(19) Евразийское патентное ведомство

(21) 202291858 (13) A1

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ЕВРАЗИЙСКОЙ ЗАЯВКЕ

(43) Дата публикации заявки
2022.09.29

(22) Дата подачи заявки
2021.02.05

(51) Int. Cl. ***C12Q 1/68*** (2018.01)
G16B 20/10 (2019.01)

(54) МОЛЕКУЛЯРНЫЕ АНАЛИЗЫ С ИСПОЛЬЗОВАНИЕМ ДЛИННЫХ ВНЕКЛЕТОЧНЫХ ФРАГМЕНТОВ ПРИ БЕРЕМЕННОСТИ

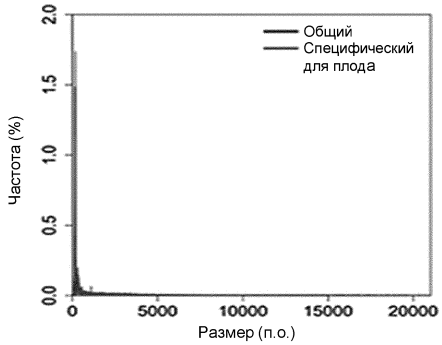
(31) **62/970,634; 63/135,486**
(32) **2020.02.05; 2021.01.08**
(33) **US**
(86) **PCT/CN2021/075394**
(87) **WO 2021/155831 2021.08.12**

(71) Заявитель:
ТЕ ЧАЙНИЗ ЮНИВЕРСИТИ ОВ ГОНКОНГ (CN)

(72) Изобретатель:
Ло Юйк-Мин Деннис, Чиу Росса Вай Квун, Чань Квань Чэ, Цзян Пэйюн, Чэн Сук Хан, Юй Чок Инь, Чхон И Тин, Пэн Вэньлэй (CN)

(74) Представитель:
Фелицына С.Б. (RU)

(57) Способы и системы, описанные в настоящем документе, включают применение длинных фрагментов внеклеточной ДНК для анализа биологического образца от беременного субъекта. Статус метилированных сайтов CpG и однонуклеотидных полиморфизмов (ОНП) часто используется для анализа фрагментов ДНК биологического образца. Сайт CpG и ОНП обычно отделены от ближайшего сайта CpG или ОНП сотнями или тысячами пар оснований. Обнаружение двух или более последовательных сайтов CpG или ОНП на большинстве фрагментов внеклеточной ДНК является маловероятным или невозможным. Фрагменты внеклеточной ДНК длиной более 600 п.о. могут включать множество сайтов CpG и/или ОНП. Наличие множества сайтов CpG и/или ОНП на длинных фрагментах внеклеточной ДНК может обеспечить проведение анализа, в сравнении с использованием только коротких фрагментов внеклеточной ДНК. Длинные фрагменты внеклеточной ДНК можно применять для идентификации ткани происхождения и/или для обеспечения информации о плоде у беременного субъекта женского пола.



202291858 A1

202291858 A1

МОЛЕКУЛЯРНЫЕ АНАЛИЗЫ С ИСПОЛЬЗОВАНИЕМ ДЛИННЫХ ВНЕКЛЕТОЧНЫХ ФРАГМЕНТОВ ПРИ БЕРЕМЕННОСТИ

Перекрестные ссылки на родственные заявки

Настоящая заявка испрашивает приоритет согласно предварительной заявке США № 62/970634, поданной 5 февраля 2020 г., и предварительной заявке США № 63/135486, поданной 8 января 2021 г., полное содержание которых включено в настоящий документ для всех целей.

Уровень техники

Сообщалось, что модальный размер циркулирующей внеклеточной ДНК при беременности составляет приблизительно 166 п.о. (Lo et al. Sci Transl Med. 2010;2:61ra91). Опубликовано очень мало данных о фрагментах размером более 600 п.о. Одним из примеров является работа Amicucci et al., которые сообщили об амплификации с использованием ПЦР фрагмента размером 8 тыс. п.о. из гена основного белка Y2 (*BPY2*) Y-хромосомы из материнской плазмы (Amicucci et al. Clin Chem 2000;40: 301-2). Неизвестно, можно ли распространить такие данные на весь геном. Действительно, существует много проблем при широкомасштабном использовании технологий параллельного секвенирования с коротким ридам, например, при использовании платформы Illumina, для детектирования таких длинных фрагментов ДНК, например, более 600 п.о. (Lo et al. Sci Transl Med. 2010;2:61ra91; Fan et al, Clin Chem. 2010;56:1278-86). Такие проблемы включают: (1) рекомендуемый диапазон размеров для платформы секвенирования Illumina обычно охватывает 100-300 п.о. (De Maio et al. Micob Genom. 2019;5(9)); (2) амплификация ДНК будет задействована в подготовке библиотеки секвенирования (за счет ПЦР) или в генерировании кластера секвенирования за счет мостиковой амплификации в проточной кювете. Такой способ амплификации может благоприятствовать амплификации более коротких фрагментов ДНК отчасти из-за того, что длинные ДНК-матрицы (например, >600 п.о.) потребуют относительно длительного времени для завершения синтеза дочерних цепей по сравнению с короткими ДНК-матрицами (например, <200 п.о.). Таким образом, в течение ограниченного периода времени для этих процессов ПЦР до или во время секвенирования на платформе Illumina те длинные молекулы ДНК, дочерние цепи которых не удалось сгенерировать полностью во время процесса ПЦР, будут недоступны для последующего анализа; (3) длинная молекула ДНК будет иметь больше шансов сформировать вторичные структуры, которые будут препятствовать амплификации; (4) при использовании технологии секвенирования

Illumina длинные молекулы ДНК с большей вероятностью будут вызывать образование кластеров, содержащих более одной клональной молекулы ДНК, по сравнению с короткими молекулами ДНК, поскольку библиотеки денатурируют, разбавляют и диффундируют на двумерной поверхности с последующей мостиковой амплификацией (Head et al. Biotechniques. 2014;56:61-4).

Краткое описание изобретения

Способы и системы, описанные в настоящем документе, включают применение длинных фрагментов внеклеточной ДНК для анализа биологического образца. Применение этих длинных фрагментов внеклеточной ДНК позволяет проводить анализ, который не предусмотрен или невозможен с более короткими фрагментами внеклеточной ДНК. Статус метилированных сайтов CpG и однонуклеотидных полиморфизмов (ОНП) часто используется для анализа фрагментов ДНК биологического образца. Сайт CpG и ОНП обычно отделены от ближайшего сайта CpG или ОНП сотнями или тысячами пар оснований. Длина большинства фрагментов внеклеточной ДНК в биологическом образце обычно менее 200 п.о. В результате обнаружение двух или более последовательных сайтов CpG или ОНП на большинстве фрагментов внеклеточной ДНК является маловероятным или невозможным. Фрагменты внеклеточной ДНК длиной более 200 п.о., включая фрагменты длиннее 600 п.о. или 1 тыс. п.о., могут включать множество сайтов CpG и/или ОНП. Наличие множества сайтов CpG и/или ОНП на длинных фрагментах внеклеточной ДНК может обеспечить более эффективный и/или точный анализ, чем при использовании только коротких фрагментов внеклеточной ДНК. Длинные фрагменты внеклеточной ДНК можно применять для идентификации ткани происхождения и/или для обеспечения информации о плоде у беременного субъекта женского пола. Кроме того, применение длинных фрагментов внеклеточной ДНК для точного анализа образцов беременных женщин является неожиданным, поскольку можно было бы ожидать, что такие длинные фрагменты внеклеточной ДНК имеют преимущественно материнское происхождение. Не ожидалось, что длинные фрагменты внеклеточной ДНК, происходящие от плода, присутствуют в количествах, достаточных для обеспечения информации о плоде.

Длинные фрагменты внеклеточной ДНК с присутствующим ОНП можно применять для определения гаплотипа, унаследованного плодом. Длинные фрагменты внеклеточной ДНК, из-за наличия множества сайтов CpG, могут иметь профиль метилирования, который указывает на ткань происхождения. Кроме того, на длинных фрагментах внеклеточной ДНК могут присутствовать тринуклеотидные повторы и другие повторяющиеся последовательности. Эти повторы можно применять для определения

вероятности генетического нарушения у плода или установления отцовства у плода. Количество длинных фрагментов внеклеточной ДНК можно применять для определения гестационного возраста. Подобным образом, мотивы на конце длинных фрагментов внеклеточной ДНК также можно применять для определения гестационного возраста. Длинные фрагменты внеклеточной ДНК (включая, например, количества, распределение длины, геномные положения, статус метилирования и т.д. таких фрагментов) можно применять для определения нарушения, ассоциированного с беременностью.

Эти и другие варианты реализации настоящего изобретения подробно описаны ниже. Например, другие варианты реализации относятся к системам, устройствам и машиночитаемым носителям, связанным со способами, описанными в настоящем документе.

Лучшее понимание сущности и преимуществ вариантов реализации настоящего изобретения можно получить со ссылкой на следующее подробное описание и прилагаемые чертежи.

Краткое описание чертежей

На фиг. 1A и 1B показано распределение размеров внеклеточной ДНК, определенное в соответствии с вариантами реализации настоящего изобретения. (A) 0-20 тыс. п.о. по линейной шкале, (B) 0-20 тыс. п.о. по логарифмической шкале.

На фиг. 2A и 2B показано распределение размеров внеклеточной ДНК, определенное в соответствии с вариантами реализации настоящего изобретения. (A) 0-5 тыс. п.о. по линейной шкале для оси Y. (B) 0-5 тыс. п.о. по логарифмической шкале для оси Y.

На фиг. 3A и 3B показано распределение размеров внеклеточной ДНК, определенное в соответствии с вариантами реализации настоящего изобретения. (A) 0-400 п.о. по линейной шкале для оси Y. (B) 0-400 п.о. по логарифмической шкале для оси Y.

На фиг. 4A и 4B показано распределение размеров внеклеточной ДНК среди фрагментов, несущих общие аллели (общие) и специфические для плода аллели (специфические для плода), определенное в соответствии с вариантами реализации настоящего изобретения. (A) 0-20 тыс. п.о. по линейной шкале для оси Y. (B) 0-20 тыс. п.о. по логарифмической шкале для оси Y. Синяя линия обозначает фрагменты, несущие общие аллели (преимущественно материнского происхождения), и красная линия обозначает фрагменты, несущие специфические для плода аллели (плацентарного происхождения).

На фиг. 5A и 5B показано распределение размеров внеклеточной ДНК среди фрагментов, несущих общие аллели (общие) и специфические для плода аллели

(специфические для плода), определенное в соответствии с вариантами реализации настоящего изобретения. (A) 0-5 тыс. п.о. по линейной шкале для оси Y. (B) 0-5 тыс. п.о. по логарифмической шкале для оси Y. Синяя линия обозначает фрагменты, несущие общие аллели (преимущественно материнского происхождения), и красная линия обозначает фрагменты, несущие специфические для плода аллели (плацентарного происхождения).

На фиг. 6A и 6B показано распределение размеров внеклеточной ДНК среди фрагментов, несущих общие аллели (общие) и специфические для плода аллели (специфические для плода), определенное в соответствии с вариантами реализации настоящего изобретения. (A) 0-1 тыс. п.о. по линейной шкале для оси Y. (B) 0-1 тыс. п.о. по логарифмической шкале для оси Y. Синяя линия обозначает фрагменты, несущие общие аллели (преимущественно материнского происхождения), и красная линия обозначает фрагменты, несущие специфические для плода аллели (плацентарного происхождения).

На фиг. 7A и 7B показано распределение размеров внеклеточной ДНК среди фрагментов, несущих общие аллели (общие) и специфические для плода аллели (специфические для плода), определенное в соответствии с вариантами реализации настоящего изобретения. (A) 0-400 п.о. по линейной шкале для оси Y. (B) 0-400 п.о. по логарифмической шкале для оси Y. Синяя линия обозначает фрагменты, несущие общие аллели (преимущественно материнского происхождения), и красная линия обозначает фрагменты, несущие специфические для плода аллели (плацентарного происхождения).

На фиг. 8 показаны уровни метилирования отдельной двухцепочечной молекулы ДНК среди фрагментов, несущих специфические для матери аллели и специфические для плода аллели, в соответствии с вариантами реализации настоящего изобретения.

На фиг. 9A и 9B показано (A) аппроксимированное распределение уровней метилирования отдельной двухцепочечной молекулы ДНК среди фрагментов, несущих специфические для матери аллели и специфические для плода аллели, и (B) анализ операционных характеристик приемника (ROC) с использованием уровней метилирования отдельной двухцепочечной молекулы ДНК в соответствии с вариантами реализации настоящего изобретения.

На фиг. 10A и 10B показана корреляция между уровнями метилирования отдельной двухцепочечной молекулы ДНК и размерами фрагментов ДНК плазмы в соответствии с вариантами реализации настоящего изобретения. (A) диапазон размеров от 0 до 20 тыс. п.о. (B) диапазон размеров от 0 до 1 тыс. п.о.

На фиг. 11A и 11B показан пример специфической для плода длинной молекулы

ДНК, идентифицированной в ДНК материнской плазмы беременной женщины в соответствии с вариантами реализации настоящего изобретения. (A) черная полоса обозначает специфическую для плода длинную молекулу ДНК, выравненную с областью в хромосоме 10 референсного генома человека. (B) Подробная иллюстрация генетической и эпигенетической информации, определенной с использованием секвенирования PacBio в соответствии с настоящим изобретением. Основание, выделенное желтым цветом (отмеченное стрелкой), вероятно, обусловлено ошибкой в последовательности, которую можно исправить в некоторых вариантах реализации.

На фиг. 12A и 12B показан пример длинной молекулы ДНК матери, несущей общие аллели, идентифицированной в ДНК материнской плазмы беременной женщины в соответствии с вариантами реализации настоящего изобретения. (A) Черная полоса обозначает специфическую для матери длинную молекулу ДНК, выравненную с областью в хромосоме 6 референса человека. (B) Подробная иллюстрация генетической и эпигенетической информации, определенной с использованием секвенирования PacBio в соответствии с вариантами реализации настоящего изобретения.

На фиг. 13 показано частотное распределение для ДНК из плацентарных клеток (красный) и материнских клеток крови (синий) в соответствии с уровнем метилирования при различных разрешениях от 1 тыс. п.о. до 20 тыс. п.о. в соответствии с вариантами реализации настоящего изобретения.

На фиг. 14A и 14B показано частотное распределение для ДНК из плацентарных клеток (красный) и материнских клеток крови (синий) в соответствии с уровнями метилирования в окнах 16 тыс. п.о. и 24 тыс. п.о. в соответствии с вариантами реализации настоящего изобретения.

На фиг. 15A и 15B показан пример специфической для матери длинной молекулы ДНК, идентифицированной в ДНК материнской плазмы беременной женщины в соответствии с вариантами реализации настоящего изобретения. (A) Черная полоса обозначает специфическую для матери длинную молекулу ДНК, выравненную с областью в хромосоме 8 референса человека. (B) Подробная иллюстрация генетической и эпигенетической информации, определенной с использованием секвенирования PacBio в соответствии с вариантами реализации настоящего изобретения.

На фиг. 16 показана иллюстрация выведения наследования по материнской линии плода в соответствии с вариантами реализации настоящего изобретения.

На фиг. 17 проиллюстрировано определение генетических/эпигенетических нарушений в молекуле ДНК плазмы с информацией о происхождении от матери и происхождении от плода в соответствии с вариантами реализации настоящего

изобретения.

На фиг. 18 проиллюстрирована идентификация аберрантных фрагментов плода в соответствии с вариантами реализации настоящего изобретения.

На фиг. 19A-19G показаны иллюстрации коррекции ошибок генотипирования внеклеточной ДНК с использованием секвенирования PacBio в соответствии с вариантами реализации настоящего изобретения. «.» представляет основание, идентичное референсному основанию в цепи по Уотсону. «,» представляет основание, идентичное референсному основанию в цепи по Крику. «Алфавитная буква» представляет альтернативный аллель, который отличается от референсного аллеля. «*» представляет вставку. «^» представляет делецию.

На фиг. 20 показан способ анализа биологического образца, полученного от субъекта женского пола, беременного плодом, в соответствии с вариантами реализации настоящего изобретения.

На фиг. 21 показан способ анализа биологического образца, полученного от субъекта женского пола, беременного плодом, для определения наследования гаплотипа в соответствии с вариантами реализации настоящего изобретения.

На фиг. 22 показаны профили метилирования для определения ткани происхождения длинной молекулы ДНК в плазме в соответствии с вариантами реализации настоящего изобретения.

На фиг. 23 показана кривая операционных характеристик приемника (ROC) для определения происхождения от матери и происхождения от плода в соответствии с вариантами реализации настоящего изобретения.

На фиг. 24 показаны парные профили метилирования в соответствии с вариантами реализации настоящего изобретения.

На фиг. 25 приведена таблица распределения отобранных маркерных областей в различных хромосомах в соответствии с вариантами реализации настоящего изобретения.

На фиг. 26 приведена таблица классификации молекул ДНК плазмы на основании их профилей метилирования отдельных молекул с использованием различного процентного содержания молекул ДНК лейкоцитарной пленки, имеющих оценку несоответствия более 0,3 в качестве критериев отбора маркерных областей, в соответствии с вариантами реализации настоящего изобретения.

На фиг. 27 показана схема способа применения гаплотипа специфического для плаценты метилирования для определения наследования у плода неинвазивным способом в соответствии с вариантами реализации настоящего изобретения.

На фиг. 28 проиллюстрирован принцип неинвазивного пренатального

детектирования синдрома ломкой Х-хромосомы с применением длинной внеклеточной ДНК в материнской плазме в соответствии с вариантами реализации настоящего изобретения.

На фиг. 29 проиллюстрировано наследование по материнской линии плода на основании профилей метилирования в соответствии с вариантами реализации настоящего изобретения.

На фиг. 30 проиллюстрирован качественный анализ наследования по материнской линии плода с использованием генетической и эпигенетической информации о молекулах ДНК плазмы в соответствии с вариантами реализации настоящего изобретения.

На фиг. 31 проиллюстрирован уровень детектирования качественного анализа наследования по материнской линии плода в масштабе всего генома с использованием генетической и эпигенетической информации о молекулах ДНК плазмы по сравнению с анализом относительной дозы гаплотипа (RHDO) в соответствии с вариантами реализации настоящего изобретения.

На фиг. 32 показана взаимосвязь между уровнем детектирования специфических для отца вариантов в масштабе всего генома и количеством секвенированных молекул ДНК плазмы разного размера, использованных для анализа в соответствии с вариантами реализации настоящего изобретения.

На фиг. 33 показан рабочий процесс для неинвазивного детектирования синдрома ломкой Х-хромосомы в соответствии с вариантами реализации настоящего изобретения.

На фиг. 34 показан профиль метилирования ДНК плазмы по сравнению с профилями метилирования ДНК плаценты и лейкоцитарной пленки в соответствии с вариантами реализации настоящего изобретения.

На фиг. 35 приведена таблица, показывающая распределение сайтов CpG в области размером 500 п.о. в геноме человека в соответствии с вариантами реализации настоящего изобретения.

На фиг. 36 приведена таблица, показывающая распределение сайтов CpG в области размером 1 тыс. п.о. в геноме человека в соответствии с вариантами реализации настоящего изобретения.

На фиг. 37 приведена таблица, показывающая распределение сайтов CpG в области размером 3 тыс. п.о. в геноме человека в соответствии с вариантами реализации настоящего изобретения.

На фиг. 38 приведена таблица, показывающая долевые вклады молекул ДНК из различных тканей в материнской плазме, проанализированных с использованием сопоставления статуса метилирования в соответствии с вариантами реализации

настоящего изобретения.

На фиг. 39A и 39B показана взаимосвязь между плацентарным вкладом и фракцией ДНК плода, выведенная с помощью подхода на основе ОНП в соответствии с вариантами реализации настоящего изобретения.

На фиг. 40 показан способ анализа биологического образца, полученного от субъекта женского пола, беременного плодом, для определения ткани происхождения с использованием анализа профиля метилирования в соответствии с вариантами реализации настоящего изобретения.

На фиг. 41A и 41B показаны распределения размеров молекул внеклеточной ДНК из образцов материнской плазмы первого, второго и третьего триместров в соответствии с вариантами реализации настоящего изобретения.

На фиг. 42 приведена таблица, показывающая долю длинных молекул ДНК плазмы в разных триместрах беременности в соответствии с вариантами реализации настоящего изобретения.

На фиг. 43A и 43B показаны распределения размеров молекул ДНК, охватывающих специфические для плода аллели, из материнской плазмы первого, второго и третьего триместров в соответствии с вариантами реализации настоящего изобретения.

На фиг. 44A и 44B показаны распределения размеров молекул ДНК, охватывающих специфические для матери аллели, из материнской плазмы первого, второго и третьего триместров в соответствии с вариантами реализации настоящего изобретения.

На фиг. 45 приведена таблица доли длинных молекул ДНК плода и матери в плазме в разных триместрах беременности в соответствии с вариантами реализации настоящего изобретения.

На фиг. 46A, 46B и 46C показаны графики долей специфических для плода фрагментов ДНК плазмы, имеющих конкретный диапазон размеров, в разных триместрах в соответствии с вариантами реализации настоящего изобретения.

На фиг. 47A, 47B и 47C показаны графики долевого состава оснований на 5′-конце молекул внеклеточной ДНК из материнской плазмы первого, второго и третьего триместров для диапазона размеров фрагментов от 0 до 3 тыс. п.о. в соответствии с вариантами реализации настоящего изобретения.

На фиг. 48 приведена таблица долей концевых нуклеотидных оснований среди коротких и длинных молекул внеклеточной ДНК из материнской плазмы первого, второго и третьего триместров в соответствии с вариантами реализации настоящего изобретения.

На фиг. 49 приведена таблица долей концевых нуклеотидных оснований среди

коротких и длинных молекул внеклеточной ДНК, охватывающих специфический для плода аллель, из материнской плазмы первого, второго и третьего триместров в соответствии с вариантами реализации настоящего изобретения.

На фиг. 50 приведена таблица долей концевых нуклеотидных оснований среди коротких и длинных молекул внеклеточной ДНК, охватывающих специфический для матери аллель, из материнской плазмы первого, второго и третьего триместров в соответствии с вариантами реализации настоящего изобретения.

На фиг. 51 проиллюстрирован иерархический кластерный анализ коротких и длинных молекул внеклеточной ДНК с использованием 256 концевых мотивов в соответствии с вариантами реализации настоящего изобретения.

На фиг. 52A и 52B показан анализ основных компонентов профилей 4-членных концевых мотивов в соответствии с вариантами реализации настоящего изобретения.

На фиг. 53 приведена таблица 25 концевых мотивов с самыми высокими частотами среди коротких молекул ДНК плазмы из материнской плазмы первого триместра в соответствии с вариантами реализации настоящего изобретения.

На фиг. 54 приведена таблица 25 концевых мотивов с самыми высокими частотами среди коротких молекул ДНК плазмы из материнской плазмы второго триместра в соответствии с вариантами реализации настоящего изобретения.

На фиг. 55 приведена таблица 25 концевых мотивов с самыми высокими частотами среди коротких молекул ДНК плазмы из материнской плазмы третьего триместра в соответствии с вариантами реализации настоящего изобретения.

На фиг. 56 приведена таблица 25 концевых мотивов с самыми высокими частотами среди длинных молекул ДНК плазмы из материнской плазмы первого триместра в соответствии с вариантами реализации настоящего изобретения.

На фиг. 57 приведена таблица 25 концевых мотивов с самыми высокими частотами среди длинных молекул ДНК плазмы из материнской плазмы второго триместра в соответствии с вариантами реализации настоящего изобретения.

На фиг. 58 приведена таблица 25 концевых мотивов с самыми высокими частотами среди длинных молекул ДНК плазмы из материнской плазмы третьего триместра в соответствии с вариантами реализации настоящего изобретения.

На фиг. 59A, 59B и 59C показаны диаграммы рассеяния частот мотивов для 16 мотивов NNXY среди коротких и длинных молекул ДНК плазмы из материнской плазмы в (A) первом триместре, (B) втором триместре, и (C) третьем триместре в соответствии с вариантами реализации настоящего изобретения.

На фиг. 60 показан способ анализа биологического образца, полученного от

субъекта женского пола, беременного плодом, для определения гестационного возраста в соответствии с вариантами реализации настоящего изобретения.

На фиг. 61 показан способ анализа биологического образца, полученного от субъекта женского пола, беременного плодом, для классификации вероятности нарушения, ассоциированного с беременностью, в соответствии с вариантами реализации настоящего изобретения.

На фиг. 62 приведена таблица, показывающая клиническую информацию о четырех случаях преэклампсии в соответствии с вариантами реализации настоящего изобретения.

На фиг. 63A-63D показаны графики распределения размеров молекул внеклеточной ДНК из преэклamптических и нормотензивных образцов материнской плазмы третьего триместра в соответствии с вариантами реализации настоящего изобретения.

На фиг. 64A-64D показаны графики распределения размеров молекул внеклеточной ДНК из преэклamптических и нормотензивных образцов материнской плазмы третьего триместра беременности в соответствии с вариантами реализации настоящего изобретения.

На фиг. 65A-65D показаны графики распределения размеров молекул ДНК, охватывающих специфические для плода аллели, из преэклamптических и нормотензивных образцов материнской плазмы третьего триместра в соответствии с вариантами реализации настоящего изобретения.

На фиг. 66A-66D показаны графики распределения размеров молекул ДНК, охватывающих специфические для плода аллели, из преэклamптических и нормотензивных образцов материнской плазмы третьего триместра в соответствии с вариантами реализации настоящего изобретения.

На фиг. 67A-67D показаны графики распределения размеров молекул ДНК, охватывающих специфические для матери аллели, из преэклamптических и нормотензивных образцов материнской плазмы третьего триместра в соответствии с вариантами реализации настоящего изобретения.

На фиг. 68A-68D показаны графики распределения размеров молекул ДНК, охватывающих специфические для матери аллели, из преэклamптических и нормотензивных образцов материнской плазмы третьего триместра в соответствии с вариантами реализации настоящего изобретения.

На фиг. 69A и 69B показаны графики доли коротких молекул ДНК, охватывающих специфические для плода аллели и специфические для матери аллели, в

преэкламптических и нормотензивных образцах материнской плазмы, секвенированных с использованием секвенирования PacBio SMRT, в соответствии с вариантами реализации настоящего изобретения.

На фиг. 70A и 70B показаны графики доли коротких молекул ДНК в преэкламптических и нормотензивных образцах материнской плазмы, секвенированных с использованием секвенирования PacBio SMRT и секвенирования Illumina, в соответствии с вариантами реализации настоящего изобретения.

На фиг. 71 показан график соотношений размеров, который показывает относительные доли коротких и длинных молекул ДНК в преэкламптических и нормотензивных образцах материнской плазмы, секвенированных с использованием секвенирования PacBio SMRT, в соответствии с вариантами реализации настоящего изобретения.

На фиг. 72A-72D показана доля различных концов молекул ДНК плазмы в преэкламптических и нормотензивных образцах материнской плазмы, секвенированных с использованием секвенирования PacBio SMRT, в соответствии с вариантами реализации настоящего изобретения.

На фиг. 73 показан иерархический кластерный анализ ДНК преэкламптических и нормотензивных образцов материнской плазмы третьего триместра с использованием частоты молекул ДНК плазмы с каждым из четырех типов концов фрагментов (первый нуклеотид на 5′-конце каждой цепи), а именно C-концом, G-концом, T-концом и A-концом, в соответствии с вариантами реализации настоящего изобретения.

На фиг. 74 показан иерархический кластерный анализ ДНК преэкламптических и нормотензивных образцов материнской плазмы третьего триместра с использованием 16 динуклеотидных мотивов XYNN (динуклеотидная последовательность первого и второго нуклеотидов с 5′-конца) в соответствии с вариантами реализации настоящего изобретения.

На фиг. 75 показан иерархический кластерный анализ ДНК преэкламптических и нормотензивных образцов материнской плазмы третьего триместра с использованием 16 динуклеотидных мотивов NNXY (динуклеотидная последовательность третьего и четвертого нуклеотидов с 5′-конца) в соответствии с вариантами реализации настоящего изобретения.

На фиг. 76 показан иерархический кластерный анализ ДНК преэкламптических и нормотензивных образцов материнской плазмы третьего триместра с использованием 256 четырехнуклеотидных мотивов (динуклеотидная последовательность с первого по четвертый нуклеотиды с 5′-конца) в соответствии с вариантами реализации настоящего изобретения.

На фиг. 77A-77D показан вклад Т-клеток среди четырех типов концов фрагментов ДНК преэклamптических и нормотензивных образцов материнской плазмы в соответствии с вариантами реализации настоящего изобретения.

На фиг. 78 показан способ анализа биологического образца, полученного от субъекта женского пола, беременного плодом, для определения вероятности нарушения, ассоциированного с беременностью, в соответствии с вариантами реализации настоящего изобретения.

На фиг. 79 показана иллюстрация выведения наследования по материнской линии плода для заболеваний, связанных с повторами, в соответствии с вариантами реализации настоящего изобретения.

На фиг. 80 показана иллюстрация выведения наследования по отцовской линии плода для заболеваний, связанных с повторами, в соответствии с вариантами реализации настоящего изобретения.

На фиг. 81, 82 и 83 приведены таблицы, показывающие примеры заболеваний, связанных с распространением повторов.

На фиг. 84 приведена таблица, показывающая примеры детектирования распространения повторов у плода и определения ассоциированного с повторами метилирования в соответствии с вариантами реализации настоящего изобретения.

На фиг. 85 показан способ анализа биологического образца, полученного от субъекта женского пола, беременного плодом, для определения вероятности генетического нарушения у плода в соответствии с вариантами реализации настоящего изобретения.

На фиг. 86 показан способ анализа биологического образца, полученного от субъекта женского пола, беременного плодом, для определения отцовства в соответствии с вариантами реализации настоящего изобретения.

На фиг. 87 показаны профили метилирования для двух типичных молекул ДНК плазмы после отбора по размеру.

На фиг. 88 приведена таблица с информацией секвенирования для образцов с отбором по размеру и без него в соответствии с вариантами реализации настоящего изобретения.

На фиг. 89A и 89B показаны графики размерных профилей ДНК плазмы для образцов с отбором по размеру на основе гранул и без него в соответствии с вариантами реализации настоящего изобретения.

На фиг. 90A и 90B показаны размерные профили молекул ДНК плода и матери в образце с отбором по размеру в соответствии с вариантами реализации настоящего

изобретения.

На фиг. 91 приведена статистическая таблица количества молекул ДНК плазмы, несущих информативные ОНП, среди образцов с отбором по размеру и без него в соответствии с вариантами реализации настоящего изобретения.

На фиг. 92 приведена таблица уровня метилирования в образцах ДНК плазмы, отобранных по размеру и без отбора по размеру, в соответствии с вариантами реализации настоящего изобретения.

На фиг. 93 приведена таблица уровня метилирования в специфических для матери или специфических для плода молекулах внеклеточной ДНК в соответствии с вариантами реализации настоящего изобретения.

На фиг. 94 приведена таблица 10 основных концевых мотивов в образцах с отбором по размеру и без него в соответствии с вариантами реализации настоящего изобретения.

На фиг. 95 показан график операционных характеристик приемника (ROC), показывающий, что длинные молекулы ДНК плазмы повышают эффективность анализа ткани происхождения в соответствии с вариантами реализации настоящего изобретения.

На фиг. 96 проиллюстрирован принцип нанопорового секвенирования молекул ДНК плазмы в соответствии с вариантами реализации настоящего изобретения.

На фиг. 97 приведена таблица процентного содержания молекул ДНК плазмы в конкретном диапазоне размеров и их соответствующих уровней метилирования в соответствии с вариантами реализации настоящего изобретения.

На фиг. 98 показан график распределения размеров и профилей метилирования по разным размерам в соответствии с вариантами реализации настоящего изобретения.

На фиг. 99 приведена таблица фракции ДНК плода, определенной с использованием нанопорового секвенирования в соответствии с вариантами реализации настоящего изобретения.

На фиг. 100 приведена таблица уровней метилирования среди специфических для плода и специфических для матери молекул ДНК в соответствии с вариантами реализации настоящего изобретения.

На фиг. 101 приведена таблица процентного содержания молекул ДНК плазмы в конкретном диапазоне размеров и их соответствующих уровней метилирования для молекул ДНК плода и матери в соответствии с вариантами реализации настоящего изобретения.

На фиг. 102A и 102B показаны графики распределения размеров молекул ДНК плода и матери, определенные с помощью нанопорового секвенирования в соответствии с

вариантами реализации настоящего изобретения.

На фиг. 103 показан график, показывающий различие в уровнях метилирования между молекулами ДНК плода и матери на основе одного информативного ОНП и двух информативных ОНП в соответствии с вариантами реализации настоящего изобретения.

На фиг. 104 приведена таблица различия в уровнях метилирования между молекулами ДНК плода и матери в соответствии с вариантами реализации настоящего изобретения.

На фиг. 105 проиллюстрирована система измерения в соответствии с вариантами реализации настоящего изобретения.

На фиг. 106 показана компьютерная система в соответствии с вариантами реализации настоящего изобретения.

Термины

«*Ткань*» соответствует группе клеток, которые группируются как функциональная единица у беременного субъекта или ее плода. В одной ткани можно найти более одного типа клеток. Различные типы ткани могут состоять из разных типов клеток (например, гепатоцитов, альвеолярных клеток или клеток крови), но также могут соответствовать ткани из разных организмов (мать в сравнении с плодом; ткани у беременного субъекта, перенесшего трансплантацию; ткани беременного организма или ее плода, которые инфицированы микроорганизмом или вирусом). «Референсные ткани» могут соответствовать тканям, используемым для определения тканеспецифических уровней метилирования. Для определения тканеспецифического уровня метилирования для этого типа ткани может быть использовано множество образцов ткани одного и того же типа от разных беременных индивидуумов или их плодов.

«*Биологический образец*» относится к любому образцу, взятому у беременного субъекта (например, человека (или другого животного), такого как беременная женщина, индивидуум с нарушением или беременный индивидуум, у которого подозревают наличие нарушения, беременный реципиент трансплантата органа или беременный субъект, у которого подозревают наличие патологического процесса, затрагивающего орган (например, сердце при инфаркте миокарда или головной мозг при инсульте, или кроветворную систему при анемии), и содержит одну или более молекул нуклеиновой кислоты, представляющих интерес. Биологический образец может представлять собой физиологическую жидкость, такую как кровь, плазма, сыворотка, моча, вагинальная жидкость, жидкости для промывания влагалища, плевральная жидкость, асцитная жидкость, спинномозговая жидкость, слюна, пот, слезы, мокрота, жидкость бронхоальвеолярного лаважа, выделяемая жидкость из соска, аспирационная жидкость из

различных частей тела (например, щитовидной железы, молочной железы), внутриглазные жидкости (например, водянистая влага) и т.д. Также можно использовать образцы фекалий. Согласно различным вариантам реализации большая часть ДНК в биологическом образце, который был обогащен внеклеточной ДНК (например, образец плазмы, полученный с помощью протокола центрифугирования), может быть внеклеточной, например, более 50%, 60%, 70%, 80%, 90%, 95% или 99% ДНК может быть внеклеточной. Протокол центрифугирования может включать, например, 3000 g x 10 минут, получение жидкой части и повторное центрифугирование, например, при 30000 g в течение еще 10 минут для удаления остаточных клеток. В качестве части анализа биологического образца может быть проанализировано статистически значимое количество молекул внеклеточной ДНК (например, для обеспечения точного измерения) для биологического образца. Согласно некоторым вариантам реализации анализируют по меньшей мере 1000 молекул внеклеточной ДНК. Согласно другим вариантам реализации может быть проанализировано по меньшей мере 10000 или 50000, или 100000, или 500000, или 1000000, или 5000000 молекул внеклеточной ДНК или большее количество. Может быть проанализировано по меньшей мере такое же количество ридов последовательности.

«*Рид последовательности*» относится к цепи нуклеотидов, секвенированной из любой части или всей молекулы нуклеиновой кислоты. Например, рид последовательности может представлять собой короткую цепь нуклеотидов (например, 20-150 нуклеотидов), секвенированную из фрагмента нуклеиновой кислоты, короткую цепь нуклеотидов на одном или обоих концах фрагмента нуклеиновой кислоты, или секвенирование всего фрагмента нуклеиновой кислоты, который существует в биологическом образце. Рид последовательности может быть получен различными способами, например, с использованием методик секвенирования или с использованием зондов, например, гибридизационные чипы или зонды для захвата, которые могут использоваться в микрочипах, или методик амплификации, таких как полимеразная цепная реакция (ПЦР), или линейная амплификация с использованием одного праймера, или изотермическая амплификация. В качестве части анализа биологического образца может быть проанализировано статистически значимое количество ридов последовательности, например, может быть проанализировано по меньшей мере 1000 ридов последовательности. В качестве другого примера, может быть проанализировано по меньшей мере 10000 или 50000, или 100000, или 500000, или 1000000, или 5000000 ридов последовательности или большее количество.

«*Сайт*» (также называемый «*геномным сайтом*») соответствует одному сайту,

который может представлять собой одно положение основания или группу соотносящихся положений оснований, например, сайт CpG или большую группу соотносящихся положений оснований. «Локус» может соответствовать области, которая включает множество сайтов. Локус может включать только один сайт, что может сделать локус эквивалентным сайту в этом контексте.

«*Статус метилирования*» относится к состоянию метилирования в определенном сайте. Например, сайт может быть метилированным, неметилированным или, в некоторых случаях, неопределенным.

«*Индекс метилирования*» для каждого геномного сайта (например, сайта CpG) может относиться к доле фрагментов ДНК (например, как определено из ридов последовательности или зондов), показывающих метилирование в данном сайте, по сравнению с общим количеством ридов, охватывающих данный сайт. «Рид» может соответствовать информации (например, статусу метилирования в сайте), полученной из фрагмента ДНК. Рид можно получить с использованием реагентов (например, праймеров или зондов), которые преимущественно гибридизуются с фрагментами ДНК с конкретным статусом метилирования в одном или более сайтах. Обычно такие реагенты применяются после обработки с помощью способа, который дифференцировано модифицирует или дифференцировано распознает молекулы ДНК в зависимости от их статуса метилирования, например, бисульфитное преобразование, или чувствительный к метилированию рестрикционный фермент, или белки, связывающие метилированные молекулы, или антитела к метилцитозину, или методики одномолекулярного секвенирования (например, одномолекулярное секвенирование в реальном времени и нанопоровое секвенирование (например, от Oxford Nanopore Technologies)), которые распознают метилцитозины и гидроксиметилцитозины.

«*Плотность метилирования*» области может относиться к количеству ридов в сайтах в пределах области, показывающих метилирование, разделенному на общее количество ридов, охватывающих сайты в этой области. Сайты могут иметь определенные характеристики, например, могут представлять собой сайты CpG. Таким образом, «плотность метилирования CpG» области может относиться к количеству ридов, показывающих метилирование CpG, разделенному на общее количество ридов, охватывающих сайты CpG в данной области (например, конкретный сайт CpG, сайты CpG в пределах CpG-островка или большей области). Например, плотность метилирования для каждой группы из 100 тыс.п.о. в геноме человека может быть определена из общего количества цитозинов, не преобразованных после обработки бисульфитом (что соответствует метилированному цитозину), в сайтах CpG, как доля всех сайтов CpG,

охваченных ридами последовательности, картированными в области из 100 тыс. п.о. Этот анализ также может быть выполнен для других размеров группы, например, 500 п.о., 5 тыс. п.о., 10 тыс. п.о., 50 тыс. п.о. или 1 млн. п.о. и т.д. Область может представлять собой весь геном или хромосому, или часть хромосомы (например, плечо хромосомы). Индекс метилирования сайта CpG аналогичен плотности метилирования для области, когда область включает только данный сайт CpG. «Доля метилированных цитозинов» может относиться к количеству цитозиновых сайтов, «С», которые, как показано, являются метилированными (например, не подверглись преобразованию после бисульфитного преобразования) по отношению к общему количеству проанализированных остатков цитозина, т.е. включая цитозины вне контекста CpG, в области. Индекс метилирования, плотность метилирования, число молекул, метилированных в одном или более сайтах, и доля молекул, метилированных (например, цитозинов) в одном или более сайтах, являются примерами «*уровней метилирования*». Помимо бисульфитного преобразования, другие способы, известные специалистам в данной области техники, могут быть использованы для исследования статуса метилирования молекул ДНК, включая, но не ограничиваясь перечисленными, ферменты, чувствительные к статусу метилирования (например, чувствительные к метилированию рестрикционные ферменты), белки, связывающие метилированные молекулы, одномолекулярное секвенирование с использованием платформы, чувствительной к статусу метилирования (например, нанопоровое секвенирование (Schreiber et al. Proc Natl Acad Sci 2013; 110: 18910-18915) и одномолекулярное секвенирование в реальном времени (например, от Pacific Biosciences) (Flusberg et al. Nat Methods 2010; 7: 461-465)).

«*Метилом*» обеспечивает меру степени метилирования ДНК во множестве сайтов или локусов в геноме. Метилом может соответствовать всему геному, значительной части генома или относительно небольшой части(частям) генома.

«*Профиль метилирования*» включает информацию, относящуюся к метилированию ДНК или РНК для множества сайтов или областей. Информация, относящаяся к метилированию ДНК, может включать, но не ограничивается перечисленными, индекс метилирования сайта CpG, плотность метилирования (сокращенно ПМ) сайтов CpG в области, распределение сайтов CpG по непрерывной области, профиль или уровень метилирования для каждого отдельного сайта CpG в пределах области, содержащей более одного сайта CpG, и метилирование не-CpG. Согласно одному варианту реализации профиль метилирования может включать профиль метилирования или неметилирования более чем одного типа основания (например, цитозина или аденина). Профиль метилирования значительной части генома можно считать эквивалентным метилому.

«Метилирование ДНК» в геномах млекопитающих обычно относится к добавлению метильной группы к 5′-углероду остатков цитозина (т.е. 5-метилцитозины) среди динуклеотидов CpG. Метилирование ДНК может происходить в цитозинах в других контекстах, например, CHG и CHH, где H представляет собой аденин, цитозин или тимин. Метилирование цитозина также может быть представлено в форме 5-гидроксиметилцитозина. Также сообщалось о нецитозиновом метилировании, таком как N^6-метиладенин.

«*Профиль метилирования*» относится к порядку метилированных и неметилированных оснований. Например, профиль метилирования может представлять собой порядок метилированных оснований на отдельной цепи ДНК, отдельной двухцепочечной молекуле ДНК или другом типе молекулы нуклеиновой кислоты. Например, три последовательных сайта CpG могут иметь любой из следующих профилей метилирования: UUU, MMM, UMM, UMU, UUM, MUM, MUU или MMU, где «U» обозначает неметилированный сайт, а «M» обозначает метилированный сайт. При расширении этой концепции до модификаций оснований, которые включают, но не ограничиваются метилированием, может использоваться термин «*профиль модификации*», который относится к порядку модифицированных и немодифицированных оснований. Например, профиль модификации может представлять собой порядок модифицированных оснований на отдельной цепи ДНК, отдельной двухцепочечной молекуле ДНК или другом типе молекулы нуклеиновой кислоты. Например, три последовательных, потенциально модифицируемых сайта могут иметь любой из следующих профилей метилирования: UUU, MMM, UMM, UMU, UUM, MUM, MUU или MMU, где «U» обозначает немодифицированный сайт, а «M» обозначает модифицированный сайт. Один из примеров модификации основания, не основанной на метилировании, представляет собой окислительные изменения, например, как в 8-оксогуанине.

Термины «*гиперметилированный*» и «*гипометилированный*» могут относиться к плотности метилирования отдельной молекулы ДНК, измеренной по уровню метилирования такой отдельной молекулы, например, к количеству метилированных оснований или нуклеотидов в молекуле, разделенному на общее количество метилируемых оснований или нуклеотидов в такой молекуле. Гиперметилированная молекула представляет собой молекулу, в которой уровень метилирования отдельной молекулы находится на уровне или выше порога, который может определяться от применения к применению. Порог может составлять 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% или 95%. Гипометилированная молекула представляет собой молекулу, в которой уровень метилирования отдельной молекулы находится на уровне или ниже

порога, который может определяться от применения к применению, и который может изменяться от применения к применению. Порог может составлять 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% или 95%.

Термины «гиперметилированный» и «гипометилированный» также могут относиться к уровню метилирования популяции молекул ДНК, измеренному на основании уровней метилирования множества молекул для этих молекул. Гиперметилированная популяция молекул представляет собой ту, в которой уровень метилирования множества молекул находится на уровне или выше порога, который может определяться от применения к применению, и который может изменяться от применения к применению. Порог может составлять 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% или 95%. Гипометилированная популяция молекул представляет собой ту, в которой уровень метилирования множества молекул находится на уровне или ниже порога, который может определяться от применения к применению. Порог может составлять 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, и 95%. Согласно одному варианту реализации популяция молекул может быть выравнена с одной или более выбранными геномными областями. Согласно одному варианту реализации выбранная геномная область(области) может быть связана с заболеванием, таким как генетическое нарушение, нарушение импринтинга, нарушение обмена веществ или неврологическое нарушение. Выбранная геномная область(области) может иметь длину 50 нуклеотидов (нт.), 100 нт., 200 нт., 300 нт., 500 нт., 1000 нт., 2 тыс. нт., 5 тыс. нт., 10 тыс. нт., 20 тыс. нт., 30 тыс. нт., 40 тыс. нт., 50 тыс. нт., 60 тыс. нт., 70 тыс. нт., 80 тыс. нт., 90 тыс. нт., 100 тыс. нт., 200 тыс. нт., 300 тыс. нт., 400 тыс. нт., 500 тыс. нт. или 1 млн. нт.

Термин «*глубина секвенирования*» относится к количеству раз охвата локуса ридом последовательности, выравненным с локусом. Локус может быть размером с нуклеотид или размером с плечо хромосомы, или размером с весь геном. Глубина секвенирования может быть выражена как 50x, 100x и т.д., где «x» относится к количеству раз охвата локуса ридом последовательности. Глубина секвенирования также может применяться к множеству локусов или ко всему геному, и в этом случае x может относиться к среднему количеству раз секвенирования локуса или гаплоидного генома, или всего генома, соответственно. Сверхглубокое секвенирование может относиться по меньшей мере к 100x глубине секвенирования.

«*Калибровочный образец*» может соответствовать биологическому образцу, в котором относительная концентрация клинически значимой ДНК (например, фракции тканеспецифической ДНК) известна или определена с помощью метода калибровки, например, с использованием аллеля, специфического для ткани, такого как при

трансплантации у беременного субъекта, при которой аллель, присутствующий в геноме донора, но отсутствующий в геноме реципиента, может быть использован в качестве маркера для трансплантируемого органа. В качестве другого примера калибровочный образец может соответствовать образцу, в котором можно определить концевые мотивы. Калибровочный образец можно использовать для обеих целей.

«*Калибровочная точка данных*» включает «*калибровочное значение*» и измеренную или известную относительную концентрацию клинически значимой ДНК (например, ДНК конкретного типа ткани). Калибровочное значение может быть определено из относительных частот (например, совокупное значение), определенных для калибровочного образца, для которого известна относительная концентрация клинически значимой ДНК. Калибровочные точки данных могут быть определены различными способами, например, как дискретные точки или как калибровочная функция (также называемая калибровочной кривой или калибровочной поверхностью). Калибровочная функция может быть получена в результате дополнительного математического преобразования калибровочных точек данных.

«*Степень разделения*» соответствует разности или соотношению, включающему два значения, например, два относительных вклада или два уровня метилирования. Степень разделения может представлять собой простую разность или соотношение. Например, прямое соотношение x/y, а также x/(x + y) представляет собой степень разделения. Степень разделения может включать другие коэффициенты, например, множители. В качестве других примеров можно использовать разность или соотношение функций значений, например, разность или соотношение натуральных логарифмов (ln) двух значений. Степень разделения может включать разность и соотношение.

«*Степень разделения*» и «*совокупное значение*» (например, относительные частоты) являются двумя примерами параметра (также называемого метрикой), который обеспечивает показатель образца, который варьируется между различными классификациями (состояниями) и, таким образом, может использоваться для определения различных классификаций. Совокупное значение может представлять собой степень разделения, например, когда получают разность между набором относительных частот образца и референсным набором относительных частот, как это может быть сделано при кластеризации.

Термин «*классификация*» в контексте настоящего документа относится к любому числу(числам) или другому символу(символам), которые связаны с конкретным свойством образца. Например, символ «+» (или слово «положительный») может означать, что образец классифицируется как имеющий делеции или амплификации. Классификация

может быть двоичной (например, положительной или отрицательной) или может иметь несколько уровней классификации (например, шкала от 1 до 10 или от 0 до 1).

Термин «*параметр*» в контексте настоящего документа означает числовое значение, которое характеризует набор количественных данных и/или числовое соотношение между наборами количественных данных. Например, соотношение (или функция соотношения) между первым количеством первой последовательности нуклеиновой кислоты и вторым количеством второй последовательности нуклеиновой кислоты представляет собой параметр.

Термин «*размерный профиль*» обычно относится к размерам фрагментов ДНК в биологическом образце. Размерный профиль может представлять собой гистограмму, обеспечивающую распределение некоторого количества фрагментов ДНК по множеству размеров. Различные статистические параметры (также называемые размерными параметрами или просто параметром) могут использоваться, чтобы отличить один размерный профиль от другого. Один параметр представляет собой процентное содержание фрагмента ДНК конкретного размера или диапазона размеров по отношению ко всем фрагментам ДНК или по отношению к фрагментам ДНК другого размера или диапазона.

Термины «*отсечение*» и «*порог*» относятся к заранее определенным числам, используемым в процессе. Например, размер отсечки может относиться к размеру, при превышении которого фрагменты исключают. Пороговое значение может представлять собой значение, выше или ниже которого применяется конкретная классификация. Любой из этих терминов может использоваться в любом из этих контекстов. Отсечение или порог может представлять собой «референсное значение» или может быть получен из референсного значения, которое является типичным для конкретной классификации или устанавливает различие между двумя или более классификациями. Такое референсное значение может быть определено различными способами, как будет понятно специалисту в данной области техники. Например, метрики могут быть определены для двух разных когорт субъектов с разными известными классификациями, и референсное значение может быть выбрано как типичное для одной классификации (например, среднее) или значение, которое находится между двумя кластерами метрик (например, выбрано для получения желаемой чувствительности и специфичности). В качестве другого примера, референсное значение может быть определено на основании статистических анализов или моделирований образцов. Конкретное значение для отсечки, порога, референса и т.д. может быть определено на основании желаемой точности (например, чувствительности и специфичности).

«*Связанное с беременностью нарушение*» включает любое нарушение, характеризующееся патологическими относительными уровнями экспрессии генов в ткани матери и/или плода или патологическими клиническими характеристиками у матери и/или плода. Эти нарушения включают, но не ограничиваются перечисленными, преэклампсию (Kaartokallio et al. Sci Rep. 2015;5:14107; Medina-Bastidas et al. Int J Mol Sci. 2020;21:3597), задержку внутриутробного развития (Faxén et al. Am J Perinatol. 1998;15:9-13; Medina-Bastidas et al. Int J Mol Sci. 2020;21:3597), инвазивную плацентацию, преждевременные роды (Enquobahrie et al. BMC Pregnancy Childbirth. 2009;9:56), гемолитическую болезнь новорожденных, плацентарную недостаточность (Kelly et al. Endocrinology. 2017;158:743-755), водянку плода (Magor et al. Blood. 2015;125:2405-17), порок развития плода (Slonim et al. Proc Natl Acad Sci USA. 2009;106:9425-9), синдром HELLP (Dijk et al. J Clin Invest. 2012;122:4003-4011), системную красную волчанку (Hong et al. J Exp Med. 2019;216:1154-1169) и другие иммунологические заболевания матери.

Аббревиатура «*п.о.*» относится к парам оснований. В некоторых случаях «п.о.» может использоваться для обозначения длины фрагмента ДНК, даже если фрагмент ДНК может быть одноцепочечным и не содержит пару оснований. В контексте одноцепочечной ДНК «п.о.» можно интерпретировать как указание длины в нуклеотидах.

Аббревиатура «*нт.*» относится к нуклеотидам. В некоторых случаях «нт.» может использоваться для обозначения длины одноцепочечной ДНК в нуклеотидных звеньях. «Нт.» также может использоваться для обозначения относительных положений, например, слева или справа от анализируемого локуса. Для двухцепочечной ДНК «нт.» может все еще относиться к длине одной цепи, а не к общему количеству нуклеотидов в двух цепях, если контекст явно не требует иного. В некоторых контекстах, касающихся технологической концептуализации, представления, обработки и анализа данных, «нт.» и «п.о.» могут использоваться взаимозаменяемо.

Термин «*модели машинного обучения*» может включать модели, основанные на использовании данных образца (например, обучающих данных) для прогнозирования тестовых данных, и, таким образом, может включать «управляемое обучение». Модели машинного обучения часто разрабатываются с использованием компьютера или процессора. Модели машинного обучения могут включать статистические модели.

Термин «*средство анализа данных*» может включать алгоритмы и/или модели, которые могут принимать данные в качестве ввода и затем выдавать прогнозируемый результат. Примеры «средств анализа данных» включают статистические модели, математические модели, модели машинного обучения, другие модели искусственного интеллекта и их комбинации.

Термин «*секвенирование в реальном времени*» может относиться к методике, которая включает сбор данных или мониторинг во время протекания реакции, вовлеченной в секвенирование. Например, секвенирование в реальном времени может включать оптический мониторинг или съемку встраивания ДНК-полимеразой нового основания.

Термин «*подпоследовательность*» может относиться к цепи оснований, которая меньше, чем полная последовательность, соответствующая молекуле нуклеиновой кислоты. Например, подпоследовательность может включать 1, 2, 3 или 4 основания, когда полная последовательность молекулы нуклеиновой кислоты включает 5 или более оснований. Согласно некоторым вариантам реализации подпоследовательность может относиться к цепи оснований, образующих звено, где звено повторяется несколько раз тандемным последовательным образом. Примеры включают звенья или подпоследовательности из 3 нуклеотидов, повторяющиеся в локусах, связанных с нарушениями, связанными с тринуклеотидным повтором, звенья или подпоследовательности из 1-6 нуклеотидов, повторяющиеся от 5 до 50 раз в виде микросателлитов, звенья или подпоследовательности из 10-60 нуклеотидов, повторяющиеся от 5 до 50 раз в виде минисателлитов, или в других генетических элементах, таких как повторы *Alu*.

Термин «*примерно*» или «*приблизительно*» может означать нахождение в пределах допустимого диапазона ошибок для конкретного значения, как определено обычным специалистом в данной области техники, что будет частично зависеть от того, как значение измеряется или определяется, т.е. ограничений системы измерения. Например, «примерно» может означать нахождение в пределах 1 или более чем 1 стандартного отклонения согласно применению в данной области техники. В качестве альтернативы, «примерно» может означать диапазон до 20%, до 10%, до 5% или до 1% от определенного значения. В качестве альтернативы, в частности, в отношении биологических систем или процессов, термин «примерно» или «приблизительно» может означать нахождение в пределах порядка величины, в пределах 5-кратного и более предпочтительно в пределах 2-кратного, значения. В тех случаях, когда конкретные значения описаны в заявке и формуле изобретения, если не указано иное, термин «примерно» следует понимать, как обозначающий нахождение в пределах допустимого диапазона ошибок для конкретного значения. Термин «примерно» может иметь значение, обычно понимаемое обычным специалистом в данной области техники. Термин «примерно» может относиться к ±10%. Термин «примерно» может относиться к ±5%.

В тех случаях, когда предложен диапазон значений, следует понимать, что также

конкретно раскрыто каждое промежуточное значение, вплоть до десятой доли единицы нижнего предела, если иное явным образом не следует из контекста, между верхним и нижним пределами этого диапазона. Каждый меньший диапазон между любым указанным значением или промежуточным значением в указанном диапазоне и любым другим указанным или промежуточным значением в указанном диапазоне охватывается вариантами реализации настоящего изобретения. Верхний и нижний пределы этих меньших диапазонов могут быть независимо включены или исключены из диапазона, и каждый диапазон, в котором в меньшие диапазоны включен один из пределов, не включен ни один из пределов или включены оба предела, также охватывается настоящим изобретением с учетом любого конкретно исключенного предела в указанном диапазоне. В тех случаях, когда указанный диапазон включает один или оба предела, диапазоны, исключающие один или оба из этих включенных пределов, также включены в настоящее изобретение.

Могут использоваться стандартные сокращения, например, п.о., пара (ы) оснований; тыс.п.о., тысяча (тысяч) п.о.; пл, пиколитр (ы); с или сек., секунда (секунды); мин, минута (минуты); ч или час, час (часы); амк., аминокислота (ы); нт., нуклеотид (ы); и тому подобное.

Если не указано иное, все технические и научные термины используются в настоящем документе в значении, соответствующем обычному пониманию специалиста в данной области техники, к которой принадлежит настоящее изобретение. Хотя любые способы и материалы, близкие или эквивалентные тем, которые описаны в настоящем документе, могут использоваться при практическом осуществлении или испытании вариантов реализации настоящего изобретения, некоторые возможные и примерные способы и материалы описаны далее.

Подробное описание изобретения

Анализ молекул внеклеточной ДНК включает преимущественно короткие фрагменты внеклеточной ДНК, часто из-за ограничений аналитических методик. Ограниченная возможность получения информации о последовательности из длинных молекул ДНК с использованием технологии секвенирования Illumina была продемонстрирована в недавних результатах секвенирования внеклеточной ДНК мыши (Serpas et al., Proc Natl Acad Sci USA. 2019;116:641-649). Только 0,02% секвенированных молекул ДНК находились в пределах диапазона от 600 п.о. до 2000 п.о. при использовании секвенирования Illumina у мышей дикого типа. Даже при использовании технологии одномолекулярного секвенирования в реальном времени (SMRT) от Pacific Biosciences (т.е. секвенирования PacBio SMRT) для секвенирования библиотек ДНК,

которые изначально были подготовлены для секвенирования Illumina, по-прежнему оставалось только 0,33% секвенированных молекул ДНК в пределах диапазона от 600 п.о. до 2000 п.о. Эти сообщенные данные свидетельствовали о том, что на этапе секвенирования будет потеряно 93% длинных молекул ДНК в пределах диапазона от 600 п.о. до 2000 п.о., присутствующих в исходной библиотеке ДНК.

Мы предположили, что на этапе подготовки библиотеки ДНК также будет потеряна значительная доля длинных молекул внеклеточной ДНК из-за ограничения ПЦР при амплификации длинных молекул ДНК, описанного выше. Jahr et al, используя гель-электрофорез, сообщили о наличии фрагментов большого размера, состоящих из многих тысяч пар оснований, например, ~10000 (Jahr et al. Cancer Res. 2001;61:1659-65). Однако полосы, показанные на изображении гель-электрофореза, не могут легко предоставить информацию о последовательности этих молекул в геле, а там более предоставить эпигенетическую информацию.

Ранее для исследования внеклеточной ДНК, экстрагированной из материнской плазмы, мы использовали платформу секвенирования Oxford Nanopore Technologies (Cheng et al Clin Chem. 2015;61:1305-6). Мы наблюдали очень небольшую долю длинной ДНК плазмы с размером более 1 тыс. п.о. (от 0,06% до 0,3%). Мы предположили, что такое низкое процентное содержание может быть результатом низкой точности секвенирования этой платформы.

В этой области внеклеточной ДНК большинство исследований было сосредоточено на коротких молекулах ДНК (например, <600 п.о.). Свойства, включая генетическую и эпигенетическую информацию, длинных молекул внеклеточной ДНК не изучены. Согласно настоящему изобретению предложен системный способ анализа длинных молекул внеклеточной ДНК, включая расшифровку их генетической и эпигенетической информации, а также их клиническую полезность в неинвазивном пренатальном тестировании, таком как, но не ограничиваясь перечисленными, неинвазивное детектирование моногенных нарушений, выяснение генома плода (например, неинвазивное секвенирование всего генома плода), детектирование мутаций *de novo* на уровне всего генома и детектирование/мониторинг связанных с беременностью нарушений, таких как преэклампсия и преждевременные роды.

I. Анализ размера внеклеточной ДНК

Были секвенированы образцы внеклеточной ДНК, полученные от беременных женщин, и было обнаружено, что значительная часть фрагментов ДНК являются длинными. Было продемонстрировано точное секвенирование длинных фрагментов внеклеточной ДНК. Были проанализированы размерные профили этих длинных молекул

внеклеточной ДНК. Было проведено сравнение количеств длинных молекул внеклеточной ДНК плода и матери. Длинные молекулы внеклеточной ДНК можно более точно выровнять с референсным геномом. Длинные молекулы внеклеточной ДНК можно применять для определения наследования гаплотипа.

Один образец ДНК плазмы беременной женщины в третьем триместре был проанализирован с использованием секвенирования PacBio SMRT. Молекулы двухцепочечной внеклеточной ДНК лигировали со шпилечными адаптерами и подвергали одномолекулярному секвенированию в реальном времени с использованием волноводов с нулевой модой и молекул одной полимеразы (Eid et al. Science. 2009;323:133-8).

Мы секвенировали 1,1 миллиарда подридов, из которых 659,3 миллиона подридов можно было выровнять с референсным геномом человека (hg19). Подриды были сгенерированы из 4,6 миллиона лунок для одномолекулярного секвенирования в реальном времени PacBio (SMRT), которые содержали по меньшей мере один подрид, который можно было выровнять с референсным геномом человека. В среднем каждая молекула в лунке SMRT была секвенирована в среднем 143 раза. В этом примере имелось 4,5 миллиона кольцевых консенсусных последовательностей (CCS), что предполагает 4,5 миллиона молекул внеклеточной ДНК, которые можно использовать для последующих анализов. Размер каждой внеклеточной ДНК определяли на основании CCS путем подсчета количества идентифицированных оснований.

На фиг. 1A и 1B показано распределение размеров внеклеточной ДНК от 0 до 20 тыс. п.о. По оси Y показана частота. По оси X показан размер в парах оснований от 0 до 20 тыс. п.о. по линейной шкале (фиг. 1A) или по логарифмической шкале (фиг. 1B). Поскольку секвенирование выполняли по всей длине молекул ДНК, размер каждой молекулы ДНК можно было определить непосредственно путем подсчета количества нуклеотидов в подриде или CCS. Измерение размера фрагмента ДНК может быть достигнуто с использованием любых платформ секвенирования, которые могут считывать по всей длине фрагментов ДНК, и не ограничивается использованием секвенаторов отдельных молекул. Например, секвенаторы по Сэнгеру могут считывать на протяжении 800 п.о. Секвенирование с коротким ридом, например, на платформах Illumina, может считывать на протяжении 250 п.о. Секвенаторы отдельных молекул, такие как Pacific Biosciences и Oxford Nanopore, могут считывать на протяжении более 10000 п.о. Размеры фрагментов ДНК также можно определить после выравнивания с референсным геномом, например, референсным геномом человека. Размеры фрагментов ДНК можно определить с помощью секвенирования спаренных концов с последующим выравниванием с референсным геномом. На фиг. 1B показан профиль с длинным хвостом. Из 4,5 миллиона

CCS было 22,5% внеклеточной ДНК более 200 п.о., 19,0% из них были более 300 п.о., 11,8% из них были более 400 п.о., 10,6% из них были более 500 п.о., 8,9% из них были более 600 п.о., 6,4% из них были более 1 тыс. п.о., 3,5% из них были более 2 тыс. п.о., 1,9% из них были более 3 тыс. п.о., 0,9% из них были более 4 тыс. п.о. и 0,04% из них были более 10 тыс. п.о. Самый длинный фрагмент, наблюдаемый в текущих результатах PacBio SMRT, представлял собой 29804 п.о.

Одну ДНК плазмы беременного субъекта также секвенировали на платформе секвенирования Illumina с использованием протокола подготовки библиотеки на основе ПЦР (Lun et al. Clin Chem. 2013;59:1583-94). Из 18,2 миллиона ридов спаренных концов было 5,3% внеклеточной ДНК более 200 п.о., 2,0% из них были более 300 п.о., 0,3% из них были более 400 п.о., 0,2% из них были более 500 п.о., 0,2% из них были более 600 п.о. (таблица 1). В качестве сравнения мы проанализировали размерные профили путем объединения данных одномолекулярного секвенирования в реальном времени (т.е. в общей сложности 4,4 миллиона CCS) от 5 беременных субъектов. Мы наблюдали большее количество молекул ДНК плазмы более 600 п.о. (28,56%) по сравнению с данными аналогичных молекул (0,2%), полученными с помощью платформы секвенирования Illumina. Эти результаты свидетельствовали о том, что секвенирование PacBio SMRT может позволить достичь в 143 раза больше длинных молекул ДНК (длиннее, чем 600 п.о.). Мы можем получить 4,77% молекул ДНК плазмы более 3 тыс. п.о., используя одномолекулярное секвенирование в реальном времени, в то время как на платформе секвенирования Illumina считывание данных отсутствовало.

В отличие от предыдущего сообщения, в котором показана очень небольшая доля длинных молекул ДНК плазмы более 1 тыс. п.о. (от 0,06% до 0,3%) при использовании платформы секвенирования Oxford Nanopore Technologies (Cheng et al Clin Chem. 2015;61:1305-6), мы смогли получить в 21 раз больше ДНК плазмы более 1 тыс. п.о. (6,4%), продемонстрировав, что секвенирование PacBio SMRT было намного более эффективным в получении информации о последовательности из популяции длинных ДНК.

По сравнению с секвенированием спаренных концов с короткими ридами, таким как платформа секвенирования Illumina, технологии секвенирования с длинными ридами, такие как технология PacBio SMRT, имеют ряд преимуществ при определении характеристик (например, длины) длинного фрагмента ДНК. Например, длинный рид в целом может обеспечить более точное выравнивание с референсным геномом человека (например, hg19). Технологии с длинными ридами также могут обеспечить точное определение длины молекулы ДНК плазмы путем прямого подсчета количества

секвенированных нуклеотидов. Напротив, оценка размера ДНК плазмы на основании коротких ридов спаренных концов является непрямым методом, который использует самые внешние координаты выравненного рида спаренных концов для выведения размера молекулы ДНК плазмы. Для такого непрямого подхода ошибки при выравнивании могут привести к точному выведению размера. В связи с этим увеличение охвата размеров между ридами спаренных концов может увеличить вероятность ошибки при выравнивании.

Таблица 1. Сравнение распределения размеров между секвенированиями внеклеточной ДНК PacBio и Illumina

Отсечение по размеру фрагмента ДНК плазмы (≥X п.о.)	Процентное содержание желаемых фрагментов, полученных с помощью одномолекулярного секвенирования в реальном времени (%)	Процентное содержание желаемых фрагментов, полученных с помощью платформы секвенирования Illumina (%)
200	50,32	5,3
300	46,43	2
400	35,05	0,3
500	32,34	0,2
600	28,56	0,2
700	26,74	0,00
800	24,50	0,00
900	23,08	0,00
1000	21,37	0,00
1100	20,06	0,00
1200	18,60	0,00
1300	17,36	0,00
1400	16,08	0,00
1500	14,94	0,00
1600	13,84	0,00
1700	12,83	0,00
1800	11,88	0,00
1900	11,00	0,00
2000	10,19	0,00
2100	9,43	0,00
2200	8,75	0,00
2300	8,10	0,00
2400	7,51	0,00
2500	6,96	0,00
2600	6,45	0,00
2700	5,99	0,00
2800	5,55	0,00
2900	5,15	0,00
3000	4,77	0,00

На фиг. 2A и 2B показано распределение размеров внеклеточной ДНК от 0 до 5 тыс. п.о. По оси Y показана частота. По оси X показан размер в парах оснований от 0 до 5 тыс. п.о. по линейной шкале (фиг. 2A) или по логарифмической шкале (фиг. 2B). Наблюдали ряд основных пиков, появляющихся с периодическими профилями. Такие

периодические профили распространялись даже на молекулы в пределах диапазона от 1 тыс. п.о. до 2 тыс. п.о. Пик с самой высокой частотой (2,6%) был при 166 п.о., что согласовывалось с предыдущим результатом с использованием технологии Illumina (Lo et al. Sci Transl Med. 2010;2:61ra91). Расстояние между соседними основными пиками на фиг. 2B составляло приблизительно 200 п.о., это указывает на то, что получение длинной внеклеточной ДНК также будет включать нуклеосомальные структуры.

На фиг. 3A и 3B показано распределение размеров внеклеточной ДНК от 0 до 400 п.о. По оси Y показана частота. По оси X показан размер в парах оснований от 0 до 400 п.о. по линейной шкале (фиг. 3A) или по логарифмической шкале (фиг. 3B). Характерные признаки с наиболее преобладающим пиком при 166 п.о. и периодичностью в 10 п.о., встречающиеся в молекулах размером менее 166 п.о., о которых сообщалось ранее (Lo et al. Sci Transl Med. 2010;2:61ra91), также воспроизводились при использовании нового способа в соответствии с настоящим изобретением. Эти результаты свидетельствовали о том, что определение размера молекулы путем подсчета количества оснований, секвенированных из отдельной молекулы в соответствии с настоящим изобретением, было надежным.

A. Анализ размера ДНК плода и матери

Были проанализированы и сравнены размеры фрагментов ДНК матери и плода. В качестве примера, были секвенированы ДНК лейкоцитной пленки одной беременной женщины и соответствующая ей плацентарная ДНК с получением 59x и 58x охвата гаплоидного генома, соответственно. Мы идентифицировали в общей сложности 822409 информативных однонуклеотидных полиморфизмов (ОНП), по которым мать была гомозиготной, а плод был гетерозиготным. Аллели, специфические для плода, определяются как те аллели, которые присутствуют в геноме плода, но отсутствуют в геноме матери. Мы идентифицировали 2652 специфических для плода фрагмента и 24837 общих фрагментов (т.е. фрагментов, несущих общий аллель; преимущественно материнского происхождения) в материнской плазме (M13160) с помощью секвенирования PacBio. Фракция ДНК плода составила 21,8%.

На фиг. 4A и 4B показано распределение размеров внеклеточной ДНК среди фрагментов, несущих общие аллели (общие) и специфические для плода аллели (специфические для плода). По оси X показан размер в парах оснований от 0 до 20 тыс. п.о. по линейной шкале (фиг. 4A) или по логарифмической шкале (фиг. 4B). Как фрагменты, несущие общие аллели (преимущественно материнского происхождения), так и фрагменты, несущие специфический для плода аллель (плацентарного происхождения), проявляли распределения с длинным хвостом, это свидетельствует о наличии длинных

молекул ДНК, происходящих как от плода, так и от матери. Наблюдали 22,6% молекул ДНК плазмы, размеры которых были более 2 тыс. п.о., для фрагментов преимущественно материнского происхождения, в то время как для фрагмента, происходящего от плода, 8,5% молекул ДНК плазмы имели размеры более 2 тыс. п.о. Эти результаты свидетельствовали о том, что молекулы ДНК плода содержали меньше длинных молекул ДНК. Процентное содержание длинной ДНК, присутствующей в этом анализе на основе ОНП, касающемся происхождения от матери и от плода ДНК плазмы, по-видимому, было намного выше, чем то, которое наблюдали в общем анализе размера. Такое расхождение, вероятно, было связано с тем, что длинная молекула ДНК имеет больше шансов охватывать один или более ОНП, чем короткая, и поэтому длинная ДНК будет преимущественно отобрана для анализа на основе ОНП. Относительная доля длинных молекул ДНК, помеченных ОНП, отклоняющаяся от соответствующей доли длинных ДНК в исходном пуле, будет определяться размерами этих молекул. Среди этих специфических для плода фрагментов ДНК самый длинный был 16186 п.о., в то время как среди фрагментов, несущих общие аллели, самый длинный был 24166 п.о.

На фиг. 5A и 5B показано распределение размеров внеклеточной ДНК среди фрагментов, несущих общие аллели (общие) и специфические для плода аллели (специфические для плода). По оси X показан размер в парах оснований от 0 до 5 тыс. п.о. по линейной шкале (фиг. 5A) или по логарифмической шкале (фиг. 5B). Наблюдали ряд основных пиков, возникающих периодическим образом, для фрагментов размером менее 2 тыс. п.о. как для специфических для плода, так и для общих фрагментов ДНК. Основные пики, вероятно, выравнивались с нуклеосомальными структурами.

На фиг. 6A и 6B показано распределение размеров внеклеточной ДНК среди фрагментов, несущих общие аллели (общие) и специфические для плода аллели (специфические для плода). По оси X показан размер в парах оснований от 0 до 1 тыс. п.о. по линейной шкале (фиг. 6A) или по логарифмической шкале (фиг. 6B). Наблюдали ряд основных пиков, возникающих периодическим образом, для фрагментов размером менее 1 тыс. п.о. как для специфических для плода, так и для общих фрагментов ДНК. Основные пики, вероятно, выравнивались с нуклеосомальными структурами. Очевидно, имел место наблюдаемый сдвиг размерного профиля ДНК плода влево от размерного профиля общих фрагментов ДНК, это указывает на то, что ДНК плода может содержать больше коротких молекул ДНК, чем ДНК матери.

На фиг. 7A и 7B показано распределение размеров внеклеточной ДНК среди фрагментов, несущих общие аллели (общие) и специфические для плода аллели (специфические для плода). По оси X показан размер в парах оснований от 0 до 400 п.о.

по линейной шкале (фиг. 7A) или по логарифмической шкале (фиг. 7B). Характерные признаки с наиболее преобладающим пиком при 166 п.о. и периодичностью в 10 п.о., встречающиеся в молекулах менее 166 п.о. как плода, так и матери, о которых сообщалось ранее (Lo et al. Sci Transl Med. 2010;2:61ra91), также воспроизводились при использовании нового способа в соответствии с настоящим изобретением. Эти результаты свидетельствовали о том, что определение размера молекулы путем подсчета количества оснований, секвенированных из отдельной молекулы в соответствии с настоящим изобретением, было надежным.

B. Анализ размера и метилирования

Были проанализированы уровни метилирования длинных молекул внеклеточной ДНК матери и плода. Было обнаружено, что уровень метилирования молекул ДНК плода ниже, чем уровень метилирования молекул ДНК матери.

При секвенировании PacBio SMRT ДНК-полимераза опосредует включение флуоресцентномеченых нуклеотидов в комплементарные цепи. Характеристики флуоресцентных импульсов, образующихся во время синтеза ДНК, включая продолжительность между импульсами и ширину импульса, будут отражать кинетику полимеразы, которую можно использовать для определения модификаций нуклеотидов, таких как, но не ограничиваясь перечисленными, 5-метилцитозин, с использованием подходов, описанных в нашем предыдущем изобретении (заявка США № 16/995607, поданная 17 августа 2020 г., озаглавленная «ОПРЕДЕЛЕНИЕ МОДИФИКАЦИЙ ОСНОВАНИЙ НУКЛЕИНОВЫХ КИСЛОТ»), полное содержание которой включено в настоящий документ посредством ссылки для всех целей.

Согласно вариантам реализации мы идентифицировали 95210 фрагментов, несущих специфические для матери аллели, и 2652 фрагмента, несущих специфические для плода аллели, соответственно. Специфические для матери аллели определяются в настоящем документе как аллели, присутствующие в геноме матери, но отсутствующие в геноме плода, которые могут быть идентифицированы по ОНП, когда мать является гетерозиготной, а плод является гомозиготным. В этом примере мы идентифицировали в общей сложности 677375 таких информативных ОНП. Мы определили размер каждой молекулы внеклеточной ДНК. Согласно одному варианту реализации, поскольку состояния метилирования в геноме вариабельны, например, уровни метилирования CpG-островков обычно ниже, чем областей без CpG-островков, чтобы свести к минимуму вариабельность, вносимую геномным контекстом, можно было *in silico* отобрать фрагменты, которые имеют размер более 1 тыс. п.о., содержат по меньшей мере 5 сайтов CpG и соответствуют плотности CpG менее 5% (т.е. количество сайтов CpG в молекуле,

деленное на общую длину этой молекулы <0,05) и использовать для последующего анализа.

На фиг. 8 показаны уровни метилирования отдельной двухцепочечной молекулы ДНК среди фрагментов, несущих специфические для матери аллели и специфические для плода аллели. По оси Y показан уровень метилирования отдельной двухцепочечной молекулы ДНК в процентах. По оси X показаны как фрагменты, несущие специфические для матери аллели, так и фрагменты, несущие специфические для плода аллели. Уровни метилирования отдельной двухцепочечной молекулы ДНК для фрагментов, несущих специфический для плода аллель (среднее: 62,7%; межквартильный диапазон, IQR: 50,0–77,2%), ниже, чем у аналогичных фрагментов, несущих специфические для матери аллели (среднее: 72,7%; IQR: 60,6% - 83,3%) (P<0,0001).

На фиг. 9A показано эмпирическое распределение уровней метилирования отдельной двухцепочечной молекулы ДНК фрагментов, аппроксимированное с использованием ядерной оценки плотности, реализованной в пакете R (r-project.org/). Частота показана по оси Y. По оси X показан уровень метилирования отдельной двухцепочечной молекулы ДНК в процентах. Распределение длинных фрагментов ДНК, специфических для плода, находится слева от распределения фрагментов, специфических для матери, это свидетельствует о более низких уровнях метилирования отдельной двухцепочечной молекулы ДНК, присутствующих в молекулах ДНК плода.

На фиг. 9B показан анализ операционных характеристик приемника (ROC) с использованием уровней метилирования отдельной двухцепочечной молекулы ДНК. По оси Y показана чувствительность. По оси X показана специфичность. Используя уровни метилирования отдельной двухцепочечной молекулы ДНК для выполнения ROC-анализа для изучения эффективности установления отличия фрагментов ДНК плода от фрагментов ДНК матери с использованием уровня метилирования отдельной двухцепочечной молекулы ДНК, было обнаружено, что площадь под ROC-кривой (AUC) была 0,62, что превышало результат случайного угадывания 0,5. Согласно вариантам реализации можно применять пространственные профили состояний метилирования, такие как последовательность состояний метилирования, относительные или абсолютные расстояния между модифицированными основаниями и геномные координаты, в отдельной молекуле для дальнейшего улучшения определения происхождения фрагментов от матери или от плода в плазме. Согласно вариантам реализации можно комбинировать профили метилирования с другими фрагментарными метриками (т.е. параметрами, относящимися к фрагментации ДНК), включая, но не ограничиваясь этим, предпочтительные концы (Chan et al. Proc Natl Acad Sci USA. 2016;113:E8159-8168),

концевые мотивы (Serpas et al. Proc Natl Acad Sci USA. 2019;116:641-649), размеры (Lo et al. Sci Transl Med. 2010;2:61ra), определение ориентации (т.е. ориентации по отношению к конкретным элементам внутри генома, например, открытым областям хроматина, профилям фрагментации (Sun et al. Genomes Res. 2019;29:418-427)), топологические формы (например, линейные в сравнении с кольцевыми молекулами ДНК (Ma et al. Clin Chem. 2019;65:1161-1170)), чтобы улучшить классификационную мощность различения фрагментов плацентарного происхождения (происходящих от плода).

На фиг. 10A и 10B показано, что уровни метилирования отдельной двухцепочечной молекулы ДНК для фрагментов ДНК как плода, так и матери варьировались в зависимости от размеров фрагментов. По оси Y показан уровень метилирования отдельной двухцепочечной молекулы ДНК в процентах. По оси X показан размер от 0 до более 20 тыс. п.о. (фиг. 10A) и от 0 до более 1 тыс. п.о. (фиг. 10B). С другой стороны, уровни метилирования отдельной двухцепочечной молекулы ДНК специфических для плода молекул ДНК обычно были ниже, чем таковые специфических для матери молекул ДНК как в длинном (фиг. 10A), так и в коротком (фиг. 10B) диапазонах. Этот результат согласовывался с имеющимися в настоящий момент данными о том, что уровень метилирования ДНК плода был ниже, чем ДНК матери в плазме беременной женщины для коротких молекул ДНК (Lun et al. Clin Chem. 2013;59:1583-94).

Согласно вариантам реализации, поскольку уровень метилирования молекул ДНК плода относительно ниже, чем уровень молекул ДНК матери, можно отобрать молекулы, у которых уровни метилирования отдельной двухцепочечной молекулы ДНК меньше определенного порога, например, но не ограничиваясь этим, 80%, 70%, 60%, 50%, 40%, 30%, 20%, 10% и 5%, для обогащения молекул внеклеточной ДНК, происходящих от плода, в пуле ДНК плазмы. Например, фракция ДНК плода составляет 2,6% для фрагментов >1 тыс. п.о. Если мы отбираем фрагменты (>1 тыс. п.о.) с уровнем метилирования отдельной двухцепочечной молекулы <50%, фракция ДНК плода этих дополнительно отобранных фрагментов >1 тыс. п.о. увеличится до 5,6% (т.е. 115,4% увеличение). В другом примере фракция ДНК плода составляет 26,2% для фрагментов <200 п.о. Если мы отбираем фрагменты (<200 п.о.) с уровнем метилирования отдельной двухцепочечной молекулы <50%, фракция ДНК плода этих дополнительно отобранных фрагментов >200 п.о. увеличится до 41,6% (т.е. 58,8%). Таким образом, при определенных обстоятельствах использование порога уровней метилирования отдельной двухцепочечной молекулы ДНК для обогащения ДНК плода будет более эффективным для длинных молекул ДНК.

C. Гаплотип и метилирование длинной внеклеточной ДНК

Согласно вариантам реализации можно получить составы оснований, размеры и модификации оснований для каждой отдельной молекулы ДНК, используя способы, описанные в настоящем изобретении. Информация о ОНП и метилировании для длинных молекул внеклеточной ДНК может быть использована для гаплотипирования. Использование длинных молекул ДНК, присутствующих в пуле внеклеточной ДНК, выявленное в настоящем изобретении, позволит фазировать варианты в геномах, эффективно используя информацию о гаплотипах, присутствующую в каждой консенсусной последовательности, в соответствии с опубликованными методами, но не ограничиваясь ими (Edge et al. Genome Res. 2017;27:801-812; Wenger et al. Nat Biotechnol. 2019;37:1155-1162). Реализация определения гаплотипов осуществляется в соответствии с информацией о последовательности внеклеточной ДНК, что отличается от предыдущих исследований, которые должны полагаться на длинную ДНК, полученную из тканевой ДНК. Гаплотип в пределах геномной области иногда может назваться гаплотипным блоком. Гаплотипный блок можно рассматривать как набор аллелей на хромосоме, которые были фазированы. Согласно некоторым вариантам реализации гаплотипный блок может быть расширен как можно дальше согласно набору информации о последовательности, которая обосновывает физическое сцепление двух аллелей на хромосоме, а также информации о перекрытии аллелей между различными последовательностями.

На фиг. 11A и 11B показан пример специфической для плода длинной молекулы ДНК, идентифицированной в ДНК материнской плазмы беременной женщины. В настоящем документе мы иллюстрируем варианты реализации нашего изобретения с использованием одной молекулы размером 16186 п.о., из числа этих специфических для плода фрагментов ДНК, которая была выравнена с областью в хромосоме 10 референсного генома человека (chr10: 56282981-56299166) (фиг. 11A) и несла 7 специфических для плода аллелей (фиг. 11B). Имелось 6 из 7 специфических для плода аллелей, которые согласовывались с информацией об аллелях, выведенной на основании глубокого секвенирования геномов матери и плода (с использованием платформы Illumina) (фиг. 11B). Было определено, что ее уровень метилирования составляет 27,1% в соответствии со способом, описанным в настоящем изобретении (фиг. 11B), что было намного ниже, чем средний уровень специфических для матери фрагментов (72,7%). Эти результаты свидетельствовали о том, что профили метилирования отдельной двухцепочечной молекулы ДНК могут служить маркерами для дифференциации происхождения молекул внеклеточной ДНК от матери и от плода.

На фиг. 12A и 12B показан пример длинной молекулы ДНК матери, несущей

общие аллели, идентифицированной в ДНК материнской плазмы беременной женщины. Из этих фрагментов, несущих общие аллели, самый длинный фрагмент имел размер 24166 п.о., который был выравнен с областью в хромосоме 6 референса человека (chr6: 111074371-111098536) (фиг. 12A) и нес 18 общих аллелей (фиг. 12B). Все эти общие аллели согласовывались с информацией об аллелях, выведенной на основании глубокого секвенирования геномов матери и плода (с использованием платформы Illumina) (фиг. 12B). Было определено, что ее уровень метилирования составляет 66,9% в соответствии со способом, описанным в настоящем изобретении (фиг. 12B). Генетическая и эпигенетическая информация о молекулах внеклеточной ДНК длиной порядка тысяч пар оснований не могла быть легко идентифицирована при использовании секвенирования с короткими ридами, такого как бисульфитное секвенирование (Illumina).

В настоящем документе мы описываем способ определения относительной вероятности происхождения молекулы от беременной женщины или плода. У беременной женщины молекулы ДНК, несущие генотипы плода, на самом деле происходят из плаценты, в то время как большинство молекул ДНК, несущих материнские генотипы, происходят из клеток крови матери. Согласно этому способу мы сначала строим кривую частотного распределения молекул ДНК в соответствии с их уровнем метилирования как для клеток плаценты, так и для материнских клеток крови. Для этого мы разделили геном человека на группы разного размера.

На фиг. 13 показано частотное распределение для ДНК из плацентарных клеток (красный) и материнских клеток крови (синий) в соответствии с уровнем метилирования при различных разрешениях от 1 тыс. п.о. до 20 тыс. п.о. Частота показана по оси Y. Уровень метилирования показан по оси X. Примеры размера групп включают, но не ограничиваются ими, 1 тыс. п.о., 2 тыс. п.о., 5 тыс. п.о., 10 тыс. п.о., 15 тыс. п.о. и 20 тыс. п.о. Уровень метилирования каждой группы определяли на основании количества метилированных сайтов CpG, деленного на общее количество сайтов CpG. После определения уровня метилирования всех групп можно построить кривую частотного распределения для каждого из плацентарного генома и генома материнских клеток крови для разных групп размеров.

Основываясь на уровне метилирования длинной молекулы ДНК, вероятность ее происхождения из клеток плаценты или материнских клеток крови может быть определена по относительной представленности двух типов молекул ДНК при таком уровне метилирования, а также по относительной концентрации ДНК плода в образце.

Пусть x и y представляют собой частоту молекул ДНК, происходящих из клеток плаценты и материнских клеток крови, соответственно, при конкретном уровне

метилирования, и f представляет собой относительную концентрацию ДНК плода в образце.

Вероятность (P) происхождения молекулы ДНК от плода может быть рассчитана следующим образом:

$$P = \frac{x \times f}{(x \times f) + y(1-f)}$$

На основании предыдущего примера рассматривается молекула ДНК плазмы размером 16 тыс. п.о. и уровень метилирования 27,1%.

На фиг. 14A и 14B показано частотное распределение для ДНК из плацентарных клеток (красный) и материнских клеток крови (синий) в соответствии с уровнями метилирования в окнах 16 тыс. п.о. (фиг. 14A) и 24 тыс. п.о. (фиг. 14B). Частота показана по оси Y. Уровень метилирования показан по оси X. На основании графика частотного распределения для фрагментов размером 16 тыс. п.о. (фиг. 14A) частоты для молекул ДНК, происходящих из клеток плаценты и материнских клеток крови, составляют 0,6% и 0,08%, соответственно. Поскольку фракция ДНК плода составляет 21,8%, вероятность происхождения этого фрагмента ДНК из плаценты составляет 64%, что предполагает повышенную вероятность плацентарного происхождения.

Вероятность того, что молекула ДНК происходит из тканей плода, также может быть рассчитана для молекулы ДНК плазмы размером 24 тыс. п.о. и уровня метилирования 66,9%. На основании графика частотного распределения для фрагментов размером 24 тыс. п.о. частоты для молекул ДНК, происходящих из клеток плаценты и материнских клеток крови, составляют 0,05% и 0,16% (фиг. 14B), соответственно. Вероятность происхождения этого фрагмента ДНК из плаценты составляет 0,8%, что предполагает очень низкую вероятность его плацентарного происхождения. Другими словами, существует высокая вероятность того, что молекула имеет материнское происхождение.

В этом расчете можно дополнительно учитывать размер молекул ДНК, обращаясь к кривым распределения размеров для ДНК плода и матери. Такой анализ может быть выполнен, например, но не ограничиваясь этим, с использованием теоремы Байеса, логистической регрессии, множественной регрессии и метода опорных векторов, анализа методом случайного леса, анализа по алгоритму построения бинарного дерева решений (CART), алгоритма K-ближайших соседей.

На фиг. 15A и 15B показано, что длинный фрагмент ДНК в плазме имеет размер 18896 п.о., который был выравнен с областью в хромосоме 8 референса человека (chr8: 108694010-108712904) (фиг. 15A) и несет 7 специфических для матери аллелей (фиг. 15B).

Все эти специфические для матери аллели согласовывались с информацией об аллелях, выведенной на основании глубокого секвенирования геномов матери и плода (технология Illumina) (фиг. 15B). Было определено, что его уровень метилирования составляет 72,6% в соответствии со способом, описанным в настоящем изобретении (фиг. 15B), что сравнимо с совокупным уровнем метилирования специфических для матери фрагментов (72,7%). Таким образом, такую молекулу с большей вероятностью можно классифицировать как фрагмент материнского происхождения. Генетическая и эпигенетическая информация о молекулах внеклеточной ДНК длиной порядка тысяч пар оснований не могла быть легко идентифицирована при использовании секвенирования с короткими ридами, такого как бисульфитное секвенирование (Illumina).

Используя способ, описанный выше, можно рассчитать вероятность происхождения этой молекулы из плаценты. На основании графика частотного распределения для фрагментов размером 19 тыс. п.о. частоты для молекул ДНК, происходящих из клеток плаценты и материнских клеток крови, составляют 0,65% и 0,23%, соответственно. Вероятность происхождения этого фрагмента ДНК из плаценты составляет 43%, что предполагает повышенную вероятность его материнского происхождения.

D. Клинические приложения гаплотипирования

Согласно вариантам реализации возможность анализировать как короткую, так и длинную молекулу ДНК в ДНК плазмы беременной женщины позволит нам провести анализ относительной дозы гаплотипа (RHDO) (Lo et al. Sci Transl Med. 2010;2:61ra91; Hui et al. Clin Chem. 2017;63:513-524) без необходимости предварительного получения информации о генотипе отца или матери, или плода из тканей. Эта возможность будет более рентабельной и клинически применимой, чем это было возможно ранее.

На фиг. 16 проиллюстрирован принцип того, как можно применять внеклеточную ДНК при беременности для проведения анализа RHDO. Внеклеточную ДНК выделяют у беременной женщины и подвергают SMRT-секвенированию на этапе 1605. Размеры, информацию об аллелях и состояния метилирования для каждой молекулы, включая длинные и короткие молекулы ДНК, можно определить в соответствии со способами, описанными в настоящем изобретении. На этапе 1610, в соответствии с информацией о размере, можно разделить секвенированные молекулы на две категории, а именно на длинные и короткие молекулы ДНК. Отсечение, используемое для определения категорий длинной и короткой ДНК, может включать, но не ограничивается перечисленными, 150 п.о., 180 п.о., 200 п.о., 250 п.о., 300 п.о., 350 п.о., 400 п.о., 450 п.о., 500 п.о., 550 п.о., 600 п.о., 650 п.о., 700 п.о., 750 п.о., 800 п.о., 850 п.о., 900 п.о., 950 п.о., 1 тыс. п.о., 1,1 тыс. п.о.,

1,2 тыс. п.о., 1,3 тыс. п.о., 1,4 тыс. п.о., 1,5 тыс. п.о., 1,6 тыс. п.о., 1,7 тыс. п.о., 1,8 тыс. п.о., 1,9 тыс. п.о., 2 тыс. п.о., 2,5 тыс. п.о., 3 тыс. п.о., 4 тыс. п.о., 5 тыс. п.о., 6 тыс. п.о., 7 тыс. п.о., 8 тыс. п.о., 9 тыс. п.о., 10 тыс. п.о., 15 тыс. п.о., 20 тыс. п.о., 30 тыс. п.о., 40 тыс. п.о., 50 тыс. п.о., 60 тыс. п.о., 70 тыс. п.о., 80 тыс. п.о., 90 тыс. п.о., 100 тыс. п.о., 200 тыс. п.о., 300 тыс. п.о., 400 тыс. п.о., 500 тыс. п.о. или 1 млн. п.о. На этапе 1615, согласно вариантам реализации, информация об аллелях, присутствующая в длинных молекулах ДНК, может быть использована для конструирования материнских гаплотипов, а именно Hap I и Hap II. Короткие молекулы ДНК могут быть выравнены с материнскими гаплотипами в соответствии с информацией об аллелях. Следовательно, можно определить количество молекул внеклеточной ДНК (например, коротких ДНК), происходящих из материнских Hap I и Hap II.

На этапе 1620 может быть проанализирован дисбаланс гаплотипов. Дисбаланс может заключаться в количествах молекул, размерах молекул или состояниях метилирования молекул. На этапе 1625 можно вывести наследование по материнской линии плода. Если доза Hap I в ДНК материнской плазмы сверхпредставлена, плод, вероятно, наследует материнский Hap I. В ином случае плод, вероятно, наследует материнский Hap II. Различные статистические подходы, включая, но не ограничиваясь перечисленными, последовательный критерий отношений вероятностей (SPRT), биномиальный критерий, критерий хи-квадрат, t-критерий Стьюдента, непараметрические критерии (например, критерий Уилкоксона) и скрытые модели Маркова, могут быть использованы для определения того, какой материнский гаплотип сверхпредставлен.

В дополнение к счетному анализу, согласно вариантам реализации, также определяют метилирование и размер короткой молекулы ДНК и относят к материнским гаплотипам. Дисбаланс метилирования между двумя гаплотипами (т.е. Hap I и Hap II) может быть использован для определения материнского гаплотипа, унаследованного плодом. Если плод унаследовал Hap I, в материнской плазме будет присутствовать больше фрагментов, несущих аллели Hap I, по сравнению с фрагментами, несущими аллели Hap II. Гипометилирование фрагментов ДНК, происходящих от плода, будет снижать уровень метилирования Hap I по сравнению с Hap II. Другими словами, если метилирование Hap I показывало более низкий уровень метилирования, чем Hap II, плод с большей вероятностью наследует материнский Hap I. В ином случае плод с большей вероятностью наследует материнский Hap II. Согласно другому варианту реализации вероятность того, что отдельные фрагменты происходят от плода или матери, можно рассчитать, как описано выше. Для всех фрагментов, выравнивающихся с Hap I, совокупная вероятность того, что эти фрагменты происходят от плода, может быть определена на основе теоремы

Байеса. Сходным образом, для Hap II можно рассчитать совокупную вероятность того, что эти фрагменты происходят от плода. Вероятность того, что плод наследует Hap I или Hap II, затем можно вывести на основании двух совокупных вероятностей.

Согласно вариантам реализации удлинение или укорочение размера между двумя гаплотипами (т.е. Hap I и Hap II) можно использовать для определения материнского гаплотипа, унаследованного плодом. Если плод унаследовал Hap I, в материнской плазме будет присутствовать больше фрагментов, несущих аллели Hap I, по сравнению с фрагментами, несущими аллели Hap II. Фрагменты ДНК, происходящие от плода, будут относительно короче, чем фрагменты, происходящие из Hap II. Другими словами, если молекулы, происходящие из Hap I, содержат больше короткой ДНК, чем Hap II, плод с большей вероятностью наследует материнский Hap I. В ином случае плод с большей вероятностью наследует материнский Hap II.

Согласно некоторым вариантам реализации можно выполнить комбинированный анализ количества, размера и метилирования между материнскими Hap I и Hap II, чтобы сделать вывод о наследовании по материнской линии плода. Например, можно использовать логистическую регрессию для объединения этих трех метрик, включая количества, размеры и состояния метилирования.

В клинической практике анализ на основании гаплотипа, касающийся количеств, размеров и состояний метилирования, может обеспечить определение того, унаследовал ли нерожденный плод материнский гаплотип, связанный с генетическими нарушениями, например, но не ограничиваясь этим, моногенными нарушениями, включая синдром ломкой Х-хромосомы, мышечную дистрофию, болезнь Хантингтона или бета-талассемию. Детектирование нарушений, в которые вовлечены повторы последовательностей ДНК, в ридах длинных внеклеточных ДНК описано отдельно в настоящем изобретении.

E. Нацеленное секвенирование длинных молекул внеклеточной ДНК

Способы, описанные в настоящем изобретении, также можно применять для анализа одного или более отобранных длинных фрагментов ДНК. Согласно вариантам реализации один или более длинных фрагментов ДНК, представляющих интерес, могут быть сначала обогащены с помощью способа гибридизации, который позволяет гибридизовать молекулы ДНК из области(областей), представляющей интерес, с синтетическими олигонуклеотидами, имеющими комплементарные последовательности. Для одновременного декодирования размера, генетической и эпигенетической информации с использованием способов, описанных в настоящем изобретении, предпочтительно не амплифицировать целевые молекулы ДНК с помощью ПЦР перед секвенированием, поскольку информация о модификации оснований в исходной молекуле

ДНК не будет переноситься в продукты ПЦР.

Было разработано несколько способов обогащения этих целевых областей без выполнения ПЦР-амплификации. Согласно другому варианту реализации одна или более целевых длинных молекул ДНК могут быть обогащены путем использования системы кластерных коротких палиндромных повторов, разделенных регулярными промежутками (CRISPR)-CRISPR-ассоциированного белка 9 (Cas9) (Stevens et al. PLOS One 2019;14(4):e0215441; Watson et al. Lab Invest 2020;100:135-146). Несмотря на то, что такие разрезы, опосредуемые CRISPR-Cas9, будут изменять размер исходных длинных молекул ДНК, их генетическая и эпигенетическая информация по-прежнему сохраняется и может быть получена с использованием способов, описанных в настоящем изобретении, включая, но не ограничиваясь перечисленными, состав оснований, информацию о гаплотипе (т.е. фазе), мутации de novo, модификации оснований (например, 4mC (N4-метилцитозин), 5hmC (5-гидроксиметилцитозин), 5fC (5-формилцитозин), 5caC (5-карбоксицитозин), 1mA (N1-метиладенин), 3mA (N3-метиладенин), 7mA (N7-метиладенин), 3mC (N3-метилцитозин), 2mG (N2-метилгуанин), 6mG (O6-метилгуанин), 7mG (N7-метилгуанин), 3mT (N3-метилтимин), 4mT (O4-метилтимин) и 8oxoG (8-оксогуанин). Согласно вариантам реализации концы молекул ДНК в образце ДНК сначала дефосфорилируют, что делает их нечувствительными к прямому лигированию с адаптерами секвенирования. Затем с помощью направляющих РНК (крРНК) белок Cas9 нацеливается на длинные молекулы ДНК, представляющие интерес, для создания двухцепочечных разрезов. Длинные молекулы ДНК, представляющие интерес, фланкированные двухцепочечными разрезами с обеих сторон, затем будут лигированы с адаптерами секвенирования, которые указываются в выбранной платформе секвенирования. Согласно другому варианту реализации ДНК можно обрабатывать экзонуклеазой, чтобы разрушить молекулы ДНК, не связанные белками Cas9 (Stevens et al. PLOS One 2019;14(4):e0215441). Поскольку эти способы не включают ПЦР-амплификацию, исходные молекулы ДНК с модификацией оснований можно секвенировать и определить модификацию основания.

Согласно вариантам реализации эти способы можно применять для нацеливания на большое количество длинных молекул ДНК, имеющих общие гомологичные последовательности, путем конструирования направляющих РНК со ссылкой на референсный геном, такой как референсный геном человека (hg19), например, длинные диспергированные ядерные повторы (LINE). В одном примере такой анализ может быть использован для анализа циркулирующей внеклеточной ДНК в материнской плазме для детектирования анеуплоидий плода (Kinde et al. PLOS One 2012;7(7):e41162). Согласно

вариантам реализации дезактивированный или «мертвый» Cas9 (dCas9) и связанная с ним одиночная направляющая РНК (онРНК) могут быть использованы для обогащения целевой длинной ДНК без разрезания двухцепочечных молекул ДНК. Например, 3′-конец онРНК может быть сконструирован так, чтобы нести дополнительную универсальную короткую последовательность. Можно использовать биотинилированные одноцепочечные олигонуклеотиды, комплементарные этой универсальной короткой последовательности, для захвата этих целевых длинных молекул ДНК, связанных dCas9. Согласно другому варианту реализации для облегчения обогащения можно использовать биотинилированный белок dCas9 или онРНК, или и то, и другое.

Согласно вариантам реализации можно выполнять отбор по размеру для обогащения длинных фрагментов ДНК, не ограничиваясь одной или более конкретными геномными областями, представляющими интерес, с использованием подходов, включающих, но не ограничивающихся перечисленными, химические, физические, ферментативные методы, методы на основе геля и методы на основе магнитных гранул, или методы, которые сочетают больше подходов, чем указано. Согласно другим вариантам реализации можно использовать иммунопреципитацию для обогащения фрагментов ДНК с определенным профилем метилирования, например, опосредуемую использованием антител к метилцитозину и белков, связывающих метилированные молекулы. Профиль метилирования связанной или захваченной ДНК может быть определен с использованием секвенирования без учета метилирования.

F. Общие принципы анализа наследственности плода на основе длинных молекул ДНК плазмы

На фиг. 17 проиллюстрировано определение генетических/эпигенетических нарушений в молекуле ДНК плазмы с информацией о происхождении от матери и от плода. Происхождение длинной молекулы ДНК плазмы от плода или от матери у беременной женщины может быть определено в соответствии с генетическим и/или эпигенетическим профилем сайтов CpG в целой молекуле или в ее части [т.е. область (a)]. Генетическая информация может представлять собой, но не ограничивается этим, информацию о последовательности, однонуклеотидных полиморфизмах, вставках, делециях, тандемных повторах, сателлитной ДНК, микросателлите, минисателлите, инверсиях и т.д. Эпигенетическая информация может представлять собой статус метилирования одного или более сайтов CpG, а также их относительные порядки в молекуле ДНК плазмы. Согласно другому варианту реализации эпигенетическая информация может представлять собой модификацию любого из A, C, G или T. Длинная ДНК плазмы с информацией о тканевом происхождении может применяться для

неинвазивного пренатального тестирования путем определения наличия генетических и/или эпигенетических нарушений в такой длинной молекуле ДНК плазмы [т.е. область (b)].

На фиг. 18 проиллюстрирована идентификация аберрантных фрагментов плода. Например, было установлено, что длинный фрагмент ДНК происходит от плода на основании профилей метилирования области (a) в соответствии с настоящим изобретением. На основании такой молекулы, происходящей от плода, можно определить вероятность того, что плод поражен генетическим или эпигенетическим нарушением. Генетические нарушения могут включать однонуклеотидные варианты, вставки, делеции, тандемные повторы, сателлитную ДНК, микросателлит, минисателлит, инверсии и т.д. Примеры генетических нарушений включают, но не ограничиваются перечисленными: бета-талассемию, альфа-талассемию, серповидно-клеточную анемию, муковисцидоз, сцепленные с полом генетические нарушения (например, гемофилию, мышечную дистрофию Дюшенна), спинальную мышечную атрофию, врожденную гиперплазию коры надпочечников и т.д. Эпигенетические нарушения могут представлять собой аберрантные уровни метилирования ДНК, например, увеличение (например, гиперметилирование) или уменьшение метилирования (гипометилирование). Примеры эпигенетических нарушений включают, но не ограничиваются перечисленными, синдром ломкой Х-хромосомы, синдром Ангельмана, синдром Прадера-Вилли, плече-лопаточно-лицевую дистрофию (FSHD), иммунодефицит, синдром центромерной нестабильности и лицевых аномалий (ICF) и т.д. Может быть установлено, что генетическое или эпигенетическое нарушение присутствует в области (b).

G. Улучшение точности секвенирования

Точность секвенирования может улучшиться при ридах последовательностей длинных фрагментов внеклеточной ДНК. На фиг. 11B из 7 аллелей в специфической для плода длинной молекуле ДНК был 1 аллель, который, по-видимому, не совпадал между секвенированиями PacBio и Illumina.

На фиг. 19A-19G показаны иллюстрации исправления ошибки генотипирования внеклеточной ДНК с использованием секвенирования PacBio. Мы визуализировали результаты выравнивания подридов для этих 7 сайтов на фиг. 11B. В 1-й строке указаны геномные координаты; во 2-й строке указана референсная последовательность. В 3-й и последующих строках указаны выравненные подриды. Например, на фиг. 19A эту область пересекают 8 подридов. «.» представляет основание, идентичное референсному основанию в цепи по Уотсону. «,» представляет основание, идентичное референсному основанию в цепи по Крику. «Алфавитная буква» представляет альтернативный аллель.

«*» представляет вставку/делецию (indel). Можно видеть, что в несовпадающем сайте, показанном на фиг. 19F, основное основание определено как «T» в консенсусной последовательности. Однако из 9 подридов в этом сайте (фиг. 19F) только 5 из 9 подридов (т.е. доля основного аллеля (MAF) 56%) были определены как «T», в то время как остальные были определены как «C». Доля основного аллеля этого сайта (фиг. 19F) была ниже, чем в других сайтах (фиг. 19A-E и фиг. 19G) (диапазон MAF: 67–89%). Таким образом, если установить строгие критерии для определения составов оснований для каждого сайта в консенсусной последовательности, например, используя MAF по меньшей мере 60%, этот ошибочный сайт будет исключен из последующей интерпретации. С другой стороны, такой ошибочный сайт оказался в пределах гомополимера (т.е. ряда последовательных одинаковых оснований «TTTTTTT»). Согласно вариантам реализации можно установить критерий, по которому варианты в пределах гомополимера были помечены как несоответствующие требованиям контроля качества и временно не использовались для последующего анализа. Согласно вариантам реализации можно применять различные требования к качеству картирования и качеству определения оснований для исправления или фильтрации оснований или подридов с низким качеством, чтобы улучшить анализ состава оснований.

С дополнительными улучшениями точности секвенирования для нанопорового секвенирования варианты реализации настоящего изобретения также можно применять с такой усовершенствованной платформой секвенирования и тем самым получить улучшенную точность.

H. Примерные способы

Длинные фрагменты внеклеточной ДНК могут быть секвенированы из биологических образцов, полученных от беременных женщин, с фрагментами внеклеточной ДНК. Эти длинные фрагменты внеклеточной ДНК можно применять для определения наследования гаплотипа плодом.

1. Секвенирование длинных фрагментов внеклеточной ДНК

На фиг. 20 показан способ 2000 анализа биологического образца беременного организма. Биологический образец может включать множество молекул внеклеточной нуклеиновой кислоты. Биологический образец может представлять собой любой биологический образец, описанный в настоящем документе. Более 20% молекул внеклеточной нуклеиновой кислоты в биологическом образце имеют размеры более 200 нт. (нуклеотидов).

В блоке 2010 секвенируют множество из множества молекул внеклеточной нуклеиновой кислоты. Секвенирование может быть выполнено с использованием

методики одномолекулярного секвенирования в реальном времени. Согласно некоторым вариантам реализации секвенирование может быть выполнено с использованием нанопоры.

Более 20% из множества секвенированных молекул внеклеточной нуклеиновой кислоты могут иметь длины более 200 нт. Согласно некоторым вариантам реализации 15-20%, 20-25%, 25-30%, 30-35% или более 35% из множества секвенированных молекул внеклеточной нуклеиновой кислоты могут иметь длины более 200 нт.

Согласно некоторым вариантам реализации более 11% из множества секвенированных молекул внеклеточной нуклеиновой кислоты могут иметь длины более 400 нт. Согласно вариантам реализации 5-10%, 10-15%, 15-20%, 20-25% или более 25% из множества секвенированных молекул внеклеточной нуклеиновой кислоты могут иметь длины более 400 нт.

Согласно некоторым вариантам реализации более 10% из множества секвенированных молекул внеклеточной нуклеиновой кислоты могут иметь длины более 500 нт. Согласно вариантам реализации 5-10%, 10-15%, 15-20%, 20-25% или более 25% из множества секвенированных молекул внеклеточной нуклеиновой кислоты могут иметь длины более 500 нт.

Согласно некоторым вариантам реализации более 8% из множества секвенированных молекул внеклеточной нуклеиновой кислоты могут иметь длины более 600 нт. Согласно вариантам реализации 5-10%, 10-15%, 15-20%, 20-25% или более 25% из множества секвенированных молекул внеклеточной нуклеиновой кислоты могут иметь длины более 600 нт.

Согласно некоторым вариантам реализации более 6% из множества секвенированных молекул внеклеточной нуклеиновой кислоты могут иметь длины более 1 тыс. нт. Согласно вариантам реализации 3-5%, 5-10%, 10-15%, 15-20%, 20-25% или более 25% из множества секвенированных молекул внеклеточной нуклеиновой кислоты могут иметь длины более 1 тыс. нт.

Согласно некоторым вариантам реализации более 3% из множества секвенированных молекул внеклеточной нуклеиновой кислоты могут иметь длины более 2 тыс. нт. Согласно вариантам реализации 1-5%, 5-10%, 10-15%, 15-20%, 20-25% или более 25% из множества секвенированных молекул внеклеточной нуклеиновой кислоты могут иметь длины более 2 тыс. нт.

Согласно некоторым вариантам реализации более 1% из множества секвенированных молекул внеклеточной нуклеиновой кислоты могут иметь длины более 3 тыс. нт. Согласно вариантам реализации 1-5%, 5-10%, 10-15%, 15-20%, 20-25% или

более 25% из множества секвенированных молекул внеклеточной нуклеиновой кислоты могут иметь длины более 3 тыс. нт.

Согласно некоторым вариантам реализации по меньшей мере 0,9% из множества секвенированных молекул внеклеточной нуклеиновой кислоты могут иметь длины более 4 тыс. нт. Согласно вариантам реализации 0,5-1%, 1-5%, 5-10%, 10-15%, 15-20% или более 20% из множества секвенированных молекул внеклеточной нуклеиновой кислоты могут иметь длины более 4 тыс. нт.

Согласно некоторым вариантам реализации по меньшей мере 0,04% из множества секвенированных молекул внеклеточной нуклеиновой кислоты могут иметь длины более 10 тыс. нт. Согласно вариантам реализации 0,01-0,1%, 0,1%-0,5%, 0,5-1%, 1-5%, 5-10%, 10-15% или более 15% из множества секвенированных молекул внеклеточной нуклеиновой кислоты могут иметь длины более 4 тыс. нт.

Множество молекул внеклеточной нуклеиновой кислоты может включать по меньшей мере 10, 50, 100, 150 или 200 молекул внеклеточной нуклеиновой кислоты. Множество молекул внеклеточной нуклеиновой кислоты может происходить из множества различных геномных областей. Например, множество хромосомных плеч или хромосом может охватываться молекулами внеклеточных нуклеиновых кислот. По меньшей мере две из множества молекул внеклеточной нуклеиновой кислоты могут соответствовать неперекрывающимся областям.

Способ секвенирования длинных фрагментов внеклеточной ДНК может быть использован любым способом, описанным в настоящем документе. Риды секвенирования можно применять для определения анеуплоидии плода, аберрации (например, аберрации числа копий), генетической мутации или вариации или наследования родительского гаплотипа. Количество ридов последовательности может быть типичным для количества фрагментов внеклеточной ДНК.

2. Наследование гаплотипа

На фиг. 21 показан способ 2100 анализа биологического образца, полученного от субъекта женского пола, беременного плодом. Субъект женского пола может иметь первый гаплотип и второй гаплотип в первой хромосомной области. Биологический образец может включать множество молекул внеклеточной ДНК плода и субъекта женского пола. Биологический образец может представлять собой любой биологический образец, описанный в настоящем документе.

В блоке 2105 могут быть получены риды, соответствующие множеству молекул внеклеточной ДНК. Риды могут представлять собой риды последовательности. Согласно некоторым вариантам реализации способ может включать выполнение секвенирования.

В блоке 2110 могут быть измерены размеры множества молекул внеклеточной ДНК. Размеры могут быть измерены путем выравнивания одного или более ридов последовательности, соответствующих концам молекулы ДНК, с референсным геномом. Размеры могут быть измерены путем полноразмерного секвенирования молекулы ДНК и последующего подсчета количества нуклеотидов в полноразмерной последовательности. Геномные координаты самых внешних нуклеотидов можно использовать для определения длины молекулы ДНК.

В блоке 2115 может быть идентифицирован первый набор молекул внеклеточной ДНК из множества молекул внеклеточной ДНК, как имеющий размеры, превышающие значение отсечки или равные ему. Значение отсечки может представлять собой любое значение отсечки, связанное с длинной ДНК. Например, значение отсечки может включать 150 п.о., 180 п.о., 200 п.о., 250 п.о., 300 п.о., 350 п.о., 400 п.о., 450 п.о., 500 п.о., 550 п.о., 600 п.о., 650 п.о., 700 п.о., 750 п.о., 800 п.о., 850 п.о., 900 п.о., 950 п.о., 1 тыс. п.о., 1,5 тыс. п.о., 2 тыс. п.о., 2,5 тыс. п.о., 3 тыс. п.о., 4 тыс. п.о., 5 тыс. п.о., 6 тыс. п.о., 7 тыс. п.о., 8 тыс. п.о., 9 тыс. п.о., 10 тыс. п.о., 15 тыс. п.о., 20 тыс. п.о., 30 тыс. п.о., 40 тыс. п.о., 50 тыс. п.о., 60 тыс. п.о., 70 тыс. п.о., 80 тыс. п.о., 90 тыс. п.о., 100 тыс. п.о., 200 тыс. п.о., 300 тыс. п.о., 400 тыс. п.о., 500 тыс. п.о. или 1 млн. п.о.

В блоке 2120 может быть определена последовательность первого гаплотипа и последовательность второго гаплотипа из ридов, соответствующих первому набору молекул внеклеточной ДНК. Определение последовательности первого гаплотипа и последовательности второго гаплотипа может включать выравнивание ридов, соответствующих первому набору молекул внеклеточной ДНК, с референсным геномом.

Согласно некоторым вариантам реализации определение последовательности первого гаплотипа и последовательности второго гаплотипа может не включать референсный геном. Определение последовательности может включать выравнивание первого поднабора ридов со вторым поднабором ридов для идентификации другого аллеля в локусе в ридах. Способ может включать определение того, что первый поднабор ридов имеет первый аллель в локусе. Способ также может включать определение того, что второй поднабор ридов имеет второй аллель в локусе. Способ может дополнительно включать определение того, что первый поднабор ридов соответствует первому гаплотипу. Кроме того, способ может включать определение того, что второй поднабор ридов соответствует второму гаплотипу. Выравнивание может быть сходным с выравниванием, описанным на фиг. 16.

В блоке 2125 второй набор молекул внеклеточной ДНК из множества молекул внеклеточной ДНК может быть выравнен с последовательностью первого гаплотипа.

Второй набор молекул внеклеточной ДНК может иметь размеры, которые меньше значения отсечки. Второй набор молекул внеклеточной ДНК может представлять собой короткие молекулы ДНК первого гаплотипа.

В блоке 2130 третий набор молекул внеклеточной ДНК из множества молекул внеклеточной ДНК может быть выравнен с последовательностью второго гаплотипа. Третий набор молекул внеклеточной ДНК может иметь размеры, которые меньше значения отсечки. Третий набор молекул внеклеточной ДНК может представлять собой короткие молекулы ДНК второго гаплотипа.

В блоке 2135 первое значение параметра может быть измерено с использованием второго набора молекул внеклеточной ДНК. Параметр может представлять собой число молекул внеклеточной ДНК, размерный профиль молекул внеклеточной ДНК или уровень метилирования молекул внеклеточной ДНК. Значения могут представлять собой исходные значения или статистические значения (например, среднее, медиану, моду, процентиль, минимум, максимум). Согласно некоторым вариантам реализации значения могут быть нормированы к значению параметра для референсного образца, другой области, обоих гаплотипов или для других диапазонов размера.

В блоке 2140 второе значение параметра может быть измерено с использованием третьего набора молекул внеклеточной ДНК. Параметр представляет собой тот же параметр, что и для второго набора молекул внеклеточной ДНК.

В блоке 2145 первое значение может быть сравнено со вторым значением. В сравнении может использоваться степень разделения. Степень разделения может быть вычислена с использованием первого значения и второго значения. Степень разделения можно сравнить со значением отсечки. Степень разделения может представлять собой любую степень разделения, описанную в настоящем документе. Значение отсечки может быть определено по референсным образцам от беременных субъектов женского пола с эуплоидными плодами. Согласно другим вариантам реализации значение отсечки может быть определено по референсным образцам от беременных субъектов женского пола с анеуплоидными плодами. Согласно некоторым вариантам реализации значение отсечки может быть определено, предполагая анеуплоидный плод. Например, данные по референсным образцам от беременных субъектов женского пола с эуплоидными плодами могут быть скорректированы для учета увеличения или уменьшения числа копий хромосомной области при анеуплоидии. Значение отсечки может быть определено путем корректировки данных.

В 2150 вероятность наследования плодом первого гаплотипа может быть определена на основании сравнения первого значения со вторым значением. Вероятность

может быть определена на основании сравнения степени разделения со значением отсечки. Когда параметр представляет собой размерный профиль молекул внеклеточной ДНК, способ может включать определение того, что плод имеет более высокую вероятность наследования первого гаплотипа, чем второго гаплотипа, когда первое значение меньше второго значения, это указывает на то, что второй набор молекул внеклеточной ДНК характеризуется меньшим размерным профилем, чем третий набор молекул внеклеточной ДНК. Когда параметр представляет собой уровень метилирования молекул внеклеточной ДНК, способ может включать определение того, что плод имеет более высокую вероятность наследования первого гаплотипа, чем второго гаплотипа, когда первое значение меньше второго значения.

Согласно некоторым вариантам реализации способы могут включать идентификацию количества повторов подпоследовательности в риде из ридов, соответствующих первому набору молекул внеклеточной ДНК. Определение последовательности первого гаплотипа может включать определение того, что последовательность включает некоторое количество повторов подпоследовательности. Первый гаплотип может включать заболевание, связанное с повторами, которое может представлять собой любое заболевание, описанное в настоящем документе. Можно определить вероятность наследования плодом заболевания, ассоциированного с повторами. Вероятность наследования плодом заболевания, ассоциированного с повторами, может быть равна или сходна с вероятностью наследования плодом первого гаплотипа. Идентификация повторов последовательностей описана далее в настоящем изобретении, включая фиг. 16.

II. Анализ ткани происхождения с использованием метилирования

Длинные молекулы внеклеточной ДНК могут иметь несколько сайтов метилирования. Как обсуждается в настоящем изобретении, уровень метилирования длинной молекулы внеклеточной ДНК у беременной женщины можно применять для определения ткани происхождения. Кроме того, профиль метилирования, присутствующего на длинной молекуле внеклеточной ДНК, можно применять для определения ткани происхождения.

Клетки из плацентарных тканей обладают уникальными метиломными профилями по сравнению с лейкоцитами и клетками из тканей, таких как, но не ограничиваясь перечисленными, печень, легкие, пищевод, сердце, поджелудочная железа, толстая кишка, тонкий кишечник, жировые ткани, надпочечники, головной мозг и т.д. (Sun et al., Proc Natl Acad Sci USA. 2015;112:E5503-12). Профили метилирования циркулирующей ДНК плода в крови беременной матери могут быть сходны с таковыми в плаценте, что обеспечивает

возможности исследования средства для разработки неинвазивных биомаркеров, специфических для плода, которые не зависят от пола или генотипа плода. Однако бисульфитное секвенирование (например, с использованием платформ секвенирования Illumina) ДНК материнской плазмы беременных женщин может быть неспособно устанавливать различие между происходящими от плода молекулами и происходящими от матери молекулами из-за ряда ограничений: (1) ДНК плазмы может быть разрушена во время обработки бисульфитом, и обычно длинная молекула ДНК расщепляется на более короткие молекулы; (2) молекулы ДНК размером более 500 п.о. могут быть неэффективно секвенированы с использованием платформ секвенирования Illumina для последующего анализа (Tan et al, Sci Rep. 2019;9:2856).

Для анализа тканей происхождения на основе метилирования можно сосредоточиться на нескольких дифференциально метилированных областях (DMR) и использовать совокупный сигнал метилирования от нескольких молекул, связанных с DMR (Sun et al, Proc Natl Acad Sci USA. 2015;112:E5503-12), вместо профилей метилирования отдельных молекул. В ряде исследований предпринимались попытки использовать подходы, основанные на чувствительных к метилированию рестрикционных ферментах (Chan et al, Clin Chem. 2006;52:2211-8) или на ПЦР, специфичной в отношении метилирования (Lo et al, Am J Hum Genet. 1998;62:768-75), чтобы оценить вклад плаценты в пул ДНК плазмы. Однако эти исследования подходили только для анализа одного или нескольких маркеров и могут сталкиваться с проблемами при использовании для анализа молекул в масштабе всего генома. Однако эти риды были выведены из амплифицированных сигналов (т.е. амплификации на основе ПЦР во время подготовки библиотеки ДНК и мостиковой амплификации во время генерирования кластеров секвенирования в проточной кювете). Такие этапы амплификации потенциально могут привести к уклону в пользу коротких молекул ДНК, что приведет к потере информации, связанной с длинными молекулами ДНК. Кроме того, Li et al. проанализировали только те риды, связанные с DMR, которые были получены заранее (Li et al., Nuclei Acids Res. 2018;46:e89).

В настоящем изобретении мы описываем новые подходы к установлению различия между молекулами ДНК плода и матери в плазме беременных женщин на основании профиля метилирования отдельной молекулы ДНК без обработки бисульфитом и амплификации ДНК. Согласно вариантам реализации для анализа можно применять одну или более длинных молекул ДНК плазмы (например, с использованием биоинформатики и/или экспериментальных анализов для отбора по размеру). Длинная молекула ДНК может быть определена как молекула ДНК, имеющая размер по меньшей мере, но не

ограничиваясь этим, 100 п.о., 200 п.о., 300 п.о., 400 п.о., 500 п.о., 600 п.о., 700 п.о., 800 п.о., 900 п.о., 1 тыс. п.о., 2 тыс. п.о., 3 тыс. п.о., 4 тыс. п.о., 5 тыс. п.о., 10 тыс. п.о., 20 тыс. п.о., 30 тыс. п.о., 40 тыс. п.о., 50 тыс. п.о., 100 тыс. п.о., 200 тыс. п.о. и т.д. Недостаточно данных о наличии и статусе метилирования более длинных молекул внеклеточной ДНК в материнской плазме. Например, неизвестно, будет ли статус метилирования таких более длинных молекул внеклеточной ДНК отражать таковой клеточной ДНК из ткани происхождения, например, поскольку такие длинные фрагменты имеют больше сайтов, статус метилирования которых может измениться после фрагментации в организме; такое изменение может произойти, пока фрагменты циркулируют в плазме. Например, исследование показало, что статус метилирования циркулирующей ДНК коррелирует с размером фрагментов ДНК (Lun et al. Clin Chem. 2013;59:1583-94). Таким образом, неизвестна выполнимость установления ткани происхождения на основании таких более длинных молекул внеклеточной ДНК. Таким образом, подходы, применяемые для идентификации сигнатур ассоциированного с тканью метилирования, и методологии, применяемые для определения и интерпретации наличия таких тканеспецифических более длинных молекул внеклеточной ДНК, существенно отличаются от тех, которые применяются для анализа короткой внеклеточной ДНК.

В соответствии с вариантами реализации настоящего изобретения можно идентифицировать короткие и длинные молекулы ДНК и определить их биологические характеристики, включая, но не ограничиваясь перечисленными, профили метилирования, концы фрагментов, размеры и составы оснований. Короткая молекула ДНК может быть определена как молекула ДНК, имеющая размер менее, но не ограничиваясь этим, 50 п.о., 60 п.о., 70 п.о., 80 п.о., 90 п.о., 100 п.о., 200 п.о., 300 п.о. и т.д. Короткая молекула ДНК может представлять собой молекулу ДНК, которая не попадает в диапазон, который считается длинным. Мы описываем новый подход к выведению тканей происхождения для циркулирующих молекул ДНК в плазме беременных женщин. В этом новом подходе применяются профили метилирования на одной или более длинных молекулах ДНК в плазме. Чем длиннее молекула ДНК, тем большее количество сайтов CpG она, вероятно, будет содержать. Наличие множества сайтов CpG на молекуле ДНК плазмы обеспечит информацию о ткани происхождения, даже если статус метилирования любого отдельного сайта CpG может быть неинформативным для определения тканей происхождения. Такие профили метилирования в длинной молекуле ДНК могут включать статус метилирования для каждого сайта CpG, порядки статуса метилирования и расстояния между любыми двумя сайтами CpG. Статус метилирования между двумя сайтами CpG может зависеть от расстояния между двумя сайтами CpG. Когда сайты CpG в пределах определенного

расстояния (например, CpG-островок) в молекуле проявляют тканеспецифический профиль, статистическая модель может придавать больший вес этим сигналам во время анализа ткани происхождения.

На фиг. 22 схематически проиллюстрирован этот принцип. На фиг. 22 показаны профили метилирования для молекул ДНК. Показаны семь сайтов CpG для различных тканей (плацента, печень, клетки крови, толстая кишка) и шесть фрагментов ДНК плазмы A-E. Метилированные сайты CpG показаны красным, а неметилированные сайты CpG показаны зеленым. В качестве примера рассмотрим 7 сайтов CpG с различным статусом метилирования в тканях плаценты, печени, клетках крови и ткани толстой кишки. Рассмотрим сценарий, при котором ни один сайт CpG не проявляет состояния метилирования, специфического для плаценты по сравнению с другими тканями. Таким образом, ткань происхождения этих молекул ДНК плазмы A, B, C, D и E с различными размерами не может быть определена только на основании состояния метилирования в одном сайте CpG. Молекулы ДНК плазмы A и B, поскольку размеры этих двух молекул относительно малы, содержат только 3 и 4 сайта CpG, соответственно. Согласно вариантам реализации профиль метилирования в молекуле ДНК, содержащей более одного сайта CpG, может быть определен как гаплотип метилирования. Как показано на фиг. 22, молекулы ДНК плазмы A и B могут происходить либо из плаценты, либо из печени на основании их гаплотипов метилирования, поскольку плацента и печень имеют один и тот же гаплотип метилирования в указанных геномных положениях, которые соответствуют молекулам A (положения 1, 2 и 3) и B (положения 1, 2, 3 и 4). Однако, если можно получить длинные молекулы ДНК в плазме, такие как молекулы C, D и E, то на основании гаплотипа метилирования можно однозначно определить, что эти молекулы C, D и E происходят из плаценты.

Референсный профиль для ткани может быть основан на профиле метилирования из референсной ткани. Согласно некоторым вариантам реализации профиль метилирования может быть основан на нескольких ридах и/или образцах. Уровень метилирования для каждого сайта CpG (также называемый индексом метилирования, MI и описанный ниже) можно применять для определения того, метилирован ли сайт.

A. Статистические модели для профилей метилирования

Согласно вариантам реализации вероятность того, что молекула ДНК плазмы происходит из плаценты, может быть определена путем сравнения гаплотипа метилирования отдельной молекулы ДНК с профилями метилирования в ряде референсных тканей. Для такого анализа могут быть предпочтительны длинные молекулы ДНК плазмы. Длинная молекула ДНК может быть определена как молекула ДНК,

имеющая размер по меньшей мере, но не ограничиваясь перечисленными, 100 п.о., 200 п.о., 300 п.о., 400 п.о., 500 п.о., 600 п.о., 700 п.о., 800 п.о., 900 п.о., 1 тыс. п.о., 2 тыс. п.о., 3 тыс. п.о., 4 тыс. п.о., 5 тыс. п.о., 10 тыс. п.о., 20 тыс. п.о., 30 тыс. п.о., 40 тыс. п.о., 50 тыс. п.о., 100 тыс. п.о., 200 тыс. п.о. и т.д. Референсные ткани могут включать, но не ограничиваются перечисленными, плаценту, печень, легкие, пищевод, сердце, поджелудочную железу, толстую кишку, тонкий кишечник, жировые ткани, надпочечники, головной мозг, нейтрофилы, лимфоциты, базофилы, эозинофилы и т.д. Согласно вариантам реализации вероятность того, что молекула ДНК плазмы происходит из плаценты можно определить путем синергического анализа гаплотипа метилирования ДНК плазмы, определенного с помощью одномолекулярного секвенирования в реальном времени, и данных метилома, основанных на полногеномном бисульфитном секвенировании референсных тканей. В качестве примера, образцы плаценты и лейкоцитарной пленки были секвенированы до среднего 94-кратного и 75-кратного геномного охвата гаплоидного генома, соответственно, с использованием бисульфитного секвенирования всего генома. Уровень метилирования каждого сайта CpG (также называемый индексом метилирования, MI) рассчитывали на основании количества секвенированных цитозинов (т.е. метилированных, обозначенных ***C***) и количества секвенированных тиминов (т.е. неметилированных, обозначенных ***T***) по следующей формуле:

$$MI = \frac{C}{C+T} \times 100\%$$

Сайты CpG были стратифицированы на три категории на основании значений MI, выведенных из ДНК плаценты:

1. Сайты CpG категории ***A***, значения MI которых были ≥70.

2. Сайты CpG категории ***B***, значения MI которых были от 30 до 70.

3. Сайты CpG категории ***C***, значения MI которых были ≤30.

Сходным образом, значения MI в сайтах CpG, выведенные из ДНК лейкоцитарной пленки, использовали для классификации сайтов CpG на три категории:

1. Сайты CpG категории ***A***, значения MI которых были ≥70.

2. Сайты CpG категории ***B***, значения MI которых были от 30 до 70.

3. Сайты CpG категории ***C***, значения MI которых были ≤30.

В категориях использовались отсечки MI 30 и 70. Отсечки могут включать другие числа, включая 10, 20, 40, 50, 60, 80 или 90. Согласно некоторым вариантам реализации эти категории можно применять для определения референсного профиля метилирования для референсной ткани (например, для использования, как описано на фиг. 22). Сайты категории A можно считать метилированными. Сайты категории C можно считать

неметилированными. Сайты категории B можно считать неинформативными и можно не включать в референсный профиль.

Для молекулы ДНК плазмы, несущей ***n*** сайтов CpG, статус метилирования для каждого сайта CpG определяли с помощью подходов, описанных в нашем предыдущем изобретении (заявка США №16/995607). Согласно некоторым вариантам реализации статус метилирования может быть определен с помощью бисульфитного секвенирования или нанопорового секвенирования. Чтобы определить вероятность того, что молекула ДНК плазмы происходит из плаценты или материнского генома, профили метилирования этой молекулы анализировали в сочетании с предшествующей информацией о метилировании в ДНК плаценты и материнской лейкоцитарной пленки. Согласно вариантам реализации мы использовали принцип, согласно которому, если сайт CpG, определенный как метилированный (M) во фрагменте ДНК плазмы, совпадал с более высоким индексом метилирования в плаценте, такое наблюдение будет указывать на то, что эта молекула с большей вероятностью произошла из плаценты. Если сайт CpG, который был определен как метилированный (M) в молекуле ДНК плазмы, совпадал с более низким индексом метилирования в плаценте, такое наблюдение будет указывать на то, что эта молекула с меньшей вероятностью произошла из плаценты; если сайт CpG, который был определен как неметилированный (U) в ДНК плазмы, совпадал с более низким индексом метилирования в плаценте. Такое наблюдение будет указывать на то, что эта молекула с большей вероятностью произошла из плаценты. Если сайт CpG, который был определен как неметилированный (U) в ДНК плазмы, совпадал с более высоким индексом метилирования в плаценте, такое наблюдение будет указывать на то, что эта молекула с меньшей вероятностью произошла из плаценты.

Мы реализовали следующую схему оценки. Начальная оценка (***S***), отражающая вероятность того, что фрагмент ДНК плазмы происходит от плода, была установлена как 0. При сравнении статуса метилирования молекулы ДНК плазмы с предшествующей информацией о метилировании ДНК плаценты,

a. если сайт CpG на молекуле ДНК плазмы был определен как «M», а его аналог в плаценте принадлежал к категории ***A***, к ***S*** будет добавлен 1 балл (т.е. увеличение балльной оценки на 1).

b. если сайт CpG на молекуле ДНК плазмы был определен как «U», а его аналог в плаценте принадлежал к категории ***A***, из ***S*** будет вычтен 1 балл (т.е. уменьшение балльной оценки на 1).

c. если сайт CpG на молекуле ДНК плазмы был определен как «M», а его аналог в плаценте принадлежал к категории ***B***, к ***S*** будет добавлено 0,5 балла.

d.если сайт CpG на молекуле ДНК плазмы был определен как «U», а его аналог в плаценте принадлежал к категории ***B***, к ***S*** будет добавлено 0,5 балла.

e.если сайт CpG на молекуле ДНК плазмы был определен как «M», а его аналог в плаценте принадлежал к категории ***C***, из ***S*** будет вычтен 1 балл.

f. если сайт CpG на молекуле ДНК плазмы был определен как «U», а его аналог в плаценте принадлежал к категории ***C***, к ***S*** будет добавлен 1 балл.

Мы называем способы выше «сопоставление статуса метилирования».

После обработки всех сайтов CpG в молекуле ДНК плазмы получали окончательную совокупную оценку ***S(плаценты)*** для этой молекулы ДНК плазмы. Согласно вариантам реализации требовалось, чтобы количество сайтов CpG составляло по меньшей мере 30, а длина молекулы ДНК плазмы должна была составлять по меньшей мере 3 тыс. п.о. Могут быть использованы другие количества сайтов CpG и длины, включая, но не ограничиваясь ими, любые из описанных в настоящем документе.

При сравнении статуса метилирования молекулы ДНК плазмы с уровнем метилирования ДНК лейкоцитарной пленки в соответствующих сайтах будет применяться сходная схема оценки. После обработки всех сайтов CpG в молекуле ДНК плазмы получали окончательную совокупную оценку ***S(лейкоцитарной пленки)*** для этой молекулы ДНК плазмы.

Если ***S(плаценты)*** > ***S(лейкоцитарной пленки)***, определяли, что молекула ДНК плазмы происходит от плода; в ином случае определяли, что молекула ДНК плазмы происходит от матери.

Имелось 17 и 405 специфических для плода и специфических для матери молекул ДНК, которые использовали для оценки эффективности выведения происхождения молекулы ДНК плазмы от плода или от матери. Молекулы, специфические для плода, представляют собой молекулы ДНК плазмы, несущие специфические для плода аллели ОНП, в то время как молекулы ДНК, специфические для матери, представляют собой молекулы ДНК, несущие специфические для матери аллели ОНП.

На фиг. 23 показана кривая операционных характеристик приемника (ROC) для определения происхождения от плода и происхождения от матери. По оси Y показана чувствительность, а по оси X показана специфичность. Красная линия представляет эффективность установления различия между молекулами, происходящими от плода, и молекулами, происходящими от матери, с использованием способа, основанного на сопоставлении статуса метилирования согласно настоящему изобретению. Синяя линия представляет эффективность установления различия между молекулами, происходящими от плода, и молекулами, происходящими от матери, с использованием уровня

метилирования отдельной молекулы (т.е. доли сайтов CpG, которые, как определено, метилированы в молекуле ДНК). На фиг. 23 показано, что площадь под кривой операционных характеристик приемника (AUC) для способа сопоставления статуса метилирования (0,94) была значительно выше, чем площадь, основанная на уровне метилирования отдельной молекулы (0,86) (значение $P<0,0001$; критерий Делонга). Это указывает на то, что анализ профилей метилирования длинной молекулы ДНК можно применять для определения происхождения от плода или от матери.

Согласно вариантам реализации величина разности (ΔS) между ***S(плаценты)*** и ***S(лейкоцитарной пленки)*** может быть принята во внимание при определении происхождения ДНК плазмы от плода или от матери. Может потребоваться, чтобы *абсолютное значение* $\Delta \boldsymbol{S}$ превышало определенный порог, например, но не ограничиваясь этим, 5, 10, 20, 30, 40, 50 и т.д. В качестве иллюстрации, когда мы использовали 10 в качестве порога $\Delta \boldsymbol{S}$, значение прогностической ценности положительного результата (PPV) при детектировании молекул ДНК плода улучшилось до 91,67% с 14,95%.

Согласно вариантам реализации на статус метилирования сайта CpG будет влиять статус метилирования соседних с ним сайтов CpG. Чем меньше нуклеотидное расстояние между любыми двумя сайтами CpG в молекуле ДНК, тем более вероятно, что два сайта CpG будут иметь одинаковый статус метилирования. Это явление было названо кометилированием. Сообщалось о ряде тканеспецифических метилирований CpG-островков; следовательно, в некоторых статистических моделях для анализа ткани происхождения больший вес будет присвоен плотным кластерам сайтов CpG (например, CpG-островкам), имеющим одинаковый статус метилирования. Для сценариев «a» и «f», если текущий рассматриваемый сайт CpG располагался на геномном расстоянии не более 100 п.о. относительно предыдущего сайта CpG и результаты способа сопоставления статуса метилирования были идентичными для этих двух последовательных сайтов CpG, к оценке ***S*** для текущего сайта CpG будет добавлен 1 дополнительный балл. Для сценариев «b» и «e», если текущий рассматриваемый сайт CpG располагался на геномном расстоянии не более 100 п.о. относительно предыдущего сайта CpG и результаты способа сопоставления статуса метилирования были идентичными для этих двух последовательных сайтов CpG, из оценки ***S*** для текущего сайта CpG будет вычтен 1 дополнительный балл. Однако если текущий рассматриваемый сайт CpG располагался на геномном расстоянии не более 100 п.о. относительно предыдущего сайта CpG, но результаты способа сопоставления статуса метилирования для этих двух последовательных сайтов CpG не совпадали, будет использована вышеупомянутая схема оценки по умолчанию. С другой стороны, если текущий рассматриваемый сайт CpG

располагался на геномном расстоянии более 100 п.о. относительно предыдущего сайта CpG, будет использована вышеупомянутая схема оценки с параметрами по умолчанию. Могут быть использованы баллы, отличные от 1, и расстояния, отличные от 100 п.о., включая любые, описанные в настоящем документе.

Согласно другим вариантам реализации сайты CpG были стратифицированы более чем на три категории на основании значений MI, выведенных из ДНК плаценты и лейкоцитарной пленки. Предварительную информацию о метилировании референсных тканей можно вывести на основании одномолекулярного секвенирования в реальном времени (т.е. нанопорового секвенирования и/или секвенирования PacBio SMRT). Может потребоваться, чтобы длина молекулы ДНК плазмы составляла по меньшей мере, но не ограничиваясь перечисленными, 100 п.о., 200 п.о., 300 п.о., 400 п.о., 500 п.о., 600 п.о., 700 п.о., 800 п.о., 900 п.о., 1 тыс. п.о., 2 тыс. п.о., 3 тыс. п.о., 4 тыс. п.о., 5 тыс. п.о., 10 тыс. п.о., 20 тыс. п.о., 30 тыс. п.о., 40 тыс. п.о., 50 тыс. п.о., 100 тыс. п.о., 200 тыс. п.о. и т.д. Может потребоваться, чтобы количество сайтов CpG составляло по меньшей мере, но не ограничиваясь перечисленными, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100 и т.д.

Согласно вариантам реализации для характеристики профилей метилирования молекулы ДНК плазмы можно использовать вероятностную модель. Статус метилирования k сайтов CpG ($k \geq 1$) на молекуле ДНК плазмы обозначили как $M = (m_1, m_2, \ldots, m_k)$, где m_i равен 0 (для неметилированного статуса) или 1 (для метилированного статуса) в сайте CpG i на молекуле ДНК плазмы. Согласно вариантам реализации вероятность M, относящаяся к молекуле ДНК плазмы, происходящей из плаценты, может зависеть от референсных профилей метилирования в тканях плаценты. Референсные профили метилирования в тканях плаценты для соответствующих сайтов CpG в $1, 2, \ldots, k$ будут подчиняться бета-распределениям. Бета-распределение параметризуется двумя положительными параметрами α и β, обозначаемыми бета(α, β). Значения, полученные из бета-распределения, будут колебаться от 0 до 1. На основании данных глубокого бисульфитного секвенирования для ткани, представляющей интерес, параметры α и β определяли по количеству секвенированных цитозинов (метилированных) и тиминов (неметилированных) в каждом сайте CpG для этой конкретной ткани, соответственно. Для плаценты такое бета-распределение было обозначено как бета(α^P, β^P). Вероятность того, что молекула ДНК плазмы происходит из плаценты, P (M|плацента), будет моделироваться следующим образом:

$$P(M \mid \text{плацента}) = \prod_{i=1}^{i=p} P\left(m_i \mid \text{бета}\ (\alpha_i^p, \beta_i^p)\right)$$

где «i» обозначало $i^{й}$ сайт CpG; Бета(α_i^P,β_i^P) обозначало бета-распределение,

относящееся к профилям метилирования в $i^{м}$ сайте CpG в плаценте; ***P*** обозначало совместную вероятность наблюдаемой молекулы ДНК плазмы с данными профилями метилирования по ***k*** сайтам CpG.

Вероятность того, что молекула ДНК плазмы происходит из лейкоцитарной пленки (т.е. лейкоцитов), P(M|лейкоцитарная пленка), будет моделироваться следующим образом:

$$P(M \mid \underset{\text{пленка}}{\text{Лейкоцитарная}}) = \prod_{i=1}^{i=p} P(m_i \mid \text{бета}\ (\alpha_i^b, \beta_i^b))$$

где «*i*» обозначало *i*^{й} сайт CpG; Бета(α_i^b,β_i^b) обозначало бета-распределение, относящееся к профилям метилирования в $i^{м}$ сайте CpG в ДНК лейкоцитарной пленки. ***P*** представляло собой совокупную вероятность наблюдаемой молекулы ДНК плазмы с данными профилями метилирования по ***k*** сайтам CpG.

Бета(α_i^p,β_i^p) и Бета(α_i^b,β_i^b) можно было определить по результатам полногеномного бисульфитного секвенирования ДНК плаценты и лейкоцитарной пленки, соответственно.

Для молекулы ДНК плазмы, если наблюдали P (M | Плацента)>P (M | лейкоцитарной пленки), такая молекула ДНК плазмы, вероятно, происходила из плаценты; в ином случае она, вероятно, происходила из лейкоцитарной пленки. Используя эту модель, мы достигли AUC 0,79.

B. Модели машинного обучения

Согласно другим вариантам реализации можно применять алгоритм машинного обучения для определения происхождения конкретной молекулы ДНК плазмы от плода или от матери. Чтобы проверить выполнимость применения подхода, основанного на машинном обучении, для классификации молекул ДНК плода и матери у беременных женщин, мы разработали графическое представление профилей метилирования для молекулы ДНК плазмы.

На фиг. 24 показано определение парных профилей метилирования. На молекуле ДНК плазмы показаны девять сайтов CpG. Метилированные сайты CpG показаны красным, а неметилированные сайты CpG показаны зеленым. Когда два сайта CpG в паре имели одинаковый статус метилирования (например, 1$^{-й}$ CpG и 5$^{-й}$ CpG), пара будет кодироваться как 1, как показано в положении, указанном стрелкой «a». Когда два сайта CpG в паре имели разный статус метилирования (например, 1$^{-й}$ CpG и 2^{й} CpG), пара будет кодироваться как 0, как показано в положении, указанном стрелкой «b». Те же самые правила кодирования применялись ко всем парам любых 2 сайтов CpG на молекуле ДНК.

В качестве примера мы использовали молекулу ДНК плазмы, содержащую 9 сайтов CpG. Профиль метилирования для этой молекулы ДНК плазмы определяли с помощью подходов, описанных в нашем предыдущем изобретении (заявка США №

16/995607), т.е. U-M-M-M-U-U-U-M-M (U и M представляли неметилированный CpG и метилированный CpG, соответственно). Попарное сравнение статуса метилирования между любыми двумя сайтами CpG можно применять для анализа на основе машинного обучения или глубокого обучения. В этом примере те же самые правила были применены в общей сложности к 36 парам. Если на молекуле ДНК плазмы было в общей сложности ***n*** сайтов CpG, то было бы n*(n-1)/2 пар сравнения. Можно использовать различное количество сайтов CpG, включая 5, 6, 7, 8, 10, 11, 12, 13 и т.д. Если молекула включает большее количество сайтов, чем используется в модели машинного обучения, то для разделения сайтов на соответствующее количество сайтов может быть использовано скользящее окно.

Мы получили одну или более молекул из образцов ДНК плаценты и лейкоцитарной пленки, соответственно. Профили метилирования для этих молекул ДНК определяли с помощью одномолекулярного секвенирования в реальном времени (SMRT) Pacific Bioscience (PacBio) в соответствии с подходами, описанными в нашем предыдущем изобретении (заявка США № 16/995607). Эти профили метилирования были преобразованы в парные профили метилирования.

Парные профили метилирования, связанные с ДНК плаценты, и те, которые связаны с ДНК лейкоцитарной пленки, использовали для обучения сверточной нейронной сети (CNN) для различения молекул, потенциально происходящих от плода и происходящих от матери. Каждому целевому выходу (т.е. аналогичному значению зависимой переменной) для фрагмента ДНК из плаценты присваивали значение «1», в то время как каждому целевому выходу для фрагмента ДНК из лейкоцитарной пленки присваивали значение «0». Парные профили метилирования использовали для обучения, чтобы определить параметры (часто называемые весами) для модели CNN. Оптимальные параметры CNN для различения происхождения фрагмента ДНК от плода или от матери были получены, когда общая ошибка прогнозирования между выходными оценками, рассчитанными с помощью сигмоидной функции, и желаемыми целевыми выходами (двоичные значения: 0 или 1) достигала минимума путем итеративной корректировки параметров модели. Общую ошибку прогнозирования измеряли с помощью сигмоидной функции потери кросс-энтропии в алгоритмах глубокого обучения (https://keras.io/). Параметры модели, полученные из обучающих наборов данных, использовали для анализа молекулы ДНК (например, молекулы ДНК плазмы) для получения на выходе вероятностной оценки, которая указывала бы на вероятность происхождения молекулы ДНК из плаценты или из лейкоцитарной пленки. Если вероятностная оценка фрагмента ДНК плазмы превышала определенный порог, считалось, что такая молекула ДНК плазмы

происходит от плода. В ином случае будет считаться, что она происходит от матери. Порог будет включать, но не ограничивается этим, 0,1, 0,2, 0,3, 0,4, 0,5, 0,6, 0,7, 0,8, 0,9, 0,95, 0,99 и т.д. В одном примере, используя эту модель CNN, мы достигли AUC 0,63 для определения, произошла ли молекула ДНК плазмы от плода или от матери, это указывает на то, что при использовании алгоритмов глубокого обучения можно вывести ткани происхождения молекул ДНК из материнской плазмы. Благодаря получению большего количества результатов одномолекулярного секвенирования в реальном времени эффективность алгоритма глубокого обучения может быть дополнительно улучшена.

Согласно некоторым другим вариантам реализации статистические модели могут включать, но не ограничиваются перечисленными, линейную регрессию, логистическую регрессию, глубокую рекуррентную нейронную сеть (например, долгая-краткосрочная память, LSTM), байесовский классификатор, скрытую модель Маркова (HMM), линейный дискриминантный анализ (LDA), кластеризацию k-средних, плотностный алгоритм кластеризации пространственных данных с присутствием шума (DBSCAN), алгоритм случайного леса и метод опорных векторов (SVM). Будут задействованы различные статистические распределения, включая, но не ограничиваясь перечисленными, биномиальное распределение, распределение Бернулли, гамма-распределение, нормальное распределение, распределение Пуассона и т.д.

C. Гаплотипы метилирования, специфические для плаценты

Статус метилирования каждого сайта CpG на отдельной молекуле ДНК может быть определен с использованием подходов, описанных в нашем предыдущем изобретении (заявка США № 16/995607), или любой методики, описанной в настоящем документе. Помимо уровня метилирования отдельной двухцепочечной молекулы ДНК можно определить профиль метилирования отдельной молекулы для каждой молекулы ДНК, который может представлять собой последовательность статуса метилирования соседних сайтов CpG вдоль отдельной молекулы ДНК.

Различные сигнатуры метилирования ДНК можно найти в разных типах тканей и клеток. Согласно вариантам реализации можно вывести ткань происхождения отдельных молекул ДНК плазмы на основании их профилей метилирования отдельных молекул.

Геномную ДНК из десяти образцов лейкоцитарной пленки и шести образцов плацентарной ткани секвенировали с использованием SMRT-секвенирования (PacBio). Путем объединения картированных высококачественных ридов секвенирования кольцевых консенсусных последовательностей (CCS) из каждого типа образцов мы смогли достичь 58,7-кратного и 28,7-кратного охвата ДНК лейкоцитарной пленки и ДНК плаценты, соответственно.

Используя подход скользящего окна, геном был разделен приблизительно на 28,2 миллиона перекрывающихся окон 5 сайтов CpG. Согласно другим вариантам реализации можно использовать различные размеры окна, такие как, но не ограничиваясь ими, 2, 3, 4, 5, 6, 7 и 8 сайтов CpG. Также можно использовать подход с неперекрывающимися окнами. Каждое окно считалось потенциальной маркерной областью. Для каждой потенциальной маркерной области мы идентифицировали преобладающий профиль метилирования отдельной молекулы среди всех секвенированных молекул ДНК плаценты, которые охватывают все 5 сайтов CpG в этой маркерной области. Сравнения могут быть проведены между сайтами CpG молекулы ДНК плазмы и соответствующими сайтами CpG отдельных молекул ДНК референсных тканей. Затем мы рассчитали оценку несоответствия для каждой молекулы ДНК лейкоцитарной пленки, охватывающей все сайты CpG в одной и той же маркерной области, путем сравнения ее профиля метилирования отдельной молекулы с преобладающим профилем метилирования отдельной молекулы в плаценте.

Оценка несоответствия = Кол-во несоответствующих сайтов CpG/общее кол-во сайтов CpG,

где количество несоответствующих сайтов CpG относится к количеству сайтов CpG, показывающих другой статус метилирования в молекуле ДНК лейкоцитарной пленки по сравнению с преобладающим профилем метилирования отдельной молекулы в плаценте.

Более высокая оценка несоответствия указывает на то, что профиль метилирования молекулы ДНК лейкоцитарной пленки больше отличается от преобладающего профиля метилирования отдельной молекулы в плаценте. Из 28,2 миллиона потенциальных маркерных областей мы отобрали те, которые показали существенное различие в профиле метилирования отдельной молекулы между пулами молекул ДНК из плаценты и лейкоцитарной пленки, используя следующие критерии: a) более 50% молекул ДНК плаценты имели преобладающий профиль метилирования отдельной молекулы; и b) более 80% молекул ДНК лейкоцитарной пленки имели оценку несоответствия более 0,3. На основании этих критериев мы отобрали 281566 маркерных областей для последующего анализа.

На фиг. 25 приведена таблица распределения отобранных маркерных областей среди различных хромосом. В первом столбике показано число хромосом. Во втором столбике показано число маркерных областей в хромосоме.

Настоящим мы иллюстрируем нашу концепцию классификации ткани происхождения для отдельных молекул ДНК плазмы, основанную на профилях

метилирования отдельных молекул с использованием молекул ДНК плазмы, секвенированных с помощью SMRT-секвенирования, которые охватывали либо специфический для плода аллель, либо специфический для матери аллель, как описано ранее в настоящем изобретении. Любая молекула ДНК плазмы, охватывающая отобранную маркерную область с профилем метилирования, идентичным преобладающему профилю метилирования отдельной молекулы в плаценте, будет классифицирована как специфическая для плаценты (т.е. специфическая для плода) молекула ДНК. Напротив, если профиль метилирования отдельной молекулы для молекулы ДНК плазмы не идентичен преобладающему профилю метилирования отдельной молекулы в плаценте, мы будем классифицировать эту молекулу как неспецифическую для плаценты. Правильная классификация в этом анализе была определена таким образом, что специфическая для плода молекула ДНК была идентифицирована как происходящая от плода (т.е. специфическая для плаценты), а молекула ДНК матери была идентифицирована как не происходящая от плода (т.е. неспецифическая для плаценты) в зависимости от того, присутствовали ли в этой молекуле гаплотипы метилирования, специфические для плаценты. Предшествующие способы анализа ткани происхождения, основанные на метилировании, обычно включали деконволюцию процентного или долевого вклада ряда тканевых источников внеклеточной ДНК в биологическом образце. Преимущество способа согласно настоящему изобретению по сравнению с предшествующими способами заключается в том, что доказательства вклада внеклеточной ДНК ткани в биологический образец, например, ДНК плацентарного происхождения в материнской плазме, могут быть определены независимо от наличия или отсутствия вкладов из других тканей. Кроме того, плацентарное происхождение любой молекулы внеклеточной ДНК может быть определено с помощью способа согласно настоящему изобретению без учета относительного вклада молекул внеклеточной ДНК из этой ткани.

Из 28 молекул ДНК, охватывающих специфический для плода аллель, 17 (61%) были классифицированы как специфические для плаценты, и 11 (39%) были классифицированы как неспецифические для плаценты. С другой стороны, из 467 молекул ДНК, охватывающих специфический для матери аллель, 433 (93%) были классифицированы как неспецифические для плаценты, и 34 (7%) были классифицированы как специфические для плаценты.

Согласно вариантам реализации можно применять различные процентные содержания молекул ДНК лейкоцитарной пленки, имеющих оценку несоответствия более 0,3 в качестве порога, включая, но не ограничиваясь перечисленными, более 60%, 70%,

75%, 80%, 85% и 90% и т.д. Общую точность классификации плацентарного или неплацентарного происхождения ДНК плазмы у беременных субъектов можно повысить путем корректировки критериев, используемых при отборе маркерной области. Это особенно важно в условиях неинвазивного пренатального тестирования, когда предпринимаются попытки определить наличие у плода вызывающей заболевание мутации или аберрации числа копий.

На фиг. 26 приведена таблица классификации молекул ДНК плазмы на основе их профилей метилирования отдельных молекул с использованием различного процентного содержания молекул ДНК лейкоцитарной пленки, имеющих оценку несоответствия более 0,3 в качестве критериев отбора маркерных областей. В первом столбике показано процентное содержание молекул ДНК лейкоцитарной пленки, имеющих оценку несоответствия более 0,3%. Во втором столбике молекулы ДНК разделены на те, которые охватывают специфический для плода аллель, и те, которые охватывают специфический для матери аллель. В третьем и четвертом столбиках показана классификация молекул ДНК как специфических для плаценты или неспецифических для плаценты на основе профиля метилирования отдельной молекулы. В пятом столбике показано процентное содержание молекул ДНК, которые были классифицированы так же, как и конкретный аллель во втором столбике.

На фиг. 27 показана схема способа применения специфического для плаценты гаплотипа метилирования для определения наследования у плода неинвазивным способом. Как показано на фиг. 27, внеклеточная ДНК из плазмы беременной женщины была экстрагирована для одномолекулярного секвенирования в реальном времени. Длинные молекулы ДНК плазмы идентифицировали в соответствии с вариантами реализации настоящего изобретения. Статус метилирования в каждом сайте CpG для каждой длинной молекулы ДНК плазмы определяли в соответствии с вариантами реализации настоящего изобретения. Гаплотип метилирования каждой длинной молекулы ДНК плазмы определяли в соответствии с вариантами реализации настоящего изобретения. Если длинная молекула ДНК плазмы была идентифицирована как несущая гаплотип метилирования, специфический для плаценты, генетическая и эпигенетическая информация, связанная с этой молекулой, считалась бы унаследованной плодом. Согласно вариантам реализации, если было определено, что одна или более длинных молекул ДНК плазмы, содержащих мутацию, вызывающую заболевание, которая аналогична мутации, вызывающей заболевание, носителем которой является беременная женщина, происходит от плода на основании информации о гаплотипе метилирования в соответствии с вариантами реализации настоящего изобретения, это указывало бы на то, что плод

унаследовал мутацию от матери.

Варианты реализации могут быть применены к генетическим заболеваниям, включая, но не ограничиваясь перечисленными, бета-талассемию, серповидно-клеточную анемию, альфа-талассемию, муковисцидоз, гемофилию A, гемофилию B, врожденную гиперплазию надпочечников, мышечную дистрофию Дюшенна, мышечную дистрофию Беккера, ахондроплазию, танатофорную дисплазию, болезнь фон Виллебранда, синдром Нунан, наследственную тугоухость и глухоту, различные врожденные нарушения обмена веществ (например, цитруллинемию типа I, пропионовую ацидемию, болезнь накопления гликогена типа Ia (болезнь фон Гирке), болезнь накопления гликогена типа Ib/c (болезнь фон Гирке), болезнь накопления гликогена типа II (болезнь Помпе), мукополисахаридоз (MPS) типа I (Гурлер/Гурлер–Шейе/Шейе), MPS типа II (синдром Хантера), MPS типа IIIA (синдром Санфилиппо A), MPS типа IIIB (синдром Санфилиппо B), MPS типа IIIC (синдром Санфилиппо C), MPS типа IIID (синдром Санфилиппо D), MPS типа IVA (синдром Моркио A), MPS типа IVB (синдром Моркио B), MPS типа VI (синдром Марото-Лами), MPS типа VII (синдром Слая), муколипидоз II (I-клеточная болезнь), метахроматическую лейкодистрофию, ганглиозидоз GM1, дефицит OTC (X-сцепленный дефицит орнитинтранскарбамилазы), адренолейкодистрофию (X-сцепленный ALD), болезнь Краббе (глобоидно-клеточную лейкодистрофию)) и т.п.

Согласно другим вариантам реализации генетическое заболевание у плода может быть связано с метилированием ДНК *de novo* в геноме плода, которое отсутствовало в родительских геномах. Примером может служить гиперметилирование гена *регулятора трансляции 1 FMRP* (*FMR1*) у плода с синдромом ломкой X-хромосомы. Синдром ломкой X-хромосомы вызывается распространением тринуклеотидного повтора CGG в 5′-нетранслируемой области гена *FMR1*. Нормальный аллель будет содержать приблизительно от 5 до 44 копий повтора CGG. Премутационный аллель будет содержать от 55 до 200 копий повтора CGG. Аллель полной мутации будет содержать более 200 копий повтора CGG.

На фиг. 28 проиллюстрирован принцип неинвазивного пренатального детектирования синдрома ломкой X-хромосомы у плода мужского пола непораженной беременной женщины, несущей либо нормальный аллель, либо премутационный аллель. На фиг. 28 «n» представляет число копий CGG в материнском геноме; «m» представляет число копий CGG в геноме плода. Геном непораженной беременной женщины будет нести гены *FMR1*, которые содержат не более 200 копий повторов CGG (т.е. $n \leq 200$) и неметилированы. Напротив, геном плода мужского пола, пораженного синдромом ломкой X-хромосомы, будет нести ген *FMR1*, который содержит более 200 копий повторов CGG

(m > 200) и метилирован. Путем выполнения одномолекулярного секвенирования ДНК материнской плазмы можно идентифицировать ряд длинных молекул ДНК из представляющей интерес геномной области (например, гена *FMR1*), число повторов и статус метилирования которых могут быть определены одновременно. При идентификации одной или более молекул ДНК, охватывающих ген *FMR1*, содержащих более 200 копий повторов CGG и метилированных, в плазме непораженной женщины, это может указывать на то, что у плода, вероятно, будет синдром ломкой Х-хромосомы. В еще одном варианте реализации можно дополнительно установить происхождение от плода таких молекул ДНК плазмы с использованием специфических для плаценты гаплотипов метилирования в соответствии с вариантами реализации настоящего изобретения. При идентификации одной или более молекул, содержащих одну или более областей в молекуле, которые несли специфические для плаценты гаплотипы метилирования, и такие молекулы охватывали ген *FMR1*, содержали более 200 копий повторов CGG и были метилированы, можно с большей уверенностью заключить, что плод имеет синдром ломкой Х-хромосомы. Напротив, при идентификации одной или более молекул, которые несли специфические для плаценты гаплотипы метилирования, и такие молекулы охватывали ген *FMR1*, содержали менее 200 копий повтора CGG и не были метилированы, это может указывать на то, что плод, вероятно, не поражен. При синдроме ломкой Х-хромосомы полная мутация (> 200 повторов) фактически приводит к метилированию всего гена и отключению функции гена. Таким образом, в частности, для ломкой Х-хромосомы детектирование длинного метилированного аллеля (в отличие от показа плацентарного профиля метилирования) может с большой вероятностью свидетельствовать о наличии у плода заболевания.

Детектирование генетических нарушений может быть выполнено с учетом информации о предшествующем статусе матери или без него. Женщины с премутацией могут не иметь никаких симптомов, но у некоторых могут быть легкие симптомы, о которых часто можно узнать только задним числом. Если нам неизвестен мутационный статус матери, один из подходов состоит в том, чтобы детектировать длинный аллель в плазме от женщины, у которой, очевидно, нет заболевания, или проанализировать лейкоцитарную пленку матери и определить, что в ней нет такого длинного аллеля. В качестве другого подхода мы можем объединить длину повтора со статусом метилирования молекулы вкДНК. Если статус метилирования указывает на профиль плода (гаплотип метилирования) и показывает длинный аллель, то плод, вероятно, поражен. Этот подход применим ко многим тринуклеотидным нарушениям, например, к болезни Хантингтона.

D. Неинвазивное конструирование генома плода с помощью длинных молекул ДНК плазмы

Профили метилирования можно применять для определения наследования гаплотипов. Определение наследования гаплотипов с помощью качественного подхода с использованием профилей метилирования может быть более эффективным, чем количественный метод, характеризующий количества определенных фрагментов. Профили метилирования можно применять для определения наследования гаплотипов по материнской и отцовской линии.

1. Наследование по материнской линии плода

Lo et al. продемонстрировали выполнимость построения полногеномной генетической карты и определения мутационного статуса плода по последовательностям ДНК материнской плазмы с использованием информации о родительских гаплотипах (Lo et al. Sci Transl Med. 2010;2:61ra91). Эта технология была названа анализом относительной дозы гаплотипов (RHDO) и является одним из подходов к решению вопроса о наследовании по материнской линии плода. Принцип был основан на том факте, что материнский гаплотип, унаследованный плодом, будет относительно сверхпредставлен в ДНК плазмы беременной женщины по сравнению с другим материнским гаплотипом, который не передается плоду. Таким образом, RHDO является количественным аналитическим методом.

Согласно вариантам реализации настоящего изобретения профили метилирования в длинной молекуле ДНК плазмы применяют для определения тканей происхождения этой молекулы ДНК плазмы. Согласно одному варианту реализации настоящего изобретения предложен качественный анализ наследования по материнской линии плода.

На фиг. 29 показан пример определения наследования по материнской линии плода. Геномное положение ***P*** было гетерозиготным в материнском геноме (A/G). Закрашенный круг обозначает метилированный сайт, а незакрашенный круг обозначает неметилированный сайт. Профиль метилирования в плаценте представлял собой «-M-U-M-M-», где «M» представляет собой метилированный цитозин, а «U» представляет собой неметилированный цитозин в сайте CpG. Согласно одному варианту реализации профиль метилирования в плаценте и соответствующих референсных тканях можно получить из данных, сгенерированных ранее в результате секвенирования (например, одномолекулярного секвенирования в реальном времени и/или бисульфитного секвенирования). Было обнаружено, что в ДНК плазмы одна неотцовская ДНК плазмы (обозначенная Z), несущая аллель A в этом конкретном геномном локусе, проявляет

профиль метилирования («-M-U-M-M-»), совместимый с профилем метилирования в плаценте, в отличие от профилей метилирования других тканей. Не были обнаружены молекулы, несущие аллель G, проявляющий профиль метилирования, совместимый с профилями метилирования в плаценте. Таким образом, на основании аллеля A и наличия профиля метилирования «-M-U-M-M-» можно определить, что плод наследует материнский аллель A.

На фиг. 30 показан качественный анализ наследования по материнской линии плода с использованием генетической и эпигенетической информации о молекулах ДНК плазмы. Как показано в верхней ветви фиг. 30, ДНК плазмы экстрагировали с последующим отбором по размеру длинной ДНК в соответствии с вариантами реализации настоящего изобретения. Отобранные по размеру молекулы ДНК плазмы подвергали одномолекулярному секвенированию в реальном времени (например, с использованием системы производства Pacific Biosciences). Генетическая и эпигенетическая информация была определена в соответствии с вариантами реализации настоящего изобретения. В иллюстративных целях молекулу (X) выравнивали с хромосомой 1 человека, содержащей аллель **G** в хромосомном положении a (chr1:a) и аллель **A** в хромосомном положении e (chr1:e). Молекула X имеет аллель **C** в хромосомном положении d.

Статус метилирования CpG этой молекулы X был определен как «-M-U-M-M-», где «M» представлял собой метилированный цитозин, а «U» представлял собой неметилированный цитозин в сайте CpG. Закрашенный круг обозначает метилированный сайт, а незакрашенный круг обозначает неметилированный сайт. В результате анализа референсного образца известно, что плацентарная ДНК имеет профиль метилирования «-M-U-M-M-» в области между положениями a и e. На основании профиля метилирования молекулы X, совпадающего с профилем метилирования плацентарной ДНК, было определено, что молекула X имеет плацентарное происхождение в соответствии с вариантами реализации настоящего изобретения.

Как показано в нижней ветви фиг. 30, ДНК из материнских лейкоцитов подвергали одномолекулярному секвенированию в реальном времени. Эпигенетическую и генетическую информацию о материнских лейкоцитах получали в соответствии с вариантами реализации настоящего изобретения. Генетические аллели были фазированы на два гаплотипа, а именно материнский гаплотип I (Hap I) и материнский гаплотип II (Hap II), с использованием методов, включающих, но не ограничивающихся перечисленными, WhatsHap (Patterson et al. J Comput Biol. 2015;22:498–509), HapCUT (Bansal et al. Bioinformatics. 2008;24:i153-9), HapCHAT (Beretta et al. BMC bioinformatics. 2018;19:252) и т.д. В данном случае мы получили два гаплотипа, а именно «-A-C-G-T-»

(Hap I) и «-G-T-A-C-» (Hap II) в материнских геномах. Hap I связан с вариантом (вариантами) дикого типа, в то время как Hap II связан с вариантом (вариантами), связанным с заболеванием. Связанный с заболеванием вариант(ы) может включать, но не ограничивается ими, однонуклеотидные варианты, вставки, делеции, транслокации, инверсии, распространения повторов и/или другие генетические структурные вариации.

В случае геномного положения e материнский генотип был определен как AA, а отцовский генотип был определен как GG. На основании профиля метилирования было определено, что молекула ДНК X плазмы имеет плацентарное происхождение. На основании наличия специфического для матери аллеля A, но отсутствия специфического для отца аллеля G, соответственно, было выведено, что молекула X унаследована от одного из материнских гаплотипов.

Чтобы далее определить, какой материнский гаплотип был передан плоду, мы сравнили информацию об аллелях в геномных положениях, отличных от положения chr1:e этой молекулы X плацентарного происхождения, с материнскими гаплотипами. Например, молекула X имеет аллель G в положении a и аллель C в положении d. Наличие любого из этих аллелей в молекуле X указывает на то, что молекула X должна быть отнесена к материнскому Hap II, включающему те же аллели.

Таким образом, можно сделать вывод, что материнский гаплотип II, связанный с вариантом(ами), связанным(и) с заболеванием, был передан плоду. Было определено, что нерожденный плод находится в группе риска развития заболевания.

Качественный анализ наследования по материнской линии плода на основе профиля метилирования может потребовать меньшего количества молекул ДНК плазмы, чтобы сделать вывод о том, какой материнский гаплотип был унаследован плодом, по сравнению с RHDO, который представлял собой подход, основанный на количественном анализе. Мы выполнили анализы методом компьютерного моделирования, чтобы оценить уровень детектирования для наследования по материнской линии плода в масштабе всего генома с различными количествами молекул ДНК плазмы, используемых для анализа.

Для анализа путем моделирования RHDO N молекул ДНК плазмы были в совокупности выравнены с M гетерозиготных ОНП в гаплотипном блоке материнского генома. Фракция ДНК плода представляла собой f. Отцовские генотипы для этих соответствующих ОНП были гомозиготными и идентичными материнскому Hap I, который был передан плоду. Из N молекул ДНК плазмы среднее число молекул ДНК плазмы, выравненных с материнским Hap I, было $N\times(0,5 + f/2)$, в то время как среднее число молекул ДНК плазмы, выравненных с материнским Hap II, было бы $N\times(0,5 - f/2)$. Мы предположили, что молекулы ДНК плазмы, взятые из гаплотипов, подчинялись

биномиальному распределению.

Ряд молекул ДНК плазмы был отнесен к Hap I (т.е. X) в соответствии со следующим распределением:

$X \sim Bin(N, 0{,}5 + f/2)$ (1),

где «Bin» обозначает биномиальное распределение.

Ряд молекул ДНК плазмы был отнесен к Hap II (т.е. Y) в соответствии со следующим распределением:

$Y \sim Bin(N, 0{,}5 - f/2)$ (2).

Таким образом, молекулы ДНК плазмы, отнесенные к материнскому Hap I, будут относительно сверхпредставлены в материнской плазме по сравнению с материнским Hap II. Чтобы определить, была ли сверхпредставленность статистически значимой, мы сравнили разницу в количествах ДНК плазмы между двумя материнскими гаплотипами с нулевой гипотезой, согласно которой два гаплотипа (обозначенные X′ и Y′) были одинаково представлены в плазме.

$X' \sim Bin(N, 0{,}5)$ (3),

$Y' \sim Bin(N, 0{,}5)$ (4).

Далее мы определили относительное различие доз между двумя гаплотипами, как показано ниже:

$D=(X-Y)/N$ (5),

$D'=(X'-Y')/N$ (6).

В одном примере статистический показатель D, отражающий относительную дозу гаплотипа, сравнивали со средним значением D′ (M), нормированным по стандартному отклонению D′ (SD), как показано ниже (т.е. z-оценка):

z-оценка = (D – M)/SD (7).

Z-оценка >3 указывала на то, что Hap I был передан плоду.

Для анализа RHDO, основанного на формулах (1)–(7), мы смоделировали 30000 гаплотипных блоков по всему геному, в которых Hap I был передан плоду. Средняя длина гаплотипных блоков была 100 тыс. п.о. Каждый гаплотипный блок содержал в среднем 100 ОНП, из которых 10 ОНП могут быть информативными, внося вклад в дисбаланс гаплотипов. В одном примере фракция ДНК плода была 10%, а медиана размеров фрагментов была 150 п.о. Мы рассчитали процентное содержание гаплотипных блоков с z-оценкой >3, именуемое в настоящем документе уровнем детектирования, путем изменения количества молекул ДНК плазмы, используемых для анализа RHDO, в диапазоне от 1 миллиона до 300 миллионов. Количество молекул ДНК плазмы в настоящем документе корректировали, исходя из вероятности, что ДНК плазмы

охватывает информативный сайт ОНП в соответствии с распределением Пуассона.

Для компьютерного моделирования, связанного с качественным анализом наследования по материнской линии плода на основе профиля метилирования, мы выдвинули предположения, показанные ниже для иллюстративных целей:

1) Для анализа использовали N молекул ДНК плазмы, охватывающих гаплотипный блок в материнском геноме.

2) Вероятность того, что фрагмент ДНК плазмы, используемый для анализа ткани происхождения, имеет длину по меньшей мере 3 тыс. п.о., была обозначена a.

3) Вероятность того, что молекула ДНК плазмы несет более 10 сайтов CpG, была обозначена b.

4) Фракция ДНК плода этих фрагментов >3 тыс. п.о. была обозначена f.

Можно достигнуть точного выведения тканей происхождения для тех молекул ДНК плазмы, размер которых превышает 3 тыс. п.о. и которые имеют по меньшей мере 10 сайтов CpG, как проиллюстрировано в одном варианте реализации настоящего изобретения. Было предположено, что количество молекул ДНК плазмы, удовлетворяющих вышеуказанным критериям (Z), подчиняется распределению Пуассона со средним значением λ (т.е. $N \times a \times b \times f$).

$Z \sim$ распределение Пуассона (λ) (8).

В одном примере на основании формулы (8) мы смоделировали 30000 гаплотипных блоков, в которых Hap I был передан плоду. Средняя длина каждого гаплотипного блока составила 100 тыс. п.о. Каждый гаплотипный блок содержал в среднем 100 ОНП, из которых 20 гетерозиготных ОНП могли быть фазированы на два материнских гаплотипа. Фракция ДНК плода составила 1%. После отбора по размеру имелось 40% молекул ДНК плазмы с размерами >3 тыс. п.о. 87,1% молекул ДНК плазмы с размерами >3 тыс. п.о. несли по меньшей мере 10 сайтов CpG. Процентное содержание гаплотипных блоков со значением $Z \geq 1$ указывало на уровень детектирования. Мы повторили несколько циклов компьютерного моделирования, варьируя количество молекул ДНК плазмы (N), использованных для анализа ткани происхождения по профилям метилирования, в диапазоне от 1 миллиона до 300 миллионов. Количество молекул ДНК плазмы в настоящем документе дополнительно корректировали, исходя из вероятности, что ДНК плазмы охватывает гетерозиготный ОНП в соответствии с распределением Пуассона.

На фиг. 31 показан уровень детектирования качественного анализа наследования по материнской линии плода в масштабе всего генома с использованием генетической и эпигенетической информации о молекулах ДНК плазмы по сравнению с анализом относительной дозы гаплотипа (RHDO). Количество молекул, использованных для

анализа, показано по оси X. Уровень детектирования наследования по материнской линии плода в процентах показан по оси Y. Уровень детектирования наследования по материнской линии плода был выше при использовании подхода, основанного на профилях метилирования, по сравнению с RHDO. Например, при использовании 100 миллионов фрагментов уровень детектирования на основе профилей метилирования составил 100%, в то время как уровень детектирования на основе RHDO составил всего 55%. Эти результаты показали, что выведение наследования по материнской линии плода с использованием способа, основанного на профилях метилирования, превосходит метод, основанный на RHDO.

2. Наследование по отцовской линии плода

Возможность получения длинных молекул ДНК плазмы для анализа можно применять для улучшения уровня детектирования специфических для отца вариантов в ДНК плазмы беременной женщины, поскольку использование длинных молекул ДНК увеличит общий геномный охват по сравнению с использованием равного количества коротких молекул ДНК. Далее мы выполнили компьютерное моделирование, основанное на следующих предположениях:

1) Фракция ДНК плода представляла собой f, зависимую от длины ДНК плазмы L. Функция была переписана как f_L, где нижний индекс L указывал на то, что для анализа использовали молекулы ДНК плазмы длиной L п.о.

2) Количество специфических для отца вариантов, которые необходимо было идентифицировать в ДНК материнской плазмы, представляло собой V.

3) Количество молекул ДНК плазмы, использованных для анализа, представляло собой N.

4) Количество молекул ДНК плазмы, происходящих из конкретного геномного локуса или области, подчинялось распределению Пуассона.

В одном примере фракции ДНК плода этих молекул ДНК плазмы размером 150 п.о., 1 тыс. п.о. и 3 тыс. п.о. составляли 10% ($f_{150\ п.о.}$ = 0,1), 2% ($f_{1\ тыс.\ п.о.}$ = 0,02) и 1% ($f_{3\ тыс.\ п.о.}$ = 0,01), соответственно. Количество специфических для отца вариантов составило 250000 (V = 250000) в геноме. Количество молекул ДНК плазмы, использованных для анализа (N), колебалось от 50 миллионов до 500 миллионов.

На фиг. 32 показана взаимосвязь между уровнем детектирования специфических для отца вариантов в масштабе всего генома и количеством секвенированных молекул ДНК плазмы разного размера, использованных для анализа. Количество секвенированных молекул, использованных для анализа, в миллионах показано по оси X. Процентное содержание детектированных специфических для отца вариантов показано по оси Y.

Различные кривые показывают фрагменты ДНК разного размера, использованные для анализа: 3 тыс. п.о. вверху, 1 тыс. п.о. в середине и 150 п.о. внизу. Чем длиннее молекулы ДНК плазмы, используемые для анализа, тем более высокий уровень детектирования специфических для отца вариантов может быть достигнут. Например, при использовании 400 миллионов молекул ДНК плазмы уровень детектирования составил 86%, 93% и 98% при фокусировании на молекулах с размерами 150 п.о., 1 тыс. п.о. и 3 тыс. п.о., соответственно.

Согласно другим вариантам реализации можно использовать другие распределения, включая, но не ограничиваясь перечисленными, распределение Бернулли, бета-нормальное распределение, нормальное распределение, распределение Конвея-Максвелла-Пуассона, геометрическое распределение и т.д. Согласно некоторым вариантам реализации для анализа наследования по материнской и отцовской линии можно использовать выборку Гиббса и теорему Байеса.

3. Анализ наследования ломкой Х-хромосомы

Согласно вариантам реализации определение наследования по материнской линии плода на основе профиля метилирования может облегчить неинвазивное детектирование синдрома ломкой Х-хромосомы с использованием одномолекулярного секвенирования в реальном времени ДНК материнской плазмы. Синдром ломкой Х-хромосомы представляет собой генетическое нарушение, обычно вызываемое распространением тринуклеотидных повторов CGG в гене *FMR1* (сцепленный с ломкой Х-хромосомой синдром умственной отсталости 1) на Х-хромосоме. Синдром ломкой Х-хромосомы и другие нарушения, вызванные распространением повторов, описаны в другом месте этой заявки. Способы детектирования синдрома ломкой Х-хромосомы у плода также можно применять к любому другому распространению повторов, раскрытому в настоящем документе.

Субъект женского пола с премутацией, которая определяется как наличие от 55 до 200 копий повторов CGG в гене *FMR1*, подвержен риску рождения ребенка с синдромом ломкой Х-хромосомы. Вероятность беременности плодом с синдромом ломкой Х-хромосомы зависит от количества повторов CGG, присутствующих в гене *FMR1*. Чем больше количество повторов у матери, тем выше риск распространения от премутации до полной мутации при передаче плоду. Образец материнской плазмы был собран в гестационном возрасте 12 недель у женщины, у которой ранее было подтверждено носительство премутационного аллеля ломкой Х-хромосомы из 115±2 повторов CGG, и которая имела сына с диагностированным синдромом ломкой Х-хромосомы (пробанд). Затем материнскую плазму подвергали одномолекулярному секвенированию в реальном

времени. В одном примере, используя одномолекулярное секвенирование в реальном времени, мы получили 3,3 миллиона кольцевых консенсусных последовательностей (CCS), выравненных с референсным геномом человека, с медианой глубины считывания 75 раз на CCS (межквартильный диапазон: 14–237 раз). Генетическая и эпигенетическая информация для каждой секвенированной ДНК плазмы может быть определена в соответствии с вариантами реализации настоящего изобретения. Чтобы получить два материнских гаплотипа хромосомы X, мы использовали чип на основе гранул Infinium Omni2.5Exome-8 в системе iScan (Illumina), которая представляла собой технологию микрочипов, для генотипирования 2000 ОНП на X-хромосоме для обеих ДНК, экстрагированных из материнской лейкоцитарной пленки и буккального мазка пробанда. Два материнских гаплотипа, а именно Hap I и Hap II, могут быть выведены на основании генотипической информации о геномах матери и пробанда.

На фиг. 33 показан рабочий процесс для неинвазивного детектирования синдрома ломкой X-хромосомы. В гетерозиготных по ОНП сайтах ДНК материнской лейкоцитарной пленки аллели, идентичные генотипам пробанда, использовали для определения гаплотипа, ассоциированного с премутационным аллелем (т.е. Hap I), который был потенциальным предшественником полной мутации в последующих поколениях. С другой стороны, аллели, отличные от генотипов пробанда, использовали для определения гаплотипа, ассоциированного с соответствующим аллелем дикого типа (Hap II). ДНК материнской плазмы от матери пробанда, беременной плодом, подвергали одномолекулярному секвенированию в реальном времени. Риды секвенирования были отнесены к материнским Hap I и Hap II в зависимости от того, была ли полученная генетическая информация идентична аллелям Hap I или Hap II в этих изучаемых геномных локусах. Профили метилирования молекул ДНК плазмы использовали для определения тканей происхождения (т.е. молекулы ДНК, идентифицированные как имеющие плацентарное происхождение на основании анализа профилей метилирования, будут определены как происходящие от плода) тех молекул ДНК плазмы, которые содержат определенное количество сайтов CpG, в соответствии с вариантами реализации настоящего изобретения.

В сценарии A, если молекулы ДНК плода (т.е. плаценты) поддавались детектированию в молекулах ДНК плазмы, отнесенных к материнскому Hap I, но не поддавались детектированию в молекулах ДНК плазмы, отнесенных к материнскому Hap II, тогда Hap I будет определен как переданный нерожденному плоду. Будет определено, что плод подвергается высокому риску поражения синдромом ломкой X-хромосомы. Плацентарное происхождение молекул ДНК плазмы будет основано на статусе

метилирования молекулы, как обсуждается ниже.

В сценарии B, если молекулы ДНК плода поддавались детектированию в молекулах ДНК плазмы, отнесенных к материнскому Hap II, но не поддавались детектированию в молекулах ДНК плазмы, отнесенных к материнскому Hap I, тогда Hap II будет определен как переданный нерожденному плоду. Будет определено, что плод не поражен синдромом ломкой X-хромосомы.

Согласно вариантам реализации определения «детектируемый» и «недетектируемый» для молекул ДНК плода могут зависеть от значений отсечки процентного содержания молекул ДНК плазмы, которые идентифицированы как происходящие от плода (т.е. плацентарные). Значения отсечки для «детектируемого» могут включать, но не ограничиваются перечисленными, более 1%, 2%, 3%, 4%, 5%, 10%, 15%, 20%, 30%, 40%, 50%, и т.д. Значения отсечки для «детектируемого» могут включать, но не ограничиваются перечисленными, менее 1%, 2%, 3%, 4%, 5%, 10%, 15%, 20%, 30%, 40%, 50% и т.д. Согласно некоторым вариантам реализации может потребоваться, чтобы разница в процентном содержании молекул ДНК плазмы, которые, как определено, происходят от плода, между Hap I и Hap II превышала, но не ограничиваясь перечисленными, 1%, 2%, 3%, 4%, 5%, 10%, 15%, 20%, 30%, 40%, 50% и т.д. Согласно некоторым другим вариантам реализации информацию о гаплотипе можно получить с помощью технологий секвенирования с длинными ридами (например, PacBio или нанопоровое секвенирование) (Edge et al. Nat Commun. 2019;10:4660), синтетических длинных ридов (например, с использованием платформы 10X Genomics) (Hui et al. Clin Chem. 2017;63:513-14), фазирования на основе нацеленной амплификации локуса (TLA) (Vermeulen et al. Am J Hum Genet. 2017; 101: 326-39) и статистического фазирования (например, Shape-IT) (Delaneau et al. Nat Method. 2011;9:179-81).

Согласно вариантам реализации можно определить происхождение от матери и происхождение от плода тех молекул ДНК плазмы, которые имеют размер по меньшей мере 200 п.о. и содержат по меньшей мере 5 сайтов CpG (или любых других отсечений для длинных молекул ДНК), в соответствии с подходом сопоставления статуса метилирования, раскрытым в данном описании. Мы идентифицировали одну молекулу ДНК плазмы, расположенную в геномном положении chrX:143,782,245 - 143,782,786 (3,2 млн. п.о. от гена *FMR1*), с аллелем (положение: chrX:143782434; номер доступа ОНП: rs6626483; генотип аллеля: C), идентичным соответствующему аллелю на материнском Hap II, но отличным от такового материнского Hap I.

На фиг. 34 показан профиль метилирования ДНК плазмы по сравнению с профилями метилирования ДНК плаценты и лейкоцитарной пленки. Молекула ДНК

плазмы содержала 5 сайтов CpG. Профиль метилирования был определен как «M-U-U-U-U». Этот профиль метилирования, полученный в результате одномолекулярного секвенирования в реальном времени, сравнивали с референсными профилями метилирования образцов ДНК из плацентарных тканей и лейкоцитарной пленки, полученными в результате бисульфитного секвенирования, в соответствии с подходом сопоставления статуса метилирования, описанным в настоящем изобретении. Оценка для этой молекулы, происходящей из плаценты [т.е. ***S(плаценты)***] была равна 2, что превышало оценку молекулы из лейкоцитарной пленки [т.е. ***S(лейкоцитарной пленки)***] в -3. Таким образом, было определено, что такая молекула ДНК плазмы (chrX:143,782,245-143,782,786) происходит от плода. Однако мы не наблюдали каких-либо молекул ДНК плазмы, несущих аллели из материнского Hap I, происходящих от плода. Таким образом, мы пришли к выводу, что плод унаследовал материнский Hap II и не был поражен синдромом ломкой Х-хромосомы.

Мы предположили, что инактивация Х-хромосомы не может существенно повлиять на эффективность подхода, описанного в настоящем документе, из-за следующих факторов:

1) Х-инактивация у человека является неполной. До 1/3 генов на Х-хромосоме проявляли вариабельное ускользание от Х-инактивации (Cotton et al. Hum Mol Genet. 2015;25:1528-1539). Сайты CpG за пределами CpG-островков (т.е. большинство сайтов CpG) были метилированы в сходной степени у представителей обоих полов, это позволяет предположить, что статус метилирования для большинства сайтов CpG в Х-хромосоме может не нарушаться инактивацией Х-хромосомы (Yasukochi et al. Proc Natl Acad Sci USA. 2010;107:3704-9).

2) Мы использовали профиль метилирования тканей плаценты, сопоставимых по полу для нерожденного плода. Эту стратегию можно применять для детектирования наследования по материнской линии плода с использованием профилей метилирования ДНК плазмы для женщины, беременной плодом мужского пола, поскольку ткани плаценты, включая плод мужского пола, которые, как предполагалось, не затрагиваются Х-инактивацией, будут нести уникальные профили метилирования, отличающиеся от других материнских тканей, которые в большей или меньшей степени вовлечены в Х-инактивацию для определенных областей.

Далее мы секвенировали ДНК, экстрагированную из образца материнской лейкоцитарной пленки, используя одномолекулярное секвенирование в реальном времени. Мы получили 2,3 миллиона CCS с медианной 5-кратной глубиной подридов на CCS. Результаты подтвердили, что материнский Hap I нес премутационный аллель со 124

повторами CGG, а материнский Нар II нес аллель дикого типа с 43 повторами CGG. Кроме того, мы дополнительно секвенировали ДНК, экстрагированную из образцов ворсинок хориона нерожденного плода, с использованием одномолекулярного секвенирования в реальном времени. Мы получили 1,1 миллиона CCS с медианной 4-кратной глубиной подридов на CCS. Результат подтвердил, что нерожденный плод нес аллель дикого типа.

E. Распределение сайтов CpG в геноме человека

Более длинные фрагменты ДНК приводят к большей вероятности того, что фрагмент имеет множество сайтов CpG. Это множество сайтов CpG можно использовать для анализа профиля метилирования или другого анализа.

На фиг. 35 показано распределение сайтов CpG в области из 500 п.о. в геноме человека. В первом столбике показано количество сайтов CpG. Во втором столбике показано количество областей из 500 п.о. с указанным количеством сайтов CpG. В третьем столбике показана доля всех областей, представленных областями, имеющими определенное количество сайтов CpG. Например, 86,14% областей размером 500 п.о. будут нести по меньшей мере 1 сайт CpG. Кроме того, 11,08% областей размером 500 п.о. будут нести по меньшей мере 10 сайтов CpG.

На фиг. 36 показано распределение сайтов CpG в области из 1 тыс. п.о. в геноме человека. В первом столбике показано количество сайтов CpG. Во втором столбике показано количество областей из 1 тыс. п.о. с указанным количеством сайтов CpG. В третьем столбике показана доля всех областей, представленных областями, имеющими определенное количество сайтов CpG. Например, 91,67% областей размером 500 п.о. будут нести по меньшей мере 1 сайт CpG. Кроме того, 32,91% областей размером 500 п.о. будут нести по меньшей мере 10 сайтов CpG.

На фиг. 37 показано распределение сайтов CpG в области из 3 тыс. п.о. в геноме человека. В первом столбике показано количество сайтов CpG. Во втором столбике показано количество областей из 3 тыс. п.о. с указанным количеством сайтов CpG. В третьем столбике показана доля всех областей, представленных областями, имеющими определенное количество сайтов CpG. Например, 92,45% областей размером 3 тыс. п.о. будут нести по меньшей мере 1 сайт CpG. Кроме того, 87,09% областей размером 3 тыс. п.о. будут нести по меньшей мере 10 сайтов CpG.

Согласно некоторым вариантам реализации для максимального повышения чувствительности и специфичности идентификации специфического для плаценты маркера и анализа ткани происхождения можно использовать другие количества сайтов CpG и другие значения отсечки по размеру. В целом сайты CpG появляются чаще, чем

ОНП. Фрагмент ДНК определенного размера, вероятно, будет иметь больше сайтов CpG, чем ОНП. Таблицы, показанные выше, могут показывать более низкие доли для областей, которые имеют то же количество ОНП, что и сайтов CpG, поскольку в области одного и того же размера имеется меньшее количество ОНП, чем сайтов CpG. В результате использование сайтов CpG позволяет использовать больше фрагментов и обеспечивает лучшую статистику, чем использование только ОНП.

F. Примеры анализа ткани происхождения

Согласно вариантам реализации анализ ткани происхождения в материнской плазме можно распространить более чем на два органа/ткани, включая Т-клетки, В-клетки, нейтрофилы, печень и плаценту. Мы секвенировали 9 образцов ДНК матери, используя одномолекулярное секвенирование в реальном времени. Мы вывели вклад плаценты в ДНК материнской плазмы, используя профили метилирования ДНК плазмы в соответствии с подходом сопоставления статуса метилирования, описанным в настоящем изобретении. Для этого анализа путем сопоставления статуса метилирования в одном варианте реализации профиль метилирования каждой из молекул ДНК, которые имели по меньшей мере 500 п.о. в длину и содержали по меньшей мере 5 сайтов CpG, в образце ДНК материнской плазмы сравнивали с профилями метилирования референсной ткани, полученными из бисульфитного секвенирования. В качестве референсных тканей использовали пять тканей, включая нейтрофилы, Т-клетки, В-клетки, печень и плаценту. Молекула ДНК плазмы будет отнесена к ткани, которая соответствовала максимальной оценке сопоставления статуса метилирования для этой молекулы ДНК плазмы. Процентное содержание молекул ДНК плазмы, отнесенных к ткани, по отношению к другим тканям будет считаться долевым вкладом этой ткани в ДНК материнской плазмы этого образца. Согласно вариантам реализации сумма долевого вклада нейтрофилов, Т-клеток и В-клеток в материнской плазме обеспечивает приближенное значение долевого вклада гемопоэтических клеток.

На фиг. 38 показаны долевые вклады молекул ДНК из разных тканей в материнской плазме с использованием анализа путем сопоставления статуса метилирования. В первом столбике показан идентификатор образца. Во втором столбике показан вклад гемопоэтических клеток в процентах. В третьем столбике показан вклад печени в процентах. В четвертом столбике показан вклад плаценты в процентах. На фиг. 38 показано, что основным источником ДНК материнской плазмы были гемопоэтические клетки (медиана: 55,9%), что согласовывалось с предыдущими сообщениями (Sun et al. Proc Natl Acad Sci USA. 2015;112:E5503-12; Zheng et al. Clin Chem. 2012;58:549-58).

На фиг. 39А и 39В показана взаимосвязь между вкладом плаценты и фракцией

ДНК плода, выведенная с помощью подхода ОНП. По оси X показана фракция плода, определенная с помощью подхода ОНП. На оси Y показан определенный вклад плаценты в материнской плазме в процентах, полученный с использованием анализа путем сопоставления статуса метилирования. На фиг. 39A показана хорошая корреляция между вкладом плаценты, определенным с помощью анализа путем сопоставления статуса метилирования, и фракцией ДНК плода, выведенной с помощью ОНП (коэф. Пирсона r = 0,95; значение P<0,0001). Далее мы выполнили деконволюционный анализ ткани для ДНК материнской плазмы путем сравнения плотности метилирования ДНК плазмы, определенной с помощью одномолекулярного секвенирования в реальном времени, с различными профилями метилирования референсной ткани, полученными с помощью бисульфитного секвенирования, в соответствии с квадратичным программированием (Sun et al. Proc Natl Acad Sci USA. 2015;112:E5503-12). На фиг. 39B показано, что при использовании подхода, основанного на плотности метилирования, корреляция между вкладом плаценты (Sun et al. Proc Natl Acad Sci USA. 2015;112:E5503-12) и фракцией ДНК плода была снижена по сравнению с использованием анализа путем сопоставления статуса метилирования (коэф. Пирсона r = 0,65; значение P = 0,059).

Эти данные свидетельствовали о выполнимости выведения долей молекул ДНК, вносимых различными тканями, в образце ДНК материнской плазмы. Согласно другому варианту реализации этот способ также можно применять для измерения молекул ДНК из различных типов клеток или тканей в образце, полученном после инвазивной биопсии солидной ткани или из солидной ткани, полученной после хирургического вмешательства. Согласно некоторым вариантам реализации применение профиля метилирования на уровне отдельной молекулы ДНК для выведения долевых вкладов различных тканей в ДНК материнской плазмы будет превосходить подходы, основанные на совокупных плотностях метилирования от всех секвенированных молекул ДНК плазмы по всему геному.

G. Примерные способы

На фиг. 40 показан способ 4000 анализа биологического образца, полученного от субъекта женского пола, беременного плодом. Биологический образец может включать множество молекул внеклеточной ДНК от плода и субъекта женского пола.

В блоке 4010 могут быть получены риды последовательности, соответствующие множеству молекул внеклеточной ДНК. Согласно некоторым вариантам реализации способ 4000 может включать выполнение секвенирования молекул внеклеточной ДНК.

В блоке 4020 могут быть измерены размеры множества молекул внеклеточной ДНК. Измерение может включать выравнивание ридов последовательности с

референсным геномом. Согласно некоторым вариантам реализации измерение может включать полноразмерное секвенирование и подсчет количества нуклеотидов в полноразмерной последовательности. Согласно некоторым вариантам реализации измерение может включать физическое отделение множества молекул внеклеточной ДНК из биологического образца от других молекул внеклеточной ДНК в биологическом образце, причем другие молекулы внеклеточной ДНК имеют размеры, которые меньше значения отсечки. Физическое отделение может включать любую методику, описанную в настоящем документе, включая использование гранул.

В блоке 4030 может быть идентифицирован набор молекул внеклеточной ДНК из множества молекул внеклеточной ДНК, имеющих размеры, превышающие значение отсечки или равные ему. Значение отсечки может быть больше или равно 200 нт. Значение отсечки может быть по меньшей мере 500 нт., включая 600 нт., 700 нт., 800 нт., 900 нт., 1 тыс. нт., 1,1 тыс. нт., 1,2 тыс. нт., 1,3 тыс. нт., 1,4 тыс. нт., 1,5 тыс. нт., 1,6 тыс. нт., 1,7 тыс. нт., 1,8 тыс. нт., 1,9 тыс. нт. или 2 тыс. нт. Значение отсечки может представлять собой любое значение отсечки, описанное в настоящем документе для длинных молекул внеклеточной ДНК. Размеры могут представлять собой количество сайтов CpG, а не длину молекулы. Например, значение отсечки может составлять 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 или более сайтов CpG.

В блоке 4040 для молекулы внеклеточной ДНК из набора молекул внеклеточной ДНК может быть определен статус метилирования в каждом сайте из множества сайтов. Множество сайтов может включать по меньшей мере 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 или более сайтов CpG. По меньшей мере один из множества сайтов может быть метилирован. Два сайта из множества сайтов могут быть разделены по меньшей мере 160 нт., 170 нт., 180 нт., 190 нт., 200 нт., 250 нт. или 500 нт. Способ может включать секвенирование множества молекул внеклеточной ДНК для получения ридов последовательности и определение статуса метилирования сайта путем измерения характеристики, соответствующей нуклеотиду сайта и нуклеотидам, соседним с сайтом. Например, метилирование можно определить, как в заявке США №16/995607.

В блоке 4050 может быть определен профиль метилирования. Профиль метилирования может свидетельствовать о статусе метилирования в каждом сайте из множества сайтов.

В блоке 4060 профиль метилирования можно сравнить с одним или более референсными профилями. Каждый из одного или более референсных профилей может быть определен для конкретного типа ткани. Согласно некоторым вариантам реализации сравнение может включать определение количества сайтов, которое соответствует

референсному профилю.

Референсный профиль из одного или более референсных профилей может быть определен путем измерения плотности метилирования в каждом референсном сайте из множества референсных сайтов с использованием молекул ДНК из референсной ткани. Плотность метилирования в каждом референсном сайте из множества референсных сайтов можно сравнить с одной или более пороговыми плотностями метилирования. Каждый референсный сайт из множества референсных сайтов может быть идентифицирован как метилированный, неметилированный или неинформативный на основании сравнения плотности метилирования с одной или более пороговыми плотностями метилирования, причем указанное множество сайтов представляет собой множество референсных сайтов, которые идентифицированы как метилированные или неметилированные. Неинформативные сайты могут включать сайты с плотностями метилирования между двумя пороговыми плотностями метилирования. Например, индекс метилирования неинформативных сайтов может быть от 30 до 70 или может представлять собой любой другой диапазон, как описано в настоящем документе.

На этапе 4070 ткань происхождения молекулы внеклеточной ДНК может быть определена с использованием профиля метилирования. Ткань происхождения может представлять собой плаценту. Ткань происхождения может быть фетальной или материнской. Способ может включать определение ткани происхождения как референсной ткани, когда профиль метилирования соответствует референсному профилю, сходно с тем, как описано на фиг. 22. Соответствие может относиться к точному соответствию. Согласно некоторым вариантам реализации определение ткани происхождения как референсной ткани может происходить, когда профиль метилирования соответствует определенному процентному содержанию сайтов референсного профиля. Например, профиль метилирования может соответствовать по меньшей мере 60%, 70%, 80%, 85%, 90%, 95%, 97% или более из сайтов референсного профиля.

Способ может включать определение ткани происхождения путем определения оценки сходства путем сравнения профиля метилирования с первым референсным профилем метилирования из первой референсной ткани из множества референсных тканей. Оценку сходства можно рассчитать с помощью способа сопоставления статуса метилирования или вероятностной модели на основе бета-распределения, описанной в настоящем документе. Оценку сходства можно сравнивать с пороговым значением. Ткань происхождения может быть определена как первая референсная ткань, когда оценка сходства превышает пороговое значение. Оценка сходства может представлять собой

первую оценку сходства. Способ может дополнительно включать вычисление порогового значения путем определения второй оценки сходства путем сравнения профиля метилирования со вторым референсным профилем метилирования из второй референсной ткани из множества референсных тканей. Первая референсная ткань и вторая референсная ткань могут представлять собой разные ткани. Пороговое значение может представлять собой вторую оценку сходства. Первая референсная ткань может иметь самую высокую оценку сходства по сравнению со всеми другими референсными тканями.

Первый референсный профиль метилирования может включать первый поднабор сайтов, имеющий по меньшей мере первую вероятность метилирования для первой референсной ткани. Например, первый поднабор сайтов может представлять собой сайты, которые считают метилированными или обычно метилированными. Первый референсный профиль метилирования может включать второй поднабор сайтов, имеющий не более чем вторую вероятность метилирования для первой референсной ткани. Например, второй поднабор сайтов может представлять собой сайты, которые считают неметилированными или обычно неметилированными. Определение оценки сходства может включать увеличение оценки сходства, когда сайт из множества сайтов метилирован и сайт из множества сайтов находится в первом поднаборе сайтов, и уменьшение оценки сходства, когда сайт из множества сайтов метилирован и сайт из множества сайтов находится во втором поднаборе сайтов. Оценка сходства может быть определена аналогично подходу сопоставления статуса метилирования, описанному в настоящем документе.

Первый референсный профиль метилирования содержит множество сайтов, при этом каждый сайт из множества сайтов характеризуется вероятностью метилирования и вероятностью неметилирования для первой референсной ткани. Оценку сходства можно определить для каждого сайта из множества сайтов, определяя вероятность в референсной ткани, соответствующую статусу метилирования сайта в молекуле внеклеточной ДНК. Оценка сходства может быть определена путем вычисления произведения множества вероятностей. Произведение может представлять собой оценку сходства. Вероятность может быть определена по бета-распределению, аналогично подходу, описанному в настоящем документе.

Способ 4000 может дополнительно включать определение ткани происхождения для каждой молекулы внеклеточной ДНК из набора молекул внеклеточной ДНК. Это определение может включать определение статуса метилирования в каждом сайте из множества соответствующих сайтов, причем указанное множество соответствующих сайтов соответствует молекуле внеклеточной ДНК. Определение ткани происхождения может дополнительно включать определение профиля метилирования. Кроме того,

определение ткани происхождения также может включать сравнение профиля метилирования по меньшей мере с одним референсным профилем из одного или более референсных профилей. Согласно некоторым вариантам реализации сравнение профиля метилирования может быть выполнено аналогично тому, как описано для фиг. 22 и сопроводительному описанию. На фиг. 22 плацента, печень, клетки крови и толстая кишка являются примерами референсных тканей, имеющих проиллюстрированные референсные профили. На фиг. 38 в качестве другого примера референсной ткани показаны гемопоэтические клетки.

Согласно некоторым вариантам реализации может быть определено количество молекул внеклеточной ДНК, соответствующих каждой ткани происхождения. Каждая ткань происхождения может включать каждую референсную ткань из множества референсных тканей. Относительный вклад ткани происхождения может быть определен с использованием количества молекул внеклеточной ДНК, соответствующих каждой ткани происхождения. Например, ткань происхождения может представлять собой плаценту. Другие ткани происхождения могут включать гемопоэтические клетки и печень. Например, относительный вклад плаценты может быть определен на основании количества молекул внеклеточной ДНК, деленного на общее количество молекул внеклеточной ДНК, соответствующих всем тканям происхождения. Согласно некоторым вариантам реализации доля, рассчитанная из количества молекул внеклеточной ДНК, деленного на общее количество молекул внеклеточной ДНК, может быть соотнесена с относительным вкладом через функцию или набор калибровочных точек данных. Как функция, так и набор калибровочных точек данных могут быть определены из множества калибровочных образцов с известными относительными вкладами ткани происхождения. Каждая калибровочная точка данных может указывать относительный вклад, соответствующий калибровочному значению доли. Функция может представлять собой линейную или нелинейную аппроксимацию калибровочных точек данных и может соотносить относительный вклад с долей ткани происхождения или другим параметром, включающим ткань происхождения. Варианты реализации определения относительного вклада могут быть аналогичны тому, что описано для фиг. 39A и 39B.

Для определения ткани происхождения можно использовать модель машинного обучения. Модель может быть обучена путем получения множества обучающих профилей метилирования, при этом каждый обучающий профиль метилирования имеет статус метилирования в одном или более сайтах из множества сайтов, каждый обучающий профиль метилирования определяется по молекуле ДНК из известной ткани. Каждая молекула из известной ткани может представлять собой клеточную ДНК. Обучение может

включать сохранение множества обучающих образцов, при этом каждый обучающий образец включает один из множества обучающих профилей метилирования и метку, указывающую на известную ткань, соответствующую обучающему профилю метилирования. Обучение может включать оптимизацию с использованием множества обучающих образцов параметров модели на основании выходных данных модели, совпадающих или не совпадающих с соответствующими метками, когда в модель вводится множество обучающих профилей метилирования. Параметры могут включать первый параметр, указывающий, имеет ли один сайт из множества сайтов тот же статус метилирования, что и другой сайт из множества сайтов. Например, модель может быть сходна с парным сравнением на фиг. 24. Параметры могут включать второй параметр, указывающий расстояние между сайтами из множества сайтов. Согласно некоторым вариантам реализации модель машинного обучения может не требовать выравнивания сайта метилирования с референсным геномом. Выходные данные модели могут указывать ткань, соответствующую входному профилю метилирования.

Модель машинного обучения может представлять собой сверточные нейронные сети (CNN) или любую модель, описанную в настоящем документе. Модель может включать, но не ограничивается перечисленными, линейную регрессию, логистическую регрессию, глубокую рекуррентную нейронную сеть (например, долгая-краткосрочная память, LSTM), байесовский классификатор, скрытую модель Маркова (HMM), линейный дискриминантный анализ (LDA), кластеризацию k-средних, плотностный алгоритм кластеризации пространственных данных с присутствием шума (DBSCAN), алгоритм случайного леса и метод опорных векторов (SVM).

Отцовство может быть определено способом 4000. Ткань происхождения может быть фетальной. Способ может дополнительно включать выравнивание рида последовательности из ридов последовательности с первой областью референсного генома, причем указанная первая область содержит множество сайтов, соответствующих аллелям, при этом указанное множество сайтов включает пороговое количество сайтов, определение первого гаплотипа с использованием соответствующего аллеля, присутствующего в каждом сайте из множества сайтов, сравнение первого гаплотипа со вторым гаплотипом, соответствующим субъекту мужского пола, и определение с использованием сравнения классификации вероятности того, что субъект мужского пола является отцом плода. Субъект мужского пола может, вероятно, считаться отцом, если гаплотипы совпадают, или, вероятно, не является отцом, если гаплотипы не совпадают. Согласно некоторым вариантам реализации первый гаплотип можно сравнить с обоими гаплотипами субъекта мужского пола.

Согласно вариантам реализации отцовство может быть протестировано, когда ткань происхождения является фетальной, путем выравнивания рида последовательности из ридов последовательности с первой областью референсного генома. Первая область может включать первое множество сайтов, соответствующих аллелям. Множество сайтов может включать пороговое количество сайтов. Пороговое количество сайтов может составлять 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 или более сайтов. Аллель в каждом сайте из множества сайтов можно сравнить с аллелем в соответствующем сайте в геноме субъекта мужского пола. Классификация вероятности того, что субъект мужского пола является отцом плода, может быть определена с использованием сравнения. Субъект мужского пола может, вероятно, считаться отцом, если совпадает определенное количество или процентное содержание аллелей, и, вероятно, не является отцом, если совпадает меньше этого количества или процентного содержания. Процент отсечки может составлять 100%, 90%, 80% или 70%.

Согласно некоторым вариантам реализации может быть определен гаплотип. Способы могут включать выравнивание рида последовательности, соответствующего молекуле внеклеточной ДНК, с референсным геномом для каждой молекулы внеклеточной ДНК из набора молекул внеклеточной ДНК. Рид последовательности может быть идентифицировано как соответствующее гаплотипу, присутствующему у субъекта женского пола. Гаплотип, присутствующий у субъекта женского пола, может быть известен из генотипирования субъекта женского пола. Согласно некоторым вариантам реализации гаплотип субъекта женского пола может быть известен путем анализа концентраций фрагментов ДНК гаплотипа в биологическом образце от субъекта женского пола. Ткань происхождения может быть определена как фетальная с использованием профиля метилирования. Гаплотип может быть определен как наследуемый по материнской линии гаплотип плода.

Наследование гаплотипа может быть определено с использованием метилирования референсных тканей, а не с использованием известных профилей метилирования, таких как профили, связанные с локусами импринтинга. Совпадение или оценка сходства профиля метилирования с референсным профилем может исключать информацию о том, метилирован ли данный аллель или сайт на основании родителя, от которого он был унаследован.

Гаплотип может быть идентифицирован как носитель вызывающей заболевание генетической мутации или вариации. Идентификация гаплотипа как носителя генетической мутации, вызывающей заболевание, может включать идентификацию генетической мутации или вариации в первом риде последовательности. Генетическая

вариация может включать однонуклеотидное различие, делецию или вставку. Может быть измерен первый уровень метилирования во втором риде последовательности, соответствующем первому геномному положению в пределах первого расстояния первого рида последовательности. Также может быть измерен второй уровень метилирования в третьем риде последовательности, соответствующем второму геномному положению в пределах второго расстояния первого рида последовательности. Первое расстояние может представлять собой 100 нт., 200 нт., 300 нт., 400 нт., 500 нт., 600 нт., 700 нт., 800 нт., 900 нт., 1 тыс. нт., 2 тыс. нт., 5 тыс. нт. или 10 тыс. нт. Второй рид последовательности и третий рид последовательности могут находиться на одном и том же плече хромосомы, что и первый рид последовательности. Первый уровень метилирования и второй уровень метилирования могут быть связаны с генетической мутацией или вариацией. Первый уровень метилирования и второй уровень метилирования могут превышать один или два пороговых уровня, связанных с генетической мутацией или вариацией. Пороговые уровни могут быть определены с использованием субъектов, о которых известно, что они имеют генетическую мутацию или вариацию или не имеют ее. Способ может включать классификацию того, что плод, вероятно, имеет заболевание, вызванное генетической мутацией или вариацией.

Могут быть определены специфические для плода профили метилирования. Способ может включать выравнивание рида последовательности, соответствующего молекуле внеклеточной ДНК, с референсным геномом для каждой молекулы внеклеточной ДНК из набора молекул внеклеточной ДНК. Способ может включать идентификацию рида последовательности как соответствующего некоторой области. Область может быть определена путем получения множества ридов последовательности плода, соответствующих множеству молекул ДНК плода из ткани плода. Способ может включать получение множества ридов последовательности матери, соответствующих множеству материнских молекул ДНК. Способ может включать определение статуса метилирования плода в каждом сайте метилирования из множества сайтов метилирования в пределах области для каждого рида последовательности плода из множества ридов последовательности плода. Способ может включать определение статуса метилирования матери в каждом сайте метилирования из множества сайтов метилирования для каждого рида последовательности матери из множества ридов последовательности матери.

Способ определения специфических для плода профилей метилирования может включать определение значения параметра, характеризующего количество сайтов, в которых статус метилирования плода отличается от статуса метилирования матери. Способ может включать сравнение значения параметра с пороговым значением. Параметр

может представлять собой долю сайтов, которые различаются между молекулами ДНК плода и молекулами ДНК матери. Доля может представлять собой оценку несоответствия, описанную в настоящем документе. Пороговое значение может указывать минимальный уровень оценки несоответствия и может составлять 0,3, 0,4, 0,5, 0,6, 0,7, 0,8, 0,9 или более. Согласно некоторым вариантам реализации пороговое значение может представлять собой среднюю оценку несоответствия для молекул ДНК матери или плода. Способ может включать определение того, что значение параметра превышает пороговое значение. Согласно некоторым вариантам реализации может потребоваться, чтобы определенное процентное содержание молекул ДНК матери или плода имело значение параметра, превышающее пороговое значение. Например, процентное содержание может составлять 50%, 60%, 70%, 80%, 90% или более. Согласно некоторым вариантам реализации может потребоваться, чтобы определенное процентное содержание молекул ДНК плода, соответствующих области, имело специфический для плода профиль метилирования. Например, процентное содержание может составлять 40%, 50%, 60%, 70%, 80% или более. Этот способ может быть сходен со способами, описанными на фиг. 25.

Способ может включать обогащение биологического образца молекулами внеклеточной ДНК из ткани происхождения. Обогащение биологического образца может включать отбор и амплификацию набора молекул внеклеточной ДНК. Обогащение может включать отбор на основе размера, как описано в настоящем документе. Согласно некоторым вариантам реализации обогащение может включать отбор на основе профиля метилирования. Например, можно использовать захват и секвенирование на основе метил-CpG-связывающего домена (MBD). Внеклеточную ДНК можно инкубировать с мечеными MBD-белками, которые могут связывать метилированные цитозины. Затем комплекс белок-ДНК можно осадить с использованием магнитных гранул, конъюгированных с антителами. Молекулы ДНК с большим количеством метилированных сайтов CpG могут быть преимущественно обогащены для последующего анализа.

III. Изменение длинных фрагментов внеклеточной ДНК в зависимости от гестационного возраста

Количество длинных фрагментов внеклеточной ДНК может изменяться в зависимости от гестационного возраста. Длинные фрагменты внеклеточной ДНК можно применять для определения гестационного возраста. Кроме того, длинные фрагменты внеклеточной ДНК могут быть более распространены в определенных концевых мотивах по сравнению с более короткими фрагментами внеклеточной ДНК, и относительное количество определенных концевых мотивов может изменяться в зависимости от

гестационного возраста. Количество концевых мотивов также можно применять для определения гестационного возраста. Отклонение гестационного возраста, определенного с использованием длинных фрагментов внеклеточной ДНК, и гестационного возраста, определенного с помощью других клинических методик, может указывать на нарушение, связанное с беременностью. Согласно некоторым вариантам реализации длинные фрагменты внеклеточной ДНК можно применять для определения вероятности нарушения, ассоциированного с беременностью, без обязательного определения гестационного возраста.

A. Анализ размера ДНК плода и матери

ДНК плазмы двух беременных женщин в первом триместре (гестационный возраст: 13 недель), двух во втором триместре (гестационный возраст: 21-22 недели) и пяти в третьем триместре (гестационный возраст: 38 недель) секвенировали с использованием одномолекулярного секвенирования в реальном времени (SMRT) (PacBio). Для каждого случая была получена медиана 176 миллионов (диапазон: 49-685 миллионов) подридов, из которых 128 миллионов (диапазон: 35-507 миллионов) подридов можно было выровнять с референсным геномом человека (hg19). Каждую молекулу в лунке SMRT секвенировали в среднем 107 раз. Медиану 965308 (диапазон: 251686-2871525) высококачественных ридов секвенирования кольцевых консенсусных последовательностей (CCS), что определяли как риды CCS по меньшей мере с 3 подридами, можно было использовать для последующего анализа.

Все секвенированные молекулы из образцов, полученных в каждом триместре беременности, объединяли для анализов размера. В общей сложности имелось 1,94 миллиона, 5,09 миллиона и 4,45 миллиона молекул внеклеточной ДНК для образцов материнской плазмы первого, второго и третьего триместров, соответственно.

На фиг. 41A и 41B показано распределение размеров молекул внеклеточной ДНК из образцов материнской плазмы первого, второго и третьего триместров с диапазоном размеров от 0 до 5 тыс. п.о. По оси X показан размер. По оси Y показана частота. Распределение размеров представлено на графике в диапазоне от 0 до 5 тыс. п.о. по линейной шкале по оси Y для фиг. 41A и от 0 до 5 тыс. п.о. по логарифмической шкале по оси Y для фиг. 41B. ДНК плазмы со всех трех триместров беременности продемонстрировала ожидаемый основной пик при 166 п.о., как показано на фиг. 41A, и ряд основных пиков, возникающих периодическим образом, которые распространялись на молекулы в пределах диапазона от 1 тыс. п.о. до 2 тыс. п.о., как показано на фиг. 41B.

На фиг. 42 приведена таблица, показывающая долю длинных молекул ДНК плазмы в разные триместры беременности. В первом столбике показан гестационный возраст,

связанный с образцом плазмы. Во втором столбике показана доля молекул ДНК длиной более 500 п.о. В третьем столбике показана доля молекул ДНК длиной более 1 тыс. п.о. По сравнению с первым и вторым триместром в третьем триместре наблюдали увеличение частоты молекул ДНК плазмы размером 500 п.о. и выше. Доли длинных молекул ДНК плазмы более 500 п.о. составляли 15,8%, 16,1% и 32,3% для первого, второго и третьего триместров, соответственно. Доли длинных молекул ДНК плазмы более 1 тыс. п.о. составляли 11,3%, 10,6% и 21,4% для первого, второго и третьего триместров, соответственно. В то время как материнская плазма первого и второго триместров показала сходную долю длинных молекул внеклеточной ДНК, материнская плазма третьего триместра содержала приблизительно в два раза больше длинных молекул ДНК.

Для всех образцов ДНК материнской плазмы, проанализированных для настоящего изобретения, ДНК, экстрагированная из их парных образцов материнской лейкоцитарной пленки и образцов плода, была генотипирована с использованием чипа на основе гранул Infinium Omni2.5Exome-8 на системе iScan (Illumina), которая представляет собой метод генотипирования, основанный на чиповой гибридизации. Образцы плода были получены путем отбора образцов ворсин хориона, амниоцентеза или отбора образцов плаценты, в зависимости от того, относился ли случай к первому, второму или третьему триместру, соответственно. Для каждого случая была установлена медиана 203647 информативных однонуклеотидных полиморфизмов (ОНП), по которым мать была гомозиготной, а плод был гетерозиготным. Мы идентифицировали в общей сложности 1362, 2984 и 6082 молекулы ДНК, охватывающие специфические для плода аллели, для первого, второго и третьего триместров, соответственно, когда секвенированные молекулы ДНК для всех случаев из каждого триместра были объединены. С другой стороны, для каждого случая была установлена медиана 210820 информативных ОНП, по которым мать была гетерозиготной, а плод был гомозиготным. Мы идентифицировали в общей сложности 30574, 65258 и 78346 молекул ДНК, охватывающих специфические для матери аллели, для первого, второго и третьего триместров, соответственно. Медиана фракции ДНК плода, определенная по данным секвенирования молекул ДНК ≤600 п.о., составила 15,6% (диапазон 7,6-26,7%) среди всех образцов материнской плазмы.

На фиг. 43A и 43B показаны распределения размеров молекул ДНК, охватывающих специфические для плода аллели, из материнской плазмы первого, второго и третьего триместров. По оси X показан размер. По оси Y показана частота. Распределение размеров представлено на графике в диапазоне от 0 до 3 тыс. п.о. по линейной шкале по оси Y для фиг. 43A и от 0 до 3 тыс. п.о. по логарифмической шкале по оси Y для фиг. 43B.

На фиг. 44A и 44B показаны распределения размеров молекул ДНК, охватывающих специфические для матери аллели, из материнской плазмы первого, второго и третьего триместров. По оси X показан размер. По оси Y показана частота. Распределение размеров представлено на графике в диапазоне от 0 до 3 тыс. п.о. по линейной шкале по оси Y для фиг. 44A и от 0 до 3 тыс. п.о. по логарифмической шкале по оси Y для фиг. 44B.

Как показано на фиг. 43A-44B, молекулы ДНК плазмы, охватывающие специфические для плода и специфические для матери аллели, всех трех триместров беременности проявляли распределения с длинными хвостами, это позволяет предположить наличие длинных молекул ДНК, происходящих как от плода, так и от матери, во всех трех триместрах.

На фиг. 45 приведена таблица доли длинных молекул ДНК плода и матери в плазме в разные триместры беременности. В первом столбике показан гестационный возраст, связанный с образцом плазмы. Во втором столбике показана доля молекул ДНК плода длиной более 500 п.о. В третьем столбике показана доля молекул ДНК матери длиной более 500 п.о. В четвертом столбике показана доля молекул ДНК плода длиной более 1 тыс. п.о. В пятом столбике показана доля молекул ДНК матери длиной более 1 тыс. п.о. В пуле молекул ДНК в материнской плазме молекулы, которые охватывают специфический для плода аллель (плацентарного происхождения), имели меньшую долю длинных молекул ДНК по сравнению с молекулами, охватывающими специфический для матери аллель. Доли длинных молекул ДНК плазмы, охватывающих специфический для плода аллель, размером более 500 п.о. были 19,8%, 23,2% и 31,7% для первого, второго и третьего триместров, соответственно. Доли длинных молекул ДНК плазмы, охватывающих специфический для плода аллель, размером более 1 тыс. п.о. были 15,2%, 16,5% и 19,9% для первого, второго и третьего триместров, соответственно.

Несмотря на то, что меньшая доля длинных молекул ДНК плазмы присутствовала в материнской плазме первого и второго триместра по сравнению с третьим триместром, и молекулы ДНК плода содержали меньше длинных молекул ДНК во всех трех триместрах, способ, описанный в нашем предыдущем изобретении и в настоящем изобретении, позволил нам проанализировать значительную долю длинных молекул ДНК плазмы, что было невозможно ранее при использовании технологий секвенирования с короткими ридами. Кроме того, можно использовать различные стратегии отбора по размеру, включая, но не ограничиваясь перечисленными, электрофоретические, хроматографические и основанные на гранулах методы обогащения длинных фрагментов ДНК в образцах плазмы.

На фиг. 46A, 46B и 46C показаны графики долей специфических для плода фрагментов ДНК плазмы, имеющих определенный диапазон размеров, в разных триместрах. Гестационные возрасты в оцениваемых случаях беременности были подтверждены с помощью ультразвуковой датировки. На фиг. 46A показаны результаты для фрагментов ДНК, которые меньше или равны 150 п.о. На фиг. 46B показаны результаты для фрагментов ДНК от 150 до 600 п.о. На фиг. 46C показаны результаты для фрагментов ДНК, которые больше или равны 600 п.о. Графики показывают долю специфических для плода фрагментов по оси Y и гестационный возраст по оси X. Как показано на графиках, доли специфических для плода фрагментов короче 150 п.о. (фиг. 46A) и длиннее 600 п.о. (фиг. 46C) обеспечат определенную дискриминационную мощность установления различия между образцами третьего триместра и образцами первого триместра и второго триместра по сравнению с долей специфических для плода фрагментов, размер которых колеблется от 150 до 600 п.о. (фиг. 46B). Доли специфических для плода фрагментов длиной более 600 п.о. могут обеспечить наилучшую дискриминационную мощность. Этот вывод подтверждается тем фактом, что абсолютное наименьшее расстояние между группой третьего триместра и объединенной группой первого и второго триместров составило 0,38 при использовании долей специфических для плода фрагментов короче 150 п.о., в то время как аналогичный показатель составил 3,76 при использовании долей специфических для плода фрагментов более 600 п.о. Эти результаты свидетельствовали о том, что применение длинных молекул ДНК для отражения патофизиологического статуса превосходит применение коротких молекул ДНК.

B. Анализ концов ДНК плазмы

В дополнение к размеру мы определили первый нуклеотид на 5′-конце обеих цепей по Уотсону и Крику отдельно для каждой секвенированной молекулы ДНК. Этот анализ состоял из 4 типов концов, а именно A-конца, C-конца, G-конца и T-конца. Рассчитывали процентное содержание молекул ДНК плазмы с конкретным концом из образцов материнской плазмы, полученных в каждом триместре. Далее анализировали процентные содержания A-конца, C-конца, G-конца и T-конца для каждого размера фрагмента.

На фиг. 47A, 47B и 47C показаны графики долевого состава оснований на 5′-конце молекул внеклеточной ДНК из материнской плазмы первого, второго и третьего триместров в диапазоне размеров фрагментов от 0 до 3 тыс. п.о. На фиг. 47A показана материнская плазма первого триместра. На фиг. 47B показана материнская плазма второго триместра. На фиг. 47C показана материнская плазма третьего триместра. Состав оснований в виде процентного содержания показан по оси Y. Размер фрагмента в парах

оснований показан по оси X. Как видно на графиках, C-конец был сверхпредставлен во многих диапазонах размера (в основном менее 1 тыс. п.о.) и варьировался в зависимости от разных диапазонов размера для образцов первого, второго и третьего триместров. Профили концов ДНК плазмы образцов третьего триместра, по-видимому, отличались от профилей образцов первого и второго триместров. Например, кривые T-конца и G-конца были смешаны при размерах, колеблющихся от 105 до 172 п.о., при этом они расходились в образцах первого и второго триместров. Для более длинных фрагментов (например, более примерно 1 тыс. п.о.) C-концевые фрагменты не являются наиболее распространенным фрагментом. G-концевые фрагменты опережают C-концевые фрагменты при примерно 1 тыс. п.о., а затем A-концевые фрагменты становятся более распространенными, чем G-концевые фрагменты, при примерно 2 тыс. п.о.

На фиг. 48 приведена таблица долей концевых нуклеотидных оснований среди коротких и длинных молекул внеклеточной ДНК из материнской плазмы первого, второго и третьего триместров. В первом столбике показано основание на конце молекулы. Во втором столбике показана ожидаемая доля и молекула. В третьем столбике показана доля концевых молекул среди фрагментов, которые меньше или равны 500 п.о., для материнской плазмы первого триместра. В четвертом столбике показана доля концевых молекул среди фрагментов более 500 п.о. для материнской плазмы первого триместра. Пятый и шестой столбик сходны с третьим столбиком и четвертым столбиком, соответственно, за исключением материнской плазмы второго триместра, вместо материнской плазмы первого триместра. Седьмой столбик и восьмой столбик сходны с третьим столбиком и четвертым столбиком, соответственно, за исключением материнской плазмы третьего триместра, вместо материнской плазмы первого триместра.

Если фрагментация внеклеточной ДНК была полностью случайной, то доли концевых нуклеотидных оснований должны отражать состав генома человека, который представляет собой 29,5% A, 29,5% T, 20,5% C и 20,5% G, как показано во втором столбике на фиг. 48. В отличие от случайной фрагментации 5′-конец коротких молекул внеклеточной ДНК размером ≤500 п.о. показал существенную сверхпредставленность C-конца (30,4%, 30,4% и 31,3% для материнской плазмы первого, второго и третьего триместров, соответственно), незначительную сверхпредставленность G-конца (27,4%, 26,9% и 25,3% для первого, второго и третьего триместров, соответственно) и недопредставленность A-конца (19,8%, 19,4% и 19,3% для первого, второго и третьего триместров, соответственно) и T-конца (22,4%, 23,3% и 24,1% для первого, второго и третьего триместров, соответственно).

Однако, по сравнению с короткими молекулами внеклеточной ДНК, длинные

молекулы внеклеточной ДНК размером >500 п.о. показали существенное увеличение доли A-концов (29,6%, 26,0% и 26,7% для материнской плазмы первого, второго и третьего триместров, соответственно), незначительное увеличение доли G-концов (31,0%, 29,5% и 29,9% для первого, второго и третьего триместров, соответственно), существенное снижение доли T-концов (13,9%, 16,9% и 16,4% для первого, второго и третьего триместров, соответственно) и незначительное снижение доли C-концов (25,5%, 27,5% и 27,1% для первого, второго и третьего триместров, соответственно).

На фиг. 49 приведена таблица долей концевых нуклеотидных оснований среди коротких и длинных молекул внеклеточной ДНК, охватывающих специфический для плода аллель, из материнской плазмы первого, второго и третьего триместров. На фиг. 50 приведена таблица долей концевых нуклеотидных оснований среди коротких и длинных молекул внеклеточной ДНК, охватывающих специфический для матери аллель, из материнской плазмы первого, второго и третьего триместров. В первом столбике показано основание на конце молекулы. Во втором столбике показана ожидаемая доля и молекула. В третьем столбике показана доля концевых молекул среди фрагментов, которые меньше или равны 500 п.о., для материнской плазмы первого триместра. В четвертом столбике показана доля концевых молекул среди фрагментов более 500 п.о. для материнской плазмы первого триместра. Пятый и шестой столбик сходны с третьим столбиком и четвертым столбиком, соответственно, за исключением материнской плазмы второго триместра, вместо материнской плазмы первого триместра. Седьмой столбик и восьмой столбик сходны с третьим столбиком и четвертым столбиком, соответственно, за исключением материнской плазмы третьего триместра, вместо материнской плазмы первого триместра. На фиг. 49 и 50 показано, что такая разница в долях концевых нуклеотидных оснований между короткими и длинными молекулами внеклеточной ДНК оставалась неизменной, даже когда мы отдельно исследовали молекулы ДНК, охватывающие специфические для плода и специфические для матери аллели.

На фиг. 51 проиллюстрирован иерархический кластерный анализ коротких и длинных молекул внеклеточной ДНК с использованием 256 4-членных концевых мотивов. В каждом столбике указан образец, использованный для анализа частоты концевого мотива на основе коротких (обозначенных голубым цветом в первом ряду) и длинных фрагментов (обозначенных желтым цветом в первом ряду), соответственно. Начиная со второго ряда, каждый ряд указывает тип концевого мотива. Частоты концевых мотивов были представлены серией цветовых градиентов в соответствии с частотами, нормированными по ряду (z-оценка) (т.е. число стандартных отклонений ниже или выше средней частоты по образцам). Более красный цвет указывает на более высокую частоту

концевого мотива, а более синий цвет указывает на меньшую частоту концевого мотива.

На фиг. 51 мы охарактеризовали короткие и длинные молекулы внеклеточной ДНК, анализируя их профили 4-членных концевых мотивов. Мы определили последовательность первых 4 нуклеотидов (4-членный мотив) на 5′-конце обеих цепей по Уотсону и Крику отдельно для каждой секвенированной молекулы ДНК. Для каждого образца материнской плазмы частоту каждого концевого мотива ДНК плазмы рассчитывали отдельно для коротких (≤500 п.о.) и длинных (>500 п.о.) молекул ДНК плазмы. Иерархический кластерный анализ, основанный на частотах 256 4-членных концевых мотивов, показал, что профили концевых мотивов длинных молекул ДНК в разных образцах материнской плазмы формировали кластер, который отличался от кластера коротких молекул ДНК. Эти результаты позволяют предположить, что длинная и короткая ДНК обладали разными свойствами фрагментации. Согласно вариантам реализации можно применять относительное отклонение этих концевых мотивов между длинными и короткими молекулами ДНК, чтобы показать вклады внеклеточной ДНК, происходящей из путей гибели клеток, таких как, но не ограничиваясь ими, апоптоз и некроз. Повышенная активность этих путей гибели клеток может быть связана с нарушениями, связанными с беременностью, и другими нарушениями.

На фиг. 52A и 52B показан анализ основных компонентов (PCA) с использованием профилей 4-членных концевых мотивов для классификационного анализа. На фиг. 52A показаны короткие молекулы внеклеточной ДНК (≤500 п.о.) из разных триместров. На фиг. 52B показаны длинные молекулы внеклеточной ДНК (>500 п.о.) образцов материнской плазмы из разных триместров. Процентные содержания в скобках по осям X и Y представляют величину изменчивости, объясняемую соответствующим компонентом. Каждая синяя точка представляет собой образец материнской плазмы первого триместра. Каждая желтая точка представляет собой образец материнской плазмы второго триместра. Каждая красная точка представляет собой образец материнской плазмы третьего триместра. Эллипс представляет собой 95% уровень достоверности для группы точек данных за конкретный триместр. По сравнению с короткими молекулами внеклеточной ДНК (фиг. 52A) (также описанными в заявке США № 15/787050) профили 4-членных концевых мотивов длинных молекул внеклеточной ДНК (фиг. 52B) давали более четкое разделение между образцами материнской плазмы первого, второго и третьего триместров. Согласно вариантам реализации можно применять профили концевых мотивов длинных молекул ДНК плазмы по отдельности или в комбинации с другими характеристиками ДНК материнской плазмы, включая, но не ограничиваясь этим, уровень метилирования и размер, для молекулярной оценки гестационного возраста.

Например, мы использовали нейронные сети для обучения модели предсказывать гестационный возраст на основе 256 концевых мотивов, общего уровня метилирования и доли фрагментов размером ≥600 п.о. Выходными переменными были 1, 2 и 3, представляющие 1й, 2й и 3й триместр. Входные переменные включали 256 концевых мотивов, общий уровень метилирования и долю фрагментов размером ≥600 п.о. Мы использовали подход «исключения по одному» для оценки эффективности предсказания гестационного возраста. Для набора данных, содержащего 9 образцов, подход «исключения по одному» выполняли таким образом, что один образец был выбран в качестве тестового образца, а оставшиеся 8 образцов использовали для обучения модели на основе нейронных сетей. Такой тестируемый образец был определен как 1, 2 или 3 на основе установленной модели. Затем мы повторили этот процесс для других образцов, которые еще не были протестированы. Всего мы повторили 9 раз такой процесс обучения-тестирования. При сравнении этих результатов тестирования с клинической информацией о гестационном возрасте 8 из 9 образцов (89%) были правильно предсказаны в отношении гестационного возраста. Согласно другому варианту реализации такой анализ может быть выполнен, например, но не ограничиваясь этим, с использованием теоремы Байеса, логистической регрессии, множественной регрессии и метода опорных векторов, анализа случайного леса, анализа по алгоритму построения бинарного дерева решений (CART), алгоритма K-ближайших соседей.

Затем все секвенированные молекулы из образцов, полученных в каждом триместре беременности, объединяли для последующего анализа концевых мотивов. 256 концевых мотивов ранжировали по их частоте среди коротких и длинных молекул ДНК плазмы.

На фиг. 53-58 приведены таблицы 25 концевых мотивов с самыми высокими частотами для определенных длин фрагментов ДНК (короче или длиннее 500 п.о.) и для разных триместров. На фиг. 53, 54 и 55 приведены таблицы с концевыми мотивами, отсортированными по их рангу в коротких фрагментах (<500 п.о.). На фиг. 53-55 в первом столбике показан концевой мотив. Во втором столбике показан частотный ранг мотива в коротких фрагментах. В третьем столбике показан частотный ранг мотива в длинных фрагментах. В четвертом столбике показана частота мотива в коротких фрагментах. В пятом столбике показана частота мотива в длинных фрагментах. В шестом столбике показано кратное изменение (частота мотива в коротких фрагментах, деленная на частоту мотива в длинных фрагментах).

На фиг. 56, 57 и 58 приведены таблицы с концевыми мотивами, отсортированными по их рангу в длинных фрагментах (>500 п.о.). На фиг. 56-58 в первом столбике показан

концевой мотив. Во втором столбике показан частотный ранг мотива в длинных фрагментах. В третьем столбике показан частотный ранг мотива в коротких фрагментах. В четвертом столбике показана частота мотива в длинных фрагментах. В пятом столбике показана частота мотива в коротких фрагментах. В шестом столбике показано кратное изменение (частота мотива в длинных фрагментах, деленная на частоту мотива в коротких фрагментах).

Фиг. 53 и 56 взяты из образцов первого триместра. Фиг. 54 и 57 взяты из образцов второго триместра. Фиг. 55 и 58 взяты из образцов третьего триместра.

Из первых 25 концевых мотивов с самыми высокими частотами среди коротких молекул ДНК плазмы 11 из них начинались с динуклеотидов CC. Концевые мотивы, начинающиеся с CC, вместе составляли 14,66%, 14,66% и 15,13% концевых мотивов коротких ДНК плазмы в материнской плазме первого, второго и третьего триместров, соответственно. Из первых 25 концевых мотивов с самыми высокими частотами среди длинных молекул ДНК плазмы 4-членные мотивы, оканчивающиеся динуклеотидами TT, составляли 9 из них в материнской плазме второго и третьего триместров и 10 из них в материнской плазме первого триместра.

Мы определили динуклеотидную последовательность третьего (X) и четвертого нуклеотидов (Y) с 5′-конца обеих цепей по Уотсону и Крику отдельно для каждой секвенированной молекулы ДНК. X и Y могут представлять собой один из четырех нуклеотидных оснований в ДНК. Имелось 16 возможных мотивов NNXY, а именно NNAA, NNAT, NNAG, NNAC, NNTA, NNTT, NNTG, NNTC, NNGA, NNGT, NNGG, NNGC, NNCA, NNCT, NNCG и NNCC.

На фиг. 59A, 59B и 59C показаны диаграммы рассеяния частот мотивов для 16 мотивов NNXY среди коротких и длинных молекул ДНК плазмы. На фиг. 59A показаны результаты для первого триместра. На фиг. 59B показаны результаты для второго триместра. На фиг. 59C показаны результаты для третьего триместра. Частота мотивов длинных фрагментов показана по оси Y. Частота мотивов коротких фрагментов показана по оси X. Каждый круг представляет собой один из 16 мотивов NNXY. Пара пунктирных линий на каждой диаграмме рассеяния обозначает 1,5-кратное увеличение (верхняя линия) и уменьшение (нижняя линия) частот мотивов в длинных молекулах ДНК плазмы (>500 п.о.) по сравнению с короткими молекулами ДНК плазмы (≤500 п.о.). Круги, расположенные за пределами заштрихованной области, представляют мотивы с кратным изменением >1,5.

В то время как концы коротких молекул ДНК плазмы показали высокие частоты 4-членных мотивов, начинающихся с динуклеотидов CC (CCNN) (Jiang et al. Cancer Discov

2020;10(5):664-673; Chan et al. Am J Hum Genet 2020;107(5):882-894), концы длинных молекул ДНК плазмы показали >1,5-кратное увеличение частот 4-членного мотива, оканчивающегося на TT (NNTT), во всех трех триместрах (фиг. 11). На мотив NNTT приходилось 18,94%, 15,22% и 15,30% концевых мотивов длинных ДНК плазмы в материнской плазме первого, второго и третьего триместров, соответственно. Напротив, на мотив NNTT приходилось только 9,53%, 9,29% и 8,91% концевых мотивов коротких ДНК плазмы в материнской плазме первого, второго и третьего триместров, соответственно.

Как сообщили ранее Han et al., новая внеклеточная ДНК, высвобождаемая из умирающих клеток в плазму, была обогащена А-концевыми фрагментами >150 п.о. Было обнаружено, что фактор фрагментации ДНК бета (DFFB), который является основной внутриклеточной нуклеазой, участвующей во фрагментации ДНК во время апоптоза, отвечает за образование таких фрагментов (Han et al. Am J Hum Genet 2020;106:202-214). В настоящем изобретении мы показали, что длинные молекулы внеклеточной ДНК размером >500 п.о. также были обогащены А-концевыми фрагментами, это указывает на то, что DFFB также может быть ответственным за образование этих фрагментов. При нормальной беременности апоптоз трофобласта увеличивается с увеличением гестационного возраста (Sharp et al. Am J Reprod Immuno 2010;64(3):159-69). Действительно, наши результаты о возрастающих долях длинных молекул ДНК, охватывающих специфический для плода аллель, на более поздних триместрах могут отражать увеличение апоптоза трофобласта на более поздних триместрах.

Согласно вариантам реализации способы, описанные в настоящем документе, можно применять для анализа длинных молекул внеклеточной ДНК в материнской плазме для прогнозирования, скрининга и мониторинга прогрессирования осложнений беременности, связанных с плацентой, включая, но не ограничиваясь перечисленными, преэклампсию, задержку внутриутробного развития (ЗВУР), преждевременные роды и гестационную трофобластическую болезнь. Сообщалось о повышенном уровне апоптоза трофобласта при осложнениях беременности, связанных с плацентой, таких как преэклампсия (Leung et al. Am J Obstet Gynecol 2001;184:1249-1250), ЗВУР (Smith et al. Am J Obstet Gynecol 1997;177:1395-1401; Levy et al. Am J Obstet Gynecol 2002;186:1056-1061) и гестационная трофобластическая болезнь. Кроме того, сообщалось о повышенном уровне ДНК плода в материнской плазме при преэклампсии (Lo et al. Clin Chem 1999;45(2):184-8; Smid et al. Ann N Y Acad Sci 2001;945:132-7), ЗВУР (Sekizawa et al. Am J Obstet Gynecol 2003;188:480-4) и преждевременных родах (Leung et al. Lancet 1998;352(9144):1904-5). Мы предположили, что при осложнениях беременности,

связанных с плацентой, будет увеличена доля длинных молекул внеклеточной ДНК плацентарного происхождения в образцах материнской плазмы из-за повышенного плацентарного апоптоза. Следовательно, длинные молекулы внеклеточной ДНК плацентарного происхождения как таковые, а также сигнатуры длинных ДНК, включая, но не ограничиваясь перечисленными, А-концевые фрагменты и мотивы NNTT, могут служить биомаркерами плацентарного апоптоза.

Хотя в приведенном выше анализе используются однонуклеотидные и 4-нуклеотидные мотивы, в других вариантах реализации можно использовать мотивы, имеющие другие длины, например, 2, 3, 5, 6, 7, 8, 9, 10 или более.

C. Примерные способы

Длинные фрагменты внеклеточной ДНК можно применять для определения гестационного возраста у субъекта женского пола, беременного плодом. Количество длинных фрагментов внеклеточной ДНК изменяется в зависимости от гестационного возраста и может быть использовано для определения гестационного возраста. Концевой мотив фрагментов внеклеточной ДНК также изменяется в зависимости от гестационного возраста и может быть использован для определения гестационного возраста. Когда гестационный возраст, определенный с использованием длинных фрагментов внеклеточной ДНК, значительно отклоняется от гестационного возраста, определенного с помощью других клинических методик, то можно считать, что беременный субъект женского пола и/или плод имеет заболевание, связанное с беременностью. Согласно некоторым вариантам реализации определение гестационного возраста может не требоваться для определения вероятности нарушения, ассоциированного с беременностью.

1. Гестационный возраст

На фиг. 60 показан способ 6000 анализа биологического образца, полученного от субъекта женского пола, беременного плодом. Гестационный возраст может быть определен и может быть использован для классификации вероятности нарушения, ассоциированного с беременностью. Биологический образец может включать множество молекул внеклеточной ДНК от плода и субъекта женского пола.

Могут быть получены риды последовательности, соответствующие множеству молекул внеклеточной ДНК. Согласно некоторым вариантам реализации для получения ридов последовательности может быть выполнено секвенирование.

В блоке 6020 могут быть измерены размеры множества молекул внеклеточной ДНК. Размеры могут быть измерены аналогично тому, как описано для фиг. 21. Размеры могут быть измерены с использованием ридов последовательности.

В блоке 6030 может быть измерено первое количество молекул внеклеточной ДНК, размеры которых превышают значение отсечки. Величина может представлять собой число, общую длину или массу молекул внеклеточной ДНК.

В блоке 6040 может быть сгенерировано значение нормированного параметра с использованием первого количества. Значение нормированного параметра может представлять собой первое количество, нормированное к общему количеству молекул внеклеточной ДНК, к количеству молекул внеклеточной ДНК от плода или от матери или к количеству молекул ДНК из определенной области. Например, нормированный параметр может представлять собой долю специфических для плода фрагментов, как описано на фиг. 46A-C.

В блоке 6050 значение нормированного параметра может сравниваться с одной или более калибровочными точками данных. Каждая калибровочная точка данных может указывать гестационный возраст, соответствующий калибровочному значению нормированного параметра. Например, гестационный возраст определенного триместра или определенного количества недель может соответствовать калибровочному значению нормированного параметра. Одна или более калибровочных точек данных могут быть определены из множества калибровочных образцов с известными гестационными возрастами и включающих молекулы внеклеточной ДНК, размеры которых превышают значение отсечки. Согласно некоторым вариантам реализации калибровочные точки данных определяются по функции, соотносящей гестационный возраст со значениями нормированного параметра.

В блоке 6060 гестационный возраст может быть определен с использованием сравнения. Гестационный возраст может считаться возрастом, соответствующим калибровочному значению, наиболее близкому к значению нормированного параметра. Согласно некоторым вариантам реализации гестационный возраст может считаться наиболее поздним возрастом, соответствующим калибровочному значению, которое превышено значением нормированного параметра.

Способ может дополнительно включать определение референсного гестационного возраста плода с использованием ультразвука или даты последней менструации субъекта женского пола. Способ также может включать сравнение гестационного возраста с референсным гестационным возрастом. Способ может дополнительно включать определение классификации вероятности нарушения, ассоциированного с беременностью, с использованием сравнения гестационного возраста с референсным гестационным возрастом. Например, расхождение между гестационным возрастом и референсным гестационным возрастом может указывать на нарушение, связанное с беременностью.

Расхождение может заключаться в другом триместре или разнице в гестационном возрасте на минимальное количество недель (например, 1, 2, 3, 4, 5, 6, 7 или более недель).

Способ может дополнительно включать использование концевых мотивов. Например, способ может включать определение первой подпоследовательности, соответствующей по меньшей мере одному концу молекул внеклеточной ДНК, имеющих размеры, превышающие значение отсечки. Первое количество может представлять собой молекулы внеклеточной ДНК, имеющие размер, превышающий значение отсечки, и имеющие первую подпоследовательность на одном или более концах соответствующей молекулы внеклеточной ДНК. Первая подпоследовательность может представлять собой или может включать 1, 2, 3, 4, 5 или 6 нуклеотидов. Концевые мотивы можно использовать для определения гестационного возраста с помощью анализа PCA, как описано на фиг. 52A и 52B. Калибровочные образцы могут быть использованы с различными концевыми мотивами и известными гестационными возрастами и подвергнуты анализу PCA. Для концевых мотивов могут использоваться другие алгоритмы классификации и регрессии, такие как линейный дискриминантный анализ, логистическая регрессия, метод опорных векторов, линейная регрессия, нелинейная регрессия и т.д. Алгоритмы классификации и регрессии могут соотносить гестационный возраст с определенными концевыми мотивами и/или фрагментами определенного размера.

Концевые мотивы могут представлять собой любой мотив, обсуждаемый на фиг. 47-59 или 94. Ранг или частоту концевого мотива можно сравнить с рангами или частотами концевого мотива в калибровочных образцах от субъектов с известными гестационными возрастами. Затем ранг или частоту концевого мотива можно использовать для определения гестационного возраста. Концевой мотив, присутствующий с рангом или частотой, отклоняющейся от ранга или частоты, определенной в референсных образцах того же гестационного возраста, может указывать на нарушение, связанное с беременностью.

Генерирование значения нормированного параметра может включать (a) нормирование первого количества к общему количеству молекул внеклеточной ДНК, имеющих размер, превышающий значение отсечки; (b) нормирование первого количества ко второму количеству молекул внеклеточной ДНК, имеющих размер больше значения отсечки и заканчивающихся второй подпоследовательностью, причем указанная вторая подпоследовательность отличается от первой подпоследовательности, или (c) нормирование первого количества к третьему количеству молекул внеклеточной ДНК, имеющих размер, который меньше значения отсечки.

2. Нарушение, связанное с беременностью

На фиг. 61 показан способ 6100 анализа биологического образца, полученного от субъекта женского пола, беременного плодом. Варианты реализации могут включать классификацию вероятности нарушения, ассоциированного с беременностью, без обязательного определения гестационного возраста. Биологический образец может включать множество молекул внеклеточной ДНК от плода и субъекта женского пола.

Могут быть получены риды последовательности, соответствующие множеству молекул внеклеточной ДНК. Согласно некоторым вариантам реализации для получения ридов последовательности может быть выполнено секвенирование.

В блоке 6120 могут быть измерены размеры множества молекул внеклеточной ДНК. Размеры могут быть получены аналогично тому, как это описано для фиг. 21. При измерении размеров можно использовать полученные риды последовательности.

В блоке 6130 может быть измерено первое количество молекул внеклеточной ДНК, размеры которых превышают значение отсечки. Значение отсечки может быть больше или равно 200 нт. Значение отсечки может представлять собой по меньшей мере 500 нт., включая 600 нт., 700 нт., 800 нт., 900 нт., 1 тыс. нт., 1,1 тыс. нт., 1,2 тыс. нт., 1,3 тыс. нт., 1,4 тыс. нт., 1,5 тыс. нт., 1,6 тыс. нт., 1,7 тыс. нт., 1,8 тыс. нт., 1,9 тыс. нт. или 2 тыс. нт. Значение отсечки может представлять собой любое значение отсечки, описанное в настоящем документе для длинных молекул внеклеточной ДНК. Первая величина может представлять собой количество или частоту.

В блоке 6140 может быть сгенерировано первое значение нормированного параметра с использованием первого количества. Генерирование значения нормированного параметра может включать измерение второго количества молекул внеклеточной ДНК, включая размеры, которые меньше значения отсечки; и вычисление соотношения первого количества и второго количества. Значение отсечки может представлять собой первое значение отсечки. Второе значение отсечки может быть меньше первого значения отсечки. Второе количество может включать молекулы внеклеточной ДНК, размеры которых меньше второго значения отсечки, или второе количество может включать все молекулы внеклеточной ДНК во множестве молекул внеклеточной ДНК. Нормированный параметр может представлять собой показатель частоты длинных молекул внеклеточной ДНК.

В блоке 6150 может быть получено второе значение, соответствующее ожидаемому значению нормированного параметра для здоровой беременности. Второе значение может зависеть от гестационного возраста плода. Второе значение может представлять собой

ожидаемое значение. Согласно некоторым вариантам реализации второе значение может представлять собой значение отсечки, отличающееся от аномального значения.

Получение второго значения может включать получение второго значения из калибровочной таблицы, соотносящей измерения у беременных субъектов женского пола с калибровочными значениями нормированного параметра. Калибровочную таблицу можно сгенерировать путем получения первой таблицы, соотносящей гестационные возрасты с измерениями у беременных субъектов женского пола. Можно получить вторую таблицу, соотносящую гестационные возрасты с калибровочными значениями нормированного параметра. Данные в первой и второй таблицах могут быть от одних и тех же субъектов или разных субъектов. Калибровочная таблица, соотносящая измерения с калибровочными значениями, может быть создана из первой таблицы и второй таблицы. Калибровочная таблица может включать функцию, которая соотносит калибровочные значения с измерениями.

Измерения у беременных субъектов женского пола могут представлять собой время, прошедшее с момента последней менструации, или характеристики изображения беременных субъектов женского пола (например, ультразвук). Измерения у беременных субъектов женского пола могут представлять собой характеристики изображений беременных субъектов женского пола. Например, характеристики изображения могут включать длину, размер, внешний вид или анатомию плода субъекта женского пола. Характеристики могут включать биометрические измерения, например, длину от темени до крестца или длину бедренной кости. Может использоваться внешний вид некоторых органов, включая внешний вид четырехкамерного сердца или позвонков на спинном мозге. Гестационный возраст может быть определен врачом по ультразвуковому изображению (например, Committee on Obstetric Practice et al., “Methods for estimating the due date,” Committee Opinion, No. 700, May 2017).

Согласно некоторым вариантам реализации модель машинного обучения может связывать одну или более калибровочных точек данных с характеристиками изображений. Модель может быть обучена путем получения множества обучающих изображений. Каждое обучающее изображение может быть получено от субъекта женского пола, о котором известно, что у него нет нарушения, ассоциированного с беременностью, или о котором известно, что он не имеет нарушения, ассоциированного с беременностью. Субъекты женского пола могут иметь диапазон гестационных возрастов. Обучение может включать сохранение множества обучающих образцов от субъектов женского пола. Каждый обучающий образец может включать известное значение нормированного параметра, ассоциированного с обучающим изображением. Модель может быть обучена

путем оптимизации с использованием множества обучающих образцов параметров модели на основании выходных данных модели, совпадающих или не совпадающих с изображением с известным значением нормированного параметра. Выходные данные модели могут указывать значение нормированного параметра, соответствующего изображению. Второе значение нормированного параметра может быть сгенерировано путем ввода изображения субъекта женского пола в модель машинного обучения.

В блоке 6160 может быть определено отклонение между первым значением нормированного параметра и вторым значением нормированного параметра. Отклонение может представлять собой степень разделения.

В блоке 6170 классификация вероятности нарушения, ассоциированного с беременностью, может быть определена с использованием отклонения. Нарушение, связанное с беременностью, может быть вероятным, когда отклонение превышает порог. Порог может указывать на статистически значимое различие. Порог может указывать на различие в 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% или 100%.

Нарушение, связанное с беременностью, может включать преэклампсию, задержку внутриутробного развития, инвазивную плацентацию, преждевременные роды, гемолитическую болезнь новорожденных, плацентарную недостаточность, водянку плода, порок развития плода, гемолиз, синдром повышенной активности печеночных ферментов и низкого уровня тромбоцитов (HELLP) или системную красную волчанку.

IV. Анализ размера и концов для нарушений, связанных с беременностью

Анализ размера и/или концов длинных молекул ДНК использовали для определения вероятности преэклампсии. Такие способы также могут быть применены к другим нарушениям, связанным с беременностью. ДНК, экстрагированная из образцов материнской плазмы четырех беременных женщин с диагнозом преэклампсия, была подвергнута одномолекулярному секвенированию в реальном времени (SMRT) (PacBio).

На фиг. 62 приведена таблица, показывающая клиническую информацию о четырех случаях преэклампсии. В первом столбике указан номер истории болезни. Во втором столбике указан гестационный возраст в неделях на момент забора крови. В третьем столбике указан пол плода. В четвертом столбике указана клиническая информация о преэклампсии (PET).

M12804 представлял собой случай преэклампсии (PET) тяжелой степени тяжести и существовавшей ранее IgA-нефропатии. M12873 представлял собой случай хронической гипертензии с наложением PET легкой степени тяжести. M12876 представлял собой случай PET тяжелой степени тяжести с поздним началом. M12903 представлял собой случай PET тяжелой степени тяжести с поздним началом с задержкой внутриутробного

развития (ЗВУР). Пять нормотензивных образцов материнской плазмы третьего триместра использовали в качестве контроля для последующих анализов согласно настоящему изобретению.

Для четырех преэклamптических и пяти нормотензивных образцов ДНК материнской плазмы третьего триместра, проанализированных для настоящего изобретения, ДНК, экстрагированная из их парных образцов материнской лейкоцитарной пленки и плаценты, была генотипирована с использованием чипа на основе гранул Infinium Omni2.5Exome-8 на системе iScan (Illumina).

Концентрацию ДНК плазмы каждого образца количественно определяли с помощью высокочувствительного анализа дцДНК Qubit с использованием флуориметра Qubit (ThermoFisher Scientific). Средние концентрации ДНК плазмы для случаев преэклампсии и третьего триместра составляли 95,4 нг/мл (диапазон 52,1–153,8 нг/мл) плазмы и 10,7 нг/мл (6,4–19,1 нг/мл) плазмы, соответственно. Средняя концентрация ДНК плазмы в случаях преэклампсии была примерно в 9 раз выше, чем в случаях третьего триместра.

Средние фракции ДНК плода, определенные по данным секвенирования молекул ДНК ≤600 п.о., которые охватывали информативные однонуклеотидные полиморфизмы (ОНП), по которым мать была гомозиготной, а плод был гетерозиготным, составили 22,6% (диапазон 16,6–25,7%) и 20,0% (диапазон 15,6–26,7%) для преэклamптических и нормотензивных образцов материнской плазмы третьего триместра, соответственно.

A. Анализ размера

Анализы размера выполняли на преэклamптических и нормотензивных образцах материнской плазмы третьего триместра в соответствии с вариантами реализации настоящего изобретения. На фиг. 63A-63D и фиг. 64A-64D показаны распределения размеров молекул ДНК плазмы в случаях преэклампсии и нормотензии в третьем триместре. По оси X показан размер. По оси Y показана частота. Распределение размеров представлено на графиках в диапазоне от 0 до 1 тыс. п.о. по линейной шкале по оси X для фиг. 63A-63D и от 0 до 5 тыс. п.о. по логарифмической шкале по оси X для фиг. 64A-64D. На фиг. 63A и 64A показан образец M12804. На фиг. 63B и 64B показан образец M12873. На фиг. 63C и 64C показан образец M12876. На фиг. 63D и 64D показан образец M12903.

Синяя линия представляет распределение размеров всех секвенированных молекул ДНК плазмы, объединенных из пяти нормотензивных случаев третьего триместра. Красная линия представляет распределение размеров секвенированных молекул ДНК плазмы в отдельных случаях преэклампсии. На фиг. 63A-63D синяя линия представляет линию более короткого пика ниже 200 п.о. и линию более высокого пика между 300 и 400

п.о. На фиг. 64A-64D синяя линия соответствует линии, которая находится выше при 1 тыс. п.о.

В целом размерные профили ДНК плазмы у пациенток с преэклампсией были короче, чем таковые у беременных женщин с нормотензией в третьем триместре, с увеличенной высотой пика 166 п.о. и увеличенной долей молекул ДНК короче 166 п.о. (фиг. 63A-63D). Эти изменения были более выражены в двух случаях тяжелой преэклампсии M12876 и M12903. Изменения были еще более значительными в случае преэклампсии M12903 с задержкой внутриутробного развития (ЗВУР).

Три из четырех преэклампсических образцов плазмы показали уменьшенные доли длинных молекул ДНК плазмы с размерами 200-5000 п.о. (фиг. 64B-64D). Доли длинных молекул ДНК плазмы >500 п.о. в M12873, M12876 и M12903 составляли 11,7%, 8,9% и 4,5%, соответственно, в то время как доля длинных молекул ДНК плазмы в объединенных данных секвенирования пяти случаев с нормотензией в третьем триместре была 32,3%. Образец плазмы для случая преэклампсии (PET) тяжелой степени тяжести с ранее существовавшей IgA-нефропатией (M12804) показал уменьшенную долю более коротких молекул ДНК менее 2000 п.о., но увеличенную долю более длинных молекул ДНК более 2000 п.о. по сравнению с объединенными данными секвенирования пяти случаев с нормотензией в третьем триместре (фиг. 2A). Доля длинных молекул ДНК плазмы у M12804 составила 34,9%.

На фиг. 65A-65D и фиг. 66A-66D показано распределение размеров молекул ДНК, охватывающих специфические для плода аллели, из преэклампсических и нормотензивных образцов материнской плазмы третьего триместра. На каждой из фигур от A до D показан отдельный преэклампсический образец. По оси X показан размер. По оси Y показана частота на фиг. 65A-65D, и суммарная частота на фиг. 66A-66D. На фиг. 66A-66D размер показан от 0 до 35 тыс. п.о.

Синяя линия на каждом графике представляет распределение размеров всех секвенированных молекул ДНК плазмы, охватывающих специфические для плода аллели, объединенных для пяти случаев с нормотензией в третьем триместре. Красная линия на каждом графике представляет распределение размеров секвенированных молекул ДНК плазмы, охватывающих специфические для плода аллели, в отдельных случаях преэклампсии. На фиг. 65A-65D синяя линия представляет собой линию более короткого пика ниже 200 п.о. и линию более высокого пика между 300 и 400 п.о. На фиг. 66A-66D синяя линия соответствует линии, которая ниже между 100 и 1000 п.о.

На фиг. 67A-67D и фиг. 68A-68D показано распределение размеров молекул ДНК, охватывающих специфические для плода аллели, из преэклампсических и

нормотензивных образцов материнской плазмы третьего триместра. На каждой из фигур от A до D показан отдельный преэклamптический образец. По оси X показан размер. По оси Y показана частота на фиг. 67A-67D и суммарная частота на фиг. 68A-68D. На фиг. 68A-68D размер показан от 0 до 35 тыс. п.о.

Синяя линия на каждом графике представляет распределение размеров всех секвенированных молекул ДНК плазмы, охватывающих специфические для матери аллели, объединенных для пяти случаев с нормотензией в третьем триместре. Красная линия на каждом графике представляет распределение размеров секвенированных молекул ДНК плазмы, охватывающих специфические для матери аллели, в отдельных случаях преэклампсии. На фиг. 67A синяя линия представляет собой линию более высокого пика ниже 200 п.о. и более высокого пика между 300 и 400 п.о. На фиг. 67B-67D синяя линия представляет собой линию более короткого пика ниже 200 п.о. На фиг. 68A синяя линия соответствует линии, которая выше между 1000 и 10000 п.о. На фиг. 68B-68D синяя линия соответствует линии, которая ниже между 100 и 1000 п.о.

Явление укорочения ДНК плазмы наблюдали как в молекулах ДНК, охватывающих специфические для плода аллели (фиг. 65B-65D и фиг. 66B-66D), так и в молекулах, охватывающих специфические для матери аллели (фиг. 67B-67D и фиг. 68B-68D), в трех из четырех преэклamптических образцов плазмы по сравнению с нормотензивными образцами материнской плазмы третьего триместра. Исключением был случай M12804 PET тяжелой степени тяжести с ранее существовавшей IgA-нефропатией, который показал повышенную долю более коротких молекул ДНК менее 1 тыс. п.о. и уменьшенную долю более длинных молекул ДНК более 1 тыс. п.о. среди тех молекул ДНК плазмы, которые охватывают специфические для плода аллели (фиг. 65A и 66A). Действительно, молекулы ДНК плазмы, охватывающие специфические для матери аллели, в случае M12804 показали удлиненный размерный профиль (фиг. 67A и 68A).

На фиг. 69A и 69B показаны графики доли коротких молекул ДНК, охватывающих (A) специфические для плода аллели и (B) специфические для матери аллели, в преэклamптических и нормотензивных образцах материнской плазмы, секвенированных с использованием секвенирования PacBio SMRT. По оси Y показана доля коротких фрагментов ДНК <150 п.о. По оси X показаны нормальные образцы и образцы PET.

Согласно вариантам реализации доля коротких молекул ДНК была определена как процентное содержание молекул ДНК материнской плазмы с размером менее 150 п.о. M12804 был исключен из этого анализа, так как в этом случае уже существовала IgA-нефропатия, а в других образцах ее не было. Группа преэклamптических образцов плазмы показала значительно увеличенные доли коротких молекул ДНК, охватывающих

специфические для плода аллели (P = 0,036, критерий суммы рангов Уилкоксона) и специфические для матери аллели (P = 0,036, критерий суммы рангов Уилкоксона), по сравнению с группой нормотензивных контрольных образцов плазмы.

На фиг. 70A и 70B показаны графики доли коротких молекул ДНК в преэклamптических и нормотензивных образцах материнской плазмы, секвенированных с использованием (A) секвенирования PacBio SMRT и (B) секвенирования Illumina. По оси Y показана доля коротких фрагментов ДНК <150 п.о.

Согласно вариантам реализации доля коротких молекул ДНК была определена как процентное содержание молекул ДНК материнской плазмы с размером менее 150 п.о. M12804 был исключен из этого анализа, так как этот случай показал другой размерный профиль по сравнению с другими случаями преэклампсии в этой когорте, вероятно, из-за ранее существовавшей IgA-нефропатии в этом случае. Группа преэклamптических образцов плазмы показала значительно повышенные доли коротких молекул ДНК (медиана: 28,0%; диапазон: 25,8–35,1%) при сравнении с группой нормотензивных контрольных образцов плазмы (медиана: 12,1%; диапазон: 8,5–15,8%) (P = 0,036, критерий суммы рангов Уилкоксона). Напротив, в предыдущей когорте из четырех преэклamптических и четырех сопоставимых по гестационному возрасту нормотензивных образцов ДНК материнской плазмы, которые были подвергнуты бисульфитному преобразованию и секвенированию Illumina, доли коротких молекул ДНК в преэклamптических образцах плазмы и контрольных образцах плазмы существенно не различались (P = 0,340, критерий суммы рангов Уилкоксона) (фиг. 70B).

Согласно некоторым вариантам реализации можно применять отсечение 20% для доли коротких молекул ДНК в образце материнской плазмы, секвенированном с использованием секвенирования PacBio SMRT, чтобы определить, была ли беременность связана с высоким риском или низким риском развития преэклампсии. Образец материнской плазмы с долей коротких молекул ДНК выше 20% будет определен как имеющий высокий риск развития преэклампсии, в то время как образец материнской плазмы с долей коротких молекул ДНК ниже 20% будет определен как имеющий низкий риск развития преэклампсии. При использовании этого отсечки как чувствительность, так и специфичность составляли 100%. Согласно некоторым другим вариантам реализации отсечение для доли используемых коротких молекул ДНК может включать, но не ограничивается перечисленными, 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60% и т.д. Согласно другому варианту реализации долю коротких молекул ДНК в образце материнской плазмы можно использовать для мониторинга и оценки тяжести преэклампсии во время беременности.

Согласно вариантам реализации соотношение размеров, указывающее относительные доли коротких и длинных молекул ДНК, рассчитывали для каждого образца с использованием следующего уравнения.

$$\text{Соотношение размеров} = \frac{P(50-150)}{P(200-1000)}$$

где P(50 – 150) обозначает долю секвенированных молекул ДНК плазмы с размерами, колеблющимися от 50 п.о. до 150 п.о.; и P(200 – 1000) обозначает долю секвенированных молекул ДНК плазмы с размерами, колеблющимися от 200 п.о. до 1000 п.о.

На фиг. 71 показан график соотношений размеров, который показывает относительные доли коротких и длинных молекул ДНК в преэклamптических и нормотензивных образцах материнской плазмы, секвенированных с использованием секвенирования PacBio SMRT. По оси Y показано соотношение размеров. По оси X показаны нормальные образцы и образцы PET. Группа преэклamптических образцов плазмы показала значительно более высокое соотношение размеров при сравнении с группой нормотензивных контрольных образцов плазмы (P = 0,016, критерий суммы рангов Уилкоксона).

Согласно вариантам реализации для прогнозирования развития и тяжести преэклампсии при беременностях можно применять размерные профили, сгенерированные на платформах секвенирования с длинными ридами, включая, но не ограничиваясь перечисленными, секвенирование PacBio SMRT и секвенирование Oxford Nanopore. Согласно некоторым вариантам реализации прогрессирование преэклампсии и развитие признаков тяжелой преэклампсии, включая, но не ограничиваясь этим, нарушения функции печени и почек, можно отслеживать путем анализа размерных профилей молекул ДНК плазмы. Согласно некоторым вариантам реализации параметры размера, используемые в анализе, могут включать, но не ограничиваются этим, долю коротких или длинных молекул ДНК и соотношение размеров, которое указывает на относительные доли коротких и длинных молекул ДНК. Отсечение, используемое для определения категорий короткой и длинной ДНК, может включать, но не ограничивается перечисленными, 150 п.о., 180 п.о., 200 п.о., 250 п.о., 300 п.о., 350 п.о., 400 п.о., 450 п.о., 500 п.о., 550 п.о., 600 п.о., 650 п.о., 700 п.о., 750 п.о., 800 п.о., 850 п.о., 900 п.о., 950 п.о., 1 тыс. п.о. и т.д. Диапазоны размеров, используемые при определении соотношения размеров коротких и длинных молекул, могут включать, но не ограничиваются перечисленными, 50 – 150 п.о., 50 – 166 п.о., 50 – 200 п.о., 200 – 400 п.о., 200 – 1000 п.о., 200 – 5000 п.о. или другие комбинации.

Анализ размера и концов может включать использование способа, описанного в способе 6100 на фиг. 61.

B. Анализ конца фрагмента

Анализы концов фрагментов выполняли на преэклamптических и нормотензивных образцах материнской плазмы третьего триместра в соответствии с вариантами реализации настоящего изобретения. Первый нуклеотид на 5′-конце обеих цепей по Уотсону и Крику определяли для каждой секвенированной молекулы ДНК плазмы. Доли Т-конца, С-конца, А-конца и G-конца фрагментов определяли для каждого образца ДНК плазмы.

На фиг. 72A-72D показана доля различных концов молекул ДНК плазмы в преэклamптических и нормотензивных образцах материнской плазмы, секвенированных с использованием секвенирования PacBio SMRT. По оси X показаны нормальные образцы третьего триместра и образцы PET. По оси Y показана доля определенного конца. На фиг. 72A показана доля Т-конца. На фиг. 72B показана доля С-конца. На фиг. 72C показана доля А-конца. На фиг. 72D показана доля G-конца. Группа преэклamптических образцов плазмы показала значительно повышенные доли молекул ДНК плазмы с Т-концом (P = 0,016, критерий суммы рангов Уилкоксона) и значительно сниженные доли молекул ДНК плазмы с G-концом (P = 0,016, критерий суммы рангов Уилкоксона) при сравнении с группой нормотензивных контрольных образцов плазмы.

На фиг. 73 показан иерархический кластерный анализ ДНК преэклamптических и нормотензивных образцов материнской плазмы третьего триместра с использованием четырех типов концов фрагментов (первый нуклеотид на 5′-конце каждой цепи), а именно С-конца, G-конца, Т-конца и А-конца. В каждом столбике указан образец ДНК плазмы. Первая строка указывает, к какой группе принадлежал каждый образец, при этом голубой цвет указывает на ДНК нормотензивного образца материнской плазмы третьего триместра, а оранжевый цвет указывает на ДНК преэклamптического образца плазмы. Голубой охватывает первые пять столбиков. Оранжевый охватывает последние четыре столбика.

Начиная со второго ряда, каждый ряд указывает тип конца фрагмента. Частоты концевых мотивов были представлены серией цветовых градиентов в соответствии с частотами, нормированными по ряду (z-оценка) (т.е. число стандартных отклонений ниже или выше средней частоты по образцам). Более красный цвет указывает на более высокую частоту концевого мотива, а более синий цвет указывает на меньшую частоту концевого мотива. Иерархический кластерный анализ, основанный на частотах 4 типов концов фрагментов, показал, что профили концов фрагментов ДНК плазмы преэклamптических

образцов образовали кластер, который отличался от кластера ДНК плазмы нормотензивных образцов третьего триместра.

Согласно вариантам реализации можно определить динуклеотидную последовательность первого (X) и второго нуклеотидов (Y) с 5′-конца обеих цепей по Уотсону и Крику отдельно для каждой секвенированной молекулы ДНК. X и Y могут представлять собой один из четырех нуклеотидных оснований в ДНК. Существуют 16 возможных динуклеотидных концевых мотивов XYNN, а именно AANN, ATNN, AGNN, ACNN, TANN, TTNN, TGNN, TCNN, GANN, GTNN, GGNN, GCNN, CANN, CTNN, CGNN и CCNN. Можно определить динуклеотидную последовательность третьего (X) и четвертого нуклеотидов (Y) с 5′-конца обеих цепей по Уотсону и Крику отдельно для каждой секвенированной молекулы ДНК в соответствии с вариантом реализации настоящего изобретения. Существуют 16 возможных динуклеотидных мотивов NNXY. Также можно определить последовательность первых четырех нуклеотидов (4-членный мотив) на 5′-конце обеих цепей по Уотсону и Крику отдельно для каждой секвенированной молекулы ДНК.

На фиг. 74 показан иерархический кластерный анализ ДНК преэкламптических и нормотензивных образцов материнской плазмы третьего триместра с использованием 16 динуклеотидных мотивов XYNN (динуклеотидная последовательность первого и второго нуклеотидов с 5′-конца). На фиг. 75 показан иерархический кластерный анализ ДНК преэкламптических и нормотензивных образцов материнской плазмы третьего триместра с использованием 16 динуклеотидных мотивов NNXY (динуклеотидная последовательность третьего и четвертого нуклеотидов с 5′-конца). На фиг. 76 показан иерархический кластерный анализ ДНК преэкламптических и нормотензивных образцов материнской плазмы третьего триместра с использованием 256 четырехнуклеотидных мотивов (динуклеотидная последовательность нуклеотидов с первого по четвертый нуклеотиды с 5′-конца).

На фиг. 74-76 первый ряд указывает на то, к какой группе принадлежал каждый образец, при этом голубой цвет указывает на ДНК нормотензивного образца материнской плазмы третьего триместра, а оранжевый цвет указывает на ДНК преэкламптического образца плазмы. Голубой охватывает первые пять столбиков. Оранжевый охватывает последние четыре столбика. Начиная со второго ряда, каждый ряд указывает тип конца фрагмента. Частоты концевых мотивов были представлены серией цветовых градиентов в соответствии с частотами, нормированными по ряду (z-оценка) (т.е. число стандартных отклонений ниже или выше средней частоты по образцам). Более красный цвет указывает на более высокую частоту концевого мотива, а более синий цвет указывает на меньшую

частоту концевого мотива.

Эти результаты свидетельствовали о том, что ДНК плазмы в преэклямптических и непреэклямптических образцах обладала различными свойствами фрагментации. Согласно одному варианту реализации для прогнозирования развития преэклампсии при беременностях можно применять профили концевых мотивов, сгенерированные на платформах секвенирования с длинными ридами, включая, но не ограничиваясь перечисленными, секвенирование PacBio SMRT и секвенирование Oxford Nanopore. Хотя в приведенном выше анализе использовали однонуклеотидные, двухнуклеотидные и четырехнуклеотидные мотивы, в других вариантах реализации можно использовать мотивы, имеющие другие длины, например, 3, 5, 6, 7, 8, 9, 10 или более.

Согласно некоторым вариантам реализации можно комбинировать анализ конца фрагмента и анализ ткани происхождения для улучшения эффективности прогнозирования, детектирования и мониторинга состояний, связанных с беременностью, включая, но не ограничиваясь этим, преэклампсию. Во-первых, можно выполнить анализ концов фрагментов для каждого образца материнской плазмы, чтобы разделить молекулы ДНК плазмы на четыре категории концов фрагментов, а именно фрагменты с Т-концом, С-концом, А-концом и G-концом. Затем отдельно можно выполнить анализ ткани происхождения, используя молекулы ДНК плазмы из каждой из категорий концов фрагментов для каждого образца ДНК материнской плазмы, используя анализ путем сопоставления статуса метилирования в соответствии с вариантами реализации настоящего изобретения. Долевой вклад различных тканей в одной из категорий концов фрагментов определяли как процентное содержание молекул ДНК плазмы в соответствующей категории концов фрагментов, которое было отнесено к соответствующей ткани, по отношению к другим тканям.

Мы проанализировали три и пять образцов ДНК плазмы беременных женщин с преэклампсией и без нее, используя одномолекулярное секвенирование в реальном времени. Мы получили медиану 658722, 889900, 851501 и 607554 фрагментов плазмы с А-концом, С-концом, G-концом и Т-концом. Для фрагментов с А-концом мы сравнили профили метилирования любого фрагмента по меньшей мере с 10 сайтами CpG с референсными профилями метилирования нейтрофилов, Т-клеток, В-клеток, печени и плаценты в соответствии с подходом сопоставления статуса метилирования, описанным в настоящем изобретении. Фрагмент ДНК плазмы может быть отнесен к ткани, которая соответствовала максимальным оценкам сопоставления статуса метилирования среди этих тканей. При использовании этого способа медиана 2,43% (диапазон: 0,73–5,50%) А-концевых фрагментов была присвоена Т-клеткам (т.е. вклад Т-клеток) среди всех

анализируемых образцов. Далее мы аналогичным образом проанализировали фрагменты с C-концом, G-концом и T-концом, соответственно. Для фрагментов с C-концом, G-концом и T-концом наблюдали медиану вклада T-клеток 3,20% (диапазон: 1,55–5,19%), 3,52% (диапазон: 1,53–6,27%) и 2,22% (0–7,79%), соответственно.

На фиг. 77A-77D показан вклад T-клеток среди молекул ДНК, принадлежащих к разным категориям концов фрагментов, а именно (A) T-концу, (B) C-концу, (C) A-концу и (D) G-концу в ДНК преэклamптических и нормотензивных образцов материнской плазмы. По оси X показаны образцы нормального третьего триместра и образцы PET. По оси Y показан вклад T-клеток в процентах. Результаты показали, что среди G-концевых фрагментов вклад T-клеток был значительно снижен в преэклamптических образцах плазмы по сравнению с нормотензивными образцами плазмы третьего триместра ($P = 0,036$, критерий суммы рангов Уилкоксона). Согласно вариантам реализации можно использовать отсечение 3% для вклада T-клеток среди всех G-концевых фрагментов в образце ДНК материнской плазмы, чтобы определить, была ли беременность связана с высоким риском или низким риском развития преэклampсии.

C. Примерные способы

На фиг. 78 показан способ 7800 анализа биологического образца, полученного от субъекта женского пола, беременного плодом. Биологический образец может включать множество молекул внеклеточной ДНК плода и субъекта женского пола. С помощью способа можно сгенерировать классификацию вероятности нарушения, ассоциированного с беременностью. Нарушение, связанное с беременностью, может представлять собой преэклampсию или любое нарушение, связанное с беременностью, описанное в настоящем документе.

Могут быть получены риды последовательности, соответствующие множеству молекул внеклеточной ДНК.

В блоке 7810 могут быть измерены размеры множества молекул внеклеточной ДНК. Размеры могут быть измерены посредством выравнивания или подсчета количества нуклеотидов или любой методики, описанной в настоящем документе, включая фиг. 21.

В блоке 7820 может быть идентифицирован набор молекул внеклеточной ДНК, размеры которых превышают значение отсечки. Значение отсечки может представлять собой любое значение отсечки для длинных фрагментов внеклеточной ДНК, включая 500 нт., 600 нт., 700 нт., 800 нт., 900 нт., 1 тыс. нт., 1,1 тыс. нт., 1,2 тыс. нт., 1,3 тыс. нт., 1,4 тыс. нт., 1,5 тыс. нт., 1,6 тыс. нт., 1,7 тыс. нт., 1,8 тыс. нт., 1,9 тыс. нт. или 2 тыс. нт. Значение отсечки может представлять собой любое значение отсечки, описанное в настоящем документе для длинных молекул внеклеточной ДНК.

В блоке 7830 может быть сгенерировано значение параметра концевого мотива с использованием первого количества. Может быть измерено первое количество молекул внеклеточной ДНК в наборе, имеющих первую подпоследовательность на одном или более концах молекул внеклеточной ДНК в наборе. Согласно некоторым вариантам реализации параметр концевого мотива может представлять собой первое количество, нормированное к общему количеству всех подпоследовательностей на конце. Согласно некоторым вариантам реализации конец может представлять собой 3′-конец. Согласно некоторым вариантам реализации конец может представлять собой 5′-конец.

Первая подпоследовательность может представлять собой 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 или более нуклеотидов в длину. Первая подпоследовательность может включать последний нуклеотид на конце соответствующей молекулы внеклеточной ДНК. Например, первая подпоследовательность может представлять собой профиль XYNN, показанный на фиг. 74. Согласно некоторым вариантам реализации первая подпоследовательность может не включать последний нуклеотид или нуклеотиды на конце соответствующей молекулы внеклеточной ДНК. Например, первая подпоследовательность может включать профиль NNXY на фиг. 75.

Может быть измерено второе количество молекул внеклеточной ДНК, имеющих подпоследовательность, отличную от первой подпоследовательности, на одном или более концах молекул внеклеточной ДНК. Значение параметра концевого мотива может быть сгенерировано с использованием соотношения второго количества и третьего количества. Например, второе количество может быть разделено на третье количество или третье количество может быть разделено на второе количество.

В блоке 7840 значение параметра концевого мотива можно сравнить с пороговым значением. Пороговое значение может представлять собой значение, которое представляет собой статистически значимое отличие от значения связанного параметра для субъекта без нарушения, ассоциированного с беременностью. Пороговое значение может быть определено у одного или более референсных субъектов с нормальными беременностями или одного или более референсных субъектов с нарушениями, связанными с беременностью.

Согласно некоторым вариантам реализации значение параметра концевого мотива можно сравнить с пороговым значением, а значение второго параметра концевого мотива можно сравнить со вторым пороговым значением. Можно измерить второе количество молекул внеклеточной ДНК, имеющих вторую подпоследовательность, отличную от первой подпоследовательности, на одном или более концах молекул внеклеточной ДНК. Таким образом, можно определить количества различных концевых мотивов. Может быть

сгенерировано значение второго параметра концевого мотива с использованием второго количества. Значение второго параметра концевого мотива можно сравнить со вторым пороговым значением. Второе пороговое значение может быть таким же, как и первое пороговое значение или может отличаться от него. Дополнительные подпоследовательности могут использоваться таким же образом, как и первая и вторая подпоследовательности. Согласно некоторым вариантам реализации все возможные подпоследовательности могут использоваться для сравнений с пороговыми значениями.

В блоке 7850 классификация вероятности нарушения, ассоциированного с беременностью, может быть определена с использованием сравнения. Нарушение, связанное с беременностью, может быть вероятным, когда значение размерного параметра или значение параметра концевого мотива превышает пороговое значение.

Согласно некоторым вариантам реализации для определения классификации вероятности нарушения, ассоциированного с беременностью, можно применять сравнение значения второго параметра концевого мотива со вторым значением отсечки. Нарушение, связанное с беременностью, может быть вероятным, когда значение первого параметра концевого мотива превышает первое пороговое значение, а значение второго параметра концевого мотива превышает второе пороговое значение.

Способ может включать использование размерного параметра в дополнение к параметру концевого мотива. Может быть идентифицирован второй набор молекул внеклеточной ДНК, имеющих размеры в пределах первого диапазона размеров. Первый диапазон размеров может включать размеры, превышающие значение отсечки. Первый диапазон размеров включает размеры, которые могут превышать значение отсечки. Первый диапазон размеров может быть менее 550 нт., 600 нт., 650 нт., 700 нт., 750 нт., 800 нт., 850 нт., 900 нт., 950 нт., 1 нт., 1,5 тыс. нт., 2 тыс. нт., 3 тыс. нт., 5 тыс. нт. или более. Значение размерного параметра может быть сгенерировано с использованием второго количества молекул внеклеточной ДНК во втором наборе. Значение размерного параметра можно сравнить со вторым пороговым значением. Для определения классификации вероятности нарушения, ассоциированного с беременностью, можно использовать сравнение значения размерного параметра со вторым пороговым значением. Классификация может указывать на вероятное наличие нарушения, ассоциированного с беременностью, когда превышено одно или оба из первого и второго пороговых значений.

Размерный параметр может представлять собой нормированный параметр. Например, может быть измерено третье количество молекул внеклеточной ДНК во втором диапазоне размеров. Второй диапазон размеров может включать размеры, которые меньше первого значения отсечки. Второй диапазон размеров может включать все

размеры. Второй диапазон размеров может включать 50 – 150 нт., 50 – 166 нт., 50 – 200 нт., 200 – 400 нт. Второй диапазон размеров может включать любые размеры для коротких фрагментов внеклеточной ДНК, описанных в настоящем документе. Второй диапазон размеров может исключать размеры в первом диапазоне размеров. Значение размерного параметра может быть сгенерировано путем определения соотношения второго количества и третьего количества. Например, второе количество может быть разделено на третье количество или третье количество может быть разделено на второе количество.

Любое из количеств молекул внеклеточной ДНК может обозначать молекулы внеклеточной ДНК из конкретной ткани происхождения. Например, ткань происхождения может представлять собой Т-клетки или другую ткань происхождения, описанную в настоящем документе. Второе количество может быть сходно с вкладом Т-клеток, описанным на фиг. 77A-77D. Вклад ткани происхождения можно определить, используя статус или профиль метилирования, как описано в настоящем изобретении.

V. Заболевания, связанные с распространением повторов

Длинные фрагменты внеклеточной ДНК, полученные от беременных женщин, можно применять для выявления распространения повторов в генах. Распространение повторов в генах может привести к нервно-мышечным заболеваниям. Распространения тандемных повторов связаны с заболеваниями человека, включая, но не ограничиваясь перечисленными, нейродегенеративные нарушения, такие как синдром ломкой Х-хромосомы, болезнь Хантингтона и спиноцеребеллярная атаксия. Распространения тандемных повторов могут возникать в областях генов, кодирующих белок (болезнь Мачадо-Джозефа, синдром Хау-Ривер, болезнь Хантингтона), или в некодирующих областях (атаксия Фридриха, миотоническая дистрофия, некоторые формы синдрома ломкой Х-хромосомы). Распространения с вовлечением минисателлитных, пентануклеотидных, тетрануклеотидных и многочисленных тринуклеотидных повторов связаны с ломкими сайтами. Распространения, связанные с этими заболеваниями, могут быть вызваны проскальзыванием при репликации или асимметричной рекомбинацией, или эпигенетическими аберрациями. Количество повторов в последовательности относится к общему количеству раз, когда появляется подпоследовательность. Например, «CAGCAG» включает два повтора. Поскольку повторы включают по меньшей мере две копии подпоследовательности, количество повторов не может быть равно 1. Подпоследовательность можно понимать как звено повтора.

Согласно вариантам реализации анализ длинной внеклеточной ДНК у беременных женщин может облегчить детектирование заболеваний, связанных с повторами.

Например, тринуклеотидный повтор представляет собой повторяющийся участок мотивов из 3 п.о. в последовательностях ДНК. Одним из примеров является то, что последовательность «CAGCAGCAG» содержит три мотива «CAG» из 3 п.о. Сообщалось, что распространение микросателлитов, обычно распространение тринуклеотидных повторов, играет решающую роль при неврологических нарушениях (Kovtun et al. Cell Res. 2008;18:198-213; McMurray et al. Nat Rev Genet. 2010;11:786-99). Одним из примеров является то, что более 55 повторов CAG (всего 165 п.о.) в гене *ATXN3* являются патогенными, что приводит к спиноцеребеллярной атаксии 3 типа (SCA3), характеризующейся прогрессирующими проблемами с двигательной активностью. Это состояние наследуется по аутосомно-доминантному типу. Таким образом, одной копии измененного гена достаточно, чтобы вызвать нарушение. Для определения количества повторов микросателлитов обычно используют полимеразную цепную реакцию (ПЦР) для амплификации представляющей интерес геномной области, а затем продукт ПЦР подвергают ряду различных методик, таких как капиллярный электрофорез (Lyon et al. J Mol Diagn. 2010;12:505-11), анализ методом Саузерн-блоттинга (Hsiao et al. J Clin Lab Anal. 1999;13:188-93), анализ кривой плавления (Lim et al. J Mol Diagn. 2014;17:302-14) и масс-спектрометрия (Zhang et al. Anal Methods. 2016;8:5039-44). Однако эти методы были трудоемкими и требовали много времени, а также их было трудно применять для высокопроизводительного скрининга в реальной клинической практике, такой как пренатальное тестирование. Секвенирование по Сэнгеру сопряжено со значительными трудностями при выявлении длинных повторов из следов сложных последовательностей при ручном исследовании. Хорошо известно, что технологии секвенирования Illumina и Ion Torrent имеют существенные трудности при секвенировании GC-богатых (или GC-бедных) областей, несущих эти повторы (Ashely et al. 2016;17:507-22), и длина ДНК, содержащей распространившиеся повторы, легко превышала длину ридов последовательности (Loomis et al. Genome Res. 2013;23:121-8).

Другим примером является миотоническая дистрофия, которая вызывается распространением повторов CTG, колеблющемся от 50 до 4000 повторов CTG, рядом с геном *DMPK*, которая также представляет собой аутосомно-доминантное нарушение. Молекулярная диагностика МД обычно выполняется при пренатальной диагностике путем инвазивного анализа числа CTG на геномной ДНК плода.

В отличие от секвенирования с короткими ридами (сотни оснований) способы, описанные в настоящем изобретении, позволяют получать длинные молекулы ДНК из ДНК материнской плазмы (несколько тысяч оснований). Используя способы, описанные в настоящем изобретении, неинвазивным путем можно определить, наследует ли

нерожденный плод это заболевание от пораженной матери.

На фиг. 79 показана иллюстрация выведения наследования по материнской линии плода для заболеваний, связанных с повторами. На этапе 7905 внеклеточную ДНК при беременности подвергали одномолекулярному секвенированию в реальном времени (например, PacBio SMRT). На этапе 7910 результаты секвенирования были разделены на категории длинной и короткой ДНК в соответствии с настоящим изобретением. На этапе 7915 информация об аллелях, присутствующих в длинных молекулах ДНК, может быть использована для конструирования материнских гаплотипов, а именно Hap I и Hap II. Каждый из Hap I и Hap II может включать распространившиеся повторы тринуклеотидной подпоследовательности (например, CTG). На этапе 7920 может быть проанализирован дисбаланс гаплотипов, аналогично тому, как это описано на фиг. 16. На этапе 7925 можно сделать вывод о наследовании по материнской линии плода. Способы, описанные в настоящем документе, позволяют не только определить гаплотипы (например, Hap I и Hap II), но также определить, какой гаплотип несет распространившиеся повторы (например, пораженный Hap I), которые вызывают нарушение, используя информацию о последовательности длинных молекул ДНК в соответствии с настоящим изобретением. Используя количества, размеры или состояния метилирования коротких молекул ДНК, распределенных по материнским Hap I и Hap II, в соответствии со способом, описанным в настоящем документе, можно определить, наследует ли плод материнский Hap I (пораженный) или Hap II (непораженный) в этом примере.

На фиг. 80 показана иллюстрация выведения наследования по отцовской линии плода для заболеваний, связанных с повторами. Используя внеклеточную ДНК при беременности, можно определить, наследует ли плод пораженный отцовский гаплотип. Как показано на фиг. 80, внеклеточная ДНК во время беременности непораженной женщины (например, 5 повторов CTG для Hap I и 6 повторов CTG для Hap II), муж которой был поражен заболеванием, связанным с распространением повторов (например, 70 повторов CTG), была подвергнута секвенированию PacBio SMRT, секвенированные длинные молекулы ДНК были идентифицированы и использованы для определения гаплотипа и количества повторов. Если гаплотип A, несущий длинный участок повтора CTG (например, 70 повторов CTG в этом примере), присутствует в материнской плазме непораженной беременной женщины, это указывает на то, что плод унаследовал пораженный отцовский гаплотип. Согласно некоторым вариантам реализации ДНК, содержащая распространившиеся повторы, также несет один или более других специфических для отца аллелей, которые отсутствуют в материнском геноме. Эту ситуацию можно применять для подтверждения наследования по отцовской линии.

Согласно другому варианту реализации можно определить, наследует ли плод пораженный отцовский гаплотип, используя внеклеточную ДНК при беременности. Как показано на фиг. 80, внеклеточная ДНК во время беременности непораженной женщины (например, 5 повторов CTG для Hap I и 6 повторов CTG для Hap II), муж которой был поражен заболеванием, связанным с распространением повторов (например, 70 повторов CTG), была подвергнута секвенированию PacBio SMRT, секвенированные длинные молекулы ДНК были идентифицированы и использованы для определения гаплотипа и количества повторов. Если гаплотип, несущий длинный участок повтора CTG (например, 70 повторов CTG в этом примере), присутствует в материнской плазме непораженной беременной женщины, это указывает на то, что плод унаследовал пораженный отцовский гаплотип. Согласно некоторым вариантам реализации ДНК, содержащая распространившиеся повторы, также несет один или более других специфических для отца аллелей, которые отсутствуют в материнском геноме. Эту ситуацию можно применять для подтверждения наследования по отцовской линии.

На фиг. 81, 82 и 83 приведены таблицы, показывающие примеры заболеваний, связанных с повторами. В первом столбике показано заболевание, связанное с распространением повторов. Во втором столбике показана подпоследовательность повтора. В третьем столбике показано количество повторов у нормальных субъектов. В четвертом столбике показано количество повторов у пораженных субъектов. В пятом столбике показаны генетические местоположения, связанные с повторами. В шестом столбике перечислены названия генов. В седьмом столбике перечислены профили наследования. Таблица получена из omicslab.genetics.ac.cn/dred/index.php.

A. Примеры детектирования распространения повторов

Сообщалось, что распространившийся повтор CAG, унаследованный от отца, может детектироваться в материнской плазме с использованием прямого подхода с помощью ПЦР и последующего анализа фрагментов на генетическом анализаторе 3130XL (Oever et al. Prenat Diagn. 2015;35:945-9). Неинвазивное пренатальное тестирование для выявления болезни Хантингтона было возможным с помощью ПЦР, поскольку размер распространившегося аллеля начинается только с >35 тринуклеотидных повторов [т.е. область ДНК с 105 п.о. (35 × 3) или более в длину, охватывающая повторы]. Многие распространившиеся повторы, в частности, при большинстве нарушений, связанных с тринуклеотидными повторами (Orr et al. Annu. Rev. Neurosci. 2007;30:575-621), будут включать повторы длиной 300 п.о. или более, что превышает размер коротких молекул ДНК плода, которые были зарегистрированы в предыдущих сообщениях. ДНК с большими распространившимися повторами может вызвать трудности при ПЦР (Orr et al.

Annu. Rev. Neurosci. 2007;30:575-621). В исследовании Oever et al. было выдвинуто предположение, что интенсивность сигнала длинных повторов CAG часто намного ниже по сравнению с сигналом меньших повторов, и это явление наблюдается как в геномной ДНК, так и в ДНК плазмы, что приводит к более низкой чувствительности для детектирования этих длинных повторов CAG (Oever et al. Prenat Diagn. 2015;35:945-9). Другим ограничением ПЦР может быть невозможность сохранения сигналов метилирования во время амплификации. Согласно одному варианту реализации одномолекулярное секвенирование в реальном времени длинных молекул ДНК позволит определить полиморфизмы тандемных повторов и связанные с ними уровни метилирования в одной или более областях.

На фиг. 84 приведена таблица, показывающая примеры детектирования распространения повторов у плода и определения ассоциированного с повторами метилирования. В первом столбике показан тип повтора в виде числа пар оснований. Во втором столбике показано звено повтора. В третьем столбике показаны геномные местоположения. В четвертом столбике показаны референсные основания, последовательности, присутствующие в референсном геноме человека. В пятом столбике показаны отцовские генотипы. В шестом столбике показаны материнские генотипы. В седьмом столбике показаны генотипы плода. В восьмом столбике показан уровень метилирования ДНК плода, связанный с отцовскими аллелями. В девятом столбике показан уровень метилирования ДНК плода, связанный с материнскими аллелями.

На фиг. 84 показан ряд примеров тандемных повторов из 1 п.о., 2 п.о., 3 п.о. и 4 п.о. Например, в геномном положении chr3:192384705-192384706 был идентифицирован тандемный повтор «GATA». Генотип отца в этом локусе был $T(GATA)_3/T(GATA)_5$, для которого аллель 1 имел 3 звена повтора, а аллель 2 имел 5 звеньев повтора. По сравнению с референсным аллелем $T(GATA)_3$ отцовский аллель 2 предполагает генетическое событие, включающее распространение повтора. Генотип матери в этом локусе был T/T, что свидетельствует о генетическом событии, включающем сокращение повтора. Генотип плода в этом локусе был $T(GATA)_5/T$, это указывает на то, что плод унаследовал отцовский аллель 2 (т.е. $T(GATA)_5$) и материнский аллель T. Уровни метилирования, связанные с отцовским аллелем и материнским аллелем, составили 50,98 и 62,8, соответственно. Эти результаты позволили предположить, что применение полиморфизмов тандемных повторов позволит определить наследование по материнской и отцовской линии плода. Эта технология позволит идентифицировать разные профили метилирования, связанные с двумя аллелями. Другой пример показывает, что в геномном положении chr4:73237157-73237158 плод унаследовал от матери распространение повтора

[$(TAAA)_3$]. Молекула плода, содержащая распространение повтора, унаследованное от матери, показала более высокий уровень метилирования (95,65%) по сравнению с молекулой плода, содержащей отцовский аллель (62,84%). Эти данные позволяют предположить, что мы можем детектировать повторы, структуры повторов и связанные с ними изменения метилирования. Согласно одному варианту реализации можно применять конкретное отсечение для определения того, является ли значимой разница в метилировании между наследованием по материнской и отцовской линии. Отсечение будет представлять собой абсолютную разницу в уровнях метилирования, превышающую, но не ограничиваясь этим, 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85% или 90% и т.д. Определение наследования по материнской линии может быть аналогично способам, описанным для способа 2100 на фиг. 21.

B. Примерные способы

Повторы подпоследовательности можно применять для определения информации о плоде. Например, наличие повторов подпоследовательности можно применять для определения фетального происхождения молекулы. Кроме того, повторы подпоследовательности могут указывать на вероятность генетического нарушения. Повторы подпоследовательности можно применять для определения наследования материнских и/или отцовских гаплотипов. Кроме того, отцовство плода может быть определено с применением повторов подпоследовательности.

1. Анализ фетального происхождения с использованием повторов подпоследовательности

На фиг. 85 показан способ 8500 анализа биологического образца, полученного от субъекта женского пола, беременного плодом, причем указанный биологический образец включает молекулы внеклеточной ДНК от плода и от субъекта женского пола. Может быть определена вероятность генетического нарушения у плода.

В блоке 8510 может быть получен первый рид последовательности, соответствующий молекуле внеклеточной ДНК из молекул внеклеточной ДНК. Молекулы внеклеточной ДНК могут иметь длину, превышающую значение отсечки. Значение отсечки может быть больше или равно 200 нт. Значение отсечки может представлять собой по меньшей мере 500 нт., включая 600 нт., 700 нт., 800 нт., 900 нт., 1 тыс. нт., 1,1 тыс. нт., 1,2 тыс. нт., 1,3 тыс. нт., 1,4 тыс. нт., 1,5 тыс. нт., 1,6 тыс. нт., 1,7 тыс. нт., 1,8 тыс. нт., 1,9 тыс. нт. или 2 тыс. нт. Значение отсечки может представлять собой любое значение отсечки, описанное в настоящем документе для длинных молекул внеклеточной ДНК.

На этапе 8520 первый рид последовательности может быть выравнен с областью

референсного генома. Может быть известно, что область потенциально включает повторы подпоследовательности. Область может соответствовать любому из местоположений или генов на фиг. 81-83. Подпоследовательность может представлять собой тринуклеотидную последовательность, включая любую из описанных в настоящем документе.

В блоке 8530 может быть идентифицировано количество повторов подпоследовательности в первом риде последовательности, соответствующем молекуле внеклеточной ДНК.

В блоке 8540 количество повторов подпоследовательности можно сравнить с пороговым количеством. Пороговое количество может представлять собой 55, 60, 75, 100, 150 или более. Пороговое количество может быть разным для разных генетических нарушений. Например, порог может отражать минимальное количество повторов у пораженных субъектов, максимальное количество повторов у здоровых субъектов или количество между этими двумя количествами (см. фиг. 81-83).

В блоке 8550 классификация вероятности наличия у плода генетического нарушения может быть определена с использованием сравнения количества повторов с пороговым количеством. Может быть определено, что плод, вероятно, имеет генетическое нарушение, когда количество повторов превышает пороговое количество. Генетическое нарушение может представлять собой синдром ломкой Х-хромосомы или любое нарушение, перечисленное на фиг. 81-83.

Согласно некоторым вариантам реализации способ может включать повторение классификации для нескольких разных целевых локусов, каждый из которых, как известно, потенциально имеет повтор подпоследовательности. Может быть получено множество ридов последовательности, соответствующих молекулам внеклеточной ДНК. Множество ридов последовательностей может быть выравнено с множеством областей референсного генома. Может быть известно, что множество областей потенциально включает повторы подпоследовательностей. Множество областей может представлять собой неперекрывающиеся области. Каждая область из множества областей может иметь разные ОНП. Множество областей может происходить из разных хромосомных плеч или хромосом. Множество областей может охватывать по меньшей мере 0,01%, 0,1% или 1% референсного генома. Количества повторов подпоследовательностей могут быть идентифицированы во множестве ридов последовательности. Количества повторов подпоследовательностей можно сравнить с множеством пороговых количеств. Каждое пороговое количество может указывать на наличие или вероятность другого генетического нарушения. Для каждого из множества генетических нарушений классификация вероятности наличия у плода соответствующего генетического нарушения

может быть определена с использованием сравнения с пороговым количеством из множества пороговых количеств.

Можно определить, что молекула внеклеточной ДНК происходит от плода. Определение происхождения от плода может включать получение второго рида последовательности, соответствующего молекуле внеклеточной ДНК материнского происхождения, полученной из лейкоцитарной пленки или образца субъекта женского пола до беременности. Второй рид последовательности может быть выравнен с областью референсного генома. Второе количество повторов подпоследовательности может быть идентифицировано во втором риде последовательности. Может быть определено, что второе количество повторов меньшее первого количества повторов.

Определение происхождения от плода может включать определение уровня метилирования молекулы внеклеточной ДНК с использованием метилированных и неметилированных сайтов молекулы внеклеточной ДНК. Уровень метилирования можно сравнить с референсным уровнем. Способ может включать определение того, что уровень метилирования превышает референсный уровень. Уровень метилирования может представлять собой количество или долю сайтов, которые метилированы.

Определение происхождения от плода может включать определение профиля метилирования множества сайтов внеклеточной молекулы. Оценка сходства может быть определена путем сравнения профиля метилирования с референсным профилем из ткани матери или плода. Оценку сходства можно сравнить с одним или более пороговыми значениями. Оценка сходства может представлять собой любую оценку сходства, описанную в настоящем документе, включая, например, описанную для способа 4000.

2. Анализ отцовства с использованием повторов подпоследовательности

На фиг. 86 показан способ 8600 анализа биологического образца, полученного от субъекта женского пола, беременного плодом, причем указанный биологический образец включает молекулы внеклеточной ДНК от плода и субъекта женского пола. Биологический образец может быть проанализирован для определения отца плода.

В блоке 8610 может быть получено первый рид последовательности, соответствующий молекуле внеклеточной ДНК из молекул внеклеточной ДНК. Способ может включать определение того, что молекула внеклеточной ДНК происходит от плода. Происхождение молекулы внеклеточной ДНК от плода можно определить с помощью любого способа, описанного в настоящем документе, включая, например, как описано в способе 8500. Молекулы внеклеточной ДНК могут иметь размеры, превышающие значение отсечки. Значение отсечки может быть больше или равно 200 нт. Значение отсечки может представлять собой по меньшей мере 500 нт., включая 600 нт., 700 нт., 800

нт., 900 нт., 1 тыс. нт., 1,1 тыс. нт., 1,2 тыс. нт., 1,3 тыс. нт., 1,4 тыс. нт., 1,5 тыс. нт., 1,6 тыс. нт., 1,7 тыс. нт., 1,8 тыс. нт., 1,9 тыс. нт. или 2 тыс. нт. Значение отсечки может представлять собой любое значение отсечки, описанное в настоящем документе для длинных молекул внеклеточной ДНК.

В блоке 8620 первый рид последовательности может быть выравнен с первой областью референсного генома. Может быть известно, что первая область имеет повторы подпоследовательности.

В блоке 8630 может быть идентифицировано первое количество повторов первой подпоследовательности в первом риде последовательности, соответствующем молекуле внеклеточной ДНК. Первая подпоследовательность может включать аллель.

В блоке 8640 данные о последовательности, полученные от субъекта мужского пола, могут быть проанализированы, чтобы определить, присутствует ли второе количество повторов первой подпоследовательности в первой области. Второе количество повторов включает по меньшей мере две копии первой подпоследовательности. Данные о последовательности могут быть получены путем экстрагирования биологического образца у субъекта мужского пола и выполнения секвенирования ДНК в биологическом образце.

В блоке 8650 классификация вероятности того, что субъект мужского пола является отцом плода, может быть определена с использованием определения того, присутствует ли второе количество повторов первой подпоследовательности. Классификация может заключаться в том, что субъект мужского пола, вероятно, является отцом, когда определено наличие второго количества повторов первой подпоследовательности. Классификация может заключаться в том, что субъект мужского пола, вероятно, не является отцом, когда определено, что второе количество повторов первой подпоследовательности отсутствует.

Способ может включать сравнение первого количества повторов со вторым количеством повторов. Определение классификации вероятности того, что субъект мужского пола является отцом, может включать использование сравнения первого количества повторов со вторым количеством повторов. Классификация может заключаться в том, что субъект мужского пола, вероятно, является отцом, когда первое количество повторов находится в пределах порогового значения второго количества повторов. Пороговое значение может находиться в пределах 10%, 20%, 30% или 40% от второго количества повторов.

Способ может включать использование множества областей повторов. Например, молекула внеклеточной ДНК представляет собой первую молекулу внеклеточной ДНК. Способ может включать получение второго рида последовательности, соответствующего

второй молекуле внеклеточной ДНК из молекул внеклеточной ДНК. Способ также может включать выравнивание второго рида последовательности со второй областью референсного генома. Способ может дополнительно включать идентификацию первого количества повторов второй подпоследовательности во втором риде последовательности, соответствующем второй молекуле внеклеточной ДНК. Способ может включать анализ данных о последовательности, полученных от субъекта мужского пола, для определения того, присутствует ли второе количество повторов второй подпоследовательности во второй области. Определение классификации вероятности того, что субъект мужского пола является отцом плода, может дополнительно включать использование определения наличия второго количества повторов второй подпоследовательности во второй области. Классификация вероятности может представлять собой более высокую вероятность того, что субъект мужского пола является отцом плода, когда повторы присутствуют как в первой области, так и во второй области в данных о последовательности субъекта мужского пола.

VI. Отбор по размеру для обогащения длинными молекулами ДНК плазмы

Согласно вариантам реализации можно физически отобрать молекулы ДНК с одним или более целевыми диапазонами размеров перед анализом (например, одномолекулярное секвенирование в реальном времени). Например, отбор по размеру может быть выполнен с использованием технологии твердофазной обратимой иммобилизации. Согласно другим вариантам реализации отбор по размеру может быть выполнен с использованием электрофореза (например, с использованием системы Coastal Genomic или системы отбора по размеру Pippin). Наш подход отличается от предыдущей работы, в которой основное внимание уделялось более короткой ДНК (Li et al. JAMA 2005; 293: 843-9), поскольку в данной области техники известно, что ДНК плода короче ДНК матери (Chan et al. Clin Chem 2004; 50: 88-92).

Методики отбора по размеру могут быть применены к любому из способов, описанных в настоящем документе, и для любых размеров, описанных в настоящем документе. Например, молекулы внеклеточной ДНК могут быть обогащены с помощью электрофореза, магнитных гранул, гибридизации, иммунопреципитации, амплификации или CRISPR. Полученный обогащенный образец может иметь большую концентрацию или более высокую долю фрагментов определенного размера, чем биологический образец до обогащения.

A. Отбор по размеру с помощью электрофореза

Согласно вариантам реализации, в которых используется электрофоретическая подвижность ДНК в зависимости от размеров ДНК, можно использовать подходы,

основанные на гель-электрофорезе, для отбора целевых молекул ДНК с желаемыми диапазонами размеров, например, но не ограничиваясь перечисленными, ≥ 100 п.о., ≥ 200 п.о., ≥ 300 п.о., ≥ 400 п.о., ≥500 п.о., ≥ 600 п.о., ≥ 700 п.о., ≥ 800 п.о., ≥ 900 п.о., ≥ 1 тыс. п.о., ≥ 2 тыс. п.о., ≥ 3 тыс. п.о., ≥ 4 тыс. п.о., ≥ 5 тыс. п.о., ≥ 6 тыс. п.о., ≥ 7 тыс. п.о., ≥ 8 тыс. п.о., ≥ 9 тыс. п.о., ≥ 10 тыс. п.о., ≥ 20 тыс. п.о., ≥ 30 тыс. п.о., ≥ 40 тыс. п.о., ≥ 50 тыс. п.о., ≥ 60 тыс. п.о., ≥ 70 тыс. п.о., ≥ 80 тыс. п.о., ≥ 90 тыс. п.о., ≥ 100 тыс. п.о., ≥ 200 тыс. п.о. или другие, включая размеры, превышающие любое значение отсечки, описанное в настоящем документе. Например, для отбора по размеру ДНК использовали автоматизированную систему гель-электрофореза LightBench (Coastal Genomics). В целом во время гель-электрофореза более короткая ДНК будет двигаться быстрее, чем более длинная. Мы применили эту технологию отбора по размеру к одному образцу ДНК плазмы (M13190) с целью отбора молекул ДНК более 500 п.о. Мы использовали 3% кассету отбора по размеру с устройством для сбора «In-Channel-Filter» (ICF) и загрузочным буфером с внутренними маркерами размера для отбора по размеру. Библиотеки ДНК загружали в гель и начинали электрофорез. Когда целевой размер был достигнут, первая фракция <500 п.о. была извлечена из ICF. Прогон возобновили и завершили электрофорез для получения второй фракции ≥500 п.о. Мы использовали одномолекулярное секвенирование в реальном времени (PacBio) для секвенирования второй фракции с размером молекул ≥500 п.о. Мы получили 1434 высококачественные кольцевые консенсусные последовательности (CCS) (т.е. 1434 молекулы). Из них 97,9% секвенированных молекул были больше 500 п.о. Такая доля молекул ДНК более 500 п.о. была намного выше, чем у аналога без отбора по размеру (10,6%). Было определено, что общее метилирование этих молекул составляет 75,5%.

На фиг. 87 показаны профили метилирования для двух типичных молекул ДНК плазмы после отбора по размеру в (I) молекуле I и (II) молекуле II. Молекула I (chr21:40,881,731-40,882,812) имела длину 1,1 тыс. нт. и несла 25 сайтов CpG. При использовании подходов, описанных в нашем предыдущем изобретении (заявка США № 16/995607), было определено, что уровень метилирования отдельной молекулы (т.е. количество метилированных сайтов, деленное на общее количество сайтов) молекулы I составляет 72,0%. Молекула II (chr12:63,108,065-63,111,674) имела 3,6 тыс. п.о. в длину и несла 34 сайта CpG. Уровень метилирования отдельной молекулы для молекулы II был определен как 94,1%. Было высказано предположение, что анализ метилирования на основе отбора по размеру позволяет эффективно анализировать метилирование длинных молекул ДНК и сравнивать статус метилирования между двумя или более молекулами.

B. Отбор по размеру с использованием гранул

В технологии твердофазной обратимой иммобилизации используются парамагнитные гранулы для селективного связывания нуклеиновых кислот в зависимости от размера молекул ДНК. Такая гранула включает полистирольное ядро, магнетит и полимерное покрытие, модифицированное карбоксилатом. Молекулы ДНК будут селективно связываться с гранулами в присутствии полиэтиленгликоля (ПЭГ) и соли в зависимости от концентрации ПЭГ и соли в реакции. ПЭГ вызывал связывание отрицательно заряженной ДНК с карбоксильными группами на поверхности гранул, которые будут собраны в присутствии магнитного поля. Молекулы с целевыми размерами элюировали с магнитных гранул, используя буферы для элюции, например, 10 мМ трис-HCl, буфер с pH 8 или воду. Объемное соотношение ПЭГ и ДНК будет определять размеры молекул ДНК, которые можно получить. Чем ниже соотношение ПЭГ:ДНК, тем более длинные молекулы будут удерживаться на гранулах.

1. Обработка образцов

Образцы периферической крови от двух беременных женщин в третьем триместре собирали в пробирки для крови с ЭДТА. Образцы периферической крови собирали и центрифугировали при 1600× *g* в течение 10 мин при 4°C. Часть плазмы дополнительно центрифугировали при 16000×*g* в течение 10 минут при 4°C для удаления остаточных клеток и дебриса. Часть лейкоцитарной пленки центрифугировали при 5000×*g* в течение 5 минут при комнатной температуре для удаления остаточной плазмы. Плацентарные ткани собирали немедленно после родов. Экстракции ДНК плазмы выполняли с использованием набора для циркулирующих нуклеиновых кислот QIAamp (Qiagen). Экстракции ДНК лейкоцитарной пленки и плацентарной ткани выполняли с использованием мини-набора для ДНК QIAamp (Qiagen).

2. Отбор ДНК плазмы по размеру

Образцы ДНК плазмы после экстракции разделяли на две аликвоты. Одну аликвоту от каждого пациента подвергали отбору по размеру с использованием гранул AMPure XP SPRI (Beckman Coulter, Inc.). 50 мкл каждого экстрагированного образца ДНК плазмы тщательно смешивали с 25 мкл раствора AMPureXP и инкубировали при комнатной температуре в течение 5 минут. Гранулы отделяли от раствора магнитами и промывали 180 мкл 80% этанола. Затем гранулы ресуспендировали в 50 мкл воды и встряхивали в течение 1 минуты для элюирования отобранной по размеру ДНК из гранул. Затем гранулы удаляли для получения раствора ДНК, отобранных по размеру.

3. Идентификация однонуклеотидного полиморфизма

Образцы геномной ДНК плода и матери генотипировали с помощью системы iScan (Illumina). Определяли однонуклеотидные полиморфизмы (ОНП). Генотипы плаценты

сравнивали с генотипами матерей, чтобы идентифицировать специфические для плода и специфические для матери аллели. Специфический для плода аллель определяли как аллель, который присутствовал в геноме плода, но отсутствовал в геноме матери. Согласно одному варианту реализации эти специфические для плода аллели можно определить путем анализа тех сайтов ОНП, по которым мать была гомозиготной, а плод был гетерозиготным. Специфический для матери аллель определяли как аллель, который присутствовал в геноме матери, но отсутствовал в геноме плода. Согласно одному варианту реализации эти специфические для плода аллели можно определить путем анализа тех сайтов ОНП, по которым мать была гетерозиготной, а плод был гомозиготным.

4. Одномолекулярное секвенирование в реальном времени

Два отобранных по размеру образца вместе с соответствующими неотобранными образцами подвергали конструированию матрицы для одномолекулярного секвенирования в реальном времени (SMRT) с использованием набора для приготовления матрицы SMRTbell 1.0—SPv3 (Pacific Biosciences). ДНК очищали с использованием 1,8× гранул AMPure PB, а размер библиотеки оценивали с использованием прибора TapeStation (Agilent). Отжиг праймеров для секвенирования и условия связывания полимеразы рассчитывали с помощью программного обеспечения SMRT Link v5.1.0 (Pacific Biosciences). В общих чертах, праймер v3 для секвенирования отжигали с матрицей для секвенирования, а затем полимеразу связывали с матрицами с использованием набора для связывания с внутренним контролем Sequel® 2.1 (Pacific Biosciences). Секвенирование выполняли на Sequel SMRT Cell 1M v2. Фильмы секвенирования были собраны на системе Sequel в течение 20 часов с помощью набора для секвенирования Sequel® 2.1 (Pacific Biosciences).

5. Анализ размера

На фиг. 88 приведена таблица с информацией о секвенировании для образцов с отбором по размеру и без него. В первом столбике показан идентификатор образца. Во втором столбике показана группа образца, независимо от того, был ли отбор по размеру. В третьем столбике показано количество секвенированных молекул. В четвертом столбике показано среднее значение глубины подрида. В пятом столбике показан медианный размер фрагмента. В шестом столбике показана доля фрагментов, которые больше или равны 500 п.о.

Мы проанализировали два образца (299 и 300) с отбором по размеру на основе гранул и без него. Как показано на фиг. 88, мы получили 2,5 миллиона и 3,1 миллиона секвенированных молекул для образцов 299 и 300, соответственно, без отбора по размеру,

используя одномолекулярное секвенирование в реальном времени (например, секвенирование PacBio SMRT). Средние глубины подридов составили 91x и 67x. Медианы размеров фрагментов составили 176 и 512 п.о.

Для парных образцов (B299 и B300) с отбором по размеру на основе твердофазной обратимой иммобилизации с целью отбора фрагментов ДНК ≥500 п.о. мы получили, соответственно, 4,1 миллиона и 2,0 миллиона секвенированных молекул со средними глубинами подридов 18x и 19x. Было обнаружено, что медианные размеры фрагментов составляли 2,5 тыс. п.о. и 2,2 тыс. п.о. для образцов B299 и B300, соответственно. Средний размер фрагмента был в 4-14 раз больше, чем у соответствующих образцов без отбора по размеру. Доля фрагментов ≥500 п.о. после отбора по размеру увеличилась с 27,3% до 97,6% для образца B299 и с 50,5% до 97,4% для образца B300.

На фиг. 89A и 89B показаны распределения размеров образцов ДНК беременных субъектов женского пола с отбором по размеру на основе гранул и без него. На фиг. 89A показан образец 299, а на фиг. 89B показан образец 300. По оси X показан размер фрагментов. По оси Y показана частота для каждого размера фрагмента по логарифмической шкале. Более высокие частоты присутствовали среди длинных молекул ДНК более 1 тыс. п.о. в образцах ДНК после отбора по размеру на основе гранул. Эти данные свидетельствовали о том, что отбор по размеру на основе гранул может обогатить большим количеством длинных молекул ДНК для последующего анализа. Такое обогащение сделает анализ более рентабельным за счет максимального увеличения количества длинных молекул ДНК, секвенируемых за цикл секвенирования. Такое обогащение длинными молекулами ДНК также улучшит информативность при анализе тканей происхождения каждой молекулы ДНК, так как больше сайтов CpG каждой молекулы ДНК плазмы будет доступно для анализа на основе сопоставления профиля метилирования. Согласно одному варианту реализации анализ метилирования может был выполнен с использованием способа, описанного в заявке США № 16/995607. Нуклеосомальные профили сохранялись в образцах с отбором по размеру, это позволяет предположить, что молекулы ДНК плазмы, отобранные по размеру, будут подходящими для исследования нуклеосомальных структур.

Для образца 299 мы получили информацию о генотипе ДНК материнской лейкоцитарной пленки и ДНК плаценты с использованием технологии микрочипов (Infinium Omni2.5). Секвенированные молекулы ДНК плазмы были дифференцированы на специфические для матери и специфические для плода молекулы ДНК в соответствии с информацией о генотипе.

На фиг. 90A и 90B показаны распределения размеров между специфическими для

плода и специфическими для матери молекулами ДНК. Размер показан по оси X. На фиг. 90A частота показана по оси Y. На фиг. 90B суммарная частота показана по оси Y. На фиг. 90A распределение размеров ДНК плода показало более высокие частоты относительно более мелких молекул по сравнению с распределением размеров ДНК матери. На фиг. 90B такое укорочение размера молекулы ДНК плода было показано на графике суммарной частоты, т.е. суммарное распределение размеров ДНК плода располагалось слева от материнского.

C. Повышение информативности ДНК плазмы с помощью отбора по размеру.

Согласно вариантам реализации информативные ОНП могут быть определены по тем ОНП, которые содержат аллель, специфический для генома плода или матери. Эти ОНП обеспечивают средство для различения молекул ДНК плода и матери. Мы идентифицировали 419539 информативных ОНП. Согласно другим вариантам реализации информативные ОНП могут быть определены по тем ОНП, которые являются гетерозиготными в материнском геноме. Согласно другим вариантам реализации информативные ОНП могут быть определены по тем ОНП в материнском геноме, которые были гетерозиготными и которые были сгруппированы в виде гаплотипа.

На фиг. 91 приведена статистическая таблица для количества молекул ДНК плазмы, несущих информативные ОНП, среди образцов с отбором по размеру и без него. В первом столбике показан идентификатор образца и группа. Во втором столбике показано общее количество анализируемых молекул ДНК плазмы. В третьем столбике показано количество молекул ДНК плазмы, несущих информативные ОНП. В четвертом столбике показано процентное содержание молекул ДНК плазмы, несущих информативные ОНП.

Как показано на фиг. 91, в образце без отбора по размеру имелось только 6,5% молекул ДНК плазмы, несущих информативные ОНП, в то время как доля молекул ДНК плазмы, несущих информативные ОНП, увеличилась до 20,6%. Таким образом, использование отбора по размеру значительно улучшит выход длинных молекул ДНК, подходящих для применения согласно настоящему изобретению. Мы идентифицировали 260 молекул ДНК плода >500 п.о. в образце 299 без отбора по размеру и 918 молекул ДНК плода >500 п.о. в образце B299 с отбором по размеру. При нормировании производительности секвенирования эти данные свидетельствовали о том, что при использовании отбора по размеру на основе гранул имело место приблизительно 3-кратное обогащение при получении специфических для плода молекул ДНК >500 п.о. Благодаря отбору по размеру мы можем существенно увеличить количество длинных молекул ДНК плода для анализа.

D. Метилирование

На фиг. 92 приведена таблица уровней метилирования в образцах ДНК плазмы, отобранных по размеру и без отбора по размеру. В первом столбике показан идентификатор образца. Во втором столбике показана группа. В третьем столбике показано количество метилированных сайтов CpG. В четвертом столбике показано количество неметилированных сайтов CpG. В пятом столбике показан уровень метилирования, основанный на количестве метилированных сайтов и общем количестве сайтов. Как показано на фиг. 92, было показано, что общий уровень метилирования выше в образцах, отобранных по размеру, по сравнению с соответствующими образцами без отбора (71,5% в сравнении с 69,1% для образца 299 и B299 во всех сайтах CpG; 71,4% в сравнении с 69,3% для образца 300 и B300).

На фиг. 93 приведена таблица уровней метилирования в специфических для матери или специфических для плода молекулах внеклеточной ДНК. В первом столбике показан идентификатор образца. Во втором столбике показана группа. В третьем столбике показано количество метилированных сайтов CpG. В четвертом столбике показано количество неметилированных сайтов CpG. В пятом столбике показан уровень метилирования, основанный на количестве метилированных сайтов и общем количестве сайтов.

Как показано на фиг. 93, увеличение уровня метилирования также наблюдали как в специфических для плода, так и в специфических для матери молекулах ДНК плазмы в образце с отбором по размеру по сравнению с образцом без отбора по размеру. Эти специфические для плода фрагменты склонны к гипометилированию по сравнению со специфическими для матери молекулами ДНК в плазме как в образцах, отобранных по размеру, так и в образцах, не отобранных по размеру.

E. Концевые мотивы

На фиг. 94 приведена таблица 10 основных концевых мотивов в образцах с отбором по размеру и без него. В первом столбике показан ранг. Столбики со второго по пятый предназначены для образцов без отбора по размеру. Столбики с шестого по девятый предназначены для образцов с отбором по размеру. Во второй строке перечислены идентификаторы образцов. Во втором, четвертом, шестом и восьмом столбиках перечислен концевой мотив. В третьем, пятом, седьмом и девятом столбиках перечислена частота концевого мотива.

Как показано на фиг. 94, без отбора по размеру молекулы ДНК плазмы, секвенированные с помощью одномолекулярного секвенирования в реальном времени,

показывали концевые мотивы, преимущественно начинающиеся с C, что указывает на сигнатуру расщепления нуклеазой DNASE1L3 (Han et al., Am J Hum Genet 2020; 106: 202-214). Напротив, для образцов с отбором по размеру ДНК плазмы, секвенированная с помощью одномолекулярного секвенирования в реальном времени, несет концевые мотивы, преимущественно начинающиеся с A или G, что указывает на сигнатуру расщепления нуклеазой DFFB (Han et al. Am J Hum Genet 2020; 106: 202-214). Эти данные свидетельствовали о том, что отбор по размеру позволит селективно обогатить молекулы ДНК плазмы, полученные в результате различных ферментативных процессов фрагментации внеклеточной ДНК. Такое селективное нацеливание можно применять при анализе, детектировании или мониторинге нарушений, связанных с аберрантными уровнями одной или более нуклеаз. Согласно одному варианту реализации отбор по размеру ДНК плазмы может повысить эффективность мониторинга активности DFFB или кинетики опосредуемой DFFB деградации ДНК.

Согласно некоторым вариантам реализации секвенировали ДНК, связанную с гранулами, обогащенную длинной ДНК плазмы, и ДНК, оставшуюся в супернатанте, обогащенную короткой ДНК плазмы. Длинную ДНК можно применять для получения информации о гаплотипе. Короткую ДНК плазмы можно применять для мониторинга активности DNASE1L3. Согласно вариантам реализации можно выполнять синергетический комбинированный анализ длинных и коротких молекул ДНК. Например, при выравнивании короткой ДНК плазмы с материнскими гаплотипами (т.е. Hap I и Hap II) один материнский гаплотип, проявляющий больше короткой ДНК и/или большее гипометилирование, и/или относительно более высокую дозу, вероятно, будет унаследован плодом по сравнению с другим гаплотипом.

Согласно некоторым вариантам реализации отбор по размеру может быть основан, но не ограничивается перечисленными, на технологиях на основе гель-электрофореза, таких как отбор по размеру ДНК PippinHT, отбор по размеру ДНК BluePippin, система отбора по размеру ДНК Pippin Prep, система фракционирования цельных образцов SageELF, импульсный электрофорез Pippin, библиотечная система SageHLS HMW и т.д.

F. Длинные молекулы ДНК плазмы повышают эффективность анализа ткани происхождения

На фиг. 95 показан график операционных характеристик приемника (ROC), показывающий, что длинные молекулы ДНК плазмы повышают эффективность анализа ткани происхождения. По оси Y показана чувствительность. По оси X показана специфичность. Разными линиями показаны результаты для фрагментов разного размера. Красная линия с наибольшей площадью под кривой (AUC) относится к фрагментам более

3000 п.о.

Как показано на фиг. 95, при установлении различия между молекулами ДНК плода и матери в плазме беременных женщин эффективность на основании длинных молекул ДНК плазмы (например, >3000 п.о.) (AUC: 0,94) в соответствии с вариантами реализации настоящего изобретения была намного выше, чем таковая для анализов, основанных на относительно коротких молекулах ДНК, таких как 100-200 п.о. (AUC: 0,66) и 200-500 п.о. (AUC: 0,67). Эти данные свидетельствовали о том, что применение длинной ДНК плазмы значительно повысит точность установления различия между молекулами ДНК плода и матери, что приведет к более высокой эффективности определения наследственности плода неинвазивным способом.

VII. Нанопоровое секвенирование для анализа длинной ДНК из ДНК материнской плазмы

В дополнение к использованию технологии одномолекулярного секвенирования в реальном времени нанопоровое секвенирование может использоваться для секвенирования длинных фрагментов внеклеточной ДНК из материнской плазмы. Информация о метилировании и ОНП может повысить точность нанопорового секвенирования длинных фрагментов внеклеточной ДНК.

На фиг. 96 показан принцип нанопорового секвенирования ДНК плазмы, полученной от беременной женщины, в котором последовательность нуклеиновых кислот выводится из изменений ионного тока через мембрану, когда отдельная молекула ДНК проходит через пору нанометрового размера. Такая пора может быть создана, например, но не ограничиваясь перечисленными, белком (например, альфа-гемолизином, аэролизином и порином А *Mycobacterium smegmatis* (MspA)) или синтетическими материалами, такими как кремний или графен (Magi et al, Brief Bioinform. 2018;19:1256-1272). Согласно вариантам реализации двухцепочечные молекулы ДНК плазмы подвергаются процессу репарации концов. Такой процесс будет превращать ДНК плазмы в ДНК с тупыми концами с последующим добавлением А-хвоста. Адаптеры последовательности, каждый из которых несет моторный белок (т.е. моторный адаптер), лигируют с любым концом молекулы ДНК плазмы, как показано на фиг. 96. Процесс секвенирования начинается, когда моторный белок раскручивает двухцепочечную ДНК, позволяя первой цепи пройти через нанопору. Когда цепь ДНК проходит через нанопору, датчик измеряет изменения ионного тока (pA) с течением времени, которые зависят от контекста последовательности и связанных модификаций оснований (называется одномерным считыванием). Согласно другим вариантам реализации для ковалентного связывания первой цепи и комплементарной цепи вместе могут использоваться адаптеры

на основе последовательностей шпилек. Во время секвенирования секвенируется цепь двухцепочечной молекулы ДНК, за которой следует комплементарная цепь (называется $1D^2$ или 2D считыванием), что потенциально может повысить точность секвенирования. Необработанные сигналы тока используются для определения основания и анализа модификации основания. Согласно другим вариантам реализации определение основания и анализы модификации основания проводят с помощью подхода машинного обучения, например, но не ограничиваясь ими, рекуррентной нейронной сети (RNN) или скрытой модели Маркова (HMM). В настоящем изобретении мы представили способы характеристики свойств молекул ДНК плазмы, включая, но не ограничиваясь перечисленными, количества молекул, составы оснований, размеры молекул, концевые мотивы и модификации оснований, с использованием нанопорового секвенирования.

В иллюстративных целях мы использовали нанопоровое секвенирование (Oxford Nanopore Technologies) для секвенирования трех образцов ДНК материнской плазмы (M12970, M12985 и M12969) беременных женщин при гестационном возрасте 38 недель. ДНК плазмы, экстрагированную из 4 мл материнской плазмы, подвергали подготовке библиотеки с использованием набора для секвенирования лигированием (Oxford Nanopore). В общих чертах, ДНК репарировали с помощью смеси для репарации FFPE (NEB), затем репарировали концы и присоединяли A-хвост с помощью модуля NEBNext End Prep (NEB). Затем к репарированной ДНК добавляли адаптерную смесь и лигировали с мастер-смесью тупой конец/TA. После очистки с использованием гранул AMPure XP (Beckman) лигированную с адаптером библиотеку смешивали с буфером для секвенирования и загрузочными гранулами и загружали в проточную кювету PromethION R9. Проточную кювету секвенировали на устройстве PromethION бета (Oxford Nanopore) в течение 64 часов.

A. Выравнивание

Секвенированные считывания выравнивали с референсным геномом человека (hg19) с использованием Minimap2 (Li H, Bioinformatics. 2018;34(18):3094-3100). Согласно некоторым вариантам реализации для выравнивания секвенированных ридов с референсным геномом можно использовать BLASR (Mark J Chaisson et al, BMC Bioinformatics. 2012; 13: 238), BLAST (Altschul SF et al, J Mol Biol. 1990;215(3):403-410), BLAT (Kent WJ, Genome Res. 2002;12(4):656-664), BWA (Li H et al, Bioinformatics. 2010;26(5):589-595), NGMLR (Sedlazeck FJ et al, Nat Methods. 2018;15(6):461-468) и LAST (Kielbasa SM et al, Genome Res. 2011;21(3):487-493). Мы получили 11,31, 12,30 и 21,28 миллиона секвенированных молекул для образцов M12970, M12985 и M12969, соответственно. Количество картированных фрагментов из

них составило 3,67, 2,63 и 4,33 миллиона, соответственно.

B. Размер и метилирование

Количество нуклеотидов в молекуле ДНК плазмы, определенное с помощью нанопорового секвенирования, использовали для выведения размера этой молекулы ДНК. Токовые сигналы молекулы ДНК можно использовать для определения модификаций оснований. Согласно вариантам реализации статус метилирования для каждого сайта CpG определяли с помощью программного обеспечения с открытым исходным кодом Nanopolish (Simpson et al, Nat Methods. 2017;14:407-410). Согласно другому варианту реализации статус метилирования может быть определен с использованием другого программного обеспечения, включая, но не ограничиваясь перечисленными, DeepMod (Liu et al, Nat Commun. 2019;10:2449), Tomo (Stoiber et al, BioRxiv. 2017:p.094672), DeepSignal (Ni et al, Bioinformatics. 2019;35:4586-4595), Guppy (github.com/nanoporetech), Megalodon (github.com/nanoporetech/megalodon) и т.д.

На фиг. 97 приведена таблица процентного содержания молекул ДНК плазмы в определенном диапазоне размеров и их соответствующих уровней метилирования. Показаны три образца: M12970, M12985 и M12969. В первом столбике показан размер фрагмента. Во втором столбике показано количество фрагментов данного размера фрагмента. В третьем столбике показана частота данного размера фрагмента. В четвертом столбике показано количество метилированных сайтов CpG данного размера фрагмента. В пятом столбике показано количество неметилированных сайтов CpG данного размера фрагмента. В шестом столбике показан уровень метилирования как процентное содержание.

Как показано на фиг. 97, доли молекул ДНК с размером ≥500 п.о. были 16,6%, 7,6% и 12,6% для образцов M12970, M12985 и M12969, соответственно. Доля молекул ДНК с размером ≥500 п.о. была намного выше, по сравнению с данными, сгенерированными с помощью секвенирования Illumina (0,2%). Уровни метилирования молекул ДНК с размером ≥500 п.о. были 64,12%, 65,05% и 63,30% для образцов M12970, M12985 и M12969, соответственно. Кроме того, уровень метилирования повышался в популяции с большим количеством длинных ДНК плазмы. Например, для образца M12970 уровень метилирования был 70,7% в молекулах с размером ≥2000 п.о., что было эквивалентно увеличению уровня метилирования на 10,3% по сравнению с молекулами с размером ≥500 п.о. Сходную тенденцию к увеличению в популяции с большим количеством длинных ДНК также наблюдали в образцах M12985 и M12969. Молекулы ДНК плазмы с разными размерами будут отражать различные пути поступления внеклеточной ДНК в кровообращение, такие как, но не ограничиваясь перечисленными, старение, апоптоз,

некроз, активная секреция и т.д. Статус метилирования длинной молекулы ДНК также позволит сделать вывод о тканях происхождения этих длинных молекул ДНК. Таким образом, комбинированный анализ профилей фрагментации длинных молекул ДНК и профилей метилирования позволит сделать вывод об относительных соотношениях старения, апоптоза, некроза и активной секреции для конкретного органа. Относительные соотношения образования внеклеточной ДНК различными путями будут отражать лежащие в основе патофизиологические состояния, такие как беременность, преэклампсия, преждевременные роды, задержка внутриутробного развития и т.д.

На фиг. 98 показан график распределения размеров и профилей метилирования по разным размерам. Размер показан по оси X. Частота показана по левой оси Y. Уровень метилирования показан по правой оси Y. Данные о распределении размеров (частоте) показаны черной линией. Показанный уровень метилирования показан желтой линией.

На фиг. 98 показано распределение размеров и уровни метилирования по фрагментам разного размера. Распределение размеров имело несколько пиков при 164 п.о., 313 п.о. и 473 п.о. со средним интервалом 154 п.о. Такие профили распределения размеров напоминали нуклеосомы, расщепленные нуклеазами, это позволяет предположить, что неслучайный процесс фрагментации ДНК плазмы может быть идентифицирован с помощью нанопорового секвенирования. В отличие от размерных профилей ДНК плазмы с основным пиком при 166 п.о. на основе данных секвенирования Illumina основной пик был при 380 п.о. Эти данные указывали на то, что нанопоровое секвенирование будет приводить к обогащению большим количеством длинных фрагментов ДНК. Такая характеристика нанопорового секвенирования ДНК плазмы будет особенно полезной для детектирования тех вариантов, которые трудно выявить с помощью технологий секвенирования с короткими ридами. Согласно вариантам реализации нанопоровое секвенирование можно применять для анализа распространения триплетных повторов. Количество тринуклеотидных повторов можно применять для прогнозирования прогрессирования, тяжести и возраста начала нарушений, связанных с тринуклеотидными повторами, таких как синдром ломкой X-хромосомы, болезнь Хантингтона, спиноцеребеллярные атаксии, миотоническая дистрофия и атаксия Фридрейха. На фиг. 98 также показано, что уровни метилирования изменялись в зависимости от разных размеров. Ряд значений пиков метилирования совпадал с пиками распределения размеров.

C. ДНК плода и матери

С помощью генотипирования ДНК, экстрагированной из материнской лейкоцитарной пленки и плаценты, с использованием платформы iScan (Illumina) мы

определили медиану 204410 информативных ОНП (диапазон: 199420–205597), по которым мать была гомозиготной (AA), а плод был гетерозиготным (AB), которые использовались для определения специфических для плода аллелей (B) и общих аллелей (A).

На фиг. 99 приведена таблица фракции ДНК плода, определенной с использованием нанопорового секвенирования. В первом столбике показан идентификатор образца. Во втором столбике показано количество молекул, несущих общие аллели. В третьем столбике показано количество молекул, несущих специфические для плода аллели. В четвертом столбике показана фракция ДНК плода, рассчитанная путем умножения значения в третьем столбике на два и деления на сумму значений второго столбика и третьего столбика. Как показано на фиг. 99, мы идентифицировали 84911, 52059 и 95273 молекулы, несущие общие аллели, и 17776, 7385 и 17007 молекул, несущих специфические для плода аллели, для образцов M12970, M12985 и M12969, соответственно. Было определено, что фракции ДНК плода составляют 34,6%, 24,9% и 30,3% для образцов M12970, M12985 и M12969, соответственно. Кроме того, мы определили медиану 212330 информативных ОНП (диапазон: 210411–214744), по которым мать была гетерозиготной (AB), а плод был гомозиготным (AA), которые были использованы для определения специфических для матери аллелей (B). Мы идентифицировали 65349, 34017 и 65481 молекул, несущих общие аллели, и 43594, 26704 и 48337 молекул, несущих специфические для матери аллели, для образцов M12970, M12985 и M12969, соответственно.

На фиг. 100 приведена таблица уровней метилирования между специфическими для плода и специфическими для матери молекулами ДНК. В первом столбике показан идентификатор образца. Во втором, третьем и четвертом столбиках показаны результаты для специфической для плода ДНК. В пятом, шестом и седьмом столбиках показаны результаты для специфической для матери ДНК. Во втором и пятом столбиках показано количество метилированных сайтов CpG. В третьем и шестом столбиках показано количество неметилированных сайтов CpG. В четвертом и седьмом столбиках показан уровень метилирования, основанный на процентном содержании метилированных сайтов.

В соответствии с вариантами реализации настоящего изобретения определяли профили метилирования для каждой специфической для плода молекулы ДНК. Доля секвенированных сайтов CpG, которые, как определено, являются метилированными (т.е. общие уровни метилирования), была 62,43%, 62,39% и 61,48% для образцов M12970, M12985 и M12969, соответственно, как показано на фиг. 100. Такие общие уровни метилирования специфической для плода ДНК были в среднем на 8% ниже, чем

аналогичной специфической для матери ДНК. Эти результаты свидетельствовали о том, что можно установить различие между молекулами ДНК плода и молекулами ДНК матери на основе дифференциальных профилей метилирования между молекулами ДНК плода и матери в соответствии с вариантами реализации настоящего изобретения с использованием результатов нанопорового секвенирования.

На фиг. 101 приведена таблица процентного содержания молекул ДНК плазмы в определенном диапазоне размеров и их соответствующих уровней метилирования для молекул ДНК плода и матери. Показаны три образца: M12970, M12985 и M12969. В первом столбике показан размер фрагмента. В столбиках со второго по шестой показаны результаты для специфической для плода ДНК. В столбиках с седьмого по одиннадцатый показаны результаты для специфической для матери ДНК. Во втором и седьмом столбиках показано количество фрагментов данного размера фрагмента. В третьем и восьмом столбиках показана частота данного размера фрагмента. В четвертом и девятом столбиках показано количество метилированных сайтов CpG для данного размера фрагмента. В пятом и десятом столбиках показано количество неметилированных сайтов CpG для данного размера фрагмента. В шестом и одиннадцатом столбиках показан уровень метилирования как процентное содержание.

Как видно на фиг. 101, свойства специфических для плода и специфических для матери молекул ДНК были проанализированы в различных диапазонах размеров, включая, но не ограничиваясь перечисленными, ≥500 п.о., ≥ 600 п.о., ≥ 1000 п.о. и ≥ 2000 п.о. По сравнению с молекулами ДНК матери мы получили относительно меньшую долю молекул ДНК плода размером более 1 тыс. п.о. Однако количество таких длинных молекул ДНК плода (например, ≥1000 п.о.) в плазме беременных женщин (диапазон: 4,9% - 9,3%) было значительно выше, чем значение, ожидаемое по результатам секвенирования Illumina (<0,2%). Такие длинные фрагменты ДНК плода нелегко выявить с помощью обычных технологий секвенирования с короткими ридами, таких как платформы секвенирования Illumina (например, но не ограничиваясь перечисленными, MiSeq, NextSeq, HiSeq, NovaSeq и т.д.), поскольку вставки в библиотеке ДНК ограничены размером менее 550 п.о. (например, система Illumina NextSeq system, support.illumina.com/sequencing/sequencing_instruments/nextseq-550/questions.html).

Согласно вариантам реализации анализ длинных фрагментов ДНК плода и матери, включая, но не ограничиваясь ими, размеры и профили метилирования, может обеспечить новый инструмент для оценки различных заболеваний. Например, дефицит DNASE1L3 вызывает моногенную системную красную волчанку. Такой дефицит DNASE1L3 может привести к образованию большего количества длинных молекул ДНК (Chan et al, Am J

Hum Genet. 2020;107:882-894). Таким образом, варианты реализации, описанные в настоящем документе, будут особенно чувствительными для мониторинга тяжести заболевания этих пациентов во время беременности и оценки того, будет ли нерожденный плод поражен этим же состоянием, путем анализа характеристик этих длинных молекул ДНК.

На фиг. 102A и 102B показаны графики распределения размеров молекул ДНК плода и матери, определенные с помощью нанопорового секвенирования. Размер фрагментов показан по оси X. Частота показана по оси Y по линейной шкале на фиг. 102A и логарифмической шкале на фиг. 102B. ДНК матери показана синей линией. ДНК плода показана красной линией.

Как показано на фиг. 102A и 102B, молекулы ДНК как матери, так и плода содержали больше длинных молекул ДНК, чем сообщалось ранее (Lo et al, Sci Transl Med. 2020;2:61ra91) на платформе секвенирования Illumina с короткими ридами. Эти результаты позволили предположить, что анализ ДНК плазмы с помощью нанопорового секвенирования выявил ряд новых характеристик внеклеточной ДНК, которые ранее не оценивались. Такие характеристики могут быть использованы в неинвазивном пренатальном тестировании.

D. Улучшенная точность определения молекул ДНК плода и матери

Поскольку нанопоровое секвенирование будет сопровождаться более высокой ошибкой секвенирования (от ~ 5% до 40%) (Goodwin et al, Genome Res. 2015;25:1750-1756), это может привести к неточной классификации молекул ДНК плода и матери на основании информации о генотипе ОНП. Согласно вариантам реализации можно применять два или более информативных ОНП для оценки фрагмента и определения того, произошел ли этот фрагмент из плаценты или нет. Например, для фрагмента, несущего два информативных ОНП, по которым мать была гомозиготной (AA), а плод был гетерозиготным (AB), только когда два информативных ОНП оба подтверждали вывод о том, что такой фрагмент происходил от плода, будет определено его происхождение от плода. Подобным образом, для фрагмента, несущего два информативных ОНП, только когда два информативных ОНП оба подтверждали, что такой фрагмент происходил от матери, будет определено его происхождение от матери.

На фиг. 103 показан график, показывающий разницу в уровнях метилирования между молекулами ДНК плода и матери на основе одного информативного ОНП и двух информативных ОНП. По оси Y показана разница в уровне метилирования как процентное содержание между молекулами ДНК плода и матери. По оси X показано использование одного информативного ОНП по сравнению с использованием двух

информативных ОНП для разницы в уровнях метилирования.

Как показано на фиг. 103, при использовании двух информативных ОНП для установления различия между молекулами ДНК плода и матери разница в уровнях метилирования между молекулами ДНК плода и матери была намного больше, чем по результатам, основанным на одном информативном ОНП. Средняя разница в уровне метилирования между специфическими для плода и специфическими для матери молекулами увеличилась с 5,4% до 11,3%, что эквивалентно увеличению на 109%. Эти результаты свидетельствовали о том, что применение нескольких ОНП значительно улучшит точность установления различия между специфическими для плода и специфическими для матери молекулами ДНК.

На фиг. 104 приведена таблица различия в уровнях метилирования между молекулами ДНК плода и матери. В первом столбике показан идентификатор образца. Во втором, третьем и четвертом столбиках показаны результаты для специфической для плода ДНК. В пятом, шестом и седьмом столбиках показаны результаты для специфической для матери ДНК. Во втором и пятом столбиках показано количество метилированных сайтов CpG. В третьем и шестом столбиках показано количество неметилированных сайтов CpG. В четвертом и седьмом столбиках показан уровень метилирования, основанный на процентном содержании метилированных сайтов.

Как видно на фиг. 104, такие общие уровни метилирования специфической для плода ДНК были в среднем на 16,3% ниже, чем аналогичной специфической для матери ДНК. Согласно вариантам реализации применение сигналов метилирования, в свою очередь, повысит точность классификации ДНК плода и матери. Например, для фрагмента, несущего предполагаемый специфический для плода аллель, когда было определено, что уровень метилирования этого фрагмента ниже порога, такой фрагмент будет иметь более высокую вероятность происхождения от плода. Такой порог может представлять собой, но не ограничивается перечисленными, 60%, 50%, 40%, 30%, 20%, 10% и т.д. Для фрагмента, несущего предполагаемый специфический для матери аллель, когда было определено, что уровень метилирования этого фрагмента превышает порог, такой фрагмент будет иметь более высокую вероятность происхождения от матери. Такой порог может представлять собой, но не ограничивается перечисленными, 90%, 80%, 70%, 60%, 50%, 40% и т.д.

В некоторых других вариантах реализации общее количество информативных ОНП должно быть по меньшей мере, например, но не ограничиваясь этим, 3, 4, 5, 6, 7, 8, 9, 10 и т.д. Количество информативных ОНП, подтверждающих происхождение фрагмента от плода, должно быть по меньшей мере, например, но не ограничиваясь этим, 3, 4, 5, 6, 7, 8,

9, 10 и т.д. Количество информативных ОНП, подтверждающих происхождение фрагмента от матери, должно быть по меньшей мере, например, но не ограничиваясь этим, 3, 4, 5, 6, 7, 8, 9, 10 и т.д. Согласно вариантам реализации процентное содержание информативных ОНП, подтверждающих происхождение фрагмента от плода, должно достигать определенного порога, например, 1%, 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% или 100%. Процентное содержание информативных ОНП, подтверждающих происхождение фрагмента от матери, должно достигать определенного порога, например, 1%, 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% или 100%.

В некоторых других вариантах реализации можно замкнуть в кольцо молекулы ДНК плазмы с последующей амплификацией по типу катящегося цикла. Амплифицированная ДНК может быть секвенирована с помощью нанопорового секвенирования, таким образом, информация о матричной ДНК может быть секвенирована несколько раз. Консенсусная последовательность может быть выведена на основании информации о многократно секвенированной последовательности.

VIII. Примеры систем

На фиг. 105 проиллюстрирована система измерения 10500 в соответствии с вариантом реализации настоящего изобретения. Показано, что система включает образец 10505, такой как молекулы внеклеточной ДНК, в устройстве анализа 10510, причем анализ 10508 может быть выполнен на образце 10505. Например, образец 10505 можно привести в контакт с реагентами анализа 10508, чтобы обеспечить сигнал физической характеристики 10515. Примером устройства анализа может быть проточная кювета, которая содержит зонды и/или праймеры для анализа, или пробирка, через которую движется капля (с каплей, содержащей систему анализа). Физическую характеристику 10515 (например, интенсивность флуоресценции, напряжение или ток) образца детектируют с помощью детектора 10520. Детектор 10520 может выполнять измерение с интервалами (например, периодическими интервалами) с получением точек данных, которые составляют сигнал данных. Согласно одному варианту реализации аналого-цифровой преобразователь несколько раз преобразует аналоговый сигнал от детектора в цифровую форму. Устройство анализа 10510 и детектор 10520 могут формировать систему анализа, например, систему секвенирования, которая выполняет секвенирование в соответствии с вариантами реализации, описанным в настоящем документе. Сигнал данных 10525 отправляется из детектора 10520 к логической системе 10530. Например, сигнал данных 10525 можно использовать для определения последовательностей и/или местоположений в референсном геноме молекул ДНК. Сигнал данных 10525 может включать различные измерения, выполненные одновременно, например, разные цвета

флуоресцентных красителей или разные электрические сигналы для разных молекул образца 10505, и, таким образом, сигнал данных 10525 может соответствовать нескольким сигналам. Сигнал данных 10525 может храниться в локальной памяти 10535, внешней памяти 10540 или запоминающем устройстве 10545.

Логическая система 10530 может представлять собой или может включать компьютерную систему, ASIC, микропроцессор, графический процессор (GPU) и т.д. Она также может содержать или может быть соединена с дисплеем (например, монитором, светодиодным дисплеем и т.д.) и устройством ввода пользователя (например, мышью, клавиатурой, кнопками и т.д.). Логическая система 10530 и другие компоненты могут быть частью автономной или подключенной к сети компьютерной системы, или они могут быть непосредственно присоединены к устройству или включены в него (например, устройство секвенирования), которое включает детектор 10520 и/или устройство анализа 10510. Логическая система 10530 также может включать программное обеспечение, которое выполняется в процессоре 10550. Логическая система 10530 может включать машиночитаемый носитель, хранящий инструкции для управления системой измерения 10500, для осуществления любого из способов, описанных в настоящем документе. Например, логическая система 10530 может подавать команды в систему, которая включает устройство анализа 10510, так что выполняется секвенирование или другие физические операции. Такие физические операции могут выполняться в определенном порядке, например, с добавлением и удалением реагентов в определенном порядке. Такие физические операции может выполнять робототехническая система, например, включая роботизированную руку, которую можно использовать для получения образца и выполнения анализа.

Система измерения 10500 также может включать устройство лечения 10560, которое может обеспечить лечение субъекта. Устройство лечения 10560 может определять лечение и/или может использоваться для выполнения лечения. Примеры такого лечения могут включать хирургическое вмешательство, лучевую терапию, химиотерапию, иммунотерапию, таргетную терапию, гормональную терапию и трансплантацию стволовых клеток. Логическая система 10530 может быть подключена к устройству лечения 10560, например, для обеспечения результатов способа, описанного в настоящем документе. Устройство лечения может получать входные данные от других устройств, таких как устройство формирования изображения, и вводимые пользователем данные (например, для управления лечением, например, для управления роботизированной системой).

Любая из упомянутых в настоящем документе компьютерных систем может

использовать любое подходящее количество подсистем. Примеры таких подсистем показаны на фиг. 106 в компьютерной системе 10. Согласно некоторым вариантам реализации компьютерная система включает одно компьютерное устройство, причем подсистемы могут быть компонентами компьютерного устройства. Согласно другим вариантам реализации компьютерная система может включать множество компьютерных устройств, каждое из которых является подсистемой, с внутренними компонентами. Компьютерная система может включать настольные и портативные компьютеры, планшеты, мобильные телефоны и другие мобильные устройства.

Подсистемы, показанные на фиг. 106, соединены между собой системной шиной 75. Показаны дополнительные подсистемы, такие как принтер 74, клавиатура 78, запоминающее устройство(а) 79, монитор 76 (например, экран дисплея, такой как светодиодный дисплей), который соединен с адаптером дисплея 82, и другие. Периферийные устройства и устройства ввода/вывода (ВВОД/ВЫВОД), которые соединены с контроллером 71 ввода/вывода, могут быть подключены к компьютерной системе с помощью любого количества средств, известных в данной области техники, таких как порт 77 ввода/вывода (ВВОД/ВЫВОД) (например, USB, FireWire®). Например, порт 77 ВВОДА/ВЫВОДА или внешний интерфейс 81 (например, Ethernet, Wi-Fi и т.д.) можно использовать для подключения компьютерной системы 10 к глобальной сети, такой как Интернет, устройству ввода с помощью мыши, или сканеру. Межкомпонентное соединение через системную шину 75 позволяет центральному процессору 73 обмениваться данными с каждой подсистемой, и контролировать выполнение множества инструкций из системной памяти 72 или запоминающего устройства(ств) 79 (например, несъемного диска, такого как жесткий диск, или оптический диск), а также обмен информацией между подсистемами. Системная память 72 и/или запоминающее устройство(а) 79 может представлять собой машиночитаемый носитель. Другой подсистемой является устройство сбора данных 85, такое как камера, микрофон, акселерометр и т.п. Любые из упомянутых в настоящем документе данных могут выводиться из одного компонента в другой компонент и могут выводиться пользователю.

Компьютерная система может включать множество одинаковых компонентов или подсистем, например, связанных друг с другом с помощью внешнего интерфейса 81, внутреннего интерфейса или через съемные запоминающие устройства, которые можно подключать и переносить от одного компонента к другому компоненту. Согласно некоторым вариантам реализации компьютерные системы, подсистемы или устройства могут обмениваться данными по сети. В таких случаях один компьютер может считаться клиентом, а другой компьютер - сервером, причем каждый из них может быть частью

одной и той же компьютерной системы. Каждый клиент и сервер могут включать множество систем, подсистем или компонентов.

Аспекты вариантов реализации можно реализовать в форме логики управления с использованием аппаратных схем (например, специализированной интегральной схемы или программируемой матрицы логических элементов) и/или с использованием компьютерного программного обеспечения с, в целом, программируемым процессором в модульном или интегрированном виде. В контексте настоящего изобретения процессор может включать одноядерный процессор, многоядерный процессор на одном интегрированном кристалле, или множество процессорных блоков на одной печатной плате или в сети, а также специализированное оборудование. На основе изобретения и идей, представленных в настоящем документе, специалист в данной области техники будет осведомлен о других путях и/или способах реализации вариантов реализации настоящего изобретения с использованием аппаратных средств и комбинации аппаратных средств и программного обеспечения и примет их во внимание.

Любой из программных компонентов или функций, описанных в данной заявке, можно реализовать в виде программного кода, который должен выполняться процессором с использованием любого подходящего языка программирования, например, Java, C, C++, C#, Objective-C, Swift или скриптового языка программирования, например, Perl или Python, с использованием, например, стандартных или объектно-ориентированных методик. Программный код может храниться в виде серии инструкций или команд на машиночитаемом носителе для хранения и/или передачи. Подходящий энергонезависимый машиночитаемый носитель может включать оперативную память (ОЗУ), постоянное запоминающее устройство (ПЗУ), магнитный носитель, например, жесткий диск или гибкий диск, или оптический носитель, например, компакт-диск (CD) или DVD (цифровой универсальный диск), или диск Blu-ray, флэш-память и т.п. Машиночитаемый носитель может быть любой комбинацией таких устройств хранения или передачи.

Такие программы также можно кодировать и передавать с использованием несущих сигналов, адаптированных для передачи через проводные, оптические и/или беспроводные сети, соответствующих множеству протоколов, включая Интернет. По существу, машиночитаемый носитель можно создать с использованием сигнала данных, закодированного с помощью таких программ. Машиночитаемый носитель, закодированный программным кодом, можно упаковать с совместимым устройством или представить отдельно от других устройств (например, путем загрузки через Интернет). Любой такой машиночитаемый носитель может поставляться с или в составе отдельного

компьютерного продукта (например, жесткого диска, компакт-диска или всей компьютерной системы), а также с или в составе различных компьютерных продуктов в пределах системы или сети. Компьютерная система может включать монитор, принтер или другой подходящий дисплей для предоставления пользователю любых результатов, упомянутых в настоящем документе.

Любой из описанных в настоящем документе способов может быть полностью или частично выполнен компьютерной системой, включающей один или более процессоров, которые можно сконфигурировать для выполнения поэтапных действий. Таким образом, варианты реализации могут быть направлены на компьютерные системы, сконфигурированные для выполнения этапов любого из способов, описанных в настоящем документе, возможно, с различными компонентами, выполняющими соответствующий этап или соответствующую группу этапов. Хотя этапы представлены в пронумерованном виде, этапы описанных в настоящем документе способов можно выполнять в одно и то же время, в разное время, или в другом порядке, который логически возможен. Кроме того, элементы этих этапов можно использовать с элементами других этапов из других способов. Кроме того, этап может быть полностью или частично необязательным. Кроме того, любой из этапов любого из способов можно выполнять с помощью модулей, блоков, схем или других средств системы для выполнения этих этапов.

Как будет понятно специалистам в данной области техники после прочтения настоящего раскрытия, каждый из отдельных вариантов реализации, описанных и проиллюстрированных в настоящем документе, имеет отдельные компоненты и признаки, которые можно легко отделить или объединить с признаками любого из других нескольких вариантов реализации, не отступая от объема и сущности настоящего изобретения.

Приведенное выше описание примерных вариантов реализации настоящего изобретения представлено в целях иллюстрации и описания и изложено, чтобы предоставить обычным специалистам в данной области техники полное раскрытие и описание того, как получить и применить варианты реализации настоящего изобретения. Не предусмотрено, что настоящее описание является исчерпывающим или ограничивает изобретение точной описанной формой, также не подразумевается, что эксперименты представляют собой все возможные или единственные выполненные эксперименты. Несмотря на то, что настоящее изобретение было подробно описано посредством иллюстрации и примера в целях ясности понимания, обычные специалисты в данной области техники легко поймут в свете принципиальных положений настоящего

изобретения, что в него могут быть внесены определенные изменения и модификации, не отступая от сущности и объема прилагаемой формулы изобретения.

Соответственно, описание выше лишь иллюстрирует принципы настоящего изобретения. Следует понимать, что специалисты в данной области техники смогут разработать различные варианты, которые, хотя и не описаны или не показаны в настоящем документе явным образом, воплощают принципы настоящего изобретения и включены в его сущность и объем. Кроме того, все приведенные в настоящем документе примеры и условные формулировки в основном предназначены для того, чтобы помочь читателю понять принципы настоящего изобретения, не ограничиваясь такими конкретно перечисленными примерами и условиями. Кроме того, все формулировки в настоящем документе, перечисляющие принципы, аспекты и варианты реализации настоящего изобретения, а также его конкретные примеры, предназначены для охвата как их структурных, так и функциональных эквивалентов. Кроме того, подразумевается, что такие эквиваленты включают как известные в настоящее время эквиваленты, так и эквиваленты, разработанные в будущем, т.е. любые разработанные элементы, выполняющие ту же функцию, независимо от структуры. Таким образом, не подразумевается, что объем настоящего изобретения ограничивается примерными вариантами реализации, показанными и описанными в настоящем документе. Предпочтительно объем и сущность настоящего изобретения реализованы в прилагаемой формуле изобретения.

Термин, обозначающий элемент в единственном или множественном числе (соотв. «a», «an» и «the» в исходном тексте на английском языке) служит для обозначения «одного или большего количества», если противоположное не указано явным образом. При использовании «или» имеется ввиду «включающее или», а не «исключающее или», если противоположное не указано явным образом. Ссылка на «первый» компонент не обязательно требует обеспечения второго компонента. Более того, ссылка на «первый» или «второй» компонент не ограничивает упомянутый компонент конкретным местоположением, если явно не указано иное. Под термином «основанный на» имеется ввиду «основанный по меньшей мере частично на».

Формула изобретения может быть составлена так, чтобы исключить любой элемент, который может быть необязательным. Таким образом, это утверждение предназначено для использования в качестве предшествующей основы для использования такой исключающей терминологии как «исключительно», «только» и т.п. в отношении перечисления элементов формулы изобретения или использования «отрицательного» ограничения.

Все патенты, заявки на патенты, публикации и описания, упомянутые в настоящем документе, настоящим полностью включены посредством ссылки для всех целей, как если бы каждая отдельная публикация или патент были конкретно и по отдельности указаны для включения посредством ссылки и включены в настоящий документ посредством ссылки для раскрытия и описания методов и/или материалов, в связи с которыми цитируются публикации. Ни один из них не рассматривается в качестве предшествующего уровня техники.

ПЕРВОНАЧАЛЬНАЯ ФОРМУЛА ИЗОБРЕТЕНИЯ, ПРЕДСТАВЛЕННАЯ ЗАЯВИТЕЛЕМ К РАССМОТРЕНИЮ

1. Способ анализа биологического образца, полученного от субъекта женского пола, беременного плодом, где указанный субъект женского пола имеет первый гаплотип и второй гаплотип в первой хромосомной области, указанный биологический образец включает множество молекул внеклеточной ДНК плода и субъекта женского пола, причем указанный способ включает:

получение ридов, соответствующих указанному множеству молекул внеклеточной ДНК;

измерение размеров указанного множества молекул внеклеточной ДНК;

идентификацию первого набора молекул внеклеточной ДНК из указанного множества молекул внеклеточной ДНК, как имеющего размеры, превышающие или равные значению отсечки;

определение последовательности первого гаплотипа и последовательности второго гаплотипа по ридам, соответствующим указанному первому набору молекул внеклеточной ДНК;

выравнивание второго набора молекул внеклеточной ДНК из указанного множества молекул внеклеточной ДНК с указанной последовательностью первого гаплотипа, причем указанный второй набор молекул внеклеточной ДНК имеет размеры, которые меньше значения отсечки;

выравнивание третьего набора молекул внеклеточной ДНК из указанного множества молекул внеклеточной ДНК с указанной последовательностью второго гаплотипа, причем указанный третий набор молекул внеклеточной ДНК имеет размеры, которые меньше значения отсечки;

измерение первого значения параметра с использованием указанного второго набора молекул внеклеточной ДНК;

измерение второго значения параметра с использованием указанного третьего набора молекул внеклеточной ДНК;

сравнение указанного первого значения с указанным вторым значением; и

определение вероятности наследования плодом указанного первого гаплотипа на основе указанного сравнения первого значения со вторым значением.

2. Способ по п. 1, отличающийся тем, что указанное значение отсечки представляет собой 600 нт.

3. Способ по п. 1, отличающийся тем, что указанное значение отсечки представляет собой 1 тыс. нт.

4. Способ по любому из пп. 1-3, отличающийся тем, что определение указанной последовательности первого гаплотипа и указанной последовательности второго гаплотипа из ридов, соответствующих указанному первому набору молекул внеклеточной ДНК, включает:

выравнивание ридов, соответствующих указанному первому набору молекул внеклеточной ДНК, с референсным геномом.

5. Способ по п. 1, отличающийся тем, что определение указанной последовательности первого гаплотипа и указанной последовательности второго гаплотипа из ридов, соответствующих указанному первому набору молекул внеклеточной ДНК, включает:

выравнивание первого поднабора ридов со вторым поднабором ридов для идентификации другого аллеля в локусе в указанных ридах,

определение того, что указанный первый поднабор ридов имеет первый аллель в указанном локусе,

определение того, что указанный второй поднабор ридов имеет второй аллель в указанном локусе,

определение того, что указанный первый поднабор ридов соответствует первому гаплотипу, и

определение того, что указанный второй поднабор ридов соответствует второму гаплотипу.

6. Способ по любому из пп. 1-5, отличающийся тем, что указанный параметр представляет собой число молекул внеклеточной ДНК, размерный профиль молекул внеклеточной ДНК или уровень метилирования молекул внеклеточной ДНК.

7. Способ по п. 6, отличающийся тем, что:

указанный параметр представляет собой число молекул внеклеточной ДНК, и

указанный способ дополнительно включает:

определение того, что плод имеет более высокую вероятность наследования указанного первого гаплотипа, чем указанного второго гаплотипа, когда указанное первое значение больше указанного второго значения.

8. Способ по п. 6, отличающийся тем, что:

указанный параметр представляет собой размерный профиль молекул внеклеточной ДНК, и

указанный способ дополнительно включает:

определение того, что плод имеет более высокую вероятность наследования указанного первого гаплотипа, чем указанного второго гаплотипа, когда указанное первое

значение меньше указанного второго значения, что указывает на то, что указанный второй набор молекул внеклеточной ДНК характеризуется меньшим размерным профилем, чем указанный третий набор молекул внеклеточной ДНК.

9. Способ по п. 6, отличающийся тем, что:

указанный параметр представляет собой уровень метилирования молекул внеклеточной ДНК, и

указанный способ дополнительно включает:

определение того, что плод имеет более высокую вероятность наследования указанного первого гаплотипа, чем указанного второго гаплотипа, когда указанное первое значение меньше указанного второго значения.

10. Способ по любому из пп. 1-9, дополнительно включающий:

вычисление степени разделения с использованием указанного первого значения и указанного второго значения;

сравнение указанной степени разделения со значением отсечки; и

определение вероятности анеуплоидии плода на основании указанного сравнения степени разделения со значением отсечки.

11. Способ по п. 10, отличающийся тем, что:

указанное значение отсечки определяют по референсным образцам от беременных субъектов женского пола с эуплоидными плодами,

указанное значение отсечки определяют по референсным образцам от беременных субъектов женского пола с анеуплоидными плодами, или

указанное значение отсечки рассчитывают, предполагая анеуплоидный плод.

12. Способ по любому из пп. 1-11, дополнительно включающий:

идентификацию количества повторов подпоследовательности в риде из ридов, соответствующих указанному первому набору молекул внеклеточной ДНК,

причем:

определение последовательности первого гаплотипа включает определение того, что последовательность первого гаплотипа содержит некоторое количество повторов подпоследовательности.

13. Способ по п. 12, отличающийся тем, что:

указанные повторы подпоследовательности ассоциированы с заболеванием, ассоциированным с повторами, и

указанный способ дополнительно включает определение вероятности наследования плодом указанного заболевания, ассоциированного с повторами.

14. Способ анализа биологического образца, полученного от субъекта женского

пола, беременного плодом, отличающийся тем, что указанный биологический образец включает множество молекул внеклеточной ДНК плода и субъекта женского пола, причем указанный способ включает:

получение ридов последовательности, соответствующих указанному множеству молекул внеклеточной ДНК;

измерение размеров указанного множества молекул внеклеточной ДНК;

идентификацию набора молекул внеклеточной ДНК из указанного множества молекул внеклеточной ДНК, имеющих размеры, превышающие или равные значению отсечки; и

для молекулы внеклеточной ДНК из указанного набора молекул внеклеточной ДНК:

определение статуса метилирования в каждом сайте из множества сайтов,

определение профиля метилирования, причем:

указанный профиль метилирования указывает на статус метилирования в каждом сайте из указанного множества сайтов с использованием одного или более ридов последовательности, соответствующих указанной молекуле внеклеточной ДНК,

сравнение указанного профиля метилирования с одним или более референсными профилями, причем каждый из указанного одного или более референсных профилей определяют для конкретного типа ткани; и

определение ткани происхождения указанной молекулы внеклеточной ДНК с использованием указанного профиля метилирования.

15. Способ по п. 14, отличающийся тем, что указанное значение отсечки представляет собой 600 нт.

16. Способ по п. 14, отличающийся тем, что указанное значение отсечки представляет собой 1 тыс. нт.

17. Способ по любому из пп. 14-16, дополнительно включающий определение ткани происхождения для каждой молекулы внеклеточной ДНК из указанного набора молекул внеклеточной ДНК путем:

определения статуса метилирования в каждом сайте из множества соответствующих сайтов, причем указанное множество соответствующих сайтов соответствует указанной молекуле внеклеточной ДНК,

определение профиля метилирования, и

сравнение указанного профиля метилирования по меньшей мере с одним референсным профилем из одного или более референсных профилей.

18. Способ по п. 17, дополнительно включающий:

определение количества молекул внеклеточной ДНК, соответствующих каждой ткани происхождения, и

определение относительного вклада указанной ткани происхождения в биологическом образце с использованием указанного количества молекул внеклеточной ДНК, соответствующих каждой ткани происхождения.

19. Способ по любому из пп. 14-18, отличающийся тем, что измерение размеров указанного множества молекул внеклеточной ДНК включает:

выравнивание указанных ридов последовательности с референсным геномом.

20. Способ по любому из пп. 14-18, отличающийся тем, что измерение размеров указанного множества молекул внеклеточной ДНК включает:

полноразмерное секвенирование указанного множества молекул внеклеточной ДНК, и

подсчет количества нуклеотидов в каждой молекуле внеклеточной ДНК из указанного множества молекул внеклеточной ДНК.

21. Способ по пп. 14-17, отличающийся тем, что измерение размеров указанного множества молекул внеклеточной ДНК включает:

физическое отделение указанного множества молекул внеклеточной ДНК из биологического образца от других молекул внеклеточной ДНК в биологическом образце, причем указанные другие молекулы внеклеточной ДНК имеют размеры, которые меньше указанного значения отсечки.

22. Способ по любому из пп. 14-21, отличающийся тем, что референсный профиль из одного или более референсных профилей определяют путем:

измерения плотности метилирования в каждом референсном сайте из множества референсных сайтов с использованием молекул ДНК из референсной ткани,

сравнения указанной плотности метилирования в каждом референсном сайте из указанного множества референсных сайтов с одной или более пороговыми плотностями метилирования, и

идентификации каждого референсного сайта из указанного множества референсных сайтов как метилированного, неметилированного или неинформативного на основании сравнения указанной плотности метилирования с одной или более пороговыми плотностями метилирования, причем указанное множество сайтов представляет собой множество референсных сайтов, которые идентифицированы как метилированные или неметилированные.

23. Способ по любому из пп. 14-22, отличающийся тем, что указанная ткань происхождения представляет собой плаценту.

24. Способ по любому из пп. 14-22, отличающийся тем, что указанная ткань происхождения является фетальной или материнской.

25. Способ по п. 24, отличающийся тем, что:

указанная ткань происхождения является фетальной,

причем указанный способ дополнительно включает:

выравнивание рида последовательности из ридов последовательности с первой областью референсного генома, причем указанная первая область содержит множество сайтов, соответствующих аллелям, при этом указанное множество сайтов включает пороговое количество сайтов,

определение первого гаплотипа с использованием соответствующего аллеля, присутствующего в каждом сайте из указанного множества сайтов,

сравнение указанного первого гаплотипа со вторым гаплотипом, соответствующим субъекту мужского пола, и

определение классификации вероятности того, что указанный субъект мужского пола является отцом плода, с использованием указанного сравнения.

26. Способ по п. 24, отличающийся тем, что:

указанная ткань происхождения является фетальной,

причем указанный способ дополнительно включает:

выравнивание рида последовательности из ридов последовательности с первой областью референсного генома, причем указанная первая область содержит первое множество сайтов, соответствующих аллелям, при этом указанное множество сайтов включает пороговое количество сайтов,

сравнение указанного аллеля в каждом сайте из указанного множества сайтов с аллелем в соответствующем сайте в геноме субъекта мужского пола, и

определение классификации вероятности того, что указанный субъект мужского пола является отцом плода, с использованием указанного сравнения.

27. Способ по п. 24, дополнительно включающий:

для каждой молекулы внеклеточной ДНК из указанного набора молекул внеклеточной ДНК:

выравнивание рида последовательности, соответствующего указанной молекуле внеклеточной ДНК, с референсным геномом,

идентификацию указанного рида последовательности как соответствующего гаплотипу, присутствующему у субъекта женского пола,

определение ткани происхождения как фетальной с использованием указанного профиля метилирования, и

определение указанного гаплотипа как унаследованного по материнской линии гаплотипа плода.

28. Способ по п. 27, дополнительно включающий:

идентификацию гаплотипа как несущего генетическую мутацию или вариацию, вызывающую заболевание, и

классификацию того, что плод, вероятно, имеет указанное заболевание, вызванное генетической мутацией или вариацией.

29. Способ по п. 28, отличающийся тем, что идентификация указанного гаплотипа как несущего генетическую мутацию, вызывающую заболевание, включает:

идентификацию указанной генетической мутации или вариации в первом риде последовательности,

измерение первого уровня метилирования во втором риде последовательности, соответствующем первому геномному положению в пределах первого расстояния указанного первого рида последовательности, и

измерение второго уровня метилирования в третьем риде последовательности, соответствующем второму геномному положению в пределах второго расстояния указанного первого рида последовательности, причем:

указанный первый уровень метилирования и указанный второй уровень метилирования связаны с указанной генетической мутацией.

30. Способ по п. 24, дополнительно включающий:

для каждой молекулы внеклеточной ДНК из указанного набора молекул внеклеточной ДНК:

выравнивание рида последовательности, соответствующего указанной молекуле внеклеточной ДНК, с референсным геномом,

идентификацию указанного рида последовательности как соответствующего области, причем указанная область определяется путем:

получения множества ридов последовательности плода, соответствующих множеству молекул ДНК плода из ткани плода,

получение множества ридов последовательности матери, соответствующих множеству молекул ДНК матери,

определения статуса метилирования плода в каждом сайте метилирования из множества сайтов метилирования в пределах указанной области для каждого рида последовательности плода из указанного множества ридов последовательности плода,

определения статуса метилирования матери в каждом сайте метилирования из множества сайтов метилирования для каждого рида последовательности матери из

указанного множества ридов последовательности матери,

определения значения параметра, характеризующего количество сайтов, в которых указанный статус метилирования плода отличается от указанного статуса метилирования матери,

сравнения указанного значения параметра с пороговым значением; и

определения того, что указанное значение параметра превышает указанное пороговое значение.

31. Способ по любому из пп. 14-28, отличающийся тем, что указанное значение отсечки представляет собой по меньшей мере 500 нт.

32. Способ по любому из пп. 14-31, отличающийся тем, что определение ткани происхождения молекулы внеклеточной ДНК включает ввод профиля метилирования в модель машинного обучения, причем указанная модель обучена путем:

получения множества обучающих профилей метилирования, при этом каждый обучающий профиль метилирования имеет статус метилирования в одном или более сайтах из множества сайтов, при этом каждый обучающий профиль метилирования определен из молекулы ДНК из известной ткани,

сохранения множества обучающих образцов, при этом каждый обучающий образец включает один из множества обучающих профилей метилирования и метку, указывающую на известную ткань, соответствующую указанному обучающему профилю метилирования, и

оптимизации с использованием указанного множества обучающих образцов параметров указанной модели на основании выходных данных модели, совпадающих или не совпадающих с соответствующими метками, когда в модель вводится указанное множество обучающих профилей метилирования, при этом выходной показатель модели определяет ткань, соответствующую введенному профилю метилирования.

33. Способ по п. 32, отличающийся тем, что указанная модель машинного обучения включает сверточные нейронные сети (CNN), линейную регрессию, логистическую регрессию, глубокую рекуррентную нейронную сеть, Байесов классификатор, скрытую Марковскую модель (HMM), линейный дискриминантный анализ (LDA), кластеризацию k-средних, плотностный алгоритм кластеризации пространственных данных с присутствием шума (DBSCAN), алгоритмом случайного леса или метод опорных векторов (SVM).

34. Способ по п. 32, отличающийся тем, что каждая молекула ДНК из известной ткани представляет собой клеточную ДНК.

35. Способ по п. 32 или 34, отличающийся тем, что указанные параметры модели

включают первый параметр, указывающий, имеет ли один сайт из множества сайтов тот же статус метилирования, что и другой сайт из множества сайтов.

36. Способ по любому из пп. 32-35, отличающийся тем, что указанные параметры модели содержат второй параметр, указывающий на расстояние между сайтами из множества сайтов.

37. Способ по любому из пп. 14-31, отличающийся тем, что референсный профиль из одного или более референсных профилей соответствует референсной ткани,

причем указанный способ дополнительно включает определение ткани происхождения как референсной ткани, когда указанный профиль метилирования соответствует указанному референсному профилю.

38. Способ по любому из пп. 14-37, отличающийся тем, что указанное множество сайтов содержит по меньшей мере 5 сайтов CpG.

39. Способ по любому из пп. 14-31, отличающийся тем, что определение указанной ткани происхождения с использованием указанного профиля метилирования включает:

определение оценки сходства путем сравнения указанного профиля метилирования с первым референсным профилем метилирования из первой референсной ткани из множества референсных тканей;

сравнение указанной оценки сходства с пороговым значением; и

определение указанной ткани происхождения как первой референсной ткани, когда указанная оценка сходства превышает указанное пороговое значение.

40. Способ по п. 39, отличающийся тем, что:

указанная оценка сходства представляет собой первую оценку сходства,

причем указанный способ дополнительно включает:

расчет указанного порогового значения путем:

определения второй оценки сходства путем сравнения указанного профиля метилирования со вторым референсным профилем метилирования из второй референсной ткани из множества референсных тканей, при этом указанная первая референсная ткань и указанная вторая референсная ткань представляют собой разные ткани, при этом указанное пороговое значение представляет собой указанную вторую оценку сходства.

41. Способ по п. 39 или 40, отличающийся тем, что:

указанный первый референсный профиль метилирования содержит первый поднабор сайтов, имеющих по меньшей мере первую вероятность метилирования для первой референсной ткани,

указанный первый референсный профиль метилирования содержит второй поднабор сайтов, имеющих не более чем вторую вероятность метилирования для первой

референсной ткани, и

определение оценки сходства включает:

увеличение указанной оценки сходства, когда сайт из указанного множества сайтов метилирован и сайт из указанного множества сайтов находится в указанном первом поднаборе сайтов, и

уменьшение указанной оценки сходства, когда сайт из указанного множества сайтов метилирован и сайт из указанного множества сайтов находится в указанном втором поднаборе сайтов.

42. Способ по п. 39 или 40, отличающийся тем, что:

указанный первый референсный профиль метилирования содержит множество сайтов, при этом каждый сайт из указанного множества сайтов характеризуется вероятностью метилирования и вероятностью неметилирования для первой референсной ткани,

указанная оценка сходства определяется путем:

для каждого сайта из указанного множества сайтов:

определения указанной вероятности в референсной ткани, соответствующей указанному статусу метилирования сайта в молекуле внеклеточной ДНК,

вычисления произведения указанного множества вероятностей, при этом указанное произведение представляет собой указанную оценку сходства.

43. Способ по п. 42, отличающийся тем, что указанная вероятность определяется с использованием бета-распределения.

44. Способ по любому из пп. 14-43, дополнительно включающий:

секвенирование указанного множества молекул внеклеточной ДНК для получения ридов последовательности, и

определение статуса метилирования сайта путем измерения характеристики, соответствующей нуклеотиду указанного сайта и нуклеотидов, соседних с указанным сайтом.

45. Способ по любому из пп. 14-44, отличающийся тем, что размеры указанного множества молекул внеклеточной ДНК включают ряд сайтов CpG.

46. Способ по любому из пп. 14-45, отличающийся тем, что по меньшей мере один сайт из указанного множества сайтов метилирован.

47. Способ по любому из пп. 14-46, отличающийся тем, что два сайта из указанного множества сайтов разделены по меньшей мере 160 нт.

48. Способ анализа биологического образца, полученного от субъекта женского пола, беременного плодом, где указанный биологический образец включает молекулы

внеклеточной ДНК плода и субъекта женского пола, причем указанный способ включает:

получение первого рида последовательности, соответствующего молекуле внеклеточной ДНК из указанных молекул внеклеточной ДНК;

выравнивание указанного первого рида последовательности с областью референсного генома, при этом известно, что указанная область потенциально включает повторы подпоследовательности;

идентификацию количества повторов указанной подпоследовательности в указанном первом риде последовательности, соответствующем указанной молекуле внеклеточной ДНК;

сравнение указанного количества повторов подпоследовательности с пороговым количеством; и

определение классификации вероятности наличия у плода генетического нарушения с использованием указанного сравнения количества повторов с пороговым количеством.

49. Способ по п. 48, отличающийся тем, что определение указанной классификации вероятности наличия у плода генетического нарушения включает:

определение того, что указанный плод, вероятно, имеет генетическое нарушение, когда указанное количество повторов превышает указанное пороговое количество.

50. Способ по п. 48 или 49, отличающийся тем, что указанное пороговое количество представляет собой 55 или больше.

51. Способ по любому из пп. 48-50, отличающийся тем, что указанное генетическое нарушение представляет собой синдром ломкой Х-хромосомы.

52. Способ по любому из пп. 48-51, отличающийся тем, что указанная подпоследовательность представляет собой тринуклеотидную последовательность.

53. Способ по любому из пп. 48-52, отличающийся тем, что указанные молекулы внеклеточной ДНК имеют длину, превышающую значение отсечки.

54. Способ по п. 53, отличающийся тем, что указанное значение отсечки представляет собой 600 нт.

55. Способ по п. 53, отличающийся тем, что указанное значение отсечки представляет собой 1 тыс. нт.

56. Способ по любому из пп. 48-54, дополнительно включающий определение того, что указанная молекула внеклеточной ДНК происходит от плода.

57. Способ по п. 56, отличающийся тем, что:

указанное количество повторов подпоследовательности в первом риде последовательности представляет собой первое количество повторов

подпоследовательности,

определение того, что указанная молекула внеклеточной ДНК происходит от плода, включает:

получение второго рида последовательности, соответствующего молекуле внеклеточной ДНК материнского происхождения, полученной из лейкоцитарной пленки или образца субъекта женского пола до беременности,

выравнивание указанного второго рида последовательности с областью референсного генома,

идентификацию второго количества повторов подпоследовательности в указанном втором риде последовательности, и

определение того, что указанное второе количество повторов меньше, чем указанное первое количество повторов.

58. Способ по п. 56, отличающийся тем, что:

определение того, что молекула внеклеточной ДНК происходит от плода, включает:

определение уровня метилирования указанной молекулы внеклеточной ДНК с использованием метилированных и неметилированных сайтов указанной молекулы внеклеточной ДНК, и

сравнение указанного уровня метилирования с референсным уровнем.

59. Способ по п. 58, дополнительно включающий определение того, что указанный уровень метилирования превышает указанный референсный уровень.

60. Способ по п. 56, отличающийся тем, что:

определение того, что молекула внеклеточной ДНК происходит от плода, включает:

определение профиля метилирования множества сайтов внеклеточной молекулы,

определение оценки сходства путем сравнения указанного профиля метилирования с референсным профилем из ткани матери или плода, и

сравнение указанной оценки сходства с одним или более пороговыми значениями.

61. Способ по п. 48, дополнительно включающий:

получение множества ридов последовательности, соответствующих молекулам внеклеточной ДНК,

выравнивание указанного множества ридов последовательности с множеством областей референсного генома, причем известно, что указанное множество областей потенциально включают повторы подпоследовательностей,

идентификацию количеств повторов подпоследовательностей в указанном

множестве ридов последовательностей;

сравнение указанных количеств повторов подпоследовательностей с множеством пороговых количеств; и

для каждого из множества генетических нарушений определение классификации вероятности наличия у плода соответствующего генетического нарушения с использованием указанного сравнения с пороговым количеством из указанного множества пороговых количеств.

62. Способ анализа биологического образца, полученного от субъекта женского пола, беременного плодом, отличающийся тем, что указанный биологический образец включает молекулы внеклеточной ДНК от плода и субъекта женского пола, причем указанный способ включает:

получение первого рида последовательности, соответствующего молекуле внеклеточной ДНК из указанных молекул внеклеточной ДНК;

выравнивание указанного первого рида последовательности с первой областью референсного генома.

идентификацию первого количества повторов первой подпоследовательности в указанном первом риде последовательности, соответствующем указанной молекуле внеклеточной ДНК;

анализ данных о последовательности, полученных от субъекта мужского пола, для определения того, присутствует ли второе количество повторов указанной первой подпоследовательности в первой области; и

определение классификации вероятности того, что указанный субъект мужского пола является отцом плода с использованием определения наличия указанного второго количества повторов указанной первой подпоследовательности.

63. Способ по п. 62, дополнительно включающий:

определение того, что указанная молекула внеклеточной ДНК происходит от плода.

64. Способ по п. 62 или 63, отличающийся тем, что указанная первая подпоследовательность содержит аллель.

65. Способ по любому из пп. 62-64, отличающийся тем, что:

указанная классификация заключается в том, что субъект мужского пола, вероятно, является отцом, если установлено наличие указанного второго количества повторов указанной первой подпоследовательности, или

указанная классификация заключается в том, что субъект мужского пола, вероятно, не является отцом, если установлено отсутствие указанного второго количества повторов

указанной первой подпоследовательности.

66. Способ по любому из пп. 62-65, дополнительно включающий:

сравнение указанного первого количества повторов с указанным вторым количеством повторов,

причем:

определение классификации вероятности того, что субъект мужского пола является отцом, включает:

использование указанного сравнения указанного первого количества повторов с указанным вторым количеством повторов, и

классификация состоит в том, что субъект мужского пола, вероятно, является отцом, когда указанное первое количество повторов находится в пределах порогового значения указанного второго количества повторов.

67. Способ по любому из пп. 62-66, отличающийся тем, что:

указанная молекула внеклеточной ДНК представляет собой первую молекулу внеклеточной ДНК;

причем указанный способ дополнительно включает:

получение второго рида последовательности, соответствующего второй молекуле внеклеточной ДНК из молекул внеклеточной ДНК;

выравнивание указанного второго рида последовательности со второй областью референсного генома,

идентификацию первого количества повторов второй подпоследовательности в указанном втором риде последовательности, соответствующем указанной второй молекуле внеклеточной ДНК;

анализ данных о последовательности, полученных от субъекта мужского пола, для определения наличия второго количества повторов второй подпоследовательности в указанной второй области; и

причем:

определение классификации вероятности того, что субъект мужского пола является отцом плода, дополнительно включает использование определения наличия указанного второго количества повторов указанной второй подпоследовательности в указанной второй области.

68. Способ по любому из пп. 62-67, отличающийся тем, что указанные молекулы внеклеточной ДНК имеют размер, превышающий значение отсечки.

69. Способ по п. 68, отличающийся тем, что указанные молекулы внеклеточной ДНК имеют размер более 600 нт.

70. Способ по п. 68, отличающийся тем, что указанные молекулы внеклеточной ДНК имеют размер более 1 тыс. нт.

71. Способ анализа биологического образца, полученного от субъекта женского пола, беременного плодом, отличающийся тем, что указанный биологический образец включает множество молекул внеклеточной ДНК плода и субъекта женского пола, причем указанный способ включает:

измерение размеров указанного множества молекул внеклеточной ДНК;

измерение первого количества внеклеточных молекул ДНК, имеющих размеры, превышающие значение отсечки;

генерирование значения нормированного параметра с использованием указанного первого количества;

сравнение указанного значения нормированного параметра с одной или более калибровочными точками данных, при этом каждая калибровочная точка данных указывает гестационный возраст, соответствующий калибровочному значению указанного нормированного параметра, и при этом указанную одну или более калибровочных точек данных определяют из множества калибровочных образцов с известными гестационными возрастами, включающих молекулы внеклеточной ДНК, размеры которых превышают указанное значение отсечки; и

определение гестационного возраста с использованием указанного сравнения.

72. Способ по п. 71, дополнительно включающий:

определение референсного гестационного возраста плода с использованием УЗИ или даты последней менструации субъекта женского пола,

сравнение указанного гестационного возраста с указанным референсным гестационным возрастом, и

определение классификации вероятности нарушения, ассоциированного с беременностью, с использованием указанного сравнения гестационного возраста с референсным гестационным возрастом.

73. Способ по п. 71, дополнительно включающий:

определение первой подпоследовательности, соответствующей по меньшей мере одному концу молекул внеклеточной ДНК, имеющих размеры, превышающие значение отсечки,

причем:

первое количество представляет собой молекулы внеклеточной ДНК, имеющие размер, превышающий значение отсечки, и имеющие указанную первую подпоследовательность на одном или более концах соответствующей молекулы

внеклеточной ДНК.

74. Способ по п. 73, отличающийся тем, что указанная первая подпоследовательность представляет собой 1, 2, 3 или 4 нуклеотида.

75. Способ по п. 73 или 74, отличающийся тем, что генерирование указанного значения нормированного параметра включает:

(a) нормирование указанного первого количества к общему количеству молекул внеклеточной ДНК, имеющих размер, превышающий значение отсечки;

(b) нормирование указанного первого количества ко второму количеству молекул внеклеточной ДНК, имеющих размер, превышающий значение отсечки, и заканчивающихся на второй подпоследовательности, причем указанная вторая подпоследовательность отличается от указанной первой подпоследовательности, или

(c) нормирование указанного первого количества к третьему количеству молекул внеклеточной ДНК, имеющих размер меньше значения отсечки.

76. Способ по любому из пп. 71-75, дополнительно включающий получение ридов последовательности, соответствующих множеству молекул внеклеточной ДНК.

77. Способ анализа биологического образца, полученного от субъекта женского пола, беременного плодом, отличающийся тем, что указанный биологический образец включает множество молекул внеклеточной ДНК от плода и субъекта женского пола, причем указанный способ включает:

измерение размеров указанного множества молекул внеклеточной ДНК;

измерение первого количества молекул внеклеточной ДНК, имеющих размеры, превышающие значение отсечки;

генерирование первого значения нормированного параметра с использованием указанного первого количества;

получение второго значения, соответствующего ожидаемому значению нормированного параметра для здоровой беременности, при этом указанное второе значение зависит от гестационного возраста плода;

определение отклонения между указанным первым значением нормированного параметра и указанным вторым значением нормированного параметра; и

определение классификации вероятности нарушения, ассоциированного с беременностью, с использованием указанного отклонения.

78. Способ по п. 77, отличающийся тем, что получение указанного второго значения включает:

получение указанного второго значения из калибровочной таблицы, соотносящей измерения у беременных субъектов женского пола с калибровочными значениями

нормированного параметра, при этом указанную калибровочную таблицу создают путем:

получения первой таблицы, соотносящей гестационные возрасты с измерениями у беременных субъектов женского пола,

получения второй таблицы, соотносящей гестационные возрасты с калибровочными значениями указанного нормированного параметра, и

создания калибровочной таблицы, соотносящей измерения с указанными калибровочными значениями из указанной первой таблицы и указанной второй таблицы.

79. Способ по п. 78, отличающийся тем, что указанные измерения у беременных субъектов женского пола представляют собой время, прошедшее с момента последней менструации.

80. Способ по п. 78, отличающийся тем, что указанные измерения у беременных субъектов женского пола представляют собой характеристики изображений беременных субъектов женского пола.

81. Способ по п. 80, отличающийся тем, что указанные характеристики изображения включают длину, размер, внешний вид или анатомию плода субъекта женского пола.

82. Способ по любому из пп. 72-81, отличающийся тем, что указанное связанное с беременностью нарушение включает преэклампсию, задержку внутриутробного развития, инвазивную плацентацию, преждевременные роды, гемолитическую болезнь новорожденных, плацентарную недостаточность, водянку плода, порок развития плода, гемолиз, синдром повышенной активности печеночных ферментов и низкого уровня тромбоцитов (HELLP) или системную красную волчанку.

83. Способ по любому из пп. 71-82, отличающийся тем, что указанное значение отсечки представляет собой 600 нт. или более.

84. Способ по любому из пп. 71-82, отличающийся тем, что указанное значение отсечки представляет собой 1000 нт. или более.

85. Способ по любому из пп. 71-84, отличающийся тем, что указанное первое количество представляет собой число или частоту.

86. Способ по любому из пп. 71-85, отличающийся тем, что генерирование указанного значения нормированного параметра с использованием указанного первого количества включает:

измерение второго количества молекул внеклеточной ДНК, включая размеры, которые меньшие значения отсечки; и

расчет соотношения указанного первого количества и указанного второго количества.

87. Способ по п. 86, отличающийся тем, что:

указанное значение отсечки представляет собой первое значение отсечки,

второе значение отсечки меньше, чем указанное первое значение отсечки, и

указанное второе количество включает молекулы внеклеточной ДНК, размеры которых меньше указанного второго значения отсечки, или указанное второе количество включает все молекулы внеклеточной ДНК во множестве молекул внеклеточной ДНК.

88. Способ анализа биологического образца, полученного от субъекта женского пола, беременного плодом, отличающийся тем, что указанный биологический образец включает множество молекул внеклеточной ДНК от плода и субъекта женского пола, причем указанный способ включает:

измерение размеров указанного множества молекул внеклеточной ДНК;

идентификацию набора молекул внеклеточной ДНК, имеющих размеры, превышающие значение отсечки;

генерирование значения параметра концевого мотива с использованием первого количества, причем генерирование указанного значения параметра концевого мотива включает:

измерение указанного первого количества молекул внеклеточной ДНК в наборе, имеющих первую подпоследовательность на одном или более концах молекул внеклеточной ДНК в указанном наборе;

сравнение указанного значения параметра концевого мотива с пороговым значением; и

определение классификации вероятности нарушения, ассоциированного с беременностью, с использованием указанного сравнения.

89. Способ по п. 88, отличающийся тем, что указанный способ дополнительно включает:

измерение второго количества молекул внеклеточной ДНК, имеющих подпоследовательность, отличную от первой подпоследовательности, на одном или более концах молекул внеклеточной ДНК, и

причем:

генерирование указанного значения параметра концевого мотива включает использование соотношения указанного первого количества и указанного второго количества.

90. Способ по п. 88, отличающийся тем, что указанная первая подпоследовательность представляет собой 1, 2, 3 или 4 нуклеотида в длину.

91. Способ по п. 90, отличающийся тем, что указанная первая

подпоследовательность содержит последний нуклеотид на конце соответствующей молекулы внеклеточной ДНК.

92. Способ по п. 88, отличающийся тем, что:

указанное пороговое значение представляет собой первое пороговое значение, и

указанный параметр концевого мотива представляет собой первый параметр концевого мотива,

причем указанный способ дополнительно включает:

измерение второго количества молекул внеклеточной ДНК, имеющих вторую подпоследовательность, отличную от первой подпоследовательности, на одном или более концах молекул внеклеточной ДНК,

генерирование значения второго параметра концевого мотива с использованием указанного третьего количества, и

сравнение указанного значения второго параметра концевого мотива со вторым пороговым значением; и

причем:

в определении классификации вероятности нарушения, ассоциированного с беременностью, используется указанное сравнение значения второго параметра концевого мотива со вторым пороговым значением, при этом указанное нарушение, связанное с беременностью, является вероятным, когда указанное значение первого параметра концевого мотива превышает указанное первое пороговое значение, а указанное значение второго параметра концевого мотива превышает указанное второе пороговое значение.

93. Способ по п. 88, отличающийся тем, что указанное первое количество молекул внеклеточной ДНК включает молекулы внеклеточной ДНК, которые, как определено, происходят из ткани происхождения.

94. Способ по п. 88, отличающийся тем, что:

указанное пороговое значение представляет собой первое пороговое значение, и

указанный набор молекул внеклеточной ДНК представляет собой первый набор молекул внеклеточной ДНК,

причем указанный способ дополнительно включает:

идентификацию второго набора молекул внеклеточной ДНК, имеющих размеры в первом диапазоне размеров, причем указанный первый диапазон размеров включает размеры, превышающие указанное значение отсечки,

генерирование значения размерного параметра с использованием второго количества молекул внеклеточной ДНК в указанном втором наборе, и

сравнение указанного значения размерного параметра со вторым пороговым

значением,

при этом определение классификации вероятности нарушения, ассоциированного с беременностью, включает использование указанного сравнения значения размерного параметра со вторым пороговым значением.

95. Способ по любому из пп. 88-94, отличающийся тем, что указанное значение отсечки представляет собой 600 нт.

96. Способ по любому из пп. 88-94, отличающийся тем, что указанное значение отсечки представляет собой 1000 нт.

97. Способ анализа биологического образца беременного организма, отличающийся тем, что указанный биологический образец включает множество молекул внеклеточной нуклеиновой кислоты, причем указанный способ включает:

секвенирование указанного множества молекул внеклеточной нуклеиновой кислоты, при этом более 20% из указанного множества секвенированных молекул внеклеточной нуклеиновой кислоты имеют длины более 200 нт.

98. Способ по п. 97, отличающийся тем, что секвенирование осуществляют с помощью методики одномолекулярного секвенирования в реальном времени.

99. Способ по п. 97 или 98, отличающийся тем, что:

более 11% из указанного множества секвенированных молекул внеклеточной нуклеиновой кислоты имеют длины более 400 нт.,

более 10% из указанного множества секвенированных молекул внеклеточной нуклеиновой кислоты имеют длины более 500 нт.,

более 8% из указанного множества секвенированных молекул внеклеточной нуклеиновой кислоты имеют длины более 600 нт.,

более 6% из указанного множества секвенированных молекул внеклеточной нуклеиновой кислоты имеют длины более 1 тыс. нт.,

более 3% из указанного множества секвенированных молекул внеклеточной нуклеиновой кислоты имеют длины более 2 тыс. нт.,

более 1% из указанного множества секвенированных молекул внеклеточной нуклеиновой кислоты имеют длины более 3 тыс. нт.,

по меньшей мере 0,9% из указанного множества секвенированных молекул внеклеточной нуклеиновой кислоты имеют длины более 4 тыс. нт., или

по меньшей мере 0,04% из указанного множества секвенированных молекул внеклеточной нуклеиновой кислоты имеют длины более 10 тыс. нт.

100. Способ по любому из пп. 97-99, отличающийся тем, что указанное множество молекул внеклеточной нуклеиновой кислоты содержит по меньшей мере 100 молекул

внеклеточной нуклеиновой кислоты.

101. Способ по любому из пп. 97-100, отличающийся тем, что указанное множество молекул внеклеточной нуклеиновой кислоты происходит из множества различных геномных областей.

102. Способ по любому из пп. 97-101, отличающийся тем, что результатом указанного секвенирования являются риды, которые используют по любому из пп. 1-94.

103. Способ по любому из пп. 97-101, отличающийся тем, что результатом указанного секвенирования являются риды,

причем указанный способ дополнительно включает:

использование указанных ридов для определения анеуплоидии, аберрации, генетической мутации или вариации или наследования родительского гаплотипа плода.

104. Способ по любому из пп. 1-103, отличающийся тем, что:

указанное множество молекул внеклеточной ДНК обогащено размерами, превышающими или равными значению отсечки, по отношению к указанному биологическому образцу, причем более 20% из указанных молекул внеклеточной нуклеиновой кислоты в биологическом образце имеют размеры более 200 нт.

105. Способ по п. 104, дополнительно включающий:

обогащение указанного множества молекул внеклеточной ДНК с помощью электрофореза.

106. Способ по п. 104, дополнительно включающий:

обогащение указанного множества молекул внеклеточной ДНК с использованием магнитных гранул для селективного связывания молекул внеклеточной ДНК на основании размера.

107. Способ по п. 104, дополнительно включающий:

обогащение указанного множества молекул внеклеточной ДНК с помощью гибридизации, иммунопреципитации, амплификации или CRISPR.

108. Способ по любому из пп. 105-107, отличающийся тем, что обогащение осуществляется для размеров более 600 нт., 700 нт., 800 нт., 900 нт. или 1 тыс. нт.

109. Способ по любому из пп. 1-103, отличающийся тем, что указанное множество молекул внеклеточной ДНК обогащено по профилю метилирования по сравнению с указанным биологическим образцом,

причем указанный способ дополнительно включает:

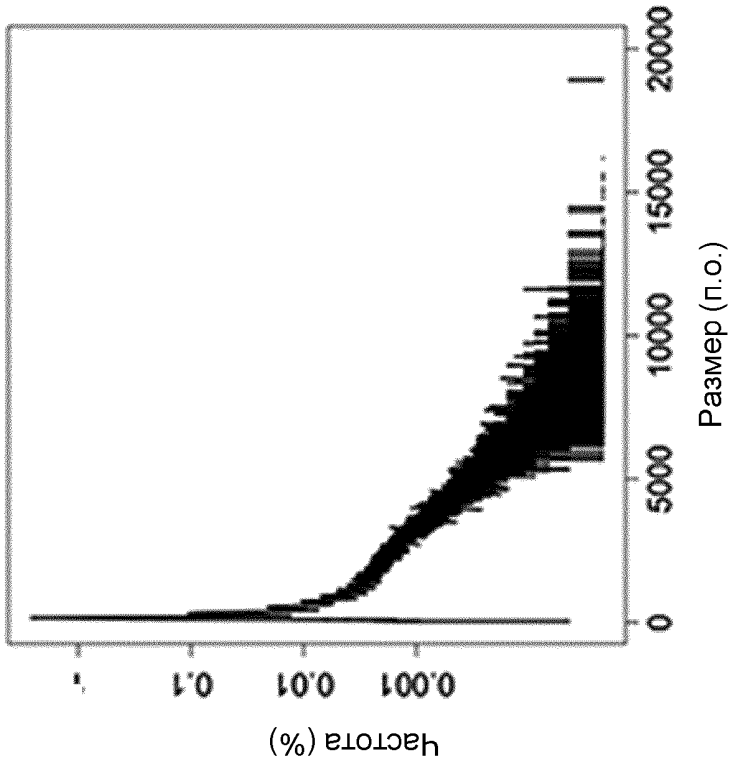
обогащение указанного множества молекул внеклеточной ДНК с использованием иммунопреципитации.

110. Компьютерный программный продукт, содержащий инструкции, которые при

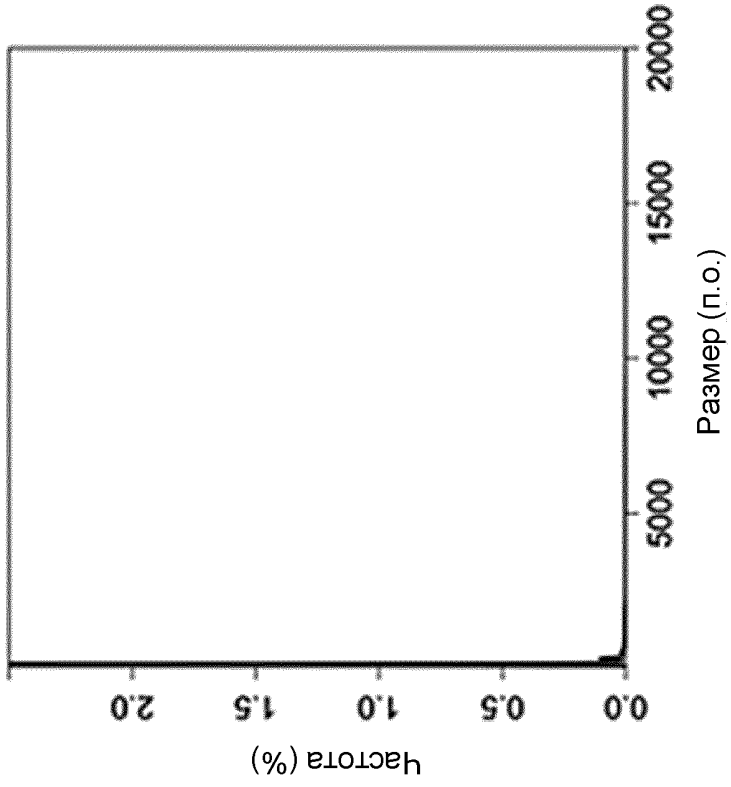
выполнении управляют вычислительной системой для выполнения способа по любому из предыдущих пунктов.

111. Машиночитаемый носитель данных, содержащий компьютерный программный продукт по п. 110.

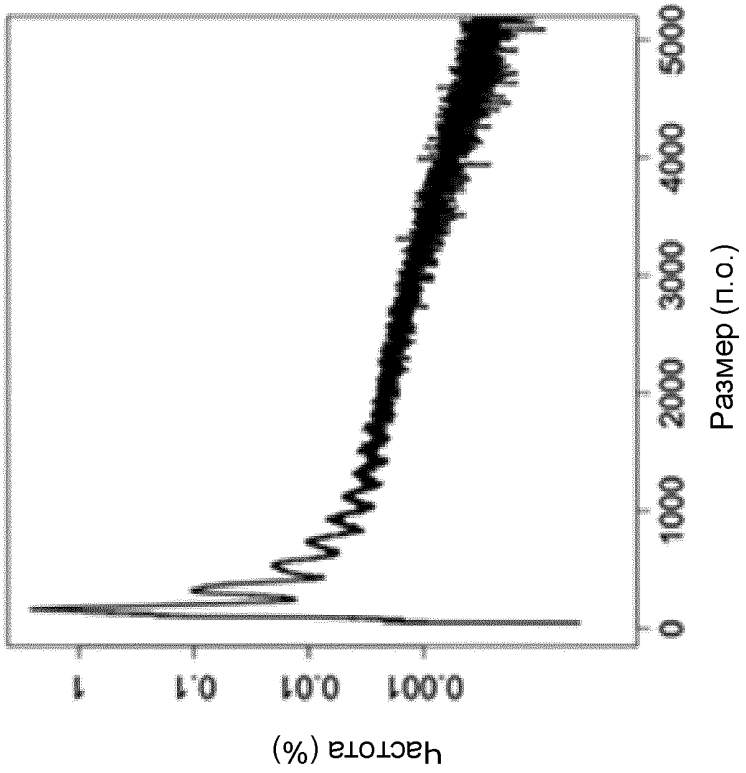
112. Вычислительная система, содержащая компьютерный программный продукт по п. 111.


Частота (%)
0,001
0,01
0,1
Размер (п.о.)
0
5000
10000
15000
20000

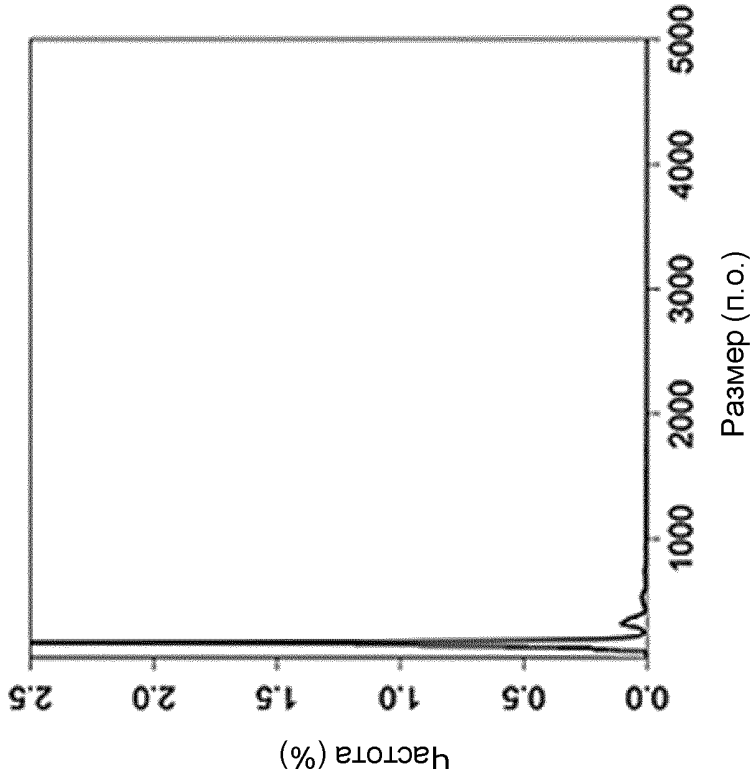
ФИГ. 1B


Частота (%)
0,0
0,5
1,0
1,5
2,0
Размер (п.о.)
5000
10000
15000
20000

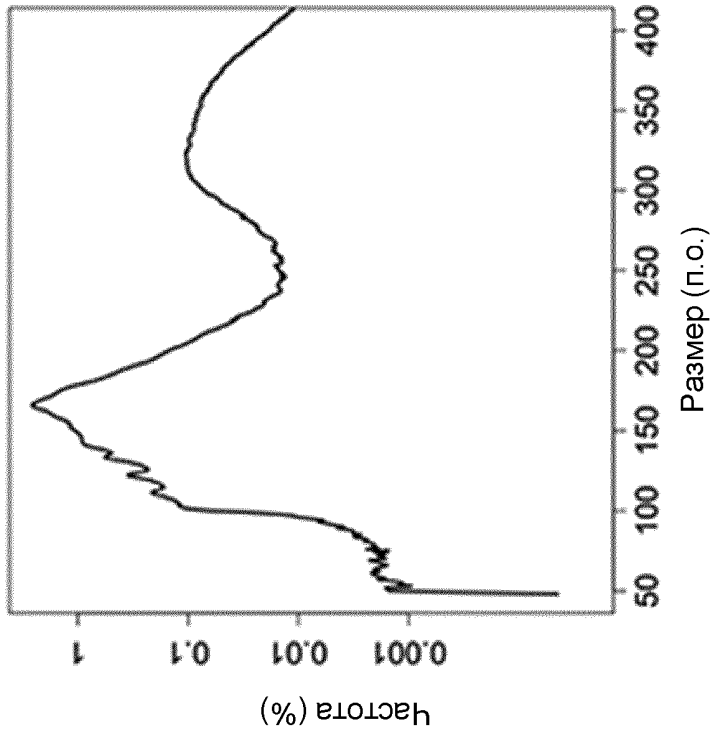
ФИГ. 1A


Частота (%)
0.001
0.01
0.1
1
Размер (п.о.)
0
1000
2000
3000
4000
5000

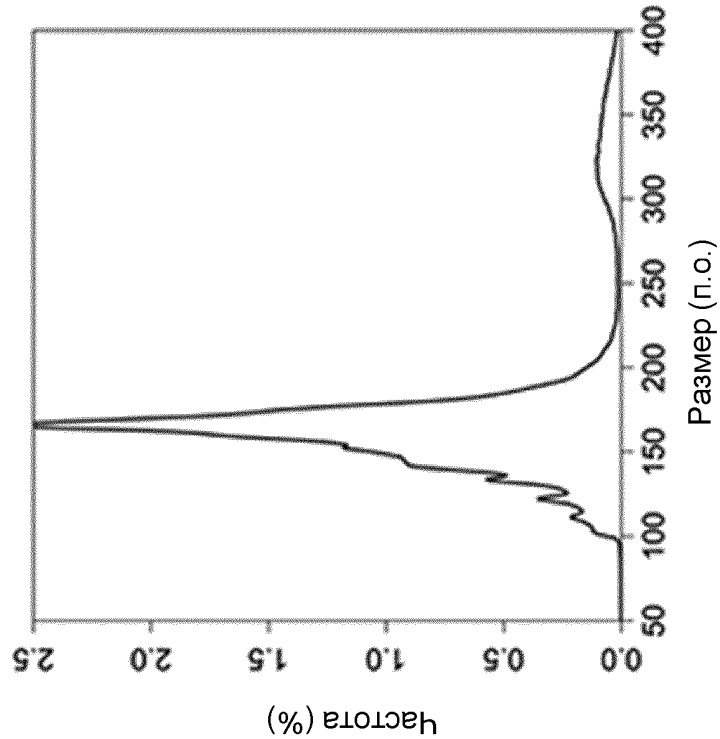
ФИГ. 2B


Частота (%)
0.0
0.5
1.0
1.5
2.0
2.5
Размер (п.о.)
1000
2000
3000
4000
5000

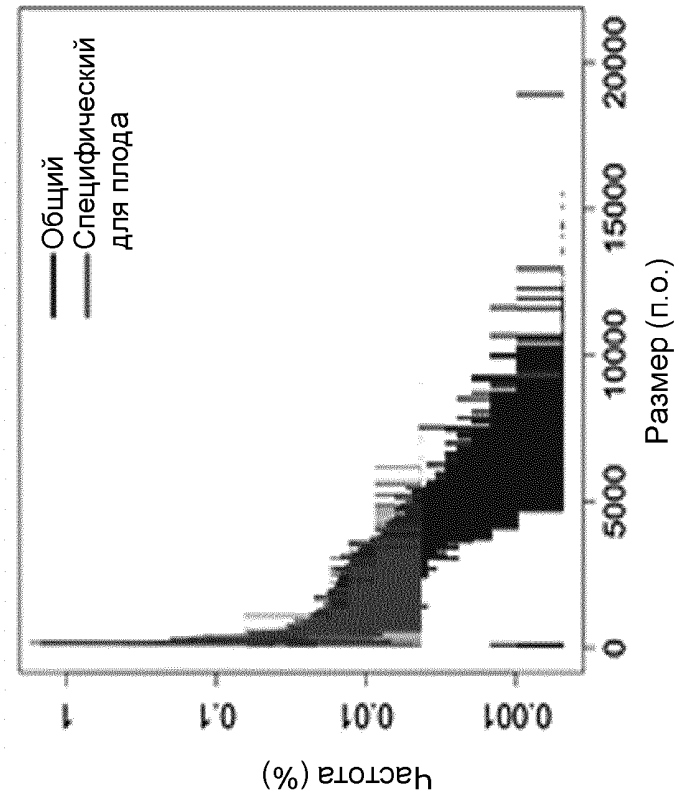
ФИГ. 2A


Частота (%)
0,001
0,01
0,1
1
Размер (п.о.)
50
100
150
200
250
300
350
400

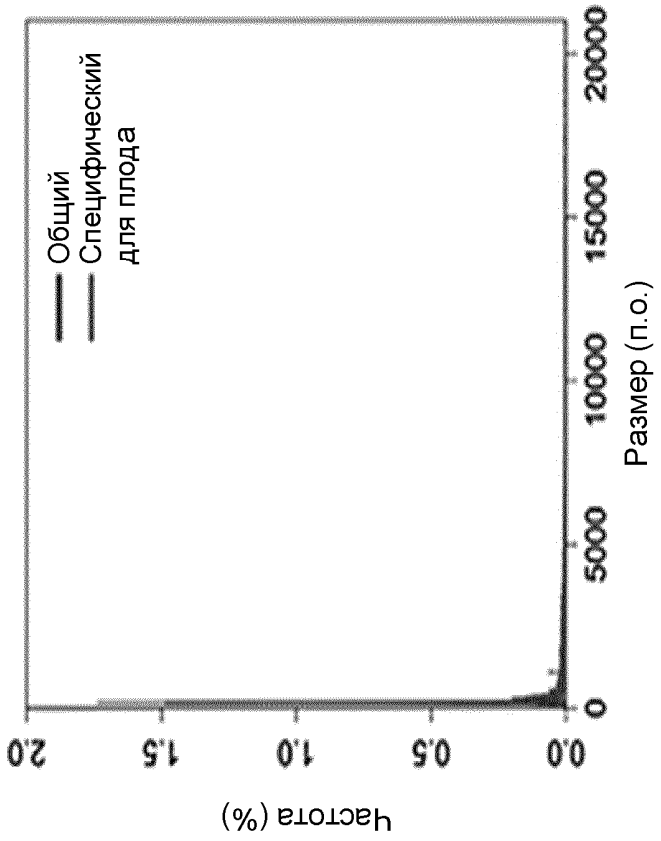
ФИГ. 3В


Частота (%)
0,0
0,5
1,0
1,5
2,0
2,5
Размер (п.о.)
50
100
150
200
250
300
350
400

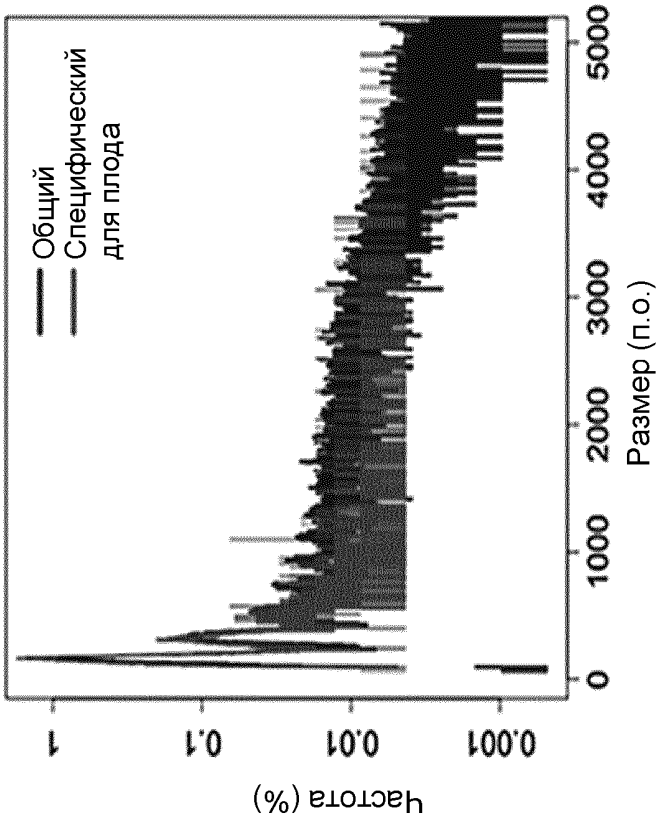
ФИГ. 3А


Общий
Специфический для плода
Частота (%)
0,001
0,01
0,1
1
Размер (п.о.)
0
5000
10000
15000
20000

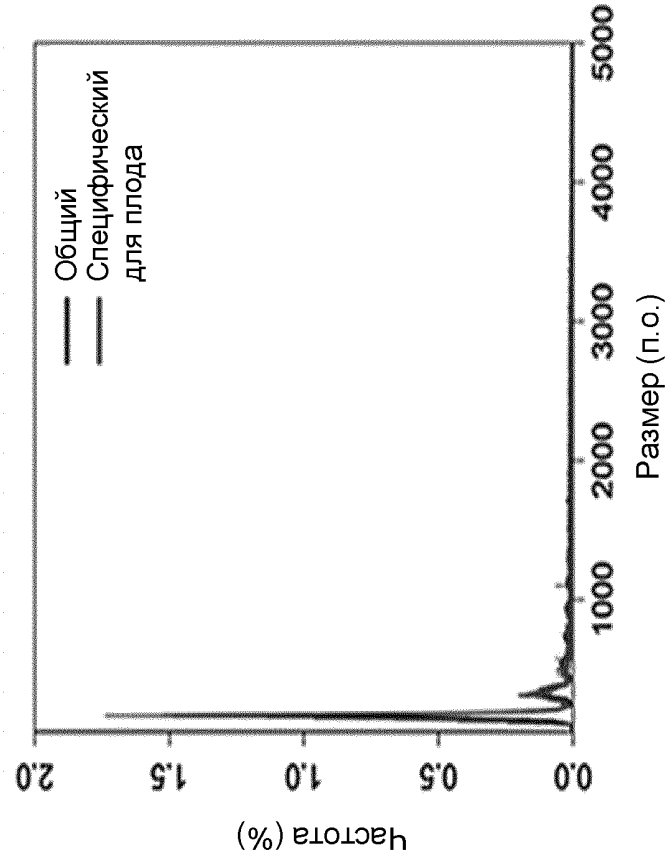
ФИГ. 4B


Общий
Специфический для плода
Частота (%)
0,0
0,5
1,0
1,5
2,0
Размер (п.о.)
0
5000
10000
15000
20000

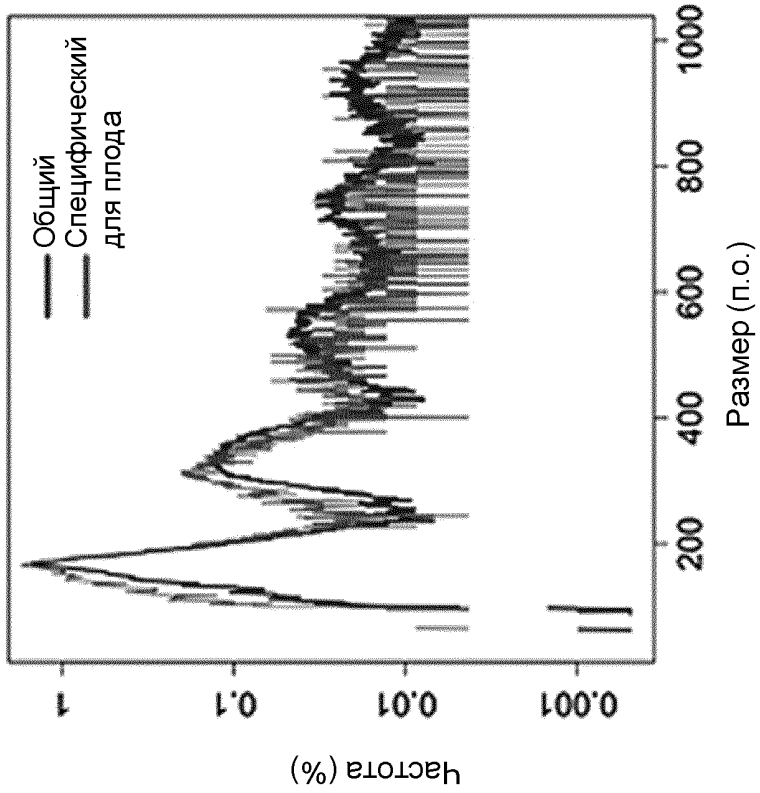
ФИГ. 4A


Общий
Специфический для плода
Частота (%)
0,001
0,01
0,1
1
Размер (п.о.)
0
1000
2000
3000
4000
5000

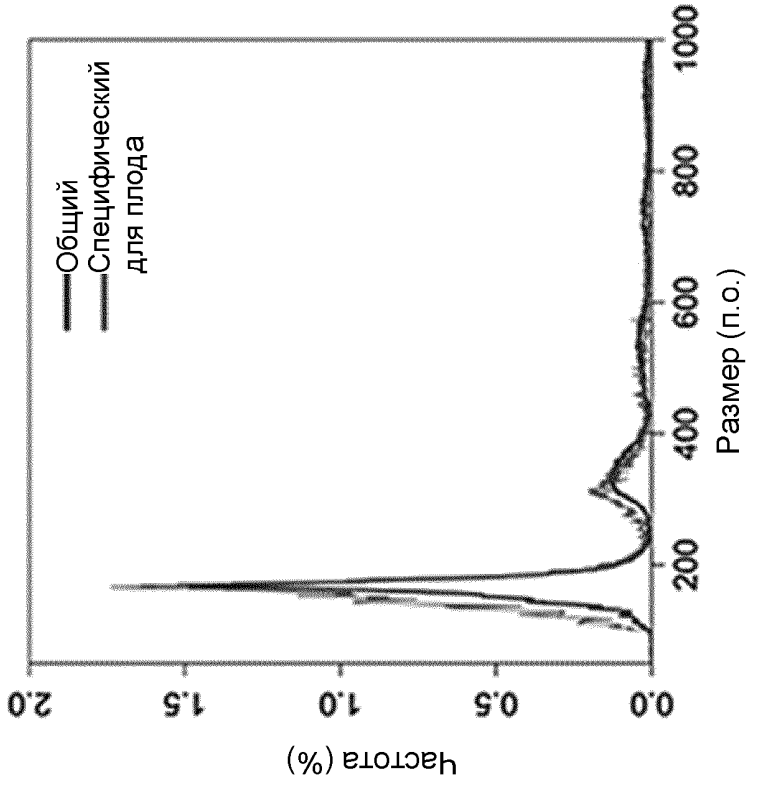
ФИГ. 5B


Общий
Специфический для плода
Частота (%)
0,0
0,5
1,0
1,5
2,0
Размер (п.о.)
1000
2000
3000
4000
5000

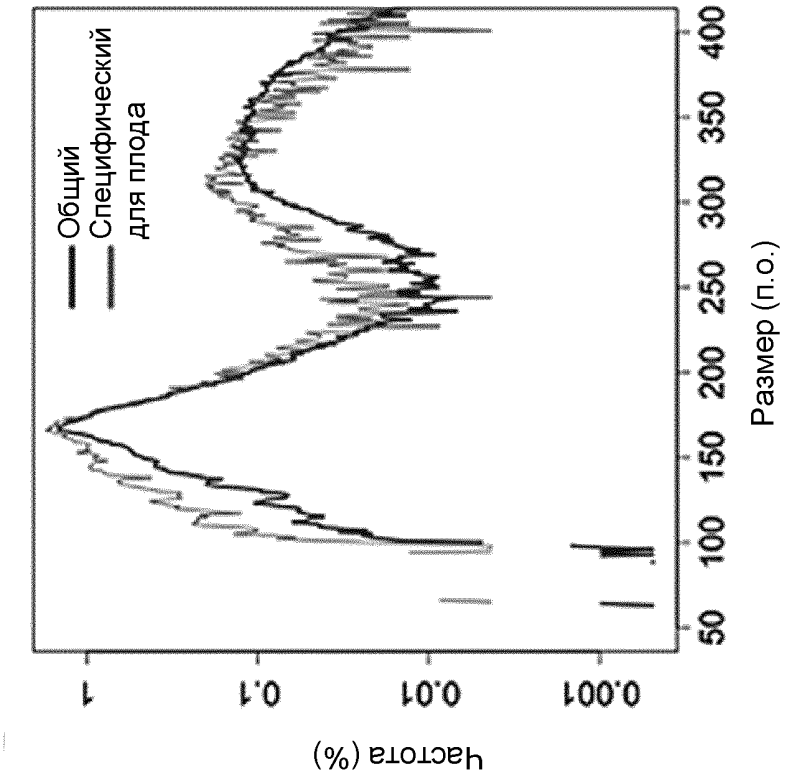
ФИГ. 5A


Частота (%)
0,001
0,01
0,1
1
Размер (п.о.)
200
400
600
800
1000
Общий
Специфический
для плода

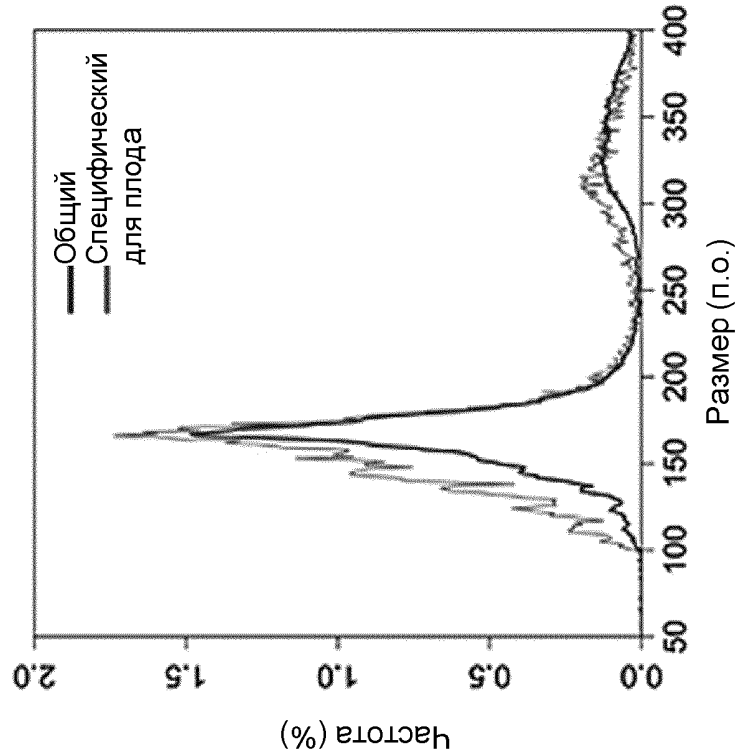
ФИГ. 6В


Частота (%)
0,0
0,5
1,0
1,5
2,0
Размер (п.о.)
200
400
600
800
1000
Общий
Специфический
для плода

ФИГ. 6А


Общий
Специфический для плода
Частота (%)
0.001
0.01
0.1
1
Размер (п.о.)
50
100
150
200
250
300
350
400

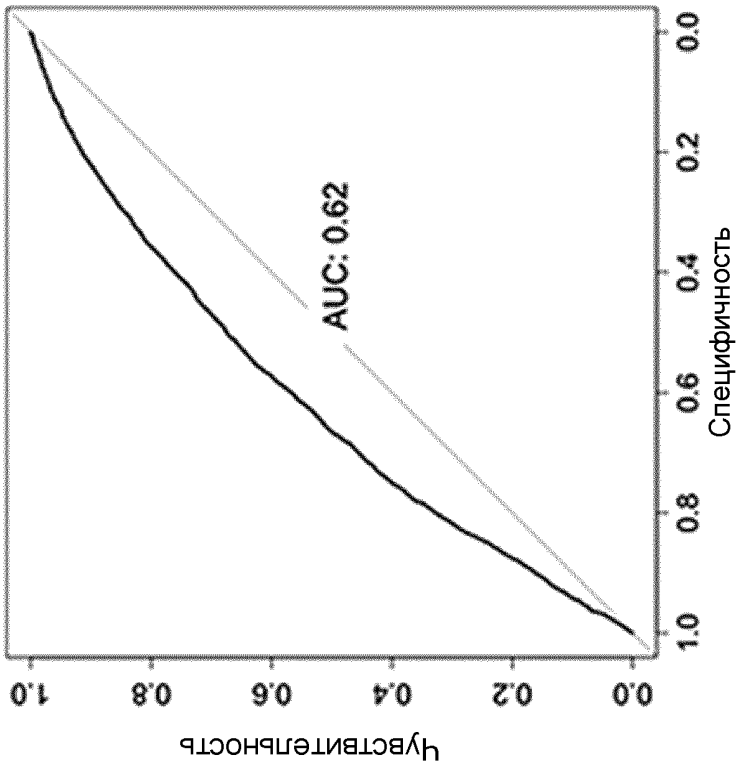
ФИГ. 7B


Общий
Специфический для плода
Частота (%)
0.0
0.5
1.0
1.5
2.0
Размер (п.о.)
50
100
150
200
250
300
350
400

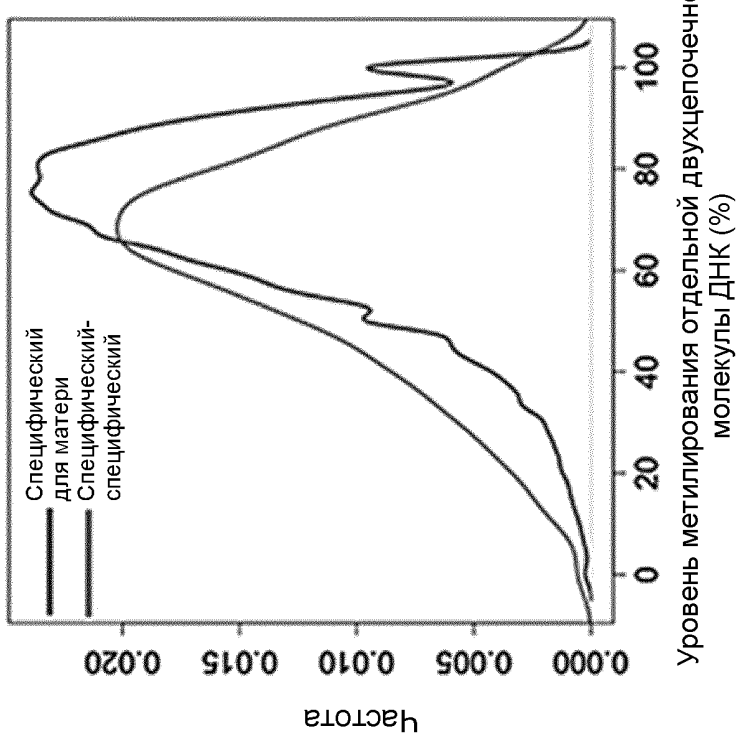
ФИГ. 7A

Уровень метилирования отдельной двухцепочечной молекулы ДНК (%)
20
40
60
80
100
Фрагменты, несущие специфические для матери аллели
Фрагменты, несущие специфические для плода аллели

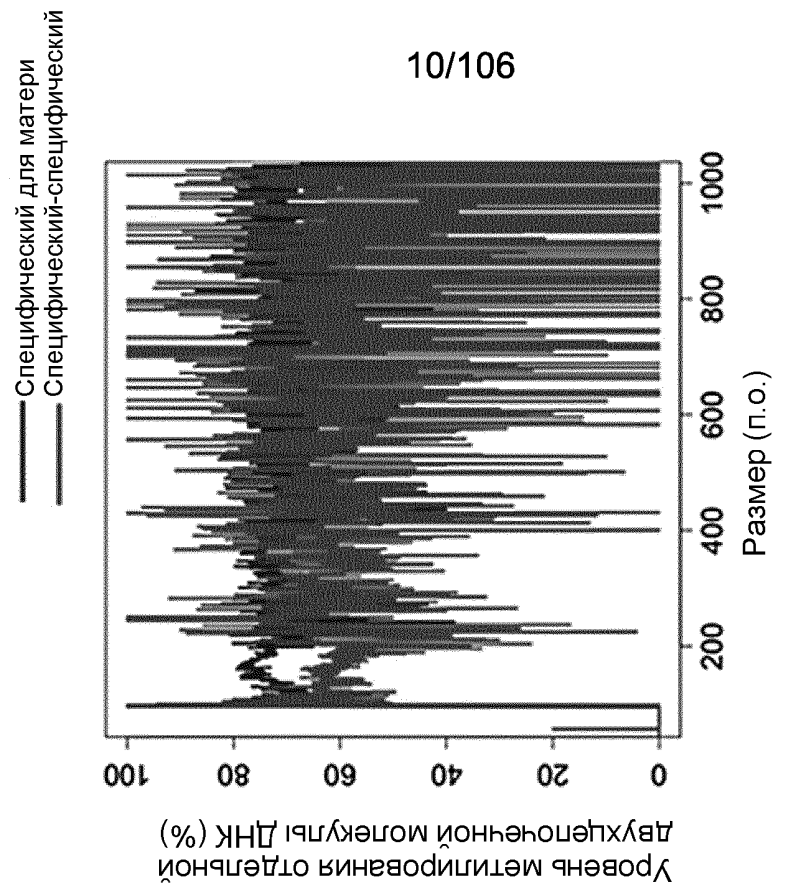
ФИГ. 8


Чувствительность
0.0
0.2
0.4
0.6
0.8
1.0
AUC: 0.62
1.0
0.8
0.6
0.4
0.2
0.0
Специфичность

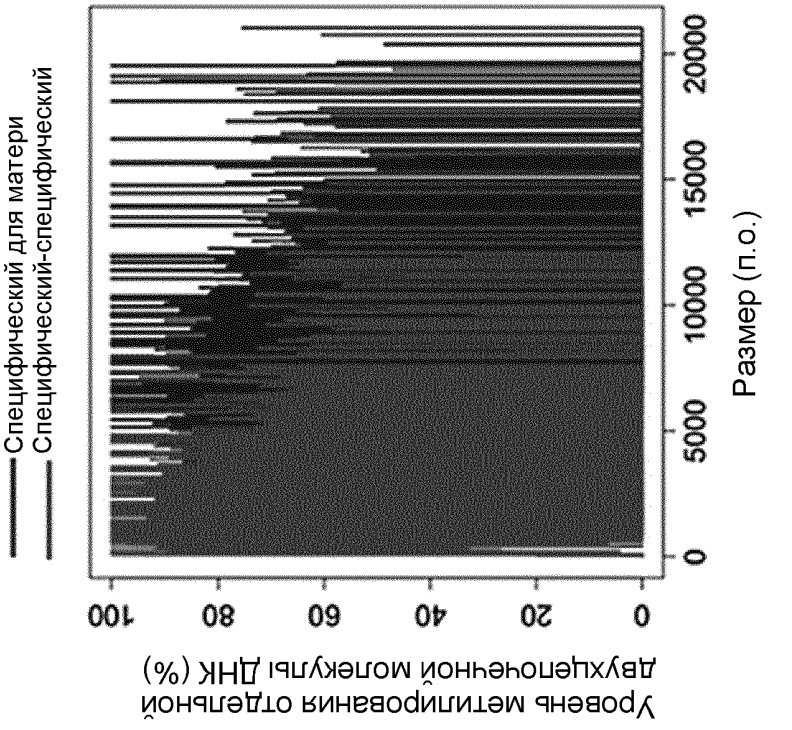
ФИГ. 9B


Частота
0.000
0.005
0.010
0.015
0.020
Специфический для матери
Специфический-специфический
0
20
40
60
80
100
Уровень метилирования отдельной двухцепочечной молекулы ДНК (%)

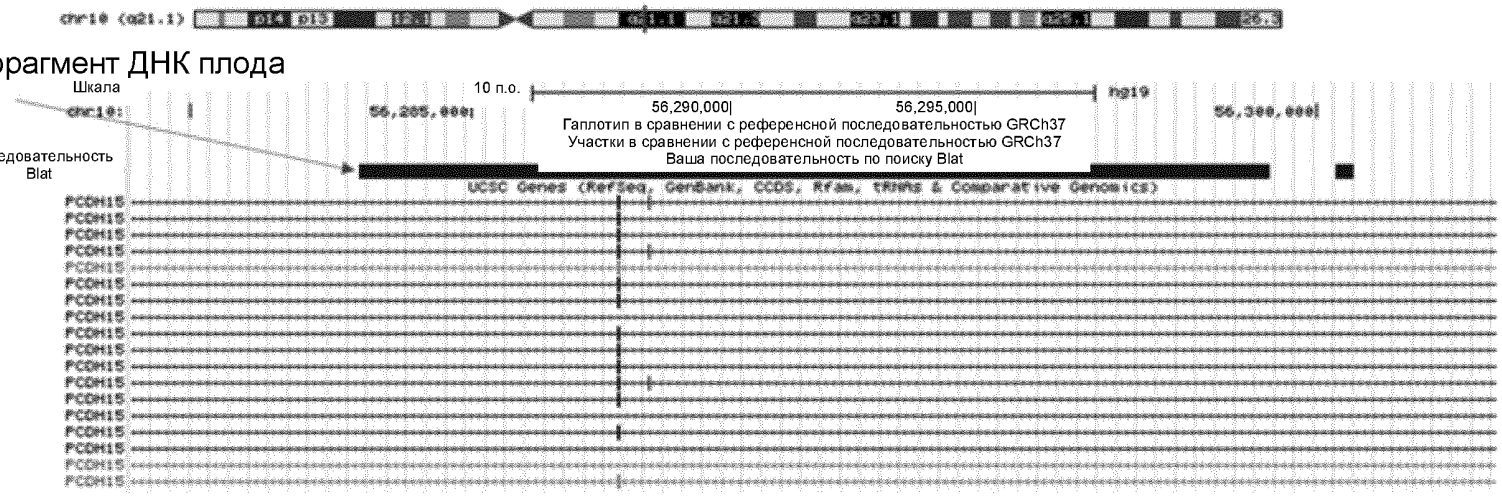
ФИГ. 9A



ФИГ. 10В



ФИГ. 10А

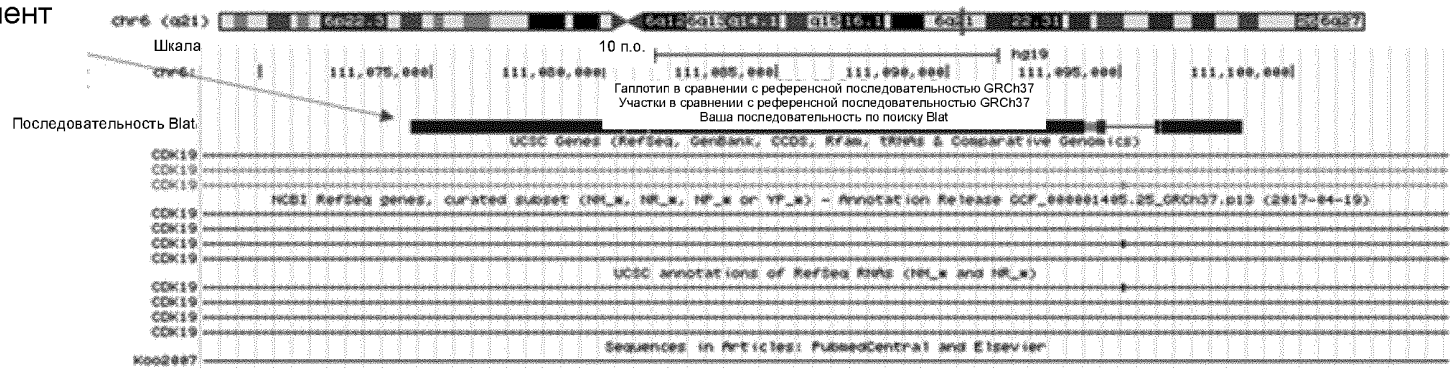


ФИГ. 11A

Секвенирование PacBio SMRT										Секвенирование Illumina тканевой ДНК			
Хромосома	Идентиф. № лунки	Координаты картирования длинного фрагмента внеклеточной ДНК		Длина	Специфические для плода аллели		Информация о последовательности	Количество сайтов CpG, классифицированных как метилированные	Количество сайтов CpG, классифицированных как неметилированные	Уровень метилирования отдельной двухцепочечной молекулы ДНК	Генотип матери	Генотип плода	Гаплотип матери	Гаплотип отца, переданный плоду
		Начало	Конец		Начало	Конец								
Хромосома10	127272846	56282981	56299166	16185	56284212	56284213	A	23	62	27.1	CC	CA	C	A
					56288865	56288866	C				TT	CT	T	C
					56290387	56290388	T				CC	CT	C	T
					56291934	56291935	C				GG	CG	G	C
					56296882	56296883	T				CC	CT	C	T
					56296930	56296931	T				TT	TC	T	C
					56296939	56296940	T				CC	CT	C	T

ФИГ. 11B

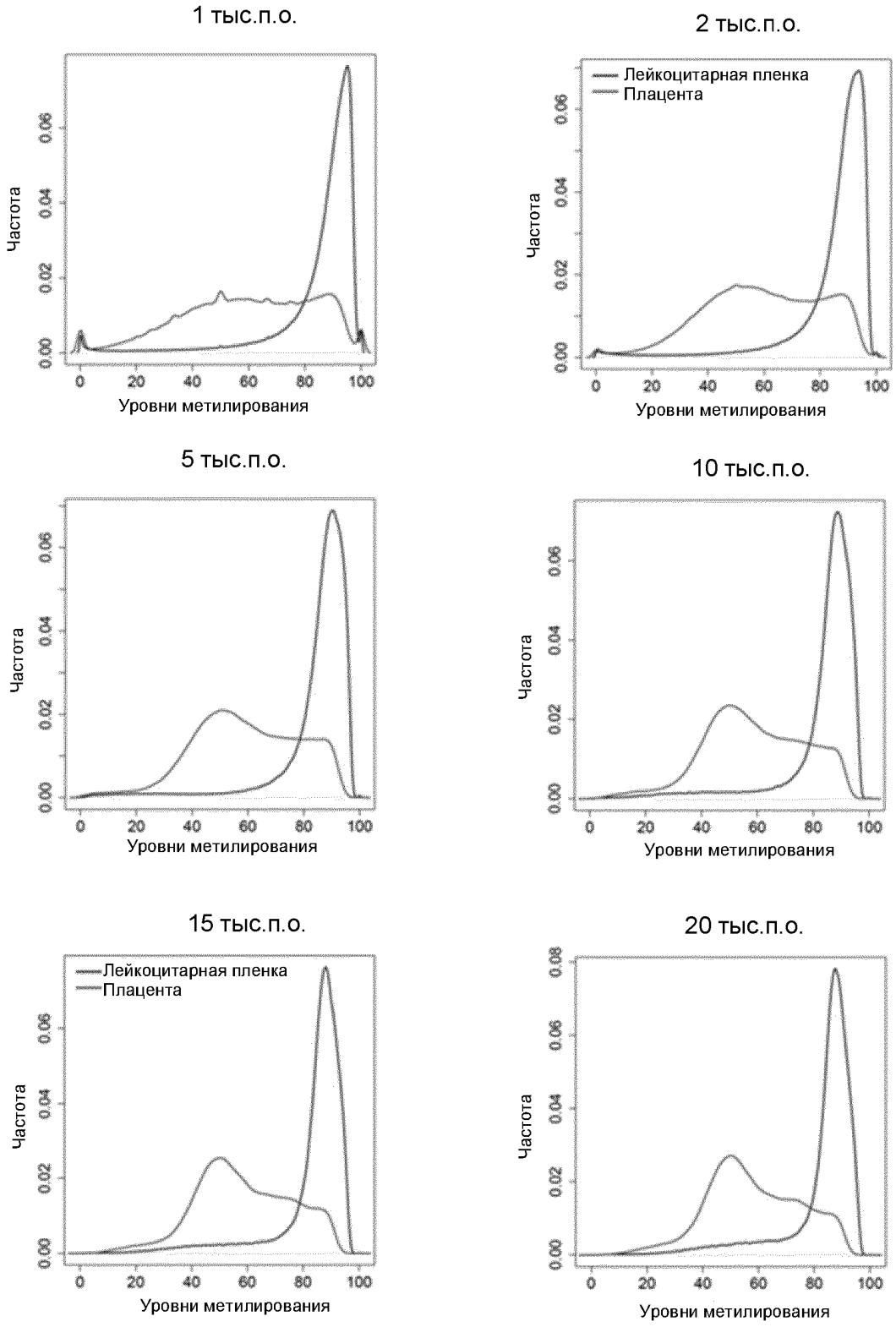
Длинный фрагмент ДНК матери



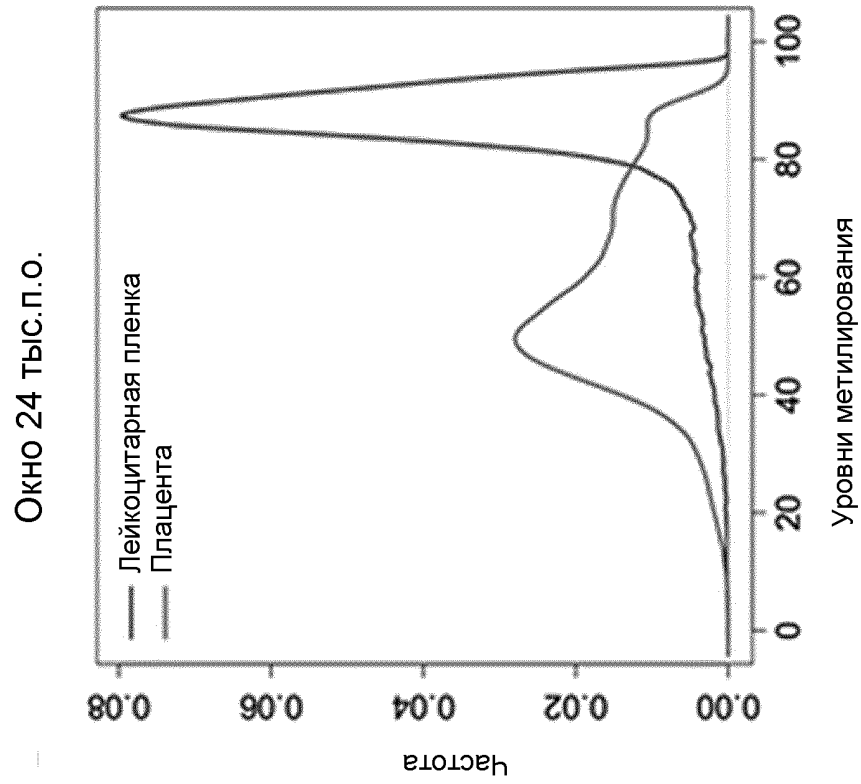
ФИГ. 12A

Секвенирование PacBio SMRT											Секвенирование Illumina тканевой ДНК			
		Координаты картирования			Специфические для плода аллели									
Хромосома	Идентиф. № лунки	Начало	Конец	Длина	Начало	Конец	Информация о последовательности	Количество сайтов CpG, классифицированных как метилированные	Количество сайтов CpG, классифицированных как неметилированные	Уровень метилирования отдельной двухцепочечной молекулы ДНК	Генотип матери	Генотип плода	Гаплотип матери	Гаплотип отца, переданный плоду
Хромосома6	128648561	111074371	111098536	24165	111075831	111075832	**A**	81	40	66.9	AA	AG	**A**	G
					111076442	111076443	**G**				GG	GT	**G**	T
					111076870	111076871	**A**				AA	AT	**A**	T
					111077506	111077507	**C**				CC	CT	**C**	T
					111078649	111078650	**T**				TT	CT	**T**	C
					111079042	111079043	**A**				AA	GA	**A**	G
					111081132	111081133	**C**				CC	CT	**C**	T
					111083336	111083337	**G**				GG	CG	**G**	C
					111084746	111084747	**G**				GG	TG	**G**	T
					111086718	111086719	**C**				CC	TC	**C**	T
					111087391	111087392	**A**				AA	AT	**A**	T
					111088191	111088192	**C**				CC	CA	**C**	A
					111091243	111091244	**C**				CC	TC	**C**	T
					111093070	111093071	**T**				TT	TC	**T**	C
					111093763	111093764	**A**				AA	GA	**A**	G
					111097401	111097402	**T**				TT	CT	**T**	C
					111097556	111097557	**C**				CC	CG	**C**	G
					111098297	111098298	**A**				AA	CA	**A**	C

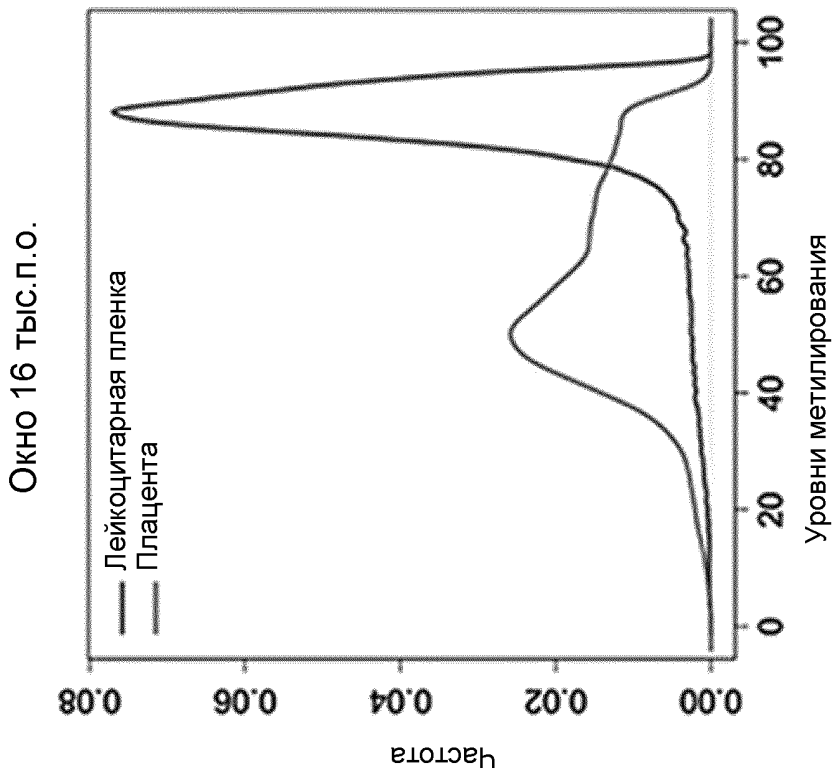
ФИГ. 12B


1 тыс.п.о.
2 тыс.п.о.
5 тыс.п.о.
10 тыс.п.о.
15 тыс.п.о.
20 тыс.п.о.
Лейкоцитарная пленка
Плацента
Частота
Уровни метилирования
0.00
0.02
0.04
0.06
0.08
0
20
40
60
80
100

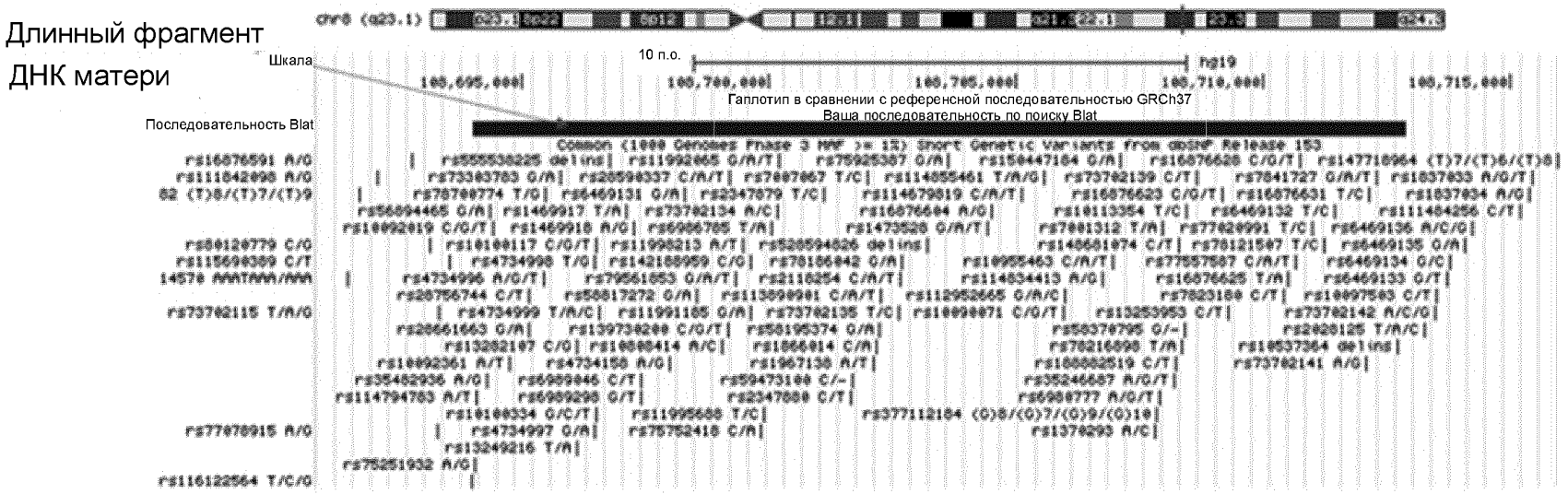
ФИГ. 13



ФИГ. 14B



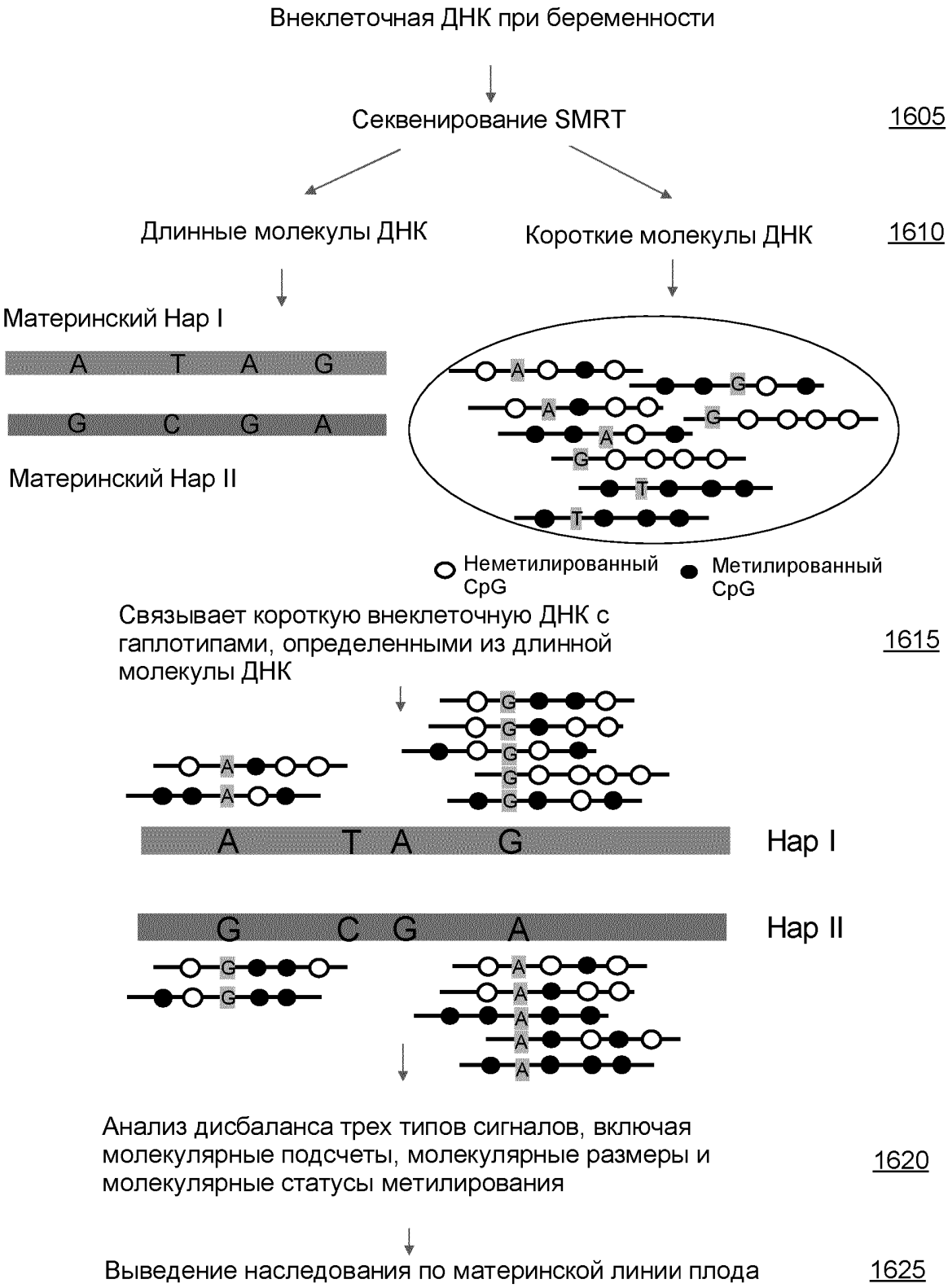
ФИГ. 14A



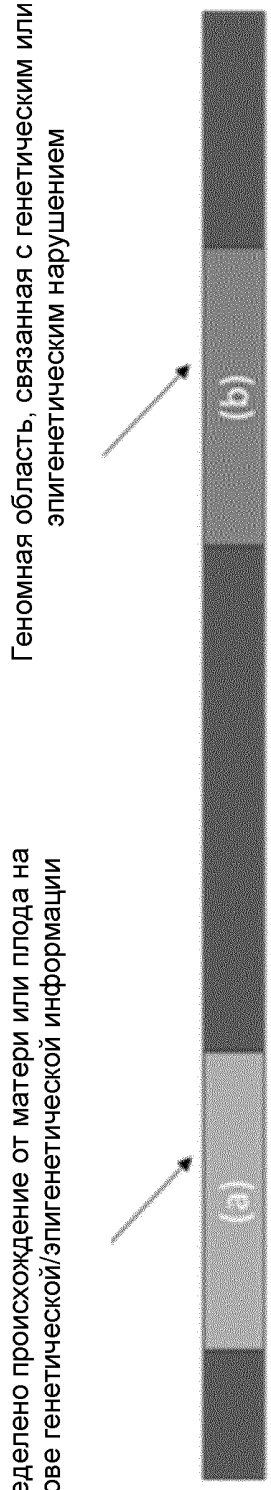
ФИГ. 15A

Секвенирование PacBio SMRT														
		Координаты картирования			Специфические для плода аллели						Секвенирование Illumina тканевой ДНК			
Хромосома	Идентиф. № лунки	Начало	Конец	Длина	Начало	Конец	Информация о последовательности	Количество сайтов CpG, классифицированных как метилированные	Количество сайтов CpG, классифицированных как неметилированные	Уровень метилирования отдельной двухцепочечной молекулы ДНК	Генотип матери	Генотип плода	Гаплотип матери	Гаплотип отца, переданный плоду
Хромосома8	107546586	108694010	108712904	18894	108695136	108695137	A	77	29	72.6	GA	GG	A	G
					108695137	108695138	T				CT	CC	T	C
					108697116	108697117	T				AT	AA	T	A
					108702099	108702100	T				AT	AA	T	A
					108706969	108706970	C				CT	TT	C	T
					108709180	108709181	G				GC	CC	G	C
					108712304	108712305	G				TG	TT	G	T

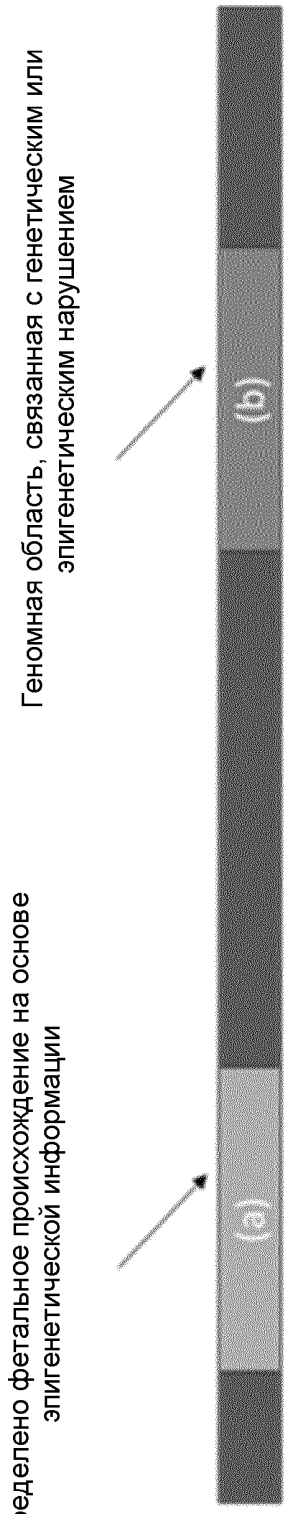
ФИГ. 15B


Внеклеточная ДНК при беременности
Секвенирование SMRT
1605
Длинные молекулы ДНК
Короткие молекулы ДНК
1610
Материнский Hap I
A T A G
G C G A
Материнский Hap II
Неметилированный CpG
Метилированный CpG
Связывает короткую внеклеточную ДНК с гаплотипами, определенными из длинной молекулы ДНК
1615
A T A G
Hap I
G C G A
Hap II
Анализ дисбаланса трех типов сигналов, включая молекулярные подсчеты, молекулярные размеры и молекулярные статусы метилирования
1620
Выведение наследования по материнской линии плода
1625

ФИГ. 16

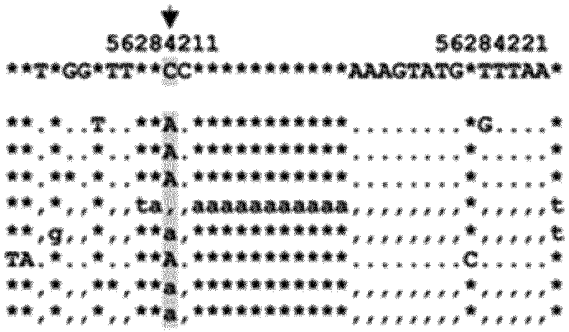

Определено происхождение от матери или плода на основе генетической/эпигенетической информации
Геномная область, связанная с генетическим или эпигенетическим нарушением
(a)
(b)
Длинная молекула ДНК плазмы

ФИГ. 17


Геномная область, связанная с генетическим или эпигенетическим нарушением
(b)
Длинная молекула ДНК плазмы
(a)
Определено фетальное происхождение на основе эпигенетической информации

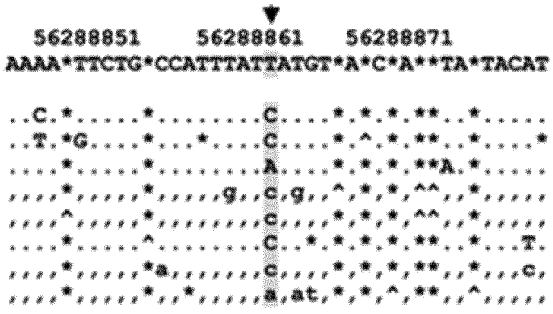
ФИГ. 18

1. Правильный ОНП, chr10:56284213, C->A



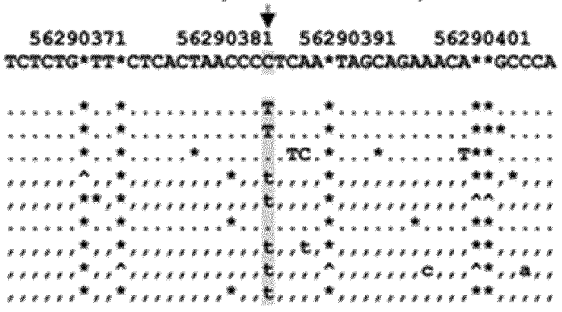
ФИГ. 19A

2. Правильный ОНП, chr10:56288866, T->C



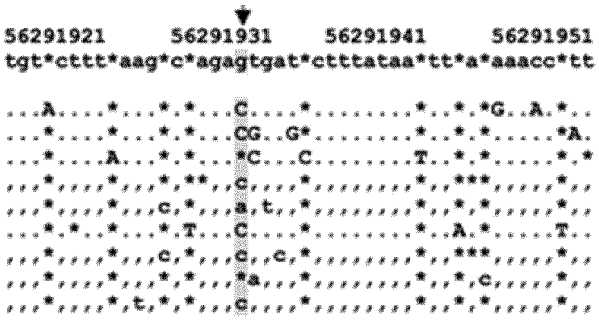
ФИГ. 19B

3. Правильный ОНП, chr10:56290388, C->T



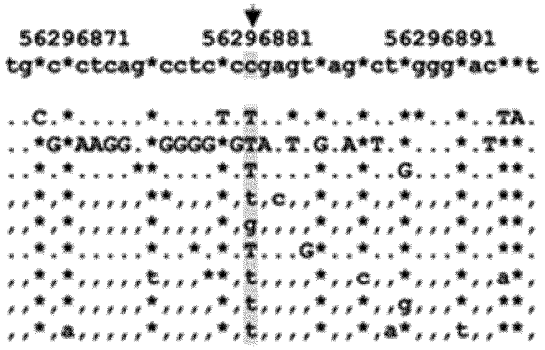
ФИГ. 19C

4. Правильный ОНП, chr10:56291935, G->C



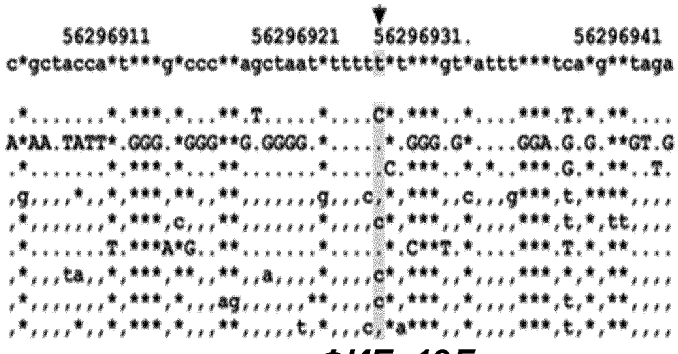
ФИГ. 19D

5. Правильный ОНП, chr10:56296883, C->T



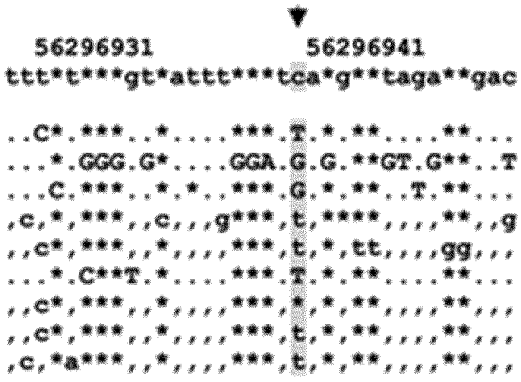
ФИГ. 19E

6. Ошибочный сайт, chr10:56296931

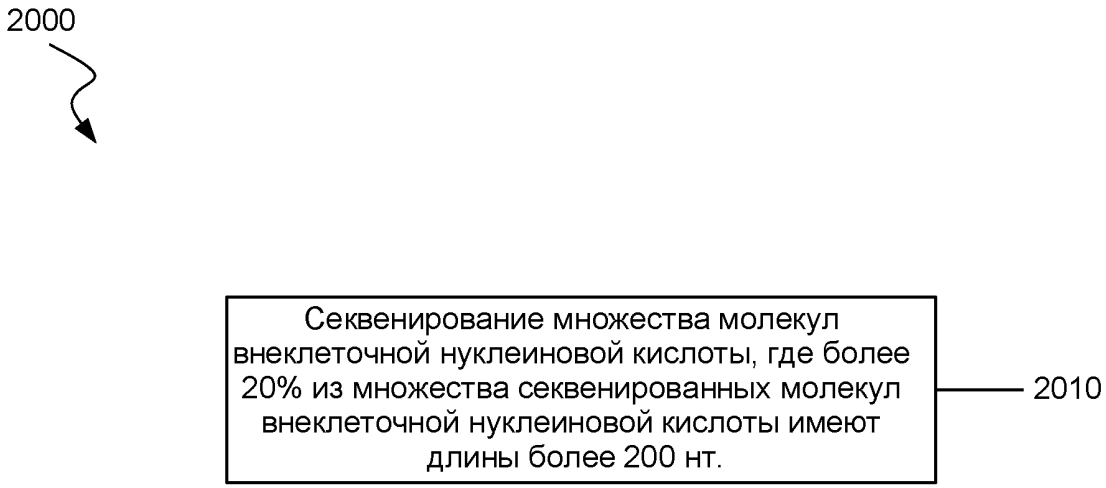


ФИГ. 19F

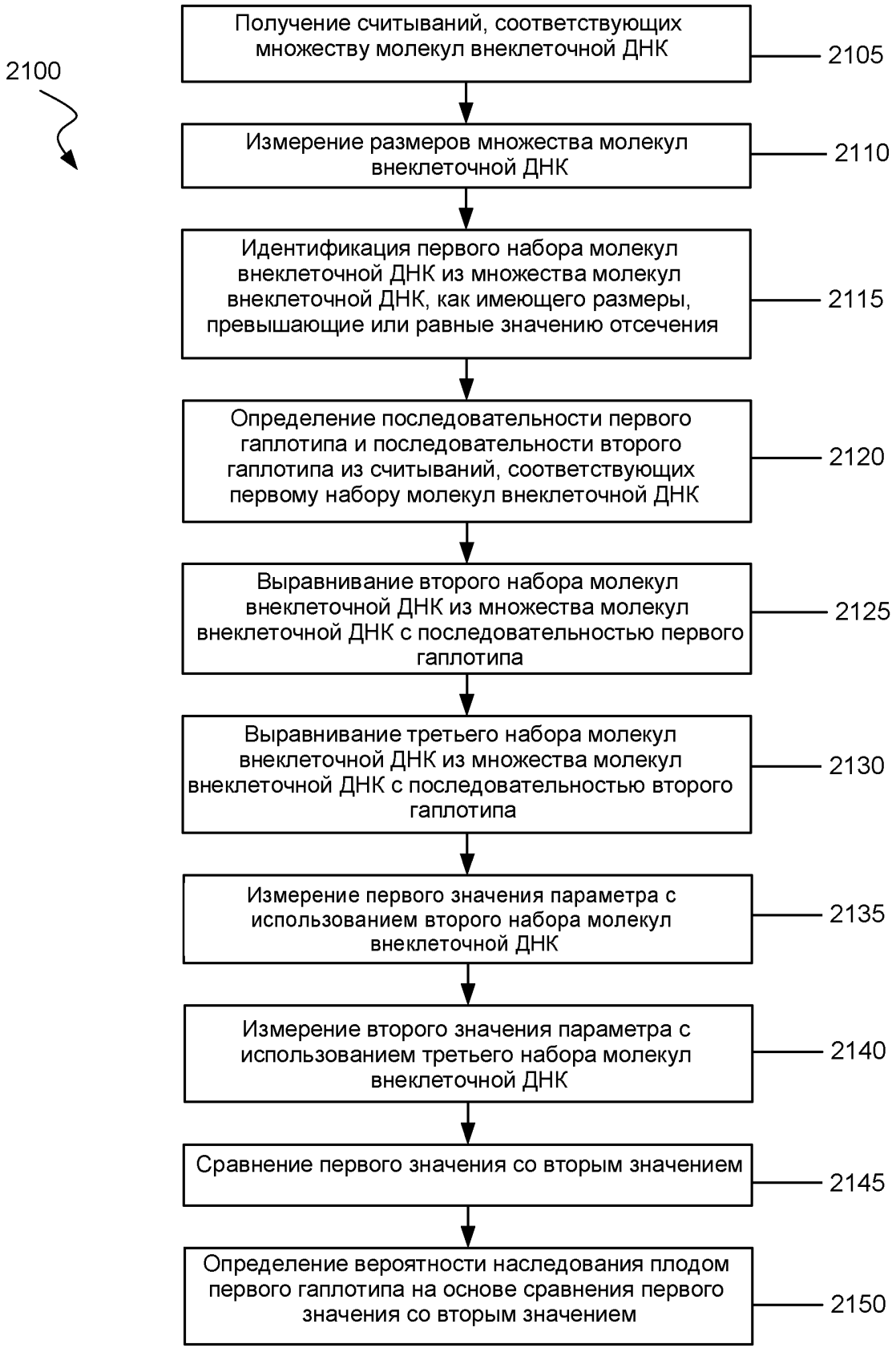
7. Правильный ОНП, chr10:56291935, C->T



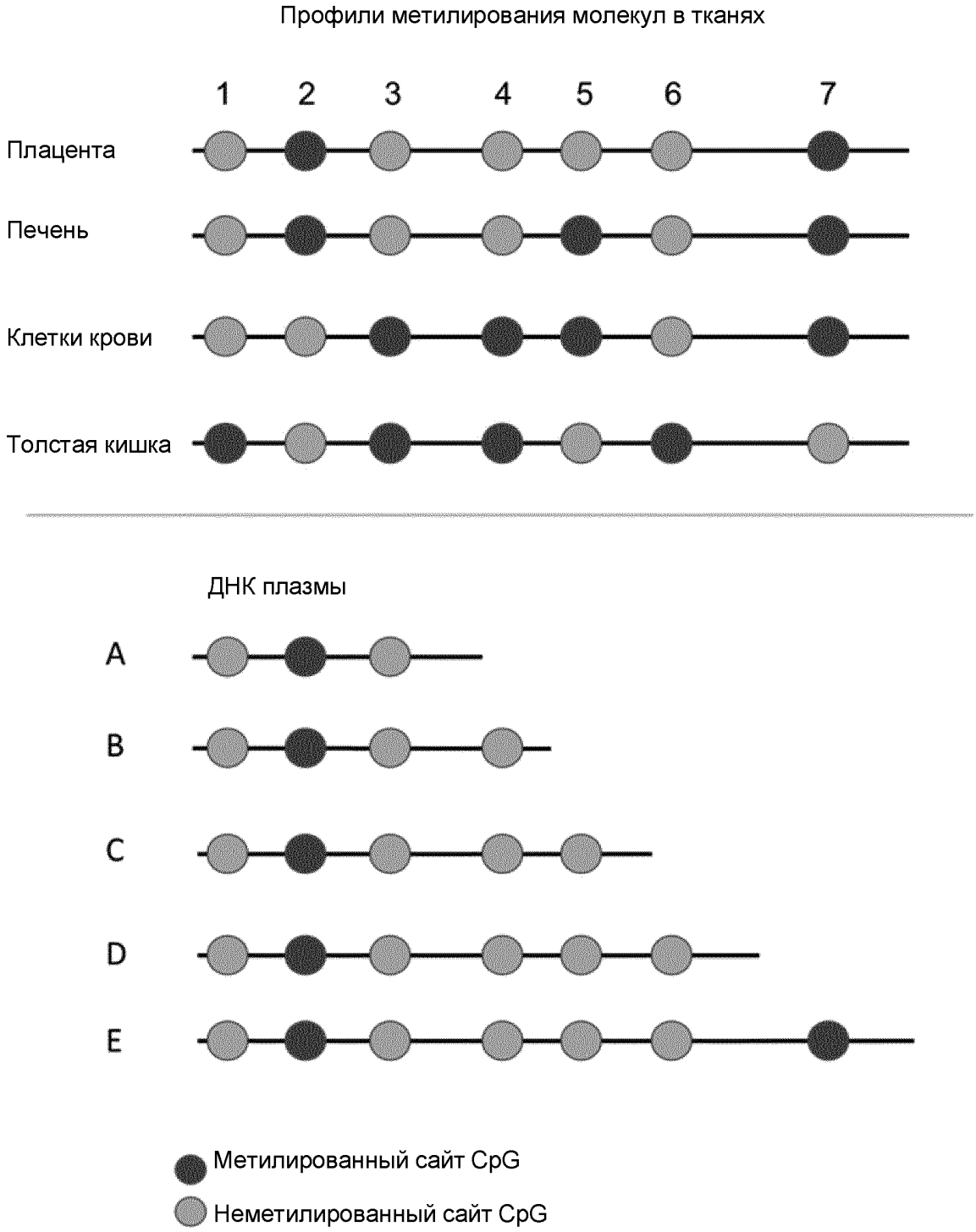
ФИГ. 19G


2000
Секвенирование множества молекул внеклеточной нуклеиновой кислоты, где более 20% из множества секвенированных молекул внеклеточной нуклеиновой кислоты имеют длины более 200 нт.
2010

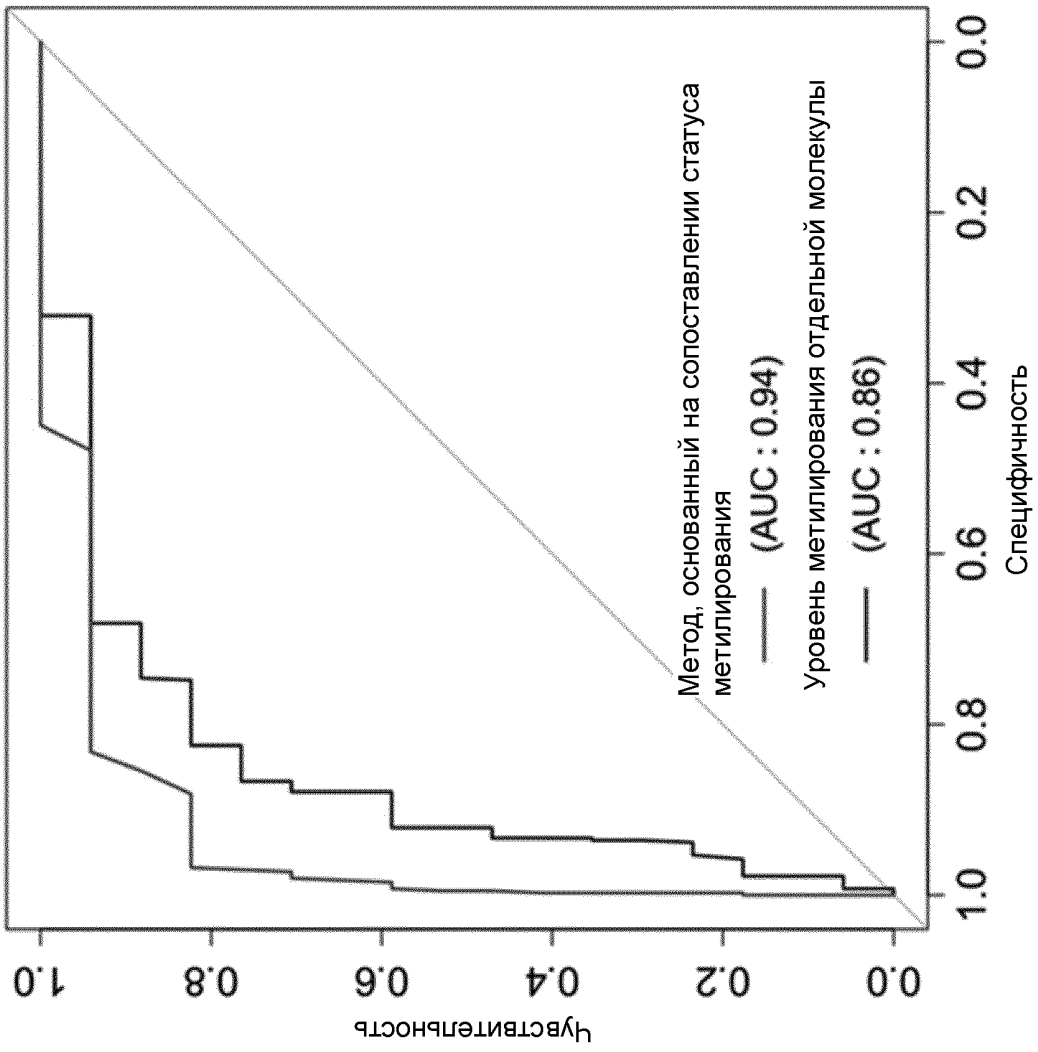
ФИГ. 20


2100
Получение считываний, соответствующих множеству молекул внеклеточной ДНК
2105
Измерение размеров множества молекул внеклеточной ДНК
2110
Идентификация первого набора молекул внеклеточной ДНК из множества молекул внеклеточной ДНК, как имеющего размеры, превышающие или равные значению отсечения
2115
Определение последовательности первого гаплотипа и последовательности второго гаплотипа из считываний, соответствующих первому набору молекул внеклеточной ДНК
2120
Выравнивание второго набора молекул внеклеточной ДНК из множества молекул внеклеточной ДНК с последовательностью первого гаплотипа
2125
Выравнивание третьего набора молекул внеклеточной ДНК из множества молекул внеклеточной ДНК с последовательностью второго гаплотипа
2130
Измерение первого значения параметра с использованием второго набора молекул внеклеточной ДНК
2135
Измерение второго значения параметра с использованием третьего набора молекул внеклеточной ДНК
2140
Сравнение первого значения со вторым значением
2145
Определение вероятности наследования плодом первого гаплотипа на основе сравнения первого значения со вторым значением
2150

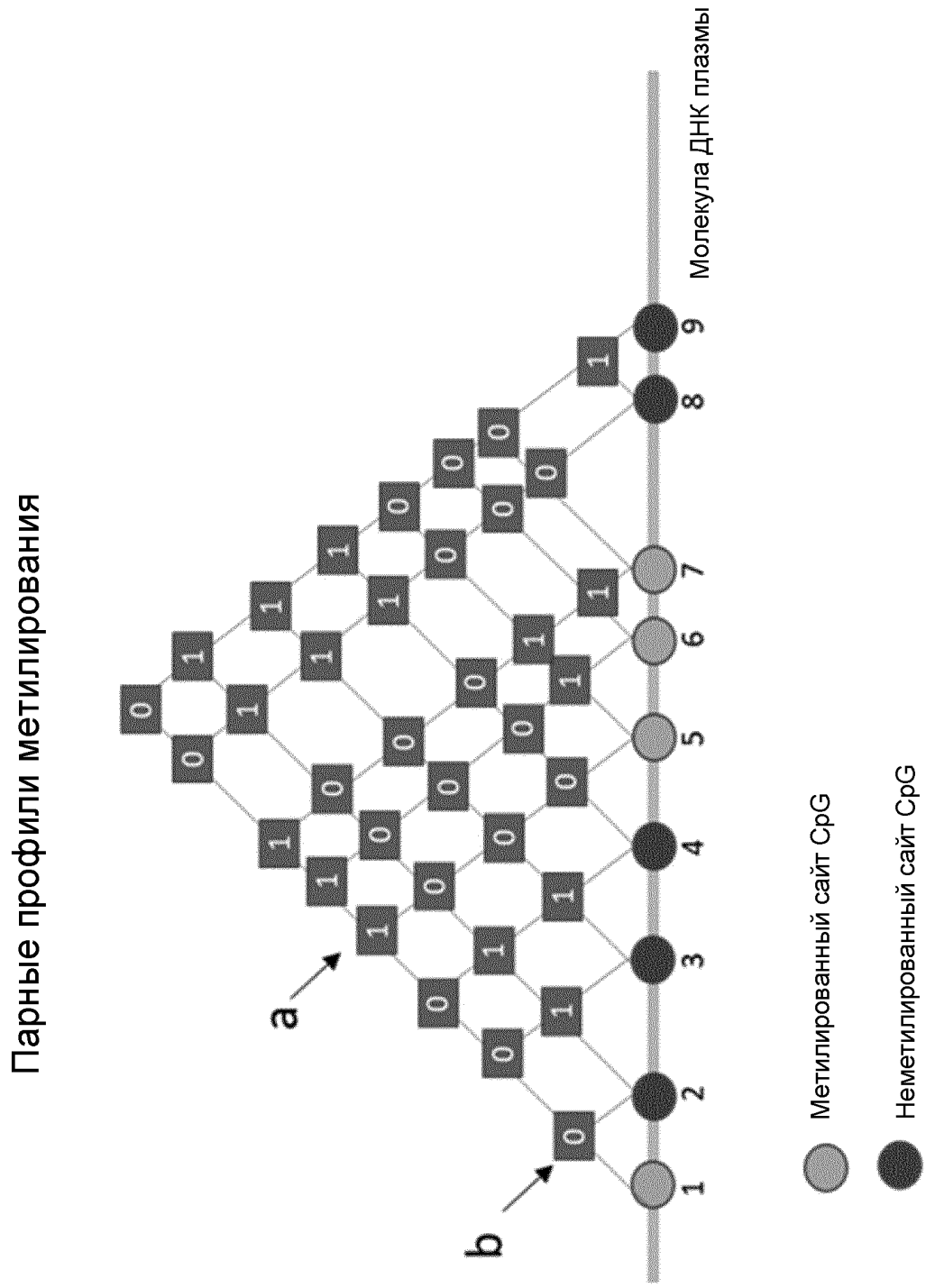
ФИГ. 21


Профили метилирования молекул в тканях
1
2
3
4
5
6
7
Плацента
Печень
Клетки крови
Толстая кишка
ДНК плазмы
A
B
C
D
E
Метилированный сайт CpG
Неметилированный сайт CpG

ФИГ. 22


Чувствительность
0.0
0.2
0.4
0.6
0.8
1.0
1.0
0.8
0.6
0.4
0.2
0.0
Специфичность
Метод, основанный на сопоставлении статуса метилирования
(AUC : 0.94)
Уровень метилирования отдельной молекулы
(AUC : 0.86)

ФИГ. 23


Парные профили метилирования
Молекула ДНК плазмы
1
2
3
4
5
6
7
8
9
a
b
Метилированный сайт CpG
Неметилированный сайт CpG

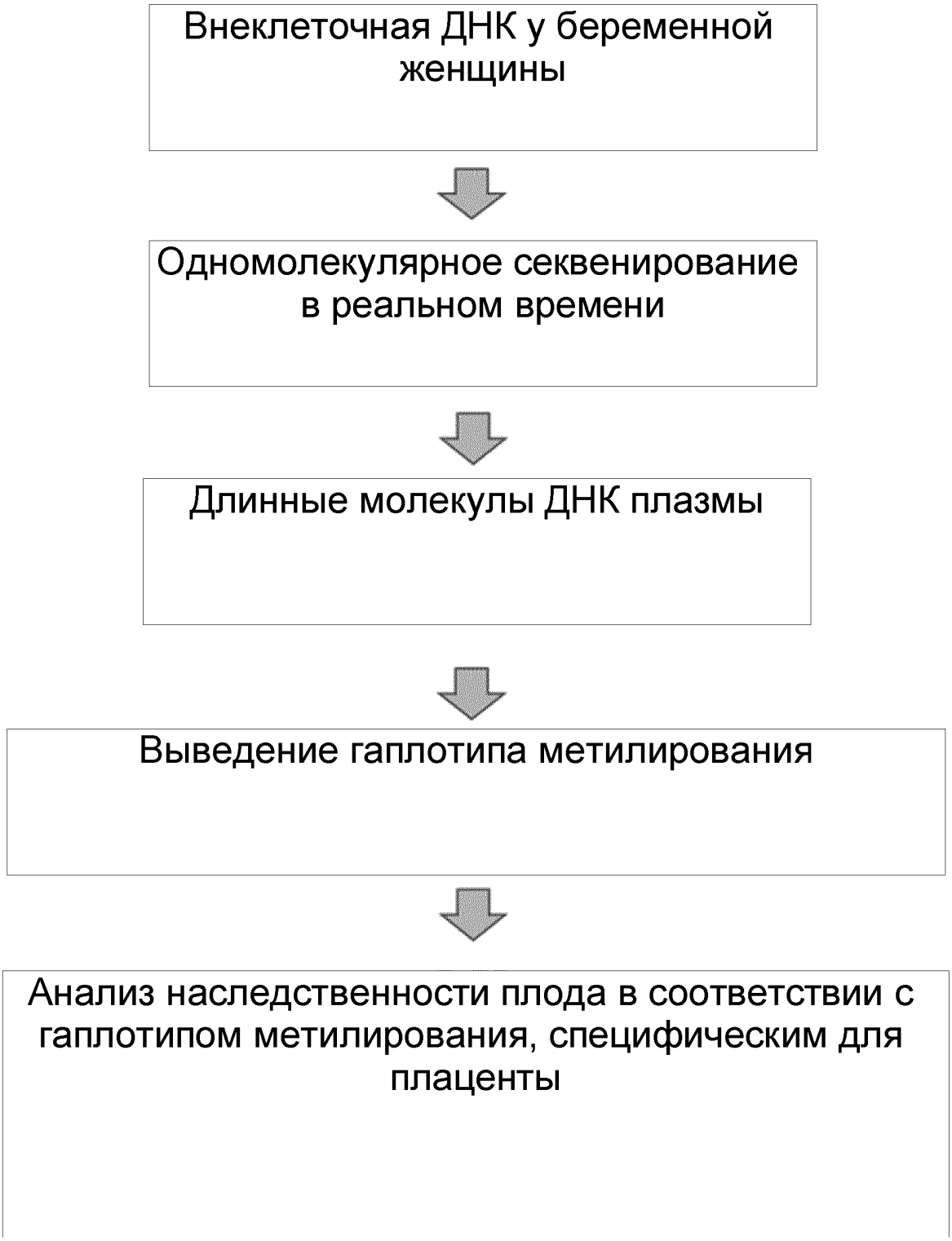
ФИГ. 24

Хромосома	Количество маркерных областей
1	22418
2	20311
3	14559
4	15700
5	15363
6	12758
7	15108
8	15172
9	11742
10	13914
11	13545
12	13697
13	10449
14	9523
15	7132
16	13962
17	11177
18	8186
19	15622
20	8264
21	5960
22	7004

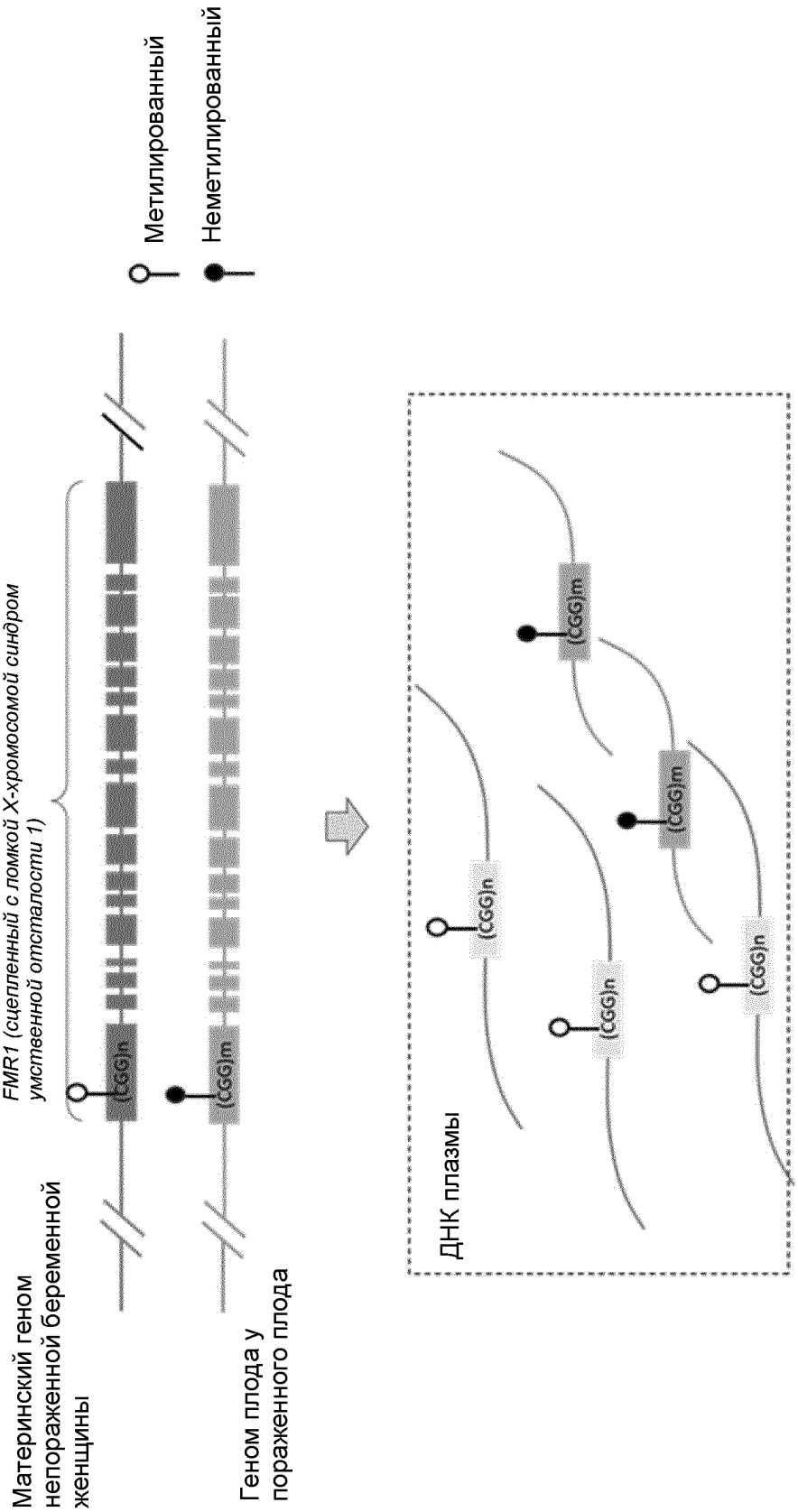
ФИГ. 25

Процентное содержание молекул ДНК лейкоцитарной пленки, имеющих оценку несоответствия более 0,3%		Классификация на основе профиля метилирования отдельной молекулы		Процентное содержание молекул ДНК с правильной классификацией (%)
		Специфич. для плаценты	Неспецифич. для плаценты	
> 60	Молекулы ДНК плазмы, охватывающие специфический для плода аллель	24	17	58.5
	Молекулы ДНК плазмы, охватывающие специфический для матери аллель	63	597	90.5
> 70	Молекулы ДНК плазмы, охватывающие специфический для плода аллель	23	14	62.2
	Молекулы ДНК плазмы, охватывающие специфический для матери аллель	53	550	91.2
> 75	Молекулы ДНК плазмы, охватывающие специфический для плода аллель	20	13	60.6
	Молекулы ДНК плазмы, охватывающие специфический для матери аллель	45	493	91.6
> 80	Молекулы ДНК плазмы, охватывающие специфический для плода аллель	17	11	60.7
	Молекулы ДНК плазмы, охватывающие специфический для матери аллель	34	433	92.7
> 85	Молекулы ДНК плазмы, охватывающие специфический для плода аллель	14	11	56.0
	Молекулы ДНК плазмы, охватывающие специфический для матери аллель	21	342	94.2
> 90	Молекулы ДНК плазмы, охватывающие специфический для плода аллель	9	7	56.3
	Молекулы ДНК плазмы, охватывающие специфический для матери аллель	10	209	95.4

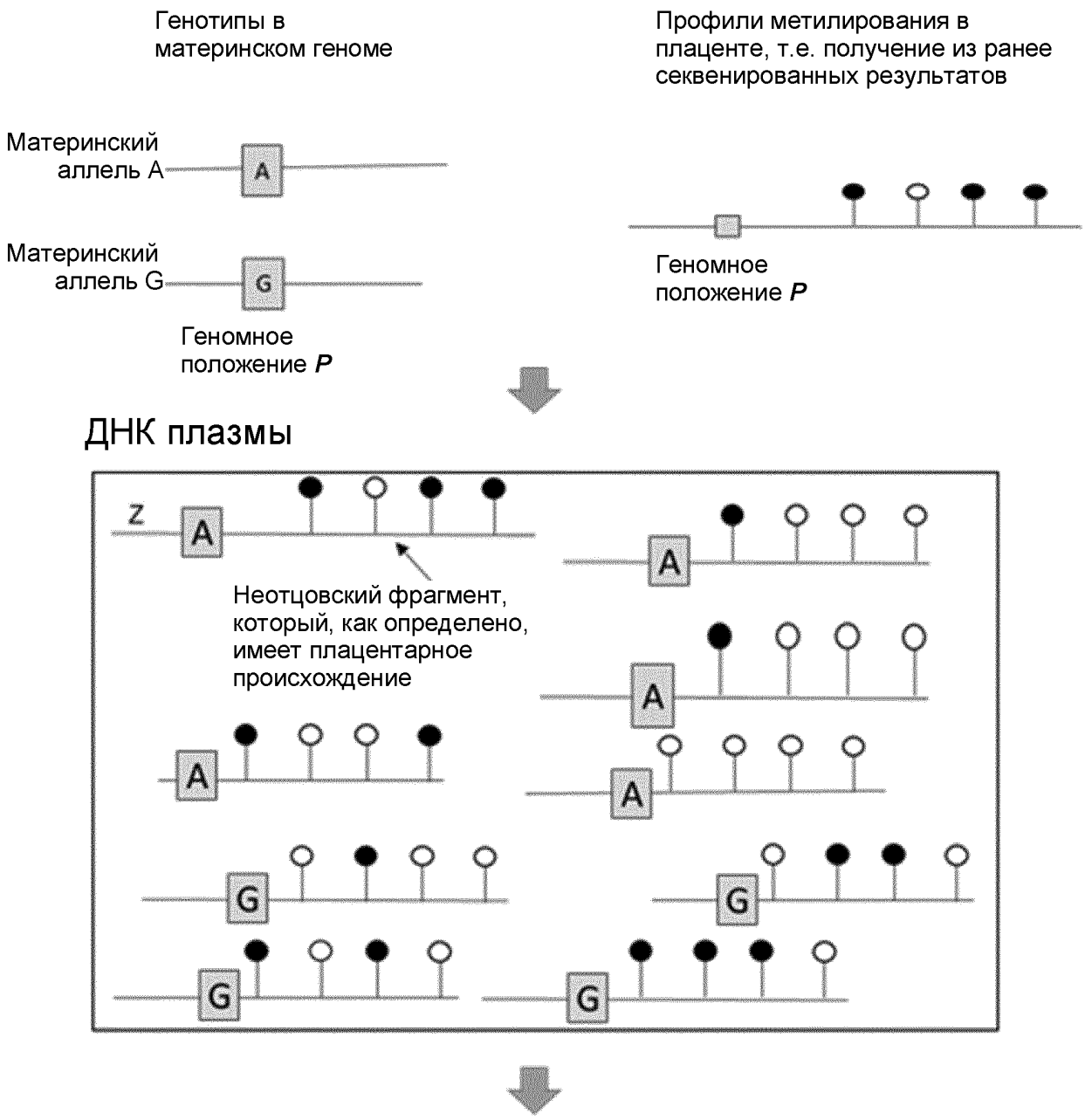
ФИГ. 26


Внеклеточная ДНК у беременной женщины
Одномолекулярное секвенирование в реальном времени
Длинные молекулы ДНК плазмы
Выведение гаплотипа метилирования
Анализ наследственности плода в соответствии с гаплотипом метилирования, специфическим для плаценты

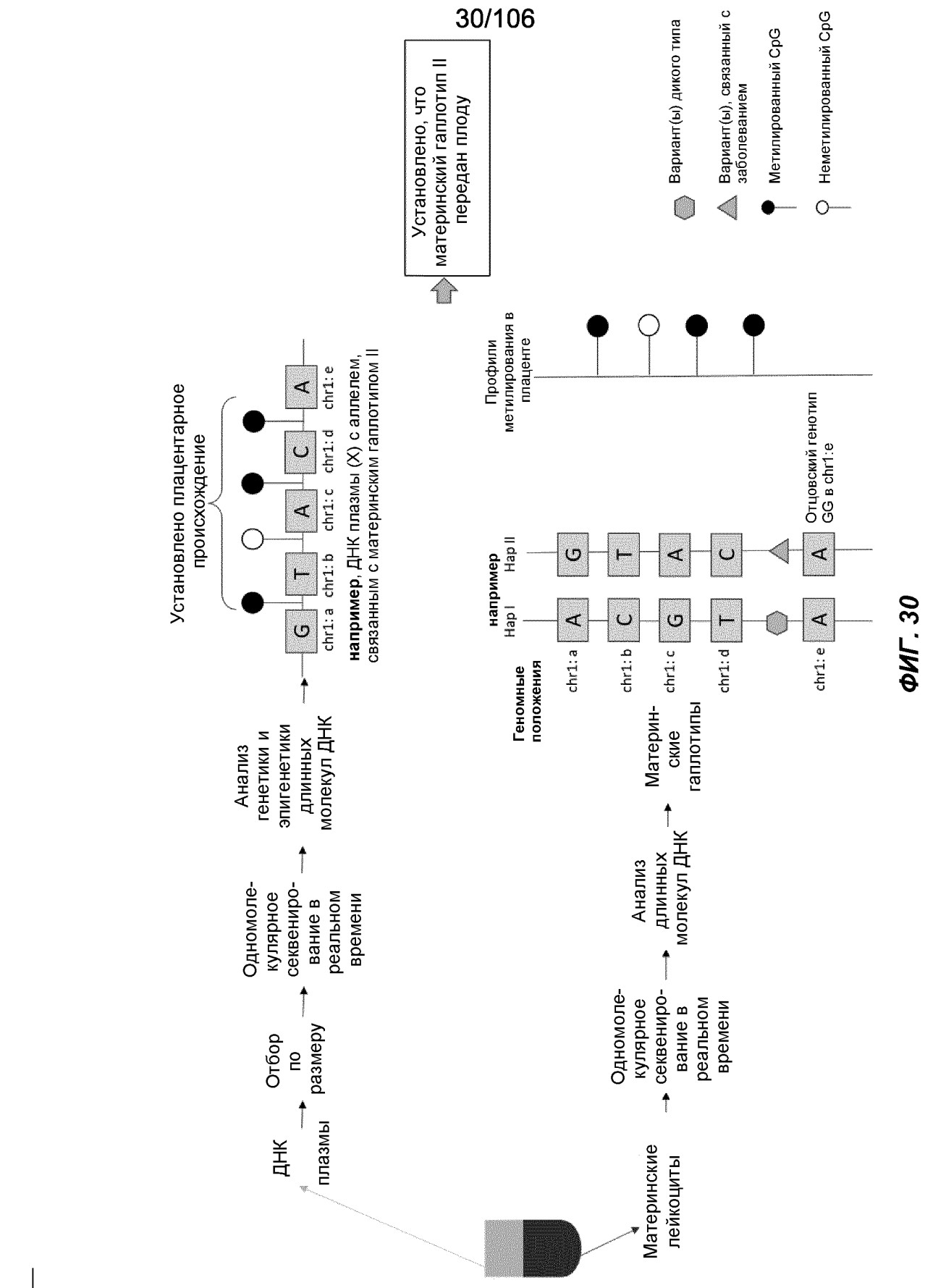
ФИГ. 27


Метилированный
Неметилированный
FMR1 (сцепленный с ломкой X-хромосомой синдром умственной отсталости 1)
(CGG)n
(CGG)m
Материнский геном непораженной беременной женщины
Геном плода у пораженного плода
ДНК плазмы
(CGG)n
(CGG)n
(CGG)m
(CGG)m
(CGG)n

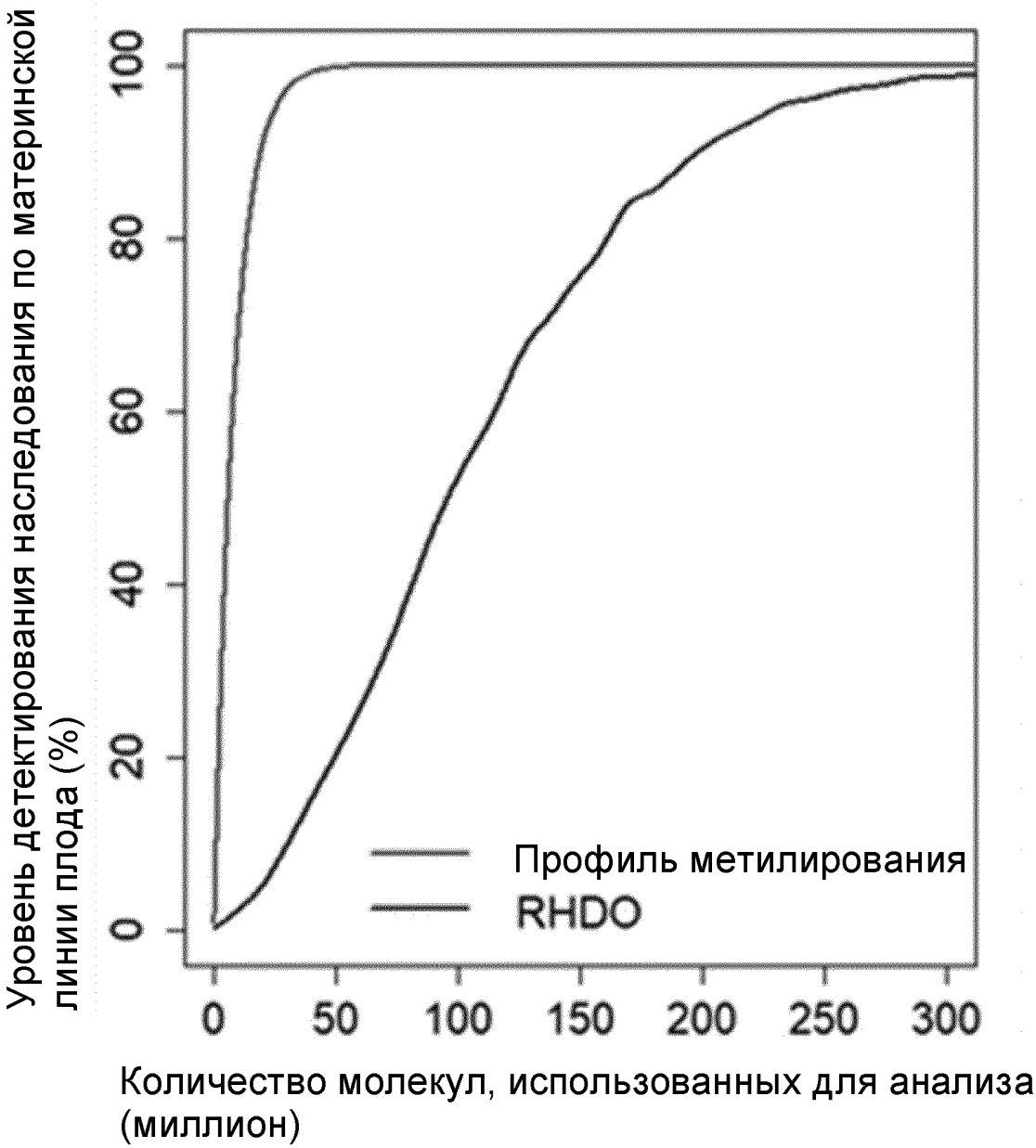
ФИГ. 28


Генотипы в материнском геноме
Профили метилирования в плаценте, т.е. получение из ранее секвенированных результатов
Материнский аллель A
A
Материнский аллель G
G
Геномное положение P
Геномное положение P
ДНК плазмы
Z
A
Неотцовский фрагмент, который, как определено, имеет плацентарное происхождение
A
A
A
G
G
G
Плод наследует материнский аллель A

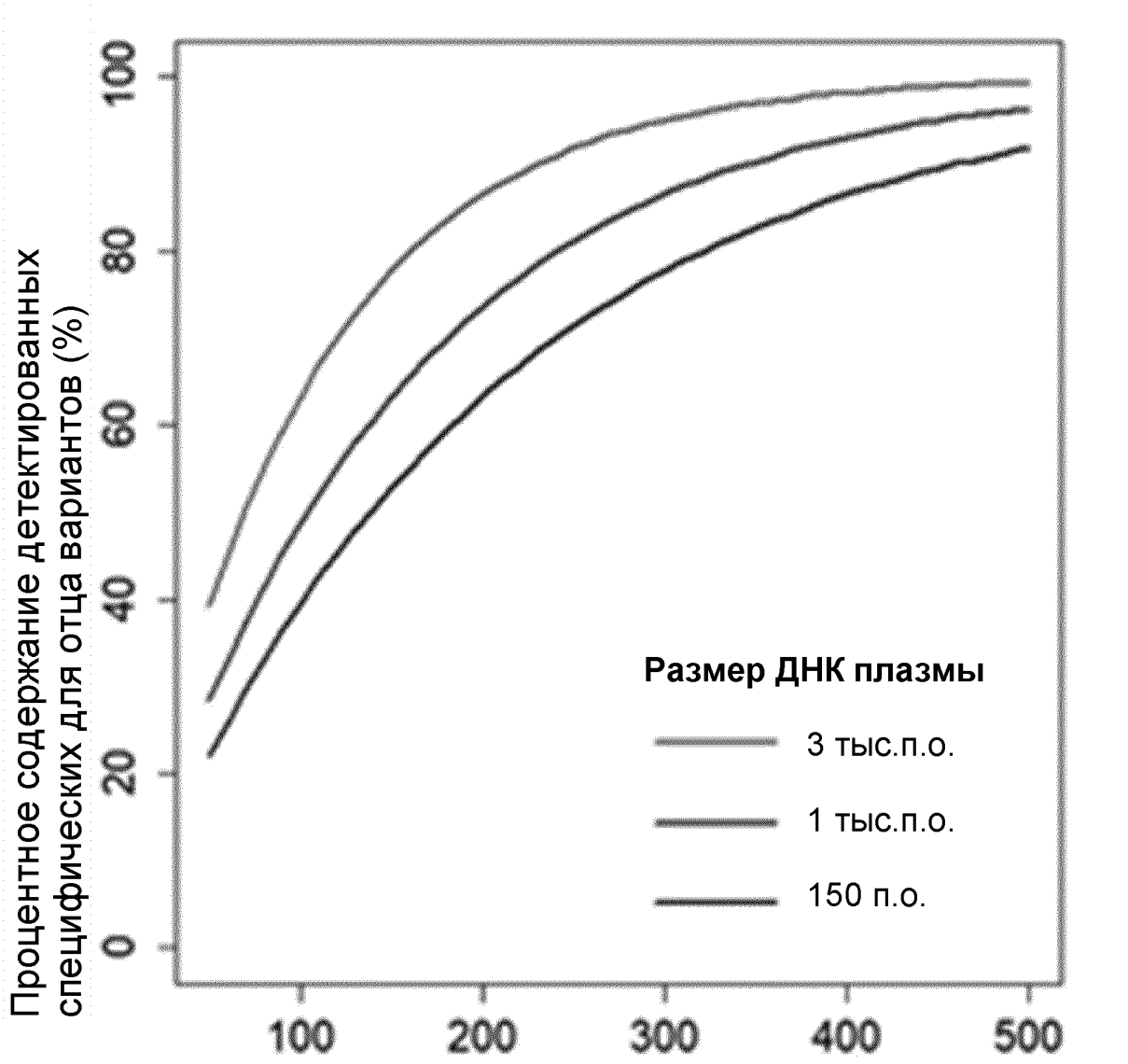
ФИГ. 29


Установлено, что материнский гаплотип II передан плоду
Вариант(ы) дикого типа
Вариант(ы), связанный с заболеванием
Метилированный CpG
Неметилированный CpG
Профили метилирования в плаценте
Установлено плацентарное происхождение
Анализ генетики и эпигенетики длинных молекул ДНК
Одномолекулярное секвенирование в реальном времени
Отбор по размеру
ДНК плазмы
Материнские лейкоциты
Одномолекулярное секвенирование в реальном времени
Анализ длинных молекул ДНК
Материнские гаплотипы
например Hap I
Hap II
Геномные положения
Отцовский генотип GG в chr1:e
например, ДНК плазмы (X) с аллелем, связанным с материнским гаплотипом II
chr1:a
chr1:b
chr1:c
chr1:d
chr1:e

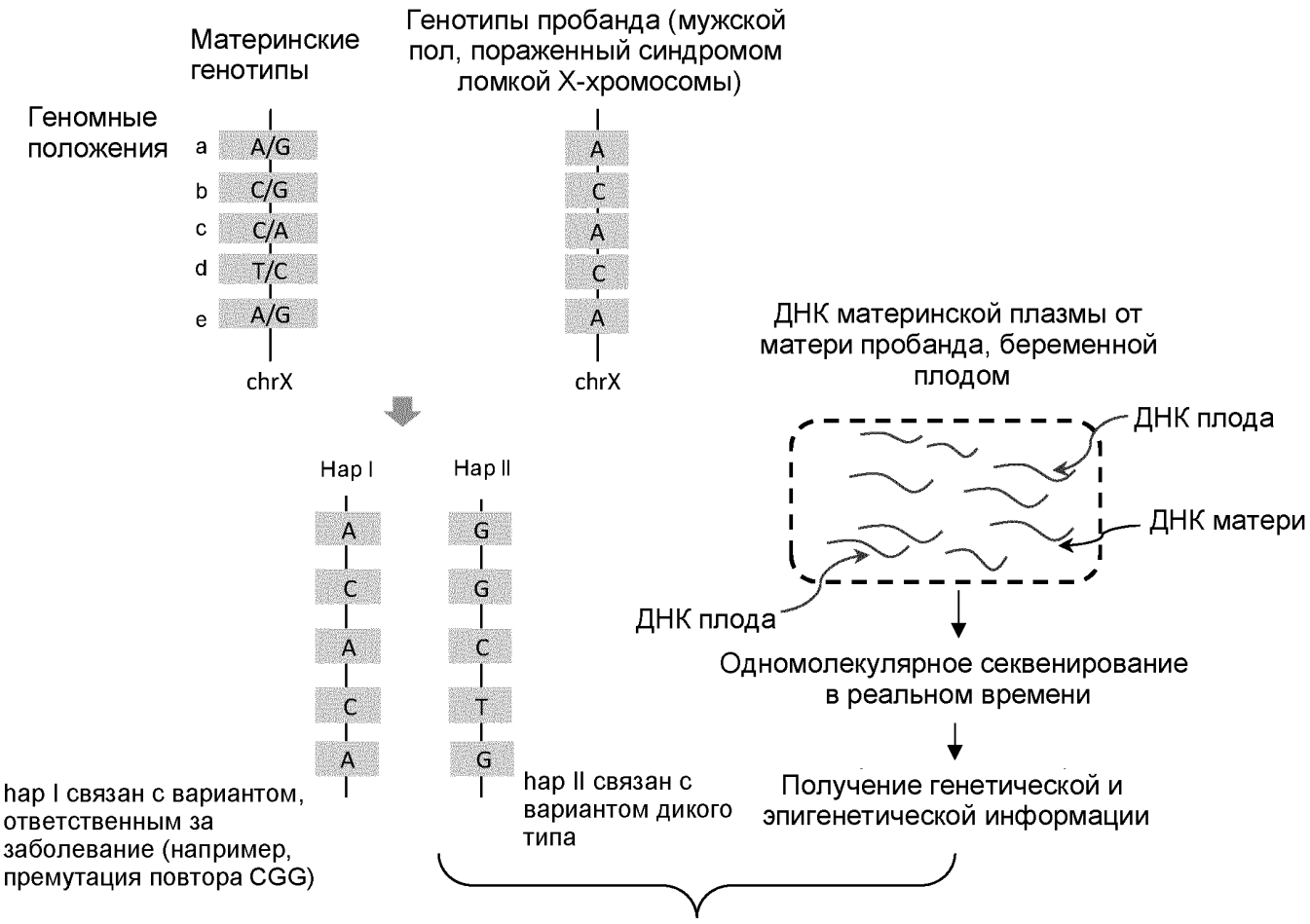
ФИГ. 30


Уровень детектирования наследования по материнской линии плода (%)
0
20
40
60
80
100
0
50
100
150
200
250
300
Профиль метилирования
RHDO
Количество молекул, использованных для анализа (миллион)

ФИГ. 31


Процентное содержание детектированных специфических для отца вариантов (%)
0
20
40
60
80
100
100
200
300
400
500
Размер ДНК плазмы
3 тыс.п.о.
1 тыс.п.о.
150 п.о.
Количество секвенированных молекул, использованных для анализа (миллион)

ФИГ. 32



Сценарий A

	Молекулы ДНК плода
Hap I	Детектируемый
Hap II	Недетектируем.

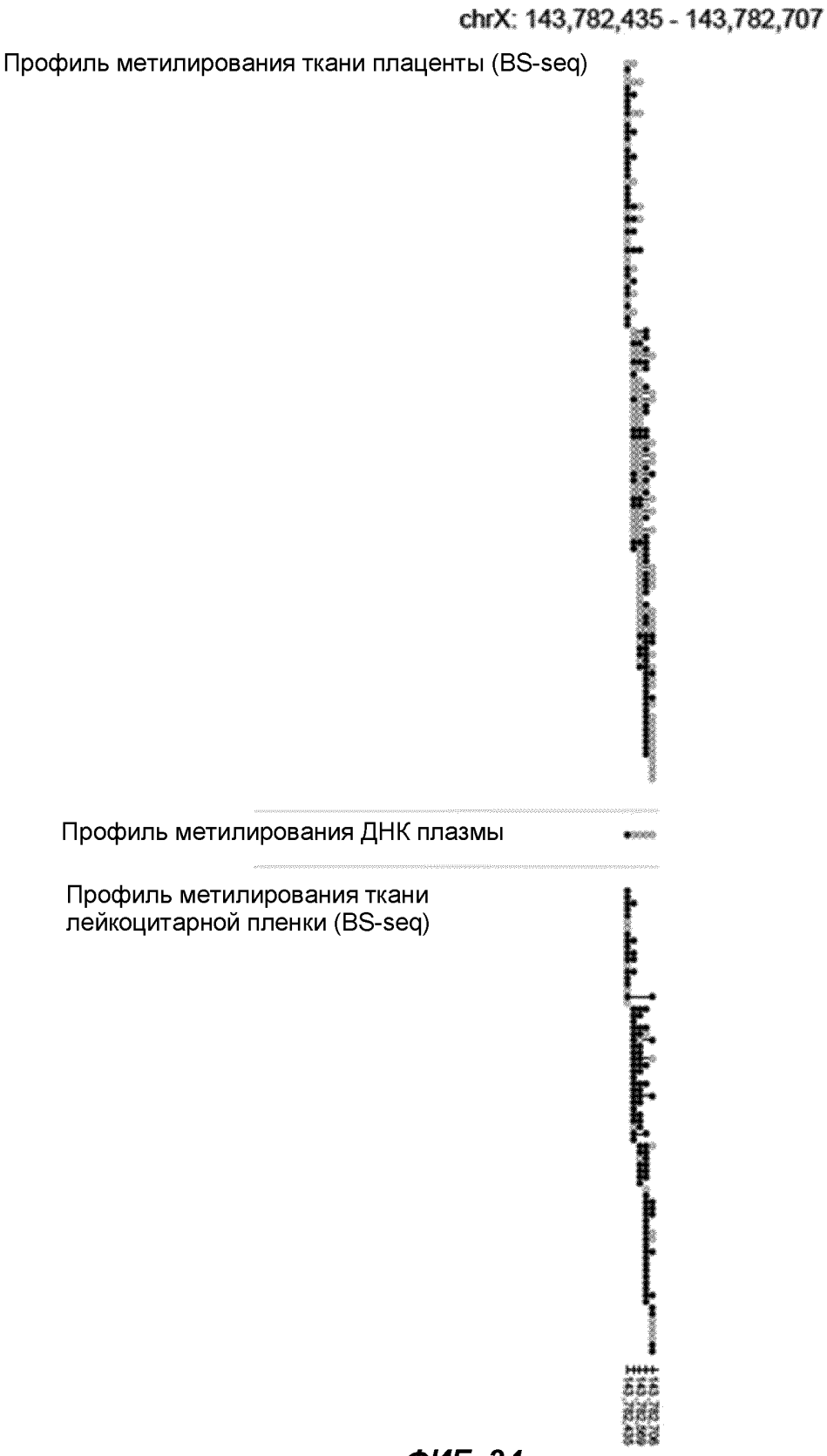
Нерожденный плод наследует материнский Hap I и, следовательно, поражен заболеванием

Сценарий B

	Молекулы ДНК плода
Hap I	Недетектируем.
Hap II	Детектируемый

Нерожденный плод наследует материнский Hap II и, следовательно, не поражен заболеванием

ФИГ. 33


chrX: 143,782,435 - 143,782,707
Профиль метилирования ткани плаценты (BS-seq)
Профиль метилирования ДНК плазмы
Профиль метилирования ткани
лейкоцитарной пленки (BS-seq)

ФИГ. 34

Кол-во сайтов CpG	Кол-во областей	Доля (%)
≥1 сайта CpG	5,333,526	86.14
≥2 сайта CpG	4,534,728	73.24
≥3 сайта CpG	3,604,867	58.22
≥4 сайта CpG	2,782,129	44.94
≥5 сайта CpG	2,141,548	34.59
≥6 сайта CpG	1,668,019	26.94
≥7 сайта CpG	1,317,812	21.28
≥8 сайта CpG	1,052,634	17.00
≥9 сайта CpG	847,632	13.69
≥10 сайта CpG	685,932	11.08

ФИГ. 35

Кол-во сайтов CpG	Кол-во областей	Доля (%)
≥1 сайта CpG	2,837,927	91.67
≥2 сайта CpG	2,752,954	88.93
≥3 сайта CpG	2,587,701	83.59
≥4 сайта CpG	2,353,438	76.02
≥5 сайта CpG	2,085,412	67.37
≥6 сайта CpG	1,817,052	58.70
≥7 сайта CpG	1,570,860	50.74
≥8 сайта CpG	1,356,211	43.81
≥9 сайта CpG	1,173,199	37.90
≥10 сайта CpG	1,018,695	32.91

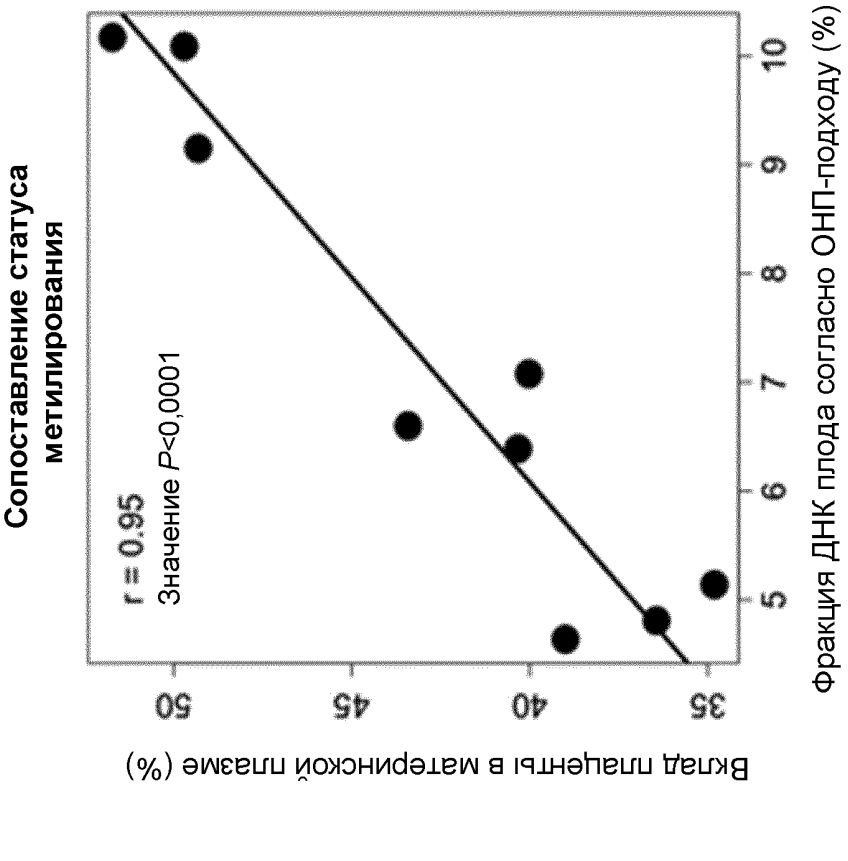
ФИГ. 36

Кол-во сайтов CpG	Кол-во областей	Доля (%)
≥1 сайта CpG	953,995	92.45
≥2 сайта CpG	953,934	92.45
≥3 сайта CpG	953,690	92.42
≥4 сайта CpG	953,021	92.36
≥5 сайта CpG	951,431	92.20
≥6 сайта CpG	948,049	91.87
≥7 сайта CpG	941,929	91.28
≥8 сайта CpG	932,143	90.33
≥9 сайта CpG	918,007	88.96
≥10 сайта CpG	898,679	87.09

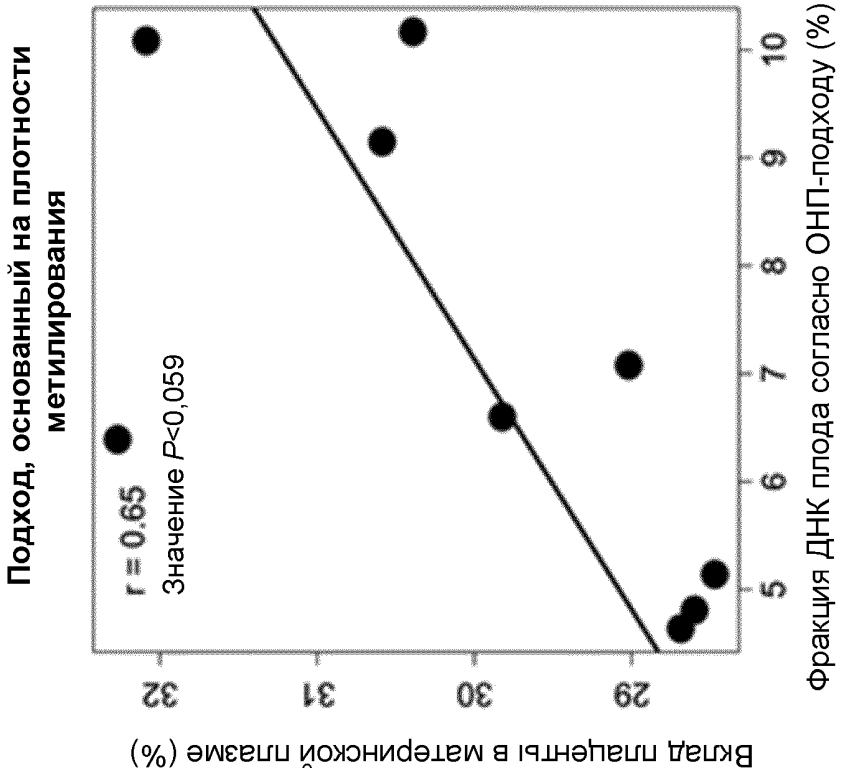
ФИГ. 37

Образец	Вклад гемопоэтических клеток (%)	Вклад печени (%)	Вклад плаценты (%)
M13323	59.31	12.1	28.6
M13324	53.12	14.61	32.27
M13315	55.15	16.16	28.69
M13318	57.26	14.28	28.47
M13211	55.4	15.58	29.02
M13230	49.41	18.51	32.09
M13304	55.03	14.57	30.39
M13199	55.86	13.55	30.59
M13198	57.72	12.46	29.82

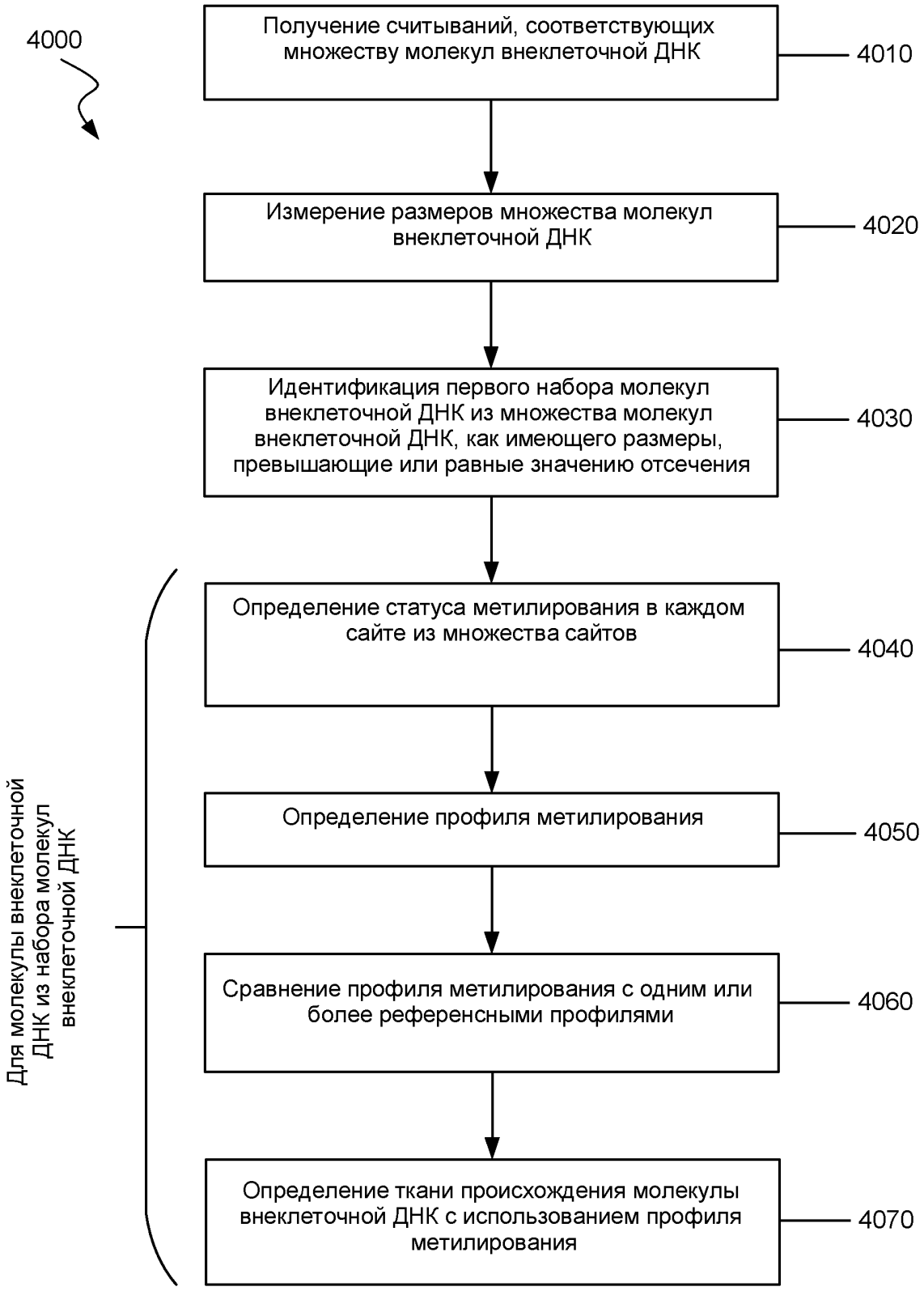
ФИГ. 38



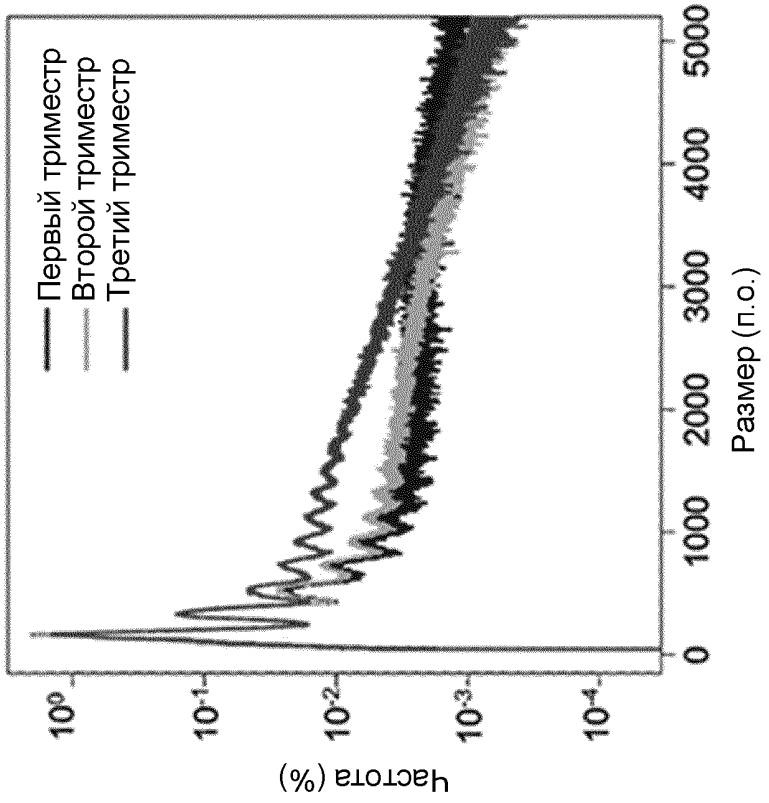
ФИГ. 39A



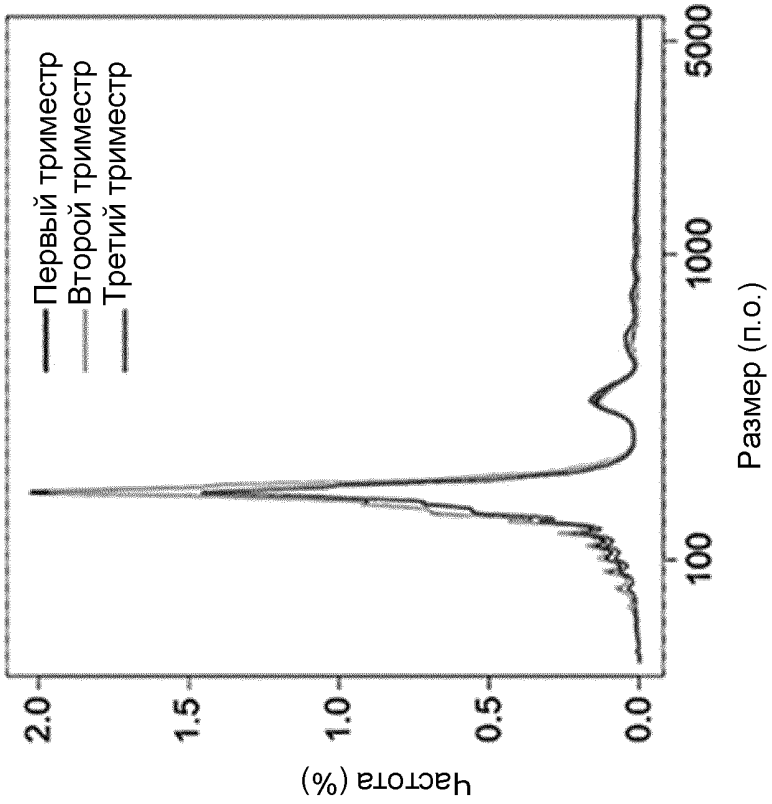
ФИГ. 39B


4000
Получение считываний, соответствующих множеству молекул внеклеточной ДНК
4010
Измерение размеров множества молекул внеклеточной ДНК
4020
Идентификация первого набора молекул внеклеточной ДНК из множества молекул внеклеточной ДНК, как имеющего размеры, превышающие или равные значению отсечения
4030
Определение статуса метилирования в каждом сайте из множества сайтов
4040
Определение профиля метилирования
4050
Сравнение профиля метилирования с одним или более референсными профилями
4060
Определение ткани происхождения молекулы внеклеточной ДНК с использованием профиля метилирования
4070
Для молекулы внеклеточной ДНК из набора молекул внеклеточной ДНК

ФИГ. 40



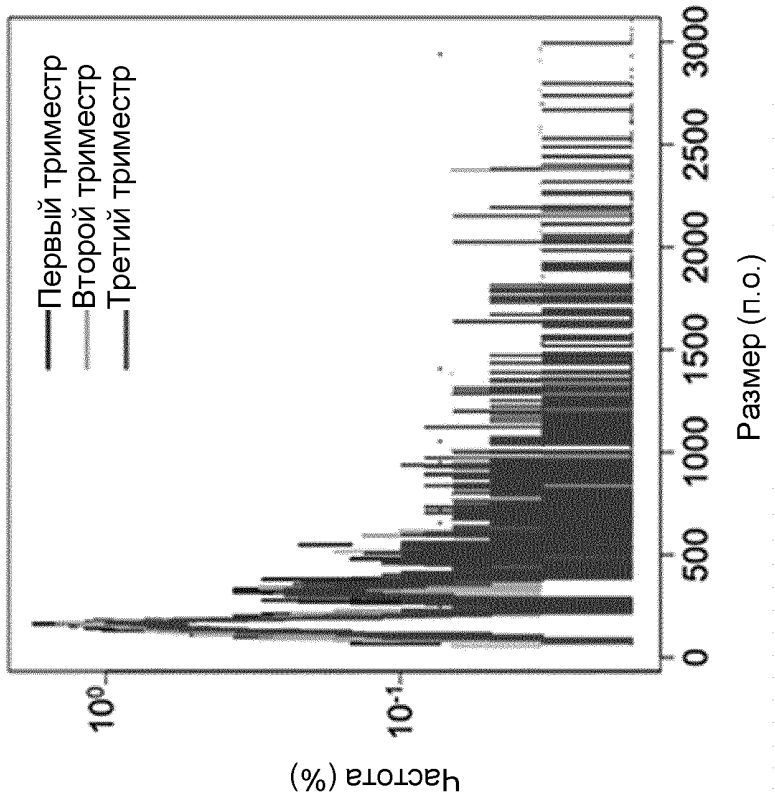
ФИГ. 41B



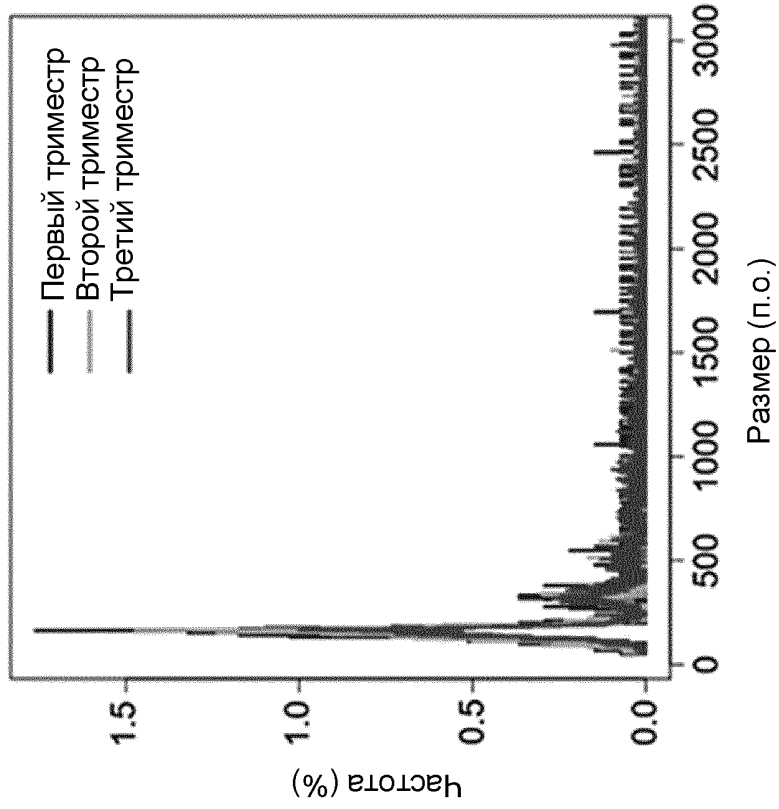
ФИГ. 41A

	Доля молекул ДНК >500 п. о. (%)	Доля молекул ДНК >1 тыс. п. о. (%)
Материнская плазма первого триместра	15.8	11.3
Материнская плазма второго триместра	16.1	10.6
Материнская плазма третьего триместра	32.3	21.4

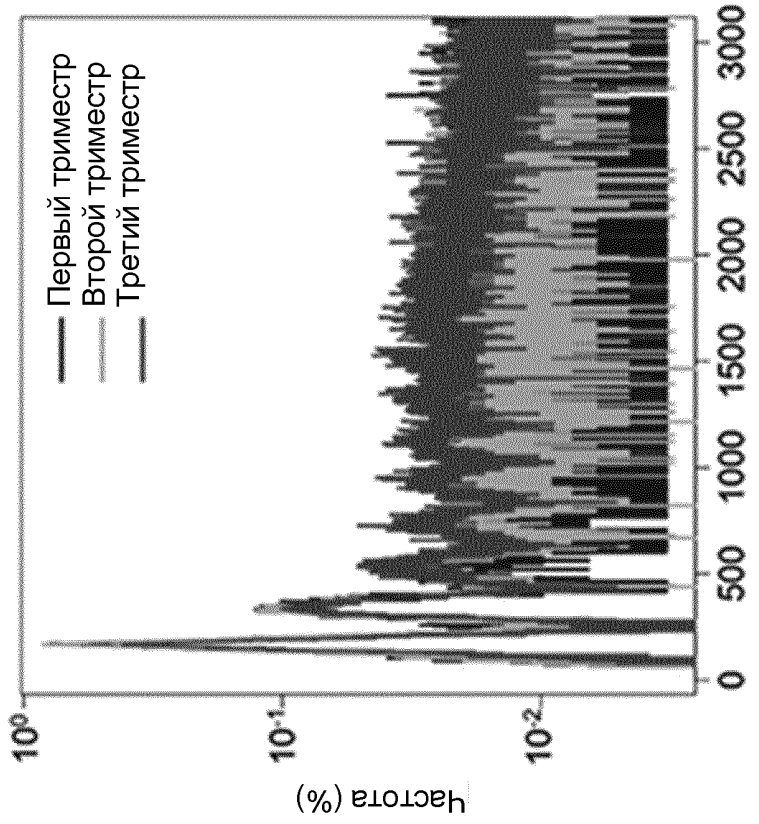
ФИГ. 42



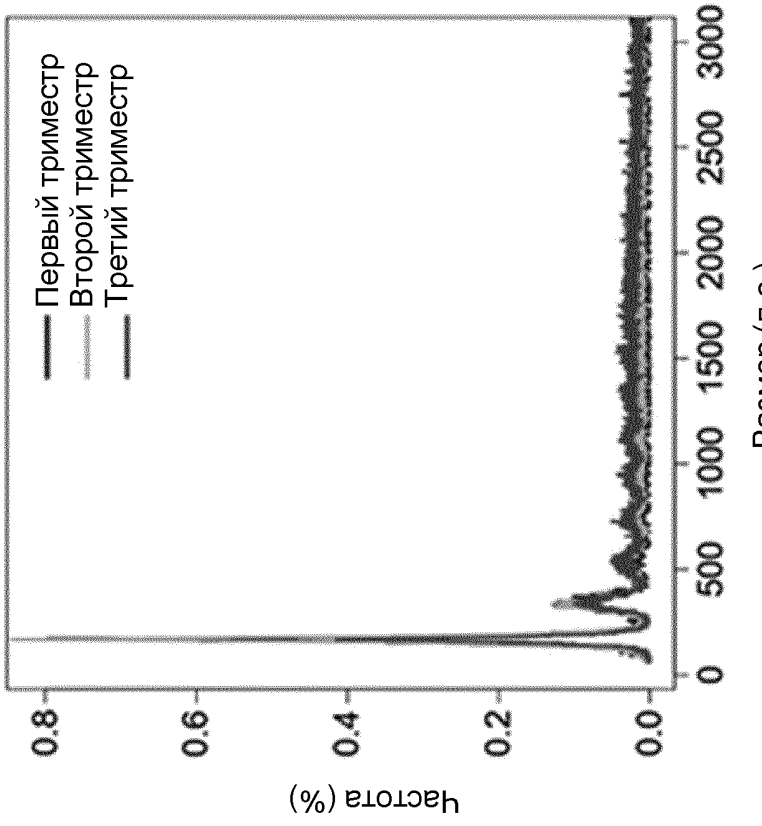
ФИГ. 43B



ФИГ. 43A



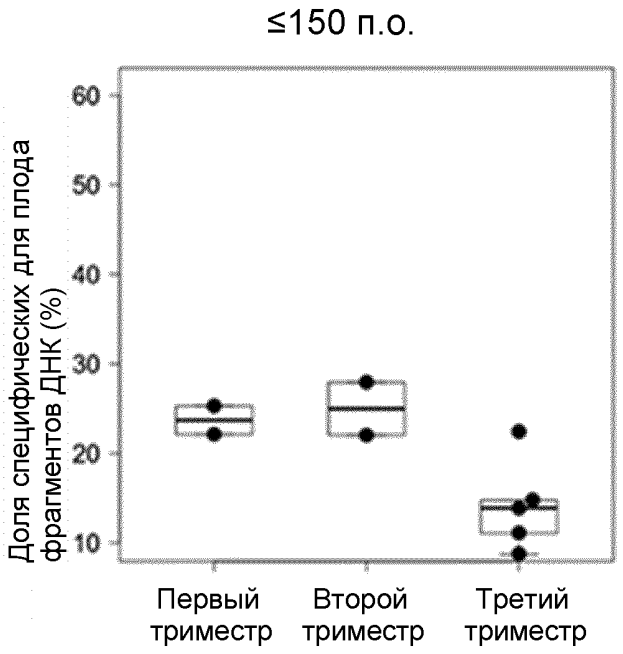
ФИГ. 44B



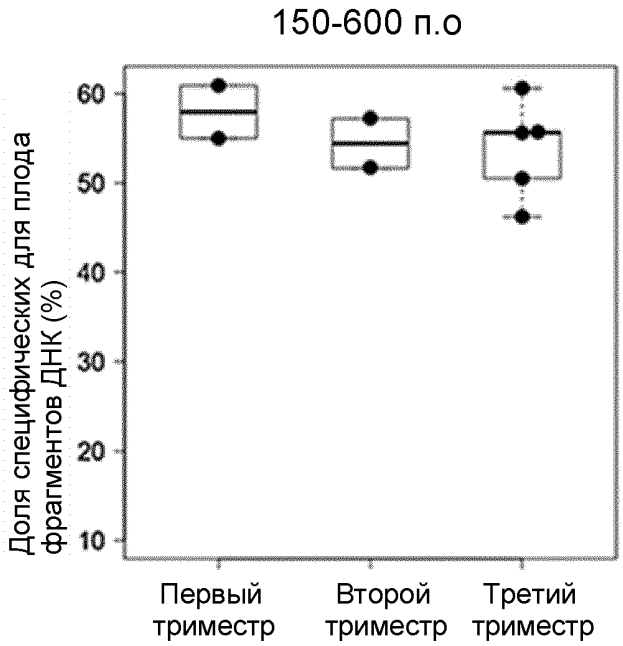
ФИГ. 44A

	Доля молекул ДНК плода >500 п. о. (%)	Доля молекул ДНК матери >500 п. о. (%)	Доля молекул ДНК плода >1 тыс. п. о. (%)	Доля молекул ДНК матери >1 тыс. п. о. (%)
Материнская плазма первого триместра	19.8	65.6	15.2	59.0
Материнская плазма второго триместра	23.2	62.6	16.5	53.9
Материнская плазма третьего триместра	31.7	76.2	19.9	64.3

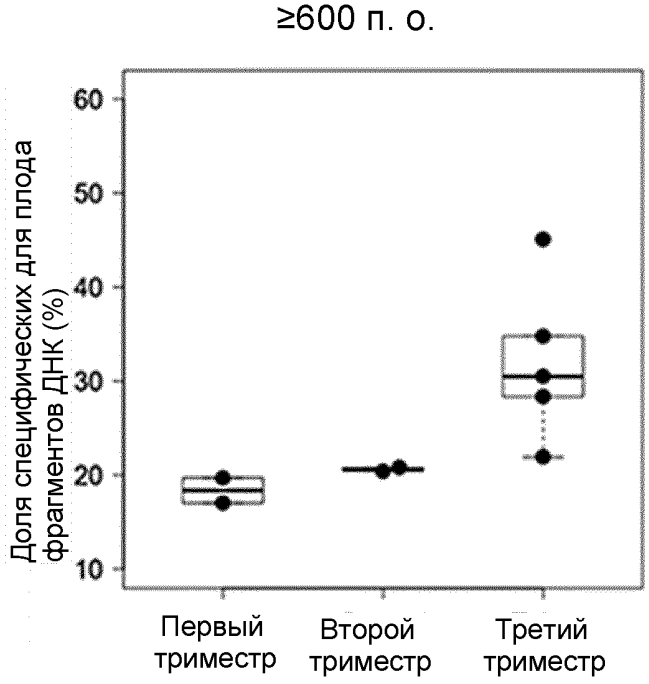
ФИГ. 45



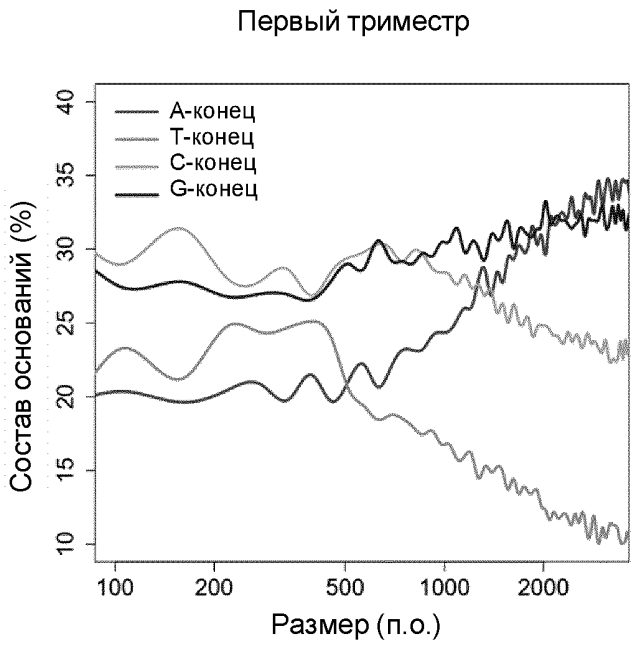
ФИГ. 46A



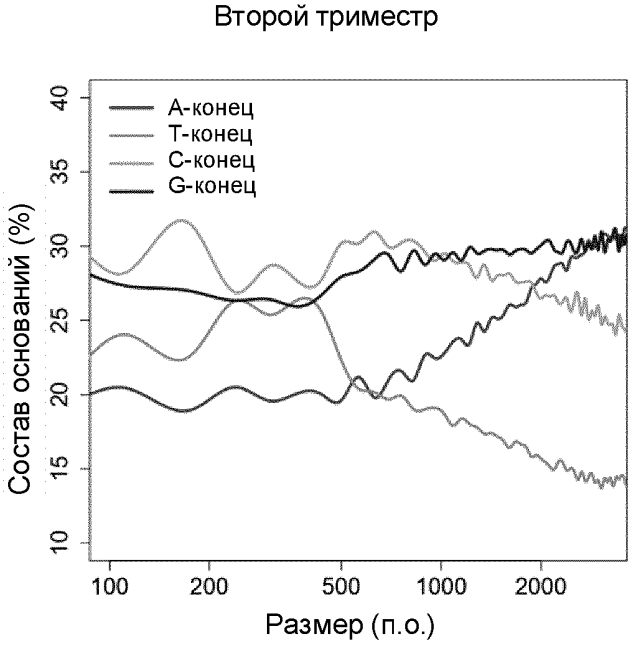
ФИГ. 46B



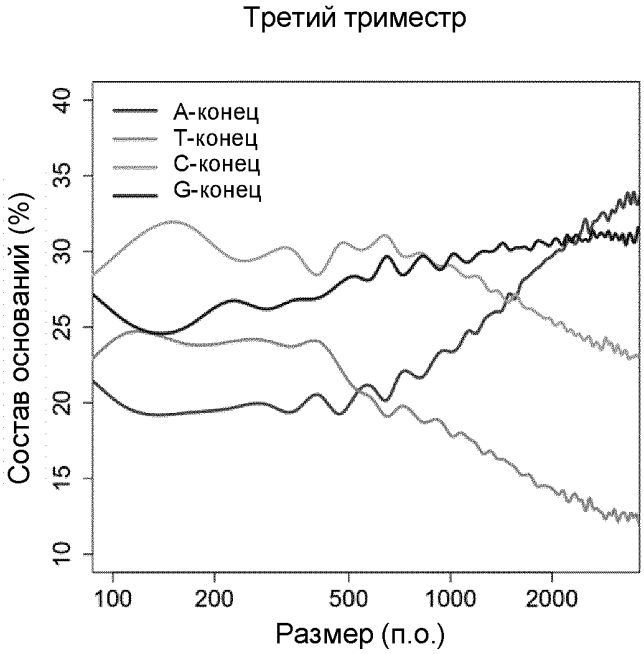
ФИГ. 46C



ФИГ. 47A



ФИГ. 47B



ФИГ. 47C

		Материнская плазма первого триместра		Материнская плазма второго триместра		Материнская плазма третьего триместра	
Концевое основание	Ожидаемая доля концевых молекул	Наблюдаемая доля концевых молекул среди фрагментов ≤ 500 п.о.	Наблюдаемая доля концевых молекул среди фрагментов > 500 п.о.	Наблюдаемая доля концевых молекул среди фрагментов ≤ 500 п.о.	Наблюдаемая доля концевых молекул среди фрагментов > 500 п.о.	Наблюдаемая доля концевых молекул среди фрагментов ≤ 500 п.о.	Наблюдаемая доля концевых молекул среди фрагментов > 500 п.о.
A-конец	29.5	19.8	29.6	19.4	26.0	19.3	26.7
T-конец	29.5	22.4	13.9	23.3	16.9	24.1	16.4
C-конец	20.5	30.4	25.5	30.4	27.5	31.3	27.1
G-конец	20.5	27.4	31.0	26.9	29.5	25.3	29.9

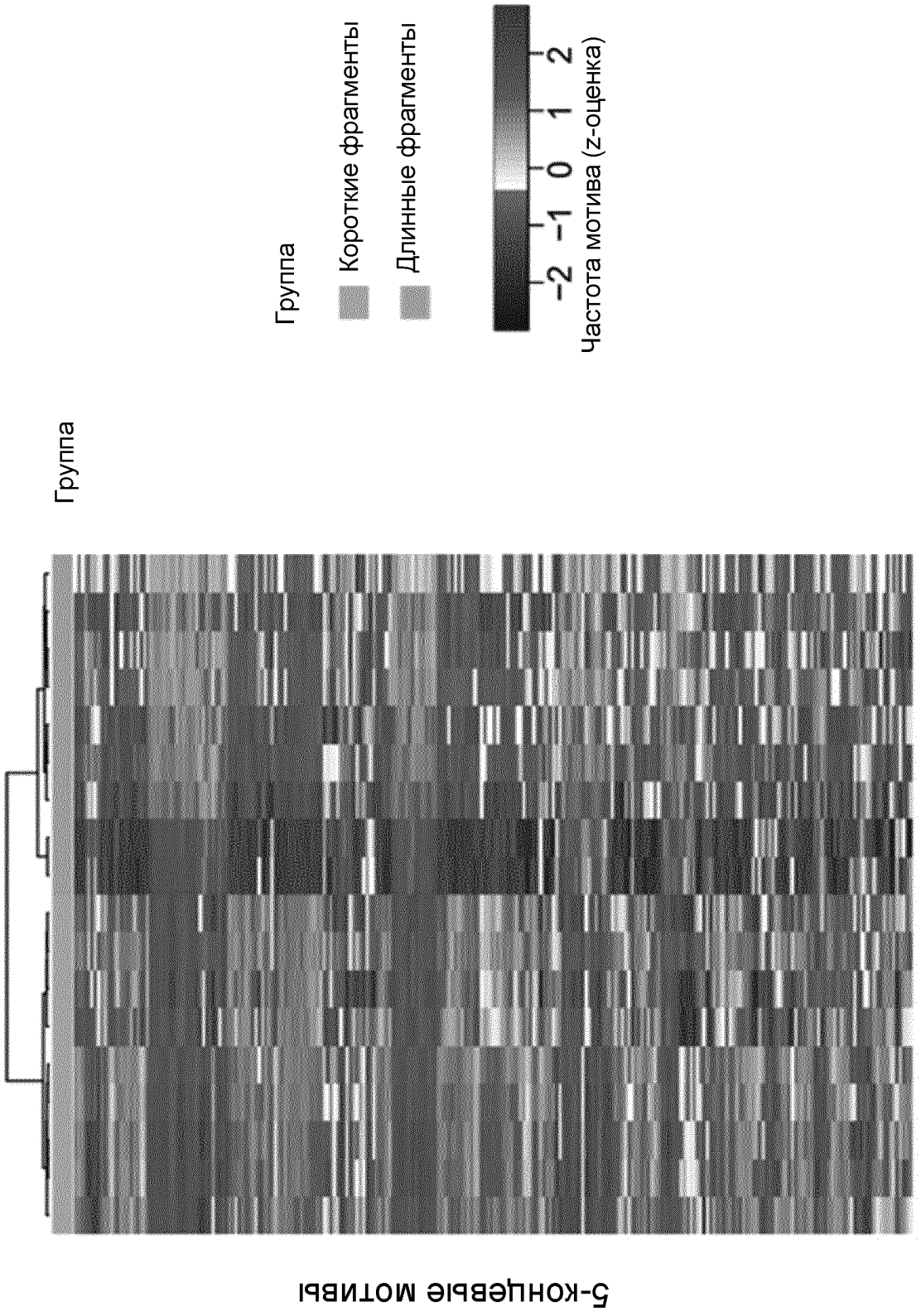
ФИГ. 48

		Материнская плазма первого триместра		Материнская плазма второго триместра		Материнская плазма третьего триместра	
Концевое основание	Ожидаемые доли концевых оснований	Доли концевых оснований среди фрагментов ≤ 500 п.о	Доли концевых оснований среди фрагментов > 500 п.о	Доли концевых оснований среди фрагментов ≤ 500 п.о	Доли концевых оснований среди фрагментов > 500 п.о	Доли концевых оснований среди фрагментов ≤ 500 п.о	Доли концевых оснований среди фрагментов > 500 п.о
A-конец	29.5	18.3	30.5	19.7	25.7	18.4	22.9
T-конец	29.5	24.2	14.7	23.2	16.0	23.0	17.7
C-конец	20.5	31.1	24.9	29.7	28.5	31.8	29.0
G-конец	20.5	26.4	29.9	27.4	29.8	26.9	30.5

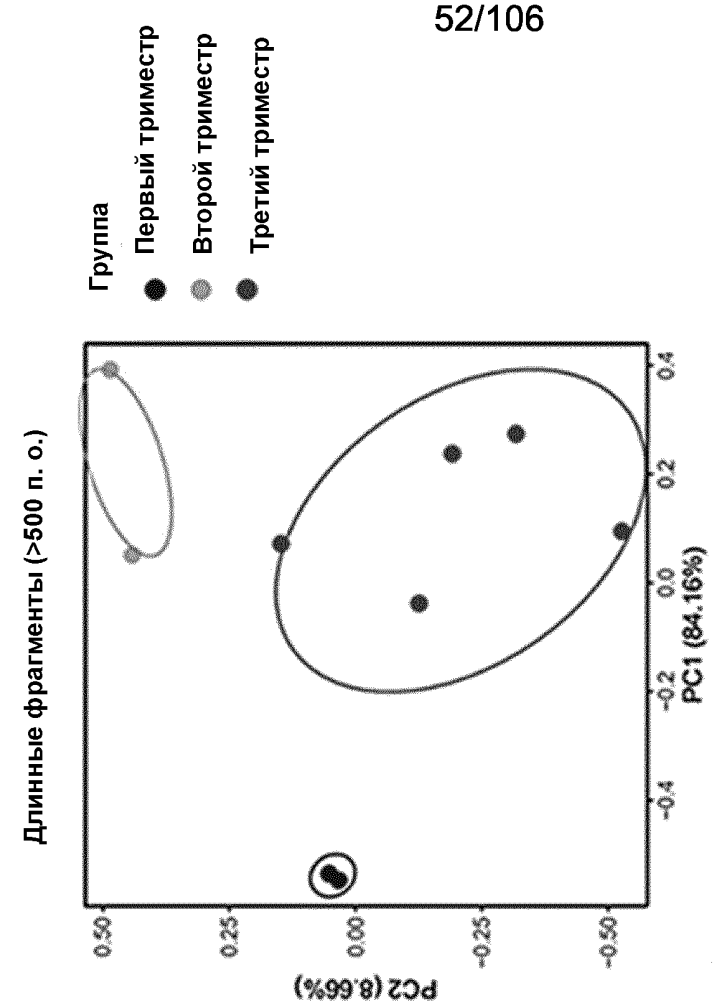
ФИГ. 49

		Материнская плазма первого триместра		Материнская плазма второго триместра		Материнская плазма третьего триместра	
Концевое основание	Ожидаемые доли концевых оснований	Доли концевых оснований среди фрагментов ≤ 500 п.о	Доли концевых оснований среди фрагментов > 500 п.о	Доли концевых оснований среди фрагментов ≤ 500 п.о	Доли концевых оснований среди фрагментов > 500 п.о	Доли концевых оснований среди фрагментов ≤ 500 п.о	Доли концевых оснований среди фрагментов > 500 п.о
A-конец	29.5	19.3	32.4	19.4	28.1	19.6	29.0
T-конец	29.5	22.0	11.4	23.0	15.1	22.2	13.5
C-конец	20.5	31.2	24.4	31.1	26.6	31.4	26.0
G-конец	20.5	27.5	31.8	26.6	30.2	26.9	31.5

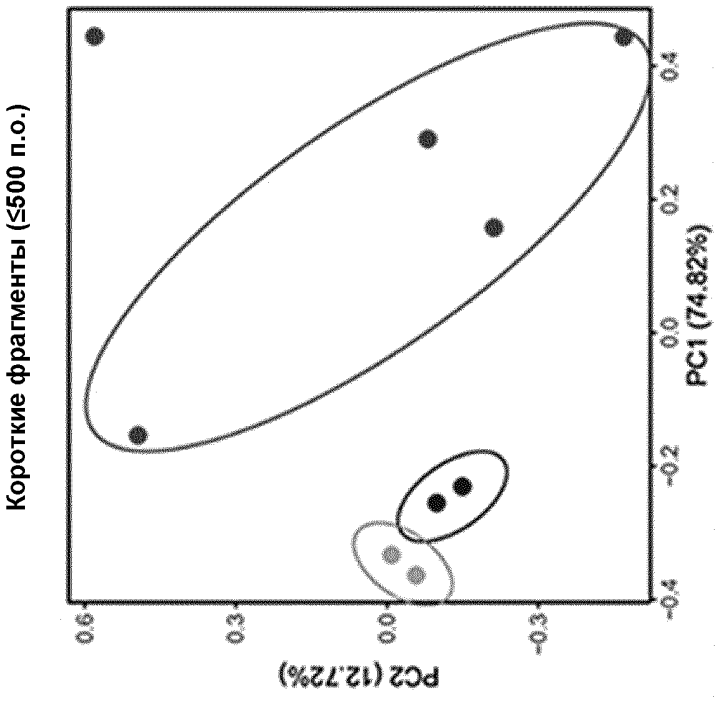
ФИГ. 50


Группа
Короткие фрагменты
Длинные фрагменты
−2 −1 0 1 2
Частота мотива (z-оценка)
Группа
5-концевые мотивы

ФИГ. 51



ФИГ. 52B



ФИГ. 52A

Мотив	Ранг в коротких фрагментах (≤500 п. о.)	Ранг в длинных фрагментах (>500 п. о.)	Частота в коротких фрагментах (%) (A)	Частота в длинных фрагментах (%) (B)	Кратное изменение (A/B)
CCCA	1	16	2.12	1.15	1.84
CCTG	2	17	1.67	1.10	1.52
CCAG	3	29	1.46	0.74	1.98
CCTC	4	7	1.35	1.43	0.94
CCTT	5	3	1.31	1.96	0.67
CCAA	6	30	1.26	0.74	1.70
AAAA	7	28	1.23	0.74	1.66
CCAC	8	38	1.07	0.68	1.57
CCCT	9	23	1.04	0.87	1.19
GCCT	10	10	1.03	1.29	0.80
CCAT	11	27	0.98	0.75	1.31
GCCA	12	33	0.97	0.72	1.35
CCCC	13	51	0.97	0.59	1.64
GCTG	14	19	0.97	1.01	0.95
GGAG	15	72	0.95	0.46	2.07
GGAA	16	49	0.93	0.61	1.52
CAAA	17	58	0.92	0.55	1.67
GAAA	18	43	0.86	0.66	1.30
TGTG	19	84	0.84	0.40	2.12
GCTT	20	2	0.83	2.30	0.36
TGTT	21	65	0.82	0.51	1.61
GGCT	22	15	0.82	1.17	0.70
CCTA	23	36	0.82	0.69	1.18
GGTG	24	48	0.81	0.61	1.33
CACA	25	40	0.81	0.67	1.20

ФИГ. 53

Мотив	Ранг в коротких фрагментах (≤500 п. о.)	Ранг в длинных фрагментах (>500 п. о.)	Частота в коротких фрагментах (%) (A)	Частота в длинных фрагментах (%) (B)	Кратное изменение (A/B)
CCCA	1	4	2.12	1.49	1.42
CCTG	2	7	1.69	1.30	1.30
CCAG	3	13	1.49	1.03	1.44
CCTC	4	5	1.28	1.33	0.96
CCTT	5	2	1.28	1.73	0.74
CCAA	6	16	1.26	0.97	1.31
AAAA	7	26	1.26	0.84	1.50
CCAC	8	28	1.07	0.82	1.31
CCCT	9	19	1.06	0.92	1.15
CCAT	10	22	1.00	0.88	1.14
GCCT	11	8	0.99	1.16	0.85
GGAG	12	46	0.97	0.63	1.54
CCCC	13	38	0.97	0.68	1.42
GCCA	14	25	0.97	0.86	1.12
CAAA	15	35	0.95	0.71	1.34
GCTG	16	14	0.94	1.01	0.93
GGAA	17	34	0.92	0.71	1.30
GAAA	18	31	0.89	0.75	1.18
TGTG	19	55	0.88	0.56	1.57
TGTT	20	48	0.85	0.61	1.40
CACA	21	32	0.82	0.75	1.09
CCTA	22	29	0.82	0.78	1.04
GCAG	23	47	0.81	0.62	1.32
TGAG	24	75	0.81	0.45	1.79
GGTG	25	40	0.80	0.66	1.21

ФИГ. 54

Мотив	Ранг в коротких фрагментах (≤500 п. о.)	Ранг в длинных фрагментах (>500 п. о.)	Частота в коротких фрагментах (%) (A)	Частота в длинных фрагментах (%) (B)	Кратное изменение (A/B)
CCCA	1	6	2.21	1.40	1.58
CCTG	2	7	1.80	1.28	1.40
CCAG	3	17	1.53	0.98	1.57
CCTC	4	4	1.39	1.40	0.99
CCTT	5	3	1.28	1.68	0.77
CCAA	6	22	1.26	0.89	1.42
CCAC	7	26	1.13	0.80	1.41
AAAA	8	29	1.12	0.76	1.47
CCCT	9	25	1.06	0.84	1.26
GCCA	10	24	1.05	0.87	1.21
GCCT	11	8	1.04	1.23	0.85
GGAG	12	46	1.02	0.63	1.63
CCCC	13	44	1.00	0.64	1.57
GCTG	14	14	0.98	1.03	0.95
CCAT	15	27	0.97	0.78	1.24
GGAA	16	35	0.95	0.72	1.32
GGCT	17	12	0.87	1.09	0.80
GGTG	18	38	0.87	0.70	1.24
GGCA	19	33	0.86	0.72	1.19
CAAA	20	42	0.86	0.64	1.33
TGTG	21	62	0.85	0.51	1.67
GAAA	22	34	0.83	0.72	1.15
GCTT	23	2	0.82	1.88	0.44
CACA	24	31	0.82	0.74	1.11
CCTA	25	32	0.82	0.74	1.11

ФИГ. 55

Мотив	Ранг в длинных фрагментах (>500 п. о.)	Ранг в коротких фрагментах (≤500 п. о.)	Частота в длинных фрагментах (%) (A)	Частота в коротких фрагментах (%) (B)	Кратное измене-ние (A/B)
ACTT	1	39	2.99	0.63	4.76
GCTT	2	20	2.30	0.83	2.76
CCTT	3	5	1.96	1.31	1.50
GTTT	4	40	1.77	0.63	2.82
ACCT	5	56	1.48	0.56	2.63
ACTG	6	53	1.47	0.57	2.57
CCTC	7	4	1.43	1.35	1.06
ACTC	8	76	1.34	0.48	2.79
GATT	9	83	1.29	0.46	2.81
GCCT	10	10	1.29	1.03	1.26
CTTT	11	31	1.24	0.69	1.79
AGTT	12	102	1.21	0.41	2.99
CATT	13	32	1.21	0.69	1.75
ACAT	14	75	1.17	0.48	2.42
GGCT	15	22	1.17	0.82	1.43
CCCA	16	1	1.15	2.12	0.54
CCTG	17	2	1.10	1.67	0.66
GGTT	18	36	1.09	0.65	1.69
GCTG	19	14	1.01	0.97	1.05
GCTC	20	49	0.99	0.58	1.70
ATTT	21	139	0.98	0.30	3.23
ACCA	22	43	0.97	0.61	1.60
CCCT	23	9	0.87	1.04	0.84
GAAT	24	34	0.82	0.67	1.23
CTTC	25	80	0.81	0.47	1.74

ФИГ. 56

Мотив	Ранг в длинных фрагментах (>500 п. о.)	Ранг в коротких фрагментах (≤500 п. о.)	Частота в длинных фрагментах (%) (A)	Частота в коротких фрагментах (%) (B)	Кратное изменение (A/B)
ACTT	1	64	2.01	0.54	3.75
CCTT	2	5	1.73	1.28	1.35
GCTT	3	29	1.68	0.76	2.20
CCCA	4	1	1.49	2.12	0.70
CCTC	5	4	1.33	1.28	1.04
GTTT	6	46	1.32	0.59	2.25
CCTG	7	2	1.30	1.69	0.77
GCCT	8	11	1.16	0.99	1.17
ACCT	9	66	1.15	0.53	2.17
ACTG	10	63	1.09	0.54	2.02
CTTT	11	32	1.07	0.69	1.56
CATT	12	31	1.05	0.69	1.52
CCAG	13	3	1.03	1.49	0.69
GCTG	14	16	1.01	0.94	1.07
ACTC	15	94	1.00	0.43	2.32
CCAA	16	6	0.97	1.26	0.77
GGCT	17	26	0.96	0.79	1.21
GATT	18	89	0.93	0.43	2.13
CCCT	19	9	0.92	1.06	0.87
ACAT	20	79	0.92	0.47	1.98
ACCA	21	44	0.88	0.60	1.47
CCAT	22	10	0.88	1.00	0.88

ФИГ. 57

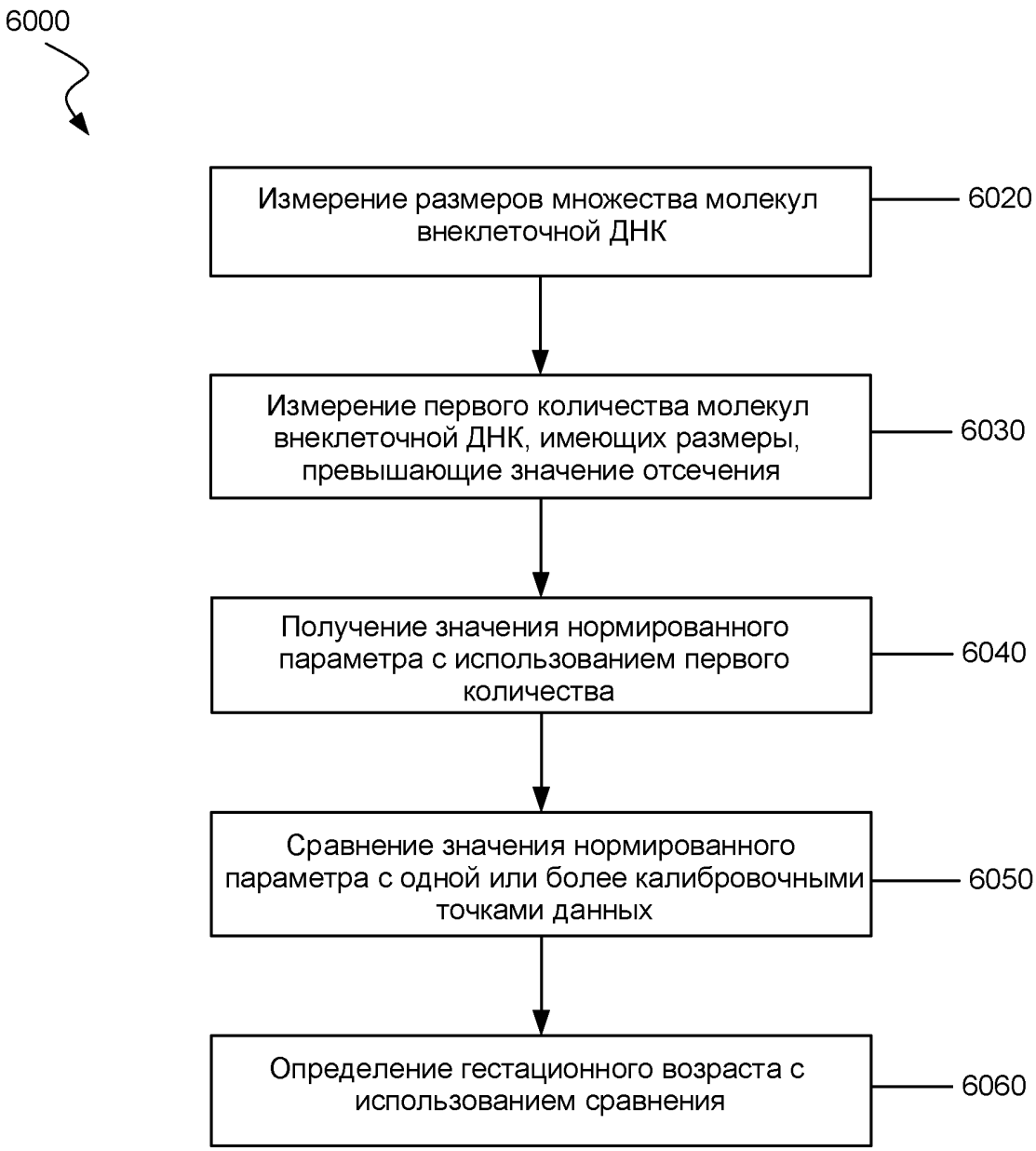
Мотив	Ранг в длинных фрагментах (>500 п. о.)	Ранг в коротких фрагментах (≤500 п. о.)	Частота в длинных фрагментах (%) (A)	Частота в коротких фрагментах (%) (A)	Кратное изменение (A/B)
ACTT	1	57	2.06	0.57	3.61
GCTT	2	23	1.88	0.82	2.29
CCTT	3	5	1.68	1.28	1.30
CCTC	4	4	1.40	1.39	1.01
GTTT	5	49	1.40	0.59	2.40
CCCA	6	1	1.40	2.21	0.63
CCTG	7	2	1.28	1.80	0.71
GCCT	8	11	1.23	1.04	1.18
ACTG	9	47	1.20	0.59	2.05
ACCT	10	59	1.18	0.56	2.09
ACTC	11	78	1.12	0.48	2.36
GGCT	12	17	1.09	0.87	1.25
CTTT	13	41	1.04	0.61	1.71
GCTG	14	14	1.03	0.98	1.05
CATT	15	34	1.00	0.64	1.57
GATT	16	92	0.99	0.43	2.30
CCAG	17	3	0.98	1.53	0.64
GGTT	18	33	0.95	0.64	1.49
ACAT	19	80	0.92	0.46	1.99
GCTC	20	40	0.92	0.61	1.50
AGTT	21	115	0.91	0.37	2.46
CCAA	22	6	0.89	1.26	0.70

ФИГ. 58

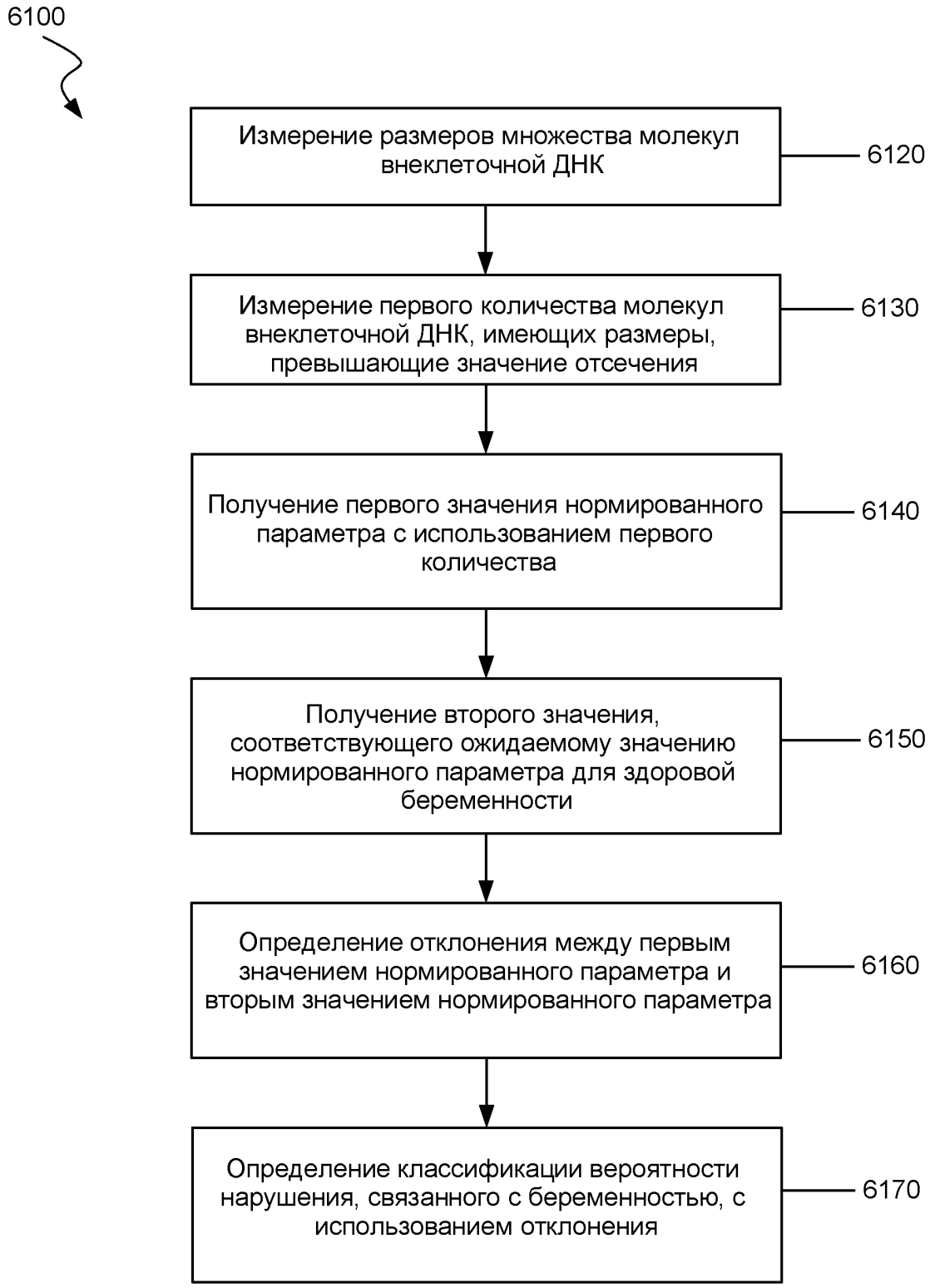
ФИГ. 59A

ФИГ. 59B

ФИГ. 59C


6000
Измерение размеров множества молекул внеклеточной ДНК
6020
Измерение первого количества молекул внеклеточной ДНК, имеющих размеры, превышающие значение отсечения
6030
Получение значения нормированного параметра с использованием первого количества
6040
Сравнение значения нормированного параметра с одной или более калибровочными точками данных
6050
Определение гестационного возраста с использованием сравнения
6060

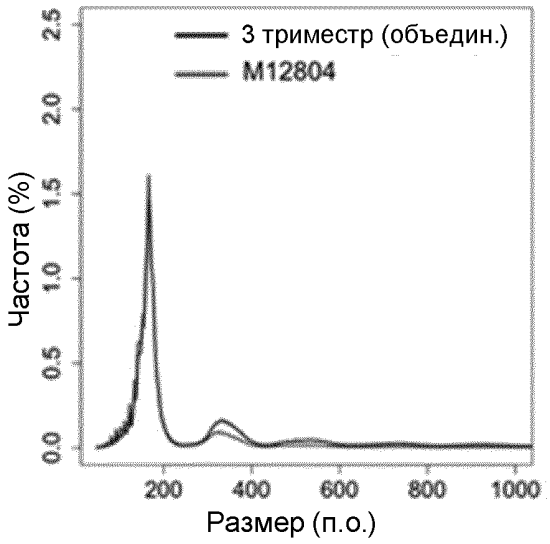
ФИГ. 60


6100
Измерение размеров множества молекул внеклеточной ДНК
6120
Измерение первого количества молекул внеклеточной ДНК, имеющих размеры, превышающие значение отсечения
6130
Получение первого значения нормированного параметра с использованием первого количества
6140
Получение второго значения, соответствующего ожидаемому значению нормированного параметра для здоровой беременности
6150
Определение отклонения между первым значением нормированного параметра и вторым значением нормированного параметра
6160
Определение классификации вероятности нарушения, связанного с беременностью, с использованием отклонения
6170

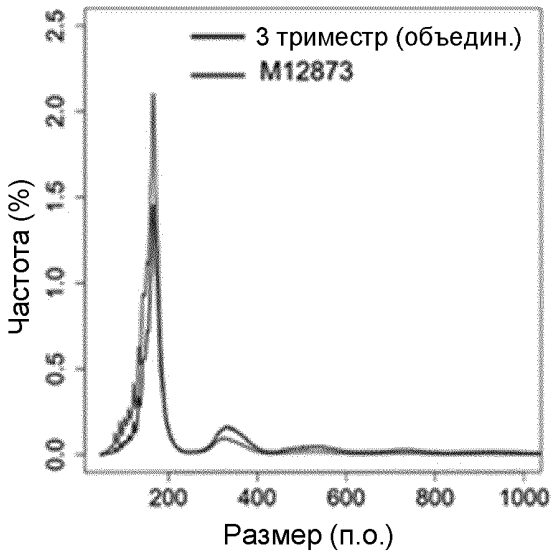
ФИГ. 61

№ истории болезни	Гестационный возраст на момент забора крови (недели)	Пол плода	Клиническая информация
M12804	34 3/7	F	Тяжелая PET с существовавшей ранее IgA-нефропатией
M12873	37	M	Хроническая гипертензия с наложением легкой PET
M12876	36	F	Тяжелая PET с поздним началом
M12903	35 5/7	M	Тяжелая PET с поздним началом с ЗВУР

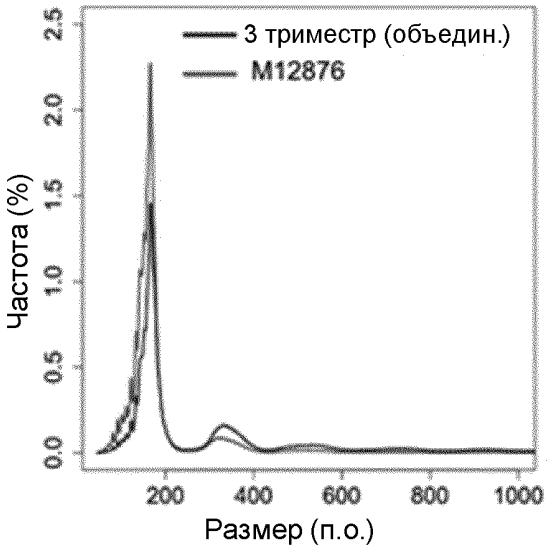
ФИГ. 62



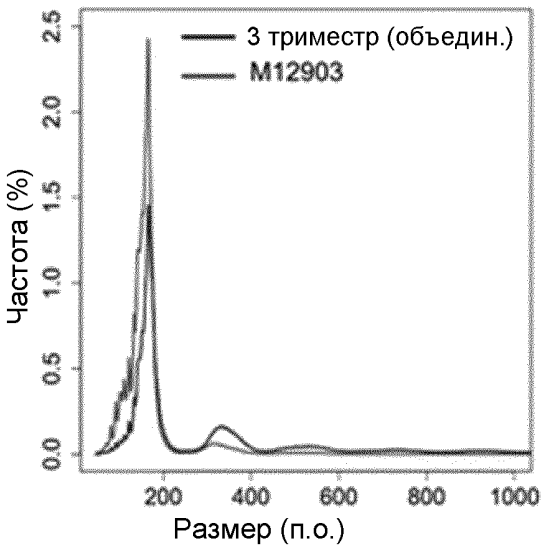
ФИГ. 63A



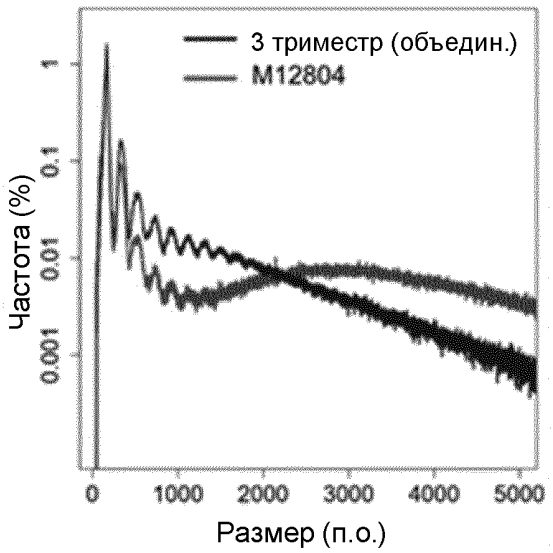
ФИГ. 63B



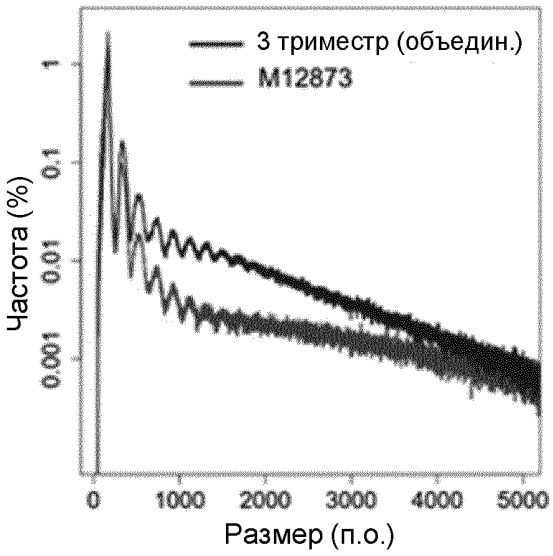
ФИГ. 63C



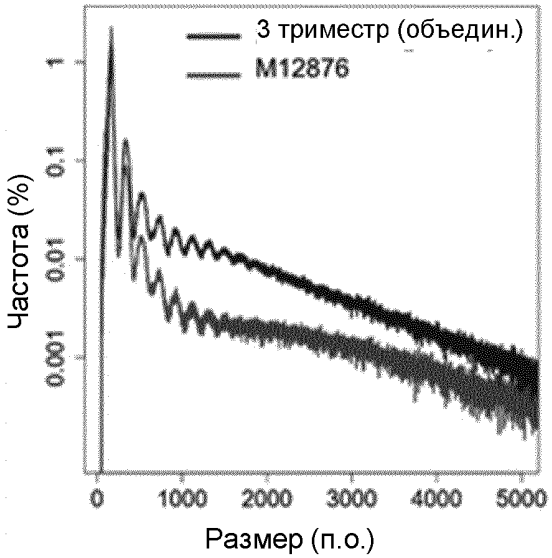
ФИГ. 63D



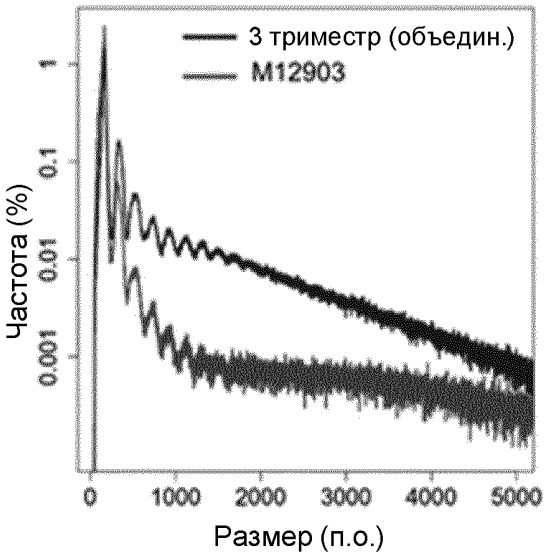
ФИГ. 64A



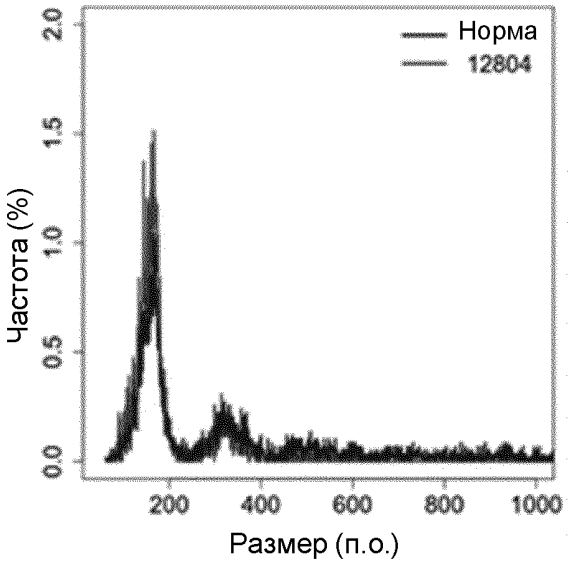
ФИГ. 64B



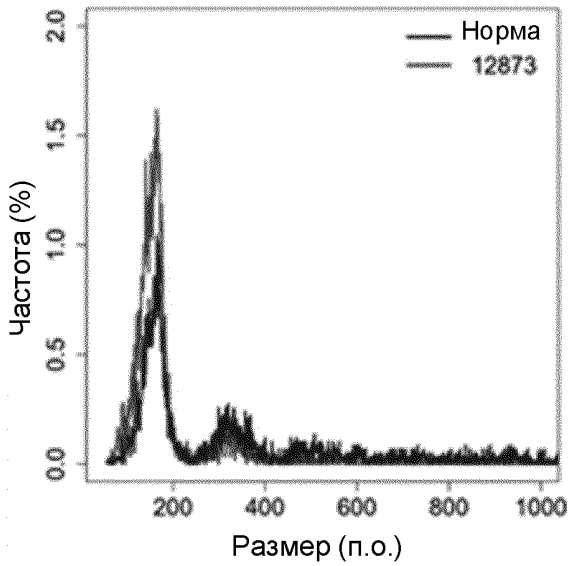
ФИГ. 64C



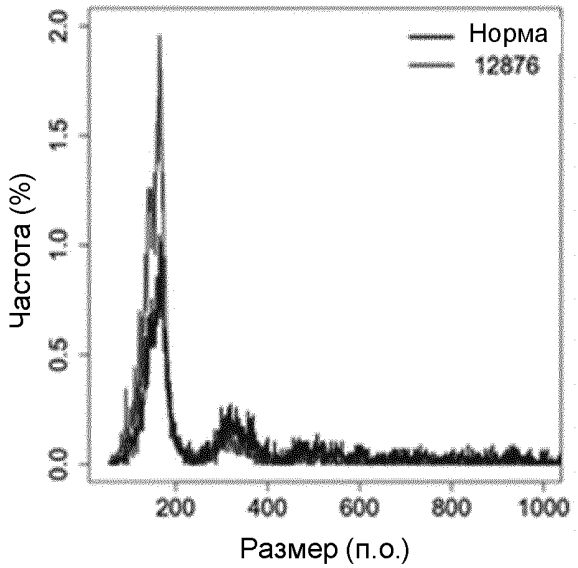
ФИГ. 64D



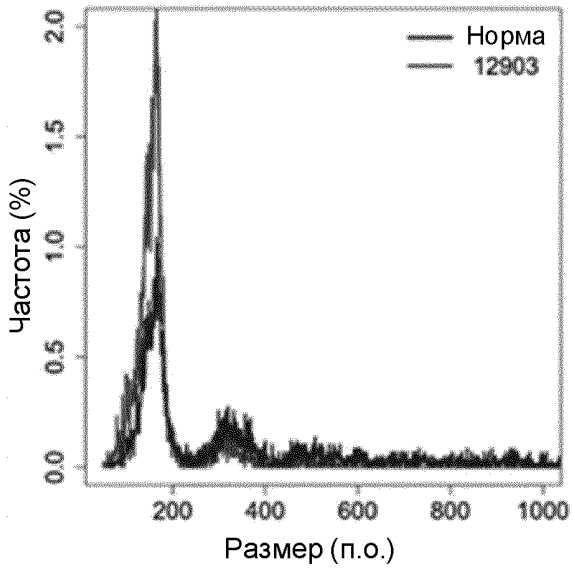
ФИГ. 65A



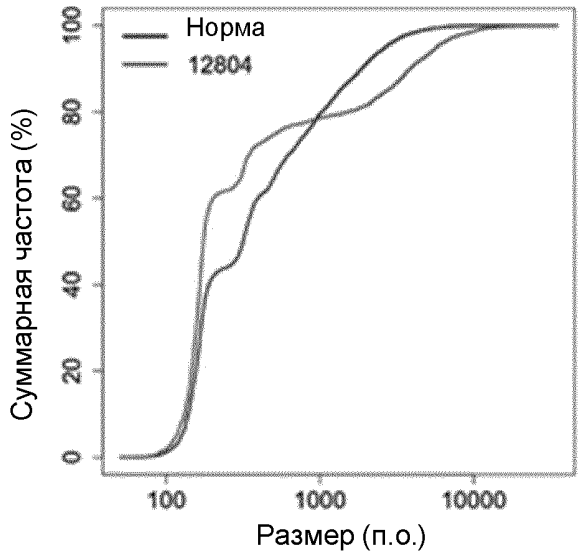
ФИГ. 65B



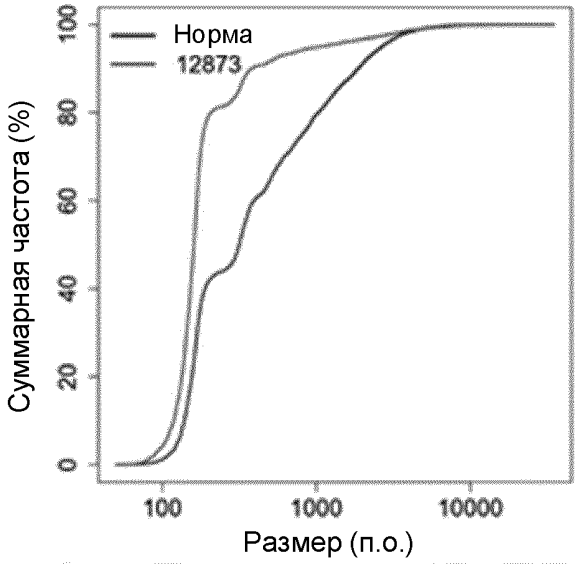
ФИГ. 65C



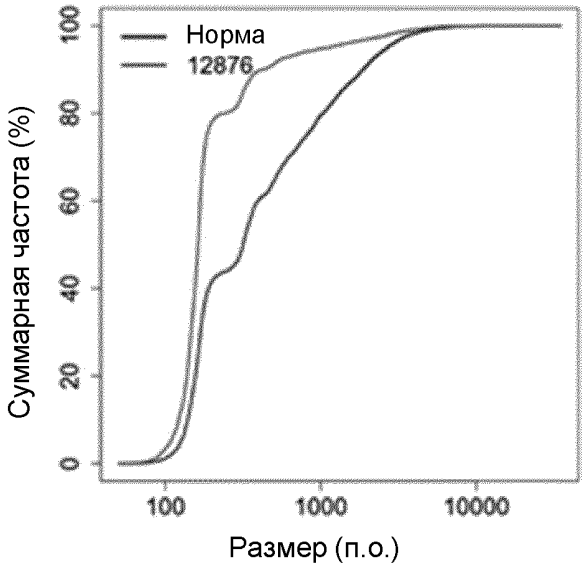
ФИГ. 65D



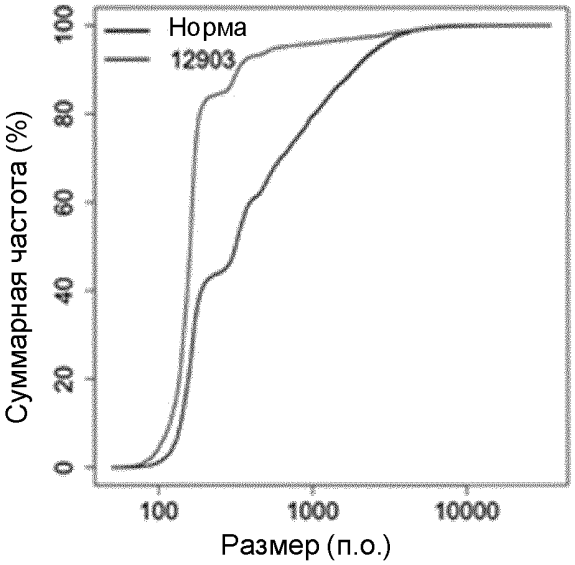
ФИГ. 66A



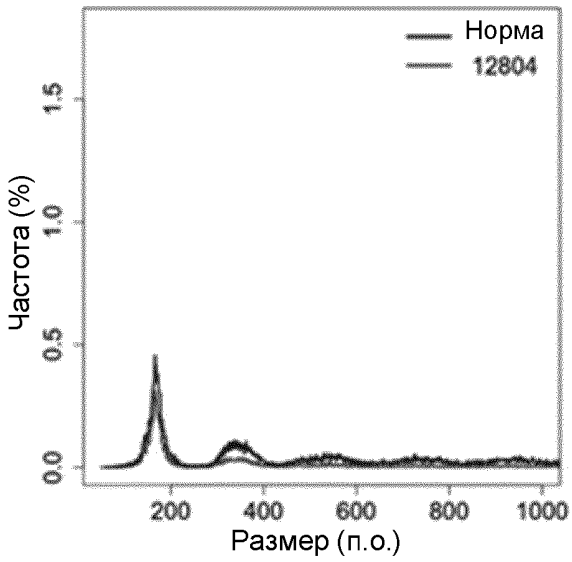
ФИГ. 66B



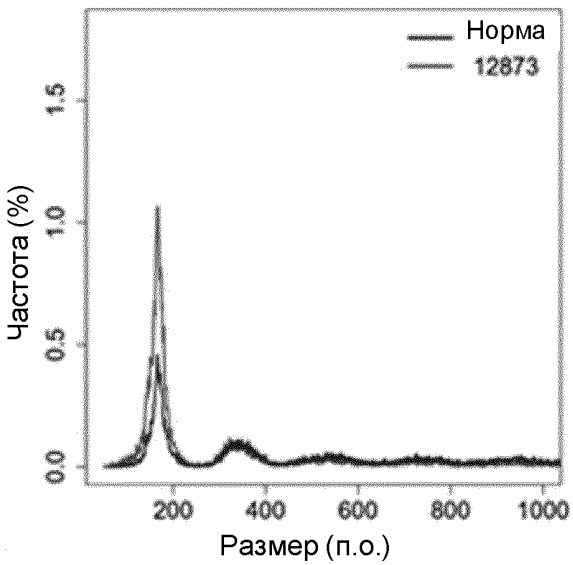
ФИГ. 66C



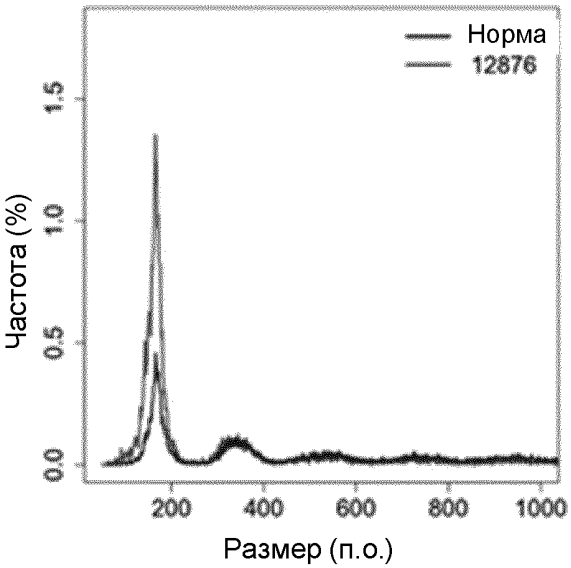
ФИГ. 66D



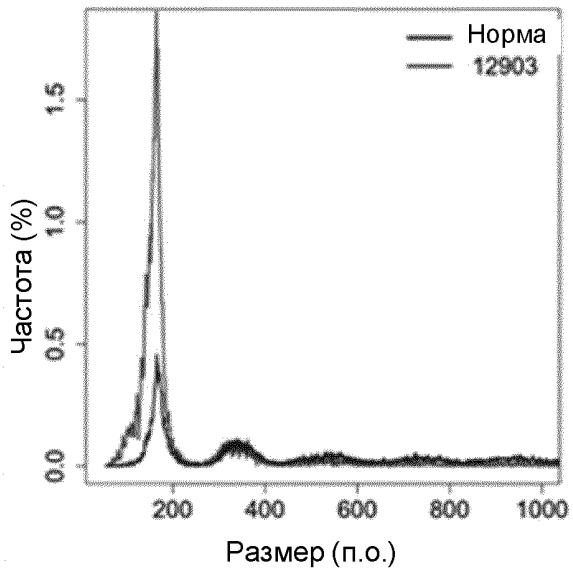
ФИГ. 67A



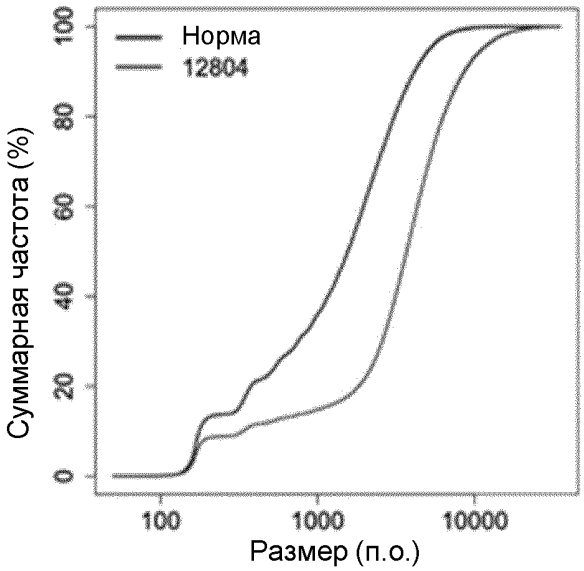
ФИГ. 67B



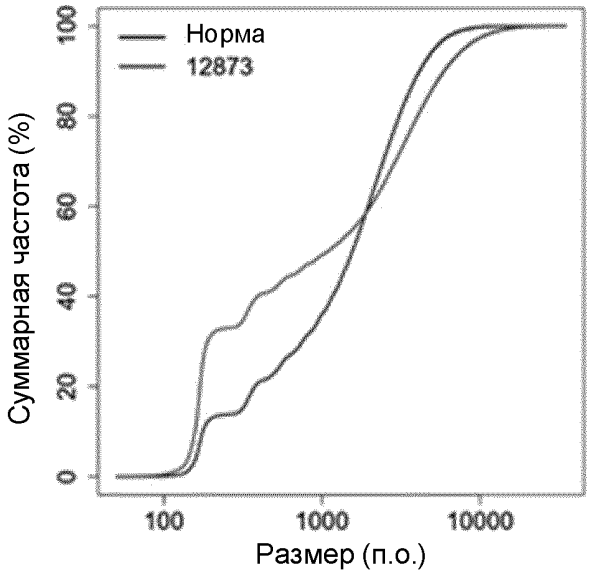
ФИГ. 67C



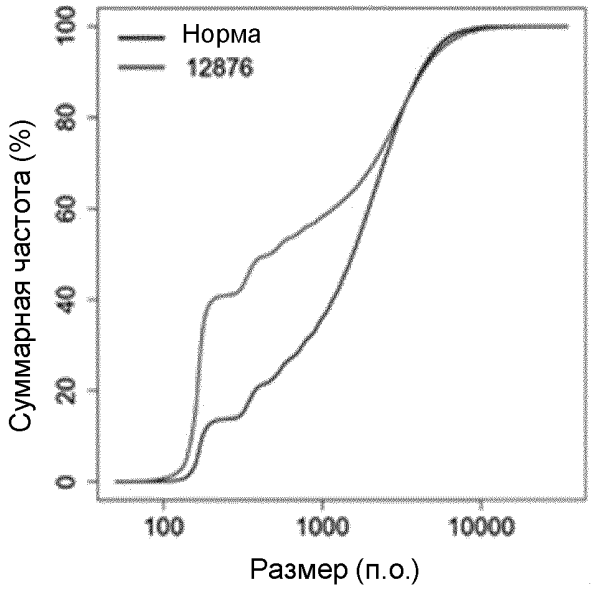
ФИГ. 67D



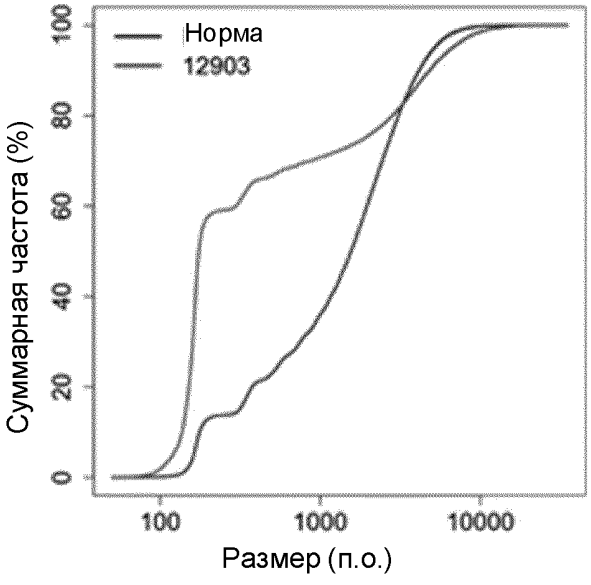
ФИГ. 68A



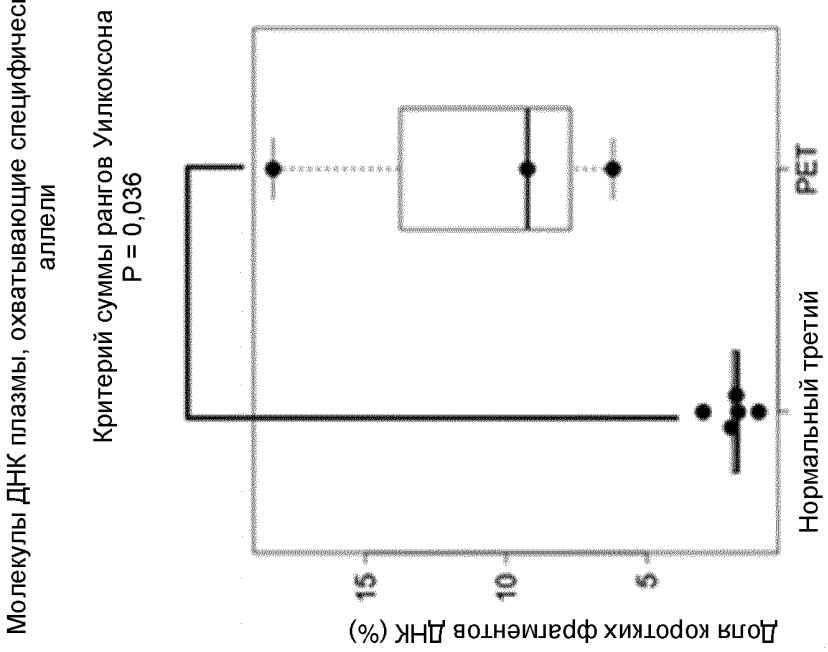
ФИГ. 68B



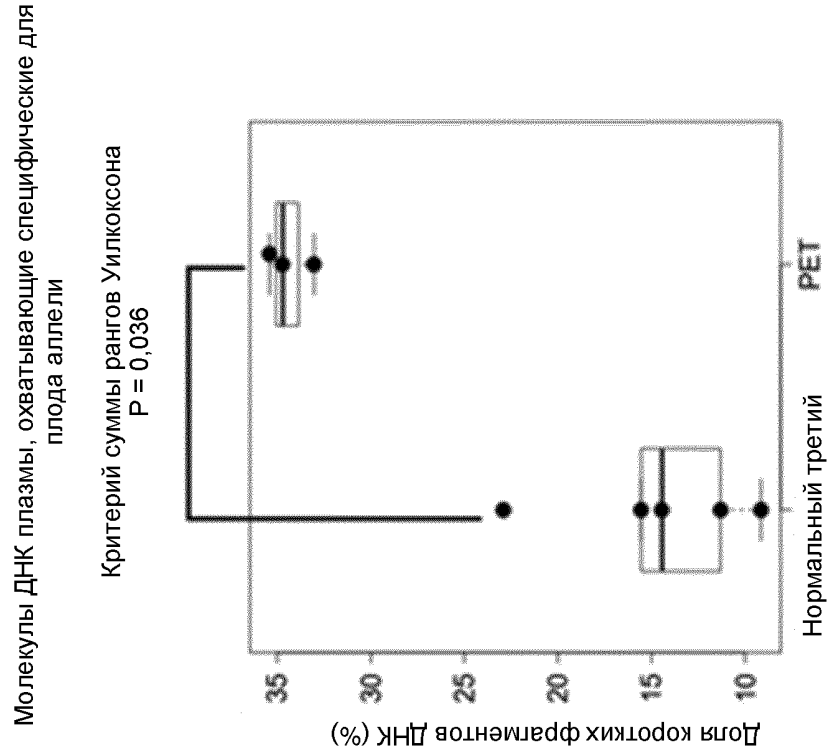
ФИГ. 68C



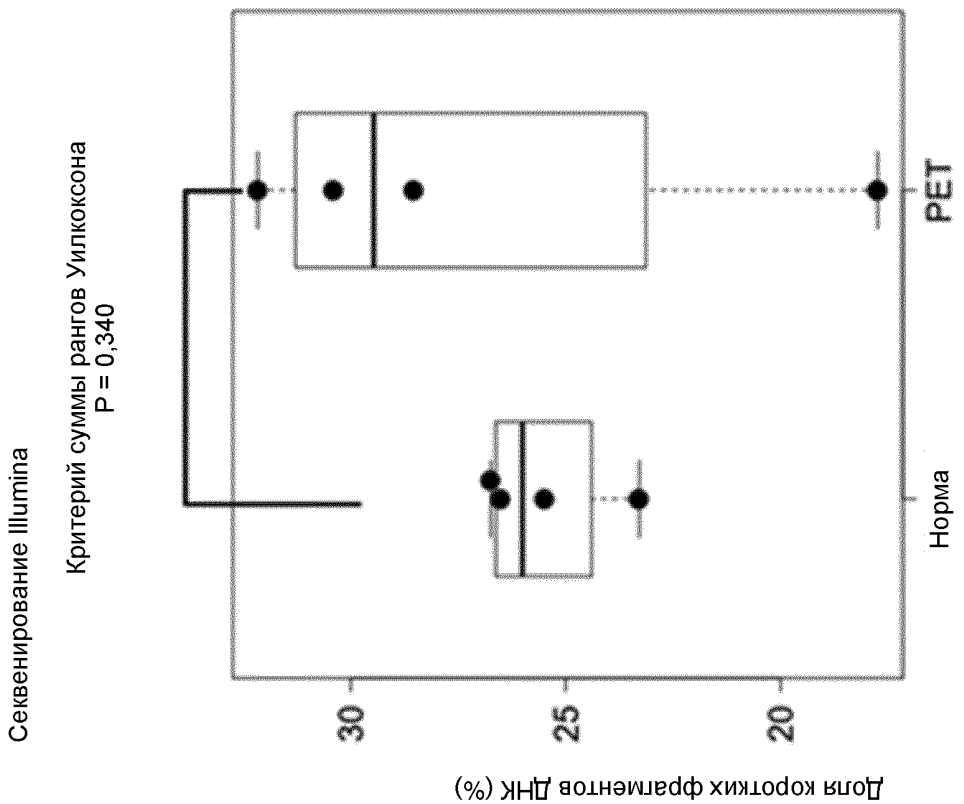
ФИГ. 68D



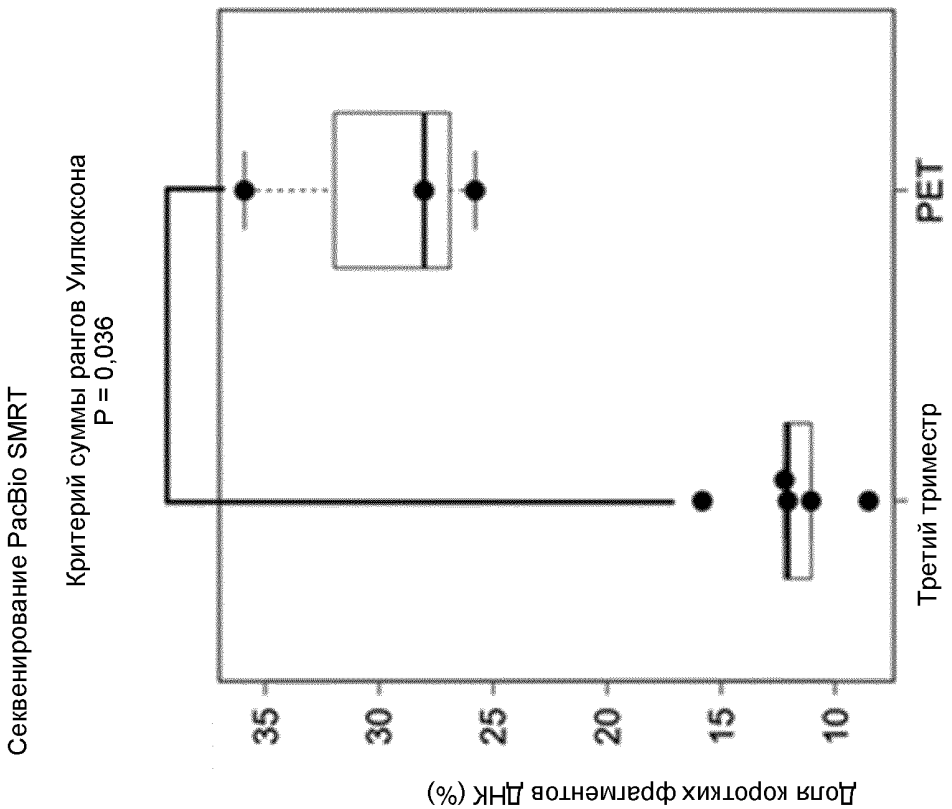
ФИГ. 69B



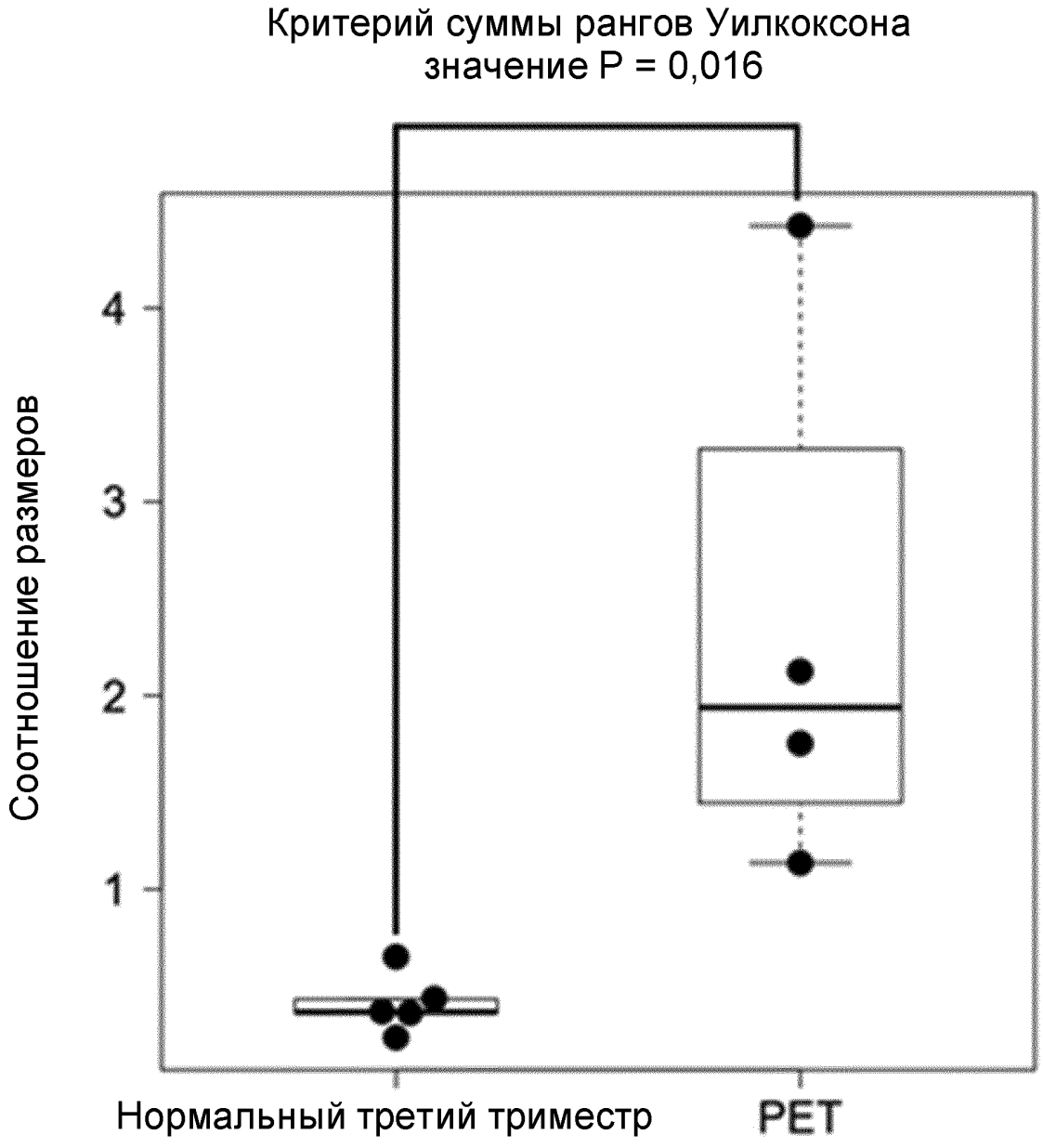
ФИГ. 69A



ФИГ. 70B



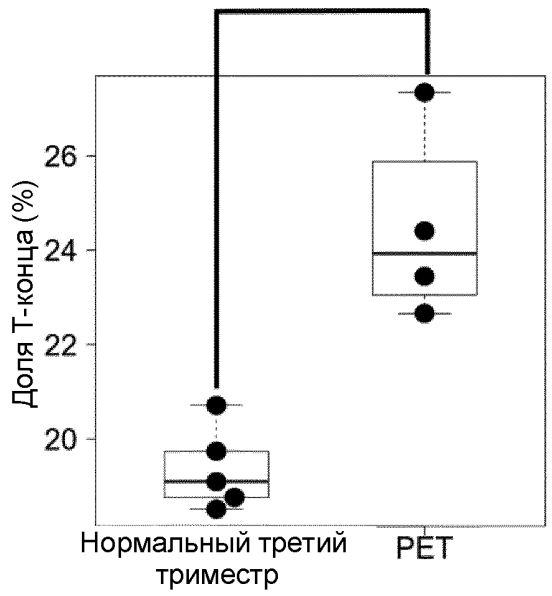
ФИГ. 70A


Критерий суммы рангов Уилкоксона
значение P = 0,016
Соотношение размеров
4
3
2
1
Нормальный третий триместр
PET

ФИГ. 71

Т-конец

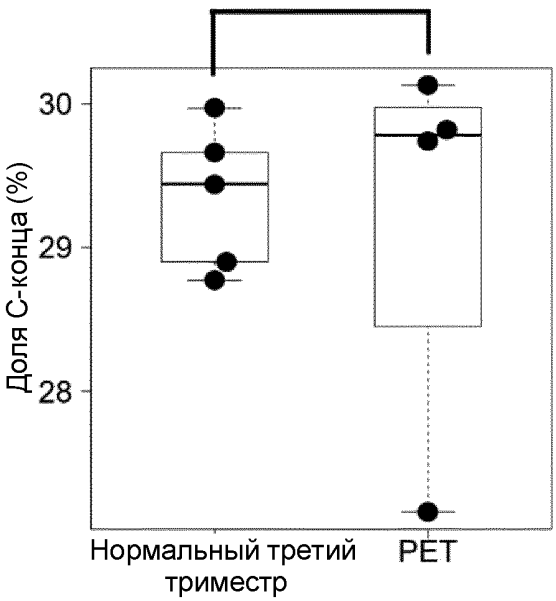
Значение *P* = 0,016*



ФИГ. 72A

С-конец

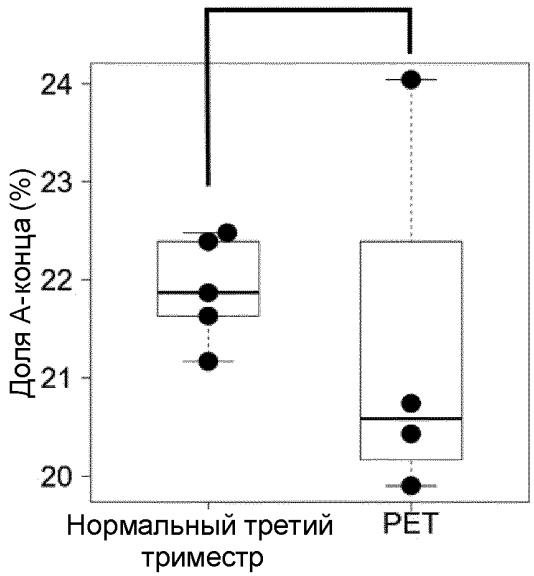
Значение *P* = 0,560



ФИГ. 72B

А-конец

Значение *P* = 0,290

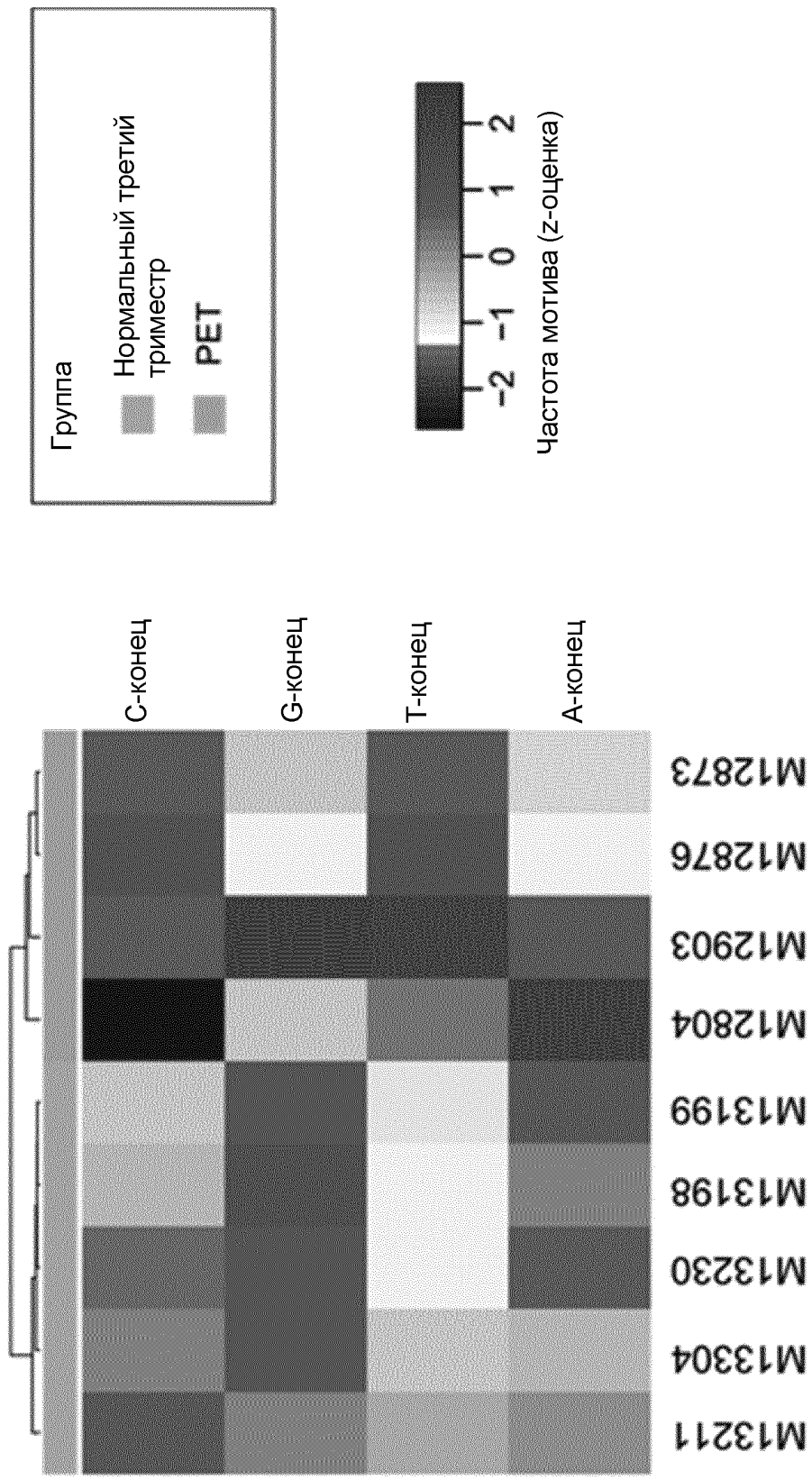


ФИГ. 72C

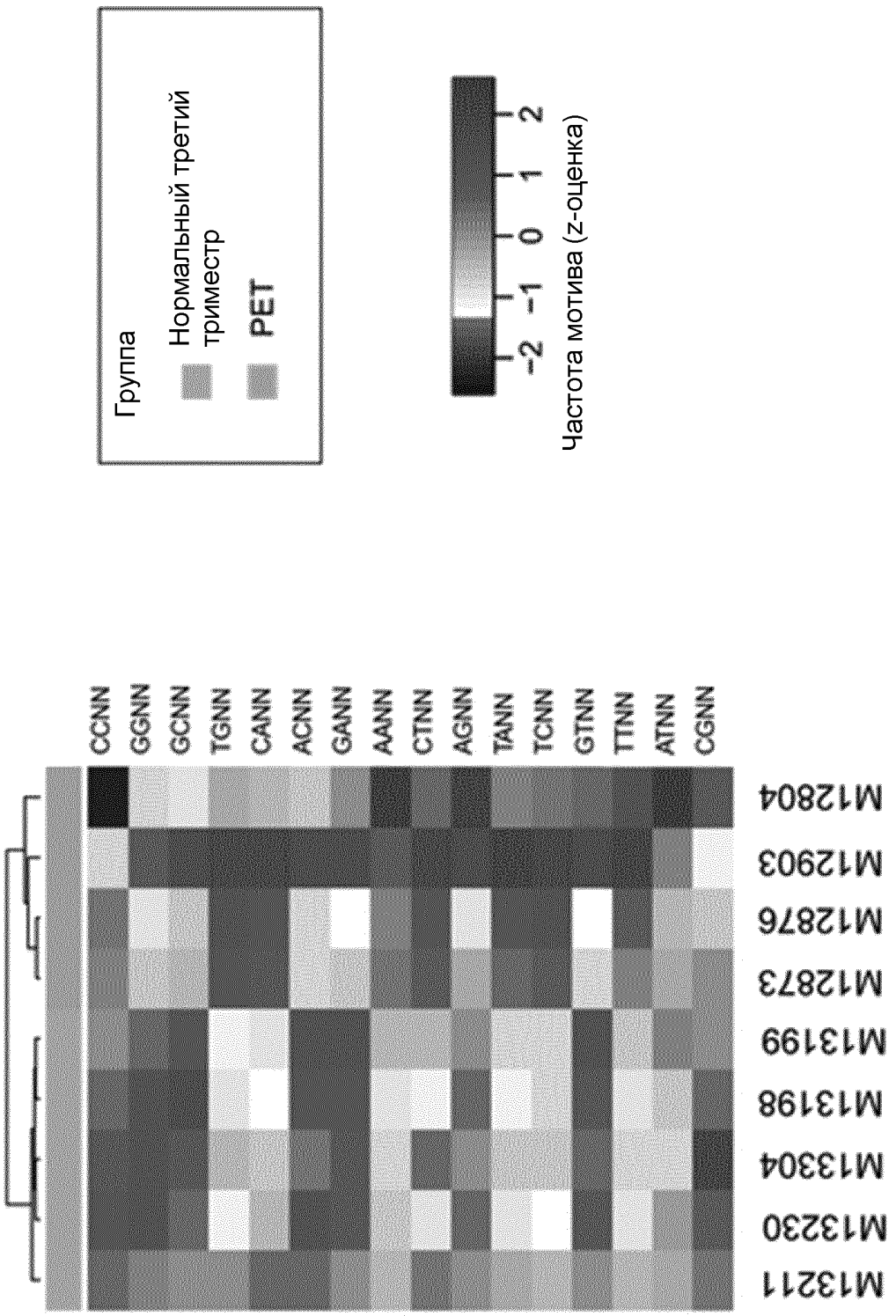
G-конец

Значение *P* = 0,016*

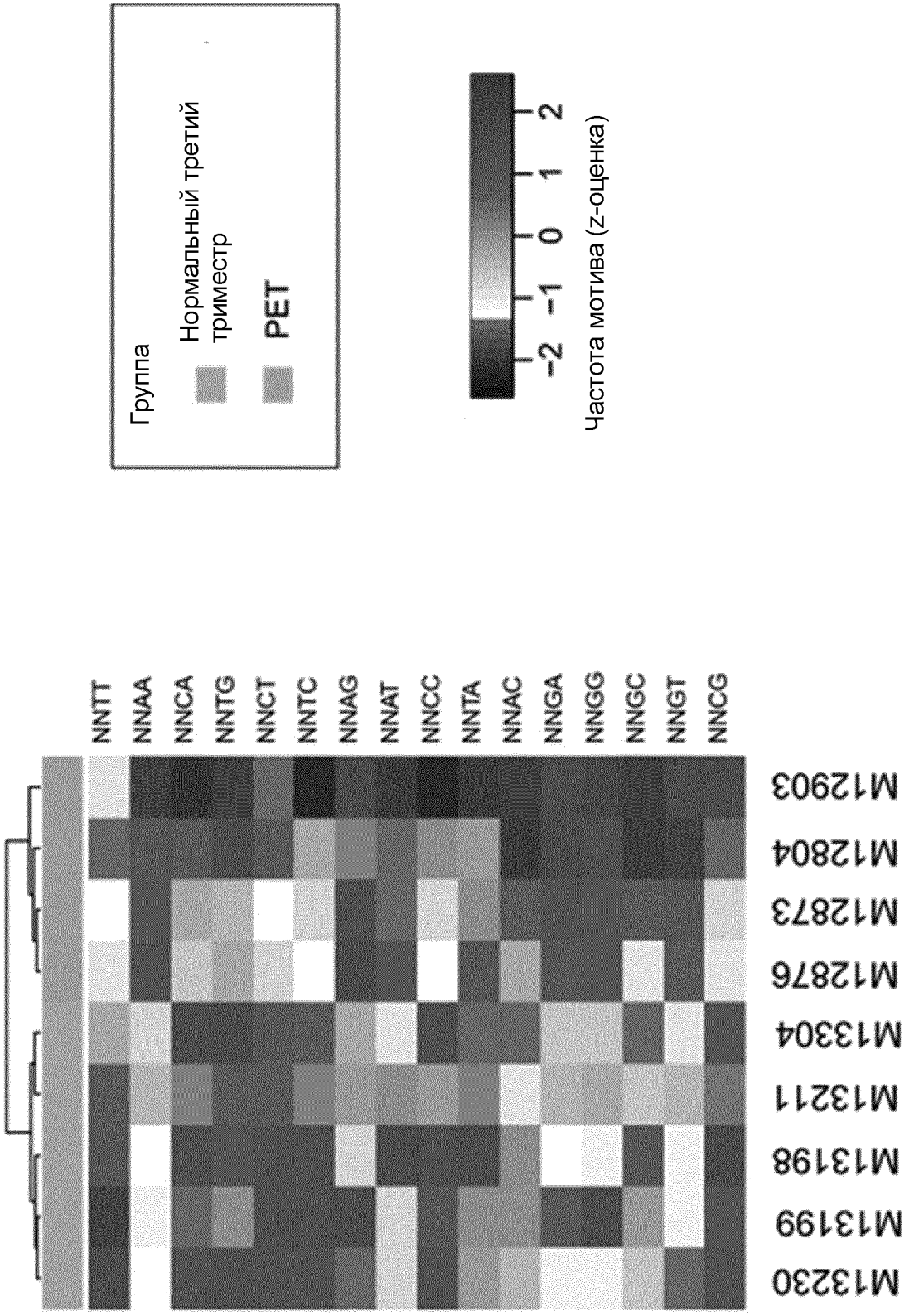
ФИГ. 72D


Группа
Нормальный третий триместр
PET
−2 −1 0 1 2
Частота мотива (z-оценка)
C-конец
G-конец
T-конец
A-конец
M12873
M12876
M12903
M12804
M13199
M13198
M13230
M13304
M13211

ФИГ. 73


Группа
Нормальный третий триместр
PET
−2 −1 0 1 2
Частота мотива (z-оценка)
CCNN
GGNN
GCNN
TGNN
CANN
ACNN
GANN
AANN
CTNN
AGNN
TANN
TCNN
GTNN
TTNN
ATNN
CGNN
M12804
M12903
M12876
M12873
M13199
M13198
M13304
M13230
M13211

ФИГ. 74


Группа
Нормальный третий триместр
PET
Частота мотива (z-оценка)
−2
−1
0
1
2
NNTT
NNAA
NNCA
NNTG
NNCT
NNTC
NNAG
NNAT
NNCC
NNTA
NNAC
NNGA
NNGG
NNGC
NNGT
NNCG
M12903
M12804
M12873
M12876
M13304
M13211
M13198
M13199
M13230

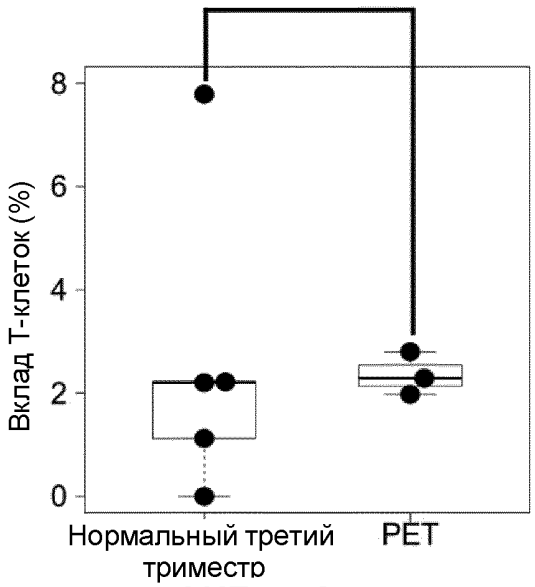
ФИГ. 75

5-концевые мотивы
M13230
M13198
M13199
M13211
M13304
M12804
M12903
M12873
M12876
Группа
Нормальный третий триместр
PET
−2 −1 0 1 2
Частота мотива (z-оценка)

ФИГ. 76

T-конец

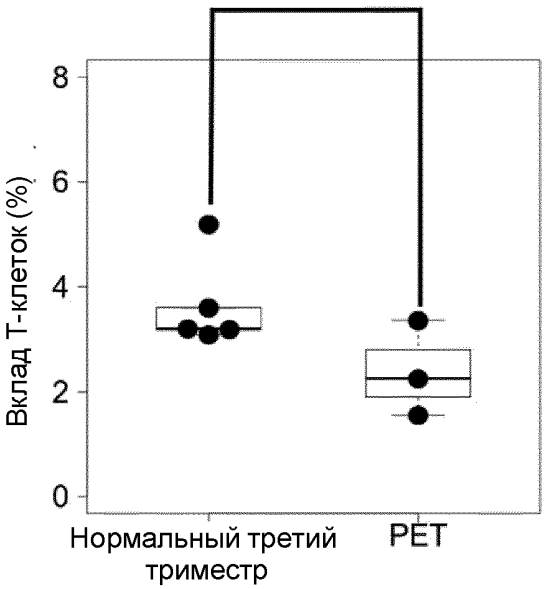
Значение *P* = 0,570



ФИГ. 77A

C-конец

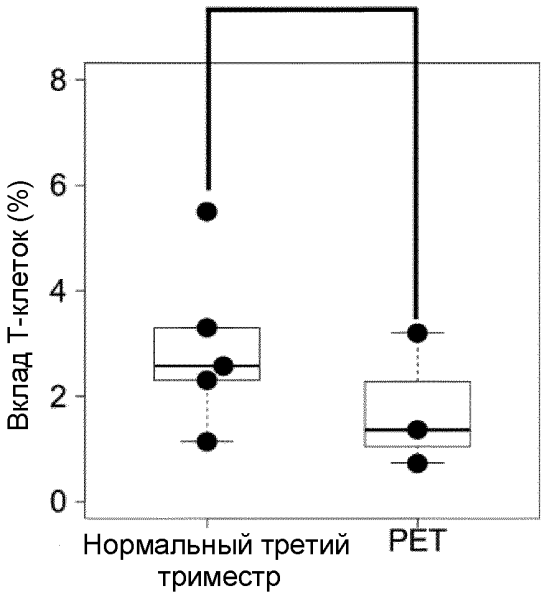
Значение *P* = 0,250



ФИГ. 77B

A-конец

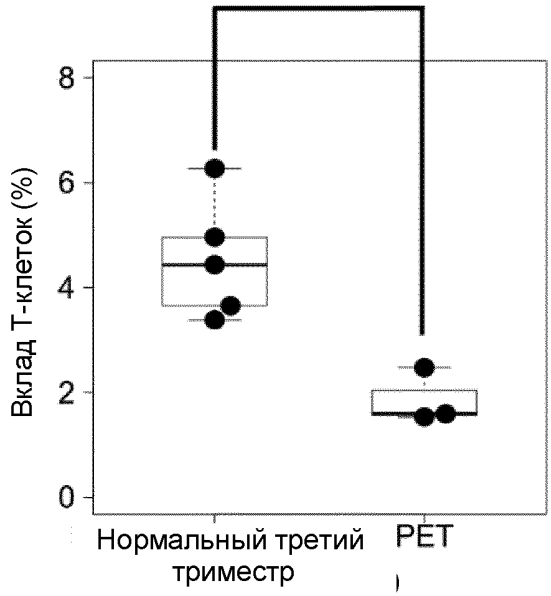
Значение *P* = 0,390



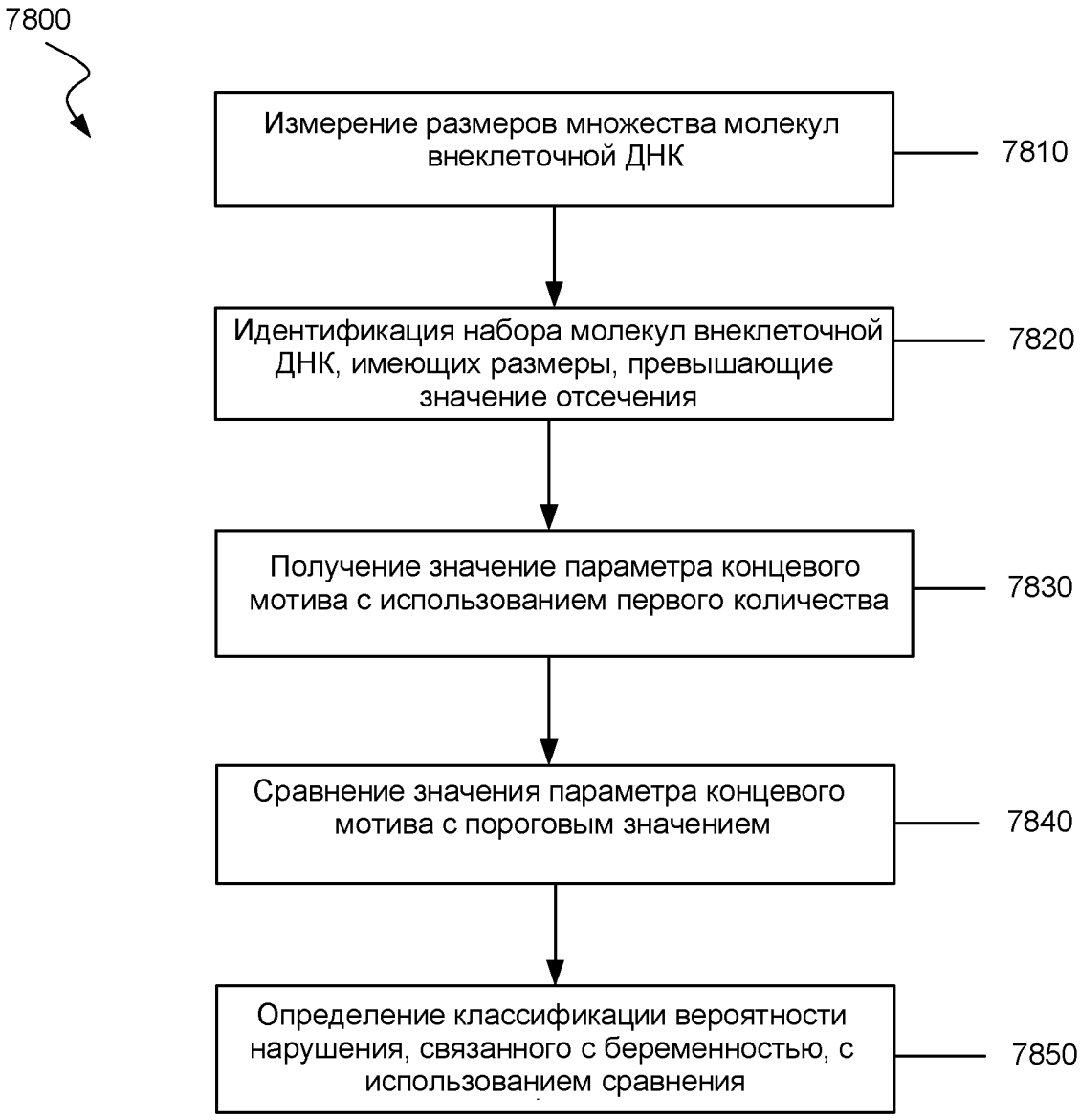
ФИГ. 77C

G-конец

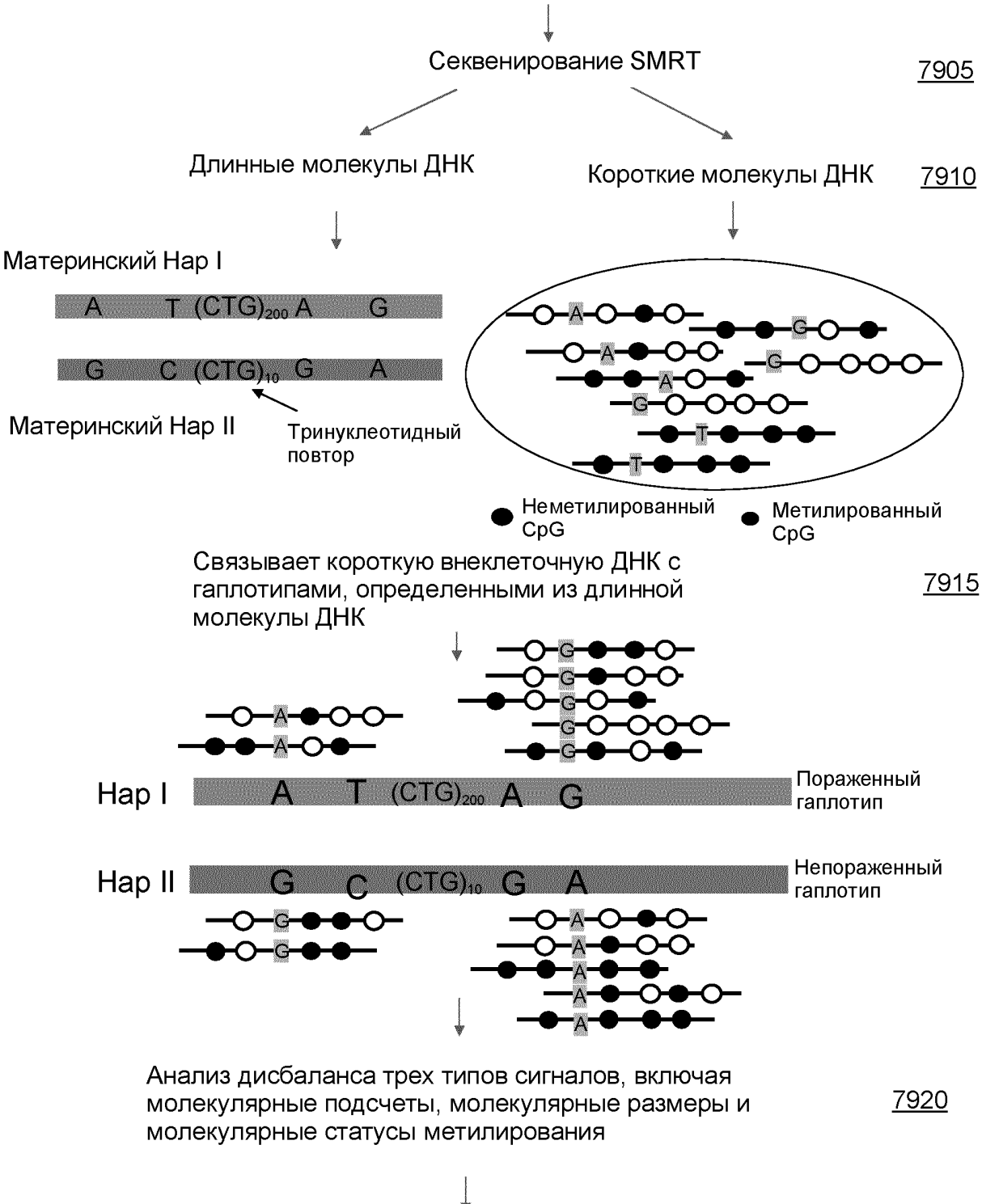
Значение *P* = 0,036*



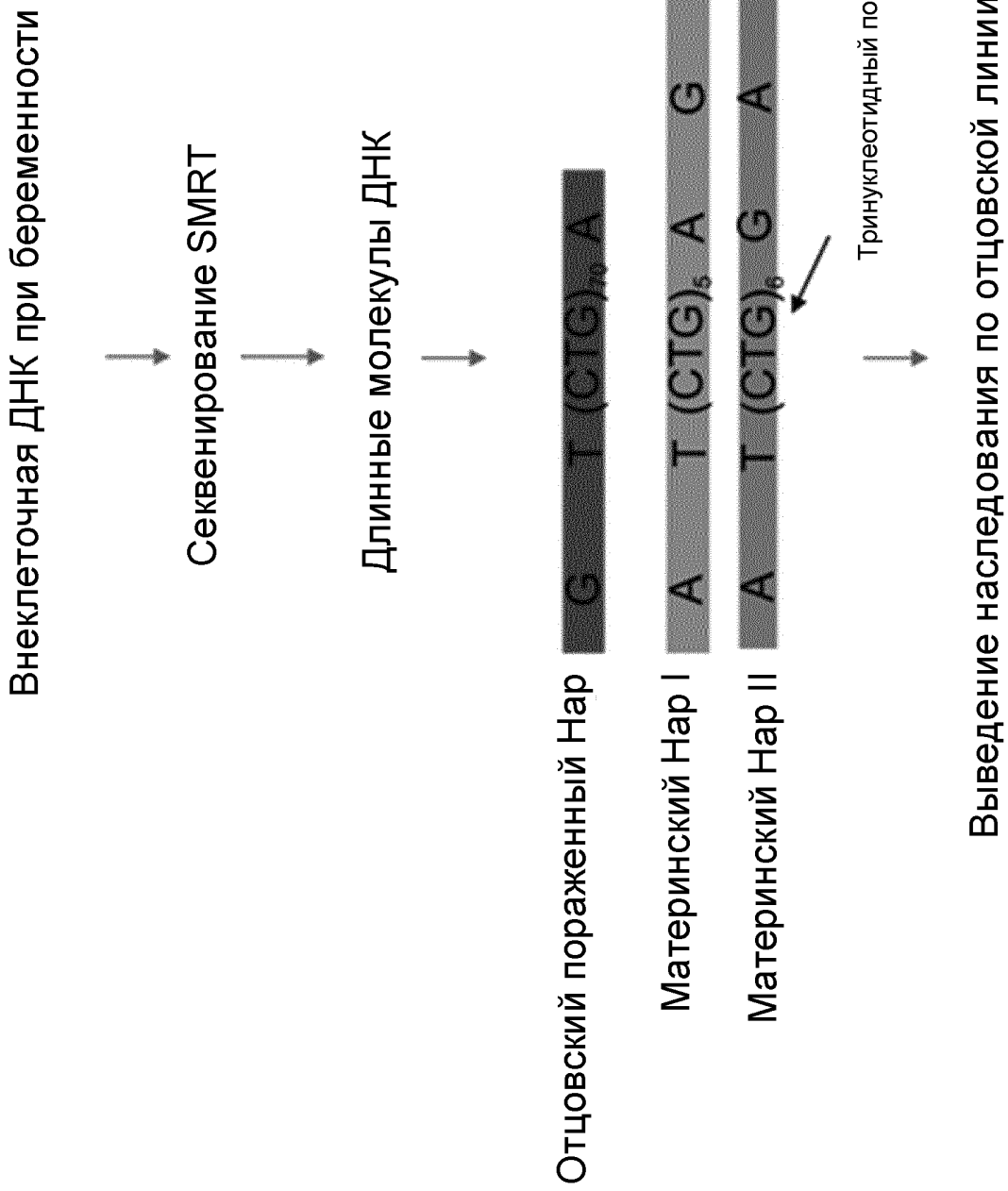
ФИГ. 77D


7800
Измерение размеров множества молекул внеклеточной ДНК
7810
Идентификация набора молекул внеклеточной ДНК, имеющих размеры, превышающие значение отсечения
7820
Получение значение параметра концевого мотива с использованием первого количества
7830
Сравнение значения параметра концевого мотива с пороговым значением
7840
Определение классификации вероятности нарушения, связанного с беременностью, с использованием сравнения
7850

ФИГ. 78


Внеклеточная ДНК при беременности
Секвенирование SMRT
7905
Длинные молекулы ДНК
Короткие молекулы ДНК
7910
Материнский Hap I
A T (CTG)200 A G
G C (CTG)10 G A
Материнский Hap II
Тринуклеотидный повтор
Неметилированный CpG
Метилированный CpG
Связывает короткую внеклеточную ДНК с гаплотипами, определенными из длинной молекулы ДНК
7915
Hap I
A T (CTG)200 A G
Пораженный гаплотип
Hap II
G C (CTG)10 G A
Непораженный гаплотип
Анализ дисбаланса трех типов сигналов, включая молекулярные подсчеты, молекулярные размеры и молекулярные статусы метилирования
7920
Выведение наследования по материнской линии плода
7925

ФИГ. 79


Внеклеточная ДНК при беременности
Секвенирование SMRT
Длинные молекулы ДНК
Отцовский пораженный Нар
G T (CTG)n A
Материнский Нар I
A T (CTG)5 A G
Материнский Нар II
A T (CTG)9 G A
Тринуклеотидный повтор
Выведение наследования по отцовской линии плода

ФИГ. 80

Заболевания, связанные с распространением повторов	Типы повторов	Количество повторов у нормальных субъектов	Количество повторов у пораженных субъектов	Генетические местоположения, связанные с повторами	Названия генов	Профили наследования
Спиноцеребеллярная атаксия 1	CAG	6-39	41-83	CDS	Атаксин 1	Аутосомно-доминантный
Спиноцеребеллярная атаксия 2	CAG	<31	33-200	CDS	Атаксин 2	Аутосомно-доминантный
Спиноцеребеллярная атаксия 3	CAG	<44	52-86	CDS	Атаксин 3	Аутосомно-доминантный
Спиноцеребеллярная атаксия 6	CAG	<18	20-33	CDS	Кальциевый канал, потенциал-зависимый, тип P/Q, субъединица альфа 1A	Аутосомно-доминантный
Спиноцеребеллярная атаксия 7	CAG	<28	>36	CDS	Атаксин 7	Аутосомно-доминантный
Спиноцеребеллярная атаксия 17	CAG	25-44	47-63	CDS	Белок, связывающий TATA-бокс	Аутосомно-доминантный
Спинальная и бульбарная мышечная атрофия	CAG	<34	>38	CDS	Андрогеновый рецептор	Сцепленный с X-хромосомой, рецессивный
Дентаторубрально-паллидолуйзийская атрофия, болезнь Найто-Оянаги	CAG	6-35	49-88	CDS	Атрофин 1	Аутосомно-доминантный
Болезнь Хантингтона	CAG	9-29	36-121	CDS	Хантингтин	Аутосомно-доминантный
Спиноцеребеллярная атаксия 8	CAG	15-50	71-1300	CDS	Атаксин 8	Аутосомно-доминантный
Спастическая параплегия 4, аутосомно-доминантная	CAG	<16	>60	CDS	Спастин	Аутосомно-доминантный
Эпифизарная дисплазия, множественная, 1	GAC	5	6-7	CDS	Олигомерный матриксный белок хряща	Аутосомно-доминантный
Ассоциация Vacterl, сцепленная с X-хромосомой, с гидроцефалией или без нее	GCC	10	12	CDS	Член семейства Zic 3	X-сцепленный рецессивный, X-сцепленный рецессивный

ФИГ. 81

Нейропатия, наследственная сенсорная и вегетативная, тип VIII	GCC	12	18-19	CDS	PR/SET домен 12	Аутосомно-рецессивный
Окулофарингеальная мышечная дистрофия	GCG	10	12-17	CDS	Поли(А) связывающий белок, ядерный 1	Аутосомно-доминантный
Синполидактилия 1	GCG	15	22-29	CDS	Гомеобокс D13	Аутосомно-доминантный
Ладонно-подошвенно-генитальный синдром	GCG	18	24-26	CDS	Гомеобокс A13	Аутосомно-доминантный
Эпилептическая энцефалопатия, ранняя инфантильная, 1	GCG	10-16	17-23	CDS	Aristaless-связанный гомеобокс	Сцепленный с X-хромосомой, рецессивный
Сцепленный с X-хромосомой синдром умственной отсталости Партингтона	GCG	Н/П	Дублирование 24 п.о.	CDS	Aristaless-связанный гомеобокс	Сцепленный с X-хромосомой, рецессивный
Центральный гиповентиляционный синдром, врожденный	GCG	20	24-33	CDS	Парноподобный гомеобокс 2b	Аутосомно-доминантный
Голопрозэнцефалия 5	GCG	15	25	CDS	Член семейства Zic 2	Аутосомно-доминантный

ФИГ. 81 (ПРОДОЛЖЕНИЕ)

Заболевания, связанные с распространением повторов	Типы повторов	Количество повторов у нормальных субъектов	Количество повторов у пораженных субъектов	Генетические местоположения, связанные с повторами	Названия генов	Профили наследования
Блефарофимоз, птоз и синдром обратного эпикантуса	GCG	14	22-24	CDS	Forkhead-бокс L2	Аутосомно-доминантный
Ключично-черепная дисплазия	GCN	17	27	CDS	Runt-связанный фактор транскрипции 2	Аутосомно-доминантный
Умственная отсталость, сцепленная с Х-хромосомой	GCN	11	15-26	CDS	SRY-бокс 3	Сцепленный с Х-хромосомой, рецессивный
Спиноцеребеллярная атаксия 12	CAG	7-32	51-78	5′-НТО	Протеинфосфатаза 2, регуляторная субъединица B, бета	Аутосомно-доминантный
Миоклоническая эпилепсия Унверрихта-Лундборга	CCCCGCCCCGC G	2-3	30-75	5′-НТО	Цистатин B	Аутосомно-рецессивный
Умственная отсталость, сцепленная с Х-хромосомой, связанная с хрупким сайтом Fraxe	CGG	4-39	>200	5′-НТО	Член семейства AF4/FMR2 2	Сцепленный с Х-хромосомой, рецессивный
Синдром Якобсена	CCG	11	>100	5′-НТО	Протоонкоген Cbl	Единичные случаи
Тремор/атаксия, связанная с ломкой Х-хромосомой	CGG	<55	>200	5′-НТО	Сцепленная с ломкой Х-хромосомой умственная отсталость 1	Х-сцепленный, доминантный
Синдром сцепленной с ломкой Х-хромосомой умственной отсталости	CGG	6-52	231-2000	5′-НТО	Сцепленная с ломкой Х-хромосомой умственная отсталость 1	Х-сцепленный, доминантный
Преждевременная недостаточность яичников 1	CGG	7-40	55-200 45-54	5′-НТО	Сцепленная с ломкой Х-хромосомой умственная отсталость 1 (FMR1)	Х-сцепленное наследование
Умственная отсталость, тип FRA12A	CGG	12-26	>150	5′-НТО	DIP2, дисковзаимодействующий белок 2, гомолог B	Аутосомно-доминантный

ФИГ. 82

Окулофарингодистальная миопатия	CGG	13-45	>51	5′-НТО	Белок 12, родственный рецептору ЛПНП	Аутосомно-доминантный
Окулофарингеальная миопатия с лейкоэнцефалопатией	CGG	11-16	>52	5′-НТО	NOTCH2NLC	Аутосомно-доминантный
Последовательность Робина с расщелиной нижней челюсти и аномалиями конечностей	CACA/CGCA	3-12	14-16	5′-НТО	Эукариотический фактор инициации трансляции 4A3	Аутосомно-рецессивный
Миотоническая дистрофия 1	CTG	5-37	50-5000	3′-НТО	Протеинкиназа DM1	Аутосомно-доминантный
Подобный болезни Хантингтона 2	CTG	6-28	>41	3′-НТО	Джанктофилин 3	Аутосомно-доминантный
Окулофарингеальная миопатия с лейкоэнцефалопатией	CGG	<7	>37	Экзон	NUTM2B-AS1	Аутосомно-доминантный
Спиноцеребеллярная атаксия 8	CTG	15-34	90-250	Экзон	Противоположная цепь ATXN8	Аутосомно-доминантный
Спиноцеребеллярная атаксия 10	ATTCT	10-29	400-4500	Интрон 9	Атаксин 10	Аутосомно-доминантный
Спиноцеребеллярная атаксия 37	ATTTC	<30	46-71	Интрон 11	DAB1; адаптерный белок рилина	Аутосомно-доминантный
Миотоническая дистрофия 2	CCTG	<30	75-11000	Интрон 1	Связывающий нуклеиновую кислоту белок с цинковым пальцем CCHC-типа	Аутосомно-доминантный

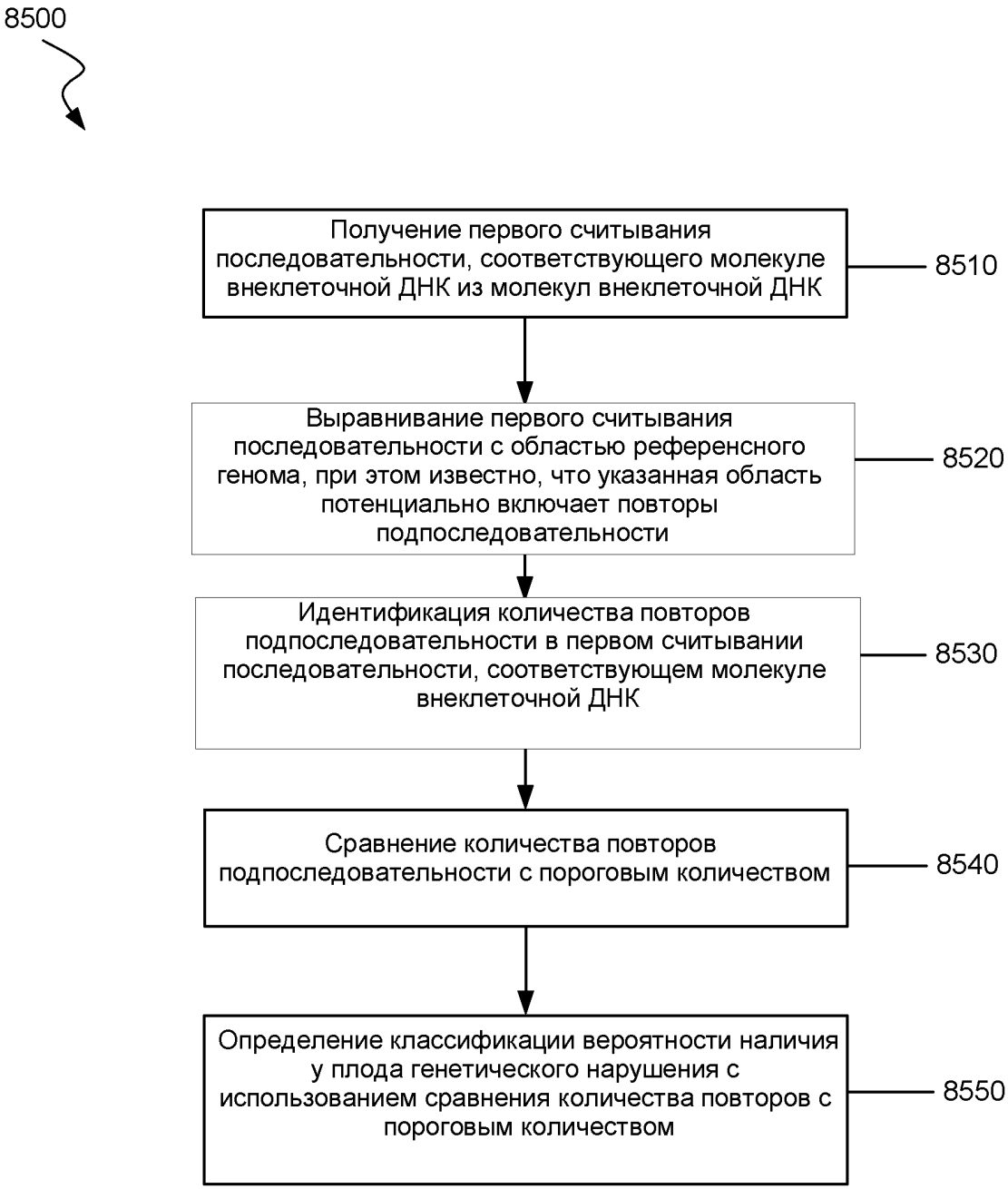
ФИГ. 82 (ПРОДОЛЖЕНИЕ)

Заболевания, связанные с распространением повторов	Типы повторов	Количество повторов у нормальных субъектов	Количество повторов у пораженных субъектов	Генетические местоположения, связанные с повторами	Названия генов	Профили наследования
Эндотелиальная дистрофия роговицы Фукса	CTG	<40	>50	Интрон 3	Фактор транскрипции 4	Аутосомно-доминантный
Мышечная дистрофия Дюшенна	GAA	11-33	59-82	Интрон 62	Дистрофин	Сцепленный с Х-хромосомой, рецессивный
Атаксия Фридриха 1	GAA	5-30	>70	Интрон 1	Фратаксин	Аутосомно-рецессивный
Спиноцеребеллярная атаксия 36	GGCCTG	3-14	650-2500	Интрон 4	Рибонуклеопротеин NOP56	Аутосомно-доминантный
Лобно-височная дегенерация и боковой амиотрофический склероз	GGGGCC	2-19	250-1600	Интрон 1	Открытая рамка считывания 72 хромосомы 9	Аутосомно-доминантный
Спиноцеребеллярная атаксия 31	TGGAA	26	2,5-3,8 тыс.п.о.	Интрон 1	Экспрессируемый в головном мозге, связанный с NEDD4, 1	Аутосомно-доминантный
Семейная миоклоническая эпилепсия взрослых-1	TTTCA/TTTTA	7-20	440-3680	Интрон	Содержащий домен Sterile Alpha-мотива 12	Аутосомно-доминантный
Доброкачественная семейная миоклоническая эпилепсия взрослых 6	TTTTA/TTTCA	18	>22	Интрон	Содержащий тринуклеотидный повтор 6A	Аутосомно-доминантный
Доброкачественная семейная миоклоническая эпилепсия взрослых 7	TTTTA/TTTCA	12	>22	Интрон	Фактор обмена гуаниновых нуклеотидов Rap 2	Аутосомно-Доминантный

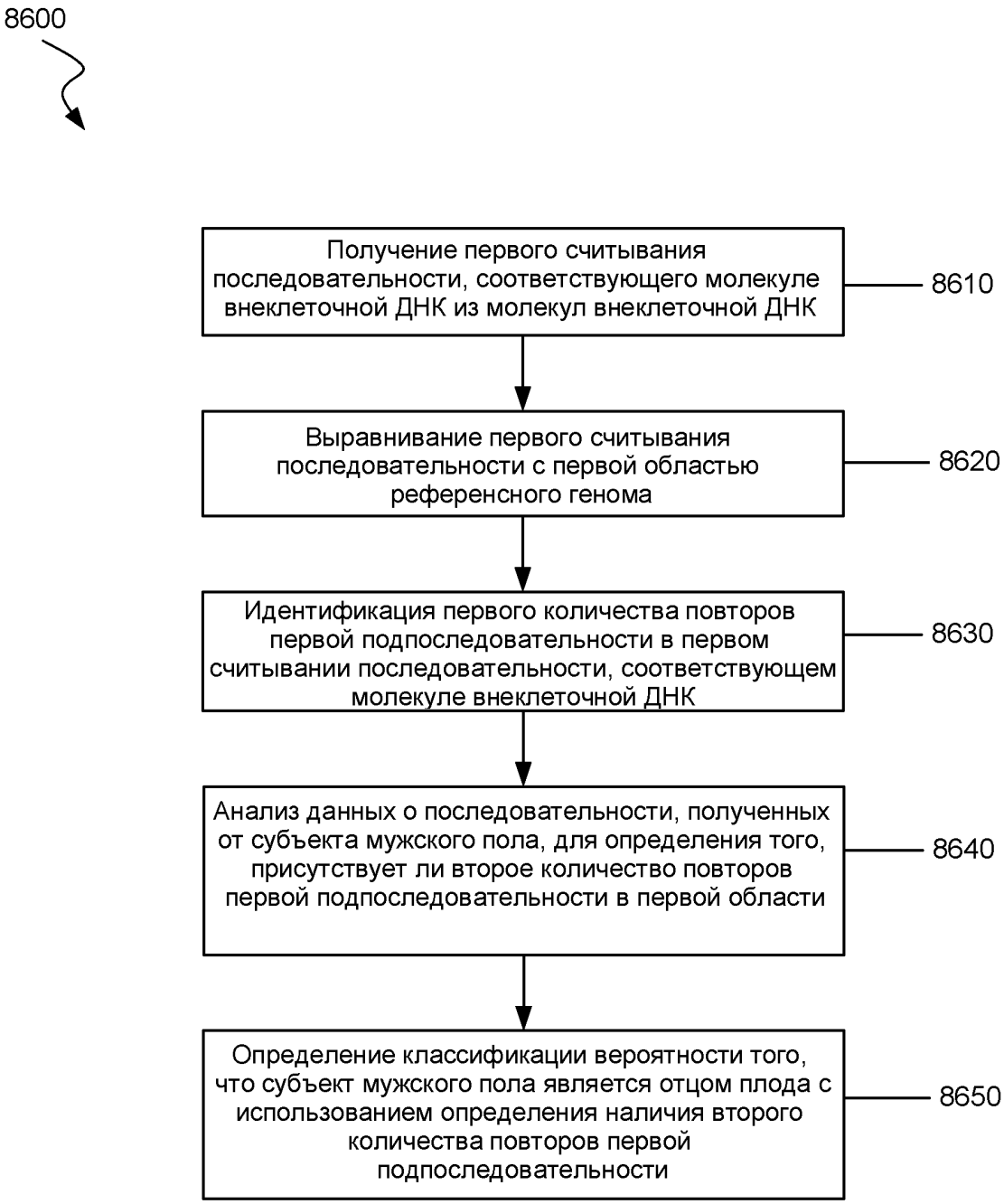
ФИГ. 83

Типы повторов	Звено повто-ра	Геномные положения	Референс-ные основания	Отцовские генотипы (аллель 1/аллель 2)	Материнские генотипы (аллель 1/аллель 2)	Фетальные генотипы (аллель 1/аллель 2)	Уровень метилирования ДНК плода, связанный с отцовскими аллелями (%)	Уровень метилир. ДНК плода, связанный с материнскими аллелями (%)
1 п.о.	A	chr1:244287246-244287247	C	$C(A)_{20}/C(A)_{20}$	C/C	$C(A)_{20}/C$	51.06	45.83
1 п.о.	T	chr2:831609-831610	C	$C(T)_{22}/C(T)_{22}$	C/C	$C(T)_{22}/C$	52.15	52.46
2 п.о.	TG	chr1:199547268-199547269	C	$C(TG)_6/C(TG)_6$	$C/C(TG)_{11}$	$C(TG)_6/C(TG)_{11}$	55.67	64.08
2 п.о.	TG	chr11:24818036-24818037	$C(TG)_5$	C/C	$C(TG)_6/C(TG)_6$	$C/C(TG)_6$	35.57	31.91
2 п.о.	AC	chr18:64408776-64408777	$A(AC)_1$	A/A	$A(AC)_3/A(AC)_3$	$A/A(AC)_3$	43.26	44.92
3 п.о.	AAT	chr22:42422276-42422277	$A(AAT)_3$	$A(AAT)_5/A(AAT)_5$	A/A	$A(AAT)_5/A$	79.78	82.43
4 п.о.	TAAA	chr4:73237157-73237158	$G(TAAA)_2$	G/G	$G(TAAA)_2/G(TAAA)_3$	$G/G(TAAA)_3$	62.84	95.65
4 п.о.	GATA	chr3:192384705-192384706	$T(GATA)_3$	$T(GATA)_3/T(GATA)_5$	T/T	$T(GATA)_5/T$	50.98	62.9

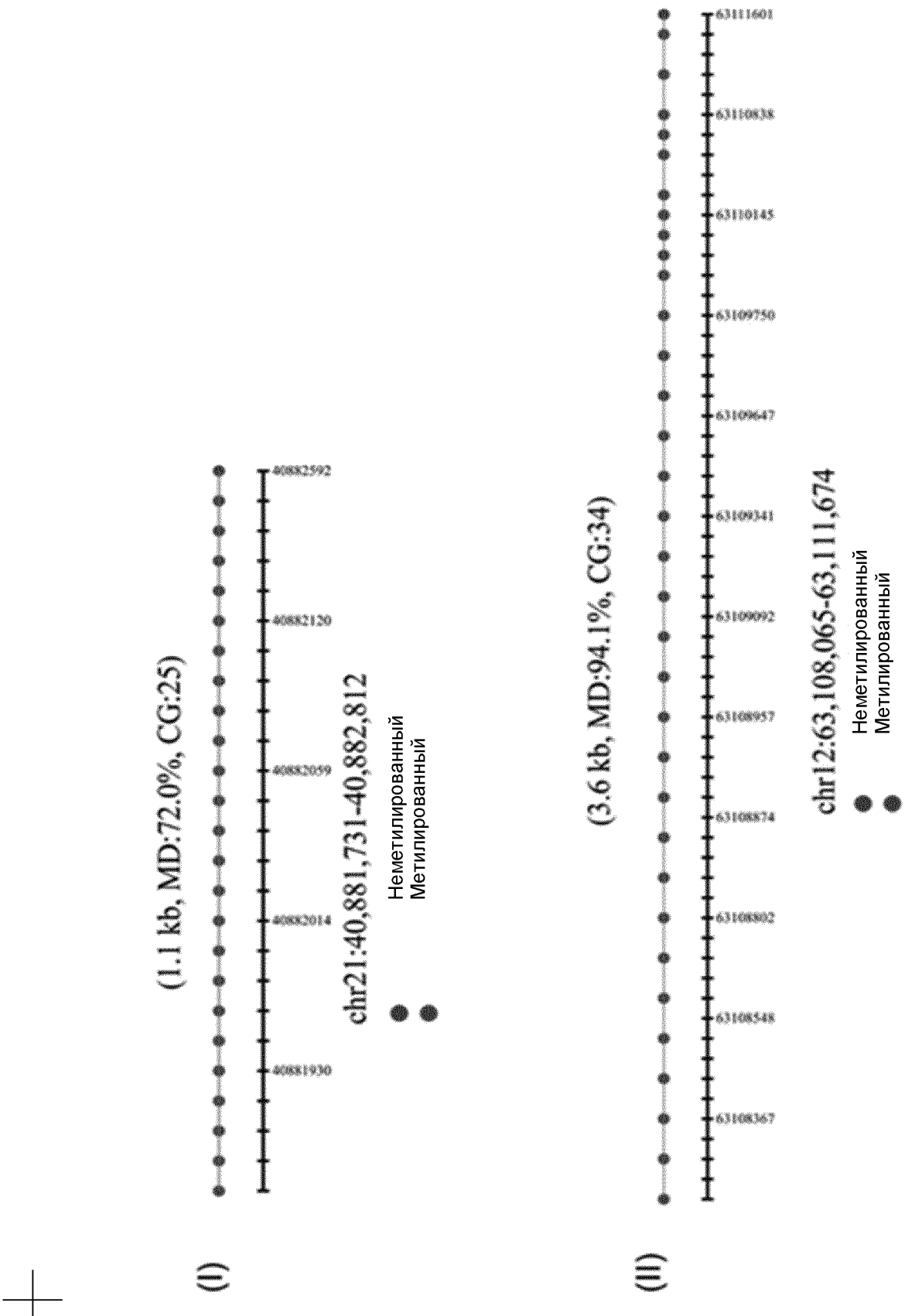
ФИГ. 84


8500
Получение первого считывания последовательности, соответствующего молекуле внеклеточной ДНК из молекул внеклеточной ДНК
8510
Выравнивание первого считывания последовательности с областью референсного генома, при этом известно, что указанная область потенциально включает повторы подпоследовательности
8520
Идентификация количества повторов подпоследовательности в первом считывании последовательности, соответствующем молекуле внеклеточной ДНК
8530
Сравнение количества повторов подпоследовательности с пороговым количеством
8540
Определение классификации вероятности наличия у плода генетического нарушения с использованием сравнения количества повторов с пороговым количеством
8550

ФИГ. 85


8600
Получение первого считывания последовательности, соответствующего молекуле внеклеточной ДНК из молекул внеклеточной ДНК
8610
Выравнивание первого считывания последовательности с первой областью референсного генома
8620
Идентификация первого количества повторов первой подпоследовательности в первом считывании последовательности, соответствующем молекуле внеклеточной ДНК
8630
Анализ данных о последовательности, полученных от субъекта мужского пола, для определения того, присутствует ли второе количество повторов первой подпоследовательности в первой области
8640
Определение классификации вероятности того, что субъект мужского пола является отцом плода с использованием определения наличия второго количества повторов первой подпоследовательности
8650

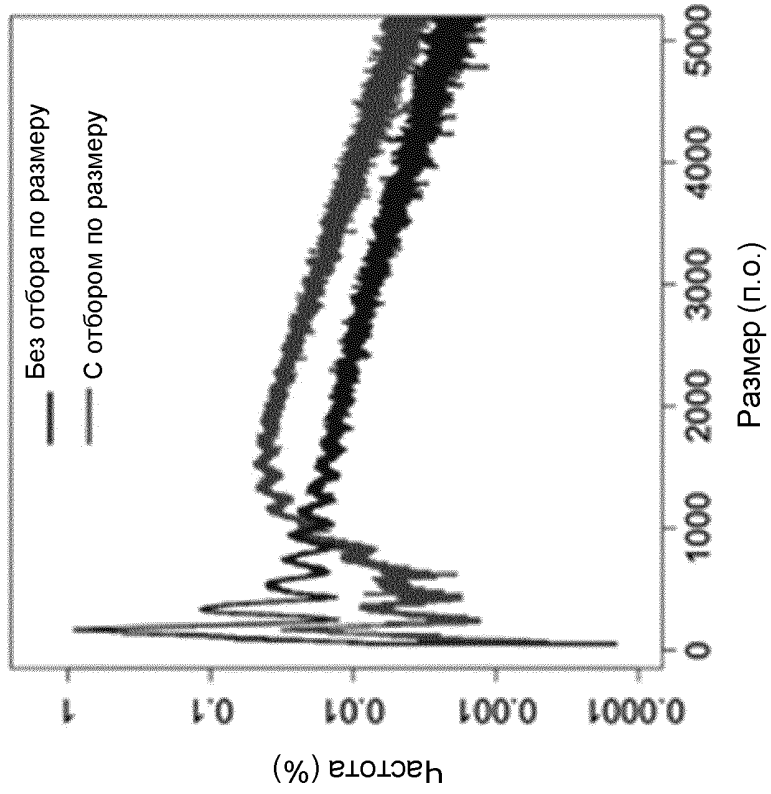
ФИГ. 86


(I)
(1.1 kb, MD:72.0%, CG:25)
40882592
40882120
40882059
40882014
40881930
chr21:40,881,731-40,882,812
Неметилированный
Метилированный
(II)
(3.6 kb, MD:94.1%, CG:34)
63111601
63110838
63110145
63109750
63109647
63109341
63109092
63108957
63108874
63108802
63108548
63108367
chr12:63,108,065-63,111,674
Неметилированный
Метилированный

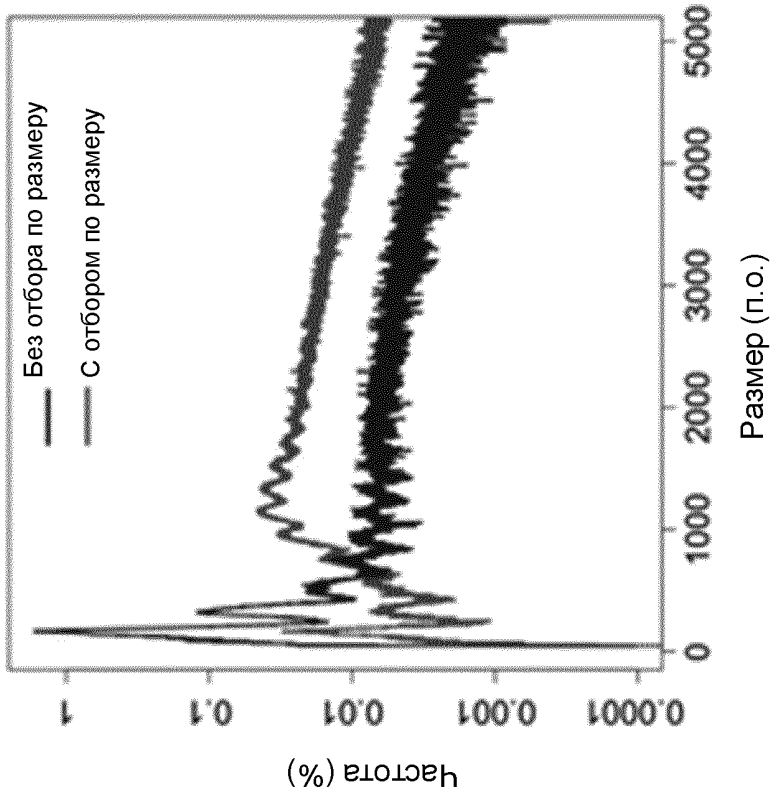
ФИГ. 87

Образцы	Группы	Количество секвенированных молекул	Средняя глубина подсчитываний (x)	Медианные размеры фрагментов (п.о.)	Доля фрагментов ≥500 п.о. (%)
299	Без отбора по размеру	2,525,216	91	176	27.3
300	Без отбора по размеру	3,057,511	67	512	50.5
B299	С отбором по размеру	4,103,718	18	2,463	97.6
B300	С отбором по размеру	1,987,264	19	2,170	97.4

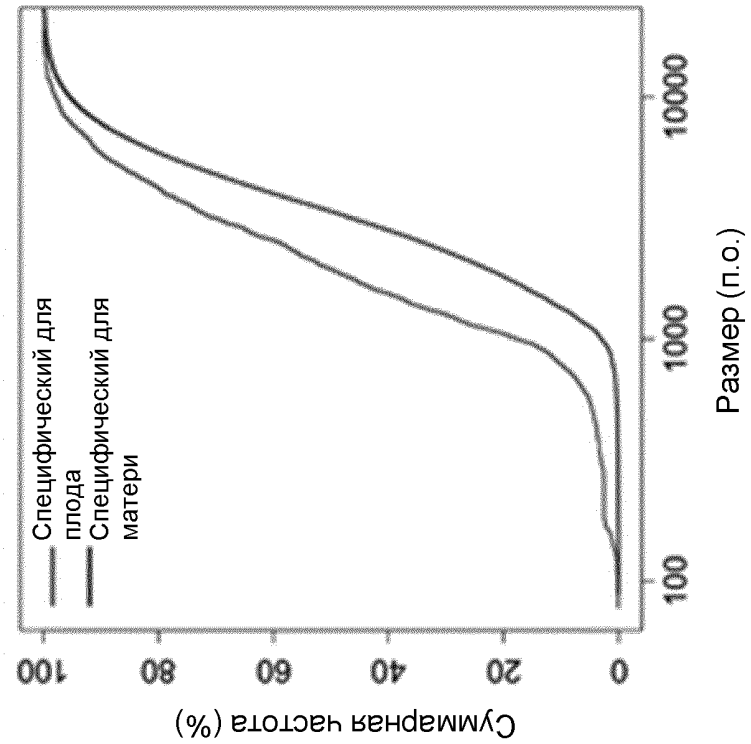
ФИГ. 88



ФИГ. 89B



ФИГ. 89A



ФИГ. 90A

ФИГ. 90B

Образец	**Общее количество анализируемых молекул ДНК плазмы**	**Количество молекул ДНК плазмы, несущих информативные ОНП**	**Процентное содержание молекул ДНК плазмы, несущих информативные ОНП**
Образец 299 (без отбора по размеру)	1,092,062	70,730	6.5%
Образец B299 (с отбором по размеру)	1,633,040	336,539	20.6%

ФИГ. 91

Образец	Группа	Метилированные сайты CpG	Неметилированные сайты CpG	Уровень метилирования (%)
299	Без отбора по размеру	600,998	268,364	69.1
300	Без отбора по размеру	934,996	413,638	69.3
B299	С отбором по размеру (>500 п.о.)	1,358,631	541,425	71.5
B300	С отбором по размеру (>500 п.о.)	817,043	327,869	71.4

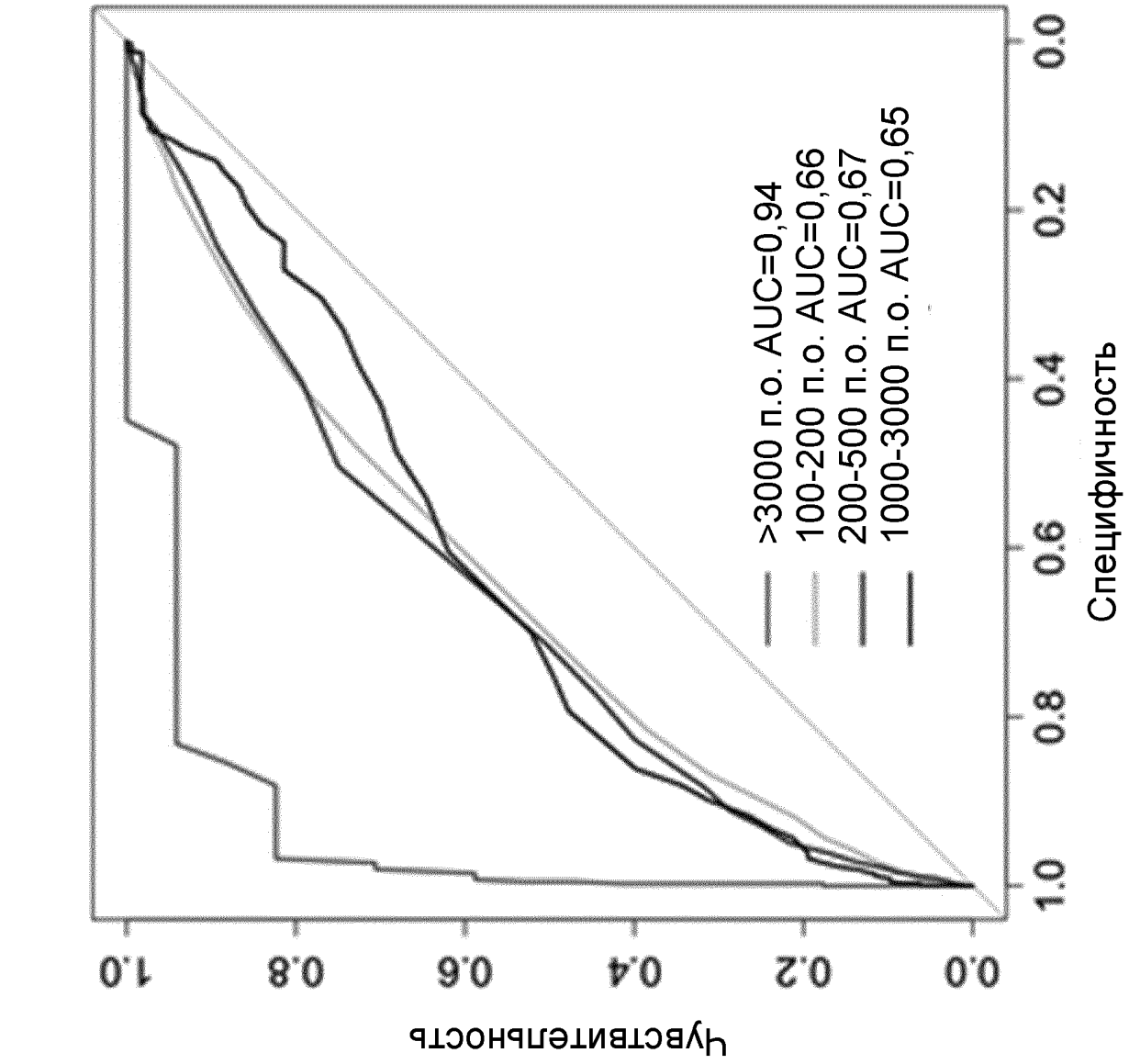
ФИГ. 92

Образец	Группа 1	Группа 2	Метилированные сайты CpG	Неметилирован-ные сайты CpG	Уровень метилирования (%)
299	Без отбора по размеру	Специфические для плода молекулы ДНК плазмы	1,277	932	57.81
		Специфические для матери молекулы ДНК плазмы	17,500	8,003	68.62
B299	С отбором по размеру (>500 п.о.)	Специфические для плода молекулы ДНК плазмы	2,682	1,570	63.08
		Специфические для матери молекулы ДНК плазмы	85,741	33,062	72.17

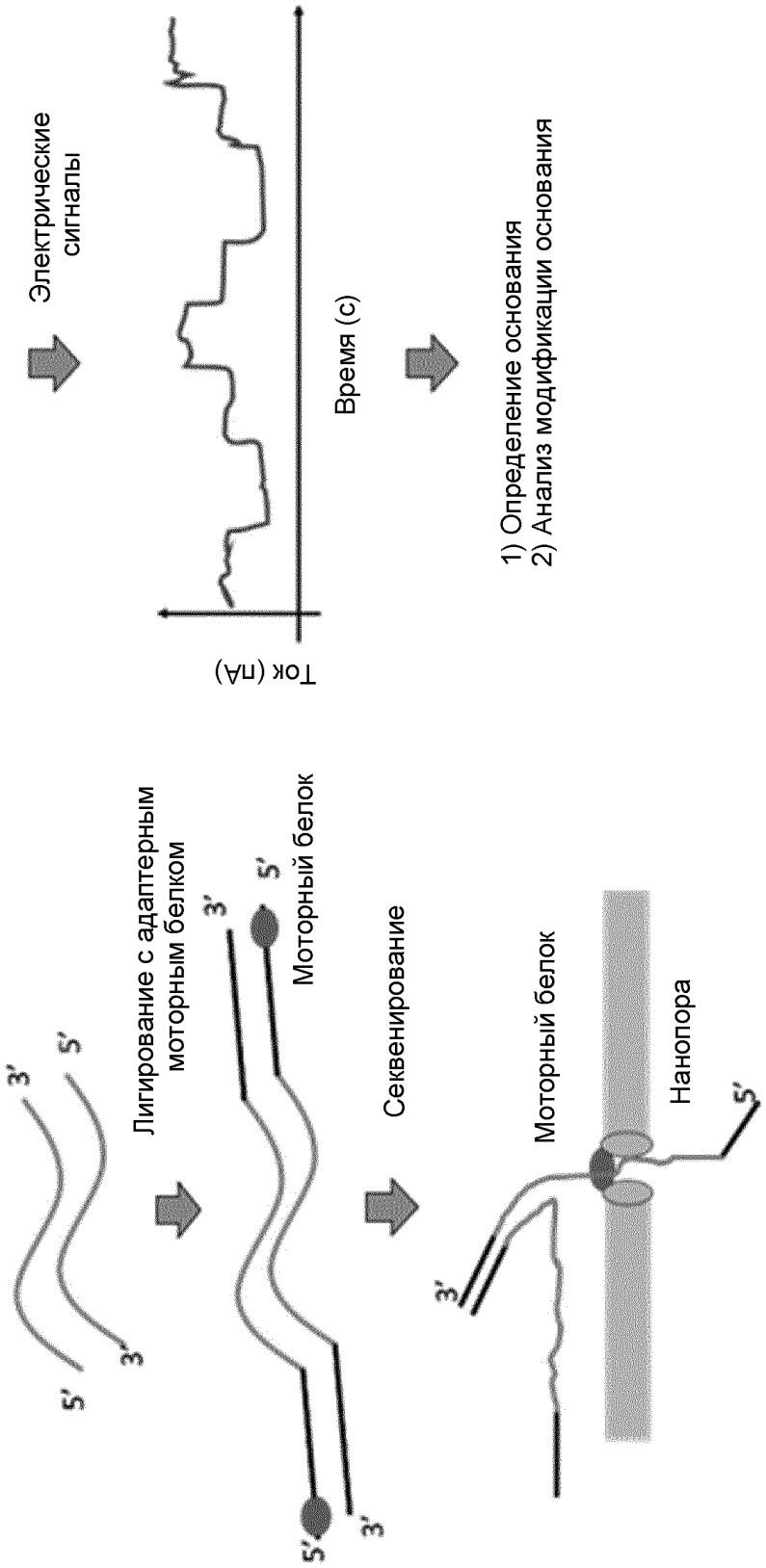
ФИГ. 93

	Без отбора по размеру				С отбором по размеру			
	M13299		M13300		B-M13299		B-M13300	
Ранг	Мотив	Частота (%)	Мотив	Частота (%)	Мотив	Частота (%)	Мотив	Частота (%)
1	CCCA	1.76	ACTT	1.87	ACTT	3.99	ACTT	3.37
2	CCTG	1.47	CCTT	1.67	GCTT	2.73	GCTT	2.27
3	CCTT	1.46	CCCA	1.66	CCTT	2.31	CCTT	2.05
4	GCTT	1.41	GCTT	1.53	ACTG	1.98	GTTT	1.80
5	CCTC	1.37	CCTG	1.47	GTTT	1.88	ACCT	1.54
6	ACTT	1.36	CCTC	1.27	ACCT	1.81	ACTG	1.49
7	CCAG	1.19	GCCT	1.15	GGCT	1.64	CTTT	1.42
8	GGCT	1.18	GTTT	1.14	ACAT	1.59	ACTC	1.38
9	GCCT	1.17	ACCT	1.14	CTTT	1.47	ATTT	1.35
10	GCTG	1.05	CCAG	1.12	GATT	1.46	ACAT	1.35

ФИГ. 94


Чувствительность
0.0
0.2
0.4
0.6
0.8
1.0
Специфичность
1.0
0.8
0.6
0.4
0.2
0.0
>3000 п.о. AUC=0,94
100-200 п.о. AUC=0,66
200-500 п.о. AUC=0,67
1000-3000 п.о. AUC=0,65

ФИГ. 95


3'
5'
5'
3'
Лигирование с адаптерным моторным белком
3'
5'
Моторный белок
5'
3'
Секвенирование
Моторный белок
Нанопора
3'
5'
Ток (пА)
Время (с)
Электрические сигналы
1) Определение основания
2) Анализ модификации основания

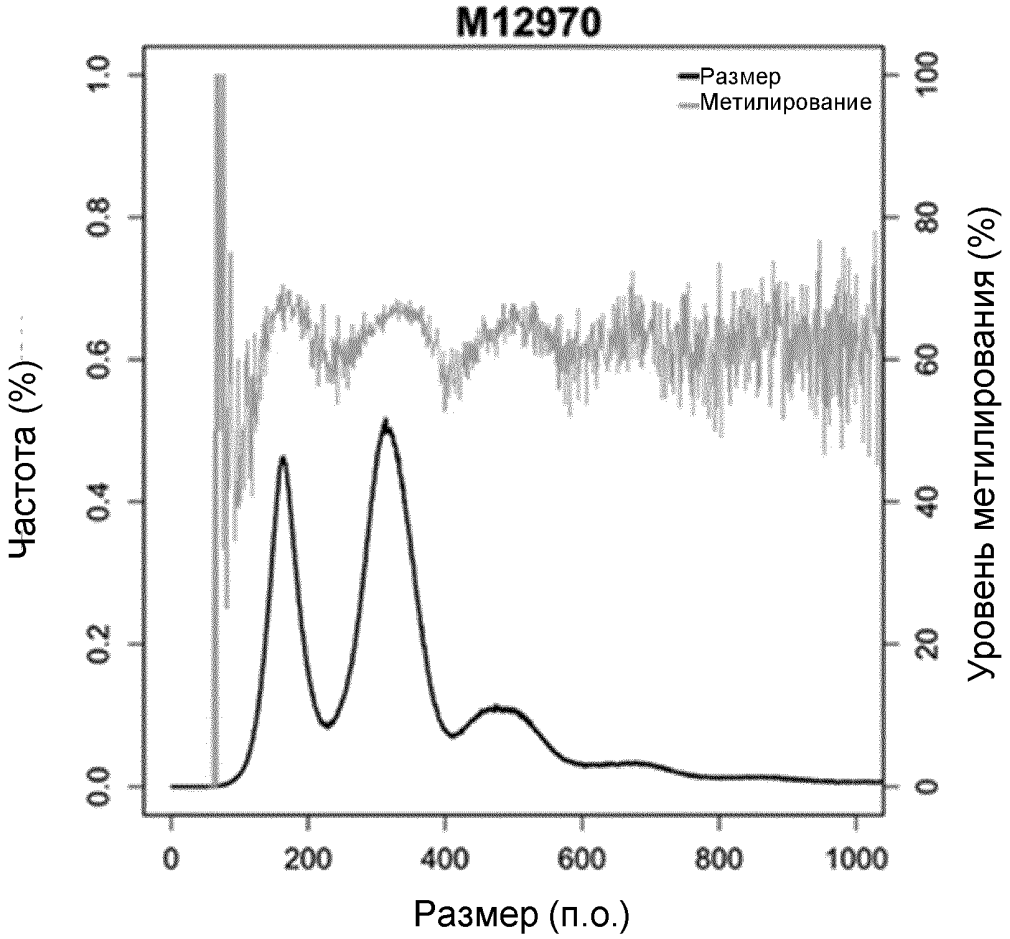
ФИГ. 96

Размер фрагмента (п.о.)	M12970				
	Кол-во фрагментов	Частота (%)	Метилир. CpG	Неметилир. CpG	Уровень метилирования (%)
≥500	636,828	16.63%	1,247,858	698,271	64.12%
≥600	393,108	10.26%	955,061	529,299	64.34%
≥1000	119,185	3.11%	463,660	231,835	66.67%
≥2000	23,518	0.61%	151,939	62,893	70.72%

Размер фрагмента (п.о.)	M12985				
	Кол-во фрагментов	Частота (%)	Метилир. CpG	Неметилир. CpG	Уровень метилирования (%)
≥500	201,170	7.63%	644,985	346,663	65.04%
≥600	104,416	3.96%	426,580	222,633	65.71%
≥1000	25,204	0.96%	168,044	72,768	69.78%
≥2000	4,090	0.16%	46,989	15,562	73.94%

Размер фрагмента (п.о.)	M12969				
	Кол-во фрагментов	Частота (%)	Метилир. CpG	Неметилир. CpG	Уровень метилирования (%)
≥500	590,634	12.55%	1,781,670	1,033,066	63.30%
≥600	350,143	7.44%	1,306,833	759,671	63.24%

ФИГ. 97


M12970
Размер
Метилирование
Частота (%)
Уровень метилирования (%)
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
0
200
400
600
800
1000
Размер (п.о.)

ФИГ. 98

Образец	Количество молекул, несущих общие аллели	Количество молекул, несущих специфические для плода аллели	Фракция ДНК плода (%)
M12970	84,911	17,776	34.6%
M12985	52,059	7,385	24.9%
M12969	95,273	17,007	30.3%

ФИГ. 99

Образец	Специфическая для плода ДНК			Специфическая для матери ДНК		
	Метилированный CpG	Неметилированный CpG	Уровень метилирования (%)	Метилированный CpG	Неметилированный CpG	Уровень метилирования (%)
M12970	17,340	10,434	62.43%	61,268	29,770	67.30%
M12985	9,426	5,682	62.39%	41,465	19,561	67.95%
M12969	26,440	16,563	61.48%	94,573	45,879	67.33%

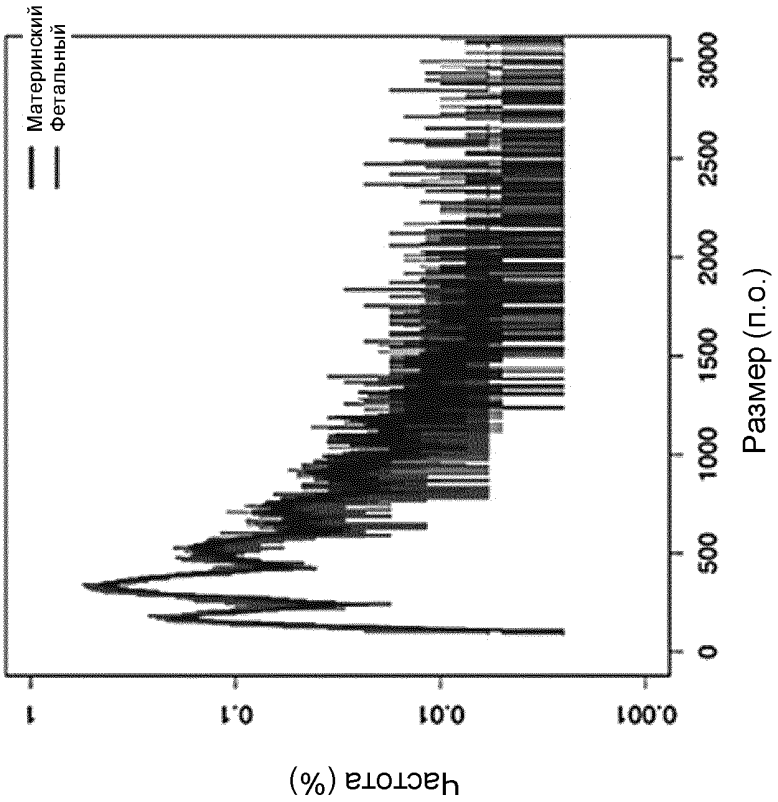
ФИГ. 100

Размер фрагмента (п.о.)	M12970									
	Специфическая для плода ДНК					Специфическая для матери ДНК				
	Кол-во фрагментов	Частота (%)	Метилированный CpG	Неметилированный CpG	Уровень метилирования (%)	Кол-во фрагментов	Частота (%)	Метилированный CpG	Неметилированный CpG	Уровень метилирования (%)
>=500	5328	31.2%	9887	5633	63.7%	16464	41.2%	39224	19781	66.5%
>=600	3715	21.8%	8366	4622	64.4%	11927	29.8%	34222	17134	66.6%
>=1000	1596	9.3%	4901	2434	66.8%	5693	14.2%	23366	10516	69.0%
>=2000	500	2.9%	2003	839	70.5%	1793	4.5%	10926	4327	71.6%

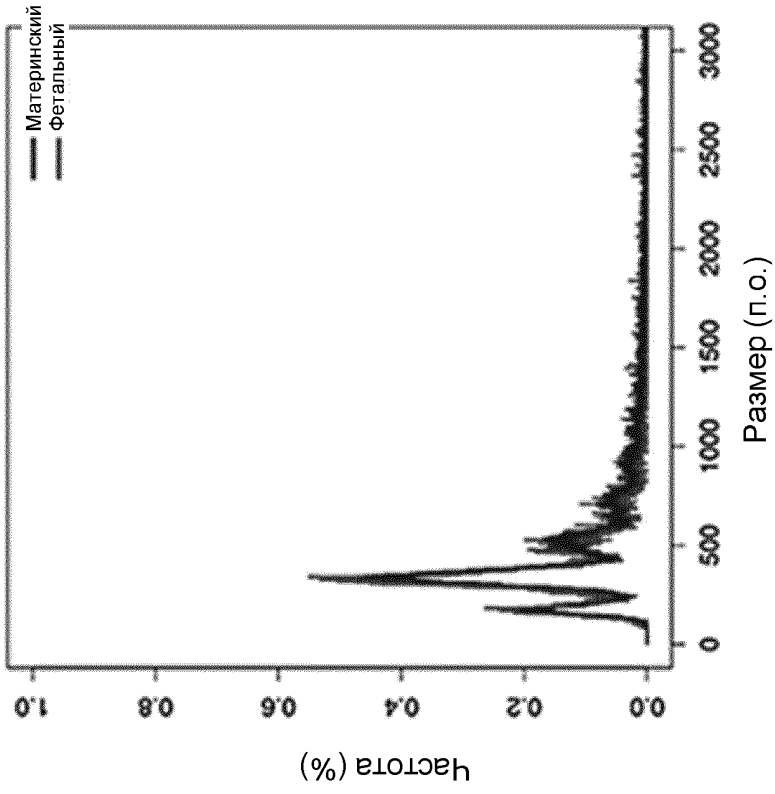
Размер фрагмента (п.о.)	M12985									
	Специфическая для плода ДНК					Специфическая для матери ДНК				
	Кол-во фрагментов	Частота (%)	Метилированный CpG	Неметилированный CpG	Уровень метилирования (%)	No. of fragments	Частота (%)	Метилированный CpG	Неметилированный CpG	Уровень метилирования (%)
>=500	1261	17.6%	3789	2092	64.4%	5132	20.4%	19235	9055	68.0%
>=600	749	10.4%	2788	1539	64.4%	3157	12.5%	15039	6870	68.6%
>=1000	290	4.0%	1502	854	63.8%	1193	4.7%	8718	3433	71.7%
>=2000	82	1.1%	597	433	58.0%	317	1.3%	3448	1174	74.6%

Размер фрагмента (п.о.)	M12969									
	Специфическая для плода ДНК					maternal-specific DNA				
	Кол-во фрагментов	Частота (%)	Метилированный CpG	Неметилированный CpG	Уровень метилирования (%)	Кол-во фрагментов	Частота (%)	Метилированный CpG	Неметилированный CpG	Уровень метилирования (%)
>=500	4713	28.8%	14412	8634	62.5%	14762	33.0%	53302	26440	66.8%
>=600	3200	19.6%	11820	7188	62.2%	10128	22.6%	44137	21631	67.1%
>=1000	1418	8.7%	7283	4380	62.4%	4600	10.3%	28579	13254	68.3%
>=2000	449	2.7%	3205	1594	66.8%	1403	3.1%	13004	4950	72.4%

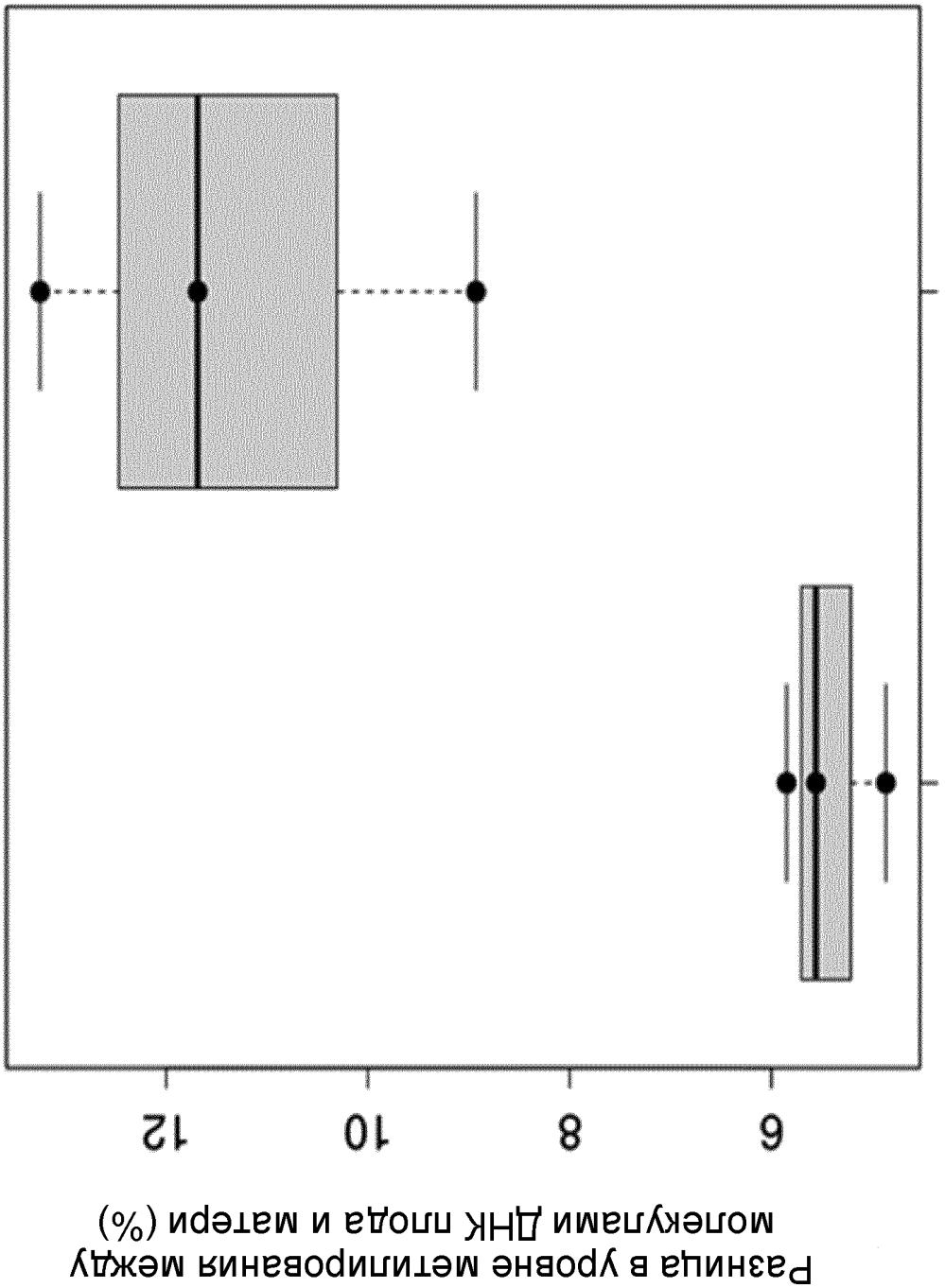
ФИГ. 101



ФИГ. 102B



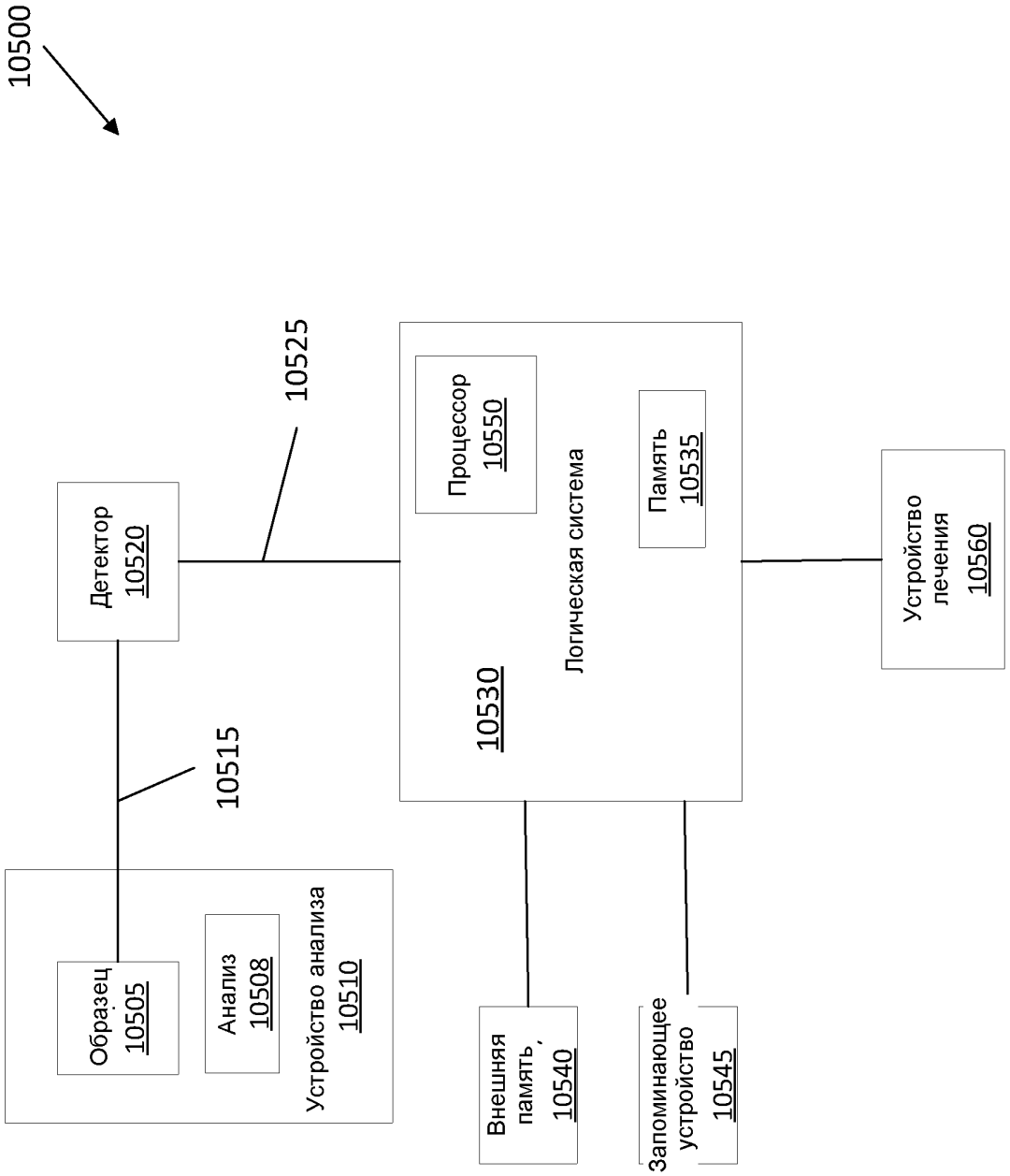
ФИГ. 102A


Разница в уровне метилирования между
молекулами ДНК плода и матери (%)
6
8
10
12
На основе одного
информативного ОНП
На основе двух
информативных ОНП

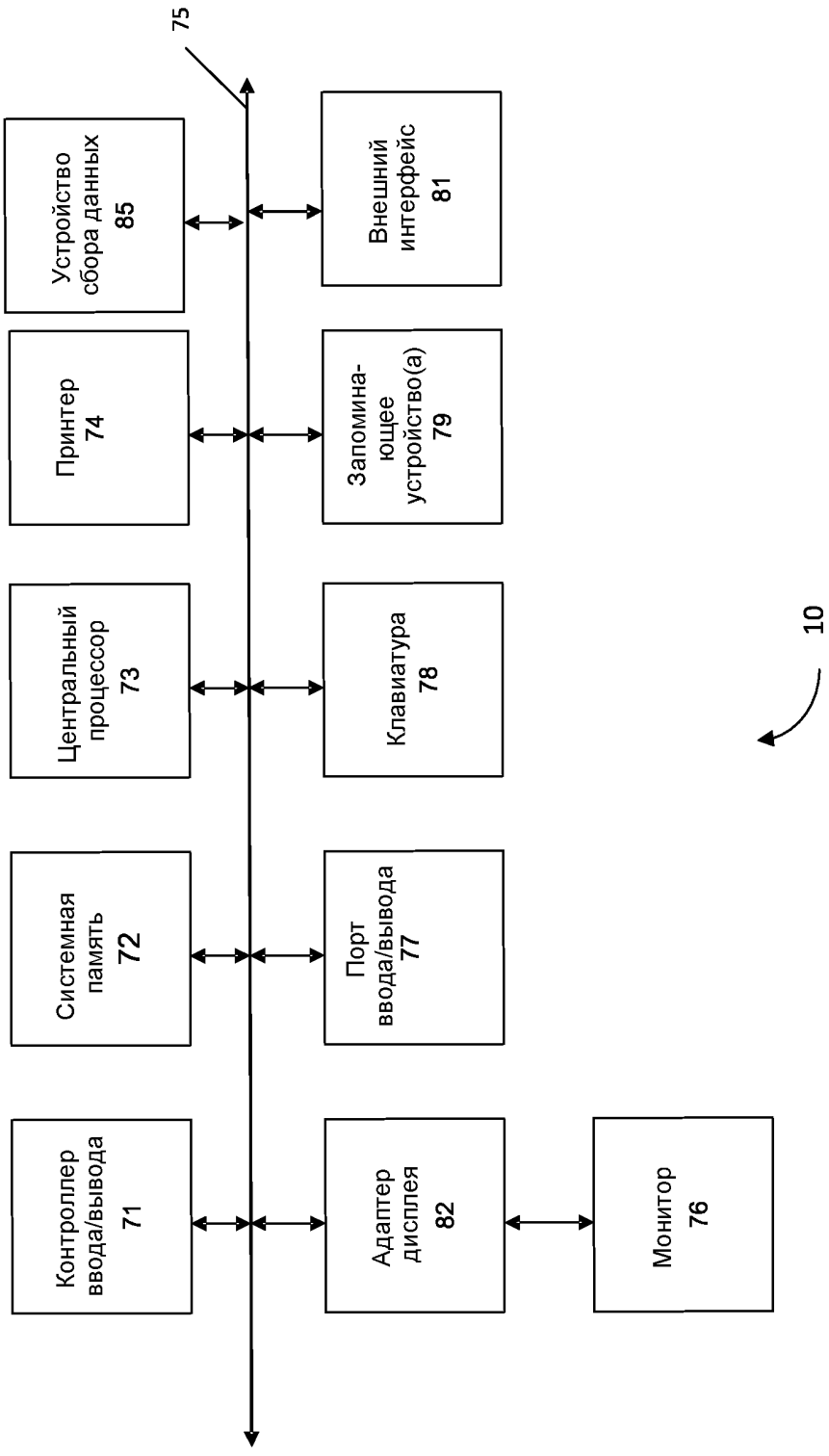
ФИГ. 103

Образец	Специфическая для плода ДНК			Специфическая для матери ДНК		
	Метилированный CpG	Неметилированный CpG	Уровень метилирования (%)	Метилированный CpG	Неметилированный CpG	Уровень метилирования (%)
M12970	682	522	56.64	8,634	4,001	68.33
M12985	245	177	58.06	4,175	1,680	71.31
M12969	1,065	751	58.65	13,399	6,429	67.58

ФИГ. 104


10500
10525
10515
Детектор
10520
Процессор
10550
Память
10535
Логическая система
10530
Устройство
лечения
10560
Образец
10505
Анализ
10508
Устройство анализа
10510
Внешняя
память,
10540
Запоминающее
устройство
10545

ФИГ. 105


75
Устройство сбора данных 85
Внешний интерфейс 81
Принтер 74
Запомина-ющее устройство(а) 79
Центральный процессор 73
Клавиатура 78
Системная память 72
Порт ввода/вывода 77
Контроллер ввода/вывода 71
Адаптер дисплея 82
Монитор 76
10

ФИГ. 106