(19) Евразийское патентное ведомство

(11) 045158 (13) B1

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ЕВРАЗИЙСКОМУ ПАТЕНТУ

(45) Дата публикации и выдачи патента
2023.10.31

(21) Номер заявки
202090277

(22) Дата подачи заявки
2016.12.20

(51) Int. Cl. ***C12Q 1/6869*** (2006.01)
C12Q 1/6883 (2006.01)
G06F 17/18 (2006.01)

(54) ПРИМЕНЕНИЕ РАЗМЕРА ФРАГМЕНТА БЕСКЛЕТОЧНОЙ ДНК ДЛЯ ОПРЕДЕЛЕНИЯ ВАРИАЦИЙ ЧИСЛА КОПИЙ

(31) **62/290,891; 15/382,508**
(32) **2016.02.03; 2016.12.16**
(33) **US**
(43) **2020.07.31**
(62) **201891580; 2016.12.20**

(71)(73) Заявитель и патентовладелец:
ВЕРИНАТА ХЭЛС, ИНК. (US)

(72) Изобретатель:
Дюнвальд Свен, Комсток Дэвид А., Барбакиору Каталин, Чудова Дарья И., Рава Ричард П., Джонс Кит В., Чэнь Гэнсинь, Скворцов Дмитрий (US)

(74) Представитель:
Нилова М.И. (RU)

(56) WO-A2-2014149134
WO-A1-2015083004

(57) Раскрыты способы определения вариации числа копий (ВЧК), которая определенно или предположительно связана с множеством медицинских состояний. Согласно некоторым вариантам реализации предложены способы определения вариации числа копий плодов с применением материнских образцов, содержащих материнскую и бесклеточную ДНК плода. Согласно некоторым вариантам реализации предложены способы определения ВЧК, которые определенно или предположительно связаны с множеством медицинских состояний. В некоторых вариантах реализации, раскрытых в настоящем документе, предложены способы улучшения чувствительности и/или специфичности анализа данных о последовательности посредством определения параметра размера фрагмента. Согласно некоторым вариантам реализации для оценки вариаций числа копий применяют информацию по фрагментам различных размеров. Согласно некоторым вариантам реализации для оценки вариаций числа копий применяют один или более параметров t-статистики, полученных из информации о перекрытии последовательности, представляющей интерес. Согласно некоторым вариантам реализации для определения вариаций числа копий одну или более оценок фракции плода объединяют с одним или более параметрами t-статистики.

045158 B1

045158 B1

Перекрестная ссылка на родственные заявки

Настоящая заявка испрашивает приоритет согласно разделу 35 § 119(e) Свода законов США на основании предварительной заявки на патент США № 62/290891, озаглавленной: "Применение размера фрагмента бесклеточной ДНК для определения вариаций числа копий", поданной 3 февраля 2016 года, и заявки на патент США № 15/382508, озаглавленной: "Применение размера фрагмента бесклеточной ДНК для определения вариаций числа копий", поданной 16 декабря 2016 года, которые полностью включены в настоящую заявку посредством ссылки для всех целей.

Уровень техники

Одним из важнейших начинаний в исследованиях в области медицины человека является открытие генетических аномалий, вызывающих неблагоприятные последствия для здоровья. Во многих случаях были идентифицированы конкретные гены и/или важные диагностические маркеры в частях генома, которые присутствуют в абнормальных количествах копий. Например, в пренатальной диагностике часто встречающимися генетическими поражениями являются дополнительные или отсутствующие копии целых хромосом. При раке частыми событиями являются делеция или умножение копий целых хромосом или сегментов хромосом и более высокий уровень амплификации определенных областей генома.

Большая часть информации о вариации числа копий (ВЧК) была получена посредством цитогенетического разрешения, которое обеспечило распознавание структурных аномалий. В общепринятых процедурах генетического скрининга и биологической дозиметрии с целью получения клеток для кариотипического анализа применяют инвазивные процедуры, например амниоцентез, кордоцентез или биопсию ворсин хориона (БВХ). В связи с осознанием потребности в более быстрых методах исследования, для которых не требуется культивирование клеток, были разработаны флуоресцентная гибридизация in situ (fluorescence in situ hybridization, FISH), количественная флуоресцентная ПЦР (КФ-ПЦР) и матриксная сравнительная геномная гибридизация (array-Comparative Genomic Hybridization, array-CGH) в качестве молекулярно-цитогенетических способов для анализа вариаций числа копий.

Появление технологий, позволяющих проводить секвенирование целых геномов в течение относительно короткого периода времени, и открытие циркулирующей бесклеточной ДНК (cell-free DNA, сцДНК) обеспечило возможность сравнивать сравниваемый генетический материал, полученный из одной хромосомы, с другим материалом другой хромосомы при отсутствии рисков, связанных с инвазивными способами отбора образца, что обеспечивает инструмент для диагностики различных видов вариаций числа копий генетических последовательностей, представляющих интерес.

Ограничения существующих способов неинвазивной пренатальной диагностики, включающие недостаточную чувствительность, которая является следствием ограниченных уровней сцДНК, и погрешность технологий секвенирования, которая является следствием природных свойств геномной информации, лежат в основе сохраняющейся потребности в неинвазивных способах, которые обеспечили бы всё или часть из специфичности, чувствительности и пригодности для надежной диагностики изменений числа копий в различных клинических условиях. Было показано, что в плазме беременных женщин средние длины фрагментов сцДНК плода являются более короткими, чем фрагментов материнской сцДНК. Эту разницу между материнской сцДНК и сцДНК плода используют в предложенном в настоящем документе решении для определения ВЧК и/или фракции плода. Варианты реализации, раскрытые в настоящем документе, удовлетворяют некоторые из указанных выше потребностей. Некоторые варианты реализации можно осуществить с применением библиотеки, полученной ПЦР, в сочетании с секвенированием спаренных концов ДНК. Некоторые варианты реализации обеспечивают высокую аналитическую чувствительность и специфичность для неинвазивной пренатальной диагностики и диагностики множества заболеваний.

Краткое описание изобретения

Согласно некоторым вариантам реализации предложены способы определения вариации числа копий (ВЧК) любой из анеуплоидий плода и ВЧК, которая определенно или предположительно связана с множеством медицинских состояний. ВЧК, которые можно определить согласно настоящему способу, включают трисомии и моносомии любой одной или более из хромосом 1-22, X и Y, другие полисомии хромосом и делеции и/или дупликации сегментов любой одной или более хромосом. Согласно некоторым вариантам реализации способы включают идентификацию ВЧК в последовательности нуклеиновой кислоты, представляющей интерес, например клинически значимой последовательности, в исследуемом образце. Способ позволяет оценить вариацию числа копий конкретной последовательности, представляющей интерес.

Согласно некоторым вариантам реализации способ реализуют с применением компьютерной системы, которая содержит один или более процессоров и системную память для оценки числа копий последовательности нуклеиновой кислоты, представляющей интерес, в исследуемом образце, содержащем нуклеиновые кислоты одного или более геномов.

Один аспект настоящего изобретения относится к способу определения вариации числа копий (ВЧК) последовательности нуклеиновой кислоты, представляющей интерес, в исследуемом образце, содержащем фрагменты бесклеточной нуклеиновой кислоты, полученные из двух или более геномов. Способ включает: (a) прием ридов последовательности, полученных в результате секвенирования фрагмен-

тов бесклеточной нуклеиновой кислоты в исследуемом образце; (b) выравнивание ридов последовательности фрагментов бесклеточной нуклеиновой кислоты или выравнивание фрагментов, содержащих риды последовательности, с блоками (bins) референсного генома, содержащего последовательность, представляющую интерес, с получением, таким образом, меток исследуемой последовательности, причем референсный геном разделен на множество блоков; (c) определение размера фрагментов по меньшей мере некоторых фрагментов бесклеточной нуклеиновой кислоты, присутствующих в исследуемом образце; (d) вычисление перекрытий меток последовательности для блоков референсного генома для каждого блока посредством: (i) определения количества меток последовательности, которые выравниваются с блоком, и (ii) нормирования количества меток последовательности, которые выравниваются с блоком, посредством вычисления межблоковых вариаций, вызванных факторами, отличными от вариации числа копий; (e) определение t-статистики для последовательности, представляющей интерес, с применением перекрытий блоков в последовательности, представляющей интерес, и перекрытий блоков в референсной области для последовательности, представляющей интерес; и (f) определение вариации числа копий в последовательности, представляющей интерес, с применением отношения правдоподобия, вычисленного t-статистике, и информации относительно размера фрагментов бесклеточной нуклеиновой кислоты.

Согласно некоторым вариантам реализации способ включает осуществление этапов (d) и (e) дважды, один раз для фрагментов в первом домене размеров и повторно для фрагментов во втором домене размеров. Согласно некоторым вариантам реализации первый домен размеров включает фрагменты бесклеточной нуклеиновой кислоты по существу всех размеров в образце, и второй домен размеров включает только фрагменты бесклеточной нуклеиновой кислоты, меньшие, чем заданный размер. Согласно некоторым вариантам реализации второй домен размеров включает только фрагменты бесклеточной нуклеиновой кислоты, меньшие, чем приблизительно 150 п.о. Согласно некоторым вариантам реализации отношение правдоподобия вычисляют по первой t-статистике для последовательности, представляющей интерес, с применением меток последовательности для фрагментов в первом диапазоне размера и из второй t-статистики для последовательности, представляющей интерес, с применением меток последовательности для фрагментов во втором диапазоне размера.

Согласно некоторым вариантам реализации отношение правдоподобия вычисляют в виде первого правдоподобия того, что исследуемый образец представляет собой анеуплоидный образец, относительно второго правдоподобия того, что исследуемый образец представляет собой эуплоидный образец.

Согласно некоторым вариантам реализации отношение правдоподобия вычисляют по одному или более значениям фракции плода в дополнение к t-статистике и информации относительно размера фрагментов бесклеточной нуклеиновой кислоты.

Согласно некоторым вариантам реализации одно или более значений фракции плода включают значение фракции плода, вычисленное с применением информации относительно размеров фрагментов бесклеточной нуклеиновой кислоты. Согласно некоторым вариантам реализации значение фракции плода вычисляют посредством: получения распределения частоты размера фрагментов; и применения распределения частоты в модели, устанавливающей взаимосвязь между фракцией плода и частотой размера фрагмента, с получением значения фракции плода. Согласно некоторым вариантам реализации модель, устанавливающая взаимосвязь между фракцией плода и частотой размера фрагмента, включает обычную линейную модель, которая содержит множество параметров и коэффициентов для множества блоков.

Согласно некоторым вариантам реализации одно или более значений фракции плода включают значение фракции плода, вычисленное с применением информации о перекрытии для блоков референсного генома. Согласно некоторым вариантам реализации значение фракции плода вычисляют посредством применения значения перекрытия множества блоков в модели, устанавливающей взаимосвязь между фракцией плода и перекрытием блока, с получением значения фракции плода. Согласно некоторым вариантам реализации модель, устанавливающая взаимосвязь между фракцией плода и перекрытием блока, включает обычную линейную модель, которая содержит множество параметров и коэффициентов для множества блоков. Согласно некоторым вариантам реализации множество блоков характеризуется высокой корреляцией между фракцией плода и перекрытием в обучающих образцах.

Согласно некоторым вариантам реализации одно или более значений фракции плода включают значение фракции плода, вычисленное с применением частот множества 8-меров, обнаруженных в ридах. Согласно некоторым вариантам реализации значение фракции плода вычисляют посредством: применения частот множества 8-меров в модели, устанавливающей взаимосвязь между фракцией плода и частотой 8-меров, с получением значения фракции плода. Согласно некоторым вариантам реализации модель, устанавливающая взаимосвязь между фракцией плода и частотой 8-мера, включает обычную линейную модель, которая содержит множество параметров и коэффициентов для множества 8-меров. Согласно некоторым вариантам реализации множество 8-меров характеризуется высокой корреляцией между фракцией плода и частотой 8-меров.

Согласно некоторым вариантам реализации одно или более значений фракции плода включают значение фракции плода, вычисленное с применением информации о перекрытии для блоков половой хромосомы.

Согласно некоторым вариантам реализации отношение правдоподобия вычисляют по фракции пло-

да, t-статистике коротких фрагментов и t-статистике всех фрагментов, причем короткие фрагменты представляют собой фрагменты бесклеточной нуклеиновой кислоты в первом диапазоне размера, меньшие, чем размер-критерий, и все фрагменты представляют собой фрагменты бесклеточной нуклеиновой кислоты, включая короткие фрагменты и фрагменты, более длинные, чем размер-критерий. Согласно некоторым вариантам реализации отношение правдоподобия (ОВ) вычисляют как

$$\text{ОВ} = \frac{\sum_{ff_{\text{суммарн.}}} q(ff_{\text{суммарн.}}) * p_1(T_{\text{коротк.}}, T_{\text{всех}} | ff_{\text{выч.}})}{p_0(T_{\text{коротк.}}, T_{\text{всех}})},$$

где p_1 представляет собой правдоподобие того, что данные получены из многомерного нормального распределения, представляющего 3-копийную или 1-копийную модель, p_0 представляет собой правдоподобие того, что данные получены из многомерного нормального распределения, представляющего 2-копийную модель, $T_{\text{коротк}}$, $T_{\text{всех}}$ представляют собой T-показатели, вычисленные по перекрытию хромосом, полученному из коротких фрагментов и всех фрагментов, и $q(ff_{\text{суммарн.}})$ представляет собой плотность распределения фракции плода.

Согласно некоторым вариантам реализации отношение правдоподобия вычисляют по одному или более значениям фракции плода в дополнение к t-статистике и информации относительно размера фрагментов бесклеточной нуклеиновой кислоты.

Согласно некоторым вариантам реализации отношение правдоподобия вычисляют для моносомии X, трисомии X, трисомии 13, трисомии 18 или трисомии 21.

Согласно некоторым вариантам реализации нормирование количества меток последовательности включает: нормирование с учетом содержания GC в образце, нормирование с учетом глобального волнового профиля вариации обучающего множества и/или нормирование с учетом одной или более компонент, полученных из анализа главных компонент.

Согласно некоторым вариантам реализации последовательность, представляющая интерес, представляет собой хромосому человека, которая выбрана из группы, состоящей из хромосомы 13, хромосомы 18, хромосомы 21, хромосомы X и хромосомы Y.

Согласно некоторым вариантам реализации референсная область представляет собой все устойчивые хромосомы, устойчивые хромосомы, не содержащие последовательность, представляющую интерес, по меньшей мере хромосому за пределами последовательности, представляющей интерес, и/или подмножество хромосом, выбранных из устойчивых хромосом. Согласно некоторым вариантам реализации референсная область содержит устойчивые хромосомы, которые были определены для обеспечения наилучшей способности обнаружения сигнала для множества обучающих образцов.

Согласно некоторым вариантам реализации способ также включает вычисление значений параметра размера для блоков для каждого блока посредством: (i) определения значения параметра размера на основании размеров фрагментов бесклеточной нуклеиновой кислоты в блоке и (ii) нормирования значения параметра размера посредством вычисления межблоковых вариаций, вызванных факторами, отличными от вариации числа копий. Способ также включает определение t-статистики на основании размера для последовательности, представляющей интерес, с применением значений параметра размера блоков в последовательности, представляющей интерес, и значений параметра размера блоков в референсной области для последовательности, представляющей интерес. Согласно некоторым вариантам реализации отношение правдоподобия (f) вычисляют по t-статистике и t-статистике на основании размера. Согласно некоторым вариантам реализации отношение правдоподобия (f) вычисляют по t-статистике на основании размера и фракции плода.

Согласно некоторым вариантам реализации способ также включает сравнение отношения правдоподобия с критерием решения для определения вариации числа копий в последовательности, представляющей интерес. Согласно некоторым вариантам реализации отношения правдоподобия преобразуют в логарифмическое отношение правдоподобия перед сравнением с критерием решения. Согласно некоторым вариантам реализации критерий решения получают посредством применения различных критериев в отношении обучающего множества обучающих образцов и выбора критерия, который обеспечивает заданную чувствительность и заданную селективность.

Согласно некоторым вариантам реализации способ также включает получение множества отношений правдоподобия и применение множества отношений правдоподобия в дереве решений для определения случая плоидности для образца.

Согласно некоторым вариантам реализации способ также включает получение множества отношений правдоподобия и одного или более значений перекрытия последовательности, представляющей интерес, и применение множества отношений правдоподобия и одного или более значений перекрытия последовательности, представляющей интерес, в дереве решений для определения случая плоидности для образца.

Другой аспект настоящего изобретения относится к способу определения вариации числа копий (ВЧК) последовательности нуклеиновой кислоты, представляющей интерес, в исследуемом образце, содержащем фрагменты бесклеточной нуклеиновой кислоты, полученные из двух или более геномов. Способ включает: (a) прием ридов последовательности, полученных в результате секвенирования фрагмен-

тов бесклеточной нуклеиновой кислоты в исследуемом образце; (b) выравнивание ридов последовательности фрагментов бесклеточной нуклеиновой кислоты или выравнивание фрагментов, содержащих риды последовательности, с блоками референсного генома, содержащего последовательность, представляющую интерес, с получением, таким образом, меток исследуемой последовательности, причем референсный геном разделен на множество блоков; (c) вычисление перекрытий меток последовательности для блоков референсного генома для каждого блока посредством: (i) определения количества меток последовательности, которые выравниваются с блоком, и (ii) нормирования количества меток последовательности, которые выравниваются с блоком, посредством вычисления межблоковых вариаций, вызванных факторами, отличными от вариации числа копий. Способ также включает: (d) определение t-статистики для последовательности, представляющей интерес, с применением перекрытий блоков в последовательности, представляющей интерес, и перекрытий блоков в референсной области для последовательности, представляющей интерес; (e) оценку одного или более значений фракции плода фрагментов бесклеточной нуклеиновой кислоты в исследуемом образце; и (f) определение вариации числа копий в последовательности, представляющей интерес, с применением t-статистики и одного или более значений фракции плода.

Согласно некоторым вариантам реализации этап (f) включает вычисление отношения правдоподобия из t-статистики и одного или более значений фракции плода. Согласно некоторым вариантам реализации отношение правдоподобия вычисляют для моносомии X, трисомии X, трисомии 13, трисомии 18 или трисомии 21.

Согласно некоторым вариантам реализации нормирование количества меток последовательности включает: нормирование с учетом содержания GC в образце, нормирование с учетом глобального волнового профиля вариации обучающего множества и/или нормирование с учетом одной или более компонент, полученных из анализа главных компонент.

Согласно некоторым вариантам реализации последовательность, представляющая интерес, представляет собой хромосому человека, которая выбрана из группы, состоящей из хромосомы 13, хромосомы 18, хромосомы 21, хромосомы X и хромосомы Y.

Следующий аспект настоящего изобретения относится к способу определения вариации числа копий (ВЧК) последовательности нуклеиновой кислоты, представляющей интерес, в исследуемом образце, содержащем фрагменты бесклеточной нуклеиновой кислоты, полученные из двух или более геномов. Способ включает: (a) прием ридов последовательности, полученных в результате секвенирования фрагментов бесклеточной нуклеиновой кислоты в исследуемом образце; (b) выравнивание ридов последовательности фрагментов бесклеточной нуклеиновой кислоты или выравнивание фрагментов, содержащих риды последовательности, с блоками референсного генома, содержащего последовательность, представляющую интерес, с получением, таким образом, меток исследуемой последовательности, причем референсный геном разделен на множество блоков; (c) определение размера фрагментов для фрагментов бесклеточной нуклеиновой кислоты, существующих в исследуемом образце; (d) вычисление перекрытий меток последовательности для блоков референсного генома с применением меток последовательности для фрагментов бесклеточной нуклеиновой кислоты, размеры которых относятся к первому домену размеров; (e) вычисление перекрытий меток последовательности для блоков референсного генома с применением меток последовательности для фрагментов бесклеточной нуклеиновой кислоты, размеры которых относятся ко второму домену размеров, причем второй домен размеров отличается от первого домена размеров; (f) вычисление характеристик размера для блоков референсного генома с применением размеров фрагментов, определенных на этапе (c); и (g) определение вариации числа копий в последовательности, представляющей интерес, с применением перекрытий, вычисленных на этапах (d) и (e), и характеристик размера, вычисленных на этапе (f).

Согласно некоторым вариантам реализации первый домен размеров включает фрагменты бесклеточной нуклеиновой кислоты по существу всех размеров в образце, и второй домен размеров включает только фрагменты бесклеточной нуклеиновой кислоты, меньшие, чем заданный размер. Согласно некоторым вариантам реализации второй домен размеров включает только фрагменты бесклеточной нуклеиновой кислоты, меньшие, чем приблизительно 150 п.о.

Согласно некоторым вариантам реализации последовательность, представляющая интерес, представляет собой хромосому человека, которая выбрана из группы, состоящей из хромосомы 13, хромосомы 18, хромосомы 21, хромосомы X и хромосомы Y.

Согласно некоторым вариантам реализации этап (g) включает вычисление t-статистики для последовательности, представляющей интерес, с применением перекрытий блоков в последовательности, представляющей интерес, вычисленных на этапе (d) и/или (e). Согласно некоторым вариантам реализации вычисление t-статистики для последовательности, представляющей интерес, включает применение перекрытий блоков в последовательности, представляющей интерес, и перекрытий блоков в референсной области для последовательности, представляющей интерес.

Согласно некоторым вариантам реализации этап (g) включает вычисление t-статистики для последовательности, представляющей интерес, с применением характеристик размера блоков в последовательности, представляющей интерес, вычисленных на этапе (f). Согласно некоторым вариантам реализа-

ции вычисление t-статистики для последовательности, представляющей интерес, включает применение характеристик размера блоков в последовательности, представляющей интерес, и характеристик размера блоков в референсной области для последовательности, представляющей интерес.

Согласно некоторым вариантам реализации характеристика размера для блока включает отношение фрагментов размера, меньших, чем заданное значение, к общему количеству фрагментов в блоке.

Согласно некоторым вариантам реализации этап (g) включает вычисление отношения правдоподобия из t-статистики.

Согласно некоторым вариантам реализации этап (g) включает вычисление отношения правдоподобия из первой t-статистики для последовательности, представляющей интерес, с применением перекрытий, вычисленных на этапе (d), и второй t-статистики для последовательности, представляющей интерес, с применением перекрытий, вычисленных на этапе (e).

Согласно некоторым вариантам реализации этап (g) включает вычисление отношения правдоподобия из первой t-статистики для последовательности, представляющей интерес, с применением перекрытий, вычисленных на этапе (d), второй t-статистики для последовательности, представляющей интерес, с применением перекрытий, вычисленных на этапе (e), и третьей t-статистики для последовательности, представляющей интерес, с применением характеристик размера, вычисленных на этапе (f).

Согласно некоторым вариантам реализации отношение правдоподобия вычисляют по одному или более значениям фракции плода в дополнение по меньшей мере к первой и второй t-статистике. Согласно некоторым вариантам реализации способ также включает вычисление одного или более значений фракции плода с применением информации относительно размеров фрагментов бесклеточной нуклеиновой кислоты.

Согласно некоторым вариантам реализации способ также включает вычисление одного или более значений фракции плода с применением информации о перекрытии для блоков референсного генома. Согласно некоторым вариантам реализации одно или более значений фракции плода включают значение фракции плода, вычисленное с применением информации о перекрытии для блоков половой хромосомы. Согласно некоторым вариантам реализации отношение правдоподобия вычисляют для моносомии X, трисомии X, трисомии 13, трисомии 18 или трисомии 21.

Согласно некоторым вариантам реализации этап (d) и/или (e) включает: (i) определение количества меток последовательности, которые выравниваются с блоком, и (ii) нормирование количества меток последовательности, которые выравниваются с блоком, посредством вычисления межблоковых вариаций, вызванных факторами, отличными от вариации числа копий. Согласно некоторым вариантам реализации нормирование количества меток последовательности включает: нормирование с учетом содержания GC в образце, нормирование с учетом глобального волнового профиля вариации обучающего множества и/или нормирование с учетом одной или более компонент, полученных из анализа главных компонент.

Согласно некоторым вариантам реализации этап (f) включает вычисление значений параметра размера для блоков для каждого блока посредством: (i) определения значения параметра размера на основании размеров фрагментов бесклеточной нуклеиновой кислоты в блоке, и (ii) нормирования значения параметра размера посредством вычисления межблоковых вариаций, вызванных факторами, отличными от вариации числа копий.

Другой аспект настоящего изобретения относится к системе для оценки числа копий последовательности нуклеиновой кислоты, представляющей интерес, в исследуемом образце, причем указанная система содержит: секвенатор для приема фрагментов нуклеиновой кислоты из исследуемого образца и обеспечения информации о последовательности нуклеиновой кислоты исследуемого образца; процессор; и один или более машиночитаемых носителей для хранения информации, на которых хранятся инструкции для выполнения на указанном процессоре. Инструкции включают инструкции для: (a) приема ридов последовательности, полученных в результате секвенирования фрагментов бесклеточной нуклеиновой кислоты в исследуемом образце; (b) выравнивания ридов последовательности фрагментов бесклеточной нуклеиновой кислоты или выравнивания фрагментов, содержащих риды последовательности, с блоками референсного генома, содержащего последовательность, представляющую интерес, с получением, таким образом, меток исследуемой последовательности, причем референсный геном разделен на множество блоков; (c) определения размеров фрагмента по меньшей мере некоторых фрагментов бесклеточной нуклеиновой кислоты, присутствующих в исследуемом образце; и (d) вычисления перекрытий меток последовательности для блоков референсного генома для каждого блока посредством: (i) определения количества меток последовательности, которые выравниваются с блоком, и (ii) нормирования количества меток последовательности, которые выравниваются с блоком, посредством вычисления межблоковых вариаций, вызванных факторами, отличными от вариации числа копий. Способ также включает: (e) определение t-статистики для последовательности, представляющей интерес, с применением перекрытий блоков в последовательности, представляющей интерес, и перекрытий блоков в референсной области для последовательности, представляющей интерес; и (f) определение вариации числа копий в последовательности, представляющей интерес, с применением отношения правдоподобия, вычисленного по t-статистике, и информации относительно размера фрагментов бесклеточной нуклеиновой кислоты.

Согласно некоторым вариантам реализации система проектирована для осуществления любого из

способов, описанных выше.

Дополнительный аспект настоящего изобретения относится к компьютерному программному продукту, который содержит один или более машиночитаемых носителей, предназначенных для долговременного хранения информации, на которых хранятся выполняемые компьютером инструкции, при выполнении которых одним или более процессорами компьютерной системы компьютерная система реализует любой из способов, описанных выше.

Несмотря на то что примеры в настоящем документе относятся к людям, и описание преимущественно направлено на проблемы человека, концепции, описанные в настоящем документе, применимы к геномам любого растения или животного. Данные и другие объекты и свойства настоящего изобретения станут более очевидными на основании следующего описания и прилагаемой формулы изобретения или могут быть выяснены при реализации настоящего изобретения на практике, как представлено ниже по тексту.

Включение посредством ссылки

Все патенты, заявки на патент и другие публикации, включая все последовательности, раскрытые в данных источниках, упомянутых в настоящем документе, явным образом включены в настоящий документ посредством ссылки в той же степени, как если бы каждая отдельная публикация, патент или заявка на патент были конкретно и индивидуально указаны как включенные посредством ссылки. Все процитированные документы в соответствующей части полностью включены в настоящий документ посредством ссылки для целей, определяемых контекстом цитирования данных источников в настоящем документе. Однако цитирование любого документа не следует толковать как признание того, что данный документ составляет предшествующий уровень техники по отношению к настоящему изобретению.

Краткое описание фигур

Фиг. 1 представляет собой структурную схему способа 100 для определения присутствия или отсутствия вариации числа копий в исследуемом образце, содержащем смесь нуклеиновых кислот.

Фиг. 2A тематически иллюстрирует, как секвенирование спаренных концов можно применять для определения как размера фрагмента, так и перекрытия последовательности.

На фиг. 2B представлена структурная схема процесса для применения перекрытия на основании размера с целью определения вариации числа копий последовательности нуклеиновой кислоты, представляющей интерес, в исследуемом образце.

На фиг. 2C представлена структурная схема процесса для определения параметра размера фрагмента для последовательности нуклеиновой кислоты, представляющей интерес, которую применяли для оценки числа копий.

На фиг. 2D представлена блок-схема двух перекрывающихся проходов рабочего процесса.

На фиг. 2E представлена блок-схема трехпроходного процесса для оценки числа копий.

На фиг. 2F представлены варианты реализации, в которых применяют t-статистику для анализа числа копий с целью улучшения точности анализа.

На фиг. 2G представлен пример процесса для определения фракции плода на основании информации о перекрытии согласно некоторым вариантам реализации настоящего изобретения.

На фиг. 2H представлен процесс для определения фракции плода на основании информации о распределении размера согласно некоторым вариантам реализации.

На фиг. 2I представлен пример процесса для определения фракции плода на основании информации о частоте 8-меров согласно некоторым вариантам реализации настоящего изобретения.

На фиг. 2J представлен рабочий процесс для обработки информации о ридах последовательности, который можно применять для получения оценок фракции плода.

На фиг. 3A представлена структурная схема примера процесса для снижения шума в данных последовательности из исследуемого образца.

На фиг. 3B-3K представлены анализы данных, полученных на различных этапах процесса, изображенного на фиг. 3A.

На фиг. 4A представлена блок-схема процесса получения маски последовательности для снижения шума в данных последовательности.

Фиг. 4B демонстрирует, что показатель MapQ характеризуется устойчивой монотонной корреляцией с KB (коэффициентом вариации) нормированных количеств перекрытия.

Фиг. 5 представляет собой блок-диаграмму дисперсной системы для процессинга (обработки) исследуемого образца и, в конечном счете, постановки диагноза.

Фиг. 6 схематично иллюстрирует, как различные операции при процессинге исследуемых образцов можно сгруппировать для манипуляции различными элементами системы.

Фиг. 7A и 7B демонстрируют электрофореграммы библиотеки секвенирования сцДНК, полученной согласно сокращенному протоколу, описанному в примере 1a (фиг. 7A), и протоколу, описанному в примере 1b (фиг. 7B).

На фиг. 8 представлен общий рабочий процесс и временные рамки для новой версии НИПТ (неинвазивного пренатального тестирования) по сравнению со стандартным лабораторным рабочим процессом.

На фиг. 9 представлен выход из библиотеки секвенирования как функция экстрагированной сцДНК на входе, которая свидетельствует об устойчивой линейной корреляции концентрации библиотеки и концентрации на входе с высокой эффективностью преобразования.

На фиг. 10 представлено распределение размера фрагментов сцДНК, измеренного в 324 образцах от беременностей плодом мужского пола.

На фиг. 11 представлена относительная фракция плода по общему подсчитанному значению картированных ридов спаренных концов по сравнению с числом ридов спаренных концов, которые составляют менее 150 п.о.

На фиг. 12 представлена объединенная t-статистика показателя анеуплоидии для обнаружения образцов трисомии 21 для (A) подсчитанных значений всех фрагментов; (B) подсчитанных значений исключительно коротких фрагментов (<150 п.о.); (C) фракции коротких фрагментов (подсчитанные значения от 80 до 150 п.о./подсчитанные значения <250 п.о.); (D) объединенной t-статистики от (B) и (C); и (E) результатов для тех же образцов, полученных с применением лабораторного процесса CLIA (Chemi Luminescent Immuno Assay, иммунохемилюминесцентный анализ) Illumina, Рэдвуд-Сити, со средним значением 16 М подсчитанных значений/образец.

На фиг. 13 представлены фракции плода, оцененные в выбранных блоках, по сравнению с таковыми, измеренными с нормированными значениями хромосом (референс, эталон), для X-хромосомы. Множество 1 применяли для калибровки значения фракции плода, и независимое множество 2 - для исследования корреляции.

Подробное описание изобретения

Определения.

Если не указано обратное, реализация на практике способа и системы, раскрытых в настоящем документе, включает общепринятые методики и аппараты, обычно применяемые в молекулярной биологии, микробиологии, очистке белка, белковой инженерии, секвенировании белка и ДНК и в области рекомбинантной ДНК, которые находятся в пределах компетенции в данной области техники. Такие методики и аппараты известны специалистам в данной области техники и описаны в многочисленных руководствах и справочных изданиях (См. например, Sambrook et al., "Molecular Cloning: A Laboratory Manual," Third Edition (Cold Spring Harbor), [2001]); и Ausubel et al., "Current Protocols in Molecular Biology" [1987]).

Числовые диапазоны включают количества, определяющие диапазон. Предполагается, что каждое максимальное количественное ограничение, приведенное на всем протяжении данной спецификации, включает каждое меньшее количественное ограничение, как если бы такие меньшие количественные ограничения были явным образом указаны в настоящем документе. Каждое минимальное количественное ограничение, приведенное на всем протяжении данной спецификации, включает каждое более высокое количественное ограничение, как если бы такие более высокие количественные ограничения были явным образом указаны в настоящем документе. Каждый числовой диапазон, приведенный на всем протяжении данной спецификации, включает каждый более узкий числовой диапазон, который попадает в такой более широкий числовой диапазон, как если бы такие более узкие числовые диапазоны были явным образом указаны в настоящем документе.

Заголовки, приведенные в настоящем документе, не предназначены для ограничения настоящего изобретения.

Если не указано обратное, в настоящем документе все технические и научные термины, используемые в настоящем документе, имеют то же значение, которое общепринято понимает средний специалист в данной области техники. Различные научные словари, которые включают термины, приведенные в настоящем документе, хорошо известны и доступны специалистам в данной области техники. Несмотря на то что любые способы и материалы, аналогичные или эквивалентные таковым, описанным в настоящем документе, находят применение при реализации на практике или исследовании вариантов реализации, раскрытых в настоящем документе, описаны некоторые способы и материалы.

Термины, определения которых приведены непосредственно ниже, более полно описаны посредством ссылки на описание в целом. Следует понимать, что настоящее изобретение не ограничено конкретными описанными методологией, протоколами и реактивами, поскольку все они могут варьировать в зависимости от контекста, в котором их применяет специалист в данной области техники. В настоящем документе термины в единственном числе включают упоминания объектов во множественном числе, если в контексте однозначно не указано обратное.

Если не указано обратное, нуклеиновые кислоты представлены слева направо в направлении от 5'- к 3'-концу, а последовательности аминокислот представлены слева направо в ориентации от амино- к карбоксиконцу, соответственно.

Термин "параметр" в настоящем документе представляет физическое свойство, значение или другую характеристику, которая оказывает влияние на соответствующее состояние, такое как вариация числа копий. В некоторых случаях термин "параметр" используют применительно к переменной, которая влияет на математическую зависимость или модель на выходе, причем данная переменная может являться независимой переменной (т.е. вводимой в модель) или промежуточной переменной, основанной на

одной или более независимых переменных. В зависимости от объема модели данные на выходе одной модели могут стать данными на входе другой модели, посредством этот став параметром для другой модели.

Термин "параметр размера фрагмента" означает параметр, который относится к размеру или длине фрагмента или совокупности фрагментов, таких как фрагменты нуклеиновой кислоты; например, фрагменты сцДНК, полученные из физиологической жидкости. В настоящем документе параметр "смещен в сторону размера фрагмента или диапазона размера", когда: 1) параметр благоприятно взвешивается по размеру фрагмента или диапазону размера, например вычисление имеет больший вес, когда связан с фрагментами размера или диапазона размера, чем для других размеров или диапазонов; или 2) параметр получен из значения, которое благоприятно взвешивается по размеру фрагмента или диапазону размера, например, соотношение получено из подсчитанного значения с большим весом, когда связан с фрагментами размера или диапазона размера. Размер фрагмента или диапазон размера может являться характеристикой генома или его части, когда геном образует фрагменты нуклеиновой кислоты, обогащенные или содержащие более высокую концентрацию размера или диапазона размера, по сравнению с фрагментами нуклеиновой кислоты из другого генома или другой части того же генома.

Термин "взвешивание" означает модификацию количества, такого как параметр или переменная, с применением одного или более значений или функций, которые считаются "весом". Согласно определенным вариантам реализации параметр или переменную умножают на вес. Согласно другим вариантам реализации параметр или переменную модифицируют экспоненциально. Согласно некоторым вариантам реализации функция может представлять собой линейную или нелинейную функцию. Примеры применимых нелинейных функций включают, без ограничения, ступенчатые функции Хевисайда, функции вагона, ступенчатые функции или сигмоидальные функции. Взвешивание исходного параметра или переменной может системно увеличить или уменьшить значение взвешенной переменной. Согласно различным вариантам реализации взвешивание может привести к получению положительных, неотрицательных или отрицательных значений.

Термин "вариация числа копий" в настоящем документе означает вариацию количества копий последовательности нуклеиновой кислоты, присутствующей в исследуемом образце, по сравнению с числом копий последовательности нуклеиновой кислоты, присутствующей в референсном образце. Согласно определенным вариантам реализации длина последовательности нуклеиновой кислоты составляет 1 т.о. (тысячу оснований) или более. В некоторых случаях последовательность нуклеиновой кислоты представляет собой целую хромосому или значительную ее часть. "Вариант числа копий" означает последовательность нуклеиновой кислоты, в которой были обнаружены различия числа копий посредством сравнения последовательности нуклеиновой кислоты, представляющей интерес, в исследуемом образце с ожидаемым уровнем последовательности нуклеиновой кислоты, представляющей интерес. Например, уровень последовательности нуклеиновой кислоты, представляющей интерес, в исследуемом образце сравнивают с присутствующим в квалификационном образце. Варианты/вариации числа копий включают делеции, в том числе микроделеции, инсерции, в том числе микроинсерции, дупликации, умножения и транслокации. ВЧК включает анеуплоидии хромосом и частичные анеуплоидии.

Термин "анеуплоидия" в настоящем документе означает дисбаланс генетического материала, вызванный утратой или добавлением целой хромосомы или части хромосомы.

Термины "анеуплоидия хромосомы" и "полная анеуплоидия хромосомы" в настоящем документе означают дисбаланс генетического материала, вызванный утратой или добавлением целой хромосомы, и включают анеуплоидию зародышевой линии и мозаичную анеуплоидию.

Термины "частичная анеуплоидия" и "частичная анеуплоидия хромосомы" в настоящем документе означают дисбаланс генетического материала, вызванный утратой или добавлением части хромосомы, например, частичную моносомию и частичную трисомию, и включают дисбаланс, который является следствием транслокации, делеции и инсерции.

Термин "множество" означает более одного элемента. Например, данный термин в настоящем документе применяют к количеству молекул нуклеиновой кислоты или меток последовательности, достаточному для идентификации значительных различий вариаций числа копий в исследуемых образцах и квалификационных образцах, с применением способов, раскрытых в настоящем документе. Согласно некоторым вариантам реализации для каждого исследуемого образца получают по меньшей мере приблизительно 3×10^6 меток последовательности длиной от приблизительно 20 до 40 п.о. Согласно некоторым вариантам реализации каждый исследуемый образец обеспечивает данные для по меньшей мере приблизительно 5×10^6, 8×10^6, 10×10^6, 15×10^6, 20×10^6, 30×10^6, 40×10^6 или 50×10^6 меток последовательности, причем каждая метка последовательности содержит приблизительно от 20 до 40 п.о.

Термин "риды спаренных концов" означает риды из секвенирования спаренных концов, которые получают один рид с каждого конца фрагмента нуклеиновой кислоты. Секвенирование спаренных концов может включать фрагментацию цепей полинуклеотидов на короткие последовательности, называемые вставками. Фрагментация является необязательной или нецелесообразной для относительно коротких полинуклеотидов, таких как молекулы бесклеточной ДНК.

Термины "полинуклеотид", "нуклеиновая кислота" и "молекулы нуклеиновой кислоты" используются взаимозаменяемо и означают ковалентным способом связанную последовательность нуклеотидов (т.е. рибонуклеотидов для РНК и дезоксирибонуклеотидов для ДНК), в которой 3'-положение пентозы одного нуклеотида соединено с помощью фосфодиэфирной группы с 5'-положением пентозы следующего нуклеотида. Нуклеотиды включают последовательности любой формы нуклеиновой кислоты, включая, без ограничения, молекулы РНК и ДНК, такие как молекулы сцДНК. Термин "полинуклеотид" включает, без ограничения, одно- и двухцепочечный полинуклеотид.

Термин "исследуемый образец" в настоящем документе означает образец, как правило, полученный из биологической жидкости, клетки, ткани, органа или организма, содержащий нуклеиновую кислоту или смесь нуклеиновых кислот, которая содержит по меньшей мере одну последовательность нуклеиновой кислоты, скрининг которой проводят в отношении вариации числа копий. Согласно определенным вариантам реализации образец содержит по меньшей мере одну последовательность нуклеиновой кислоты, число копий которой, как ожидается, подверглось вариации. Такие образцы включают, без ограничения, мокроту/жидкость ротовой полости, амниотическую жидкость, кровь, фракцию крови или образцы тонкоигольной биопсии (например, хирургической биопсии, тонкоигольной биопсии и т.д.), мочу, перитонеальную жидкость, плевральную жидкость и т.п. Несмотря на то что образец часто отбирают от субъекта-человека (например, пациента), анализы можно применять для оценки вариаций числа копий (ВЧК) в образцах от любого млекопитающего, включая, без ограничения, собак, кошек, лошадей, коз, овец, крупный рогатый скот, свиней и т.д. Образец можно применять непосредственно в том виде, в котором он был получен из биологического источника, или после предварительной обработки для модификации характера образца. Например, такая предварительная обработка может включать получение плазмы из крови, разведение вязких жидкостей и т.д. Способы предварительной обработки могут также включать, без ограничения, фильтрацию, преципитацию, разведение, дистилляцию, перемешивание, центрифугирование, замораживание, лиофилизацию, концентрирование, амплификацию, фрагментацию нуклеиновой кислоты, инактивацию интерферирующих соединений, добавление реактивов, лизис и т.д. Если такие способы предварительной обработки применяют в отношении образца, такие способы предварительной обработки, как правило, являются таковыми, при которых нуклеиновая кислота или кислоты, представляющие интерес, остаются в исследуемом образце, иногда в концентрации, пропорциональной таковой в необработанном исследуемом образце (например, а именно, в образце, который не подвергали какому-либо из таких способов предварительной обработки). Такие "обработанные" или "процессированные" образцы все еще считают биологическими "исследуемыми" образцами применительно к способам, описанным в настоящем документе.

Термин "квалификационный образец" или "непораженный образец" в настоящем документе означает образец, содержащий смесь нуклеиновых кислот, которые присутствуют в известном числе копий, с которыми будут сравнивать нуклеиновые кислоты в исследуемом образце, и представляет собой образец, который является нормальным, т.е. не анеуплоидным, в отношении последовательности нуклеиновой кислоты, представляющей интерес. Согласно некоторым вариантам реализации квалификационные образцы применяют в качестве непораженных обучающих образцов обучающего множества для получения масок последовательности или профилей последовательности. Согласно определенным вариантам реализации квалификационные образцы применяют для идентификации одной или более нормирующих хромосом или сегментов для рассматриваемой хромосомы. Например, квалификационные образцы можно применять для идентификации нормирующей хромосомы для хромосомы 21. В таком случае квалификационный образец представляет собой образец, отличный от образца трисомии 21. Другой пример включает применение в качестве квалификационных образцов для хромосомы X исключительно образцов женского пола. Квалификационные образцы можно также применять для других целей, таких как определение порогов для принятия решения о пораженных образцах, идентификация порогов для определения областей масок на референсной последовательности, определение ожидаемого перекрытия количеств для различных областей генома и т.п.

Термин "обучающее множество" в настоящем документе означает множество обучающих образцов, которое может содержать пораженные и/или непораженные образцы и которое применяют для разработки модели для анализа исследуемых образцов. Согласно некоторым вариантам реализации обучающее множество содержит непораженные образцы. Согласно данным вариантам реализации пороги для определения ВЧК устанавливают с применением обучающих множеств образцов, которые являются непораженными в отношении вариации числа копий, представляющей интерес. Непораженные образцы в обучающем множестве можно применять в качестве квалификационных образцов для идентификации нормирующих последовательностей, например нормирующих хромосом, и дозы хромосом непораженных образцов применяют для установления порогов для каждой из последовательностей, например, хромосом, представляющих интерес. Согласно некоторым вариантам реализации обучающее множество содержит пораженные образцы. Пораженные образцы в обучающем множестве можно применять для подтверждения того, что пораженные исследуемые образцы можно с легкостью отличить от непораженных образцов.

Обучающее множество представляет собой также статистический образец в популяции, представ-

ляющей интерес, причем статистический образец не стоит путать с биологическим образцом. Статистический образец часто содержит образцы от нескольких индивидуумов, и данные от этих индивидуумов применяют для определения одного или более количественных значений, представляющих интерес, обобщаемых на популяцию. Статистический образец представляет собой подмножество индивидуумов в популяции, представляющей интерес. Индивидуумы могут представлять собой лиц, животных, ткани, клетки, другие биологические образцы (т.е. статистический образец может включать несколько биологических образцов) и других индивидуальных субъектов, обеспечивающих данные наблюдений для статистического анализа.

Обычно обучающее множество применяют в сочетании с валидационным множеством. Термин "валидационное множество" применяют для обозначения множества индивидуумов в статистическом образце, причем данные от этих индивидуумов применяют для валидации или оценки количественных значений, представляющих интерес, определенных с применением обучающего множества. Согласно некоторым вариантам реализации, например, обучающее множество обеспечивает данные для вычисления маски для референсной последовательности, тогда как валидационное множество обеспечивает данные для оценки правильности или эффективности маски.

"Оценку числа копий" используют в настоящем документе применительно к статистической оценке статуса генетической последовательности в отношении числа копий последовательности. Например, согласно некоторым вариантам реализации оценка включает определение присутствия или отсутствия генетической последовательности. Согласно некоторым вариантам реализации оценка включает определение частичной или полной анеуплоидии генетической последовательности. Согласно другим вариантам реализации оценка включает установление отличий между двумя или более образцами на основании числа копий генетической последовательности. Согласно некоторым вариантам реализации оценка включает статистические анализы, например, нормирование и сравнение, на основании числа копий генетической последовательности.

Термин "квалификационная нуклеиновая кислота" применяют взаимозаменяемо с "квалификационной последовательностью", которая представляет собой последовательность, с которой сравнивают количество последовательности или нуклеиновой кислоты, представляющей интерес. Квалификационная последовательность представляет собой таковую, присутствующую в биологическом образце, предпочтительно, с известной представленностью, т.е. количество квалификационной последовательности известно. Как правило, квалификационная последовательность представляет собой последовательность, присутствующую в "квалификационном образце". "Квалификационная последовательность, представляющая интерес", представляет собой квалификационную последовательность, количество которой в квалификационном образце известно, и представляет собой последовательность, которая связана с отличием последовательности, представляющей интерес, между контрольным субъектом и индивидуумом с медицинским состоянием.

Термин "последовательность, представляющая интерес", или "последовательность нуклеиновой кислоты, представляющая интерес", в настоящем документе означает последовательность нуклеиновой кислоты, которая связана с отличием в представленности последовательности между здоровыми и заболевшими индивидуумами. Последовательность, представляющая интерес, может представлять собой последовательность на хромосоме, которая при заболевании или генетическом состоянии представлена в искаженном виде, т.е. чрезмерно или недостаточно представлена. Последовательность, представляющая интерес, может представлять собой часть хромосомы, т.е. сегмент хромосомы, или целую хромосому. Например, последовательность, представляющая интерес, может представлять собой хромосому, которая чрезмерно представлена при состоянии анеуплоидии, или ген, кодирующий супрессор опухоли, который недостаточно представлен при раке. Последовательности, представляющие интерес, включают последовательности, которые чрезмерно или недостаточно представлены в общей популяции или субпопуляции клеток субъекта. "Квалификационная последовательность, представляющая интерес", представляет собой последовательность, представляющую интерес, в квалификационном образце. "Исследуемая последовательность, представляющая интерес", представляет собой последовательность, представляющую интерес, в исследуемом образце.

Термин "нормирующая последовательность" в настоящем документе означает последовательность, которую применяют для нормирования количества меток последовательности, картированных на последовательности, представляющей интерес, связанной с нормирующей последовательностью. Согласно некоторым вариантам реализации нормирующая последовательность содержит устойчивую хромосому. "Устойчивая хромосома" представляет собой хромосому, которая с низкой долей вероятности является анеуплоидной. В некоторых случаях, относящихся к хромосоме человека, устойчивая хромосома представляет собой любую хромосому, отличную от X-хромосомы, Y-хромосомы, хромосомы 13, хромосомы 18 и хромосомы 21. Согласно некоторым вариантам реализации нормирующая последовательность демонстрирует вариабельность в отношении количества меток последовательности, которые картируются на нее среди образцов и серий секвенирования, которая аппроксимирует вариабельность последовательности, представляющей интерес, для которой ее применяют в качестве параметра нормирования. Нормирующая последовательность может отличить пораженный образец от одного или более непораженных

образцов. Согласно некоторым вариантам реализации нормирующая последовательность лучше или более эффективно отличает пораженный образец от одного или более непораженных образцов по сравнению с другими потенциальными нормирующими последовательностями, такими как другие хромосомы. Согласно некоторым вариантам реализации вариабельность нормирующей последовательности вычисляют как вариабельность дозы хромосомы для последовательности, представляющей интерес, среди образцов и серий секвенирования. Согласно некоторым вариантам реализации нормирующие последовательности идентифицируют во множестве непораженных образцов.

"Нормирующая хромосома", "нормирующая хромосома в знаменателе" или "последовательность нормирующей хромосомы" представляет собой пример "нормирующей последовательности". "Последовательность нормирующей хромосомы" может состоять из одной хромосомы или из группы хромосом. Согласно некоторым вариантам реализации нормирующая последовательность содержит две или более устойчивых хромосом. Согласно определенным вариантам реализации устойчивые хромосомы представляют собой все аутосомные хромосомы, отличные от хромосом X, Y, 13, 18 и 21. "Нормирующий сегмент" представляет собой другой пример "нормирующей последовательности". "Последовательность нормирующего сегмента" может состоять из одного сегмента хромосомы или может состоять из двух или более сегментов одной и той же или различных хромосом. Согласно определенным вариантам реализации нормирующая последовательность предназначена для нормирования в отношении вариабельности, такой как связанная с процессом, межхромосомная (в одной серии определений) вариабельность и вариабельность между секвенированиями (в нескольких сериях определений).

Термин "дифференцируемость" в настоящем документе означает характеристику нормирующей хромосомы, которая позволяет отличить один или более непораженных, т.е. нормальных, образцов от одного или более пораженных, т.е. анеуплоидных, образцов. Нормирующая хромосома, демонстрирующая наибольшую "дифференцируемость", представляет собой хромосому или группу хромосом, которые обеспечивают наибольшее статистическое различие между распределением доз хромосом для хромосомы, представляющей интерес, во множестве квалификационных образцов и дозы хромосомы для той же хромосомы, представляющей интерес, в соответствующей хромосоме в одном или более пораженных образцах.

Термин "вариабельность" в настоящем документе означает другую характеристику нормирующей хромосомы, которая позволяет отличить один или более непораженных, т.е. нормальных, образцов от одного или более пораженных, т.е. анеуплоидных, образцов. Вариабельность нормирующей хромосомы, которую измеряют во множестве квалификационных образцов, означает вариабельность количества меток последовательности, которые картируются на нее, которое аппроксимирует вариабельность количества меток последовательности, которые картируются на хромосому, представляющую интерес, для которой она выступает как параметр нормирования.

Термин "плотность метки последовательности" в настоящем документе означает количество ридов последовательности, которые картируются на последовательность референсного генома, например, плотность метки последовательности для хромосомы 21 представляет собой количество ридов последовательности, полученных посредством способа секвенирования, которые картируются на хромосому 21 референсного генома.

Термин "соотношение плотности метки последовательности" в настоящем документе означает соотношение количества меток последовательности, которые картируются на хромосому референсного генома, например, хромосому 21, и длины хромосомы референсного генома.

Термин "доза последовательности" в настоящем документе означает параметр, который соотносит количество меток последовательности или другой параметр, идентифицированный для последовательности, представляющей интерес, с количеством меток последовательности или другим параметром, идентифицированным для нормирующей последовательности. В некоторых случаях доза последовательности представляет собой соотношение перекрытия метки последовательности или другого параметра для последовательности, представляющей интерес, и перекрытия метки последовательности или другого параметра для нормирующей последовательности. В некоторых случаях доза последовательности означает параметр, который соотносит плотность метки последовательности для последовательности, представляющей интерес, с плотностью метки последовательности нормирующей последовательности. "Доза исследуемой последовательности" представляет собой параметр, который соотносит плотность метки последовательности или другой параметр последовательности, представляющей интерес, например, хромосомы 21, с таковой нормирующей последовательности, например, хромосомы 9, определенной в исследуемом образце. Аналогично, "доза квалификационной последовательности" представляет собой параметр, который соотносит плотность метки последовательности или другой параметр последовательности, представляющей интерес, с таковой нормирующей последовательности, определенной в квалификационном образце.

Термин "перекрытие" означает изобилие меток последовательности, картированных на заданную последовательность. Перекрытие можно количественно определить на основании плотности метки последовательности (или подсчета меток последовательности), соотношения плотности метки последовательности, количества нормированного перекрытия, подогнанных значений перекрытия и т.д.

Термин "количество перекрытия" означает модификацию первичного перекрытия и часто представляет собой относительное количество меток последовательности (иногда называемое результатами подсчета) в области генома, такой как блок. Количество перекрытия можно получить посредством нормирования, подгонки и/или исправления первичного перекрытия или подсчитанного значения для области генома. Например, нормированное количество перекрытия для области можно получить посредством деления подсчитанного значения метки последовательности, картированной на области, на суммарное количество меток последовательности, картированных на целом геноме. Нормированное количество перекрытия позволяет проводить сравнение перекрытия блока между различными образцами, которые могут характеризоваться различными глубинами секвенирования. Нормированное количество перекрытия отличается от дозы последовательности тем, что последнюю, как правило, получают посредством деления на вычисление меток, картированных на подмножество целого генома. Подмножество представляет собой один или более нормирующих сегментов или хромосом. Количества перекрытия, будь то нормированные или не нормированные, можно корректировать с учетом глобального профиля вариации от области к области в геноме, вариаций фракции G-C, выпадающих показателей в устойчивых хромосомах и т.д.

Термин "секвенирование нового поколения (СНП)" в настоящем документе означает способы секвенирования, позволяющие проводить широкомасштабное параллельное секвенирование клонально амплифицированных молекул и отдельных молекул нуклеиновой кислоты. Неограничивающие примеры СНП включают секвенирование посредством синтеза с применением обратимых красителей-терминаторов и секвенирование посредством лигирования.

Термин "параметр" в настоящем документе означает числовое значение, которое характеризует свойство системы. Часто параметр численно характеризует множество количественных данных и/или числовую взаимосвязь между множествами количественных данных. Например, соотношение (или функцию соотношения) между количеством меток последовательности, картированных на хромосому, и длиной хромосомы, на которую картированы метки, представляет собой параметр.

Термины "пороговое значение" и "квалификационное пороговое значение" в настоящем документе означают любое число, которое применяют в качестве предела для характеризации образца, такого как исследуемый образец, содержащий нуклеиновую кислоту из организма, который, как подозревают, страдает от медицинского состояния. Порог можно сравнить со значением параметра для определения того, способствует ли образец возникновению такого значения параметра, который свидетельствует, что организм страдает от медицинского состояния. Согласно определенным вариантам реализации квалификационное пороговое значение вычисляют с применением множества квалификационных данных, и квалификационное пороговое значение выступает в качестве предела диагностики вариации числа копии, например, анеуплоидии, в организме. Если результаты, полученные в результате способов, раскрытых в настоящем документе, превосходят порог, у субъекта можно диагностировать вариацию числа копий, например, трисомию 21. Соответствующие пороговые значения для способов, описанных в настоящем документе, можно идентифицировать посредством анализа нормированных значений (например, доз хромосомы, NCV (normalized chromosome value, нормированного значения хромосомы) или NSV (normalized segment value, нормированного значения сегмента)), вычисленных для обучающего множества образцов. Пороговые значения можно идентифицировать с применением квалификационных (т.е. непораженных) образцов в обучающем множестве, которое содержит как квалификационные (т.е. непораженные) образцы, так и пораженные образцы. Образцы в обучающем множестве, которые установленно содержат анеуплоидии хромосом (т.е. пораженные образцы), можно применять для подтверждения того, что выбранные пороги являются подходящими для установления отличия пораженных от непораженных образцов в исследуемом множестве (см. примеры, представленные в настоящем документе). Выбор порога зависит от уровня достоверности, который выбирает пользователь для проведения классификации. Согласно некоторым вариантам реализации обучающее множество, применяемое для идентификации соответствующих пороговых значений, содержит по меньшей мере 10, по меньшей мере 20, по меньшей мере 30, по меньшей мере 40, по меньшей мере 50, по меньшей мере 60, по меньшей мере 70, по меньшей мере 80, по меньшей мере 90, по меньшей мере 100, по меньшей мере 200, по меньшей мере 300, по меньшей мере 400, по меньшей мере 500, по меньшей мере 600, по меньшей мере 700, по меньшей мере 800, по меньшей мере 900, по меньшей мере 1000, по меньшей мере 2000, по меньшей мере 3000, по меньшей мере 4000 или более квалификационных образцов. Для улучшения диагностической значимости пороговых значений может характеризоваться преимуществом применение больших множеств квалификационных образцов.

Термин "блок" означает сегмент последовательности или сегмент генома. Согласно некоторым вариантам реализации блоки являются непрерывными друг относительно друга в пределах генома или хромосомы. Каждый блоки может определять последовательность нуклеотидов в референсном геноме. Размеры блока могут составлять 1 т.о., 100 т.о., 1 Мб (мегабазу) и т.д. в зависимости от анализа, который требуется для конкретных применений, и плотности метки последовательности. В дополнение к положениям в пределах референсной последовательности блоки могут обладать другими характеристиками, такими как перекрытие образца и структурные характеристики последовательности, такие как фракция

G-C.

Термин "порог маскирования" в настоящем документе означает количество, с которым сравнивают значение, основанное на количестве меток последовательности, в блоке последовательности, причем блоки, который характеризуется значением, превосходящим порог маскирования, маскируют. Согласно некоторым вариантам реализации порог маскирования может представлять собой процентильный ранг, абсолютное количество, показатель качества картирования или другие подходящие значения. Согласно некоторым вариантам реализации порог маскирования можно задать как процентильный ранг коэффициента вариации среди множества непораженных образцов. Согласно другим вариантам реализации порог маскирования можно задать как показатель качества картирования, например, показатель MapQ, который относится к надежности выравнивания ридов последовательности с референсным геномом. Отметим, что пороговое значение маскирования отличается от порогового значения вариации числа копий (ВЧК), причем последнее представляет собой предел, характеризующий образец, содержащий нуклеиновую кислоту из организма, который, как подозревают, страдает от медицинского состояния, связанного с ВЧК. Согласно некоторому варианту реализации пороговое значение ВЧК задают по сравнению с нормированным значением хромосомы (normalized chromosome value, NCV) или нормированным значением сегмента (normalized segment value, NSV), описанными в настоящем документе в другом месте.

Термин "нормированное значение" в настоящем документе означает числовое значение, которое соотносит количество меток последовательности, идентифицированных для последовательности (например, хромосомы или сегмента хромосомы), представляющей интерес, с количеством меток последовательности, идентифицированных для нормирующей последовательности (например, нормирующей хромосомы или нормирующего сегмента хромосомы). Например, "нормированное значение" может представлять собой дозы хромосомы, описанные в настоящем документе в другом месте, или может представлять собой NCV, или может представлять собой NSV, описанные в настоящем документе в другом месте.

Термин "рид" означает последовательность, полученную из части образца нуклеиновой кислоты. Как правило, хотя и не обязательно, рид представляет собой короткую последовательность непрерывных пар оснований в образце. Рид может быть представлен символически последовательностью пар оснований (в A, T, C или G) части образца. Рид может храниться на запоминающем устройстве и обрабатываться соответствующим образом для определения того, соответствует ли оно референсной последовательности или соответствует ли другим критериям. Рид можно получить непосредственно из аппарата секвенирования или опосредованно из хранящейся информации о последовательности образца. В некоторых случаях рид представляет собой последовательность ДНК достаточной длины (например, по меньшей мере приблизительно 25 п.о.), которую можно применять для идентификации большей последовательности или области, например, которую можно выровнять и специфично отнести к хромосоме или геномной области или гену.

Термин "геномный рид" используют применительно к ридам любых сегментов в целом геноме индивидуума.

Термин "метка последовательности" в настоящем документе используется взаимозаменяемо с термином "метка картированной последовательности" и означает рид последовательности, которое было специфично отнесено, т.е. картировано, к большей последовательности, например, референсному геному, посредством выравнивания. Метки картированной последовательности являются уникально картированными на референсный геном, т.е. они отнесены к одному расположению в референсном геноме. Если не указано обратное, метки, которые картируются на одну и ту же последовательность на референсной последовательности, подсчитывают один раз. Метки могут быть предложены в виде структур данных или других совокупностей данных. Согласно определенным вариантам реализации метка содержит последовательность рида и связанную информацию для данного рида, такую как расположение последовательности в геноме, например, положение на хромосоме. Согласно определенным вариантам реализации положение указано для положительной ориентации цепи. Можно задать метку, чтобы обеспечить ограниченное количество несоответствия при выравнивании с референсным геномом. Согласно некоторым вариантам реализации метки, которые можно картировать на более чем одно расположение на референсном геноме, т.е. метки, которые не картируются уникально, можно не включать в анализ.

Термин "не повторяющаяся метка последовательности" означает метки последовательности, которые не картируются на один и тот же сайт, которые подсчитывают с целью определения нормированных значений хромосом (NCV) согласно некоторым вариантам реализации. Иногда несколько ридов последовательности выравниваются с одними и теми же расположениями на референсном геноме с получением повторяющихся или дублирующихся меток последовательности. Согласно некоторым вариантам реализации дублирующиеся метки последовательности, которые картируются на одно и то же положение, опускают или подсчитывают как одну "не повторяющуюся метку последовательности" с целью определения NCV. Согласно некоторым вариантам реализации не повторяющиеся метки последовательности, выровненные с неисключенными сайтами, подсчитывают для получения "подсчитанного значения неисключенных сайтов" (подсчитанных значений NES, non-excluded site) для определения NCV.

Термин "сайт" означает уникальное положение (т.е. идентификатор хромосомы, положение и ори-

ентацию хромосомы) на референсном геноме. Согласно некоторым вариантам реализации сайт может обеспечить положение для остатка, метки последовательности или сегмента на последовательности.

"Исключенные сайты" представляют собой сайты, обнаруженные в областях референсного генома, которые были исключены из подсчитанного значения меток последовательности. Согласно некоторым вариантам реализации исключенные сайты обнаружены в областях хромосом, которые содержат повторяющиеся последовательности, например, центромеры и теломеры, и в областях хромосом, которые являются общими для более одной хромосомы, например, в областях, присутствующих на Y-хромосоме, которые также присутствуют на X-хромосоме.

"Неисключенные сайты" (NES) представляют собой сайты, которые не исключены в референсном геноме при подсчете меток последовательности.

"Подсчитанные значения неисключенных сайтов" (подсчитанные значения NES) представляют собой количества меток последовательности, которые картируются на NES на референсном геноме. Согласно некоторым вариантам реализации NES представляют собой количества не повторяющихся меток последовательности, картированых на NES. Согласно некоторым вариантам реализации перекрытие и связанные параметры, такие как нормированные количества перекрытия, глобальный профиль с устранением количеств перекрытия и доза хромосомы, основаны на подсчитанных значениях NES. В одном примере дозу хромосомы вычисляют как соотношение подсчитанного значения NES для хромосомы, представляющей интерес, и подсчитанного значения для нормирующей хромосомы.

Нормированное значение хромосомы (NCV) представляет собой соотношение перекрытия исследуемого образца и перекрытий множества обучающих/квалификационных образцов. Согласно некоторым вариантам реализации NCV основано на дозе хромосомы. Согласно некоторым вариантам реализации NCV относится к различию между дозой хромосомы для хромосомы, представляющей интерес, в исследуемом образце и средним значением соответствующей дозы хромосомы во множестве квалификационных образцов и может быть вычислено как

$$NCV_{ij} = \frac{x_{ij} - \hat{\mu}_j}{\hat{\sigma}_j},$$

где $\hat{\mu}_j$ и $\hat{\sigma}_j$ представляют собой вычисленное среднее значение и стандартное отклонение, соответственно, для дозы j-й хромосомы во множестве квалификационных образцов, и x_{ij} представляет собой наблюдаемое соотношение j-й хромосомы (дозы) для исследуемого образца i.

Согласно некоторым вариантам реализации NCV может быть вычислено "на ходу" посредством соотнесения дозы хромосомы для хромосомы, представляющей интерес, в исследуемом образце, к медиане соответствующей дозы хромосомы в мультиплексных образцах, секвенированных в одних и тех же проточных ячейках, как

$$NCV_{ij} = \frac{x_{ij} - \mathrm{M}_j}{\hat{\sigma}_j},$$

где M_j представляет собой вычисленную медиану для дозы j-й хромосомы во множестве мультиплексных образцов, секвенированных в одной и той же проточной ячейке;

$\hat{\sigma}_j$ представляет собой стандартное отклонение для дозы j-й хромосомы в одном или более множествах мультиплексных образцов, секвенированных в одной или более проточных ячейках, и X_{ij} представляет собой наблюдаемую дозу j-й хромосомы для исследуемого образца i. Согласно данному варианту реализации исследуемый образец i представляет собой один из мультиплексных образцов, секвенированных в одной и той же проточной ячейке, из которой определяют M_j.

Например, для хромосомы 21, представляющей интерес, в исследуемом образце A, который секвенирован как один из 64 мультиплексных образцов в одной проточной ячейке, NCV для хромосомы 21 в исследуемом образце A вычисляют как дозу хромосомы 21 в образце A минус медиана дозы для хромосомы 21, определенной в 64 мультиплексных образцах, разделенную на стандартное отклонение дозы для хромосомы 21, определенной для 64 мультиплексных образцов в проточной ячейке 1 или в дополнительных проточных ячейках.

В настоящем документе термины "выровненный" или "выравнивание" означают процесс сравнения риды или метки с референсной последовательностью и посредством этого определения того, содержит ли референсная последовательность последовательность рида. Если референсная последовательность содержит рид, рид можно картировать на референсную последовательность или согласно определенным вариантам реализации на конкретное расположение в референсной последовательности. В некоторых случаях выравнивание просто показывает, является ли рид членом конкретной референсной последовательности или нет (т.е. присутствует или отсутствует рид в референсной последовательности). Например, выравнивание риды с референсной последовательностью для хромосомы человека 13 продемонстрирует, присутствует ли рид в референсной последовательности для хромосомы 13. Инструмент, который обеспечивает данную информацию, можно назвать определителем принадлежности множеству. В некоторых случаях выравнивание дополнительно указывает на расположение в референсной последовательности, на которое картируется рид или метка. Например, если референсная последовательность

представляет собой последовательность целого генома человека, выравнивание может указать на то, что рид присутствует на хромосоме 13, и может также указать на то, что рид находится на конкретной цепи и/или сайте хромосомы 13.

Выровненные риды или метки представляют собой одну или более последовательностей, которые идентифицированы как совпадение применительно к порядку их молекул нуклеиновой кислоты с известной последовательностью из референсного генома. Выравнивание можно выполнить вручную, несмотря на то, что выравнивание, как правило, осуществляют с помощью компьютерного алгоритма, поскольку для реализации способов, раскрытых в настоящем документе, невозможно выровнять риды в течение разумного периода времени. Примером алгоритма для выравнивания последовательностей является компьютерная программа Efficient Local Alignment of Nucleotide Data (ELAND, Эффективное локальное выравнивание нуклеотидных данных), которую распространяют как часть ассортимента программ Genomics Analysis (Геномный анализ) компании Illumina. В качестве альтернативы, для выравнивания риды с референсными геномами можно применять фильтр Bloom или аналогичный определитель принадлежности множеству. См. заявку на патент США № 61/552374, поданную 27 октября 2011 года, которая полностью включена в настоящий документ посредством ссылки. Совпадение последовательности риды в выравнивании может составлять 100% совпадения последовательности или менее 100% (неидеальное совпадение).

Термин "картирование" в настоящем документе означает специфичное отнесение последовательности рида к большей последовательности, например, референсному геному, посредством выравнивания.

В настоящем документе термин "референсный геном" или "референсная последовательность" означает любую известную конкретную последовательность генома, будь то частичную или полную, любого организма или вируса, которую можно применять для сравнения с идентифицированными последовательностями от субъекта. Например, референсный геном, применяемый в случае субъектов-людей, а также многих других организмов, можно найти в Национальном центре биотехнологической информации (National Center for Biotechnology Information) по адресу ncbi.nlm.nih.gov. "Геном" означает полную генетическую информацию организма или вируса, выраженную в последовательностях нуклеиновой кислоты.

Согласно различным вариантам реализации референсная последовательность является значительно большей, чем риды, которые с ней выравнивают. Например, референсная последовательность может быть по меньшей мере приблизительно в 100 раз большей, или по меньшей мере приблизительно в 1000 раз большей, или по меньшей мере приблизительно в 10000 раз большей, или по меньшей мере приблизительно в 10^5 раз большей, или по меньшей мере приблизительно в 10^6 раз большей, или по меньшей мере приблизительно в 10^7 раз большей.

В одном примере референсная последовательность представляет собой таковую полной длины генома человека. Такие последовательности можно назвать геномными референсными последовательностями. В другом примере референсная последовательность ограничена конкретной хромосомой человека, такой как хромосома 13. Согласно некоторым вариантам реализации референсная Y-хромосома представляет собой последовательность Y-хромосомы из генома человека версии hg19. Такие последовательности можно назвать референсными последовательностями хромосомы. Другие примеры референсных последовательностей включают геномы других видов, а также хромосомы, субхромосомные области (такие как цепи) и т.д. любого вида.

Согласно различным вариантам реализации референсная последовательность представляет собой консенсусную последовательность или другую комбинацию, полученную от нескольких индивидуумов. Однако в определенных вариантах применения референсная последовательность может быть получена от конкретного индивидуума.

Термин "клинически значимая последовательность" в настоящем документе означает последовательность нуклеиновой кислоты, которая, как известно или как предполагают, связана с генетическим состоянием или состоянием заболевания или вовлечена в такое состояние. Определение отсутствия или присутствия клинически значимой последовательности может являться подходящим при определении диагноза или при подтверждении диагноза медицинского состояния либо при составлении прогноза развития заболевания.

Термин "полученный" при использовании в контексте нуклеиновой кислоты или смеси нуклеиновых кислот в настоящем документе означает средства, посредством которых нуклеиновую кислоту или кислоты получают из источника, из которого они происходят. Например, согласно одному варианту реализации смесь нуклеиновых кислот, которая получена из двух различных геномов, означает, что нуклеиновые кислоты, например, сцДНК, были природным путем высвобождены клетками в результате встречающихся в природе процессов, таких как некроз или апоптоз. Согласно другому варианту реализации смесь нуклеиновых кислот, которая получена из двух различных геномов, означает, что нуклеиновые кислоты были экстрагированы из двух различных типов клеток от субъекта.

Термин "основано на" при использовании в контексте получения конкретного количественного значения в настоящем документе означает применение другого количества в качестве входных данных для вычисления конкретного количественного значения в качестве выходных данных.

Термин "образец от пациента" в настоящем документе означает биологический образец, полученный от пациента, т.е. реципиента медицинского обслуживания, помощи или лечения. Образец от пациента может представлять собой любые образцы, описанные в настоящем документе. Согласно определенным вариантам реализации образец от пациента получен в результате неинвазивных процедур, например образец периферической крови или образец стула. Способы, описанные в настоящем документе, не следует ограничивать людьми. Таким образом, предусмотрены различные ветеринарные варианты применения, в случае которых образец от пациента может представлять собой образец от млекопитающего, отличного от человека (например, кошек, свиней, лошадей, крупного рогатого скота и т.п.).

Термин "смешанный образец" в настоящем документе означает образец, содержащий смесь нуклеиновых кислот, полученных из различных геномов.

Термин "материнский образец" в настоящем документе означает биологический образец, полученный от беременного субъекта, например, женщины.

Термин "биологическая жидкость" в настоящем документе означает жидкость, отобранную из биологического источника, и включает, например, кровь, сыворотку, плазму, мокроту, промывную жидкость, спинномозговую жидкость, мочу, семенную жидкость, пот, слезы, слюну и т.п. В настоящем документе термины "кровь", "плазма" и "сыворотка" однозначно включают фракции или их обработанные части. Аналогично, когда образец отбирают из биопсии, мазка, соскоба и т.д., "образец" однозначно включает процессированную фракцию или часть, полученную из биопсии, мазка, соскоба и т.д.

Термины "материнские нуклеиновые кислоты" и "нуклеиновые кислоты плода" в настоящем документе означают нуклеиновые кислоты беременного субъекта женского пола и нуклеиновые кислоты плода, вынашиваемого беременным субъектом женского пола, соответственно.

В настоящем документе термин "соответствующий" иногда означает последовательность нуклеиновой кислоты, например, ген или хромосому, которая присутствует в геноме различных субъектов и которая необязательно характеризуются одинаковой последовательностью во всех геномах, но которая выступает для обеспечения идентичности вместо генетической информации последовательности, представляющей интерес, например, гена или хромосомы.

В настоящем документе термин "фракция плода" означает фракцию нуклеиновых кислот плода, присутствующую в образце, содержащем нуклеиновые кислоты плода и матери. Фракции плода часто применяют для характеризации сцДНК в крови матери.

В настоящем документе термин "хромосома" означает обеспечивающий наследственность носитель генов живой клетки, который получен из цепей хроматина, содержащих ДНК и белковые компоненты (в частности, гистоны). В настоящем документе применяют общепринятую международно признанную систему нумерации отдельных хромосом генома человека.

В настоящем документе термин "длина полинуклеотида" означает абсолютное количество нуклеотидов в последовательности или в области референсного генома. Термин "длина хромосомы" означает известную длину хромосомы, приведенную в парах оснований, например, представленную в сборке NCBI36/hg18 хромосом человека, которую можно найти в сети Интернет по адресу: |genome|.|ucsc|.|edu/cgi-bin/hgTracks?hgsid=167155613&chromInfoPage=.

Термин "субъект" в настоящем документе означает субъекта-человека, а также субъекта, отличного от человека, такого как млекопитающее, беспозвоночное животное, позвоночное животное, грибы, дрожжи, бактерии и вирус. Несмотря на то что примеры в настоящем документе относятся к людям, и описание преимущественно направлено на проблемы человека, концепции, раскрытые в настоящем документе, применимы к геномам любого растения или животного и являются подходящими в областях ветеринарной медицины, наук о животных, в исследовательских лабораториях и т.п.

Термин "состояние" в настоящем документе означает "медицинское состояние" как широкий термин, который включает все заболевания и нарушения, но может включать поражения и нормальные состояния здоровья, такие как беременность, которые могут оказывать влияние на здоровье субъекта, получать пользу от медицинской помощи или иметь последствия для медицинского лечения.

Термин "полная" при использовании применительно к анеуплоидии хромосом в настоящем документе означает добавление или утрату целой хромосомы.

Термин "частичная" при использовании применительно к анеуплоидии хромосом в настоящем документе означает добавление или утрату части, т.е. сегмента, хромосомы.

Термин "мозаик" в настоящем документе означает присутствие у одного индивидуума, который развился из одной оплодотворенной яйцеклетки, двух популяций клеток с различными кариотипами. Мозаицизм может являться следствием мутации в процессе развития, которая передалась исключительно подмножеству взрослых клеток.

Термин "немозаичный" в настоящем документе означает организм, например, плод человека, состоящий из клеток одного кариотипа.

Термин "чувствительность" в настоящем документе означает вероятность того, что результаты анализа будут положительными, если присутствует состояние, представляющее интерес. Чувствительность можно вычислить как количество истинно положительных результатов, разделенное на сумму истинно положительных и ложноотрицательных результатов.

Термин "специфичность" в настоящем документе означает вероятность того, что результаты анализа будут отрицательными, если отсутствует состояние, представляющее интерес. Специфичность можно вычислить как количество истинно отрицательных результатов, разделенное на сумму истинно отрицательных и ложноположительных результатов.

Термин "обогащать" в настоящем документе означает процесс амплификации полиморфных целевых нуклеиновых кислот, которые содержатся в части материнского образца, и объединения амплифицированного продукта с оставшимся материнским образцом, из которого была отобрана часть. Например, оставшийся материнский образец может представлять собой исходный материнский образец.

Термин "исходный материнский образец" в настоящем документе означает необогащенный биологический образец, полученный от беременного субъекта, например, женщины, выступающего в качестве источника, от которого отбирают часть для амплификации полиморфных целевых нуклеиновых кислот. "Исходный образец" может представлять собой любой образец, полученный от беременного субъекта, и процессированные фракции данного образца, например, очищенный образец сцДНК, экстрагированный из образца материнской плазмы.

Термин "праймер" в настоящем документе означает выделенный олигонуклеотид, который способен выступать в качестве точки инициации синтеза при помещении в условия, вызывающие синтез продукта удлинения (например, условия включают нуклеотиды, индуцирующий агент, такой как ДНК-полимераза, и подходящие температуру и pH). Праймер предпочтительно является одноцепочечным для максимальной эффективности при амплификации, но, в качестве альтернативы, может являться двухцепочечным. В случае двухцепочечного праймера праймер сначала обрабатывают с целью разделения его цепей перед применением для получения продуктов удлинения. Предпочтительно, праймер представляет собой олигодезоксирибонуклеотид. Праймер должен быть достаточно длинным, чтобы запускать синтез продуктов удлинения в присутствии индуцирующего агента. Точные длины праймеров зависят от множества факторов, включая температуру, источник праймера, применение способа и параметры, применяемые для разработки праймера.

Введение и контекст.

ВЧК в геноме человека в значительной степени влияет на этническое разнообразие и предрасположенность человека к заболеваниям (Redon et al., Nature 23:444-454 [2006], Shaikh et al., Genome Res 19:1682-1690 [2009]). Такие заболевания включают, без ограничения, рак, инфекционные и аутоиммунные заболевания, заболевания нервной системы, метаболические и/или сердечно-сосудистые заболевания и т.п.

Известно, что ВЧК способствует генетическим заболеваниям посредством различных механизмов, которые приводят в большинстве случаев к дисбалансу дозы гена или к разрушению гена. Известно, что в дополнение к прямой корреляции с генетическими нарушениями ВЧК опосредуют фенотипические изменения, которые могут быть пагубными. Недавно в нескольких исследованиях было сообщено об увеличенной нагрузке редкой ВЧК или ВЧК de novo при комплексных нарушениях, таких как аутизм, СДВГ (синдром дефицита внимания при гиперактивности) и шизофрения, по сравнению с нормальными контролями, что подчеркивает потенциальную патогенность редкой или уникальной ВЧК (Sebat et al., 316:445 - 449 [2007]; Walsh et al., Science 320:539 - 543 [2008]). ВЧК возникают в результате геномных реаранжировок, преимущественно, вследствие явлений делеции, дупликации, вставки и несбалансированной транслокации.

Было показано, что фрагменты сцДНК плодного происхождения в среднем являются более короткими, чем таковые материнского происхождения. Успешно применяли НИПТ (неинвазивное пренатальное тестирование), основанное на данных СНП. В применяемых на сегодняшний день методологиях используют секвенирование материнских образцов с применением коротких ридов (25 п.о. - 36 п.о.), выравнивание с геномом, компьютеризированное вычисление и нормирование субхромосомного перекрытия и, наконец, оценку чрезмерной представленности целевых хромосом (13/18/21/X/Y) по сравнению с ожидаемым нормированным перекрытием, связанным с нормальным диплоидным геномом. Таким образом, традиционный анализ и исследование НИПТ основаны на подсчитанных значениях или перекрытии для оценки правдоподобия анеуплоидии плода.

Поскольку образцы материнской плазмы представляют собой смесь материнской и плодной сцДНК, успех любого данного способа НИПТ зависит от его чувствительности для обнаружения изменений числа копий в незначительных образцах фракции плода. Для способов, основанных на подсчитанном значении, чувствительность определяется (a) глубиной секвенирования и (b) способностью нормирования данных снижать техническую дисперсию. В настоящем изобретении предложена аналитическая методология для НИПТ и других вариантов применения посредством получения информации о размере фрагмента из, например, риды спаренных концов, и применение данной информации в анализе ассортимента. Улучшенная аналитическая чувствительность обеспечивает способность применять способы НИПТ при сниженном перекрытии (например, сниженной глубине секвенирования), что делает возможным применение технологии для недорогостоящего исследования среднего риска беременностей.

В настоящем документе раскрыты способы, аппараты и системы для определения числа копий и вариаций числа копий (ВЧК) различных последовательностей, представляющих интерес, в исследуемом

образце, который содержит смесь нуклеиновых кислот, полученную из двух или более различных геномов, и который, как известно или как предполагают, отличается количеством одной или более последовательностей, представляющих интерес. Вариации числа копий, определенные с применением способов и аппаратов, раскрытых в настоящем документе, включают добавления или утраты целых хромосом, изменения, затрагивающие очень большие сегменты хромосом, которые являются видимыми под микроскопом, и изобилие субмикроскопических вариаций числа копий сегментов ДНК, варьирующих по размеру от одного нуклеотида до тысяч оснований (т.о.) и мегабаз (Мб).

Согласно некоторым вариантам реализации предложены способы определения вариации числа копий (ВЧК) плодов с применением материнских образцов, содержащих материнскую и бесклеточную ДНК плода. В некоторых вариантах реализации применяют длину фрагмента (или размер фрагмента) сцДНК для улучшения чувствительности и специфичности с целью обнаружения анеуплоидии плода из сцДНК в материнской плазме. Некоторые варианты реализации осуществляют с получением библиотек без применения ПНР в сочетании с секвенированием спаренных концов ДНК. Согласно некоторым вариантам реализации для усиления обнаружения анеуплоидии плода применяют как размер фрагмента, так и перекрытие. Согласно некоторым вариантам реализации способы включают объединение независимого подсчета более коротких фрагментов с относительной фракцией более коротких фрагментов в блоках в пределах генома.

В некоторых вариантах реализации, раскрытых в настоящем документе, предложены способы улучшения чувствительности и/или специфичности анализов данных последовательности посредством устранения внутривыборочной погрешности содержания GC. Согласно некоторым вариантам реализации устранение внутривыборочной погрешности содержания GC основано на данных последовательности, откорректированных с учетом систематической вариации, распространенной в пределах непораженных обучающих образцов.

В некоторых раскрытых вариантах реализации предложены способы получения параметров с высоким соотношением сигнал/шум из фрагментов бесклеточной нуклеиновой кислоты для определения различных генетических состояний, связанных с числом копий и ВЧК, с улучшенной чувствительностью, селективностью и/или эффективностью по сравнению с общепринятыми способами. Параметры включают, без ограничения, перекрытие, взвешенное по размеру фрагмента перекрытие, фракцию или отношение фрагментов в заданном диапазоне, уровень метилирования фрагментов, t-статистику, полученную из перекрытия, оценки фракции плода, полученные из информации о перекрытии, и т.д. Было установлено, что представленный процесс является в особенности эффективным для улучшения сигнала в образцах, содержащих относительно низкие фракции ДНК из рассматриваемого генома (например, генома плода). Пример такого образца представляет собой образец материнской крови от индивидуума, беременного разнояйцевыми близнецами, тройней и т.д., при котором процесс оценивает вариацию числа копий в геноме одного из плодов.

Согласно некоторым вариантам реализации высоких аналитических чувствительностей и специфичностей можно достичь при простом получении библиотеки с применением очень низкого количества сцДНК на входе, для которого не требуется ПЦР-амплификация. Способ без применения ПЦР упрощает рабочий процесс, улучшает время оборота и устраняет погрешности, присущие способам на основе ПЦР. Согласно некоторым вариантам реализации обнаружение анеуплоидии плода из материнской плазмы можно провести более надежным и эффективным способом, чем общепринятые способы, с потребностью в меньшем количестве уникальных фрагментов сцДНК. В комбинации улучшенную аналитическую чувствительность и специфичность достигают с очень быстрым временем оборота при значительно меньшем количестве фрагментов сцДНК. Это потенциально позволяет проводить НИПТ со значительно меньшими затратами для облегчения применения в общей популяции беременных.

Согласно различным вариантам реализации с помощью раскрытых способов возможно получение библиотеки без применения ПЦР. Некоторые варианты реализации устраняют погрешности, присущие способам ПЦР, снижают сложность анализа, снижают требуемую глубину секвенирования (в 2,5 раза), обеспечивают более быстрое время оборота, например, оборот за один день, делают возможным внутреннее измерение фракции плода (ФЭ), облегчают установление отличий между материнской и плодной/плацентарной сцДНК с применением информации о размере фрагмента.

Оценка ВЧК.

Способы определения ВЧК.

С применением значения перекрытия последовательности, параметров размера фрагментов и/или уровней метилирования, обеспеченных в способах, раскрытых в настоящем документе, можно определить различные генетические состояния, связанные с числом копий и ВЧК последовательностей, хромосом или сегментов хромосом с улучшенной чувствительностью, селективностью и/или эффективностью по сравнению с применением значений перекрытия последовательности, полученных общепринятыми способами. Например, согласно некоторым вариантам реализации для определения присутствия или отсутствия любых двух или более различных полных анеуплоидий хромосом плода в исследуемом материнском образце, содержащем молекулы плодной и материнской нуклеиновой кислоты, применяют маскированные референсные последовательности. В иллюстративных способах, предложенных ниже, риды

выравнивают с референсными последовательностями (включая референсные геномы). Выравнивание можно осуществить с немаскированной или маскированной референсной последовательностью, посредством этого получая метки последовательности, картированные на референсной последовательности. Согласно некоторым вариантам реализации для определения вариации числа копий учитывают исключительно метки последовательности, попадающие в немаскированные сегменты референсной последовательности.

Согласно некоторым вариантам реализации оценка образца нуклеиновой кислоты в отношении ВЧК включает характеризацию статуса анеуплоидий хромосом или анеуплоидий сегмента с помощью одного из трех типов решений: "нормальный" или "непораженный", "пораженный" и "решение отсутствует". Пороги для принятия решения о нормальных и пораженных образцах, как правило, установлены. В образце измеряют параметр, связанный с анеуплоидией или другой вариацией числа копий, и измеренное значение сравнивают с порогами. Для анеуплоидий типа дупликации решение о пораженном образце принимают, если хромосома или доза сегмента (или другое измеренное значение содержания последовательности) превышает определенный порог, заданный для пораженных образцов. Для таких анеуплоидий решение о нормальных образцах принимают, если доза хромосомы или сегмента ниже порога, заданного для нормальных образцов. Напротив, для анеуплоидий типа делеции решение о пораженных образцах принимают, если доза хромосомы или сегмента ниже определенного порога для пораженных образцов, и решение о нормальных образцах принимают, если доза хромосомы или сегмента выше порога, заданного для нормальных образцов. Например, в случае присутствия трисомии решение "нормальный" определяют значением параметра, например, дозы исследуемой хромосомы, который ниже заданного пользователем порога надежности, и решение "пораженный" определяют на основании параметра, например, дозы исследуемой хромосомы, который выше заданного пользователем порога надежности. Результат "решение отсутствует" определяют на основании параметра, например, дозы исследуемой хромосомы, который лежит между порогами для принятия решения "нормальный" или "пораженный". Термин "решение отсутствует" применяют взаимозаменяемо с термином "неклассифицированный".

Параметры, которые можно применять для определения ВЧК, включают, без ограничения, перекрытие, смещенное/взвешенное по размеру фрагмента перекрытие, фракцию или отношение фрагментов в заданном диапазоне размера и уровень метилирования фрагментов. Как обсуждается в настоящем документе, перекрытие получают из подсчитанных значений ридов, выровненных с областью референсного генома и необязательно нормированных для получения подсчитанных значений метки последовательности. Согласно некоторым вариантам реализации подсчитанные значения метки последовательности можно взвесить по размеру фрагмента.

Согласно некоторым вариантам реализации параметр размера фрагмента смещен в сторону характеристики размера фрагментов одного из геномов. Параметр размера фрагмента представляет собой параметр, который относится к размеру фрагмента. Параметр смещен в сторону размера фрагмента, когда: 1) параметр благоприятно взвешивается по размеру фрагмента, например, вычисление имеет больший вес по размеру, чем для других размеров; или 2) параметр получен из значения, которое благоприятно взвешивается по размеру фрагмента, например, соотношение получено из подсчитанного значение, который имеет больший вес по размеру. Размер представляет собой характеристику генома, когда геном характеризуется обогащенной или большей концентрацией размера нуклеиновой кислоты по сравнению с другим геномом или другой частью того же генома.

Согласно некоторым вариантам реализации способ определения присутствия или отсутствия любых полных анеуплоидий хромосом плода в материнском исследуемом образце включает (a) прием информации о последовательности для нуклеиновых кислот плода и матери в материнском исследуемом образце; (b) применение информации о последовательности и способа, описанного выше, для идентификации количества меток последовательности, количества перекрытия последовательности, параметра размера фрагмента или другого параметра для каждой из хромосом, представляющих интерес, выбранных из хромосом 1-22, X и Y, и для идентификации количества меток последовательности или другого параметра для одной или более последовательностей нормирующей хромосомы; (c) применение количества меток последовательности или другого параметра, идентифицированного для каждой из хромосом, представляющих интерес, и количества меток последовательности или другого параметра, идентифицированного для каждой из нормирующих хромосом, для вычисления дозы одной хромосомы для каждой из хромосом, представляющих интерес; и (d) сравнение каждой дозы хромосомы с пороговым значением и посредством этого определение присутствия или отсутствия любых полных анеуплоидий хромосом плода в материнском исследуемом образце.

Согласно некоторым вариантам реализации этап (a), описанный выше, может включать секвенирование по меньшей мере части молекул нуклеиновой кислоты исследуемого образца для получения указанной информации о последовательности для молекул нуклеиновой кислоты плода и матери исследуемого образца. Согласно некоторым вариантам реализации этап (c) включает вычисление дозы одной хромосомы для каждой из хромосом, представляющих интерес, как соотношения количества меток последовательности или другого параметра, идентифицированного для каждой из хромосом, представляющих интерес, и количества меток последовательности или другого параметра, идентифицированного для

последовательности или последовательностей нормирующей хромосомы. Согласно некоторым другим вариантам реализации доза хромосомы основана на количествах перекрытия процессированной последовательности, полученных из количества меток последовательности или другого параметра. Согласно некоторым вариантам реализации для вычисления количеств перекрытия процессированной последовательности или другого параметра применяют исключительно уникальные, не повторяющиеся метки последовательности. Согласно некоторым вариантам реализации количество перекрытий процессированной последовательности представляет собой соотношение плотности метки последовательности, которое представляет собой количество метки последовательности, стандартизированное по длине последовательности. Согласно некоторым вариантам реализации количество перекрытия процессированной последовательности или другой параметр представляет собой нормированную метку последовательности или другой нормированный параметр, который представляет собой количество меток последовательности или другой параметр последовательности, представляющей интерес, разделенный на таковой всего генома или значительной его части. Согласно некоторым вариантам реализации количество перекрытия процессированной последовательности или другой параметр, такой как параметр размера фрагмента, подгоняют в соответствии с глобальным профилем последовательности, представляющей интерес. Согласно некоторым вариантам реализации количество перекрытия процессированной последовательности или другой параметр подгоняют в соответствии с внутривыборочной корреляцией между содержанием GC и перекрытием последовательности для образца, исследование которого проводят. Согласно некоторым вариантам реализации количество перекрытия процессированной последовательности или другой параметр получают из комбинаций данных процессов, которые дополнительно описаны в настоящем документе в другом месте.

Согласно некоторым вариантам реализации дозу хромосомы вычисляют в виде соотношения перекрытия процессированной последовательности или другого параметра для каждой из хромосом, представляющих интерес, и такового для последовательности или последовательностей нормирующей хромосомы.

Согласно любому из вариантов реализации, описанных выше, полные анеуплоидии хромосом выбраны из полных трисомий хромосом, полных моносомий хромосом и полных полисомий хромосом. Полные анеуплоидии хромосом выбраны из полных анеуплоидии любой из хромосом 1-22, X и Y. Например, указанные различные полные анеуплоидии хромосом плода выбраны из трисомий 2, трисомий 8, трисомий 9, трисомий 20, трисомий 21, трисомий 13, трисомий 16, трисомий 18, трисомий 22, 47, XXX, 47, XYY и моносомий X.

Согласно любому из вариантов реализации, описанных выше, этапы (a)-(d) повторяют для исследуемых образцов от различных материнских субъектов, и способ включает определение присутствия или отсутствия любых двух или более различных полных анеуплоидий хромосом плода в каждом из исследуемых образцов.

Согласно любому из вариантов реализации, описанных выше, способ может также включать вычисление нормированного значения хромосомы (NCV), где NCV представляет собой отношение дозы хромосомы к среднему значению соответствующей дозы хромосомы во множестве квалификационных образцов, вычисленное как

$$NCV_{ij} = \frac{x_{ij} - \hat{\mu}_j}{\hat{\sigma}_j},$$

где $\hat{\mu}_j$ и $\hat{\sigma}_j$ представляют собой вычисленное среднее значение и стандартное отклонение, соответственно, для дозы j-й хромосомы во множестве квалификационных образцов, и x_{ij} представляет собой наблюдаемую дозу j-й хромосомы для исследуемого образца i.

Согласно некоторым вариантам реализации NCV может быть вычислено "на ходу" посредством соотнесения дозы хромосомы для хромосомы, представляющей интерес, в исследуемом образце к медиане соответствующей дозы хромосомы в мультиплексных образцах, секвенированных в одних и тех же проточных ячейках, как

$$NCV_{ij} = \frac{x_{ij} - \mathrm{M}_j}{\hat{\sigma}_j},$$

где M_j представляет собой вычисленную медиану для дозы j-й хромосомы во множестве мультиплексных образцов, секвенированных в одной и той же проточной ячейке;

$\hat{\sigma}_j$ представляет собой стандартное отклонение для дозы j-й хромосомы в одном или более множествах мультиплексных образцов, секвенированных в одной или более проточных ячейках, и x_i представляет собой наблюдаемую дозу j-й хромосомы для исследуемого образца i. Согласно данному варианту реализации исследуемый образец i представляет собой один из мультиплексных образцов, секвенированных в одной и той же проточной ячейке, для которого определяют M_j.

Согласно некоторым вариантам реализации предложен способ определения присутствия или отсутствия различных частичных анеуплоидий хромосом плода в материнском исследуемом образце, содержащем нуклеиновых кислот плода и матери. Способ включает процедуры, аналогичные способу обнару-

жения полной анеуплоидий, изложенному выше. Однако вместо анализа полной хромосомы анализируют сегмент хромосомы. См. публикацию заявки на патент США № 2013/0029852, которая включена в настоящую заявку посредством ссылки.

На фиг. 1 представлен способ определения присутствия вариации числа копий согласно некоторым вариантам реализации. В процессе 100, проиллюстрированном на фиг. 1, для определения ВЧК применяют перекрытие метки последовательности, основанное на количестве меток последовательности (т.е. на подсчитанном значении метки последовательности). Однако аналогично описанию выше для определения NCV, вместо перекрытия можно применять другие переменные или параметры, такие как размер, соотношение размера и уровень метилирования. Согласно некоторым вариантам реализации для определения ВЧК объединяют две или более переменных. Более того, перекрытие и другие параметры можно взвесить по размеру фрагментов, из которых были получены метки. Для удобства чтения в процессе 100, проиллюстрированном на фиг. 1, упомянуто исключительно перекрытие, но следует отметить, что вместо перекрытия можно применять другие параметры, такие как размер, соотношение размера и уровень метилирования, подсчитанное значение, взвешенное по размеру, и т.д.

В операциях 130 и 135 определяют перекрытия квалификационной метки последовательности (или значения другого параметра) и перекрытия метки исследуемой последовательности (или значения другого параметра). В настоящем изобретении предложены процессы для определения количеств перекрытия, которые обеспечивают улучшенную чувствительность и селективность по сравнению с общепринятыми способами. Операции 130 и 135 отмечены звездочками и выделены рамками с жирными линиями для указания на то, что данные операции способствуют улучшению по сравнению с предшествующим уровнем техники. Согласно некоторым вариантам реализации количества перекрытия метки последовательности нормируют, подгоняют, цензурируют и иным способом процессируют для улучшения чувствительности и селективности анализа. Данные процессы дополнительно описаны в настоящем документе в другом месте.

С точки зрения общего представления в способе при определении ВЧК исследуемых образцов применяют нормирующие последовательности квалификационных обучающих образцов. Согласно некоторым вариантам реализации квалификационные обучающие образцы являются непораженными и характеризуются нормальным числом копий. Нормирующие последовательности обеспечивают механизм для нормирования измерений с целью определения вариабельностей в одной серии определений и в нескольких сериях определений. Нормирующие последовательности идентифицируют с применением информации о последовательности из множества квалификационных образцов, полученных от субъектов, которые установленно содержат клетки, характеризующиеся нормальным числом копий любой одной последовательности, представляющей интерес, например, хромосомы или ее сегмента. Определение нормирующих последовательностей представлено на этапах 110, 120, 130, 145 и 146 варианта реализации способа, изображенного на фиг. 1. Согласно некоторым вариантам реализации нормирующие последовательности применяют для вычисления дозы последовательности для исследуемых последовательностей. См. этап 150. Согласно некоторым вариантам реализации нормирующие последовательности также применяют для вычисления порога, с которым сравнивают дозу последовательности исследуемых последовательностей. См. этап 150. Информацию о последовательности, полученную из нормирующей последовательности и исследуемой последовательности, применяют для определения статистически значимой идентификации анеуплоидий хромосом в исследуемых образцах (этап 160).

Переходя к деталям способа определения присутствия вариации числа копий, согласно некоторым вариантам реализации на фиг. 1 представлена блок-схема 100 варианта реализации для определения ВЧК последовательности, представляющей интерес, например, хромосомы или ее сегмента, в биологическом образце. Согласно некоторым вариантам реализации биологический образец получен от субъекта и содержит смесь нуклеиновых кислот, полученных из различных геномов. Различные геномы могут быть внесены в образец двумя индивидуумами, например различные геномы были внесены плодом и матерью, вынашивающей плод. Также различные геномы могут быть внесены в образец тремя или более индивидуумами, например, различные геномы были внесены двумя или более плодами и матерью, вынашивающей плоды. В качестве альтернативы, геномы внесены в образец анеуплоидными раковыми клетками и нормальными эуплоидными клетками от того же субъекта, например, образец плазмы от пациента, страдающего от рака.

Помимо анализа исследуемого образца от пациента для каждой возможной хромосомы, представляющей интерес, выбирают одну или более нормирующих хромосом или один или более сегментов нормирующих хромосом. Нормирующие хромосомы или сегменты идентифицируют асинхронно из нормального исследования образцов от пациента, которое может проходить в клинических условиях. Другими словами, нормирующие хромосомы или сегменты идентифицируют перед исследованием образцов от пациента. Взаимосвязи между нормирующими хромосомами или сегментами и хромосомами или сегментами, представляющими интерес, сохраняют для использования в ходе анализа. Как объяснено ниже, такая взаимосвязь, как правило, сохраняется в течение периодов времени, которые охватывают исследование многих образцов. Следующее обсуждение затрагивает варианты реализации для выбора нормирующих хромосом или сегментов хромосомы для индивидуальных хромосом или сегментов, представ-

ляющих интерес.

Множество квалификационных образцов получают для идентификации квалификационных нормирующих последовательностей и для обеспечения значений дисперсии с целью применения при определении статистически значимой идентификации ВЧК в исследуемых образцах. На этапе 110 множество биологических квалификационных образцов получают из множества субъектов, которые установленно содержат клетки, характеризующиеся нормальным числом копий любой одной последовательности, представляющей интерес. Согласно одному варианту реализации квалификационные образцы получают от матерей, беременных плодом, который, как было подтверждено с применением цитогенетических способов, характеризуется нормальным числом копий хромосом. Биологические квалификационные образцы могут представлять собой биологическую жидкость, например плазму, или любой подходящий образец, описанный ниже. Согласно некоторым вариантам реализации квалификационный образец содержит смесь молекул нуклеиновой кислоты, например, молекул сцДНК. Согласно некоторым вариантам реализации квалификационный образец представляет собой образец материнской плазмы, который содержит смесь плодных и материнских молекул сцДНК. Информацию о последовательности для нормирующих хромосом и/или их сегментов получают в результате секвенирования по меньшей мере части нуклеиновых кислот, например, нуклеиновых кислот плода и матери, с применением любого известного способа секвенирования. Предпочтительно, в отношении последовательности нуклеиновых кислот плода и матери в качестве отдельных или клонально амплифицированных молекул применяют любой из способов секвенирования нового поколения (СНП), описанных в настоящем документе в другом месте. Согласно различным вариантам реализации квалификационные образцы процессируют, как раскрыто ниже, перед и в течение секвенирования. Квалификационные образцы можно процессировать с применением аппарата, системы и наборов, раскрытых в настоящем документе.

На этапе 120 по меньшей мере часть каждой из всех квалификационных нуклеиновых кислот, содержащихся в квалификационных образцах, секвенируют для получения миллионов ридов последовательности, например, ридов 36 п.о., которые выравнивают с референсным геномом, например, hg18. Согласно некоторым вариантам реализации риды последовательности содержат приблизительно 20 п.о., приблизительно 25 п.о., приблизительно 30 п.о., приблизительно 35 п.о., приблизительно 40 п.о., приблизительно 45 п.о., приблизительно 50 п.о., приблизительно 55 п.о., приблизительно 60 п.о., приблизительно 65 п.о., приблизительно 70 п.о., приблизительно 75 п.о., приблизительно 80 п.о., приблизительно 85 п.о., приблизительно 90 п.о., приблизительно 95 п.о., приблизительно 100 п.о., приблизительно 110 п.о., приблизительно 120 п.о., приблизительно 130, приблизительно 140 п.о., приблизительно 150 п.о., приблизительно 200 п.о., приблизительно 250 п.о., приблизительно 300 п.о., приблизительно 350 п.о., приблизительно 400 п.о., приблизительно 450 п.о. или приблизительно 500 п.о. Ожидают, что технологические преимущества сделают возможным получение ридов одиночных концов длиной более 500 п.о., что сделает возможным получение риды длиной более приблизительно 1000 п.о., когда получают риды спаренных концов. Согласно одному варианту реализации картированные риды последовательности содержат 36 п.о. Согласно другому варианту реализации картированные риды последовательности содержат 25 п.о.

Риды последовательности выравнивают с референсным геномом, и риды, которые уникально картируются на референсный геном, известны как метки последовательности. Метки последовательности, попадающие на маскированные сегменты маскированной референсной последовательности, не подсчитывают для анализа ВЧК.

Согласно одному варианту реализации по меньшей мере приблизительно 3×10^6 квалификационных меток последовательности, по меньшей мере приблизительно 5×10^6 квалификационных меток последовательности, по меньшей мере приблизительно 8×10^6 квалификационных меток последовательности, по меньшей мере приблизительно 10×10^6 квалификационных меток последовательности, по меньшей мере приблизительно 15×10^6 квалификационных меток последовательности, по меньшей мере приблизительно 20×10^6 квалификационных меток последовательности, по меньшей мере приблизительно 30×10^6 квалификационных меток последовательности, по меньшей мере приблизительно 40×10^6 квалификационных меток последовательности или по меньшей мере приблизительно 50×10^6 квалификационных меток последовательности, содержащих риды длиной от 20 до 40 п.о., получают из риды, который уникально картируется на референсном геноме.

На этапе 130 все метки, полученные из секвенирования нуклеиновых кислот в квалификационных образцах, подсчитывают для получения перекрытия метки квалификационной последовательности. Аналогично, в операции 135 все метки, полученные из исследуемого образца, подсчитывают для получения перекрытия метки исследуемой последовательности. В настоящем изобретении предложены процессы для определения количеств перекрытия, которые обеспечивают улучшенные чувствительность и селективность по сравнению с общепринятыми способами. Операции 130 и 135 отмечены звездочками и выделены рамками с жирными линиями для указания на то, что данные операции способствуют улучшению по сравнению с предшествующим уровнем техники. Согласно некоторым вариантам реализации количества перекрытия метки последовательности нормируют, подгоняют, цензурируют и иным способом про-

цессируют для улучшения чувствительности и селективности анализа. Данные процессы дополнительно описаны в настоящем документе в другом месте.

После того как все квалификационные метки последовательности были картированы и подсчитаны в каждом из квалификационных образцов, определяют перекрытие метки последовательности для последовательности, представляющей интерес, например клинически значимой последовательности, в квалификационных образцах, равно как и перекрытия метки последовательности для дополнительных последовательностей, из которых затем идентифицируют нормирующие последовательности.

Согласно некоторым вариантам реализации последовательность, представляющая интерес, представляет собой хромосому, которая связана с полной анеуплоидией хромосом, например, хромосому 21, и квалификационная нормирующая последовательность представляет собой полную хромосому, которая не связана с анеуплоидией хромосом и вариация которой в перекрытии метки последовательности приблизительно равна таковой последовательности (т.е. хромосомы), представляющей интерес, например хромосомы 21. Выбранная нормирующая хромосома или хромосомы могут представлять собой одну хромосому или группу хромосом, которые наилучшим способом приблизительно равны вариации в перекрытии метки последовательности для последовательности, представляющей интерес. Любая одна или более из хромосом 1-22, X и Y может представлять собой последовательность, представляющую интерес, и одну или более хромосом можно идентифицировать как нормирующую последовательность для каждой из любой из хромосом 1-22, X и Y в квалификационных образцах. Нормирующая хромосома может представлять собой отдельную хромосому или может представлять собой группу хромосом, как описано в настоящем документе в другом месте.

Согласно другому варианту реализации последовательность, представляющая интерес, представляет собой сегмент хромосомы, связанный с частичной анеуплоидией, например делецией или вставкой хромосомы, или несбалансированной хромосомной транслокацией, и нормирующая последовательность представляет собой сегмент хромосомы (или группу сегментов), который не связан с частичной анеуплоидией и вариация которого в перекрытии метки последовательности приблизительно равна таковой сегмента хромосомы, связанного с частичной анеуплоидией. Выбранный сегмент или сегменты нормирующей хромосомы могут представлять собой один или более сегментов, которые наилучшим способом приблизительно равны вариации в перекрытии метки последовательности для последовательности, представляющей интерес. Любой один или более сегментов любой одной или более из хромосом 1-22, X и Y может представлять собой последовательность, представляющую интерес.

Согласно другим вариантам реализации последовательность, представляющая интерес, представляет собой сегмент хромосомы, связанный с частичной анеуплоидией, и нормирующая последовательность представляет собой целую хромосому или хромосомы. Согласно еще одним вариантам реализации последовательность, представляющая интерес, представляет собой целую хромосому, связанную с анеуплоидией, и нормирующая последовательность представляет собой сегмент или сегменты хромосомы, которые не связаны с анеуплоидией.

Независимо от того, одну последовательность или группу последовательностей идентифицируют в квалификационных образцах как нормирующую последовательность или последовательности для любой одной или более последовательностей, представляющих интерес, можно выбрать квалификационную нормирующую последовательность так, чтобы она характеризовалась вариацией перекрытия метки последовательности или параметра размера фрагмента, которая наилучшим или наиболее эффективным способом приблизительно равна таковой последовательности, представляющей интерес, как определено в квалификационных образцах. Например, квалификационная нормирующая последовательность представляет собой последовательность, которая вызывает наименьшую вариабельность среди квалификационных образцов при применении для нормирования последовательности, представляющей интерес, т.е. вариабельность нормирующей последовательности является наиболее близкой к таковой последовательности, представляющей интерес, определенной в квалификационных образцах. Говоря иначе, квалификационная нормирующая последовательность представляет собой последовательность, выбранную для образования наименьшей вариации в дозе последовательности (для последовательности, представляющей интерес) в пределах квалификационных образцов. Таким образом, в процессе выбирают последовательность, которая при применении в качестве нормирующей хромосомы, как ожидают, образует наименьшую вариабельность в дозе хромосомы от серии к серии для последовательности, представляющей интерес.

Нормирующая последовательность, идентифицированная в квалификационных образцах для любой одной или более последовательностей, представляющих интерес, остается нормирующей последовательностью, предпочтительной для определения присутствия или отсутствия анеуплоидии в исследуемых образцах в течение дней, недель, месяцев и, возможно, лет, при условии, что процедуры, необходимые для получения библиотек секвенирования и секвенирования образцов, по существу не меняются с течением времени. Как описано выше, нормирующие последовательности для определения присутствия анеуплоидий выбирают для обеспечения (возможно, также среди других причин) вариабельности количества меток последовательности или значений параметра размера фрагмента, которые картируются на последовательность среди образцов, например различных образцов, и серий секвенирования, например се-

рий секвенирования, которые происходят в один и тот же день и/или в различные дни, которые наилучшим способом приблизительно равны вариабельности последовательности, представляющей интерес, для которой ее применяют в качестве нормирующего параметра. Существенные изменения в данных процедурах будут влиять на количество меток, которые картируются на все последовательности, что, в свою очередь, будет определять, какая одна последовательность или группа последовательностей будет характеризоваться вариабельностью среди образцов в одной и той же и/или в различных сериях секвенирования, в тот же день или в различные дни, которая наиболее точно приблизительно равна таковой последовательности или последовательностей, представляющих интерес, для чего будет требоваться, чтобы множество нормирующих последовательностей было повторно определено. Существенные изменения в процедурах включают изменения в лабораторном протоколе, применяемом для получения библиотеки секвенирования, которые включают изменения в отношении получения образцов для мультиплексного секвенирования вместо синглплексного секвенирования и изменения платформ секвенирования, которые включают изменения в химии, применяемой для секвенирования.

Согласно некоторым вариантам реализации нормирующая последовательность, выбранная для нормирования конкретной последовательности, представляющей интерес, представляет собой последовательность, которая наилучшим способом позволяет отличить один или более квалификационных образцов от одного или более пораженных образцов, что подразумевает, что нормирующая последовательность представляет собой последовательность, которая характеризуется наивысшей дифференцируемостью, т.е. дифференцируемость нормирующей последовательности является таковой, что она обеспечивает оптимальное установление отличии последовательности, представляющей интерес, в пораженном исследуемом образце, чтобы с легкостью отличить пораженный исследуемый образец от других непораженных образцов. Согласно другим вариантам реализации нормирующая последовательность представляет собой последовательность, которая характеризуется комбинацией наименьшей вариабельности и наибольшей дифференцируемости.

Уровень дифференцируемости можно определить как статистическое различие между дозами последовательности, например дозами хромосомы или дозами сегмента, в популяции квалификационных образцов и дозой или дозами хромосом в одном или более исследуемых образцах, как описано ниже и показано в примерах. Например, дифференцируемость можно численно представить как значение t-критерия, которое представляет статистическое различие между дозами хромосомы в популяции квалификационных образцов и дозой или дозами хромосомы в одном или более исследуемых образцах. Аналогично, дифференцируемость может быть основана на дозах сегмента вместо доз хромосом. В качестве альтернативы, дифференцируемость можно представить численно как нормированное значение хромосомы (NCV), которое представляет собой z-показатель для доз хромосомы, при условии, что распределение для NCV является нормальным. Аналогично, в случае, когда сегменты хромосомы являются последовательностями, представляющими интерес, дифференцируемость доз сегмента можно численно представить как нормированное значение сегмента (NSV), которое представляет собой z-показатель для доз сегмента хромосомы при условии, что распределение для NSV является нормальным. При определении z-показателя можно применять среднее значение и стандартное отклонение доз хромосомы или сегмента во множестве квалификационных образцов. В качестве альтернативы, можно применять среднее значение и стандартное отклонение доз хромосомы или сегмента в обучающем множестве, содержащем квалификационные образцы и пораженные образцы. Согласно другим вариантам реализации нормирующая последовательность представляет собой последовательность, которая характеризуется наименьшей вариабельностью и наивысшей дифференцируемостью или оптимальной комбинацией низкой вариабельности и высокой дифференцируемости.

Способ идентифицирует последовательности, которые по своей природе обладают аналогичными характеристиками и которые склонны к аналогичным вариациям среди образцов и серий секвенирования, и которые являются подходящими для определения доз последовательности в исследуемых образцах.

Определение доз последовательности.

Согласно некоторым вариантам реализации дозы хромосомы или сегмента для одной или более хромосом или сегментов, представляющих интерес, определяют во всех квалификационных образцах, как описано на этапе 146, представленном на фиг. 1, и последовательность нормирующей хромосомы или сегмента идентифицируют на этапе 145. Некоторые нормирующие последовательности предложены до того, как вычисляют дозы последовательности. Затем одну или более нормирующих последовательностей идентифицируют согласно различным критериям, которые дополнительно описаны ниже, см. этап 145. Согласно некоторым вариантам реализации, например, идентифицированная нормирующая последовательность приводит к наименьшей вариабельности дозы последовательности для последовательности, представляющей интерес, среди всех квалификационных образцов.

На этапе 146 на основании вычисленных плотностей квалификационной метки определяют дозу квалификационной последовательности, т.е. дозу хромосомы или дозу сегмента, для последовательности, представляющей интерес, в виде соотношения перекрытия метки последовательности для последовательности, представляющей интерес, и перекрытия метки квалификационной последовательности для дополнительных последовательностей, из которых затем на этапе 145 идентифицируют нормирующие

последовательности. После этого идентифицированные нормирующие последовательности применяют для определения доз последовательности в исследуемых образцах.

Согласно одному варианту реализации доза последовательности в квалификационных образцах представляет собой дозу хромосомы, которую вычисляют как соотношение количества меток последовательности или параметра размера фрагмента для хромосомы, представляющей интерес, и количества меток последовательности для последовательности нормирующей хромосомы в квалификационном образце. Последовательность нормирующей хромосомы может представлять собой одну хромосому, группу хромосом, сегмент одной хромосомы или группу сегментов от различных хромосом. Соответственно, дозу хромосомы для хромосомы, представляющей интерес, определяют в квалификационном образце как соотношение количества меток для хромосомы, представляющей интерес, и количества меток для (i) последовательности нормирующей хромосомы, состоящей из одной хромосомы, (ii) последовательности нормирующей хромосомы, состоящей из двух или более хромосом, (iii) последовательности нормирующего сегмента, состоящей из одного сегмента хромосомы, (iv) последовательности нормирующего сегмента, состоящей из двух или более сегментов из одной хромосомы, или (v) последовательности нормирующего сегмента, состоящей из двух или более сегментов двух или более хромосом. Примеры для определения дозы хромосомы для хромосомы 21, представляющей интерес, согласно (i)-(v) являются следующими: дозы хромосом для хромосомы, представляющей интерес, например, хромосомы 21, определяют как соотношение перекрытия метки последовательности хромосомы 21 и одного из следующих перекрытий меток последовательности: (i) каждая из всех оставшихся хромосом, т.е. хромосом 1 -20, хромосомы 22, хромосомы X и хромосомы Y; (ii) все возможные комбинации двух или более оставшихся хромосом; (iii) сегмент другой хромосомы, например, хромосомы 9; (iv) два сегмента другой хромосомы, например, два сегмента хромосомы 9; (v) два сегмента двух различных хромосом, например, сегмент хромосомы 9 и сегмент хромосомы 14.

Согласно другому варианту реализации доза последовательности в квалификационных образцах представляет собой дозу сегмента вместо дозы хромосомы, причем дозу сегмента вычисляют как соотношение количества меток последовательности для сегмента, представляющего интерес, который не представляет собой целую хромосому, и количества меток последовательности для последовательности нормирующего сегмента в квалификационном образце. Последовательность нормирующего сегмента может представлять собой любую из последовательностей нормирующей хромосомы или сегмента, которые обсуждаются выше.

Идентификация нормирующих последовательностей.

На этапе 145 идентифицируют нормирующую последовательность для последовательности, представляющей интерес. Согласно некоторым вариантам реализации, например, нормирующая последовательность представляет собой последовательность на основании вычисленных доз последовательности, например, которая приводит к наименьшей вариабельности дозы последовательности для последовательности, представляющей интерес, среди всех квалификационных обучающих образцов. Способ идентифицирует последовательности, которые по своей природе обладают аналогичными характеристиками и склонны к аналогичным вариациям среди образцов и серий секвенирования, и которые являются подходящими для определения доз последовательности в исследуемых образцах.

Нормирующие последовательности для одной или более последовательностей, представляющих интерес, можно идентифицировать во множестве квалификационных образцов, и затем последовательности, которые идентифицированы в квалификационных образцах, применяют для вычисления доз последовательностей для одной или более последовательностей, представляющих интерес, в каждом из исследуемых образцов (этап 150) для определения присутствия или отсутствия анеуплоидии в каждом из исследуемых образцов. Нормирующая последовательность, идентифицированная для хромосом или сегментов, представляющих интерес, может отличаться, если применяют различные платформы секвенирования, и/или если существуют отличия в очистке нуклеиновой кислоты, которая подлежит секвенированию и/или получению библиотеки секвенирования. Применение нормирующих последовательностей согласно способам, описанным в настоящем документе, обеспечивает специфичный и чувствительный критерий вариации числа копий хромосомы или ее сегмента независимо от получения образца и/или платформы секвенирования, которую применяют.

Согласно некоторым вариантам реализации идентифицируют более одной нормирующей последовательности, т.е. для одной последовательности, представляющей интерес, можно определить различные нормирующие последовательности, и для одной последовательности, представляющей интерес, можно определить несколько доз последовательности. Например, вариация, например, коэффициент вариации (КВ=стандартное отклонение/среднее значение) дозы хромосомы для хромосомы 21, представляющей интерес, является наименьшей, когда применяют перекрытие метки последовательности хромосомы 14. Однако можно идентифицировать две, три, четыре, пять, шесть, семь, восемь или более нормирующих последовательностей для применения при определении дозы последовательности для последовательности, представляющей интерес, в исследуемом образце. В качестве примера, вторую дозу для хромосомы 21 в любом исследуемом образце можно определить с применением хромосомы 7, хромосомы 9, хромосомы 11 или хромосомы 12 в качестве последовательности нормирующей хромосомы, поскольку все

данные хромосомы характеризуются КВ, близким к таковому для хромосомы 14.

Согласно некоторым вариантам реализации, когда одну хромосому выбрали в качестве последовательности нормирующей хромосомы для хромосомы, представляющей интерес, последовательность нормирующей хромосомы будет представлять собой хромосому, которая приводит к получению доз хромосомы для хромосомы, представляющей интерес, которые характеризуются наименьшей вариабельностью среди всех исследованных образцов, например, квалификационных образцов. В некоторых случаях наилучшая нормирующая хромосома может не характеризоваться наименьшей вариацией, но может характеризоваться распределением квалификационных доз, которое наилучшим способом позволяет отличить исследуемый образец или образцы от квалификационных образцов, т.е. наилучшая нормирующая хромосома может не характеризоваться наименьшей вариацией, но может характеризоваться наибольшей дифференцируемостью.

Согласно некоторым вариантам реализации нормирующие последовательности включают одну или более последовательностей устойчивых аутосом или их сегментов. Согласно некоторым вариантам реализации устойчивые аутосомы включают все аутосомы, за исключением хромосомы или хромосом, представляющих интерес. Согласно некоторым вариантам реализации устойчивые аутосомы включают все аутосомы, за исключением хромосом X, Y, 13, 18 и 21. Согласно некоторым вариантам реализации устойчивые аутосомы включают все аутосомы, за исключением таковых, определенных из образца, который отклоняется от нормального диплоидного состояния, и который может являться подходящим при определении геномов рака, характеризующихся аномальным числом копий по сравнению с нормальным диплоидным геномом.

Определение анеуплоидий в исследуемых образцах.

На основании идентификации нормирующей последовательности или последовательностей в квалификационных образцах определяют дозу последовательности для последовательности, представляющей интерес, в исследуемом образце, содержащем смесь нуклеиновых кислот, полученных из геномов, которые отличаются одной или более последовательностями, представляющими интерес.

На этапе 115 исследуемый образец получают от субъекта, который, как подозревают или как известно, несет клинически значимые ВЧК последовательности, представляющей интерес. Исследуемый образец может представлять собой биологическую жидкость, например, плазму, или любой подходящий образец, как описано ниже. Как объяснено, образец можно получить с применением неинвазивной процедуры, такой как простой забор крови. Согласно некоторым вариантам реализации исследуемый образец содержит смесь молекул нуклеиновой кислоты, например, молекул сцДНК. Согласно некоторым вариантам реализации исследуемый образец представляет собой образец материнской плазмы, который содержит смесь молекул сцДНК плода и матери.

На этапе 125 по меньшей мере часть исследуемых нуклеиновых кислот в исследуемом образце секвенируют, как описано для квалификационных образцов, с целью получения миллионов ридов последовательности, например, ридов длиной 36 п.о. Согласно различным вариантам реализации риды спаренных концов 2×36 п.о. применяют для секвенирования спаренных концов. Как на этапе 120, риды, полученные с помощью секвенирования нуклеиновых кислот в исследуемом образце, уникально картируют или выравнивают с референсным геномом для получения меток. Как описано на этапе 120, риды по меньшей мере приблизительно 3×10^6 квалификационных меток последовательности, по меньшей мере приблизительно 5×10^6 квалификационных меток последовательности, по меньшей мере приблизительно 8×10^6 квалификационных меток последовательности, по меньшей мере приблизительно 10×10^6 квалификационных меток последовательности, по меньшей мере приблизительно 15×10^6 квалификационных меток последовательности, по меньшей мере приблизительно 20×10^6 квалификационных меток последовательности, по меньшей мере приблизительно 30×10^6 квалификационных меток последовательности, по меньшей мере приблизительно 40×10^6 квалификационных меток последовательности или по меньшей мере приблизительно 50×10^6 квалификационных меток последовательности, содержащие от 20 до 40 п.о., получают из ридов, которые уникально картируются на референсный геном. Согласно определенным вариантам реализации риды, образованные с помощью аппарата секвенирования, предложены в электронном формате. Выравнивание осуществляют с применением компьютерного аппарата, как обсуждается ниже. Отдельные риды сравнивают с референсным геномом, который часто является обширным (миллионы пар оснований), для идентификации сайтов, в которых риды уникально соответствуют референсному геному. Согласно некоторым вариантам реализации процедура выравнивания обеспечивает ограниченное несоответствие между ридами и референсным геномом. В некоторых случаях допускается, что 1, 2 или 3 пары оснований в риде не соответствуют соответствующим парам оснований в референсном геноме, и при этом все равно проводят картирование.

На этапе 135 все или большинство меток, полученных в результате секвенирования нуклеиновых кислот в исследуемых образцах, подсчитывают для определения перекрытия метки исследуемой последовательности с применением компьютерного аппарата, как описано ниже. Согласно некоторым вариантам реализации каждый рид выравнивают с конкретной областью референсного генома (в большинстве случаев, хромосомой или сегментом), и рид преобразуют в метку посредством добавления к риду ин-

формации о сайте. По мере того как протекает данный процесс, компьютерный аппарат может проводить непрерывный вычисление количества картирования меток/ридов на каждую область референсного генома (в большинстве случаев, хромосому или сегмент). Подсчитанные значения хранят для каждой хромосомы или сегмента, представляющих интерес, и для каждой соответствующей нормирующей хромосомы или сегмента.

Согласно определенным вариантам реализации референсный геном содержит одну или более исключенных областей, которые являются частью истинного биологического генома, но не включены в референсный геном. Риды, потенциально выравнивающиеся с данными исключенными областями, не подсчитывают. Примеры исключенных областей включают области длинных повторяющихся последовательностей, области подобия между X- и Y-хромосомами и т.д. С применением маскированной референсной последовательности, полученной с помощью методик маскирования, описанных выше, для анализа ВЧК учитывают исключительно метки на немаскированных сегментах референсной последовательности.

Согласно некоторым вариантам реализации способ определяет, следует ли подсчитывать метку более одного раза при выравнивании множественных ридов с одним и тем же сайтом на референсном геноме или последовательности. Существуют случаи, когда две метки содержат одну и ту же последовательность, и вследствие этого выравниваются с идентичным сайтом на референсной последовательности. Способ, применяемый для вычисления меток, может при определенных обстоятельствах исключать из подсчета идентичные метки, полученные из одного и того же секвенированного образца. Если в данном образце непропорциональное количество меток является идентичным, это свидетельствует о том, что существует значительная погрешность или другой дефект процедуры. Вследствие этого согласно определенным вариантам реализации в способе подсчета не учитывают метки из данного образца, идентичные меткам из образца, которые были подсчитаны ранее.

Для выбора ситуации, когда следует пренебречь идентичной меткой из одного образца, можно задать различные критерии. Согласно определенным вариантам реализации заданный процент меток, которые подсчитывают, должен являться уникальным. Если большее, чем данный порог, число меток не являются уникальными, ими пренебрегают. Например, если заданный процент требует, чтобы по меньшей мере 50% являлись уникальными, идентичные метки не подсчитывают до тех пор, пока процент уникальных меток не превысит 50% для образца. Согласно другим вариантам реализации пороговое количество уникальных меток составляет по меньшей мере приблизительно 60%. Согласно другим вариантам реализации пороговый процент уникальных меток составляет по меньшей мере приблизительно 75%, или по меньшей мере приблизительно 90%, или по меньшей мере приблизительно 95%, или по меньшей мере приблизительно 98%, или по меньшей мере приблизительно 99%. Порог может быть задан на уровне 90% для хромосомы 21. Если 30М меток выравниваются с хромосомой 21, тогда по меньшей мере 27М из них должны быть уникальными. Если 3М подсчитанных меток не являются уникальными, и первая после 30 миллионов метка не являются уникальной, ее не подсчитывают. Выбор конкретного порога или другого критерия, используемого для определения ситуации, когда следует пренебречь подсчетом следующих идентичных меток, можно осуществить с применением соответствующего статистического анализа. Одним из факторов, влияющих на данный порог или другой критерий, является относительное количество секвенированного образца по отношению к размеру генома, с которым можно выровнять метки. Другие факторы включают размер ридов и аналогичные соображения.

Согласно одному варианту реализации количество меток исследуемой последовательности, картированных на последовательность, представляющую интерес, нормируют к известной длине последовательности, представляющей интерес, на которую они картируются, для получения соотношения плотности метки исследуемой последовательности. Как описано для квалификационных образцов, нормирование к известной длине последовательности, представляющей интерес, не требуется, и может быть включено как этап для снижения количества цифр в числе для упрощения интерпретации человеком. После того как в исследуемом образце подсчитывают все картированные метки исследуемой последовательности, определяют перекрытие метки последовательности для последовательности, представляющей интерес, например, клинически значимой последовательности, в исследуемых образцах, равно как и перекрытия метки последовательности для дополнительных последовательностей, которые соответствуют по меньшей мере одной нормирующей последовательности, идентифицированной в квалификационных образцах.

На этапе 150 на основании идентичности по меньшей мере одной нормирующей последовательности в квалификационных образцах определяют дозу исследуемой последовательности для последовательности, представляющей интерес, в исследуемом образце. Согласно различным вариантам реализации дозу исследуемой последовательности определяют компьютерным способом с применением перекрытий метки последовательности для последовательности, представляющей интерес, и соответствующей нормирующей последовательности, как описано в настоящем документе. Компьютерный аппарат, служащий для данной процедуры, электронным способом оценивает взаимосвязь между последовательностью, представляющей интерес, и связанной с ней нормирующей последовательностью, которая может храниться в базе данных, таблице, графике или может быть включена как код в инструкции программы.

Как описано в настоящем документе в другом месте, по меньшей мере одна нормирующая последовательность может представлять собой одну последовательность или группу последовательностей. Доза последовательности для последовательности, представляющей интерес, в исследуемом образце представляет собой соотношение перекрытия метки последовательности, определенное для последовательности, представляющей интерес, в исследуемом образце, и перекрытия метки последовательности по меньшей мере одной нормирующей последовательности, определенной в исследуемом образце, причем нормирующая последовательность в исследуемом образце соответствует нормирующей последовательности, идентифицированной в квалификационных образцах для конкретной последовательности, представляющей интерес. Например, если нормирующая последовательность, идентифицированная в квалификационных образцах для хромосомы 21, как определено, является хромосомой, например, хромосомой 14, тогда дозу исследуемой последовательности для хромосомы 21 (последовательности, представляющей интерес) определяют в виде соотношения перекрытия метки последовательности для хромосомы 21 и перекрытия метки последовательности для хромосомы 14, каждое из которых определяют в исследуемом образце. Аналогично определяют дозы хромосом для хромосом 13, 18, X, Y и других хромосом, связанных с анеуплоидиями хромосом. Нормирующая последовательность для хромосомы, представляющей интерес, может представлять собой одну хромосому или группу хромосом, или один сегмент или группу сегментов хромосомы. Как описано ранее, последовательность, представляющая интерес, может представлять собой часть хромосомы, например, сегмент хромосомы. Соответственно, дозу для сегмента хромосомы можно определить в виде соотношения перекрытия метки последовательности, определенного для сегмента в исследуемом образце, и перекрытия метки последовательности для сегмента нормирующей хромосомы в исследуемом образце, причем нормирующий сегмент в исследуемом образце соответствует нормирующему сегменту (одному сегменту или группе сегментов), идентифицированному в квалификационных образцах для конкретного сегмента, представляющего интерес. Размер сегментов хромосомы может варьировать от килобаз (т.о.) до мегабаз (Мб) (например, приблизительно от 1 т.о. до 10 т.о., или приблизительно от 10 т.о. до 100 т.о., или приблизительно от 100 т.о. до 1 Мб).

На этапе 155 из значений стандартного отклонения, установленных для доз квалификационной последовательности, определенных во множестве квалификационных образцов, и доз последовательности, определенных для образцов, которые установленно являются анеуплоидными для последовательности, представляющей интерес, получают пороговые значения. Отметим, что данную операцию, как правило, осуществляют асинхронно с анализом исследуемых образцов от пациента. Данную операцию можно осуществить, например, одновременно с выбором нормирующих последовательностей из квалификационных образцов. Точная классификация зависит от различий между распределениями вероятностей для различных классов, т.е. типа анеуплоидии. В некоторых примерах пороги выбирают из эмпирического распределения для каждого типа анеуплоидии, например, трисомии 21. Возможные пороговые значения, которые были установлены для классификации анеуплоидии трисомии 13, трисомии 18, трисомии 21 и моносомии X, описаны в примерах, в которых описано применение способа для определения анеуплоидии хромосом посредством секвенирования сцДНК, экстрагированной из материнского образца, содержащего смесь нуклеиновых кислот плода и матери. Пороговое значение, которое определяют, чтобы отличить образцы, пораженные анеуплоидией хромосомы, может быть таким же или может отличаться от порога для другой анеуплоидии. Как показано в примерах, пороговое значение для каждой хромосомы, представляющей интерес, определяют из вариабельности дозы хромосомы, представляющей интерес, среди образцов и серий секвенирования. Чем менее вариабельна доза хромосомы для любой хромосомы, представляющей интерес, тем уже распространение дозы хромосомы, представляющей интерес, среди всех непораженных образцов, которые применяют, чтобы задать порог для определения различных анеуплоидий.

Возвращаясь к потоку процесса, связанного с классификацией исследуемого образца пациента, на этапе 160 определяют вариацию числа копий последовательности, представляющей интерес, в исследуемом образце посредством сравнения дозы исследуемой последовательности для последовательности, представляющей интерес, с по меньшей мере одним пороговым значением, установленным из доз квалификационной последовательности. Данную операцию можно осуществить с помощью того же компьютерного аппарата, применявшегося для измерения перекрытий метки последовательности, и/или вычисления доз сегмента.

На этапе 160 вычисленную дозу для исследуемой последовательности, представляющей интерес, сравнивают с таковой, заданной в качестве пороговых значений, которые выбраны согласно заданному пользователем "порогу надежности" для классификации образца как "нормального", "пораженного" или "решение отсутствует". Образцы "решение отсутствует" представляют собой образцы, для которых окончательный диагноз не может быть поставлен с надежностью. Каждый тип пораженного образца (например, трисомия 21, частичная трисомия 21, моносомия X) характеризуется своими собственными порогами, одним - для принятия решения о нормальных (непораженных) образцах и другим - для принятия решения о пораженных образцах (несмотря на то, что в некоторых случаях два порога совпадают). Как описано в настоящем документе в другом месте, в некоторых обстоятельствах результат "решение отсутствует" можно преобразовать в решение (пораженный или нормальный), если фракция нуклеиновой ки-

слоты плода в исследуемом образце является в достаточной степени высокой. Классификация исследуемой последовательности может сообщаться компьютерным аппаратом, применяемым в других операциях данного потока процесса. В некоторых случаях классификацию сообщают в электронном формате, и классификация может быть выведена на экран, отправлена по электронной почте, представлена в текстовом виде и т.д. заинтересованным лицам.

Согласно некоторым вариантам реализации определение ВЧК включает вычисление NCV или NSV, которые представляют собой отношение дозы хромосомы или сегмента к среднему значению соответствующей дозы хромосомы или сегмента во множестве квалификационных образцов, как описано выше. Затем можно определить ВЧК посредством сравнения NCV/NSV с определенным ранее пороговым значением для оценки числа копий.

Порог для оценки числа копий можно выбрать для оптимизации доли ложноположительных и ложноотрицательных результатов. Чем выше порог оценки числа копий, тем менее вероятно появление ложноположительных результатов. Аналогично, чем ниже порог, тем менее вероятно появление ложноотрицательных результатов. Таким образом, существует компромисс между первым идеальным порогом, выше которого классифицируют исключительно истинно положительные результаты, и вторым идеальным порогом, ниже которого классифицируют исключительно истинно отрицательные результаты.

Пороги задают, главным образом, в зависимости от вариабельности доз хромосом для конкретной хромосомы, представляющей интерес, которая определена во множестве непораженных образцов. Вариабельность зависит от большого числа факторов, включая фракцию кДНК плода, присутствующей в образце. Вариабельность (КВ) определяют по среднему значению или медиане и стандартному отклонению для доз хромосомы среди популяции непораженных образцов. Таким образом, в пороге (s) для классификации анеуплоидии используют NCV согласно уравнению:

$$\boldsymbol{NCV_{ij}} = \frac{x_{ij}-\hat{\mu}_j}{\hat{\sigma}_j},$$

где $\hat{\mu}_j$ и $\hat{\sigma}_j$ представляют собой вычисленное среднее значение и стандартное отклонение, соответственно, для дозы j-й хромосомы во множестве квалификационных образцов, и x_{ij} представляет собой наблюдаемую дозу j-й хромосомы для исследуемого образца i,

со связанной фракцией плода в виде

$$\text{ФЭ}_{ij} = \mathbf{2} \times \left|\frac{\boldsymbol{NCV}_{ij}\times\hat{\sigma}_j}{\hat{\mu}_j}\right| = \mathbf{2} \times \boldsymbol{NCV} \times \text{КВ}$$

Таким образом, для каждой NCV хромосомы, представляющей интерес, ожидаемая фракция плода, связанная с данным значением NCV, может быть вычислена по КВ на основании среднего значения и стандартного отклонения соотношения хромосомы для хромосомы, представляющей интерес, среди популяции непораженных образцов.

Затем на основании взаимосвязи между фракцией плода и значениями NCV можно выбрать границу принятия решения, выше которой образцы определяют как положительные (пораженные), на основании нормальных квантилей распределения. Как описано выше, согласно некоторым вариантам реализации задают порог для оптимального компромисса между обнаружением истинно положительных и долей ложноотрицательных результатов. А именно, выбирают порог для максимизации суммы истинно положительных и истинно отрицательных результатов или минимизации суммы ложноположительных и ложноотрицательных результатов.

В определенных вариантах реализации предложен способ обеспечения пренатальной диагностики анеуплоидии хромосомы плода в биологическом образце, содержащем молекулы нуклеиновой кислоты плода и матери. Диагноз ставят на основании получения информации о последовательности по меньшей мере из части смеси молекул нуклеиновой кислоты плода и матери, полученных из биологического исследуемого образца, например, образца материнской плазмы, компьютеризированного вычисления из данных секвенирования дозы нормирующей хромосомы для одной или более хромосом, представляющих интерес, и/или дозы нормирующего сегмента для одного или более сегментов, представляющих интерес, и определения статистически значимого различия между дозой хромосомы для хромосомы, представляющей интерес, и/или дозой сегмента для сегмента, представляющего интерес, соответственно, в исследуемом образце и пороговым значением, установленным во множестве квалификационных (нормальных) образцов, и обеспечения пренатальной диагностики на основании статистического различия.

Как описано на этапе 160 способа, ставят диагноз нормальных или пораженных образцов. Результат "решение отсутствует" предложен в случае, если диагноз нормальных или пораженных образцов не может быть поставлен с уверенностью.

Согласно некоторым вариантам реализации можно выбрать два порога. Первый порог выбирают для минимизации доли ложноположительных результатов, выше которого образцы будут классифицированы как "пораженные", и второй порог выбирают для минимизации доли ложноотрицательных результатов, ниже которого образцы будут классифицированы как "непораженные". Образцы с NCV выше второго порога, но ниже первого порога можно классифицировать как образцы "с подозрением на анеуплоидию" или "решение отсутствует", для которых присутствие или отсутствие анеуплоидии можно подтвердить независимыми способами. Область между первым и вторым порогами можно обозначить как

область "решение отсутствует".

Согласно некоторым вариантам реализации пороги подозрения и результата "решение отсутствует" представлены в табл. 1. Как видно, пороги NCV варьируют между различными хромосомами. Согласно некоторым вариантам реализации пороги варьируют в зависимости от ФЭ для образца, как объяснено выше. Методики порога, применяемые в настоящем документе, способствуют улучшению чувствительности и селективности согласно некоторым вариантам реализации.

Таблица 1. Пороги NCV подозрения и поражения, определяющие диапазоны результата "решение отсутствует"

	Подозрение	**Пораженные**
Хр. 13	3,5	4,0
Хр. 18	3,5	4,5
Хр. 21	3,5	4,0
Хр. X (XO, XXX)	4,0	4,0
Хр. Y (XX по сравнению с XY)	6,0	6,0

Анализы размера фрагмента и перекрытия последовательности.

Как упомянуто выше, для оценки ВЧК можно применять параметры размера фрагментов, а также перекрытие. Размер фрагмента для фрагмента бесклеточной нуклеиновой кислоты, например фрагмента сцДНК, можно получить посредством секвенирования спаренных концов, электрофореза (например, капиллярного электрофореза на основе микрочипов) и других способов, известных в данной области техники. На фиг. 2A тематически проиллюстрировано, как секвенирование спаренных концов можно применять для определения как размера фрагмента, так и перекрытия последовательности.

В верхней половине фиг. 2A представлена диаграмма фрагмента бесклеточной ДНК плода и фрагмента материнской бесклеточной ДНК, обеспечивающего матрицу для процесса секвенирования спаренных концов. Обычно длинные последовательности нуклеиновой кислоты фрагментируют на более короткие последовательности для рида в процессе секвенирования спаренных концов. Такие фрагменты также называют вставками. Фрагментация является нецелесообразной для бесклеточной ДНК, поскольку бесклеточная ДНК уже существуют в виде фрагментов, по большей части более коротких, чем 300 пар оснований. Было показано, что фрагменты бесклеточной ДНК плода в материнской плазме являются более длинными, чем фрагменты материнской бесклеточной ДНК. Как показано в верхней части фиг. 2A, бесклеточные ДНК плодного происхождения характеризуются средней длиной приблизительно 167 пар оснований, в то время как бесклеточные ДНК материнского происхождения характеризуются средней длиной приблизительно 175 пар оснований. При секвенировании спаренных концов на определенных платформах, таких как платформа Illumina для секвенирования посредством синтеза, как описано подробнее ниже по тексту, с двумя концами фрагмента лигируют адаптерные последовательности, индексные последовательности и/или праймерные последовательности (не представлено на фиг. 2A). Фрагмент сначала прочитывают в одном направлении, получая рид 1 с одного конца фрагмента. Затем начинают второй рид с противоположного конца фрагмента, получая последовательность рида 2. Соответствие между ридом 1 и ридом 2 можно идентифицировать посредством их координат в проточной ячейке. Затем рид 1 и рид 2 картируют на референсную последовательность в виде пары меток, которые находятся поблизости друг от друга, как показано в нижней половине фиг. 2A. Согласно некоторым вариантам реализации, если риды являются достаточно длинными, два рида могут перекрываться в средней части вставки. После того как пару выравнивают с референсной последовательностью, относительное расстояние между двумя ридами и длину фрагмента можно определить на основании положений двух ридов. Поскольку риды спаренных концов обеспечивают в два раза больше пар оснований, чем риды одиночных концов при той же длине рида, они способствуют улучшению качества выравнивания, в особенности для последовательностей со многими повторами или для неуникальных последовательностей. Согласно многим вариантам реализации референсную последовательность подразделяют на блоки, такие как блоки по 100 тысяч пар оснований. После того как риды спаренных концов выравнивают с референсной последовательностью, можно определить количество ридов, выровненных с блоком. Также для блока можно определить количество, а также длины вставок (например, фрагментов сцДНК). Согласно некоторым вариантам реализации, если вставка одновременно попадает в два блока, половины вставки можно отнести к каждому блокиу.

На фиг. 2B представлен вариант реализации, обеспечивающий процесс 220 для применения перекрытия на основании размера с целью определения вариации числа копий последовательности нуклеиновой кислоты, представляющей интерес, в исследуемом образце, содержащем фрагменты бесклеточной нуклеиновой кислоты, полученные из двух или более геномов. Как раскрыто в настоящем документе, параметр "смещен в сторону размера фрагмента или диапазона размера", когда: 1) параметр благоприятно взвешивается по размеру фрагмента или диапазону размера, например, вычисление имеет больший вес, когда связан с фрагментами размера или диапазона размера, чем для других размеров или диапазонов; или 2) параметр получен из значения, которое благоприятно взвешивается по размеру фрагмента

или диапазону размера, например соотношение получено из подсчета, который имеет больший вес, когда связан с фрагментами размера или диапазона размера. Размер фрагмента или диапазон размера может являться характеристикой генома или его части, когда геном образует фрагменты нуклеиновой кислоты, обогащенные или содержащие более высокую концентрацию размера или диапазона размера по сравнению с фрагментами нуклеиновой кислоты из другого генома или другой части того же генома.

Процесс 220 начинается с получения ридов последовательности, полученных в результате секвенирования фрагментов бесклеточной нуклеиновой кислоты в исследуемом образце. См. блок 222. Два или более геномов в исследуемом образце могут представлять собой геном беременной матери и геном плода, вынашиваемого беременной матерью. В других вариантах применения исследуемый образец включает бесклеточную ДНК из опухолевых клеток и непораженных клеток. Согласно некоторым вариантам реализации в связи с высоким соотношением сигнал/шум, обеспеченным перекрытием на основании размера, секвенирование фрагментов бесклеточной нуклеиновой кислоты проводят без необходимости в амплификации фрагментов нуклеиновой кислоты с применением ПЦР. Процесс 200 также включает выравнивание ридов последовательности фрагментов бесклеточной нуклеиновой кислоты с референсным геномом, который содержит последовательность, представляющую интерес, и разделен на множество блоков. Успешное выравнивание приводит к получению меток исследуемой последовательности, которые включают последовательность и ее расположение на референсной последовательности. См. блок 224. Затем процесс 220 продолжается определением размеров фрагментов бесклеточной нуклеиновой кислоты, существующих в исследуемом образце. Некоторые варианты реализации, в которых применяют секвенирование спаренных концов, обеспечивают длину вставки, связанной с меткой последовательности. См. блок 226. Термины "размер" и "длина" используются взаимозаменяемо, когда их используют применительно к последовательностям или фрагментам нуклеиновой кислоты. Согласно варианту реализации, проиллюстрированному в настоящем документе, процесс 220 также включает взвешивание меток исследуемой последовательности на основании размеров фрагментов бесклеточной нуклеиновой кислоты, из которых получают метки. См. блок 228. В настоящем документе "взвешивание" означает модификацию количества с применением одной или более переменных или функций. Одну или более переменных или функций считают "весом". Согласно многим вариантам реализации переменную умножают на вес. Согласно другим вариантам реализации переменную можно модифицировать экспоненциально или иным способом. Согласно некоторым вариантам реализации взвешивание меток исследуемой последовательности осуществляют посредством смещения перекрытий в сторону меток исследуемой последовательности, полученных из фрагментов бесклеточной нуклеиновой кислоты размера или диапазона размера, характерного для одного генома в исследуемом образце. Как раскрыто в настоящем документе, размер представляет собой характеристику генома, когда геном содержит обогащенную или более высокую концентрацию нуклеиновой кислоты указанного размера по сравнению с другим геномом или другой частью того же генома.

Согласно некоторым вариантам реализации функция взвешивания может представлять собой линейную или нелинейную функцию. Примеры применимых нелинейных функций включают, без ограничения, ступенчатые функции Хевисайда, функции вагона, ступенчатые функции или сигмоидальные функции. Согласно некоторым вариантам реализации используют функцию Хевисайда или функцию вагона, в результате чего метку в конкретном диапазоне размера умножают на вес 1, и метки за пределами диапазона умножают на вес 0. Согласно некоторым вариантам реализации фрагментам от 80 до 150 пар оснований присваивают вес 1, тогда как фрагментам за пределами данного диапазона присваивают вес 0. В данных примерах взвешивание является дискретным, представляя собой ноль или единицу в зависимости от того, попадает ли параметр всего значения в пределы или за пределы конкретного диапазона. В качестве альтернативы, вес вычисляют как непрерывную функцию размера фрагмента или другого аспекта связанного значения параметра.

Согласно некоторым вариантам реализации вес для фрагментов в одном диапазоне размера является положительным, и вес в другом диапазоне является отрицательным. Данный факт можно применять, чтобы способствовать усилению сигнала, когда направления различия между двумя геномами характеризуются противоположными знаками. Например, подсчитанные значения рида имеют вес 1 для вставки 80 - 150 пар оснований и вес -1 для вставки 160 - 200 пар оснований.

Вес может быть присвоен подсчетам, а также другим параметрам.

Например, взвешивание можно также применять в отношении дробных параметров или параметров соотношения, в которых используется размер фрагмента. Например, соотношение может присваивать фрагментам в определенных поддиапазонах больший вес, чем фрагментам и блокам другого размера.

Затем вычисляют перекрытия для блоков на основании взвешенных меток исследуемой последовательности. См. блок 230. Такие перекрытия считают смещенными в сторону размера. Как объяснено выше, значение смещено в сторону размера фрагмента или диапазона размера, если параметр благоприятно взвешивается по размеру фрагмента или диапазону размера. Процесс 200 также включает идентификацию вариации числа копий в последовательности, представляющей интерес, из вычисленных перекрытий. См. блок 232. Согласно некоторым вариантам реализации, как подробнее объяснено ниже по тексту применительно к фиг. 2C, 3A-3K и 4, перекрытия можно подогнать или откорректировать для удаления

шума в данных, и посредством этого увеличить соотношение сигнал/шум. В некоторых вариантах применения перекрытие на основании взвешенных меток, полученных в процессе 220, обеспечивает более высокую чувствительность и/или более высокую селективность по сравнению с невзвешенными перекрытиями при определении вариации числа копий. В некоторых вариантах применения пример рабочего процесса, предложенный ниже, может дополнительно улучшить чувствительность и селективность анализа ВЧК.

Пример рабочего процесса для анализа размера фрагмента и/или перекрытия последовательности.

В некоторых раскрытых вариантах реализации предложены способы определения количеств перекрытия последовательности с низким шумом и/или высоким сигналом, которые обеспечивают данные для определения различных генетических состояний, связанных с числом копий и ВЧК, с улучшенной чувствительностью, селективностью и/или эффективностью по сравнению с количествами перекрытия последовательности, полученными общепринятыми способами. Согласно определенным вариантам реализации последовательности из исследуемого образца процессируют для получения количеств перекрытия последовательности.

В процессе применяют определенную информацию, доступную из других источников. Согласно некоторым вариантам реализации всю данную информацию получают из обучающего множества образцов, которые установленно являются непораженными (например, не анеуплоидными). Согласно другим вариантам реализации некоторую часть или всю информацию получают от других исследуемых образцов, которые могут быть предложены "на ходу", поскольку в одном и том же процессе анализируют несколько образцов.

Согласно определенным вариантам реализации для снижения шума данных применяют маски последовательности. Согласно некоторым вариантам реализации как последовательность, представляющая интерес, так и ее нормирующие последовательности являются маскированными. Согласно некоторым вариантам реализации можно применять различные маски, когда рассматривают различные хромосомы или сегменты, представляющие интерес. Например, одну маску (или группу масок) можно применять, когда хромосома 13 представляет собой хромосому, представляющую интерес, и отличную маску (или группу масок) можно применять, когда хромосома 21 представляет собой хромосому, представляющую интерес. Согласно определенным вариантам реализации маски задают при разрешении блоков. Вследствие этого в одном примере разрешение маски составляет 100 т.о. Согласно некоторым вариантам реализации в отношении хромосомы Y можно применять отдельную маску. Маскированные области исключения могут быть предложены для хромосомы Y при более высоком разрешении (1 т.о.), чем для других хромосом, представляющих интерес, как описано в предварительной заявке на патент США № 61/836057, поданной 17 июня 2013 года [номер патентного реестра ARTEP008P]. Маски предложены в форме файлов, идентифицирующих исключенные геномные области.

Согласно определенным вариантам реализации для устранения межблоковой вариации в профиле последовательности, представляющей интерес, в процессе применяют ожидаемое значение нормированного перекрытия, причем вариация является неинформативной для определения ВЧК для исследуемого образца. Процесс подгоняет нормированные количества перекрытия в соответствии с ожидаемым значением нормированного перекрытия для каждого блока по всему геному или по меньшей мере для блоков устойчивых хромосом в референсном геноме (для применения в операции 317 ниже). В ходе данного процесса также можно улучшить параметры, отличные от перекрытия. Ожидаемое значение можно определить из обучающего множества непораженных образцов. В качестве примера, ожидаемое значение может представлять собой медианное значение в пределах образцов обучающего множества. Ожидаемые значения перекрытия образцов можно определить как количество уникальных не повторяющихся меток, выровненных с блоком, разделенное на суммарное количество уникальных не повторяющихся меток, выровненных со всеми блоками в устойчивых хромосомах референсного генома.

На фиг. 2C представлена структурная схема процесса 200 для определения параметра размера фрагмента для последовательности, представляющей интерес, причем параметр применяют для оценки числа копий последовательности, представляющей интерес, в исследуемом образце в блоке 214. Данный процесс устраняет систематическую вариацию, общую среди непораженных обучающих образцов, причем вариация увеличивает шум в анализе для оценки ВЧК. Данный процесс также устраняет погрешности GC, присущие исследуемому образцу, посредством этого увеличивая соотношение сигнал/шум в данных анализа. Следует отметить, что процесс 200 можно также применять в отношении перекрытия вне зависимости от того, смещено ли перекрытие в сторону размера или нет. Аналогично, процессы на фиг. 2D, 3 и 4 являются в равной степени применимыми к перекрытию, взвешенному по размеру фрагмента перекрытию, размеру фрагмента, фракции или соотношению фрагментов в заданном диапазоне размера, уровню метилирования фрагментов и т.д.

Процесс 200 начинается с обеспечения ридов последовательности исследуемого образца, как указано в блоке 202. Согласно некоторым вариантам реализации риды последовательности получают в результате секвенирования сегментов ДНК, полученных из крови беременной женщины, включая сцДНК матери и плода. Процесс продолжается выравниванием ридов последовательности с референсным геномом, содержащим последовательность, представляющую интерес, с обеспечением меток исследуемой

последовательности. Блок 204. Согласно некоторым вариантам реализации риды, которые выравниваются с более одним сайтом, исключают. Согласно некоторым вариантам реализации несколько ридов, которые выравниваются с одним и тем же сайтом, исключают или снижают до подсчета единичного рида. Согласно некоторым вариантам реализации риды, которые выравниваются с исключенными сайтами, также исключают. Вследствие этого согласно некоторым вариантам реализации для обеспечения подсчета неисключенных сайтов (подсчета NES) с целью определения перекрытия или других параметров каждого блока подсчитывают исключительно уникально выровненные не повторяющиеся метки, выровненные с неисключенными сайтами.

Процесс 200 обеспечивает размеры фрагментов бесклеточной нуклеиновой кислоты, существующих в исследуемом образце. Согласно некоторым вариантам реализации с применением секвенирования спаренных концов можно получить размер/длину вставки из расположений пары ридов на концах вставки. Для определения размера фрагмента можно применять другие методики. См. блок 205. Затем в блоках референсного генома, включая блоки в последовательности, представляющей интерес, процесс 200 определяет значения параметра размера фрагмента, смещенного в сторону характеристики размеров фрагмента одного из геномов. Термин "параметр размера фрагмента" означает параметр, который относится к размеру или длине фрагмента или совокупности фрагментов для фрагментов нуклеиновой кислоты; например, фрагментов сцДНК, полученных из физиологической жидкости. В настоящем документе параметр "смещен в сторону размера фрагмента или диапазона размера", когда:

1) параметр благоприятно взвешивается по размеру фрагмента или диапазону размера, например, вычисление имеет больший вес, когда связан с фрагментами размера или диапазона размера, чем для других размеров или диапазонов; или

2) параметр получен из значения, которое благоприятно взвешивается по размеру фрагмента или диапазону размера, например, соотношение получено из подсчета, который имеет больший вес, когда связан с фрагментами размера или диапазона размера.

Размер фрагмента или диапазон размера может являться характеристикой генома или его части, когда геном образует фрагменты нуклеиновой кислоты, обогащенные или содержащие более высокую концентрацию размера или диапазона размера, по сравнению с фрагментами нуклеиновой кислоты из другого генома или другой части того же генома.

Согласно некоторым вариантам реализации параметр размера фрагмента представляет собой взвешенный по размеру подсчет. Согласно некоторым вариантам реализации фрагмент весит 1 в диапазоне и 0 за пределами диапазона. Согласно другим вариантам реализации параметр размера фрагмента представляет собой фракцию или отношение фрагментов в диапазоне размера. См. блок 206. Согласно некоторым вариантам реализации значение параметра размера фрагмента (или перекрытие, как отмечено выше) каждого блока делят на значение параметра нормирующей последовательности в одном и том же образце, получая нормированный параметр.

После этого процесс 200 обеспечивает глобальный профиль последовательности, представляющей интерес. Глобальный профиль содержит значение ожидаемого параметра в каждом блоке, полученное из обучающего множества непораженных обучающих образцов. Блок 208. Процесс 200 устраняет вариацию, обычную для обучающего образца, посредством подгонки значений нормированного параметра меток исследуемой последовательности в соответствии с ожидаемыми значениями параметра для получения откорректированных с учетом глобального профиля значений параметра для последовательности, представляющей интерес. Блок 210. Согласно некоторым вариантам реализации ожидаемое значение параметра, полученное из обучающего множества, обеспеченного в блоке 208, представляет собой медиану среди обучающих образцов. Согласно некоторым вариантам реализации операция 2010 подгоняет нормированное значение параметра посредством вычитания ожидаемого значения параметра из нормированного значения параметра. Согласно другим вариантам реализации операция 210 делит нормированное значение параметра на ожидаемое значение параметра каждого блока для получения откорректированного с учетом глобального профиля значения параметра.

В дополнение к корректировке с учетом глобального профиля или вместо нее процесс 200 устраняет погрешности GC, присущие исследуемому образцу, посредством подгонки значения параметра. Как показано в блоке 212, процесс подгоняет откорректированное с учетом глобального профиля значение параметра на основании взаимосвязи между уровнем содержания GC и откорректированным с учетом глобального профиля перекрытием, существующем в исследуемом образце, посредством этого получая откорректированное с учетом GC в образце значение параметра размера фрагмента. После подгонки с учетом систематической вариации, обычной для непораженных обучающих образцов, и внутрисубъектных погрешностей GC процесс обеспечивает значение размера фрагмента, откорректированное с учетом глобального профиля и/или дисперсии GC, причем значение применяют для оценки ВЧК образца с улучшенной чувствительностью и специфичностью. Согласно некоторым вариантам реализации значение размера фрагмента можно подогнать с применением способа анализа главных компонент для устранения компонентов дисперсии, не связанных с вариацией числа копий последовательности, представляющей интерес, как далее описано применительно к блоку 719 фиг. 2F. Согласно некоторым вариантам реализации значение размера фрагмента можно подобрать посредством устранения выпадающих значе-

ний блоков в пределах образца, как описано применительно к блоку 321 фиг. 3А.

Многопроходный процесс для определения числа копий с применением нескольких параметров.

Как подчеркнуто выше, процессы, раскрытые в настоящем документе, являются подходящими для определения ВЧК с применением нескольких параметров, включая, без ограничения, перекрытие, взвешенное по размеру фрагмента перекрытие, размер фрагмента, фракцию или отношение фрагментов в заданном диапазоне размера, уровень метилирования фрагментов и т.д. Каждый из данных параметров можно отдельно процессировать, чтобы параметр индивидуально внес вклад в определение итоговой вариации числа копий.

Согласно некоторым вариантам реализации аналогичные процессы можно применять в отношении анализа взвешенного по размеру перекрытия и анализа размера фрагмента, оба из которых являются параметрами размера фрагментов. На фиг. 2D представлена блок-схема двух перекрывающихся проходов рабочего процесса 600, проход 1 для взвешенного по размеру перекрытия и проход 2 для анализа размера фрагмента. Согласно другому варианту реализации, не показанному в настоящем документе, уровень метилирования можно процессировать в одном дополнительном проходе. Два прохода могут включать сравнимые операции для получения подогнанной информации о перекрытии, на которой основано определение ВЧК.

Исходная однопроходная часть процесса начинается с получения данных секвенирования, см. блок 602, и продолжается компьютеризированным вычислением подсчитанных значений, как описано выше, см. блок 612. После данной точки изображенный процесс разделяется на два прохода, как описано выше. Возвращаясь к начальной части процесса, рабочий процесс преобразует данные секвенирования в риды последовательности. Если данные секвенирования получены из мультиплексного секвенирования, риды последовательности также демультиплексируют для идентификации источника данных. См. блок 604. Затем риды последовательности выравнивают с референсной последовательностью, причем выровненные риды последовательности предложены в виде меток последовательности. См. блок 606. После этого метки последовательности фильтруют для получения неисключенных сайтов (NES), которые представляют собой однозначно картированные недублирующиеся метки последовательности. Метки последовательности организованы в блоки конкретной длины последовательности, такой как 1 т.о., 100 т.о. или 1 Мб. См. блок 610. Согласно некоторым вариантам реализации, включающим анализ синдром-специфичных областей, длина блоков составляет 100 т.о. Согласно некоторым вариантам реализации блоки, демонстрирующие высокую вариабельность, можно маскировать с применением маски последовательности, полученной из множества непораженных образцов способом, описанным на фиг. 3А, блок 313. Затем метки в NES подсчитывают для получения перекрытий, подлежащих нормированию и подгонке для анализа ВЧК. См. блок 612.

Согласно представленному варианту реализации операции 604, 606, 610 и 612 осуществляют один раз, и большинство из остальных операций осуществляют дважды, один раз для анализа взвешенного по размеру перекрытия (проход 1) и один раз для анализа размера фрагмента (проход 2). Согласно другим вариантам реализации одну или более операций, которые показаны как осуществляемые в двух проходах, осуществляют исключительно один раз, и результаты используют в обоих процессах. Примеры таких совместно используемых операций включают операции 614, 616 и 618.

Согласно представленным вариантам реализации полученные перекрытия (взвешенные по размеру подсчитанные значения) или параметр размера фрагмента (фракции или соотношения размера) NES нормируют посредством, например, деления значения NES блока на суммарные NES генома или множества нормирующих хромосом. Согласно некоторым вариантам реализации нормируют исключительно перекрытие, в то время как нет необходимости нормировать параметр размера фрагмента, поскольку глубина секвенирования не влияет на него таким же образом, как на перекрытие. См. блок 614. Затем согласно некоторым вариантам реализации устраняют дисперсию, общую для обучающего множества, включая непораженные образцы, причем дисперсия не связана с ВЧК, представляющей интерес. Согласно представленному варианту реализации общая дисперсия представлена как глобальный волновой профиль, полученный из непораженных образцов образом, аналогичным получению глобального волнового профиля, описанному выше. Согласно некоторым вариантам реализации, как проиллюстрировано на фиг. 6, непораженные образцы, применяемые для получения глобального волнового профиля, включают образцы из одной и той же проточной ячейки или процессированной партии. См. блок 616. Вычисление глобальной волны, специфичной к проточной ячейке, подробнее объяснен ниже по тексту. Согласно представленному варианту реализации после того, как был устранен глобальный волновой профиль, перекрытия корректируют с учетом уровня GC образец-специфичным образом. См. блок 616. Некоторые алгоритмы коррекции GC описаны более подробно ниже по тексту в описании, связанном с фиг. 3А, блок 319.

Согласно представленному варианту реализации как в проходе 1 для анализа взвешенного перекрытия, так и в проходе 2 для анализа размера фрагмента затем данные можно отфильтровать с учетом шума, специфичного индивидуальному образцу, например из анализа можно удалить резко отклоняющиеся данные блоков, которые характеризуются перекрытиями, чрезвычайно отличающимися от других блоков, причем различие нельзя отнести к вариации числа копий, представляющей интерес. См. блок 622.

Данная операция внутривыборочного фильтрования может соответствовать блоку 321 на фиг. 3А.

Согласно некоторым вариантам реализации после фильтрования одного образца взвешенные значения перекрытия прохода 1 и параметра размера фрагмента прохода 2 обогащают в целевой сигнал по сравнению с референсом. См. блоки 624 и 628. Затем перекрытие и параметр размера фрагмента для хромосомы применяют для вычисления дозы хромосомы и нормированного значения хромосомы (NCV), как описано выше. После этого NCV можно сравнить с критерием для определения показателя, свидетельствующего о вероятности ВЧК. См. блоки 626 и 630. Затем показатели из двух проходов можно объединить с получением комплексного итогового показателя, который определяет, следует ли принять решение об анеуплоидии. Согласно некоторым вариантам реализации показатели 626 и 630 представляют собой статистические данные t-критерия или Z-значения. Согласно некоторым вариантам реализации итоговый показатель представляет собой значение хи-квадрат. Согласно другим вариантам реализации итоговый показатель представляет собой среднее квадратичное значение двух t-значений или z-показателей. Другой способ объединения двух показателей от двух проходов можно применять для улучшения общей чувствительности и селективности при обнаружении ВЧК. В качестве альтернативы, можно объединить два показателя из двух проходов посредством логических операции, например операции И или операции ИЛИ. Например, когда для обеспечения низкой доли ложноотрицательных результатов предпочтительной является высокая чувствительность, решение о ВЧК можно принять, когда показатель из прохода 1 ИЛИ прохода 2 соответствует критерию решения. С другой стороны, если для обеспечения низкой доли ложноположительных результатов желательной является высокая селективность, решение о ВЧК можно принять исключительно если показатель прохода 1 И прохода 2 соответствует критерию решения.

Примечательно, что существует компромисс между чувствительностью и селективностью с применением таких логических операций, описанных выше. Согласно некоторым вариантам реализации применяют подход двухэтапного секвенирования, чтобы преодолеть компромисс, как далее описано ниже по тексту. Вкратце, исходное определение показателя для образца сравнивают с относительно низким первым порогом, предназначенным для увеличения чувствительности, и, если показатель образца превышает первый порог, его направляют на второй раунд секвенирования, который является более глубоким, чем первый. Такой образец затем повторно процессируют и анализируют в рабочем процессе, аналогичном таковому, описанному выше. Затем полученный в результате показатель сравнивают с относительно высоким вторым порогом, предназначенным для улучшения чувствительности. Согласно некоторым вариантам реализации образцы, которые подвергают второму раунду секвенирования, характеризуются показателем относительно более низким среди образцов, показатель которых превышает первый порог, посредством чего снижается количество образцов, которые необходимо повторно секвенировать.

Согласно некоторым вариантам реализации можно применять 3-й проход с применением 3-го параметра. Примером данного 3-го прохода является метилирование. Метилирование можно определить напрямую посредством измерения метилирования нуклеиновых кислот из образца или опосредованно как параметр, который коррелирует с размером фрагмента бесклеточных нуклеиновых кислот.

Согласно некоторым вариантам реализации данный 3-й параметр представляет собой 2-е перекрытие или параметр на основании подсчета, причем подсчитанные значения основаны на размере фрагментов за пределами размера первичного фрагмента, который использовали в первом параметре на основании подсчета. Когда для получения подсчета или параметра перекрытия применяют фрагменты от 80 до 150 пар оснований, они исключают приблизительно 70% ридов из секвенирования. В той степени, в которой данные исключенные риды все еще характеризуются некоторым потенциально подходящим сигналом, их можно применять в 3-м параметре, который включает исключенные риды или риды во фракции на основании размера, которая находится за пределами или перекрывается с фракцией на основании размера, использованной в первом параметре. В этой связи ридам и связанным значениям перекрытия, взятым из исключенных фрагментов, может быть присвоен меньший вес. Другими словами, параметру вариации числа копий, вычисленному с применением данных ридов, можно приписать меньшую важность при принятии итогового решения о вариации числа копий. В качестве альтернативы, как описано выше, метки за пределами диапазона размера в первом параметре могут принимать отрицательное значение, когда два генома имеют противоположные характеристики в двух диапазонах размера.

Согласно различным вариантам реализации перекрытия в процессах 200, 220 и 600 смещены в сторону меток из фрагментов на более короткой границе спектра размера фрагмента. Согласно некоторым вариантам реализации перекрытия смещены в сторону меток из фрагментов размеров, более коротких, чем указанное значение. Согласно некоторым вариантам реализации перекрытия смещены в сторону меток из фрагментов в диапазоне размеров фрагмента, и верхняя граница диапазона составляет приблизительно 150 пар оснований или менее.

Согласно различным вариантам реализации процессов 200, 220 и 600 риды последовательности получают в результате секвенирования фрагментов бесклеточной нуклеиновой кислоты без первоначального применения ПЦР для амплификации нуклеиновых кислот фрагментов бесклеточной нуклеиновой кислоты. Согласно различным вариантам реализации риды секвенирования получают в результате секвенирования фрагментов бесклеточной нуклеиновой кислоты до глубины не более чем приблизительно 6М

фрагментов на образец. Согласно некоторым вариантам реализации глубина секвенирования составляет не более чем приблизительно 1M фрагментов на образец. Согласно некоторым вариантам реализации риды секвенирования получают посредством мультиплексного секвенирования, и количество мультиплексированных образцов составляет по меньшей мере приблизительно 24.

Согласно различным вариантам реализации процессов 200, 220 и 600 исследуемый образец содержит плазму от индивидуума. Согласно некоторым вариантам реализации процессы также включают получение бесклеточной нуклеиновой кислоты из исследуемого образца. Согласно некоторым вариантам реализации процессы также включают секвенирование фрагментов бесклеточной нуклеиновой кислоты, полученных из двух или более геномов.

Согласно различным вариантам реализации процессов 200, 220 и 600 два или более геномов включают геномы матери и плода. Согласно некоторым вариантам реализации вариация числа копий в последовательности, представляющей интерес, включает анеуплоидию в геноме плода.

Согласно некоторым вариантам реализации процессов 200, 220 и 600 два или более геномов включают геномы раковых и соматических клеток. Согласно некоторым вариантам реализации процессы включают применение вариации числа копий в раковом геноме для диагностики рака, контроля прогрессирования рака и/или определения лечения рака. Согласно некоторым вариантам реализации вариация числа копий вызывает генетическую аномалию.

Согласно некоторым вариантам реализации процессов 200, 220 и 600 перекрытия смещены в сторону меток из фрагментов на более длинной границе спектра размера фрагментов. Согласно некоторым вариантам реализации перекрытия смещены в сторону меток из размеров фрагментов, более длинных, чем указанное значение. Согласно некоторым вариантам реализации перекрытия смещены в сторону меток из фрагментов в диапазоне размеров фрагментов, причем более низкая граница диапазона составляет приблизительно 150 пар оснований или более.

Согласно некоторым вариантам реализации процессов 200, 220 и 600 процессы также включают: определение в блоках референсного генома, содержащих последовательность, представляющую интерес, уровней метилирования фрагментов бесклеточной нуклеиновой кислоты в указанных блоках и применение уровней метилирования в дополнение к вычисленным перекрытиям или значениям параметра размера фрагмента либо вместо них для идентификации вариации числа копий. Согласно некоторому варианту реализации применение уровней метилирования для идентификации вариации числа копий включает обеспечение глобального профиля метилирования для блоков последовательности, представляющей интерес. Глобальный профиль метилирования включает ожидаемые уровни метилирования по меньшей мере в блоках последовательности, представляющей интерес. Согласно некоторым вариантам реализации ожидаемые уровни метилирования получают из длин фрагментов бесклеточной нуклеиновой кислоты в обучающем множестве непораженных обучающих образцов, содержащих нуклеиновые кислоты, секвенированные и выровненные по существу тем же способом, что и фрагменты нуклеиновой кислоты исследуемого образца, причем ожидаемые уровни метилирования демонстрируют межблоковую вариацию. Согласно некоторым вариантам реализации процессы включают подгонку значения уровней метилирования с применением ожидаемых уровней метилирования в блоках по меньшей мере последовательности, представляющей интерес, и посредством этого получение откорректированных с учетом глобального профиля значений уровней метилирования для последовательности, представляющей интерес. Процессы также включают идентификацию вариации числа копий с применением откорректированных с учетом глобального профиля перекрытий и откорректированных с учетом глобального профиля уровней метилирования. Согласно некоторым вариантам реализации идентификация вариации числа копий с применением откорректированных с учетом глобального профиля перекрытий и откорректированных с учетом глобального профиля уровней метилирования также включает: подгонку откорректированных с учетом глобального профиля перекрытий и откорректированных с учетом глобального профиля уровней метилирования, основанную на уровне содержания GC, и посредством этого получение откорректированных с учетом GC перекрытий и откорректированных с учетом GC значений уровней метилирования для последовательности, представляющей интерес; и идентификацию вариации числа копий с применением откорректированных с учетом GC перекрытий и откорректированных с учетом GC уровней метилирования.

Согласно некоторым вариантам реализации процессов 200, 220 и 600 параметр размера фрагмента включает фракцию или соотношение, включая часть фрагментов бесклеточной нуклеиновой кислоты в исследуемом образце, размер фрагментов которых является более коротким или более длинным, чем пороговое значение. Согласно некоторым вариантам реализации параметр размера фрагмента включает фракцию, содержащую (i) количество фрагментов в исследуемом образце в пределах первого диапазона размера, содержащего 110 пар оснований, и (ii) количество фрагментов в исследуемом образце в пределах второго диапазона размера, содержащего первый диапазон размера и размеры за пределами первого диапазона размера.

Определение числа копий с применением трехпроходного процесса, отношений правдоподобия, t-статистики и/или фракций плода.

На фиг. 2E представлена блок-схема трехпроходного процесса для оценки числа копий. Данный

процесс включает три перекрывающихся прохода рабочего процесса 700, который включает проход 1 (или 713A) анализа перекрытия ридов, связанных с фрагментами всех размеров, проход 2 (или 713B) анализа перекрытия ридов, связанных с более короткими фрагментами, и проход 3 (или 713C) анализа относительной частоты более коротких ридов по сравнению со всеми ридами.

Процесс 700 аналогичен процессу 600 по своей общей организации. Операции, указанные в блоках 702, 704, 706, 710, 712, можно осуществить тем же или аналогичным способом, как и операции, указанные в блоках 602, 604, 606 и 610 и 612. После получения подсчитанных значений ридов определяют перекрытие с применением ридов из фрагментов всех размеров в проходе 713A. Перекрытие определяют с применением ридов из коротких фрагментов в проходе 713B. Частоту ридов из коротких фрагментов по сравнению со всеми ридами определяют в проходе 713C. Относительную частоту в другом месте в настоящем документе также называют соотношением размера или фракцией размера. Относительная частота представляет собой пример характеристики размера фрагмента. Согласно некоторым вариантам реализации короткие фрагменты представляют собой фрагменты, более короткие, чем приблизительно 150 пар оснований. Согласно различным вариантам реализации короткие фрагменты могут находиться в диапазонах размера приблизительно 50-150, 80-150 или 110-150 пар оснований. Согласно некоторым вариантам реализации третий проход, или проход 713C, является необязательным.

Все данные из трех проходов 713A, 713B и 713C подвергают операции нормирования 714, 716, 718, 719 и 722 для устранения дисперсии, не связанной с числом копий последовательности, представляющей интерес. Данные операции нормирования ограничены в блоке 723. Операция 714 включает нормирование проанализированного количества последовательности, представляющей интерес, посредством деления проанализированного количества на суммарное значение количества референсной последовательности. На данном этапе нормирования используют значения, полученные из исследуемого образца. Аналогично, операции 718 и 722 нормируют проанализированное количество с применением значений, полученных из исследуемого образца. В операциях 716 и 719 используют значения, полученные из обучающего множества непораженных образцов.

Операция 716 устраняет глобальную волновую дисперсию, полученную из обучающего множества непораженных образцов, в которых используют те же или аналогичные способы, описанные применительно к блоку 616. Операция 718 устраняет дисперсию специфичной индивидууму дисперсии GC с применением того же или аналогичных способов, описанных применительно к блоку 618.

Операция 719 устраняет дополнительную дисперсию с применением способа анализа главных компонент (АГК). Дисперсия, устраняемая методами АГК, обусловлена факторами, не связанными с числом копий последовательности, представляющей интерес. Проанализированное количество в каждом блоке (перекрытие, соотношение размера фрагмента и т.д.) обеспечивает независимую переменную для АГК, и образцы непораженного обучающего множества обеспечивают значения для данных независимых переменных. Все образцы обучающего множества включают образцы, которые характеризуются тем же числом копий последовательности, представляющей интерес, например, двумя копиями соматической хромосомы, одной копией X-хромосомы (когда в качестве непораженных образцов применяют образцы мужского пола) или двумя копиями X-хромосомы (когда в качестве непораженных образцов применяют образцы женского пола). Таким образом, дисперсия в образцах не является следствием анеуплоидии или другого отличия в числе копий. АГК обучающего множества позволяет получить главные компоненты, которые не связаны с числом копий последовательности, представляющей интерес. Затем главные компоненты можно использовать для устранения дисперсии в исследуемом образце, не связанной с числом копий последовательности, представляющей интерес.

Согласно определенным вариантам реализации дисперсию одной или более главных компонент устраняют из данных исследуемого образца с применением коэффициентов, вычисленных из данных непораженных образцов в области за пределами последовательности, представляющей интерес. Согласно некоторым вариантам реализации область представляет собой все устойчивые хромосомы. Например, АГК осуществляют на нормированных данных перекрытия блока обучающих нормальных образцов с получением, таким образом, главных компонент, соответствующих размерам, при которых может быть зафиксирована наибольшая дисперсия в данных. Дисперсия, зафиксированная таким образом, не связана с вариацией числа копий в последовательности, представляющей интерес. После того как из обучающих нормальных образцов были получены главные компоненты, их применяют в отношении исследуемых данных. В пределах блоков из области за пределами последовательности, представляющей интерес получают модель линейной регрессии с исследуемым образцом в качестве переменной ответа и с главными компонентами в качестве зависимых переменных. Полученные в результате коэффициенты регрессии применяют для нормирования перекрытия блока области, представляющей интерес, посредством вычитания линейной комбинации главных компонент, заданных посредством вычисленных коэффициентов регрессии. Это позволяет устранить дисперсию, не связанную с ВЧК, из последовательности, представляющей интерес. См. блок 719. Для последующего анализа применяют остаточные данные. Дополнительно, операция 722 устраняет резко отклоняющиеся значения данных наблюдений с применением способов, описанных применительно к блоку 622.

После проведения операций нормирования в блоке 723 значения перекрытия всех блоков были

"нормированы" для устранения источников вариации, отличных от анеуплоидии или другой вариаций числа копий. В некотором смысле блоки последовательности, представляющей интерес, обогащены или изменены по сравнению с другими блоками с целью обнаружения вариации числа копий. См. блок 724, который представляет собой не операцию, но представляет полученные в результате значения перекрытия. Операции нормирования в большом блоке 723 могут увеличить сигнал и/или снизить шум для количества, которое анализируют. Аналогично, значения перекрытия коротких фрагментов для блоков нормировали с целью устранения источников вариации, отличной от анеуплоидии или других вариаций числа копий, как показано в блоке 728, и относительную частоту коротких фрагментов (или соотношение размера) для блоков нормировали аналогичным способом для устранения источников вариации, отличных от анеуплоидии, или других вариаций числа копий, как показано в блоке 732. Как и в случае блока 724, блоки 728 и 732 представляют собой не операции, но представляют перекрытие и значения относительной частоты после обработки большого блока 723. Следует понимать, что операции в большом блоке 723 можно модифицировать, реорганизовать или удалить. Например, согласно некоторым вариантам реализации операцию АГК 719 не осуществляют. Согласно другим вариантам реализации операцию корректировки с учетом GC 718 не осуществляют. Согласно другим вариантам реализации порядок операций изменен; например, операцию АГК 719 осуществляют перед операцией корректировки с учетом GC 718.

Перекрытие всех фрагментов после нормирования и удаления дисперсии, представленного в блоке 724, применяют для получения t-статистики в блоке 726. Аналогично, перекрытие коротких фрагментов после нормирования и удаления дисперсии, представленного в блоке 728, применяют для получения t-статистики в блоке 730, и относительную частоту коротких фрагментов после нормирования и удаления дисперсии, представленного в блоке 732, применяют для получения t-статистики в блоке 734.

На фиг. 2F представлено, почему применение t-статистики в отношении анализа числа копий может способствовать улучшению точности анализа. На фиг. 2F представлены, на каждом чертеже, распределения частоты нормированного перекрытия блока последовательности, представляющей интерес, и референсной последовательности, причем распределение последовательности, представляющей интерес, перекрывает и ограничивает распределение референсной последовательности. На верхнем чертеже представлено перекрытие блока для образца, который характеризуется более высоким перекрытием и который содержит свыше 6 миллионов ридов; на нижнем чертеже представлено перекрытие блока для образца, который характеризуется более низким перекрытием и который содержит менее 2 миллионов ридов. На горизонтальной оси указано перекрытие, нормированное по сравнению со средним значением перекрытия референсной последовательности. На вертикальной оси указана относительная плотность вероятности в отношении количеств блоков, которые характеризуются средними значениями перекрытия. Таким образом, фиг. 2F представляет собой разновидность гистограммы. Распределение для последовательности, представляющей интерес, представлено спереди, и распределение для референсной последовательности представлено сзади. Среднее значение распределения последовательности, представляющей интерес, является более низким, чем таковое для референсной последовательности, что свидетельствует о меньшем числе копий в образце. Среднее значение разницы между последовательностью, представляющей интерес, и референсной последовательностью аналогично для образца с высоким перекрытием на верхнем чертеже и для образца с низким перекрытием на нижнем чертеже. Таким образом, согласно некоторым вариантам реализации можно использовать отличие среднего значения для идентификации вариации числа копий в последовательности, представляющей интерес. Отметим, что распределения образца с высоким перекрытием характеризуются дисперсиями, меньшими, чем таковые образца с низким перекрытием. Применение исключительно среднего значения для установления отличия между двумя распределениями не фиксирует отличие между двумя распределениями, а также применение среднего значения и дисперсии. Т-статистика может отражать как среднее значение, так и дисперсию распределения.

Согласно некоторым вариантам реализации операция 726 вычисляет t-статистику следующим образом:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

где x_1 представляет собой перекрытие блока последовательности, представляющей интерес, x_2 представляет собой перекрытие блока референсной области/последовательности, s_1 представляет собой стандартное отклонение перекрытий последовательности, представляющей интерес, s_2 представляет собой стандартное отклонение перекрытий референсной области, n_1 представляет собой количество блоков последовательности, представляющей интерес; и n_2 представляет собой количество блоков референсной области.

Согласно некоторым вариантам реализации референсная область содержит все устойчивые хромосомы (например, хромосомы, отличные от хромосом, которые, наиболее вероятно, несут анеуплоидию).

Согласно некоторым вариантам реализации референсная область содержит по меньшей мере одну хромосому за пределами последовательности, представляющей интерес. Согласно некоторым вариантам реализации референсная область содержит устойчивые хромосомы, не содержащие последовательность, представляющую интерес. Согласно другим вариантам реализации референсная область содержит множество хромосом (например, подмножество хромосом, выбранных из устойчивых хромосом), которые были определены для обеспечения наилучшей способности обнаружения сигнала для множества обучающих образцов. Согласно некоторым вариантам реализации способность обнаружения сигнала основана на способности референсной области устанавливать отличие между блоками, которые несут вариации числа копий, и блоками, которые не несут вариации числа копий. Согласно некоторым вариантам реализации референсную область идентифицируют способом, аналогичным таковому, который применяют для определения "нормирующей последовательности" или "нормирующей хромосомы", как описано в разделе, озаглавленном "Идентификация нормирующих последовательностей".

Возвращаясь к фиг. 2E, одну или более оценок фракции плода (блок 735) можно объединить с любой t-статистикой в блоке 726, 730 и 734 с целью получения оценки правдоподобия для случая плоидности. См. блок 736. Согласно некоторым вариантам реализации одну или более фракций плода блока 740 получают посредством любого из процесса 800 на фиг. 2G, процесса 900 на фиг. 2H или процесса 1000 на фиг. 2I. Процессы можно осуществлять параллельно с применением рабочего процесса, такого как рабочий процесс 1100 на фиг. 2J.

На фиг. 2G представлен пример процесса 800 для определения фракции плода из информации о перекрытии согласно некоторым вариантам реализации настоящего изобретения. Процесс 800 начинается с получения информации о перекрытии (например, значений дозы последовательности) обучающих образцов из обучающего множества. См. блок 802. Каждый образец обучающего множества получен от беременной женщины, которая установленно вынашивает плод мужского пола. А именно, образец содержит сцДНК плода мужского пола. Согласно некоторым вариантам реализации операция 802 может получить перекрытие последовательности, нормированное способами, отличными от дозы последовательности, как описано в настоящем документе, или может получить другие значения перекрытия.

Затем процесс 800 включает вычисление фракций плода обучающих образцов. Согласно некоторым вариантам реализации фракции плода можно вычислить по значениям дозы последовательности

$$\text{ФЭ}_{\boldsymbol{j}} = -\mathbf{2} \times \frac{\boldsymbol{Rx_j} - \text{медиана}(\boldsymbol{Rx_i})}{\text{медиана}(\boldsymbol{Rx_i})},$$

где Rx_j представляет собой дозу последовательности для образца мужского пола, медиана (Rx_i) представляет собой медиану доз последовательности для образцов женского пола. Согласно другим вариантам реализации можно применять среднее значение или другой показатель главной тенденции. Согласно некоторым вариантам реализации ФЭ можно получить другими способами, такими как относительная частота X- и Y-хромосомы. См. блок 804.

Процесс 800 также включает разделение референсной последовательности на несколько блоков субпоследовательностей. Согласно некоторым вариантам реализации референсная последовательность представляет собой полный геном. Согласно некоторым вариантам реализации блоки представляют собой блоки длиной 100 т.о. Согласно некоторым вариантам реализации геном разделяют приблизительно на 25000 блоков. После этого процесс получает перекрытия блоков. См. блок 806. Согласно некоторым вариантам реализации перекрытия, используемые в блоке 806, получают после осуществления операций нормирования, продемонстрированных в блоке 1123 фиг. 2J. Согласно другим вариантам реализации можно применять перекрытия из отличного диапазона размера.

Каждый блок связан с перекрытиями образцов в обучающем множестве. Вследствие этого для каждого блока можно получить корреляцию между перекрытием образцов и фракциями плода образцов. Процесс 800 включает получение корреляций между фракцией плода и перекрытием для всех блоков. См. блок 808. Затем процесс выбирает блоки, значения корреляции которых превышают порог. См. блок 810. Согласно некоторым вариантам реализации выбирают блоки, которые характеризуются 6000 наивысшими значениями корреляции. Целью является идентификация блоков, которые демонстрируют высокую корреляцию между перекрытием и фракцией плода в обучающих образцах. Затем блоки можно применять для прогнозирования фракции плода в исследуемом образце. Несмотря на то что обучающие образцы представляют собой образцы мужского пола, можно обобщить корреляцию между фракцией плода и перекрытием на исследуемые образцы мужского и женского пола.

С применением выбранных блоков, которые характеризуются высокими значениями корреляции, процесс позволяет получить линейную модель, устанавливающую взаимосвязь между фракцией плода и перекрытием. См. блок 812. Каждый выбранный блок обеспечивает независимую переменную для линейной модели. Вследствие этого полученная линейная модель также включает параметр или вес для каждого блока. Вес блоков подгоняют для аппроксимации модели к данным. После получения линейной модели процесс 800 включает применение данных перекрытия исследуемого образца в модели для определения фракции плода для исследуемого образца. См. блок 814. Применяемые данные перекрытия исследуемого образца предназначены для блоков, которые характеризуются высокими корреляциями между фракциями плода и перекрытием.

На фиг. 2J представлен рабочий процесс 1100 для обработки информации о ридах последовательности, который можно применять для получения оценок фракции плода. Рабочий процесс 1100 характеризуется аналогичными этапами процессинга, что и рабочий процесс 600 на фиг. 2D. Блоки 1102, 1104, 1106, 1110, 1112, 1123, 1114, 1116, 1118 и 1122, соответственно, соответствуют блокам 602, 604, 606, 610, 612, 623, 614, 616, 618 и 622. Согласно некоторым вариантам реализации одна или более операций нормирования в блоке 123 являются необязательными. Проход 1 обеспечивает информацию о перекрытии, которую можно применять в блоке 806 процесса 800, представленного на фиг. 2G. Процесс 800 затем может позволить получить оценку фракции плода 1150 на фиг. 2J.

Согласно некоторым вариантам реализации можно объединить множество оценок фракции плода (например, 1150 и 1152 на фиг. 2J) с получением комплексной оценки фракции плода (например, 1154). Для получения оценок фракции плода можно применять различные способы. Например, фракцию плода можно получить из информации о перекрытии. См. блок 1150 фиг. 2J и процесс 800 фиг. 2G. Согласно некоторым вариантам реализации фракцию плода можно также вычислить по распределению размера фрагментов. См. блок 1152 фиг. 2J и процесс 900 фиг. 2H. Согласно некоторым вариантам реализации фракцию плода можно также вычислить по распределению частоты 8-меров. См. блок 1152 фигуры 2J и процесс 1000 фиг, 2I.

В исследуемом образце, содержащем сцДНК плода мужского пола, фракцию плода можно также рассчитать из перекрытия Y-хромосомы и/или X-хромосомы. Согласно некоторым вариантам реализации комплексную оценку фракции плода (см., например, блок 1155) для плода предположительно мужского пола получают посредством применения информации, которая выбрана из группы, состоящей из: фракции плода, полученной из информации о перекрытии блоков, фракции плода, полученной из информации о размере фрагмента, фракции плода, полученной из перекрытия Y-хромосомы, фракции плода, полученной из X-хромосомы, и любой комбинации указанных фракций плода. Согласно некоторым вариантам реализации предполагаемый пол плода определяют посредством применения перекрытия Y-хромосомы. Две или более фракций плода (например, 1150 и 1152) можно объединить различными способами с получением комплексной оценки фракции плода (например, 1155). Например, можно применять подход среднего или взвешенного среднего согласно некоторым вариантам реализации, при котором взвешивание может быть основано на статистической достоверности оценки фракции плода.

Согласно некоторым вариантам реализации комплексную оценку фракции плода для плода предположительно женского пола получают посредством применения информации, выбранной из группы, состоящей из фракции плода, полученной из информации о перекрытии блоков, фракции плода, полученной из информации о размере фрагмента, и любой комбинации указанных фракций плода.

На фиг. 2H представлен процесс для определения фракции плода из информации о распределении размера согласно некоторым вариантам реализации. Процесс 900 начинается с получения информации о перекрытии (например, значений дозы последовательности) обучающих образцов мужского пола из обучающего множества. См. блок 902. Процесс 900 затем включает вычисление фракций плода обучающих образцов с применением способов, описанных выше применительно к блоку 804. См. блок 904.

Процесс 900 продолжается разделением диапазона размера на множество блоков для обеспечения блоков на основании размера фрагмента и для определения частот ридов для блоков на основании размера фрагмента. См. блок 906. Согласно некоторым вариантам реализации частоты блоков на основании размера фрагмента получают без нормирования с учетом факторов, продемонстрированных в блоке 1123. См. путь 1124 фиг. 2J. Согласно некоторым вариантам реализации частоты блоков на основании размера фрагмента получают после необязательного осуществления операции нормирования, продемонстрированной в блоке 1123 фиг. 2J. Согласно некоторым вариантам реализации диапазон размера разделяют на 40 блоков. Согласно некоторым вариантам реализации блок на нижней границе содержит фрагменты размера, меньшие, чем приблизительно 55 пар оснований. Согласно некоторым вариантам реализации блок на нижней границе содержит фрагменты размера в диапазоне приблизительно 50-55 пар оснований, что исключает информацию для ридов более коротких, чем 50 п.о. Согласно некоторым вариантам реализации блок на верхней границе содержит фрагменты размером более чем приблизительно 245 пар оснований. Согласно некоторым вариантам реализации блок на верхней границе содержит фрагменты размером в диапазоне приблизительно 245-250 пар оснований, что исключает информацию для ридов более длинных, чем 250 п.о.

Процесс 900 продолжается получением линейной модели, устанавливающей взаимосвязь между фракцией плода и частотами ридов для блоков на основании размера фрагмента, с применением данных обучающих образцов. См. блок 908. Полученная линейная модель включает независимые переменные для частот ридов блоков на основании размера. Модель также включает параметр или вес для каждого блока на основании размера. Вес блоков подгоняют для аппроксимации модели к данным. После получения линейной модели процесс 900 включает применение данных о частоте риды исследуемого образца в модели для определения фракции плода для исследуемого образца. См. блок 910.

Согласно некоторым вариантам реализации для вычисления фракции плода можно применять частоту 8-меров. На фиг. 2I представлен пример процесса 1000 для определения фракции плода из информации о частоте 8-меров согласно некоторым вариантам реализации настоящего изобретения. Процесс

1000 начинается с получения информации о перекрытии (например, значений дозы последовательности) обучающих образцов мужского пола из обучающего множества. См. блок 1002. Затем процесс 1000 включает вычисление фракций плода обучающих образцов с применением любого из способов, описанных для блока 804. См. блок 1004.

Процесс 1000 также включает получение частот 8-меров (например, все возможные пермутации 4 нуклеотидов в 8 положениях) из ридов каждого обучающего образца. См. блок 1006. Согласно некоторым вариантам реализации получают вплоть до 65536 или приблизительно данное количество 8-меров и их частот. Согласно некоторым вариантам реализации частоты 8-меров получают без нормирования с учетом факторов, продемонстрированных в блоке 1123. См. путь 1124 фиг. 2J. Согласно некоторым вариантам реализации частоты 8-меров получают после необязательного осуществления операции нормирования, продемонстрированной в блоке 1123 фиг. 2J.

Каждый 8-мер связан с частотами образцов в обучающем множестве. Вследствие этого для каждого 8-мера можно получить корреляцию между частотой 8-мера образцов и фракциями плода образцов. Процесс 1000 включает получение корреляций между фракцией плода и частотами 8-меров для всех 8-меров. См. блок 1008. Затем процесс позволяет выбрать 8-меры, которые характеризуются значениями корреляции выше порога. См. блок 1010. Целью является идентификация 8-меров, которые демонстрируют высокую корреляцию между частотой 8-мера и фракцией плода в обучающих образцах. Затем можно применять блоки для прогнозирования фракции плода в исследуемом образце. Несмотря на то что обучающие образцы представляют собой образцы мужского пола, корреляцию между фракцией плода и частотой 8-мера можно обобщить на исследуемые образцы мужского и женского пола.

С применением выбранных 8-меров, которые характеризуются высокими значениями корреляции, процесс позволяет получить линейную модель, устанавливающую взаимосвязь фракции плода с частотой 8-мера. См. блок 1012. Каждый выбранный блок обеспечивает независимую переменную для линейной модели. Вследствие этого полученная линейная модель также включает параметр или вес для каждого блока. После получения линейной модели процесс 1000 включает применение данных о частоте 8-меров исследуемого образца в модели для определения фракции плода для исследуемого образца. См. блок 1014.

Возвращаясь к фиг. 2E, согласно некоторым вариантам реализации процесс 700 включает получение итогового правдоподобия плоидности в операции 736 с применением t-статистики на основании перекрытия всех фрагментов, обеспеченных операцией 726, оценки фракции плода, обеспеченной операцией 726, и t-статистики на основании перекрытия коротких фрагментов, обеспеченных операцией 730. В данных вариантах реализации сочетают результаты прохода 1 и прохода 2 с применением многомерных нормальных моделей. Согласно некоторым вариантам реализации для оценки ВЧК правдоподобие плоидности представляет собой правдоподобие анеуплоидии, которое представляет собой правдоподобие модели, которая характеризуется анеуплоидным допущением (например, трисомия или моносомия) минус правдоподобие модели, которая характеризуется эуплоидным допущением, причем модель использует на входе t-статистику на основании перекрытия всех фрагментов, оценку фракции плода и t-статистику на основании перекрытия коротких фрагментов, и выдает правдоподобие.

Согласно некоторым вариантам реализации правдоподобие плоидности выражают в виде отношения правдоподобия. Согласно некоторым вариантам реализации отношение правдоподобия моделируют в виде

$$\mathrm{OB} = \frac{\sum_{ff_{\text{суммарн.}}} \boldsymbol{q}(\boldsymbol{ff}_{\text{суммарн.}}) * \boldsymbol{p_1}(\boldsymbol{T}_{\text{коротк.}}, \boldsymbol{T}_{\text{всех}} | \boldsymbol{ff}_{\text{рассч.}})}{\boldsymbol{p_0}(\boldsymbol{T}_{\text{коротк.}}, \boldsymbol{T}_{\text{всех}})},$$

где p_1 представляет собой правдоподобие того, что данные получены из многомерного нормального распределения, представляющего 3-копийную или 1-копийную модель, p_0 представляет собой правдоподобие того, что данные получены из многомерного нормального распределения, представляющего 2-копийную модель, $T_{\text{коротк.}}$, $T_{\text{всех}}$ представляют собой Т-показатели, вычисленные по хромосомному перекрытию, полученному из коротких и всех фрагментов, тогда как $q(ff_{\text{суммарн.}})$ представляет собой плотность распределения фракции плода (вычисленного из обучающих данных) с учетом ошибки, связанной с оценкой фракции плода. Модель сочетает в себе перекрытие, полученное из коротких фрагментов, с перекрытием, полученным из всех фрагментов, что помогает улучшить разделение между показателями перекрытия пораженных и непораженных образцов. Согласно представленному варианту реализации модель также использует фракцию плода, в результате чего дополнительно улучшается способность устанавливать отличие между пораженными и непораженными образцами. В настоящем документе отношение правдоподобия вычисляют с применением t-статистики, основанной на перекрытии всех фрагментов (726), t-статистики, основанной на перекрытии коротких фрагментов (730), и оценки фракции плода, обеспеченной процессами 800 (или блоком 726), 900 или 1000, как описано выше. Согласно некоторым вариантам реализации данное отношение правдоподобия применяют для анализа хромосом 13, 18 и 21.

Согласно некоторому варианту реализации правдоподобие плоидности, полученная посредством операции 736, использует исключительно t-статистику, полученную на основании относительной частоты коротких фрагментов, обеспеченных операцией 734 прохода 3, и оценку фракции плода, обеспечен-

ную операцией 726 процессов 800, 900 или 1000. Отношение правдоподобия можно вычислить согласно следующему уравнению:

$$\text{ОВ} = \frac{\Sigma_{ff_{\text{суммарн.}}} q(ff_{\text{суммарн.}}) * p_1(T_{\text{част._коротк.}} | ff_{\text{рассч.}})}{p_0(T_{\text{част._коротк}})},$$

где p_1 представляет собой правдоподобие того, что данные получены из многомерного нормального распределения, представляющего 3-копийную или 1-копийную модель, p_0 представляет собой правдоподобие того, что данные получены из многомерного нормального распределения, представляющего 2-копийную модель, $T_{\text{част._коротк.}}$ представляет собой T-показатель, вычисленный из относительной частоты коротких фрагментов, тогда как $q(ff_{\text{суммарн.}})$ представляет собой распределение плотности фракции плода (вычисленной из обучающих данных) с учетом ошибки, связанной с оценкой фракции плода. В настоящем документе отношение правдоподобия вычисляют с применением t-статистики, основанной на относительной частоте коротких фрагментов (734), и оценки фракции плода, обеспеченной процессами 800 (или блок 726), 900 или 1000, как описано выше. Согласно некоторым вариантам реализации данное отношение правдоподобия применяют для анализа хромосомы X.

Согласно некоторым вариантам реализации отношение правдоподобия вычисляют с применением t-статистики, основанной на перекрытии всех фрагментов (726), t-статистики, основанной на перекрытии коротких фрагментов (730), и относительной частоты коротких фрагментов (734). Более того, фракцию плода, полученную, как описано выше, можно объединить с t-статистикой для вычисления отношения правдоподобия. Различительную способность оценки плоидности можно улучшить посредством объединения информации из любого из трех проходов 713A, 713B и 713C. См., например, пример 2 и фиг. 12. Согласно некоторым вариантам реализации для получения отношений правдоподобия для хромосомы можно применять различные комбинации, например t-статистику из всех трех проходов, t-статистику из первого и второго проходов, фракцию плода и три параметра t-статистики, фракцию плода и один параметр t-статистики и т.д. Затем на основании рабочих характеристик моделей можно выбрать оптимальную комбинацию.

Согласно некоторым вариантам реализации для оценки аутосом смоделированное отношение правдоподобия представляет правдоподобие смоделированных данных, которые были получены из образца трисомии или моносомии, по сравнению с правдоподобием смоделированных данных, которые были получены из диплоидного образца. Такое отношение правдоподобия можно применять для определения трисомии или моносомии аутосом согласно некоторым вариантам реализации.

Согласно некоторым вариантам реализации для оценки половой хромосомы оценивают отношение правдоподобия для моносомии X и отношение правдоподобия для трисомии X. Более того, также оценивают измерение перекрытия хромосомы (например, ВЧК или z-показатель перекрытия) для хромосомы X и одно - для хромосомы Y. Согласно некоторым вариантам реализации для определения числа копий половой хромосомы оценивают четыре значения с применением дерева решений. Согласно некоторым вариантам реализации дерево решений позволяет определить случай плоидности XX, XY, X, XXY, XXX или XYY.

Согласно некоторым вариантам реализации отношение правдоподобия преобразуют в логарифм отношения правдоподобия, и критерий или порог для принятия решения об анеуплоидии или вариации числа копий можно задать эмпирически для получения конкретной чувствительности и селективности. Например, для принятия решения о трисомии 13 или трисомии 18 можно задать логарифм отношения правдоподобия 1,5 на основании чувствительности и селективности модели при использовании в отношении обучающего множества. Более того, например, в некоторых вариантах применения для трисомии хромосомы 21 можно задать значение критерия решения 3.

Детали иллюстративного процесса для определения перекрытия последовательности.

На фиг. 3A представлен пример процесса 301 для снижения шума в данных последовательности из исследуемого образца. На фиг. 3B-3J представлены данные анализов на различных этапах процесса. На фигуре представлен пример потока процесса, который можно применять в многопроходном процессе, таком как представленный на фиг. 2D.

В процессе 301, проиллюстрированном на фиг. 3A, для оценки числа копий применяют перекрытие метки последовательности, основанное на количестве меток последовательности. Однако аналогично описанию, приведенному выше относительно процесса 100 для определения ВЧК применительно к фиг. 1, для процесса 400 вместо перекрытия можно применять другие переменные или параметры, такие как размер, соотношение размера и уровень метилирования. Согласно некоторым вариантам реализации две или более переменных можно отдельно подвергать одному процессу для получения двух показателей, свидетельствующих о вероятности ВЧК, как показано выше применительно к фиг. 2D. Затем два показателя можно объединить для определения ВЧК. Более того, перекрытие и другие параметры можно взвесить по размеру фрагментов, из которых были получены метки. Для удобства чтения в процессе 300 упомянуто исключительно перекрытие, но следует отметить, что вместо перекрытия можно применять другие параметры, такие как размер, соотношение размера и уровень метилирования, подсчет, взвешенный по размеру, и т.д.

Как представлено на фиг. 3А, изображенный процесс начинается с экстракции сцДНК из одного или более образцов. См. блок 303. Подходящие процессы и аппараты для экстракции описаны в другом месте в настоящем документе. Согласно некоторым вариантам реализации сцДНК экстрагируют в процессе, описанном в заявке на патент США № 61/801126, поданной 15 марта 2013 года (полностью включена в настоящий документ посредством ссылки). Согласно некоторым вариантам реализации аппарат процессирует сцДНК из нескольких образцов в совокупности для обеспечения мультиплексных библиотек и данных последовательности. См. блоки 305 и 307 на фиг. 3А. Согласно некоторым вариантам реализации аппарат процессирует сцДНК из восьми или более исследуемых образцов параллельно. Как описано в настоящем документе в другом месте, система секвенирования может процессировать экстрагированную сцДНК для получения библиотеки кодированных (например, штриховым кодом) фрагментов сцДНК. Секвенатор секвенирует библиотеку сцДНК для получения очень большого количества ридов последовательности. Кодирование на образец позволяет демультиплексировать риды в мультиплексных образцах. Каждый из восьми или более образцов может характеризоваться сотнями тысяч или миллионами ридов. Процесс может фильтровать риды перед дополнительными операциями на фиг. 3А. Согласно некоторым вариантам реализации фильтрование рида представляет собой процесс фильтрования качества, осуществляемый программами системы программного обеспечения, встроенными в секвенатор, для отфильтровывания ошибочных и низкокачественных ридов. Например, программное обеспечение Sequencing Control Software (SCS, программное обеспечение для контроля секвенирования) и Consensus Assessment of Sequence and Variation (консенсусная оценка последовательности и вариации) программ системы Illumina отфильтровывает ошибочные и низкокачественные риды посредством преобразования первичных данных изображений, полученных посредством реакций секвенирования, в показатели интенсивности, основные отклики, оцененные по качеству выравнивания, и дополнительные форматы для обеспечения биологически значимой информации для последующего анализа.

После того как секвенатор или другой аппарат получает риды для образца, элемент системы компьютерным способом выравнивает риды с референсным геномом. См. блок 309. Выравнивание описано в другом месте в настоящем документе. Выравнивание позволяет получить метки, которые содержат риды последовательностей с аннотированной информацией о расположении, указывающей на уникальные положения в референсном геноме. Согласно определенным вариантам реализации система проводит первый проход выравнивания без учета дублирующихся ридов - двух или более ридов, которые содержат идентичные последовательности, - а затем устраняет дублирующиеся риды или подсчитывает дублирующиеся риды как один рид для получения недублирующихся меток последовательности. Согласно другим вариантам реализации система не устраняет дублирующиеся риды. Согласно некоторым вариантам реализации процесс устраняет из рассмотрения риды, которые выравниваются с несколькими расположениями в геноме, для получения уникально выровненных меток. Согласно некоторым вариантам реализации уникально выровненные, не повторяющиеся метки последовательности, картированные на неисключенные сайты (NES), подсчитывают для получения подсчета неисключенных сайтов (подсчитанных значений NES), которые обеспечивают данные для оценки перекрытия.

Как объяснено в другом месте, исключенные сайты представляют собой сайты, обнаруженные в областях референсного генома, которые были исключены с целью подсчета меток последовательности. Согласно некоторым вариантам реализации исключенные сайты обнаружены в областях хромосом, которые содержат повторяющиеся последовательности, например центромеры и теломеры, и в областях хромосом, которые являются общими для более чем одной хромосомы, например области, присутствующие на Y-хромосоме, которые также присутствуют на X-хромосоме. Неисключенные сайты (NES) представляют собой сайты, которые не исключены в референсном геноме с целью подсчета меток последовательности.

Затем система разделяет выровненные метки на блоки в референсном геноме. См. блок 311. Блоки расположены по всей длине референсного генома. Согласно некоторым вариантам реализации весь референсный геном разделяют на непрерывные блоки, которые могут характеризоваться заданным равным размером (например, 100 т.о.). В качестве альтернативы, блоки могут характеризоваться длиной, определенной динамически, возможно, для каждого образца. Глубина секвенирования влияет на выбор оптимального размера блока. Блоки с динамически определенными размерами могут характеризоваться размером, определяемым размером библиотеки. Например, можно определить размер блока, который представляет собой длину последовательности, требуемую, в среднем, для размещения 1000 меток.

Каждый блок содержит некоторое количество меток из рассматриваемого образца. Данное количество меток, которое отражает "перекрытие" выровненной последовательности, выступает в качестве исходной точки для фильтрования и очистки данных образца иным способом для достоверного определения вариации числа копий в образце. На фиг. 3А представлены операции очистки в блоках 313-321.

Согласно варианту реализации, изображенному на фиг. 3А, в процессе применяют маску в отношении блоков референсного генома. См. блок 313. Система может исключать из рассмотрения перекрытие в маскированных блоках в некоторых или всех из следующих операций процесса. Во многих случаях в любой из оставшихся операций на фиг. 3А не учитывают значения перекрытия из маскированных блоков.

Согласно различным вариантам реализации для устранения блоков из областей генома, которые, как было обнаружено, демонстрируют высокую внутривыборочную вариабельность, применяют одну или более масок. Такие маски предложены как для хромосомы, представляющей интерес (например, хр13, 18 и 21), так и для других хромосом. Как объяснено в другом месте, хромосома, представляющая интерес, представляет собой хромосому, рассматриваемую как потенциально несущую вариацию числа копий или другую аберрацию.

Согласно некоторым вариантам реализации маски идентифицируют из обучающего множества квалификационных образцов с применением следующего подхода. Сначала каждый образец обучающего множества процессируют и фильтруют согласно операциям с 315 по 319 на фиг. 3А. Затем нормированные и откорректированные количества перекрытия отмечают для каждого блока, и для каждого блока вычисляют статистику, такую как стандартное отклонение, медиана абсолютного отклонения и/или коэффициент вариации. Различные комбинации фильтра можно оценить для каждой хромосомы, представляющей интерес. Комбинации фильтра обеспечивают один фильтр для блоков хромосомы, представляющей интерес, и отличный фильтр для блоков всех других хромосом.

Согласно некоторым вариантам реализации выбор нормирующей хромосомы (или группы хромосом) пересматривают после получения масок (например, посредством выбора пределов для хромосомы, представляющей интерес, как описано выше). После применения маски последовательности можно осуществить процесс выбора нормирующей хромосомы или хромосом, как описано в другом месте в настоящем документе. Например, все возможные комбинации хромосом оценивают в качестве нормирующих хромосом и ранжируют в зависимости от их способности различать пораженные и непораженные образцы. В ходе данного процесса можно найти (или можно не находить) различные оптимальные нормирующие хромосомы или группу хромосом. Согласно другим вариантам реализации нормирующие хромосомы представляют собой таковые, которые приводят к наименьшей вариабельности в дозе последовательности для последовательности, представляющей интерес, среди всех квалификационных образцов. Если идентифицируют отличную нормирующую хромосому или группу хромосом, процесс необязательно выполняет вышеописанную идентификацию блоков для фильтрования. Возможно, новая нормирующая хромосома или хромосомы приведут к отличным пределам.

Согласно определенным вариантам реализации для хромосомы Y применяют отличную маску. Пример подходящей маски для хромосомы Y описан в предварительной заявке на патент США № 61/836057, поданной 17 июня 2013 года [номер патентного реестра ARTEP008P], которая включена в настоящий документ посредством ссылки для всех целей.

После того как система компьютерным способом маскирует блока, она компьютерным способом нормирует значения перекрытия в блоках, которые не исключены масками. См. блок 315. Согласно определенным вариантам реализации система нормирует значения перекрытия исследуемого образца в каждом блоке (например, подсчитанные значения NES на блок) к большинству или всем перекрытиям в референсном геноме или его части (например, перекрытие в устойчивых хромосомах референсного генома). В некоторых случаях система нормирует значения перекрытия исследуемого образца (на блок) посредством деления подсчета для рассматриваемого блока на суммарное количество всех неисключенных сайтов, выровненных со всеми устойчивыми хромосомами в референсном геноме. Согласно некоторым вариантам реализации система нормирует значения перекрытия исследуемого образца (на блок) посредством осуществления линейной регрессии. Например, система сначала вычисляет перекрытия для подмножества блоков в устойчивых хромосомах как

$$y_a = \text{отсекаемый отрезок} + \text{наклон} \times gwp_a,$$

где y_a представляет собой перекрытие для блока a, и gwp_a представляет собой глобальный профиль для этого же блока. Затем система вычисляет нормированные перекрытия z_b как:

$$z_b = y_b/(\text{отсекаемый отрезок} + \text{наклон} \times gwp_b) - 1.$$

Как объяснено выше, устойчивая хромосома представляет собой хромосому, которая вряд ли является анеуплоидной. Согласно определенным вариантам реализации устойчивые хромосомы представляют собой все аутосомные хромосомы, отличные от хромосом 13, 18 и 21. Согласно некоторым вариантам реализации устойчивые хромосомы представляют собой все аутосомные хромосомы, отличные от хромосом, которые, как было определено, отклоняются от нормального диплоидного генома.

Значение трансформированного подсчета блока или перекрытия называют "нормированным количеством перекрытия" для последующего процессинга. Нормирование осуществляют с применением информации, уникальной для каждого образца. Как правило, не применяют информацию из обучающего множества. Нормирование позволяет обеспечить количества перекрытия из образцов, которые характеризуются различными размерами библиотеки (и, следовательно, различными количествами ридов и меток), которые подлежат обработке в равных условиях. В некоторых из последующих операций процесса применяют количества перекрытия, полученные из обучающих образцов, которые могут быть секвенированы из библиотек, которые являются большими или меньшими, чем библиотеки, применяемые для рассматриваемого исследуемого образца. Согласно некоторым вариантам реализации без нормирования, основанного на количестве ридов, выровненных со всем референсным геномом (или по меньшей мере с

устойчивыми хромосомами), обработка с применением параметров, полученных из обучающего множества, может быть ненадежной или не поддающейся обобщению.

Фиг. 3B иллюстрирует перекрытие в пределах хромосом 21, 13 и 18 для многих образцов. Некоторые образцы процессировали отлично друг от друга. Как следствие, в любом данном геномном положении наблюдается широкая внутривыборочная вариация. Нормирование устраняет некоторую часть внутривыборочной вариации. На левом чертеже фиг. 3C представлены нормированные количества перекрытия по всему геному.

Согласно варианту реализации фиг. 3A система устраняет или снижает "глобальный профиль" из нормированных количеств перекрытия, полученных в операции 315. См. блок 317. Данная операция устраняет систематические погрешности в нормированных количествах перекрытия, возникающие вследствие структуры генома, процесса получения библиотеки и процесса секвенирования. Помимо этого, данная операция предназначена для корректировки с учетом любого систематического линейного отклонения от ожидаемого профиля в любом данном образце.

Согласно некоторым вариантам реализации устранение глобального профиля включает деление нормированного количества перекрытия каждого блока на соответствующее ожидаемое значение каждого блока. Согласно другим вариантам реализации устранение глобального профиля включает вычитание ожидаемого значения каждого блока из нормированного количества перекрытия каждого блока. Ожидаемое значение может быть получено из обучающего множества непораженных образцов (или непораженных образцов женского пола для X-хромосомы). Непораженные образцы представляют собой образцы от индивидуумов, которые установленно не характеризуются анеуплоидией по хромосоме, представляющей интерес. Согласно некоторым вариантам реализации устранение глобального профиля включает вычитание ожидаемого значения каждого блока (полученного из обучающего множества) из нормированного количества перекрытия каждого блока. Согласно некоторым вариантам реализации процесс применяет медианные значения нормированных количеств перекрытия для каждого блока, как определено с применением обучающего множества. Другими словами, медианные значения представляют собой ожидаемые значения.

Согласно некоторым вариантам реализации устранение глобального профиля осуществляют с применением линейной корректировки для зависимости перекрытия образца от глобального профиля. Как указано, глобальный профиль представляет собой ожидаемое значение для каждого блока, как определено из обучающего множества (например, медианное значение для каждого блока). В данных вариантах реализации можно применять устойчивую линейную модель, полученную посредством аппроксимации нормированных количеств перекрытия исследуемого образца к медиане глобального профиля, полученной для каждого блока. Согласно некоторым вариантам реализации линейную модель получают посредством регрессирования наблюдаемых нормированных количеств перекрытия образца по сравнению с глобальной медианой (или другим ожидаемым значением) профиля.

Линейная модель основана на допущении, что количества перекрытия образца характеризуются линейной взаимосвязью со значениями глобального профиля, причем линейная взаимосвязь должна сохраняться как для устойчивых хромосом/областей, так и для последовательности, представляющей интерес. См. фиг. 3D. В таком случае регрессия нормированных количеств перекрытия образца на ожидаемые количества перекрытия глобального профиля позволит получить линию, которая характеризуется наклоном и отсекаемым отрезком. Согласно определенным вариантам реализации наклон и отсекаемый отрезок такой линии применяют для вычисления "предсказанного" количества перекрытия из значения глобального профиля для блока. Согласно некоторым вариантам реализации корректировка с учетом глобального профиля включает моделирование нормированного количества перекрытия каждого блока посредством предсказанных количеств перекрытия для блока. Согласно некоторым вариантам реализации перекрытия меток исследуемой последовательности подгоняют посредством: (i) получения математической зависимости между перекрытием меток исследуемой последовательности по сравнению с ожидаемым перекрытием во множестве блоков в одной или более устойчивых хромосомах или областях, и (ii) применения математической зависимости в отношении блоков в последовательности, представляющей интерес. Согласно некоторым вариантам реализации перекрытия в исследуемом образце являются откорректированными с учетом вариации с применением линейной взаимосвязи между ожидаемыми значениями перекрытия из непораженных обучающих образцов и значениями перекрытия для исследуемого образца в устойчивых хромосомах или других устойчивых областях генома. Подгонка приводит к получению перекрытий, откорректированных с учетом глобального профиля. В некоторых случаях подгонка включает получение перекрытий для исследуемого образца для подмножества блоков в устойчивых хромосомах или областях следующим образом:

$$y_a = \text{отсекаемый отрезок} + \text{наклон} * gwp_a ,$$

где y_a представляет собой перекрытие блока a для исследуемого образца в одной или более устойчивых хромосомах или областях, и gwp_a представляет собой глобальный профиль для блока a для непораженных обучающих образцов. Затем процесс вычисляет откорректированное с учетом глобального профиля перекрытие z_b для последовательности или области, представляющей интерес, как

$$z_b = y_b / (\text{отсекаемый отрезок} + \text{наклон} * gwp_b) - 1,$$

где y_b представляет собой наблюдаемое перекрытие блока b для исследуемого образца в последовательности, представляющей интерес (которая может располагаться за пределами устойчивой хромосомы или области), и gwp_b представляет собой глобальный профиль для блока b для непораженных обучающих образцов. Знаменатель (отсекаемый отрезок+наклон×gwp_b) представляет собой перекрытие для блока b, которое, как было предсказано, наблюдается в непораженных исследуемых образцах на основании взаимосвязи, вычисленной из устойчивых областей генома. В случае последовательности, представляющей интерес, несущей вариацию числа копий, наблюдаемое перекрытие и, следовательно, откорректированное с учетом глобального профиля значение перекрытия для блока b будет в значительной степени отклоняться от перекрытия непораженного образца. Например, в случае трисомического образца откорректированное перекрытие z_b будет являться пропорциональным фракции плода для блоков на пораженной хромосоме. Данный процесс проводит нормирование в пределах образца посредством компьютеризированного вычисления отсекаемого отрезка и наклона на устойчивых хромосомах, а затем оценивает, как геномная область, представляющая интерес, отклоняется от взаимосвязи (которая описывается наклоном и отсекаемым отрезком), справедливой для устойчивых хромосом в пределах одного образца.

Наклон и отсекаемый отрезок получают из линии, как представлено на фиг. 3D. Пример устранения глобального профиля представлен на фиг. 3C. На левом чертеже представлена высокая межблоковая вариация в нормированных количествах перекрытия среди множества образцов. На правом чертеже представлены те же нормированные количества перекрытия после устранения глобального профиля, как описано выше.

После того как система устраняет или снижает глобальный профиль вариаций в блоке 317, она проводит корректировку с учетом вариаций содержания GC (гуанин-цитозин) в образце. См. блок 319. Каждый блок характеризуется своим собственным относительным вкладом в GC. Фракцию определяют посредством деления количества нуклеотидов G и C в блоке на суммарное количество нуклеотидов в блоке (например, 100000). Некоторые блоки будут характеризоваться большими фракциями GC, чем другие. Как представлено на фиг. 3E и 3F, различные образцы демонстрируют различные погрешности GC. Данные различия и их корректировки дополнительно пояснены ниже. На фиг. 3E-G представлено откорректированное с учетом глобального профиля нормированное количество перекрытия (на блок) как функция от фракции GC (на блок). Неожиданно было установлено, что различные образцы демонстрируют различную GC-зависимость. Некоторые образцы демонстрируют монотонно убывающую зависимость (как на фиг. 3E), тогда как другие демонстрируют зависимость в виде запятой (как на фиг. 3F и 3G). Поскольку данные профили могут являться уникальными для каждого образца, коррекцию, описанную на данном этапе, осуществляют отдельно и уникально для каждого образца.

Согласно некоторым вариантам реализации система компьютерным способом упорядочивает блоки в зависимости от фракции GC, как проиллюстрировано на фиг. 3E-G. Затем система корректирует откорректированное с учетом глобального профиля нормированное количество перекрытия блока с применением информации от других блоков с аналогичным содержанием GC. Данную коррекцию применяют в отношении каждого немаскированного блока.

В некоторых процессах каждый блок корректируют с учетом содержания GC следующим образом. Система компьютерным способом выбирает блоки, которые характеризуются фракциями GC, аналогичными таковым рассматриваемого блока, а затем определяет параметр коррекции из информации в выбранных блоках. Согласно некоторым вариантам реализации те блоки, которые характеризуются аналогичными фракциями GC, выбирают с применением произвольно заданного значения предела подобия. В одном примере выбирают 2% всех блоков. Данные блоки представляют собой 2% блоков, которые характеризуются содержанием GC, максимально аналогичным рассматриваемому блоку. Например, выбирают 1% блоков, которые характеризуются незначительно большим содержанием GC, и 1%, которые характеризуются незначительно меньшим содержанием GC.

С применением выбранных блоков система компьютерным способом определяет параметр коррекции. В одном примере параметр коррекции представляет собой репрезентативное значение нормированных количеств перекрытия (после устранения глобального профиля) в выбранных блоках. Примеры такого репрезентативного значения включают медиану или среднее значение нормированных количеств перекрытия в выбранных блоках. Система применяет вычисленный параметр коррекции для рассматриваемого блока в отношении нормированного количества перекрытия (после устранения глобального профиля) для рассматриваемого блока. Согласно некоторым вариантам реализации репрезентативное значение (например, медианное значение) вычитают из нормированного количества перекрытия рассматриваемого блока. Согласно некоторым вариантам реализации медианное значение (или другое репрезентативное значение) нормированных количеств перекрытия выбирают исключительно с применением количеств перекрытия для устойчивых аутосомных хромосом (всех аутосом, отличных от хромосом 13, 18 и 21).

В одном примере с применением, например, блоков длиной 100 т.о. каждый блок будет характеризоваться уникальным значением фракции GC, и блоки разделяют на группы в зависимости от содержа-

ния в них фракции GC. Например, блоки разделяют на 50 групп, причем границы групп соответствуют (0, 2, 4, 6, ... и 100) квантилям распределения %GC. Медианное нормированное количество перекрытия вычисляют для каждой группы блоков из картирования устойчивых аутосом на ту же группу GC (в образце), а затем из нормированных количеств перекрытия вычитают медианное значение (для всех блоков по всему геному в той же группе GC). При этом применяют коррекцию GC, вычисленную из устойчивых хромосом в пределах любого данного образца, в отношении потенциально пораженных хромосом в пределах того же образца. Например, все блоки на устойчивых хромосомах, которые характеризуются содержанием GC от 0,338660 до 0,344720, группируют, вычисляют медиану для данной группы и вычитают из нормированного перекрытия блоков в пределах данного диапазона GC, причем блоки могут быть обнаружены где-либо в геноме (за исключением хромосом 13, 18, 21 и X). Согласно определенным вариантам реализации хромосому Y исключают из данного процесса коррекции с учетом GC.

На фиг. 3G представлено использование коррекции с учетом GC с применением медианы нормированных количеств перекрытия в качестве параметра коррекции, которая была только что описана. На левом чертеже представлены неоткорректированные количества перекрытия по сравнению с профилем фракции GC. Как показано, профиль характеризуется нелинейной формой. На правом чертеже представлены откорректированные количества перекрытия. На фиг. 3H представлены нормированные перекрытия для многих образцов до коррекции с учетом фракции GC (левый чертеж) и после коррекции с учетом фракции GC (правый чертеж). На фиг. 3I представлен коэффициент вариации (KB) нормированных перекрытий для многих исследуемых образцов до коррекции с учетом фракции GC (красный) и после коррекции с учетом фракции GC (зеленый), причем коррекция с учетом фракции GC приводит к по существу меньшей вариации в нормированных перекрытиях.

Описанный выше процесс является относительно простым вариантом реализации коррекции с учетом GC. В альтернативных подходах для коррекции погрешности GC применяют сплайн-функцию или другую нелинейную методику аппроксимации, которую можно применять в непрерывном пространстве GC и которая не включает сортировку количеств перекрытия по содержанию GC. Примеры подходящих методик включают непрерывную коррекцию локальных полиномиальных регрессий (loess) и гладкую сплайн-коррекцию. Функцию аппроксимации можно получить из нормированного от блока к блоку количества перекрытия по сравнению с содержанием GC для рассматриваемого образца. Коррекцию для каждого блока вычисляют посредством применения содержания GC для рассматриваемого блока в отношении функции аппроксимации. Например, нормированное количество перекрытия можно подогнать посредством вычитания ожидаемого значения перекрытия сплайна при содержании GC рассматриваемого блока. В качестве альтернативы, подгонку можно обеспечить посредством деления ожидаемого значения перекрытия согласно аппроксимации с помощью сплайн-функции.

После корректировки GC-зависимости в операции 319 система компьютерным способом устраняет резко отклоняющиеся блоки в рассматриваемом образце - см. блок 321. Данную операцию можно назвать фильтрованием или цензурированием единичного образца. Фиг. 3J демонстрирует, что даже после коррекции с учетом GC перекрытие все еще характеризуется образец-специфичной вариацией в пределах небольшой области. См., например, перекрытие в положении 1.1 e8 на хромосоме 12 с неожиданно высоким отклонением от ожидаемых результатов значения. Возможно, данное отклонение является следствием небольшой вариации числа копий в материнском геноме. В качестве альтернативы, отклонение может быть обусловлено техническими причинами при секвенировании, не связанными с вариацией числа копий. Как правило, данную операцию применяют исключительно в отношении устойчивых хромосом.

В качестве примера, системы компьютерным способом фильтруют любые блоки, которые содержат откорректированное с учетом GC нормированное количество перекрытия, составляющее более чем 3 медианы абсолютных отклонений от медианы откорректированного с учетом GC нормированного количества перекрытия, по всем блокам в хромосоме, несущей рассматриваемый блок, для фильтрования. В одном примере предельное значение задают как 3 медианы абсолютных отклонений, которое подгоняют для соответствия стандартному отклонению, поэтому фактически предел составляет 1,4826×медиану абсолютных отклонений от медианы. Согласно определенным вариантам реализации данную операцию применяют в отношении всех хромосом в образце, включая как устойчивые хромосомы, так и хромосомы, для которых подозревают анеуплоидию.

Согласно определенным вариантам реализации осуществляют дополнительную операцию, которую можно охарактеризовать как контроль качества. См. блок 323. Согласно некоторым вариантам реализации метрика контроля качества включает обнаружение того, являются ли какие-либо потенциальные хромосомы в знаменателе, т.е. "нормирующие хромосомы" или "устойчивые хромосомы", анеуплоидными или по другой причине не соответствующими для определения того, характеризуется ли исследуемый образец вариацией числа копий в последовательности, представляющей интерес. Когда процесс определяет, что устойчивая хромосома является не соответствующей, процесс может пренебречь исследуемым образцом и выдать результат "решение отсутствует". В качестве альтернативы, несостоятельность данной метрики КК (контроля качества) может способствовать применению альтернативного множества нормирующих хромосом для принятия решения. В одном примере способ контроля качества сравнивает фактические нормированные значения перекрытия для устойчивых хромосом с ожидаемыми значениями

для устойчивых аутосомных хромосом. Ожидаемые значения можно получить посредством аппроксимации многомерной нормальной модели к нормированным профилям непораженных обучающих образцов, выбора наилучшей структуры модели согласно правдоподобию данных или байесовского критерия (например, модель выбрана с применением информационного критерия Акаике или, возможно, байесовского информационного критерия), и фиксации наилучшей модели для применения в КК. Нормальные модели устойчивых хромосом можно получить посредством, например, применения приемов группирования, идентифицирующих функцию вероятности, которая характеризуется средним значением и стандартным отклонением для перекрытий хромосомы в нормальных образцах. Разумеется, можно применять другие формы модели. Процесс оценивает правдоподобие наблюдаемого нормированного перекрытия в любом исследуемом образце на входе, принимая во внимание параметры фиксированной модели. Процесс может выполнять данную функцию посредством оценки каждого исследуемого образца на входе с моделью для получения правдоподобия и посредством этого идентифицировать резко отклоняющиеся показатели по сравнению с множеством нормальных образцов. Отклонение правдоподобия исследуемого образца от таковой обучающих образцов может свидетельствовать об аномалии в нормирующих хромосомах или артефакте при обращении с образцом/при подготовке к анализу, который может привести к неправильному классифицированию образца. Данную метрику КК можно применять для снижения ошибок в классификации, связанных с любым из данных артефактов образца. На фиг. 3K, правый чертеж, на оси x представлено количество хромосом, а на оси y представлено нормированное перекрытие хромосом, основанное на сравнении с моделью КК, полученной, как описано выше. Графики демонстрируют один образец с избыточным перекрытием для хромосомы 2 и другой образец с избыточным перекрытием для хромосомы 20. Данные образцы будут устранены с применением метрики КК, описанной в настоящем документе, или отклонены для применения альтернативного множества нормирующих хромосом. На левом чертеже фиг. 3K представлено NCV по сравнению с правдоподобием для хромосомы.

Последовательность, представленную на фиг. 3A, можно применять для всех блоков всех хромосом в геноме. Согласно определенным вариантам реализации для хромосомы Y применяют отличный процесс. Для вычисления дозы хромосомы или сегмента NCV и/или NSV применяют откорректированные нормированные количества перекрытия (как определено на фиг. 3A) из блоков в хромосомах или сегментах, использованных в выражениях для дозы, NCV и/или NSV. См. блок 325. Согласно определенным вариантам реализации среднее значение нормированного количества перекрытия вычисляют по всем блокам в хромосоме, представляющей интерес, нормирующей хромосоме, сегменте, представляющем интерес, и/или для вычисления дозы последовательности, NCV и/или NSV, применяют нормирующий сегмент, как описано в другом месте в настоящем документе.

Согласно определенным вариантам реализации хромосому Y обрабатывают иным способом. Хромосому Y можно фильтровать посредством маскирования множества блоков, уникальных для Y-хромосомы. Согласно некоторым вариантам реализации фильтр Y-хромосомы определяют согласно процессу, описанному в предварительной заявке на патент США № 61/836057, ранее включенной в настоящий документ посредством ссылки. Согласно некоторым вариантам реализации фильтр маскирует блоки, которые являются меньшими, чем таковые в фильтре другой хромосомы. Например, маска Y-хромосомы может фильтровать на уровне 1 т.о., тогда как маски другой хромосомы могут фильтровать на уровне 100 т.о. Несмотря на это Y-хромосому можно нормировать в том же блоке размера, что и другие хромосомы (например, 100 т.о.).

Согласно определенным вариантам реализации отфильтрованную Y-хромосому нормируют, как описано выше в операции 315 фиг. 3A. Однако, в отличие от указанной операции, Y-хромосому дополнительно не корректируют. Таким образом, в блоках Y-хромосомы не устраняют глобальный профиль. Аналогично, блоки Y-хромосомы не подвергают коррекции с учетом GC или другим этапам фильтрования, которые выполняют впоследствии. Это обусловлено тем, что, когда образец процессируют, процессу не известно, является ли образец образцом мужского или женского пола. Образец женского пола не должен характеризоваться ридами, выравнивающимися с референсной Y-хромосомой.

Создание маски последовательности.

В некоторых вариантах реализации, раскрытых в настоящем документе, применяют стратегию фильтрования (или маскирования) недискриминантных ридов последовательности на последовательности, представляющей интерес, с применением масок последовательности, что приводит к увеличению сигнала и снижению шума по сравнению со значениями, вычисленными общепринятыми способами, в значениях перекрытия, применяемых для оценки ВЧК. Такие маски можно идентифицировать посредством различных методик. Согласно одному варианту реализации маску идентифицируют с применением методики, проиллюстрированной на фиг. 4A-4B, как объяснено более подробно ниже.

Согласно некоторым вариантам реализации маску идентифицируют с применением обучающего множества репрезентативных образцов, которые установленно содержат нормальное число копий последовательности, представляющей интерес. Маски можно идентифицировать с применением методики, которая сначала нормирует обучающее множество образцов, затем проводит корректировку с учетом систематической вариации в диапазоне последовательности (например, профиля), а затем корректирует их с учетом вариабельности GC, как описано ниже. Нормирование и коррекцию осуществляют в отно-

шении образцов из обучающего множества, а не исследуемых образцов. Маску идентифицируют один раз, а затем применяют в отношении множества исследуемых образцов.

На фиг. 4А представлена блок-схема процесса 400 для создания такой маски последовательности, которую можно применять в отношении одного или более исследуемых образцов для устранения из рассмотрения блоков на последовательности, представляющей интерес, при оценке числа копий. Процесс 400, проиллюстрированный на фиг. 4, применяет перекрытие метки последовательности, основанное на количестве меток последовательности, для получения маски последовательности. Однако аналогично приведенному выше описанию процесса 100 для определения ВЧК применительно к фиг. 1, для процесса 400 в дополнение к перекрытию или вместо него можно применять другие переменные или параметры, такие как размер, соотношение размера и уровень метилирования. Согласно некоторым вариантам реализации для каждого из двух или более параметров получают одну маску. Более того, перекрытие и другие параметры можно взвесить по размеру фрагментов, из которых были получены метки. Для удобства чтения в процессе 400 упомянуто исключительно перекрытие, но следует отметить, что вместо перекрытия можно применять другие параметры, такие как размер, соотношение размера и уровень метилирования, подсчет, взвешенный по размеру, и т.д.

Процесс 400 начинается с обеспечения обучающего множества, содержащего риды последовательности из множества непораженных обучающих образцов. Блок 402. Затем процесс выравнивает риды последовательности обучающего множества с референсным геномом, содержащим последовательность, представляющую интерес, с получением, таким образом, меток обучающей последовательности для обучающих образцов. Блок 404. Согласно некоторым вариантам реализации для последующего анализа применяют исключительно уникально выровненные не повторяющиеся метки, картированные с неисключенными сайтами. Процесс включает разделение референсного генома на множество блоков и определение для каждого непораженного обучающего образца перекрытия меток обучающей последовательности в каждом блоке для каждого обучающего образца. Блок 406. Процесс также определяет для каждого блока ожидаемое перекрытие меток обучающей последовательности среди всех обучающих образцов. Блок 408. Согласно некоторым вариантам реализации ожидаемое перекрытие каждого блока представляет собой медиану или средние значения в пределах обучающих образцов. Ожидаемые перекрытия составляют глобальный профиль. Затем процесс подгоняет перекрытие меток обучающей последовательности в каждом блоке для каждого обучающего образца посредством устранения вариации в глобальном профиле, и посредством этого получает откорректированные с учетом глобального профиля перекрытия меток обучающей последовательности в блоках для каждого обучающего образца. Затем процесс создает маску последовательности, содержащую немаскированные и маскированные блоки в пределах референсного генома. Каждый маскированный блок обладает характеристикой распределения, превышающей порог маскирования. Характеристика распределения предложена для подогнанных перекрытий меток обучающей последовательности в блоке в пределах обучающих образцов. Согласно некоторым вариантам реализации порог маскирования может относиться к наблюдаемой вариации нормированного перекрытия в блоке в пределах обучающих образцов. Блоки с высокими коэффициентами вариации или медианой абсолютного отклонения нормированного перекрытия среди образцов можно идентифицировать на основании эмпирического распределения соответствующих метрик. Согласно некоторым альтернативным вариантам реализации порог маскирования может относиться к наблюдаемой вариации в нормированном перекрытии в блоке в пределах обучающих образцов. Блоки с высокими коэффициентами вариации или медианой абсолютного отклонения нормированного перекрытия среди образцов можно маскировать на основании эмпирического распределения соответствующих метрик.

Согласно некоторым вариантам реализации для хромосомы, представляющей интерес, и для всех других хромосом задают отдельные пределы для идентификации маскированных блоков, т.е. пороги маскирования. Также можно задать отдельные пороги маскирования для каждой хромосомы, представляющей интерес, отдельно, и один порог маскирования для множества всех непораженных хромосом. В качестве примера для хромосомы 13 задают маску на основании определенного порога маскирования, и с целью определения маски для других хромосом применяют другой порог маскирования. Непораженные хромосомы могут также характеризоваться своими порогами маскирования, заданными для хромосомы.

Можно оценить различные комбинации порога маскирования для каждой хромосомы, представляющей интерес. Комбинации порога маскирования обеспечивают одну маску для блоков хромосомы, представляющей интерес, и отличную маску для блоков всех других хромосом.

Согласно одному подходу диапазон значений для коэффициента вариации (КВ) или критерий пределов распределения образца задают в виде процентилей (например, 95, 96, 97, 98, 99) эмпирического распределения значений КВ блока, и данные значения предела применяют в отношении всех аутосом, за исключением хромосом, представляющих интерес. Также задают диапазон процентиля значений предела для КВ для эмпирического распределения КВ, и данные значения предела применяют в отношении хромосомы, представляющей интерес (например, хр. 21). Согласно некоторым вариантам реализации хромосомы, представляющие интерес, представляют собой Х-хромосому и хромосомы 13, 18 и 21. Разумеется, можно принимать во внимание другие подходы; например, для каждой хромосомы можно осуществить отдельную оптимизацию. Взятые вместе, диапазоны, которые подлежат оптимизации параллельно (на-

пример, один диапазон для рассматриваемой хромосомы, представляющей интерес, и другой диапазон для всех других хромосом), определяют сеть комбинаций предела КВ. См. фиг. 4В. Рабочие характеристики системы в отношении обучающего множества оценивают в двух пределах (один - для нормирующих хромосом (или аутосом, отличных от хромосомы, представляющей интерес) и второй - для хромосомы, представляющей интерес), и для итоговой конфигурации выбирают демонстрирующую наилучшие рабочие характеристики комбинацию. Данная комбинация может являться отличной для каждой из хромосом, представляющих интерес. Согласно определенным вариантам реализации рабочие характеристики оценивают на валидационном множестве вместо обучающего множества, а именно, для оценки рабочих характеристик применяют перекрестную валидацию.

Согласно некоторым вариантам реализации рабочие характеристики, оптимизированные для определения диапазонов предела, представляют собой коэффициент вариации доз хромосом (основанный на экспериментальном выборе нормирующих хромосом). Процесс выбирает комбинацию пределов, которая минимизирует КВ дозы хромосомы (например, соотношение) для хромосомы, представляющей интерес, с применением выбранной на сегодняшний день нормирующей хромосомы (или хромосом). Согласно одному подходу процесс исследует рабочие характеристики каждой комбинации пределов в сети следующим образом: (1) применяет комбинацию пределов для определения масок для всех хромосом и применяет данные маски для фильтрации меток обучающего множества; (2) вычисляет нормированные перекрытия в пределах обучающего множества непораженных образцов посредством применения процесса фиг. 3А в отношении отфильтрованных меток; (3) определение репрезентативного нормированного перекрытия на хромосому посредством, например, суммирования нормированных перекрытий блока для рассматриваемой хромосомы; (4) вычисляет дозы хромосом с применением выбранных на сегодняшний день нормирующих хромосом, и (5) определение КВ доз хромосом. Процесс может оценивать рабочие характеристики выбранных фильтров посредством применения их в отношении исследуемых образцов, отделенных от исходной части обучающего множества. То есть процесс разделяет исходное обучающее множество на обучающее и исследуемое подмножества. Обучающее подмножество применяют для определения пределов маски, как описано выше.

Согласно альтернативным вариантам реализации вместо определения масок, основанного на КВ перекрытий, маски можно задать посредством распределения показателей качества картирования из результатов выравнивания в пределах обучающих образцов в блоках. Показатель качества картирования отражает уникальность, с которой рид картируется на референсный геном. Другими словами, показатели качества картирования количественно определяют вероятность того, что рид неправильно выровнен. Низкий показатель качества картирования связан с низкой уникальностью (высокой вероятностью неправильного выравнивания). Уникальность соответствует одной или нескольким ошибкам в риде последовательности (полученной посредством секвенатора). Подробное описание показателей качества картирования можно найти в публикации Li H., Ruan J., Durbin R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Research 18:1851-8, которая полностью включена в настоящий документ посредством ссылки. Согласно некоторому варианту реализации показатель качества картирования в настоящем документе называют показателем MapQ. Фиг. 4В демонстрирует, что показатель MapQ характеризуется устойчивой монотонной корреляцией с КВ процессированных перекрытий. Например, блоки с КВ выше 0,4 практически полностью группируются в левой части графика на фиг. 4В, характеризуясь показателями MapQ ниже приблизительно 4. Вследствие этого, маскирование блоков с небольшим MapQ может позволить получить маску, вполне аналогичную таковой, заданной посредством маскирования блоков с высоким КВ.

Образцы и процессинг образцов.

Образцы.

Образцы, которые применяют для определения ВЧК, например анеуплоидий хромосом, частичных анеуплоидий и т.п., могут включать образцы, отобранные от любой клетки, ткани или органа, в которых необходимо определить вариации числа копий для одной или более последовательностей, представляющих интерес. Предпочтительно, образцы содержат нуклеиновые кислоты, которые присутствуют в клетках, и/или нуклеиновые кислоты, которые являются "бесклеточными" (например, сцДНК).

Согласно некоторым вариантам реализации предпочтительным является получение бесклеточных нуклеиновых кислот, например бесклеточной ДНК (сцДНК). Бесклеточные нуклеиновые кислоты, включая бесклеточную ДНК, можно получить различными способами, известными в данной области техники, из биологических образцов, включая, без ограничения, плазму, сыворотку и мочу (см., например, публикации Fan et al., Proc Natl Acad Sci 105:16266-16271 [2008]; Koide et al., Prenatal Diagnosis 25:604-607 [2005]; Chen et al., Nature Med. 2: 1033-1035 [1996]; Lo et al., Lancet 350: 485-487 [1997]; Botezatu et al., Clin Chem. 46: 1078-1084, 2000; и Su et al., J Mol. Diagn. 6: 101-107 [2004]). Для отделения бесклеточной ДНК от клеток в образце можно применять различные способы, включая, без ограничения, фракционирование, центрифугирование (например, центрифугирование в градиенте плотности), ДНК-специфичную преципитацию или высокопроизводительную сортировку клеток и/или другие способы разделения. Существуют коммерчески доступные наборы для ручного и автоматического разделения сцДНК (Roche Diagnostics, Индианаполис, Индиана, Qiagen, Валенсия, Калифорния, Macherey-Nagel, Дюрен, Делавер).

Биологические образцы, содержащие сцДНК, применяли в анализах для определения присутствия или отсутствия аномалий хромосом, например, трисомии 21, посредством анализов секвенирования, которые могут обнаружить анеуплоидии и/или различные полиморфизмы хромосом.

Согласно различным вариантам реализации сцДНК, присутствующую в образце, перед применением можно обогатить, специфично или неспецифично (например, перед получением библиотеки секвенирования). Неспецифичное обогащение образца ДНК означает амплификацию целого генома фрагментов геномной ДНК образца, которую можно применять для увеличения уровня образца ДНК перед получением библиотеки секвенирования сцДНК. Неспецифичное обогащение может представлять собой селективное обогащение одного из двух геномов, присутствующих в образце, который содержит более одного генома. Например, неспецифичное обогащение может являться селективным в отношении генома плода в материнском образце, которое можно получить известными способами для увеличения относительной доли ДНК плода по сравнению с материнской ДНК в образце. В качестве альтернативы, неспецифичное обогащение может представлять собой неселективную амплификацию обоих геномов, присутствующих в образце. Например, можно осуществить неспецифичную амплификацию ДНК плода и материнской ДНК в образце, содержащем смесь ДНК из геномов плода и матери. Способы амплификации целого генома известны в данной области техники. ПЦР с дегенеративными олигонуклеотидными праймерами (Degenerate oligonucleotide-primed PCR, DOP), методика ПЦР с достройкой праймера (primer extension PCR, PEP) и амплификация с множественным вытеснением цепи (multiple displacement amplification, MDA) являются примерами способов амплификации целого генома. Согласно некоторым вариантам реализации образец, содержащий смесь сцДНК из различных геномов, является необогащенным сцДНК геномов, присутствующих в смеси. Согласно другим вариантам реализации образец, содержащий смесь сцДНК из различных геномов, неспецифично обогащен любым из геномов, присутствующих в образце.

Образец, содержащий нуклеиновую кислоту или кислоты, в отношении которых применяют способы, описанные в настоящем документе, как правило, включает биологический образец ("исследуемый образец"), например, как описано выше. Согласно некоторым вариантам реализации нуклеиновую кислоту или кислоты, скрининг которых проводят в отношении одной или более ВЧК, очищают или выделяют любым из множества хорошо известных способов.

Соответственно, согласно определенным вариантам реализации образец содержит или состоит из очищенного или выделенного полинуклеотида или может включать образцы, такие как образец ткани, образец биологической жидкости, образец клетки и т.п. Подходящие образцы биологической жидкости включают, без ограничения, кровь, плазму, сыворотку, пот, слезы, мокроту, мочу, слюну, ушную жидкость, лимфу, слюну, спинномозговую жидкость, жидкость после лаважа, суспензию костного мозга, влагалищную жидкость, жидкость после трансцервикального лаважа, жидкость головного мозга, асцит, молоко, секреты дыхательных, кишечных и мочеполовых путей, амниотическую жидкость, молоко и образцы лейкафереза. Согласно некоторым вариантам реализации образец представляет собой образец, который с легкостью получают в результате неинвазивных процедур, например, кровь, плазму, сыворотку, пот, слезы, мокроту, мочу, мокроту, ушную жидкость, слюну или фекалии. Согласно определенным вариантам реализации образец представляет собой образец периферической крови или фракцию плазмы и/или сыворотки образца периферической крови. Согласно другим вариантам реализации биологический образец представляет собой мазок или соскоб, образец биопсии или культуру клеток. Согласно другому варианту реализации образец представляет собой смесь двух или более биологических образцов, например, биологический образец может содержать два или более образцов биологической жидкости, образцов ткани и образцов культуры клеток. В настоящем документе термины "кровь", "плазма" и "сыворотка" однозначно включают фракции или процессированные части указанных образцов. Аналогично, когда образец отбирают из биопсии, мазка, соскоба и т.д., "образец" однозначно включает процессированную фракцию или часть, полученную из биопсии, мазка, соскоба и т.д.

Согласно определенным вариантам реализации образцы можно получить из источников, включая, без ограничения, образцы от различных индивидуумов, образцы от различных этапов развития одного и того же или различных индивидуумов, образцы от различных страдающих от заболевания индивидуумов (например, индивидуумов, страдающих от рака, или индивидуумов, которые, как подозревают, страдают от генетического нарушения), от нормальных индивидуумов, образцы, полученные на различных стадиях заболевания индивидуума, образцы, полученные от индивидуума, который получает различные варианты лечения заболевания, образцы от индивидуумов, которые подвергаются воздействию различных факторов окружающей среды, образцы от индивидуумов с предрасположенностью к патологии, образцы от индивидуумов, которые подвергаются воздействию возбудителя инфекционного заболевания (например, ВИЧ), и т.п.

Согласно одному иллюстративному, но неограничивающему варианту реализации образец представляет собой материнский образец, который получают от беременного субъекта женского пола, например, беременной женщины. В данном случае образец можно анализировать с применением способов, описанных в настоящем документе, с обеспечением пренатальной диагностики потенциальных хромосомных аномалий у плода. Материнский образец может представлять собой образец ткани, образец биологической жидкости или образец клетки. Биологическая жидкость включает, в качестве неограничи-

вающих примеров, кровь, плазму, сыворотку, пот, слезы, мокроту, мочу, мокроту, ушную жидкость, лимфу, слюну, спинномозговую жидкость, жидкость после лаважа, суспензию костного мозга, влагалищную жидкость, жидкость после трансцервикального лаважа, жидкость головного мозга, асцит, молоко, секреты дыхательных, кишечных и мочеполовых путей и образцы лейкафереза.

Согласно другому иллюстративному, но неограничивающему варианту реализации материнский образец представляет собой смесь двух или более биологических образцов, например биологический образец может содержать два или более образцов биологической жидкости, образцов ткани и образцов культуры клеток. Согласно некоторым вариантам реализации образец представляет собой образец, который с легкостью получают в результате неинвазивных процедур, например кровь, плазму, сыворотку, пот, слезы, мокроту, мочу, молоко, мокроту, ушную жидкость, слюну и фекалии. Согласно некоторым вариантам реализации биологический образец представляет собой образец периферической крови и/или фракцию плазмы и сыворотки указанного образца. Согласно другим вариантам реализации биологический образец представляет собой мазок или соскоб, образец биопсии или образец культуры клеток. Как раскрыто выше, "кровь", "плазма" и "сыворотка" однозначно включают фракции или процессированные части указанных образцов. Аналогично, когда образец отбирают из биопсии, мазка, соскоба и т.д., "образец" однозначно включает процессированную фракцию или часть, полученную из биопсии, мазка, соскоба и т.д.

Согласно определенным вариантам реализации образцы также можно получить из культивируемых in vitro тканей, клеток или других источников, содержащих полинуклеотиды. Культивируемые образцы могут быть получены из источников, включая, без ограничения, культуры (например, ткани или клеток), поддерживаемые в различных средах и условиях (например, pH, давление или температура), культуры (например, ткани или клеток), поддерживаемые в течение различных периодов времени, культуры (например, ткани или клеток), обработанные различными факторами или реактивами (например, потенциальным лекарственным средством или модулятором), или культуры различных типов ткани и/или клеток.

Способы выделения нуклеиновых кислот из биологических источников хорошо известны и будут отличаться в зависимости от природы источника. Специалист в данной области техники может с легкостью выделить нуклеиновую кислоту или кислоты из источника, как требуется для способа, описанного в настоящем документе. В некоторых случаях может характеризоваться преимуществом фрагментация молекулы нуклеиновой кислоты в образце нуклеиновой кислоты. Фрагментация может быть случайной или специфичной, которую достигают, например, с применением расщепления рестрикционными эндонуклеазами. Способы случайной фрагментации хорошо известны в данной области техники и включают, например, ограниченное расщепление ДНКазой, обработку щелочью и физическое разрезание. Согласно одному варианту реализации образец нуклеиновых кислот получают из сцДНК, которую не подвергают фрагментации.

Получение библиотеки секвенирования.

Согласно одному варианту реализации в способах, описанных в настоящем документе, можно применять технологии секвенирования нового поколения (СНП), позволяющие секвенировать множество образцов по отдельности в виде геномных молекул (т.е. сингплексное секвенирование) или в виде объединенных образцов, содержащих индексированные геномные молекулы (например, мультиплексное секвенирование), в ходе одной серии секвенирования. Данные способы могут позволить получить вплоть до нескольких сотен миллионов ридов последовательностей ДНК. Согласно различным вариантам реализации последовательности геномных нуклеиновых кислот и/или индексированных геномных нуклеиновых кислот можно определить с применением, например, технологий секвенирования нового поколения (СНП), описанных в настоящем документе. Согласно различным вариантам реализации анализ значительного количества данных последовательности, полученных с применением СНП, можно осуществить с применением одного или более процессоров, как описано в настоящем документе.

Согласно различным вариантам реализации применение таких технологий секвенирования не включает получение библиотек секвенирования.

Однако согласно определенным вариантам реализации способы секвенирования, предусмотренные в настоящем документе, включают получение библиотек секвенирования. Согласно одному иллюстративному подходу получение библиотеки секвенирования включает получение случайного набора модифицированных адаптерами фрагментов ДНК (например, полинуклеотидов), уже готовых к секвенированию. Библиотеки секвенирования полинуклеотидов можно получить из ДНК или РНК, включая эквиваленты, аналоги ДНК или кДНК, например ДНК или кДНК, которая является комплементарной или представляет собой копию ДНК, полученной с матрицы РНК под действием обратной транскриптазы. Полинуклеотиды можно получить в двухцепочечной форме (например, дсДНК, такая как фрагменты геномной ДНК, кДНК продукты ПНР-амплификации и т.п.), или согласно определенным вариантам реализации полинуклеотиды можно получить в одноцепочечной форме (например, осДНК, РНК и т.д.) и можно преобразовать в форму дсДНК. В качестве примера, согласно определенным вариантам реализации одноцепочечные молекулы мРНК можно копировать в двухцепочечные кДНК, подходящие для применения при получении библиотеки секвенирования. Точная последовательность первичных полинуклеотидных молекул, как правило, не имеет значения для способа получения библиотеки и может быть извест-

ной или не известной. Согласно одному варианту реализации полинуклеотидные молекулы представляют собой молекулы ДНК. Более конкретно, согласно определенным вариантам реализации полинуклеотидные молекулы представляют весь генетический набор организма или по существу весь генетический набор организма и представляют собой геномные молекулы ДНК (например, клеточную ДНК, бесклеточные ДНК (сцДНК) и т.д.), которые, как правило, включают как последовательность интрона, так и последовательность экзона (кодирующую последовательность), а также некодирующие регуляторные последовательности, такие как последовательности промотора и энхансера. Согласно определенным вариантам реализации первичные полинуклеотидные молекулы содержат молекулы ДНК генома человека, например, молекулы сцДНК, присутствующие в периферической крови беременного субъекта.

Получение библиотек секвенирования для некоторых платформ секвенирования СНП облегчается посредством применения полинуклеотидов, содержащих конкретный диапазон размеров фрагментов. Получение таких библиотек, как правило, включает фрагментацию больших полинуклеотидов (например, клеточной геномной ДНК) для получения полинуклеотидов желаемого диапазона размера.

Фрагментацию можно обеспечить любым из множества способов, известных специалистам в данной области техники. Например, фрагментацию можно обеспечить механическими способами, включая, без ограничения, пульверизацию, обработку ультразвуком и гидросдвиг. Однако механическая фрагментация, как правило, расщепляет остов ДНК по связям C-O, P-O и C-C, что приводит к получению гетерогенной смеси тупых и 3'- и 5'-выступающих концов с разорванными связями C-O, P-O и C-C (см., например, публикации Alnemri and Liwack, J Biol. Chem 265:17323-17333 [1990]; Richards and Boyer, J Mol Biol 11:327-240 [1965]), которые, возможно, необходимо восстановить, поскольку в них может отсутствовать необходимый для последующих ферментативных реакций 5'-фосфат, например, для лигирования адаптеров секвенирования, которые необходимы для подготовки ДНК к секвенированию.

Напротив, сцДНК, как правило, существует в виде фрагментов размером менее приблизительно 300 пар оснований и, следовательно, для получения библиотеки секвенирования с применением образцов сцДНК фрагментация, как правило, не требуется.

Как правило, вне зависимости от того, были ли полинуклеотиды фрагментированы принудительно (например, фрагментированы in vitro), или существуют в природе в виде фрагментов, их преобразуют в ДНК с тупыми концами, содержащую 5'-фосфаты и 3'-гидроксильную группу. Стандартные протоколы, например, протоколы для секвенирования с применением, например, платформы Illumina, описанные в другом месте в настоящем документе, инструктируют пользователей восстанавливать концы образца ДНК, очищать продукты с восстановленными концами перед присоединением dA-"хвоста" и очищать продукты с присоединением dA-"хвоста" перед этапами лигирования адаптеров для получения библиотеки.

В различных вариантах реализации способов получения библиотеки последовательности, описанных в настоящем документе, избегают необходимости осуществлять один или более этапов, как правило, предписываемых стандартными протоколами для получения модифицированного продукта ДНК, который можно секвенировать посредством СНП. Сокращенный (abbreviated, ABB) способ, 1-этапный способ и 2-этапный способ являются примерами способов получения библиотеки секвенирования, которые можно найти в заявке на патент 13/555037 поданной 20 июля 2012 года, которая полностью включена в настоящей документ посредством ссылки.

Маркерные нуклеиновые кислоты для отслеживания и подтверждения целостности образца.

Согласно различным вариантам реализации подтверждение целостности образцов и отслеживание образца можно осуществить посредством секвенирования смесей образца геномных нуклеиновых кислот, например, сцДНК, и сопутствующих маркерных нуклеиновых кислот, которые были введены в образцы, например, перед процессингом.

Маркерные нуклеиновые кислоты можно объединить с исследуемым образцом (например, образцом из биологического источника) и подвергнуть процессам, которые включают, например, один или более этапов фракционирования образца из биологического источника, например получение по существу бесклеточной фракции плазмы из образца цельной крови, очистку нуклеиновых кислот из фракционированного образца, например, плазмы, или нефракционированного образца из биологического источника, например, образца ткани, и секвенирование. Согласно некоторым вариантам реализации секвенирование включает получение библиотеки секвенирования. Последовательность или комбинацию последовательностей маркерных молекул, которые объединяют с образцом из источника, выбирают так, чтобы она была уникальной в отношении образца из источника. Согласно некоторым вариантам реализации все уникальные маркерные молекулы в образце содержат одну и ту же последовательность. Согласно другим вариантам реализации уникальные маркерные молекулы в образце представляют собой множество последовательностей, например, комбинацию двух, трех, четырех, пяти, шести, семи, восьми, девяти, десяти, пятнадцати, двадцати или более различных последовательностей.

Согласно одному варианту реализации целостность образца можно подтвердить с применением множества маркерных молекул нуклеиновой кислоты, которые содержат идентичные последовательности. В качестве альтернативы, подлинность образца можно подтвердить с применением множества маркерных молекул нуклеиновой кислоты, которые содержат по меньшей мере две, по меньшей мере три, по

меньшей мере четыре, по меньшей мере пять, по меньшей мере шесть, по меньшей мере семь, по меньшей мере восемь, по меньшей мере девять, по меньшей мере десять, по меньшей мере 11, по меньшей мере 12, по меньшей мере 13, по меньшей мере 14, по меньшей мере 15, по меньшей мере 16, по меньшей мере 17, по меньшей мере 18, по меньшей мере 19, по меньшей мере 20, по меньшей мере 25, по меньшей мере 30, по меньшей мере 35, по меньшей мере 40, по меньшей мере 50 или более различных последовательностей. Для подтверждения целостности множества биологических образцов, т.е. двух или более биологических образцов, требуется, чтобы каждый из двух или более образцов был маркирован маркерными нуклеиновыми кислотами, которые содержат последовательности, уникальные для каждого из множества исследуемых образцов, которые являются маркированными. Например, первый образец может быть маркирован маркерной нуклеиновой кислотой, содержащей последовательность А, и второй образец может быть маркирован маркерной нуклеиновой кислотой, содержащей последовательность В. В качестве альтернативы, первый образец может быть маркирован молекулами маркерной нуклеиновой кислоты, все из которых содержат последовательность А, и второй образец может быть маркирован смесью последовательностей В и С, причем последовательности А, В и С представляют собой маркерные молекулы, содержащие различные последовательности.

Маркерную нуклеиновую кислоту или кислоты можно добавить к образцу на любом этапе получения образца, который происходит перед получением библиотеки (если необходимо получить библиотеки) и секвенирования. Согласно одному варианту реализации маркерные молекулы можно объединить с непроцессированным образцом из источника. Например, маркерная нуклеиновая кислота может быть обеспечена в пробирке для сбора образцов, которую применяют для забора образца крови. В качестве альтернативы, маркерные нуклеиновые кислоты можно добавить к образцу крови после забора крови. Согласно одному варианту реализации маркерную нуклеиновую кислоту добавляют в сосуд, который применяют для сбора образца биологической жидкости, например маркерную нуклеиновую кислоту или кислоты добавляют в пробирку для забора крови, которую применяют для забора образца крови. Согласно другому варианту реализации маркерную нуклеиновую кислоту или кислоты добавляют во фракцию образца биологической жидкости. Например, маркерную нуклеиновую кислоту добавляют во фракцию плазмы и/или сыворотки образца крови, например, в образец материнской плазмы. Согласно еще одному варианту реализации маркерные молекулы добавляют в очищенный образец, например, образец нуклеиновых кислот, которые были очищены из биологического образца. Например, маркерные нуклеиновые кислоты добавляют в образец очищенной материнской и плодной сцДНК. Аналогично, маркерные нуклеиновые кислоты можно добавить в образец биопсии перед процессированием образца. Согласно некоторым вариантам реализации маркерные нуклеиновые кислоты можно объединить с носителем, который доставляет маркерные молекулы в клетки биологического образца. Носители для доставки клеток включают pH-чувствительные и катионные липосомы.

Согласно различным вариантам реализации маркерные молекулы содержат антигеномные последовательности, которые представляют собой последовательности, отсутствующие в геноме образца из биологического источника. Согласно иллюстративному варианту реализации маркерные молекулы, которые применяют для подтверждения целостности образца из биологического источника человека, содержат последовательности, отсутствующие в геноме человека. Согласно альтернативному варианту реализации маркерные молекулы содержат последовательности, которые отсутствуют в образце из источника и в любом одном или более других известных геномов. Например, маркерные молекулы, которые применяют для подтверждения целостности образца из биологического источника человека, содержат последовательности, отсутствующие в геноме человека и в геноме мыши. Альтернативный вариант позволяет подтверждать целостность исследуемого образца, который содержит два или более геномов. Например, целостность образца бесклеточной ДНК человека, полученного от субъекта, пораженного патогеном, например, бактерией, можно подтвердить с применением маркерных молекул, которые содержат последовательности, отсутствующие как в геноме человека, так и в геноме поражающей бактерии. Последовательности геномов многочисленных патогенов, например, бактерий, вирусов, дрожжей, грибов, простейших и т.д., являются общедоступными в сети Интернет по адресу: ncbi.nlm.nih.gov/genomes. Согласно другому варианту реализации маркерные молекулы представляют собой нуклеиновые кислоты, которые содержат последовательности, отсутствующие в любом известном геноме. Последовательности маркерных молекул можно получить случайным образом алгоритмически.

Согласно различным вариантам реализации маркерные молекулы могут представлять собой встречающиеся в природе дезоксирибонуклеиновые кислоты (ДНК), рибонуклеиновые кислоты или искусственные аналоги нуклеиновой кислоты (миметики нуклеиновой кислоты), включая пептидные нуклеиновые кислоты (ПНК), морфолиновые нуклеиновые кислоты, запертые нуклеиновые кислоты, гликолевые нуклеиновые кислоты и треозные нуклеиновые кислоты, которые отличаются от встречающихся в природе ДНК или РНК изменениями в остове молекулы, или миметики ДНК, которые не содержат фосфодиэфирный остов. Дезоксирибонуклеиновые кислоты могут происходить из встречающихся в природе геномов или могут быть получены в лаборатории посредством применения ферментов или посредством твердофазного химического синтеза. Химические способы также можно применять для получения миметиков ДНК, не обнаруженных в природе. Доступны производные ДНК, в которых фосфодиэфирная связь

была заменена, но в которых сохранена дезоксирибоза, и которые включают, без ограничения, миметики ДНК, содержащие остовы, образованные триоформацетальной или карбоксамидной связью, которые, как было показано, являются хорошими структурными миметиками ДНК. Другие миметики ДНК включают морфолиновые производные и пептидные нуклеиновые кислоты (ПНК), которые содержат псевдопептидный остов на основе N-(2-аминоэтил)глицина (Ann Rev Biophys Biomol Struct 24:167-183 [1995]). ПНК представляет собой чрезвычайно хороший структурный миметик ДНК (или рибонуклеиновой кислоты [РНК]), и олигомеры ПНК способны образовать весьма стабильные дуплексные структуры с комплементарными согласно принципу Уотсона-Крика олигомерами ДНК и РНК (или ПНК), и могут также связываться с мишенями в дуплексной ДНК посредством внедрения в спираль (Mol Biotechnol 26:233-248 [2004]. Другой хороший структурный миметик/аналог ДНК, который можно применять в качестве маркерной молекулы, представляет собой фосфотиоатную ДНК, в которой один из немостиковых кислородов заменен серой. Данная модификация снижает действие эндо- и экзонуклеаз 2, включая от5'-3'и3'-5' ДНК POL 1 экзонуклеазу, нуклеазы S1 и P1, РНКазы, сывороточные нуклеазы и фосфодиэстеразу змеиного яда.

Длина маркерных молекул может являться отличной или такой же, как длина нуклеиновых кислот образца, т.е. длина маркерных молекул может являться аналогичной таковой геномных молекул образца или может быть большей или меньшей, чем таковая геномных молекул образца. Длину маркерных молекул измеряют по количеству оснований нуклеотидов или аналогов нуклеотидов, которые составляют маркерную молекулу. Маркерные молекулы, длины которых отличаются от таковых геномных молекул образца, можно отличить от нуклеиновых кислот из источника с применением способов разделения, известных в данной области техники. Например, различия в длине молекул маркерных нуклеиновых кислот и нуклеиновых кислот образца можно определить посредством электрофоретического разделения, например, капиллярного электрофореза. Установление отличий в размере может характеризоваться преимуществом для количественного определения и оценки качества маркера и нуклеиновых кислот образца. Предпочтительно, маркерные нуклеиновые кислоты являются более короткими, чем геномные нуклеиновые кислоты, и характеризуются достаточной длиной, чтобы исключить их из картирования на геном образца. Например, необходима последовательность человека длиной 30 оснований, чтобы уникально картировать ее на геном человека. Соответственно, согласно определенным вариантам реализации маркерные молекулы, которые применяют в биоанализах секвенирования образцов человека, должны составлять по меньшей мере 30 п.о. в длину.

Выбор длины маркерной молекулы определяют преимущественно с применением технологии секвенирования, которую используют для подтверждения целостности образца из источника. Также можно принимать во внимание длину геномных нуклеиновых кислот образца, секвенирование которого проводят. Например, в некоторых технологиях секвенирования применяют клональную амплификацию полинуклеотидов, для которой может требоваться, чтобы геномные полинуклеотиды, которые необходимо клонально амплифицировать, характеризовались минимальной длиной. Например, секвенирование с применением анализатора последовательностей Illumina GAII включает клональную амплификацию in vitro методом мостиковой ПЦР (также известна как кластерная амплификация) полинуклеотидов, которые характеризуются минимальной длиной 110 п.о., с которыми лигируют адаптеры с получением нуклеиновой кислоты размером по меньшей мере 200 п.о. и менее 600 п.о., которую можно клонально амплифицировать и секвенировать. Согласно некоторым вариантам реализации длина лигированной с адаптерами маркерной молекулы составляет от приблизительно 200 п.о. до приблизительно 600 п.о., от приблизительно 250 п.о. до 550 п.о., от приблизительно 300 п.о. до 500 п.о. или от приблизительно 350 до 450. Согласно другим вариантам реализации длина лигированной с адаптерами маркерной молекулы составляет приблизительно 200 п.о. Например, при секвенировании сцДНК плода, которая присутствует в материнском образце, длину маркерной молекулы можно выбрать так, чтобы она являлась аналогичной таковой молекул сцДНК плода. Таким образом, согласно одному варианту реализации длина маркерной молекулы, применяемой в анализе, который включает широкомасштабное параллельное секвенирование сцДНК в материнском образце для определения присутствия или отсутствия анеуплоидии хромосомы плода, может составлять приблизительно 150 п.о., приблизительно 160 п.о., 170 п.о., приблизительно 180 п.о., приблизительно 190 п.о. или приблизительно 200 п.о.; предпочтительно, длина маркерной молекулы составляет приблизительно 170 п.о. В других подходах секвенирования, например секвенировании SOLiD, полони-секвенировании и секвенировании 454, для клональной амплификации молекул ДНК с целью секвенирования применяют эмульсионную ПЦР, и каждая технология диктует минимальную и максимальную длину молекул, которые необходимо амплифицировать. Длина маркерных молекул, секвенирование которых проводят в виде клонально амплифицированных нуклеиновых кислот, может составлять вплоть до приблизительно 600 п.о. Согласно некоторым вариантам реализации длина маркерных молекул, секвенирование которых проводят, может составлять более 600 п.о.

В случае технологий одномолекулярного секвенирования, в которых не применяют клональную амплификацию молекул и которые способны к секвенированию нуклеиновых кислот в пределах очень широкого диапазона длин матриц, в большинстве ситуаций не требуется, чтобы молекулы, секвенирование которых проводят, характеризовались любой конкретной длиной. Однако выход последовательно-

стей на единицу массы зависит от количества 3'-концевых гидроксильных групп и, таким образом, наличие относительно коротких матриц для секвенирования является более эффективным, чем наличие длинных матриц. Если начинать с нуклеиновых кислот, более длинных, чем 1000 нуклеотидов, как правило, рекомендуют разрезать нуклеиновые кислоты до средней длины 100-200 нуклеотидов для того, чтобы с помощью той же массы нуклеиновых кислот можно было получить больше информации о последовательности. Таким образом, длина маркерной молекулы может варьировать от десятков оснований до тысяч оснований. Длина маркерных молекул, применяемых для одномолекулярного секвенирования, может составлять вплоть до приблизительно 25 п.о., вплоть до приблизительно 50 п.о., вплоть до приблизительно 75 п.о., вплоть до приблизительно 100 п.о., вплоть до приблизительно 200 п.о., вплоть до приблизительно 300 п.о., вплоть до приблизительно 400 п.о., вплоть до приблизительно 500 п.о., вплоть до приблизительно 600 п.о., вплоть до приблизительно 700 п.о., вплоть до приблизительно 800 п.о., вплоть до приблизительно 900 п.о., вплоть до приблизительно 1000 п.о. или более.

Длина, выбранная для маркерной молекулы, также определяется длиной геномной нуклеиновой кислоты, секвенирование которой проводят. Например, сцДНК циркулирует в сосудистом русле человека в виде геномных фрагментов клеточной геномной ДНК. Молекулы сцДНК плода, обнаруженные в плазме беременных женщин, как правило, более короткие, чем молекулы материнской сцДНК (Chan et al., Clin Chem 50:8892 [2004]). Фракционирование циркулирующей ДНК плода по размеру подтвердило, что средняя длина фрагментов циркулирующей ДНК плода составляет <300 п.о., тогда как длина материнской ДНК, согласно оценкам, составляет приблизительно от 0,5 до 1 т.о. (Li et al., Clin Chem, 50: 1002-1011 [2004]). Данные результаты согласуются с таковыми Fan et al., которые определили с применением СНП, что длина сцДНК плода редко превышает 340 п.о. (Fan et al., Clin Chem 56:1279-1286 [2010]). ДНК, выделенная из мочи стандартным способом на основе диоксида кремния, состоит из двух фракций, высокомолекулярной ДНК, которая происходит из выделенных клеток, и низкомолекулярной (150 - 250 пар оснований) фракции трансренальной ДНК (Tr-DNA) (Botezatu et al., Clin Chem. 46: 1078-1084, 2000; и Su et al., J Mol. Diagn. 6: 101-107, 2004). Применение недавно разработанной методики для выделения бесклеточных нуклеиновых кислот из жидкостей для выделения трансренальных нуклеиновых кислот позволило обнаружить присутствие в моче фрагментов ДНК и РНК в значительной степени более коротких, чем 150 пар оснований (публикация заявки на патент США № 20080139801). Согласно вариантам реализации, в которых сцДНК представляет собой геномную нуклеиновую кислоту, которую секвенируют, выбранная длина маркерных молекул может составлять вплоть до приблизительно длины сцДНК. Например, длина маркерных молекул, применяемых в образцах материнской сцДНК, секвенирование которых проводят в виде единичных молекул нуклеиновой кислоты или в виде клонально амплифицированных нуклеиновых кислот, может составлять от приблизительно 100 до 600 п.о. Согласно другим вариантам реализации геномные нуклеиновые кислоты образца представляют собой фрагменты больших молекул. Например, геномная нуклеиновая кислота образца, которую секвенируют, представляет собой фрагментированную клеточную ДНК. Согласно вариантам реализации, в которых секвенируют фрагментированную клеточную ДНК, длина маркерных молекул может составлять вплоть до длины фрагментов ДНК. Согласно некоторым вариантам реализации длина маркерных молекул составляет по меньшей мере минимальную длину, необходимую для уникального картирования рида последовательности на соответствующий референсный геном. Согласно другим вариантам реализации длина маркерной молекулы составляет минимальную длину, необходимую для исключения маркерной молекулы из картирования на референсный геном образца.

Помимо этого, маркерные молекулы можно применять для подтверждения образцов, которые не анализируют посредством секвенирования нуклеиновой кислоты и которые можно подтвердить посредством общепринятых биологических методик, отличных от секвенирования, например ПЦР в режиме реального времени.

Контрольные образцы (например, внутренние положительные контроли для секвенирования и/или анализа).

Согласно различным вариантам реализации маркерные последовательности, вводимые в образцы, например, как описано выше, могут выступать в качестве положительных контролей для подтверждения точности и эффективности секвенирования и последующего процессинга и анализа.

Соответственно, предложены композиции и способ обеспечения внутреннего положительного контроля (ВПК) для секвенирования ДНК в образце. Согласно определенным вариантам реализации предложены положительные контроли для секвенирования сцДНК в образце, содержащем смесь геномов. ВПК можно применять для установления связи сдвигов базовой линии в информации о последовательности, полученной из различных множеств образцов, например образцов, которые секвенируют в различные времена в различных сериях секвенирования. Таким образом, например, ВПК может устанавливать связь информации о последовательности, полученной для материнского исследуемого образца, с информацией о последовательности, полученной из множества квалификационных образцов, которые секвенировали в отличное время.

Аналогично, в случае анализа сегментов ВПК может устанавливать связь между информацией о последовательности, полученной от субъекта для конкретного сегмента или сегментов, с информацией о

последовательности, полученной из множества квалификационных образцов (аналогичных последовательностей), которые секвенировали в отличное время. Согласно определенным вариантам реализации ВПК может устанавливать связь информации о последовательности, полученной от субъекта для конкретного связанного с раком локуса, с информацией о последовательности, полученной из множества квалификационных образцов (например, из известной амплификации/делеции и т.п.).

Помимо этого, ВПК можно применять в качестве маркеров для отслеживания образца или образцов в течение процесса секвенирования. ВПК могут также обеспечить качественное значение положительной дозы последовательности, например, NCV, для одной или более анеуплоидий хромосом, представляющих интерес, например, трисомии 21, трисомии 13, трисомии 18, для обеспечения надлежащей интерпретации и для гарантирования достоверности и точности данных. Согласно определенным вариантам реализации можно создать ВПК, которые содержат нуклеиновые кислоты из мужского и женского геномов, с целью обеспечения доз для хромосом X и Y в материнском образце для определения того, является ли плод плодом мужского пола.

Тип и количества внутренних контролей зависят от типа или природы необходимого анализа. Например, для анализа, для которого требуется секвенирование ДНК из образца, содержащего смесь геномов, с целью определения присутствия анеуплоидии хромосомы, внутренний контроль может содержать ДНК, полученную из образца, который установленно содержит ту же хромосомную анеуплоидию, исследование которой проводят. Согласно некоторым вариантам реализации ВПК содержит ДНК из образца, который установленно содержит анеуплоидию хромосомы, представляющей интерес. Например, ВПК для анализа с целью определения присутствия или отсутствия трисомии у плода, например, трисомии 21, в материнском образце содержит ДНК, полученную от индивидуума с трисомией 21. Согласно некоторым вариантам реализации ВПК содержит смесь ДНК, полученной от двух или более индивидуумов с различными анеуплоидиями. Например, для анализа с целью определения присутствия или отсутствия трисомии 13, трисомии 18, трисомии 21 и моносомии X ВПК содержит комбинацию образцов ДНК, полученной от беременных женщин, каждая из которых вынашивает плод с одной из трисомий, исследование которой проводят. В дополнение к полным анеуплоидиям хромосом можно создать ВПК для обеспечения положительных контролей для анализов с целью определения присутствия или отсутствия частичных анеуплоидий.

ВПК, который выступает в качестве контроля для обнаружения единичной анеуплоидии, можно создать с применением смеси клеточной геномной ДНК, полученной из двух субъектов, один из которых является источником анеуплоидного генома. Например, ВПК, который создают в качестве контроля для анализа с целью определения трисомии у плода, например трисомии 21, можно создать посредством объединения геномной ДНК из мужского или женского субъекта, несущего трисомическую хромосому, с геномной ДНК субъекта женского пола, который установленно не несет трисомическую хромосому. Геномную ДНК можно экстрагировать из клеток обоих субъектов и разрезать для обеспечения фрагментов длиной от приблизительно 100-400 п.о., от приблизительно 150-350 п.о. или от приблизительно 200-300 п.о. для имитации циркулирующих фрагментов сцДНК в материнских образцах. Долю фрагментированной ДНК из субъекта, несущего анеуплоидию, например, трисомию 21, выбирают для имитации доли циркулирующей сцДНК плода, обнаруженной в материнских образцах, с получением ВПК, содержащего смесь фрагментированной ДНК, которая содержит приблизительно 5%, приблизительно 10%, приблизительно 15%, приблизительно 20%, приблизительно 25%, приблизительно 30% ДНК от субъекта, несущего анеуплоидию. ВПК может содержать ДНК от различных субъектов, каждый из которых несет отличную анеуплоидию. Например, ВПК может содержать приблизительно 80% непораженной женской ДНК, и оставшиеся 20% могут представлять собой ДНК от трех различных субъектов, каждый из которых несет трисомическую хромосому 21, трисомическую хромосому 13 и трисомическую хромосому 18. Для секвенирования готовят смесь фрагментированной ДНК. Процессинг смеси фрагментированной ДНК может включать получение библиотеки секвенирования, которую можно секвенировать с применением любого широкомасштабного параллельного способа в синглплексном или мультиплексном режиме. Базовые растворы геномного ВПК можно хранить и применять во множестве диагностических анализов.

В качестве альтернативы, можно создать ВПК с применением сцДНК, полученной от матери, которая установленно вынашивает плод с известной анеуплоидией хромосомы. Например, сцДНК можно получить от беременной женщины, которая вынашивает плод с трисомией 21. сцДНК экстрагируют из материнского образца и клонируют в бактериальном векторе и выращивают в бактериях с получением постоянного источника ВПК. ДНК можно экстрагировать из бактериального вектора с применением рестрикционных ферментов. В качестве альтернативы, клонированную сцДНК можно амплифицировать посредством, например, ПЦР. ДНК ВПК можно процессировать для секвенирования в одной и той же серии, что и сцДНК из исследуемых образцов, которые анализируют в отношении присутствия или отсутствия анеуплоидий хромосом.

Несмотря на то что создание ВПК описано выше применительно к трисомии, следует принимать во внимание, что ВПК можно создать для отражения других частичных анеуплоидий, включая, например, различные амплификации и/или делеции сегментов. Таким образом, например, когда известно, что различные типы рака связаны с конкретными амплификациями (например, рак молочной железы, связанный

с 20Q13), можно создать ВПК, которые содержат данные известные амплификации.

Способы секвенирования.

Как указано выше, полученные образцы (например, библиотеки секвенирования) секвенируют как часть процедуры идентификации вариации или вариаций числа копий. Можно применять любую из множества технологий секвенирования.

Некоторые технологии секвенирования доступны коммерчески, такие как платформа секвенирования посредством гибридизации от компании Affymetrix Inc. (Саннивейл, Калифорния) и платформы для секвенирования посредством синтеза от компаний 454 Life Sciences (Брэдфорд, Коннектикут), Illumina/Solexa (Хейвард, Калифорния) и Helicos Biosciences (Кембридж, Массачусетс), а также платформа для секвенирования посредством лигирования от компании Applied Biosystems (Фостер Сити, Калифорния), описанная ниже. В дополнение к одномолекулярному секвенированию, которое осуществляют с применением секвенирования посредством синтеза от компании Helicos Biosciences, другие технологии одномолекулярного секвенирования включают, без ограничения, технологию SMRT™ от компании Pacific Biosciences, технологию ION TORRENT™ и нанопоровое секвенирование, разработанное, например, компанией Oxford Nanopore Technologies.

Несмотря на то что автоматизированный способ Сэнджера считают технологией "первого поколения", секвенирование по Сэнджеру, включая автоматизированное секвенирование по Сэнджеру, можно также применять в способах, описанных в настоящем документе. Дополнительные подходящие способы секвенирования включают, без ограничения, технологии визуализации нуклеиновой кислоты, например, атомно-силовую микроскопию (АСМ) или трансмиссионную электронную микроскопию (ТЭМ). Иллюстративные технологии секвенирования более подробно описаны ниже.

Согласно одному иллюстративному, но неограничивающему варианту реализации способы, описанные в настоящем документе, включают получение информации о последовательности нуклеиновых кислот в исследуемом образце, например сцДНК в материнском образце, сцДНК или клеточной ДНК у субъекта, скрининг которого проводят в отношении рака, и т.п., с применением секвенирования посредством синтеза Illumina и химии секвенирования на основе обратимого терминатора (например, как описано в публикации Bentley et al., Nature 6:53-59 [2009]). Матрица ДНК может представлять собой геномную ДНК, например, клеточную ДНК или сцДНК. Согласно некоторым вариантам реализации в качестве матрицы применяют геномную ДНК из выделенных клеток, и ее фрагментируют на длины в несколько сотен пар оснований. Согласно другим вариантам реализации сцДНК применяют в качестве матрицы, и фрагментация не требуется, поскольку сцДНК существует в виде коротких фрагментов. Например, сцДНК плода циркулирует в сосудистом русле в виде фрагментов длиной приблизительно 170 пар оснований (п.о.) (Fan et al., Clin Chem 56:1279-1286 [2010]), и перед секвенированием фрагментация ДНК не требуется. Технология секвенирования Illumina основана на присоединении фрагментированной геномной ДНК к плоской, оптически прозрачной поверхности, с которой связывают олигонуклеотидные якоря. В матрице ДНК восстанавливают концы для получения 5'-фосфорилированных тупых концов, и полимеразную активность фрагмента Кленова применяют для добавления одного основания А к 3'-концу тупых фосфорилированных фрагментов ДНК. Данное добавление подготавливает фрагменты ДНК к лигированию с олигонуклеотидными адаптерами, которые содержат выступ одного основания Т на 3'-конце для увеличения эффективности лигирования. Адаптерные олигонуклеотиды комплементарны якорным олигонуклеотидам проточной ячейки (не путать с якорными/заякоренными ридами в анализе экспансии повторов). В условиях серийных разведений модифицированную адаптерами одноцепочечную матрицу ДНК добавляют в проточную ячейку и иммобилизуют посредством гибридизации с якорными олигонуклеотидами. Присоединенные фрагменты ДНК удлиняют и амплифицируют посредством мостиковой амплификации для получения секвенирования сверхвысокой плотности проточной ячейки с сотнями миллионов кластеров, каждый из которых содержит приблизительно 1000 копий одной и той же матрицы. Согласно одному варианту реализации случайным образом фрагментированную геномную ДНК амплифицируют с применением ПНР до того, как ее подвергнут кластерной амплификации. В качестве альтернативы, применяют получение геномной библиотеки без амплификации (например, без применения ПЦР), и случайным образом фрагментированную геномную ДНК обогащают с применением кластерной амплификации самой по себе (Kozarewa et al., Nature Methods 6:291-295 [2009]). Матрицы секвенируют с применением устойчивой четырехцветной технологии секвенирования ДНК посредством синтеза, в которой применяют обратимые терминаторы с удаляемыми флуоресцентными красителями. Высокочувствительное флуоресцентное обнаружение обеспечивают с применением возбуждения лазером и оптики полного внутреннего отражения. Короткие риды последовательности длиной приблизительно от десяти до нескольких сотен пар оснований выравнивают с референсным геномом, и уникальные картирования коротких ридов последовательности на референсный геном идентифицируют с применением специально разработанного ассортимента программного обеспечения для анализа данных. После завершения первого рида матрицы можно регенерировать in situ, что делает возможным получение второго рида с противоположных концов фрагментов. Таким образом, можно применять секвенирование одиночных концов или спаренных концов фрагментов ДНК.

В различных вариантах реализации настоящего изобретения можно применять секвенирование посредством синтеза, позволяющее проводить секвенирование спаренных концов. Согласно некоторым вариантам реализации платформа Illumina для секвенирования посредством синтеза включает кластеризацию фрагментов. Кластеризация представляет собой процесс, в котором каждую молекулу фрагмента изотермически амплифицируют. Согласно некоторым вариантам реализации в качестве примера, описанного в настоящем документе, фрагмент содержит два различных адаптера, присоединенных к двум концам фрагмента, причем адаптеры позволяют фрагменту гибридизоваться с двумя различными олигонуклеотидами на поверхности дорожки проточной ячейки. Фрагмент также содержит на двух своих концах две индексные последовательности или присоединен к ним, причем индексные последовательности обеспечивают метки для идентификации различных образцов в мультиплексном секвенировании. В некоторых платформах секвенирования фрагмент, секвенирование которого проводят, также называют вставкой.

Согласно некоторому варианту реализации проточная ячейка для кластеризации в платформе Illumina представляет собой стеклянную пластинку с дорожками. Каждая дорожка представляет собой стеклянный канал, на который нанесено покрытие из двух типов олигонуклеотидов. Гибридизация обеспечивается благодаря первому из двух типов олигонуклеотидов на поверхности. Данный олигонуклеотид комплементарен первому адаптеру на одном конце фрагмента. Полимераза создает комплементарную цепь гибридизированного фрагмента. Двухцепочечную молекулу денатурируют, и цепь исходной матрицы смывают. Оставшуюся цепь, параллельно со многими другими оставшимися цепями, клонально амплифицируют посредством мостиковой амплификации.

При мостиковой амплификации цепь сворачивается, и вторая адаптерная область на втором конце цепи гибридизуется со вторым типом олигонуклеотидов на поверхности проточной ячейки. Полимераза создает комплементарную цепь, образуя двухцепочечную мостиковую молекулу. Данную двухцепочечную молекулу денатурируют, что приводит к получению двух одноцепочечных молекул, присоединенных к проточной ячейке посредством двух различных олигонуклеотидов. Затем процесс повторяют снова и снова, и он происходит одновременно для миллионов кластеров, что приводит к клональной амплификации всех фрагментов. После мостиковой амплификации обратные цепи отщепляют и смывают, оставляя исключительно прямые цепи. 3'-концы блокируют для предотвращения нежелательного праймирования.

После кластеризации секвенирование начинают с удлинения первого праймера секвенирования для получения первого рида. С каждым циклом флуоресцентно меченные нуклеотиды конкурируют за добавление к растущей цепи. На основании последовательности матрицы встраивается исключительно один нуклеотид. После добавления каждого нуклеотида кластер возбуждают источником света, и испускается характерный флуоресцентный сигнал. Количество циклов определяет длину рида. Длина волны испускания и интенсивность сигнала определяют основной отклик. Для данного кластера все идентичные цепи прочитываются одновременно. Сотни миллионов кластеров секвенируют широкомасштабным параллельным способом. После завершения первого рида прочитанный продукт смывают.

На следующем этапе протоколов, включающих два индексных праймера, праймер индекс 1 вводят и гибридизуют с областью индекс 1 на матрице. Индексные области обеспечивают идентификацию фрагментов, которые являются подходящими для демультиплексирования образцов в процессе мультиплексного секвенирования. Рид индекс 1 получают аналогично первому риду. После завершения рида индекс 1 прочитанный продукт смывают, и с 3'-конца цепи снимают защиту. Затем цепь матрицы сворачивается и связывается со вторым олигонуклеотидом на проточной ячейке. Последовательность индекс 2 прочитывают тем же способом, что и индекс 1. Затем прочитанный продукт индекс 2 смывают после завершения этапа.

После прочитывания двух индексов начинают рид 2 посредством применения полимераз для удлинения олигонуклеотидами второй проточной ячейки с образованием двухцепочечного мостика. Данную двухцепочечную ДНК денатурируют, и 3'-конец блокируют. Исходную прямую цепь отщепляют и смывают, оставляя обратную цепь. Рид 2 начинают с введения праймера секвенирования рида 2. Как и в случае рида 1, этапы секвенирования повторяют до достижения желаемой длины. Продукт рида 2 смывают. Весь данный процесс позволяет получить миллионы ридов, которые представляют все фрагменты. Последовательности из библиотек объединенных образцов разделяют на основании уникальных индексов, введенных в процессе получения образца. Для каждого образца риды аналогичных протяженностей основных откликов локально кластеризуют. Прямые и обратные риды располагают парами, создавая непрерывные последовательности. Данные непрерывные последовательности выравнивают с референсным геномом для идентификации варианта.

Пример секвенирования посредством синтеза, описанный выше, включает риды спаренных концов, которые применяют во множестве вариантов реализации раскрытых способов. Секвенирование спаренных концов включает 2 рида с двух концов фрагмента. Когда пару ридов картируют на референсную последовательность, можно определить расстояние между парами оснований между двумя ридами, и затем данное расстояние можно применять для определения длины фрагментов, из которых были получены риды. В некоторых случаях у фрагмента, расположенного в двух блоках, одно из ридов парных концов

будет выровнено с одним блоком, а другое - с прилежащим блоком. Это происходит реже по мере того как блоки становятся более длинными или риды становятся более короткими. Для определения принадлежности данных фрагментов к блокам можно применять различные способы. Например, фрагменты можно опустить при определении частоты размера фрагмента блока; можно вычислить для обоих из прилежащих блоков; можно отнести к блоку, который охватывает большее количество пар оснований, из двух блоков; или фрагменты можно отнести к обоим блокам с весом в отношении части пар оснований в каждом блоке.

В ридах спаренных концов можно применять вставки различных длин (т.е. различный размер фрагмента, секвенирование которого проводят). В качестве значения по умолчанию в настоящем изобретении применяют риды спаренных концов для обозначения ридов, полученных от различных длин вставок. В некоторых случаях, чтобы различить риды спаренных концов короткой вставки от ридов спаренных концов длинной вставки, последнюю также называют ридами сопряженной пары. Согласно некоторым вариантам реализации, включающим риды сопряженной пары, сначала к двум концам относительно длинной вставки (например, несколько т.о.) присоединяют два соединительных адаптера на основе биотина. После этого соединительные адаптеры на основе биотина соединяют с двумя концами вставки с образованием циркуляризованной молекулы. Затем можно получить субфрагмент, содержащий соединительные адаптеры на основе биотина, посредством последующей фрагментации циркуляризованной молекулы. После этого можно секвенировать субфрагмент, содержащий два конца исходного фрагмента в противоположном порядке последовательности, посредством той же процедуры, что и для секвенирования спаренных концов короткой вставки, описанного выше. Дополнительные подробности секвенирования сопряженной пары с применением платформы Illumina представлены в онлайн-публикации по следующему электронному адресу, которая полностью включена в настоящий документ посредством ссылки: res|.|illumina|.|com/documents/products/technotes/technote_nextera_matepair_data_processing. Дополнительную информацию относительно секвенирования спаренных концов можно найти в патенте США № 7601499 и публикации патента США № 2012/0053063, которые включены в настоящий документ посредством ссылки применительно к материалам о способах и аппаратах для секвенирования спаренных концов.

После секвенирования фрагментов ДНК риды последовательности заранее определенной длины, например, 100 п.о., картируют или выравнивают с известным референсным геномом. Картированные или выровненные риды и их соответствующие расположения на референсной последовательности также называют метками. Согласно одному варианту реализации последовательность референсного генома представляет собой последовательность NCBI36/hg18, которая доступна в сети Интернет по адресу: genome dot ucsc dot edu/cgi-bin/hgGateway?org=Human&db=hg18&hgsid=166260105. В качестве альтернативы, последовательность референсного генома представляет собой GRCh37/hg19, которая доступна в сети Интернет по адресу: genome dot ucsc dot edu/cgi-bin/hgGateway. Другие источники общедоступной информации о последовательности включают GenBank, dbEST, dbSTS, EMBL (European Molecular Biology Laboratory, Европейская лаборатория по молекулярной биологии) и DDBJ (DNA Databank of Japan, База данных ДНК Японии). Для выравнивания последовательностей доступно множество компьютерных алгоритмов, включая, без ограничения, BLAST (Altschul et al., 1990), BLITZ (MPsrch) (Sturrock & Collins, 1993), FASTA (Person & Lipman, 1988), BOWTIE (Langmead et al., Genome Biology 10:R25.1-R25.10 [2009]) или ELAND (Illumina, Inc., Сан-Диего, Калифорния, США). Согласно одному варианту реализации один конец клонально удлиненных копий молекул сцДНК плазмы секвенируют и процессируют посредством анализа биоинформатического выравнивания для геномного анализатора Genome Analyzer Illumina, в котором используется программное обеспечение Efficient Large-Scale Alignment of Nucleotide Databases (Эффективное крупномасштабное выравнивание нуклеотидных данных, ELAND).

Согласно одному иллюстративному, но неограничивающему варианту реализации способы, описанные в настоящем документе, включают получение информации о последовательности для нуклеиновых кислот в исследуемом образце, например сцДНК в материнском образце, сцДНК или клеточной ДНК у субъекта, скрининг которого проводят в отношении рака, и т.п., с применением метода одномолекулярного секвенирования на основе технологии истинного одномолекулярного секвенирования (True Single Molecule Sequencing, tSMS) компании Helicos (например, как описано в публикации Harris T.D. et al., Science 320:106-109 [2008]). В методиках tSMS образец ДНК расщепляют на цепи длиной приблизительно от 100 до 200 нуклеотидов, и к 3'-концу каждой цепи ДНК добавляют последовательность polyA. Каждую цепь метят посредством добавления флуоресцентно меченного нуклеотида аденозина. Затем цепи ДНК гибридизуют с проточной ячейкой, которая содержит миллионы захватывающих сайтов олигонуклеотид-Т, которые иммобилизуют на поверхности проточной ячейки. Согласно определенным вариантам реализации матрицы могут характеризоваться плотностью приблизительно 100 миллионов матриц/см^2. Затем проточную ячейку помещают в прибор, например, секвенатор HeliScope™, и лазер освещает поверхность проточной ячейки, позволяя определить положение каждой матрицы. Камера на ПЗС (приборах с зарядовой связью) может картировать положение матриц на поверхности проточной ячейки. Затем флуоресцентную метку матрицы отщепляют и смывают. Реакцию секвенирования начинают посредством введения ДНК-полимеразы и флуоресцентно меченного нуклеотида. Нуклеиновая кислота

олигонуклеотид-Т выступает в качестве праймера. Полимераза встраивает меченные нуклеотиды к праймеру управляемым матрицей способом. Полимеразу и невстроенные нуклеотиды удаляют. Матрицы, которые направляли встраивание флуоресцентно меченного нуклеотида, распознают посредством визуализации поверхности проточной ячейки. После визуализации на этапе отщепления удаляют флуоресцентную метку, и процесс повторяют с другими флуоресцентно меченными нуклеотидами до достижения желаемой длины рида. Информацию о последовательности собирают для каждого этапа добавления нуклеотида. При секвенировании целого генома посредством технологий одномолекулярного секвенирования исключают или, как правило, избегают амплификации на основе ПЦР при получении библиотек секвенирования, и способы позволяют проводить прямое измерение образца вместо измерения копий данного образца.

Согласно другому иллюстративному, но неограничивающему варианту реализации способы, описанные в настоящем документе, включают получение информации о последовательности для нуклеиновых кислот в исследуемом образце, например, сцДНК в материнском исследуемом образце, сцДНК или клеточной ДНК у субъекта, скрининг которого проводят в отношении рака, и т.п., с применением секвенирования 454 (Roche) (например, как описано в публикации Margulies, M. et al. Nature 437:376-380 [2005]). Секвенирование 454, как правило, включает два этапа. На первом этапе ДНК разрезают на фрагменты длиной приблизительно 300-800 пар оснований, причем фрагменты содержат тупые концы. Затем олигонуклеотидные адаптеры лигируют с концами фрагментов. Адаптеры выступают в качестве праймеров для амплификации и секвенирования фрагментов. Фрагменты можно присоединить к бусинам, захватывающим ДНК, например, бусинам, покрытым стрептавидином, с применением, например, Адаптера B, который содержит 5'-биотиновую метку. Фрагменты, присоединенные к бусинам, ПЦР-амплифицируют в каплях эмульсии "масло в воде". Результат представляет собой множество копий клонально амплифицированных фрагментов ДНК на каждой бусине. На втором этапе бусины захватывают в лунки (например, лунки пиколитрового объема). Пиросеквенирование осуществляют на каждом фрагменте ДНК параллельно. Добавление одного или более нуклеотидов позволяет получить световой сигнал, который записывает камера на ПЗС в приборе для секвенирования. Сила сигнала пропорциональна количеству встроенных нуклеотидов. При пиросеквенировании применяют пирофосфат (PPi), который высвобождается после добавления нуклеотида. PPi преобразуется в АТФ под действием АТФ-сульфурилазы в случае присутствия аденозин-5'-фосфосульфата. Люцифераза использует АТФ для преобразования люциферина в оксилюциферин, и данная реакция позволяет получить свет, который измеряют и анализируют.

Согласно другому иллюстративному, но неограничивающему варианту реализации способы, описанные в настоящем документе, включают получение информации о последовательности для нуклеиновых кислот в исследуемом образце, например, сцДНК в материнском исследуемом образце, сцДНК или клеточной ДНК у субъекта, скрининг которого проводят в отношении рака, и т.п., с применением технологии SOLiD™ (Applied Biosystems). В секвенировании посредством лигирования SOLiD™ геномную ДНК разрезают на фрагменты, и к 5'- и 3'-концам фрагментов присоединяют адаптеры для получения библиотеки фрагментов. В качестве альтернативы, внутренние адаптеры можно ввести посредством лигирования адаптеров с 5'- и 3'-концами фрагментов, циркуляризации фрагментов, расщепления циркуляризованного фрагмента для получения внутреннего адаптера и присоединения адаптеров к 5'-и 3'-концам полученных в результате фрагментов для получения библиотеки сопряженной пары. Затем в микрореакторах, содержащих бусины, праймеры, матрицу и компоненты ПНР, получают популяции клональных бусин. После проведения ПНР матрицы денатурируют, и бусины обогащают отдельными бусинами с удлиненными матрицами. Матрицы на выбранных бусинах подвергают 3'-модификации, которая обеспечивает образование связей со стеклянной пластинкой. Последовательность можно определить посредством последующей гибридизации и лигирования частично случайных олигонуклеотидов с центральным определенным основанием (или парой оснований), которые идентифицируют с помощью специфичного флуорофора. После регистрации цвета лигированный олигонуклеотид отщепляют и удаляют, а затем повторяют процесс.

Согласно другому иллюстративному, но неограничивающему варианту реализации способы, описанные в настоящем документе, включают получение информации о последовательности для нуклеиновых кислот в исследуемом образце, например, сцДНК в материнском исследуемом образце, сцДНК или клеточной ДНК у субъекта, скрининг которого проводят в отношении рака, и т.п., с применением технологии одномолекулярного секвенирования в режиме реального времени (single molecule, real-time, SMRT™) от компании Pacific Biosciences. В секвенировании SMRT непрерывное встраивание меченных красителем нуклеотидов визуализируют в течение синтеза ДНК. Одну молекулу ДНК-полимеразы присоединяют ко дну поверхности отдельных детекторов длины волны ноль-режима (ZMW, zero-mode wavelength), которые получают информацию о последовательности, когда фосфосвязанные нуклеотиды встраиваются в растущую цепь праймера. Детектор ZMW содержит разграниченные структуры, которые позволяют проводить наблюдение за встраиванием отдельного нуклеотида ДНК-полимеразой по сравнению с фоном флуоресцентных нуклеотидов, которые быстро диффундируют в ZMW и из него (напри-

мер, за микросекунды). Как правило, для встраивания нуклеотида в растущую цепь требуется несколько миллисекунд. В течение данного времени флуоресцентная метка возбуждается и образует флуоресцентный сигнал, и флуоресцентную метку отщепляют. Измерение соответствующей флуоресценции красителя указывает на то, какое основание было встроено. Процесс повторяют с получением последовательности.

Согласно другому иллюстративному, но неограничивающему варианту реализации способы, описанные в настоящем документе, включают получение информации о последовательности для нуклеиновых кислот в исследуемом образце, например, сцДНК в материнском исследуемом образце, сцДНК или клеточной ДНК у субъекта, скрининг которого проводят в отношении рака, и т.п., с применением нанопорового секвенирования (например, как описано в публикации Soni G.V. and Meller A. Clin Chem 53: 1996-2001 [2007]). Методики анализа нанопорового секвенирования ДНК разработаны множеством компаний, включая, например, Oxford Nanopore Technologies (Оксфорд, Великобритания), Sequenom, NABsys и т.п. Нанопоровое секвенирование представляет собой технологию одномолекулярного секвенирования, посредством которой одну молекулу ДНК секвенируют напрямую, по мере того как она проходит через нанопору. Нанопора представляет собой небольшое отверстие, как правило, порядка 1 нм в диаметре. Погружение нанопоры в проводящую жидкость и наложение на нее потенциала (напряжения) приводит к появлению небольшого электрического тока благодаря проводимости ионов через нанопору. Количество тока, которое протекает, чувствительно к размеру и форме нанопоры. По мере того как молекула ДНК проходит через нанопору, каждый нуклеотид на молекуле ДНК в различной степени преграждает нанопору, в различной степени изменяя магнитуду тока через нанопору. Таким образом, данное изменение тока по мере того как молекула ДНК проходит через нанопору обеспечивает прочитывание последовательности ДНК.

Согласно другому иллюстративному, но неограничивающему варианту реализации способы, описанные в настоящем документе, включают получение информации о последовательности для нуклеиновых кислот в исследуемом образце, например сцДНК в материнском исследуемом образце, сцДНК или клеточной ДНК у субъекта, скрининг которого проводят в отношении рака, и т.п., с применением матрицы химически-чувствительного полевого транзистора (chemical-sensitive field effect transistor, chemFET) (например, как описано в публикации заявки на патент США № 2009/0026082). В одном примере данной методики молекулы ДНК можно поместить в реакционные камеры, и молекулы матрицы можно гибридизовать с праймером секвенирования, связанным с полимеразой. Встраивание одного или более трифосфатов в новую цепь нуклеиновой кислоты на 3'-конце праймера секвенирования можно выявить с помощью chemFET как изменение тока. Матрица может содержать множество сенсоров chemFET. В другом примере отдельные нуклеиновые кислоты можно присоединить к бусинам, и нуклеиновые кислоты можно амплифицировать на бусине, и отдельные бусины можно перенести в отдельные реакционные камеры на матрице chemFET, причем каждая камера содержит сенсор chemFET, и можно секвенировать нуклеиновые кислоты.

Согласно другому варианту реализации настоящий способ включает получение информации о последовательности для нуклеиновых кислот в исследуемом образце, например, сцДНК в материнском исследуемом образце, с применением способа трансмиссионной электронной микроскопии (ТЭМ). Способ, называемый быстрым нанопереносом положения отдельной молекулы (Individual Molecule Placement Rapid Nano Transfer, IMPRNT), включает визуализацию с применением трансмиссионного электронного микроскопа с одноатомным разрешением высокомолекулярной (150 т.о. или более) ДНК, селективно меченной маркерами тяжелых металлов, и упорядочивание данных молекул на ультратонких пленках в сверхплотных (расстояние от цепи до цепи 3 нм) параллельных матрицах с последовательным расположением от основания к основанию. Для визуализации молекул на пленках с целью определения положения маркеров тяжелых атомов и извлечения информации о последовательности оснований из ДНК применяют электронный микроскоп. Способ дополнительно описан в публикации патента PCT WO 2009/046445. Способ позволяет секвенировать полные геномы человека менее чем за десять минут.

Согласно другому варианту реализации технология секвенирования ДНК представляет собой одномолекулярное секвенирование Ion Torrent, которое объединяет полупроводниковую технологию с простой химией секвенирования, чтобы напрямую транслировать закодированную химическим способом информацию (A, C, G, T) в цифровую информацию (0, 1) на полупроводниковом чипе. В природе, когда нуклеотид встраивается полимеразой в цепь ДНК, в качестве побочного продукта высвобождается ион водорода. В технологии Ion Torrent для осуществления данного биохимического процесса широкомасштабным параллельным способом применяют матрицу микрообработанных лунок высокой плотности. Каждая лунка содержит отличную молекулу ДНК. Ниже лунок расположен ион-чувствительный слой, а ниже него - ионный сенсор. Когда к матрице ДНК добавляют нуклеотид, например, C, а затем нуклеотид встраивается в цепь ДНК, высвобождается ион водорода. Заряд данного иона изменит pH раствора, что может обнаружить ионный сенсор Ion Torrent. Секвенатор по существу, наименьший в мире твердофазный pH-метр называет основание, переходя непосредственно от химической информации к цифровой информации. Затем секвенатор Ion personal Genome Machine (PGM™) последовательно заливает чип нуклеотидами друг за другом. Если следующий нуклеотид, который заливают на чип, не соответствует,

не будет зафиксировано какого-либо изменения напряжения, и какое-либо основание не будет определено. Если на цепи ДНК присутствуют два идентичных основания, напряжение увеличится вдвое, и чип зафиксирует определение двух идентичных оснований. Прямое обнаружение позволяет регистрировать встраивание нуклеотида в течение секунд.

Согласно другому варианту реализации настоящий способ включает получение информации о последовательности для нуклеиновых кислот в исследуемом образце, например сцДНК в материнском исследуемом образце, с применением секвенирования посредством гибридизации. Секвенирование посредством гибридизации включает осуществление контакта множества полинуклеотидных последовательностей с множеством полинуклеотидных зондов, причем каждый из множества полинуклеотидных зондов может быть необязательно присоединен к субстрату. Субстрат может представлять собой плоскую поверхность, содержащую матрицу известных нуклеотидных последовательностей. Характер гибридизации с матрицей можно использовать для определения полинуклеотидных последовательностей, присутствующих в образце. Согласно другим вариантам реализации каждый зонд присоединен к бусине, например, магнитной бусине или т.п. Гибридизацию к бусинам можно определить и использовать для идентификации множества полинуклеотидных последовательностей в образце.

Согласно некоторым вариантам реализации способов, описанных в настоящем документе, картированные метки последовательности содержат риды последовательности длиной приблизительно 20 п.о., приблизительно 25 п.о., приблизительно 30 п.о., приблизительно 35п.о., приблизительно 40 п.о., приблизительно 45 п.о., приблизительно 50 п.о., приблизительно 55 п.о., приблизительно 60 п.о., приблизительно 65 п.о., приблизительно 70 п.о., приблизительно 75 п.о., приблизительно 80 п.о., приблизительно 85 п.о., приблизительно 90 п.о., приблизительно 95 п.о., приблизительно 100 п.о., приблизительно 110 п.о., приблизительно 120 п.о., приблизительно 130, приблизительно 140 п.о., приблизительно 150 п.о., приблизительно 200 п.о., приблизительно 250 п.о., приблизительно 300 п.о., приблизительно 350 п.о., приблизительно 400 п.о., приблизительно 450 п.о. или приблизительно 500 п.о. Ожидают, что технологические преимущества сделают возможными риды одиночных концов длиной более 500 п.о., что сделает возможными риды длиной более приблизительно 1000 п.о. при получении ридов спаренных концов. Согласно одному варианту реализации картированные метки последовательности содержат последовательности ридов, которые составляют 36 п.о. Картирования меток последовательности достигают посредством сравнения последовательности метки с последовательностью референса для определения хромосомного происхождения молекулы секвенированной нуклеиновой кислоты (например, сцДНК), и конкретная генетическая информация о последовательности не является необходимой. Можно допустить небольшую степень несоответствия (0-2 несовпадения на метку последовательности), которая приходится на минимальные полиморфизмы, которые могут существовать между референсным геномом и геномами в смешанном образце.

На образец, как правило, получают множество меток последовательности. Согласно некоторым вариантам реализации из картирования ридов с референсным геномом получают по меньшей мере приблизительно 3×10^6 меток последовательности, по меньшей мере приблизительно 5×10^6 меток последовательности, по меньшей мере приблизительно 8×10^6 меток последовательности, по меньшей мере приблизительно 10×10^6 меток последовательности, по меньшей мере приблизительно 15×10^6 меток последовательности, по меньшей мере приблизительно 20×10^6 меток последовательности, по меньшей мере приблизительно 30×10^6 меток последовательности, по меньшей мере приблизительно 40×10^6 меток последовательности или по меньшей мере приблизительно 50×10^6 меток последовательности, содержащих риды длиной от 20 до 40 п.о., например, 36 п.о., на образец. Согласно одному варианту реализации все риды последовательности картируют на все области референсного генома. Согласно одному варианту реализации подсчитывают метки, которые были картированы на все области, например все хромосомы, референсного генома, и определяют ВЧК, т.е. чрезмерную или недостаточную представленность последовательности, представляющей интерес, например, хромосомы или ее части, в смешанном образце ДНК. Способу не требуется установление отличий между двумя геномами.

Точность, необходимая для правильного определения того, присутствует или отсутствует ВЧК, например, анеуплоидия, в образце, основывается на вариации количества меток последовательности, которые картируются на референсный геном, среди образцов в пределах серии секвенирования (внутрихромосомная вариабельность) и вариации количества меток последовательности, которые картируются на референсный геном, в различных сериях секвенирования (вариабельность между секвенированиями). Например, вариации могут быть в особенности ярко выраженными для меток, которые картируются на GC-обогащенные или GC-обедненные референсные последовательности. Другие вариации могут являться следствием применения различных протоколов для экстракции и очистки нуклеиновых кислот, получения библиотек секвенирования и применения различных платформ секвенирования. В настоящем способе применяют дозы последовательностей (дозы хромосомы или дозы сегмента) на основании знания нормирующих последовательностей (последовательностей нормирующей хромосомы или последовательностей нормирующего сегмента), чтобы в действительности учесть накопленную вариабельность, которая является следствием межхромосомной вариабельности (в одной серии определений) и вариа-

бельности между секвенированиями (в нескольких сериях определений) и зависимой от платформы вариабельности. Дозы хромосом основаны на знании последовательности нормирующей хромосомы, которая может состоять из одной хромосомы или двух или более хромосом, выбранных из хромосом 1-22, X и Y. В качестве альтернативы, последовательности нормирующей хромосомы могут состоять из одного сегмента хромосомы или двух или более сегментов одной хромосомы либо двух или более хромосом. Дозы сегмента основаны на знании последовательности нормирующего сегмента, которая может состоять из одного сегмента любой одной хромосомы либо двух или более сегментов любых двух или более из хромосом 1-22, X и Y.

ВЧК и пренатальная диагностика.

Бесклеточную ДНК и РНК плода, циркулирующую в материнской крови, можно применять для ранней неинвазивной пренатальной диагностики (НПД) увеличивающегося количества генетических состояний как для ведения беременности, так и для способствования принятию решений в области репродукции. О присутствии бесклеточной ДНК, циркулирующей в сосудистом русле, стало известно более 50 лет назад. Совсем недавно в материнском сосудистом русле в течение беременности было обнаружено присутствие небольших количеств циркулирующей ДНК плода (Lo et al., Lancet 350:485-487 [1997]). Было показано, что бесклеточная ДНК плода (сцДНК), как считают, происходящая из гибнущих плацентарных клеток, состоит из коротких фрагментов, как правило, менее 200 п.о. в длину (Chan et al., Clin Chem 50:88-92 [2004]), которые можно выявить еще на 4 неделе гестации (Illanes et al., Early Human Dev 83:563-566 [2007]) и которые установленно выводятся из материнского сосудистого русла в течение часов после появления (Lo et al., Am J Hum Genet 64:218-224 [1999]). В дополнение к сцДНК, в материнском сосудистом русле также можно распознать фрагменты бесклеточной РНК плода (cfRNA), полученные из генов, которые транскрибируются в плоде или плаценте. Экстракция и последующий анализ данных генетических элементов плода из материнского образца крови предоставляет новые возможности для НПД.

Настоящий способ представляет собой независимый от полиморфизма способ, который предназначен для применения в НПД и для которого не требуется установления отличий сцДНК плода от материнской сцДНК, что делает возможным определение анеуплоидии плода. Согласно некоторым вариантам реализации анеуплоидия представляет собой полную трисомию или моносомию хромосомы либо частичную трисомию или моносомию. Частичные анеуплоидии вызваны утратой или приобретением части хромосомы и включают хромосомный дисбаланс, который является следствием несбалансированной транслокации, несбалансированных инверсий, делеций и инсерций. Безусловно, наиболее распространенная известная анеуплоидия, совместимая с жизнью, представляет собой трисомию 21, т.е. синдром Дауна (СД), вызванный присутствием части или всей хромосомы 21. Редко СД может быть вызван врожденным или спорадическим дефектом, в результате которого к другой хромосоме (обычно хромосоме 14) присоединяется дополнительная копия всей или части хромосомы 21 с образованием одной аберрантной хромосомы. СД связан с умственным расстройством, серьезными затруднениями в учебе и повышенной смертностью, вызванной длительными нарушениями состояния здоровья, такими как заболевание сердца. Другие анеуплоидии с известной клинической значимостью включают синдром Эдвардса (трисомию 18) и синдром Патау (трисомию 13), которые часто являются смертельными в течение первых нескольких месяцев жизни. Аномалии, связанные с количествами половых хромосом, также известны и включают моносомию X, например, синдром Тернера (XO) и синдром тройной X (XXX) у новорожденных женского пола и синдром Клайнфельтера (XXY) и синдром XYY у новорожденных мужского пола, все из которых связаны с различными фенотипами, включая бесплодие и снижение интеллектуальных способностей.

Моносомия X [45, X] является частой причиной ранней потери беременности, на которую приходится приблизительно 7% самопроизвольных абортов. На основании частоты живорожденных с 45,X (также называется синдромом Тернера) 1 - 2/10000, согласно оценкам, менее 1% оплодотворенных яйцеклеток 45,X выживут к сроку. Приблизительно 30% пациентов с синдромом Тернера представляют собой мозаики с линией клеток 45,X и линией клеток 46,XX или одной линией клеток, содержащей реорганизованную X-хромосому (Hook and Warburton 1983). Фенотип у живорожденных младенцев является относительно умеренным, учитывая высокую смертность плодов, и было высказано предположение, что, возможно, все живорожденные младенцы женского пола с синдромом Тернера несут линию клеток, содержащих две половые хромосомы. Моносомия X может возникнуть у субъектов женского пола в виде 45,X или в виде 45,X/46XX, и у субъектов мужского пола - в виде 45,X/46XY. Аутосомные моносомии у человека, как правило, считают несовместимыми с жизнью; однако существует достаточно много цитогенетических сообщений, в которых описана полная моносомия одной хромосомы 21 у живорожденных детей (Vosranova Iet al., Molecular Cytogen. 1:13 [2008]; Joosten et al., Prenatal Diagn. 17:271-5 [1997]. Способ, описанный в настоящем документе, можно применять для диагностики данных и других аномалий хромосом пренатальным способом.

Согласно некоторым вариантам реализации способы, раскрытые в настоящем документе, могут определить присутствие или отсутствие трисомий хромосом любой одной из хромосом 1-22, X и Y. Примеры трисомий хромосом, которые можно обнаружить согласно настоящему способу, включают, без огра-

ничения, трисомию 21 (T21; синдром Дауна), трисомию 18 (T18; синдром Эдвардса), трисомию 16 (T16), трисомию 20 (T20), трисомию 22 (T22; синдром кошачьего глаза), трисомию 15 (T15; синдром Прадера-Вилли), трисомию 13 (T13; синдром Патау), трисомию 8 (T8; синдром Варкани), трисомию 9 и XXY (синдром Клайнфельтера), трисомию XYY или XXX. Полные трисомий других аутосом, существующие в немозаичном состоянии, являются смертельными, но могут быть совместимы с жизнью, когда присутствуют в мозаичном состоянии. Следует принимать во внимание, что различные полные трисомий, будь то существующие в мозаичном или немозаичном состоянии, и частичные трисомий можно определить в сцДНК плода в соответствии с идеями, представленными в настоящем документе.

Неограничивающие примеры частичных трисомий, которые можно определить настоящим способом, включают, без ограничения, частичную трисомию 1q32-44, трисомию 9p, трисомию 4 с мозаицизмом, трисомию 17p, частичную трисомию 4q26-qter, частичную трисомию 2p, частичную трисомию 1q и/или частичную трисомию 6p/моносомию 6q.

Способы, раскрытые в настоящем документе, также можно применять для определения моносомии хромосомы X, моносомии хромосомы 21 и частичной моносомии, такой как моносомия 13, моносомия 15, моносомия 16, моносомия 21 и моносомия 22, которые установленно связаны с выкидышем при беременности. Частичную моносомию хромосом, которая, как правило, причастна к полной анеуплоидии, можно также определить способом, описанным в настоящем документе. Неограничивающие примеры синдромов делеции, которые можно определить согласно настоящему способу, включают синдромы, вызванные частичными делециями хромосом. Примеры частичных делеций, которые можно определить согласно способам, описанным в настоящем документе, включают, без ограничения, частичные делеции хромосом 1, 4, 5, 7, 11, 18, 15, 13, 17, 22 и 10, которые описаны ниже.

Синдром делеции 1q21.1 или микроделеции 1q21.1 (рецидивирующий) представляет собой редкую аберрацию хромосомы 1. Наряду с синдромом делеции, существует также синдром дупликации 1q21.1. Хотя существует часть ДНК, не содержащая синдром делеции в конкретной точке, существует две или три копии аналогичной части ДНК в той же точке с синдромом дупликации. Литература относит как делецию, так и дупликацию к вариациям числа копий (ВЧК) 1q21.1. Делеция 1q21.1 может быть связана с синдромом TAR (Thrombocytopenia with Absent radius, тромбоцитопения с отсутствием лучевой кости).

Синдром Вольфа-Хиршхорна (Wolf-Hirschhorn syndrome, WHS) (OMIN (Online Mendelian Inheritance in Man, онлайн-каталог фенетических маркеров у человека) № 194190) представляет собой синдром сплошной делеции гена, связанный с гемизиготной делецией хромосомы 4p16.3. Синдром Вольфа-Хиршхорна представляет собой синдром врожденного порока развития, который характеризуется пре- и постнатальной недостаточностью роста, расстройством развития различной степени тяжести, характерными черепно-мозговыми чертами (внешний вид носа по типу "шлема греческого воина", высокий лоб, выдающиеся глабеллы, гипертелоризм, высокие дугообразные брови, выпуклые глаза, эпикантальные складки, короткий подносовой желобок, четко очерченный рот с опущенными вниз уголками и микрогнатия) и эпилепсией.

Частичная делеция хромосомы 5, также известная как 5p- или 5p минус и названная синдромом Cris du Chat (OMIN№ 123450), вызвана делецией короткого плеча (p-плеча) хромосомы 5 (5p15.3-p15.2). У младенцев с данным состоянием часто наблюдается пронзительный крик, который часто похож на крик кошки. Расстройство характеризуется умственной отсталостью и задержкой развития, небольшим размером головы (микроцефалия), низкой массой тела при рождении и слабым тонусом мышц (гипотония) в младенческом возрасте, характерными чертами лица и, возможно, пороками сердца.

Синдром Уильямса-Бойрена, также известный как синдром делеции хромосомы 7ql 1.23 (OMIN 194050), представляет собой синдром сплошной делеции гена, который приводит к мультисистемному нарушению, вызванному гемизиготной делецией размером от 1,5 до 1,8 Мб на хромосоме 7q11.23, которая содержит приблизительно 28 генов.

Синдром Якобсена, также известный как нарушение, вызванное делецией 11q, представляет собой редкий врожденный порок развития, который является следствием делеции концевой области хромосомы 11, содержащей полосу 11q24.1. Данный синдром может вызывать умственную отсталость, характерные черты лица и множество физических нарушений, включая пороки сердца и нарушение свертываемости крови.

Частичная моносомия хромосомы 18, известная как моносомия 18p, представляет собой редкое хромосомное нарушение, при котором вся хромосома 18 или часть ее короткого плеча (p) делетированы (моносомический). Расстройство, как правило, характеризуется низким ростом, различной степенью задержки умственного развития, задержкой развития речи, врожденными пороками черепа и области лица (черепно-лицевой области) и/или дополнительными физическими аномалиями. Связанные черепно-лицевые дефекты могут в значительной степени варьировать от случая к случаю по диапазону и тяжести.

Состояния, вызванные изменениями структуры или количества копий хромосомы 15, включают синдром Эйнджелмена и синдром Прадера-Вилли, которые включают утрату активности гена в одной и той же части хромосомы 15, области 15q11-q13. Следует принимать во внимание, что у родителя-носителя несколько транслокаций и микроделеций могут являться бессимптомными, и при этом они могут вызвать значительное генетическое заболевание у потомства. Например, здоровая мать, которая не-

сет микроделецию 15q11-q13, может родить ребенка с синдромом Эйнджелмена, тяжелым нейродегенеративным нарушением. Таким образом, способы, аппараты и системы, описанные в настоящем документе, можно применять для идентификации такой частичной делеции и других делеций у плода.

Частичная моносомия 13q представляет собой редкое хромосомное нарушение, которое возникает, когда утрачивается часть длинного плеча (q) хромосомы 13 (моносомический). Младенцы, рожденные с частичной моносомией 13q, могут демонстрировать низкую массу тела при рождении, врожденные пороки головы и лица (черепно-лицевой области), аномалии скелета (в особенности, рук и стоп) и другие физические аномалии. Для данного состояния характерны задержки умственного развития. Среди индивидуумов, рожденных с данным нарушением, высок уровень смертности в младенчестве. Практически все случаи частичной моносомии 13q возникают случайным образом по неясным причинам (спорадически).

Синдром Смита-Магениса (Smith-Magenis syndrome, SMS - OMIM № 182290) вызван делецией или утратой генетического материала на одной копии хромосомы 17. Данный хорошо известный синдром связан с задержкой в развитии, задержкой умственного развития, врожденными пороками развития, такими как пороки сердца и почек, и нейроповеденческими аномалиями, такими как тяжелые нарушения сна и самотравмирующее поведение. Синдром Смита-Магениса (SMS) в большинстве случаев (90%) вызван внутренней делецией размером 3,7 Мб в хромосоме 17p11.2.

Синдром делеции 22q11.2, также известный как синдром Ди Георге, представляет собой синдром, вызванный делецией небольшого фрагмента хромосомы 22. Делеция (22 q11.2) возникает возле середины хромосомы на длинном плече одной из пары хромосом. Признаки данного синдрома широко варьируют даже среди членов одной семьи и затрагивают многие части тела. Характерные черты и симптомы могут включать врожденные пороки развития, такие как врожденные заболевание сердца, патологии нёба, наиболее часто, в отношении нейромышечных нарушений смыкания (небно-глоточная недостаточность), нарушение обучаемости, незначительные отличия черт лица и рецидивирующие инфекции. Микроделеции в хромосомной области 22q11.2 связаны с в 20 - 30 раз увеличенным риском шизофрении.

Делеции в коротком плече хромосомы 10 связаны с фенотипом, подобным синдрому Ди Георге. Частичная моносомия хромосомы 10p является редкой, но свойственной части пациентов, у которых наблюдались черты синдрома Ди Георге.

Согласно одному варианту реализации способы, аппараты и системы, описанные в настоящем документе, применяют для определения частичной моносомии, включая, без ограничения, частичную моносомию хромосом 1, 4, 5, 7, 11, 18, 15, 13, 17, 22 и 10, например, частичную моносомию 1q21.11, частичную моносомию 4p16.3, частичную моносомию 5p15.3-p15.2, частичную моносомию 7q11.23, частичную моносомию 11q24.1, частичную моносомию 18p, частичную моносомию хромосомы 15 (15q11-q13), частичную моносомию 13q, частичную моносомию 17p11.2, частичную моносомию хромосомы 22 (22q11.2), и частичную моносомию 10p также можно определить с применением данного способа.

Другие частичные моносомии, которые можно определить согласно способам, описанным в настоящем документе, включают несбалансированную транслокацию t(8;11)(p23.2;p15.5); микроделецию 11q23; делецию 17p11.2; делецию 22q13.3; микроделецию Xp22.3; делецию 10p14; микроделецию 20p, [del(22)(q11.2q11.23)], делецию 7q11.23 и 7q36; делецию 1p36; микроделецию 2p; нейрофиброматоз типа 1 (микроделецию 17q11.2), делецию Yq; микроделецию 4p16.3; микроделецию 1p36.2; делецию 11q14; микроделецию 19q13.2; синдром Рубинштейна-Тэйби (микроделецию 16 p13.3); микроделецию 7p21; синдром Миллера-Дикера (17p13.3); и микроделецию 2q37. Частичные делеции могут представлять собой небольшую делецию части хромосомы или могут представлять собой микроделеции хромосомы, когда может возникнуть делеция одного гена.

Было идентифицировано несколько синдромов дупликации, вызванных дупликациями части плеч хромосомы (см. OMIN [Online Mendelian Inheritance in Man, онлайн-каталог фенетических маркеров у человека, доступный онлайн по адресу: ncbi.nlm.nih.gov/omim]). Согласно одному варианту реализации настоящий способ можно применять для определения присутствия или отсутствия дупликации и/или умножения сегментов любой из хромосом 1 -22, X и Y. Неограничивающие примеры синдромов дупликации, которые можно определить согласно настоящему способу, включают дупликации части хромосом 8, 15, 12 и 17, которые описаны ниже.

Синдром дупликации 8p23.1 представляет собой редкое генетическое нарушение, вызванное дупликацией области хромосомы 8 человека. Данный синдром дупликации характеризуется оцениваемой распространенностью 1 на 64000 рождений и противоположен синдрому делеции 8p23.1. Дупликация 8p23.1 связана с различными фенотипами, включая один или более фенотипов, которые выбраны из задержки развития речи, задержки в развитии, умеренного дизморфизма с выпуклым лбом и изогнутыми бровями и врожденного заболевания сердца (ВЗС).

Синдром дупликации хромосомы 15q (Dup15q) представляет собой идентифицируемый клиническим способом синдром, который является следствием дупликации хромосомы 15q11-13.1 Младенцы с Dup15q обычно характеризуются гипотонией (слабым тонусом мышц), отставанием в росте; они могут родиться с расщепленной губой и/или нёбом либо с врожденными пороками сердца, почек или других органов; у них наблюдается некоторая степень задержки развития когнитивных функций/ограничение

когнитивных функций (задержки умственного развития), задержки развития речи и понимания языка и дисфункция сенсорной интеграции.

Синдром Паллистера-Киллиана является следствием дополнительного материала хромосомы № 12. Обычно существует смесь клеток (мозаицизм), некоторые из которых содержат дополнительный материал № 12, а некоторые являются нормальными (46 хромосом без дополнительного материала № 12). Младенцы с данным синдромом характеризуются многими нарушениями, включая тяжелые задержки умственного развития, слабый тонус мышц, "грубые" черты лица и выпуклый лоб. У них наблюдается тенденция иметь очень тонкую верхнюю губу с более толстой нижней губой и коротким носом. Другие нарушения здоровья включают эпилепсию, плохое усваивание питания, онемение суставов, катаракту во взрослом возрасте, потерю слуха и пороки сердца. Лица с синдромом Паллистера-Киллиана характеризуются укороченной продолжительностью жизни.

Индивидуумы с генетическим состоянием, обозначаемым как dup(17)(p11.2p11.2) или dup17p, несут дополнительную генетическую информацию (известна как дупликация) на коротком плече хромосомы 17. Дупликация хромосомы 17p11.2 лежит в основе синдрома Потоцки-Лупски (Potocki-Lupski syndrome, PTLS), который представляет собой недавно обнаруженное генетическое состояние с исключительно несколькими десятками случаев, о которых сообщалось в медицинской литературе. Пациенты, у которых присутствует данная дупликация, часто характеризуются низким тонусом мышц, плохим усваиванием питания и отсутствием прибавки в весе в младенчестве, а также характеризуются задержкой развития моторных и вербальных показателей развития. Многие индивидуумы с PTLS страдают от трудностей с произношением и с обработкой лингвистической информации. Помимо этого, пациенты могут характеризоваться поведенческими характеристиками, аналогичными таковым, наблюдаемым у лиц с аутизмом или нарушениями аутического спектра. Индивидуумы с PTLS могут характеризоваться пороками сердца и апноэ во сне. Дупликация большой области в хромосоме 17p12, которая содержит ген PMP22, как известно, вызывает заболевание Шарко-Мари-Тута.

ВЧК связана с рождением мертвого плода. Однако в связи с ограничениями, присущими общепринятой цитогенетике, вклад ВЧК в рождение мертвого плода, как считают, является недостаточно представленным (Harris et al., Prenatal Diagn 31:932-944 [2011]). Как показано в примерах и описано в другом месте в настоящем документе, настоящий способ позволяет определять присутствие частичных анеуплоидий, например, делеций и умножений сегментов хромосомы, и данный способ можно применять для идентификации и определения присутствия или отсутствия ВЧК, которые связаны с рождением мертвого плода.

Определение ВЧК при клинических нарушениях.

В дополнение к раннему определению врожденных пороков развития способы, описанные в настоящем документе, можно применять для определения любых аномалий представления генетических последовательностей в геноме. Множество аномалий представления генетических последовательностей в геноме связаны с различными патологиями. Такие патологии включают, без ограничения, рак, инфекционные и аутоиммунные заболевания, заболевания нервной системы, метаболические и/или сердечно-сосудистые заболевания и т.п.

Соответственно, согласно различным вариантам реализации предусмотрено применение способов, описанных в настоящем документе, при диагностике и/или мониторинге и/или лечении таких патологий. Например, способы можно применять для определения присутствия или отсутствия заболевания, для контроля прогрессирования заболевания и/или эффективности режима лечения, для определения присутствия или отсутствия нуклеиновых кислот патогена, например, вируса; для определения хромосомных аномалий, связанных с реакцией "трансплантат против хозяина" (РТПХ), и для определения причастности индивидуумов в криминалистических анализах.

ВЧК при раке.

Было показано, что ДНК плазмы и сыворотки крови от пациентов, страдающих от рака, содержит поддающиеся измерению количества опухолевой ДНК, которую можно восстановить и применять в качестве заменителя источника опухолевой ДНК, и опухоли характеризуются анеуплоидией или несоответствующими количествами последовательностей гена или даже целых хромосом. Таким образом, определение отличия количества данной последовательности, т.е. последовательности, представляющей интерес, в образце от индивидуума можно применять при прогнозировании или диагностике медицинского состояния. Согласно некоторым вариантам реализации настоящий способ можно применять для определения присутствия или отсутствия анеуплоидии хромосом у пациента, который, как подозревают или как известно, страдает от рака.

Согласно некоторым вариантам реализации в настоящем документе предложены способы обнаружения рака, отслеживания терапевтического ответа и минимального остаточного заболевания на основании образцов циркулирующей сцДНК с применением неглубокого секвенирования образцов с помощью методологии спаренных концов и с применением информации о размере фрагмента, доступной из ридов спаренных концов, для идентификации присутствия избирательно метилированной апоптотической ДНК из раковых клеток на фоне нормальных клеток. Было показано, что при некоторых типах рака полученная из опухоли сцДНК является более короткой, чем сцДНК, полученная не из опухоли. Вследствие это-

го способ на основании размера, описанный в настоящем документе, можно применять для определения ВЧК, включая анеуплоидии, связанные с данными типами рака, который делает возможным (a) обнаружение опухоли, присутствующей в условиях скрининга или диагностики; (b) контроль ответа на терапию; (c) контроль минимального остаточного заболевания.

Согласно определенным вариантам реализации анеуплоидия является характерной для генома субъекта и приводит, как правило, к увеличенной предрасположенности к раку. Согласно определенным вариантам реализации анеуплоидия является характерной для конкретных клеток (например, опухолевых клеток, предопухолевых неопластических клеток и т.д.), которые являются или характеризуются увеличенной предрасположенностью к неоплазии. Конкретные анеуплоидии связаны с конкретными типами рака или с предрасположенностью к конкретным типам рака, как описано ниже. Согласно некоторым вариантам реализации для обнаружения/контроля присутствия рака экономически выгодным способом можно применять подход очень неглубокого секвенирования спаренных концов.

Соответственно, различные варианты реализации способов, описанных в настоящем документе, обеспечивают определение вариации числа копий последовательности или последовательностей, представляющих интерес, например, клинически значимой последовательности или последовательностей, в исследуемом образце от субъекта, причем определенные вариации числа копий обеспечивают свидетельство присутствия и/или предрасположенности к раку. Согласно определенным вариантам реализации образец содержит смесь нуклеиновых кислот, полученных из двух или более типов клеток. Согласно одному варианту реализации смесь нуклеиновых кислот получена из нормальных и раковых клеток, полученных от субъекта, страдающего от медицинского состояния, например, рака.

Развитие рака часто сопровождается изменением количества целых хромосом, т.е. полной анеуплоидией хромосом, и/или изменением количества сегментов хромосом, т.е. частичной анеуплоидией, вызванной процессом, известным как нестабильность хромосом (НХ) (Thoma et al., Swiss Med Weekly 2011:141:wl3170). Считают, что многие солидные опухоли, такие как рак молочной железы, прогрессируют от возникновения к метастазированию вследствие накопления нескольких генетических аберраций. [Sato et al., Cancer Res., 50: 7184-7189 [1990]; Jongsma et al., J Clin Pathol: Mol Path 55:305-309 [2002])]. Такие генетические аберрации, по мере того как они накапливаются, могут обеспечить пролиферативные преимущества, генетическую нестабильность и сопутствующую способность быстро развивать устойчивость к лекарственным средствам и усиленный ангиогенез, протеолиз и метастазирование. Генетические аберрации могут затрагивать рецессивные "гены-онкосупрессоры" или доминантно функционирующие онкогены. Делеции и рекомбинация, приводящие к потере гетерозиготности (ПГ), как считают, играют основную роль в прогрессировании опухоли в результате выявления мутированных аллелей онкосупрессора.

сцДНК была обнаружена в сосудистом русле пациентов, у которых были диагностированы злокачественные новообразования, включая, без ограничения, рак легких (Pathak et al., Clin Chem 52:1833-1842 [2006]), рак предстательной железы (Schwartzenbach et al., Clin Cancer Res 15:1032-8 [2009]) и рак молочной железы (Schwartzenbach et al., публикация доступна онлайн по адресу: breast-cancer-research.com/content/11/5/R71 [2009]). Идентификация геномных нестабильностей, связанных с типами рака, которые можно определить в циркулирующей сцДНК у пациентов, страдающих от рака, является перспективным диагностическим и прогностическим инструментом. Согласно одному варианту реализации способы, описанные в настоящем документе, применяют для определения ВЧК одной или более последовательностей, представляющих интерес, в образце, например, образце, содержащем смесь нуклеиновых кислот, полученных от субъекта, который, как предполагают или как известно, страдает от рака, например, карциномы, саркомы, лимфомы, лейкоза, герминогенных опухолей и бластомы. Согласно одному варианту реализации образец представляет собой образец плазмы, полученный (процессированный) из периферической крови, который может содержать смесь сцДНК, полученной из нормальных и раковых клеток. Согласно другому варианту реализации биологический образец, необходимый для определения присутствия ВЧК, получен из клеток, которые в случае присутствия рака, включают смесь раковых и нераковых клеток от других биологических тканей, включая, без ограничения, биологические жидкости, такие как сыворотка, пот, слезы, мокрота, моча, мокрота, ушная жидкость, лимфа, слюна, спинномозговая жидкость, жидкость после лаважа, суспензия костного мозга, влагалищная жидкость, жидкость после трансцервикального лаважа, жидкость головного мозга, асцит, молоко, секреты дыхательных, кишечных и мочеполовых путей и образцы лейкафереза, или в биопсиях ткани, мазках или соскобах. Согласно другим вариантам реализации биологический образец представляет собой образец стула (фекалий).

Способы, описанные в настоящем документе, не ограничены анализом сцДНК. Следует понимать, что аналогичные анализы можно проводить в отношении образцов клеточной ДНК.

Согласно различным вариантам реализации последовательность или последовательности, представляющие интерес, содержат последовательность нуклеиновой кислоты или кислот, которые, как известно или как предполагают, играют роль в развитии и/или прогрессировании рака. Примеры последовательности, представляющей интерес, включают последовательности нуклеиновых кислот, например, полных хромосом и/или сегментов хромосом, которые амплифицированы или делетированы в раковых клетках,

как описано ниже.

Суммарное количество ВЧК и риск развития рака.

Каждые из общих ОНП (однонуклеотидных полиморфизмов) рака - и по аналогии общих ВЧК рака - могут вызывать исключительно незначительное повышение риска развития заболевания. Однако в совокупности ОНП и ВЧК могут вызывать по существу повышенный риск типов рака. В этой связи следует отметить, что добавления и утраты больших сегментов ДНК зародышевой линии, как сообщалось, являются факторами, предрасполагающими индивидуумов к нейробластоме, раку предстательной железы и толстой и прямой кишок, раку молочной железы и BRCA1-ассоциированному раку яичников (см., например, публикации Krepischi et al. Breast Cancer Res., 14: R24 [2012]; Diskin et al., Nature 2009, 459:987-991; Liu et al. Cancer Res 2009, 69: 2176-2179; Lucito et al., Cancer Biol Ther 2007, 6:1592-1599; Thean et al., Genes Chromosomes Cancer 2010, 49:99-106; Venkatachalam et al., Int J Cancer 2011, 129:1635-1642; и Yoshihara et al., Genes Chromosomes Cancer 2011, 50:167-177). Следует отметить, что ВЧК, часто обнаруживаемые в здоровой популяции (общие ВЧК), как считают, играют роль в этиологии рака (см., например, публикацию Shlien and Malkin (2009) Genome Medicine, 1(6): 62). В одном исследовании, в котором изучали предположение, что общие ВЧК связаны со злокачественным новообразованием (Shlien et al., Proc Natl Acad Sci USA 2008, 105:11264-11269), получили карту каждой известной ВЧК, локус которой совпадает с таковым подлинных связанных с раком генов (каталог которых составлен в публикации Higgins et al., Nucleic Acids Res 2007, 35:D721-726). Данные ВЧК были названы "ВЧК рака". В исходном анализе (Shlien et al., Proc Natl Acad Sci USA 2008, 105:11264-11269) 770 здоровых геномов оценивали с применением набора матриц Affymetrix 500K, которые характеризуются средним расстоянием между зондами 5,8 т.о. Поскольку считают, что ВЧК, как правило, истощены в областях генов (Redon et al. (2006) Nature 2006, 444:444-454), было неожиданно обнаружить 49 раковых генов, которые были непосредственно включены в ВЧК или перекрывались ВЧК, у более одного лица в большой референсной популяции. Среди первых десяти генов ВЧК рака можно было обнаружить у четырех или более человек.

Таким образом, считают, что частоту ВЧК можно применять в качестве критерия риска развития рака (см., например, публикацию патента США № 2010/0261183 A1). Частоту ВЧК можно определить простым способом на основании конститутивного генома организма, или она может представлять фракцию, полученную из одной или более опухолей (неопластических клеток), если таковые присутствуют.

Согласно определенным вариантам реализации количество ВЧК в исследуемом образце (например, образце, содержащем конститутивные (зародышевой линии) нуклеиновые кислоты) или смеси нуклеиновых кислот (например, нуклеиновая кислота зародышевой линии и нуклеиновая кислота или кислоты, полученные из неопластических клеток) определяют с применением способов, описанных в настоящем документе для вариаций числа копий. Идентификация увеличенного количества ВЧК в исследуемом образце, например, по сравнению с референсным значением, свидетельствует о риске или предрасположенности к раку у субъекта. Следует принимать во внимание, что референсное значение в данной популяции может варьировать. Также следует принимать во внимание, что абсолютное значение увеличения частоты ВЧК будет варьировать в зависимости от разрешения способа, применяемого для определения частоты ВЧК, и других параметров. Как правило, увеличение частоты ВЧК по меньшей мере приблизительно в 1,2 раза по сравнению с референсным значением будет определено как свидетельствующее о риске развития рака (см., например, публикацию патента США № 2010/0261183 A1), например, увеличение частоты ВЧК по меньшей мере или приблизительно в 1,5 раза по сравнению с референсным значением или более, такое как увеличение в 2-4 раза по сравнению с референсным значением, является свидетельством повышенного риска развития рака (например, по сравнению с нормальной здоровой референсной популяцией).

Считают, что определение структурной вариации в геноме млекопитающего по сравнению с референсным значением также свидетельствует о риске развития рака. В данном контексте согласно одному варианту реализации термин "структурная вариация" можно обозначить как частоту ВЧК у млекопитающего, умноженную на средний размер ВЧК (в и.о.) у млекопитающего. Таким образом, высокие показатели структурной вариации будут являться следствием увеличения частоты ВЧК и/или возникновения больших делеций или дупликаций геномной нуклеиновой кислоты. Соответственно, согласно определенным вариантам реализации количество ВЧК в исследуемом образце (например, образце, содержащем конститутивную (зародышевой линии) нуклеиновую кислоту) определяют с применением способов, описанных в настоящем документе для определения размера и количества вариаций числа копий. Согласно определенным вариантам реализации суммарный показатель структурной вариации в геномной ДНК более приблизительно 1 мегабазы, или более приблизительно 1,1 мегабаз, или более приблизительно 1,2 мегабаз, или более приблизительно 1,3 мегабаз, или более приблизительно 1,4 мегабаз, или более приблизительно 1,5 мегабаз, или более приблизительно 1,8 мегабаз или более приблизительно 2 мегабаз ДНК свидетельствует о риске развития рака.

Считают, что данные способы обеспечивают критерий риска развития любого рака, включая, без ограничения, острый и хронический лейкозы, лимфомы, многочисленные солидные опухоли мезенхимальной или эпителиальной ткани, рак головного мозга, молочной железы, печени, желудка, толстой кишки, В-клеточную лимфому, рак легких, рак бронхов, рак толстой и прямой кишок, рак предстатель-

ной железы, рак молочной железы, рак поджелудочной железы, рак желудка, рак яичников, рак мочевого пузыря, рак головного мозга или центральной нервной системы, рак периферической нервной системы, рак пищевода, рак шейки матки, меланому, рак матки или эндометрия, рак ротовой полости или гортани, рак печени, рак почек, рак желчных путей, рак тонкого кишечника или аппендикса, рак слюнных желез, рак щитовидной железы, рак надпочечников, остеосаркому, хондросаркому, липосаркому, рак яичек и злокачественную фиброзную гистиоцитому, а также другие типы рака.

Анеуплоидии целых хромосом.

Как указано выше, при раке наблюдается высокая частота анеуплоидии. В определенных исследованиях, в которых изучали распространенность изменений соматического числа копий (somatic copy number alterations, SCNA) при раке, было обнаружено, что одна четверть генома типичной раковой клетки поражена SCNA целого плеча или анеуплоидией SCNA целой хромосомы (см., например, публикацию Beroukhim et al., Nature 463: 899-905 [2010]). Изменения целой хромосомы периодически наблюдают при нескольких типах рака. Например, в 10 - 20% случаев острого миелоидного лейкоза (ОМЛ), а также некоторых солидных опухолей, включая саркому Юинга и десмоидные опухоли, наблюдается добавление хромосомы 8 (см., например, публикации Barnard et al., Leukemia 10: 5-12 [1996]; Maurici et al., Cancer Genet. Cytogenet. 100: 106-110 [1998]; Qi et al., Cancer Genet. Cytogenet. 92: 147-149 [1996]; Barnard, D. R. et al., Blood 100: 427-434 [2002]; и т.п. Иллюстративный, но неограничивающий перечень добавлений и утрат хромосом при типах рака человека представлен в табл. 2.

Таблица 2. Иллюстративные специфичные рецидивирующие добавления и утраты хромосом при раке человека (см., например, публикацию Gordon et al. (2012) Nature Rev. Genetics, 13: 189-203)

Хромосома	**Добавления** **Тип рака**	**Утраты** **Тип рака**
1	Множественная миелома Аденокарцинома (молочной железы)	Аденокарцинома (почек)
2	Гепатобластома Саркома Юинга	
3	Множественная миелома Диффузная В-крупноклеточная лимфома	Меланома Аденокарцинома (почек)
4	Острый лимфобластный лейкоз	Аденокарцинома (почек)
5	Множественная миелома Аденокарцинома (почек)	
6	Острый лимфобластный лейкоз Опухоль Вильмса	Аденокарцинома (почек)
7	Аденокарцинома (почек) Аденокарцинома (кишечника)	Острый миелоидный лейкоз Ювенильный миеломоноцитарный лейкоз
8	Острый миелоидный лейкоз Хронический миелоидный лейкоз Саркома Юинга	Аденокарцинома (почек)
9	Множественная миелома Истинная полицитемия	

10	Острый лимфобластный лейкоз Аденокарцинома (матки)	Астроцитома Множественная миелома
11	Множественная миелома	
12	Хронический лимфоцитарный лейкоз Опухоль Вильмса	Множественная миелома
13	Острый миелоидный лейкоз Опухоль Вильмса	Множественная миелома
14	Острый лимфобластный лейкоз	Аденокарцинома (почек) Менингиома
15	Множественная миелома	
16	Аденокарцинома (почек)	Множественная миелома
17	Аденокарцинома (почек) Острый лимфобластный лейкоз	
18	Острый лимфобластный лейкоз Опухоль Вильмса	Аденокарцинома (почек)
19	Множественная миелома Хронический миелоидный лейкоз	Аденокарцинома (молочной железы) Менингиома
20	Гепатобластома Аденокарцинома (почек)	
21	Острый лимфобластный лейкоз Острый мегакариобластный лейкоз	
22	Острый лимфобластный лейкоз	Менингиома
X	Острый лимфобластный лейкоз Фолликулярная лимфома	
Y		

Согласно различным вариантам реализации способы, описанные в настоящем документе, можно применять для обнаружения и/или количественного определения анеуплоидий целой хромосомы, которые связаны с раком в целом и/или которые связаны с конкретными типами рака. Таким образом, например, согласно определенным вариантам реализации предусмотрено обнаружение и/или количественное определение анеуплоидий целой хромосомы, которые характеризуются добавлением или утратой, представленной в табл. 2.

Вариации числа копий сегментов хромосомы на уровне плеча.

Во многих исследованиях сообщалось о паттернах вариаций числа копий на уровне плеча в пределах большого количества образцов рака (Lin et al., Cancer Res 68, 664-673 (2008); George et al., PLoS ONE 2, e255 (2007); Demichelis et al., Genes Chromosomes Cancer 48: 366-380 (2009); Beroukhim et al., Nature. 463(7283): 899-905 [2010). Дополнительно наблюдали, что частота вариаций числа копий на уровне плеча снижается с уменьшением длины плеч хромосом. С учетом данной тенденции для большинства плеч хромосом наблюдается весомое доказательство преимущественного добавления или утраты, но редко и того, и другого, в пределах множества линий рака (см., например, публикацию Beroukhim et al., Nature. 463(7283): 899-905 [2010]).

Соответственно, согласно одному варианту реализации способы, описанные в настоящем документе, применяют для определения ВЧК на уровне плеча (ВЧК, включающие одно плечо хромосомы или по существу одно плечо хромосомы) в образце. ВЧК можно определить в ВЧК в исследуемом образце, содержащем конститутивную (зародышевой линии) нуклеиновую кислоту, и ВЧК на уровне плеча можно идентифицировать в таких конститутивных нуклеиновых кислотах. Согласно определенным вариантам реализации ВЧК на уровне плеча идентифицируют (в случае наличия) в образце, содержащем смесь нуклеиновых кислот (например, нуклеиновые кислоты, полученные из нормальных, и нуклеиновые кислоты, полученные из неопластических клеток). Согласно определенным вариантам реализации образец получают от субъекта, который, как предполагают или как известно, страдает от рака, например, карциномы, саркомы, лимфомы, лейкоза, герминогенных опухолей, бластомы и т.п. Согласно одному варианту реализации образец представляет собой образец плазмы, полученный (процессированный) из периферической крови, который может содержать смесь сцДНК, полученной из нормальных и раковых клеток. Согласно другому варианту реализации биологический образец, который применяют для определения присутствия ВЧК, получают из клеток, которые, если рак присутствует, содержат смесь раковых и нераковых клеток из других биологических тканей, включая, без ограничения, биологические жидкости, такие как сыворотка, пот, слезы, мокрота, моча, мокрота, ушная жидкость, лимфа, слюна, спинномозговая жидкость, жидкость после лаважа, суспензия костного мозга, влагалищная жидкость, жидкость после трансцервикального лаважа, жидкость головного мозга, асцит, молоко, секреты дыхательных, кишечных

и мочеполовых путей и образцы лейкафереза, или в биопсиях ткани, мазках или соскобах. Согласно другим вариантам реализации биологический образец представляет собой образец стула (фекалий).

Согласно различным вариантам реализации ВЧК, идентифицированные как свидетельствующие о присутствия рака или о повышенном риске развития рака, включают, без ограничения, ВЧК на уровне плеча, перечисленные в табл. 3. Как проиллюстрировано в табл. 3, определенные ВЧК, которые включают существенное добавление на уровне плеча, свидетельствуют о присутствии рака или о повышенном риске развития определенных типов рака. Таким образом, например, добавление в 1q свидетельствует о присутствии или повышенном риске развития острого лимфобластного лейкоза (ОЛЛ), рака молочной железы, ЖКСО (желудочно-кишечной стромальной опухоли), ПКК (печеночноклеточной карциномы), НПК (неплоскоклеточной карциномы) легких, медуллобластомы, меланомы, МПС (миелопролиферативного синдрома), рака яичников и/или рака предстательной железы. Добавление в 3q свидетельствует о присутствии или повышенном риске развития плоскоклеточного рака пищевода, ПК (плоскоклеточной карциномы) легких и/или МПС. Добавление в 7q свидетельствует о присутствии или повышенном риске развития рака толстой и прямой кишок, глиомы, ПКК, НПК легких, медуллобластомы, меланомы, рака предстательной железы и/или ренального рака. Добавление в 7p свидетельствует о присутствии или повышенном риске развития рака молочной железы, рака толстой и прямой кишок, аденокарциномы пищевода, глиомы, ПКК, НПК легких, медуллобластомы, меланомы и/или ренального рака. Добавление в 20q свидетельствует о присутствии или повышенном риске развития рака молочной железы, рака толстой и прямой кишок, дедифференцированной липосаркомы, аденокарциномы пищевода, плоскоклеточного рака пищевода, рака глиомы, ПКК, НПК легких, меланомы, рака яичников и/или ренального рака и т.д.

Аналогично, как проиллюстрировано в табл. 3, определенные ВЧК, которые включают существенную утрату на уровне плеча, свидетельствуют о присутствии и/или о повышенном риске развития определенных типов рака. Таким образом, например, утрата в 1p свидетельствует о присутствии или повышенном риске развития желудочно-кишечной стромальной опухоли. Утрата в 4q свидетельствует о присутствии или повышенном риске развития рака толстой и прямой кишок, аденокарциномы пищевода, ПК легких, меланомы, рака яичников и/или ренального рака. Утрата в 17p свидетельствует о присутствии или повышенном риске развития рака молочной железы, рака толстой и прямой кишок, аденокарциномы пищевода, ПКК, НПК легких, ПК (плоскоклеточной карциномы) легких и/или рака яичников и т.п.

Таблица 3. Значительные изменения числа копий сегментов хромосом на уровне плеча в каждом из 16 подтипов рака (рак молочной железы, толстой и прямой кишок, дедифференцированная липосаркома, аденокарцинома пищевода, плоскоклеточный рак пищевода, ЖКСО (желудочно-кишечная стромальная опухоль), глиома, ПКК (печеночноклеточная карцинома), НПК легких, ПК легких, медуллобластома, меланома, МПС (миелопролиферативный синдром), рак яичников, предстательной железы, острый лимфобластный лейкоз (ОЛЛ) и ренальный рак) (см., например, публикацию Beroukhim et al., Nature (2010) 463(7283): 899-905)

Плечо	**Типы рака Значительное добавление в**	**Типы рака Значительная утрата в**	**Известный онкоген/ген-онкосупрессор**
1p	----	ЖКСО	
1q	ОЛЛ, молочной железы, ЖКСО, ПКК, НПК легких, медуллобластома, меланома, МПС, яичников, предстательной железы	----	
3p	----	Плоскоклеточный пищевода, НПК легких, ПК легких, ренальный	*VHL*
3q	Плоскоклеточный пищевода, ПК легких, МПС	----	
4p	ОЛЛ	Молочной железы, аденокарцинома пищевода, ренальный	
4q	ОЛЛ	Толстой и прямой кишок, аденокарцинома пищевода, ПК легких,	

		меланома, яичников, ренальный	
5p	Плоскоклеточный пищевода, ПКК, НПК легких, ПК легких, ренальный	----	*TERT*
5q	ПКК, ренальный	Аденокарцинома пищевода, НПК легких	*APC*
6p	ОЛЛ, ПКК, НПК легких, меланома	----	
6q	ОЛЛ	Меланома, ренальный	
7p	Молочной железы, толстой и прямой кишок, аденокарцинома пищевода, глиома, ПКК, НПК легких, медуллобластома, меланома, ренальный	----	*EGFR*
7q	Толстой и прямой кишок, глиома, ПКК, НПК легких, медуллобластома, меланома, предстательной железы, ренальный	----	*BRAF, MET*
8p	ОЛЛ, МПС	Молочной железы, ПКК, НПК легких, медуллобластома, предстательной железы, ренальный	
8q	ОЛЛ, молочной железы, толстой и прямой кишок, аденокарцинома пищевода, плоскоклеточный пищевода, ПКК, НПК легких, МПС, яичников, предстательной железы	Медуллобластома	*MYC*
9p	МПС	ОЛЛ, молочной железы, аденокарцинома пищевода, НПК легких, меланома, яичников, ренальный	*CDKN2A/B*
9q	ОЛЛ, МПС	НПК легких, меланома, яичников, ренальный	
10p	ОЛЛ	Глиома, ПК легких, меланома	

10q	ОЛЛ	Глиома, ПК легких, медуллобластома, меланома	*PTEN*
11p	----	Медуллобластома	*WT1*
11q	----	Дедифференцированная липосаркома, медуллобластома, меланома	*ATM*
12p	Толстой и прямой кишок, ренальный	----	*KRAS*
12q	Ренальный	----	
13q	Толстой и прямой кишок	Молочной железы, дедифференцированная липосаркома, глиома, НПК легких, яичников	*RB1/BRCA2*
14q	ОЛЛ, НПК легких, ПК легких, предстательной железы	ЖКСО, меланома, ренальный	
15q	----	ЖКСО, НПК легких, ПК легких, яичников	
16p	Молочной железы	----	
16q	----	Молочной железы, ПКК, медуллобластома, яичников, предстательной железы	
17p	ОЛЛ	Молочной железы, толстой и прямой кишок, аденокарцинома пищевода, ПКК, НПК легких, ПК легких, яичников	*TP53*
17q	ОЛЛ, ПКК, НПК легких, медуллобластома	Молочной железы, яичников	*ERBB2, NF1/BRCA1*
18p	ОЛЛ, медуллобластома	Толстой и прямой кишок, НПК легких	
18q	ОЛЛ, медуллобластома	Толстой и прямой кишок, аденокарцинома пищевода, НПК легких	*SMAD2, SMAD4*
19p	Глиома	Аденокарцинома пищевода, НПК легких, меланома, яичников	
19q	Глиома, ПК легких	Аденокарцинома пищевода, НПК легких	
20p	Молочной железы, толстой и прямой кишок, аденокарцинома пищевода, плоскоклеточный	----	

	пищевода, ЖКСО, глиома, ПКК, НПК легких, меланома, ренальный		
20q	Молочной железы, толстой и прямой кишок, дедифференцированная липосаркома, аденокарцинома пищевода, плоскоклеточный пищевода, глиома, ПКК, НПК легких, меланома, яичников, ренальный	----	
21q	ОЛЛ, ЖКСО, МПС	----	
22q	Меланома	Молочной железы, толстой и прямой кишок, дедифференцированная липосаркома, аденокарцинома пищевода, ЖКСО, НПК легких, ПК легких, яичников, предстательной железы	*NF2*

Примеры взаимосвязи между вариациями числа копий на уровне плеча являются иллюстративными, а не ограничивающими. Специалистам в данной области техники известны другие вариации числа копий на уровне плеча и их взаимосвязи с раком.

Меньшие, например фокальные, вариации числа копий.

Как указано выше, согласно определенным вариантам реализации способы, описанные в настоящем документе, можно применять для определения присутствия или отсутствия амплификации хромосом. Согласно некоторым вариантам реализации амплификация хромосом представляет собой добавление одной или более целых хромосом. Согласно другим вариантам реализации амплификация хромосом представляет собой добавление одного или более сегментов хромосомы. Согласно третьим вариантам реализации амплификация хромосом представляет собой добавление двух или более сегментов двух или более хромосом. Согласно различным вариантам реализации амплификация хромосом может включать добавление одного или более онкогенов.

Доминантно функционирующие гены, связанные с солидными опухолями человека, как правило, оказывают свое влияние посредством сверхэкспрессии или изменения экспрессии. Амплификация гена является частым механизмом, приводящим к повышающей регуляции экспрессии гена. Доказательства, полученные в цитогенетических исследованиях, указывают на то, что при более 50% типов рака молочной железы человека наблюдается значительная амплификация. Главным образом, амплификация протоонкогена рецептора эпидермального фактора роста 2 (HER2) человека, расположенного на хромосоме 17 (17(17q21-q22)), приводит к сверхэкспрессии рецепторов HER2 на поверхности клеток, приводящей к чрезмерной и неуправляемой передаче сигналов при раке молочной железы и других злокачественных новообразованиях (Park et al., Clinical Breast Cancer 8:392-401 [2008]). Было обнаружено, что множество онкогенов амплифицируются при других злокачественных новообразованиях человека. Примеры амплификации клеточных онкогенов в опухолях человека включают амплификации c-tyc в линии клеток промиелоцитарного лейкоза HL60 и в линиях клеток мелкоклеточной карциномы легких, N-myc в первичных нейробластомах (стадии III и IV), линиях клеток нейробластомы, линиях клеток ретинобластомы и первичных опухолей и линиях и опухолях мелкоклеточной карциномы легких, L-myc в линиях клеток и опухолях мелкоклеточной карциномы легких, c-myb в линиях клеток острого миелоидного лейкоза и карциномы толстой кишки, c-erbb в клетках эпидермоидной карциномы и первичных глиомах, c-K-ras-2 в первичных карциномах легких, толстой кишки, мочевого пузыря и прямой кишки, N-ras в линии клеток карциномы молочной железы (Varmus H., Ann Rev Genetics 18: 553-612(1984) [по данным Watson et al., Molecular Biology of the Gene (4^{th} ed.; Benjamin/Cummings Publishing Co. 1987)].

Дупликации онкогенов являются распространенной причиной многих типов рака, как и в случае амплификации P70-S6 киназы 1 и рака молочной железы. В таких случаях генетическая дупликация возникает в соматической клетке и поражает исключительно геном раковых клеток самих по себе, но не всего организма, а тем более какого-либо последующего потомства.

Другие примеры онкогенов, которые амплифицируются при типах рака человека, включают MYC, ERBB2 (EFGR), CCND1 (циклин D1), FGFR1 и FGFR2 при раке молочной железы, MYC и ERBB2 при раке шейки матки, HRAS, KRAS и MYB при раке толстой и прямой кишок, MYC, CCND1 и MDM2 при

раке пищевода, CCNE, KRAS и MET при гастрическом раке, ERBB1 и CDK4 при глиобластоме, CCND1, ERBB1 и MYC при раке головы и шеи, CCND1 при печеночноклеточном раке, MYCB при нейробластоме, MYC, ERBB2 и AKT2 при раке яичников, MDM2 и CDK4 при саркоме и MYC при мелкоклеточном раке легких. Согласно одному варианту реализации настоящий способ можно применять для определения присутствия или отсутствия амплификации онкогена, связанного с раком. Согласно некоторым вариантам реализации амплифицируется онкоген, связанный с раком молочной железы, раком шейки матки, раком толстой и прямой кишок, раком пищевода, гастрическим раком, глиобластомой, раком головы и шеи, печеночноклеточным раком, нейробластомой, раком яичников, саркомой и мелкоклеточным раком легких.

Согласно одному варианту реализации настоящий способ можно применять для определения присутствия или отсутствия делеции хромосомы. Согласно некоторым вариантам реализации делеция хромосомы представляет собой утрату одной или более целых хромосом. Согласно другим вариантам реализации делеция хромосомы представляет собой утрату одного или более сегментов хромосомы. Согласно третьим вариантам реализации делеция хромосомы представляет собой утрату двух или более сегментов двух или более хромосом. Делеция хромосомы может включать утрату одного или более генов-онкосупрессоров.

Делеции хромосом, в которых участвуют гены-онкосупрессоры, как считают, играют важную роль в развитии и прогрессировании солидных опухолей. Ген-онкосупрессор ретинобластомы (Rb-1), расположенный в хромосоме 13q14, представляет собой наиболее полно охарактеризованный ген-онкосупрессор. Продукт гена Rb-1, ядерный фосфопротеин 105 кДа, по-видимому, играет важную роль в регуляции клеточного цикла (Howe et al., Proc Natl Acad Sci (USA) 87:5883-5887 [1990]). Изменение или утрату экспрессии белка Rb вызывает инактивация обоих аллелей гена в результате точечной мутации или делеции хромосомы. Было обнаружено, что изменения гена Rb-i наблюдаются не только в ретинобластомах, но также в других злокачественных новообразованиях, таких как остеосаркомы, мелкоклеточный рак легких (Rygaard et al., Cancer Res 50: 5312-5317 [1990)]) и рак молочной железы. В исследованиях полиморфизма длины фрагментов рестрикции (Restriction fragment length polymorphism, RFLP) было установлено, что такие типы опухолей часто утратили гетерозиготность по 13q; это свидетельствует, что один из аллелей гена Rb-1 был утрачен в связи с крупной делецией хромосомы (Bowcock et al., Am J Hum Genet, 46: 12 [1990]). Аномалии хромосомы 1, включая дупликации, делеции и несбалансированные транслокации, с участием хромосомы 6 и другой хромосомы-партнера, свидетельствуют, что области хромосомы 1, в частности, 1q21-1q32 и 1p11-13, могут нести онкогены или гены-онкосупрессоры, которые являются значимыми с патогенетической точки зрения как на хронической, так и на прогрессирующей стадиях миелопролиферативных новообразований (Caramazza et al., Eur J Hematol 84:191-200 [2010]). Миелопролиферативные новообразования также связаны с делециями хромосомы 5. Полная утрата или внутренние делеции хромосомы 5 являются наиболее частыми аномалиями кариотипа при миелодиспластических синдромах (МДС). Пациенты с выделенным del(5q)/5q- МДС характеризуются более благоприятным прогнозом, чем таковые с дополнительными дефектами кариотипа, у которых наблюдается тенденция к развитию миелопролиферативных новообразований (МПН) и острого миелоидного лейкоза. Частота несбалансированных делеции хромосомы 5 привела к возникновению мнения, что 5q несет один или более супрессоров опухолевых генов, которые играют фундаментальные роли в контроле роста гематопоэтических стволовых клеток/клеток предшественников (ГСК/ГКП). Цитогенетическое картирование часто делетированных областей (commonly deleted regions, CDR) сосредоточилось на идентифицированных кандидатах 5q31 и 5q32 супрессоров опухолевых генов, включая рибосомальную субъединицу RPS14, фактор транскрипции Egr1/Krox20 и белок ремоделирования цитоскелета, альфа-катенин (Eisenmann et al., Oncogene 28:3429-3441 [2009]). Цитогенетические исследования и исследования по аллелетипированию свежих опухолей и опухолевых линий клеток продемонстрировали, что утрата аллелей из нескольких различных областей на хромосоме 3p, включая 3p25, 3p21-22, 3p21.3, 3p12-13 и 3p14, представляет собой самые ранние и наиболее частые геномные аномалии, вовлеченные в широкий спектр большинства эпителиальных типов рака легких, молочной железы, почек, головы и шеи, яичников, шейки матки, толстой кишки, поджелудочной железы, пищевода, мочевого пузыря и других органов. Несколько генов-онкосупрессоров было картировано на области 3p хромосомы, и считают, что внутренние делеции или гиперметилирование промотора предшествуют утрате 3p или целой хромосомы 3 при развитии карцином (Angeloni D., Briefings Functional Genomics 6:19-39 [2007]).

У новорожденных и детей с синдромом Дауна (СД) часто наблюдается врожденный транзиторный лейкоз, и такие новорожденные и дети характеризуются повышенным риском развития острого миелоидного лейкоза и острого лимф областного лейкоза. Хромосома 21, несущая приблизительно 300 генов, может быть вовлечена в многочисленные структурные аберрации, например транслокации, делеции и амплификации, при лейкозах, лимфомах и солидных опухолях. Более того, были идентифицированы гены, расположенные на хромосоме 21, которые играют важную роль в онкогенезе. Соматические численные, а также структурные аберрации хромосомы 21 связаны с лейкозами, и конкретные гены, включая RUNX1, TMPRSS2, и TFF, которые расположены в 21q, играют роль в онкогенезе (Fonatsch C Gene Chromosomes Cancer 49:497-508 [2010]).

С учетом вышеизложенного, согласно различным вариантам реализации способы, описанные в настоящем документе, можно применять для определения ВЧК сегмента, который, как известно, содержит один или более онкогенов или генов-онкосупрессоров, и/или, как известно, связан с раком или с увеличенным риском развития рака. Согласно определенным вариантам реализации ВЧК можно определить в исследуемом образце, содержащем конститутивную (зародышевой линии) нуклеиновую кислоту, и сегмент можно идентифицировать в данных конститутивных нуклеиновых кислотах. Согласно определенным вариантам реализации ВЧК сегмента идентифицируют (в случае наличия) в образце, содержащем смесь нуклеиновых кислот (например, нуклеиновые кислоты, полученные из нормальных, и нуклеиновые кислоты, полученные из неопластических клеток). Согласно определенным вариантам реализации образец получают от субъекта, который, как предполагают или как известно, страдает от рака, например, карциномы, саркомы, лимфомы, лейкоза, герминогенных опухолей, бластомы и т.п. Согласно одному варианту реализации образец представляет собой образец плазмы, полученный (процессированный) из периферической крови, который может содержать смесь сцДНК, полученной из нормальных и раковых клеток. Согласно другому варианту реализации биологический образец, который применяют для определения присутствия ВЧК, получают из клеток, которые, если рак присутствует, содержат смесь раковых и нераковых клеток из других биологических тканей, включая, без ограничения, биологические жидкости, такие как сыворотка, пот, слезы, мокрота, моча, мокрота, ушная жидкость, лимфа, слюна, спинномозговая жидкость, жидкость после лаважа, суспензия костного мозга, влагалищная жидкость, жидкость после трансцервикального лаважа, жидкость головного мозга, асцит, молоко, секреты дыхательных, кишечных и мочеполовых путей и образцы лейкафереза, или в биопсиях ткани, мазках или соскобах. Согласно другим вариантам реализации биологический образец представляет собой образец стула (фекалий).

ВЧК, которую используют для определения присутствия рака и/или повышенного риска развития рака, может включать амплификацию или делеции.

Согласно различным вариантам реализации ВЧК, идентифицированные как свидетельствующие о присутствии рака или о повышенном риске развития рака, включают одну или более амплификаций, представленных в табл. 4.

Таблица 4. Иллюстративные, но неограничивающие сегменты хромосом, которые характеризуются амплификациями, связанными с типами рака. Перечисленные типы рака представляют собой таковые, идентифицированные в публикации Beroukhim et al., Nature 18: 463: 899-905

Максимальная область	**Длина (Мб)**	**Типы рака, идентифицированные в данном анализе, но не в предшествующих публикациях**
chr1:119996566-120303234	0,228	Молочной железы, ПК легких, меланома
chr1:148661965-149063439	0,35	Молочной железы, дедифференцированная липосаркома, аденокарцинома пищевода,

		печеночноклеточный, ПК легких, меланома, яичников, предстательной железы, ренальный
chr1:1-5160566	4,416	Аденокарцинома пищевода, яичников
chr1:158317017-159953843	1,627	Дедифференцированная липосаркома, аденокарцинома пищевода, предстательной железы, ренальный
chr1:169549478-170484405	0,889	Толстой и прямой кишок, дедифференцированная липосаркома, предстательной железы, ренальный
chr1:201678483-203358272	1,471	Предстательной железы
chr1:241364021-247249719	5,678	НПК легких, меланома, яичников
chr1:39907605-40263248	0,319	Острый лимфобластный лейкоз, молочной железы, НПК легких, ПК легких
chr1:58658784-60221344	1,544	Молочной железы, дедифференцированная липосаркома, ПК легких
chr3:170024984-173604597	3,496	Молочной железы, аденокарцинома пищевода, глиома
chr3:178149984-199501827	21,123	Плоскоклеточный пищевода, НПК легких
chr3:86250885-95164178	8,795	ПК легких, меланома
chr4:54471680-55980061	1,449	НПК легких
chr5:1212750-1378766	0,115	Дедифференцированная липосаркома
chr5:174477192-180857866	6,124	Молочной железы, НПК легких
chr5:45312870-49697231	4,206	ПК легких
chr6:1-23628840	23,516	Аденокарцинома пищевода
chr6:135561194-135665525	0,092	Молочной железы, аденокарцинома пищевода
chr6:43556800-44361368	0,72	Аденокарцинома пищевода, печеночноклеточный, яичников
chr6:63255006-65243766	1,988	Аденокарцинома пищевода, НПК легких
chr7:115981465-116676953	0,69	Аденокарцинома пищевода, НПК легких, меланома, яичников
chr7:54899301-55275419	0,363	Аденокарцинома пищевода, плоскоклеточный пищевода
chr7:89924533-98997268	9,068	Молочной железы, аденокарцинома пищевода, плоскоклеточный пищевода, яичников

chr8:101163387-103693879	2,516	НПК легких, меланома, яичников
chr8:116186189-120600761	4,4	Молочной железы, печеночноклеточный, НПК легких, яичников
chr8:128774432-128849112	0,009	Аденокарцинома пищевода, плоскоклеточный пищевода, печеночноклеточный, ПК легких, медуллобластома, миелопролиферативное нарушение, яичников
chr8:140458177-146274826	5,784	НПК легких, медуллобластома, меланома, яичников
chr8:38252951-38460772	0,167	Толстой и прямой кишок, аденокарцинома пищевода, плоскоклеточный пищевода
chr8:42006632-42404492	0,257	Аденокарцинома пищевода, НПК легких, ПК легких, яичников, предстательной железы
chr8:81242335-81979194	0,717	Молочной железы, меланома
chr9:137859478-140273252	2,29	Толстой и прямой кишок, дедифференцированная липосаркома
chr10:74560456-82020637	7,455	Молочной железы, яичников, предстательной железы
chr11:101433436-102134907	0,683	НПК легких, ПК легких
chr11:32027116-37799354	5,744	Молочной железы, дедифференцированная липосаркома, НПК легких, ПК легких
chr11:69098089-69278404	0,161	Дедифференцированная липосаркома, аденокарцинома пищевода, печеночноклеточный, ПК легких, яичников
chr11:76699529-78005085	1,286	Дедифференцированная липосаркома, аденокарцинома пищевода, ПК легких, яичников
chr12:1-1311104	1,271	НПК легких
chr12:25189655-25352305	0,112	Острый лимфобластный лейкоз, аденокарцинома пищевода, плоскоклеточный пищевода, яичников
chr12:30999223-32594050	1,577	Острый лимфобластный лейкоз, толстой и прямой кишок, аденокарцинома пищевода, плоскоклеточный пищевода, НПК легких, ПК легких
chr12:38788913-42596599	3,779	Молочной железы, толстой и прямой кишок, дедифференцированная

		липосаркома, плоскоклеточный пищевода, НПК легких, ПК легких
chr12:56419524-56488685	0,021	Дедифференцированная липосаркома, меланома, ренальный
chr12:64461446-64607139	0,041	Дедифференцированная липосаркома, ренальный
chr12:66458200-66543552	0,058	Дедифференцированная липосаркома, плоскоклеточный пищевода, ренальный
chr12:67440273-67566002	0,067	Молочной железы, дедифференцированная липосаркома, плоскоклеточный пищевода, меланома, ренальный
chr12:68249634-68327233	0,06	Молочной железы, дедифференцированная липосаркома, плоскоклеточный пищевода, ренальный
chr12:70849987-70966467	0,036	Дедифференцированная липосаркома, ренальный
chr12:72596017-73080626	0,23	Ренальный
chr12:76852527-77064746	0,158	Дедифференцированная липосаркома
chr12:85072329-85674601	0,272	Дедифференцированная липосаркома
chr12:95089777-95350380	0,161	Дедифференцированная липосаркома
chr13:108477140-110084607	1,6	Молочной железы, аденокарцинома пищевода, НПК легких, ПК легких
chr13:1-40829685	22,732	Острый лимфобластный лейкоз, аденокарцинома пищевода
chr13:89500014-93206506	3,597	Молочной железы, аденокарцинома пищевода, медуллобластома
chr14:106074644-106368585	0,203	Плоскоклеточный пищевода
chr14:1-23145193	3,635	Острый лимфобластный лейкоз, плоскоклеточный пищевода, печеночноклеточный, ПК легких
chr14:35708407-36097605	0,383	Молочной железы, аденокарцинома пищевода, плоскоклеточный пищевода, печеночноклеточный, предстательной железы
chr15:96891354-97698742	0,778	Молочной железы, толстой и прямой кишок, аденокарцинома пищевода, НПК легких, медуллобластома, меланома
chr17:18837023-19933105	0,815	Молочной железы, печеночноклеточный

chr17:22479313-22877776	0,382	Молочной железы, НПК легких
chr17:24112056-24310787	0,114	Молочной железы, НПК легких
chr17:35067383-35272328	0,149	Толстой и прямой кишок, аденокарцинома пищевода, плоскоклеточный пищевода
chr17:44673157-45060263	0,351	Меланома
chr17:55144989-55540417	0,31	НПК легких, медуллобластома, меланома, яичников
chr17:62318152-63890591	1,519	Молочной железы, НПК легких, меланома, яичников
chr17:70767943-71305641	0,537	Молочной железы, НПК легких, меланома, яичников
chr18:17749667-22797232	5,029	Толстой и прямой кишок, аденокарцинома пищевода, яичников
chr19:34975531-35098303	0,096	Молочной железы, аденокарцинома пищевода, плоскоклеточный пищевода
chr19:43177306-45393020	2,17	НПК легких, яичников
chr19:59066340-59471027	0,321	Молочной железы, НПК легких, яичников
chr2:15977811-16073001	0,056	ПК легких
chr20:29526118-29834552	0,246	Яичников
chr20:51603033-51989829	0,371	Печеночноклеточный, НПК легких, яичников
chr20:61329497-62435964	0,935	Печеночноклеточный, НПК легких
chr22:19172385-19746441	0,487	Толстой и прямой кишок, меланома, яичников
chrX:152729030-154913754	1,748	Молочной железы, НПК легких, ренальный
chrX:66436234-67090514	0,267	Яичников, предстательной железы

Согласно определенным вариантам реализации в сочетании с амплификациями, описанными выше (в настоящем документе), или отдельно от них ВЧК, идентифицированные как свидетельствующие о присутствии рака или о повышенном риске развития рака, включают одну или более делеций, представленных в табл. 5.

Таблица 5. Иллюстративные, но неограничивающие сегменты хромосом, которые характеризуются делениями, связанными с типами рака. Перечисленные типы рака представляют собой таковые, идентифицированные в публикации Beroukhim et al., Nature 18: 463: 899-905

Максимальная область	**Длина (Мб)**	**Типы рака, идентифицированные в данном анализе, но не в предшествующих публикациях**
chr1:110339388-119426489	1p13.2	Острый лимфобластный лейкоз, аденокарцинома пищевода, НПК легких, ПК легких, меланома, яичников, предстательной железы
chr1:223876038-247249719	1q43	Острый лимфобластный лейкоз, молочной железы, ПК легких, меланома, предстательной железы
chr1:26377344-27532551	1p36.11	Молочной железы, аденокарцинома пищевода, плоскоклеточный пищевода, НПК легких, ПК легких, медуллобластома, миелопролиферативное нарушение, яичников, предстательной железы
chr1:3756302-6867390	1p36.31	Острый лимфобластный лейкоз, молочной железы, плоскоклеточный пищевода, печеночноклеточный, НПК легких, ПК легких, медуллобластома, миелопролиферативное нарушение, яичников, предстательной железы, ренальный
chr1:71284749-74440273	1p31.1	Молочной железы, аденокарцинома пищевода, глиома, печеночноклеточный, НПК легких, ПК легких, меланома, яичников, ренальный
chr2:1-15244284	2p25.3	НПК легких, яичников
chr2:138479322-143365272	2q22.1	Молочной железы, толстой и прямой кишок, аденокарцинома пищевода, плоскоклеточный пищевода, печеночноклеточный, НПК легких, яичников, предстательной железы, ренальный
chr2:204533830-206266883	2q33.2	Аденокарцинома пищевода, печеночноклеточный, НПК легких, медуллобластома, ренальный
chr2:241477619-242951149	2q37.3	Молочной железы, дедифференцированная липосаркома, аденокарцинома пищевода, плоскоклеточный пищевода, печеночноклеточный, НПК легких, ПК легких, медуллобластома, меланома, яичников, ренальный

chr3:116900556-120107320	3q13.31	Дедифференцированная липосаркома, аденокарцинома пищевода, печеночноклеточный, НПК легких, меланома, миелопролиферативное нарушение, предстательной железы
chr3:1-2121282	3p26.3	Толстой и прямой кишок, дедифференцированная липосаркома, аденокарцинома пищевода, НПК легких, меланома, миелопролиферативное нарушение
chr3:175446835-178263192	3q26.31	Острый лимфобластный лейкоз, дедифференцированная липосаркома, аденокарцинома пищевода, НПК легких, меланома, миелопролиферативное нарушение, предстательной железы
chr3:58626894-61524607	3p14.2	Молочной железы, толстой и прямой кишок, дедифференцированная липосаркома, аденокарцинома пищевода, плоскоклеточный пищевода, печеночноклеточный, НПК легких, ПК легких, медуллобластома, меланома, миелопролиферативное нарушение, яичников, предстательной железы, ренальный
chr4:1-435793	4p16.3	Миелопролиферативное нарушение
chr4:186684565-191273063	4q35.2	Молочной железы, аденокарцинома пищевода, плоскоклеточный пищевода, НПК легких, медуллобластома, меланома, предстательной железы, ренальный
chr4:91089383-93486891	4q22.1	Острый лимфобластный лейкоз, аденокарцинома пищевода, печеночноклеточный, НПК легких, ренальный
chr5:177541057-180857866	5q35.3	Молочной железы, НПК легких, миелопролиферативное нарушение, яичников
chr5:57754754-59053198	5q11.2	Молочной железы, толстой и прямой кишок, дедифференцированная липосаркома, аденокарцинома пищевода, плоскоклеточный пищевода, ПК легких, меланома, миелопролиферативное нарушение, яичников, предстательной железы
chr5:85837489-133480433	5q21.1	Толстой и прямой кишок, дедифференцированная липосаркома, НПК легких, ПК легких,

		миелопролиферативное нарушение, яичников
chr6:101000242-121511318	6q22.1	Толстой и прямой кишок, НПК легких, ПК легких
chr6:1543157-2570302	6p25.3	Толстой и прямой кишок, дедифференцированная липосаркома, аденокарцинома пищевода, НПК легких, ПК легких, яичников, предстательной железы
chr6:161612277-163134099	6q26	Толстой и прямой кишок, аденокарцинома пищевода, плоскоклеточный пищевода, НПК легких, ПК легких, яичников, предстательной железы
chr6:76630464-105342994	6q16.1	Толстой и прямой кишок, печеночноклеточный, НПК легких
chr7:141592807-142264966	7q34	Молочной железы, толстой и прямой кишок, аденокарцинома пищевода, плоскоклеточный пищевода, печеночноклеточный, НПК легких, яичников, предстательной железы, ренальный
chr7:144118814-148066271	7q35	Молочной железы, аденокарцинома пищевода, плоскоклеточный пищевода, НПК легких, меланома, миелопролиферативное нарушение, яичников
chr7:156893473-158821424	7q36.3	Молочной железы, аденокарцинома пищевода, плоскоклеточный пищевода, НПК легких, меланома, миелопролиферативное нарушение, яичников, предстательной железы
chr7:3046420-4279470	7p22.2	Меланома, миелопролиферативное нарушение, яичников
chr7:65877239-79629882	7q21.11	Молочной железы, медуллобластома, меланома, миелопролиферативное нарушение, яичников
chr8:1-392555	8p23.3	Острый лимфобластный лейкоз, молочной железы, миелопролиферативное нарушение
chr8:2053441-6259545	8p23.2	Острый лимфобластный лейкоз, дедифференцированная липосаркома, аденокарцинома пищевода, плоскоклеточный пищевода, печеночноклеточный, НПК легких, миелопролиферативное нарушение
chr8:22125332-30139123	8p21.2	Острый лимфобластный лейкоз, дедифференцированная липосаркома, печеночноклеточный,

		миелопролиферативное нарушение, яичников, ренальный
chr8:39008109-41238710	8p11.22	Острый лимфобластный лейкоз, молочной железы, дедифференцированная липосаркома, плоскоклеточный пищевода, печеночноклеточный, НПК легких, миелопролиферативное нарушение, ренальный
chr8:42971602-72924037	8q11.22	Молочной железы, дедифференцированная липосаркома, плоскоклеточный пищевода, печеночноклеточный, НПК легких, миелопролиферативное нарушение, ренальный
chr9:1-708871	9p24.3	Острый лимфобластный лейкоз, молочной железы, НПК легких, миелопролиферативное нарушение, яичников, предстательной железы
chr9:21489625-22474701	9p21.3	Толстой и прямой кишок, аденокарцинома пищевода, плоскоклеточный пищевода, миелопролиферативное нарушение, яичников
chr9:36365710-37139941	9p13.2	Миелопролиферативное нарушение
chr9:7161607-12713130	9p24.1	Острый лимфобластный лейкоз, молочной железы, толстой и прямой кишок, аденокарцинома пищевода, печеночноклеточный, ПК легких, медуллобластома, меланома, миелопролиферативное нарушение, яичников, предстательной железы, ренальный
chr10:1-1042949	10p15.3	Толстой и прямой кишок, НПК легких, ПК легких, яичников, предстательной железы, ренальный
chr10:129812260-135374737	10q26.3	Молочной железы, толстой и прямой кишок, глиома, НПК легких, ПК легких, меланома, яичников, ренальный
chr10:52313829-53768264	10q11.23	Толстой и прямой кишок, НПК легких, ПК легких, яичников, ренальный
chr10:89467202-90419015	10q23.31	Молочной железы, ПК легких, яичников, ренальный
chr11:107086196-116175885	11q23.1	Аденокарцинома пищевода, медуллобластома, ренальный
chr11:1-1391954	11p15.5	Молочной железы, дедифференцированная липосаркома,

		аденокарцинома пищевода, НПК легких, медуллобластома, яичников
chr11:130280899-134452384	11q25	Аденокарцинома пищевода, плоскоклеточный пищевода, печеночноклеточный, НПК легких, медуллобластома, ренальный
chr11:82612034-85091467	11q14.1	Меланома, ренальный
chr12:11410696-12118386	12p13.2	Молочной железы, печеночноклеточный, миелопролиферативное нарушение, предстательной железы
chr12:131913408-132349534	12q24.33	Дедифференцированная липосаркома, НПК легких, миелопролиферативное нарушение
chr12:97551177-99047626	12q23.1	Молочной железы, толстой и прямой кишок, плоскоклеточный пищевода, НПК легких, миелопролиферативное нарушение
chr13:111767404-114142980	13q34	Молочной железы, печеночноклеточный, НПК легких
chr13:1-23902184	13q12.11	Молочной железы, ПК легких, яичников
chr13:46362859-48209064	13q14.2	Печеночноклеточный, ПК легких, миелопролиферативное нарушение, предстательной железы
chr13:92308911-94031607	13q31.3	Молочной железы, печеночноклеточный, НПК легких, ренальный
chr14:1-29140968	14q11.2	Острый лимфобластный лейкоз, аденокарцинома пищевода, миелопролиферативное нарушение
chr14:65275722-67085224	14q23.3	Дедифференцированная липосаркома, миелопролиферативное нарушение
chr14:80741860-106368585	14q32.12	Острый лимфобластный лейкоз, дедифференцированная липосаркома, меланома, миелопролиферативное нарушение
chr15:1-24740084	15q11.2	Острый лимфобластный лейкоз, молочной железы, аденокарцинома пищевода, НПК легких, миелопролиферативное нарушение, яичников
chr15:35140533-43473382	15q15.1	Аденокарцинома пищевода, НПК легких, миелопролиферативное нарушение
chr16:1-359092	16p13.3	Аденокарцинома пищевода, печеночноклеточный, НПК легких, ренальный

chr16:31854743-53525739	16q11.2	Молочной железы, печеночноклеточный, НПК легких, меланома, ренальный
chr16:5062786-7709383	16p13.3	Печеночноклеточный, НПК легких, медуллобластома, меланома, миелопролиферативное нарушение, яичников, ренальный
chr16:76685816-78205652	16q23.1	Молочной железы, толстой и прямой кишок, аденокарцинома пищевода, печеночноклеточный, НПК легких, ПК легких, медуллобластома, ренальный
chr16:80759878-82408573	16q23.3	Толстой и прямой кишок, печеночноклеточный, ренальный
chr16:88436931-88827254	16q24.3	Толстой и прямой кишок, печеночноклеточный, НПК легких, предстательной железы, ренальный
chr17:10675416-12635879	17p12	НПК легких, ПК легких, миелопролиферативное нарушение
chr17:26185485-27216066	17q11.2	Молочной железы, толстой и прямой кишок, дедифференцированная липосаркома, НПК легких, ПК легких, меланома, миелопролиферативное нарушение, яичников
chr17:37319013-37988602	17q21.2	Молочной железы, толстой и прямой кишок, дедифференцированная липосаркома, ПК легких, меланома, миелопролиферативное нарушение, яичников
chr17:7471230-7717938	17p13.1	ПК легких, миелопролиферативное нарушение
chr17:78087533-78774742	17q25.3	Толстой и прямой кишок, миелопролиферативное нарушение
chr18:1-587750	18p11.32	Миелопролиферативное нарушение
chr18:46172638-49935241	18q21.2	Аденокарцинома пищевода, НПК легких
chr18:75796373-76117153	18q23	Толстой и прямой кишок, аденокарцинома пищевода, плоскоклеточный пищевода, яичников, предстательной железы
chr19:1-526082	19p13.3	Печеночноклеточный, НПК легких, ренальный
chr19:21788507-34401877	19p12	Печеночноклеточный, НПК легких, ренальный
chr19:52031294-53331283	19q13.32	Молочной железы, печеночноклеточный, НПК легких, медуллобластома, яичников, ренальный
chr19:63402921-63811651	19q13.43	Молочной железы, толстой и прямой кишок, дедифференцированная

		липосаркома, печеночноклеточный, НПК легких, медуллобластома, яичников, ренальный
chr20:1-325978	20p13	Молочной железы, дедифференцированная липосаркома, НПК легких
chr20:14210829-15988895	20p12.1	Аденокарцинома пищевода, НПК легких, медуллобластома, меланома, миелопролиферативное нарушение, предстательной железы, ренальный
chr21:38584860-42033506	21q22.2	Молочной железы
chr22:20517661-21169423	22q11.22	Острый лимфобластный лейкоз, аденокарцинома пищевода
chr22:45488286-49691432	22q13.33	Молочной железы, печеночноклеточный, НПК легких, ПК легких
chrX:1-3243111	Xp22.33	Аденокарцинома пищевода, НПК легких, ПК легких
chrX:31041721-34564697	Xp21.2	Острый лимфобластный лейкоз, аденокарцинома пищевода, глиома

Анеуплоидии, идентифицированные как характерные для различных типов рака (например, анеуплоидии, идентифицированные в табл. 4 и 5), могут содержать гены, которые, как известно, вовлечены в этиологию рака (например, онкосупрессоры, онкогены и т.д.). Данные анеуплоидии можно также анализировать для идентификации значимых, но ранее не известных генов.

Например, в публикации Beroukhim et al., ссылка выше, оценивали потенциальные вызывающие рак гены при изменениях числа копий с применением алгоритма GRAIL (Gene Relationships Among Implicated $Loci_{20}$, взаимосвязь генов среди вовлеченного локуса $_{20}$), который проводит поиск в отношении функциональных взаимосвязей среди геномных областей. GRAIL подсчитывает каждый ген в совокупности геномных областей в отношении его "связанности" с генами в другой области на основании текстового подобия между опубликованными тезисами всех научных статей, в которых упоминаются гены, на основании того, что некоторые целевые гены будут функционировать в общих путях. Данные способы позволяют проводить идентификацию/характеризацию генов, ранее не связанных с конкретными типами рака, о которых идет речь. Табл. 6 иллюстрирует целевые гены, которые, как известно, находятся в пределах идентифицированного амплифицированного сегмента и прогнозированных генов, а табл. 7 иллюстрирует целевые гены, которые, как известно, находятся в пределах идентифицированного делегированного сегмента и прогнозированных генов.

Таблица 6. Иллюстративные, но неограничивающие сегменты хромосом и гены, которые, как известно или как прогнозируют, присутствуют в областях, которые характеризуются амплификацией при различных типах рака (см., например, публикацию Beroukhim et al., ссылка выше)

Хромосома и полоса	**Максимальная область**	**Кол-во генов**	**Известная мишень**	**Приоритетная мишень GRAIL**
8q24.21	chr8:128774432-128849112	1	*MYC*	*MYC*
11q13.2	chr11:69098089-69278404	3	*CCND1*	*ORAOV1*
17q12	chr17:35067383-35272328	6	*ERBB2*	*ERBB2, C17orf37*
12q14.1	chr12:56419524-56488685	7	*CDK4*	*TSPAN31*
14q13.3	chr14:35708407-36097605	3	*NKX2-1*	*NKX2-1*
12q15	chr12:67440273-67566002	1	*MDM2*	*MDM2*
7p11.2	chr7:54899301-55275419	1	*EGFR*	*EGFR*
1q21.2	chr1:148661965-149063439	9	*MCL1*	*MCL1*
8p12	chr8:38252951-38460772	3	*FGFR1*	*FGFR1*
12p12.1	chr12:25189655-25352305	2	*KRAS*	*KRAS*
19q12	chr19:34975531-35098303	1	*CCNE1*	*CCNE1*
22q11.21	chr22:19172385-19746441	11	*CRKL*	*CRKL*
12q15	chr12:68249634-68327233	2		*LRRC10*
12q14.3	chr12:64461446-64607139	1	*HMGA2*	*HMGA2*
Xq28	chrX:152729030-154913754	53		*SPRY3*
5p15.33	chr5:1212750-1378766	3	*TERT*	*TERT*
3q26.2	chr3:170024984-173604597	22	*PRKCI*	*PRKCI*

15q26.3	chr15:96891354-97698742	4	*IGF1R*	*IGF1R*
20q13.2	chr20:51603033-51989829	1		*ZNF217*
8p11.21	chr8:42006632-42404492	6		*PLAT*
1p34.2	chr1:39907605-40263248	7	*MYCL1*	*MYCL1*
17q21.33	chr17:44673157-45060263	4		*NGFR, PHB*
2p24.3	chr2:15977811-16073001	1	*MYCN*	*MYCN*
7q21.3	chr7:89924533-98997268	62	*CDK6*	*CDK6*
13q34	chr13:108477140-110084607	4		*IRS2*
11q14.1	chr11:76699529-78005085	14		*GAB2*
20q13.33	chr20:61329497-62435964	38		*BIRC7*
17q23.1	chr17:55144989-55540417	5		*RPS6KB1*
1p12	chr1:119996566-120303234	5		*REG4*
8q21.13	chr8:81242335-81979194	3		*ZNF704, ZBTB10*
6p21.1	chr6:43556800-44361368	18		*VEGFA*
5p11	chr5:45312870-49697231	0		
20q11.21	chr20:29526118-29834552	5	*BCL2L1*	*BCL2L1, ID1*
6q23.3	chr6:135561194-135665525	1	*MYB*	*hsa-mir-548a-2*
1q44	chr1:241364021-247249719	71		*AKT3*
5q35.3	chr5:174477192-180857866	92		*FLT4*
7q31.2	chr7:115981465-116676953	3	*MET*	*MET*
18q11.2	chr18:17749667-22797232	21		*CABLES1*
17q25.1	chr17:70767943-71305641	13		*GRB2, ITGB4*
1p32.1	chr1:58658784-60221344	7	*JUN*	*JUN*

17q11.2	chr17:24112056-24310787	5		*DHRS13, FLOT2, ERAL1, PHF12*
17p11.2	chr17:18837023-19933105	12		*MAPK7*
8q24.11	chr8:116186189-120600761	13		*NOV*
12q15	chr12:66458200-66543552	0		
19q13.2	chr19:43177306-45393020	60		*LGALS7, DYRK1B*
11q22.2	chr11:101433436-102134907	8	*BIRC2, YAP1*	*BIRC2*
4q12	chr4:54471680-55980061	7	*PDGFRA, KIT*	*KDR, KIT*
12p11.21	chr12:30999223-32594050	9		*DDX11, FAM60A*
3q28	chr3:178149984-199501827	143	*PIK3CA*	*PIK3CA*
1p36.33	chr1:1-5160566	77		*TP73*
17q24.2	chr17:62318152-63890591	12		*BPTF*
1q23.3	chr1:158317017-159953843	52		*PEA15*
1q24.3	chr1:169549478-170484405	6		*BAT2D1, MYOC*
8q22.3	chr8:101163387-103693879	14		*RRM2B*
13q31.3	chr13:89500014-93206506	3		*GPC5*
12q21.1	chr12:70849987-70966467	0		
12p13.33	chr12:1-1311104	10		*WNK1*
12q21.2	chr12:76852527-77064746	0		
1q32.1	chr1:201678483-203358272	21	*MDM4*	*MDM4*
19q13.42	chr19:59066340-59471027	19		*PRKCG, TSEN34*
12q12	chr12:38788913-42596599	12		*ADAMTS20*
12q23.1	chr12:95089777-95350380	2		*ELK3*
12q21.32	chr12:85072329-85674601	0		
10q22.3	chr10:74560456-82020637	46		*SFTPA1B*
3p11.1	chr3:86250885-95164178	8		*POU1F1*
17q11.1	chr17:22479313-22877776	1		*WSB1*
8q24.3	chr8:140458177-146274826	97		*PTP4A3, MAFA, PARP10*
Xq12	chrX:66436234-67090514	1	*AR*	*AR*
6q12	chr6:63255006-65243766	3		*PTP4A1*
14q11.2	chr14:1-23145193	95		*BCL2L2*
9q34.3	chr9:137859478-140273252	76		*NRARP, MRPL41, TRAF2, LHX3*
6p24.1	chr6:1-23628840	95		*E2F3*
13q12.2	chr13:1-40829685	110		*FOXO1*
12q21.1	chr12:72596017-73080626	0		
14q32.33	chr14:106074644-106368585	0		
11p13	chr11:32027116-37799354	35		*WT1*

Таблица 7. Иллюстративные, но неограничивающие сегменты хромосом и гены, которые, как известно или как прогнозируют, присутствуют в областях, которые характеризуются амплификацией при различных типах рака (см., например, публикацию Beroukhim et al., ссылка выше)

Хромосома и полоса	**Максимальная область**	**Кол-во генов**	**Известная мишень**	**Приоритетная мишень GRAIL**
9p21.3	chr9:21489625-22474701	5	*CDKN2A/B*	*CDKN2A*
3p14.2	chr3:58626894-61524607	2	*FHIT*	*FHIT*
16q23.1	chr16:76685816-78205652	2	*WWOX*	*WWOX*
9p24.1	chr9:7161607-12713130	3	*PTPRD*	*PTPRD*
20p12.1	chr20:14210829-15988895	2	*MACROD2*	*FLRT3*
6q26	chr6:161612277-163134099	1	*PARK2*	*PARK2*
13q14.2	chr13:46362859-48209064	8	*RB1*	*RB1*
2q22.1	chr2:138479322-143365272	3	*LRP1B*	*LRP1B*
4q35.2	chr4:186684565-191273063	15		*FRG2, TUBB4Q*
5q11.2	chr5:57754754-59053198	5	*PDE4D*	*PLK2, PDE4D*
16p13.3	chr16:5062786-7709383	2	*A2BP1*	*A2BP1*
7q34	chr7:141592807-142264966	3	*TRB*	*PRSS1*
2q37.3	chr2:241477619-242951149	19		*TMEM16G, ING5*
19p13.3	chr19:1-526082	10		*GZMM, THEG, PPAP2C, C19orf20*
10q23.31	chr10:89467202-90419015	4	*PTEN*	*PTEN*
8p23.2	chr8:2053441-6259545	1	*CSMD1*	*CSMD1*
1p36.31	chr1:3756302-6867390	23		*DFFB, ZBTB48, AJAP1*
4q22.1	chr4:91089383-93486891	2		*MGC48628*
18q23	chr18:75796373-76117153	4		*PARD6G*
6p25.3	chr6:1543157-2570302	2		*FOXC1*
19q13.43	chr19:63402921-63811651	17		*ZNF324*
Xp21.2	chrX:31041721-34564697	2	*DMD*	*DMD*
11q25	chr11:130280899-134452384	12	*OPCML, HNT*	*HNT*
13q12.11	chr13:1-23902184	29		*LATS2*
22q13.33	chr22:45488286-49691432	38		*TUBGCP6*
15q11.2	chr15:1-24740084	20		*A26B1*
22q11.22	chr22:20517661-21169423	3		*VPREB1*
10q26.3	chr10:129812260-135374737	35		*MGMT, SYCE1*
12p13.2	chr12:11410696-12118386	2	*ETV6*	*ETV6*
8p23.3	chr8:1-392555	2		*ZNF596*
1p36.11	chr1:26377344-27532551	24		*SFN*
11p15.5	chr11:1-1391954	49		*RASSF7*

17q11.2	chr17:26185485-27216066	10	*NF1*	*NF1*
11q23.1	chr11:107086196-116175885	61	*ATM*	*CADM1*
9p24.3	chr9:1-708871	5		*FOXD4*
10q11.23	chr10:52313829-53768264	4	*PRKG1*	*DKK1, PRKG1*
15q15.1	chr15:35140533-43473382	109		*TUBGCP4*
1p13.2	chr1:110339388-119426489	81		*MAGI3*
Xp22.33	chrX:1-3243111	21		*SHOX*
3p26.3	chr3:1-2121282	2		*CHL1*
9p13.2	chr9:36365710-37139941	2	*PAX5*	*MELK*
17p13.1	chr17:7471230-7717938	10	*TP53*	*ATP1B2*
12q24.33	chr12:131913408-132349534	7		*CHFR*
7q36.3	chr7:156893473-158821424	7	*PTPRN2*	*NCAPG2*
6q16.1	chr6:76630464-105342994	76		*FUT9, C6orf165, C6orf162, GJA10*
5q21.1	chr5:85837489-133480433	142	*APC*	*APC*
8p11.22	chr8:39008109-41238710	7		*C8orf4, ZMAT4*
19q13.32	chr19:52031294-53331283	25		*BBC3*
10p15.3	chr10:1-1042949	4		*TUBB8*
1p31.1	chr1:71284749-74440273	4	*NEGR1*	*NEGR1*
13q31.3	chr13:92308911-94031607	2	*GPC6*	*GPC6, DCT*
16q11.2	chr16:31854743-53525739	37		*RBL2*
20p13	chr20:1-325978	10		*SOX12*
5q35.3	chr5:177541057-180857866	43		*SCGB3A1*
1q43	chr1:223876038-247249719	173	*RYR2*	*FH, ZNF678*
16p13.3	chr16:1-359092	16		*HBZ*
17q21.2	chr17:37319013-37988602	22		*CNP*
2p25.3	chr2:1-15244284	51		*MYT1L*

3q13.31	chr3:116900556-120107320	1		*LSAMP*
7q21.11	chr7:65877239-79629882	73	*MAGI2*	*CLDN4*
7q35	chr7:144118814-148066271	3	*CNTNAP2*	*CNTNAP2*
14q32.12	chr14:80741860-106368585	154		*PRIMA1*
16q24.3	chr16:88436931-88827254	9		*C16orf3*
3q26.31	chr3:175446835-178263192	1	*NAALADL2*	*NAALADL2*
17q25.3	chr17:78087533-78774742	8		*ZNF750*
19p12	chr19:21788507-34401877	12		*ZNF492, ZNF99*
12q23.1	chr12:97551177-99047626	3	*ANKS1B*	*ANKS1B*
4p16.3	chr4:1-435793	4		*ZNF141*
18p11.32	chr18:1-587750	4		*COLEC12*
2q33.2	chr2:204533830-206266883	1	*PARD3B*	*PARD3B*
8p21.2	chr8:22125332-30139123	63		*DPYSL2, STMN4*
8q11.22	chr8:42971602-72924037	86	*SNTG1*	*FLJ23356, ST18, RB1CC1*
16q23.3	chr16:80759878-82408573	2	*CDH13*	*CDH13*
11q14.1	chr11:82612034-85091467	6	*DLG2*	*CCDC89, CCDC90B, TMEM126A*
14q23.3	chr14:65275722-67085224	7		*GPHN, MPP5*
7p22.2	chr7:3046420-4279470	1	*SDK1*	*SDK1*
13q34	chr13:111767404-114142980	25		*TUBGCP3*
17p12	chr17:10675416-12635879	5	*MAP2K4*	*MAP2K4, ZNF18*
21q22.2	chr21:38584860-42033506	19	*DSCAM, TMPRSS2/ERG*	*DSCAM*
18q21.2	chr18:46172638-49935241	7	*SMAD4, DCC*	*DCC*
6q22.1	chr6:101000242-121511318	87		*GTF3C6, TUBE1, ROS1*
14q11.2	chr14:1-29140968	140		*ZNF219, NDRG2*

Согласно различным вариантам реализации предусмотрено применение способов, рассмотренных в настоящем документе, для идентификации ВЧК сегментов, содержащих амплифицированные области или гены, указанные в табл. 6, и/или применение способов, рассмотренных в настоящем документе, для идентификации ВЧК сегментов, содержащих делетированные области или гены, идентифицированные в табл. 7.

Согласно одному варианту реализации способы, описанные в настоящем документе, обеспечивают средства для оценки взаимосвязи между амплификацией гена и степенью развития опухоли. Корреляция между амплификацией и/или делецией и стадией или степенью злокачественности рака может являться важной с прогностической точки зрения, поскольку такая информация может способствовать определению степени злокачественности опухоли на генетической основе, которая лучше предскажет будущее течение заболевания, причем более прогрессирующие опухоли будут характеризоваться наихудшим прогнозом. Помимо этого, информация относительно событий ранней амплификации и/или делеции может быть подходящей при соотнесении данных событий как предвестников последующего прогрессирования заболевания.

Амплификация и делеции гена, идентифицированные данным способом, могут быть связаны с другими известными параметрами, такими как степень злокачественности опухоли, гистология, уровень

включения метки Brd/Urd, гормональный статус, поражение лимфоузлов, размер опухоли, продолжительность выживания и другие свойства опухоли, доступные из эпидемиологических и биостатистических исследований. Например, ДНК опухоли, которую исследуют данным способом, может включать атипическую гиперплазию, протоковую карциному in situ, рак I-III стадии и метастатические лимфатические узлы для того, чтобы обеспечить идентификацию взаимосвязей между амплификациями и делециями и стадией. Установленные взаимосвязи могут сделать возможным эффективное терапевтическое вмешательство. Например, систематически амплифицированные области могут содержать сверхэкспрессированный ген, на продукт которого можно воздействовать терапевтическим способом (например, рецепторной тирозинкиназой фактора роста, $p185^{HER2}$).

Согласно различным вариантам реализации способы, описанные в настоящем документе, можно применять для идентификации событий амплификации и/или делеции, связанных с устойчивостью к лекарственным средствам, посредством определения вариации числа копий последовательностей нуклеиновой кислоты из первичных типов рака в отношении таких клеток, которые метастазировали в другие участки. Если амплификация и/или делеция гена представляет собой манифестацию кариотипической нестабильности, обеспечивающую быстрое развитие устойчивости к лекарственным средствам, можно ожидать больше амплификации и/или делеций в первичных опухолях от не восприимчивых к химиотерапии пациентов, чем в опухолях восприимчивых к химиотерапии пациентов. Например, если амплификация конкретных генов отвечает за развитие устойчивости к лекарственным средствам, области, окружающие данные гены, как ожидается, будут систематически амплифицированы в опухолевых клетках от плеврального выпота не восприимчивых к химиотерапии пациентов, но не в первичных опухолях. Открытие взаимосвязей между амплификацией и/или делецией гена и развитием устойчивости к лекарственным средствам может позволить идентифицировать пациентов, которые получат или не получат пользу от адъювантной терапии.

По аналогии с описанными для определения присутствия или отсутствия полных и/или частичных анеуплоидий хромосом плода в материнском образце, способы, аппараты и системы, описанные в настоящем документе, можно применять для определения присутствия или отсутствия полных и/или частичных анеуплоидий хромосом в любом образце от пациента, содержащем нуклеиновые кислоты, например, ДНК или сцДНК (включая образцы пациента, которые не представляют собой материнские образцы). Образец от пациента может представлять собой любой тип биологического образца, как описано в другом месте в настоящем документе. Предпочтительно, образец получают в результате неинвазивных процедур. Например, образец может представлять собой образец крови или фракцию сыворотки и плазмы крови. В качестве альтернативы, образец может представлять собой образец мочи или образец фекалий. Согласно третьим вариантам реализации образец представляет собой образец биопсии ткани. Во всех случаях образец содержит нуклеиновые кислоты, например, сцДНК или геномную ДНК, которую очищают и секвенируют с применением любого из способов секвенирования СНП, описанных ранее.

Согласно настоящему способу можно определить как полные, так и частичные анеуплоидии хромосом, связанные с возникновением и прогрессированием рака.

Согласно различным вариантам реализации при применении способов, описанных в настоящем документе, для определения присутствия и/или повышенного риска развития рака можно осуществить нормирование данных в отношении хромосомы или хромосом, для которых определяют ВЧК. Согласно определенным вариантам реализации нормирование данных можно осуществить в отношении плеча или плеч хромосомы, для которой определяют ВЧК. Согласно определенным вариантам реализации нормирование данных можно осуществить в отношении конкретного сегмента или сегментов, для которых определяют ВЧК.

В дополнение к роли ВЧК при раке ВЧК связаны с увеличивающимся количеством распространенных комплексных заболеваний, включая вирус иммунодефицита человека (ВИЧ), аутоиммунные заболевания и спектр нейропсихиатрических нарушений.

ВЧК при инфекционных и аутоиммунных заболеваниях.

На сегодняшний день во многих исследованиях сообщалось о взаимосвязи между ВЧК в генах, связанных с воспалением и иммунным ответом, и ВИЧ, астмой, болезнью Крона и другими аутоиммунными нарушениями (Fanciulli et al., Clin Genet 77:201-213 [2010]). Например, ВЧК в CCL3L1 была вовлечена в предрасположенность к ВИЧ/СПИДу (CCL3L1, делеция 17q11.2), ревматоидному артриту (CCL3L1, делеция 17q11.2) и болезни Кавасаки (CCL3L1, дупликация 17q11.2); ВЧК в HBD-2, как сообщалось, предрасполагает к болезни Крона толстого кишечника (HDB-2, делеция 8p23.1) и псориазу (HDB-2, делеция 8p23.1); ВЧК в FCGR3B, как показано, предрасполагает к гломерулонефриту при системной красной волчанке (FCGR3B, делеция 1q23, дупликация 1q23), васкулиту, ассоциированному с антителами против цитоплазмы нейтрофилов (anti-neutrophil cytoplasmic antibody, ANCA) (FCGR3B, делеция 1q23), и повышает риск развития ревматоидного артрита. Существует по меньшей мере два воспалительных или аутоиммунных заболевания, которые, как было показано, связаны с ВЧК в различных локусах гена. Например, болезнь Крона связана с низким числом копий в HDB-2, но также с общим делеционным полиморфизмом выше по течению от гена IGRM, который кодирует члена семейства связанных с иммунитетом p47 ГТФаз. В дополнение к взаимосвязи с числом копий FCGR3B также сообщалось о в значительной

степени повышенной предрасположенности к СКВ (системной красной волчанке) среди субъектов с более низким количеством копий компонента комплемента С4.

Во многих независимых исследованиях сообщалось о взаимосвязях между геномными делециями в локусах GSTM1 (GSTM1, делеция 1q23) и GSTT1 (GSTT1, делеция 22q11.2) и повышенным риском развития атопической бронхиальной астмы. Согласно некоторым вариантам реализации способы, описанные в настоящем документе, можно применять для определения присутствия или отсутствия ВЧК, связанной с воспалением и/или аутоиммунными заболеваниями. Например, данные способы можно применять для определения присутствия ВЧК у пациента, который, как подозревают, страдает от ВИЧ, астмы или болезни Крона. Примеры ВЧК, связанной с такими заболеваниями, включают, без ограничения, делеции в 17q11.2, 8p23.1, 1q23 и 22q11.2 и дупликации в 17q11.2 и 1q23. Согласно некоторым вариантам реализации настоящий способ можно применять для определения присутствия ВЧК в генах, включая, без ограничения, CCL3L1, HBD-2, FCGR3B, GSTM, GSTT1, C4 и IRGM.

Заболевания ВЧК нервной системы.

Сообщалось о взаимосвязях между de novo и врожденной ВЧК и несколькими распространенными неврологическими и психиатрическими заболеваниями при аутизме, шизофрении и эпилепсии и некоторых случаях нейродегенеративных заболеваний, таких как болезнь Паркинсона, амиотрофический латеральный склероз (АЛС) и аутосомная доминантная болезнь Альцгеймера (Fanciulli et al., Clin Genet 77:201-213 [2010]). Цитогенетические аномалии наблюдались у пациентов с аутизмом и расстройствами аутического спектра (РАС) с дупликациями в 15q11-q13. Согласно Консорциуму Геномного проекта аутизма (Autism Genome project Consortium) 154 ВЧК, включая несколько рецидивирующих ВЧК, на хромосоме 15q11-q13 или в новом геномном расположении, включая хромосому 2p16, 1q21 ив 17p12, в области, связанной с синдромом Смита-Магениса, перекрываются с РАС. Рецидивирующие микроделеции или микродупликации на хромосоме 16p11.2 подчеркнули наблюдение о том, что ВЧК de novo обнаружены в локусах генов, таких как SHANK3 (делеция 22q13.3), нейрексина 1 (NRXN1, делеция 2p16.3) и нейроглинов (NLGN4, делеция Xp22.33), которые, как известно, регулируют синаптическую дифференциацию и регулируют высвобождение глутаминэргического нейротрансмиттера. Шизофрения также связана с множеством ВЧК de novo. Микроделеции и микродупликации, связанные с шизофренией, включают чрезмерную представленность генов, принадлежащих к нейроонтогенетическому и глутаминэргическому путям; это свидетельствует, что множество ВЧК, поражающих данные гены, могут напрямую способствовать патогенезу шизофрении, например, ERBB4, делеция 2q34, SLC1A3, делеция 5p13.3; RAPEGF4, делеция 2q31.1; CIT, делеция 12.24; и множество генов с ВЧК de novo. ВЧК также были связаны с другими неврологическими нарушениями, включая эпилепсию (CHRNA7, делеция 15q13.3), болезнь Паркинсона (SNCA, дупликация 4q22) и АЛС (SMN1, делеция 5q12.2.-q13.3; и делеция SMN2). Согласно некоторым вариантам реализации способы, описанные в настоящем документе, можно применять для определения присутствия или отсутствия ВЧК, связанной с заболеваниями нервной системы. Например, данные способы можно применять для определения присутствия ВЧК у пациента, который, как подозревают, страдает от аутизма, шизофрении, эпилепсии, нейродегенеративных заболеваний, таких как болезнь Паркинсона, амиотрофический латеральный склероз (АЛС) или аутосомная доминантная болезнь Альцгеймера. Данные способы можно применять для определения ВЧК генов, связанных с заболеваниями нервной системы, включая, без ограничения, любое из расстройств аутического спектра (РАС), шизофрению и эпилепсию, и ВЧК генов, связанных с нейродегенеративными нарушениями, такими как болезнь Паркинсона. Примеры ВЧК, связанных с такими заболеваниями, включают, без ограничения, дупликации в 15q11-q13, 2p16, 1q21, 17p12, 16p11.2 и 4q22 и делеции в 22q13.3, 2p16.3, Xp22.33, 2q34, 5p13.3, 2q31.1, 12.24, 15q13.3 и 5q12.2. Согласно некоторым вариантам реализации данные способы можно применять для определения присутствия ВЧК в генах, включая, без ограничения, SHANK3, NLGN4, NRXN1, ERBB4, SLC1A3, RAPGEF4, CIT, CHRNA7, SNCA, SMN1 HSMN2.

ВЧК и метаболические или сердечно-сосудистые заболевания.

Во многих исследованиях сообщалось о взаимосвязи между метаболическими и сердечно-сосудистыми характеристиками, такими как семейная гиперхолестеринемия (СГ), атеросклероз и заболевание коронарной артерии, и ВЧК (Fanciulli et al., Clin Genet 77:201-213 [2010]). Например, реаранжировки зародышевой линии, главным образом, делеции, наблюдались в гене LDLR (LDLR, делеция/дупликация 19p13.2) у некоторых пациентов с СГ, которые не несли другие мутации LDLR. Другой пример представляет собой ген LPA, кодирующий аполипопротеин(а) (Апо(а)), концентрация которого в плазме связана с риском заболевания коронарной артерии, инфаркта миокарда (ИМ) и инсульта. Концентрации Апо(а), содержащего липопротеин Lp(a), в плазме отличаются в 1000 раз среди индивидуумов, и 90% данной вариабельности генетически определено в локусе LPA, причем концентрация в плазме и размер изоформ Lp(a) пропорциональны в высокой степени варьирующему количеству последовательностей повтора "kringle 4" (диапазон 5-50). Эти данные свидетельствуют, что ВЧК по меньшей мере в двух генах может быть связана с риском развития сердечно-сосудистых заболеваний. Способы, описанные в настоящем документе, можно применять в крупных исследованиях для специфичного поиска взаимосвязи ВЧК с сердечно-сосудистыми нарушениями. Согласно некоторым вариантам реализации настоящий способ можно применять для определения присутствия или отсутствия ВЧК, связанной с метаболиче-

ским или сердечно-сосудистым заболеванием. Например, настоящий способ можно применять для определения присутствия ВЧК у пациента, который, как подозревают, страдает от семейной гиперхолестеринемии. Способы, описанные в настоящем документе, можно применять для определения ВЧК генов, связанных с метаболическим или сердечно-сосудистым заболеванием, например, гиперхолестеринемией. Примеры ВЧК, связанной с такими заболеваниями, включают, без ограничения, делецию/дупликацию 19p13.2 гена LDLR и умножения в гене LPA.

Аппараты и системы для определения ВЧК.

Анализ данных секвенирования и диагноза, поставленного на основании этих данных, как правило, проводят с применением различных выполняемых компьютером алгоритмов и программ. Вследствие этого в определенных вариантах реализации применяют процессы с использованием данных, которые хранят в одной или более компьютерных системах или других системах обработки информации или передают посредством данных систем. Варианты реализации, раскрытые в настоящем документе, также относятся к аппарату для осуществления данных операций. Данный аппарат может быть специально сконструирован для требуемых целей или он может представлять собой компьютер общего назначения (или группу компьютеров), избирательно активируемый или реконфигурируемый компьютерной программой и/или структурой данных, которые хранят на компьютере. Согласно некоторым вариантам реализации группа процессоров осуществляет некоторые или все из перечисленных аналитических операций совместно (например, посредством сети или компьютеризированного вычисления в облаке) и/или параллельно. Процессор или группа процессоров для осуществления способов, описанных в настоящем документе, могут относиться к различным типам, включая микроконтроллеры и микропроцессоры, такие как программируемые устройства (например, CPLD, Complex Programmable Logic Devices, сложные устройства с программируемой логикой, и FPGA, Field Programmable Gate Array, программируемая логическая интегральная схема) и непрограммируемые устройства, такие как логическая матрица ASIC (Application Specific Integrated Circuit, специализированная заказная интегральная схема), или микропроцессоры общего назначения.

Помимо этого, определенные варианты реализации относятся к материальному и/или энергонезависимому машиночитаемому носителю информации или продуктам компьютерной программы, которые включают инструкции и/или данные программы (включая структуры данных) для осуществления различных компьютеризированных операций. Примеры машиночитаемых носителей информации включают, без ограничения, полупроводниковые запоминающие устройства, магнитные носители информации, такие как дисковые накопители, магнитную ленту, оптические носители информации, такие как CD, магнитооптические носители информации и электронные устройства, которые специальным образом конфигурированы для хранения и выполнения программных инструкций, таких как постоянно запоминающие устройства (Read-Only Memory Devices, ROM) и запоминающее устройство с произвольным порядком выборки (Random Access Memory, RAM). Конечный пользователь может контролировать машиночитаемый носитель информации напрямую либо конечный пользователь может контролировать носитель информации опосредованно. Примеры носителей информации с прямым контролем включают носитель информации, расположенный на пользовательском оборудовании, и/или носитель информации, который не является общим с другими структурами. Примеры носителя информации с опосредованным контролем включают носитель информации, который является опосредованно доступным для пользователя через внешнюю сеть и/или посредством обеспечивающих сервис общих ресурсов, таких как "облако". Примеры программных инструкций включают как машинный код, такой как образованный с помощью компилятора, так и файлы, содержащие код более высокого уровня, который может быть выполнен компьютером с применением интерпретатора.

Согласно различным вариантам реализации данные или информация, применяемые в раскрытых способах и аппаратах, предложены в электронном формате. Такие данные или информация могут включать риды и метки, полученные из образца нуклеиновой кислоты, подсчитанные значения или плотности таких меток, которые выравниваются с конкретными областями референсной последовательности (например, которые выравниваются с хромосомой или сегментом хромосомы), референсные последовательности (включая референсные последовательности, обеспечивающие исключительно или преимущественно полиморфизмы), дозы хромосомы и сегмента, решения, такие как решения об анеуплоидии, нормированные значения хромосомы и сегментов, пары хромосом или сегментов и соответствующих нормирующих хромосом или сегментов, консультационные рекомендации, диагнозы и т.п. В настоящем документе данные или другая информация, предоставленная в электронном формате, доступна для хранения в машине и для передачи между машинами. Обычно данные в электронном формате предложены в цифровой форме и могут храниться в виде битов и/или байтов в различных структурах данных, перечнях, базах данных и т.д. Данные можно реализовать электронным, оптическим способом и т.д.

В одном варианте реализации обеспечен продукт компьютерной программы для получения выходного сигнала, свидетельствующего о присутствии или отсутствии анеуплоидии, например, анеуплоидии плода или рака, в исследуемом образце. Компьютерный продукт может содержать инструкции для осуществления любого одного или более вышеописанных способов для определения хромосомной аномалии. Как объяснено, компьютерный продукт может содержать энергонезависимый и/или материальный

машиночитаемый носитель, содержащий выполняемую или компилируемую компьютером логическую схему (например, инструкции), записанную на нем для включения процессора для определения дозы хромосом и, в некоторых случаях, присутствия или отсутствия анеуплоидии плода. В одном примере компьютерный продукт содержит машиночитаемый носитель, содержащий выполняемую или компилируемую компьютером логическую схему (например, инструкции), записанную на нем, для включения процессора для диагностики анеуплоидии плода, включающей: процедуру получения для получения данных секвенирования по меньшей мере из части молекул нуклеиновой кислоты из материнского биологического образца, причем указанные данные секвенирования содержат вычисленную дозу хромосомы и/или сегмента; компьютеризированную логическую схему для анализа анеуплоидии плода на основании указанных полученных данных; и процедуры на выходе для получения выходного сигнала, свидетельствующего о присутствии, отсутствии или типе указанной анеуплоидии плода.

Информацию о последовательности из рассматриваемого образца можно картировать на референсные последовательности хромосомы для идентификации количества меток последовательности для каждой из любой одной или более хромосом, представляющих интерес, и для идентификации количества меток последовательности для последовательности нормирующего сегмента для каждой из указанных любой одной или более хромосом, представляющих интерес. Согласно различным вариантам реализации референсные последовательности хранят в базе данных, такой как, например, реляционная или объектная база данных.

Следует понимать, что в большинстве случаев для человека непрактично или даже невозможно без посторонней помощи осуществить компьютерные операции способов, раскрытых в настоящем документе. Например, для картирования одного рида длиной 30 п.о. из образца на любую из хромосом человека без помощи компьютерного аппарата могут потребоваться годы усилий. Разумеется, проблема усугубляется тем, что для принятия надежных решений об анеуплоидии, как правило, требуется картирование тысяч (например, по меньшей мере приблизительно 10000) или даже миллионов ридов на одну или более хромосом.

Способы, раскрытые в настоящем документе, можно осуществить с применением системы для оценки числа копий генетической последовательности, представляющей интерес, в исследуемом образце. Система содержит: (a) секвенатор для получения нуклеиновых кислот из исследуемого образца, который обеспечивает информацию о последовательности нуклеиновой кислоты из образца; (b) процессор; и (c) один или более машиночитаемых носителей информации, на которых хранятся инструкции для выполнения на указанном процессоре с целью осуществления способа для идентификации любой ВЧК, например, анеуплоидий хромосом или частичных анеуплоидий.

Согласно некоторым вариантам реализации способы инструктируются машиночитаемым носителем, на котором хранятся машиночитаемые инструкции для осуществления способа с целью идентификации любой ВЧК, например, анеуплоидий хромосом или частичных анеуплоидий. Таким образом, в одном варианте реализации предложен продукт компьютерной программы, содержащий один или более машиночитаемых носителей, предназначенных для долговременного хранения информации, на которых хранятся выполняемые компьютером инструкции, которые при выполнении одним или более процессорами компьютерной системы заставляют компьютерную систему реализовать способ для оценки числа копий последовательности, представляющей интерес, в исследуемом образце, содержащем плодные и материнские бесклеточные нуклеиновые кислоты. Способ включает: (a) прием ридов последовательности, полученных в результате секвенирования фрагментов бесклеточной нуклеиновой кислоты в исследуемом образце; (b) выравнивание ридов последовательности фрагментов бесклеточной нуклеиновой кислоты с референсным геномом, содержащим последовательность, представляющую интерес, с получением, таким образом, меток исследуемой последовательности, причем референсный геном разделен на множество блоков; (c) определение размеров фрагментов бесклеточной нуклеиновой кислоты, существующих в исследуемом образце; (d) взвешивание меток исследуемой последовательности на основании размеров фрагментов бесклеточной нуклеиновой кислоты, из которых получают метки; (e) вычисление перекрытий для блоков на основании взвешенных меток (d); и (f) идентификацию вариации числа копий в последовательности, представляющей интерес, из вычисленных перекрытий. Согласно некоторым вариантам реализации взвешивание меток исследуемой последовательности включает смещение перекрытий в сторону меток исследуемой последовательности, полученной из фрагментов бесклеточной нуклеиновой кислоты размера или диапазона размера, характерного для одного генома в исследуемом образце. Согласно некоторым вариантам реализации взвешивание меток исследуемой последовательности включает присвоение значения 1 меткам, полученным из фрагментов бесклеточной нуклеиновой кислоты размера или диапазона размера, и присвоение значения 0 другим меткам. Согласно некоторым вариантам реализации способ также включает определение в блоках референсного генома, содержащих последовательность, представляющую интерес, значений параметра размера фрагмента, включая количество фрагментов бесклеточной нуклеиновой кислоты в исследуемом образце, размер фрагментов которых является более коротким или более длинным, чем пороговое значение. В настоящем документе идентификация вариации числа копии в последовательности, представляющей интерес, включает применение значений параметра размера фрагмента, а также перекрытий, вычисленных на этапе (e). Согласно некоторым вари-

антам реализации система спроектирована для оценки числа копий в исследуемом образце с применением различных способов и процессов, которые обсуждаются выше.

Согласно некоторым вариантам реализации инструкции могут также содержать автоматически регистрируемую информацию, относящуюся к способу, такую как дозы хромосом и присутствие или отсутствие анеуплоидии хромосомы плода в медицинской карте пациента для субъекта-человека, от которого был получен материнский исследуемый образец. Медицинская карта пациента может храниться, например, в лаборатории, кабинете врача, больнице, организации медицинского обеспечения, страховой компании или на веб-сайте с персональными медицинскими картами. Также на основании результатов осуществляемого процессором анализа способ может дополнительно включать назначение, начало и/или изменение лечения субъекта-человека, от которого был получен материнский исследуемый образец. Способ может включать осуществление одного или более дополнительных исследований или анализов дополнительных образцов, отобранных от субъекта.

Раскрытые способы можно также осуществлять с применением системы компьютерной обработки информации, которая приспособлена или конфигурирована для осуществления способа с целью идентификации любой ВЧК, например, анеуплоидий хромосом или частичных анеуплоидий. В одном варианте реализации предложена система компьютерной обработки информации, которая приспособлена или конфигурирована для осуществления способа, как описано в настоящем документе. Согласно одному варианту реализации аппарат содержит устройство для секвенирования, приспособленное или конфигурированное для секвенирования по меньшей мере части молекул нуклеиновой кислоты в образце для получения типа информации о последовательности, описанной в другом месте в настоящем документе. Аппарат может также содержать компоненты для процессинга образца. Такие компоненты описаны в другом месте в настоящем документе.

Последовательность или другие данные можно вводить в компьютер или хранить на машиночитаемом носителе напрямую или опосредованно. Согласно одному варианту реализации компьютерная система напрямую соединена с устройством для секвенирования, которое прочитывает и/или анализирует последовательности нуклеиновых кислот из образцов. Последовательности или другую информацию от таких инструментов получают через интерфейс компьютерной системы. В качестве альтернативы, последовательности, процессированные системой, получают из источника хранения последовательностей, такого как база данных или другой репозиторий. После того как запоминающее устройство или устройство памяти большой емкости становится доступным аппарату для обработки информации, запоминающее устройство хранит в буфере или хранит по меньшей мере временно последовательности нуклеиновых кислот. Помимо этого, запоминающее устройство может хранить подсчитанные значения меток для различных хромосом или геномов и т.д. Память может также хранить различные последовательности команд и/или программы для анализа представленной последовательности или картированных данных. Такие программы/последовательности команд могут включать программы для осуществления статистических анализов и т.д.

В одном примере пользователь вносит образец в аппарат для секвенирования. Аппарат для секвенирования, который присоединен к компьютеру, собирает и/или анализирует данные. Программное обеспечение на компьютере позволяет собирать и/или анализировать данные. Данные можно хранить, демонстрировать (с помощью монитора или другого аналогичного устройства) и/или направлять в другое месторасположение. Компьютер может быть присоединен к интернету, который применяют для передачи данных на карманное устройство, используемое удаленным пользователем (например, врачом, исследователем или аналитиком). Следует понимать, что перед передачей данные можно хранить и/или анализировать. Согласно некоторым вариантам реализации первичные данные собирают и отправляют удаленному пользователю или аппарату, который будет анализировать и/или хранить данные. Передачу можно осуществить через Интернет, но можно также осуществить через спутник или другое соединение. В качестве альтернативы, данные можно хранить на машиночитаемом носителе, и носитель можно доставить конечному пользователю (например, по почте). Удаленный пользователь может находиться в том же или в отличном географическом месторасположении, включая, без ограничения, здание, город, штат, страну или континент.

Согласно некоторым вариантам реализации данные способы также включают сбор данных относительно множества полинуклеотидных последовательностей (например, ридов, меток и/или последовательностей референсных хромосом) и отправку данных на компьютер или другую компьютерную систему. Например, компьютер может быть присоединен к лабораторному оборудованию, например, аппарату для сбора образца, аппарату для амплификации нуклеотидов, аппарату для секвенирования нуклеотидов или аппарату для гибридизации. Затем компьютер может собирать применимые данные, полученные лабораторным устройством. Данные можно хранить на компьютере на любом этапе, например в течение сбора в режиме реального времени, перед отправкой, в течение или в сочетании с отправкой или после отправки. Данные можно хранить на машиночитаемом носителе, который можно отделить от компьютера. Собранные или хранимые данные можно передать от компьютера в удаленное месторасположение, например, через локальную сеть или сеть связи для обширной области, такую как Интернет. В удаленном месторасположении можно осуществить различные операции с переданными данными, как описано ниже.

Среди типов форматированных электронным способом данных, которые можно хранить, передавать, анализировать и/или с которыми можно манипулировать в системах, аппаратах и способах, раскрытых в настоящем документе, присутствуют следующие:

риды, полученные в результате секвенирования нуклеиновых кислот в исследуемом образце,

метки, полученные посредством выравнивания ридов с референсным геномом или другой референсной последовательностью или последовательностями,

референсный геном или последовательность,

плотность метки последовательности. - Подсчитанные значения или количества меток для каждой из двух или более областей (как правило, хромосом или сегментов хромосом) референсного генома или других референсных последовательностей,

идентичности нормирующих хромосом или сегментов хромосом для конкретных хромосом или сегментов хромосом, представляющих интерес,

дозы хромосом или сегментов хромосом (или других областей), полученные из хромосом или сегментов, представляющих интерес, и соответствующие нормирующие хромосомы или сегменты,

пороги для принятия решения о дозах хромосом как пораженных, непораженных или "решение отсутствует",

фактические решения о дозах хромосом,

диагнозы (клиническое состояние, связанное с принятым решением),

рекомендации относительно последующих исследований, полученные из решений и/или диагнозов,

планы лечения и/или наблюдения, полученные из решений и/или диагнозов.

Эти различные типы данных можно получить, хранить, передавать, анализировать и/или манипулировать с ними в одном или более месторасположениях с применением различных аппаратов. Варианты обработки данных охватывают широкий спектр. На одном конце спектра всю или большую часть данной информации хранят и применяют в месторасположении, в котором процессируют исследуемый образец, например, в кабинете врача или других клинических условиях. В другом противоположном варианте образец получают в одном месторасположении, процессируют и необязательно секвенируют в отличном месторасположении, риды выравнивают и решения принимают в одном или более отличных месторасположениях, и диагнозы, рекомендации и/или планы готовят в третьем месторасположении (которое может представлять собой месторасположение, в котором был получен образец).

Согласно различным вариантам реализации риды получают с помощью аппарата для секвенирования, а затем передают в удаленный пункт, в котором их процессируют для принятия решений об анеуплоидии. В данном удаленном месторасположении, в качестве примера, риды выравнивают с референсной последовательностью для получения меток, которые подсчитывают и относят к хромосомам или сегментам, представляющим интерес. Также в удаленном месторасположении подсчитанные значения преобразуют в дозы с применением связанных нормирующих хромосом или сегментов. В еще более удаленном месторасположении дозы используют для принятия решений об анеуплоидии.

Среди операций процессинга, которые можно применять в различных месторасположениях, выделяют следующие:

сбор образца,

процессинг образца до секвенирования,

секвенирование,

анализ данных последовательности и принятие решений об анеуплоидии,

диагностика,

предоставление отчета о диагнозе и/или решении пациенту или медицинскому работнику,

разработка плана последующего лечения, исследования и/или наблюдения,

выполнение плана,

консультирование.

Любую одну или более из данных операций можно автоматизировать, как описано в другом месте в настоящем документе. Как правило, секвенирование и анализ данных последовательности и принятие решений об анеуплоидии будут осуществлять компьютерным способом. Другие операции можно осуществлять вручную или автоматически.

Примеры месторасположений, в которых можно осуществлять сбор образца, включают кабинеты практикующих врачей, клиники, дома пациентов (в которых обеспечен инструмент или набор для сбора образца) и мобильный медицинский автотранспорт. Примеры месторасположений, в которых можно осуществлять процессинг образца перед секвенированием, включают кабинеты практикующих врачей, клиники, дома пациентов (в которых обеспечен аппарат или набор для процессинга образца), мобильный медицинский автотранспорт и помещения поставщиков услуг по анализу анеуплоидии. Примеры месторасположений, в которых можно осуществлять секвенирование, включают кабинеты практикующих врачей, клиники, кабинеты практикующих врачей, клиники, дома пациентов (в которых обеспечен аппарат или набор для секвенирования образца), мобильный медицинский автотранспорт и помещения поставщиков услуг по анализу анеуплоидии. Может быть предложено месторасположение, в котором проводят секвенирование, соединенное с выделенной сетью для передачи данных последовательности (как прави-

ло, ридов) в электронном формате. Такое соединение может быть проводным или беспроводным и может быть конфигурировано для отправки данных в пункт, в котором данные можно обрабатывать и/или агрегировать перед передачей в пункт процессинга. Агрегацию данных могут проводить организации здравоохранения, такие как организации медицинского обеспечения (ОМО).

Операции анализа и/или получения можно осуществить в любом из вышеупомянутых месторасположений или, в качестве альтернативы, в еще одном удаленном участке, предназначенном для вычисления и/или проведения анализа данных о последовательности нуклеиновой кислоты. Такие месторасположения включают, например, кластеры, такие как парки серверов общего назначения, помещения предприятий-поставщиков услуг по анализу анеуплоидии и т.п. Согласно некоторым вариантам реализации компьютерный аппарат, применяемый для осуществления анализа, берут во временное пользование или аренду. Компьютерные ресурсы могут являться частью доступной через Интернет совокупности процессоров, такие ресурсы для обработки информации в разговорной речи известны как "облако". В некоторых случаях вычисления осуществляют с помощью параллельной или широкомасштабной параллельной группы процессоров, которые связаны или не связаны друг с другом. Обработку данных можно осуществлять с применением распределенной обработки данных, такой как кластерное компьютеризированное вычисление, сетевое компьютеризированное вычисление и т.п. Согласно таким вариантам реализации кластер или сеть компьютерных ресурсов в совокупности образуют виртуальный суперкомпьютер, состоящий из множества процессоров или компьютеров, которые совместно функционируют для осуществления анализа и/или получения, описанных в настоящем документе. Данные технологии, а также более общепринятые суперкомпьютеры, можно применять для обработки данных последовательности, как описано в настоящем документе. Каждый из них представляет собой форму параллельного компьютеризированного вычисления, основанного на процессорах или компьютерах. В случае сетевого компьютеризированного вычисления данные процессоры (часто целые компьютеры) соединены сетью (частной, общественной или Интернет) с помощью общепринятого сетевого протокола, такого как Ethernet. Напротив, суперкомпьютеры содержат множество процессоров, соединенных локальной высокоскоростной компьютерной шиной.

Согласно определенным вариантам реализации диагноз (например, что плод страдает от синдрома Дауна или что пациент страдает от конкретного типа рака) ставят в том же месторасположении, где проводят операцию анализа. Согласно другим вариантам реализации диагноз ставят в отличном месторасположении. В некоторых примерах предоставление отчета о диагнозе осуществляют в месторасположении, в котором был получен образец, несмотря на то, что это не всегда должно обязательно выполняться на практике. Примеры месторасположений, в которых можно поставить или предоставить отчет о диагнозе и/или в которых осуществляют разработку плана, включают кабинеты практикующих врачей, клиники, интернет-сайты, доступные на компьютере, и карманные устройства, такие как мобильные телефоны, планшетные устройства, смартфоны и т.д., которые имеют проводное или беспроводное соединение с сетью. Примеры месторасположений, в которых осуществляют консультирование, включают кабинеты практикующих врачей, клиники, интернет-сайты, доступные на компьютере, карманные устройства и т.д.

Согласно некоторым вариантам реализации сбор образца, процессинг образца и операции секвенирования осуществляют в первом месторасположении, и операции анализа и получения осуществляют во втором месторасположении. Однако в некоторых случаях сбор образца осуществляют в одном месторасположении (например, кабинете практикующего врача или в клинике), а процессинг и секвенирование образца осуществляют в отличном месторасположении, которое необязательно представляет собой то же месторасположение, в котором происходит анализ и принятие решения.

Согласно различным вариантам реализации последовательность вышеперечисленных операций может запускаться пользователем или субъектом, начинающим сбор образца, процессинг и/или секвенирование образца. После того как одна или более из данных операций были начаты, выполнение другой операции может последовать естественным образом. Например, операция секвенирования может вызвать автоматический сбор ридов и их отправку в аппарат для обработки информации, который затем проводит, часто автоматически и, возможно, без дополнительного вмешательства пользователя, анализ последовательности и операцию принятия решения об анеуплоидии. Согласно некоторым вариантам реализации результат данной операции обработки данных затем автоматически передают, возможно, с переформатированием в виде диагноза, в компонент системы или учреждение, которое обрабатывает и сообщает информацию медицинскому специалисту и/или пациенту. Как объяснено, такая информация также может быть автоматически обработана для получения плана лечения, исследования и/или наблюдения, возможно, вместе с консультационной информацией. Таким образом, начало операции ранней стадии может запустить непрерывную последовательность, в которой медицинскому специалисту, пациенту или другой заинтересованной стороне будет обеспечен диагноз, план, консультирование и/или другая информация, подходящая для воздействия на физическое состояние. Процесс осуществляют, даже если части общей системы физически разделены и, возможно, удалены от месторасположения, например, аппарата для образца и последовательности.

На фиг. 5 представлен один вариант реализации дисперсной системы для получения решения или диагностики исследуемого образца. Месторасположение для сбора образца 01 используют для получения

исследуемого образца от пациента, такого как беременный субъект женского пола или предполагаемый пациент, страдающий от рака. Затем образцы направляют в месторасположение для процессинга и секвенирования 03, где исследуемый образец может быть процессирован и секвенирован, как описано выше. Месторасположение 03 содержит аппарат для процессинга образца, а также аппарат для секвенирования процессированного образца. Результатом секвенирования, как описано в другом месте в настоящем документе, является совокупность ридов, которые, как правило, предложены в электронном формате и которые направляют в сеть, такую как Интернет, отмеченную учетным номером 05 на фиг. 5.

Данные о последовательности направляют в удаленное месторасположение 07, в котором осуществляют анализ и принятие решения. Данное месторасположение может содержать одно или более мощных компьютерных устройств, таких как компьютеры или процессоры. После того как компьютерные ресурсы в месторасположении 07 завершили анализ и получили из полученной информации о последовательности решение, решение передают назад в сеть 05. Согласно некоторым вариантам реализации в месторасположении 07 получают не только решение, но также ставят связанный диагноз. Затем решение и/или диагноз передают по сети назад в месторасположение для сбора образца 01, как проиллюстрировано на фиг. 5. Как объяснено, описанная процедура является лишь одним из множества вариантов того, как различные операции, связанные с получением решения или диагноза, могут быть разделены на различные месторасположения. Один распространенный вариант включает проведение сбора и процессинга и секвенирования образца в одном месторасположении. Другой вариант включает проведение процессинга и секвенирования в одном и том же месторасположении, что и анализ и принятие решения.

Фиг. 6 конкретизирует варианты для осуществления различных операций в различных месторасположениях. В наиболее детализированном случае, представленном на фиг. 6, каждую следующую операцию осуществляют в отдельном месторасположении: сбор образца, процессинг образца, секвенирование, выравнивание ридов, принятие решения, диагностику и предоставление отчета и/или разработку плана.

Согласно одному варианту реализации, который объединяет некоторые из данных операций, процессинг и секвенирование образца осуществляют в одном месторасположении, а выравнивание ридов, принятие решения и диагностику осуществляют в отдельном месторасположении. См. часть фиг. 6, отмеченную условным обозначением A. Согласно другому варианту реализации, который обозначен на фиг. 6 символом B, сбор образца, процессинг и секвенирование образца осуществляют в одном и том же месторасположении. Согласно данному варианту реализации выравнивание ридов и принятие решения осуществляют во втором месторасположении. Наконец, диагностику и предоставление отчета и/или разработку плана осуществляют в третьем месторасположении. Согласно варианту реализации, обозначенному на фиг. 6 символом C, сбор образца осуществляют в первом месторасположении, процессинг образца, секвенирование, выравнивание ридов, принятие решения и диагностику осуществляют совместно во втором месторасположении, и предоставление отчета и/или разработку плана осуществляют в третьем месторасположении. Наконец, согласно варианту реализации, обозначенному на фиг. 6 символом D, сбор образца осуществляют в первом месторасположении, процессинг образца, секвенирование, выравнивание ридов и принятие решения осуществляют во втором месторасположении, и диагностику и предоставление отчета и/или управление планом осуществляют в третьем месторасположении.

В одном варианте реализации предложена система для применения при определении присутствия или отсутствия любой одной или более различных полных анеуплоидий хромосом плода в материнском исследуемом образце, содержащем нуклеиновых кислот плода и матери, причем указанная система содержит секвенатор для получения образца нуклеиновой кислоты и обеспечения информации о последовательности нуклеиновой кислоты плода и матери из образца; процессор; и машиночитаемый носитель для хранения информации, содержащий инструкции для выполнения на указанном процессоре, причем указанные инструкции содержат:

(a) код для получения информации о последовательности для указанных нуклеиновых кислот плода и матери в образце;

(b) код для применения указанной информации о последовательности для идентификации компьютерным способом количества меток последовательности из нуклеиновых кислот плода и матери для каждой любой одной или более хромосом, представляющих интерес, которые выбраны из хромосом 1-22, X и Y, и для идентификации количества меток последовательности для по меньшей мере одной последовательности нормирующей хромосомы или последовательности нормирующего сегмента хромосомы для каждой из указанной любой одной или более хромосом, представляющих интерес;

(c) код для применения указанного количества меток последовательности, идентифицированных для каждой из указанных любой одной или более хромосом, представляющих интерес, и указанного количества меток последовательности, идентифицированных для каждой последовательности нормирующей хромосомы или последовательности нормирующего сегмента хромосомы, для вычисления единичной дозы хромосомы для каждой из любой одной или более хромосом, представляющих интерес; и

(d) код для сравнения каждой из единичных доз хромосом для каждой из любой одной или более хромосом, представляющих интерес, с соответствующим пороговым значением для каждой из одной или более хромосом, представляющих интерес, и посредством этого определение присутствия или отсутствия любой одной или более полных различных анеуплоидий хромосом плода в образце.

Согласно некоторым вариантам реализации код для вычисления единичной дозы хромосомы для каждой из любой одной или более хромосом, представляющих интерес, содержит код для вычисления дозы хромосомы для выбранной одной из хромосом, представляющих интерес, в виде соотношения количества меток последовательности, идентифицированных для выбранной хромосомы, представляющей интерес, и количества меток последовательности, идентифицированных для соответствующей последовательности по меньшей мере одной нормирующей хромосомы или последовательности нормирующего сегмента хромосомы для выбранной хромосомы, представляющей интерес.

Согласно некоторым вариантам реализации система также содержит код для повторяющегося вычисления дозы хромосомы для каждого из любых оставшихся сегментов хромосомы любого одного или более сегментов любой одной или более хромосом, представляющих интерес.

Согласно некоторым вариантам реализации одна или более хромосом, представляющих интерес, выбранных из хромосом 1-22, X и Y, содержит по меньшей мере двадцать хромосом, выбранных из хромосом 1-22, X и Y, и причем инструкции содержат инструкции для определения присутствия или отсутствия по меньшей мере двадцати различных полных анеуплоидии хромосом плода.

Согласно некоторым вариантам реализации по меньшей мере одна последовательность нормирующей хромосомы представляет собой группу хромосом, выбранных из хромосом 1-22, X и Y. Согласно другим вариантам реализации по меньшей мере одна последовательность нормирующей хромосомы представляет собой одну хромосому, которая выбрана из хромосом 1-22, X и Y.

Согласно другому варианту реализации предложена система для применения при определении присутствия или отсутствия любой одной или более различных частичных анеуплоидий хромосом плода в материнском исследуемом образце, содержащем нуклеиновых кислот плода и матери, причем указанная система содержит: секвенатор для получения образца нуклеиновой кислоты и обеспечения информации о последовательности нуклеиновой кислоты плода и матери из образца; процессор; и машиночитаемый носитель для хранения, содержащий инструкции для выполнения на указанном процессоре, причем указанные инструкции содержат:

(a) код для получения информации о последовательности для указанных нуклеиновых кислот плода и матери в указанном образце;

(b) код для применения указанной информации о последовательности для идентификации компьютерным способом количества меток последовательности из нуклеиновых кислот плода и матери для каждого из любого одного или более сегментов любой одной или более хромосом, представляющих интерес, выбранных из хромосом 1-22, X и Y, и для идентификации количества меток последовательности для по меньшей мере одной последовательности нормирующего сегмента для каждого из указанных любого одного или более сегментов любой одной или более хромосом, представляющих интерес;

(c) код с применением указанного количества меток последовательности, идентифицированных для каждого из указанных любого одного или более сегментов любой одной или более хромосом, представляющих интерес, и указанного количества меток последовательности, идентифицированных для указанной последовательности нормирующего сегмента, для вычисления единичной дозы сегмента хромосомы для каждого из указанных любого одного или более сегментов любой одной или более хромосом, представляющих интерес; и

(d) код для сравнения каждой из указанных единичных доз сегмента хромосомы для каждого из указанного любого одного или более сегментов любой одной или более хромосом, представляющих интерес, с соответствующим пороговым значением для каждого из указанного любого одного или более сегментов хромосомы любой одной или более хромосом, представляющих интерес, и посредством этого определения присутствия или отсутствия одной или более различных частичных анеуплоидий хромосом плода в указанном образце.

Согласно некоторым вариантам реализации код для вычисления единичной дозы сегмента хромосомы содержит код для вычисления дозы сегмента хромосомы для выбранного одного из сегментов хромосомы как соотношение количества меток последовательности, идентифицированных для выбранного сегмента хромосомы, и количества меток последовательности, идентифицированных для последовательности соответствующего нормирующего сегмента для выбранного сегмента хромосомы.

Согласно некоторым вариантам реализации система также содержит код для повторения вычисления дозы сегмента хромосомы для каждого из любых оставшихся сегментов хромосомы любого одного или более сегментов любой одной или более хромосом, представляющих интерес.

Согласно некоторым вариантам реализации система также содержит (i) код для повторения этапов (a)-(d) для исследуемых образцов от различных материнских субъектов, и (ii) код для определения присутствия или отсутствия любой одной или более различных частичных анеуплоидий хромосом плода в каждом из указанных образцов.

Согласно другим вариантам реализации любой из систем, предложенных в настоящем документе, код также содержит код для автоматической регистрации присутствия или отсутствия анеуплоидий хромосомы плода, которую определяют на этапе (d), в медицинской карте пациента для субъекта-человека, от которого был получен материнский исследуемый образец, причем указанную регистрацию осуществляют с применением процессора.

Согласно некоторым вариантам реализации любой из систем, предложенных в настоящем документе, секвенатор спроектирован для осуществления секвенирования нового поколения (СНП). Согласно некоторым вариантам реализации секвенатор спроектирован для осуществления широкомасштабного параллельного секвенирования с применением секвенирования посредством синтеза с обратимыми красителями-терминаторами. Согласно другим вариантам реализации секвенатор спроектирован для осуществления секвенирования посредством лигирования. Согласно третьим вариантам реализации секвенатор спроектирован для осуществления одномолекулярного секвенирования.

Примеры

Пример 1.

Получение и секвенирование первичных и обогащенных библиотек секвенирования.

a. Получение библиотек секвенирования - сокращенный протокол (ABB).

Все библиотеки секвенирования, т.е. первичные и обогащенные библиотеки, получали из приблизительно 2 нг очищенной сцДНК, которую экстрагировали из материнской плазмы. Получение библиотеки осуществляли с применением набора реактивов NEBNext™ DNA Sample Prep DNA Reagent Set 1 (номер по каталогу E6000L; New England Biolabs, Ипсвич, Массачусетс) для Illumina® следующим образом. Поскольку бесклеточная ДНК плазмы в природе является фрагментированной, какую-либо дополнительную фрагментацию образцов ДНК плазмы посредством пульверизации или обработки ультразвуком не проводили. Выступающие концы приблизительно 2 нг очищенных фрагментов сцДНК в 40 мкл преобразовывали в фосфорилированные тупые концы согласно модулю NEBNext® End Repair Module посредством инкубации сцДНК в микроцентрифужной пробирке объемом 1,5 мл с 5 мкл 10× буфера для фосфорилирования, 2 мкл смеси дезоксинуклеотидов в растворе (10 мМ каждого дНТФ), 1 мкл ДНК-полимеразы I в разведении 1:5, 1 мкл ДНК-полимеразы T4 и 1 мкл полинуклеотидкиназы T4 из набора реактивов NEBNext™ DNA Sample Prep DNA Reagent Set 1 в течение 15 мин при температуре 20°C. Затем ферменты инактивировали нагреванием посредством инкубации реакционной смеси при температуре 75°C в течение 5 мин. Смесь охлаждали до температуры 4°C, и проводили присоединение dA-"хвоста" к тупым концам ДНК с применением 10 мкл мастер-микс для присоединения dA-"хвоста", содержащей фрагмент Кленова (от 3'- к 5'-экзо минус) (NEBNext™ DNA Sample Prep DNA Reagent Set 1), и инкубации в течение 15 мин при температуре 37°C. Затем фрагмент Кленова инактивировали нагреванием посредством инкубации реакционной смеси при температуре 75°C в течение 5 мин. После инактивации фрагмента Кленова 1 мкл смеси Genomic Adaptor Oligo Mix (номер по каталогу 1000521; Illumina Inc., Хейвард, Калифорния) Illumina в разведении 1:5 применяли для лигирования адаптеров Illumina (неиндексные Y-адаптеры) с ДНК с присоединенным dA-"хвостом" с применением 4 мкл ДНК-лигазы T4 из набора реактивов NEBNext™ DNA Sample Prep DNA Reagent Set 1 посредством инкубации реакционной смеси в течение 15 мин при температуре 25°C. Смесь охлаждали до температуры 4°C, и лигированную с адаптерами сцДНК очищали от нелигированных адаптеров, димеров адаптеров и других реактивов с применением магнитных бусин из системы для очистки продуктов ПНР Agencourt AMPure XP (номер по каталогу A63881; Beckman Coulter Genomics, Денверс, Массачусетс). Для селективного обогащения лигированной с адаптерами сцДНК (25 мкл) проводили восемнадцать циклов ПНР с применением мастер-микс Phusion® High-Fidelity Master Mix (25 мкл; Finnzymes, Уоберн, Массачусетс) и праймеров для ПНР Illumina (0,5 мкМ каждого), комплементарных адаптерам (номер по каталогу 1000537 и 1000537). Проводили ПНР лигированной с адаптерами ДНК (98°C в течение 30 с; 18 циклов 98°C в течение 10 с, 65°C в течение 30 с и 72°C в течение 30; конечная элонгация при температуре 72°C в течение 5 мин, хранение при температуре 4°C) с применением праймеров Genomic PCR Primers (№№ по каталогу 100537 и 1000538) Illumina и мастер-микс Phusion HF PCR Master Mix из набора реактивов NEBNext™ DNA Sample Prep DNA Reagent Set 1 в соответствии с инструкциями производителя. Амплифицированный продукт очищали с применением системы для очистки продуктов ПЦР Agencourt AMPure XP (Agencourt Bioscience Corporation, Беверли, Массачусетс) в соответствии с инструкциями производителя, доступными по адресу: www.beckm angenomics.com/products/AMPureXPProtocol_000387v001.pdf. Очищенный амплифицированный продукт элюировали в 40 мкл буфера EB Qiagen, и концентрацию и размер распределения амплифицированных библиотек анализировали с применением набора Agilent DNA 1000 Kit для биоанализатора 2100 Bioanalyzer (Agilent technologies Inc., Санта-Клара, Калифорния).

b. Получение библиотек секвенирования - полный протокол.

Полный протокол, описанный в настоящем документе, представляет собой по существу стандартный протокол, предложенный компанией Illumina, и отличается от протокола Illumina исключительно очисткой амплифицированной библиотеки. Протокол Illumina информирует, что очистку амплифицированной библиотеки проводят с применением гель-электрофореза, в то время как в протоколе, описанном в настоящем документе, для данного этапа очистки применяют магнитные бусины. Для получения первичной библиотеки секвенирования с применением набора реактивов NEBNext™ DNA Sample Prep DNA Reagent Set 1 (номер по каталогу E6000L; New England Biolabs, Ипсвич, Массачусетс) для Illumina® по существу в соответствии с инструкциями производителя применяли приблизительно 2 нг очищенной сцДНК, экстрагированной из материнской плазмы. Все этапы, за исключением итоговой очистки лиги-

рованных с адаптерами продуктов, которую осуществляли с применением магнитных бусин и реактивов Agencourt вместо колонки для очистки, проводили согласно протоколу, сопутствующему реактивам NEBNext™ Reagents for Sample Preparation для библиотеки геномной ДНК, которую секвенируют с применением GAII Illumina®. Протокол NEBNext™ по существу соответствует таковому, предложенному Illumina, который доступен по адресу: grcf.jhml.edu/hts/protocols/11257047_ChIP_Sample_Prep.pdf.

Выступающие концы приблизительно 2 нг очищенных фрагментов сцДНК в 40 мкл преобразовывали в фосфорилированные тупые концы согласно модулю NEBNext® End Repair Module посредством инкубации 40 мкл сцДНК с 5 мкл 10× буфера для фосфорилирования, 2 мкл смеси дезоксинуклеотидов в растворе (10 мМ каждого дНТФ), 1 мкл ДНК-полимеразы I в разведении 1:5, 1 мкл ДНК-полимеразы T4 и 1 мкл полинуклеотидкиназы T4 из набора реактивов NEBNext™ DNA Sample Prep DNA Reagent Set 1 в микроцентрифужной пробирке объемом 200 мкл в термоциклере в течение 30 мин при температуре 20°C. Образец охлаждали до температуры 4°C и очищали с применением колонки QIAQuick из набора QIAQuick PCR Purification Kit (QIAGEN Inc., Валенсия, Калифорния) следующим образом. 50 мкл реакционной смеси переносили в микроцентрифужную пробирку объемом 1,5 мл и добавляли 250 мкл буфера PB Qiagen. Полученные в результате 300 мкл переносили на колонку QIAquick, которую центрифугировали в микроцентрифуге при 13000 об/мин в течение 1 мин. Колонку промывали 750 мкл буфера PE Qiagen и повторно центрифугировали. Остаточный этанол удаляли посредством дополнительного центрифугирования в течение 5 мин при 13000 об/мин ДНК элюировали в 39 мкл буфера EB Qiagen посредством центрифугирования. Присоединение dA-"хвоста" к 34 мкл ДНК с тупыми концами проводили с применением 16 мкл мастер-микс для присоединения dA-"хвоста", содержащей фрагмент Кленова (от 3'- к 5'-экзо минус) (NEBNext™ DNA Sample Prep DNA Reagent Set 1), и инкубации в течение 30 мин при температуре 37°C согласно инструкции производителя модуля NEBNext® dA-Tailing Module. Образец охлаждали до температуры 4°C и очищали с применением колонки из набора MinElute PCR Purification Kit (QIAGEN Inc., Валенсия, Калифорния) следующим образом. 50 мкл реакционной смеси переносили в микроцентрифужную пробирку объемом 1,5 мл и добавляли 250 мкл буфера PB Qiagen. 300 мкл переносили на колонку MinElute, которую центрифугировали в микроцентрифуге при 13000 об/мин в течение 1 мин. Колонку промывали 750 мкл буфера PE Qiagen и повторно центрифугировали. Остаточный этанол удаляли посредством дополнительного центрифугирования в течение 5 мин при 13000 об/мин ДНК элюировали в 15 мкл буфера EB Qiagen посредством центрифугирования. Десять микролитров элюата ДНК инкубировали с 1 мкл смеси Genomic Adapter Oligo Mix (номер по каталогу 1000521) Illumina в разведении 1:5, 15 мкл 2× буфера Quick Ligation Reaction Buffer и 4 мкл ДНК-лигазы Quick T4 DNA Ligase в течение 15 мин при температуре 25°C согласно инструкции модуля NEBNext® Quick Ligation Module. Образец охлаждали до температуры 4°C и очищали с применением колонки MinElute следующим образом. К 30 мкл реакционной смеси добавляли сто пятьдесят микролитров буфера PE Qiagen, и весь объем переносили на колонку MinElute переносили на колонку MinElute, которую центрифугировали в микроцентрифуге при 13000 об/мин в течение 1 мин. Колонку промывали 750 мкл буфера PE Qiagen и повторно центрифугировали. Остаточный этанол удаляли посредством дополнительного центрифугирования в течение 5 мин при 13000 об/мин ДНК элюировали в 28 мкл буфера EB Qiagen посредством центрифугирования. Двадцать три микролитра элюата лигированной с адаптерами ДНК подвергали 18 циклам ПНР (98°C в течение 30 с; 18 циклов 98°C в течение 10 с, 65°C в течение 30 с и 72°C в течение 30; конечная элонгация при температуре 72°C в течение 5 мин, хранение при температуре 4°C) с применением праймеров Genomic PCR Primers Illumina (№№ по каталогу 100537 и 1000538) и мастер-микс Phusion HF PCR Master Mix из набора реактивов NEBNext™ DNA Sample Prep DNA Reagent Set 1 в соответствии с инструкциями производителя. Амплифицированный продукт очищали с применением системы для очистки продуктов ПНР Agencourt AMPure XP (Agencourt Bioscience Corporation, Беверли, Массачусетс) в соответствии с инструкциями производителя, доступными по адресу www.beckmangenomics.com/products/AMPureXPProtocol_000387v001.pdf. Система для очистки продуктов ПНР Agencourt AMPure XP удаляет невстроенные дНТФ, праймеры, димеры праймеров, соли и другие загрязняющие вещества и восстанавливает ампликоны размером более 100 п.о. Очищенный амплифицированный продукт элюировали с бусин Agencourt в 40 мкл буфера EB Qiagen, и распределение размера библиотек анализировали с применением набора Agilent DNA 1000 Kit для биоанализатора 2100 Bioanalyzer (Agilent technologies Inc., Санта-Клара, Калифорния).

c. Анализ библиотек секвенирования, полученных согласно сокращенному (a) и полному (b) протоколам.

Электрофореграммы, полученные с помощью прибора Bioanalyzer, представлены на фиг. 7A и 7B. На фиг. 7A представлена электрофореграмма библиотеки ДНК, полученной из сцДНК, очищенной из образца плазмы M24228 с применением полного протокола, описанного в пункте (a), и на фиг. 7B представлена электрофореграмма библиотеки ДНК, полученной из сцДНК, очищенной из образца плазмы M24228, с применением полного протокола, описанного в пункте (b). На обеих фигурах пики 1 и 4 представляют нижний маркер 15 п.о. и верхний маркер 1500, соответственно; числа над пиками указывают времена миграции фрагментов библиотеки; и горизонтальные линии указывают заданный порог для ин-

тегрирования. На электрофореграмме на фиг. 7B представлен второстепенный пик фрагментов из 187 п.о. и основной пик фрагментов из 263 п.о., тогда как электрофореграмма на фиг. 7A демонстрирует исключительно один пик при 265 п.о. Интегрирование площадей пиков позволило рассчитать концентрацию 0,40 нг/мкл для ДНК пика 187 п.о. на фиг. 7B, концентрацию 7,34 нг/мкл для ДНК пика 263 п.о. на фиг. 7B и концентрацию 14,72 нг/мкл для ДНК пика 265 п.о. на фиг. 7A. Адаптеры Illumina, которые лигировали с сцДНК, как известно, составляют 92 п.о. в длину, что после вычитания из 265 п.о. позволило определить, что размер пика сцДНК составляет 173 п.о. Возможно, что второстепенный пик при 187 п.о. представляет собой фрагменты двух праймеров, которые были лигированы конец к концу. Линейные фрагменты двух праймеров удаляли из итоговой библиотеки продукта, когда применяли сокращенный протокол. В сокращенном протоколе также удаляли другие меньшие фрагменты, длина которых составляла менее 187 п.о. В данном примере концентрация очищенной лигированной с адаптерами сцДНК в два раза превышает таковую лигированной с адаптерами сцДНК, полученной с применением полного протокола. Было отмечено, что концентрация лигированных с адаптерами фрагментов сцДНК всегда превышала таковую, полученную с применением полного протокола (данные не показаны).

Таким образом, преимущество получения библиотеки секвенирования с применением сокращенного протокола заключается в том, что полученная библиотека систематически содержит исключительно один основной пик в диапазоне 262-267 п.о., в то время как качество библиотеки, полученной с применением полного протокола, варьирует, что отражено количествами и подвижностью пиков, отличных от таковых, представляющих сцДНК. Продукты, отличные от сцДНК, займут пространство в проточной ячейке и снизят качество кластерной амплификации и последующей визуализации реакций секвенирования, что лежит в основе общей оценки статуса анеуплоидии. Сокращенный протокол, как было показано, не влияет на секвенирование библиотеки.

Другое преимущество получения библиотеки секвенирования с применением сокращенного протокола заключается в том, что три ферментативных этапа - добавления тупых концов, присоединения d-A-"хвоста" и лигирования с адаптерами, - завершаются в течение менее часа, что способствует валидации и внедрению быстрого сервиса по диагностике анеуплоидии.

Другое преимущество заключается в том, что три ферментативных этапа - добавления тупых концов, присоединения d-A-"хвоста" и лигирования с адаптерами, - осуществляют в одной реакционной пробирке, таким образом, избегая многочисленных переносов образца, которые потенциально могут привести к потере материала и, что более важно, к возможному перепутыванию образцов и загрязнению образца.

Пример 2.

Неинвазивное пренатальное тестирование с применением размера фрагмента.

Введение.

С момента коммерческого внедрения в конце 2011 - начале 2012 года неинвазивное пренатальное тестирование (НИПТ) бесклеточной ДНК (сцДНК) в материнской плазме быстро стало способом, предпочтительным для скрининга беременных женщин, подверженных высокому риску анеуплоидий плода. Способы преимущественно основаны на выделении и секвенировании сцДНК в плазме беременных женщин и подсчете количества фрагментов сцДНК, которые выравниваются с конкретной областью референсного генома человека (источники: Fan et al., Lo et al.). Данные способы секвенирования ДНК и молекулярного подсчета обеспечивают высокоточное определение относительного числа копий для каждой из хромосом в пределах генома. Высокие чувствительности и специфичности обнаружения трисомии 21, 18 и 13 были воспроизводимым образом достигнуты во многих клинических исследованиях (ссылки, цитата метаанализ Gil/Nicolaides).

Совсем недавно дополнительные клинические исследования продемонстрировали, что данный подход можно распространить на общую популяцию беременных. Между популяциями высокого и среднего риска поддающиеся обнаружению различия во фракциях плода отсутствуют (ссылки). Результаты клинических исследований демонстрируют, что НИПТ с применением молекулярного подсчета посредством секвенирования сцДНК осуществляется одинаково в обеих популяциях. Было продемонстрировано статистически значимое улучшение положительной прогностической значимости (positive predictive value, PPV) по сравнению со стандартным скринингом сыворотки (ссылки). Более низкая доля ложноположительных результатов анализа по сравнению с биохимическим анализом сыворотки и измерением толщины воротникового пространства в значительной степени снизила потребность в инвазивных диагностических процедурах (см. публикацию Larion et al., ссылки из группы Abuhamad).

Учитывая хорошие рабочие характеристики НИПТ в общей популяции беременных, на сегодняшний день простота и стоимость рабочего процесса стали основным соображением для внедрения секвенирования сцДНК с целью обнаружения анеуплоидии целой хромосомы в общей популяции беременных (ссылка: ISPD Debate 1, Brisbane). В большинстве лабораторных способов НИПТ применяют этап амплификации посредством полимеразной цепной реакции (ПНР) после получения библиотеки и секвенирование одиночных концов, для которого требуется 10-20 миллионов уникальных фрагментов сцДНК для достижения приемлемой чувствительности с целью обнаружения анеуплоидии. Сложность рабочего процесса на основе ПЦР и потребность в более глубоком секвенировании ограничивали потенциал ана-

лиза НИПТ и привели к увеличению затрат.

В настоящем документе продемонстрировано, что высоких аналитических чувствительностей и специфичностей можно достичь при простом получении библиотеки с применением очень низкого количества сцДНК на входе, для которого не требуется ПЦР-амплификация. Способ без применения ПЦР упрощает рабочий процесс, улучшает время оборота и устраняет погрешности, присущие способам на основе ПЦР. Рабочий процесс без амплификации можно сочетать с секвенированием спаренных концов для обеспечения определения длины фрагмента для каждой метки и для суммарной фракции плода в каждом образце. Поскольку фрагменты сцДНК плода являются более короткими, чем материнские фрагменты [ссылка Quake 2010, также следует процитировать статью Lo Science Clin Translation], обнаружение анеуплоидии плода из материнской плазмы можно осуществить значительно более надежным и эффективным способом, для чего потребуется меньшее количество уникальных фрагментов сцДНК. В сочетании улучшенная аналитическая чувствительность и специфичность достигаются с очень быстрым временем оборота при в значительной степени меньшем количестве фрагментов сцДНК. Это потенциально позволяет проводить НИПТ со значительно меньшими затратами для облегчения применения в общей популяции беременных.

Способы.

Образцы периферической крови помещали в пробирки BCT (Streck, Омаха, Небраска, США) и перевозили в лабораторию CLIA Illumina в Рэдвуд-Сити для коммерческого исследования НИПТ. Подписанные формы согласия пациента позволяли деидентифицировать вторые аликвоты плазмы и применять для клинического исследования, за исключением образцов от пациента, отправленных из штата Нью-Йорк. Образцы плазмы для данной работы выбирали так, чтобы они включали как непораженные, так и анеуплоидные плоды с диапазоном концентраций сцДНК и фракций плода.

Упрощение процессинга библиотеки.

сцДНК экстрагировали из 900 мкл материнской плазмы с применением 96-луночного набора для очистки крови NucleoSpin (Macherey-Nagel, Дюрен, Германия) с незначительными изменениями для использования большего количества лизата на входе. Выделенную сцДНК непосредственно помещали в процесс библиотеки секвенирования без какого-либо нормирования сцДНК на входе. Библиотеки секвенирования готовили с помощью набора для библиотеки TruSeq PCR Free DNA (Illumina, Сан-Диего, Калифорния, США) с двойными индексами для штрих-кодирования фрагментов сцДНК с целью идентификации образцов. Следующие изменения в протоколе библиотеки использовали для улучшения совместимости получения библиотеки с низкой концентрацией сцДНК на входе. Объем матрицы на входе увеличивали, тогда как объем мастер-микс для восстановления конца, присоединения A-"хвоста" и лигирования и концентрации адаптера снижали. Дополнительно, после восстановления концов вводили этап уничтожения нагреванием для деактивации ферментов, удаляли этап очистки после восстановления концов с помощью бусин SPRI (поставщик услуг), и элюирование в течение этапа очистки после лигирования бусин SPRI проводили с применением буфера HT1 (Illumina).

Один жидкостный манипулятор MICROLAB® STAR (Hamilton, Рино, Невада, США), конфигурированный с 96-канальной головкой и 8 каналами для пипетирования объемом 1 мл, применяли для обработки 96 образцов плазмы партиями в течение времени. Жидкостный манипулятор процессировал каждый индивидуальный образец плазмы посредством экстракции ДНК, получения библиотеки секвенирования и количественного определения. Библиотеки индивидуальных образцов количественно определяли с помощью AccuClear (Biotium, Хейвард, Калифорния, США) и получали объединенные 48 образцов с нормированными количествами на входе, что приводило к получению конечной концентрации 32 пМ для секвенирования.

Секвенированж спаренных концов.

Секвенирование ДНК проводили на приборе NextSeq 500 Illumina с применением секвенирования спаренных концов 2×36 п.о., плюс дополнительные 16 циклов для секвенирования штрих-кодов образцов. В сумме в 8 независимых партиях секвенирования анализировали 364 образца.

Парные последовательности ДНК демультиплексировали с применением bcl2fastq (Illumina) и картировали на референсный геном человека (hg19) с применением алгоритма выравнивания bowtie2 [ссылка Landmead]. Парные риды должны были соответствовать прямой и обратной цепям для их подсчета. Все подсчитанные картированные пары, превышающие показатель качества картирования 10 (Ruan et al.), с глобально уникальными первыми ридами относили к неперекрывающимся последовательным геномным блокам фиксированной ширины размером 100 т.о. Приблизительно 2% генома продемонстрировало в высокой степени варьирующее перекрытие среди независимого множества образцов НИПТ, и их исключали из последующего анализа.

С применением информации о геномном расположении и размере фрагмента, доступной из картированных расположений каждого из двух концов секвенированных фрагментов сцДНК, получали две переменные для каждого окна 100 т.о.: (a) общие подсчитанные значения коротких фрагментов менее 150 пар оснований в длину, и (b) фракцию фрагментов от 80 до 150 пар оснований в пределах множества всех фрагментов длиной менее 250 пар оснований. Ограничение размера фрагментов менее 150 пар ос-

нований обогащает фрагментами, полученными из плаценты, которые являются заменителем ДНК плода. Фракция коротких фрагментов характеризует относительные количества сцДНК плода в смеси плазме. CfDNA из трисомического плода, как ожидается, будет содержать большую фракцию коротких ридов, картирующихся на трисомическую хромосому, по сравнению с эуплоидным плодом, который является дисомическим по данной хромосоме.

Подсчитанные значения и фракции коротких фрагментов независимо нормировали для устранения систематических погрешностей анализа и образец-специфичных вариаций, присущих геномному содержанию гуанина и цитозина (GC), с применением процесса, представленного на фиг. 2D. Нормированные значения цензурировали посредством удаления блоков, отклонявшихся от медианы целой хромосомы на более чем 3 устойчивых измерения стандартного отклонения. Наконец, для каждой из двух переменных цензурированные нормированные значения, связанные с целевой хромосомой, сравнивали с таковыми нормирующих референсных хромосом для составления t-статистики.

Данные от каждой серии секвенирования спаренных концов проходили четыре этапа анализа: 1) преобразование рида, 2) разделение характеристик на блоки при разрешении 100 т.о., 3) нормирование каждой характеристики (подсчитанные значения и фракция) при разрешении 100 т.о. и 4) объединение характеристик и определение показателей для обнаружения анеуплоидии. На этапе 1 данные образца демультиплексировали из индивидуальных штрих-кодов, выравнивали с геномом и фильтровали для качества последовательности. На этапе 2 проводили общий вычисление коротких фрагментов менее 150 пар оснований в длину и определяли фракции фрагментов от 80 до 150 пар оснований в пределах множества всех фрагментов длиной менее 250 пар оснований для каждого блока. Погрешности анализа и образец-специфичные вариации устраняли на этапе 3. Наконец, определяли обогащение по сравнению с референсом и определяли показатель с применением t-критерия для каждого из подсчитанных значений и фракции и объединяли для получения итогового показателя с целью обнаружения анеуплоидии.

Обнаружение анеуплоидии целой хромосомы у плода.

Авторы настоящего изобретения провели исследование, можно ли объединить подсчитанные значения и фракцию данных для усиления способности обнаружения трисомии 21 у плода. Шестнадцать образцов плазмы от беременных женщин, вынашивающих плоды с кариотипически подтвержденной трисомией 21, и 294 образцов от непораженных беременностей случайным образом распределяли в партии процессинга, что позволило получить девять проточных ячеек для секвенирования. Каждый этап алгоритма исследовали отдельно для определения способности каждого этапа и комбинации этапов обнаруживать анеуплоидию. Итоговый показатель для обнаружения анеуплоидии плода в объединенном случае задавали как квадратный корень из суммы квадратов двух индивидуальных t-статистик, и единичный порог применяли для получения решения "анеуплоидия обнаружена" по сравнению с "анеуплоидия не обнаружена".

Вычисление фракции плода.

Для каждого образца фракцию плода оценивали с применением соотношения общего количества фрагментов размера [111136 п.о.] к общему количеству фрагментов размера [165175 п.о.] в пределах подмножества геномных блоков 100 т.о. С применением образцов от женщин, вынашивающих плод известного мужского пола, определяли первые 10% геномных блоков, которые характеризовались наивысшими корреляциями с фракцией плода, полученной из количества копий X-хромосомы [ссылка Rava]. Корреляцию между оценками фракции плода на основании размера фрагмента и таковыми, полученными из X-хромосомы в плодах известного мужского пола, вычисляли компьютерным способом с применением анализа перекрестной валидации с исключением по одному [ссылка], который включал как выбор блока, так и оценку параметра регрессионной модели. Затем вычисленную фракцию плода получали из соотношений размера фрагмента с применением модели линейной регрессии.

Результаты.

Упрощение процессинга библиотеки.

На фиг. 8 представлен общий рабочий процесс и временные рамки данной новой версии НИПТ по сравнению со стандартным лабораторным рабочим процессом. Весь рабочий процесс получения 96 образцов для выделения плазмы, экстракции сцДНК, конструирования библиотеки, количественного определения и объединения позволил процессировать образцы в течение общего времени менее 6 ч на одной системе Hamilton STAR. Это значительно отличается от 9 ч на двух системах Hamilton STAR с применением способов на основе ПНР, которые использовали в лаборатории CLIA. Количество сцДНК, экстрагированной на образец, в среднем составило 60 пг/мкл, и выход библиотеки секвенирования на выходе демонстрировал линейную корреляцию (R^2=0,94) с сцДНК на входе, как представлено на фиг. 9. Среднее восстановление составило более 70% (добавить диапазон), что свидетельствует о высокоэффективном восстановлении сцДНК после очистки на бусинах SPRI. В каждой серии секвенирования применяли нормированные количества 48 мультиплексированных образцов, и для завершения серии требовалось приблизительно 14 ч.

Секвенирование спаренных кониов.

Общее время секвенирования партии из 48 образцов на секвенаторе NextSeq 500 составило менее 14 ч. Это значительно отличается от 40 ч (1 проточная ячейка, 96 образцов) или 50 ч (2 проточных ячейки,

192 образцов) для лабораторного процесса на секвенаторе HiSeq 2500.

Картированные геномные расположения обоих концов фрагментов сцДНК обеспечили информацию о размере фрагментов сцДНК. На фиг. 10 представлено распределение размера фрагмента сцДНК, измеренного из 324 образцов от беременностей плодом мужского пола. Размер фрагментов, которые картировались на аутосомные хромосомы, которые установленно являются эуплоидными и преимущественно представляют материнские хромосомы, представлен тонкой кривой. Средний размер вставки составлял 175 п.о. причем XX% фрагментов были измерены от 100 п.о. до 200 п.о. Толстая кривая представляет размер фрагмента, который возникает исключительно из Y-хромосомы, представляя собой только фрагменты сцДНК плода. Распределение размера специфичных Y-хромосоме последовательностей было меньшим, в среднем составляло 167 п.о. с периодичностью оснований 10 при более коротких размерах фрагмента.

Поскольку более короткие фрагменты сцДНК обогащены ДНК плода, селективный анализ с применением исключительно более коротких фрагментов, как ожидается, увеличит относительное представление плода в связи с преимущественным выбором ридов плода. На фиг. 11 представлена относительная фракция плода из общего подсчета картированных ридов спаренных концов по сравнению с подсчетом от ридов спаренных концов, которые составляют менее 150 п.о. В целом, медиана фракции плода увеличивается в 2 раза по сравнению с общим подсчетом, хоть и с некоторым увеличением дисперсии. Было установлено, что предел размера 150 п.о. обеспечивал оптимальный компромисс для вычисления с увеличением представленности плода по сравнению с дисперсией в подсчетах.

Обнаружение анеуплоидии целой хромосомы у плода.

Каждую из доступных метрик, общий подсчет, подсчитанные значения менее 150 п.о., фракции подсчитанных значений, обогащенных сцДНК плода (подсчитанные значения от 80 до 150 п.о./подсчитанные значения <250 п.о.), и комбинацию подсчитанных значений более коротких фрагментов с фракцией исследовали в отношении способности устанавливать отличия образцов с трисомией 21 от образцов, эуплоидных по хромосоме 21. На фиг. 12 представлены результаты для каждой из данных метрик. Как видно из фиг. 12A и 12B, меньшие подсчитанные значения продемонстрировали лучшее разделение между трисомией 21 и эуплоидией, преимущественно, поскольку данная метрика обогащена сцДНК плода. Фракция сама по себе являлась приблизительно так же эффективной, как и общий подсчет, для отличия анеуплоидии (фиг. 12C), но при применении в комбинации с подсчетами коротких фрагментов (фиг. 12D) обеспечивала улучшенное установление отличий по сравнению с подсчетами коротких фрагментов самими по себе. Это указывает на то, что фракция обеспечивает независимую информацию, которая улучшает обнаружение трисомии 21. По сравнению с используемым на сегодняшний день рабочим процессом лаборатории CLIA с применением получения библиотеки с ПЦР-амплификацией и медианы 16 M подсчитанных значений/образец, рабочий процесс на основе секвенирования без применения ПЦР спаренных концов демонстрирует эквивалентные рабочие характеристики с в значительной степени меньшим количеством подсчитанных значений/образец (например, 6 M подсчитанных значений/образец или менее) и более простой, более короткий рабочий процесс получения образца.

Вычисление фракции плода.

С применением результатов для X-хромосомы от беременностей плодом мужского пола можно использовать нормированные значения хромосом с целью определения фракций плода для подсчитанных значений (ссылка ClinChem) и проводить сравнение для различных размеров фрагментов сцДНК. Фракции плода, полученные из X-хромосомы, использовали для калибровки соотношений для множества из 140 образцов и оценивали рабочие характеристики с применением перекрестной валидации с исключением по одному. На фиг. 13 представлены результаты перекрестно валидированных предсказаний фракции плода и корреляция между двумя данными множествами, которая свидетельствует, что оценки фракции плода можно получить из любых образцов, включая таковые от женщин, вынашивающих плод женского пола, после того как было измерено калибровочное множество.

Обсуждение.

Было продемонстрировано, что можно достичь высокой аналитической чувствительности и специфичности обнаружения анеуплоидии плода из сцДНК в материнской плазме при получении библиотеки без применения ПЦР в сочетании с секвенированием спаренных концов ДНК. Данный способ упрощает рабочий процесс, улучшает время оборота (фиг. 8) и должен устранить некоторые погрешности, присущие способам на основе ПЦР. Секвенирование спаренных концов позволяет определить размеры длины фрагмента и фракцию плода, которые затем можно применять для усиления обнаружения анеуплоидии при значительно меньшем подсчете метки по сравнению с применяемыми на сегодняшний день коммерческими способами. Рабочие характеристики варианта реализации спаренных концов без применения ПЦР, как представляется, аналогичны способам секвенирования одиночных концов, в которых применяют вплоть до в три раза большее количество меток.

Упрощение процессинга библиотеки.

Рабочий процесс без применения ПЦР характеризуется несколькими преимуществами для клинических лабораторий. Благодаря высокому выходу и линейным законам получения библиотеки нормированные пулы образцов для секвенирования можно получить непосредственно из концентраций библио-

теки индивидуального образца. В результате этого устраняются погрешности, присущие ПЦР-амплификации процесса получения библиотеки. Помимо этого, отсутствует потребность в выделении отдельных жидкостных манипуляторов для активностей до и после ПЦР, что снижает материальную нагрузку на лабораторию. Это упрощает рабочий процесс, позволяет готовить партии образцов в клинической лаборатории в одну смену, а затем секвенировать и анализировать в течение ночи. В целом, снижение капитализируемых расходов, уменьшение времени "работы руками" и быстрый оборот потенциально позволяют в значительной степени снизить стоимость и, в целом, устойчивость НИПТ.

Секвенирование спаренных концов.

Применение секвенирования спаренных концов на системе NextSeq 500 характеризуется несколькими преимуществами при подсчете фрагментов сцДНК. Во-первых, с применением двойных индексных штрих-кодов образцы можно мультиплексировать на высоком уровне, что позволяет проводить нормирование и коррекцию вариации от серии к серии с высокой статистической достоверностью. Помимо этого, поскольку на серию мультиплексируют 48 образцов, и количество, необходимое для кластеризации на проточной ячейке, ограничено, требование к образцу на входе в значительной степени снижается, что позволяет использовать рабочий процесс библиотеки без применения ПЦР. При типичном выходе сцДНК приблизительно 5 нг на образец исследователям удалось получить 2-3 серии секвенирования на образец даже без применения ПЦР-амплификации. Это в значительной степени отличается от других подходов, для которых требуются значительные количества плазмы на входе из множества пробирок для сбора крови с целью получения выхода достаточного количества сцДНК для определения анеуплоидии (ссылка). Наконец, секвенирование спаренных концов позволяет проводить определение размера фрагмента сцДНК и аналитическое обогащение сцДНК плода.

Обнаружение анеуплоидии целой хромосомы у плода.

Результаты, полученные авторами настоящего изобретения, демонстрируют, что подсчитанные значения фрагментов сцДНК менее 150 п.о. способны лучше установить отличия анеуплоидии от эуплоидных хромосом, чем общие подсчитанные значения. Данное наблюдение отличается от результатов Fan et al., которые предположили, что при применении более коротких фрагментов точность статистики подсчета будет снижена (Fan et al.) вследствие снижения количества доступных подсчитанных значений. Фракция коротких фрагментов также обеспечивает установление некоторых отличий для обнаружения трисомии 21, как было установлено Yu et al., хотя и с меньшим динамическим диапазоном, чем подсчитанные значения. Однако объединение подсчета и метрик фракции приводит к наилучшему отделению образцов трисомии 21 от эуплоидных и подразумевает, что две данные метрики являются комплементарными измерениями для представленности хромосомы. Другие биологические метрики, например, метилирование, могут также обеспечить ортогональную информацию, которая может усилить соотношение сигнал/шум для обнаружения анеуплоидии.

Вычисление фракции плода.

Способы, представленные в настоящем документе, также позволяют оценить фракцию плода в каждом образце без проведения дополнительной лабораторной работы. С применением множества образцов в каждой проточной ячейке, приблизительно половина из которых являются образцами мужского пола, можно получить точную оценку фракции плода для всех образцов посредством калибровки измерения фракции плода из информации о размере фрагмента с таковой, определенной для мужских образцов. В коммерческих условиях клинический опыт исследователей продемонстрировал, что стандартные способы подсчета с применением большего количества меток одиночных концов привели к очень низкой доле ложноотрицательных результатов даже при отсутствии специфичных измерений фракции плода (ссылка). С учетом аналогичного предела обнаружения, наблюдаемого в настоящем документе, вычисляют получить эквивалентные рабочие характеристики тестирования.

Заключение.

Было продемонстрировано, что высокой аналитической чувствительности и специфичности обнаружения анеуплоидии плода из сцДНК в материнской плазме можно достичь с получением библиотеки без применения ПЦР в сочетании с секвенированием спаренных концов ДНК. Данный упрощенный рабочий процесс характеризуется очень быстрым временем оборота, которое потенциально позволяет проводить НИПТ со значительно меньшими финансовыми затратами для применения в общей популяции беременных. Помимо этого, методики секвенирования спаренных концов характеризуются потенциалом измерять другой биологический феномен, а также обеспечивать другие клинические варианты применения. Например, информация о размере из метилированных конкретных областей генома или CpG-островков может обеспечить другую ортогональную метрику для усиления обнаружения вариантов числа копий в пределах генома.

Настоящее изобретение можно реализовать в других конкретных формах, не выходя за пределы духа или существенных характеристик изобретения. Описанные варианты реализации следует считать во всех отношениях исключительно иллюстративными, а не ограничивающими. Вследствие этого объем настоящего изобретения определен прилагаемой формулой изобретения, а не вышеупомянутым описанием. Все изменения, которые попадают в значение и диапазон эквивалентности формулы изобретения, подлежат охвату ее объемом.

ФОРМУЛА ИЗОБРЕТЕНИЯ

1. Способ определения вариации числа копий (ВЧК) последовательности нуклеиновой кислоты, представляющей интерес, в исследуемом образце, содержащем фрагменты бесклеточной нуклеиновой кислоты, происходящие из двух или более геномов, причем указанный способ включает:

(a) прием ридов последовательности, полученных в результате секвенирования указанных фрагментов бесклеточной нуклеиновой кислоты в исследуемом образце;

(b) выравнивание указанных ридов последовательности фрагментов бесклеточной нуклеиновой кислоты или выравнивание фрагментов, содержащих указанные риды последовательности, с блоками референсного генома, содержащего последовательность, представляющую интерес, с получением, таким образом, меток исследуемой последовательности, причем референсный геном разделен на множество блоков;

(c) определение размеров фрагментов указанных фрагментов бесклеточной нуклеиновой кислоты, присутствующих в исследуемом образце;

(d) получение первых перекрытий меток последовательности для блоков референсного генома с применением меток последовательности для фрагментов бесклеточной нуклеиновой кислоты, имеющих размеры в первом диапазоне размеров;

(e) получение вторых перекрытий меток последовательности для блоков референсного генома с применением меток последовательности для фрагментов бесклеточной нуклеиновой кислоты, имеющих размеры во втором диапазоне размеров, причем второй диапазон размеров отличается от первого диапазона размеров, указанный первый диапазон размеров содержит фрагменты бесклеточной нуклеиновой кислоты всех размеров в образце, а указанный второй диапазон размеров содержит только фрагменты бесклеточной нуклеиновой кислоты, меньшие, чем заданный размер; и

(f) определение вариации числа копий в последовательности, представляющей интерес, с применением указанных первых перекрытий и вторых перекрытий.

2. Способ по п.1, дополнительно включающий вычисление характеристик размера для блоков указанного референсного генома с применением размеров фрагментов, определенных на этапе (c), где вариацию числа копий определяют с применением указанных первых перекрытий, вторых перекрытий и характеристик размера.

3. Способ по п.1, характеризующийся тем, что указанный второй диапазон размеров содержит только фрагменты бесклеточной нуклеиновой кислоты, меньшие чем 150 п.о.

4. Способ по п.1, характеризующийся тем, что этап (f) включает вычисление правдоподобия плоидности по первым перекрытиям и вторым перекрытиями, где правдоподобие плоидности включает первую вероятность того, что указанные первые перекрытия и указанные вторые перекрытия или полученные по ним статистики соответствуют модели, которая характеризуется анеуплоидным допущением, и вторую вероятность того, что указанные первые перекрытия и указанные вторые перекрытия или полученные по ним статистики соответствуют модели, которая характеризуется эуплоидным допущением.

5. Способ по п.4, характеризующийся тем, что указанные статистики содержат первую t-статистику для указанной последовательности, представляющей интерес, с применением указанных первых перекрытий и вторую t-статистику для указанной последовательности, представляющей интерес, с применением указанных вторых перекрытий.

6. Способ по п.5, характеризующийся тем, что t-статистику для последовательности, представляющей интерес, вычисляют с применением перекрытий блоков в указанной последовательности, представляющей интерес, и перекрытий блоков в референсной области для последовательности, представляющей интерес.

7. Способ по п.6, характеризующийся тем, что правдоподобие плоидности содержит отношение правдоподобия между указанной первой вероятностью и указанной второй вероятностью.

8. Способ по п.7, характеризующийся тем, что указанное отношение правдоподобия вычисляют по одному или более значениям фракции плода в дополнение к указанным первой и второй t-статистикам.

9. Способ по п.8, дополнительно включающий вычисление указанных одного или более значений фракции плода с применением информации относительно размеров фрагментов бесклеточной нуклеиновой кислоты.

10. Способ по п.8, характеризующийся тем, что указанное отношение правдоподобия содержит

$$\text{ОВ} = \frac{\Sigma_{ff_{\text{суммарн.}}} q(ff_{\text{суммарн.}}) * p_1(T_{\text{коротк.}}, T_{\text{всех}} | ff_{\text{выч.}})}{p_0(T_{\text{коротк.}}, T_{\text{всех}})},$$

где p_1 представляет собой правдоподобие того, что данные получены из многомерного нормального распределения, представляющего 3-копийную или 1-копийную модель, p_0 представляет собой правдоподобие того, что данные получены из многомерного нормального распределения, представляющего 2-копийную модель, $T_{\text{коротк.}}$, $T_{\text{всех}}$ представляют собой T-показатели, вычисленные по хромосомному перекрытию, полученному с помощью коротких фрагментов и всех фрагментов, и $q(ff_{\text{суммарн.}})$ представляет собой плотность распределения фракции плода.

11. Способ по п.10, характеризующийся тем, что указанное отношение правдоподобия вычисляют

для моносомии X, трисомии X, трисомии 13, трисомии 18 или трисомии 21.

12. Способ по п.1, характеризующийся тем, что этап (d) и/или (e) включает:

(i) определение количества меток последовательности, которые выравниваются с блоком; и

(ii) нормирование количества меток последовательности, которые выравниваются с блоком, посредством подсчета межблоковых вариаций, вызванных факторами, отличными от вариации числа копий.

13. Способ по п.12, характеризующийся тем, что нормирование количества меток последовательности включает: нормирование с учетом содержания GC в образце, нормирование с учетом глобального волнового профиля вариации обучающего множества и/или нормирование с учетом одной или более компонент, полученных из анализа главных компонент.

14. Способ по п.2, характеризующийся тем, что указанная характеристика размера для блока включает отношение фрагментов размера, меньшего, чем заданное значение, к общему количеству фрагментов в блоке.

15. Способ по п.2, дополнительно включающий вычисление третьей t-статистики для последовательности, представляющей интерес, с применением характеристик размера блоков в последовательности, представляющей интерес.

16. Способ по п.15, характеризующийся тем, что этап (f) включает вычисление первого отношения правдоподобия по первой t-статистике для последовательности, представляющей интерес, с применением перекрытий, вычисленных на этапе (d), и второй t-статистики для последовательности, представляющей интерес, с применением перекрытий, вычисленных на этапе (e), и указанной третьей t-статистики.

17. Система для оценки числа копий последовательности нуклеиновой кислоты, представляющей интерес, в исследуемом образце, содержащем фрагменты бесклеточной нуклеиновой кислоты, содержащая

секвенатор для приема фрагментов нуклеиновой кислоты из исследуемого образца и обеспечения информации о последовательности нуклеиновой кислоты исследуемого образца;

процессор; и

один или более машиночитаемых носителей для хранения информации, на которых хранятся инструкции для исполнения на указанном процессоре для:

(a) приема ридов последовательности, полученных в результате секвенирования указанных фрагментов бесклеточной нуклеиновой кислоты в исследуемом образце;

(b) выравнивания указанных ридов последовательности указанных фрагментов бесклеточной нуклеиновой кислоты или выравнивания фрагментов, содержащих указанные риды последовательности, с блоками референсного генома, содержащего последовательность, представляющую интерес, с получением, таким образом, меток исследуемой последовательности, причем референсный геном разделен на множество блоков;

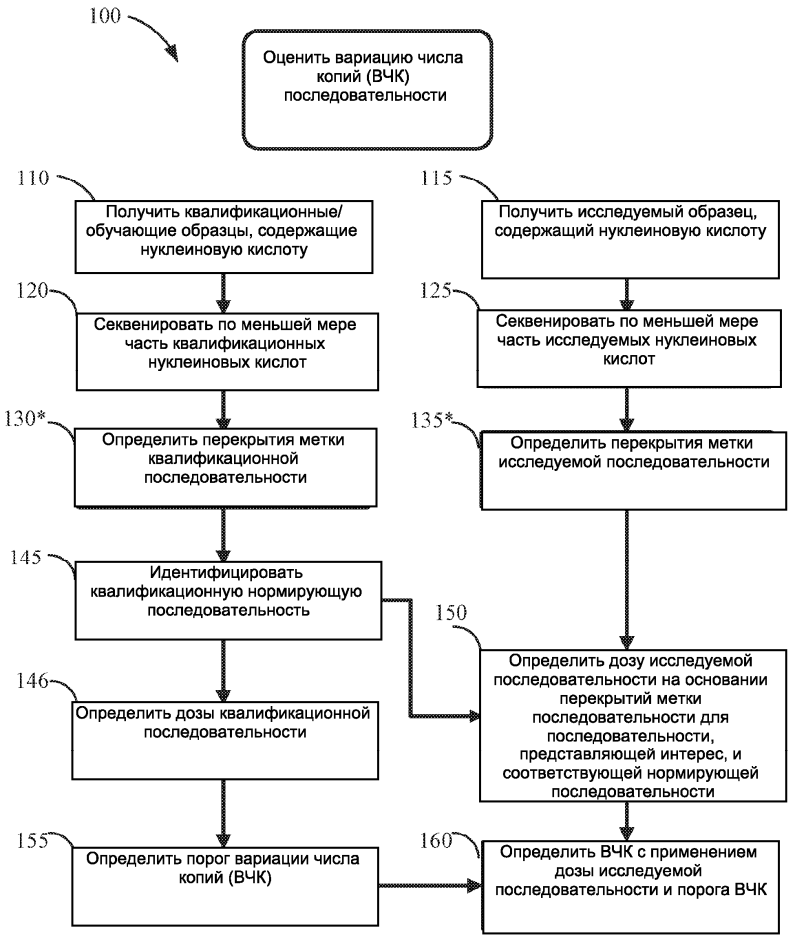
(c) определения размеров фрагментов бесклеточной нуклеиновой кислоты, присутствующих в исследуемом образце;

(d) получения первых перекрытий меток последовательности для блоков референсного генома с применением меток последовательности для фрагментов бесклеточной нуклеиновой кислоты, имеющих размеры в первом диапазоне размеров;

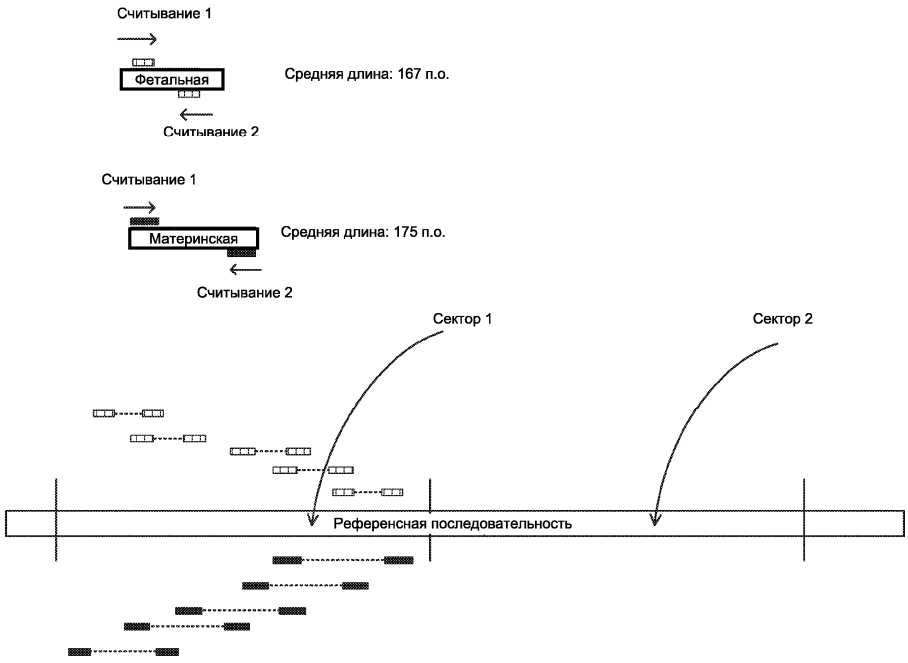
(e) получения вторых перекрытий меток последовательности для блоков референсного генома с применением меток последовательности для фрагментов бесклеточной нуклеиновой кислоты, имеющих размеры во втором диапазоне размеров, где указанный второй диапазон размеров отличается от указанного первого диапазона размеров, указанный первый диапазон размеров содержит фрагменты бесклеточной нуклеиновой кислоты всех размеров в образце, а указанный второй диапазон размеров содержит только фрагменты бесклеточной нуклеиновой кислоты, меньшие, чем заданный размер; и

(f) определения вариации числа копий в последовательности, представляющей интерес, с применением указанных первых перекрытий и указанных вторых перекрытий.

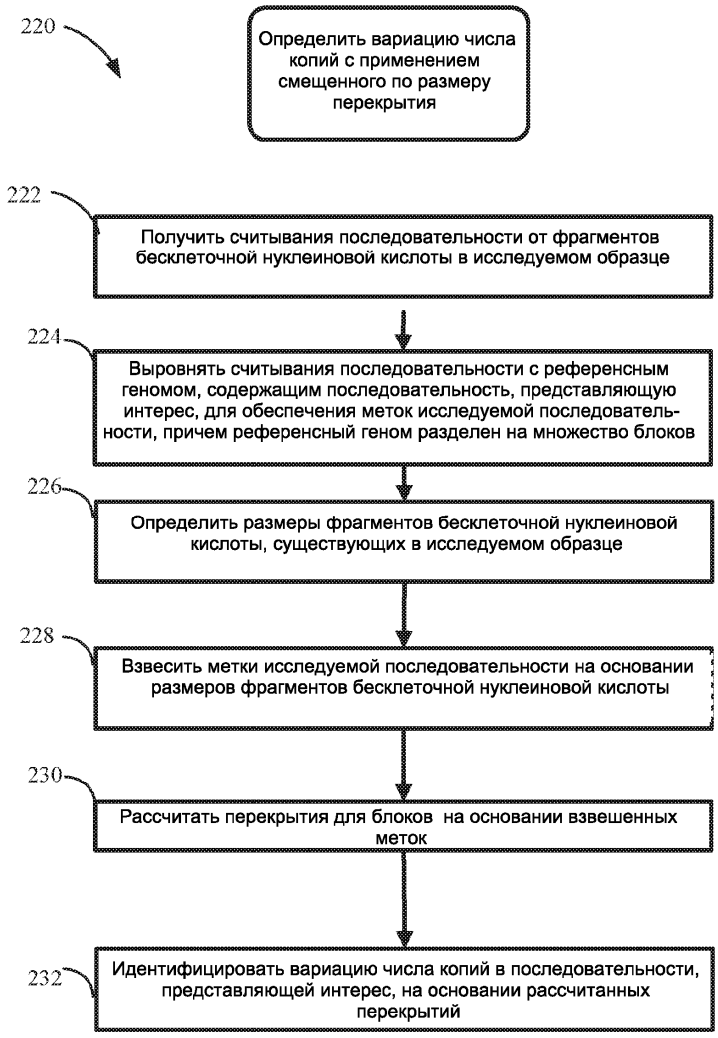
18. Машиночитаемый носитель, предназначенный для долговременного хранения информации, на котором хранится программный код в форме инструкций для компьютерной системы, запрограммированный для осуществления способа по любому из пп.1-17.


100
Оценить вариацию числа копий (ВЧК) последовательности
110
Получить квалификационные/ обучающие образцы, содержащие нуклеиновую кислоту
115
Получить исследуемый образец, содержащий нуклеиновую кислоту
120
Секвенировать по меньшей мере часть квалификационных нуклеиновых кислот
125
Секвенировать по меньшей мере часть исследуемых нуклеиновых кислот
130*
Определить перекрытия метки квалификационной последовательности
135*
Определить перекрытия метки исследуемой последовательности
145
Идентифицировать квалификационную нормирующую последовательность
150
Определить дозу исследуемой последовательности на основании перекрытий метки последовательности для последовательности, представляющей интерес, и соответствующей нормирующей последовательности
146
Определить дозы квалификационной последовательности
155
Определить порог вариации числа копий (ВЧК)
160
Определить ВЧК с применением дозы исследуемой последовательности и порога ВЧК

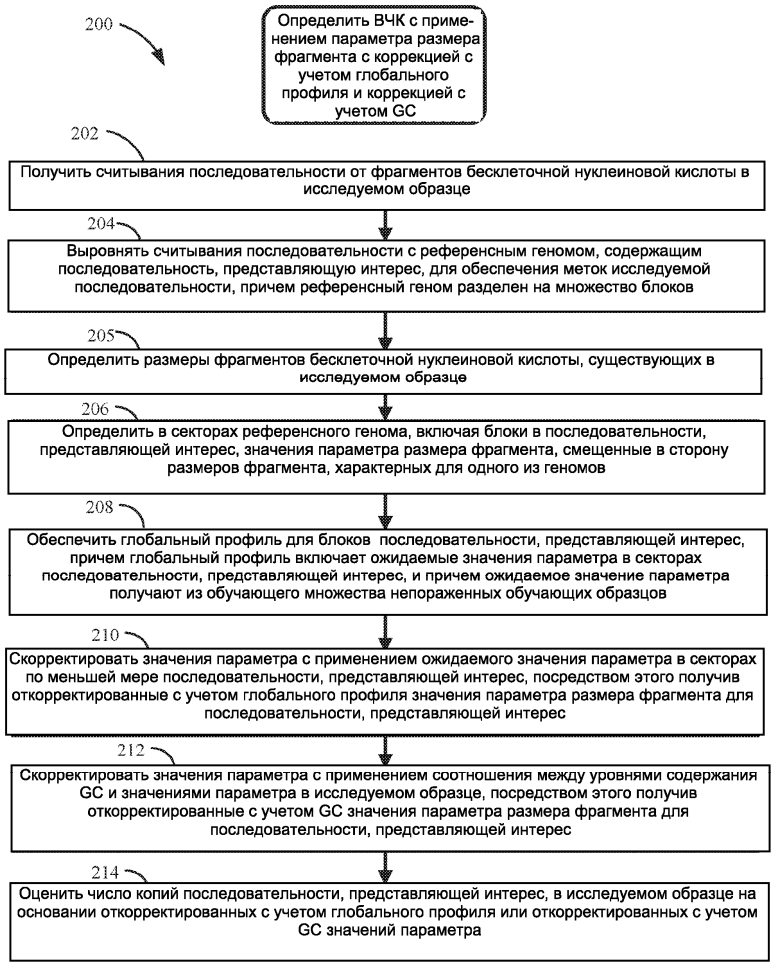
Фиг. 1


Считывание 1
Фетальная
Средняя длина: 167 п.о.
Считывание 2
Считывание 1
Материнская
Средняя длина: 175 п.о.
Считывание 2
Сектор 1
Сектор 2
Референсная последовательность

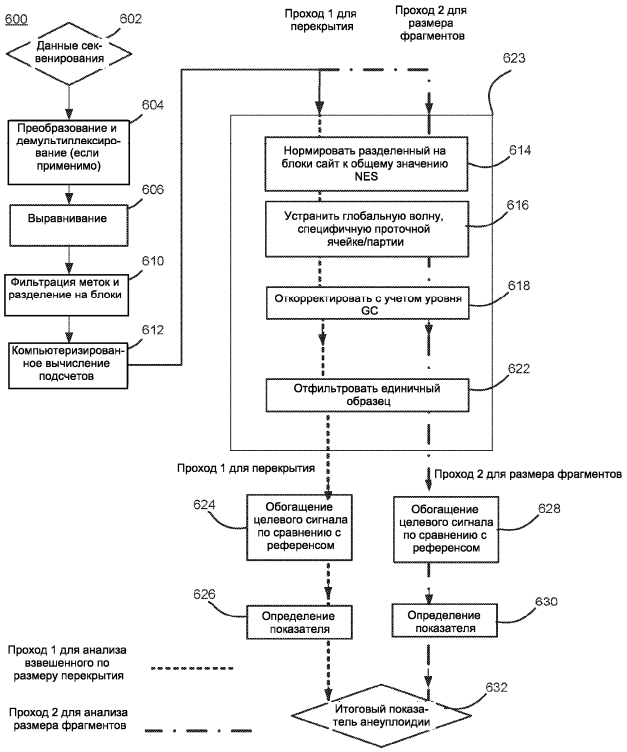
Фиг. 2А


220
Определить вариацию числа копий с применением смещенного по размеру перекрытия
222
Получить считывания последовательности от фрагментов бесклеточной нуклеиновой кислоты в исследуемом образце
224
Выровнять считывания последовательности с референсным геномом, содержащим последовательность, представляющую интерес, для обеспечения меток исследуемой последовательности, причем референсный геном разделен на множество блоков
226
Определить размеры фрагментов бесклеточной нуклеиновой кислоты, существующих в исследуемом образце
228
Взвесить метки исследуемой последовательности на основании размеров фрагментов бесклеточной нуклеиновой кислоты
230
Рассчитать перекрытия для блоков на основании взвешенных меток
232
Идентифицировать вариацию числа копий в последовательности, представляющей интерес, на основании рассчитанных перекрытий

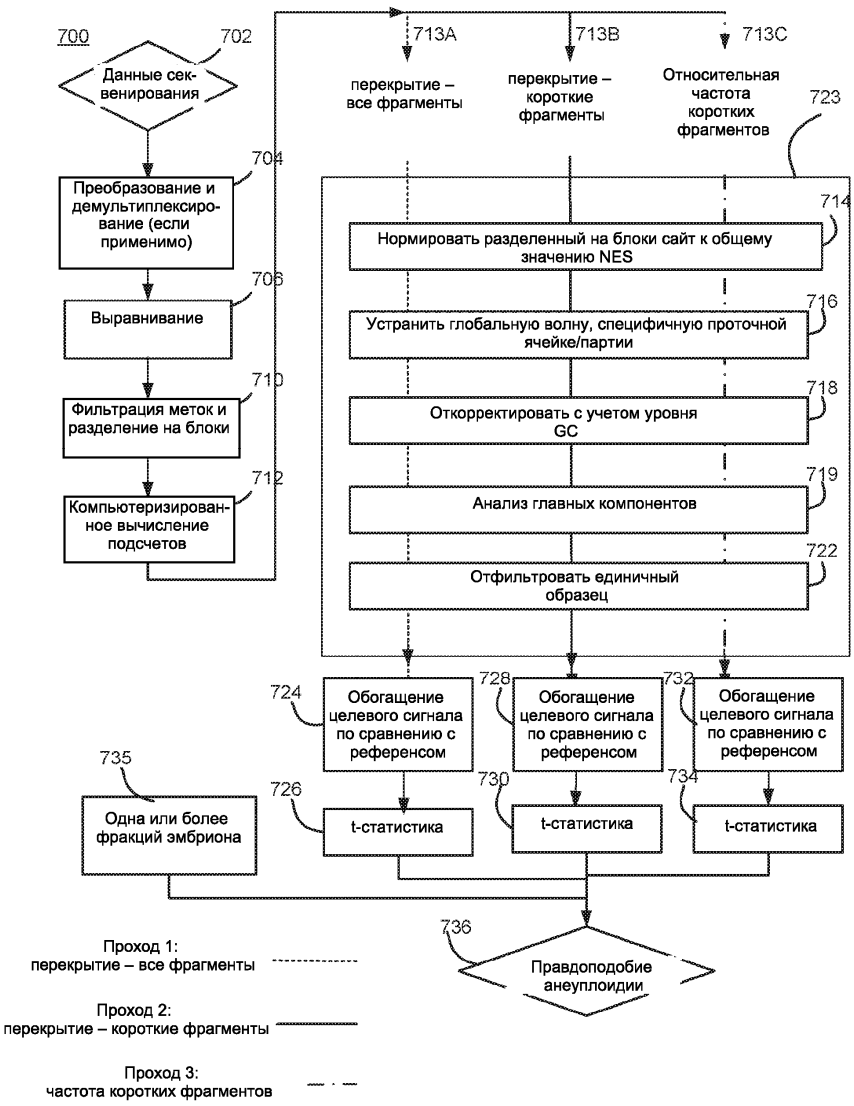
Фиг. 2В


200
Определить ВЧК с применением параметра размера фрагмента с коррекцией с учетом глобального профиля и коррекцией с учетом GC
202
Получить считывания последовательности от фрагментов бесклеточной нуклеиновой кислоты в исследуемом образце
204
Выровнять считывания последовательности с референсным геномом, содержащим последовательность, представляющую интерес, для обеспечения меток исследуемой последовательности, причем референсный геном разделен на множество блоков
205
Определить размеры фрагментов бесклеточной нуклеиновой кислоты, существующих в исследуемом образце
206
Определить в секторах референсного генома, включая блоки в последовательности, представляющей интерес, значения параметра размера фрагмента, смещенные в сторону размеров фрагмента, характерных для одного из геномов
208
Обеспечить глобальный профиль для блоков последовательности, представляющей интерес, причем глобальный профиль включает ожидаемые значения параметра в секторах последовательности, представляющей интерес, и причем ожидаемое значение параметра получают из обучающего множества непораженных обучающих образцов
210
Скорректировать значения параметра с применением ожидаемого значения параметра в секторах по меньшей мере последовательности, представляющей интерес, посредством этого получив откорректированные с учетом глобального профиля значения параметра размера фрагмента для последовательности, представляющей интерес
212
Скорректировать значения параметра с применением соотношения между уровнями содержания GC и значениями параметра в исследуемом образце, посредством этого получив откорректированные с учетом GC значения параметра размера фрагмента для последовательности, представляющей интерес
214
Оценить число копий последовательности, представляющей интерес, в исследуемом образце на основании откорректированных с учетом глобального профиля или откорректированных с учетом GC значений параметра

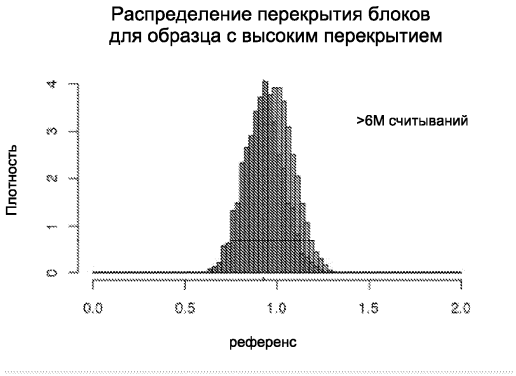
Фиг. 2C

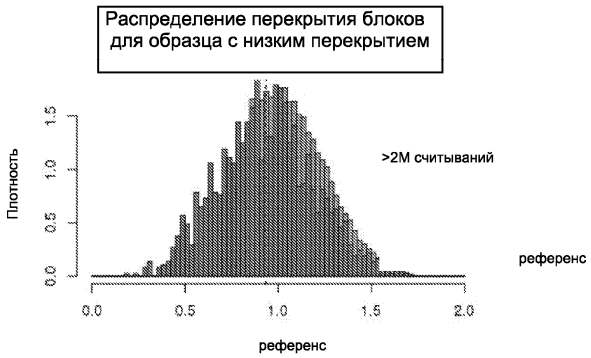

600
602
Данные секвенирования
604
Преобразование и демультиплексирование (если применимо)
606
Выравнивание
610
Фильтрация меток и разделение на блоки
612
Компьютеризированное вычисление подсчетов
Проход 1 для перекрытия
Проход 2 для размера фрагментов
623
614
Нормировать разделенный на блоки сайт к общему значению NES
616
Устранить глобальную волну, специфичную проточной ячейке/партии
618
Откорректировать с учетом уровня GC
622
Отфильтровать единичный образец
Проход 1 для перекрытия
Проход 2 для размера фрагментов
624
Обогащение целевого сигнала по сравнению с референсом
628
Обогащение целевого сигнала по сравнению с референсом
626
Определение показателя
630
Определение показателя
Проход 1 для анализа взвешенного по размеру перекрытия
Проход 2 для анализа размера фрагментов
632
Итоговый показатель анеуплоидии

Фиг. 2D


700
Данные секвенирования
702
Преобразование и демультиплексирование (если применимо)
704
Выравнивание
706
Фильтрация меток и разделение на блоки
710
Компьютеризированное вычисление подсчетов
712
713A
перекрытие – все фрагменты
713B
перекрытие – короткие фрагменты
713C
Относительная частота коротких фрагментов
723
Нормировать разделенный на блоки сайт к общему значению NES
714
Устранить глобальную волну, специфичную проточной ячейке/партии
716
Откорректировать с учетом уровня GC
718
Анализ главных компонентов
719
Отфильтровать единичный образец
722
Обогащение целевого сигнала по сравнению с референсом
724
Обогащение целевого сигнала по сравнению с референсом
728
Обогащение целевого сигнала по сравнению с референсом
732
Одна или более фракций эмбриона
735
t-статистика
726
t-статистика
730
t-статистика
734
Правдоподобие анеуплоидии
736
Проход 1: перекрытие – все фрагменты
Проход 2: перекрытие – короткие фрагменты
Проход 3: частота коротких фрагментов

Фиг. 2E

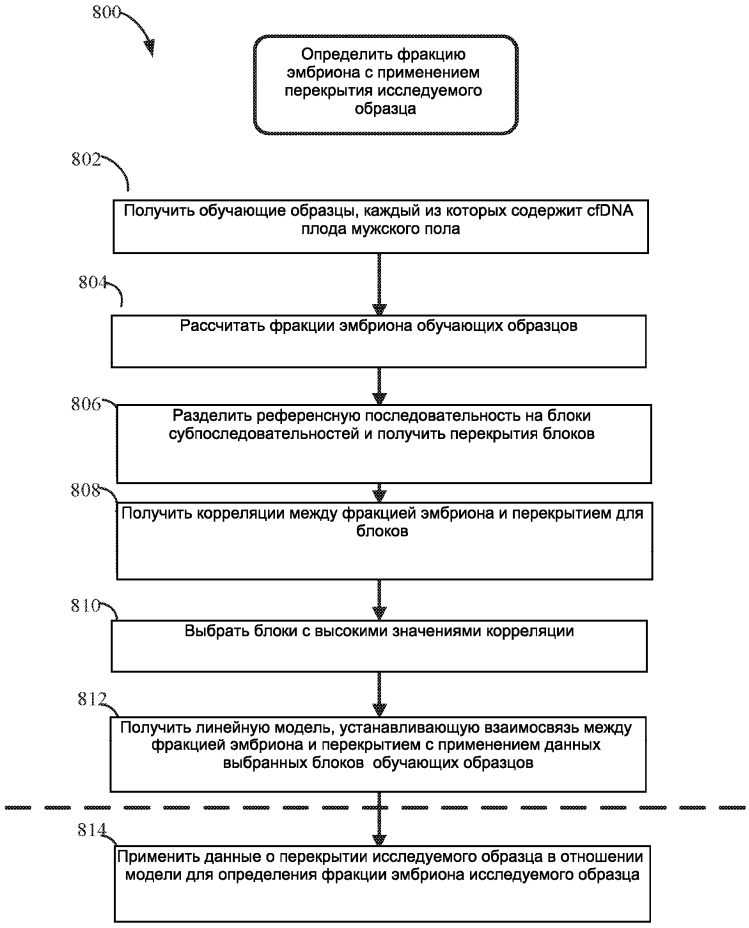




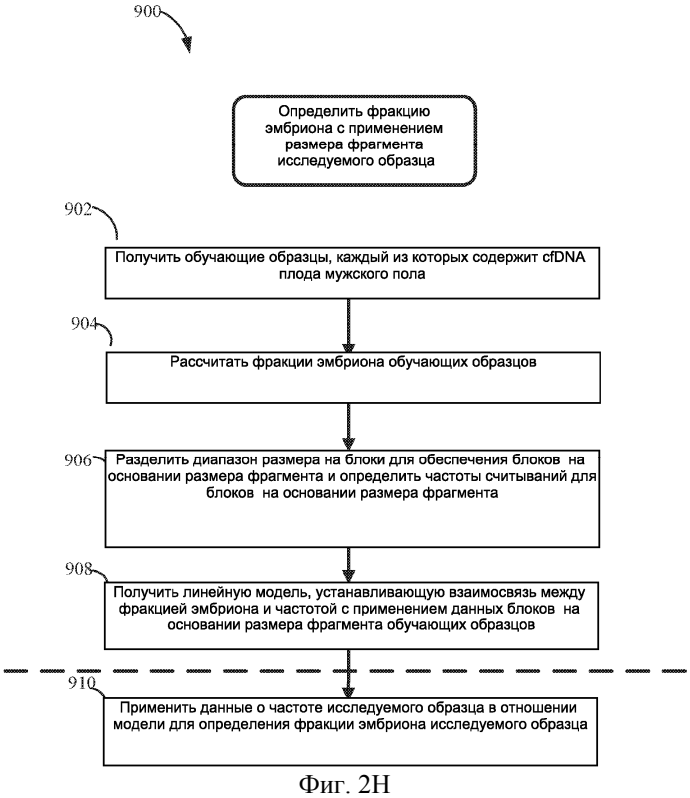
$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$\bar{X}_1$ = перекрытие блоков для хромосомы, представляющей интерес
$\bar{X}_2$ = перекрытие блоков референсной области

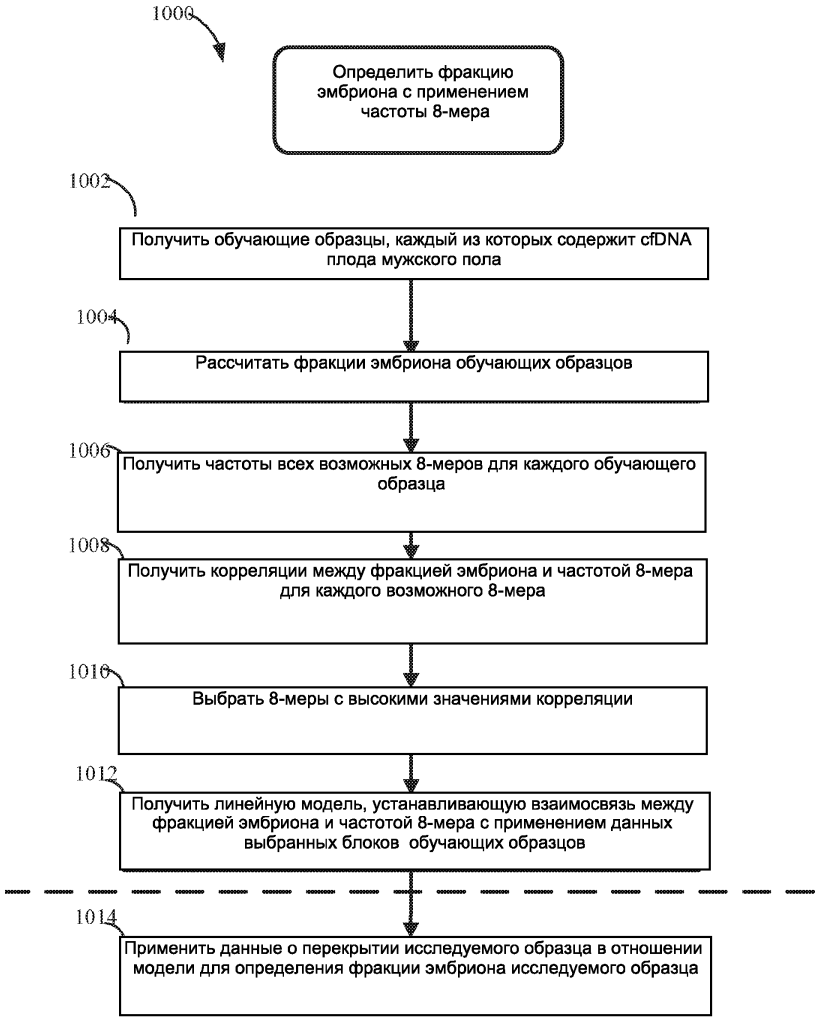
Фиг. 2F


800
Определить фракцию эмбриона с применением перекрытия исследуемого образца
802
Получить обучающие образцы, каждый из которых содержит cfDNA плода мужского пола
804
Рассчитать фракции эмбриона обучающих образцов
806
Разделить референсную последовательность на блоки субпоследовательностей и получить перекрытия блоков
808
Получить корреляции между фракцией эмбриона и перекрытием для блоков
810
Выбрать блоки с высокими значениями корреляции
812
Получить линейную модель, устанавливающую взаимосвязь между фракцией эмбриона и перекрытием с применением данных выбранных блоков обучающих образцов
814
Применить данные о перекрытии исследуемого образца в отношении модели для определения фракции эмбриона исследуемого образца

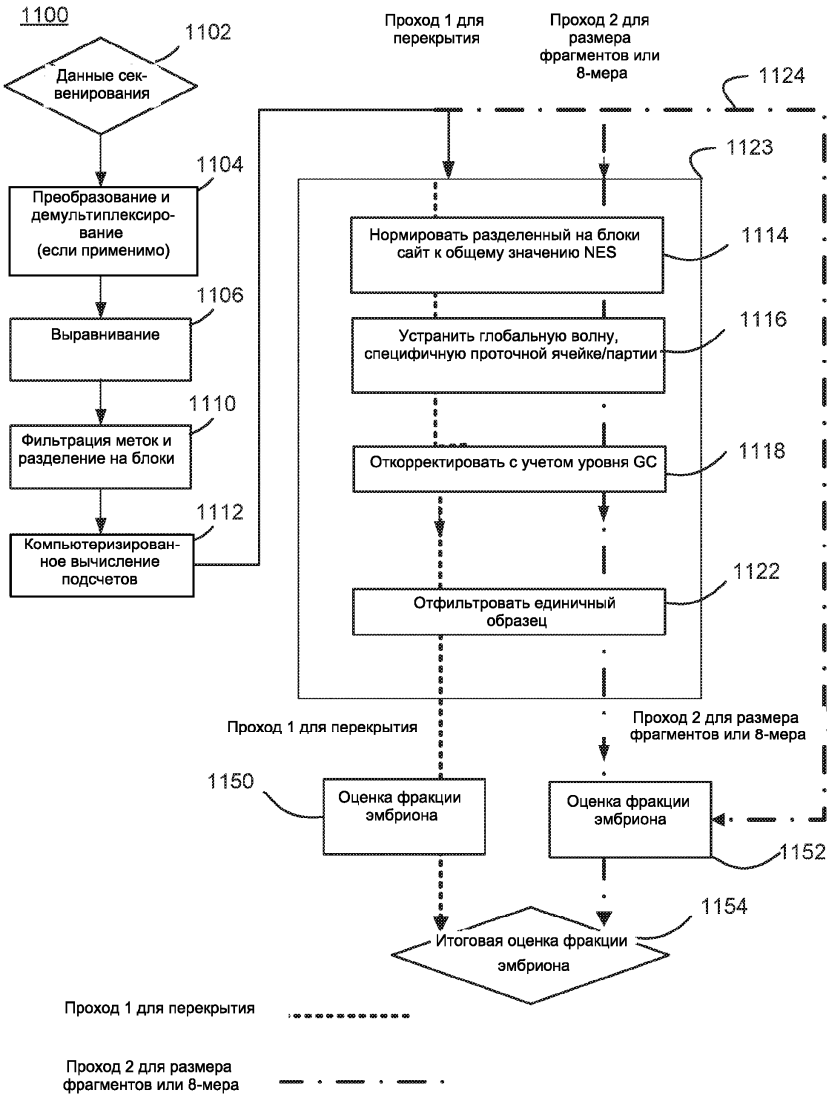
Фиг. 2G


900
Определить фракцию эмбриона с применением размера фрагмента исследуемого образца
902
Получить обучающие образцы, каждый из которых содержит cfDNA плода мужского пола
904
Рассчитать фракции эмбриона обучающих образцов
906
Разделить диапазон размера на блоки для обеспечения блоков на основании размера фрагмента и определить частоты считываний для блоков на основании размера фрагмента
908
Получить линейную модель, устанавливающую взаимосвязь между фракцией эмбриона и частотой с применением данных блоков на основании размера фрагмента обучающих образцов
910
Применить данные о частоте исследуемого образца в отношении модели для определения фракции эмбриона исследуемого образца

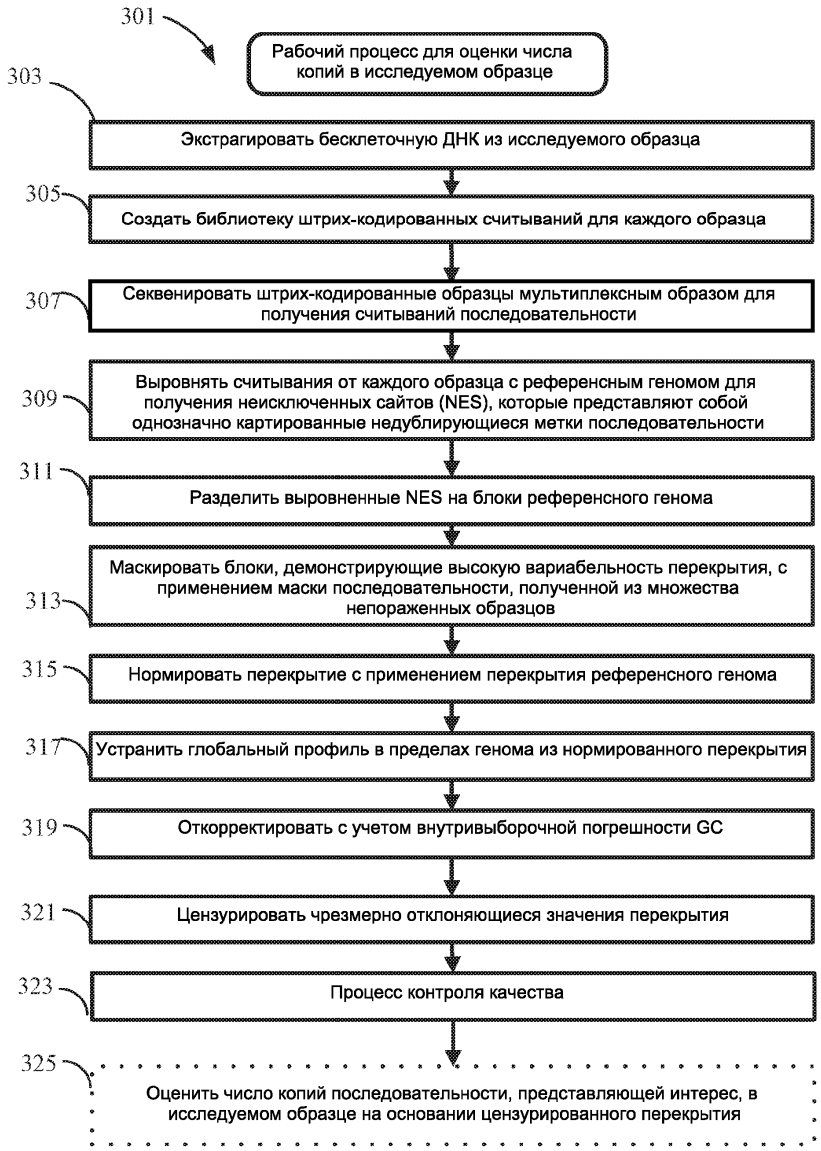
Фиг. 2H


1000
Определить фракцию эмбриона с применением частоты 8-мера
1002
Получить обучающие образцы, каждый из которых содержит cfDNA плода мужского пола
1004
Рассчитать фракции эмбриона обучающих образцов
1006
Получить частоты всех возможных 8-меров для каждого обучающего образца
1008
Получить корреляции между фракцией эмбриона и частотой 8-мера для каждого возможного 8-мера
1010
Выбрать 8-меры с высокими значениями корреляции
1012
Получить линейную модель, устанавливающую взаимосвязь между фракцией эмбриона и частотой 8-мера с применением данных выбранных блоков обучающих образцов
1014
Применить данные о перекрытии исследуемого образца в отношении модели для определения фракции эмбриона исследуемого образца

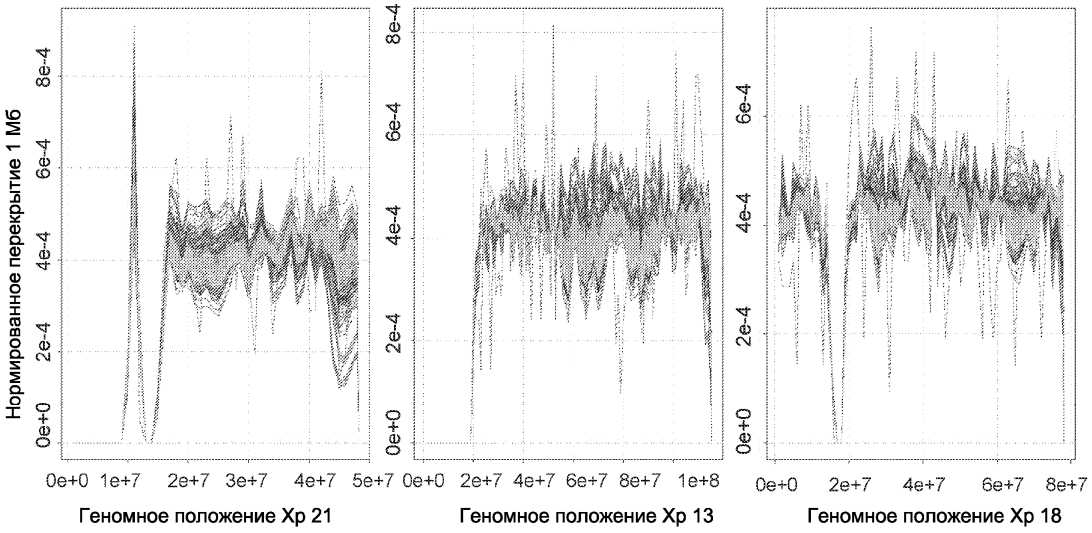
Фиг. 2I


1100
1102
Данные секвенирования
1104
Преобразование и демультиплексирование (если применимо)
1106
Выравнивание
1110
Фильтрация меток и разделение на блоки
1112
Компьютеризированное вычисление подсчетов
Проход 1 для перекрытия
Проход 2 для размера фрагментов или 8-мера
1124
1123
Нормировать разделенный на блоки сайт к общему значению NES
1114
Устранить глобальную волну, специфичную проточной ячейке/партии
1116
Откорректировать с учетом уровня GC
1118
Отфильтровать единичный образец
1122
Проход 1 для перекрытия
Проход 2 для размера фрагментов или 8-мера
1150
Оценка фракции эмбриона
Оценка фракции эмбриона
1152
1154
Итоговая оценка фракции эмбриона
Проход 1 для перекрытия
Проход 2 для размера фрагментов или 8-мера

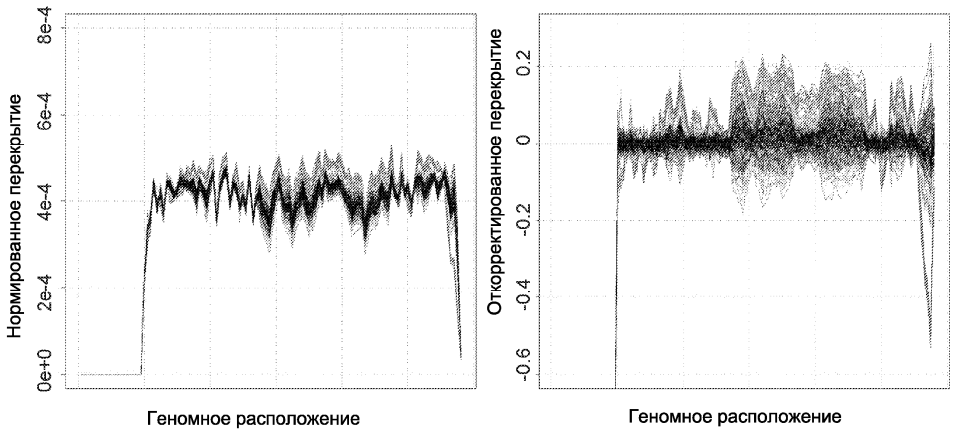
Фиг. 2J


301
Рабочий процесс для оценки числа копий в исследуемом образце
303
Экстрагировать бесклеточную ДНК из исследуемого образца
305
Создать библиотеку штрих-кодированных считываний для каждого образца
307
Секвенировать штрих-кодированные образцы мультиплексным образом для получения считываний последовательности
309
Выровнять считывания от каждого образца с референсным геномом для получения неисключенных сайтов (NES), которые представляют собой однозначно картированные недублирующиеся метки последовательности
311
Разделить выровненные NES на блоки референсного генома
313
Маскировать блоки, демонстрирующие высокую вариабельность перекрытия, с применением маски последовательности, полученной из множества непораженных образцов
315
Нормировать перекрытие с применением перекрытия референсного генома
317
Устранить глобальный профиль в пределах генома из нормированного перекрытия
319
Откорректировать с учетом внутривыборочной погрешности GC
321
Цензурировать чрезмерно отклоняющиеся значения перекрытия
323
Процесс контроля качества
325
Оценить число копий последовательности, представляющей интерес, в исследуемом образце на основании цензурированного перекрытия

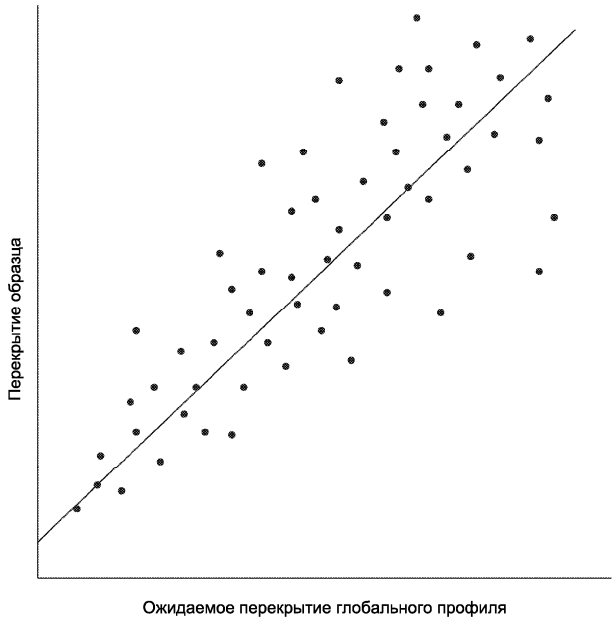
Фиг. 3A


Нормированное перекрытие 1 Мб
0e+0
2e-4
4e-4
6e-4
8e-4
0e+0 1e+7 2e+7 3e+7 4e+7 5e+7
Геномное положение Xp 21
0e+0 2e+7 4e+7 6e+7 8e+7 1e+8
Геномное положение Xp 13
0e+0 2e+7 4e+7 6e+7 8e+7
Геномное положение Xp 18

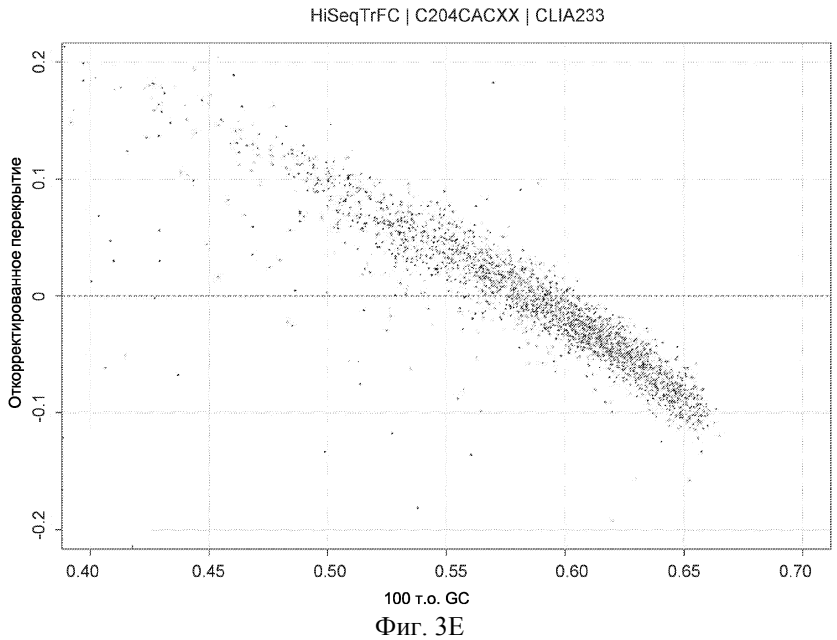
Фиг. 3B



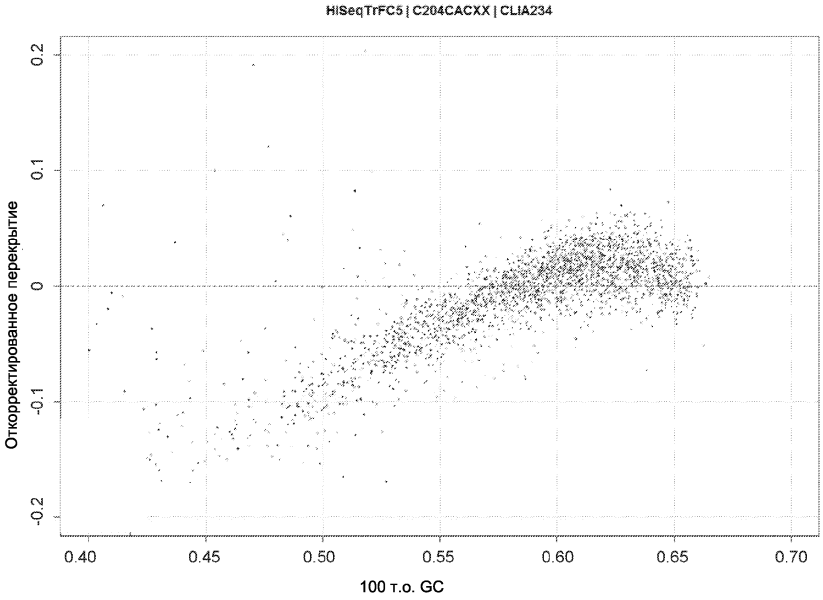
Фиг. 3C



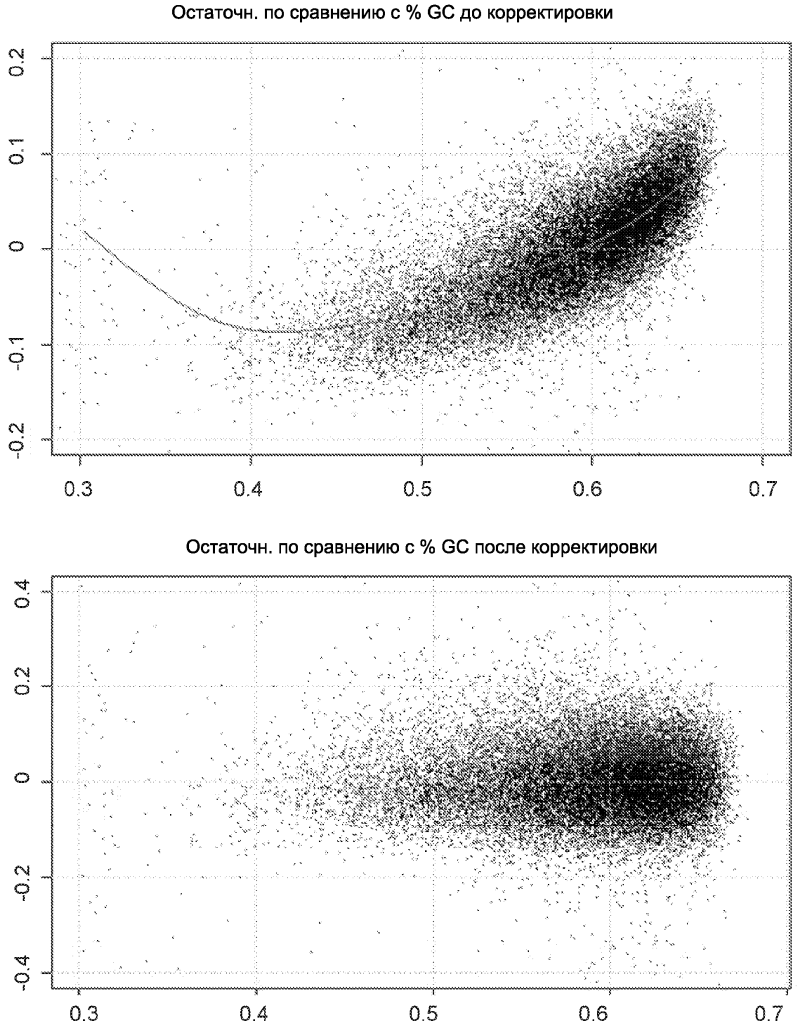
Фиг. 3D



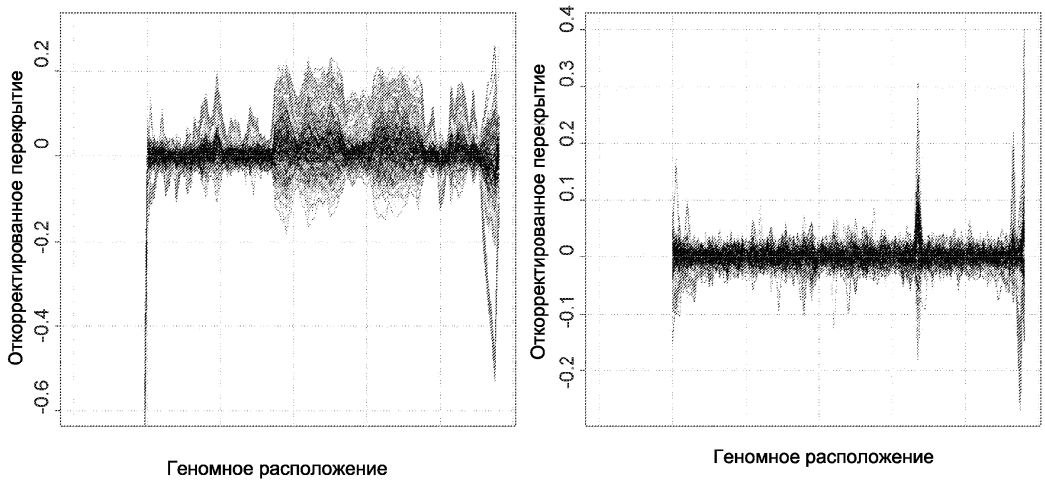
Фиг. 3E


HiSeqTrFC5 | C204CACXX | CLIA234
Откорректированное перекрытие
0.2
0.1
0
-0.1
-0.2
0.40
0.45
0.50
0.55
0.60
0.65
0.70
100 т.о. GC

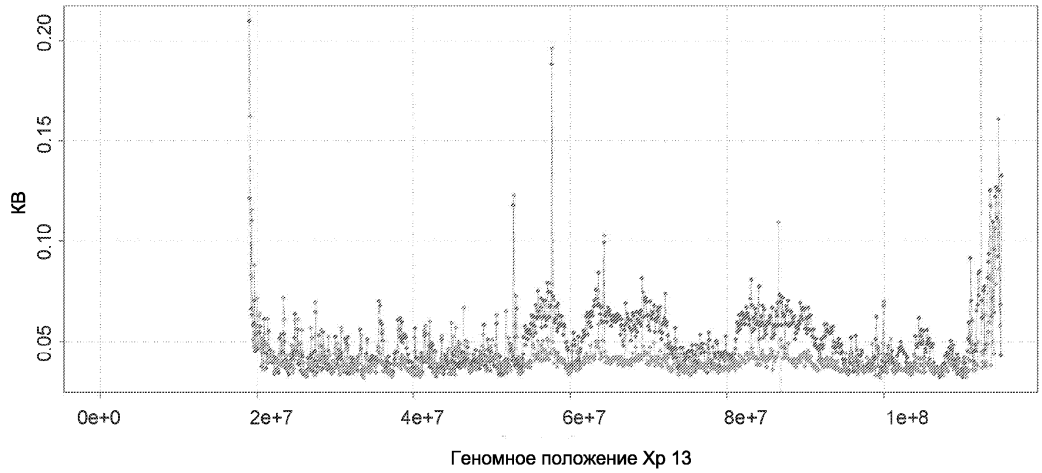
Фиг. 3F


Остаточн. по сравнению с % GC до корректировки
0.2
0.1
0
-0.1
-0.2
0.3
0.4
0.5
0.6
0.7
Остаточн. по сравнению с % GC после корректировки
0.4
0.2
0
-0.2
-0.4
0.3
0.4
0.5
0.6
0.7

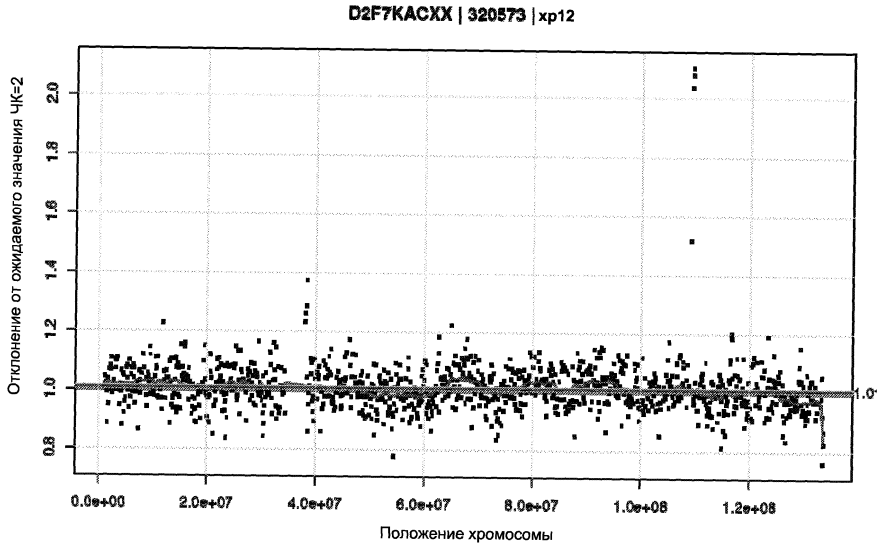
Фиг. 3G



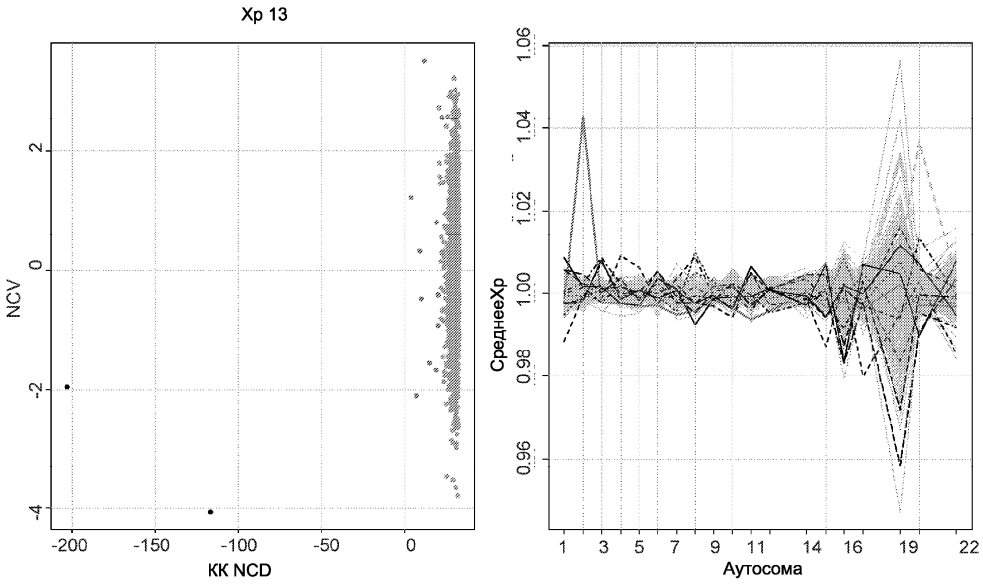
Фиг. 3H



Фиг. 3I



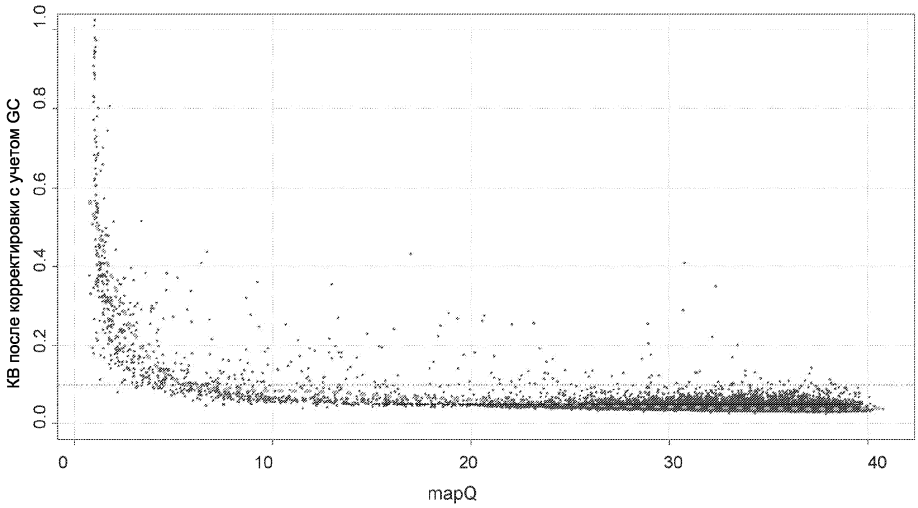
Фиг. 3J


Хр 13
NCV
2
0
-2
-4
-200
-150
-100
-50
0
КК NCD
СреднееХр
1.06
1.04
1.02
1.00
0.98
0.96
1
3
5
7
9
11
14
16
19
22
Аутосома

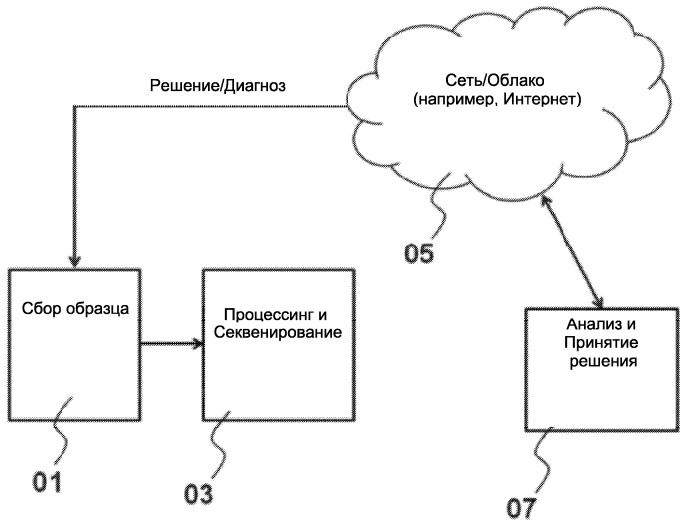
Фиг. 3K


400
Получить маску последовательности
402
Обеспечить геномные считывания множества непораженных обучающих образцов
404
Выровнять геномные считывания с референсным геномом, содержащим последовательность, представляющую интерес, для получения меток обучающей последовательности для обучающих образцов
406
Определить для каждого обучающего образца значение параметра меток обучающей последовательности, расположенных в каждом из множества блоков , на которые разделен референсный геном
408
Определить ожидаемое значение параметра среди обучающих образцов для каждого сектора
410
Скорректировать значение параметра в каждом секторе для каждого обучающего образца в соответствии с ожидаемым перекрытием в каждом секторе, посредством этого получив значение параметра с устраненным глобальным профилем для каждого обучающего образца
412
Скорректировать перекрытия с устраненным глобальным профилем для каждого обучающего образца на основании взаимосвязи между уровнем содержания GC и перекрытием с устраненным глобальным профилем, существующим в каждом обучающем образце, посредством этого получив перекрытия, откорректированные с учетом GC в образце, меток обучающей последовательности для каждого обучающего образца
414
Создать маску последовательности, содержащую немаскированные и маскированные блоки, причем каждому маскированному сектору свойственна характеристика распределения, превышающая порог маскирования

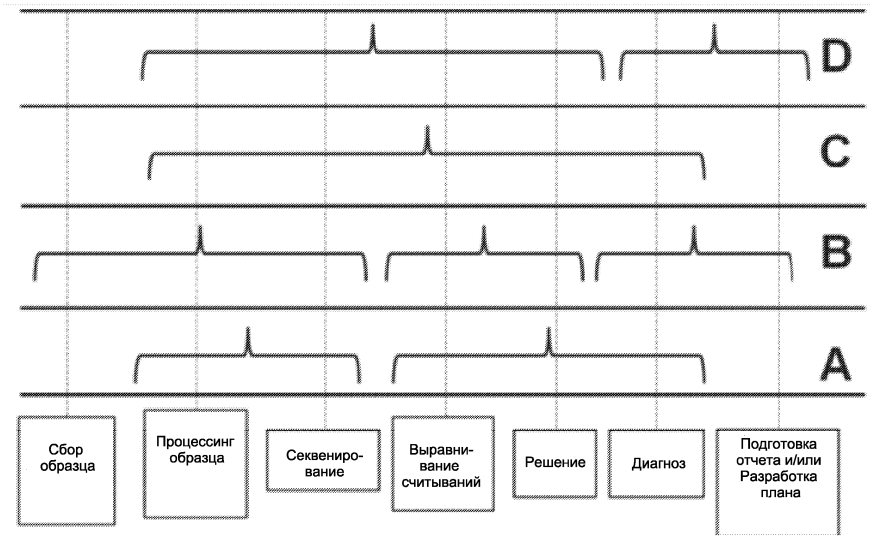
Фиг. 4A



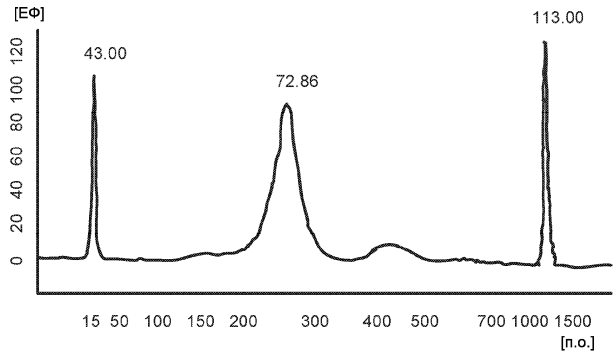
Фиг. 4B



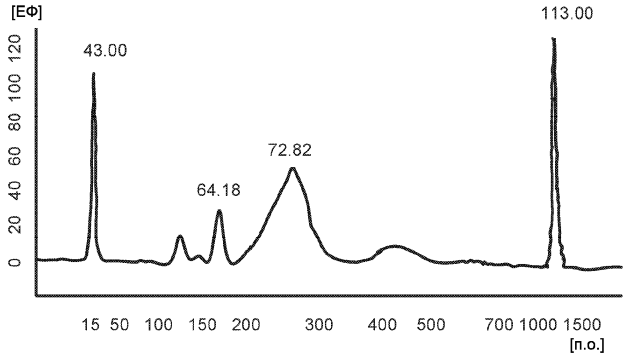
Фиг. 5



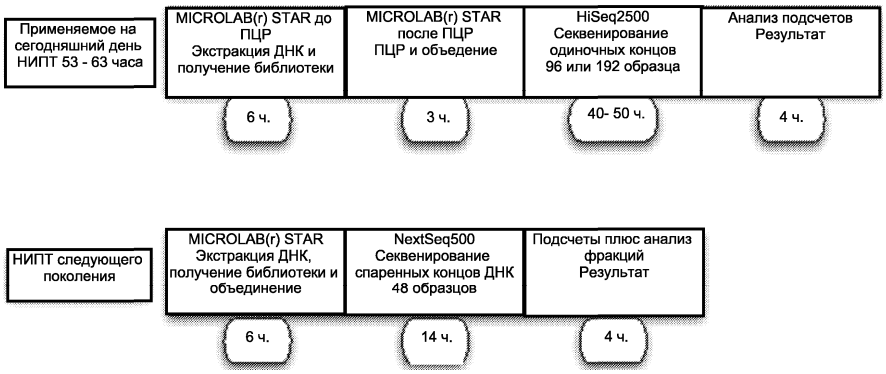
Фиг. 6



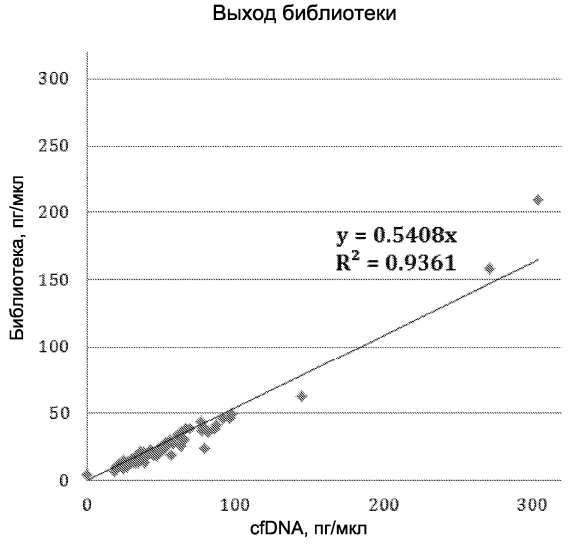
Фиг. 7A



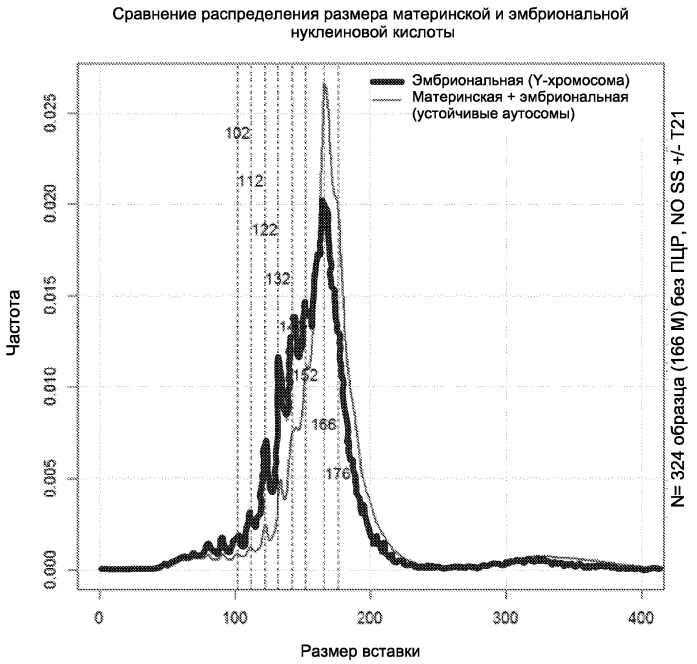
Фиг. 7B



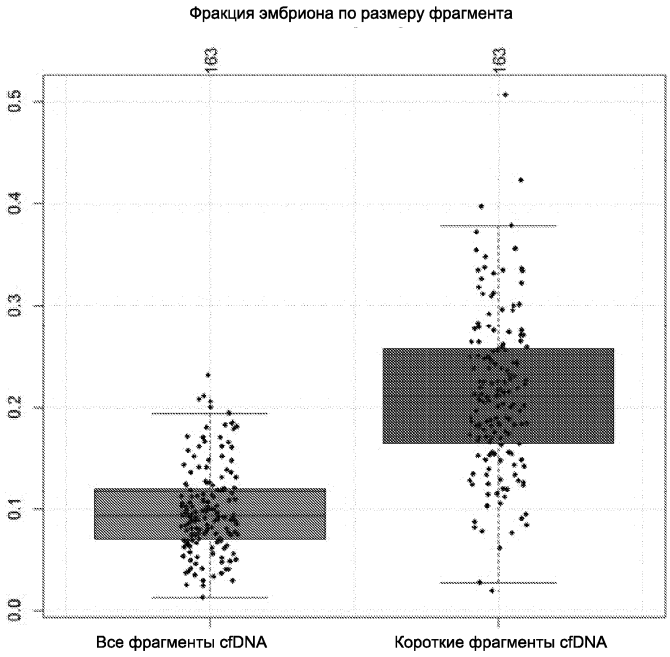
Фиг. 8



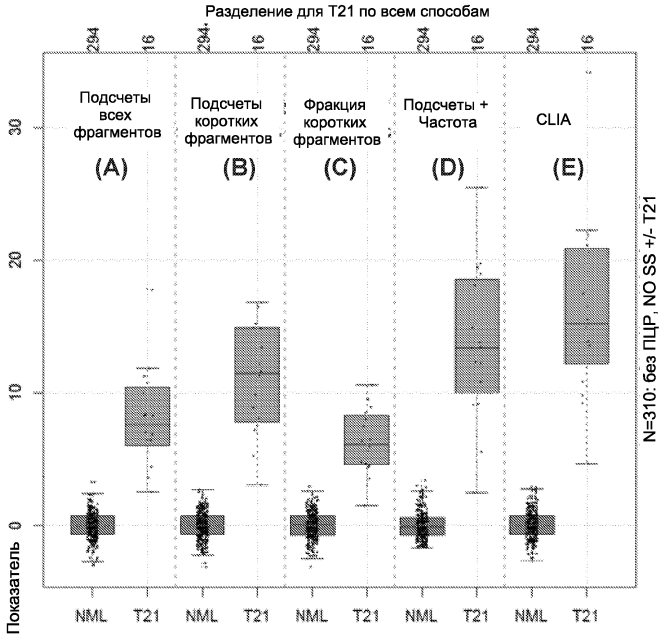
Фиг. 9


Сравнение распределения размера материнской и эмбриональной нуклеиновой кислоты
Эмбриональная (Y-хромосома)
Материнская + эмбриональная (устойчивые аутосомы)
N= 324 образца (166 M) без ПЦР, NO SS +/- T21
Частота
Размер вставки
0.000
0.005
0.010
0.015
0.020
0.025
0
100
200
300
400
102
112
122
132
152
166
176

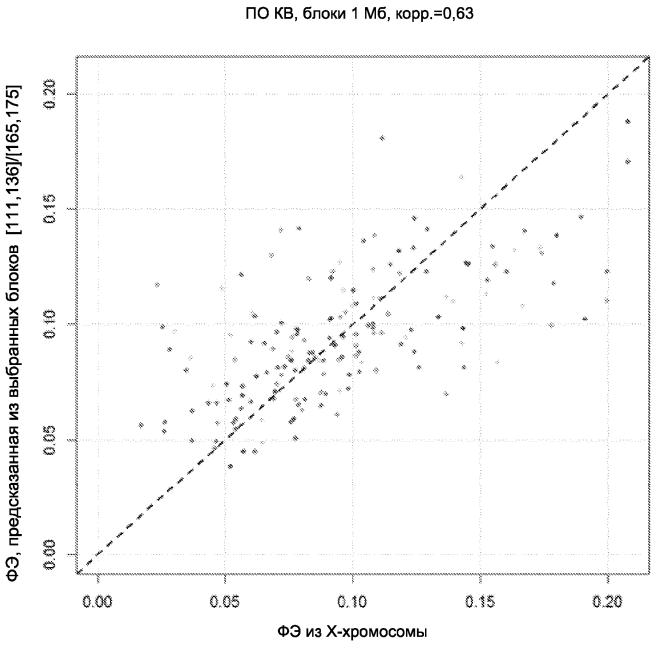
Фиг. 10


Фракция эмбриона по размеру фрагмента
163
163
0.0
0.1
0.2
0.3
0.4
0.5
Все фрагменты cfDNA
Короткие фрагменты cfDNA

Фиг. 11


Разделение для T21 по всем способам
Подсчеты всех фрагментов
(A)
Подсчеты коротких фрагментов
(B)
Фракция коротких фрагментов
(C)
Подсчеты + Частота
(D)
CLIA
(E)
294
16
Показатель
0
10
20
30
NML
T21
N=310: без ПЦР, NO SS +/- T21

Фиг. 12


ПО КВ, блоки 1 Мб, корр.=0,63
ФЭ, предсказанная из выбранных блоков [111,136]/[165,175]
0.00
0.05
0.10
0.15
0.20
ФЭ из X-хромосомы

Фиг. 13