целях общей обработки.

[00564] Например, архитектура может быть выполнена с возможностью максимального повышения эффективности и сокращения задержки путем комбинирования различных экземпляров на различных виртуальных уровнях. В частности, множество, например значительная часть и/или все, экземпляров ЦП уровня 1 могут быть выполнены с возможностью снабжения различных экземпляров ППВМ уровня 2, например F12X, которые специально выполнены с возможностью осуществления специальных функций, таких как картирование, выравнивание, операция Смита-Ватермана, HMM, определение вариантов и тому подобного. Следовательно, ЦП уровня 1 могут быть задействованы для формирования первого уровня конвейера для геномного анализа для выполнения этапов общей обработки и постановки в очередь на подготовку данных для конвейерного анализа, причем данные после обработки одним из множества ЦП могут быть поданы на специально предназначенный экземпляр ППВМ на уровне 2, например F12X, например где экземпляр ППВМ F12X выполнен с возможностью исполнения вычислительноемких функций, таких как функции картирования и/или выравнивания и т.д.

[00565] Таким образом, в конкретном варианте осуществления экземпляры ЦП в конвейере направляют свои данные, после их подготовки, в один или два экземпляра ППВМ, предназначенных для картирования и выравнивания. После выполнения картирования результирующие данные могут быть сохранены в запоминающем устройстве и/или затем поданы в экземпляр выравнивания, где может быть выполнено выравнивание, например по меньшей мере одним специально предназначенным экземпляром FPGA уровня 2. Аналогичным образом обработанные картированные и выровненные данные затем могут быть сохранены в запоминающем устройстве и/или направлены в экземпляр ЦП уровня 3 для дальнейшей обработки, который может быть тем же самым, что и на уровне 1, или другим экземпляром, например для выполнения менее ресурсоемкой функции обработки геномного анализа, например для выполнения функции сортировки. Кроме того, после того, как ЦП уровня 3 выполнили свою

обработку, результирующие данные могут быть направлены либо обратно в другие экземпляры ППВМ уровня 2, либо в экземпляры ППВМ уровня 4, например для ресурсоемких функций дальнейшей геномной обработки, таких как функции обработки Смита-Ватермана (SW), например в экземпляре ППВМ, специально предназначенном для SW. Аналогичным образом, после того, как анализ SW выполнен, например с помощью ППВМ F12X, специально предназначенной для SW, обработанные данные могут быть отправлены в одно или более связанных запоминающих устройств и/или далее по конвейеру для обработки, например в экземпляр ЦП и/или ППВМ уровня 4 или 5, или обратно на уровень 1 или 3, например для выполнения анализа HMM и/или определения вариантов, например в специально предназначенной ППВМ и/или ядре обработки ЦП дальнейшего уровня.

[00566] Подобным образом можно преодолеть проблемы задержки и эффективности за счет сочетания всевозможных разных экземпляров на одном или более разных уровнях, для обеспечения конвейерной платформы для геномной обработки. Такая конфигурация может подразумевать больше, чем масштабирующие и/или комбинирующие экземпляры, экземпляры могут быть выполнены так, что они специализируются на осуществлении специально назначенных функций. В таком случае экземпляр ППВМ для картирования выполняет только картирование, и, аналогично, экземпляр ППВМ для выравнивания выполняет только выравнивание и т. д., вместо того, чтобы один экземпляр выполнял обработку в конвейере от начала до конца. Хотя в других конфигурациях одна или более ППВМ могут быть по меньшей мере частично переконфигурированы, например между выполнением конвейерных задач.

[00567] Следовательно, диспетчер конвейера может управлять очередью запросов на геномную обработку, создаваемых экземплярами ЦП уровня 1, чтобы разбивать их на дискретные задания, агрегировать и направлять в соответствующий специфичные для задания ЦП/F1 для обработки, например для картирования и/или выравнивания, причем данные F1 после обработки могут быть отправлены назад или вперед на следующий уровень обработки ЦП/ППВМ результирующих данных, например для выполнения различных этапов в модуле определения вариантов. Например, функция определения вариантов может быть разбита на множество операций, которые могут быть выполнены в программном обеспечении, затем направлены на обработку HMM в один или более экземпляров ППВМ аппаратного обеспечения, а затем могут быть отправлены в ЦП для продолжения операций определения вариантов, например когда вся платформа эластично подобрана по размеру и реализована для сведения к минимуму стоимости дорогих экземпляров ППВМ при максимальном повышении использования, максимальном снижении задержки и, следовательно, оптимизации операций. Соответственно, таким образом требуется меньше экземпляров аппаратного обеспечения вследствие их абсолютных возможностей обработки и специфичности аппаратной реализации, и поэтому количество ППВМ относительно ЦП может быть сведено к минимуму, а их использование, например ППВМ, может быть максимально повышено, и поэтому система может быть оптимизирована, чтобы поддерживать постоянную занятость всех экземпляров. Такая конфигурация оптимально спроектирована для анализа геномной обработки, особенно для картирования, выравнивания и определения вариантов.

[00568] Дополнительный структурный элемент, который можно добавить, например в качестве вспомогательного устройства, в архитектуру конвейера, описанную в настоящем документе, это один или более модулей эластичной памяти, которые могут быть выполнены с возможностью функционирования для обеспечения блочного

хранения данных, например результирующих данных, по мере их передачи по всему конвейеру. Соответственно, одно или более модулей эластичного хранилища блоков данных (EBDS) могут быть вставлены между одним или более уровнями обработки, например между различными экземплярами и/или уровнями экземпляров, так что по мере обработки данных и получения результатов, обработанные результаты могут быть направлены в устройство EBDS для хранения перед направлением на обработку следующего уровня, например посредством специально предназначенного модуля обработки ППВМ. То же самое EBDS может быть использовано между всеми экземплярами или уровнями экземпляров, или множество EBDS могут быть использованы между различными экземплярами и/или уровнями экземпляров, например для хранения, компиляции и/или постановки в очередь результирующих данных.

[00569] В данной конфигурации перед отправкой данных непосредственно из одного экземпляра и/или с одного уровня обработки на другой, данные могут быть направлены в EBDS или другое запоминающее устройство или структуру для хранения и последующего направления в надлежащий модуль аппаратной обработки. А именно, модуль хранения блоков может быть присоединен к узлу в качестве запоминающего устройства, где данные могут быть записаны в блочное хранилище данных (BDS) для хранения на одном уровне, а BDS может быть переключено на другой узел, чтобы направить сохраненные данные на следующий уровень обработки. Таким образом, один или более, например несколько, модулей BDS могут быть включены в конвейер и выполнены с возможностью переключения с одного узла на другой для участия в переходе данных по всему конвейеру. Кроме того, как указано выше, может быть использовано более гибкое устройство хранения файлов, такое как устройство, которое выполнено с возможностью его связывания с одним или более экземплярами одновременно, например без необходимости переключения с одного на другой.

[00570] Соответственно, в конвейере для обработки существует множество этапов, например на его обслуживающих узлах, при подготовке данных к обработке, например предварительной обработке, причем после подготовки эти данные направляют в соответствующие экземпляры обработки на одном уровне, где могут быть сгенерированы результирующие данные, затем результиру.щие данные могут быть сохранены, например на устройстве EDS, поставлены в очередь и подготовлены для следующей стадии обработки путем переключения на следующий узел экземпляров и направления в следующий экземпляр для обработки с использованием экземпляров обработки ППВМ и/или ЦП следующего порядка, где могут быть сгенерированы дальнейшие результирующие данные, и опять после генерации результирующие данные могут быть направлены обратно на тот же самый или вперед на следующий уровень EDS для хранения перед продвижением на следующую стадию обработки.

[00571] В частности, в одном конкретном варианте осуществления поток через конвейер может выглядеть следующим образом: ЦП: данные подготовлены (поставлены в очередь и/или сохранены); ППВМ: Картирование, временное хранение, ППВМ: выравнивание, временное хранение; ЦП: сортировка, временное хранение, ЦП: удаление дубликатов, временное хранение; ППВМ: HMM, временное хранение, ЦП: определение 1 вариантов, временное хранение, ППВМ: Смит-Ватерман, временное хранение, ЦП: определение 2 вариантов, временное хранение, ЦП: VCGF, временное хранение и т. д. Следует отметить, что один или более из этих этапов могут быть выполнены в любом логическом порядке и могут быть реализованы любыми подходящим образом сконфигурированными ресурсами, например реализованы в программном обеспечении и/или аппаратном обеспечении во множестве различных сочетаний. Кроме того, одно

или более EDS или другие подходящим образом сконфигурированные устройства хранения данных и/или файлов могут быть присоединены к одному или более из различных узлов, например между различными уровнями экземпляров, например для временного хранения между множеством разных этапов обработки. Соответственно, подобным образом каждый уровень экземпляров обработки может быть эластично масштабирован по мере надобности, например между каждым из разных узлов или уровней узлов, например для обработки одного или нескольких геномов.

[00572] В соответствии с другим аспектом, как показано на ФИГ. 16, предложен способ использования системы для генерации одного или более файлов данных, над которыми можно выполнять один или более протоколов вторичной и/или третичной обработки. Например, способ может включать обеспечение геномной инфраструктуры, например для одной или более из локальной, облачной и/или гибридной генерации, обработки и/или анализа в области генома и/или биоинформатики.

[00573] В таком случае геномная инфраструктура может включать биоинформационную платформу обработки, имеющую одно или более хапоминающих устройств, которые выполнены с возможностью хранения одной или более выполненных с возможностью конфигурирования обрабатывающих структур для конфигурирования системы с целью обеспечения возможности выполнения одной или более функций аналитической обработки над данными, такими как данные, содержащие интересующую геномную последовательность или относящиеся к ней обработанные результирующие данные. Память может содержать интересующую геномную последовательность, которую необходимо обработать, например после того, как она сгенерирована и/или получена, одну или более контрольных генетических последовательностей и/или может дополнительно содержать индекс одной или более контрольных генетических последовательностей и/или список относящихся к ним границ сплайсинга. Система может также содержать устройство ввода, имеющее программный интерфейс приложения (API) платформы для выбора из списка вариантов одной или более структур обработки, выполненных с возможностью конфигурирования, например дял конфигурирования системы, например путем выбора функций обработки системы, которые будут исполняться над данными, например предварительной или последующей обработки геномных последовательностей, представляющих интерес. Возможно также наличие графического пользовательского интерфейса (ГПИ), которые выполнен с возможностью функционального связывания с API, например для предоставления меню, с помощью которого пользователь может выбирать, какие из имеющихся вариантов требуется выполнить над данными.

[00574] Система может быть реализована на одной или более интегральных схем, которые могут быть сформированы из одного или более наборов выполненных с возможностью конфигурирования, например предварительного конфигурирования, или жестко смонтированных цифровых логических схем, которые могут быть взаимно соединены посредством множества физических электрических соединителей. В таком примере интегральная схема может иметь вход, например, интерфейс запоминающего устройства, для приема одного или множества протоколов структуры, выполненных с возможностью конфигурирования, например из запоминающего устройства, и может быть также выполнена с возможностью реализации одной или более структур на интегральной схеме в соответствии с протоколами структуры обработки, выполненными с возможностью конфигурирования. Интерфейс запоминающего устройства входа может быть также выполнен с возможностью приема данных геномной последовательности, которые могут быть представлены в виде множества ридов

геномных данных. Интерфейс может быть также выполнен с возможностью доступа к одной или более генетических контрольных последовательностей и индексу (индексам).

[00575] В различных примерах цифровые логические схемы могут быть выполнены в виде набора модулей обработки, каждый из которых сформирован из подмножества цифровых логических схем. Цифровые логические схемы и/или движки обработки могут быть выполнены с возможностью осуществления одного или более предварительно конфигурируемых этапов протокола первичной, вторичной и/или третичной обработки для генерации множества ридов данных геномной последовательности и/или обработки множества ридов геномных данных, например в соответствии с контрольной(ыми) генетической(ими) последовательностью(ями) или другой информацией, полученной из генетической последовательности. Интегральная схема может также иметь выход, чтобы выводить результирующие данные первичной, вторичной и/или третичной обработки, например в соответствии с программным интерфейсом приложения (API) платформы.

[00576] В частности, в различных вариантах реализации цифровые логические схемы и/или наборы модулей обработки могут образовывать множество конвейеров для геномной обработки, например когда каждый конвейер может иметь вход, который определен в соответствии с программным интерфейсом приложения платформы для приема платформой биоинформационной обработки результирующих данных первичной и/или вторичной обработки и для выполнения на них одного или более аналитических процессов с целью получения результирующих данных. Кроме того, множество конвейеров для геномной обработки могут иметь общий API конвейеров, который определяет операцию вторичной и/или третичной обработки, которую нужно исполнить над результирующими данными первичной и/или вторичной обработки, например когда каждый из множества конвейеров для геномной обработки выполнен с возможностью осуществления подмножества операций вторичной и/или третичной обработки и вывода результирующих данных вторичной и/или третичной обработки в соответствии с API конвейеров.

[00577] В таких случаях в памяти и/или связанном хранилище приложений, выполненном с возможностью поиска, могут храниться множество приложений геномного анализа, например когда каждое из множества приложений геномного анализа может быть доступно компьютеру посредством электронного носителя, например для исполнения процессором компьютера, с целью осуществления целевого анализа геномных данных из предварительной или последующей обработки результирующих данных первичной, вторичной и/или третичной обработки, например одним или более из множества конвейеров для геномной обработки. В конкретных примерах каждое из множества приложений геномного анализа может быть определено интерфейсом API и может быть выполнено с возможностью приема результирующих данных первичной, вторичной и/или третичной обработки и/или осуществления целевого анализа геномных данных предварительной или последующей обработки и вывода результирующих данных целевого анализа в одну или более геномных баз данных.

[00578] Способ может дополнительно включать выбор, например в меню ГПИ, одного или более конвейеров для геномной обработки из множества доступных конвейеров для геномной обработки системы; выбор одного или более приложений геномного анализа из множества приложений геномного анализа, которые хранятся в хранилище приложений; и исполнение с помощью процессора компьютера одного или более выбранных приложений геномного анализа для осуществления целевого анализа геномных данных из результирующих данных первичной, вторичной и/или третичной

обработки.

[00579] Кроме того, в различных вариантах реализации картирование, выравнивание, сортировка и определение вариантов могут происходить на микросхеме, и в различных вариантах реализации местное повторное выравнивание, маркировка дубликатов, перекалибровка оценки качества оснований и/или один или более протоколов и/или конвейеров для третичной обработки тоже выполняются на микросхеме, и в различных примерах различные протоколы сжатия, такие как BAM и CRAM, тоже могут выполняться на микросхеме. Однако после того как данные в результате первичной, вторичной и/или третичной обработки созданы, они могут быть сжаты, например перед передачей, например оправкой по всей системе, отправкой в облако, например для выполнения модуля определения вариантов, платформы вторичной, третичной и/или другой обработки, например включая протокол анализа интерпретации и/или совместной работы. Это может быть полезно, особенно с учетом того факта, что определение вариантов, включая их третичную обработку, может быть «стрельбой по движущейся мишени», например отсутствует стандартизованный согласованный алгоритм, используемый в данной отрасли.

[00580] Поэтому при необходимости для достижения различных типов результатов могут использоваться различные алгоритмы, например удаленными пользователями, и, следовательно, полезно иметь модуль на облачной основе для выполнения данной функции, чтобы обеспечить гибкость при выборе алгоритма, полезного в любой заданный конкретный момент, а также последовательной и/или параллельной обработки. Соответственно, любой из модулей, описанных в настоящем документе, может быть реализован либо аппаратно, например на микросхеме, либо программно, например в облаке, но в определенных вариантах реализации все модули могут быть выполнены с возможностью осуществления их функций на микросхеме, или все модули могут быть выполнены с возможностью осуществления их функций удаленно, например в облаке, или может быть обеспечена комбинация модулей, так что некоторые из них находятся на одной или более микросхем, а другие расположены в облаке. Кроме того, как было указано, в различных вариантах реализации сама микросхема или сами микросхемы могут быть выполнены с возможностью функционирования совместно, а в некоторых вариантах реализации, в непосредственном взаимодействии с генетическим секвенатором, таким как NGS и/или секвенатор на микросхеме.

[00581] Более конкретно, в различных вариантах реализации устройство по настоящему изобретению может быть микросхемой, такой как микросхема, которая выполнена с возможностью обработки данных генома, например путем использования модулей конвейеров для анализа данных. Соответственно, как показано на ФИГ. 17-19, предложена геномная конвейерная процессорная микросхема 100 вместе со связанным аппаратным обеспечением геномной конвейерной процессорной системы 10. Микросхема 100 имеет одно или более соединений с внешним запоминающим устройством 102 («Управляющее устройство памяти DDR3») и соединение 104 (например, интерфейс PCIe) с внешним миром, таким как, например, главный компьютер 106. Коммутатор 108 (например, переключатель) обеспечивает доступ к интерфейсам запоминающего устройства различным инициаторам запросов. Движки 110 DMA передают данные с высоко скоростью между главным устройством и внешними запоминающими устройствами 102 процессорной микросхемы 100 (через коммутатор 108) и/или между главным устройством и центральным управляющим устройством 112. Центральное управляющее устроойство 112 управляет операциями микросхемы, в частности координирует действия нескольких модулей 13 обработки. Движки обработки

сформированы из набора жестко смонтированных цифровых логических схем, которые взаимно связаны физическими электрическими соединениями и организованы в кластеры 114 модулей. В некоторых выраинтах осуществления движки в одном кластере совместно используют один порт коммутатора посредством арбитратора. Центральный контроллер 112 имеет соединения с каждым из кластеров модулей. Каждый кластер 114 модулей имеет ряд модулей обработки для обработки геномных данных, в том числе картировщик 120 (или модуль картирования), выравниватель 122 (или модуль выравнивания) и сортировщик 124 (или модуль сортировки). Кластер 114 модулей может содержать также другие движки или модули.

[00582] В соответствии с одной моделью потока данных, согласующейся с вариантами осуществления, описанными в настоящем документе, главное устройство посылает команды и данные через движки 110 DMA в центральное управляющее устройство 112, которое равномерно распределяет данные между модулями обработки. Движки обработки возвращают обработанные данные в центральное управляющее устройство 112, который передает их в потоковом режиме обратно в главное устройство посредством модулей 110 DMA. Эта модель потока данных приспособлена для картирования и выравнивания.

[00583] В соответствии с альтернативной моделью потока данных, согласующейся с вариантами осуществления, описанными в настоящем документе, главное устройство передает в потоковом режиме данные во внешнее запоминающее устройство, либо напрямую посредством модулей 110 DMA и комутатора 108, или посредством центрального управляющего устройства 112. Главное устройство отправляет команды в центральное управляющее устройство 112, которое отправляет в движки обработки команды, указывающие модулям обработки, какие данные обрабатывать. Движки обработки осуществляют доступ к входным данным из внешнего запоминающего устройства, обрабатывают их и записывают результаты обратно во внешнее запоминающее устройство, сообщая статус в центральное управляющее устройство 112. Центральное управляющее устройство 112 либо отправляет результирующие данные в потоковом режиме обратно в главное устройство из внешнего запоминающего устройства, либо уведомляет главное устройство, чтобы оно само извлекало результирующие данные посредством модулей 110 DMA.

[00584] На ФИГ. 17 и 18 показана геномная конвейерная процессорная система, изображающая полный комплект модулей обработки внутри кластера 114/214 модулей. Конвейерная процессорная система может содержать один или более кластеров 114/214 модулей. В некоторых вариантах осуществления конвейерная процессорная система 20 содержит четыре или более кластеров 114/214 модулей. В число модулей обработки или типов модулей обработки могут входить, без ограничений, картировщик, выравниватель, сортировщик, местный перевыравниватель, перекалибровщик оценки качества оснований, маркировщик дубликатов, определитель вариантов, сжимающие средства и/или распаковывающие средства. В некоторых вариантах осуществления каждый кластер 114/214 модулей имеет по одному движку обработки каждого типа. Соответственно, все движки обработки одного типа могут осуществлять доступ к коммутатору 108 одновременно через разные порты коммутатора, так как каждый из них находится в разных кластерах 114/214 модулей. Формирование в каждом кластере 114/214 модулей обработки каждого типа не требуется. Типы модулей обработки, которые требуют огромной параллельной обработки или пропускной способности запоминающего устройства, такие как картировщик (с прикрепленным(и) выравнивателем(ями)) и сортировщик, могут содержаться в каждом кластере модулей

конвейерной процессорной системы 20. Движки других типов могут появляться только в одном или нескольких кластерах 114/214 модулей по мере необходимости для удовлетворения требований к их производительности или производительности конвейерной процессорной системы.

[00585] На ФИГ. 19 приведена геномная конвейерная процессорная система, показывающая в дополнение к кластерам модулей, описанным выше, одно или более внедренных центральных процессоров (ЦП) 202. В число примеров таких внедренных ЦП входят ядра Snapdragon® или стандартные ядра ARM®. Эти ЦП исполняют полностью программируемые биоинформационные алгоритмы, такие как улучшенное определение вариантов. Такую обработку ускоряют с помощью вычислительных функций в кластерах модулей, которые могут быть вызваны ядрами 202 ЦП по мере надобности. Кроме того, даже ориентированная на движки обработка, такая как картирование или выравнивание, может выполняться ядрами 202 ЦПУ, обеспечивая их повышенную программируемость.

[00586] На ФИГ. 20 показан поток обработки для геномных конвейерных процессорных системы и способа. В некоторых предпочтительных вариантах осуществления данные обрабатывают в три прохода. Первый проход включает в себя картирование 402 и выравнивание 404, причем через движки прогоняют полный набор ридов. Второй проход включает в себя сортировку 406, где один большой блок, подлежащий сортировке (например, существенную часть всех ридов, ранее картированных на одну хромосому) загружают в запоминающее устройство, сортируют с помощью модулей обработки и возвращают в центральный компьютер. Третий проход включает в себя следующие по цепочке стадии (местное повторное выравнивание 408, маркировку 410 дубликатов, перекалибровку 412 оценки качества оснований (BQSR), вывод 414 BAM, вывод 416 редуцированного BAM и/или сжатие 418 CRAM). Этапы и функции третьего прохода могут быть выполнены в любой комбинации или подкомбинации и в любом порядке за один проход. Архитектуру с виртуальным конвейером, например как описана выше, используют для потоковой передачи ридов из главного устройства в циклические буферы в запоминающем устройстве через один модуль обработки за другим последовательно и обратно в главное устройство. В некоторых вариантах осуществления распаковка CRAM может быть отдельной функцией потоковой передачи. В некоторых вариантах осуществления вывод 414 BAM, вывод 416 редуцированного BAM и/или сжатие 418 CRAM могут быть заменены определением вариантов, сжатием и распаковкой.

[00587] В различных примерах описана аппаратная реализация конвейера для анализа последовательности. На фиг. 21 показана общая блок-схема варианта осуществления настоящего изобретения. В блоке 1 описана аппаратная реализация конвейера для анализа последовательности. Это можно сделать множеством различных способов, напимер при помощи варианта осуществления с использованием ППВМ, ИССН или структурированной ИССН. Функциональные блоки, которые реализованы с помощью ППВМ, ИССН или структурированной ИССН, показаны на фиг. 6 и 7. Фиг. 6 и 7 включают ряд блоков или модулей для выполнения анализа последовательности. Входными данными аппаратной реализации может быть файл в формате FASTQ, но этот формат не является ограничением. Помимо файла в формате FASTQ входные данные ППВМ, ИССН или структурированной ИССН содержат вспомогательную информацию, например Информацию об объеме потока (Flow Space Information), относящуюся к технологии, такой как Ion Torrent. В число блоков или модулей по фиг. 21 могут входить следующие блоки: исправление ошибок, картирование, выравнивание,

сортировка, местное повторное выравнивание, маркировка дубликатов, перекалибровка оценки качества оснований, сокращение BAM и побочной информации и/или определение вариантов.

[00588] Эти блоки или модули могут присутствовать внутри, или могут быть реализованы аппаратным обеспечением, но для достижения цели реализации конвейера для анализа последовательности некоторые из указанных блоков могут быть опущены, а другие блоки добавлены. Блоки 2 и 3 описывают два альтернативных варианта конвейерной платформы для анализа последовательности. Конвейерная платформа для анализа последовательности содержит ППВМ, ИССН или структурированную ИССН и программное обеспечение, поддерживаемое главным устройством (например, ПК, сервером, кластером или средствами облачного вычисления) с помощью облачного и/или кластерного хранилища. Блоки 4-7 описывают различные интерфейсы, которые может иметь конвейер для анализа последовательности. В блоках 4 и 6 интерфейс может представлять собой интерфейс PCIe, но не ограничивается интерфейсом PCIe. В блоках 5 и 7 аппаратное обеспечение (ППВМ, ИССН или структурированная ИССН) может быть непосредственно интегрировано в секвенатор. Блоки 8 и 9 описывают интеграцию аппаратного конвейера для анализа последовательности, интегрированного в главную систему, такую как ПК, кластер серверов или секвенатор. Аппаратное обеспечение ППВМ, ИССН или структурированной ИССН окружают множество элементов памяти DDR3 и интерфейса PCIe. Плата с ППВМ/ИССН/сИССН соединена с главным компьютером, состоящим из главного ЦП, который может быть маломощным ЦП, таким ка ARM®, Snapdragon® или любой другой процессор. Блок 10 показывает API аппаратный конвейер для анализа последовательности, который может быть доступен приложениям третьих лиц для выполнения третичного анализа.

[00589] Соответственно, в различных вариантах реализации устройство по настоящему изобретению может содержать вычислительную архитектуру, например внедренную в ППВМ или кремниевую интегральную схему специального назначения (ИССН) 100, как показано на фиг. 6 и 7. Интегральная схема 100 может быть интегрирована в печатную плату (ПП) 104, такую как плата интерфейса периферийных компонентов типа экспресс (PCIe), которую можно вставить в вычислительную платформу. В различных примерах, как показано на ФИГ. 6, плата 104 PCIe может содержать одну ППВМ или ИССН 100, причем интегральная схема может быть окружена локальными запоминающими устройствами 105, однако в различных вариантах реализации, как показано на ФИГ. 7, плата 104 PCIe может содержать множество ППВМ и/или ИССН 100A, 100B и 100C. В различных примерах плата PCI может также содержать шину PCIe. Плата 104 PCIe может быть добавлена на вычислительную платформу для исполнения алгоритмов над чрезвычайно большими наборами данных. Соответственно, в различных примерах весь рабочий поток секвенирования генома, задействующий интегральную схему, может включать в себя следующее: подготовку образца, выравнивание (включая картирование и выравнивание), анализ вариантов, биологическую интерпретацию и/или специальные приложения.

[00590] Следовательно, в различных вариантах реализации устройство по настоящему изобретению может содержать вычислительную архитектуру, которая достигает высокоэффективного исполнения алгоритмов, таких как алгоритмы картирования и выравнивания, которые работают над чрезвычайно большими наборами данных, например когда наборы данных проявляют плохую локальность ссылок (LOR). Эти алгоритмы предназначены для реконструкции всего генома из миллионов последовательностей коротких ридов из современных, так называемых, секвенаторов

нового поколения, требующей многогигабайтные структуры данных с произвольным доступом. По достижении реконструкции, как описано выше в настоящем документе, используют дальнейшие алгоритмы с аналогичными характеристиками для сравнения одного генома с библиотеками других, выполнения функционального анализа генов и т. д.

[00591] В нестоящее время используются два основных подхода, многоядерные ЦП общего назначения и графические процессоры общего назначения (ГПОН). В таком примере каждый ЦП в многоядерной системе может иметь классическую архитектуру на основе кэша, где инструкции и данные извлекаются из кэша уровня 1 (кэш L1), который мал, но обладает чрезвычайно быстрым доступом. Множество кэшей L1 могут быть соединены с более крупным, но более медленным кэшем L2. Кэш L2 может быть соединен с большой, но более медленной системой памяти DRAM (динамическое оперативное запоминающее устройство), или может быть соединен с еще более крупным, но более медленным кэшем L3, который может быть затем соединен с DRAM. Преимущество такой компоновки может заключаться в том, что приложения, в которых программы и данные проявляют локальность ссылок, ведут себя почти так, как если бы они исполнялись на компьютере с одним запоминающим устройством, большим как DRAM, но быстрым как кэш L1. Поскольку полностью заказные в высшей степени оптимизированные ЦП работают при очень высоких тактовых частотах, например от 2 до 4 ГГц, эта архитектура может быть существенна для достижения хороших рабочих характеристик.

[00592] Кроме того, эту архитектуру можно расширить с помощью ГПОН, например за счет реализации очень большого количества малых ЦП, каждый со своим малым кэшем L1, причем каждый ЦП исполняет одни и те же инструкции над различными подмножествами данных. Это представляет собой так называемую архитектуру SIMD (один поток команд, много потоков данных). Экономии можно достичь за счет совместного использования логики выборки и декодирования команды по всем большому количеству ЦП. Каждый кэш имеет доступ к множеству больших внешних DRAM через сеть межсоединений. Предполагая, что вычисление должно выполняться с возможностью высокого распараллеливания, ГПОН имеют значительное преимущество над ЦП общего назначения благодаря большому количеству вычислительных ресурсов. Тем не менее, они все равно имеют архитектуру с кэшированием и их рабочие характеристики ухудшаются приложениями, которые не обладают достаточной высокой степенью локальности ссылок. Это приводит к высокому коэффициенту непопадания при обращении к кэшу и простою процессоров в ожидании поступления данных из внешнего DRAM.

[00593] Например, в различных примерах в качестве системной памяти могут использоваться динамические RAM, поскольку они более экономичные, чем статические RAM (SRAM). Приблизительно можно считать, что при одинаковой стоимости объем DRAM в 4 раза превосходит объем SRAM. Однако, ввиду падения спроса на SRAM в пользу DRAM, разрыв между ними значительно увеличился из-за экономии пространства, которую предлагают DRAM, которые пользуются большим спросом. Независимо от стоимости DRAM в 4 раза плотнее, чем SRAM, размещенные на одинаковой площади кремния, так как для них требуются только один транзистор и одна емкость на бит по сравнению с 4 транзисторами на бит для реализации триггера в SRAM. DRAM представляет один бит информации как наличие или отсутствие заряда в емкости. Проблема с такой структурой состоит в том, что заряд ослабевает со временем, поэтому его нужно периодически обновлять. Такая необходимость привела

к архитектурам, организующим запоминающие устройства в виде независимых блоков и механизмов доступа, которые выдают множество слов памяти на каждый запрос. Это компенсирует время, когда данный блок недоступен во время обновления. Идея состоит в перемещении огромного количества данных, пока данный блок доступен. В этом заключается отличие от SRAM, в котором любое место в запоминающем устройстве доступно за одно обращение в течение постоянного количества времени. Данная характеристика позволяет при обращении к запоминающему устройству ориентироваться на одно слово, а не блок. DRAM работают хорошо в архитектуре с кэшированием, так как каждое непопадание в кэш приводит к блоку памяти, считываемому из DRAM. Теория локальности ссылок состоит в том, что сразу после обращения к слову N, вероятно, последует обращение к словам N + 1, N + 2, N + 3 и т. д.

[00594] На фиг. 7 показан альтернативный вариант реализации по фиг. 6, имеющий множество микросхем 100A, 100B, 100C, где каждая микросхема может включать один или более различных модулей геномной и/или биоинформационной обработки, например приведенное в качестве примера конвейерное устройство для анализа, как раскрыто в настоящем документе.

[00595] На фиг. 17 показана система 100 для исполнения конвейера анализа последовательности над данными генетической последовательности. Система 100 содержит диспетчер 102 конфигураций, который содержит вычислительную систему. Вычислительная система диспетчера 102 конфигураций может содержать персональный компьютер или другую компьютерную рабочую станцию, или может быть реализована комплектом сетевых компьютеров. Диспетчер 102 конфигураций может также содержать одно или более приложений третьих лиц, соединенных с вычислительной системой посредством одного или более API, которые вместе с одним или более частных приложений генерирует конфигурацию для обработки данных генома из секвенатора или другого источника данных генома. Диспетчер 102 конфигураций также содержит драйверы, которые загружают конфигурацию в конвейерную процессорную систему 10 для генома. Конвейерная процессорная система 10 для генома может выводить результирующие данные в сеть Интернет 504 или другую сеть, или быть доступна через них, для сохранения результирующих данных в электронной записи 200 о здоровья или другой базе 400 данных знаний.

[00596] Как не раз отмечалось выше в настоящем документе, микросхема, реализующая конвейерный процессор для генома, может быть соединена или интегрирована с секвенатором. Эта микросхема может быть также соединена или интегрирована на плате расширения, такой как плата PCIe, а плата расширения может быть соединена или интегрирована с секвенатором. В других вариантах осуществления микросхема может быть соединена или интегрирована с серверным компьютером, который соединен с секвенатором, чтобы передавать риды геномов из секвенатора на сервер. В еще одних вариантах осуществления микросхема может быть соединена или интегрирована с сервером в облачном вычислительном кластере компьютеров и серверов. Система может содержать один или более секвенаторов, подключенных (например, посредством Ethernet) к серверу, содержащему микросхему, причем риды геномов генерируются множеством секвенаторов, передаются на сервер, и затем картируются и выравниваются в микросхеме.

[00597] Например, обычно в конвейерах данных секвенатора ДНК нового поколения (NGS) обработка стадии первичного анализа, как правило, специфична для данной технологии секвенирования. Эту стадию первичного анализа выполняют для перевода

физических сигналов, обнаруживаемых внутри секвенатора, в «риды» нуклеотидных последовательностей вместе с соответствующими оценками качества (достоверности), например в файлах в формате FASTQ или в других форматах, содержащих последовательность и, как правило, информацию о качестве. Кроме того, после получения такого формата переходят к вторичному анализу, как описано в настоящем документе, чтобы определить содержимое секвенированного образца ДНК (или РНК и т.д.), например с помощью картирования и выравнивания ридов на контрольный геном, сортировки, маркировки дубликатов, перекалибровки оценки качества оснований, местного повторного выравнивания и определения вариантов. Затем может следовать третичный анализ для извлечения медицинских и исследовательских заключений из определенного содержимого ДНК.

[00598] Однако первичный анализ, как упоминалось выше, по своей природе часто довольно специфичен для используемой технологии секвенирования. В различных секвенаторах нуклеотиды обнаруживаются путем измерения электрических зарядов, электрического тока или излучаемого света. Некоторые конвейеры для первичного анализа часто включают в себя: обработку сигнала для усиления, фильтрации, разделения и измерения выходного сигнала датчика; уменьшение объема данных, например путем разбиения на подгруппы, прореживания, усреднения, преобразования и т. д.; обработку изображения или цифровую обработку с целью выявления и усиления имеющих значение сигналов и связывания их с конкретными ридами и нуклеотидами (например, вычисление смещения изображения, идентификацию кластера); алгоритмическую обработку и эвристическую процедуру для компенсации артефактов технологии секвенирования (например, оценку фазирования, матрицы перекрестных помех); вычисления байесовской вероятности; скрытые марковские модели; определение оснований (выбор наиболее вероятного нуклеотида в каждой позиции в последовательности); оценку качества (достоверности) определения оснований и т. п.

[00599] Первичный анализ может быть сильно емким с коммутативной точки зрения, иногда таким же емким, как и вторичный анализ. Например, в существующих технологиях секвенирования первичный анализ зачастую использует ППВМ и ГП для ускорения процесса выше возможностей ЦП. Но эти ускоренные функции могут быть выполнены значительно более эффективно в определенной интегральной схеме, например такой, как описана в настоящем документе. Например они могут быть реализованы в структурированной ИССН с использованием конфигурируемых металлических слоев, так как им не требуется столько точности при физическом размещении, сколько требуют встроенные ядра процессоров, однако существенные параллельные вычисления, реализованные в больших ППВМ и ГП может быть сложно осуществить в конфигурируемых ресурсах структурированной ИССН. В качестве альтернативы можено реализовать логику ускорения первичной обработки в базовой пластине структурированной ИССН, обеспечивая преимущество пространственной эффективности стандартной ячейки в базовой пластине.

[00600] Причина того, что функции вторичной обработки могут быть реализованы в конфигурируемых металлических слоях структурированной ИССН, заключается в том, что алгоритмы вторичной обработки данных генома все еще подвергаются активным исследованиям. Таким образом, преимуществом может быть возможность периодически выполнять обновленную конструкцию структурированной ИССН недорогим образом, например каждый год или два года для применения наиболее новых алгоритмов. В противоположность этому алгоритмы первичного анализа в настоязее, используемые в настоящее время, являются более проработанными,

обязательные этапы обраюотки были исследованы и определены соответствующими производителями секвенаторов. Несмотря на то, что они все еще в некоторой мере подвергнуты изменениям, указанные алгоритмы в большей степени представляют собой родовой сигнал и числовую обработку, чем в случае вторичного анализа, так что надлежащая конфигурируемость и микропрограммирование модулей ускорения первичной обработки может сделать их достаточно гибкими для значительных изменения. При наличии встроенные ядра процессоры повышают эту гибкость еще сильнее. Поэтому есть смысл интегрировать методы ускорения первичной обработки в базовую пластину ППВМ и/или структурированной ИССН, как описано в настоящем документе.

[00601] Также преимуществом является интеграция ускорения первичной обработки и ускорения вторичной обработки в единую интегральную схему ППВМ или ИССН (стандартную ячейку или структурированную ИССН) с встроенными процессорами, или без них. Это может быть полезным, так как секвенаторы вырабатывают данные, которым требуется первичный и вторичный анализ, и интеграция их в единое устройство является наиболее эффекктивной с точки зрения стоимости, пространства, энергии и совместного использования ресурсов. Если также присутствуют встроенные процессоры, они могут быть использованы для повышения скорости и гибкости как первичной, так и вторичной обработки.

[00602] Эти три компонента - первичные ускорители, вторичные ускорители и встроенные процессоры - могут быть реализованы в базовой пластине ППВМ или структурированной ИССН и/или с использованием конфигурируемых металлических слоев, или их комбинации. Все три могут быть в базовой пластине, или все три могут использовать конфигурируемые металлические слои, или любые два из них могут быть в базовой пластине, а другие использовать конфигурируемые металлические слои. В любой из указанных конфигураций все три могут осуществлять связь друг с другом в любой комбинации непрямую и/или через запоминающее устройство, и взаимодействовать в общих задачах. Одной предпочтительной конфигурацией является реализация первичного ускорения и встроенных процессоров в базовой пластине и реализация вторичного ускорения с использованием конфигурируемых металлических слоев.

[00603] Кроме того, как указано выше, микросхема, реализованная как ИССН, ППВМ или структурированная ИССН, может включать или другим образом быть связана с одной или более архитектурами памяти. Например, архитектура памяти может состоять из M модулей памяти, которые взаимодействуют с микросхемой, например с ИССН. ИССН может быть реализована с использованием множества различных технологий, включая ППВМ (Программируемые пользователем вентильные матрицы) или структурированную ИССН, стандартные ячейки или полностью заказную логику. В ИССН присутствуют Подсистема памяти (MSS) и функциональные процессоры (ФП). MSS содержит M контроллеров памяти (MC) для модулей памяти, N системных интерфейсов памяти (SMI) для ФП и N×M коммутаторов, которые обеспечивают возможность доступа любых SMI к любыт MC. В случае конфликтов обеспечено разрешение конфликтных ситуаций.

[00604] Каждый модуль памяти сконструирован из микросхем DRAM, к которым можно обращаться посредством A_{MM}-битного слова и поддерживают передачу данных шириной D_{MM} бит. Указанная память имеет $2^{A_{MM}}$ адресных местоположений. Ключевая

характеристика DRAM в том, что оно реализует считывания/записи пакетами по W слов с использованием переданного адреса в качестве базового адреса, B, а также извлекая или сохраняя местоположения B+1, B+2, … B+W-1. Обычно W равно 8.

[00605] В MSS ИССН каждый контроллер памяти подает требуемый управляющий сигнал и выполняет при необходимости мультиплексирование/демультиплексирование между разрядностью системного слова, D_{SYS}, и разрядностью слова памяти, D_{MM}, а также выполняет требования для пакетов считывания/записи. Он может иметь дополнительную буферизацию, так что множество запросов к памяти могут быть поставлены в очередь и обработаны конвейерным образом для максимизации пропускной способности. Это компенсирует множество тактовых циклов задержки между предоставлением адреса и завершением работы памяти (записи или считывания).

[00606] МС могет работать со скоростью прикрепленного DRAM в модуле памяти. Предположим, что его тактовая частота C_{MM}. Она обычно в несколько раз быстрее, чем скорость ядра, на которой работает большая часть логики в ИССН, которая равна C_{SYS}. Следовательно, логика мультиплексирования/демультиплексирования расположена рядом со связанными с ней контактами интерфейса для сведения к минимуму расстояний, на которые проходят сигналы. Демультиплексирование является первой операцией, выполняемой над входящими данными, а мультиплексирования является последней операцией, выполняемой над выходящими данными. Оставшаяся часть MSS работает над данными с разрядностью D_{SYS}, которая больше D_{MM}, обеспечивая использование более медленной тактовой частоты C_{SYS}.

[00607] Каждый интерфейс памяти системы в MSS представляет шину с A_{SYS} -битным адресом и шину с D_{SYS}-битными данными для любого прикрепленного ФП. SMI спроектирован так, чтобы прикрепленному ФП казалось, что он имеет произвольный доступ к единой быстрой памяти большой емкости. ФП не осведомлен о существовании отдельных модулей памяти. Значение A_{SYS} является достаточно большим для обеспечения доступа к местоположению памяти в любом прикрепленном модуле памяти. Картирование из адресного пространства системы в адресное пространство модуля памяти описано ниже.

[00608] N системных интерфейсов памяти перекрестно соединены с M модулей памяти посредством коммутатора N×M. Указанный коммутатор обеспечивает min(M,N) одновременных соединений между SMI и МС, обеспечивает разрешение конфликтов и способствует трансляции адресного пространства системы в адресное пространство модуля памяти.

[00609] Организация ФП является в значительной степени гибкой. Один или более ФП могут хранить один и тот же интерфейс памяти системы. Для максимизации рабочих характеристик ФП, которые не работают в одно и то же время должны совместно использоватьб SMI. А те, что работают одновременно, должны быть соединены с различными SMI. ФП, который работает над структурами данных, большими, чем D_{SYS}, может использовать множество SMI для доступа ко всей структуре данных за одну операцию с памятью. Следовательно, указанная архитектура данных поддерживает широкий диапазон вычсислительных архитектур. Каждый ФП может быть идентичным и, следовательно, массив из них может быть реализован в двумерной структуре. Это проиллюстрировано, когда FPU(i,j) является j-тым компонентом, прикрепленным к SMI i, $0 \le i < N$, $0 \le j < k_i$. В данном случае все k_i имеют один размер, а k_i может быть малым и равным 1. Это поддерживает архитектуры SIMD (один поток команд, много потоков

данных) и MIMD (много потоков команд, много потоков данных) в зависимости от того, извлекают ли ФП команды из одних и тех же или отдельных командных запоминающих устройств.

[00610] Согласно одному частному аспекту, настоящее изобретение направлено на систему, такую как система для исполнения конвейера для анализа последовательности над данными генетической последовательности. В различных примерах указанная система может содержать источник электронных данных, такой как источник данных, который обеспечивает цифровые сигналы, например цифровые сигналы, представляющие множество ридов геномных данных, причем каждый из множества ридов геномных данных содержит последовательность нуклеотидов. Система может содержать одно или более запоминающих устройств, таких как память, хранящих одну или более контрольных генетических последовательностей и/или индекс одной или более контрольных генетических последовательностей; и/или система может содержать микросхему, такую как ИССН, ППВМ или сИССН.

[00611] Более конкретно, в различных конкретных вариантах реализации система может содержать структурированную интегральную схему специального назначения (ИССН), например такую, в которой микросхема образована набором жестко смонтированных цифровых логических схем с масочным программированием, которые могут быть взаимно соединены посредством множества физических электрических межсоединений. В различных примерах одно или более из множество физических электрических межсоединений содержат вход в структурированную ИССН, который соединен с источником электронных данных, например для приема множества ридом геномных данных. В таком примере одно или более из множества физических электрических межсоединений могут содержать интерфейс памяти для доступа структурированной ИССН к памяти. Соответственно жестко смонтированные цифровые логические схемы могут быть выполнены в виде набора модулей обработки, такого где каждый модуль обработки может быть сформирован подмножеством жестко смонтированных цифровых логических схем для выполнения одного или более этапов в конвейере анализа последовательностей на множестве ридов геномных данных. Например, одно или более, например каждое, из подмножеств жестко смонтированных цифровых логических схем может быть в монтажной конфигурации для выполнения одного или более этапов в конвейере анализа последовательностей. Например, набор модулей обработки может быть выполнен с возможностью содержания одного или более из модуля картирования, модуля выравнивания и/или модуля сортировки.

[00612] Например, набор модулей обработки может включать в себя модуль картирования в монтажной конфигурации, выполненный с возможностью доступа, в соответствии по меньшей мере с частью последовательности нуклеотидов в риде из множества ридов, к индексу одной или более контрольных генетических последовательностей из памяти через интерфейс памяти для картирования рида на один или более сегментов одной или более контрольных генетических последовательностей на основе индекса. Например, в некоторых вариантах реализации индекс одной или более контрольных генетических последовательностей может содержать хэш-таблицу и/или модуль картирования может применять хэш-функцию по меньшей мере к части последовательности нуклеотидов для доступа к хэш-таблице индекса.

[00613] Движки обработки могут также или в качестве альтернативы содержать модуль выравнивания в монтажной конфигурации, выполненный с возможностью доступа к одной или более контрольных генетических последовательностей из памяти, например через интерфейс памяти для выравнивания рида на одно или более положений

в одном или более сегментов одной или более контрольных генетических последовательностей, например как получено из модуля картирования. Движки обработки могут также или в качестве альтернативы содержать модуль сортировки в монтажной конфигурации, выполненный с возможностью доступа к одному или более выровненным ридам из памяти, например через интерфейс памяти для сортировки рида на одно или более положений, например положений хромосом, в одной или более контрольных генетических последовательностей, например как получено из модуля выравнивания.

[00614] В различных примерах структурированная ИССН может содержать базовую пластину, которая содержит по меньшей мере некоторые из жестко смонтированных цифровых логических схем, и в некоторых примерах может содержать один или более конфигурируемых металлических слоев, образованных на базовой пластине, например где каждый из одного или более конфигурируемых металлических слоев может иметь по меньшей мере некоторые из множества физических электрических межсоединений, которые обеспечивают взаимное соединение по меньшей мере некоторых из жестко смонтированных цифровых логических схем для формирования набора модулей обработки. В некоторых вариантах реализации одно или более из множества физических электрических межсоединений может содержать выход из структурированной ИССН, например для осуществления передачи результирующих данных из модуля картирования, модуля выравивания и/или модуля сортировки.

[00615] В различных примерах структурированная ИССН может содержать главное управляющее устройство для установления монтажной конфигурации для каждого поднабора жестко смонтированных цифровых логических схем для выполнения одного или более этапов в конвейере анализа последовательностей. В различных вариантах реализации монтажная конфигурация выполнена при производстве интегральной схемы и являются некратковременными. В некоторых вариантах реализации структурированная ИССН и/или запоминающее устройство расположены в плате расширения, например в плате интерфейса периферийных компонентов (PCI). Как указано выше, в различных вариантах реализации указанная система может содержать секвенатор, например который содержит источник электронных данных, который обеспечивает цифровые сигналы, представляющие множество ридов геномных данных. И в таком примере плата расширения может быть физически интегрирована в секвенатор.

[00616] Кроме того, в различных вариантах реализации может быть обеспечена структурированная интегральная схема специального назначения (ИССН), например для анализа данных генетической последовательности, например когда данные генетической последовательности хранятся в запоминающем устройстве, например хранящем одну или более контрольных генетических последовательностей, связанных с геномными данными, и/или индекс одной или более контрольных генетических последовательностей. В таком примере структурированная ИССН может содержать базовую пластину, которая также содержит набор цифровых логических схем, и может также содержать один или более конфигурируемых металлических слоев, образованных на базовой пластине, например где каждый из одного или более конфигурируемых металлических слоев может иметь набор монтажных соединений, которые обеспечивают взаимное соединение поднабора цифровых логических схем для формирования набора модулей обработки. В таком примере набор модулей обработки может содержать модуль картирования, модуль выравнивания и/или модуль сортировки. В различных примерах часть набора цифровых логических схем в базовой пластине является жестко

смонтированной в качестве модуля определения оснований.

[00617] Например, набор модулей обработки может включать в себя модуль картирования для доступа, в соответствии по меньшей мере с частью последовательности нуклеотидов в риде из множества ридов, к индексу одной или более контрольных генетических последовательностей, хранящихся в памяти, для картирования рида на один или более сегментов одной или более контрольных генетических последовательностей, например на основе индекса. Дополнительно или в качестве альтернативы набор модулей обработки может содержать модуль выравнивания для доступа к одной или более контрольных генетических последовательностей из памяти, например через интерфейс памяти, для выравнивания рида на одно или более положений в одном или более сегментов одной или более контрольных генетических последовательностей из модуля картирования. Дополнительно или в качестве альтернативы набор модулей обработки может содержать модуль сортировки для сортировки каждого выровненного рида в соответствии с одним или более положениями в одной или более контрольных генетических последовательностей.

[00618] В одном варианте реализации обеспечена система для реализации конвейера для анализа последовательности над данными генетической последовательности, причем система содержит источник электронных данных, который обеспечивает цифровые сигналы, представляющие множество ридов геномных данных, причем каждый из множества ридов геномных данных содержит последовательность нуклеотидов. Система может содержать одно или более запоминающих устройств, например для хранения одной или более контрольных генетических последовательностей; и/или система может содержать интегральную схему, имеющую базовую пластину, например базовую пластину, образованную посредством фотолитаграфической маски, которая определяет набор цифровых логических схем. В таком примере базовая пластина может быть выполнена с возможностью обладания одной или более функций, как описаны выше в настоящем документе, интегрированных в нее. Например, базовая пластина может иметь один или более конфигурируемых металлических слоев, например когда каждый из одного или более конфигурируемых металлических слоев имеет одно или более межсоединений, которые соединяют поднабор набора цифровых логических схем в монтажной конфигурации для выполнения указанных выше функций.

[00619] Согласно различным аспектам, обеспечен способ выполнения структурированной интегральной схемы специального назначения (ИССН) для анализа данных генетической последовательности. В некоторых вариантах реализации способ включает одно или более из обеспечения множества фотолитографических масок, например масок, которые определяют набор цифровых логических схем базовой пластины; формирования набора цифровой логики, например с использованием множества фотолитографических масок для формирования базовой пластины; обеспечения двух или более различных наборов специфичных для конструкции масок конфигурируемых металлических слоев, например масок, которые определяют соответствующие две или более цифровых логик для осуществления набора модулей обработки; образования двух или более конфигурируемых металлических слоев, например использующих две или более специфичных для конструкции масок конфигурируемых металлических слоев, например в которых каждый из двух или более конфигурируемых металлических слоев имеет набор монтажных соединений, которые могут быть выполнены в соответствии с конструкцией масок конфигурируемых металлических слоев, например для взаимного соединения поднабора цифровых логических схем для формирования наборы модулей обработки; и/или обеспечения

двух или более конфигурируемых металлических слоев на базовой пластине для формирования набора модулей обработки.

[00620] Согласно одному или более аспектов или признаков объекта изобретения, описанного в настоящем документе, могут быть реализованы в цифровой электронной схеме, интегрированной схеме, специально разработанных интегральных схем специального назначения (ИССН), программируемых пользователем вентильных матрицах (ППВМ) или структурированной ИССН, компьютерном аппаратном обеспечении, встроенном программном обеспечении, программном обеспечении и/или их комбинациях.

[00621] Указанные различные аспекты или признаки могут включать реализацию в одной или более компьютерных программ, которые могут исполняться и/или интерпретироваться в программируемой системе, включая по меньшей мере один программируемый процессор, который может быть специализированным или общего назначения, связанный с возможностью приема данных и команд от системы хранения, по меньшей мере одного устройства ввода и по меньшей мере одного устройства вывода, и передачи данных и команд в них. Выполненная с возможностью программирования система или вычислительная система может содержать клиенты или серверы. Клиент и сервер обычно удалены друг от друга и, как правило, взаимодействуют посредством сети связи. Взаимоотношение клиента и сервера появляется благодаря компьютерным программам, исполняемым на соответствующих компьютерах и имеющих взаимосвязь клиент-сервер друг с другом.

[00622] Эти компьютерные программы, которые также можно назвать программами, программным обеспечением, программными приложениями, приложениями, компонентами или кодом, содержат машинные команды для выполненного с возможностью программирования процессора и могут быть реализованы на процедурном и/или объектно-ориентированном языке программирования высокого уровня и/или на ассемблере/машинном языке. Используемый в настоящем документе термин «машиночитаемый носитель» относится к любому компьютерному программному продукту, прибору и/или устройству, к такому как, например, магнитные диски, оптические диски, запоминающие устройства, и программируемые логические устройства (PLD), используемые для предоставления машинных команд и/или данных в выполненный с возможностью программирования процессор, в том числе машиночитаемый носитель, который принимает машинные команды как машиночитаемый сигнал. Термин «машиночитаемый сигнал» относится к любому сигналу, используемому для предоставления машинных инструкций и/или данных в процессор, выполненный с возможностью программирования. Машиночитаемый носитель может хранить такие машинные команды некратковременным образом, например как это делают некратковременное твердотельное запоминающее устройство, накопитель на жестких магнитных дисках или любой эквивалентный запоминающий носитель. Машиночитаемый носитель может в качестве альтернативы или дополнительно хранить такие машинные команды кратковременным образом, например как это делает кэш процессора или другое запоминающее устройство с произвольным доступом, связанное с одним или более физическими ядрами процессора.

[00623] Для обеспечения взаимодействия с пользователем можно реализовать один или более аспектов или признаков объекта изобретения, описанных в настоящем документе, на компьютере, имеющем устройство отображения, такое как, например, электронно-лучевая трубка (CRT), жидкокристаллический дисплей (LCD) или монитор на светоизлучающих диодах (LED) для отображения информации пользователю, и

клавиатуру или указательное устройство, такое как, например, мышь или трекбол, с помощью которого пользователь может осуществлять ввод в компьютер. Для обеспечения взаимодействия с пользователем можно также использовать устройства других видов. Например, обратная связь с пользователем может быть любой формой сенсорной обратной связи, такой как, например, визуальная обратная связь, звуковая обратная связь или тактильная обратная связь; а ввод пользователя может приниматься в любой форме, включая, без ограничений, звуковой, речевой или тактильный ввод. В число других возможных устройств ввода входят, без ограничений, сенсорные экраны или другие сенсорные устройства, такие как одно- или многоточечные резистивные или емкостные сенсорные панели, аппаратное или программное обеспечение распознавания голоса, оптические сканеры, оптические указатели, устройства захвата цифровых изображений со связанным программным обеспечением для интерпретации и т. п.

[00624] Объект изобретения, описанный в настоящем документе, может быть реализован в системах, устройствах, способах и/или изделиях в зависимости от требуемой конфигурации. Варианты осуществления, приведенные в вышеизложенном описании, представляют не все вариантах осуществления, соответствующие объекту изобретения, описанному в настоящем документе. Напротив, это всего лишь некоторые примеры, соответствующие аспектам, связанным с описанным объектом изобретения. Хотя выше приведено подробное описание нескольких вариантов, возможны другие модификации и дополнения. В частности, описанные в настоящем документе признаки и/или варианты могут быть дополнены другими признаками и/или вариантами. Например, варианты осуществления, описанные выше, могут относиться к различным комбинациям и подкомбинациям описанных признаков и/или к комбинациям и подкомбинациям нескольких дополнительных признаков, описанных выше. Кроме того, логические потоки, показанные на прилагаемых фигурах и/или описанные в настоящем документе, не обязательно требуют соблюдения конкретного показанного порядка или последовательного порядка для достижения требуемых результатов. Возможны другие варианты осуществления в объеме прилагаемой формулы изобретения.

(57) Формула изобретения

1. Платформа для анализа геномных данных, причем указанная платформа для анализа геномных данных содержит:

графический пользовательский интерфейс, который представляет множество выбираемых пользователем вариантов, где один или более из указанного множества выбираемых пользователем вариантов соответствует конкретному конвейеру обработки геномных данных; и

программный интерфейс приложения (API) платформы, который

(i) получает данные, представляющие выбор одного или более из множества выбираемых пользователем вариантов, соответствующих конкретному конвейеру обработки геномных данных, и

(ii) конфигурирует одну или более схем программируемого логического устройства для исполнения набора из одного или более конвейеров обработки геномных данных на основе полученных данных, представляющих выбор, где конфигурирование включает определение программным интерфейсом приложения (API) входов для каждого из указанных одного или более конвейеров обработки геномных данных, исполняемых указанными сконфигурированными схемами программируемого логического устройства, на основе полученных данных, представляющих выбор.

2. Платформа для анализа геномных данных по п. 1, где один или более из множества выбираемых пользователем вариантов соответствуют конкретному приложению анализа геномных данных, которое хранится в одном или более хранилищах приложений; и

где API дополнительно (iii) получает вторые данные, представляющие выбор одного или более из множества выбираемых пользователем вариантов, соответствующих конкретному приложению анализа геномных данных, и (iv) конфигурирует один или более входов для соответствующих приложений анализа геномных данных, хранящихся в одном или более хранилищах приложений, на основе полученных вторых данных.

3. Платформа для анализа геномных данных по п. 1, где указанное множество выбираемых пользователем вариантов включает один или более из выбираемого пользователем варианта, соответствующего конвейеру обработки генома, выбираемого пользователем варианта, соответствующего конвейеру обработки эпигенома, выбираемого пользователем варианта, соответствующего конвейеру обработки метагенома, выбираемого пользователем варианта, соответствующего конвейеру обработки совместного генотипирования, или выбираемого пользователем варианта, соответствующего конвейеру обработки программного обеспечения геномного анализа (GATK).

4. Платформа для анализа геномных данных по п. 2, где указанное множество выбираемых пользователем вариантов включает один или более из выбираемого пользователем варианта, соответствующего приложению неинвазивного пренатального тестирования, выбираемого пользователем варианта, соответствующего приложению отделения интенсивной терапии новорожденных, выбираемого пользователем варианта, соответствующего приложению анализа рака, выбираемого пользователем варианта, соответствующего приложению проводимых в лаборатории исследований (LDT), или выбираемого пользователем варианта, соответствующего приложению сельскохозяйственного и биологического анализа.

5. Способ анализа геномных данных, включающий:

получение, посредством программного интерфейса приложения (API), реализуемого одним или более компьютерами, первых данных, представляющих выбор одного или более из множества выбираемых пользователем вариантов, вводимых посредством графического пользовательского интерфейса, где один или более из множества выбираемых пользователем вариантов идентифицируют конкретный конвейер обработки геномных данных;

конфигурирование, с использованием API, реализуемого одним или более компьютерами, схем программируемого логического устройства для исполнения конвейера обработки геномных данных на основе указанных первых данных, где конфигурирование конвейера обработки геномных данных включает использование API для определения каждого из одного или более конвейеров обработки геномных данных, идентифицируемых первыми данными и исполняемых указанными сконфигурированными схемами программируемого логического устройства;

получение, посредством одного или более компьютеров, вторых данных, представляющих набор геномных данных или набор данных, полученных из геномных данных;

использование, посредством указанных одного или более компьютеров, указанного конвейера обработки геномных данных, сконфигурированного в указанных схемах программируемого логического устройства, на основе первых данных, для обработки полученных вторых данных;

получение, посредством указанных одного или более компьютеров, результирующих данных, генерируемых указанным конвейером обработки геномных данных, сконфигурированным в схемах программируемого логического устройства, на основе обработки конвейером обработки геномных данных полученных вторых данных; и

предоставление, указанными одним или более компьютерами, выходных данных на основе указанных результирующих данных.

6. Способ по п. 5, где указанный набор геномных данных включает одну или более геномных последовательностей, сгенерированных секвенатором нуклеиновых кислот.

7. Способ по п. 5, где указанный набор данных, полученных из геномных данных, включает набор из одного или более вариантов.

8. Способ по п. 5,

где один или более из множества выбираемых пользователем вариантов соответствуют конкретному приложению анализа геномных данных, хранящемуся в одном или более хранилищах приложений,

при этом указанные первые данные дополнительно включают данные, представляющие выбор одного или более выбираемых пользователем вариантов, каждый из которых идентифицирует одно или более приложений анализа геномных данных, и

при этом способ дополнительно включает:

конфигурирование, посредством API, входов для одного или более соответствующих приложений анализа геномных данных, хранящихся в одном или более хранилищах приложений, на основе полученных первых данных.

9. Способ по п.5, где указанное множество выбираемых пользователем вариантов включает один или более из выбираемого пользователем варианта, соответствующего конвейеру обработки генома, выбираемого пользователем варианта, соответствующего конвейеру обработки эпигенома, выбираемого пользователем варианта, соответствующего конвейеру обработки метагенома, выбираемого пользователем варианта, соответствующего конвейеру обработки совместного генотипирования, или выбираемого пользователем варианта, соответствующего конвейеру обработки программного обеспечения геномного анализа (GATK).

10. Способ по п. 9, где указанное множество выбираемых пользователем вариантов включает один или более из выбираемого пользователем варианта, соответствующего приложению неинвазивного пренатального тестирования, выбираемого пользователем варианта, соответствующего приложению отделения интенсивной терапии новорожденных, выбираемого пользователем варианта, соответствующего приложению анализа рака, выбираемого пользователем варианта, соответствующего приложению проводимых в лаборатории исследований (LDT), или выбираемого пользователем варианта, соответствующего приложению сельскохозяйственного и биологического анализа.

11. Энергонезависимый машиночитаемый носитель для анализа геномных данных, на котором хранится программное обеспечение, содержащее инструкции, которые могут быть выполнены одним или более компьютерами, которые, при таком выполнении, заставляют указанные один или более компьютеров выполнять операции, включающие:

получение, посредством программного интерфейса приложения (API), первых данных, представляющих выбор одного или более из множества выбираемых пользователем вариантов, вводимых посредством графического пользовательского интерфейса, где один или более из множества выбираемых пользователем вариантов идентифицируют

конкретный конвейер обработки геномных данных;

конфигурирование, с использованием указанного API, схем программируемого логического устройства для исполнения конвейера обработки геномных данных на основе указанных первых данных, где конфигурирование конвейера обработки геномных данных включает использование API для определения входов каждого из одного или более конвейеров обработки геномных данных, идентифицируемых первыми данными, и исполняемых схемами программируемого логического устройства;

получение вторых данных, представляющих набор геномных данных или набор данных, полученных из геномных данных;

использование указанного конвейера обработки геномных данных, сконфигурированного в схемах программируемого логического устройства на основе первых данных, для обработки полученных вторых данных;

получение результирующих данных, генерируемых указанным конвейером обработки геномных данных, сконфигурированным в схемах программируемого логического устройства, на основе обработки конвейером обработки геномных данных полученных вторых данных; и

предоставление выходных данных, которые основаны на указанных результирующих данных.

12. Машиночитаемый носитель по п. 11, где указанный набор геномных данных включает одну или более геномных последовательностей, сгенерированных секвенатором нуклеиновых кислот.

13. Машиночитаемый носитель по п. 11, где указанный набор данных, полученных из геномных данных, включает набор из одного или более вариантов.

14. Машиночитаемый носитель по п. 11,

где один или более из множества выбираемых пользователем вариантов соответствуют конкретному приложению анализа геномных данных, хранящемуся в одном или более хранилищ приложений,

где указанные первые данные дополнительно включают данные, представляющие выбор одного или более выбираемых пользователем вариантов, каждый из которых идентифицирует одно или более приложений анализа геномных данных, и

при этом указанные операции дополнительно включают:

конфигурирование, посредством API, входов для одного или более соответствующих приложений анализа геномных данных, хранящихся в одном или более хранилищах приложений, на основе полученных первых данных.

15. Машиночитаемый носитель по п. 11, где указанное множество выбираемых пользователем вариантов включает один или более из выбираемого пользователем варианта, соответствующего конвейеру обработки генома, выбираемого пользователем варианта, соответствующего конвейеру обработки эпигенома, выбираемого пользователем варианта, соответствующего конвейеру обработки метагенома, выбираемого пользователем варианта, соответствующего конвейеру обработки совместного генотипирования, или выбираемого пользователем варианта, соответствующего конвейеру обработки программного обеспечения геномного анализа (GATK).

16. Машиночитаемый носитель по п. 14, где указанное множество выбираемых пользователем вариантов включает один или более из выбираемого пользователем варианта, соответствующего приложению неинвазивного пренатального тестирования, выбираемого пользователем варианта, соответствующего приложению отделения интенсивной терапии новорожденных, выбираемого пользователем варианта,

соответствующего приложению анализа рака, выбираемого пользователем варианта, соответствующего приложению проводимых в лаборатории исследований (LDT), или выбираемого пользователем варианта, соответствующего приложению сельскохозяйственного и биологического анализа.

17. Система для анализа геномных данных, содержащая:

один или более компьютеров и одно или более запоминающих устройств, хранящих инструкции, которые при их исполнении указанными одним или более компьютерами могут заставлять указанные один или более компьютеров выполнять операции, включающие:

получение, посредством программного интерфейса приложения (API), размещенного на указанных одном или более компьютерах, первых данных, представляющих выбор одного или более из множества выбираемых пользователем вариантов, вводимых посредством графического пользовательского интерфейса, где один или более из множества выбираемых пользователем вариантов идентифицируют конкретный конвейер обработки геномных данных;

конфигурирование, с использованием указанного API, размещенного на указанных одном или более компьютерах, схем программируемого логического устройства для исполнения конвейера обработки геномных данных на основе указанных первых данных, где конфигурирование конвейера обработки геномных данных включает использование API для определения входов каждого из одного или более конвейеров обработки геномных данных, идентифицируемых первыми данными и исполняемых схемами программируемого логического устройства;

получение, посредством одного или более компьютеров, вторых данных, представляющих набор геномных данных или набор данных, полученных из геномных данных;

использование, посредством указанных одного или более компьютеров, указанного конвейера обработки геномных данных, сконфигурированного в схемах программируемого логического устройства, на основе первых данных, для обработки полученных вторых данных;

получение результирующих данных, генерируемых указанным конвейером обработки геномных данных, сконфигурированным в схемах программируемого логического устройства, на основе обработки конвейером обработки геномных данных полученных вторых данных; и

предоставление, указанными одним или более компьютерами, выходных данных на основе указанных результирующих данных.

18. Система по п. 17, где указанный набор геномных данных включает одну или более геномных последовательностей, сгенерированных секвенатором нуклеиновых кислот.

19. Система по п. 17, где указанный набор данных, полученных из геномных данных, включает набор из одного или более вариантов.

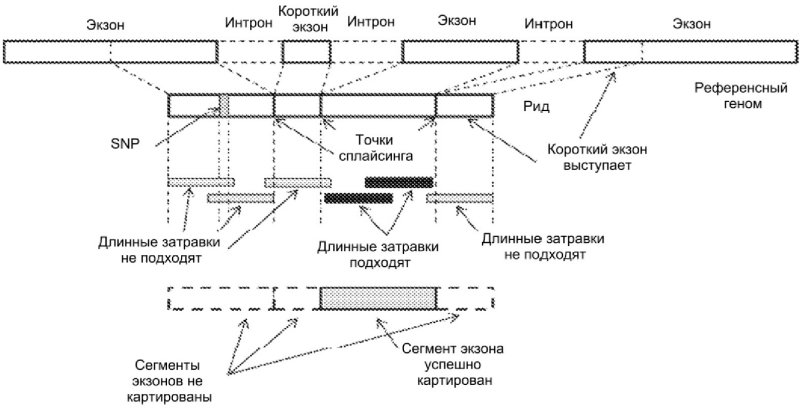
20. Система по п. 17, где один или более из множества выбираемых пользователем вариантов соответствуют конкретному приложению анализа геномных данных, хранящемуся в одном или более хранилищ данных, где указанные первые данные дополнительно включают данные, представляющие выбор одного или более выбираемых пользователем вариантов, каждый из которых идентифицирует одно или более приложений анализа геномных данных, и при этом указанные операции дополнительно включают: конфигурирование, посредством API, входов для одного или более соответствующих приложений анализа геномных данных, хранящихся в

одном или более хранилищах приложений, на основе полученных первых данных.

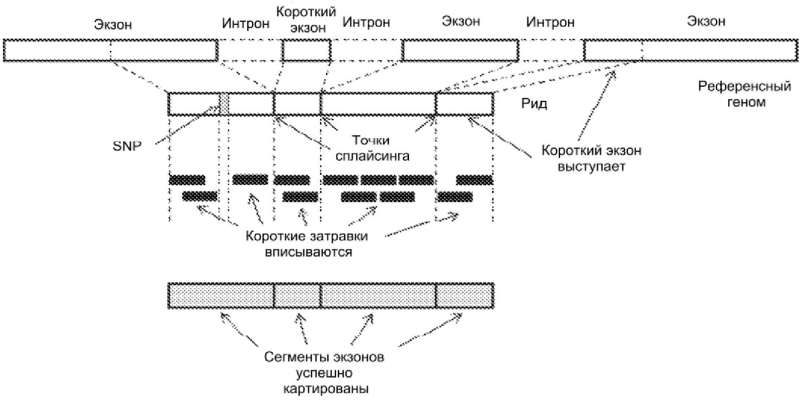
21. Система по п. 17, где указанное множество выбираемых пользователем вариантов включает один или более из выбираемого пользователем варианта, соответствующего конвейеру обработки генома, выбираемого пользователем варианта, соответствующего конвейеру обработки эпигенома, выбираемого пользователем варианта, соответствующего конвейеру обработки метагенома, выбираемого пользователем варианта, соответствующего конвейеру обработки совместного генотипирования, выбираемого пользователем варианта, соответствующего конвейеру обработки программного обеспечения геномного анализа (GATK).

22. Система по п. 20, где указанное множество выбираемых пользователем вариантов включает один или более из выбираемого пользователем варианта, соответствующего приложению неинвазивного пренатального тестирования, выбираемого пользователем варианта, соответствующего приложению отделения интенсивной терапии новорожденных, выбираемого пользователем варианта, соответствующего приложению анализа рака, выбираемого пользователем варианта, соответствующего приложению проводимых в лаборатории исследований (LDT), или выбираемого пользователем варианта, соответствующего приложению сельскохозяйственного и биологического анализа.

1

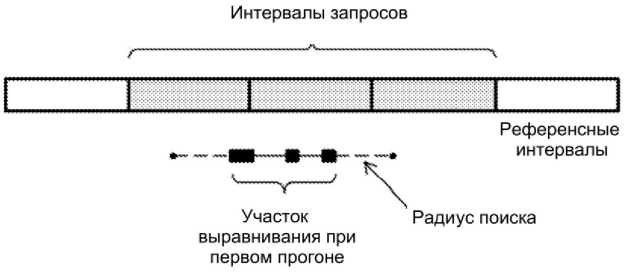

Экзон
Интрон
Короткий экзон
Интрон
Экзон
Интрон
Экзон
Референсный геном
Рид
SNP
Точки сплайсинга
Короткий экзон выступает
Длинные затравки не подходят
Длинные затравки подходят
Длинные затравки не подходят
Сегменты экзонов не картированы
Сегмент экзона успешно картирован

ФИГ. 1

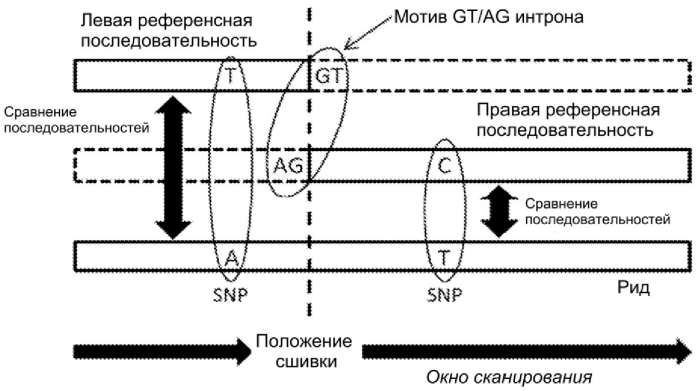

Экзон
Интрон
Короткий экзон
Интрон
Экзон
Интрон
Экзон
Референсный геном
Рид
SNP
Точки сплайсинга
Короткий экзон выступает
Короткие затравки вписываются
Сегменты экзонов успешно картированы

ФИГ. 2

2

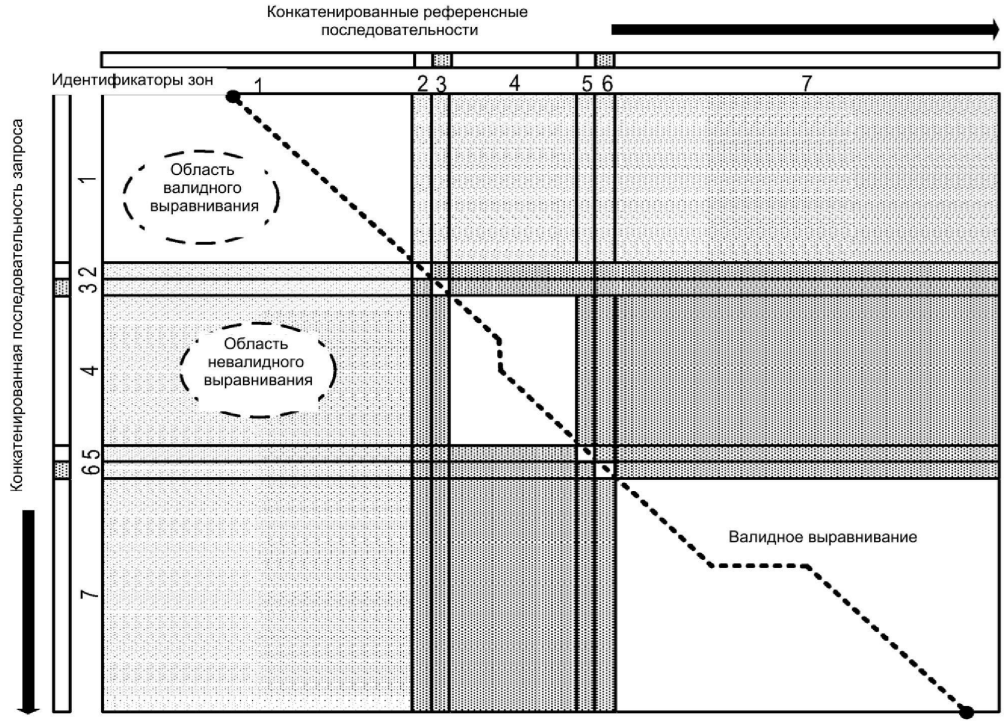

Интервалы запросов
Референсные интервалы
Участок выравнивания при первом прогоне
Радиус поиска

ФИГ. 3


Левая референсная последовательность
Мотив GT/AG интрона
T
GT
Сравнение последовательностей
Правая референсная последовательность
AG
C
Сравнение последовательностей
A
T
SNP
SNP
Рид
Положение сшивки
Окно сканирования

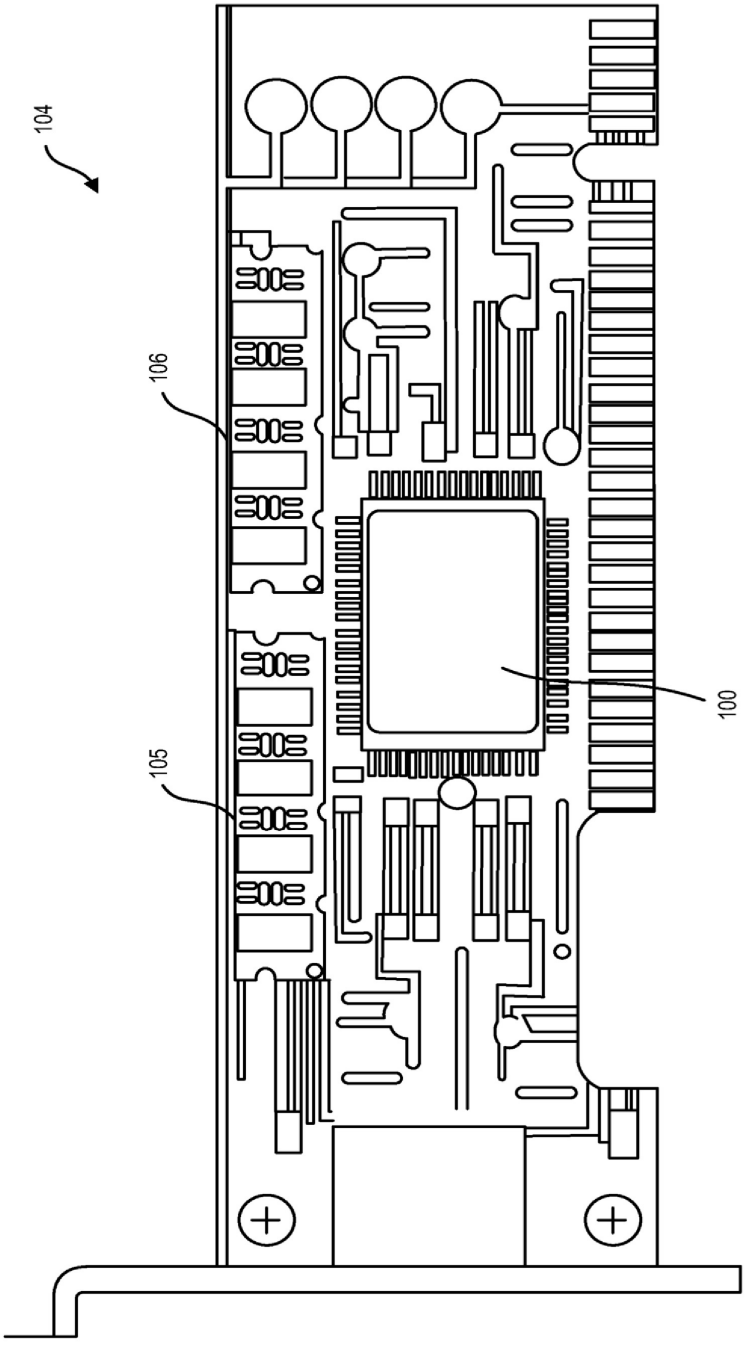
ФИГ. 4

3

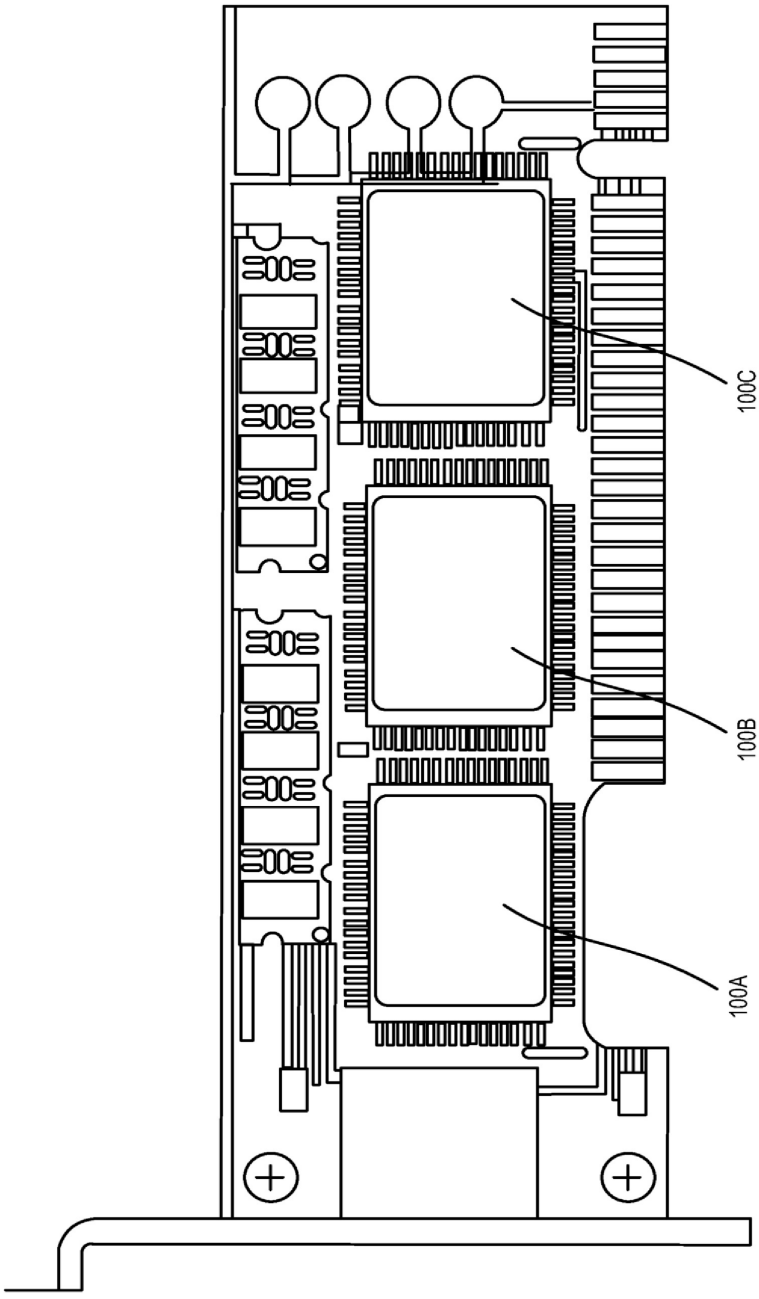

Конкатенированные референсные последовательности
Идентификаторы зон
1
2 3
4
5 6
7
Конкатенированная последовательность запроса
1
3 2
4
6 5
7
Область валидного выравнивания
Область невалидного выравнивания
Валидное выравнивание

ФИГ. 5

4

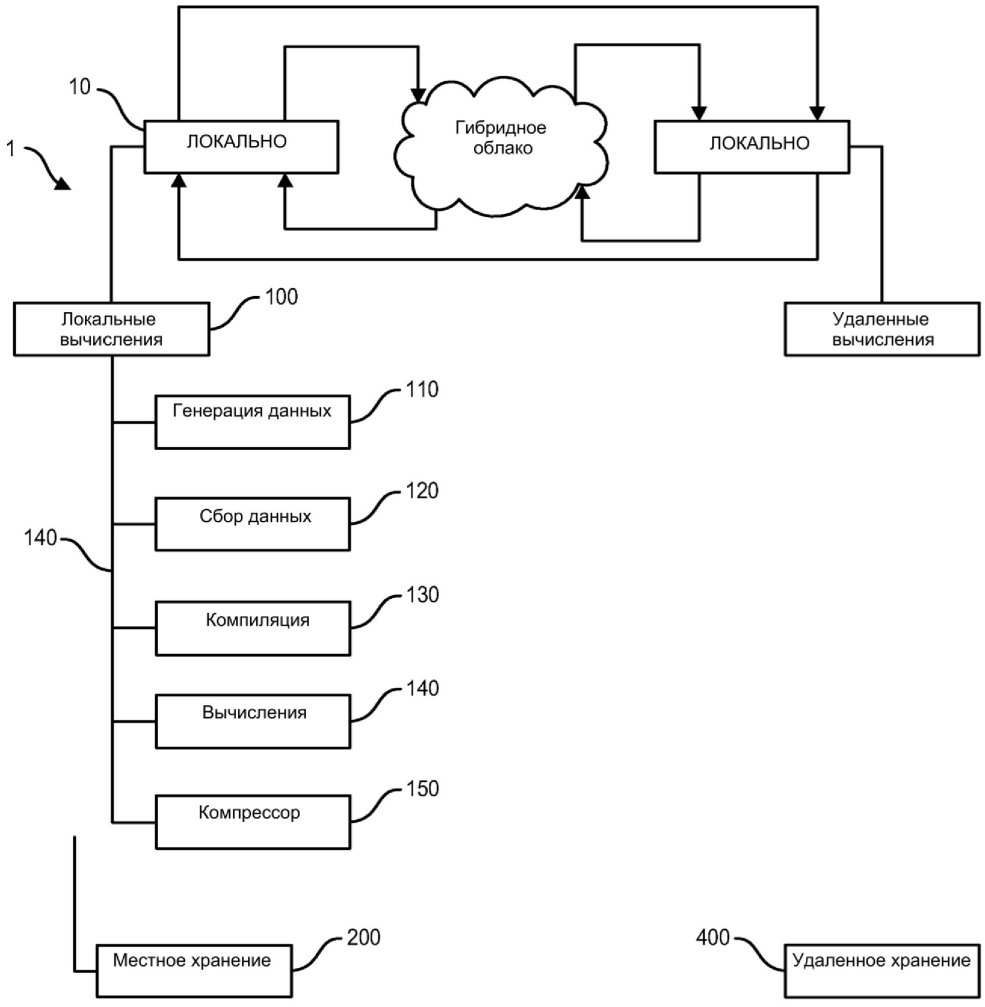


ФИГ. 6



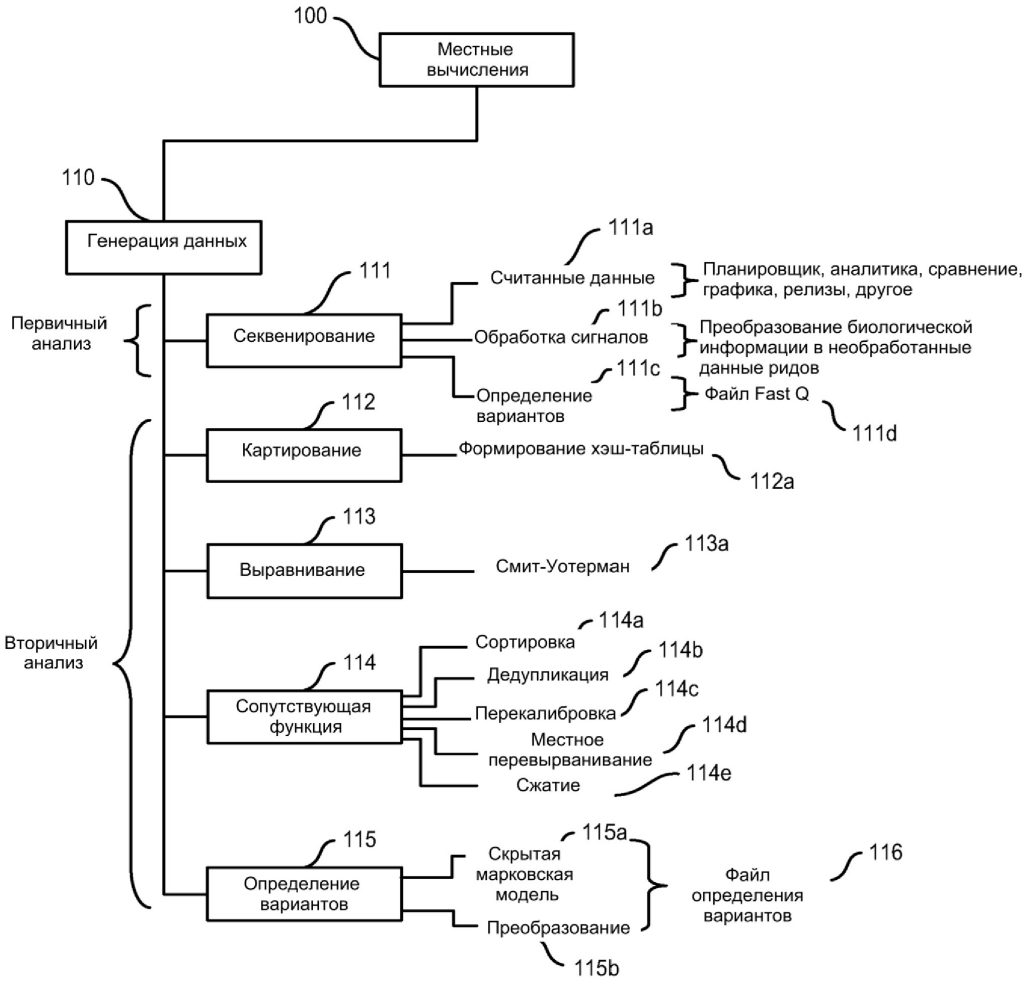
ФИГ. 7

6

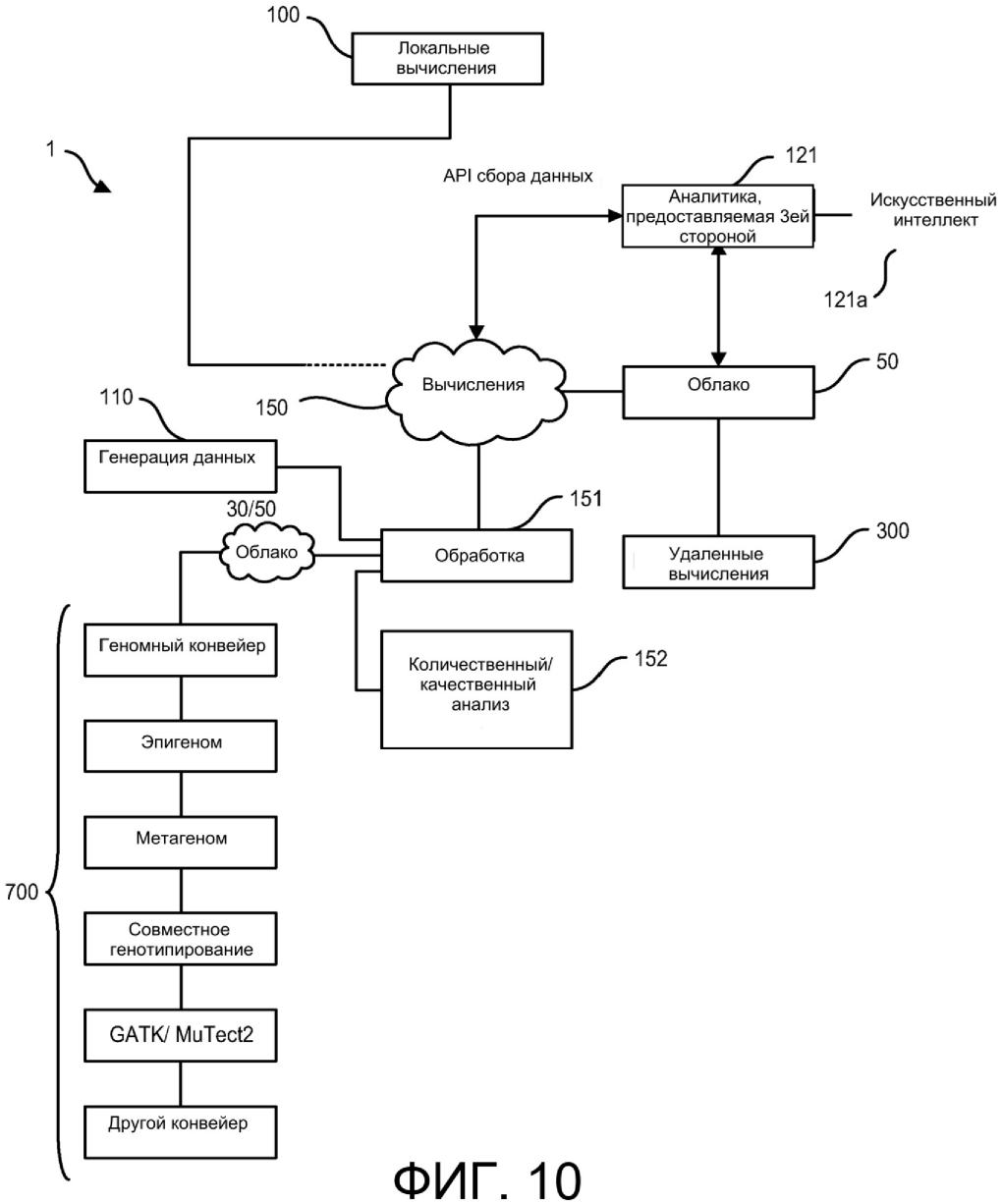

10
1
ЛОКАЛЬНО
Гибридное облако
ЛОКАЛЬНО
Локальные вычисления
100
Удаленные вычисления
Генерация данных
110
Сбор данных
120
140
Компиляция
130
Вычисления
140
Компрессор
150
Местное хранение
200
400
Удаленное хранение

ФИГ. 8

7

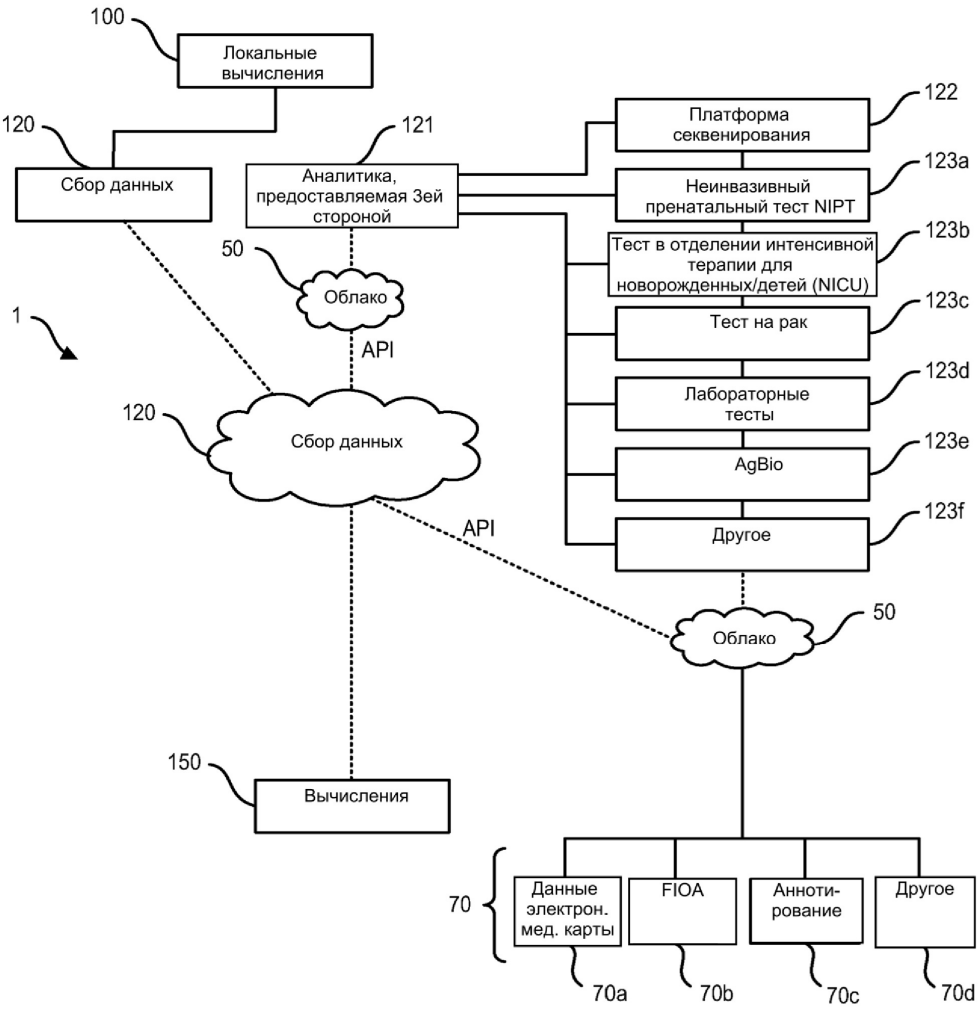

100
Местные вычисления
110
Генерация данных
Первичный анализ
111
Секвенирование
111a
Считанные данные
Планировщик, аналитика, сравнение, графика, релизы, другое
111b
Обработка сигналов
Преобразование биологической информации в необработанные данные ридов
111c
Определение вариантов
Файл Fast Q
111d
Вторичный анализ
112
Картирование
Формирование хэш-таблицы
112a
113
Выравнивание
Смит-Уотерман
113a
114
Сопутствующая функция
114a
Сортировка
114b
Дедупликация
114c
Перекалибровка
114d
Местное перевырванивание
114e
Сжатие
115
Определение вариантов
115a
Скрытая марковская модель
Преобразование
115b
116
Файл определения вариантов

ФИГ. 9


100
Локальные вычисления
1
API сбора данных
121
Аналитика, предоставляемая 3ей стороной
Искусственный интеллект
121a
Вычисления
150
Облако
50
110
Генерация данных
30/50
Облако
151
Обработка
Удаленные вычисления
300
Количественный/ качественный анализ
152
700
Геномный конвейер
Эпигеном
Метагеном
Совместное генотипирование
GATK/ MuTect2
Другой конвейер

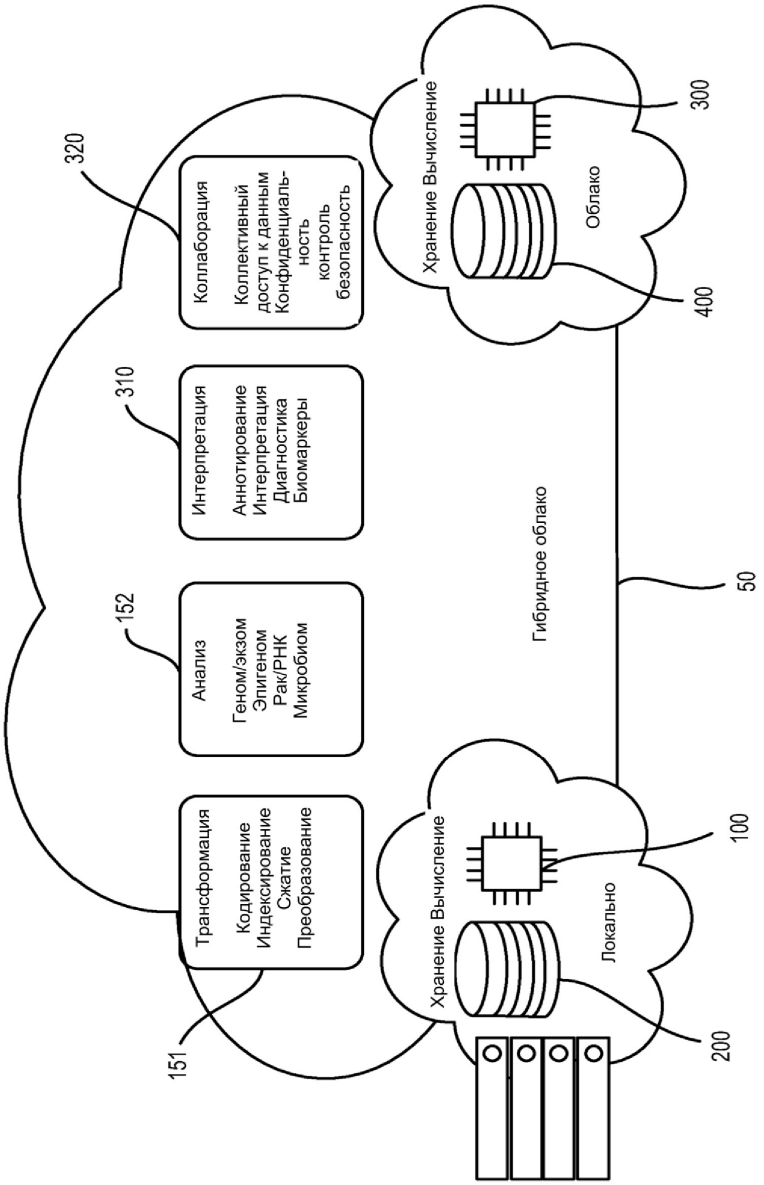
ФИГ. 10

9


100
Локальные вычисления
120
Сбор данных
121
Аналитика, предоставляемая Зей стороной
122
Платформа секвенирования
123a
Неинвазивный пренатальный тест NIPT
123b
Тест в отделении интенсивной терапии для новорожденных/детей (NICU)
123c
Тест на рак
123d
Лабораторные тесты
123e
AgBio
123f
Другое
50
Облако
1
API
120
Сбор данных
API
50
Облако
150
Вычисления
70
Данные электрон. мед. карты
FIOA
Аннотирование
Другое
70a
70b
70c
70d

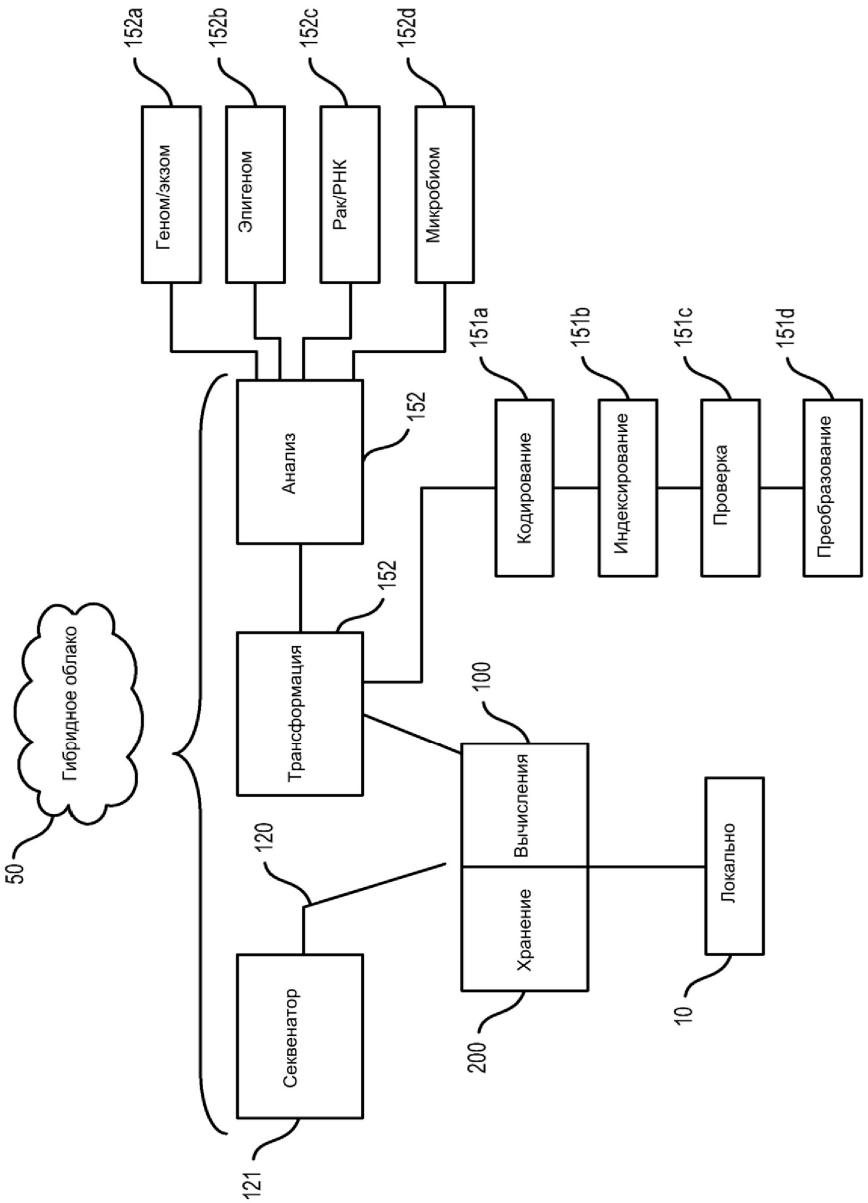
ФИГ. 11

10

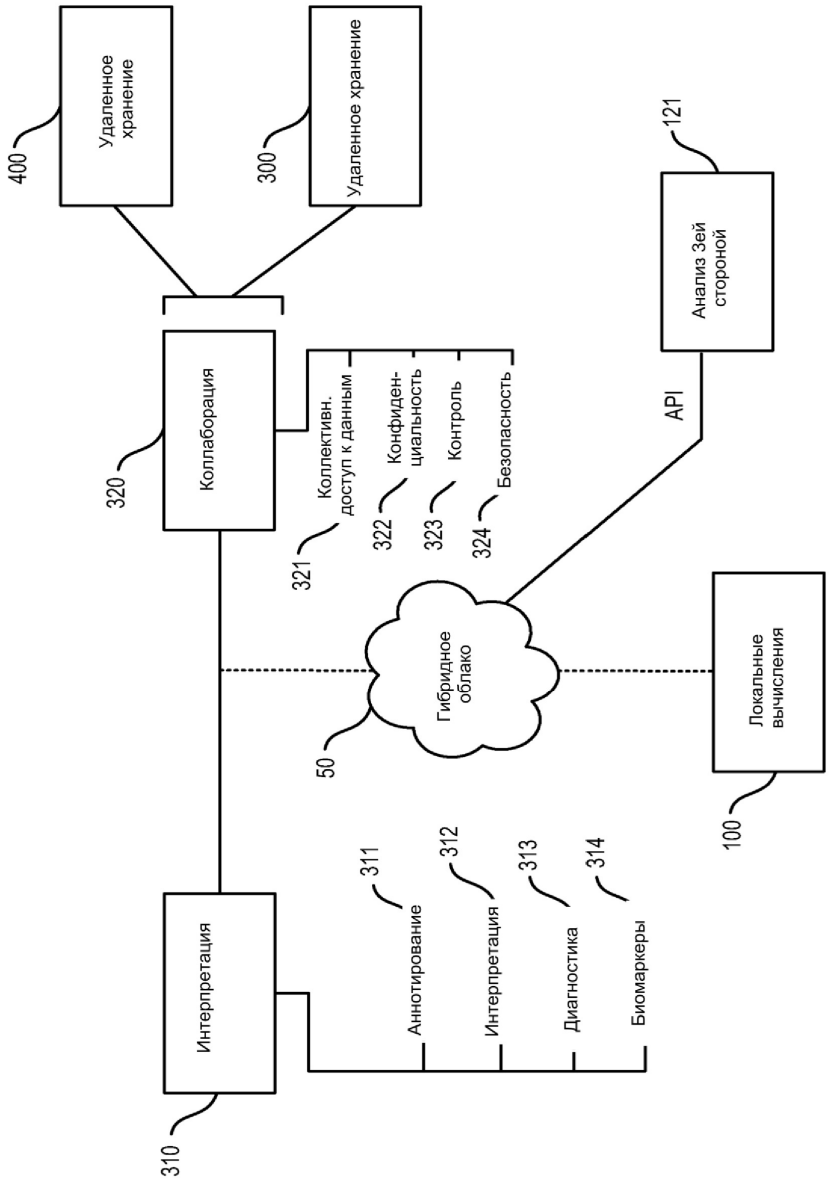

320
Коллаборация
Коллективный доступ к данным
Конфиденциальность
контроль безопасность
310
Интерпретация
Аннотирование
Интерпретация
Диагностика
Биомаркеры
152
Анализ
Геном/экзом
Эпигеном
Рак/РНК
Микробиом
Трансформация
Кодирование
Индексирование
Сжатие
Преобразование
151
Гибридное облако
50
Хранение Вычисление
Облако
300
400
Хранение Вычисление
Локально
100
200

ФИГ. 12

11

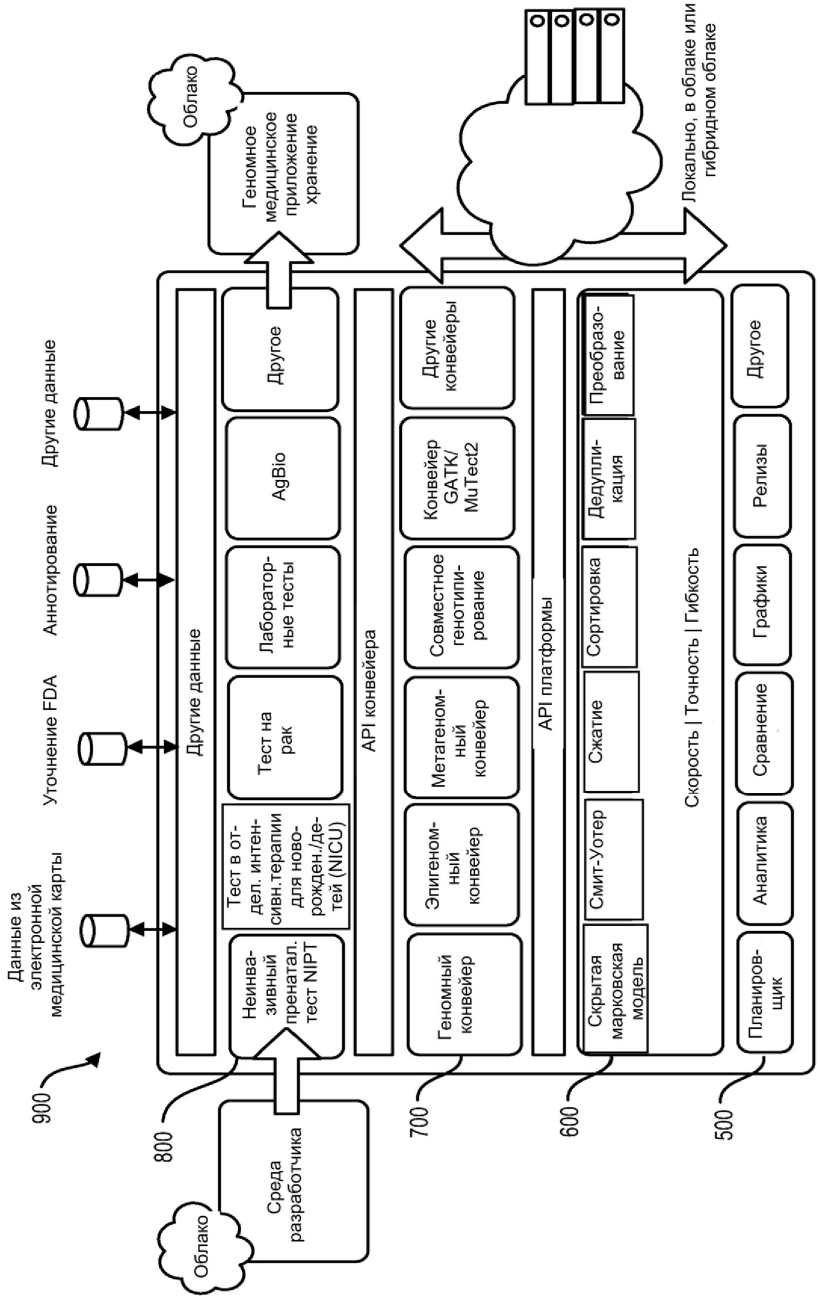

152a
152b
152c
152d
Геном/экзом
Эпигеном
Рак/РНК
Микробиом
151a
151b
151c
151d
Анализ
152
Кодирование
Индексирование
Проверка
Преобразование
Гибридное облако
50
Трансформация
152
100
Вычисления
Хранение
Локально
120
Секвенатор
121
200
10

ФИГ. 13

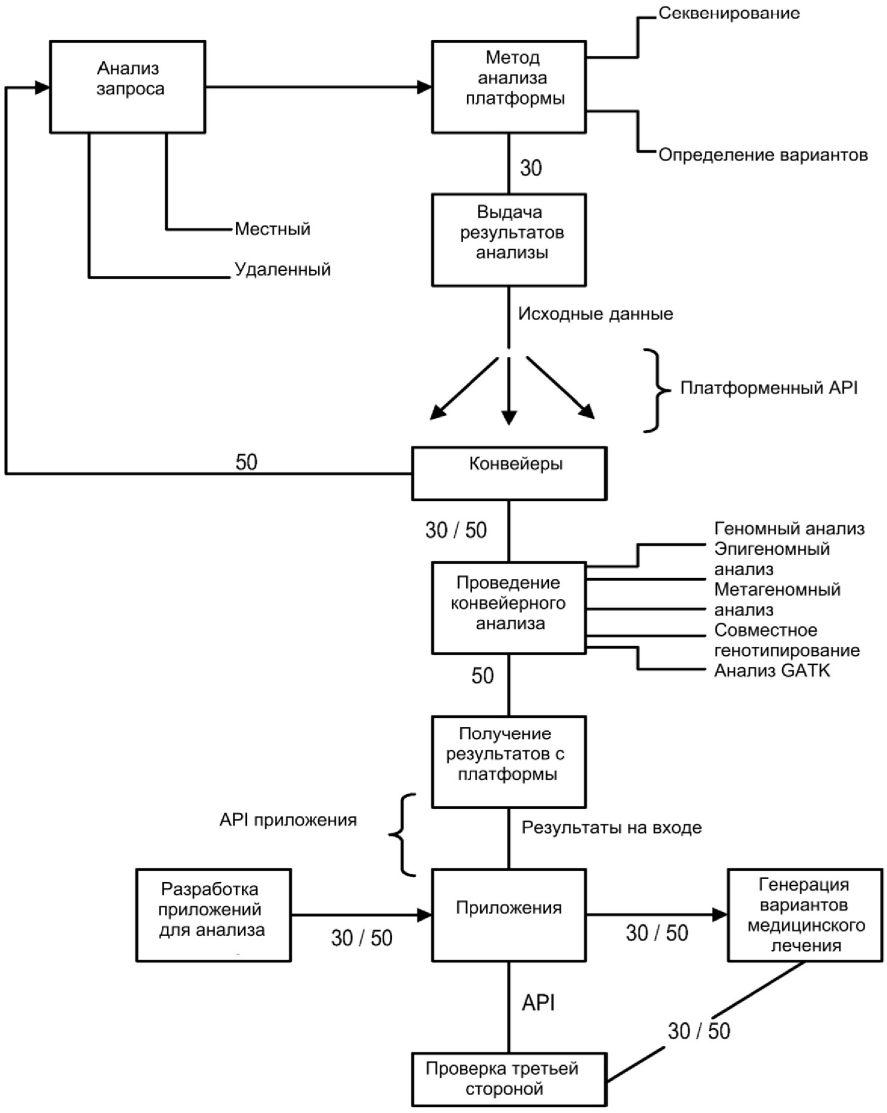

Удаленное хранение
400
Удаленное хранение
300
121
Анализ Зей стороной
Коллаборация
320
Коллективн. доступ к данным
321
Конфиден-циальность
322
Контроль
323
Безопасность
324
API
Гибридное облако
50
Локальные вычисления
100
Интерпретация
310
311
Аннотирование
312
Интерпретация
313
Диагностика
314
Биомаркеры

ФИГ. 14

13

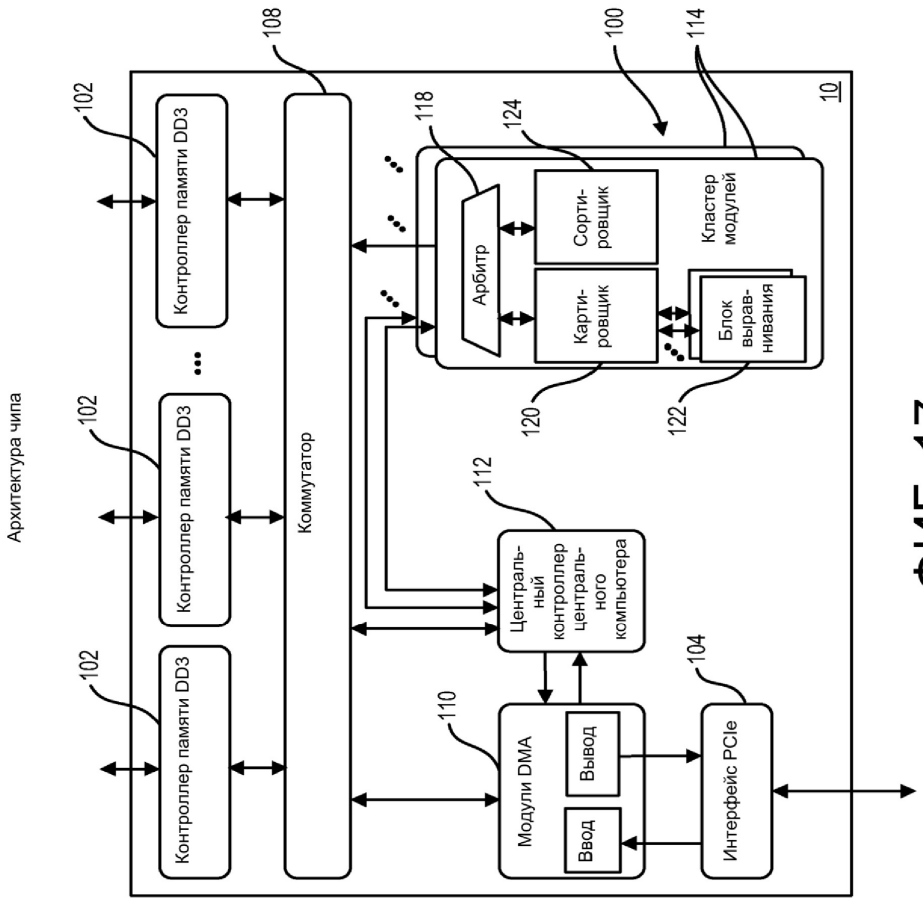

Данные из электронной медицинской карты
Уточнение FDA
Аннотирование
Другие данные
Облако
Геномное медицинское приложение хранение
Локально, в облаке или гибридном облаке
Другое
AgBio
Лабораторные тесты
Тест на рак
Тест в отдел. интенсивн.терапии для новорожден./детей (NICU)
Неинвазивный пренатал. тест NIPT
Другие данные
API конвейера
Другие конвейеры
Конвейер GATK/ MuTect2
Совместное генотипирование
Метагеномный конвейер
Эпигеномный конвейер
Геномный конвейер
API платформы
Преобразование
Дедупликация
Сортировка
Сжатие
Смит-Уотер
Скрытая марковская модель
Скорость | Точность | Гибкость
Другое
Релизы
Графики
Сравнение
Аналитика
Планировщик
900
800
700
600
500
Облако
Среда разработчика

ФИГ. 15


Анализ запроса
Метод анализа платформы
Секвенирование
Определение вариантов
30
Выдача результатов анализы
Местный
Удаленный
Исходные данные
Платформенный API
50
Конвейеры
30 / 50
Геномный анализ
Эпигеномный анализ
Метагеномный анализ
Совместное генотипирование
Анализ GATK
Проведение конвейерного анализа
50
Получение результатов с платформы
API приложения
Результаты на входе
Разработка приложений для анализа
30 / 50
Приложения
30 / 50
Генерация вариантов медицинского лечения
API
30 / 50
Проверка третьей стороной

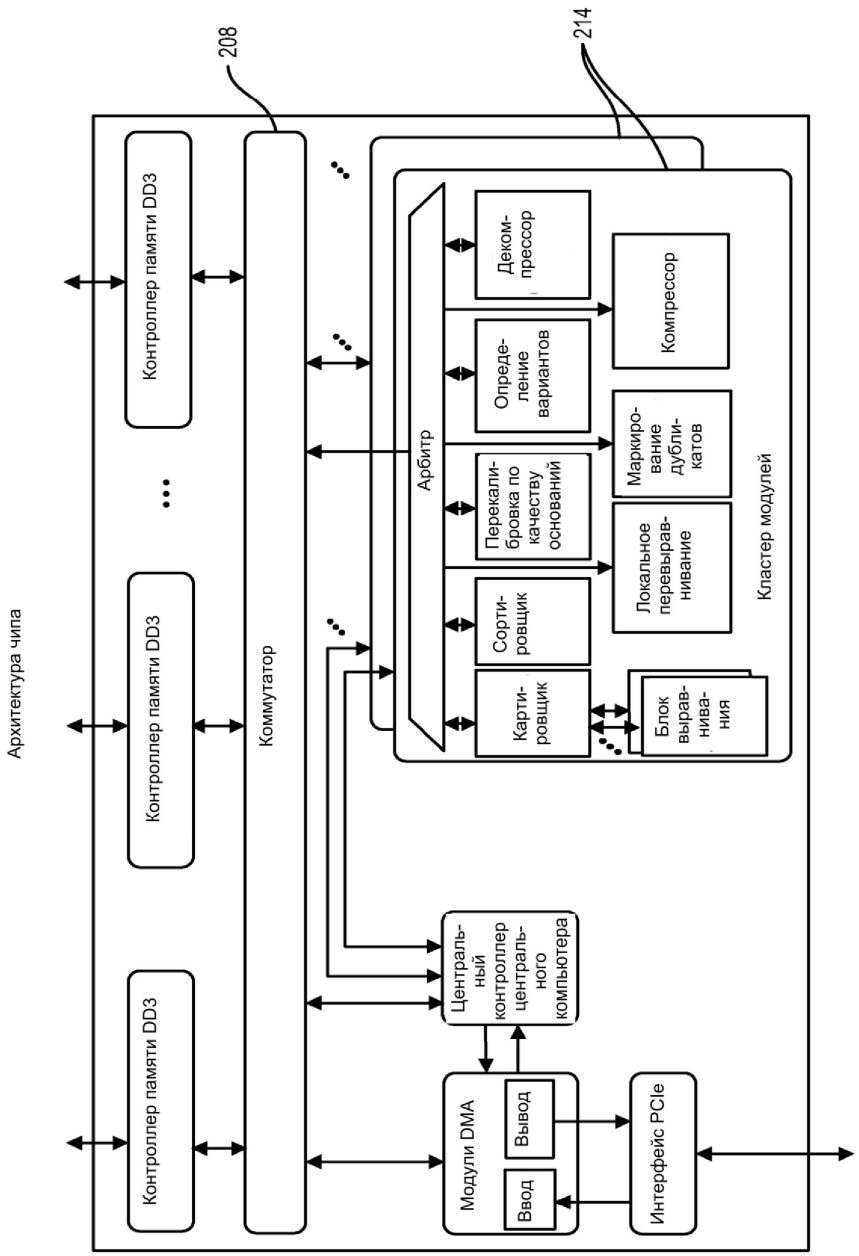
ФИГ. 16

15


Архитектура чипа
10
100
102
102
102
104
108
110
112
114
118
120
122
124
Контроллер памяти DD3
Контроллер памяти DD3
Контроллер памяти DD3
Коммутатор
Арбитр
Сорти-ровщик
Карти-ровщик
Блок выров-нивания
Кластер модулей
Централь-ный контроллер централь-ного компьютера
Модули DMA
Вывод
Ввод
Интерфейс PCIe

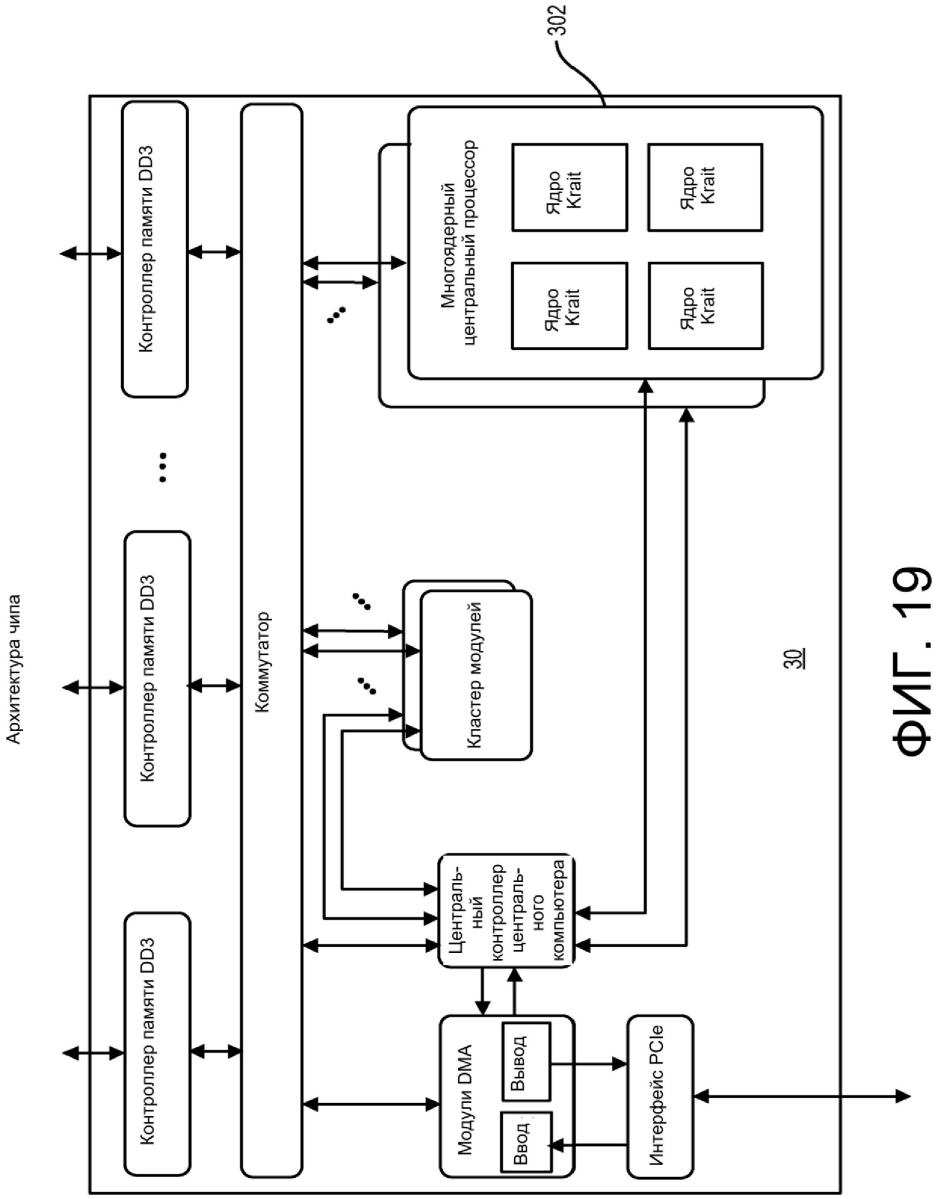
ФИГ. 17

16


Архитектура чипа
208
214
Контроллер памяти DD3
Контроллер памяти DD3
Контроллер памяти DD3
Коммутатор
Арбитр
Декомпрессор
Определение вариантов
Перекалибровка по качеству оснований
Сортировщик
Картировщик
Компрессор
Маркирование дубликатов
Локальное перевыравнивание
Блок выравнивания
Кластер модулей
Центральный контроллер центрального компьютера
Модули DMA
Вывод
Ввод
Интерфейс PCIe

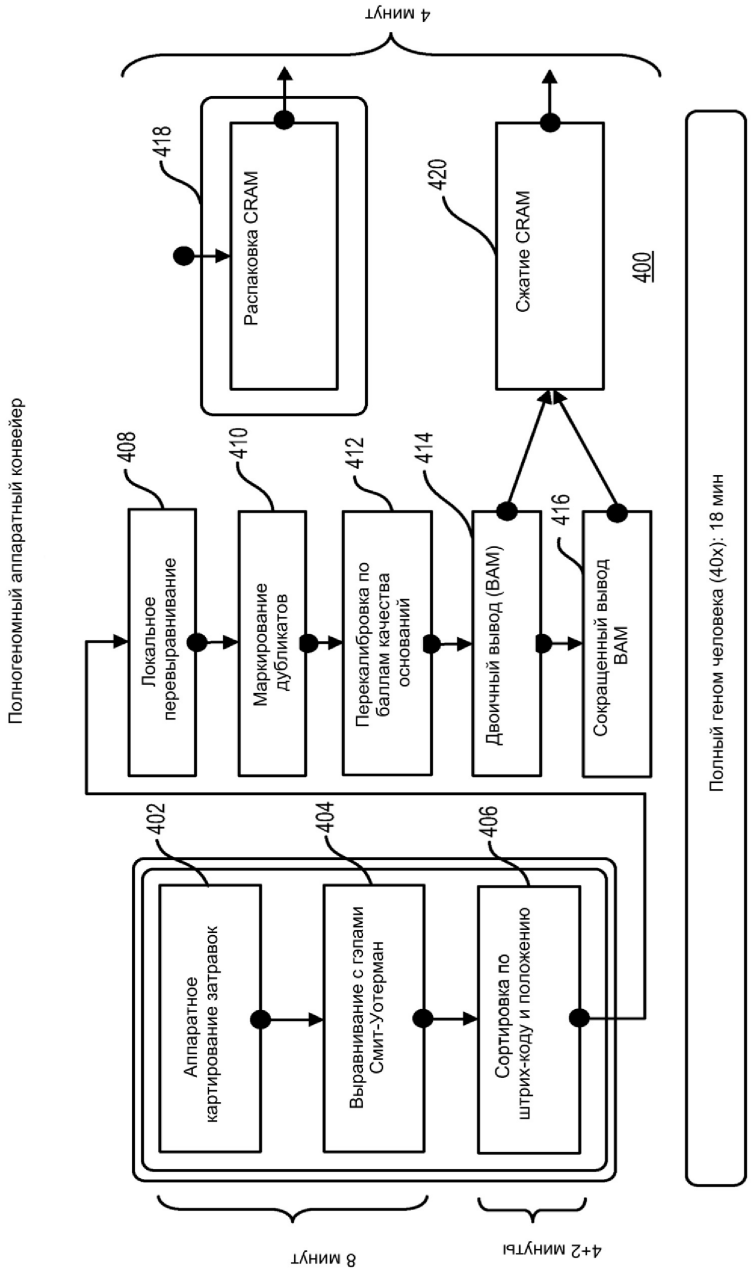
ФИГ. 18

17


302
Архитектура чипа
Контроллер памяти DD3
Контроллер памяти DD3
Контроллер памяти DD3
Коммутатор
Многоядерный центральный процессор
Ядро Krait
Ядро Krait
Ядро Krait
Ядро Krait
Кластер модулей
Централь-ный контроллер централь-ного компьютера
Модули DMA
Вывод
Ввод
Интерфейс PCIe
30

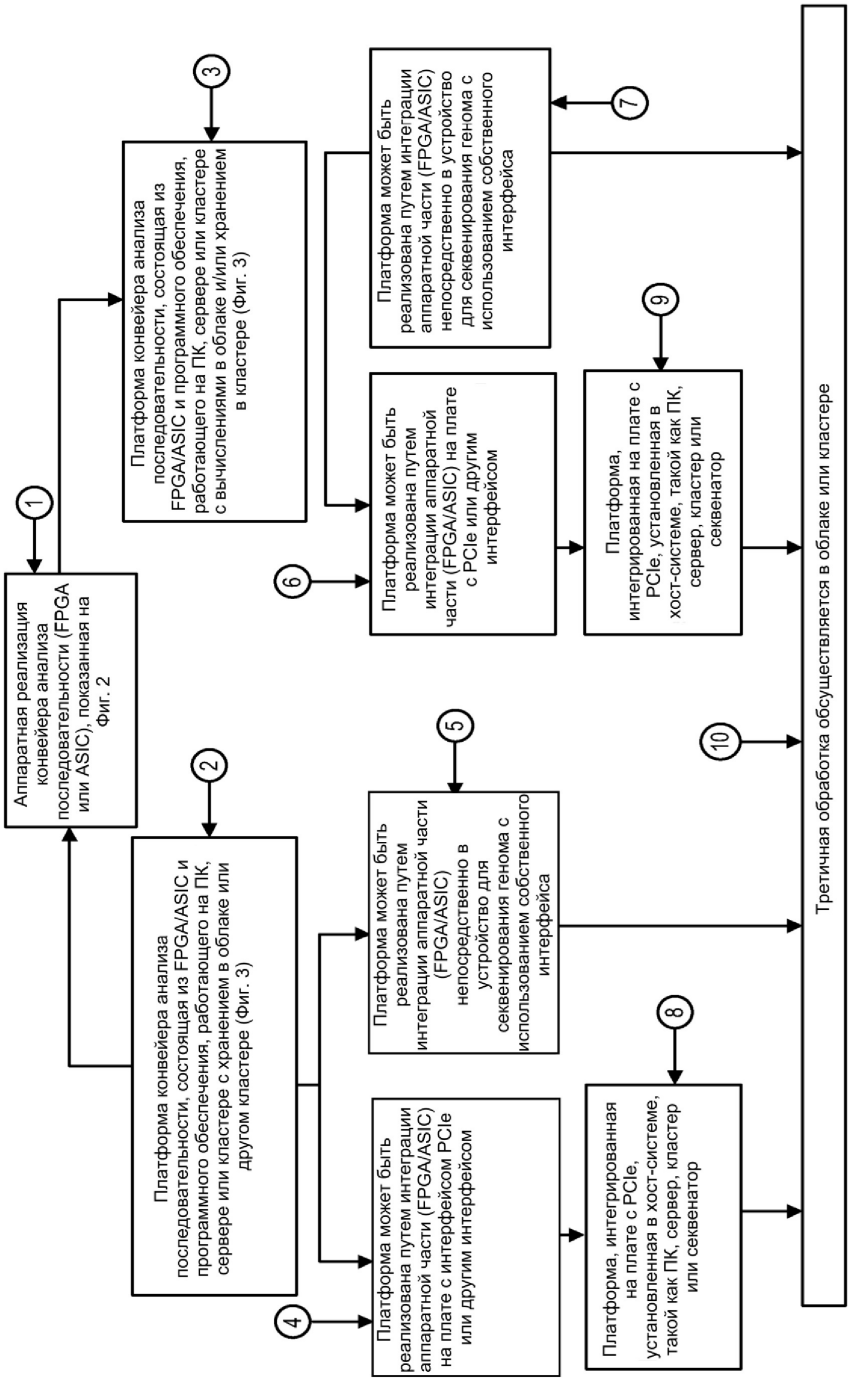
ФИГ. 19

18


Полногеномный аппаратный конвейер
4 минут
418
Распаковка CRAM
420
Сжатие CRAM
400
408
Локальное перевыравнивание
410
Маркирование дубликатов
412
Перекалибровка по баллам качества оснований
414
Двоичный вывод (BAM)
416
Сокращенный вывод BAM
402
Аппаратное картирование затравок
404
Выравнивание с гэпами Смит-Уотерман
406
Сортировка по штрих-коду и положению
8 минут
4+2 минуты
Полный геном человека (40x): 18 мин

ФИГ. 20

19


Аппаратная реализация конвейера анализа последовательности (FPGA или ASIC), показанная на Фиг. 2
Платформа конвейера анализа последовательности, состоящая из FPGA/ASIC и программного обеспечения, работающего на ПК, сервере или кластере с вычислениями в облаке и/или хранением в кластере (Фиг. 3)
Платформа конвейера анализа последовательности, состоящая из FPGA/ASIC и программного обеспечения, работающего на ПК, сервере или кластере с хранением в облаке или другом кластере (Фиг. 3)
Платформа может быть реализована путем интеграции аппаратной части (FPGA/ASIC) на плате с интерфейсом PCIe или другим интерфейсом
Платформа может быть реализована путем интеграции аппаратной части (FPGA/ASIC) непосредственно в устройство для секвенирования генома с использованием собственного интерфейса
Платформа может быть реализована путем интеграции аппаратной части (FPGA/ASIC) на плате с PCIe или другим интерфейсом
Платформа может быть реализована путем интеграции аппаратной части (FPGA/ASIC) непосредственно в устройство для секвенирования генома с использованием собственного интерфейса
Платформа, интегрированная на плате с PCIe, установленная в хост-системе, такой как ПК, сервер, кластер или секвенатор
Платформа, интегрированная на плате с PCIe, установленная в хост-системе, такой как ПК, сервер, кластер или секвенатор
Третичная обработка обсуществляется в облаке или кластере
1
2
3
4
5
6
7
8
9
10

ФИГ. 21