(19) Евразийское патентное ведомство

(11) 047318 (13) B1

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ЕВРАЗИЙСКОМУ ПАТЕНТУ

(45) Дата публикации и выдачи патента
2024.07.01

(21) Номер заявки
202190610

(22) Дата подачи заявки
2020.08.17

(51) Int. Cl. ***C12Q 1/68*** (2018.01)
C12N 15/11 (2006.01)

(54) ОПРЕДЕЛЕНИЕ МОДИФИКАЦИЙ ОСНОВАНИЙ НУКЛЕИНОВЫХ КИСЛОТ

(31) **62/887,987; 62/970,586; 62/991,891; 63/019,790; 63/051,210**
(32) **2019.08.16; 2020.02.05; 2020.03.19; 2020.05.04; 2020.07.13**
(33) **US**
(43) **2021.06.02**
(86) **PCT/CN2020/109602**
(87) **WO 2021/032060 2021.02.25**

(71)(73) Заявитель и патентовладелец:
ТЕ ЧАЙНИЗ ЮНИВЕРСИТИ ОВ ГОНКОНГ (CN)

(72) Изобретатель:
Ло Юйк-Мин Деннис, Чиу Росса Вай Квун, Чань Квань Чэ, Цзян Пэйюн, Чэн Сук Хан, Пэн Вэньлэй, Це Он И (CN)

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

(56) WO-A2-2010068289
WO-A2-2010027484
WO-A1-2014153757
WO-A1-2013188846
CN-A-104053789
WO-A1-2013170429
US-A1-2013230909
Benjamin A.F. "Structural and Mechanistic Basis for Anaerobic Ergothioneine Biosynthesis". NATURE METHODS, Vol. 7, No. 6, 09 May 2010 (2010-05-09), abstract, page 462, left column, paragraph 2 - page 465, left column, paragraph 2

(57) В данном документе описаны системы и способы использования определения модификации оснований при анализе молекул нуклеиновых кислот и сборе данных для анализа молекул нуклеиновых кислот. Модификации оснований могут включать в себя метилирование. Способы определения модификаций оснований могут включать в себя использование характеристик, полученных в результате секвенирования. Эти характеристики могут включать в себя ширину импульса оптического сигнала от оснований, которые секвенируют, межимпульсный период оснований и тип оснований. Модели машинного обучения можно обучить обнаруживать модификации оснований с помощью этих характеристик. Относительные уровни модификации или метилирования между гаплотипами могут указывать на нарушение. Статусы модификаций или метилирования также могут быть использованы для обнаружения химерных молекул.

047318 B1

047318 B1

Заявка на данный патент заявляет приоритет по предварительной заявке США № 63/051210, озаглавленной как "ОПРЕДЕЛЕНИЕ МОДИФИКАЦИЙ ОСНОВАНИЙ НУКЛЕИНОВЫХ КИСЛОТ", поданной 13 июля 2020 г.; предварительной заявке США № 63/019790, озаглавленной как "ОПРЕДЕЛЕНИЕ МОДИФИКАЦИЙ ОСНОВАНИЙ НУКЛЕИНОВЫХ КИСЛОТ", поданной 4 мая 2020 г.; предварительной заявке США № 62/991891, озаглавленной как "ОПРЕДЕЛЕНИЕ МОДИФИКАЦИЙ ОСНОВАНИЙ НУКЛЕИНОВЫХ КИСЛОТ", поданной 19 марта 2020 г.; предварительной заявке США № 62/970586, озаглавленной как "ОПРЕДЕЛЕНИЕ МОДИФИКАЦИЙ ОСНОВАНИЙ НУКЛЕИНОВЫХ КИСЛОТ", поданной 5 февраля 2020 г.; и предварительной заявке США № 62/887987, озаглавленной как "ОПРЕДЕЛЕНИЕ МОДИФИКАЦИЙ ОСНОВАНИЙ НУКЛЕИНОВЫХ КИСЛОТ", поданной 16 августа 2019 г., полное содержание которых включено в данный документ посредством ссылки для всех целей.

Уровень техники

Присутствие модификаций оснований в нуклеиновых кислотах различается у разных организмов, включая вирусы, бактерии, растения, грибы, нематоды, насекомые и позвоночные (например, люди) и т.д. Наиболее распространенные модификации оснований представляют собой добавление метильной группы к различным основаниям ДНК в разных позициях, так называемое метилирование. Метилирование было обнаружено на цитозинах, аденинах, тиминах и гуанинах, например, 5mC (5-метилцитозин), 4mC (N4-метилцитозин), 5hmC (5-гидроксиметилцитозин), 5fC (5-формилцитозин), 5caC (5-карбоксилцитозин), 1mA (N1-метиладенин), 3mA (N3-метиладенин), 7mA (N7-метиладенин), 3mC (N3-метилцитозин), 2mG (N2-метилгуанин), 6mG (O6-метилгуанин), 7mG (N7-метилгуанин), 3mT (N3-метилтимин) и 4mT (O4-метилтимин). В геномах позвоночных 5mC является наиболее распространенным типом метилирования оснований, за ним следует метилирование гуанина (т.е. в контексте CpG).

Метилирование ДНК является важным для развития млекопитающих и играет значимую роль в генной экспрессии и сайленсинге, эмбриональном развитии, транскрипции, структуре хроматина, инактивации X-хромосомы, защите от активности повторяющихся элементов, поддержании стабильности генома во время митоза, и регуляции геномного импринтинга, родительского по происхождению.

Метилирование ДНК играет важную роль в скоординированном сайленсинге промоторов и энхансеров (Robertson, 2005; Smith and Meissner, 2013). Было обнаружено, что многие заболевания человека связаны с аберрациями метилирования ДНК, включающие в себя, но не ограниченные лишь этими: процесс канцерогенеза, нарушения импринтинга (например, синдром Беквита-Видемана и синдром Прадера-Вилли), болезни нестабильности повторов (например, синдром ломкой X-хромосомы), аутоиммунные заболевания (например, системная красная волчанка), нарушения обмена веществ (например, диабет типа I и типа II), неврологические нарушения, старение и т.д.

Точное определение метильной модификации на молекулах ДНК будет иметь множество клинических применений. Один из широко используемых способов определения метилирования ДНК представляет собой бисульфитное секвенирование (БС-секв.) (Lister et al., 2009; Frommer et al., 1992). В этом подходе образцы ДНК сначала обрабатывают бисульфитом, который преобразовывает неметилированный цитозин (т.е. C) в урацил. Напротив, метилированный цитозин остается неизменным. Затем модифицированную бисульфитом ДНК анализируют с помощью секвенирования ДНК. В другом подходе, после преобразования бисульфитом, модифицированную ДНК затем подвергают амплификации с помощью полимеразной цепной реакции (ПНР) с использованием праймеров, которые могут дифференцировать ДНК, преобразованную бисульфитом, с разными профилями метилирования (Herman et al., 1996). Последний подход называют специфичной к метилированию ПЦР.

Одним из недостатков таких подходов на основе бисульфита является то, что, как сообщается, стадия преобразования бисульфитом значительно разрушает большую часть обработанной ДНК (Grunau, 2001). Другой недостаток заключается в том, что стадия преобразования бисульфитом может создавать сильное CG-смещение (Olova et al., 2018), что приводит к снижению соотношения сигнал/шум, как правило, для смесей ДНК с гетерогенными состояниями метилирования. Кроме того, бисульфитное секвенирование не способно секвенировать длинные молекулы ДНК из-за деградации ДНК во время обработки бисульфитом. Таким образом, существует потребность в определении модификации оснований нуклеиновых кислот без предварительного химического воздействия (например, преобразования бисульфитом) и амплификации нуклеиновой кислоты (например, с использованием ПЦР).

Краткое описание сущности изобретения

Мы разработали новый способ, который в одном варианте осуществления, позволяет определять модификации оснований, такие как 5mC, в нуклеиновых кислотах без предварительной обработки ДНК-матрицы, например ферментативного и/или химического преобразования, или связывания с белками и/или антителами. Хотя такая предварительная обработка ДНК-матрицы не является необходимой для определения модификаций оснований, в приведенных примерах определенная предварительная обработка (например, расщепление ферментами рестрикции) может служить для улучшения аспектов изобретения (например, позволяя обогащать CpG сайты для анализа). Варианты осуществления, представленные в данном изобретении, могут использоваться для обнаружения различных типов модификации оснований, например, включающих в себя, но не ограничивающихся лишь этими: 4mC, 5hmC, 5fC и 5caC, 1mA, 3mA, 7mA, 3mC, 2mG, 6mG, 7mG, 3mT, и 4mT, и т.д. Такие варианты осуществления могут использовать

характеристики, полученные в результате секвенирования, например, кинетические характеристки, на которые влияют различные модификации оснований, а также тип нуклеотидов в окне вокруг целевой позиции, чей статус метилирования определяют.

Варианты осуществления данного изобретения могут быть использованы для, но не ограничиваются секвенированием отдельной молекулы. Одним из типов секвенирования отдельной молекулы является секвенирование отдельной молекулы в реальном времени, при котором в реальном времени отслеживают прогресс секвенирования отдельной молекулы ДНК. Один из типов секвенирования отдельной молекулы в реальном времени является тот, который выведен на рынок Pacific Biosciences, использующей их систему отдельной молекулы в реальном времени (SMRT). Способы могут использовать ширину импульса сигнала от оснований, которые секвенируют, межимпульсный период (МИП) оснований, и идентичность оснований для обнаружения модификации в основании или в соседнем основании. Еще одна система отдельной молекулы основана на нанопоровом секвенировании. Одним из примеров системы нанопорового секвенирования является система, выпускаемая Oxford Nanopore Technologies.

Разработанные нами способы могут служить в качестве инструментов для обнаружения модификаций оснований в биологических образцах для оценки профилей метилирования в образцах для различных целей, включающих в себя, но неограниченных исследованиями и диагностическими целями. Обнаруженные профили метилирования можно использовать для различного анализа. Профили метилирования могут использоваться для определения источника ДНК (например, материнского или эмбрионального, тканевого, бактериального, или ДНК, полученной из опухолевых клеток, обогащенных из крови больного раком). Обнаружение профилей аберрантного метилирования в тканях помогает идентифицировать нарушения развития у индивидов, выявлять и прогнозировать опухоли или злокачественные новообразования.

Варианты осуществления данного изобретения могут включать в себя анализ относительных уровней метилирования гаплотипов организма. Дисбаланс в уровнях метилирования между двумя гаплотипами может быть использован для определения классификации нарушения. Более высокий дисбаланс может указывать на наличие нарушения или более серьезного нарушения. Нарушение может включать в себя рак.

Паттерны метилирования в одной молекуле позволяют идентифицировать химерную и гибридную ДНК. Химерные и гибридные молекулы могут включать в себя последовательности из двух разных генов, хромосом, органелл (например, митохондрий, ядра, хлоропластов), организмов (млекопитающих, бактерий, вирусов и т.д.) и/или видов. Обнаружение областей соединения химерных или гибридных молекул ДНК может сделать возможным обнаружение слияния генов для различных нарушений или заболеваний, включая рак, пренатальные или врожденные нарушения.

Лучшее понимание природы и преимуществ вариантов осуществления данного изобретения может быть получено путем отсылки к следующему подробному описанию и прилагаемым графическим материалам.

Краткое описание графических материалов

Фиг. 1 иллюстрирует SMRT-секвенирование молекул, несущих модификации оснований, согласно вариантам осуществления данного изобретения.

Фиг. 2 иллюстрирует SMRT-секвенирование молекул, несущих метилированные и неметилированные сайты CpG, согласно вариантам осуществления данного изобретения.

Фиг. 3 иллюстрирует межимпульсный период и ширину импульса согласно вариантам осуществления данного изобретения.

Фиг. 4 демонстрирует пример окна измерения цепи Уотсона ДНК для обнаружения модификации основания согласно вариантам осуществления данного изобретения.

Фиг. 5 демонстрирует пример окна измерения цепи Крика ДНК для обнаружения модификации основания согласно вариантам осуществления данного изобретения.

Фиг. 6 демонстрирует пример окна измерения путем объединения данных для цепи Уотсона ДНК и ее комплементарной цепи Крика для обнаружения любой модификации основания согласно вариантам осуществления данного изобретения.

Фиг. 7 демонстрирует пример окна измерения путем объединения данных для цепи Уотсона ДНК и ее близлежащей области цепи Крика для обнаружения любой модификации основания согласно вариантам осуществления данного изобретения.

Фиг. 8 демонстрирует примеры окон измерения цепи Уотсона, цепи Крика и обеих цепей для определения состояний метилирования в сайтах CpG согласно вариантам осуществления данного изобретения.

Фиг. 9 демонстрирует общую процедуру построения аналитических, вычислительных, математических или статистических моделей для классификации модификаций оснований согласно вариантам осуществления данного изобретения.

Фиг. 10 демонстрирует общую процедуру классификации модификаций оснований согласно вариантам осуществления данного изобретения.

Фиг. 11 демонстрирует общую процедуру построения аналитических, вычислительных, математи-

ческих или статистических моделей для классификации состояний метилирования в сайтах CpG с использованием образцов с известными состояниями метилирования цепи Уотсона согласно вариантам осуществления данного изобретения.

Фиг. 12 демонстрирует общую процедуру классификации состояний метилирования цепи Уотсона для неизвестного образца согласно вариантам осуществления данного изобретения.

Фиг. 13 демонстрирует общую процедуру построения аналитических, вычислительных, математических или статистических моделей для классификации состояний метилирования в сайтах CpG с использованием образцов с известными состояниями метилирования цепи Крика согласно вариантам осуществления данного изобретения.

Фиг. 14 демонстрирует общую процедуру классификации состояний метилирования цепи Крика для неизвестного образца согласно вариантам осуществления данного изобретения.

Фиг. 15 демонстрирует общую процедуру построения статистических моделей для классификации состояний метилирования в сайтах CpG с использованием образцов с известными состояниями метилирования из обеих цепей - Уотсона и Крика, согласно вариантам осуществления данного изобретения.

Фиг. 16 демонстрирует общую процедуру классификации состояний метилирования неизвестного образца из обеих цепей - Уотсона и Крика, согласно вариантам осуществления данного изобретения.

Фиг. 17A и 17B демонстрируют эффективность для обучающего набора данных и тестового набора данных для определения метилирования согласно вариантам осуществления данного изобретения.

Фиг. 18 демонстрирует эффективность для обучающего набора данных и тестового набора данных для определения метилирования согласно вариантам осуществления данного изобретения.

Фиг. 19 демонстрирует эффективность для обучающего набора данных и тестового набора данных при различных глубинах секвенирования для определения метилирования согласно вариантам осуществления данного изобретения.

Фиг. 20 демонстрирует эффективность для обучающего набора данных и тестового набора данных для различных цепей для определения метилирования согласно вариантам осуществления данного изобретения.

Фиг. 21 демонстрирует эффективность для обучающего набора данных и тестового набора данных для различных окон измерения для определения метилирования согласно вариантам осуществления данного изобретения.

Фиг. 22 демонстрирует эффективность для обучающего набора данных и тестового набора данных для различных окон измерения с использованием только нисходящих оснований для определения метилирования согласно вариантам осуществления данного изобретения.

Фиг. 23 демонстрирует эффективность для обучающего набора данных и тестового набора данных для различных окон измерения с использованием только восходящих оснований для определения метилирования согласно вариантам осуществления данного изобретения.

Фиг. 24 демонстрирует эффективность анализа метилирования с использованием кинетических паттернов, связанных с нисходящими и восходящими основаниями, с использованием асимметричных фланкирующих форматов в обучающем наборе данных согласно вариантам осуществления данного изобретения.

Фиг. 25 демонстрирует эффективность анализа метилирования с использованием кинетических паттернов, связанных с нисходящими и восходящими основаниями, с использованием асимметричных фланкирующих форматов в тестовом наборе данных согласно вариантам осуществления данного изобретения.

Фиг. 26 демонстрирует относительную важность признаков в отношении классификации состояний метилирования в сайтах CpG согласно вариантам осуществления данного изобретения.

Фиг. 27 демонстрирует эффективность анализа МИП на основе мотивов для обнаружения метилирования без использования сигнала ширины импульса согласно вариантам осуществления данного изобретения.

Фиг. 28 представляет собой график методики анализа главных компонентов с использованием 2 нуклеотидов выше и 6 нуклеотидов ниже цитозина, который подвергается анализу метилирования согласно вариантам осуществления данного изобретения.

Фиг. 29 представляет собой график сравнения эффективности между способом, использующим анализ главных компонентов, и способом, использующим сверточную нейронную сеть, согласно вариантам осуществления данного изобретения.

Фиг. 30 демонстрирует эффективность для обучающего набора данных и тестового набора данных для различных аналитических, вычислительных, математических и статистических моделей с использованием только восходящих оснований для определения метилирования согласно вариантам осуществления данного изобретения.

Фиг. 31A демонстрирует пример одного подхода для создания молекул с неметилированными аденинами путем амплификации всего генома согласно вариантам осуществления данного изобретения.

Фиг. 31B демонстрирует пример одного подхода для создания молекул с метилированными аденинами путем амплификации всего генома согласно вариантам осуществления данного изобретения.

Фиг. 32A и 32B демонстрируют значения межимпульсного периода (МИП) для секвенированных оснований A в ДНК-матрице цепи Уотсона между неметилированными и метилированными наборами данных согласно вариантам осуществления данного изобретения.

Фиг. 32C демонстрирует кривую операционных характеристик приёмника для определения метилирования в цепи Уотсона согласно вариантам осуществления данного изобретения.

Фиг. 33A и 33B демонстрируют значения межимпульсного периода (МИП) для секвенированных оснований A в ДНК-матрице цепи Крика между неметилированными и метилированными наборами данных согласно вариантам осуществления данного изобретения.

Фиг. 33C демонстрирует кривую операционных характеристик приёмника для определения метилирования в цепи Крика согласно вариантам осуществления данного изобретения.

Фиг. 34 иллюстрирует определение 6mA цепи Уотсона согласно вариантам осуществления данного изобретения.

Фиг. 35 иллюстрирует определение 6mA цепи Крика согласно вариантам осуществления данного изобретения.

Фиг. 36A и фиг. 36B демонстрируют выясненную вероятность метилирования для секвенированных оснований A цепи Уотсона между наборами данных uA и mA с использованием модели сверточной нейронной сети на основе окна измерения согласно вариантам осуществления данного изобретения.

Фиг. 37 демонстрирует кривую ROC для определения 6mA с использованием модели CHC на основе окна измерения для секвенированных оснований A цепи Уотсона согласно вариантам осуществления данного изобретения.

Фиг. 38 демонстрирует сравнение эффективности между обнаружением 6mA на основе метрики МИП и обнаружением 6mA на основе окна измерения согласно вариантам осуществления данного изобретения.

Фиг. 39A и 39B демонстрируют выясненную вероятность метилирования для тех секвенированных оснований A цепи Крика между наборами данных uA и mA с использованием модели CHC на основе окна измерения согласно вариантам осуществления данного изобретения.

Фиг. 40 демонстрирует эффективность определения 6mA при применнии модели CHC на основе окна измерения к секвенированным основаниям A цепи Крика согласно вариантам осуществления данного изобретения.

Фиг. 41 демонстрирует примеры состояний метилирования для оснований A в молекуле, включающей в себя цепи Уотсона и Крика, согласно вариантам осуществления данного изобретения.

Фиг. 42 демонстрирует пример улучшенного обучения путем выборочного использования оснований A в наборе данных mA со значениями МИП, превышающими его 10-й процентиль, согласно вариантам осуществления данного изобретения.

Фиг. 43 представляет собой график процентного содержания неметилированных аденинов в наборе данных mA в зависимости от количества субпрочтений (subreads) в каждой ячейке согласно вариантам осуществления данного изобретения.

Фиг. 44 демонстрирует метиладениновые паттерны между цепями Уотсона и Крика двухцепочечной молекулы ДНК в тестовом наборе данных согласно вариантам осуществления данном изобретения.

Фиг. 45 представляет собой таблицу, показывающую процент полностью неметилированных молекул, полуметилированных молекул, полностью метилированных молекул и молекул с чередующимися метиладениновыми паттернами в обучающих и тестовых наборах данных, согласно вариантам осуществления данного изобретения.

Фиг. 46 иллюстрирует показательные примеры молекул с полностью неметилированными молекулами в отношении аденинового сайтов, полуметилированных молекул, полностью метилированных молекул и молекул с чередующимися метиладениновыми паттернами согласно вариантам осуществления данного изобретения.

Фиг. 47 демонстрирует пример длинного прочтения (read) (6265 п.о.), несущего островок CpG (заштрихованный желтым) согласно вариантам осуществления данного изобретения.

Фиг. 48 представляет собой таблицу, показывающую, что с помощью SMRT-секвенирования Pacific Biosciences были секвенированы 9 молекул ДНК и они перекрывались с областями импринтинга согласно вариантам осуществления данного изобретения.

Фиг. 49 демонстрирует пример геномного импринтинга согласно вариантам осуществления данного изобретения.

Фиг. 50 демонстрирует пример определения паттернов метилирования в области импринтинга согласно вариантам осуществления данного изобретения.

Фиг. 51 демонстрирует сравнение уровней метилирования, выясненных с помощью нового подхода и обычного бисульфитного секвенирования согласно вариантам осуществления данного изобретения.

Фиг. 52 демонстрирует эффективность обнаружения метилирования ДНК плазмы согласно вариантам осуществления данного изобретения. (A) Взаимосвязь между предсказанной вероятностью метилирования в сравнении с диапазонами уровней метилирования, определенными количественно с помощью бисульфитного секвенирования. (B) Корреляция между уровнями метилирования, определенными с по-

мощью секвенирования Pacific Biosciences (PacBio) согласно вариантам осуществления, представленным в данном изобретении (ось ординат), и уровнями метилирования, количественно определенными с помощью бисульфитного секвенирования (ось абсцисс) с разрешением 10 млн.п.о.

Фиг. 53 демонстрирует корреляцию геномного представления (ГП) Y-хромосомы между SMRT-секвенированием Pacific Biosciences и БС-секв. согласно вариантам осуществления данного изобретения.

Фиг. 54 демонстрирует пример обнаружения метилирования на основе блока CpG с использованием блоков CpG, каждый из которых содержит серию сайтов CpG согласно вариантам осуществления данного изобретения. 5mC: метилирование; C: неметилирование.

Фиг. 55 демонстрирует обучение и тестирование распознавания метилирования для молекул ДНК человека, с использованием подхода, основанного на блоках CpG, согласно вариантам осуществления данного изобретения. (A) Эффективность в обучающем наборе данных. (B) Эффективность в независимом тестовом наборе данных.

Фиг. 56A и 56B демонстрируют изменения числа копий в опухолевой ткани согласно вариантам осуществления данного изобретения.

Фиг. 57A и 57B демонстрируют изменения числа копий в опухолевой ткани согласно вариантам осуществления данного изобретения.

Фиг. 58 демонстрирует схематическую иллюстрацию тканевого картирования ДНК плазмы из плазмы беременной женщины с использованием уровней метилирования, выявленных согласно вариантам осуществления данного изобретения.

Фиг. 59 демонстрирует корреляцию между вкладом плаценты в ДНК материнской плазмы и фракцией ДНК плода, как определено по прочтениям Y-хромосомы согласно вариантам осуществления данного изобретения.

Фиг. 60 демонстрирует таблицу, обобщающую данные секвенирования из различных образцов ДНК тканей человека согласно вариантам осуществления данного изобретения.

Фиг. 61 демонстрирует иллюстрацию различных способов анализа паттернов метилирования согласно вариантам осуществления данного изобретения.

Фиг. 62A и 62B демонстрируют сравнение степеней метилирования на уровне всего генома, количественно определяемых бисульфитным секвенированием и секвенированием отдельной молекулы в реальном времени, согласно вариантам осуществления данного изобретения.

Фиг. 63A, 63B и 63C демонстрируют различные корреляции общих уровней метилирования, количественно определяемых бисульфитным секвенированием и секвенированием отдельной молекулы в реальном времени согласно вариантам осуществления данного изобретения.

Фиг. 64A и 64B демонстрируют паттерны метилирования при разрешении 1-Мнт (млн. нуклеотидов) для линии клеток гепатоцеллюлярной карциномы (ГНК) и образца лейкоцитарного слоя из здорового контрольного субъекта, с уровнями метилирования, определенными с помощью бисульфитного секвенирования и секвенирования отдельной молекулы в реальном времени, согласно вариантам осуществления данного изобретения.

Фиг. 65A и 65B демонстрируют диаграммы рассеяния уровней метилирования с разрешением 1-Мнт, определенных с помощью бисульфитного секвенирования и секвенирования отдельной молекулы в реальном времени согласно вариантам осуществления данного изобретения, для линии клеток ГЦК (HepG2) и образца лейкоцитарного слоя из здорового контрольного субъекта.

Фиг. 66A и 66B демонстрируют диаграммы рассеяния уровней метилирования с разрешением 100-кнт, определенных с помощью бисульфитного секвенирования и секвенирования отдельной молекулы в реальном времени согласно вариантам осуществления данного изобретения, для линии клеток ГЦК (HepG2) и образца лейкоцитарного слоя из здорового контрольного субъекта.

Фиг. 67A и 67B демонстрируют паттерны метилирования при разрешении 1-Мнт для опухолевой ткани ГЦК и прилегающей нормальной ткани с уровнями метилирования, определенными с помощью бисульфитного секвенирования и секвенирования отдельной молекулы в реальном времени, согласно вариантам осуществления данного изобретения.

Фиг. 68A и 68B демонстрируют диаграммы рассеяния уровней метилирования при разрешении 1-Мнт, определенных с помощью бисульфитного секвенирования и секвенирования отдельной молекулы в реальном времени, согласно вариантам осуществления данного изобретения, для опухолевой ткани ГЦК и прилегающей нормальной ткани.

Фиг. 69A и 69B демонстрируют диаграммы рассеяния уровней метилирования при разрешении 100-кнт, определенных с помощью бисульфитного секвенирования и секвенирования отдельной молекулы в реальном времени, согласно вариантам осуществления данного изобретения, для опухолевой ткани ГЦК и прилегающей нормальной ткани.

Фиг. 70A и 70B демонстрируют паттерны метилирования при разрешении 1-Мнт для опухолевой ткани ГЦК и прилегающей нормальной ткани с уровнями метилирования, определенными с помощью бисульфитного секвенирования и секвенирования отдельной молекулы в реальном времени, согласно вариантам осуществления данного изобретения.

Фиг. 71A и 71B демонстрируют диаграммы рассеяния уровней метилирования при разрешении 1-

Мнт, определенных с помощью бисульфитного секвенирования и секвенирования отдельной молекулы в реальном времени, согласно вариантам осуществления данного изобретения, для опухолевой ткани ГЦК и прилегающей нормальной ткани.

Фиг. 72A и 72B демонстрируют диаграммы рассеяния уровней метилирования при разрешении 100-кнт, определенных с помощью бисульфитного секвенирования и секвенирования отдельной молекулы в реальном времени, согласно вариантам осуществления данного изобретения, для опухолевой ткани ГЦК и прилегающей нормальной ткани.

Фиг. 73 демонстрирует пример аберрантного паттерна метилирования рядом с геном-супрессором опухоли CDKN2A согласно вариантам осуществления данного изобретения.

Фиг. 74A и 74B демонстрируют области неодинакового метилирования, обнаруженные с помощью секвенирования отдельной молекулы в реальном времени согласно вариантам осуществления данного изобретения.

Фиг. 75 демонстрирует паттерны метилирования ДНК вируса гепатита B между тканями ГНК и прилегающими неопухолевыми тканями с использованием секвенирования отдельной молекулы в реальном времени согласно вариантам осуществления данного изобретения.

Фиг. 76A демонстрирует уровни метилирования ДНК вируса гепатита B в тканях печени пациентов с циррозом, но без ГЦК, с использованием бисульфитного секвенирования согласно вариантам осуществления данного изобретения.

Фиг. 76B демонстрирует уровни метилирования ДНК вируса гепатита B в тканях ГЦК с использованием бисульфитного секвенирования согласно вариантам осуществления данного изобретения.

Фиг. 77 иллюстрирует анализ гаплотипа метилирования согласно вариантам осуществления данного изобретения.

Фиг. 78 демонстрирует распределение по размеру секвенированных молекул, как определено из консенсусных последовательностей согласно вариантам осуществления данного изобретения.

Фиг. 79A, 79B, 79C и 79D демонстрируют примеры паттернов аллельного метилирования в областях импринтинга согласно вариантам осуществления данного изобретения.

Фиг. 80A, 80B, 80C и 80D демонстрируют примеры паттернов аллельного метилирования в неподверженных импринтингу областях согласно вариантам осуществления данного изобретения.

Фиг. 81 демонстрирует таблицу уровней метилирования аллель-специфичных фрагментов согласно вариантам осуществления данного изобретения.

Фиг. 82 демонстрирует пример определения плацентарного происхождения ДНК из плазмы во время беременности с использованием профилей метилирования согласно вариантам осуществления данного изобретения.

Фиг. 83 иллюстрирует анализ метилирования плод-специфичной ДНК согласно вариантам осуществления данного изобретения.

Фиг. 84A, 84B и 84C демонстрируют эффективность различных размеров окна измерения для разных наборов реагентов для SMRT-секв. согласно вариантам осуществления данного изобретения.

Фиг. 85A, 85B и 85C демонстрируют эффективность различных размеров окна измерения для разных наборов реагентов для SMRT-секв. согласно вариантам осуществления данного изобретения.

Фиг. 86A, 86B и 86C демонстрируют корреляцию общих уровней метилирования, количественно определенных с помощью бисульфитного секвенирования и SMRT-секв. (Sequel II Sequencing Kit 2.0) согласно вариантам осуществления данного изобретения.

Фиг. 87A и 87B демонстрируют сравнение общего уровня метилирования между различными опухолевыми тканями и прилегающими неопухолевыми тканями в паре с ними, согласно вариантам осуществления данного изобретения.

Фиг. 88 демонстрирует определение статуса метилирования с использованием контекста последовательности, определенного из кольцевой консенсусной последовательности (ККП), согласно вариантам осуществления данного изобретения.

Фиг. 89 демонстрирует кривую ROC для обнаружения метилированных сайтов CpG с использованием контекста последовательности, определенного из ККП, согласно вариантам осуществления данного изобретения.

Фиг. 90 демонстрирует кривую ROC для обнаружения метилированных сайтов CpG без информации ККП и без предварительного выравнивания с эталонным геномом согласно вариантам осуществления данного изобретения.

Фиг. 91 демонстрирует пример получения молекул для секвенирования отдельной молекулы в реальном времени согласно вариантам осуществления данного изобретения.

Фиг. 92 представляет иллюстрацию системы CRISPR/Cas9 согласно вариантам осуществления данного изобретения.

Фиг. 93 демонстрирует пример комплекса Cas9 для внесения двух разрезов, охватывающих молекулу интереса с заблокированными концами, согласно вариантам осуществления данного изобретения.

Фиг. 94 демонстрирует распределение метилирования областей Alu, определенное с помощью бисульфитного секвенирования и секвенирования отдельной молекулы в реальном времени согласно вари-

антам осуществления данного изобретения.

Фиг. 95 демонстрирует распределение уровней метилирования областей Alu, определенных с помощью модели с использованием результатов секвенирования отдельной молекулы в реальном времени согласно вариантам осуществления данного изобретения.

Фиг. 96 представляет таблицу тканей и уровней метилирования областей Alu в тканях согласно вариантам осуществления данного изобретения.

Фиг. 97 демонстрирует кластерный анализ для различных типов рака с использованием сигналов метилирования, относящихся к повторам Alu, согласно вариантам осуществления данного изобретения.

Фиг. 98A и 98B демонстрируют влияние глубины прочтения на общую количественную оценку уровня метилирования в тестовых наборах данных, которые были получены путем амплификации всего генома и обработки M.SsssI согласно вариантам осуществления данного изобретения.

Фиг. 99 демонстрирует сравнение между общими уровнями метилирования, определенными с помощью SMRT-секв. (Sequel II Sequencing Kit 2.0) и БС-секв. с использованием различных пороговых значений субпрочтений согласно вариантам осуществления данного изобретения.

Фиг. 100 представляет таблицу, демонстрирующую влияние глубины субпрочтений на корреляцию уровней метилирования между двумя типами измерения - с помощью SMRT-секв. (Sequel II Sequencing Kit 2.0) и БС-секв., согласно вариантам осуществления данного изобретения.

Фиг. 101 демонстрирует распределение глубины субпрочтений относительно размеров фрагментов в данных, сгенерированных с помощью Sequel II Sequencing Kit 2.0 согласно вариантам осуществления данного изобретения.

Фиг. 102 демонстрирует способ обнаружения модификации нуклеотида в молекуле нуклеиновой кислоты согласно вариантам осуществления данного изобретения.

Фиг. 103 демонстрирует способ обнаружения модификации нуклеотида в молекуле нуклеиновой кислоты согласно вариантам осуществления данного изобретения.

Фиг. 104 иллюстрирует анализ относительного дисбаланса метилирования на основе гаплотипов согласно вариантам осуществления данного изобретения.

Фиг. 105A и 105B представляют собой таблицу гаплотипных блоков, демонстрирующую отличающиеся уровни метилирования между Гапл I и Гапл II в ДНК опухоли по сравнению с ДНК прилегающей неопухолевой ткани для случая TBR3033 согласно вариантам осуществления данного изобретения.

Фиг. 106 представляет собой таблицу гаплотипных блоков, демонстрирующую отличающиеся уровни метилирования между Гапл I и Гапл II в ДНК опухоли по сравнению с ДНК прилегающей нормальной ткани для случая TBR3032 согласно вариантам осуществления данного изобретения.

Фиг. 107A представляет собой таблицу, резюмирующую количество гаплотипных блоков, демонстрирующих дисбаланс метилирования между двумя гаплотипами между опухолью и прилегающими неопухолевыми тканями на основе данных, полученных с помощью Sequel II Sequencing Kit 2.0 согласно вариантам осуществления данного изобретения.

Фиг. 107B представляет собой таблицу, резюмирующую количество гаплотипных блоков, демонстрирующих дисбаланс метилирования между двумя гаплотипами в опухолевых тканях для разных стадий опухоли на основе данных, полученных с помощью Sequel II Sequencing Kit 2.0 согласно вариантам осуществления данного изобретения.

Фиг. 108 иллюстрирует анализ относительного дисбаланса метилирования на основе гаплотипов согласно вариантам осуществления данного изобретения.

Фиг. 109 демонстрирует способ классификации нарушения в организме, имеющем первый гаплотип и второй гаплотип, согласно вариантам осуществления данного изобретения.

Фиг. 110 иллюстрирует создание гибридных фрагментов человек-мышь, у которых человеческая часть является метилированной, в то время как мышиная часть является неметилированной, согласно вариантам осуществления данного изобретения.

Фиг. 111 иллюстрирует создание гибридных фрагментов человек-мышь, у которых человеческая часть является неметилированной, в то время как мышиная часть является метилированной, согласно вариантам осуществления данного изобретения.

Фиг. 112 демонстрирует распределение длин молекул ДНК в смеси ДНК (образец MIX01) после лигирования согласно вариантам осуществления данного изобретения.

Фиг. 113 иллюстрирует область соединения, с помощью которой первую ДНК (A) и вторую ДНК (B) соединяют вместе согласно вариантам осуществления данного изобретения.

Фиг. 114 иллюстрирует анализ метилирования смеси ДНК согласно вариантам осуществления данного изобретения.

Фиг. 115 демонстрирует диаграмму вероятностей метилирования для сайтов CpG в образце MIX01 согласно вариантам осуществления данного изобретения.

Фиг. 116 демонстрирует распределение длин молекул ДНК в смеси ДНК после перекрестного лигирования образца MIX02 согласно вариантам осуществления данного изобретения.

Фиг. 117 демонстрирует диаграмму вероятностей метилирования для сайтов CpG в образце MIX02 согласно вариантам осуществления данного изобретения.

Фиг. 118 представляет собой таблицу, в которой сравнивается метилирование, определенное бисульфитным секвенированием и секвенированием Pacific Biosciences для MIX01 согласно вариантам осуществления данного изобретения.

Фиг. 119 представляет собой таблицу, в которой сравнивается метилирование, определенное бисульфитным секвенированием и секвенированием Pacific Biosciences для MIX02 согласно вариантам осуществления данного изобретения.

Фиг. 120A и 120B демонстрируют уровни метилирования в 5-млн.п.о. группах для ДНК только человека и только мыши для MIX01 и MIX02 согласно вариантам осуществления данного изобретения.

Фиг. 121A и 121B демонстрируют уровни метилирования в 5-млн.п.о. группах для части человека и части мыши гибридных человеко-мышиных фрагментов ДНК для MIX01 и MIX02 согласно вариантам осуществления данного изобретения.

Фиг. 122A и 122B представляют собой иллюстративные графики, показывающие состояния метилирования в одной гибридной человеко-мышиной молекуле согласно вариантам осуществления данного изобретения.

Фиг. 123 демонстрирует способ обнаружения химерных молекул в биологическом образце согласно вариантам осуществления данного изобретения.

Фиг. 124 иллюстрирует систему измерения согласно вариантам осуществления данного изобретения.

Фиг. 125 демонстрирует блок-схему иллюстративной компьютерной системы, которую можно использовать с системами и способами согласно вариантам осуществления данного изобретения.

Фиг. 126 демонстрирует целевое секвенирование отдельной молекулы в реальном времени с обработкой MspI, с использованием репарации концов ДНК и добавления A-хвоста согласно вариантам осуществления данного изобретения.

Фиг. 127A и 127B демонстрируют распределение по размеру фрагментов, расщепленных MspI, согласно вариантам осуществления данного изобретения.

Фиг. 128 демонстрирует таблицу с количеством молекул ДНК для определенных выбранных диапазонов размеров согласно вариантам осуществления данного изобретения.

Фиг. 129 представляет собой график зависимости процента покрытия CpG-сайтов в пределах CpG-островков от размера фрагментов ДНК после расщепления рестрикционным ферментом согласно вариантам осуществления данного изобретения.

Фиг. 130 демонстрирует целевое секвенирование отдельной молекулы в реальном времени с обработкой MspI, без использования репарации концов ДНК и добавления A-хвоста согласно вариантам осуществления данного изобретения.

Фиг. 131 демонстрирует целевое секвенирование отдельной молекулы в реальном времени с обработкой MspI, со сниженной вероятностью самолигирования адаптера согласно вариантам осуществления данного изобретения.

Фиг. 132 представляет собой график совокупных уровней метилирования между образцами ДНК плаценты и лейкоцитарного слоя, определенных с помощью целевого секвенирования отдельной молекулы в реальном времени с обработкой MspI, согласно вариантам осуществления данного изобретения.

Фиг. 133 демонстрирует кластерный анализ плацентарных образцов и образцов лейкоцитарного слоя с использованием их профилей метилирования ДНК, определенных с помощью целевого секвенирования отдельной молекулы в реальном времени с обработкой MspI, согласно вариантам осуществления данного изобретения.

Термины.

"Ткань" соответствует группе клеток, которые объединяются вместе в функциональную единицу. В одной ткани можно найти больше чем один тип клеток. Различные типы тканей могут состоять из разных типов клеток (например, гепатоцитов, альвеолярных клеток или клеток крови), но также могут соответствовать тканям из разных организмов (мать против плода; ткани субъекта, перенесшего трансплантацию; ткани организма, инфицированного микроорганизмом или вирусом) или здоровым клеткам в сравнении с опухолевыми клетками. "Эталонные ткани" могут соответствовать тканям, используемым для определения тканеспецифичных уровней метилирования. Для определения тканеспецифичного уровня метилирования для этого типа ткани может быть использовано множество образцов ткани одного и того же типа от разных индивидов.

"Биологический образец" относится к любому образцу, взятому у человека. Биологический образец может представлять собой биопсию ткани, аспират тонкой иглой или клетки крови. Образец также может быть, например, плазмой, или сывороткой, или мочой беременной женщины. Также могут быть использованы образцы стула. В различных вариантах осуществления, большая часть ДНК в биологическом образце из беременной женщины, который был обогащен внеклеточной ДНК (например, образец плазмы, полученный с помощью протокола центрифугирования), может быть внеклеточной, например, больше чем 50, 60, 70, 80, 90, 95 или 99% ДНК могут быть внеклеточными. Протокол центрифугирования может включать в себя, например, 3000 g x10 мин, с получением жидкой части, и повторное центрифугирование, например, 30000 g в течение еще 10 мин для удаления остаточных клеток. В некоторых вариантах

осуществления, после стадии центрифугирования 3000 g может быть выполнена фильтрация жидкой части (например, с использованием фильтра с размером пор 5 мкм, или меньше, в диаметре).

"Прочтение последовательности" относится к цепочке нуклеотидов, секвенированной из любой части или всей молекулы нуклеиновой кислоты. Например, прочтение последовательности может представлять собой короткую цепочку нуклеотидов (например, 20-150), секвенированную из фрагмента нуклеиновой кислоты, короткую цепочкую нуклеотидов на одном или обеих концах фрагмента нуклеиновой кислоты, или секвенирование всего фрагмента нуклеиновой кислоты, который находится в биологическом образце. Прочтение последовательности может быть получен различными способами, например, с использованием методов секвенирования или с использованием зондов, например, в матрицах гибридизации или зондах захвата, или методов амплификации, таких как полимеразная цепная реакция (ПНР), или линейная амплификация с использованием единственного праймера, или изотермическая амплификация.

"Субпрочтение" представляет собой последовательность, генерируемую из всех оснований в одной цепи замкнутой в кольцо ДНК-матрицы, которая была скопирована в одну непрерывную цепь ДНК-полимеразой. Например, субпрочтение может соответствовать одной цепи замкнутой в кольцо ДНК-матрице ДНК. В таком случае, после замыкания в кольцо, одна двухцепочечная молекула ДНК будет иметь два субпрочтения: по одному на каждый проход секвенирования. В некоторых вариантах осуществления, генерируемая последовательность может включать в себя подмножество всех оснований в одной цепи, например, из-за наличия ошибок секвенирования.

"Сайт" (также называемый "геномным сайтом") соответствует одному сайту, который может представлять собой одну позицию основания или группу соотносящихся позиций оснований, например, сайт CpG или большую группу соотносящихся позиций оснований. "Локус" может соответствовать области, которая содержит множество сайтов. Локус может содержать только один сайт, что сделало бы локус эквивалентным сайту в этом контексте.

"Статус метилирования" относится к состоянию метилирования в данном 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 прочтений последовательности или зондов), показывающих метилирование в данном сайте, по сравнению с общим числом прочтений, охватывающих данный сайт. "Прочтение" может соответствовать информации (например, статусу метилирования в сайте), полученной из фрагмента ДНК. Прочтение можно получить с использованием реагентов (например, праймеров или зондов), которые предпочтительно гибридизируются с фрагментами ДНК с определенным статусом метилирования в одном или большем количестве сайтов. Обычно такие реагенты применяются после обработки с помощью процесса, который избирательно модифицирует или избирательно 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 методы секвенирования отдельной молекулы (например, секвенирования отдельной молекулы в реальном времени и нанопоровое секвенирование (например, от Oxford Nanopore Technologies)), которые распознают метилцитозины и гидроксиметилцитозины.

"Плотность метилирования" области может относиться к количеству прочтений в сайтах в пределах области, демонстрирующей метилирование, разделенному на общее количество прочтений, охватывающих сайты в этой области. Сайты могут иметь определенные характеристики, например, быть сайтами CpG. Таким образом, "плотность метилирования CpG" области может относиться к количеству прочтений, демонстрирующих метилирование CpG, разделенному на общее количество прочтений, покрывающих сайты CpG в области (например, конкретный сайт CpG, сайты CpG внутри CpG-островка или большую область). Например, плотность метилирования для каждой 100-т.п.о. группы в геноме человека может быть определена из общего числа цитозинов, не преобразованных после обработки бисульфитом (что соответствует метилированному цитозину) в сайтах CpG, как доля всех сайтов CpG, охваченных прочтениями последовательности, сопоставленными с 100-т.п.о. областью. Этот анализ также может быть выполнен для других размеров группы, например, 500 п.о., 5 т.п.о., 10 т.п.о., 50 т.п.о. или 1 млн.п.о. и т.д. Областью может быть весь геном, хромосома или часть хромосомы (например, плечо хромосомы). Индекс метилирования сайта CpG является таким же, как плотность метилирования для области, когда область включает в себя только данный сайт CpG. "Доля метилированных цитозинов" может относиться к количеству цитозиновых сайтов, "C", которые, как показано, являются метилированными (например, непреобразованными после преобразования бисульфитом) по отношению к общему количеству проанализированных остатков цитозина, т.е. включая цитозины вне контекста CpG, в области. Индекс метилирования, плотность метилирования, количество молекул, метилированных в одном или большем количестве сайтов, и доля молекул, метилированных (например, цитозинов) в одном или большем количестве сайтов, являются примерами "уровней метилирования". Помимо преобразования бисульфитом, другие процессы, известные специалистам в данной области техники, могут быть использованы для исследования статуса метилирования молекул ДНК, включая, без ограничения, ферменты, чувствительные к стату-

су метилирования (например, чувствительные к метилированию рестрикционные ферменты), белки, связывающие метилированные молекулы, секвенирование отдельной молекулы с использованием платформы, чувствительной к статусу метилирования (например, нанопоровое секвенирование (Schreiber et al. Proc Natl Acad Sci 2013; 110: 18910-18915) и секвенирование отдельной молекулы в реальном времени (например, от Pacific Biosciences) (Flusberg et al. Nat Methods 2010; 7: 461-465)).

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

"Метилом плазмы беременной" представляет собой метилом, определенный по плазме или сыворотке беременных животных (например, человека). Метилом плазмы беременной является примером внеклеточного метилома, поскольку плазма и сыворотка содержат внеклеточную ДНК. Метилом плазмы беременной также является примером смешанного метилома, поскольку он представляет собой смесь ДНК из разных органов, тканей или клеток в организме. В одном варианте осуществления, такие клетки представляют собой гемопоэтические клетки, включающие в себя, но не ограничивающиеся клетками эритроидной (т.е. эритроцитарной) линии, миелоидной линии (например, нейтрофилов и их предшественников) и мегакариоцитарной линии. Во время беременности метилом плазмы может содержать метиломную информацию от плода и матери. "Клеточный метилом" соответствует метилому, определенному по клеткам (например, клеткам крови) пациента. Метилом клеток крови называют метиломом клеток крови (или метиломом крови).

"Профиль метилирования" включает в себя информацию, относящуюся к метилированию ДНК или РНК для множества сайтов или областей. Информация, относящаяся к метилированию ДНК, может включать в себя, помимо прочего, индекс метилирования сайта CpG, плотность метилирования (сокращенно ПМ) сайтов CpG в области, распределение сайтов CpG в смежной области, структуру или уровень метилирования для каждого отдельного сайта CpG в области, содержащей больше чем один сайт CpG, и метилирование не-CpG. В одном варианте осуществления, профиль метилирования может включать в себя характер метилирования или неметилирования больше 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 цитозиновых остатков (т.е. 5-метилцитозины) из числа 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.

"Паттерн метилирования" относится к порядку метилированных и неметилированных оснований. Например, паттерн метилирования может представлять собой порядок метилированных оснований на одной цепи ДНК, одной двухцепочечной молекуле ДНК или другом типе молекулы нуклеиновой кислоты. Например, три последовательных сайта CpG могут иметь любой из следующих паттернов метилирования: UUU, MMM, UMM, UMU, UUM, MUM, MUU, или MMU, где "U" обозначает неметилированный сайт, а "M" обозначает метилированный сайт. Когда кто-то расширяет эту концепцию до модификации оснований, которые включают в себя, но не ограничиваются метилированием, может использоваться термин "паттерн модификации", который относится к порядку модифицированных и немодифицированных оснований. Например, паттерн модификаций может представлять собой порядок модифицированных оснований на одной цепи ДНК, одной двухцепочечной молекуле ДНК или другом типе молекулы нуклеиновой кислоты. Например, три последовательных, потенциально модифицируемых сайта могут иметь любой из следующих паттернов 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, количеству метилированных оснований или нуклеотидов в молекуле, разделенному на общее количество метилируемых оснований или нуклеотидов в такой молекуле. Гиперметилированная молекула - это молекула, в которой уровень метилирования отдельной молекулы находится на уровне или выше порога, который может определятся от применения к применению. Порог может составлять 5, 10, 20, 30, 40, 50, 60, 70, 80, 90 или 95%. Гипометилированная молекула - это молекула, в которой уровень метилирования отдельной молекулы находится на уровне или ниже порога, который может определятся от применения к применению, и который может изменятся 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 по уровням метилирования множества молекул данных молекул. Гиперметилированная популяция молекул представляет собой ту, в которой уровень метилирования множества молекул находится на уровне или ниже порога, который может определятся

от применения к применению, и который может изменятся 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 определятся от применения к применению. Порог может составлять 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, и 95%. В одном варианте осуществления, популяция молекул может быть выровнена с одной или большим количеством выбранных геномных областей. В одном варианте осуществления, выбранная область(области) генома может быть связана с таким заболеванием, как рак, генетическое нарушение, нарушение импринтинга, нарушение обмена веществ или неврологическое нарушение. Выбранная область(области) генома может иметь длину 50 нуклеотидов (нт), 100, 200, 300, 500, 1000 нт, 2, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500 кнт, или 1 Мнт.

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

В контексте данного документа, термин "классификация" 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 иметь несколько уровней классификации (например, шкала от 1 до 10, или от 0 до 1).

Термины "пороговое значение" и "порог" относятся к заранее определенным числам, используемым в процессе. Например, размер порогового значения может относиться к размеру, выше которого исключают фрагменты. Пороговое значение может быть значением, выше или ниже которого применяется конкретная классификация. Любой из этих терминов может использоваться в любом из этих контекстов. Пороговое значение или порог может представлять собой "эталонное значение" или производное от эталонного значения, которое является представителем конкретной классификации или проводит границу между двумя или более классификациями. Такое эталонное значение может быть определено различными способами, как будет понятно специалисту в данной области техники. Например, параметры могут быть определены для двух разных групп субъектов с разными известными классификациями, а эталонное значение может быть выбрано как репрезентативное для одной классификации (например, среднее) или значение, которое находится между двумя кластерами параметров (например, выбрано для получения желаемой чувствительности и специфичности). В качестве другого примера, эталонное значение может быть определено на основе статистического анализа или моделирования образцов.

Термин "уровень рака" может относиться к тому, есть ли рак (т.е. присутствует или отсутствует), стадии рака, размеру опухоли, наличию метастазов, общей опухолевой нагрузке на организм, реакции рака на лечение и/или другому показателю степени тяжести рака (например, рецидиву рака). Уровень рака может быть представлен числом или другими обозначениями, такими как символы, буквы алфавита и цвета. Уровень может быть нулевым. Уровень рака может также включать в себя предзлокачественные или предраковые патологии (состояния). Уровень рака можно использовать по-разному. Например, скрининг может проверить, присутствует ли рак у кого-то, для кого ранее не было известно, что он болен раком. Оценивание может исследовать кого-либо, у кого был диагностирован рак, чтобы отслеживать прогрессирование рака с течением времени, изучать эффективность лечения или определять прогноз. В одном варианте осуществления, прогноз может быть выражен как вероятность смерти пациента от рака, или вероятность прогрессирования рака после определенного периода или времени, или вероятность или степень метастазирования рака. Обнаружение может означать "скрининг" или может означать проверку того, есть ли у кого-то с признаками рака (например, симптомами или другими положительными результатами) рак.

"Уровень патологии" (или уровень нарушения) может относиться к количеству, степени или серьезности патологии, связанной с организмом, где уровень может быть таким, как описано выше для рака. Еще один пример патологии - отторжение пересаженного органа. Другие примеры патологий могут включать в себя нарушения импринтинга генов, аутоиммунную атаку (например, волчаночный нефрит, поражающий почки или рассеянный склероз), воспалительные заболевания (например, гепатит), фиброзные процессы (например, цирроз), жировую инфильтрацию (например, жировые заболевания печени), дегенеративные процессы (например, болезнь Альцгеймера) и ишемическое повреждение ткани (например, инфаркт миокарда или инсульт). Состояние здоровья пациента можно рассматривать как классификацию отсутствия патологии.

"Связанное с беременностью нарушение" включает в себя любое нарушение, характеризующееся аномальными относительными уровнями экспрессии генов в ткани матери и/или плода. Эти нарушения включают в себя, помимо прочего, преэклампсию, задержку внутриутробного развития, инвазивную

плаценту, преждевременные роды, гемолитическую болезнь новорожденных, плацентарную недостаточность, водянку плода, пороки развития плода, HELLP-синдром, системную красную волчанку 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 содержать 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 базовых единицах. Также, "нт" может использоваться для обозначения относительных позиций, например, выше или ниже анализируемого локуса. В некоторых контекстах, касающихся технологической концептуализации, представления, обработки и анализа данных, "нт" и "п.о." могут использоваться взаимозаменяемо.

Термин "контекст последовательности" может относиться к композициям оснований (A, C, G или T) и порядкам оснований в участке ДНК. Такой участок ДНК может соседствовать с основанием, которое подвергается анализу модификации основания или является его целью. Например, контекст последовательности может относиться к основаниям выше и/или ниже основания, которое подвергается анализу модификации основания.

Термин "кинетические характеристики" может относиться к признакам, полученным в результате секвенирования, включая секвенирование отдельной молекулы в реальном времени. Такие признаки можно использовать для анализа модификации оснований. Примеры кинетических характеристик включают в себя контекст восходящей и нисходящей последовательности, информацию о цепях, межимпульсный период, ширину импульса и силу импульса. При секвенировании отдельной молекулы в режиме реального времени постоянно отслеживают влияние активностей полимеразы на ДНК-матрицу. Следовательно, параметры, полученные в результате такого секвенирования, можно рассматривать как кинетические характеристики, например, нуклеотидные последовательности.

Термин "модели машинного обучения" могут включать в себя модели, основанные на использовании данных образцов (например, обучающих данных) для прогнозирования тестовых данных, и, таким образом, может включать в себя "обучение с учителем". Модели машинного обучения часто разрабатываются с использованием компьютера или процессора. Модели машинного обучения могут включать в себя статистические модели.

Термин "средство анализа данных" может включать в себя алгоритмы и/или модели, которые могут принимать данные в качестве ввода и затем выдавать прогнозируемый результат. Примеры "средства анализа данных" включают в себя 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 в себя сбор данных или мониторинг во время протекания реакции, связанной с секвенированием. Например, секвенирование в реальном времени может включать в себя оптический мониторинг или съемку того как ДНК-полимераза добавляет новое основание.

Термин "около" или "примерно" может означать нахождение в пределах допустимого диапазона ошибок для конкретного значения, как определено специалистом в данной области техники, что будет частично зависеть от того, как значение измеряется или определяется, т.е. ограничений системы измерения. Например, "около" может означать нахождение в пределах 1-го или больше чем 1-го стандартного отклонения согласно применению в данной области техники. В альтернативном варианте, "около" может означать диапазон вплоть до 20%, вплоть до 10%, плоть до 5% или вплоть до 1% от заданного значения. В альтернативном варианте, в частности в отношении биологических систем или процессов, термин "около" или "примерно" может означать нахождение в пределах порядка величины, в пределах 5-кратного и более предпочтительно в пределах 2-кратного значения. Если конкретные значения описаны в заявке и формуле изобретения, если не указано иное, следует понимать термин "около" как обозначающий нахождение в пределах допустимого диапазона ошибок для конкретного значения. Термин "около" может иметь значение, обычно понимаемое специалистом в данной области техники. Термин "около" может относиться к ±10%. Термин "около" может относиться к ±5%.

Подробное описание сущности изобретения

Достижение определения модификации основания без бисульфита, включая метилированное основание, является предметом усилий различных исследований, но ни одно из них не было коммерчески жизнеспособным. Недавно был опубликован способ определения 5mC и 5hmC без использования бисульфита (Y. Liu et al, 2019) с использованием мягких условий для преобразования оснований 5mC и 5hmC. Этот способ включает в себя множество этапов ферментативных и химических реакций, включая окисление с транслокацией десять-одиннадцать (TET), пиридин борановое восстановление и ПЦР. Эффективность каждой стадии реакции преобразования, а также смещение ПЦР отрицательно повлияют на конечную точность анализа 5mC. Например, как сообщалось, коэффициент преобразования 5mC составляет около 96%, а коэффициент ложноотрицательных результатов составляет около 3%. Такая эффективность потенциально ограничивает способность обнаруживать некоторые тонкие изменения метили-

рования в геноме. С другой стороны, ферментативное преобразование не сможет одинаково хорошо показать себя для всего генома. Например, коэффициент конверсии 5hmC был на 8,2% ниже, чем для 5mC, а коэффициент конверсии для не-CpG был на 11,4% ниже, чем для контекстов CpG (Y. Liu et al., 2019). Таким образом, идеальной ситуацией является разработка подходов к определению модификаций оснований нативной молекулы ДНК без какой-либо предварительной стадии преобразования (химической, ферментативной или их комбинации) и даже без стадии амплификации.

Был проведен ряд исследований, подтверждающих концепцию (Q. Liu et al., 2019; Ni et al., 2019), в которых электрические сигналы, генерируемые подходом в виде нанопорового секвенирования с длинными прочтениями (например, с использованием системы, разработанной Oxford Nanopore Technologies), позволили обнаружить состояния метилирования с использованием способа глубокого обучения. Помимо Oxford Nanopore, существуют и другие подходы к секвенированию отдельных молекул, которые делают возможными длинные прочтения. Один пример представляет собой секвенирование отдельной молекулы в реальном времени. Один пример секвенирования отдельной молекулы в реальном времени является выведенной на рынок системой SMRT Pacific Biosciences. Поскольку принцип секвенирования отдельной молекулы в реальном времени (например, системы SMRT Pacific Biosciences) отличается от принципа неоптической нанопоровой системы (например, Oxford Nanopore Technologies), подходы к обнаружению модификации основания, разработанные для такой неоптической нанопоровой системы, не могут применятся для секвенирования отдельной молекулы в реальном времени. Например, неоптическая нанопоровая система не предназначена для захвата паттернов флуоресцентных сигналов, производимых синтезом ДНК на основе иммобилизованной ДНК-полимеразы (применяется в секвенировании отдельной молекулы в реальном времени, например, в системе SMRT Pacific Biosciences). В качестве дополнительного примера, в платформе секвенирования Oxford Nanopore каждое измеренное электрический событие ассоциируется с k-мером (например, 5-мером) (Q. Liu et al., 2019). Однако в платформе секвенирования SMRT Pacific Biosciences каждое флуоресцентное событие в целом ассоциируется с одним добавленным основанием. Кроме того, одна молекула ДНК будет многократно секвенирована в SMRT-секвенировании Pacific Biosciences, включая цепи Уотсона и Крика. И наоборот, для подхода секвенирования Oxford Nanopore с длинными прочтениями, считывание последовательности выполняют один раз для каждой из цепей Уотсона и Крика.

Сообщалось, что кинетика полимеразы будет зависеть от состояний метилирования в последовательностях E. coli (Flusberg et al., 2010). Предыдущие исследования показали, что по сравнению с обнаружением 6mA, 4mC, 5hmC и 8-оксогуанина гораздо сложнее использовать кинетику полимеразы в секвенировании отдельной молекулы в реальном времени для определения состояний метилирования (5mC в сравнении C) определенного CpG в отдельной молекуле. Причина в том, что метильная группа мала и ориентирована в сторону большой бороздки и не участвует в спаривании оснований, что приводит к очень тонкому нарушению кинетики, вызванному 5mC. (Clark et al., 2013). Следовательно, существует мало подходов для определения состояний метилирования цитозинов на уровне отдельной молекулы.

Suzuki et al. разработали алгоритм (Suzuki et al., 2016), пытающийся объединить соотношения межимпульсного периода (МИП) для соседних сайтов CpG, чтобы повысить достоверность идентификации состояний метилирования этих сайтов. Однако этот алгоритм позволял только предсказать то, что геномная область является полностью метилированной или полностью неметилированной, но не был способен определять промежуточные паттерны метилирования.

Что касается секвенирования отдельной молекулы в реальном времени, современные подходы используют только один или два параметра независимо, что позволяет достичь очень ограниченной точности измерения 5mC из-за разницы в определении между 5-метилцитозином и цитозином. Например, Flusberg et al. продемонстрировали, что МИП был изменен в модификациях оснований, включая N6-метиладенозин, 5-метилцитозин и 5-гидроксиметилцитозин. Однако не было обнаружено, что ширина импульса (ШИ) кинетики секвенирования оказывает существенное влияние. Следовательно, в способе, который они использовали для прогнозирования модификации основания, используя в качестве примера определение N6-метиладенозина, использовали только МИП, но не ШИ.

В последующих публикациях той же группы (Clark et al., 2012; Clark et al. 2013), МИП, но не ШИ, внедряли в алгоритмы для определения 5-метилцитозина. В Clark et al. 2012, уровень обнаружения 5-метилцитозина без его преобразования в 5-метилцитозин составлял всего от 1,9% до 4,3%. Кроме того, в Clark et al. 2013, авторы дополнительно повторно подтвердили тонкость кинетической сигнатуры 5-метилцитозина. Чтобы преодолеть низкую чувствительность определения 5-метилцитозина, Clark et al. дополнительно разработал способ, который преобразовывал 5-метилцитозин в 5-карбоксилметилцитозин с использованием белков транслокации десять-одиннадцать (Tet), чтобы улучшить чувствительность 5-метилцитозина (Clark et al. 2013), поскольку изменение МИП, вызванное 5-карбоксилцитозином, было намного большим, чем 5-метилцитозином.

В более позднем отчете Blow et al., способ на основе соотношения МИП, ранее описанный Flusberg et al. был использован для обнаружения модификаций оснований в 217 бактериальных и 13 архейных видов с 130-кратным покрытием считывания для каждого организма (Blow et al., 2016). Среди всех идентифицированных модификаций оснований только 5% составляют 5-метилцитозин. Они приписали этот

низкий уровень обнаружения 5-метилцитозина низкой чувствительности секвенирования отдельной молекулы в реальном времени для обнаружения 5-метилцитозина. У большинства бактерий набор мотивов последовательностей стал целью метилирования ДНК-метилтрансферазой (МТазы) (например, 5'-GmATC-3' для Dam или 5'-CmCWGG-3' для Dcm у E. coli) почти во всех этих мотивах в геноме, и только небольшая часть этих мотивов остается неметилированной (Beaulaurier et al. 2019). Кроме того, использование способа на основе МИП для классификации статуса метилирования второго C в мотиве 5'-CCWGG-3' с обработкой или без обработки белками Tet дало уровни обнаружения 5-метилцитозина, составляющие 95,2% и 1,9%, соответственно (Clark et al. 2013). В целом способ МИП без предварительного преобразования оснований (например, применяя белки Tet) пропускает большинство 5-метилцитозинов.

В упомянутых выше исследованиях (Clark et al., 2012; Clark et al., 2013; Blow et al., 2016), алгоритмы на основе МИП использовались без учета контекста последовательности, в которой находилась модификация-кандидат основания. Другие группы попытались принять во внимание контекст последовательности возле нуклеотида для обнаружения модификации основания. Например, Feng et al. использовали иерархическую модель для анализа МИП для обнаружения 4-метилцитозина и 6-метиладенозина в соответствующем контексте последовательности (Feng et al. 2013). Однако в своем способе они учитывали только МИП по основанию интереса и контекст последовательности, смежный с этим основанием, но не использовали информацию МИП всех соседних оснований, смежных с основанием интереса. Кроме того, ШИ не учитывалась в алгоритме, и они не представили каких-либо данных по обнаружению 5-метилцитозина.

В другом исследовании Schadt et al. разработали статистический способ, называемый условным случайным полем, для анализа информации МИП основания интереса и соседних оснований, чтобы определить, является ли интересующее основание 5-метилцитозином (Schadt et al., 2012). В данной работе они также рассмотрели взаимодействие МИП между этими основаниями, введя их в уравнение. Однако они не вносили нуклеотидную последовательность, а именно A, T, G или C, в свое уравнение. Когда они применили этот способ для определения статуса метилирования плазмиды M.Sau3AI, площадь под кривой ROC была близка к 0,5 даже при 800-кратном охвате последовательности плазмидной последовательности. Более того, в их способе, они не учитывали ШИ в своем анализе.

В еще одном исследовании Beckman et al. они сравнили МИП всех последовательностей, которые имеют один и тот же 4-нуклеотидный или 6-нуклеотидный мотив в геноме между целевым бактериальным геномом и полностью неметилированным геномом, например, полученным путем амплификации всего генома (Beckman et al. 2014). Целью такого анализа было только выявление мотивов, которые чаще всего бы подвергались модификациям оснований. В исследовании они учитывали только МИП потенциально модифицированного основания, но не МИП соседнего основания или ШИ. Их способ не был информативен относительно статуса метилирования отдельного нуклеотида.

В целом, эти предыдущие попытки использовать только МИП или в комбинации с информацией последовательностей соседних нуклеотидов для группирования данных не сделали возможным определение модификации основания 5-метилцитозина с убедительной или практической точностью. В недавнем обзоре Gouil et al. авторы пришли к выводу, что из-за низкого соотношения сигнал/шум обнаружение 5-метилцитозина в одной молекуле с использованием секвенирования отдельной молекулы в реальном времени является неточным (Gouil et al., 2019). В этих предыдущих исследованиях остается неизвестным то, возможно ли использовать кинетические характеристики для полногеномного метиломного анализа, особенно для сложных геномов, таких как геномы людей, геномы видов рака или геномы зародышей.

В отличие от предыдущих исследований, некоторые варианты осуществления способов, описанных в этом изобретении, основаны на измерении и использовании МИП, ШИ и контекста последовательности для каждого основания в пределах окна измерения. Мы рассудили, что если мы сможем использовать комбинацию нескольких показателей, например, параллельно используя характеристики, включая контекст восходящей и нисходящей последовательности, информацию о цепи, МИП, ширину импульса, а также силу импульса, мы сможем достичь точного измерения модификаций оснований (например, обнаружения mC) при разрешении, составляющим одно основание. Контекст последовательности относится к композициям оснований (A, C, G или T) и порядкам оснований в участке ДНК. Такой участок ДНК может прилегать к основанию, которое подвергается анализу модификации основания или является его целью. В одном варианте осуществления, участок ДНК может быть проксимальным относительно основания, которое подвергается анализу модификации основания. В другом варианте осуществления, участок ДНК может находиться далеко от основания, которое подвергается анализу модификации основания. Участок ДНК может располагаться выше и/или ниже основания, которое подвергается анализу модификации основания.

В одном варианте осуществления, характеристики - контекст восходящей и нисходящей последовательностей, информация о цепях, МИП, ширина импульса, а также сила импульса, которые используются для анализа модификации основания, называются кинетическими характеристиками.

Варианты осуществления, представленные в данном изобретении, могут быть использованы для ДНК, полученной из, но не ограничиваясь лишь этими: клеточных линий, образцов из организма (на-

пример, твердых органов, твердых тканей, образца, полученного с помощью эндоскопии, крови или плазмы, или сыворотки, или мочи беременной женщины, биопсии ворсинок хориона и т.д.), образцов, полученных из окружающей среды (например, бактерии, клеточных загрязнителей), продуктов питания (например, мяса). В некоторых вариантах осуществления, способы, представленные в данном изобретении, также могут применяться после стадии, на которой сначала обогащается часть генома, например, с использованием гибридизационных зондов (Albert et al., 2007; Okou et al., 2007; Lee et al., 2011) или подходов, основанных на физическом разделении (например, на основе размеров и т.д.) или после расщепления рестрикционным ферментом (например, MspI), или обогащения на основе Cas9 (Watson et al., 2019). Хотя изобретение не требует для реализации ферментативного или химического преобразования, в некоторых вариантах осуществления такая стадия преобразования может быть включена для дополнительного повышения эффективности изобретения.

Варианты осуществления данного изобретения позволяют улучшить точность, практичность, или удобство обнаружения модификаций оснований или определения уровней модификации. Модификация может быть обнаружена напрямую. В вариантах осуществления может избегаться ферментативное или химическое преобразование, которое может не сохранять всю информацию о модификации для обнаружения.

Дополнительно, определенные ферментативные или химические преобразования могут быть несовместимы с определенными типами модификаций. Варианты осуществления данного изобретения также могут избегать амплификации с помощью ПЦР, которая может не переносить информацию об модификации основания в продукты ПЦР. Дополнительно, обе цепи ДНК могут быть секвенированы вместе, тем самым позволяя спаривать последовательность из одной цепи с ее комплементарной последовательностью другой цепи. Напротив, амплификация ПЦР разделяет две цепи двухцепочечной ДНК, поэтому такое спаривание последовательностей затруднено.

Профили метилирования, определенные с ферментативным или химическим преобразованием, или без него, могут использоваться для анализа биологических образцов. В одном варианте осуществления, профили метилирования могут использоваться для определения происхождения клеточной ДНК (например, материнской или зародышевой, тканевой, вирусной или опухолевой). Определение профилей аберрантного метилирования в тканях помогает идентифицировать нарушения развития у индивидов, а также выявлять и прогнозировать опухоли или злокачественные новообразования. Дисбаланс в уровнях метилирования между гаплотипами можно использовать для выявления нарушений, включая рак. Паттерны метилирования в отдельной молекуле могут идентифицировать химерную (например, между вирусом и человеком) и гибридную ДНК (например, между двумя генами, обычно не слитыми в природном геноме); или между двумя видами (например, посредством генетических или геномных манипуляций).

Анализ метилирования может быть улучшен за счет дополнительного обучения, которое может включать в себя сужение данных, используемых в обучающем наборе. Для анализа могут быть выбраны определенные области. В вариантах осуществления такое нацеливание может включать в себя фермент, который либо сам по себе, либо в комбинации с другим реагентом(ами) может расщеплять последовательность ДНК или геном на основе его последовательности. В некоторых вариантах осуществления, фермент представляет собой рестрикционный фермент, который распознает и расщепляет конкретную последовательность(и) ДНК. В других вариантах осуществления, в комбинации может быть использован больше чем один рестрикционный фермент с разными распознаваемыми последовательностями. В некоторых вариантах осуществления, рестрикционный фермент может расщеплять или не расщеплять в зависимости от статуса метилирования распознаваемых последовательностей. В некоторых вариантах осуществления, фермент относится к семейству CRISPR/Cas. Например, геномные области интереса могут быть сделаны целями с использованием системы CRISPR/Cas9 или другой системы, основанной на направляющей РНК (т.е. коротких последовательностях РНК, которые связываются с комплементарными целевыми последовательностями ДНК и в процессе направляют фермент для воздействия на целевой геномный участок). В некоторых случаях, анализ метилирования может быть возможен без выравнивания с эталонным геномом.

I. Обнаружение метилирования с помощью секвенирования отдельной молекулы в реальном времени.

Варианты осуществления данного изобретения позволяют прямо обнаруживать модификации оснований без ферментативного или химического преобразования. Кинетические характеристики (например, контекст последовательности, МИП и ШИ) полученные с помощью секвенирования отдельной молекулы в реальном времени могут быть проанализированы с помощью машинного обучения для разработки модели для обнаружения модификации или отсутствия модификации. Уровни модификации можно использовать для определения происхождения молекул ДНК, или наличия или уровня нарушения.

Используя SMRT-секвенирование Pacific Biosciences в качестве примера секвенирования отдельной молекулы в реальном времени в целях иллюстрации, молекулу ДНК-полимеразы располагают на дне лунок, которые служат в качестве волноводов с нулевой модой (ZMW). ZMW - это нанофотонное устройство для заключения света в небольшом объеме, находящемся под наблюдением, который может представлять собой отверстие с очень маленьким диаметром и не позволяющее распространяться свету в

диапазоне длин волн, используемом для обнаружения, так что только излучение оптических сигналов от меченого красителем нуклеотида, встраиваемого иммобилизованной полимеразой, обнаруживается на фоне низкого и постоянного фонового сигнала (Eid et al., 2009). ДНК-полимераза катализирует встраивание флуоресцентно меченых нуклеотидов в комплементарные цепи нуклеиновых кислот.

Фиг. 1 демонстрирует пример молекул, несущих модификации оснований, которые были секвенированы с помощью консенсусного секвенирования отдельной кольцевой молекулы. Молекулы 102, 104 и 106 несут модификации оснований. Молекулы ДНК (например, молекула 106) могут быть лигированы адаптерами в виде шпилек для формирования лигированной молекулы 108. Затем лигированная молекула 108 может формировать замкнутую в кольцо молекулу 110. Замкнутые в кольцо молекулы могут связываться с иммобилизованной ДНК-полимеразой и могут инициировать синтез ДНК. Также могут быть секвенированы молекулы, не несущие модификации оснований.

Фиг. 2 демонстрирует пример молекул, несущих метилированные и/или неметилированные сайты CpG, которые были секвенированы путем секвенирования отдельной молекулы в реальном времени. Сначала молекулы ДНК были лигированы с помощью адаптеров в виде шпилек, чтобы сформировать замкнутые в кольца молекулы, которые будут связываться с иммобилизованной ДНК-полимеразой и инициировать синтез ДНК. На фиг. 2, молекулу ДНК 202 лигируют с адаптерами в виде шпилек с формированием лигированной молекулы 204. Затем лигированная молекула 204 формирует замкнутую в кольцо молекулу 206. Также могут быть секвенированы молекулы без сайтов CpG. Замкнутая в кольцо молекула 206 включает в себя неметилированный сайт 208 CpG, который все еще может быть секвенирован.

Как только синтез ДНК был инициирован, нуклеотиды, меченные флуоресцентным красителем, будут встраиваться иммобилизованной полимеразой в вновь синтезированную цепь на основе кольцевой ДНК-матрицы, что приведет к излучению оптических сигналов. Поскольку ДНК-матрицы были замкнуты в кольца, вся кольцевая ДНК-матрица будет проходить через полимеразу множество раз (т.е. один нуклеотид в ДНК-матрице будет секвенирован множество раз). Последовательность, сгенерированная в результате процесса, в котором все основания в замкнутой в кольцо ДНК-матрице полностью проходят через ДНК-полимеразу, называется субпрочтением. Одна молекула в ZMW будет генерировать множество субпрочтений, потому что полимераза может проходить по всей кольцевой ДНК-матрице множество раз. В одном варианте осуществления, субпрочтение может содержать только часть последовательности, модификаций оснований или другой молекулярной информации из кольцевой ДНК-матрицы из-за того, что, в одном варианте осуществления, существуют ошибки секвенирования.

Как продемонстрировано на фиг. 3, время проявления и длительность результирующих импульсов флуоресценции позволяет измерить кинетику полимеразы. Межимпульсный период (МИП) - это показатель длительности периода времени между двумя импульсами излучения, каждый из которых может указывать на встроенный в формирующуюся цепь флуоресцентно меченый нуклеотид фиг. 3). Как продемонстрировано на фиг. 3, ширина импульса (ШИ) является еще одним показателем, отражающим кинетику полимеразы в сочетании с длительностью импульсов, связанных с сигналом от основания. ШИ может представлять собой длительность импульса при 0% высоты пика сигнала (т.е. интенсивностью флуоресценции меченного красителем нуклеотида при встраивании). В одном варианте осуществления, ШИ может определяться, например, но без ограничения, длительностью импульса при 5, 10, 20, 30, 40, 50, 60, 70, 80 или 90% высоты пика сигнала. В некоторых вариантах осуществления, ШИ может быть площадью под пиком, разделенной на высоту пика сигнала.

Было показано, что такой кинетический параметр полимеразы, как МИП, зависит от модификаций оснований, таких как N6-метиладенин (6mA), 5-метилцитозин (5mC) и 5-гидроксиметилцитозин (5hmC) в синтетических и микробных последовательностях (например, E. coli) (Flusberg et al., 2010). Flusberg et al. 2010 не использовали контекст последовательности и МИП в качестве независимых входных данных для обнаружения модификаций, что привело к модели, которой не хватало практически эффективной точности для обнаружения. Flusberg et al. использовали только контекст последовательности для подтверждения 6mA, встречающейся в GATC. Flusberg et al. не упоминают об использовании контекста последовательности в сочетании с МИП в качестве входных данных для определения статуса метилирования.

Слабые прерывания, связанные с добавлением нового основания в пару к 5-метилцитозину в комплементарных цепях, делают получение сигнала от метилирования чрезвычайно сложным даже для относительно простых микробных геномов при использовании только сигналов МИП, поскольку сообщалось, что обнаружение мотива метилирования $C^{m}CWGG$ варьировалось в пределах от 1,9 до 4,3% (Clark et al., 2013). Например, пакет аналитического программного обеспечения (SMRT Link v6.0.0), предоставляемый Pacific Biosciences, не способен выполнять анализ 5mC. Кроме того, предыдущая версия SMRT Link v5.1.0 требовала использования фермента Tet1 для преобразования 5mC в 5-карбоксилцитозин (5caC) перед анализом метилирования, поскольку сигналы МИП, связанные с 5caC, были бы усилены (Clark et al., 2013). Таким образом, неудивительно, что нет исследований, показывающих возможность использования секвенирования отдельной молекулы в реальном времени для анализа нативной ДНК в масштабе всего генома для генома человека.

II. Паттерны окна измерения и модели машинного обучения.

Являются желательными методы обнаружения модификаций в основаниях без ферментативного или химического преобразования модификации и/или основания. Как описано в данном документе, модификации в основании-цели могут быть обнаружены с использованием данных кинетических характеристик, полученных из секвенирования отдельной молекулы в реальном времени для оснований, окружающих целевое основание. Кинетические характеристики могут включать в себя межимпульсный период, ширину импульса и контекст последовательности. Эти кинетические характеристики могут быть получены для окна измерения определенного количества нуклеотидов выше и ниже целевого основания. Эти характеристики (например, в определенных местах в окне измерения) можно использовать для обучения модели машинного обучения. В качестве примера приготовления образца две цепи молекулы ДНК могут быть соединены адаптерами в виде шпилек, формируя тем самым кольцевую молекулу ДНК. Кольцевая молекула ДНК позволяет получить кинетические характеристики одной или обеих цепей Уотсона и Крика. Средство анализа данных может быть разработано на основе кинетических характеристик в окнах обнаружения. Это средство анализа данных может затем использоваться для обнаружения модификаций, включая метилирование. В разделе описаны различные методы обнаружения модификаций.

A. Использование одной цепи.

Как показано на фиг. 4, в качестве примера, мы получили субпрочтения цепи Уотсона из секвенирования SMRT Pacific Biosciences для анализа одного конкретного основания относительно состояний модификаций оснований. На фиг. 4, 3 основания с каждой стороны от основания, которое было подвергнуто анализу модификации основания, будут обозначены как окно измерения 400. В одном варианте осуществления, контекст последовательности, МИП и ШИ для этих 7 оснований (т.е. 3-нуклеотидная (нт) восходящая и нисходящая последовательности и один нуклеотид для анализа модификации основания) были скомпилированы в 2-размерную (т.е. 2-мерную) матрицу в виде окна измерения. В показанном примере окно измерения 400 предназначено для одного субпрочтения цепи Уотсона. В данном документе описаны другие варианты.

Первая строка 402 матрицы обозначала исследуемую последовательность. Во второй строке 404 матрицы позиция 0 представляла основание для анализа модификации основания. Относительные позиции -1, -2 и -3 обозначали позиции 1-нт, 2-нт и 3-нт, соответственно, выше основания, которое подвергали анализу модификации основания. Относительные позиции +1, +2 и +3 обозначали позиции 1-нт, 2-нт и 3-нт, соответственно, ниже основания, которое подвергали анализу модификации основания. Каждая позиция включает в себя 2 столбца, которые содержат соответствующие значения МИП и ШИ Следующие 4 строки (строки 408, 412, 416 и 420) соответствуют 4 типам нуклеотидов (A, C, G и T) в цепи (например, цепи Уотсона) соответственно. Наличие значений МИП и ШИ в матрице зависело от того, какой соответствующий тип нуклеотида был секвенирован в конкретной позиции. Как показано на Фиг. 4, в относительной позиции 0 значения МИП и ШИ были показаны в строке, обозначающей "G" в цепи Уотсона, что позволяет предположить, что в последовательности-результате в той позиции был сигнал гуанина. Другие ячейки в столбце, которые не соответствуют секвенированному основанию, будут кодироваться как "0". В качестве примера, информация о последовательности, соответствующая 2-мерной цифровой матрице (фиг. 4), будет представлять собой 5'-GATGACT-3' для цепи Уотсона.

Как показано в одном варианте осуществления, изображенном на фиг. 5, окно измерения может применяться к данным из цепи Крика. Мы получили субпрочтения цепи Крика из секвенирования отдельной молекулы в реальном времени для анализа одного конкретного основания относительно состояний модификаций оснований. На фиг. 5, 3 основания с каждой стороны от основания, которое было подвергнуто анализу модификации основания и основание, подвергнутое анализу модификации основания, были обозначены как окно измерения. В одном варианте осуществления, контекст последовательности, МИП, ШИ для этих 7 оснований (т.е. 3-нуклеотидная (нт) восходящая и нисходящая последовательности и один нуклеотид для анализа модификации основания) были скомпилированы в 2-размерную (т.е. 2-мерную) матрицу в виде окна измерения. Первая строка матрицы обозначала исследуемую последовательность. Во второй строке матрицы позиция 0 представляла основание для анализа модификации основания. Относительные позиции -1, -2 и -3 обозначали позиции 1-нт, 2-нт и 3-нт, соответственно, выше основания, которое подвергали анализу модификации основания. Относительные позиции +1, +2 и +3 обозначали позиции 1-нт, 2-нт и 3-нт, соответственно, ниже основания, которое подвергали анализу модификации основания. Каждая позиция включает в себя 2 столбца, которые содержат соответствующие значения МИП и ШИ. Следующие 4 строки соответствуют 4 типам нуклеотидов (A, C, G и T) в данной цепи (например, цепи Крика). Наличие значений МИП и ШИ в матрице зависело от того, какой соответствующий тип нуклеотида был секвенирован в конкретной позиции. Как показано на фиг. 5, в относительной позиции О значения МИП и ШИ были показаны в строке, обозначающей "T" в цепи Крика, что позволяет предположить, что в последовательности-результате в той позиции был сигнал тимина. Другие ячейки в столбце, которые не соответствуют секвенированному основанию, будут кодироваться как "О". В качестве примера, информация о последовательности, соответствующая 2-мерной цифровой матрице (фиг. 5), будет представлять собой 5'-ACTTAGC-3' для цепи Крика.

B. Использование обеих цепей Уотсона и Крика.

Фиг. 6 демонстрирует вариант осуществления, в котором окно измерения может быть реализовано таким образом, чтобы можно было комбинировать данные из цепи Уотсона и ее комплементарной цепи Крика. Как показано на фиг. 6, мы получили субпрочтения цепей Уотсона и Крика из секвенирования отдельной молекулы в реальном времени для анализа одного конкретного основания на предмет модификаций. В одном варианте осуществления, окно измерения для цепи Крика кольцевой ДНК-матрицы было комплементарным окну измерения для цепи Уотсона, которую подвергали анализу модификации основания. На фиг. 6, 3 основания с каждой стороны от первого основания цепи Уотсона, которые была подвергнуты анализу модификации основания и первое основание были обозначены как первое окно измерения. 3 основания с каждой стороны второго основания в цепи Крика и второе основание будут определены как второе окно измерения. Второе основание было комплементарно первому основанию. В одном варианте осуществления, контекст последовательности, МИП, ШИ для этих 7 оснований (т.е. 3-нуклеотидная (нт) восходящая и нисходящая последовательности и один нуклеотид для анализа модификации основания) из цепей Уотсона и Крика были скомпилированы в 2-размерную (т.е. 2-мерную) матрицу. Эти окна измерения из цепей Уотсона и Крика рассматривали как первое и второе окна измерения, соответственно.

Первая строка матрицы цепей Уотсона и Крика обозначала исследуемую последовательность. Во второй строке матрицы цепи Уотсона позиция 0 представляла первое основание для анализа модификации основания. Позиция 0, показанная во второй строке матрицы цепи Крика, представляла второе основание, комплементарное первому основанию. Относительные позиции -1, -2 и -3 обозначали позиции 1-нт, 2-нт и 3-нт, соответственно, выше первого и второго оснований. Относительные позиции +1, +2 и +3 обозначали позиции 1-нт, 2-нт и 3-нт, соответственно, ниже первого и второго оснований. Каждая позиция, полученная из цепей Уотсона и Крика, буде соответствовать 2 столбцам, которые содержат соответствующие значения МИП и ШИ. Следующие 4 строки в матрицах цепей Уотсона и Крика соответствовали 4 типам нуклеотидов (A, C, G и T) в конкретной цепи (например, цепи Крика), соответственно. Наличие значений МИП и ШИ в матрице зависело от того, какой соответствующий тип нуклеотида был секвенирован в конкретной позиции.

Как показано на фиг. 6, в относительной позиции 0 значения МИП и ШИ были показаны в строке, обозначающей "A" в цепи Уотсона и "T" в цепи Крика, что позволяет предположить, что в последовательности-результате в той позиции цепей Уотсона и Крика были сигналы аденина и тимина, соответственно. Другие ячейки в столбце, которые не соответствовали секвенированному основанию, кодировались как "O". В качестве примера, информация о последовательности, соответствующая 2-мерной цифровой матрице цепи Уотсона (фиг. 6), будет представлять собой 5'-ATAAGTT-3'. Информация о последовательности, соответствующая 2-мерной цифровой матрице цепи Крика (фиг. 6), будет представлять собой 5'-AACTTAT-3'.

Как показано в этом примере, данные из цепей Уотсона и Крика можно объединить для формирования новой матрицы, которую также можно рассматривать как окно измерения. Эту новую матрицу можно использовать в качестве единого образца, который используется для обучения модели машинного обучения. Таким образом, все значения в новой матрице можно рассматривать как отдельные характеристики, хотя может иметь влияние конкретное размещение в 2-мерной матрице, например, когда используется сверточная нейронная сеть (CHC). Контекст последовательности в различных позициях для разных цепей может быть передан через ненулевые записи в матрице.

Фиг. 7 демонстрирует, что окно измерения может быть реализовано таким образом, чтобы данные из цепей Уотсона и Крика не являлись позициями, точно комплементарными друг другу. Как показано на фиг. 7 первое окно измерения представляло собой 5'-ATAAGTT-3'; и второе окно измерения представляло собой 5'-GTAACGC-3'. В некоторых вариантах осуществления, цепи Уотсона и Крика могут быть смещены относительно друг друга, так что позиции не являются комплементарными.

Фиг. 8 демонстрирует, что окно измерения может быть использовано для анализа состояний метилирования в сайтах CpG. Позиция 0 соответствует цитозину сайта CpG, и, таким образом, существует сдвиг на одну позицию между двумя цепями, так что C находится в позиции 0 для обеих цепей. Соответственно, только часть последовательностей, включенных в окно измерения из цепей Уотсона и Крика, являются комплементарными друг другу. В других вариантах осуществления, все последовательности в окне измерения из цепей Ватсона и Крика могут быть комплементарны друг другу. В еще других вариантах осуществления ни одна из последовательностей в окне измерения из цепей Уотсона и Крика не является комплементарной друг другу.

В одном варианте осуществления для окна измерения длина участка ДНК, окружающего основание, которое подвергали анализу модификации основания, может быть асимметричной. Например, для анализа модификации основания можно использовать X-нт выше и Y-нт ниже этого основания. X может включать в себя, но не ограничивается: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 100, 150, 200, 300, 400, 500, 1000, 2000, 4000, 5000, и 10000; Y может включать в себя, но не ограничивается: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32,

33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 100, 150, 200, 300, 400, 500, 1000, 2000, 4000, 5000, и 10000.

C. Тренировка моделей и обнаружение модификаций.

Фиг. 9 демонстрирует общую процедуру использования окна измерения для определения любых модификаций оснований. Образцы ДНК, для которых известно, что они немодифицированы и модифицированы, были подвергнуты секвенированию отдельной молекулы в реальном времени. Модифицированная ДНК (например, модифицированная молекула 902) означает, что основание (например, основание 904) имеет модификацию (например, метилирование) в данном сайте. Немодифицированная ДНК (например, немодифицированная молекула 906) означает, что основание (например, основание 908) не имеет модификации в данном сайте. Оба набора ДНК могут быть искусственно созданы или обработаны для формирования модифицированной/немодифицированной ДНК.

На стадии 910 образцы затем могут подвергаться секвенированию отдельной молекулы в реальном времени. В рамках секвенирования SMRT кольцевые молекулы можно секвенировать множество раз, путем их многократного пропускания через иммобилизованную ДНК-полимеразу. Информация о последовательности, полученная каждый раз, будет считаться субпрочтением. Таким образом, одна кольцевая ДНК-матрица будет генерировать множество субпротчений. Субпрочтения из секвенирования могут быть выровнены с эталонным геномом, например, с использованием, но не ограничиваясь BLASR (Mark J Chaisson et al., BMC Bioinformatics. 2012; 13: 238). В различных других вариантах осуществления, для выравнивания субпрочтений с эталонным геномом могут быть использованы 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), LAST (Kielbasa SM et al., Genome Res. 2011;21(3):487-493) и Minimap2 (Li H, Bioinformatics. 2018;34(18):3094-3100). Выравнивание может позволить объединить данные из множества субпрочтений (например, усреднить), поскольку данные в каждом субпрочтении могут быть идентифицированы для одной и той же позиции.

На стадии 912 из результата выравнивания были получены МИП, ШИ и контекст последовательности, окружающий основание, которое было подвергнуто анализу модификации основания. На стадии 914 МИП, ШИ и контекст последовательности были записаны в виде определенной структуры, например, но не ограничиваясь 2-мерной матрицей, как показано на фиг. 9.

На стадии 916 для обучения аналитической, вычислительной, математической или статистической модели(моделей) использовался ряд 2-мерных матриц, содержащих эталонные кинетические паттерны полученных молекул с известными модификациями оснований. На стадии 918 создавали статистическую модель, полученную в результате обучения. Для простоты, фиг. 9 демонстрирует только статистическую модель, разработанную путем обучения, но может быть разработана любая модель или платформа анализа данных. Примеры платформ анализа данных включают в себя модели машинного обучения, статистические модели и математические модели. Статистические модели могут включать в себя, помимо прочего, 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 машину опорных векторов (SVM). Участок ДНК, окружающий основание, которое было подвергнуто анализу модификации основания, может быть X-нт выше и Y-нт ниже данного основания, а именно "окном измерения".

В процессе обучения могут использоваться структуры данных, поскольку известны правильные выходные данные (т.е. статус модификации). Например, МИП, ШИ и контекст последовательности, соответствующий 3-нт выше и ниже основания из цепи(цепей) Уотсона и/или Крика, могут использоваться для построения 2-мерной матрицы, которая будет использоваться для обучения статистической модели(моделей) для классификации модификаций оснований. Таким образом, обучение может предоставить модель, которая может классифицировать модификацию основания в позиции нуклеиновой кислоты с ранее известным статусом.

Фиг. 10 демонстрирует общую процедуру того, как статистические модели, обученные на основе образцов ДНК, которые несут известные состояния модификаций оснований, могут обнаруживать модификации оснований. Образец с неизвестными состояниями модификаций оснований был подвергнут SMRT-секвенированию. Субпрочтения из секвенирования были выровнены с эталонным геномом с использованием, например, методов, упомянутых выше. В дополнение или вместо этого, субпрочтения могут быть выровнены относительно друг друга. Еще другие варианты осуществления могут использовать только одно субпрочтение или анализировать их независимо, так что выравнивание не выполняется.

Для основания, которое подвергалось анализу модификации основания, можно было получить МИП, ШИ и контекст последовательности из цепи(цепей) Уотсона и/или Крика из результатов выравнивания с использованием сопоставимого окна измерения, как использовали на стадии обучения (фиг. 9) и связать с этим основанием. В другом варианте осуществления, окна измерения для процедур обучения и тестирования будут разными. Например, размер окон измерения для процедур обучения и тестирования может отличаться. Такие МИП, ШИ и контекст последовательности будут преобразованы в 2-мерную

матрицу. Такую 2-мерную матрицу исследуемого образца можно будет сравнить с эталонными кинетическими характеристиками для определения модификаций оснований. Например, 2-мерную матрицу тестового образца можно сравнить с эталонными кинетическими характеристиками с помощью статистических моделей(модели), которые были обучены на обучающих выборках, так что могут быть определены модификации оснований в сайтах в молекулах нуклеиновых кислот в тестовом образце. Статистические модели могут включать в себя, помимо прочего, 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 машину опорных векторов (SVM).

Фиг. 11 демонстрирует общую процедуру того, как может быть создан способ для классификации состояний метилирования в сайтах CpG. Образцы ДНК, о которых известно, что они неметилированы и метилированы в сайтах CpG, были подвергнуты секвенированию отдельной молекулы в реальном времени. Субпрочтения из секвенирования были выровнены с эталонным геномом. Использовались данные цепи Уотсона.

Из результатов выравнивания были получены МИП, ШИ и контекст последовательности, окружающий цитозин в сайте CpG, который был подвергнут анализу метилирования, и они были записаны в виде определенной структуры, например, без ограничения, 2-мерной матрицы, как показано на фиг. 11. Ряд 2-мерных матриц, содержащих эталонные кинетические паттерны полученных молекул с известными состояниями метилирования, использовали для обучения статистической модели(моделей). Участок ДНК, окружающий исследуемое основание, может быть X-нт выше и Y-нт ниже данного основания, а именно "окном измерения". X может включать в себя, но не ограничивается: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 100, 150, 200, 300, 400, 500, 1000, 2000, 4000, 5000, и 10000; Y может включать в себя, но не ограничивается: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 100, 150, 200, 300, 400, 500, 1000, 2000, 4000, 5000, и 10000. В одном варианте осуществления, МИП, ШИ и контекст последовательности, соответствующий 3-нт выше и ниже основания из цепи Уотсона, могут использоваться для построения 2-мерной матрицы, которая будет использоваться для обучения статистической модели(моделей) для классификации модификаций оснований.

Фиг. 12 демонстрирует общую процедуру классификации состояний метилирования неизвестного образца. Образец с неизвестным состоянием метилирования подвергали секвенированию отдельной молекулы в реальном времени. Субпрочтения из секвенирования были выровнены с эталонным геномом.

Для цитозина сайта CG в результате выравнивания можно получить МИП, ШИ и контекст последовательности из цепи Уотсона, используя сопоставимое окно измерения, которое применялось на стадии обучения (фиг. 11), связанное с тем основанием, модификацию которого выясняли. Эти МИП, ШИ и контекст последовательности могут быть преобразованы в 2-мерную матрицу. Такую 2-мерную матрицу тестового образца можно было бы сравнить с эталонными кинетическими паттернами, показанными на фиг. 11 для определения состояний метилирования. XI1.

Фиг. 13 и фиг. 14 демонстрируют, что кинетические характеристики из цепи Крика могут быть использованы для процедур обучения и тестирования, как описано выше, аналогично процедурам с цепью Уотсона. Статистические модели могут быть одинаковыми или разными. Когда модели разные, их можно использовать для получения независимых классификаций, которые можно сравнивать, например, если они согласуются, то идентифицируют статус модификации. Если они не согласуются, то может быть определен неклассифицируемый статус. Когда они представляют собой одну и ту же модель, данные могут быть объединены в единую структуру данных, например, матрицу на фиг. 6.

Фиг. 15 и фиг. 16 демонстрируют, что кинетические характеристики цепей Уотсона и Крика могут быть использованы для процедур обучения и тестирования, как описано выше. Образцы ДНК, о которых известно, что они неметилированы и метилированы в сайтах CpG, были подвергнуты секвенированию отдельной молекулы в реальном времени. Субпрочтения из секвенирования были выровнены с эталонным геномом, хотя возможно выравнивание субпрочтений друг с другом, как это может быть сделано для других способов, описанных в данном документе.

Для субпрочтения в результате выравнивания были получены МИП, ШИ и контекст последовательности, окружающий цитозин сайта CpG, который подвергался анализу метилирования. Поскольку молекулы ДНК были замкнуты в кольцо с помощью двух адаптеров в виде шпильки (например, следуя протоколу подготовки матрицы SMRTBell), кольцевые молекулы можно было секвенировать более одного раза, тем самым создавая множество субпрочтений молекулы. Субпрочтения могут использоваться для создания кольцевых консенсусных последовательных (ККП) субпрочтений. В целом, для всех описанных в данном документе способов один ZMW может генерировать множество субпрочтений, но соответствовать только одному ККП прочтению.

В некоторых вариантах осуществления, полностью неметилированный набор данных может быть создан с помощью ПЦР на фрагментах ДНК человека. Например, полностью метилированный набор

данных может быть получен с помощью фрагментов ДНК человека, обработанных CpG-метилтрансферазой M.SssI, для которых предполагается, что все сайты CpG метилированы. В других примерах, может использоваться другая CpG-метилтрансфераза, такая как M.MpeI. В других вариантах осуществления, синтетические последовательности с известными состояниями метилирования или ранее существовавшие образцы ДНК с разными уровнями метилирования, или гибридными состояниями метилирования, создаваемые рестрикционным ферментом, расщепляющим метилированные и неметилированные молекулы ДНК с последующим лигированием (что могло бы создать часть химерных метилированных/неметилированных молекулы ДНК) могут быть использованы для обучения моделей предсказания или классификаторов метилирования.

Трансформация кинетических паттернов, включая контекст последовательности, МИП и ширину импульса (ШИ), может представлять собой 2-мерную матрицу, содержащую характеристики из цепей Уотсона и Крика для анализа состояний метилирования в сайтах CG, как показано на фиг. 15. Данный подход позволил нам точно уловить тонкие кинетические изменения, вызванные метилированными цитозинами, а также контекст их ближайшей последовательности. Как и в случае любого из различных способов, описанных в данном документе, для каждого CpG, присутствующего в субпрочтении, окно измерения (например, 3 основания выше и ниже цитозина сайта CpG) может быть использовано для последующих анализов, что приводит к тому, что совокупно 7 нуклеотидов (включая цитозин сайта CpG) анализируются вместе. Могут быть вычислены МИП и ШИ для каждого основания из этих 7 нуклеотидов. Чтобы уловить контекст последовательности, относящийся к кинетическим изменениям, сигналы МИП и ШИ могут быть скомпилированы в конкретный сигнал основания, относительные секвенированные позиции и информацию о цепи, как показано на фиг. 15. Такая структура данных для простоты называется 2-мерной цифровой матрицей кинетики.

Такая 2-мерная цифровая матрица аналогична "2-мерному цифровому изображению". Например, первая строка 2-мерной цифровой матрицы содержала относительные позиции, окружающие цитозин локуса CpG, который был подвергнут анализу метилирования, с 3-нт выше и ниже данного сайта цитозина. Позиция 0 соответствует сайту цитозина, метилирование которого необходимо было определить. Относительные позиции -1 и -2 обозначали 1-нт и 2-нт выше цитозина, о котором идет речь. Относительные позиции +1 и +2 обозначали 1-нт и 2-нт ниже цитозина, который будут использовать. Каждая позиция будет соответствовать 2 столбцам, которые содержат соответствующие значения МИП и ШИ. Каждая строка соответствует 4 типам нуклеотидов (A, C, G и T) в цепях Уотсона и Крика. Заполнение значений МИП и ШИ в матрице зависело от того, какой соответствующий тип нуклеотида был представлен в результате секвенирования (т.е. субпрочтении) в конкретной позиции.

Как показано на фиг. 15, в относительной позиции 0, значения МИП и ШИ были показаны в строке "C" в цепи Уотсона, предполагая, что был сигнал цитозина в той позиции. Другие ячейки в столбце, которые не соответствуют секвенированному основанию, будут кодироваться как "0". В качестве примера информация о последовательности, соответствующая 2-мерной цифровой матрице (фиг. 15), будет выглядеть следующим образом: 5'-ATACGTT-3' и 5'-TAACGTA-3' для цепей Уотсона и Крика соответственно. В этом контексте будут разными восходящая и нисходящая последовательности, фланкирующие цитозин сайта CpG в цепях Уотсона и Крика. Поскольку метилирование по сайтам CpG должно быть симметричным между цепями Уотсона и Крика (Lister et al., 2009), кинетику в обеих цепях использовали для обучения модели прогнозирования метилирования в одном предпочтительном варианте. В другом варианте осуществления, цепи Уотсона и Крика по отдельности могут использоваться для обучения модели прогнозирования метилирования.

Принимая во внимание высокую пропускную способность секвенирования отдельной молекулы в реальном времени, в одном варианте осуществления, алгоритм глубокого обучения (например, сверточные нейронные сети (CHC)) (LeCun et al., 1989) может подходить для различения метилированных CpG от неметилированных CpG. Также могут быть использованы другие алгоритмы в дополнение к или вместо, например, но не ограничиваясь лишь этими: линейной регрессии, логистической регрессии, глубокой рекуррентной нейронной сети (например, долгая-краткосрочная память, LSTM), байесовскому классификатору, скрытой модели Маркова (HMM), линейному дискриминантному анализу (LDA), кластеризации k-средних, плотностному алгоритму кластеризации пространственных данных с присутствием шума (DBSCAN), алгоритму случайного леса и машине опорных векторов (SVM), и т.д. При обучении можно использовать цепи Уотсона и Крика по отдельности или в скомбинированной новой матрице, как описано на фиг. 6-8.

Другое преобразование кинетических паттернов может представлять собой N-мерную матрицу. N может представлять собой, например, 1, 3, 4, 5, 6 и 7. Например, 3-мерная матрица будет представлять собой стек 2-мерных матриц, распределенных согласно количеству тандемных сайтов CG для анализируемого участка ДНК, в котором 3-м измерением будет количество тандемных сайтов CG в данном участке ДНК. Сила импульса или величина импульса (например, измеренная по высоте пика импульса или по площади под сигналом импульса) также могут быть включены в матрицу в некоторых вариантах осуществления. Сила импульса (показатель амплитуды пика импульса, фиг. 3) может быть либо добавлена в дополнительный столбец рядом со столбцами в ассоциации со значениями ШИ и МИП поверх исходной

2-мерной матрицы, либо добавлена в 3-тье измерение для формирования 3-мерной матрицы.

В качестве дополнительных примеров, 2-мерная матрица размером 8(строка) x21 (столбец) может быть преобразована в 1-мерную матрицу (т.е. вектор), содержащую 168 элементов. И мы можем сканировать эту 1-мерную матрицу, например, для реализации CHC или другого моделирования. В качестве другого примера, способы могут разбить 2-мерную матрицу 8x21 на несколько меньших матриц, например, две 2-мерные матрицы 4x21. Объединение этих двух меньших матриц вместе в вертикальном направлении дает 3-мерную матрицу (т.е. x=21, y=4, z=2). Способы могут сканировать 1-ую 2-мерную матрицу, а затем 2-ую 2-мерную матрицу, чтобы сформировать представление данных для машинного обучения. Данные могут быть дополнительно разделены, чтобы сформировать матрицу более высокой размерности. Дополнительно к структуре данных может быть добавлена информация о вторичной структуре, например, дополнительная матрица (1-мерная матрица) поверх 2-мерной матрицы. Такая дополнительная матрица может кодировать то, участвует ли каждое основание в пределах окна измерения в вторичной структуре (например, структуре "петля-на-стебле"), например, основание, задействованное в "стебле", кодируется как 0, а основание, задействованное в "петле", кодируется как 1.

В одном варианте осуществления, статус метилирования сайта CpG в пределах одной молекулы ДНК может быть выражен как вероятность подвергнутся метилированию на основе статистической модели, а не в виде предоставления качественного результата - "метилирован" или "неметилирован". Вероятность, равная 1, указывает на то, что на основании статистической модели, сайт CpG можно считать метилированным. Вероятность, равная 0, указывает на то, что на основании статистической модели, сайт CpG можно считать неметилированным. В последующем нисходящем анализе пороговое значение можно использовать для классификации того, классифицируется ли конкретный сайт CpG как "метилированный" или "неметилированный" на основе вероятности.

Возможные значения порога включают в себя 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90 или 95%. Прогнозируемая вероятность подвергнутся метилированию для сайта CpG, превышающая заранее определенное пороговое значение, может быть классифицирована как "метилировано", в то время как вероятность метилирования для сайта CpG не выше заранее определенного порогового значения может быть классифицирована как "неметилировано". Желаемое пороговое значение может быть получено из набора обучающих данных с использованием, например, анализа кривой операционных характеристик приемника (ROC).

Фиг. 16 демонстрирует общую процедура классификации состояний метилирования неизвестного образца из цепей Уотсона и Крика. Образец с неизвестным состоянием метилирования был подвергнут секвенированию отдельной молекулы в реальном времени. Субпрочтения секвенирования могут быть выровнены с эталонным геномом или друг с другом, как и с помощью других способов, для определения консенсусных значений (например, среднего, медианы, моды или другого статистического параметра) для данной позиции. Как показано, измеренные значения для двух цепей можно объединить в единую 2-мерную матрицу.

Для цитозина сайта CG в результате выравнивания можно было получить МИП, ШИ и контекст последовательности из цепи Уотсона с использованием сопоставимого окна измерения (3-нт выше и ниже цитозина сайта CpG), как применяется в стадии обучения (фиг. 16), связаннного с тем основанием, чья модификация подлежала определению, хотя могут использоваться окна разного размера. Такую 2-мерную матрицу тестового образца можно сравнить с эталонными кинетическими паттернами, показанными на фиг. 16, для определения состояний метилирования.

III. Пример обучения модели для обнаружения метилирования.

Чтобы проверить осуществимость и правильность предложенных подходов, мы подготовили библиотеку ДНК плаценты с обработкой M.SssI (метилированная библиотека) и амплификацией ПЦР (неметилированная библиотека) перед секвенированием отдельной молекулы в реальном времени. Мы получили 44799736 и 43580452 субпрочтений для метилированной и неметилированной библиотек, соответственно, что соответствует 421614 и 446285 кольцевым консенсусным последовательностям (ККП). В результате каждая молекула была секвенирована в среднем 34 и 32 раза в метилированной и неметилированной библиотеках. Набор данных был получен из ДНК, подготовленной с помощью набора Sequel Sequencing Kit 3.0 от Pacific Biosciences. Этот набор был разработан для использования с оригинальным секвенатором Pacific Biosciences Sequel. Чтобы отличить Sequel от его преемника, Sequel II, в данном документе мы будем обозначать исходный Sequel как Sequel I. Следовательно, набор Sequel Sequencing Kit 3.0 будет упоминаться в данном документе как Sequel I Sequencing Kit 3.0. Наборы для секвенирования, разработанные для секвенатора Sequel II, включают в себя Sequel II Sequencing Kit 1.0 и Sequel II Sequencing Kit 2.0, которые также описаны в данном раскрытии изобретения.

Мы использовали 50% секвенированных молекул, сгенерированных из метилированной и неметилированной библиотек, для обучения статистической модели (и использовали оставшиеся 50% для проверки), которая в данном случае является моделью сверточной нейронной сети (CHC). Например, модель CHC может иметь один или большее количество сверточных слоев (например, 1-мерный или 2-мерные слои). Сверточный слой может использовать один или большее количество различных фильтров, причем каждый фильтр использует ядро, которое работает с значениями матрицы, локальными (например, со-

седними или окружающими) для конкретного элемента матрицы, тем самым обеспечивая новое значение для конкретного элемента матрицы. В одном варианте осуществления, использовались два 1-мерных сверточных слоя (каждый со 100 фильтрами с размером ядра 4). Фильтры можно применять по отдельности, а затем комбинировать (например, с использованием взвешенного среднего). Результирующая матрица может быть меньше входной матрицы.

За сверточными слоями может следовать слой ReLU (блок линейной ректификации), за которым может следовать слой отсева с величиной отсева 0,5. ReLU является примером функции активации, которая может работать с отдельными значениями, в результате чего получается новая матрица (изображение) из сверточного слоя(слоев). Также можно использовать другие функции активации (например, сигмоидальную, многопеременную логистическую (softmax) и т.д.). Можно использовать один или большее количество таких слоев. Слой отсева может использоваться на слое ReLU или на слое максимума подвыборки (субдискретизации, пулинга), и служить для регуляризации для предотвращения переобучения. Слой отсева может использоваться в процессе обучения, чтобы игнорировать различные (например, случайные) значения во время различных итераций процесса оптимизации (например, для уменьшения функции затрат/потерь), который выполняется как часть обучения.

После слоя ReLU может использоваться слой максимума подвыборки (например, размер подвыборки 2). Слой максимума подвыборки может действовать аналогично сверточному слою, но вместо скалярного произведения между входными данными и ядром может быть взят максимум области из входных данных, перекрываемых ядром. Может использоваться дополнительный сверточный слой(слои). Например, данные из слоя подвыборки могут представлять собой входные данные для других двух 1-мерных сверточных слоев (например, каждый со 128 фильтрами с размером ядра 2, за которым следует слой ReLU), дополнительно используя слой отсева с значением отсева 0,5. Использовался слой максимума подвыборки с размером подвыборки 2. Наконец, может быть использован полносвязный слой (например, с 10 нейронами, за которыми следует слой ReLU). За выходным слоем с одним нейроном может следовать сигмоидальный слой, и как результат получают вероятность метилирования. Могут быть адаптированы различные настройки слоев, фильтров и размеров ядра. В данном наборе обучающих данных мы использовали 468596 и 432761 сайтов CpG из метилированной и неметилированной библиотек.

А. Результаты обучающих и тестовых наборов данных.

Фиг. 17А демонстрирует вероятность метилирования для каждого сайта CpG в каждой отдельной молекуле ДНК в наборе обучающих данных. Вероятность метилирования была намного выше в метилированной библиотеке, чем в неметилированной. При пороговом значении 0,5 для вероятности метилирования, 94,7% неметилированных сайтов CpG были правильно предсказаны как неметилированные, а 84,7% метилированных CpG были правильно предсказаны как метилированные.

Фиг. 17В демонстрирует эффективность тестового набора данных. Мы использовали модель, обученную на обучающем наборе данных, для прогнозирования состояний метилирования 469729 и 432024 сайтов CpG в независимом тестовом наборе данных из метилированной и неметилированной библиотек. Для порогового значения 0,5 для вероятности подвергнутся метилированию, 94,0% неметилированных сайтов CpG были правильно предсказаны как неметилированные, а 84,1% метилированных CpG были правильно предсказаны как метилированные. Эти результаты предполагают, что использование нового преобразования кинетики в комбинации с контекстом последовательности может сделать возможным определение состояний метилирования в ДНК (например, у людей).

Мы оценили эффективность каждой характеристики (контекст последовательности, МИП и ШИ) в прогнозировании состояния метилирования CpG путем включения подмножества характеристик в модель. В обучающем наборе данных модели с (i) только контекстом последовательности, (ii) только МИП и (iii) только ШИ дали значения площади под кривой (AUC), составляющие 0,5, 0,74 и 0,86 соответственно. Тогда как при комбинировании МИП и контекста последовательности эффективность улучшилась, с AUC, составляющей 0,86. Комбинированный анализ контекста последовательности ("Seq"), МИП и ШИ существенно улучшил эффективность, с AUC, составляющей 0,94 (фиг. 18А). Эффективность для независимого тестового набора данных была сопоставима с таковой для обучающего набора данных (фиг. 18В).

Мы определили глубину субпрочтения сайта CpG как среднее количество субпрочтений, покрывающих его и окружающие его на 10 п.н. Как показано на фиг. 19А и 19В, чем больше глубина субпрочтений сайта CpG, тем более высокой точности обнаружения метилирования мы могли достичь. Например, как показано в тестовом наборе данных (фиг. 19В), если глубина каждого сайта CpG составляла по меньшей мере 10, AUC предсказания состояний метилирования составляла бы 0,93. Хотя, если глубина субпрочтений каждого сайта CpG составляет по меньшей мере 300, AUC предсказания состояний метилирования будет 0,98. С другой стороны, даже для глубины, составляющей 1, мы могли достичь AUC 0,9, предполагая, что наш подход может обеспечить предсказание метилирования с использованием малой глубины секвенирования.

Чтобы проверить влияние информации о цепи на эффективность анализа метилирования, использовали контекст последовательности, МИП и ШИ, полученные из цепей Уотсона и Крика, для обучения согласно вариантам осуществления, представленным в данном раскрытии изобретения, соответственно.

Фиг. 20A и 20B демонстрируют, что можно использовать одну цепь, а именно цепь Уотсона или Крика, для обучения и тестирования, поскольку может быть достигнута AUC, составляющая вплоть до 0,91 и 0,87 в обучающих и тестовых наборах данных. Использование обеих цепей (например, как описано на фиг. 6-8), включая цепи Уотсона и Крика, обеспечило бы лучшую эффективность (AUC: 0,94 и 0,90 в обучающих и тестовых наборах данных, соответственно), предполагая, что информация о цепях будет важной для достижения оптимальной эффективности.

Мы дополнительно протестировали различное количество нуклеотидов выше и ниже сайта CpG, чтобы изучить, как этот параметр влияет на эффективность согласно вариантам осуществления, представленным в данном раскрытии изобретения, разработанным в данном изобретении. Фиг. 21A и 21B демонстрируют, что количество нуклеотидов выше и ниже цитозина в контексте CpG может влиять на точность предсказания метилирования. Например, в целях иллюстрации, рассматривая, но не ограничиваясь лишь 2 нуклеотидами (нт), 3, 4, 6, 8, 10, 15 и 20 нт выше и ниже анализируемого цитозина, AUC способа, использующего 2 нт выше и ниже исследуемого цитозина, который подвергается анализу, будет составлять только 0,50 в обучающем и тестовом наборах данных, тогда как AUC способа с использованием 15 нт выше и ниже исследуемого цитозина будет увеличиваться до 0,95 и 0,92 в обучающем и тестовом наборах данных. Эти результаты предполагают, что изменение длины восходящих и нисходящих областей, фланкирующих анализируемые цитозины, позволит определить оптимальную эффективность. В одном варианте осуществления, как показано на фиг. 21B, можно было бы использовать 3 нт выше и ниже цитозина для определения состояний метилирования, что могло бы дать AUC, составляющее 0,89.

В одном варианте осуществления можно использовать асимметричные последовательности, фланкирующие исследуемый цитозин, для выполнения анализа, согласно вариантами осуществления, представленными в данном раскрытие изобретения. Например, могут использоваться 2 нт выше, скомбинированные с 1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 30, 35, и 40 нт ниже цитозина; могут использоваться 3 нт выше, скомбинированные с 1, 2, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 30, 35, и 40 нт ниже цитозина; могут использоваться 4 нт выше, скомбинированные с 1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 30, 35, и 40 нт ниже цитозина. Например, могут использоваться 2 нт ниже, скомбинированные с 1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 30, 35, и 40 нт выше цитозина; могут использоваться 3 нт ниже, скомбинированные с 1, 2, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 30, 35, и 40 нт выше цитозина; могут использоваться 4 нт ниже, скомбинированные с 1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 30, 35, и 40 нт выше цитозина. Использование преимуществ МИП, ШИ, информации о цепи и контекста последовательности в комбинации с n-нуклеотидами выше и m-нуклеотидами ниже цитозина может обеспечить повышенную точность определения состояний метилирования в некоторых вариантах осуществления. Такие различные окна измерения могут применяться к другим типам анализа модификаций оснований, таким как 5hmC, 6mA, 4mC и oxoG, или к любой модификации, описанной в данном документе. Такие различные окна измерения могут включать в себя анализ вторичной структуры ДНК, например, структуры G-квадруплекс и петля-на-стебле. Такой пример объясняется выше. Такую информацию о вторичной структуре можно также добавить в качестве другого столбца в матрицу.

Фиг. 22A и 22B демонстрируют, что возможно определить состояния метилирования с использованием кинетических паттернов, связанных только с нисходящими основаниями, составляющими по меньше мере 3 основания. Согласно вариантам осуществления, представленным в данном раскрытии изобретения, с использованием характеристик, связанных с цитозином и его нисходящими 3, 4, 6, 8 и 10 основаниями, AUC определения состояний метилирования в обучающем наборе данных составляли 0,91, 0,92, 0,94, 0,94 и 0,94, соответственно, в обучающем наборе данных; AUC составляли 0,87, 0,88, 0,90, 0,90 и 0,90, соответственно, в тестовом наборе данных.

Фиг. 23A и 23B демонстрируют, однако, что, если использовать только характеристики, связанные с восходящими основаниями, эффективность классификации, по-видимому, уменьшается в способности различать состояния метилирования. Все значения AUC в обучающем наборе данных и тестовом наборе данных составляли 0,50 для от 2 до 10 восходящих оснований.

Фиг. 24 и 25 демонстрируют, что различные комбинации восходящих и нисходящих оснований позволят достичь оптимальной эффективности классификации при определении состояний метилирования. Например, характеристики, связанные с 8 основаниями выше и 8 основаниями ниже цитозина, будут иметь наилучшую эффективность в этом наборе данных с AUC, составляющей 0,94 и 0,91 в обучающем и тестовом наборах данных, соответственно.

Фиг. 26 демонстрирует относительную важность характеристик в отношении классификации состояний метилирования в сайтах CpG. "W" и "C" в скобках обозначают информацию о цепи, "W" для цепи Уотсона и "C" для цепи Крика. Важность каждой характеристики, включая контекст последовательности, МИП и ШИ определялась с использованием алгоритма случайного леса. Анализ с помощью алгоритма случайного леса продемонстрировал, что значимость характеристик МИП и ШИ достигает пика ниже исследуемого цитозина, показывая, что основной вклад в способность к классификации вносят МИП и ШИ ниже исследуемого цитозина.

Случайный лес состоял из нескольких деревьев решений. При построении дерева решений исполь-

зовали примесь Джини, чтобы определить, какая должна быть принята логика решения для узлов решений. Важные характеристики, которые имеют большее влияние на окончательный результат классификации, вероятно, были в узлах ближе к корню дерева решений, в то время как неважные характеристики, которые имеют меньшее влияние на окончательный результат классификации, скорее всего, находятся в узлах дальше от корня. Таким образом, важность характеристики может быть оценена путем вычисления среднего расстояния относительно корней всех деревьев решений в случайном лесу.

В некоторых вариантах осуществления, консенсус сигналов метилирования в сайтах CpG между цепями Уотсона и Крика может быть дополнительно использован для улучшения специфичности. Например, может потребоваться, чтобы обе цепи, демонстрирующие метилирование, обозначались как метилированное состояние, а обе цепи, демонстрирующие неметилирование, обозначались как неметилированное состояния. Поскольку известно, что метилирование в сайтах CpG обычно симметрично, подтверждение из каждой цепи может улучшить специфичность.

В различных вариантах осуществления, совокупные кинетические характеристики всей молекулы можно использовать для определения состояний метилирования. Например, метилирование всей молекулы будет влиять на кинетику всей молекулы во время секвенирования отдельной молекулы в реальном времени. Моделируя кинетику секвенирования всей молекулы ДНК-матрицы, включая МИП, ШИ, размеры фрагментов, информацию о цепи и контекст последовательности, можно повысить точность классификации в отношении того, метилирована ли молекула или нет. Например, окна измерения могут представлять собой всю матричную молекулу. Статистические параметры (например, среднее значение, медиана, мода, процентиль и т.д.) для МИП, ШИ или других кинетических характеристик могут использоваться для определения метилирования всей молекулы.

B. Ограничения других методов анализа.

Сообщалось, что обнаружение метилирования на основе МИП для конкретного C в конкретном мотиве последовательности было очень низким, например, чувствительность составляла всего 1,9% (Clark et al., 2013). Мы также попытались воспроизвести такой анализ, комбинируя различные мотивы последовательностей с МИП без использования метрики ШИ, и просто используя пороговое значение для МИП, а не структуры данных, как описано в данном документе. Например, были извлечены 3-нт выше и ниже, фланкирующие исследуемый CpG. МИП этого CpG были разделены на разные группы (4096 групп для 6 позиций) в зависимости от контекста фланкирующих последовательностей из 6-нт (т.е. 3 восходящих и нисходящих нуклеотида, соответственно), для которых центром был этот CpG. МИП для метилированных и неметилированных CpG в пределах одного и того же мотива последовательности изучали с помощью ROC. Например, сравнивали МИП CpG в неметилированном мотиве "AATCGGAC" и метилированном мотиве "AAT^{m}CGGAC", показывая AUC, составляющее 0,48. Таким образом, использование пороговых значений в конкретной группе последовательностей работало плохо по сравнению с вариантами осуществления, в которых используются различные.

Фиг. 27 демонстрирует эффективность вышеупомянутого анализа МИП на основе мотивов (Beckmann et al. BMC Bioinformatics. 2014) для обнаружения метилирования без использования сигнала ширины импульса. Диаграммы с вертикальными столбиками представляют собой усредненные значения AUC для различных k-мерных мотивов, фланкирующих изучаемые сайты CpG (т.е. количество оснований, окружающих исследуемые сайты CpG). Фиг. 27 продемонстрировала, что усредненные значения AUC для дискриминирующих способностей на основе МИП между метилированными и неметилированными цитозинами среди различных k-мерных мотивов (например, 2-мер, 3-мер, 4-мер, 6-мер, 8-мер, 10-мер, 15-мер, 20-мер, окружающий затрагиваемые сайты CpG) оказались меньше чем 60%. Эти результаты предполагают, что рассмотрение МИП нуклеотида-кандидата в данном контексте мотива без учета МИП соседних нуклеотидов (Flusberg et al., 2010) будет менее выигрышным, чем способы, описанные в данном документе для определения метилирования CpG.

Мы также протестировали способ, представленный в исследовании Flusberg et al. (Flusberg et al., 2010). Мы проанализировали в общей сложности 5948348 сегментов ДНК, которые находились 2-нт выше и 6-нт ниже цитозина, который подвергался анализу метилирования. Было 2828848 сегментов, которые были метилированы, и 3119500 сегментов были неметилированы. Как показано на фиг. 28, было обнаружено, что сигналы, выведенные из анализа главных компонентов с использованием МИП и ШИ, в значительной степени перекрываются между фрагментами с метилированными цитозинами (mC) и неметилированными цитозинами (C), что позволяет предположить, что способу, описанному Flusberg et al. недостает соответствующей точности. Эти результаты свидетельствуют о том, что анализ главных компонентов, который линейно комбинирует значения ШИ и МИП по основаниям и соседним основаниям, используемый в исследовании Flusberg et al. (Flusberg et al., 2010), не может надежно или достоверно дифференцировать 5-метилцитозин и неметилированные цитозины.

Фиг. 29 демонстрирует, что AUC способа, основанного на анализе главных компонентов, для которого в исследовании Flusberg et al. (Flusberg et al., 2010) использовались два основных компонента, включающих в себя МИП и ШИ, были намного менее точными (AUC: 0,55), чем подход на основе сверточной нейронной сети, использующей МИП и ШИ, а также контекст последовательности, как показано в нашем изобретении (AUC: 0,94).

C. Другие математические/статистические модели.

В другом варианте осуществления, другие математические/статистические модели, например, включающие в себя, но не ограниченные случайным лесом и логистической регрессией, могут быть обучены путем адаптации характеристик, разработанных выше.

Что касается модели CHC, обучающие и тестовые наборы данных были построены из ДНК с обработкой M.SssI (метилированная) и амплификацией ПЦР (неметилированная), которые использовались для обучения алгоритма случайного леса (Breiman, 2001). В этом анализе случайным лесом мы описали каждый нуклеотид с 6 характеристиками: МИП, ШИ и 4-компонентный бинарный вектор, кодирующий обозначение основания. В таком бинарном векторе A, C, G и T были закодированы с помощью [1,0,0,0], [0,1,0,0], [0,0,1,0] и [0,0,0,1], соответственно. Для каждого анализируемого сайта CpG мы включили информацию о его 10 нуклеотидах выше и ниже в обеих цепях, формируя 252-размерный (252-мерный) вектор, где каждая характеристика представляет одно измерение. Обучающий набор данных, описанный выше с помощью 252-мерных векторов, использовался для обучения модели случайного леса, а также модели логистической регрессии. Обученная модель была использована для прогнозирования состояний метилирования в независимом тестовом наборе данных. Случайный лес состоял из 100 деревьев решений. При построении дерева использовались бутстрэп-выборки. При разбиении узла каждого дерева решений для определения наилучшего разбиения использовалась примесь Джини, и в каждом разбиении учитывалось максимум 15 характеристик. Кроме того, каждый лист дерева решений должен был содержать не меньше чем 60 образцов.

Фиг. 30A и 30B демонстрируют эффективность способа с использованием случайного леса и логистической регрессии для прогнозирования метилирования. Фиг. 30A демонстрирует значения AUC в обучающем наборе данных для CHC, случайного леса и логистической регрессии. Фиг. 30B демонстрирует значения AUC в тестовом наборе данных для CHC, случайного леса и логистической регрессии. AUC способа с использованием случайного леса достигало 0,93 и 0,86 в обучающем и тестовом наборах данных, соответственно.

Обучающий набор данных, описанный с помощью тех же 252-мерных векторов, использовался для обучения модели логистической регрессии. Обученная модель была использована для прогнозирования состояний метилирования в независимом тестовом наборе данных. Модель логистической регрессии с регуляризацией L2 (Ng and Y., 2004) была подогнана с помощью обучающего набора данных. Как показано на фиг. 30A и 30B, AUC способа, использующего логистическую регрессию, достигнет 0,87 и 0,83 в обучающем и тестовом наборе данных, соответственно.

Таким образом, эти результаты предполагают, что определенные модели (например, но не без ограничения, случайный лес и логистическая регрессия), отличные от CHC, могут быть использованы для анализа метилирования с использованием признаков и аналитических протоколов, которые мы разработали в данном изобретении. Эти результаты также предполагают, что CHC, реализованная в соответствии с вариантами осуществления в данном изобретении, с значением AUC, составляющим 0,90 в тестовом наборе данных (фиг. 30B) превосходила как случайный лес (AUC: 0,86), так и логистическую регрессию (AUC: 0,83).

D. Определение 6mA модификаций нуклеиновых кислот.

Помимо метилированного CpG, описанные в данном документе способы могут также обнаруживать другие модификации оснований ДНК. Например, может быть обнаружен метилированный аденин, в том числе в форме 6mA.

1. Обнаружение 6mA с использованием кинетических характеристик и контекста секвенирования.

Чтобы оценить эффективность и полезность вариантов осуществления, раскрытых касательно определения модификаций оснований нуклеиновых кислот, мы дополнительно проанализировали метилирование N6-аденина (6mA). В одном варианте осуществления, примерно 1 нг ДНК человека (например, экстрагированной из тканей плаценты) амплифицировали с получением 100 нг продукта ДНК путем амплификации всего генома с неметилированным аденином (uA), неметилированным цитозином (C), неметилированным гуанином (G) и неметилированным тимин (T).

Фиг. 31A демонстрирует пример одного подхода для генерации молекул с неметилированными аденинами путем амплификации всего генома. На фигуре "uA" обозначает неметилированный аденин, а "mA" обозначает метилированный аденин. Полногеномную амплификацию выполняли с использованием устойчивых к экзонуклеазам тиофосфат-модифицированных случайных гексамеров в качестве праймеров, которые случайным образом связываются по всему геному, что позволяет полимеразе (например, ДНК-полимеразе Phi29) амплифицировать ДНК (например, путем изотермической линейной амплификации). На стадии 3102 денатурируют двухцепочечную ДНК. На стадии 3106 инициируют реакцию амплификации, когда ряд случайных гексамеров (например, 3110) отжигается с денатурированной ДНК-матрицей (т.е. одноцепочечной ДНК). Как показано в 3114, когда опосредованный гексамером синтез ДНК цепи 3118 продолжался в направлении от 5' до 3' и достигал следующего сайта синтеза ДНК, опосредованного гексамером, полимераза вытесняла вновь синтезированную цепь ДНК (3122) и продолжила удлинение цепи. Вытесненные цепи снова становились одноцепочечными матрицами ДНК для связывания случайных гексамеров и могли инициировать новый синтез ДНК. Повторный отжиг гексамера и вы-

теснение цепи в изотермическом процессе может привести к высокому выходу амплифицированных продуктов ДНК. Это амплификация, описанная в данном документе, может быть частным случаем метода амплификации множественного вытеснения (MDA).

Амплифицированные продукты ДНК дополнительно фрагментировали, например, но без ограничений, на фрагменты с размером 100, 200, 300, 400, 500, 600, 700, 800, 900 п.о., 1, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100 т.п.о., или с другими желательными диапазонами размеров. Процесс фрагментации может включать в себя ферментативное расщепление, распыление, гидродинамическое фрагментирование, обработку ультразвуком и т.д. В результате, исходные модификации оснований, такие как 6mA, могут быть почти полностью устранены путем амплификации всего генома с неметилированным A (uA). Фиг. 31A демонстрирует возможные фрагменты (3126, 3130 и 3134) продуктов ДНК, причем обе цепи имеют неметилированный A. Такие продукты ДНК амплификации всего генома без mA подвергали секвенированию отдельной молекулы в реальном времени для создания набора данных uA.

Фиг. 31B демонстрирует пример одного подхода для генерирования молекул с метилированными аденинами путем амплификации всего генома. На фигуре "uA" обозначает неметилированный аденин, а "mA" обозначает метилированный аденин. Примерно 1 нг ДНК человека амплифицировали для получения 10 нг продукта ДНК путем амплификации всего генома с 6 мА и неметилированными C, G и T. Метилированные аденины могут быть получены посредством ряда химических реакций (J D Engel et al. J Biol Chem. 1978;253:927-34). Как проиллюстрировано на фиг. 31B, амплификацию всего генома проводили с использованием устойчивых к экзонуклеазам тиофосфат-модифицированных случайных гексамеров в качестве праймеров, которые случайным образом связываются по всему геному, что позволяет полимеразе (например, ДНК-полимеразе Phi29) амплифицировать ДНК (например, путем изотермической линейной амплификации), аналогично фиг. 31 A. Устойчивые к экзонуклеазам тиофосфат-модифицированные случайные гексамеры устойчивы к 3'→5' экзонуклеазной активности корректирующих ДНК-полимераз. Таким образом, во время амплификации случайные гексамеры будут защищены от деградации.

Реакция амплификации инициировалась, когда ряд случайных гексамеров отжигали с денатурированной ДНК-матрицей (т.е. одноцепочечной ДНК). Когда опосредованный гексамером синтез ДНК продолжался в направлении от 5' до 3' и достигал следующего сайта синтеза ДНК, опосредованного гексамером, полимераза вытесняла вновь синтезированную цепь ДНК и продолжала удлинение цепи. Вытесненные цепи становились одноцепочечными матрицами ДНК для связывания случайных гексамеров и могли инициировать новый синтез ДНК. Повторный отжиг гексамера и вытеснение цепи в изотермическом процессе может привести к высокому выходу амплифицированных продуктов ДНК.

Амплифицированные продукты ДНК дополнительно фрагментировали, например, но без ограничений, на фрагменты с размером 100, 200, 300, 400, 500, 600, 700, 800, 900 п.о., 1, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100 т.п.о., или с другими комбинациям по длине. Как показано на фиг. 31B, амплифицированные продукты ДНК будут включать в себя различные формы паттернов метилирования по сайтам аденина в каждой цепи. Например, обе цепи двухцепочечной молекулы могут быть метилированы по отношению к аденинам (молекула I), которые будут образовываться, когда две цепи получают в результате синтеза ДНК во время амплификации всего генома.

В качестве другого примера, одна цепь двухцепочечной молекулы может содержать чередующиеся паттерны метилирования по сайтам аденина (молекула II). Паттерн чередующегося метилирования определяется как паттерн, который включает в себя смесь метилированных и неметилированных оснований, присутствующих в цепи ДНК. В следующих примерах мы используем чередующийся паттерн метилирования аденина, который включает в себя смесь метилированных и неметилированных аденинов, присутствующих в цепи ДНК. Этот тип двухцепочечной молекулы (молекула II) возможно было генерировать потому, что неметилированный гексамер, содержащий неметилированные аденины, был связан с цепью ДНК и инициировал удлинение ДНК. Мог быть секвенирован такой амплифицированный продукт ДНК, содержащий гексамер с неметилированными аденинами. В альтернативном варианте, этот тип двухцепочечной молекулы (молекула II) мог быть инициирован фрагментированной ДНК из исходной ДНК-матрицы, содержащей неметилированные аденины, поскольку такую фрагментированную ДНК можно было связать с цепью ДНК в качестве праймера. Мог быть секвенирован такой амплифицированный продукт ДНК, содержащий часть исходной ДНК с неметилированными аденинами в цепи. Поскольку неметилированные гексамерные праймеры составляют лишь небольшую часть образующихся цепей ДНК, большинство фрагментов все равно будут содержать 6mA.

В качестве другого примера, одна цепь двухцепочечной молекулы ДНК может быть метилирована по сайтам аденина, но другая цепь может быть неметилированной (Молекула III). Этот тип двухцепочечной молекулы может быть получен, когда исходная цепь ДНК без метилированных аденинов предоставляется в качестве молекулы ДНК-матрицы для получения новой цепи с метилированными аденинами.

Обе цепи могут быть неметилированными (молекула IV). Этот тип двухцепочечной молекулы может быть результатом повторного отжига двух исходных цепей ДНК без метилированных аденинов.

Процесс фрагментации может включать в себя ферментативное расщепление, распыление, гидродинамическое фрагментирование, обработку ультразвуком и т.д. Такие продукты ДНК амплификации

всего генома могут быть почти целиком метилированы касаемо сайтов A. Эту ДНК с mA подвергали секвенированию отдельной молекулы в реальном времени для создания набора данных mA.

Для набора данных uA мы секвенировали 262608 молекул со средней длиной 964 п.о. с использованием секвенирования отдельной молекулы в реальном времени. Медиана глубины субпрочтений составляла 103x. 48% субпрочтений можно было выровнять с эталонным геномом человека с помощью программы выравнивания BWA (Li H et al. Bioinformatics. 2009;25:1754-60). В качестве примера можно использовать Sequel II System (Pacific Biosciences) для выполнения секвенирования отдельной молекулы в реальном времени. Фрагментированные молекулы ДНК были подвергнуты построению матрицы для секвенирования отдельной молекулы в реальном времени (SMRT) с использованием SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences). Отжиг праймеров для секвенирования и условия связывания полимеразы рассчитывали с помощью программного обеспечения SMRT Link v8.0 (Pacific Biosciences). Вкратце, праймер v2 для секвенирования отжигали с матрицей для секвенирования, а затем полимеразу связывали с матрицами с использованием Sequel II Binding and Internal Control Kit 2.0 (Pacific Biosciences). Секвенирование выполняли на Sequel II SMRT Cell 8M. Фильмы секвенирования были собраны на системе Sequel II в течение 30 ч с помощью Sequel II Sequencing Kit 2.0 (Pacific Biosciences).

Для набора данных mA мы секвенировали 804469 молекул со средней длиной 826 п.о. с использованием секвенирования отдельной молекулы в реальном времени. Медиана глубины субпрочтений составляла 34x. 27% субпрочтений можно было выровнять с эталонным геномом человека с помощью программы выравнивания BWA (Li H et al. Bioinformatics. 2009;25:1754-60).

В одном варианте осуществления, кинетические характеристики, включая, но не ограничиваясь, МИП и ШИ, были проанализированы специфичным для цепи способом. Для результатов секвенирования, полученных из цепи Уотсона, 644318 сайтов A без метилирования, случайно выбранных из набора данных uA, и 718586 сайтов A с метилированием, случайно выбранных из набора данных mA, были использованы для создания обучающего набора данных. Такой обучающий набор данных использовался для налаживания моделей классификации и/или пороговых значений для различения метилированных и неметилированных аденинов. Тестовый набор данных был составлен из 639702 сайтов A без метилирования и 723320 сайтов A с метилированием. Такой тестовый набор данных использовался для проверки эффективности модели/порога, установленных с помощью обучающего набора данных.

Мы проанализировали результаты секвенирования, полученные из цепей Уотсона. Фиг. 32A демонстрирует значения межимпульсного периода (МИП) в обучающем наборе данных для наборов данных uA и mA. Для обучающего набора данных значения МИП для секвенированных сайтов A были выше в наборе данных mA (медиана: 1,09; диапазон: 0-9,52), чем в наборе данных uA (медиана: 0,20; диапазон: 0-9,52) (значение $P<0,0001$; U-критерий Манна-Уитни).

Фиг. 32B демонстрирует МИП для тестового набора данных наборов данных uA и mA. Когда мы изучали значения МИП для секвенированных сайтов A в тестовом наборе данных, мы заметили, что значения МИП были выше в наборе данных mA, чем в наборе данных uA (медиана 1,10 в сравнении с 0,19; значение $P<0,0001$; U-критерий Манна-Уитни).

Фиг. 32C демонстрирует область под кривой операционных характеристик приёмника (ROC) с использованием порогового значения МИП. Частота истинно-положительных результатов представлена на оси ординат, а частота ложно-положительных результатов - на оси абсцисс. Площадь под кривой операционных характеристик приёмника (AUC) при различении секвенированных оснований A в матричных молекулах ДНК с метилированием и без метилирования с использованием соответствующих значений МИП составляла 0,86 как для обучающего, так и для тестового наборов данных.

Помимо результатов по цепям Уотсона, мы проанализировали результаты секвенирования, полученные по цепям Крика. Фиг. 33A демонстрирует значения МИП для обучающего набора данных uA и mA. Для обучающего набора данных значения МИП для секвенированных сайтов A были выше в наборе данных mA (медиана: 1,10; диапазон: 0-9,52), чем в наборе данных uA (медиана: 0,19; диапазон: 0-9,52) (значение $P<0,0001$; U-критерий Манна-Уитни).

Фиг. 34B демонстрирует значения МИП для тестового набора данных наборов данных uA и mA. Также наблюдали более высокие значения МИП для секвенированных сайтов A для тестового набора данных, по сравнению с набором данных uA (медиана 1,10 в сравнении с 0,19; значение $P<0,0001$; U-критерий Манна-Уитни).

Фиг. 33C демонстрирует площадь под кривой ROC. Частота истинно-положительных результатов представлена на оси ординат, а частота ложно-положительных результатов - на оси абсцисс. Значение площади под кривой ROC (AUC) при различении секвенированных оснований A в молекулах ДНК-матрицы с метилированием и без метилирования с использованием соответствующих значений МИП составляло 0,86 и 0,87 для обучающего и тестового наборов данных, соответственно.

Фиг. 34 демонстрирует иллюстрацию для определения 6mA цепи Уотсона с использованием окна измерения согласно вариантам осуществления данного изобретения. Такое окно измерения может включать в себя кинетические характеристики, такие как МИП и ШИ, и контекст близлежащей последовательности. Определение 6mA может быть выполнено так же, как определение метилированного CpG.

Фиг. 35 демонстрирует иллюстрацию для определения 6mA цепи Крика с использованием окна из-

мерения согласно вариантам осуществления данного изобретения. Такое окно измерения может включать в себя кинетические характеристики, такие как МИП и ШИ, и контекст близлежащей последовательности.

Например, 10 оснований с каждой стороны секвенированного основания A в исследуемой ДНК-матрице были использованы для создания окна измерения. Значения характеристик, включая МИП, ШИ и контекст последовательности, использовали для обучения модели с использованием сверточной нейронной сети (CNC) согласно раскрытым в данном документе способам. В других вариантах осуществления, статистические модели могут включать в себя, но без ограничения, 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 машину опорных векторов (SVM).

Фиг. 36A и 36B демонстрируют выясненную вероятность метилирования для секвенированных оснований A цепи Уотсона между наборами данных uA и mA с использованием окна измерения на основе модели CNC. Фиг. 36A демонстрирует, что модель CNC была обучена на обучающем наборе данных. Например, модель CNC использует два 1-мерных сверточных слоя (каждый с 64 фильтрами с размером ядра 4, за которым следует слой ReLU), с последующим слоем отсева с значением отсева 0,5. Использовался слой максимума подвыборки с размером подвыборки 2. Затем он переходил в два 1-мерных сверточных слоя (каждый с 128 фильтрами с размером ядра 2, за которыми следовал слой ReLU), далее использовался слой отсева с значением отсева 0,5. Использовался слой максимума подвыборки с размером подвыборки 2. Наконец, полносвязный слой с 10 нейронами, за которым следовал слой ReLU, с выходным слоем с одним нейроном, за которым следовал сигмоидальный слой, тем самым давая вероятность метилирования. Могут быть адаптированы другие настройки слоев, фильтров, размеров ядра, например, как описано в данном документе для другого метилирования (например, CpG). В этом обучающем наборе данных, касающемся результатов секвенирования цепи Уотсона, мы использовали 644318 и 718586 оснований A из неметилированной и метилированной библиотек.

Основываясь на модели CNC, для данных, связанных с цепью Уотсона, секвенированные основания A в молекулах ДНК-матрицы из набора данных mA служили источником гораздо более высокой вероятности метилирования как в обучающем, так и в тестовом наборах данных по сравнению с основаниями A, присутствующими в наборе данных uA (значение $P<0,0001$; U-критерий Манна-Уитни). Для обучающего набора данных средняя вероятность метилирования по сайтах A в наборе данных uA составляла 0,13 (межквартильный диапазон, МКД: 0,09-0,15), тогда как это значение в наборе данных mA составляло 1,000 (IQR: 0,998-1,000).

Фиг. 36A демонстрирует вероятность метилирования, определенную для тестового набора данных. Для обучающего набора данных средняя вероятность метилирования по сайтах A в наборе данных uA составляла 0,13 (межквартильный диапазон, МКД: 0,10-0,15), тогда как это значение в наборе данных mA составляло 1,000 (IQR: 0,997-1,000). Фиг. 36A и 36B демонстрируют, что модель CNC на основе окна измерения может быть обучена обнаружению метилирования в тестовом наборе данных.

Фиг. 37 представляет собой кривую ROC для обнаружения 6mA с использованием модели CNC на основе окна измерения для секвенированных оснований A цепи Уотсона. Частота истинно-положительных результатов представлена по оси ординат, а частота ложно-положительных - по оси абсцисс. Фигура демонстрирует, что значение AUC для различения секвенированных сайтов A с метилированием и без метилирования с использованием модели CNC составляло 0,94 и 0,93 для обучающего и тестового наборов данных, которые состояли из результатов секвенирования цепи Уотсона, соответственно. Это предполагает то, что было возможно использовать раскрытие изобретения в данном документе для определения состояний метилирования в сайтах A с использованием данных цепи Уотсона. Если бы мы использовали определенную вероятность метилирования с значением 0,5 в качестве порога, можно было бы достичь 99,3% специфичности и 82,6% чувствительности для обнаружения 6mA. Фиг. 37 демонстрирует, что модель CNC на основе окна измерения может использоваться для обнаружения 6mA с высокой специфичностью и чувствительностью. Точность модели можно сравнить с методом, использующим только показатель МИП.

Фиг. 38 демонстрирует сравнение эффективности между обнаружением 6mA на основе показателя МИП и обнаружением 6mA на основе окна измерения. Чувствительность отложена по оси ординат, а специфичность отложена по оси абсцисс. Фиг. 38 демонстрирует, что эффективность с использованием классификации 6mA на основе окна измерения согласно раскрытию изобретения в данном документе (AUC: 0,94) превосходит этот традиционный способ, использующий только показатель МИП (AUC: 0,87) (значение $P<0,0001$; тест Делонга). Модель CNC на основе окна измерения превзошла обнаружение на основе показателя МИП.

Фиг. 39A и 39B демонстрируют выясненную вероятность метилирования для тех секвенированных оснований A цепи Крика между наборами данных uA и mA с использованием модели CNC на основе окна измерения. Фиг. 39A демонстрирует обучающий набор данных, а фиг. 39B демонстрирует тестовый набор данных. На обоих фигурах по оси ординат отложена вероятность метилирования. Фиг. 39A и 39B

демонстрируют то, что на основе модели CHC, для данных, связанных с цепью Крика, секвенированные основания A в молекулах ДНК-матрицы из набора данных mA служили источником гораздо более высокой вероятности метилирования как в обучающем, так и в тестовом наборах данных по сравнению с основаниями A, присутствующими в наборе данных uA (значение P<0,0001; U-критерий Манна-Уитни).

Фиг. 40 демонстрирует эффективность определения 6mA с использованием модели CHC на основе окна измерения на секвенированных основаниях A цепи Крика. Частота истинно-положительных результатов представлена на оси ординат. Частота ложно-положительных результатов представлена на оси абсцисс. Фиг. 40 демонстрирует, что значение AUC при различении секвенированных сайтов A с метилированием и без метилирования с использованием модели CHC составляло 0,95 и 0,94 для обучающего и тестового наборов данных, которые состояли из результатов секвенирования цепи Крика, соответственно. Также было показано, что эффективность с использованием подхода CHC, раскрытого в данном документе (AUC: 0,94), превосходит таковую способа с использованием только показателя МИП (0,87) (значение P<0,0001). Результаты свидетельствуют о возможности использования раскрытого в данном документе изобретения для определения состояний метилирования в сайтах A с использованием данных цепи Крика. Если бы мы использовали определенную вероятность метилирования с значением 0,5 в качестве порога, можно было бы достичь 99,3% специфичности и 83,0% чувствительности для обнаружения 6mA. Фиг. 40 демонстрирует, что модель CHC на основе окна измерения может использоваться для обнаружения 6mA с высокой специфичностью и чувствительностью.

Фиг. 41 демонстрирует примеры состояний метилирования для оснований A в молекуле, включающей в себя цепи Уотсона и Крика. Белые точки представляют неметилированные аденины. Черные точки представляют метилированные аденины. Горизонтальные линии с точками представляют собой цепь двухцепочечной молекулы ДНК. Молекула 1 демонстрирует, что обе - цепи Уотсона и Крика, как определено, являются неметилированными по основаниям A. Молекула 2 демонстрирует, что цепь Уотсона была почти полностью неметилированной, тогда как цепь Крика была почти полностью метилированной. Молекула 3 демонстрирует, что обе - цепи Уотсона и Крика, как было установлено, почти полностью метилированы по основаниям A.

2. Расширенное обучение с использованием выборочного набора данных.

Как показано на фиг. 36A, 36B, 39A и 39B, наблюдали бимодальное распределение вероятности метилирования по секвенированными основаниями A в молекулах ДНК-матриц в наборе данных mA. Другими словами, в наборе данных uA существовали некоторые молекулы с сигналами uA. Это дополнительно подтверждается наличием полностью неметилированных молекул и полуметилированных молекул в наборе данных mA (фиг. 41). Одна из возможных причин может заключаться в том, что молекулы с uA в ДНК-матрицах по-прежнему будут составлять значительную часть набора данных mA после амплификации всего генома, поскольку молекулы с 6mA приведут к снижению эффективности амплификации ДНК на стадии амплификации всего генома. Это объяснение было подтверждено тем фактом, что 1 нг геномной ДНК, амплифицированный с 6mA, в результате даст только 10 нг продуктов ДНК, тогда как 1 нг геномной ДНК, амплифицированный с неметилированным A, приведет к образованию 100 нг продуктов ДНК при тех же условиях амплификации. Следовательно, для набора данных mA, исходные молекулы ДНК-матриц, аденины которых обычно неметилированы (например, 0,051%) (Xiao CL et al. Mol Cell. 2018;71:306-318) будут составлять примерно 10% от общего количества аденинов.

В одном варианте осуществления, при попытке обучить модель CHC различать mA и uA, можно выборочно использовать те основания A с относительно более высокими значениями МИП в наборе данных mA так, чтобы уменьшить влияние данных uA на обучение модели для обнаружения mA. Могут использоваться только основания A со значениями МИП выше определенного порогового значения. Значение порога может соответствовать процентилю. В одном варианте осуществления, можно использовать эти основания A в наборе данных mA со значениями МИП, превышающими значение в 10-м процентиле. В некоторых вариантах осуществления, можно использовать те A со значениями МИП, превышающими значение в 1-м, 5-м, 15-м, 20-м, 30-м, 40-м, 50-м, 60-м, 70-м, 80-м, 90-м или 95-м процентилях. Процентиль может быть основан на данных по всем молекулам нуклеиновых кислот в эталонном образце или множестве эталонных образцов.

Фиг. 42 демонстрирует эффективность с улучшенным обучением за счет выборочного использования оснований A в наборе данных mA со значениями МИП, превышающими его 10-й процентиль. Фиг. 42 демонстрирует частоту истинно-положительных результатов по оси ординат и частоту ложно-положительных результатов по оси абсцисс. Фигура демонстрирует, что при использовании оснований A в наборе данных mA со значениями МИП, превышающими 10-й процентиль для обучения модели CHC, AUC при различении оснований mA и uA увеличится до 0,98, что превосходит модель (AUC: 0,94), обученную на данных без отбора на основании значений МИП до обучения. Было высказано предположение, что выбор сайтов mA с использованием значений МИП для создания обучающего набора данных поможет улучшить различительную способность.

Чтобы дополнительно подтвердить существование молекул с основаниями uA в наборе данных mA, мы предположили, что процентное содержание uA в наборе данных mA будет обогащено в ячейках с большим количеством субпрочтений, поскольку 6mA, присутствующие в молекуле, будут замедлять

элонгацию полимеразы при создании новой цепи, по сравнению с молекулой без 6mA.

Фиг. 43 демонстрирует график процентного содержания неметилированных аденинов в наборе данных mA в зависимости от количества субпрочтений (subread) в каждой ячейке. Ось ординат показывает процент uA в наборе данных mA. По оси абсцисс показано количество субпрочтений в каждой ячейке. Тестовый набор данных был повторно проанализирован с использованием расширенной модели, которая была обучена с использованием сайтов mA после удаления сайтов A, значения МИП которых были ниже 10-го процентиля. Постепенное увеличение uA (т.е. рост с 14,6 до 55,05%) наблюдали по мере увеличения количества субпрочтений на одну ячейку, включая от 1 до 10 субпрочтений на одну ячейку секвенирования, от 10 до 20 субпрочтений на ячейку, от 40 до 50 субпрочтений на ячейку, от 60 до 70 субпрочтений на ячейку и больше 70. Таким образом, ячейки с большим количеством субпрочтений обычно имеют низкое содержание mA. Метилирование A может замедлить прохождение реакции секвенирования. Следовательно, более вероятно, что ячейки секвенирования с большой глубиной субпрочтения будут неметилированы в отношении A. Это поведение можно использовать для обнаружения неметилированных молекул с использованием порогового значения для количества субпрочтений, связанных с молекулой, например, больше чем 70 субпрочтений могут быть идентифицированы как большая часть является неметилированной.

Фиг. 44 демонстрирует метиладениновые паттерны между цепями Уотсона и Крика двухцепочечной молекулы ДНК в тестовом наборе данных. Метилирование A является асимметричным, и поэтому поведение этих двух цепей различается. Большинство молекул были метилированы из-за включения mA, с некоторыми остаточными неметилированными A. По оси ординат показан уровень метиладенина в цепи Крика. По оси абсцисс показан уровень метиладенина в цепи Уотсона. Каждая точка представляет собой двухцепочечную молекулу. Используя расширенную модель, которая была обучена с помощью выбранных сайтов mA, двухцепочечные молекулы можно разделить на разные группы в соответствии с уровнем метилирования каждой цепи следующим образом.

(a) Для двухцепочечной молекулы ДНК уровни метиладенина в цепях Уотсона и Крика были выше чем 0,8. Такая двухцепочечная молекула была определена как полностью метилированная молекула в отношении аденинових сайтов (фиг. 44, область A). Уровень метиладенина в цепи определяли, как процент сайтов A, которые были определены как метилированные, среди общего количества сайтов A в этой цепи.

(b) Для двухцепочечной молекулы ДНК уровень метиладенина в одной цепи был больше чем 0,8, тогда как в другой цепи был меньше чем 0,2. Такая молекула была определена как полуметилированная молекула в отношении сайтов аденина (фиг. 44, области B1 и B2).

(c) Для двухцепочечной молекулы ДНК уровни метиладенина в цепях Уотсона и Крика были меньше чем 0,2. Такая двухцепочечная молекула была определена как полностью неметилированная молекула в отношении аденинових сайтов (фиг. 44, область C).

(d) Для двухцепочечной молекулы ДНК уровни метиладенина в цепях Уотсона и Крика не принадлежали группам a, b и c. Такая двухцепочечная молекула была определена как молекула с чередующимися паттернами метилирования в отношении сайтов аденина (фиг. 44, область D). Чредующиеся паттерны метилирования были определены как смесь метилированных и неметилированных аденинов, присутствующих в цепи ДНК.

В некоторых других вариантах осуществления, пороговые значения уровней метиладенина для определения неметилированной цепи могут представлять собой, но без ограничения, меньше чем 0,01, 0,05, 0,1, 0,2, 0,3, 0,4 и 0,5. Пороговые значения уровней метиладенина для определения метилированной цепи могли представлять собой, но без ограничения, больше чем 0,5, 0,6, 0,7, 0,8, 0,9, 0,95 и 0,99.

Фиг. 45 представляет собой таблицу, показывающую процент полностью неметилированных молекул, полуметилированных молекул, полностью метилированных молекул и молекул с чередующимися метиладениновыми паттернами в обучающем и тестовом наборах данных. Молекулы в тестовом наборе данных могут быть классифицированы на полностью неметилированные молекулы (7,0%) в отношении сайтов аденина, полуметилированные молекулы (9,8%), полностью метилированные молекулы (79,4%) и молекулы с чередующимися паттернами метиладенина (3,7%). Эти результаты были сопоставимы с результатами, продемонстрированными в обучающем наборе данных, для которого представлены полностью неметилированные молекулы (7,0%) по отношению к сайтам аденина, полуметилированные молекулы (10,0%), полностью метилированные молекулы (79,4%) и молекулы с чередующимися паттернами метиладенина (3,6%).

Фиг. 46 иллюстрирует показательные примеры молекул с полностью неметилированными молекулами в отношении аденинових сайтов, полуметилированных молекул, полностью метилированных молекул и молекул с чередующимися метиладениновыми паттернами.Белые точки представляют неметилированные аденины. Черные точки представляют метилированные аденины. Горизонтальные линии с точками представляют собой цепь двухцепочечной молекулы ДНК.

В вариантах осуществления, можно улучшить эффективность различения метилированных и неметилированных аденинов за счет увеличения чистоты оснований 6mA, которые использовались для обучения модели CHC. С этой целью можно увеличить продолжительность реакции амплификации ДНК

так, чтобы увеличение количества вновь продуцируемых продуктов ДНК могло ослабить эффект неметилированных аденинов, привнесенных из исходных ДНК-матриц. В других вариантах осуществления, можно вставлять биотинилированные основания во время амплификации ДНК с 6mA. Вновь продуцированные продукты ДНК с 6mA могут быть извлечены и обогащены с помощью магнитных шариков, покрытых стрептавидином.

3. Использование профилей метилирования 6 mA.

Модификация ДНК 6mA присутствует в геномах бактерий, архей, протистов и грибов. (Didier W et al. Nat Rev Micorbiol. 2009;4:183-192). Также сообщалось, что в геноме человека присутствуют 6mA, составляющие 0,051% от общего количества аденинов (Xiao CL et al. Mol Cell. 2018;71:306-318). Учитывая низкое содержание 6mA в геноме человека, в одном варианте осуществления, можно создать обучающий набор данных, отрегулировав соотношение 6mA в смеси дНТФ (N представляет немодифицированные A, C, G и T) на этапе амплификации всего генома. Например, можно использовать соотношение 6mA к дНТФ, составляющее 1:10, 1:100, 1:1000, 1:10000, 1:100000 или 1:1000000. В другом варианте осуществления, аденин-ДНК-метилтрансфераза M.EcoGII может использоваться для создания обучающего набора данных 6mA.

Количество 6mA было ниже в тканях рака желудка и печени, и это подавление 6mA коррелировало с повышенным онкогенезом (Xiao CL et al. Mol Cell. 2018;71:306-318). С другой стороны, сообщалось, что в глиобластоме присутствовали более высокие уровни 6mA (Xie et al. Cell. 2018;175:1228-1243). Таким образом, раскрытый в данном документе подход для 6mA может быть полезен для изучения геномики рака (Xiao CL et al. Mol Cell. 2018;71:306-318; Xie et al. Cell. 2018;175:1228-1243). Кроме того, было обнаружено, что 6mA более распространен и широко представлен в митохондриальной ДНК млекопитающих, что связано с гипоксией (Hao Z et al. Mol Cell. 2020; doi:10.1016/j.molcel.2020.02.018). Таким образом, подход к обнаружению 6mA в этом раскрытии изобретения будет полезен для изучения митохондриальной стрессовой реакции в различных клинических условиях, таких как беременность, рак и аутоиммунные заболевания.

IV. Результаты и применения.

A. Обнаружение метилирования.

Обнаружение метилирования в сайтах CpG с использованием описанных выше способов было выполнено для различных биологических образцов и областей генома. В качестве примера, определение метилирования с помощью внеклеточной ДНК в плазме беременных женщин с использованием секвенирования отдельной молекулы в реальном времени было проверено в сравнении с определением метилирования с использованием бисульфитного секвенирования. Результаты метилирования могут использоваться для различных применений, включая определение числа копий и диагностику нарушений. Описанные ниже способы не ограничиваются сайтами CpG и также могут применяться к любой модификации, описанной в данном документе.

1. Обнаружение метилирования длинных молекул ДНК в ткани плаценты.

Секвенирование отдельной молекулы в реальном времени может секвенировать молекулы ДНК длиной в несколько тысяч пар оснований (Nattestad et al., 2018). Расшифровка состояний метилирования для сайтов CpG с использованием описанного в данном документе изобретения позволит получить информацию по гаплотипу состояний метилирования путем синергетического использования информации из долгих прочтений секвенирования отдельной молекулы в реальном времени. Чтобы продемонстрировать возможность получения вывода о состояниях метилирования длинных прочтений, а также информации по его гаплотипу, мы секвенировали ДНК ткани плаценты при наличии 478739 молекул, которые охватывались 28913838 субпрочтениями. Было 7 молекул размером более 5 т.п.о. В среднем каждая из них было охватывалась 3 субпрочтениями.

Фиг. 47 демонстрирует состояния метилирования вдоль длинной молекулы ДНК размером 6265 п.о. (т.е. гаплотипного блока), которая была секвенирована в ZMW с номером ячейки ZMW m54276_180626_162240/40763503 и картирована в геномном локусе chr1:113246546-113252811 в человеческом геноме. "-" представляет не-CpG нуклеотид; "U" представляет неметилированное состояние в сайте CpG; и "M" представляет метилированное состояние в сайте CpG. Область 4710, выделенная желтым цветом, указывает на область островка CpG, который, как известно, в целом неметилирован (фиг. 47). Было установлено, что большинство сайтов CpG в этом островке CpG неметилированы (96%). Напротив, было сделано вывод, что 75% сайтов CpG за пределами островка CpG неметилированы. Эти результаты свидетельствуют о том, что уровень метилирования за пределами островка CpG (например, "берег/шельф" островка CpG) был выше, чем у островка CpG. Смесь метилированных и неметилированных состояний в компоновке гаплотипов в областях за пределами этого CpG-островка могла указывать на вариабельность паттернов метилирования. Такие наблюдения в целом соответствовали нынешнему пониманию (Zhang et al., 2015; Feinberg and Irizarry, 2010). Таким образом, это изобретение сделало возможным получение сигнала различных состояний метилирования вдоль длинной молекулы, включая состояния метилирования и неметилирования, что подразумевает то, что информация о гаплотипах состояний метилирования может быть оценена. Информация о гаплотипах относится к привязке состояний метилирования сайтов CpG к непрерывному участку ДНК.

В одном варианте осуществления, мы могли бы в данном документе использовать этот подход для анализа состояний метилирования вдоль гаплотипа для обнаружения и анализа областей импринтинга. Области импринтинга подвергаются эпигенетической регуляции, которая обуславливает состояния метилирования по типу родительского происхождения. Например, одна важная, облас^{ть} импринтинга расположена на хромосоме 11p15.5 человека и содержит подверженные импринтингу гены IGF2, H19 и CDKN1C ($P57^{kip2}$), которые являются сильными регуляторами роста плода (Brioude et al., Nat Rev Endocrinol. 2018;14:229-249). Генетические и эпигенетические аберрации в областях импринтинга могут быть связаны с болезнями. Синдром Беквита-Видемана (BWS) - это синдром чрезмерного роста, при котором пациенты часто проявляют макроглоссию, дефекты брюшной стенки, гемигиперплазию, увеличенные органы брюшной полости и повышенный риск эмбриональных опухолей в раннем детстве. Считается, что BWS вызывается генетическими или эпигенетическими дефектами в областях 11p15.5 (Brioude et al., Nat Rev Endocrinol. 2018;14:229-249). Область, называемая ICR1 (область контроля импринтинга 1), которая расположена между H19 и IGF2, выборочно метилирована по отцовскому аллелю. ICR1 управляет экспрессией IGF2, зависящей от родительского источника. Таким образом, генетические и эпигенетические аберрации в ICR1 могут привести к аберрантной экспрессии IGF2, что является одной из возможных причин, вызывающих BWS. Таким образом, обнаружение состояний метилирования вдоль областей импринтинга может иметь клиническое значение.

Мы загрузили данные для 92 подверженных импринтингу генов из общедоступной базы данных, в которой собраны данные о подверженных импринтингу генах (http://www.geneimprint.org/). Области 5 т.п.о. выше и ниже этих подверженных импринтингу генов использовали для дополнительного анализа. Среди этих областей 160 CpG-островков связаны с этими подверженными импринтингу генами. Мы получили 324248 кольцевых консенсусных последовательностей из образца плаценты. После удаления кольцевых консенсусных последовательностей с низким качеством и короткими перекрывающимися областями с островками CpG (например, меньше 50% длины соответствующего CpG-островка), мы получили 9 кольцевых консенсусных последовательностей, перекрывающихся с 9 островками CpG, которые соответствуют 8 подверженным импринтингу генам.

Фиг. 48 представляет собой таблицу, показывающую, что 9 молекул ДНК были секвенированы с помощью секвенирования отдельной молекулы в реальном времени и перекрываются с областями импринтинга, включающими в себя H19, WT1-AS, WT1, DLK1, MEG3, ATP10A, LRRTM1, и MAGI2. Шестой столбец содержал участки ДНК, перекрывающиеся с островками CpG, включая области импринтинга. "U" представляет неметилированный цитозин в контексте CpG; "M" представляет метилированный цитозин в контексте CpG. "*" представляет сайт CpG, который не был охвачен результатом секвенирования; "-" представляет нуклеотид из не-CpG сайтов; генотип указывается в скобках, если молекула частично совпадает с однонуклеотидным полиморфизмом (ОНП). В 7-м столбце указаны состояния метилирования для всей молекулы. Молекула может быть названа метилированной, если было показано, что большинство сайтов CpG (например, больше чем 50%) метилированы согласно вариантам осуществления, представленным в данном изобретении; в противном случае она была бы названа неметилированной.

Среди 9 молекул ДНК 5 молекул ДНК (55,6%) были названы метилированными, что незначительно отклонялось от ожидаемого, согласно которому 50% молекул ДНК будут метилированными. Как показано в 6-м столбце таблицы на фиг. 48, было показано, что большинство сайтов CpG метилируются или неметилируются согласованным образом, то есть как гаплотип метилирования. Один вариант осуществления заключается в том, что молекула будет называться метилированной, если будет показано, что большинство сайтов CpG (например, больше чем 50%) метилировано согласно вариантам осуществления, представленным в данном изобретении, в противном случае она будет называться неметилированной. Могут быть использованы другие пороговые значения для определения того, является ли молекула метилированной или нет, например, но не ограничивающиеся лишь этими: по меньшей мере 10, 20, 30, 40, 50, 60, 70, 80, 90 и 100% сайтов CpG в молекуле, которые были проанализированы, признаются метилированными.

В другом варианте осуществления, мы могли бы использовать молекулы, одновременно содержащие по меньшей мере один ОНП и по меньшей мере одну оценку сайта CpG, чтобы определить, может ли область быть связана с областью импринтинга, или может ли известный подверженный импринтингу ген быть аберрантным (например, потеря импринтинга). В целях иллюстрации фиг. 49 демонстрирует, что первая молекула из области импринтинга несла аллель "A"; и вторая молекула из этой области импринтинга несла аллель "G". Предполагая, что область импринтинга была подвергнута отцовскому импринтингу, первая молекула материнского гаплотипа была полностью неметилирована; а вторая молекула отцовского гаплотипа была полностью метилирована. В одном варианте осуществления, такое предположение обеспечило бы данные о состояниях метилирования, делая возможным исследование эффективности обнаружения модификации основания согласно вариантам осуществления, представленным в данном раскрытии изобретения.

Фиг. 49 демонстрирует пример определения паттернов метилирования в области импринтинга. ДНК из биологического образца была извлечена и лигирована с адаптерами в виде шпилек для формиро-

вания кольцевых молекул ДНК. Информация о последовательностях и модификации оснований (например, состояниях метилирования в сайтах CpG) в отношении этих кольцевых молекул ДНК были неизвестны. Эти кольцевые молекулы ДНК были подвергнуты секвенированию отдельной молекулы в реальном времени. МИП, ШИ и контекст последовательностей для оснований в каждом суброчтении, происходящем из этих кольцевых молекул ДНК, были определены после того, как субпрочтения были сопоставлены с эталонным геномом. Кроме того, были определены генотипы этих молекул. МИП, ШИ и контекст последовательности в окне измерения, связанном с сайтами CG, будут сравниваться с эталонными кинетическими паттернами согласно вариантам осуществления, представленным в данном раскрытии изобретения, для определения состояний метилирования для каждого CpG. Если две молекулы с разными аллелями демонстрируют разные паттерны метилирования таким образом, что одна полностью неметилирована, а другая полностью метилирована, геномная область, связанная с этими двумя молекулами, будет областью импринтинга. В одном варианте осуществления, если такая геномная область оказалась известной областью импринтинга, например, как показано на фиг. 49, паттерны метилирования для этих двух молекул соответствовали ожидаемым паттернам метилирования (т.е. наблюдаемое) в нормальной ситуации. Это может свидетельствовать о точности способов классификации состояний метилирования согласно вариантам осуществления, представленным в данном изобретении. В одном варианте осуществления, расхождение между определенными паттернами метилирования согласно вариантам осуществления, представленным в данном изобретении, и ожидаемыми паттернами метилирования будет указывать на аберрации импринтинга, например, потерю импринтинга.

Фиг. 50 демонстрирует пример определения паттернов метилирования в области импринтинга. В одном варианте осуществления, паттерн импринтинга может быть дополнительно определен посредством анализа паттернов метилирования данной области в пределах определенного родословного дерева. Например, может быть выполнен анализ паттернов метилирования и аллельной информации в отцовском, материнском геномах и потомстве. Такое древо родословной может дополнительно включать в себя геном деда по отцовской или материнской линии, бабушки по отцовской или материнской линии, или другие соответствующие геномы. В другом варианте осуществления, такой анализ может быть расширен до наборов данных трех членов семьи (мать, отец и ребенок) в определенной популяции, например, получая информацию о метилировании и генотипе для каждого индивида согласно вариантам осуществления, представленным в данном документе.

Как показано после классификации, можно определить как генотип (аллель в рамке), так и статус метилирования. Для каждой из молекул, может быть предоставлен паттерн метилирования в каждом сайте (например, все метилированные или все неметилированные), чтобы идентифицировать от какого из родителей унаследована молекула. Или может быть определена плотность метилирования, и одно или большее количество пороговых значений могут классифицировать, является ли молекула гиперметилированной (например, >80% или другой % и от одного родителя) или гипометилированной (например, <20% или другой % и от другого родителя).

2. Обнаружение метилирования молекул вкДНК.

В качестве другого примера, метилирование внеклеточной ДНК (вкДНК) также все чаще признается как важный молекулярный признак для неинвазивного пренатального тестирования. Например, мы показали, что молекулы вкДНК из областей, несущих тканеспецифическое метилирование, могут использоваться для определения пропорционального вклада из различных тканей, например, нейтрофилов, Т-лимфоцитов, В-лимфоцитов, печени, плаценты, в плазму беременных женщин (Sun et al., 2015). Также была продемонстрирована возможность использования метилирования ДНК плазмы беременных женщин для обнаружения трисомии 21 (Lun et al., 2013). Молекулы вкДНК в материнской плазме были фрагментированы со средним размером 166 п.о., что намного короче, чем искусственно фрагментированная ДНК E. coli с размером примерно 500 п.о. Сообщалось, что вкДНК фрагментирована неслучайно, например, концевые мотивы ДНК в плазме связаны с тканевым происхождением, например, происходят из плаценты. Такие характерные свойства внеклеточной ДНК дают совершенно иной контекст последовательности по сравнению с искусственно фрагментированной ДНК E. coli. Таким образом, остается неизвестным, позволит ли такая кинетика полимеразы количественно определить уровни метилирования, типичные для внеклеточных молекул ДНК. Публикуемые в данной патентной заявке сведения могут применятся в, но не ограничиваются лишь этим: анализе метилирования внеклеточной ДНК в плазме беременных женщин, например, с использованием модели предсказания метилирования, обученной на основе молекул ДНК из упомянутых выше тканей.

Используя секвенирование отдельной молекулы в реальном времени было секвенировано шесть образцов ДНК из плазмы беременных женщин с плодом мужского пола с медианным значением, составляющим 30738399 субпрочтений (диапазон: 1431215-105835846), что соответствует медианному значению, составляющему 111834 ККП (диапазон: 61010-503582). Каждое ДНК плазмы секвенировали такое количество раз, которое имело медианное значение, составляющее 262 раза (диапазон: 173-320). Набор данных был получен из ДНК, приготовленной с помощью Sequel I Sequencing Kit 3.0.

Чтобы оценить обнаружение метилирования молекул вкДНК, мы использовали бисульфитное секвенирование (Jiang et al., 2014) для анализа метилирования вышеупомянутых 6 образцов ДНК плазмы

беременных женщин. Мы получили в среднем 66 миллионов прочтений с одинаковыми концами (58-82 миллиона прочтений с одинаковыми концами). Медиана совокупного метилирования составила 69,6% (67,1-72,0%).

Фиг. 51 демонстрирует сравнение уровней метилирования, полученных с помощью нового подхода и обычного бисульфитного секвенирования. По оси ординат отложены уровни метилирования, предсказанные в соответствии с вариантами осуществления, представленными в данной патентной заявке. По оси абсцисс отложены уровни метилирования, полученные с помощью бисульфитного секвенирования. Была проанализирована медиана 314675 сайтов CpG (диапазон: 144 546-1 382 568) для результатов из ДНК плазмы, полученных с помощью секвенирования отдельной молекулы в реальном времени. Медианная доля CpG-сайтов, которые, как предполагалось, будут метилированы, составила 64,7% (диапазон: 60,8-68,5%), что оказалось сопоставимым с результатами, полученными при бисульфитном секвенировании. Как показано на фиг. 51, наблюдалась высокая корреляция (r: 0,96, p-значение=0,0023) между совокупными уровнями метилирования, рассчитанными с помощью секвенирования отдельной молекулы в реальном времени и представленного подхода предопределения метилирования, и бисульфитного секвенирования.

Из-за малой глубины бисульфитного секвенирования оно может оказаться ненадежным для определения уровней метилирования (т.е. доли секвенированных CpG, которые были метелированы) для каждого CpG в геноме человека. Вместо этого мы рассчитали уровни метилирования в некоторых областях с множеством сайтов CpG путем агрегирования сигналов прочтений, покрывающих сайты CpG в геномной области, в которой любые два последовательных сайта CpG находятся в пределах 50 нуклеотидов, а количество сайтов CpG составляет по меньшей мере 10. Процент секвенированных цитозинов среди суммы секвенированных цитозинов и тиминов в сайтах CpG в области указывает на уровни метилирования этой области. Области были разделены на разные группы в соответствии с уровнями метилирования областей. Вероятность метилирования, предсказанная с помощью модели, обученной на предыдущих обучающих наборах данных (т.е. тканевой ДНК), соответственно повышалась по мере увеличения уровней метилирования (фиг. 52A). Эти результаты также подтверждают возможность и обоснованность использования секвенирования отдельной молекулы в реальном времени для прогнозирования состояний метилирования молекул вкДНК у беременных женщин. Фиг. 52B продемонстрировала, что уровень метилирования в геномном окне размером 10 млн.п.о., оцененный с использованием секвенирования отдельной молекулы в реальном времени в соответствии с вариантами осуществления, представленными в данном изобретении, был хорошо скорректирован с помощью бисульфитного секвенирования (r=0,74; p-значение <0,0001).

Фиг. 53 продемонстрировала, что геномные представления (GR) Y-хромосомы в материнской плазме беременных женщин, определенные с помощью секвенирования отдельной молекулы в реальном времени, хорошо коррелировали с теми, что были определены с помощью БС-секв. (r=0,97; P-значение =0,007). Эти результаты свидетельствуют о том, что секвенирование отдельной молекулы в реальном времени также позволило точно определить количество молекул ДНК, происходящих из негематопоэтических тканей, таких как плацента, чья привнесенная ДНК обычно составляла меньшую часть. Другими словами, это изобретение продемонстрировало возможность одновременного анализа аберраций числа копий и состояний метилирования для нативных молекул без каких-либо преобразований и амплификаций оснований до секвенирования.

3. Способ на основе блок CpG.

В некоторых вариантах осуществления, можно выполнить анализ метилирования ряда геномных областей, содержащих множество сайтов CpG, например, но без ограничения, 2, 3, 4, 5, 10, 20, 30, 40, 50, 100 сайтов CpG и т.д. Размер такой геномной области может составлять, например, но без ограничения, 50, 100, 200, 300 и 500 нуклеотидов и т.д. Расстояние между сайтами CpG в этой области может составлять, например, но без ограничения, 10, 20, 30, 40, 50, 100, 200, 300 нуклеотидов и т.д. В одном варианте осуществления, мы могли бы объединить любые два последовательных сайта CpG в пределах 50 нуклеотидов, чтобы сформировать блок CpG так, чтобы количество сайтов CpG в этом блоке было больше чем 10. В таком способе на основе блоков, множество областей могут быть объединены в одно окно, представленное в виде единой матрицы, эффективно обрабатывая области вместе.

В качестве примера, как показано на фиг. 54, для анализа метилирования использовали кинетику всех субпрочтений, связанных с блоком CpG. Предсказанные профили МИП восходящих и нисходящих фланкирующих 10 нуклеотидов по каждому CpG в этом блоке были искусственно выровнены относительно сайтов CpG для расчета среднего профиля МИП (фиг. 54). Слово "спроецированный" означает, что мы выровняли кинетические сигналы суброчтений с каждым соответствующим рассматриваемым сайтом CpG. Средние профили МИП для блока CpG использовались для обучения модели (например, с использованием искусственной нейронной сети, сокращенно ИНС) для определения состояний метилирования для каждого блока. Анализ ИНС включал в себя входной слой, два скрытых слоя и выходной слой. Каждый блок CpG характеризовался вектором характеритик из 21 значения МИП, который вводился в ИНС. Первый скрытый слой включал в себя 10 нейронов с ReLu в качестве функции активации. Второй скрытый слой включал в себя 5 нейронов с ReLu в качестве функции активации. Наконец, вы-

ходной слой включал в себя 1 нейрон с сигмоидной функцией в качестве функции активации, которая выдавала вероятность метилирования. Сайт CpG, показывающий вероятность метилирования >0,5, считали метилированным, в противном случае он считался неметилированным. Средний профиль МИП можно использовать для анализа состояния метилирования целой молекулы. Вся молекула может считаться метилированной, если определенное количество сайтов выше порогового значения (например, 0, 1, 2, 3 и т.д.) метилировано или если молекула имеет определенную плотность метилирования.

В неметилированной и метилированной библиотеках было 9678 и 9020 блоков CpG, каждый из которых содержал по меньшей мере 10 сайтов CpG. Эти блоки CpG охватывают 176048 и 162943 сайта CpG для неметилированной и метилированной библиотек. Как показано на фиг. 55A и 55B, мы могли бы достичь больше чем 90% общей точности в прогнозировании состояний метилирования как в обучающем, так и тестовом наборе данных. Однако, такой вариант осуществления, основанный на блоках CpG, значительно снизил бы количество CpG, которые можно было бы оценить. По определению, требование в виде наименьшего количества сайтов CpG ограничит анализ метилирования до некоторых конкретных областей генома (например, предпочтительно до анализа островков CpG).

B. Определение источника или нарушения.

Профили метилирования могут использоваться для определения ткани-источника или определения классификации заболевания. Анализ профиля метилирования может использоваться в сочетании с другими клиническими данными, включая изображения, стандартные анализы крови и другую медицинскую диагностическую информацию. Профили метилирования могут быть определены с использованием любого способа, описанного в данном документе.

1. Определение аберрации числа копий.

В этом разделе показано, что SMRT является точным для определения числа копий, и, таким образом, профиль метилирования и профиль числа копий можно анализировать одновременно.

Было показано, что аберрации числа копий могут быть обнаружены путем секвенирования опухолевых тканей (Chan (2013)). В данном документе мы демонстрируем, что связанные с раком аберрации числа копий могут быть идентифицированы путем секвенирования опухолевых тканей с использованием секвенирования отдельной молекулы в реальном времени. Например, для случая TBR3033 мы получили 589435 и 1495225 консенсусных последовательностей (минимальное требование по субпрочтениям, используемым для построения каждой консенсусной последовательности, составляло 5) для опухолевой ДНК и ДНК прилегающей неопухолевой ткани печени в паре с ней, соответственно. Набор данных был сгенерирован из ДНК, полученной с помощью Sequel II Sequencing Kit 1.0. В одном варианте осуществления, геном был разделен, in silico, на окна размером 2 млн.п.о. Был рассчитан процент сопоставления консенсусных последовательностей с каждым окном, что дало геномное представление (GR) с разрешением 2 млн.п.о. GR можно определить по количеству субпрочтений в позиции, нормализованному по совокупному количеству субпрочтений последовательности по всему геному.

Фиг. 56A демонстрирует соотношение GR для ДНК опухоли и ДНК прилегающей неопухолевой ткани в паре с ней, с использованием секвенирования отдельной молекулы в реальном времени. Соотношение числа копий между опухолевой ДНК и ДНК нормальной ткани в паре с ней показано на оси ординат, индекс геномного группирования для каждого окна размером 2 млн.п.о., включая хромосомы с 1 по 22, показан на оси абсцисс. Для этой фигуры, область, имеющая соотношение GR выше 95-го процентиля всех окон размером 2 млн.п.о, была классифицирована как имеющая прирост числа копий, тогда как область, имеющая соотношение GR ниже 5-го процентиля всех окон размером 2 млн.п.о, была классифицирована как имеющая потерю числа копии. Мы заметили, что на хромосоме 13 наблюдается потеря числа копий, а на хромосоме 20 - прирост числа копий. Такие приросты и потери - правильный результат.

Фиг. 56B демонстрирует соотношение GR для опухоли и прилегающей неопухолевой ткани в паре с ней с использованием бисульфитного секвенирования. Соотношение числа копий между опухолевой ДНК и ДНК нормальной ткани в паре с ней показано на оси ординат, а индекс геномного группирования для каждого окна размером 2 млн.п.о., включая хромосомы с 1 по 22, показан на оси абсцисс. Изменения числа копий, идентифицированные с помощью секвенирования отдельной молекулы в реальном времени на фиг. 56A были подтверждены результатами соответствующего бисульфитного секвенирования на фиг. 56B.

Для случая TBR3032 мы получили 413982 и 2396054 консенсусных последовательностей (минимальное требование по субпрочтениям, используемым для построения каждой консенсусной последовательности, составляло 5) для опухолевой ДНК и ДНК прилегающей неопухолевой ткани в паре с ней, соответственно. В одном варианте осуществления, геном был разделен, in silico, на окна размером 2 млн.п.о. Был рассчитан процент сопоставления консенсусных последовательностей с каждым окном, а именно геномное представление (GR) с разрешением 2 млн.п.о.

Фиг. 57A демонстрирует соотношение GR для ДНК опухоли и ДНК прилегающей неопухолевой ткани в паре с ней, с использованием секвенирования отдельной молекулы в реальном времени. Соотношение числа копий между опухолевой ДНК и ДНК нормальной ткани в паре с ней показано на оси ординат, а индекс геномного группирования для каждого окна размером 2 млн.п.о., включая хромосомы с 1

по 22, показан на оси абсцисс. Для этой фигуры, область, имеющая соотношение GR выше 95-го процентиля всех окон размером 2 млн.п.о, была классифицирована как имеющая прирост числа копий, тогда как область, имеющая соотношение GR ниже 5-го процентиля всех окон размером 2 млн.п.о, была классифицирована как имеющая потерю числа копии. Мы заметили, что на хромосомах 4, 6, 11, 13, 16 и 17 наблюдается потеря числа копий, в то время как на хромосомах 5 и 7 - прирост числа копий.

Фиг. 57B демонстрирует соотношение GR для ДНК опухоли и ДНК прилегающей неопухолевой ткани в паре с ней, с использованием бисульфитного секвенирования. Соотношение числа копий между ДНК опухоли и ДНК нормальной ткани в паре с ней показано на оси ординат, а индекс геномного группирования для каждого окна размером 2 млн.п.о., включая хромосомы с 1 по 22, показан на оси абсцисс. Изменения числа копий, идентифицированные с помощью секвенирования отдельной молекулы в реальном времени на фиг. 57A были подтверждены результатами соответствующего бисульфитного секвенирования на фиг. 57B.

Соответственно, профиль метилирования и профиль числа копий можно анализировать одновременно. В этом примере, поскольку чистота опухоли опухолевой ткани как правило не всегда составляет 100%, амплифицированные области будут относительно увеличивать вклад опухолевой ДНК, в то время как удаленные области будут относительно уменьшать вклад опухолевой ДНК. Поскольку геном опухоли характеризуется глобальным гипометилированием, амплифицированные области будут дополнительно снижать уровни метилирования по сравнению с удаленными областями. В качестве иллюстрации, для случая TBR3033 уровень метилирования хромосомы 22 (прирост числа копий), измеренный с использованием данного изобретения, составил 48,2%, что ниже, чем у хромосомы 3 (потеря числа копий) (уровень метилирования: 54,0%). Для случая TBR3032 уровень метилирования плеча хромосомы 5p (прирост числа копий), измеренный с использованием данного изобретения, составил 46,5%, что ниже, чем у плеча хромосомы 5q (потеря числа копий) (уровень метилирования: 54,9%).

2. Картирование ДНК из плазмы по тканям у беременной женщины.

Как показано на фиг. 58, мы пришли к выводу, что точность анализа метилирования позволит нам сравнить профили метилирования ДНК из плазмы беременной женщины с профилями метилирования различных референсных тканей (например, печени, нейтрофилов, лимфоцитов, плаценты, Т-лимфоцитов, В-лимфоцитов, сердца, мозга и т.д.). Таким образом, вклад ДНК в пул ДНК из плазмы беременной женщины из разных типов клеток можно было определить с помощью следующих процедур. Уровни метилирования CpG смеси ДНК (например, ДНК плазмы), определенные в соответствии с вариантами осуществления, представленными в данном раскрытии изобретения, были записаны в векторе (X), и полученные эталонные уровни метилирования в различных тканях были записаны в матрице (M), которые можна было количественно определить бисульфитным секвенированием, но не ограничиваясь лишь им. Пропорциональные вклады (p) из разных тканей в смесь ДНК могут быть рассчитаны с помощью квадратичного программирования, но не ограничиваются лишь им. В данном документе мы используем математические уравнения, чтобы проиллюстрировать вычисление пропорционального вклада различных органов в анализируемую смесь ДНК. Математическая взаимосвязь между плотностями метилирования различных сайтов в смеси ДНК и плотностями метилирования соответствующих сайтов в разных тканях может быть выражена как:

$$\bar{X}_i = \sum_k (p_k \times M_{ik}),$$

где $\bar{X}_i$ представляет плотность метилирования сайта CpG i в смеси ДНК; p_k представляет пропорциональный вклад типа клеток k в смесь ДНК; M_{ik} представляет плотность метилирования сайта CpG i в типе клеток k. Когда количество сайтов такое же или больше, чем количество органов, могут быть определены значения отдельного p_k.

Для повышения информативности исключали сайты CpG, которые продемонстрировали небольшую вариабельность уровней метилирования среди всех типов референсных тканей. В одном варианте осуществления, мы использовали конкретный набор сайтов CpG для выполнения анализа. Например, эти сайты CpG характеризовались большим чем 30% коэффициентом вариации (CV) уровней метилирования среди разных тканей, и большей чем 25% разницей между максимальным и минимальным уровнями метилирования среди тканей. В некоторых других вариантах осуществления, может быть использован CV, составляющий 5, 10, 20, 30, 40, 50, 60, 80, 90, 100, 110, 200, 300% и т.д.; и может быть использована разница между максимальным и минимальным уровнями метилирования среди тканей большая чем 5, 10, 15, 20, 25, 30, 40, 50, 60, 70, 80, 90, 100% и т.д.

В алгоритм могут быть включены дополнительные критерии для повышения точности. Например, совокупный вклад всех типов клеток будет ограничен 100%, т.е.:

$$\sum_k p_k = 100\%.$$

Кроме того, необходимо, чтобы вклад всех органов был неотрицательным:

$$p_k \geq 0, \forall k$$

Из-за биологических вариаций наблюдаемый совокупный паттерн метилирования может не полностью совпадать с паттерном метилирования, установленным по метилированию тканей. В таких обстоятельствах потребуется математический анализ для определения наиболее вероятного пропорционального

вклада отдельных тканей. В этом отношении, разница между наблюдаемым паттерном метилирования в ДНК и предполагаемым паттерном метилирования из тканей обозначается W:

$$W = \bar{X}_i - \sum_k (p_k \times M_{ik})$$

Наиболее вероятное значение каждого $p_к$ может быть определено путем минимизации W, что представляет собой разницу между наблюдаемыми и предполагаемыми паттернами метилирования. Это уравнение может быть решено с использованием математических алгоритмов, например, но не ограничиваясь лишь этими: с использованием квадратичного программирования, линейной/нелинейной регрессии, алгоритма максимизации ожидания (EM), алгоритма максимального правдоподобия, максимальной апостериорной оценки и способа наименьших квадратов.

Как показано на фиг. 59, мы наблюдали, что вклад плацентарной ДНК в материнскую плазму беременных женщин, несущих плод мужского пола, с использованием метода картирования ДНК из плазмы по тканям, представленного на фиг. 58, хорошо коррелировал с фракциями ДНК плода, которые оценивали по прочтениям Y-хромосомы. Этот результат свидетельствует о возможности использования кинетических характеристик для отслеживания тканевого происхождения ДНК из плазмы у беременных женщин.

3. Количественная оценка уровня метилирования по областям.

В этом разделе описаны методы определения характерного уровня метилирования для выбранных областей генома, которые могут быть реализованы с использованием относительно низкого уровня секвенирования. Уровни метилирования могут быть определены для каждой цепи или для каждой молекулы, или для каждой области, используя количество метилированных сайтов и общее количество метилированных сайтов. Также анализируют уровни метилирования различных тканей.

Мы секвенировали 11 тканевых образцов ДНК человека с средним количеством субпрочтений - 30,7 млн. (диапазон: 9,1-88,6 млн.) для каждого образца, которые можно было выровнять с эталонным геномом человека (hg19). Субпрочтения из каждого образца были сгенерированы из в среднем 3,8 млн. ячеек Pacific Biosciences Single Molecular Real-Time (SMRT) Sequencing (диапазон: 1,1-11,5 млн.), каждая из которых содержала по меньшей мере один фрагмент, каждый из которых можно было выровнять с эталонным геномом человека. В среднем каждая молекула в ячейке SMRT была секвенирована в среднем 9,9 раза (диапазон: 6,5-13,4 раза). Тканевые образцы ДНК человека включали в себя 1 материнский образец лейкоцитарного слоя беременного субъекта, 1 образец плаценты, 2 образца опухолевых тканей гепатоцеллюлярной карциномы (ГНК), 2 образца прилегающих неопухолевых тканей в паре с 2-мя ранее упомянутыми тканями ГЦК, 4 образца лейкоцитарного слоя здоровых контрольных субъектов (M1 и M2 были из субъектов мужского пола; F1 и F2 были из субъектов женского пола), 1 образец из линии клеток ГЦК (HepG2). Подробности итогов по данным секвенирования показаны на фиг. 60.

Фиг. 60 демонстрирует различные группы тканей в первом столбце и названия образцов во втором столбце. "Совокупное количество субпрочтений" обозначает совокупное количество последовательностей, сгенерированных из ячеек SMRT, включая таковые из цепей Уотсона и Крика. "Картированные субпрочтения" приводят количество субпрочтений, которые можно выровнять с эталонным геномом человека. "Картированность субпрочтений" относится к доле субпрочтений, которые можно выровнять с эталонным геномом человека. "Средняя глубина субпрочтений на ячейку SMRT" указывает среднее количество субпрочтений, генерируемых из каждой ячейки SMRT. "Количество ячеек SMRT" относится к количеству ячеек SMRT, которые произвели детектируемые субпрочтения. Термин "картируемые ячейки" обозначает количество ячеек, содержащих по меньшей мере одно выравниваемое субпрочтение. "Процент картируемых ячеек (%)" - это процент ячеек, которые содержали по меньшей мере одно выравниваемое субпрочтение.

а) Методы анализа уровня и паттерна метилирования.

В одном варианте осуществления, можно измерить плотность метилирования одной цепи нуклеиновой кислоты (например, ДНК или РНК), которая определяется как количество метилированных оснований в цепи, разделенное на совокупное количество метилируемых оснований в пределах данной цепи. Это измерение также обозначают как "уровень метилирования одной цепи". Это измерение отдельной цепи в частности осуществимо в контексте данного раскрытия изобретения, поскольку платформа для секвенирования отдельной молекулы в реальном времени может получать информацию о секвенировании из каждой из двух цепей двухцепочечной молекулы ДНК. Этому способствует использование адаптеров в виде шпильки при подготовке библиотек секвенирования, так что цепи Уотсона и Крика двухцепочечной молекулы ДНК будут объединены в формате кольца и секвенированы вместе. Фактически, эта структура также позволяет секвенировать находящиеся в паре цепи Уотсона и Крика одной и той же двухцепочечной молекулы ДНК в одной и той же реакции, так что статус метилирования соответствующих комплементарных сайтов на цепях Уотсона и Крика любой двухцепочечной ДНК молекулы может отдельно определятся и напрямую сравниваться (например, фиг. 20A и 20B).

Эти анализы метилирования на основе цепей не могут быть легко выполнены с помощью других технологий. Поскольку без использования способа прямого анализа метилирования, раскрытого в данной

заявке, необходимо было бы применить другие средства для дифференциации метилированных оснований от неметилированных, например, с помощью преобразования бисульфитом. Преобразование бисульфитом требует обработки ДНК бисульфитом натрия, чтобы метилированные цитозины и неметилированные цитозины можно было различить как цитозины и тимины соответственно. В денатурирующих условиях многих протоколов преобразования бисульфитом две цепи двухцепочечной молекулы ДНК диссоциируют друг от друга. Во многих применениях секвенирования, например, с использованием платформы Illumina, преобразованную бисульфитом ДНК затем амплифицируют с помощью полимеразной цепной реакции (ПЦР), которая включает в себя диссоциацию двухцепочечной ДНК на отдельные цепи.

При секвенировании Illumina можно приготовить библиотеки для секвенирования без ПЦР, используя метилированные адаптеры до преобразования бисульфитом. Даже при использовании этой стратегии каждая цепь ДНК двухцепочечной молекулы ДНК будет случайным образом выбрана для мостиковой амплификации в проточной ячейке. Из-за случайного характера секвенирования маловероятно, что каждая цепь одной и той же молекулы ДНК секвенируется в одной и той же реакции. Даже если в одном и том же цикле анализируют больше чем одно субпрочтение последовательности из одного и того же локуса, не существует простых средств для определения того, происходят ли эти два прочтения из каждой из пары цепей Уотсона и Крика из одной двухцепочечной молекулы ДНК, или происходят из двух разных двухцепочечных молекул ДНК. Такие факторы важны, потому что в некоторых вариантах осуществления данного изобретения две цепи двухцепочечной молекулы ДНК могут проявлять разные паттерны метилирования. Когда определяют плотности метилирования одной цепи множества цепей нуклеиновых кислот (например, ДНК или РНК), также можно определить "уровень метилирования множества цепей" на основе концепций и уравнения, касающихся "уровня метилирования интересующей геномной области" на фиг. 61.

Фиг. 61 демонстрирует различные способы анализа паттернов метилирования. Двухцепочечная молекула ДНК (X) с неизвестной последовательностью и информацией о метилировании лигируется с адаптерами, которые в одном примере формируют структуру шпилька-петля. В результате две отдельные цепи молекулы ДНК, включая цепи Уотсона X (a) и Крика X (b), физически объединяют в пару в виде кольцевой формы в данном примере. Статусы метилирования сайтов в обеих цепях Уотсона и Крика могут быть получены с использованием способов, описанных в данном раскрытии изобретения (например, с использованием кинетических, электронных, электромагнитных, оптических сигналов или других типов физических сигналов от секвенсора). Цепи Уотсона и Крика в кольцевой молекуле ДНК можно исследовать в той же реакции. После секвенирования последовательности адаптеров обрезаются.

Различные уровни метилирования могут быть определены с помощью анализа. В (I) на фиг. 61, можно проанализировать паттерн метилирования только одноцепочечной молекулы, такой как X (a) или X (b). Этот анализ можно назвать анализом паттерна метилирования одной цепи. Анализ может включать в себя, но не ограничивается определением статуса метилирования сайтов или паттерна метилирования. На фиг. 61, одноцепочечная молекула X(a) демонстрирует паттерн метилирования 5'-UMMUU-3', где "U" обозначает неметилированный сайт, а "M" обозначает метилированный сайт, в то время как комплементарная одноцепочечная молекула X(b) демонстрирует паттерн метилирование 3'-UMUUU-5'. Таким образом, X(b) имеет паттерн метилирования, отличный от X(a). Соответствующие уровни метилирования одной цепи для X(a) и X(b) составляют 40 и 20%, соответственно.

Напротив, как показано в (II), можно анализировать паттерны метилирования на уровне одной двухцепочечной молекулы ДНК (т.е. принимать во внимание паттерны метилирования обеих цепей Уотсона и Крика. Этот анализ можно назвать анализом паттерна метилирования одной молекулы двухцепочечной ДНК. Уровень метилирования одной молекулы двухцепочечной ДНК молекулы X данного примера составляет 30%. В одном из вариантов этого анализа кинетические сигналы из обеих цепей Уотсона и Крика могут быть объединены для анализа модификации. В частности, поскольку метилирование по сайтах CpG обычно симметрично, кинетические сигналы от цепей Уотсона и Крика могут быть объединены для сайта до определения статусов метилирования сайтов. В некоторых ситуациях, эффективность определения модификаций оснований с использованием кинетических сигналов, объединенных из цепей Уотсон и Крика молекулы, будет лучше, чем определение, которое независимо использует кинетические сигналы одной цепи. Например, как показано на фиг. 20B, комбинированное использование кинетических сигналов из обеих цепей, включая цепи Уотсона и Крика, приведет к увеличению AUC (0,90) в тестовом наборе данных по сравнению с независимым использованием одной цепи (AUC: 0,85).

В (III) на фиг. 61, определяют уровень метилирования интересующей области генома, где разные молекулы ДНК, имеющие разные размеры молекул и разное количество метилируемых сайтов (например, сайтов CpG), могут вносить вклад в интересующую область генома. Этот анализ можно назвать анализом уровня метилирования множества цепей. Термин "множество цепей" может относиться к множеству одноцепочечных молекул ДНК, или множеству двухцепочечных молекул ДНК, или любой их комбинации. В данном примере имеется три молекулы двухцепочечной ДНК, покрывающие интересующую область генома: молекулы "X", "Y" и "Z", каждая из которых имеет цепи "a" и "b". Соответствующий уровень метилирования данной области составляет 9/28, т.е. 32%. Размер анализируемой области генома

может составлять 1, 10, 20, 30, 40, 50, 100 нт, 1 кнт (килонуклеотидов, т.е. одну тысячу нуклеотидов), 2, 3, 4, 5, 10, 20, 30, 40, 50, 100, 200, 300, 400, 500 кнт, 1 Мнт (мегануклеотидов, т.е. 1 миллион нуклеотидов), 2, 3, 4, 5, 10, 20, 30, 40, 50, 100 или 200 Мнт. Геномная область может представлять собой хромосомное плечо или весь геном.

Паттерн метилирования также может быть определен после определения статусов метилирования сайтов в молекуле. Например, в одном сценарии, где имеется три последовательных сайта CpG на одной двухцепочечной молекуле ДНК, паттерн метилирования на каждой из цепей - Уотсона и Крика может быть определен как метилировано (M), неметилировано (N) и метилировано (M) для указанных трех сайтов. Этот паттерн, MNM, например, для цепи Уотсона может быть обозначен как "гаплотип метилирования" для цепи Уотсона для данной области. Из-за присутствия активности поддержания метилирования ДНК, паттерны метилирования цепи Уотсона и цепи Крика двухцепочечной молекулы ДНК могут быть комплементарными. Например, если сайт CpG метилирован на цепи Уотсона, комплементарный сайт CpG на цепи Крика также может быть метилирован. Подобным образом, неметилированный сайт CpG на цепи Уотсона может быть комплементарным неметилированному сайту CpG на цепи Крика.

В одном варианте осуществления, можно измерить уровень метилирования отдельной молекулы ДНК, который определяется как количество метилированных оснований или нуклеотидов в молекуле, разделенное на совокупное количество метилируемых оснований или нуклеотидов в данной молекуле. Это измерение также называют "уровнем метилирования отдельной молекулы". Это измерение отдельной молекулы может быть особенно полезным в контексте данного раскрытия изобретения из-за большой длины прочтения, что возможно с помощью платформы для секвенирования отдельной молекулы в реальном времени. Когда измеряют уровни метилирования отдельной молекулы из множества молекул ДНК, можно также определить "уровень метилирования множества молекул" на основе концепций и уравнения на фиг. 61. Например, "уровень метилирования множества молекул" может быть средним или медианным уровнями метилирования отдельных молекул.

В некоторых вариантах осуществления, один или большее количество генетических полиморфизмов (например, однонуклеотидных полиморфизмов (ОНП)) могут быть проанализированы в молекуле ДНК вместе со статусом метилирования сайта в молекуле, таким образом выявляя как генетическую, так и эпигенетическую информацию о данной молекуле. Такой анализ покажет "оценку гаплотипа метилирования" анализируемой молекулы ДНК. Анализ оценки гаплотипа метилирования полезен, например, при исследовании геномного импринтинга и внеклеточных нуклеиновых кислот в материнской плазме (содержащей смесь внеклеточных молекул ДНК, несущих генетические и эпигенетические сигнатуры матери и плода).

b) Сравнение результатов метилирования.

Плотности метилирования на уровне всего генома для тканей в таблице на фиг. 60 определены с использованием бисульфитного секвенирования и секвенирования отдельной молекулы в реальном времени, как описано в данном изобретении. Фиг. 62A демонстрирует плотность метилирования, количественно определенную бисульфитным секвенированием, по оси ординат и тип ткани по оси абсцисс. Фиг. 62B демонстрирует плотность метилирования, количественно определенную путем секвенирования отдельной молекулы в реальном времени, как описано в данном изобретении, по оси ординат и тип ткани по оси абсцисс.

Фиг. 62A демонстрирует плотности метилирования в различных тканях с использованием бисульфитного секвенирования (т.е. образцы были преобразованы с помощью бисульфита, а затем подвергались секвенированию Illumina) (Lister et al. Nature. 2009; 462: 315-322), включая HepG2, опухолевые ткани ГНК, сопоставленные с нормальной тканью печени, прилегающей к опухоли ГЦК (т.е. соседние нормальные ткани), образцы плацентарной ткани и лейкоцитарного слоя. HepG2 показала самый низкий уровень метилирования, с уровнем метилирования 40,4%. Образцы лейкоцитарного слоя показали самый высокий уровень метилирования, с уровнем метилирования 76,5%. Средняя плотность метилирования опухолевых тканей ГЦК (51,2%) оказалась ниже, чем у соответствующих прилегающих нормальных тканей (71,0%). Это согласуется с ожиданием, что опухоли ГЦК являются гипометилированными на уровне всего генома по сравнению с прилегающими нормальными тканями (Ross et al. Epigenomics. 2010;2:245-69). Набор данных был сгенерирован из ДНК, полученной с помощью Sequel II Sequencing Kit 1.0.

Части одних и тех же тканей подвергали анализу метилирования с использованием секвенирования отдельной молекулы в реальном времени и способов согласно данному изобретению. Результаты показаны на фиг. 62B. Анализ метилирования с использованием способов секвенирования отдельной молекулы в реальном времени согласно данному изобретению показал, что линия клеток HepG2 была наиболее гипометилированной, за ней следовала проанализированная опухолевая ткань ГЦК, а затем - плацентарная ткань. Образец прилегающей неопухолевой ткани печени был более метилирован, чем другие ткани, включая ГЦК и ткани плаценты, при этом образец лейкоцитного слоя был наиболее гиперметилирован.

Фиг. 63A, 63B и 63C демонстрируют корреляцию совокупных уровней метилирования, количественно определенных бисульфитным секвенированием и секвенированием отдельной молекулы в реальном времени согласно описанным в данном документе способам. Фиг. 63A демонстрирует уровень метилирования, количественно определенный с помощью бисульфитного секвенирования, по оси абсцисс, и

уровень метилирования, количественно определенный с помощью секвенирования отдельной молекулы в реальном времени, с использованием описанных в данном документе способов, по оси ординат. Сплошная черная линия - это подобранная линия регрессии. Пунктирная линия соответствует равности двух измерений.

Наблюдалась очень высокая корреляция уровней метилирования между бисульфитным секвенированием и секвенированием отдельной молекулы в реальном времени согласно раскрытому в данном документе изобретению (r=0,99; значение P<0,0001). Эти данные показали, что анализ метилирования с использованием способов секвенирования отдельной молекулы в реальном времени, раскрытых в данном документе, был эффективным инструментом для определения уровней метилирования между тканями и позволял сравнивать состояния и профили метилирования между этими тканями. Для двух определений уровней метилирования мы отметили, что наклон линии регрессии на фиг. 63A отклонился от единицы. Эти результаты предполагают, что существует отклонение между двумя измерениями (в некотором контексте это отклонение можно назвать систематической ошибкой), которое может присутствовать при определении уровней метилирования с использованием секвенирования отдельной молекулы в реальном времени согласно изобретению по сравнению с стандартным масштабным параллельным бисульфитным секвенированием.

В одном варианте осуществления, мы могли количественно оценить смещение, используя линейную регрессию или регрессию LOESS (локально взвешенное сглаживание). В качестве примера, если бы мы считали эталоном масштабное параллельное бисульфитное секвенирование (Illumina), результаты, определенные с помощью секвенирования отдельной молекулы в реальном времени согласно данному раскрытию изобретения, можно было бы преобразовать с использованием коэффициентов регрессии, таким образом согласовав субпрочтения между различными платформами. На фиг. 63A формула линейной регрессии представлена как Y=aX+b, где "Y" представляло уровни метилирования, определенные с помощью секвенирования отдельной молекулы в реальном времени согласно раскрытию изобретения; "X" представляло уровни метилирования, определенные с помощью бисульфитного секвенирования; "a" представляло наклон линии регрессии (например, a=0,62); "b" представляло точку пересечения с осью ординат (например, b=17,72). В данной ситуации согласованные значения метилирования, определенные с помощью секвенирования отдельной молекулы в реальном времени, буду рассчитывать как (Y-b)/a. В другом варианте осуществления, можно использовать соотношение отклонения между двумя измерениями (ΔM) и соответствующее среднее значение двух измерений ($\bar{M}$), которые определяются формулами (1) и (2) ниже:

$$\Delta M = S - \text{Обусловленное бисульфитом метилирование}, \quad (1)$$

$$\bar{M} = \frac{S + \text{Обусловленное бисульфитом метилирование}}{2}, \quad (2)$$

где "S" представляет уровень метилирования, определенный с помощью секвенирования отдельной молекулы в реальном времени согласно данному изобретению, а "Обусловленное бисульфитом метилирование" представляет уровень метилирования, определенный с помощью бисульфитного секвенирования.

Фиг. 63B демонстрирует взаимосвязь между ΔM и $\bar{M}$. Среднее значение двух измерений (ччч) отложено по оси абсцисс, а отклонение между двумя измерениями (ΔM) отложено по оси ординат. Пунктирная линия представляет собой линию, проходящую по горизонтали через ноль, на которой точка данных указывает на отсутствие разницы между двумя измерениями. Эти результаты предполагают, что отклонение варьировалось в зависимости от усредненных значений. Чем выше среднее значение двух измерений, тем больше будет отклонение. Медиана значений ΔM составляла -8,5% (диапазон: от -12,6% до +2,5%), подразумевая то, что существует расхождение между способами.

Фиг. 63C демонстрирует среднее двух измерений ($\bar{M}$) по оси абсцисс и относительное отклонение (RD) по оси ординат. Относительное отклонение определяется по формуле ниже:

$$RD = \frac{\Delta M}{\bar{M}} \times 100\%, \quad (3).$$

Пунктирная линия представляет собой линию, проходящую по горизонтали через ноль, на которой точка данных указывает на отсутствие разницы между двумя измерениями. Эти результаты предполагают, что относительное отклонение варьировало в зависимости от усредненных значений. Чем больше среднее двух определений, тем больше по величине будет относительное отклонение. Медиана значений RD составила -12,5% (диапазон: от -18,1% до +6,0%).

Сообщалось, что стандартное полногеномное бисульфитное секвенирование (Illumina) приводит к значительному отклонению выходных последовательностей и завышению оценки глобального метилирования, с существенными вариациями в количественной оценке уровней метилирования между способами в конкретных областях генома (Olova et al. Genome Biol. 2018;19:33). Раскрытые в данном документе способы могут быть выполнены без преобразования бисульфитом, которое может значительно разрушить ДНК, и могут выполняться без амплификации ПЦР, которая может усложнить процесс или может внести дополнительную ошибку в определение уровней метилирования.

Фиг. 64A и 64B демонстрируют паттерны метилирования при разрешении 1 млн.п.о. Фиг. 64A демонстрирует паттерн метилирования для линии клеток ГПК (HepG2). Фиг. 64B демонстрирует паттерн метилирования для образца лейкоцитного слоя из здорового контрольного субъекта. Идеограммы хромосом (внешнее кольцо на каждой фигуре) организованы от короткого плеча хромосомы к длинному по часовой стрелке. Второе кольцо от внешнего (также описанное как среднее кольцо) демонстрирует уровни метилирования, определенные бисульфитным секвенированием. Самое внутреннее кольцо демонстрирует уровни метилирования, определенные с помощью секвенирования отдельной молекулы в реальном времени согласно данному изобретению. Уровни метилирования подразделяют на 5 уровней, а именно: 0-20% (светло-зеленый), 20-40% (зеленый), 40-60% (синий), 60-80% (светло-красный) и 80-100% (красный). Как показано на фиг. 64A и 64B, профили метилирования при разрешении 1 млн.п.о. были сопоставимыми между бисульфитным секвенированием (средний круг) и секвенированием отдельной молекулы в реальном времени (самый внутренний круг) согласно данному изобретению. Показано, что уровень метилирования материнского образца лейкоцитарного слоя выше, чем у линии клеток ГЦК (HepG2).

Фиг. 65A и 65B демонстрируют диаграммы разброса уровней метилирования, измеренных с разрешением 1 млн.п.о. Фиг. 65A демонстрирует уровни метилирования для линии клеток ГЦК (HepG2). Фиг. 65B демонстрирует уровни метилирования для образца лейкоцитного слоя из здорового контрольного субъекта. Для обеих Фиг. 65A и 65B, уровни метилирования, измеренные с помощью бисульфитного секвенирования, приведены на оси абсцисс, а уровни метилирования, измеренные с помощью секвенирования отдельной молекулы в реальном времени согласно данному изобретению, приведены на оси ординат. Сплошная линия - это подобранная линия регрессии. Пунктирная линия соответствует двум методам измерения. Для линии клеток ГЦК уровень метилирования, определенный с помощью секвенирования отдельной молекулы в реальном времени с разрешением 1 млн.п.о., хорошо коррелировал с уровнем, измеренным с помощью бисульфитного секвенирования ($r=0,99$; $P<0,0001$) (фиг. 65A). Корреляцию также наблюдали для данных из образца лейкоцитарного слоя ($r=0,87$, $P<0,0001$) (фиг. 65B).

Фиг. 66A и 66B демонстрируют диаграммы разброса уровней метилирования, измеренных с разрешением 100 т.п.о. Фиг. 66A демонстрирует уровни метилирования для линии клеток ГЦК (HepG2). Фиг. 66B демонстрирует уровни метилирования для образца лейкоцитного слоя из здорового контрольного субъекта. Для обеих фиг. 66A и 66, уровни метилирования, измеренные с помощью бисульфитного секвенирования, приведены на оси абсцисс, а уровни метилирования, измеренные с помощью секвенирования отдельной молекулы в реальном времени согласно данному изобретению, приведены на оси ординат. Сплошная линия - это подобранная линия регрессии. Пунктирная линия соответствует двум методам измерения. Высокая степень корреляции между количественными измерениями метилирования между двумя способами при разрешении 1 млн.п.о. (или 1 Мнт) также наблюдалась, когда разрешение анализа увеличивалось до каждого окна в 100 т.п.о. (или 100 кнт). Все эти данные указывают на то, что подход для отдельной молекулы в реальном времени согласно данному изобретению представляет собой эффективный инструмент для количественной оценки уровней метилирования или плотности метилирования в геномных областях, различающихся при разных порядках разрешения, например, при 1 млн.п.о. (или 1 Мнт) или 100 т.п.о. (или 100 кнт). Данные также указывают на то, что данное изобретение является эффективным инструментом для оценки профилей метилирования или паттернов метилирования между областями или между образцами.

Фиг. 67A и 67B демонстрируют паттерны метилирования при разрешении 1 млн.п.о. Фиг. 67A демонстрирует паттерн метилирования опухолевой ткани ГЦК (TBR3033T). Фиг. 67B демонстрирует паттерн метилирования прилегающей нормальной ткани (TBR3033N). Идеограммы хромосом (внешнее кольцо на каждой фигуре) организованы от короткого плеча хромосомы к длинному по часовой стрелке. Второе кольцо от внешнего (также описанное как среднее кольцо) демонстрирует уровни метилирования, определенные бисульфитным секвенированием. Самое внутреннее кольцо демонстрирует уровни метилирования, определенные с помощью секвенирования отдельной молекулы в реальном времени согласно данному изобретению. Уровни метилирования подразделяют на 5 уровней, а именно: 0-20% (светло-зеленый), 20-40% (зеленый), 40-60% (синий), 60-80% (светло-красный) и 80-100% (красный). Как показано на фиг. 67A, мы могли обнаружить гипометилирование в ДНК опухолевой ткани ГЦК (TBR3033T), которую можно было отличить от ДНК прилегающей нормальной ткани печени (TBR3033N) на фиг. 67B. Были сопоставимы уровни и паттерны метилирования, определенные с помощью бисульфитного секвенирования (средний круг) и секвенирования отдельной молекулы в реальном времени (внутренний круг) согласно данному изобретению. Показано, что уровень метилирования ДНК прилегающей нормальной ткани выше, чем у ДНК опухолевой ткани ГЦК.

Фиг. 68A и 68B демонстрируют диаграммы разброса уровней метилирования, определенных с разрешением 1 млн.п.о. Фиг. 68A демонстрирует уровни метилирования опухолевой ткани ГЦК (TBR3033T). Фиг. 68B демонстрирует уровни метилирования для прилегающей нормальной ткани. Для обеих фиг. 68A и 68B, уровни метилирования, измеренные с помощью бисульфитного секвенирования, приведены на оси абсцисс, а уровни метилирования, измеренные с помощью секвенирования отдельной молекулы в реальном времени согласно данному изобретению, приведены на оси ординат. Сплошная

линия - это подобранная линия регрессии. Пунктирная линия соответствует двум методам измерения. Для ДНК опухолевой ткани ГЦК уровень метилирования, определенный с помощью секвенирования отдельной молекулы в реальном времени с разрешением 1 млн.п.о., хорошо коррелировал с уровнем, измеренным с помощью бисульфитного секвенирования (r=0,96; P<0,0001) (фиг. 68A). Данные для образца прилегающей нормальной ткани печени также были сопоставимы (r=0,83, значение P<0,0001) (фиг. 68B).

Фиг. 69A и 69B демонстрируют диаграммы разброса уровней метилирования, измеренных с разрешением 100 т.п.о. Фиг. 69A демонстрирует уровни метилирования опухолевой ткани ГЦК (TBR3033T). Фиг. 69B демонстрирует уровни метилирования для прилегающей нормальной ткани (TBR3033N). Для обеих фиг. 69A и 69B, уровни метилирования, измеренные с помощью бисульфитного секвенирования, приведены на оси абсцисс, а уровни метилирования, измеренные с помощью секвенирования отдельной молекулы в реальном времени согласно данному изобретению, приведены на оси ординат. Сплошная линия - это подобранная линия регрессии. Пунктирная линия соответствует двум методам измерения. Также наблюдали такую высокую степень корреляции количественных данных по метилированию между двумя способами при разрешении 1 млн.п.о., когда измерение уровней метилирования выполняли при более высоком разрешении, например, с окнами в 100 т.п.о.

Фиг. 70A и 70B демонстрируют паттерны метилирования при разрешении 1 млн.п.о. для другой опухолевой ткани и нормальной ткани. Фиг. 70A демонстрирует паттерн метилирования опухолевой ткани ГЦК (TBR3032T). Фиг. 70B демонстрирует паттерн метилирования прилегающей нормальной ткани (TBR3032N). Идеограммы хромосом (внешнее кольцо на каждой фигуре) организованы от короткого плеча хромосомы к длинному по часовой стрелке. Второе кольцо от внешнего (также описанное как среднее кольцо) демонстрирует уровни метилирования, определенные бисульфитным секвенированием. Самое внутреннее кольцо демонстрирует уровни метилирования, определенные с помощью секвенирования отдельной молекулы в реальном времени согласно данному изобретению. Уровни метилирования подразделяют на 5 уровней, а именно: 0-20% (светло-зеленый), 20-40% (зеленый), 40-60% (синий), 60-80% (светло-красный) и 80-100% (красный). Как показано на фиг. 70A, мы могли обнаружить гипометилирование в ДНК опухолевой ткани ГЦК (TBR3032T), которую можно было отличить от ДНК прилегающей нормальной ткани печени (TBR3032N) на фиг. 70B. Были сопоставимы уровни и паттерны метилирования, определенные с помощью бисульфитного секвенирования (средний круг) и секвенирования отдельной молекулы в реальном времени (внутренний круг) с использованием данного изобретения. Показано, что уровень метилирования ДНК прилегающей нормальной ткани выше, чем у ДНК опухолевой ткани ГЦК.

Фиг. 71A и 71B демонстрируют диаграммы разброса уровней метилирования, определенных при разрешении 1 млн.п.о. Фиг. 71A демонстрирует уровни метилирования опухолевой ткани ГЦК (TBR3032T). Фиг. 71B демонстрирует уровни метилирования для прилегающей нормальной ткани. Для обеих фиг. 71A и 71B, уровни метилирования, измеренные с помощью бисульфитного секвенирования, приведены на оси абсцисс, а уровни метилирования, измеренные с помощью секвенирования отдельной молекулы в реальном времени согласно данному изобретению, приведены на оси ординат. Сплошная линия - это подобранная линия регрессии. Пунктирная линия соответствует двум методам измерения. Для ДНК опухолевой ткани ГЦК уровень метилирования, определенный с помощью секвенирования отдельной молекулы в реальном времени с разрешением 1 млн.п.о., хорошо коррелировал с уровнем, определенным с помощью бисульфитного секвенирования (r=0,98; P<0,0001) (фиг. 71A). Данные для образца прилегающей нормальной ткани печени также были сопоставимы (r=0,87, значение P<0,0001) (фиг. 71B).

Фиг. 72A и 72B демонстрируют диаграммы разброса уровней метилирования, определенных с разрешением 100 т.п.о. Фиг. 72A демонстрирует уровни метилирования опухолевой ткани ГЦК (TBR3032T). Фиг. 72B демонстрирует уровни метилирования для прилегающей нормальной ткани (TBR3032N). Для обеих фиг. 72A и 72B, уровни метилирования, измеренные с помощью бисульфитного секвенирования, приведены на оси абсцисс, а уровни метилирования, измеренные с помощью секвенирования отдельной молекулы в реальном времени согласно данному изобретению, приведены на оси ординат. Сплошная линия - это подобранная линия регрессии. Пунктирная линия соответствует двум методам измерения. Также наблюдали такую высокую степень корреляции количественных данных по метилированию между двумя способами при разрешении 1 млн.п.о., когда измерения уровней метилирования выполняли при более высоком разрешении, например, с окнами в 100 т.п.о.

4. Области с различным метилирования между опухолевыми и прилегающими нормальными тканями.

В областях геномов злокачественных новообразований часто обнаруживают метиломные аберрации. Одним из примеров таких аберраций является гипометилирование и гиперметилирование выбранных областей генома (Cadieux et al. Cancer Res. 2006;66:8469-76; Graff et al. Cancer Res. 1995;55:5195-9; Costello et al. Nat Genet. 2000;24:132-8). Другой пример - аберрантный паттерн метилированных и неметилированных оснований в выбранных геномных областях. В этом разделе показано, что методы определения метилирования могут использоваться при выполнении количественного анализа и диагностики

при анализе опухолей.

Фиг. 73 демонстрирует пример аберрантного паттерна метилирования рядом с геном-супрессором опухоли CDKN2A. Координаты, выделенные синим и подчеркнутые, указывают на островки CpG. Черные закрашенные точки указывают на метилированные сайты. Незакрашенные точки указывают на неметилированные сайты. Цифры в круглых скобках справа от каждой горизонтальной линии с точками обозначают размер фрагмента, плотность метилирования отдельной молекулы и количество сайтов CpG. Например, (3,3 т.п.н., ПМ:17,9%, CG:39) означает, что размер этого фрагмента составляет 3,3 т.п.н., уровень метилирования этого фрагмента составляет 17,9%, а количество сайтов CpG составляет 39. ПМ представляет собой плотность метилирования.

Как показано на фиг. 73, ген CDKN2A (ингибитор циклинзависимой киназы 2A) кодирует два белка, включая INK4A (p16) и ARF (p14), действующих как опухолевые супрессоры. Было две молекулы (молекула 7301 и молекула 7302), покрывающие область, перекрывающуюся с геном CDKN2A в неопухолевой ткани, прилегающей к опухолевой ткани. Уровни метилирования отдельной двухцепочечной молекулы ДНК для молекулы 7301 и молекулы 7302 составили 17,9% и 7,6% соответственно. Напротив, было обнаружено то, что уровень метилирования отдельной двухцепочечной молекулы ДНК для молекулы 7303, присутствующей в опухолевой ткани, составил 93,9%, что намного выше, чем у молекул, присутствующих в прилегающих неопухолевых тканях, поставленных в пару с опухолевой. С другой стороны, можно также рассчитать уровень метилирования множества цепей, используя молекулы 7301 и 7302, присутствующие в неопухолевой ткани, прилегающей к опухолевой ткани. Как результат, уровень метилирования множества цепей составил 9,7%, что ниже, чем таковой в опухолевой ткани (93,9%). Различные уровни метилирования предполагают, что можно использовать уровень метилирования одной двухцепочечной молекулы и/или уровень метилирования множества цепей для обнаружения или мониторинга таких заболеваний, как рак.

Фиг. 74A и 74B демонстрируют области неодинакового метилирования, обнаруженные с помощью секвенирования отдельной молекулы в реальном времени согласно вариантам осуществления данного изобретения. Фиг. 74A демонстрирует гипометилирование в геноме рака. Фиг. 74B демонстрирует гиперметилирование в геноме рака. Ось x обозначает координаты сайтов CpG. Координаты, выделенные синим и подчеркнутые, указывают на островки CpG. Черные закрашенные точки указывают на метилированные сайты. Незакрашенные точки указывают на неметилированные сайты. Цифры в круглых скобках справа от каждой горизонтальной линии с точками обозначают размер фрагмента, плотность метилирования на уровне фрагмента, и количество сайтов CpG. Например, (3,1 т.п.н., ПМ:88,9%, CG:180) означает, что размер этого фрагмента составляет 3,1 т.п.н., плотность метилирования этого фрагмента составляет 88,9%, а количество сайтов CpG составляет 180.

Фиг. 74A демонстрирует область, близкую к гену GNAS, демонстрирующую больше гипометилированных фрагментов в опухолевой ткани ГЦК по сравнению с прилегающей нормальной тканью печени. Фиг. 74B демонстрирует область, близкую к гену ESR1, демонстрирующую гиперметилированный фрагмент в ткани ГЦК, но фрагмент ДНК из поставленной в пару прилегающей неопухолевой ткани, выравненной с соответствующей областью, вместо этого показал гипометилирование. Как показано на фиг. 74B, профили метилирования или гаплотипы метилирования отдельных молекул ДНК были пригодны для выявления аберрантного статуса метилирования данных геномных областей, а именно GNAS и ESR1, когда образцы рака сравнивают с нераковыми образцами.

Эти данные показывают, что описанный в данном документе анализ метилирования с секвенированием отдельной молекулы в реальном времени может определять статус метилирования в каждом сайте CpG (метилирован или не метилирован) на отдельных фрагментах ДНК. Длина прочтения для секвенирования отдельной молекулы в реальном времени намного больше (порядка килобаз в длину), чем у секвенирования Illumina, которое обычно может охватить 100-300 нуклеотидов за одно прочтение (De Maio et al. Micob Genom. 2019;5(9)). Комбинируя свойство большой длины прочтения секвенирования отдельной молекулы в реальном времени с способом анализа метилирования, который мы раскрыли, можно легко определить гаплотип метилирования множества сайтов CpG, которые присутствуют на всей любой отдельной молекуле ДНК. Профиль метилирования относится к статусу метилирования сайтов CpG от одной координаты генома к другой координате на непрерывном участке ДНК (например, на той же хромосоме, или в бактериальной плазмиде, или на одном участке ДНК в вирусном геноме).

Поскольку секвенирование отдельной молекулы в реальном времени анализирует каждую молекулу ДНК отдельно без необходимости в предварительной амплификации, профиль метилирования, определенный для любой отдельной молекулы ДНК, фактически является гаплотипом метилирования, что означает статус метилирования сайтов CpG от одного конца до другого конца той же молекулы ДНК. Если одна или большее количество молекул секвенированы из одной и той же области генома, % метилирования (а именно уровень метилирования или плотность метилирования) каждого сайта CpG по всем секвенированным сайтам CpG в геномной области можно агрегировать из данных множества фрагментов ДНК с использованием формулы, которая показана на фиг. 61. Процент метилирования каждого сайта CpG может быть представлен для всех секвенированных сайтов CpG, с предоставлением профиля метилирования секвенированной области генома. В альтернативном варианте, данные могут быть агрегированы из

всех прочтений и всех сайтов в секвенированной области генома, чтобы предоставить значение % метилирования области, а именно таким же образом, как были рассчитаны уровни метилирования для областей 1 млн.п.о. или 1 т.п.о., как показано на фиг. 64-72.

5. Анализ метилирования вирусной ДНК.

В этом разделе показано, что методы метилирования согласно данному раскрытию изобретения могут использоваться для точного определения уровней метилирования в вирусной ДНК.

Фиг. 75 демонстрирует паттерны метилирования ДНК вируса гепатита В между двумя парами образцов ткани ГЦК и образцами прилегающей неопухолевой ткани с использованием секвенирования отдельной молекулы в реальном времени. Каждая стрелка представляет собой аннотацию гена в геноме HB V. Стрелки с "P", "S", "X" и "C" обозначают аннотацию генов для генома HBV: кодирующие полимеразу, поверхностный антиген, X-белок и коровый белок, соответственно. Мы идентифицировали один фрагмент (молекула I) размером 1183 п.о., происходящий из прилегающих неопухолевых тканей, охватывающий геном HBV от 2278 до 3141, выделенный пунктирным прямоугольником, показывающий уровень метилирования 12%. Мы также идентифицировали три фрагмента (молекулы II, III и IV) с 3215 п.о., 2961 п.о. и 3105 п.о., происходящих из опухолевых тканей. Среди них два фрагмента (молекула III и IV) в опухолях ГЦК перекрываются с геномными областями HBV, охватываемыми молекулой I в неопухолевых тканях. В отличие от низкого уровня метилирования (12%) в области HBV, выделенной пунктирным прямоугольником (позиции в геноме HBV: 2278-3141), уровни метилирования были выше для данных фрагментов (молекулы III и IV) в тканях ГЦК (т.е. 24% и 30%). Эти результаты свидетельствуют о том, что подход с использованием секвенирования отдельной молекулы в реальном времени был выполним для определения паттернов метилирования в вирусном геноме и позволял идентифицировать неодинаково метилированную область (HMO) HBV между ГЦК и не ГЦК тканями. Следовательно, определение состояний метилирования в вирусных геномах с использованием секвенирования отдельной молекулы в реальном времени согласно данному изобретению могло бы предоставить новый инструмент для изучения клинической значимости с использованием биопсии ткани.

Эта область HMO перекрывалась с генами P, C и S. Сообщалось, что эта область также была гиперметилирована в тканях ГЦК по сравнению с таковой в тканях печени с инфекцией HBV, но без рака (Jain et al. Sci Rep. 2015;5:10478; Fernandez et al. Genome Res. 2009;19:438-51).

Мы объединили результаты бисульфитного секвенирования тканей печени из четырех пациентов с циррозом, но без ГЦК, получив 1156 фрагментов HBV для анализа метилирования. Фиг. 76A демонстрирует уровни метилирования ДНК вируса гепатита В в тканях печени пациентов с циррозом, но без ГЦК. Кроме того, мы объединили результаты бисульфитного секвенирования опухолевых тканей ГЦК из 15 пациентов, получив 736 фрагментов HBV для анализа метилирования. Фиг. 76B демонстрирует уровни метилирования ДНК вируса гепатита В в опухолевой ткани ГЦК. Как показано на фиг. 76A и 76B, мы также наблюдали HMO-область HBV (позиции в геноме HBV: 1982-2435), которая имела более высокий уровень метилирования в тканях ГЦК, чем в тканях с циррозом печени, с помощью масштабного параллельного бисульфитного секвенирования. Эти результаты предполагают, что подход к определению статуса метилирования вирусных геномов был обоснованным.

6. Вариант-ассоциированный анализ метилирования.

Разные аллели могут быть связаны с разными профилями метилирования. Например, подверженные импринтингу гены могут иметь один аллель с более высоким уровнем метилирования, чем другой аллель. В этом разделе показано, что профили метилирования можно использовать для различения аллелей в определенных областях генома.

Одна ячейка секвенирования отдельной молекулы в реальном времени, содержащая одну ДНК-матрицу, может генерировать множество субпрочтений. Супрочтения включают в себя кинетические характеристики [например, межимпульсный период (МИП) и ширину импульса (ШИ)] и нуклеотидный состав. В одном варианте осуществления, субпрочтения из одной ячейки секвенирования отдельной молекулы в реальном времени могут использоваться для создания консенсусной последовательности (также называемой кольцевой консенсусной последовательностью, ККП), которая может значительно уменьшить ошибки секвенирования (например, несовпадения, вставки или делеции). Дополнительные сведения о ККП описаны в данном документе. В одном варианте осуществления, консенсусная последовательность может быть сконструирована с использованием этих субпрочтений, выровненных с эталонным геномом человека. В другом варианте осуществления, консенсусная последовательность может быть сконструирована путем картирования субпрочтений на самом длинном субпрочтении в той же ячейке секвенирования отдельной молекулы в реальном времени.

Фиг. 77 демонстрирует принцип анализа оценки гаплотипа метилирования. Закрашенные кружки на палочке представляют сайты CpG, которые классифицированы как метилированные. Незакрашенные кружки на палочке представляют сайты CpG, которые классифицированы как неметилированные.

Как показано в одном варианте осуществления на фиг. 77, субпрочтения были выровнены с эталонным геномом человека. Выровненные субпрочтения из одной ячейки секвенирования отдельной молекулы в реальном времени были сложены, чтобы сформировать консенсусную последовательность. Консенсусную последовательность в целом можно определить с использованием наиболее часто встречающихся

нуклеотидов, присутствующих в субпрочтениях в каждой выровненной позиции. Следовательно, варианты нуклеотидов, включая, помимо прочего, однонуклеотидные варианты, вставки и делеции, могут быть идентифицированы по консенсусным последовательностям. Усредненные МИП и ШИ в одной и той же молекуле, помеченные нуклеотидным вариантом, могут быть использованы для определения паттернов метилирования согласно данному изобретению. Таким образом, мы могли бы дополнительно определить паттерны вариант-ассоциированного метилирования. Состояния метилирования в одной и той же молекуле можно рассматривать как гаплотип метилирования. Гаплотип метилирования не может быть легко и напрямую сконструирован из двух или большего количества коротких молекул ДНК, поскольку может отсутствовать молекулярный маркер, позволяющий дифференцировать, происходят ли две или большее количество фрагментированных коротких молекулы ДНК из исходной одной молекулы, или происходят из двух или большее разных исходных молекул. Технологии синтетического длинного прочтения (такие как секвенирование с соединенным прочтением, разработанное 10X Genomics) предлагают возможность поместить одну длинную молекулу ДНК в ячейку (например, каплю) и разметить короткие молекулы ДНК, происходящие из этой длинной молекулы ДНК, с помощью одинаковых молекулярных последовательностей бар-кодов. Однако, этот этап баркодирования включает в себя амплификацию ПНР, которая не сохраняет исходные состояния метилирования.

Кроме того, если кто-то пытается использовать бисульфит для обработки длинных молекул ДНК, первый шаг перед обработкой бисульфитом включает в себя денатурацию ДНК в деструктивных условиях, превращающих двухцепочечную ДНК в одноцепочечную ДНК, поскольку бисульфит может действовать только на одноцепочечные ДНК молекулы в определенных химических условиях. Эта стадия денатурации ДНК приведет к разрушению длинных молекул ДНК на короткие фрагменты, что приведет к потере информации об исходном гаплотипе метилирования. Второй недостаток анализа метилирования на основе бисульфита заключается в денатурации двухцепочечной ДНК в одноцепочечную ДНК на стадии преобразования бисульфитом, а именно цепи Уотсона и Крика. Для молекулы существует 50%-ная вероятность секвенирования цепи Уотсона и 50%-ная вероятность секвенирования цепи Крика. Среди миллионов цепей Уотсона и Крика вероятность одновременного секвенирования цепей Уотсона и Крика крайне мала. Хотя предполагается, что обе цепи, Уотсона и Крика, в молекуле будут секвенированы, все еще невозможно однозначно определить, происходят ли такие цепи Уотсона и Крика из одного исходного фрагмента, или происходят из двух или больше разных исходных фрагментов. Liu et al недавно представили способ секвенирования без бисульфита для обнаружения метилированных цитозинов и гидроксиметилцитозина (Liu et al. Nat Biotechnol. 2019;37:424-429) с использованием ферментативного преобразования белком транслокации десять-одиннадцать (TET) в мягких условиях, что приводит к меньшей деградации ДНК. Однако он включает в себя две последовательные стадии ферментативных реакций. Низкая скорость преобразования на каждой стадии ферментативной реакции значительно влияет на общую скорость преобразования. Кроме того, даже для этого способа секвенирования без бисульфита, для обнаружения метилированных цитозинов, все еще существует трудность в распознавании цепей Уотсона и Крика молекулы в результатах секвенирования.

Напротив, в вариантах осуществления данного изобретения, цепи Уотсона и Крика молекулы ковалентно лигируют через адаптеры в форме колокольчика с образованием кольцевых молекул ДНК. В результате, цепи Уотсона и Крика молекулы секвенируют в одной и той же реакционной ячейке, и могут быть определены состояния метилирования для каждой цепи.

Одним из преимуществ вариантов осуществления данного изобретения является возможность получить информацию о метилировании и генетическую информацию (т.е. о последовательности) из длинной непрерывной молекулы ДНК (например, с длиной, составляющей килооснования или килонуклеотиды). Сгенерировать такую информацию сложнее, используя технологии секвенирования с короткими прочтениями Для технологий секвенирования с короткими прочтениями необходимо комбинировать информацию о секвенировании для множества коротких прочтений с использованием каркасов генетических или эпигенетических сигнатур для того, чтобы установить информацию о метелировании и генетическую информацию для длинного участка. Однако это может оказаться сложной задачей во многих сценариях из-за расстояний между такими генетическими или эпигенетическими ориентирами. Например, в среднем на 1 т.п.о. приходится один ОНП, в то время как современные технологии секвенирования с коротким прочтением обычно могут секвенировать вплоть до 300 нт за одно прочтение, что дает 600 нт даже в формате парных концов.

В одном варианте осуществления, анализ гаплотипа метилирования, связанного с вариантами, можно использовать для изучения паттернов метилирования в подверженных импринтингу генах. Подверженные импринтингу области подлежат эпигенетической регуляции (например, метилированию CpG) по принципу родительского происхождения. Например, один образец ДНК лейкоцитарного слоя (N2) в таблице на фиг. 60 был секвенирован, чтобы получить около 152 миллионов субпрочтений. Для этого образца 53% ячеек секвенирования отдельной молекулы в реальном времени генерировали по меньшей мере одно субпрочтение, которое могло быть выравнено с эталонным геномом человека. Средняя глубина субпрочтений для каждой ячейки SMRT составляла 7,7x. Совокупно мы получили около 3 миллионов консенсусных последовательностей. Консенсусные последовательности по меньшей мере один раз охва-

тывали около 91% эталонного генома. Для охваченных областей глубина секвенирования составляла 7,9x. Набор данных был сгенерирован из ДНК, полученной с помощью Sequel II Sequencing Kit 1.0.

Фиг. 78 демонстрирует распределение размеров секвенированных молекул, определенное на основе консенсусных последовательностей, с медианным размером 6289 п.о. (диапазон: 66-198109 п.о.). Размер фрагмента (п.о.) показан на оси абсцисс, а частота (%), связанная с размером фрагмента, показана на оси ординат.

Фиг. 79A, 79B, 79C и 79D демонстрируют примеры паттернов аллельного метилирования в областях импринтинга. Ось x обозначает координаты сайтов CpG. Координаты, выделенные синим и подчеркнутые, указывают на островки CpG. Черные закрашенные точки указывают на метилированные сайты CpG. Незакрашенные точки указывают на неметилированные сайты CpG. Буквы, помещенные между каждой горизонтальной серией закрашенных и не закрашенных точек (т.е., сайтов CpG), обозначают аллель в сайте ОНП. Цифры в круглых скобках справа от каждой горизонтальной серии точек обозначают размер фрагмента, плотность метилирования на уровне фрагмента, и количество сайтов CpG. Например, (10,0 т.п.н., ПМ:79,1%, CG:139) предполагает, что размер этого фрагмента составляет 10,0 т.п.н., плотность метилирования этого фрагмента составляет 79,1%, а количество сайтов CpG составляет 139. Пунктирными прямоугольниками обозначены наиболее дифференцированно метилированные области в пределах каждого гена.

Фиг. 79A демонстрирует 11 секвенированных фрагментов со средним размером 11,2 т.п.н. (диапазон: 1,3-25 т.п.н.), происходящих из гена SNURF. Ген SNURF был подвергнут материнскому импринтингу, что означает, что копия гена, унаследованная человеком от матери, метилирована и заглушена транскрипционно. Как показано на фиг. 79A, в пунктирном прямоугольнике, фрагменты, связанные с аллелем C, были крайне метилированы, тогда как фрагменты, связанные с аллелем T, были крайне неметилированы. Значительное метилирование может обозначать то, что более 70, 80, 90, 95 или 99% сайтов метилированы. Паттерны аллель-специфического метилирования можно было наблюдать в других подверженных импринтингу генах, включая PLAGL1 (фиг. 79B), NAP1L5 (фиг. 79C) и ZIM2 (фиг. 79D). Фиг. 79B демонстрирует, что для PLAGL1 фрагменты, ассоциированные с аллелем T, были крайне неметилированы, в то время как фрагменты, ассоциированные с аллелем C, были крайне метилированы. Фиг. 79C демонстрирует, что для NAP1L5 фрагменты, ассоциированные с аллелем C, были крайне неметилированы, в то время как фрагменты, ассоциированные с аллелем T, были крайне метилированы. Фиг. 79D демонстрирует, что для ZIM2 фрагменты, ассоциированные с аллелем C, были крайне неметилированы, в то время как фрагменты, ассоциированные с аллелем T, были крайне метилированы.

Фиг. 80A, 80B, 80C и 80D демонстрируют примеры паттернов аллельного метилирования в не подверженных импринтингу областях.Ось x обозначает координаты сайтов CpG. Координаты, выделенные синим и подчеркнутые, указывают на островки CpG. Черные закрашенные точки указывают на метилированные сайты CpG. Незакрашенные точки указывают на неметилированные сайты CpG. Буквы, помещенные между каждой горизонтальной серией закрашенных и не закрашенных точек (т.е., сайтов CpG), обозначают аллель в сайте однонуклеотидного полиморфизма (ОНП). Цифры в круглых скобках справа от каждой горизонтальной серии точек обозначают размер фрагмента, плотность метилирования на уровне фрагмента, и количество сайтов CpG. Пунктирные прямоугольники обозначают случайно выбранные области для расчета плотностей метилирования, указанных в скобках. В отличие от результатов на фиг. 79A-79D, не было обнаружено таких наблюдаемых паттернов аллельного метилирования в не подверженных импринтингу генах. Фиг. 80A демонстрирует то, что нет неодинакового паттерна аллельного метилирования в области chr7. Фиг. 80B демонстрирует то, что нет неодинакового паттерна аллельного метилирования в области chr12. Фиг. 80C демонстрирует то, что нет неодинакового паттерна аллельного метилирования в области chrl. Фиг. 80D демонстрирует то, что нет неодинакового паттерна аллельного метилирования в области chr1.

Фиг. 81 демонстрирует таблицу с уровнями метилирования аллель-специфичных фрагментов. В первом столбце указаны категории "подверженные импринтингу гены" и "случайно выбранные области". Во втором столбце указан конкретный ген. В третьем столбце указан первый аллель в ОНП гена. В четвертом столбце указан второй аллель в ОНП гена. В пятом столбце показан уровень метилирования фрагментов, связанных с первым аллелем. В шестом столбце показан уровень метилирования фрагментов, связанных со вторым аллелем. Уровни метилирования фрагментов, связанных с аллелем 2 (среднее значение: 88,6%; диапазон 84,6-91,1%), намного выше, чем фрагментов, связанных с аллелем 1 (среднее значение: 12,2%; диапазон 7,6-15,7%) для этих подверженных импринтингу генов (значение P=0,03), что указывает на наличие аллель-специфичного метилирования. Напротив, нет значительных отличий в уровнях метилирования между этими случайно выбранными областями (значение P=1), что указывает на отсутствие аллель-специфичного метилирования.

7. Анализ внеклеточной ДНК при беременности.

В этом примере показано, что раскрытые в данном документе способы применимы для анализа внеклеточных нуклеиновых кислот в плазме или сыворотке, полученных от женщин, беременных по меньшей мере одним плодом. Во время беременности молекулы внеклеточной ДНК и внеклеточной РНК из клеток плаценты обнаруживаются в кровотоке матери. Такие внеклеточные молекулы нуклеиновых ки-

слот, происходящие из плаценты, также называют внеклеточными нуклеиновыми кислотами плода в плазме матери или циркулирующими внеклеточными нуклеиновыми кислотами плода. Внеклеточные нуклеиновые кислоты плода присутствуют в материнской плазме на фоне материнских внеклеточных нуклеиновых кислот. Например, циркулирующие внеклеточные молекулы ДНК плода присутствуют в качестве второстепенных видов на фоне внеклеточной материнской ДНК в плазме и сыворотке матери.

Известно, что для того, чтобы отличить внеклеточную ДНК плода от внеклеточной материнской ДНК в плазме или сыворотке матери, можно использовать генетические или эпигенетические методы, или их комбинацию. Согласно генетике, геном плода может отличаться от материнского генома наследуемыми по отцу специфическими для плода аллелями ОНП, наследуемыми по отцу мутациями или мутациями de novo. Согласно эпигенетике, метилом плаценты в целом гипометилирован по сравнению с метиломом клеток материнской крови (Lun et al. Clin Chem. 2013; 59:1583-94). Поскольку плацента является основным источником внеклеточной ДНК плода, в то время как клетки материнской крови являются основным источником внеклеточной материнской ДНК в кровотоке (плазме или сыворотке), молекулы внеклеточной ДНК плода в целом гипометилированы по сравнению внеклеточной материнской ДНК в плазме или сыворотке. Существуют специфические геномные локусы, в которых плацента гиперметилирована по сравнению с материнскими клетками крови. Например, промотор и область экзона 1 RASSF1A более метилированы в плаценте, чем в материнских клетках крови (Chiu et al. Am J Pathol. 2007; 170:941-950). Таким образом, циркулирующая внеклеточная ДНК плода, полученная из данного локуса RASSFIA, будет гиперметилирована по сравнению с циркулирующей внеклеточной материнской ДНК из того же локуса.

В вариантах осуществления, внеклеточную ДНК плода можно отличить от внеклеточных молекул материнской ДНК на основании статуса неодинакового метилирования между двумя пулами циркулирующих нуклеиновых кислот. Например, обнаружено, что сайты CpG вдоль внеклеточной молекулы ДНК в основном неметилированы, и эта молекула, вероятно, происходит из плода. Если будет обнаружено, что сайты CpG вдоль внеклеточной молекулы ДНК в основном метилированы, эта молекула, скорее всего, принадлежит матери. Специалистам в данной области техники известно несколько способов, позволяющих установить, действительно ли такие молекулы принадлежат плоду или матери. Один из подходов состоит в сравнении паттерна метилирования секвенированной молекулы с известным профилем метилирования соответствующего локуса в клетках плаценты или материнских клетках крови.

Фиг. 82 демонстрирует пример определения плацентарного происхождения ДНК из плазмы во время беременности с использованием профилей метилирования. Координаты, выделенные синим и подчеркнутые, указывают на островки CpG. Черные закрашенные точки указывают на метилированные сайты. Незакрашенные точки указывают на неметилированные сайты. Цифры в круглых скобках возле каждой горизонтальной линии с точками обозначают размер фрагмента, плотность метилирования отдельной молекулы и количество сайтов CpG.

Как показано на фиг. 82, если внеклеточная молекула ДНК из материнской плазмы выравнивается с промоторной областью RASSF1A (областью, которая, как известно, специфически метилирована в тканях плаценты), и данные секвенирования, полученные с использованием способов согласно данному изобретению, указывают на гиперметилирование, эта молекула, вероятно, происходит из плода или плаценты. Напротив, молекулы, демонстрирующие гипометилирование, вероятно, происходят из материнской фоновой ДНК (преимущественно гематопоэтического происхождения).

Фиг. 83 иллюстрирует подход для анализа метилирования плода. Подход включает в себя использование секвенированной молекулы, которая содержит специфичный для плода аллель ОНП или специфическую для плода мутацию (например, наследуемую по отцу или de novo по своей природе). Когда идентифицируют такие специфичные для плода генетические признаки, статус метилирования оснований, представленных в одной и той же внеклеточной молекуле ДНК, отражает профиль метилирования внеклеточной ДНК плода или метилома плаценты. Специфические для плода генетические признаки могут быть обнаружены, когда секвенирование внеклеточной ДНК выявляет аллели или мутации, отсутствующие в материнском геноме (например, путем анализа материнской геномной ДНК), или путем анализа отцовской ДНК или для которых известна передача в семье (например, путем анализа ДНК пробанда).

Метилирование специфичных для плода молекул ДНК может быть определено путем анализа тех фрагментов ДНК, несущих аллели, которые отличались от гомозиготных аллелей в материнском геноме. Можно ожидать, что метилирование молекул ДНК плода будет ниже, чем у молекул материнской ДНК.

В качестве примера, были секвенированы ДНК лейкоцитного слоя одной беременной женщины и соответствующая ей плацентарная ДНК для получения гаплоидного генома 59x и 58x соответственно. Мы идентифицировали совокупно 822409 информативных ОНП, по которым мать была гомозиготной, а плод - гетерозиготным. Мы обнаружили 2652 специфичных для плода фрагментов и 24837 общих фрагментов (т.е. фрагментов, несущих общий аллель; преимущественно материнского происхождения) в материнской плазме (M13160) с помощью секвенирования отдельной молекулы в реальном времени. Фракция ДНК плода составила 19,3%. Согласно данному изобретению были получены профили метилирования этих специфичных для плода фрагментов и общих фрагментов. Как результат, уровень метилирова-

ния специфичных для плода фрагментов составил 57,4%, а уровень метилирования общих фрагментов - 69,9%. Это открытие согласуется с текущими знаниями о том, что уровень метилирования ДНК плода ниже, чем материнской ДНК в плазме беременной женщины (Lun et al., Clin Chem. 2013;59:1583-94).

Паттерны метилирования могут использоваться для целей диагностики или мониторинга. Например, профиль метилирования образца материнской плазмы был использован для определения гестационного возраста (https://www.ncbi.nlm.nih.gov/pubmed/27979959). Одно применение представляют собой этап контроля качества. Еще одно потенциальное применение - мониторинг "биологического" возраста по сравнению с "хронологическим" возрастом беременности. Это применение может использоваться для выявления или оценки риска преждевременных родов. Другие варианты осуществления могут быть использованы для анализа клеток плода в материнской крови. В еще других вариантах осуществления, такие клетки плода могут быть идентифицированы с помощью подходов на основе антител или селективного окрашивания с использованием клеточных маркеров (например, на поверхности клетки или в цитоплазме), или обогащены с помощью проточной цитометрии, микроманипулирования, микродиссекции или физических способов (например, потоком с дифференциальной скоростью через камеру, поверхность или контейнер).

C. Обнаружение метилирования с использованием различных реагентов.

В этом разделе показано, что методы метилирования не ограничиваются конкретной системой реагентов.

Анализ метилирования был выполнен с использованием различных систем реагентов, чтобы подтвердить возможность применения методов. Например, SMRT-секв. было выполнено с использованием системы Sequel II (Pacific Biosciences) для реализации секвенирования отдельной молекулы в реальном времени. Фрагментированные гидродинамическим сдвигом молекулы ДНК были подвергнуты построению матрицы для секвенирования отдельной молекулы в реальном времени (SMRT) с использованием SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences). Отжиг праймеров для секвенирования и условия связывания полимеразы рассчитывали с помощью программного обеспечения SMRT Link v8.0 (Pacific Biosciences). Вкратце, праймер v2 для секвенирования отжигали с матрицей для секвенирования, а затем полимеразу связывали с матрицами с использованием Sequel II Binding and Internal Control Kit 2.0 (Pacific Biosciences). Секвенирование выполняли на Sequel II SMRT Cell 8M. Фильмы секвенирования были собраны на системе Sequel II в течение 30 часов с помощью Sequel II Sequencing Kit 2.0 (Pacific Biosciences). В других вариантах осуществления, для SMRT-секв. могут использоваться другие химические реагенты и реакционные буферы. В одном варианте осуществления, полимераза будет иметь разные кинетические характеристики включения нуклеотидов в матричную цепь ДНК в зависимости от ее статуса метилирования (Huber et al. Nucleic Acids Res. 2016;44:9881-9890). В этом раскрытии изобретения, результаты получены с использованием праймера v1 для секвенирования, если не указано иное.

Чтобы продемонстрировать использование изобретения в раскрытии изобретения, описанном в данном документе, с использованием различных реагентов, мы проанализировали данные SMRT-секв., полученные на основе различных наборов для секвенирования, включающих в себя, но не ограничивающихся, Sequel I Sequencing Kit 3.0, RS II, Sequel II Sequencing Kit 1.0 и Sequel II Sequencing Kit 2.0. RS II включает в себя 150000 ZMW на секцию SMRT. В Sequel используется 1000000 ZMW на секцию SMRT. В Sequel II используется 8 миллионов ZMW на секцию SMRT с двумя наборами для секвенирования (1.0 и 2.0). Этот анализ включал в себя два набора данных. Первый набор данных был подготовлен на основе ДНК после амплификации всего генома, представляющий неметилированный статус. Набор данных второго типа был подготовлен на основе ДНК после обработки метилтрансферазой M.SsssI, представляющий метилированный статус. Эти данные были получены с использованием Sequel Sequencing Kit 3.0 в секвенаторе Sequel; и Sequel II Sequencing Kit 1.0, и Sequel II Sequencing Kit 2.0 в секвенаторе Sequel II. Таким образом, мы получили три набора данных с кинетическими профилями, сгенерированными с использованием различных реагентов (например, полимераз). Каждый набор данных был разделен на обучающий набор данных и тестовый набор данных для оценки эффективности с использованием моделей CHC согласно данному изобретению.

1. Окна определения.

Фиг. 84A, 84B и 84C демонстрируют эффективность разных размеров окон измерения для разных наборов реагентов для SMRT-секв. в обучающих наборах данных, содержащих данные всего амплифицированного генома (неметилированные сайты CpG) и данные при обработке M.SsssI (метилированные сайты CpG). Частота истинно-положительных результатов отложена по оси ординат, а частота ложных-положительных результатов - по оси абсцисс. Фиг. 84A демонстрирует данные SMRT-секв., сгенерированные на основе Sequel Sequencing Kit 3.0. Фиг. 84B демонстрирует данные SMRT-секв., сгенерированные на основе Sequel II Sequencing Kit 1.0. Фиг. 84C демонстрирует данные SMRT-секв., сгенерированные на основе Sequel II Sequencing Kit 2.0. На фигурах "-" обозначают сигналы выше анализируемого сайта цитозина CpG. "+" обозначает сигналы ниже анализируемого сайта цитозина CpG. Например, "-6 нт" представляет собой сигналы 6 нт выше анализируемого сайта цитозина CpG. "+6 нт" представляет собой сигналы 6 нт ниже анализируемого сайта цитозина CpG. "±6 нт" обозначает включение как сигналов 6 нт выше, так и сигналов 6 нт ниже анализируемого сайта цитозина CpG (т.е. всего 12 нт последова-

тельность, фланкирующая сайт цитозина CpG).

Для обучающего набора данных на основе Sequel Sequencing Kit 3.0, как показано на фиг. 84A, используя окно измерения, включающее в себя сигналы от анализируемого цитозина CpG и сигналы 6 нт выше (например, МИП, ШИ относительные позиции и состав последовательностей) данного сайта цитозина (обозначенного -6 нт), значение AUC 0,50 предполагает отсутствие способности отличать метилированные цитозины CpG от неметилированных. Однако для обучающих наборов данных на основе Sequel II Sequencing Kit 1.0 и 2.0, соответствующие значения AUC составляли 0,62 (фиг. 84B) и 0,75 (фиг. 84C). Эти данные продемонстрировали, что существуют разные кинетические профили, присущие разным реагентам, используемым в SMRT-секв. Эти данные показывают, что раскрытые в данном документе способы легко адаптировать к использованию различных реагентов. Кроме того, точность обнаружения модификаций оснований потенциально может быть улучшена с помощью дополнительных доработок реагентов, например, применяя различные полимеразы и другие химические реагенты.

В качестве другого примера, для обучающего набора данных на основе Sequel Sequencing Kit 3.0, как показано на фиг. 84A, используя окно измерения, включающее в себя сигналы 10 нт выше от анализируемого сайта цитозина CpG (обозначенного -10 нт), значение AUC 0,50 предполагает отсутствие способности отличать метилированные цитозины CpG от неметилированных. Однако для обучающих наборов данных на основе Sequel II Sequencing Kit 1.0 и 2.0, соответствующие значения AUC составляли 0,66 (фиг. 84B) и 0,79 (фиг. 84C), которые, как было показано, улучшились по сравнению с окном измерения, включающим в себя сигналы восходящих 6 нт. Эти данные подтвердили, что существуют разные кинетические профили, присущие разным реагентам, используемым в SMRT-секв. Эти данные показывают, что раскрытые в данном документе способы легко адаптировать к использованию различных реагентов.

В отличие от окна измерения с восходящими сигналами, окно определение с нисходящими сигналами может дать большее улучшение эффективности классификации. Например, для обучающего набора данных на основе Sequel Sequencing Kit 3.0, как показано на фиг. 84A, используя окно измерения, включающее в себя сигналы 6 нт ниже сайта цитозина CpG (+6 нт), значение AUC 0,94 было намного больше, чем при использовании сигналов восходящих 6 нт (AUC: 0,5). Для обучающих наборов данных на основе Sequel II Sequencing Kit 1.0 и 2.0, соответствующие значения AUC составляли 0,95 (фиг. 84B) и 0,92 (фиг. 84C), соответственно, демонстрируя улучшение по сравнению с окном измерения, включающим в себя сигналы восходящих 6 нт. Эти данные предполагают, что кинетические характеристики, связанные с контекстом последовательности, улучшат классификационную способность при использовании моделей CHC, но не ограничиваясь лишь ними. Эти данные также предполагают, что описанное в данном документе изобретение будет применимо к наборам данных, полученным с помощью разных реагентов и условий секвенирования (например, разных полимераз, других химических реагентов, их концентраций и параметров реакции секвенирования (например, продолжительности)) посредством настройки окна измерения. Аналогичный вывод можно сделать из анализа с использованием окна измерения, включающего в себя сигналы 10 нт ниже от сайта цитозина CpG (фиг. 84A, 84B и 84C).

В другом варианте осуществления, можно использовать окно измерения, содержащее сигналы анализируемого цитозина, и сигналы как выше, так и ниже данного цитозина. Например, как показано на фиг. 84A, 84B и 84C, используя окно измерения, включающее в себя сигналы восходящих 6 нт и сигналы нисходящих 6 нт (обозначены как ±6 нт), было определенно, что значения AUC составили 0,94, 0,95 и 0,92 для обучающего набора данных на основе Sequel Sequencing Kit 3.0, Sequel II Sequencing Kit 1.0 и 2.0, соответственно. Используя окно измерения, включающее в себя сигналы восходящих 10 нт и сигналы нисходящих 10 нт (обозначены как ±10 нт), было определенно, что значения AUC составили 0,94, 0,95 и 0,94 для обучающего набора данных на основе Sequel Sequencing Kit 3.0, Sequel II Sequencing Kit 1.0 и 2.0, соответственно. Эти данные предполагают, что описанное в данном документе изобретение будет более применимо к наборам данных, полученным с помощью разных реагентов и параметров секвенирования.

Фиг. 85A, 85B и 85C продемонстрировали, что результаты были получены из тестовых наборов данных с разными окнами измерения с разными наборами секвенирования при применении моделей CHC, обученных на основе обучающих наборов данных. Частота истинно-положительных результатов отложена по оси ординат, а частота ложных-положительных результатов - по оси абсцисс. Обозначения в описании эквивалентны обозначениям, используемым на фиг. 84A, 84B и 84C. Фиг. 85A демонстрирует данные SMRT-секв., сгенерированные на основе Sequel Sequencing Kit 3.0. Фиг. 85B демонстрирует данные SMRT-секв., сгенерированные на основе Sequel II Sequencing Kit 1.0. Фиг. 85C демонстрирует данные SMRT-секв., сгенерированные на основе Sequel II Sequencing Kit 2.0. Все выводы, сделанные на основе обучающих наборов данных, можно было проверить в этих независимых тестовых наборах данных, которые не были задействованы в процессе обучения. Кроме того, среди трех независимых тестовых наборов данных, анализ двух наборов данных (2/3) с использованием Sequel II Sequencing Kit 1.0 и 2.0 продемонстрировал, что применение окна измерения, включающего в себя сигналы восходящих и нисходящих 10 нт (обозначенных как ±10 нт), показало себя лучше других.

2. Сравнение с бисульфитным секвенированием.

Фиг. 86A, 86B и 86C демонстрируют корреляцию совокупных уровней метилирования, количественно определенных с помощью бисульфитного секвенирования и SMRT-секв. (Sequel II Sequencing Kit 2.0). Фиг. 86A демонстрирует уровень метилирования в виде процента, количественно определенного с помощью SMRT-секв., на оси ординат. Фиг. 86B демонстрирует уровень метилирования в виде процента, количественно определенного с помощью бисульфитного секвенирования, на оси абсцисс. Черная линия - это подобранная линия регрессии. Пунктирная линия - это диагональная линия, на которой два определения равны. Фиг. 86B демонстрирует график Бланда-Альтмана. По оси абсцисс показано среднее уровней метилирования, определенных количественно с помощью SMRT-секв., согласно данному изобретению, и бисульфитного секвенирования. Ось ординат показывает разницу в уровне метилирования между SMRT-секв. согласно данному изобретению и бисульфитным секвенированием (т.е. метилированием Pacific Biosciences - обусловленным бисульфитом метилированием). Пунктирная линия соответствует горизонтальной линии, проходящей через ноль, которая обозначает отсутствие разницы между двумя определениями. Точки данных, отклонившиеся от пунктирной линии, предполагают, что существуют отклонения между определениями. Фиг. 86C демонстрирует процентное изменение относительно значения, количественно определенного с помощью бисульфитного секвенирования. По оси абсцисс показано среднее уровней метилирования, определенных количественно с помощью SMRT-секв., согласно данному изобретению, и бисульфитного секвенирования. По оси ординат показан процент разницы в уровнях метилирования между двумя определениями по сравнению со средним уровнем метилирования. Пунктирная линия соответствует горизонтальной линии, проходящей через ноль, которая обозначает отсутствие разницы между двумя определениями. Точки данных, отклонившиеся от пунктирной линии, предполагают, что существуют отклонения между определениями.

На фиг. 86A формула линейной регрессии представлена как Y=aX+b, где "Y" представляет уровни метилирования, определенные с помощью SMRT-секв., согласно раскрытию изобретения; "X" представляет уровни метилирования, определенные с помощью бисульфитного секвенирования; "a" представляет наклон линии регрессии (например, a=1,45); "b" представляет точку пересечения с осью ординат (например, b=-20,98). В этой ситуации, значения метилирования, определенные с помощью SMRT-секв., будут рассчитываться как (Y-b)/a. Этот график демонстрирует, что уровни метилирования, определенные с помощью SMRT-секв., могут быть преобразованы в уровни метилирования, определенные с помощью бисульфитного секвенирования, и наоборот для Sequel II Sequencing Kit 2.0, как и для Sequel II Sequencing Kit 1.0.

Фиг. 86B представляет собой график Бланда-Альтмана, который показывает смещение количественной оценки метилирования между SMRT-секв. согласно данному изобретению и бисульфитным секвенированием, на котором ось абсцисс показывает среднее уровней метилирования, количественно определенных с помощью SMRT-секв. согласно данному изобретению и бисульфитного секвенирования, а ось ординат показывает разницу в уровнях метилирования, количественно определенных с помощью SMRT-секв. согласно данному изобретению и бисульфитного секвенирования. Медианная разница между двумя измерениями составила -6,85% (диапазон: -10,1 - 1,7%). Медианный процент изменение уровня метилирования, количественно определенного согласно данному изобретению, по сравнению со значением, полученным с помощью бисульфитного секвенирования, составил -9,96% (диапазон: -14,76 - 3,21%). Разница варьировалась в зависимости от усредненных значений. Чем выше среднее значение двух определений, тем выше смещение.

Фиг. 86C демонстрирует те же данные, что и на фиг. 86B, но с разницей в уровнях метилирования, разделенной на среднее значение двух уровней метилирования. Фиг. 86C также демонстрирует, что чем больше среднее значение двух определений, тем больше смещение.

Ошибка может быть связана с бисульфитным секвенированием и не связана с способами с SMRT-секв. Сообщалось, что стандартное полногеномное бисульфитное секвенирование (Illumina) приводит к значительному отклонению выходных последовательностей и завышению оценки глобального метилирования, с существенными вариациями в количественной оценке уровней метилирования между способами в конкретных областях генома (Olova et al. Genome Biol. 2018;19:33). Раскрытые в данном документе варианты осуществления имеют ряд иллюстративных преимуществ, благодаря которым их можно выполнять без преобразования бисульфитом, которое могло бы резко разрушить ДНК, и могут быть выполнены без ПЦР-амплификации.

3. Тканевое происхождение.

Мы выполнили анализ метилирования для различных типов рака согласно вариантам осуществления в данном раскрытии изобретения, используя секвенирование отдельной молекулы в реальном времени (SMRT-секв., Pacific Biosciences). Типы рака, использованные для SMRT-секв., включали в себя, без ограничения, колоректальный рак (n=3), рак пищевода (n=2), рак груди (n=2), почечно-клеточную карциному (n=2), рак легких (n=2), рак яичников (n=2), рак простаты (n=2), рак желудка (n=2) и рак поджелудочной железы (n=1). Соответствующие прилегающие неопухолевые ткани также были включены для SMRT-секв. Набор данных был сгенерирован из ДНК, полученной с помощью Sequel II Sequencing Kit 2.0.

Фиг. 87A и 87B демонстрируют сравнение совокупного уровня метилирования между различными

опухолевыми тканями и прилегающими неопухолевыми тканями в паре с ними. Уровень метилирования в процентах отложен по оси ординат. На фиг. 87A, уровень метилирования количественно определяется с помощью SMRT-секв. На фиг. 87B, уровни метилирования количественно определены с помощью бисульфитного секвенирования. Тип ткани (т.е. опухолевая ткань или прилегающая неопухолевая ткань) обозначена на оси абсцисс. Различные символы обозначают разные ткани по происхождению.

Фиг. 87A демонстрирует, что совокупные уровни метилирования опухолевых тканей, включая рак молочной железы, колоректальный рак, рак пищевода, рак печени, рак легкого, рак яичников, рак поджелудочной железы, почечно-клеточную карциному и рак желудка, были значительно ниже, чем соответствующих неопухолевых тканей (значение P=0,006, знаковый критерий Уилкоксона парных выборок), включая молочную железу, толстую кишку, пищевод, печень, легкое, яичники, поджелудочную железу, предстательную железу, почки и желудок, соответственно. Медианная разница в уровне метилирования между опухолью и неопухолевыми тканями в паре с ней составляла -2,7% (IQR: -6,4 ~ -0,8%).

Фиг. 84B подтверждает более низкие уровни метилирования в опухолевых тканях. Таким образом, эти результаты предполагают, что паттерны метилирования для различных типов рака и тканей могут быть точно определены с помощью SMRT-секв. согласно данному изобретению, что подразумевает широкое применение этого изобретения для раннего обнаружения, прогнозирвоания, диагностики и лечения рака, на основе биопсии ткани. Различные величины снижения уровня метилирования в различных типах опухолей, вероятно, предполагают, что паттерны метилирования были связаны с типами рака, что позволяет определить тканевое происхождение рака.

D. Улучшение обнаружения и другие методы.

В некоторых вариантах осуществления, анализ модификации основания (например, метилирования) может быть выполнен с использованием одного или большего количества из следующих параметров: контекст последовательности, МИП и ШИ МИП и ШИ можно определить из реакции секвенирования без выравнивания с эталонным геномом. Аспекты подхода секвенирования отдельной молекулы в реальном времени могут дополнительно повысить точность определения контекста последовательности, МИП и ШИ. Один из аспектов представляет собой выполнение кольцевого консенсусного секвенирования, при котором конкретная часть матрицы секвенирования может быть определена множество раз, что делает возможным определение контекста последовательности, МИП и ШИ на основе среднего или распределения значений с помощью этих множественных прочтений. В некоторых вариантах осуществления, анализ модификации основания без процесса выравнивания может повысить эффективность вычислений, сократить время обработки и может снизить затраты на анализ. Хотя варианты осуществления могут быть реализованы без процесса выравнивания, в еще других вариантах осуществления может быть использован и может быть предпочтительным процесс выравнивания, например, если процесс выравнивания используется для установления клинических или биологических последствий обнаруженной модификации основания (например, если супрессор опухолей гиперметилирован); или если процесс выравнивания используется для выбора подмножества данных секвенирования, которое соответствует определенным областям интереса генома для дополнительного анализа. Для вариантов осуществления, в которых желательны данные из выбранных областей генома, эти варианты осуществления могут включать в себя нацеливание на такие области с использованием одного или большего количества ферментов, или методов на основе ферментов, которые могут расщеплять области интереса генома, например, рестрикционнго фермента или системы CRISPR-Cas9. Система CRISPR-Cas9 может быть предпочтительнее способа на основе ПЦР, поскольку амплификация ПЦР обычно не сохраняет информацию о модификациях оснований ДНК. Уровни метилирования таких выбранных (либо биоинформатически [например, путем выравнивания], либо с помощью таких способов, как CRISPR-Cas9) областей могут быть проанализированы для получения информации о тканевом происхождении, нарушениях плода, нарушениях беременности и раке.

1. Анализ метилирования с использованием субпрочтений без выравнивания с эталонным геномом.

В вариантах осуществления, анализ метилирования может быть выполнен с использованием окон измерения, содержащих кинетические характеристики и контекст последовательности из субпрочтений, без выравнивания с эталонным геномом. Как показано на фиг. 88, субпрочтения, производимые волноводом с нулевой модой (ZMW), были использованы для создания консенсусной последовательности 8802 (также известной как кольцевая консенсусная последовательность, ККП). Были рассчитаны средние кинетические значения в каждой позиции в ККП, включая, помимо прочего, значения ШИ и МИП. Контекст последовательности, окружающей сайт CpG, определяли из ККП на основании последовательностей, расположенных выше и ниже этого сайта CpG. Следовательно, окно измерения, как указанно в данном раскрытии изобретения, будет построено для обучения, с окном измерения, включающим в себя значения ШИ, МИП, и контекст последовательности согласно субпрочтениям с кинетическими характеристиками относительно ККП. Эта процедура позволяет избежать выравнивания субпрочтений с эталонным геномом.

Чтобы проверить принцип, продемонстрированный на фиг. 88, мы использовали 601942 неметилированных сайтов CpG, которые происходили из амплифицированной ДНК всего генома, и 163527 метилированных сайтов CpG, которые происходили из ДНК, обработанной CpG-метилтрансферазой (напри-

мер, M.SssI), формируя обучающий набор данных. Мы использовали 546393 неметилированных сайтов CpG, которые происходили из амплифицированной ДНК всего генома, и 193641 метилированных сайтов CpG, которые происходили из ДНК, обработанной CpG-метилтрансферазой (например, M.SssI), формируя тестовый набор данных. Набор данных был сгенерирован из ДНК, полученной с помощью Sequel II Sequencing Kit 2.0.

Как показано на фиг. 89, в одном варианте осуществления, используя кинетические характеристики и контекст последовательности, ассоциированные с субпрочтениями и ККП, для обучения модели сверточной нейронной сети (СНС) для определения метилирования, можно достичь значения AUC 0,94 и 0,95 для дифференциации метилированных сайтов CpG от неметилированных сайтов CpG в тестовых и обучающих наборах данных, соответственно. В других вариантах осуществления, могут использоваться другие модели нейронных сетей, алгоритмы глубокого обучения, искусственный интеллект и/или алгоритмы машинного обучения.

Если мы установим пороговое значение 0,2 для вероятности метилирования, мы сможем получить чувствительность 82,4% и специфичность 91,7% при обнаружении метилированных сайтов CpG. Эти результаты продемонстрировали, что можно дифференцировать метилированные и неметилированные сайты CpG с помощью субпрочтений с кинетическими характеристиками без предварительного выравнивания с эталонным геномом.

В другом варианте осуществления, для определения статуса метилирования в сайтах CpG можно также использовать кинетические характеристики вместе с контекстом последовательности прямо из субпрочтений без информации КПП и предварительного выравнивания с эталонным геномом. Мы использовали кинетические характеристики, включая значения ШИ и МИП в позициях, простирающихся 20 нт выше и 20 нт ниже от CpG, представленного в субпрочтении, для обучения модели СНС для определения статуса метилирования. Как показано на фиг. 90, согласно вариантам осуществления в данном раскрытии изобретения, AUC кривой ROC с использованием кинетических характеристик, относящихся к субпрочтениям, составляла 0,70 и 0,69 для обнаружения метилированных сайтов CpG в обучающем и тестовом наборе данных, соответственно. Эти данные предполагают, что возможно использовать варианты осуществления из данного раскрытия изобретения, чтобы получить паттерны метилирования для молекулы ДНК с использованием кинетических характеристик, связанных с субпрочтениями, но без предварительного выравнивания и построения консенсусных последовательностей. Однако эффективность определения метилирования в данном варианте осуществления была хуже, чем в вариантах осуществления, комбинаторно использующих информацию о выравнивании или консенсусных последовательностях, как описано в данном раскрытии изобретения. Можно было бы предположить, что повышенная точность при генерации субпрочтений и кинетических значений улучшит эффективность определения модификаций оснований с использованием субпрочтений и связанных с ними кинетических характеристик.

2. Анализ метилирования удаленных областей с использованием целевого секвенирования отдельной молекулы в реальном времени.

Описанные в данном документе способы также можно применять для анализа одной или большего количества выбранных геномных областей. В одном варианте осуществления, область(ти) интереса может быть сначала обогащена с помощью способа гибридизации, который позволяет гибридизовать молекулы ДНК из области(тей) интереса с синтетическими олигонуклеотидами, имеющими комплементарные последовательности. Для анализа модификаций оснований с использованием описанных в данном документе способов целевые молекулы ДНК не могут быть амплифицированы с помощью ПНР перед секвенированием, поскольку информация о модификациях оснований в исходной молекуле ДНК не будет перенесена в продукты ПНР Было разработано несколько способов обогащения этих целевых областей без выполнения ПЦР-амплификации.

В другом варианте осуществления целевая область(ти) может быть обогащена за счет использования системы CRISPR-Cas9 (Stevens et al. PLOS One 2019;14(4):e0215441; Watson et al. Lab Invest 2020; 100:135-146). В одном варианте осуществления, концы молекул ДНК в образце ДНК сначала дефосфорилируют, что делает их нечувствительными к прямому лигированию с адаптерами секвенирования. Затем белок Cas9 нацеливает на область(ти) интереса направляющие РНК (крРНК) для создания двухцепочечных разрезов. Области интереса, фланкированные двухцепочечными разрезами с обеих сторон, затем будут лигированы с адаптерами секвенирования, которые указываются в выбранной платформе секвенирования. В другом варианте осуществления, ДНК можно обрабатывать экзонуклеазой, чтобы разрушались молекулы ДНК, не связанные белками Cas9 (Stevens et al. PLOS One 2019;14(4):e0215441). Поскольку эти способы не включают в себя ПЦР-амплификацию, исходные молекулы ДНК с модификацией оснований можно секвенировать и определить модификацию основания. В одном варианте осуществления, данный способ можно использовать для нацеливания на большое количество областей, имеющих общие гомологичные последовательности, например, длинные диспергированные ядерные повторы (LINE). В одном примере, такой анализ может быть использован для анализа внеклеточной ДНК, которую замыкают в кольцо, в материнской плазме для выявления анеуплоидий плода (Kinde et al. PLOS One 2012;7(7):e41162.

Как показано на фиг. 91, целевое секвенирование отдельной молекулы в реальном времени может быть реализовано с помощью системы CRISPR (короткие палиндромные повторы, регулярно расположенные группами)/Cas9 (связанный с CRISPR белок 9). Фрагменты ДНК (например, молекула 9102), несущие 5'-фосфорильные группы (т.е. 5'-P) и 3'-гидроксильные группы (то есть 3'-OH), были подвергнуты процессу концевой блокады, в результате которого 5'-P был удален и 3'-OH был лигирован с дидезоксинуклеотидами (т.е. ддНТФ). Следовательно, полученные молекулы (например, молекула 9104), концы которых были модифицированы, не могли быть лигированы с адаптерами для последующего приготовления библиотеки ДНК. Однако, молекулы с заблокированными концами подвергались целенаправленному расщеплению, опосредованному системой CRISPR/Cas9, внося концы 5'-P и 3'-OH в представляющие интерес молекулы. Такие вновь расщепленные молекулы ДНК (например, молекула 9106), несущие 5'-P и 3'-OH-концы, обрели способность лигироваться с адаптерами в виде шпильки с образованием кольцевой молекулы 9108. Нелигированные адаптеры, линейную ДНК и молекулы, несущие только один сайт расщепления, подвергали расщеплению экзонуклеазами III и VII. В результате молекулы, лигированные с двумя адаптерами в виде шпильки, были обогащены и подвергнуты секвенированию отдельной молекулы в реальном времени. Эти целевые молекулы подходили для анализа модификации оснований согласно вариантам осуществления, представленным в данном раскрытии изобретения (т.е. целевое секвенирование отдельной молекулы в реальном времени).

Как показано на фиг. 92, белок Cas9 в системе CRISPR/Cas9 взаимодействовал с направляющей РНК (т.е. нРНК), которая включает в себя РНК CRISPR - (крРНК, отвечающую за нацеливание на ДНК) и транс-активирующую крРНК (тракрРНК, отвечающую за формирование комплекса с Cas9) (Pickar-Oliver et al. Nat Rev Mol Cell biol. 2019;20:490-507). Изогнутая фигура представляет белок Cas9, который является ферментом, который использует последовательности CRISPR в качестве ориентира для распознавания и разрезания определенных цепей ДНК, которые комплементарны одной части последовательностей CRISPR. крРНК отжигалсь с тракрРНК. В одном варианте осуществления, синтетическая единая последовательность РНК содержала последовательности как крРНК, так и тракрРНК, называемая единой направляющей РНК (онРНК). Сегмент крРНК, называемый спейсерной последовательностью, будет направлять белок Cas9 в распознавании и разрезании определенных цепей двухцепочечной ДНК (дцДНК) посредством комплементарного спаривания оснований с целевой областью. В одном варианте осуществления, не существовало ошибочных спариваний, вовлеченных в комплементарность между последовательностью спейсера и целевой дцДНК. В другом варианте осуществления, комплементарное спаривание оснований между спейсерной последовательностью и целевой дцДНК может допускать ошибочные спаривания. Например, количество ошибочных спариваний составляет, но не ограничивается, 1, 2, 3, 4, 5, 6, 7, 8 и т.д. В одном варианте осуществления, последовательности CRISPR могут программироваться, в зависимости от эффективности разрезания, специфичности, чувствительности и способности к мультиплексированию для различных сложных конструкций CRISPR/Cas.

Как показано на фиг. 93, мы разработали пару комплексов CRISPR/Cas9, нацеленных на два разреза, охватывающих элемент Alu в геноме человека. "XXX" обозначает три нуклеотида, фланкирующие сайт разрезания нуклеазы Cas9. "YYY" обозначает три соответствующих нуклеотида, комплементарных "XXX". 5'-NGG представляет собой последовательность смежного мотива протоспейсера (PAM). В других системах CRISPR/Cas последовательность PAM может быть другой, и последовательности, фланкирующие сайт разрезания нуклеазы Cas, могут быть разными. На данной фигуре размер области Alu составляет 223 п. о. Насчитывалось 1175329 областей Alu, каждая из которых содержала гомологи такого элемента Alu в геноме человека. Медианное значение сайтов CpG, находящихся в данном элементе Alu, составляло 5 (диапазон: 0-34). В качестве примера, эта конструкция содержала крРНК из 36 нт, которая содержала спейсерную последовательность из 20 нт. Подробная информация о последовательности нРНК показана ниже:

Первый комплекс CRISPR/Cas9 для внесения первого разреза: (все последовательности от 5 'до 3').

крРНК:GCCUGUAAUCCCAGCACUUUGUUUUAGAGCUAUGCU.

тракрРНК:

AGCAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACC GAGUCGGUGCUUU.

Второй комплекс CRISPR/Cas9 для внесения второго разреза:

крРНК: AGGGUCUCGCUCUGUCGCCCGUUUUAGAGCUAUGCU.

тракрРНК:

AGCAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACC GAGUCGGUGCUUU.

Молекулы крРНК отжигали с тракрРНК (например, 67 нт), чтобы сформировать основу нРНК. Нуклеаза Cas9 с сконструированной нРНК может расщеплять обе цепи молекул с блокированными концами, несущих целевые сайты разрезания, с определенным уровнем специфичности. В геноме человека было 116184 представляющих интерес областей Alu, которые, как предполагалось, будут разрезаться разработанными комплексами CRISPR/Cas9. Следовательно, эти области Alu после целенаправленного разрезания комплексами Cas9 могут быть лигированы с адаптерами в виде шпильки. Эти молекулы, лигирован-

ные с адаптерами в виде шпильки, можно секвенировать с помощью секвенирования отдельной молекулы в реальном времени. Для этих областей Alu целенаправленно могут быть определены паттерны метилирования. В одном варианте осуществления, спейсерные последовательности из двух комплексов Cas9 могут быть спарены по основаниям с той же цепью (например, цепью Уотсона или цепью Крика) двухцепочечного ДНК-субстрата. В одном варианте осуществления, спейсерные последовательности в нРНК из двух комплексов Cas9 могут быть спарены по основаниям с разными цепями двухцепочечного ДНК-субстрата. Например, одна спейсерная последовательность в комплексе Cas9 была комплементарна цепи Уотсона двухцепочечного ДНК-субстрата, а другая спейсерная последовательность в комплексе Cas9 была комплементарна цепи Крика двухцепочечного ДНК-субстрата, или наоборот.

В одном варианте осуществления, молекулы ДНК, лигированные с адаптерами в виде шпильки, имели кольцевую форму, которая была устойчивой к расщеплению экзонуклеазами. Следовательно, можно обработать экзонуклеазой продукт ДНК с лигированными адаптерами (например, экзонуклеазой III и VII) для удаления линейной ДНК (например, нецелевых молекул ДНК). Этот шаг с использованием экзонуклеаз может дополнительно обогатить целевые молекулы. Размеры целевых молекул, подлежащих секвенированию, зависели от размера протяженности между двумя сайтами разрезания, внесенными одной или большим количеством нуклеаз Cas9, например, включающие, но не ограниченные лишь этими: 10, 20, 30, 40, 50, 100, 200, 300, 400, 500, 1000, 2000, 3000, 4000, 5000 п.о., 10, 20, 30, 40, 50, 100, 200, 300, 500 т.п.о., и 1 млн.п.о.

В качестве примера, используя Cas9 с нРНК, нацеленной на области Alu, мы секвенировали 187010 молекул из образца опухолевой ткани гепатоцеллюлярной карциномы человека (ГЦК), используя секвенирование отдельной молекулы в реальном времени. Среди них 113491 молекула несли целевые разрезы (т.е. доля целевого расщепления составила около 60,7% молекул). Набор данных был сгенерирован из ДНК, полученной с помощью Sequel II Sequencing Kit 2.0. Другими словами, специфичность сайтов разрезания, внесенных в молекулы интереса с помощью комплексов Cas9, в данном примере составила 60,7%. В других вариантах осуществления, специфичность сайтов разрезания, вносимых в молекулы интереса с помощью Cas9 или других комплексов Cas, может варьировать, включая, но без ограничения: 1, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, и 100%. Значения МИП, ШИ и контекст последовательности, полученные из ККП и субпрочтений без выравнивания с эталонным геномом, использовали для определения статуса метилирования в сайтах CpG в последовательностях Alu.

Как показано на фиг. 94, мы наблюдали аналогичное распределение метилирования между уровнями метилирования, определенными с помощью бисульфитного секвенирования, и секвенирования отдельной молекулы в реальном времени согласно данному изобретению. Фиг. 94 демонстрирует гистограммы плотностей метилирования (в процентах) для бисульфитного секвенирования и секвенирования отдельной молекулы в реальном времени (Pacific Biosciences). Ось ординат показывает долю молекул в образце с конкретной плотностью метилирования, показанной на оси абсцис. Этот результат предполагает, что возможно определить паттерны метилирования с помощью Cas9-опосредованного целевого секвенирования отдельной молекулы в реальном времени. Этот результат также предполагает, что можно определить метилирование, используя ассоциированные с субпрочтениями кинетические характеристики, включая значения ШИ и МИП, без выравнивания с эталонным геномом. Как показано на фиг. 94, мы наблюдали значительное количество областей Alu, демонстрирующих гипометилирование, что согласуется с предшествующими знаниями о том, что геном рака будет деметилироваться в областях Alu-повторов (Rodriguez et al. Nucleic Acids Res. 2008; 36:770-784).

Фиг. 95 демонстрирует распределение уровней метилирования, как определенно с помощью секвенирования отдельной молекулы в реальном времени согласно данному изобретению, по оси ординат, и плотность метилирования, как определено с помощью бисульфитного секвенирования, по оси абсцисс. Как показано на фиг. 95, уровни метилирования в областях Alu были разделены на 5 категорий, а именно, 0-20%, 20-40%, 40-60%, 60-80% и 80-100% в соответствии с результатами бисульфитного секвенирования. Уровни метилирования одного и того же набора областей Alu были дополнительно определены с помощью нашей модели с использованием окон измерения, включающих в себя кинетические характеристики и контекст последовательности (ось ординат) для каждой категории Alu-регионов. Рассеянье уровней метилирования, определенное с помощью нашей модели, постепенно увеличивалось в соответствии с возрастающим порядком уровней метилирования в сгруппированных категориях. И снова, данные результаты предполагают, что возможно определить паттерны метилирования с помощью Cas9-опосредованного целевого секвенирования отдельной молекулы в реальном времени. Можно определить метилирование, используя ассоциированные с субпрочтениями кинетические характеристики, включая значения ШИ и МИП, без выравнивания с эталонным геномом.

В еще одном варианте осуществления, можно использовать другие типы систем CRISPR/Cas, например, но без ограничения, Cas12a, Cas3 и другие ортологи (например, Staphylococcus aureus Cas9) или сконструированные белки Cas (усиленные Acidaminococcus spp Cas12a) для выполнения целевого секвенирования отдельной молекулы в реальном времени.

В одном варианте осуществления, можно использовать деактивированный Cas9 (dCas9) без нуклеазной активности для обогащения целевых молекул без расщепления. Например, целевые молекулы ДНК

были связаны комплексом, содержащим биотинилированный dCas9 и специфичные к целевой последовательности нРНК. Такие целевые молекулы ДНК могут быть не разрезаны dCas9, поскольку dCas9 не обладал нуклеазной активностью. За счет использования магнитных гранул, покрытых стрептавидином, целевые молекулы ДНК могут быть обогащены.

В одном варианте осуществления, можно использовать экзонуклеазы для расщепления смеси ДНК после инкубации с белками Cas. Экзонуклеазы могут разрушать молекулы ДНК, не связанные с белками Cas, в то время как экзонуклеазы не могут разрушать или могут быть в значительной степени менее эффективными в разрушении молекул ДНК, связанных с белками Cas. Следовательно, информация о целевых молекулах, связанных с белками Cas, может быть дополнительно пополнена в конечных результатах секвенирования.

Фиг. 96 демонстрирует таблицу тканей и уровней метилирования областей Alu в тканях. Многие ткани демонстрируют уровни метилирования в диапазоне 85-92%, в том числе в диапазоне от 88% до 92%. Опухолевая ткань ГЦК и ткань плаценты показали уровни метилирования ниже 80%. Как видно на фиг. 96, было показано, что опухоль ГЦК часто гипометилируется в областях Alu, на которые были нацелены наши конструкции. Следовательно, определение метилирования областей Alu, представленное в данном раскрытии изобретения, можно использовать для обнаружения, определения стадии и мониторинга рака во время прогрессирования опухоли или лечения с использованием ДНК, выделенной из биопсий опухоли, или других тканей или клеток.

Гипометилирование тканей плаценты в областях Alu может быть использовано для проведения неинвазивного пренатального тестирования с использованием ДНК плазмы беременных женщин. Например, более высокая степень гипометилирования может указывать на более высокую фракцию ДНК плода у беременной женщины. В другом примере, если женщина беременна плодом с хромосомной анеуплоидией, количество фрагментов Alu, происходящих из затронутой хромосомы, обнаруженных с помощью данного подхода, может быть количественно другим (т.е. либо увеличиваться, либо уменьшаться), чем у женщин, беременных эуплоидными плодами. Следовательно, если у плода есть трисомия 21, то количество фрагментов Alu, происходящих из хромосомы 21, обнаруженных с помощью данного подхода, может быть увеличено по сравнению с женщинами, беременными эуплоидными плодами. С другой стороны, если у плода есть моносомная хромосома, то количество фрагментов Alu, происходящих из этой хромосомы, обнаруженных с помощью данного подхода, может быть уменьшено по сравнению с женщинами, беременными эуплоидными плодами. По сравнению с незатронутыми хромосомами, определение проявления дополнительного гипометилирования затронутой хромосомы (13, 18 или 21) в плазме может использоваться в качестве молекулярного индикатора для дифференциации женщин, беременных нормальным и аномальным плодами.

3. Анализ метилирования в областях Alu, на которые нацелен комплекс Cas9, для различных типов рака.

Несмотря на то, что повторы Alu, которые были для нас мишенями, были сильно метилированы в разных тканях, мы предположили, что разные типы рака будут иметь разные паттерны деметилирования для данных повторов Alu. В одном варианте осуществления, можно использовать Cas9 опосредованное целевое секвенирование отдельной молекулы в реальном времени для анализа паттернов метилирования с целью определения различных типов рака согласно данному изобретению.

Фиг. 97 демонстрирует кластерный анализ сигналов метилирования, относящихся к повторам Alu, для различных типов рака. Субъекты с раком из базы данных TCGA (www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga) имели статус метилирования в сайтах CpG, проанализированных с использованием технологии микрочипов (Infinium HumanMethylation450 BeadChip, Illumina Inc). Были проанализированы статусы метилирования 3024 сайтов CpG, присутствующих на микрочипе и перекрывающихся с областями Alu, на которые были нацелены комплексы CRISPR/Cas9. Существует ряд CpG, происходящих из областей Alu интереса у пациента. Уровень метилирования каждого CpG определяли количественно с помощью микрочипа (также называемый индексом метилирования или бета-величиной). Мы выполнили иерархический кластерный анализ на основе набора уровней метилирования в тех сайтах CpG у пациентов. Следовательно, пациенты со сходным паттерном уровней метилирования в этих сайтах CpG будут группироваться вместе, образуя кладу. Сходство паттернов метилирования у разных пациентов будет обозначаться значениями высоты в дендрограмме кластеризации. В этом примере высота была рассчитана в соответствии с евклидовыми расстояниями. В других вариантах осуществления, могут использоваться другие метрики расстояния, включающие в себя, но не ограничивающиеся лишь этими: Минковского, Чебичева, Махаланобиса, Манхэттена, Отиаи, корреляции, Спирмена, Хэмминга, Жаккара и т.д. Высота, используемая в данном документе, представляет собой значение метрики расстояния между кластерами, отражающее взаимосвязь между кластерами. Например, если наблюдали, что два кластера слились на высоте x, это предполагает, что расстояние между этими кластерами было x (например, среднее расстояние между всеми межкластерными пациентами).

С использованием статусов метилирования в сайтах CpG пациенты были разгруппированы в различные отдельные группы в зависимости от типа рака по результатам кластерного анализа. Типы рака включали в себя уротелиальную карциному мочевого пузыря (BLCA), инвазивную карциному молочной

железы (BRCA), серозную цистаденокарциному яичников (OV), аденокарциному поджелудочной железы (PAAD), HCC (ГНК), аденокарциному легких (LUAD), аденокарциному желудка (STAD), накожную меланому кожи (SKCM), и карциносаркому матки (UCS). Число после типа рака на фигуре обозначает пациента. Следовательно, кластеризация предполагает, что сигналы метилирования в выбранных нами повторах Alu были информативными для классификации типов рака, включая типы рака, не показанные на фиг. 97. В одном варианте осуществления, можно дифференцировать первичные и вторичные опухоли на основе паттернов метилирования при биопсии ткани.

4. Пороговые значения глубины и размера субпрочтения.

В этом разделе показано, что пороговые значения глубины и/или размера субпрочтения могут использоваться для повышения точности и/или эффективности обнаружения метилирования. Приготовление библиотеки может быть изменено для тестирования определенных глубин или размеров субпрочтений.

На основе Sequel II Sequencing Kit 2.0 мы проанализировали влияние глубины прочтения на количественную оценку совокупного уровня метилирования в тестовых наборах данных, которые были сгенерированные из образцов после амплификации всего генома или обработки M.SsssI. Мы изучали геномные сайты, которые были охвачены субпрочтениями, с по меньшей мере определенным пороговым значением, например, но не ограниченным лишь этими: ≥1x, 10x, 20x, 30x, 40x, 50x, 60x, 70x, 80x, 90x, 100x и т.д.

Фиг. 98A демонстрирует влияние глубины прочтения на количественную оценку совокупного уровня метилирования в тестовых наборах данных, которые были получены с использованием амплификации всего генома. Фиг. 98B демонстрирует влияние глубины считывания на количественную оценку совокупного уровня метилирования в тестовых наборах данных, которые были получены с использованием обработки M.SsssI. На оси ординат показан совокупный уровень метилирования в процентах. По оси абсцисс показана глубина субпрочтений. Пунктирными линиями показаны ожидаемые значения совокупных уровней метилирования.

Как показано на фиг. 98A, для набора данных, с вовлечением амплификации всего генома, совокупное метилирование снижалось для начальных нескольких пороговых значений, таких как, но не ограничиваясь, 1x, 10x, 20x, 40x, 50x, в диапазоне от 5,7 до 5,2%. Уровни метилирования постепенно стабилизировались на уровне около 5% при пороговом значении 50x или выше.

С другой стороны, на фиг. 98B, для набора данных, сгенерированного из образцов после обработки M.SsssI, совокупное метилирование увеличивалось для начальных нескольких пороговых значениях, таких как, но не ограничиваясь, 1x, 10x, 20x, 40x, 50x, в диапазоне от 70 до 83%. Уровни метилирования постепенно стабилизировались на примерно 83% при пороговом значении 50x или выше.

В одном варианте осуществления, можно было бы отрегулировать пороговые значения глубины субпрочтений, сделав выполнение анализа модификации оснований подходящим для различных применений. В других вариантах осуществления, можно было бы использовать менее жесткое ограничение глубины субпрочтений для получения большего количества ZMW (т.е. количества молекул), которые подходят для последующего анализа. В еще одном варианте осуществления, можно откалибровать считывание уровней метилирования, определенных с помощью SMRT-секв. согласно данному изобретению для второго способа измерения, например, но не ограничиваясь, БС-секв., цифровой капельной ПЦР (на образцах, конвертированных бисульфитом), специфичной к метилированию ПЦР, или антителами или другими белками, связывающимися с метилированными цитозинами. В другом варианте осуществления, второе измеренение может быть реализовано путем анализа молекул ДНК, после амплификации всего генома, сохранившей 5mC, с помощью БС-секв., цифровой капельной ПЦР (на образцах, конвертированных бисульфитом), специфичной к метилированию ПЦР, или геномного секвенирования с обогащением посредством метил-CpG связывающего белка. Например, амплификация всего генома с сохранением 5mC может быть опосредована ДНК-примазой TthPrimPol, полимеразой phi29 и DNMT1 (ДНК-метилтрансфераза 1).

Мы проанализировали уровни метилирования в различных типах рака и неопухолевых тканях для разной глубины субпрочтений. Уровни метилирования, определенные с помощью SMRT-секв. согласно данному изобретению, также сравнивали с результатами секвенирования БС-секв. Используя Sequel II Sequencing Kit 2.0, мы получили медианное значение, составляющее 43 миллиона субпрочтений (межквартильный диапазон (МКД): 30-52 млн.), что позволило сгенерировать 4,6 млн. (медианное значение) кольцевых консенсусных последовательностей (ККП), которые были выровнены с эталонный геном человека (МКД: 2,8 - 5,8 млн.). Среди этих образцов 22 образца также были подвергнуты хорошо зарекомендовавшему себя масштабному параллельному бисульфитному секвенированию (БС-секв.) для определения паттернов метилирования, обеспечивающего второе измерение для сравнения уровней метилирования.

Фиг. 99 демонстрирует сравнение между совокупными уровнями метилирования, определенными с помощью SMRT-секв. (Sequel II Sequencing Kit 2.0) согласно данному изобретению, и БС-секв. с использованием различных пороговых значений глубины субпрочтений. Уровень метилирования в процентах, определенный с помощью SMRT-секв., показан на оси ординат. Уровень метилирования в процентах,

определенный с помощью бисульфитного секвенирования, отложен по оси абсцисс. Символы указывают на разную глубину субпрочтений, составляющую 1х, 10х и 30х. Три диагональные линии показывают подогнанные линии для разной глубины суброчтений.

Фиг. 99 продемонстрировала, что уровни метилирования по сайтах CpG, определенные с помощью SMRT-секв. согласно данному изобретению, хорошо коррелировали с (r=0,8; значение P<0,0001) уровнями, определенными с помощью БС-секв., при анализе геномных сайтов, которые хотя бы один раз покрывались субпрочтениями (т.е. пороговое значение глубины субпрочтений ≥1х). Эти результаты предполагают, что варианты осуществления, представленные в данном раскрытии изобретения, могут быть использованы для измерения уровней метилирования для различных типов тканей, включающих в себя, но не ограничивающихся лишь этими: колоректальный рак, колоректальные ткани, рак пищевода, ткани пищевода, рак молочной железы, нераковые ткани молочной железы, почечно-клеточную карциному, ткани почек, рак легкого и ткани легкого. Мы также заметили, что корреляция между этими двумя измерениями улучшилась до 0,87 (значение P<0,0001) и 0,95 (значение P<0,0001), по мере того как пороговые значения глубины субпрочтений были увеличены до 10х и 30х, соответственно. В некоторых вариантах осуществления, увеличение глубины субпрочтений или выбор геномных областей с охватом большим количеством субпрочтений улучшит эффективность определения метилирования на основе SMRT-секв. согласно данному изобретению.

Фиг. 100 представляет собой таблицу, демонстрирующую влияние глубины субпрочтений на корреляцию уровней метилирования между двумя измерениями с помощью SMRT-секв. (Sequel II Sequencing Kit 2.0) и БС-секв. В первом столбце показано пороговое значение глубины субпрочтений. Во втором столбце показано r Пирсона -коэффициент корреляции. В третьем столбце показано количество сайтов CpG, связанных с пороговым значением, с диапазоном количества сайтов в скобках.

Как показано на фиг. 100, корреляция уровней метилирования между двумя измерениями с помощью SMRT-секв. и БС-секв. варьировалась в соответствии с разными пороговыми значениями глубины субпрочтений. В одном варианте осуществления, можно использовать взаимосвязь между пороговыми значениями глубины субпрочтений и коэффициентами корреляции (например, коэффициентом корреляции Пирсона) между двумя измерениями для определения оптимального порогового значения глубины субпрочтений для дифференциации метилированных цитозинов и неметилированных цитозинов. Фиг. 100 демонстрирует, что при пороговом значении глубины субпрочтений 30х (т.е. ≥30х) уровни метилирования, определенные с помощью SMRT-секв. согласно данному изобретению, давали самую высокую корреляцию с результатами, полученными с помощью БС-секв. (r Пирсона=0,952). В других вариантах осуществления, можно использовать, но не без ограничения, пороговые значения глубины субпрочтений, составляющие 1х, 10х, 30х, 40х, 50х, 60х, 70х, 80х, 900х, 100х, 200х, 300х, 400х, 500х, 600х, 700х, 800х и т.д.

Количество CpG-сайтов, используемых для анализа метилирования, уменьшается с увеличением порогового значения глубины субпрочтений, как показано на фиг. 100. При пороговом значении глубины субпрочтений, составляющем 100х, наблюдали более низкую корреляцию (r Пирсона=0,875) между двумя измерениями уровней метилирования, по сравнению с пороговым значением глубины субпрочтений 30х (r Пирсона=0,952). Более низкая корреляция для более высокого порогового значения субпрочтений может быть связана с меньшим количеством сайтов CpG, которые соответствовали более строгим пороговым значениям глубины субпрочтений. В одном варианте осуществления, можно рассмотреть компромисс между требованием по глубине субпрочтений и количеством молекул, которые можно использовать для анализа метилирования. Например, если требуется просканировать весь геном на предмет паттернов метилирования, может потребоваться больше молекул. Если целью была выбрана конкретная область с использованием целевого SMRT-секв., для получения паттернов метилирования для этой области может быть желательной более высокая глубина субпрочтений.

Фиг. 101 демонстрирует распределение глубины субпрочтений относительно размеров фрагментов в данных, сгенерированных с помощью Sequel II Sequencing Kit 2.0. Глубина субпрочтений отображается на оси ординат, а длина молекулы ДНК - на оси абсцисс. Длины молекул ДНК были выведены из размера кольцевых консенсусных последовательностей (ККП).

Так как глубина субпрочтений может влиять на эффективность определения метилирования с использованием данных SMRT-секв., и глубина субпрочтений является функцией длины секвенируемой молекулы ДНК, то размеры молекул ДНК могут иметь решающее значение для получения оптимальной глубины субпрочтений для анализа паттернов метилирования в образце. Как показано на фиг. 101, чем длиннее ДНК, тем меньше глубина субпрочтений. Например, для популяции молекул размером 1 т.п.н. средняя глубина субпрочтений составляла 50х. Для популяции молекул размером 10 т.п.н. средняя глубина субпрочтений составляла 15х.

В одном варианте осуществления, как показано на фиг. 100, оптимальное пороговое значение глубины субпрочтений может составлять по меньшей мере 30х, что дает самый высокий коэффициент корреляции. Для дополнительного улучшения пропускной способности для молекул, которые бы соответствовали оптимальному пороговому значению глубины субпрочтений, составляющему 30х, можно исполь-

зовать взаимосвязь между глубиной субпрочтений и длиной молекул-матриц ДНК. Например, на фиг. 101, 30х - это медианная глубина субпрочтений для молекул длиной около 4 т.п.н. Таким образом, можно фракционировать молекулы ДНК с размером 4 т.п.н. перед приготовлением библиотеки SMRT-секв. и ограничить секвенирование молекулами ДНК с размером 4 т.п.н. В других вариантах осуществления, могут использоваться другие пороговые значения размера для фракционирования молекул ДНК, включающие в себя, но не ограничивающиеся лишь этими: 100, 200, 300, 400, 500, 600, 700, 800 п.о., 900 п.н., 1, 2, 3, 4, 5, 6, 7, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 500 т.п.н., 1 млн.п.о. или различные комбинации пороговых значений размера.

5. Целевове секвенирование отдельной молекулы в реальном времени на основе фермента рестрикции.

В данном разделе описывается использование рестрикционных ферментов для улучшения осуществимости и/или производительности и/или рентабильности обнаружения модификаций. Фрагменты ДНК, сгенерированные с помощью рестрикционных ферментов, можно использовать для определения происхождения образца.

а) Использование ферментов рестрикции для расщепления молекул ДНК.

В вариантах осуществления, может использоваться один или большее количество ферментов рестрикции для расщепления молекул ДНК перед секвенированием отдельной молекулы в реальном времени (например, с использованием системы Pacific Biosciences). Поскольку распределение сайтов распознавания ферментов рестрикции будет неравномерным в геноме человека, ДНК, расщепленная ферментами рестрикции, может давать смещенное распределение по размерам. Геномные области с большим количеством сайтов распознавания ферментов рестрикции могут быть расщеплены на более мелкие фрагменты, тогда как геномные области с меньшим количеством сайтов распознавания ферментов рестрикции могут быть расщеплены на более длинные фрагменты. В вариантах осуществления, в соответствии с диапазонами размеров, можно выборочно получать молекулы ДНК, происходящие из одной или большего количества областей, которые имеют аналогичные схемы разрезания одного или большего количества ферментов рестрикции. Желательные диапазоны размеров для выбора размера могут быть определены с помощью разрезания in silico для одного или большего количества рестрикционных ферментов. Можно использовать компьютерную программу для определения количества сайтов распознавания рестрикционных ферментов интереса в эталонном геноме (например, эталонном геноме человека). Такой эталонный геном был разрезан in silico на фрагменты в соответствии с такими сайтами распознавания, для которых предоставлена информация о размерах для геномных областей интереса.

Фиг. 126 демонстрирует способ целевого секвенирования отдельной молекулы в реальном времени с обработкой MspI, с применением восстановления концов ДНК и с добавлением А-хвоста. В вариантах осуществления, показанных на фиг. 126, можно использовать MspI, который распознает сайты 5'C^CGG3', для расщепления образца ДНК организма, например, но без ограничения, образца ДНК человека. Расщепленные фрагменты ДНК с выступами 5 'CG подвергали отбору по размеру, обогащая молекулы ДНК, происходящие из островков CpG. Геномные области, которые обогащают по остаткам G и C (также называемые содержанием GC), могут генерировать более короткие фрагменты. Таким образом, можно определить диапазон размеров фрагментов для выполнения отбора на основе содержания GC в интересующих областях. Специалистам в данной области техники доступны различные инструменты отбора фрагментов ДНК по размеру, которые включают в себя, помимо прочего, гель-электрофорез, электрофорез исключения по размеру, капиллярный электрофорез, хроматографию, масс-спектрометрию, фильтрационные подходы, подходы на основе осаждения, микрогидродинамические подходы и наногидродинамические подходы. Фракционированные по размеру молекулы ДНК подвергали репарации концов ДНК и формированию А-хвоста таким образом, чтобы желаемый продукт ДНК можно было лигировать с адаптерами в виде шпильки, которые несли 5' Т-выступ, с формированием кольцевых ДНК-матриц.

После удаления нелигированных адаптеров, линейной ДНК и не полностью кольцевой ДНК, например, но без ограничения, с использованием экзонуклеаз (например, экзонуклеазы III и VII), молекулы ДНК, лигированные с адаптерами в виде шпильки, могут быть использованы для секвенирования отдельной молекулы в реальном времени з целью определения МИП, ШИ и контекста последовательности при определении профилей метилирования, как раскрыто в данном документе. Путем анализа геномных областей, обогащенных по CpG, ДНК, полученная из различных тканей, или тканей с различными заболеваниями и/или физиологическими патологиями, или биологических образцов, может быть дифференциирована и классифицирована по ее профилю метилирования, определенному с помощью способов анализа данных секвенирования данного изобретения.

Для этапа, включающего в себя выбор размера на фиг. 126, в вариантах осуществления, диапазоны желаемых размеров могут быть определены с помощью анализа разрезания MspI in silico. Мы определили в совокупности 2286541 сайтов разрезания MspI у человека. Эталонный геном человека был разрезан in silico на фрагменты в соответствии с этими сайтами разрезания MspI. Всего нами было получено 2286565 фрагментов. Размер каждого отдельного фрагмента определи с помощью совокупного числа нуклеотидов данного фрагмента.

Фиг. 127A и 127B демонстрируют распределение по размеру фрагментов, расщепленных MspI. По оси ординат на этих фигурах отложена частота в процентах для конкретного размера фрагмента. Фиг. 127A имеет логарифмическую шкалу для оси абсцисс в диапазоне от 50 до 500000 п.о. Фиг. 127B имеет линейную шкалу по оси абсцисс от 50 до 1000 п.н.

Как показано на фиг. 127A и 127B, молекулы ДНК, расщепленные MspI, имеют смещенное распределение по размерам. Медианный размер фрагментов, расщепленных MspI, составлял 404 п.о. (МКД: 98-1411 п.о.). Около 53% этих фрагментов, расщепленных MspI, были меньше 1 т.п.о. В профиле размера имелась серия острых пиков, причиной которых могли служить повторяющиеся элементы. Определенные повторяющиеся элементы могут иметь сходные паттерны сайтов разрезания MspI, что дает набор молекул, полученных в результате расщепления MspI, которые обладают сходными размерами фрагментов. Например, пик с самой высокой частотой (т.е., в совокупности 49079) соответствовал размеру 64 п.о. Из них 45894 (94%) перекрывались с повторами Alu. Можно отобрать молекулы ДНК размером 64 п.о. для обогащения молекул ДНК, происходящих из повторов Alu. Данные предполагают, что выбор размера может быть использован для обогащения желаемых молекул ДНК для последующего анализа метилирования согласно данному изобретению.

Фиг. 128 демонстрирует таблицу с количеством молекул ДНК для определенных выбранных диапазонов размеров. Первый столбец показывает диапазоны размеров в парах оснований. Второй столбец показывает процент молекул в пределах диапазона размеров по отношению к совокупному количеству фрагментов. Третий столбец показывает количество молекул в пределах диапазона размеров, перекрывающихся с островками CpG. Четвертый столбец показывает процент молекул в пределах диапазона размеров, перекрывающихся с островками CpG. Пятый столбец показывает количество секвенированных сайтов CpG. Шестой столбец показывает количество сайтов CpG, попадающих в пределы островков CpG. Седьмой столбец показывает процент сайтов CpG, которые подпадали под отбор по размеру и которые попадают в пределы островков CpG. Как показано на фиг. 128, количество молекул ДНК, генерируемых из генома человека, подвергнутого расщеплению MspI, варьировалось в зависимости от рассматриваемых диапазонов размеров. Число молекул ДНК, перекрывающихся с островками CpG, варьировалось в зависимости от диапазонов размеров.

Поскольку мотив CCGG присутствует преимущественно в островках CpG, отбор молекул с размером меньшим, чем определенное пороговое значение, может сделать возможным обогащение молекул ДНК, происходящих из островков CpG. Например, для диапазона размеров от 50 до 200 п.о. количество молекул составляло 526543, что составляло 23,03% от всех фрагментов ДНК, полученных из генома человека, подвергнутого расщеплению MspI. Среди 526543 молекул ДНК 104079 (19,76%) перекрывались с островками CpG. Для диапазона размеров от 600 до 800 п.о. количество молекул составляло 133927, что составляло 5,86% всех фрагментов ДНК, полученных из генома человека, подвергнутого расщеплению MspI. Из 133927 молекул 3673 (2,74%) молекулы перекрывалось с островками CpG. В качестве примера можно выбрать размер от 50 до 200 п.о. для обогащения по фрагментам ДНК, происходящим из CpG-островков.

Чтобы рассчитать степень обогащения сайтов CpG, перекрывающихся с островками CpG, с помощью целевого секвенирования отдельной молекулы в реальном времени с обработкой MspI, мы выполнили моделирование для ДНК, фрагментированной ультразвуком, и мы смоделировали 526543 фрагментов, сгенерированных из ZMW, с медианным размером 200 п.о. и стандартным отклонением 20 п.о. на основе нормального распределения. Только 0,88% молекул ДНК перекрывались с островками CpG. Всего 71495 сайтов CpG перекрывались с островками CpG. Как показано на фиг. 128, отбор MspI-расщепленных фрагментов в диапазоне от 50 до 200 п.о. приведет к тому, что 19,8% фрагментов будут перекрываться с островками CpG. Таким образом, эти данные позволяют предположить, что ДНК, полученная путем расщепления MspI, может иметь в 22,5 раза больше фрагментов ДНК, происходящих из островков CpG, по сравнению с ДНК, полученной путем обработки ультразвуком. Кроме того, мы проанализировали сайты CpG, обогащенные в островках CpG путем расщепления MspI. Отбор расщепленных MspI фрагментов в диапазоне от 50 до 200 п.о. может привести к появлению 885041 сайтов CpG, перекрывающихся с островками CpG, что составляет 37,5% от совокупного количества сайтов CpG из секвенированных фрагментов в этом диапазоне размеров. Наблюдали 12,3-кратное (т.е. 885041/71495) обогащение сайтов CpG, перекрывающихся с островками CpG, по сравнению с ДНК, полученной путем обработки ультразвуком. На основании информации, показанной на фиг. 128, можно выбрать подходящий диапазон размеров, включающий в себя желаемое количество сайтов CpG и желаемое кратное обогащение сайтов CpG в пределах островков CpG.

Фиг. 129 представляет собой график зависимости процента покрытия CpG-сайтов в пределах CpG-островков от размера фрагментов ДНК после расщепления ферментом рестрикции. По оси ординат показан процент сайтов CpG в пределах островков CpG, покрытых фрагментами с заданными размерами. По оси абсцисс показан верхний предел диапазона размеров фрагментов ДНК после расщепления ферментом рестрикции. Фиг. 129 продемонстрировала процент сайтов CpG в пределах островков CpG, который будет покрываться за счет расширения диапазона отбора по размеру. На Фиг. 129 диапазон размеров составляет от 50 п.о. до размера, показанного на оси абсцисс. В других вариантах осуществления, нижний

предел диапазона размеров может быть адаптирован, например, но без ограничения, до 60, 70, 80, 90, 100, 200, 300, 400 и 500 п.о. С расширением диапазона размеров за счет увеличения верхнего предела размера, мы можем наблюдать, что процентное покрытие сайтов CpG в пределах островков CpG постепенно увеличивается и достигает плато на 65%. Некоторые из сайтов CpG не охвачены, потому что они находятся в фрагментах ДНК с размером меньше чем 50 п.о., или они находятся в фрагментах в пределах очень длинных молекул (например, >100000 п.о.).

В некоторых вариантах осуществления, образец ДНК можно анализировать с использованием двух или большего количества различных ферментов рестрикции (с разными сайтами рестрикции), чтобы увеличить охват сайтов CpG в пределах островков CpG. Расщепление образца ДНК различными ферментами можно проводить в отдельных реакциях, так что в каждой реакции присутствует только один фермент рестрикции. Например, AccII, который распознает сайты CG^CG, может быть использован для предпочтительного разрезания островков CpG. В других вариантах осуществления, можно использовать другие ферменты рестрикции с динуклеотидами CG в виде части сайта распознавания. В геноме человека насчитывалось 678669 сайтов разрезания AccII. Мы выполнили разрезание in silico эталонного генома человека с использованием рестрикции AccII и получили в совокупности 678693 фрагмента. Затем мы выполнили отбор по размеру этих фрагментов in silico и рассчитали процент покрытия сайтов CpG в пределах островков CpG согласно способу, описанному выше для расщепления MspI. Мы можем наблюдать постепенное увеличение процента покрытия сайтов CpG при расширении диапазона отбора по размеру. Плато процентного покрытия наступает на около 50%. Покрытие сайтов CpG дополнительно увеличивается при объединении данных из двух экспериментов по расщеплению ферментами, а именно, по расщеплению MspI и расщеплению AccII. 80% сайтов CpG в пределах островков CpG покрываются путем отбора фрагментов ДНК с размером от 50 до 400 п.о. Этот процент выше, чем соответствующие числа для экспериментов по расщеплению любым из двух ферментов по отдельности. Покрытие можно дополнительно увеличить за счет анализа образца ДНК с использованием других ферментов рестрикции. Если образец ДНК разделен на две аликвоты, то одну аликвоту расщепляют MspI, а другую - AccII. Два образца расщепленной ДНК смешивают вместе в равном молярном соотношении и секвенируют с использованием секвенирования отдельной молекулы в реальном времени с 5 миллионами ZMW. Основываясь на анализе in silico, 83% сайтов CpG в пределах островков CpG (т.е. 1734345) будут секвенированы по меньшей мере 4 раза в контексте кольцевых консенсусных последовательностей.

Фиг. 130 демонстрирует целевое секвенирование отдельной молекулы в реальном времени с обработкой MspI без использования репарации концов ДНК и добавления A-хвоста. В вариантах осуществления, лигирование между расщепленными молекулами ДНК и адаптерами в виде шпильки может быть выполнено без процесса репарации концов ДНК и добавления A-хвоста. Можно напрямую лигировать расщепленные молекулы ДНК, несущие 5 'CG-выступы, с адаптерами в виде шпильки, несущими 5' CG-выступы, формируя кольцевую ДНК-матрицу для секвенирования отдельной молекулы в реальном времени. После очистки от нелигированных адаптеров и димеров самолигируемых адаптеров, и в некоторых вариантах осуществления после очистки от нелигированных адаптеров, линейной ДНК и неполностью кольцевой ДНК, молекулы ДНК, лигированные с помощью адаптеров в виде шпильки, могут быть пригодны для секвенирования отдельной молекулы в реальном времени для получения МИП, ШИ и контекста последовательности. Профиль метилирования отдельной молекулы может быть определен с использованием МИП, ШИ и контекста последовательности согласно данному изобретению.

Фиг. 131 демонстрирует целевое секвенирование отдельной молекулы в реальном времени с обработкой MspI с пониженной вероятностью самолигирования адаптера. Нижележащее основание цитозина обозначает основание без 5'-фосфатных групп. В некоторых вариантах осуществления, для минимизации возможности формирования самолигирующихся димеров адаптеров, которое может происходить в процессе лигирования адаптера, можно использовать дефосфорилированные адаптеры в виде шпильки для выполнения лигирования адаптера с этими молекулами ДНК, расщепленными MspI. Эти дефосфорилированные адаптеры в виде шпильки могут не формировать самолигирующиеся димеры адаптеров из-за отсутствия 5'-фосфатных групп. После лигирования продукт подвергали стадии очистки от адаптера для очистки молекул ДНК, лигированных с адаптерами в виде шпильки. Молекулы ДНК, лигированные с адаптерами в виде шпильки, которые могут нести разрывы, дополнительно подвергали фосфорилированию (например, полинуклеотидкиназой T4) и восстановлению разрыва ДНК-лигазой (например, ДНК-лигазой T4). В вариантах осуществления, можно дополнительно выполнить удаление нелигированных адаптеров, линейной ДНК и неполностью кольцевой ДНК. Молекулы ДНК, лигированные с адаптерами в виде шпильки, подходили для секвенирования отдельной молекулы в реальном времени для получения МИП, ШИ и контекста последовательности. Профиль метилирования отдельной молекулы может быть определен с использованием МИП, ШИ и контекста последовательности согласно данному изобретению.

Помимо MspI, также можно использовать другие ферменты рестрикции, такие как SmaI, с сайтом распознавания CCCGGG.

В некоторых вариантах осуществления, процесс отбора по желаемому размеру может быть выполнен после стадии репарации концов ДНК. В некоторых вариантах осуществления, процесс отбора по желаемому размеру может быть выполнен после лигирования адаптеров в виде шпильки, когда было опре-

делено влияние адаптеров в виде шпильки на результат отбора по размеру. В этих и других вариантах осуществления, порядок этапов процедуры, включающей в себя целевое секвенирование отдельной молекулы в реальном времени с обработкой MspI, может изменяться в зависимости от экспериментальных условий.

В вариантах осуществления, отбор по размеру может осуществляться с использованием способов на основе гель-электрофореза и/или способов на основе магнитных гранул. В вариантах осуществления, ферменты рестрикции могут включать в себя, но не ограничиваются: BgIII, EcoRI, EcoRII, BamHI, HindIII, TaqI, NotI, HinFI, PvuII, Sau3AI, SmaI, HaeIII, HgaI, HpaII, AluI, EcoRV, EcoP15I, KpnI, PstI, SacI, SalI, ScaI, SpeI, SphI, StuI, XbaI, и их комбинации.

b) Различение типов биологических образцов по метилированию.

В этом разделе описывается использование профилей метилирования, определенных с использованием фрагментов, сгенерированных путем расщепления ферментами рестрикции, для облегчения различения отличающихся биологических образцов.

Мы оценили различия в профилях метилирования между биологическими образцами, используя профили метилирования, определенные с помощью секвенирования отдельной молекулы в реальном времени с обработкой MspI согласно вариантам осуществления, приведенным в данном раскрытии изобретения. В качестве примера мы взяли образцы ДНК плацентарной ткани и ДНК лейкоцитарного слоя. Мы выполнили компьютерное моделирование для получения данных, касающихся образца ДНК плаценты и лейкоцитарного слоя, на основе целевого секвенирования отдельной молекулы в реальном времени с обработкой MspI. Моделирование было основано на кинетических показателях, включая МИП и ШИ для каждого нуклеотида, ранее сгенерированных с помощью SMRT-секвенирования ДНК плацентарной ткани и ДНК лейкоцитарного слоя до покрытия всего генома с использованием Sequel II Sequencing Kit 1.0. Затем мы смоделировали условия, при которых образцы плацентарной ДНК и ДНК лейкоцитарного слоя подвергали расщеплению MspI с последующим отбором по размеру с применением геля, используя диапазон размеров от 50 до 200 п.о. Отобранные молекулы ДНК были лигированы с адаптерами в виде шпильки для формирования кольцевых ДНК-матриц.

Кольцевые ДНК-матрицы подвергали секвенированию отдельной молекулы в реальном времени для получения информации, касающейся МИП, ШИ и контекста последовательности.

Исходя из того, что было 500000 ZMW, генерирующих субпрочтения секвенирования SMRT, эти субпрочтения следовали геномному распределению MspI-расщепленных фрагментов в диапазоне размеров от 50 до 200 п.о., как показано в Таблице 1. Предполагалось, что глубина субпрочтений составляет 30x для образцов ДНК плаценты и лейкоцитарного слоя. Мы повторили моделирование 10 раз для образца ДНК плаценты и образца ДНК лейкоцитарного слоя, соответственно. Таким образом, набор данных, сгенерированный in silico с помощью целевого секвенирования отдельной молекулы в реальном времени, с расщеплением MspI, содержал всего 10 образцов плацентарной ДНК и 10 образцов ДНК лейкоцитарной пленки. Набор данных был дополнительно проанализировали с помощью CHC, определив профили метилирования для каждого образца согласно данному изобретению. Мы получили медианное значение сайтов CpG-9198, из островков CpG (диапазон: 5497-13928), что составило 13,6% от совокупного числа секвенированных сайтов CpG (диапазон: 45304-90762). Статус метилирования для каждого сайта CpG в каждой молекуле определяли с помощью модели CHC согласно данному изобретению.

Фиг. 132 представляет собой график совокупных уровней метилирования между образцами ДНК плаценты и лейкоцитарного слоя, определенных с помощью целевого секвенирования отдельной молекулы в реальном времени с обработкой MspI. По оси ординат отложен уровень метилирования в процентах. Тип образцов перечислен на оси абсцисс. Фиг. 132 демонстрирует, что совокупные уровни метилирования (медиана: 57,6%; диапазон: 56,9-59,1%) были ниже в образцах плаценты по сравнению с образцами лейкоцитарного слоя (медиана: 69,5%; диапазон: 68,9-70,4%) (значение $P<0,0001$, U-критерий Манна-Уитни).Эти результаты свидетельствуют о том, что профили метилирования, определенные с помощью секвенирования отдельной молекулы в реальном времени с обработкой MspI можно использовать для дифференциации образцов тканей или биологических образцов на основе их различий в метилировании. Поскольку эти данные показывают, что ДНК плаценты можно отличить от ДНК лейкоцитного слоя из-за различий в метилировании, обнаруживаемых с помощью секвенирования отдельной молекулы в реальном времени с обработкой MspI, этот способ можно применить для измерения фракции ДНК плода в материнской плазме. Фракцию ДНК плода можно определить с помощью метилирования, поскольку ДНК плода в материнской плазме или материнской сыворотке происходит из плаценты, в то время как остальные молекулы ДНК в образце в основном происходят из материнских клеток лейкоцитарного слоя. В вариантах осуществления, эта технология может быть полезным инструментом для дифференциации различных тканей или тканей с различными заболеваниями и/или физиологическими патологиями, или биологических образцов.

Чтобы выполнить кластерный анализ для образцов ДНК плаценты и образца ДНК лейкоцитарного слоя с использованием профилей метилирования островков CpG, мы рассчитали уровни метилирования ДНК островка CpG, используя долю сайтов CpG, классифицированных как метилированные, среди всех сайтов CpG данного островка CpG. Мы использовали уровни метилирования из областей островков CpG

для выполнения кластерного анализа в целях иллюстрации.

Фиг. 133 демонстрирует кластерный анализ образцов плаценты и лейкоцитарного слоя с использованием их профилей метилирования ДНК, определенных с помощью целевого секвенирования отдельной молекулы в реальном времени с обработкой MspI. Сходство паттернов метилирования из разных островков CpG среди разных пациентов будет обозначаться значениями высоты в дендрограмме кластеризации. В этом примере высоту рассчитывали в соответствии с евклидовыми расстояниями. В одном варианте осуществления, можно использовать пороговое значение 100 по высоте, чтобы рассечь дерево кластеризации на две группы, что позволяет дифференцировать образцы плаценты и лейкоцитарного слоя с 100% чувствительностью и специфичностью. В других вариантах осуществления, можно использовать другие пороговые значения высоты, включающие в себя, но не ограничивающиеся 50, 60, 70, 80, 90, 120, 130, 140, 150 и т.д. Фиг. 133 продемонстрировала, что 10 образцов ДНК плаценты и 10 образцов ДНК лейкоцитарного слоя были четко сгруппированы раздельно в две группы с использованием профилей метилирования островков CpG, определенных с помощью секвенирования отдельной молекулы в реальном времени с обработкой MspI согласно данному изобретению.

V. Способы обучения и обнаружения.

В этом разделе показаны примеры способов обучения модели машинного обучения для обнаружения модификации основания и использования модели машинного обучения для обнаружения модификации основания.

A. Обучение модели.

Фиг. 102 демонстрирует иллюстративный способ 1020 обнаружения модификации нуклеотида в молекуле нуклеиновой кислоты. Иллюстративный способ 1020 может быть способом обучения модели для обнаружения модификации. Модификация может включать в себя метилирование. Метилирование может включать в себя любое метилирование, описанное в данном документе. Модификация может иметь дискретные состояния, такие как метилировано и неметилировано, и потенциально определять тип метилирования. Таким образом, у нуклеотида может быть больше чем два состояния (групп классификаций).

На этапе 1022 принимается множество первых структур данных. В данном документе описаны различные примеры структур данных, например, на фиг. 4-16. Каждая из первых структур данных первого множества первых структур данных может соотносится с соответствующим окном нуклеотидов, секвенированных в соответствующей молекуле нуклеиновой кислоты из множества первых молекул нуклеиновой кислоты. Каждое окно, связанное с первым множеством структур данных, может включать в себя 4 или большее количество последовательных нуклеотидов, включая 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21 или больше последовательных нуклеотидов. Каждое окно может иметь одинаковое количество последовательных нуклеотидов. Окна могут перекрываться друг с другом. Каждое окно может включать в себя нуклеотиды из первой цепи первой молекулы нуклеиновой кислоты и нуклеотиды из второй цепи первой молекулы нуклеиновой кислоты. Первая структура данных может также включать в себя для каждого нуклеотида в пределах окна значение свойства цепи. Свойство цепи может указывать на наличие нуклеотида, либо первой цепи, либо второй цепи. Окно может включать в себя нуклеотиды во второй цепи, которые не комплементарны нуклеотиду в соответствующей позиции в первой цепи. В некоторых вариантах осуществления, все нуклеотиды во второй цепи комплементарны нуклеотидам в первой цепи. В некоторых вариантах осуществления, каждое окно может включать в себя нуклеотиды только из одной цепи первой молекулы нуклеиновой кислоты.

Первая молекула нуклеиновой кислоты может быть кольцевой молекулой ДНК. Кольцевая молекула ДНК может быть образована путем разрезания двухцепочечной молекулы ДНК с использованием комплекса Cas9 для формирования разрезанной двухцепочечной молекулы ДНК. Адаптер в виде шпильки может быть лигирован к концу разрезанной двухцепочечной молекулы ДНК. В вариантах осуществления, оба конца двухцепочечной молекулы ДНК могут быть разрезаны и лигированы. Например, разрезание, лигирование и последующий анализ могут выполняться, как описано на фиг. 91.

Первое множество первых структур данных может включать в себя от 5000 до 10000, от 10000 до 50000, от 50000 до 100000, от 100000 до 200000, от 200000 до 500000, от 500000 до 1000000, или 1000000, или больше первых структур данных. Множество первых молекул нуклеиновой кислоты может включать по меньшей мере 1000, 10000, 50000, 100000, 500000, 1000000, 5000000 или больше молекул нуклеиновой кислоты. В качестве дополнительного примера может быть сгенерировано по меньшей мере 10000, 50000, 100000, или 500000, или 1000000, или 5000000 прочтений последовательности.

Каждую из первых молекул нуклеиновой кислоты секвенируют путем измерения импульсов в сигнале, соответствующем нуклеотидам. Сигнал может быть сигналом флуоресценции или оптическим сигналом другого типа (например, хемилюминесцентным, фотометрическим). Сигнал может возникать из-за нуклеотидов или меток, связанных с нуклеотидами.

Модификация имеет известное первое состояние для нуклеотида в целевой позиции в каждом окне каждой первой молекулы нуклеиновой кислоты. Первое состояние может заключаться в том, что модификация отсутствует на нуклеотиде или может состоять в том, что модификация присутствует на нуклеотиде. Может быть известно, что модификация отсутствует в первых молекулах нуклеиновой кислоты,

или первые молекулы нуклеиновой кислот могут подвергаться такой обработке, при которой модификация отсутствует. Может быть известно, что модификация присутствует в первых молекулах нуклеиновой кислоты, или первые молекулы нуклеиновой кислот могут подвергаться такой обработке, при которой модификация присутствует. Если первое состояние заключается в том, что модификация отсутствует, модификация может отсутствовать в каждом окне каждой первой молекулы нуклеиновой кислоты, а не только в целевой позиции. Известные первые состояния могут включать в себя метилированное состояние для первой части первых структур данных и неметилированное состояние для второй части первых структур данных.

Целевая позиция может быть центром соответствующего окна. Для окна, охватывающего четное число нуклеотидов, целевой позицией может быть позиция сразу выше или сразу ниже центра окна. В некоторых вариантах осуществления, целевая позиция может быть в любой другой позиции соответствующего окна, включая первую позицию или последнюю позицию. Например, если окно охватывает n нуклеотидов одной цепи от 1-ой позиции до n-ой позиции (либо выше, либо ниже), целевая позиции может находиться в любой позиции от 1-ой до n-ой позиции.

Каждая первая структура данных включает в себя значения свойств в пределах окна. Свойства могут быть представлены для каждого нуклеотида в пределах окна. Свойства могут включать в себя тип нуклеотида. Тип может включать в себя основание (например, A, T, C или G). Свойства могут также включать в себя позицию нуклеотида по отношению к позиции цели в соответствующем окне. Например, позиция может представлять собой нуклеотидное расстояние относительно целевой позиции.

Позиция может быть +1, когда нуклеотид находится на расстоянии одного нуклеотида от целевой позиции в одном направлении, и позиция может быть -1, когда нуклеотид находится на расстоянии одного нуклеотида от целевой позиции в противоположном направлении.

Свойства могут включать в себя ширину импульса, соответствующего нуклеотиду. Ширина импульса может представлять собой ширину импульса, составляющую половину максимального значения импульса. Свойства могут дополнительно включать в себя межимпульсный период (МИП), представляющий время между импульсом, соответствующим нуклеотиду, и импульсом, соответствующим соседнему нуклеотиду. Межимпульсный период может представлять собой время между максимальным значением импульса, связанного с нуклеотидом, и максимальным значением импульса, связанного с соседним нуклеотидом. Соседний нуклеотид может представлять собой прилегающий нуклеотид. Свойства могут также включать в себя высоту импульса, соответствующего каждому нуклеотиду в пределах окна. Свойства могут дополнительно включать в себя значение свойства цепи, которое указывает на то, присутствует ли нуклеотид в первой или второй цепи первой молекулы нуклеиновой кислоты. Обозначение цепи может быть аналогично матрице, показанной на фиг. 6.

Каждая структура данных из множества первых структур данных может исключать первые молекулы нуклеиновой кислоты с МИП или шириной ниже порогового значения. Например, могут быть использованы только первые молекулы нуклеиновой кислоты со значением МИП, превышающим 10-й процентиль (или 1-й, 5-й, 15-й, 20-й, 30-й, 40-й, 50-й, 60-й, 70-й, 80-й, 90-й или 95-й процентиль). Процентиль может быть основан на данных по всем молекулам нуклеиновой кислоты в эталонном образце или эталонных образцах. Пороговое значение ширины также может соответствовать процентилю.

На этапе 1024 сохраняется множество первых обучающих выборок. Каждая первая обучающая выборка включает в себя одну из первого множества первых структур данных и первую метку, обозначающую первое состояние модификации нуклеотида в целевой позиции.

На этапе 1026 принимается второе множество вторых структур данных. Этап 1026 может быть необязательным. Каждая из вторых структур данных второго множества вторых структур данных соотносится с соответствующим окном нуклеотидов, секвенированных в соответствующей молекуле нуклеиновой кислоты из множества вторых молекул нуклеиновой кислоты. Второе множество молекул нуклеиновой кислоты может быть таким же или отличаться от множества первых молекул нуклеиновой кислоты. Модификация имеет известное второе состояние нуклеотида в целевой позиции в каждом окне каждой из вторых молекул нуклеиновой кислоты. Второе состояние отличается от первого состояния. Например, если первое состояние - это то, что модификация присутствует, то второе состояние - то, что модификация отсутствует, и наоборот. Каждая из вторых структур данных включает в себя значения тех же свойств, что и первое множество первых структур данных.

Множество первых обучающих выборок может быть сгенерировано с использованием амплификации множественного вытеснения (MDA). В некоторых вариантах осуществления, множество первых обучающих выборок может быть сгенерировано путем амплификации первого множества молекул нуклеиновой кислоты с использованием набора нуклеотидов. Набор нуклеотидов может включать в себя первый тип метилирования (например, 6mA или любое другое метилирование [например, CpG]) в определенном соотношении. Указанное соотношение может включать в себя 1:10, 1:100, 1:1000, 1:10000, 1:100000, или 1:1000000 по отношению к неметилированным нуклеотидам. Множество вторых молекул нуклеиновой кислоты может быть получено с использованием амплификации множественного вытеснения с неметилированными нуклеотидами первого типа.

На этапе 1028 сохраняется множество вторых обучающих выборок. Этап 1028 может быть необяза-

тельным. Каждая из вторых обучающих выборок включает в себя одну из второго множества вторых структур данных и вторую метку, обозначающую второе состояние для модификации нуклеотида в целевой позиции.

На этапе 1029 модель обучают с использованием множества первых обучающих выборок и, необязательно, множества вторых обучающих выборок. Обучение выполняют путем оптимизации параметров модели на основе выходных данных модели, совпадающих или не совпадающих с соответствующими метками из первых меток и, необязательно, вторых меток, когда первое множество первых структур данных и, необязательно, второе множество вторых структур данных является входными данными для модели. Выходные данные модели указывают на то, имеет ли нуклеотид в целевой позиции в соответствующем окне модификацию. Способ может включать в себя только множество первых обучающих выборок, поскольку модель может идентифицировать выброс как состояние, отличное от первого состояния. Модель может быть статистической моделью, также называемой моделью машинного обучения.

В некоторых вариантах осуществления, выходные данные модели могут включать в себя вероятность нахождения в каждом из множества состояний. За состояние может быть принято состояние с наибольшей вероятностью.

Модель может включать в себя сверточную нейронную сеть (СНС). СНС может включать в себя набор сверточных фильтров, сконфигурированных для фильтрации первого множества структур данных и, необязательно, второго множества структур данных. Фильтр может быть любым фильтром, описанным в данном документе. Количество фильтров для каждого слоя может быть от 10 до 20, от 20 до 30, от 30 до 40, от 40 до 50, от 50 до 60, от 60 до 70, от 70 до 80, от 80 до 90, от 90 до 100, от 100 до 150, от 150 до 200 или больше. Размер ядра для фильтров может составлять 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, от 15 до 20, от 20 до 30, от 30 до 40, или больше. СНС может включать в себя входной слой, сконфигурированный для приема отфильтрованного первого множества структур данных и, необязательно, отфильтрованного второго множества структур данных. СНС также может включать в себя множество скрытых слоев, включающих в себя множество узлов. Первый слой из множества скрытых слоев связан с входным слоем. СНС может дополнительно включать в себя выходной слой, связанный с последним слоем из множества скрытых слоев и сконфигурированный для вывода выходной структуры данных. Выводимая структура данных может содержать свойства.

Модель может включать в себя модель обучения с учителем. Модели обучения с учителем могут включать в себя различные подходы и алгоритмы, включающие в себя аналитическое обучение, искусственную нейронную сеть, обратное распространение, усиление (бустинг) (мета-алгоритм), байесовскую статистику, формирование рассуждений по прецедентам, обучение дерева решений, индуктивное логическое программирование, регрессию на основе гауссовских процессов, генетическое программирование, способ группового учёта аргументов, ядерные оценки, обучающиеся автоматы, обучающиеся системы классификации, минимальную длину сообщения (деревья решений, графы решений и т.д.), многолинейное подпространственное обучение, наивный байесовский классификатор, классификатор максимальной энтропии, условное случайное поле, алгоритм ближайшего соседа, вероятностно-корректное в смысле аппроксимации обучение (PAC), правила приобретения знаний на основе текущего контекста, методологию получения знаний, символьные алгоритмы машинного обучения, субсимвольные алгоритмы машинного обучения, машину опорных векторов, машины минимальной сложности (МСМ), классификаторы на основе комитета деревьев принятия решений, ансамбли классификаторов, порядковую классификацию, предварительную обработку данных, обработку несбалансированных наборов данных, статистическое реляционное обучение, или Proaftn, алгоритм классификации по множеству признаков. Модель может представлять собой 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 машину опорных векторов (SVM), или любую, описанную в данном документе.

В рамках обучения модели машинного обучения, параметры модели машинного обучения (такие как веса, пороговые значения, например, которые могут использоваться для функций активации в нейронных сетях и т.д.) могут быть оптимизированы на основе обучающих выборок (обучающего набора), чтобы обеспечить оптимальную точность классификации модификации нуклеотида в целевой позиции. Могут быть выполнены различные формы оптимизации, например, обратное распространение, минимизация эмпирического риска и минимизация структурного риска. Набор образцов для проверки (структура данных и метка) может использоваться для проверки точности модели. Перекрестная проверка может выполняться с использованием различных частей обучающего набора для обучения и проверки. Модель может содержать множество субмоделей, тем самым предоставляя ансамблевую модель. Субмодели могут быть более слабыми моделями, которые после объединения дают более точную окончательную модель.

В некоторых вариантах осуществления, для проверки модели можно использовать химерные или гибридные молекулы нуклеиновой кислоты. По меньшей мере, некоторые из множества первых молекул нуклеиновой кислоты каждая содержат первую часть, соответствующую первой эталонной последова-

тельности, и вторую часть, соответствующую второй эталонной последовательности. Первая эталонная последовательность может происходить из другой хромосомы, ткани (например, опухолевой или неопухолевой), организма или вида, в отличие от второй эталонной последовательности. Первая эталонная последовательность может быть человеческой, а вторая эталонная последовательность может быть из другого животного. Каждая химерная молекула нуклеиновой кислоты может содержать первую часть, соответствующую первой эталонной последовательности, и вторую часть, соответствующую второй эталонной последовательности. Первая часть может иметь первый паттерн метилирования, а вторая часть может иметь второй паттерн метилирования. Первая часть может быть обработана метилазой. Вторая часть может быть не обработана метилазой и может соответствовать неметилированной части второй эталонной последовательности.

В. Обнаружение модификаций.

Фиг. 103 демонстрирует способ 1030 для обнаружения модификации нуклеотида в молекуле нуклеиновой кислоты. Модификация может представлять собой любую модификацию, описанную с помощью способа 1020 на фиг. 102.

На этапе 1032 принимается входная структура данных. Входная структура данных может соответствовать окну нуклеотидов, секвенированных в молекуле нуклеиновой кислоты образца. Молекула нуклеиновой кислоты образца может быть секвенирована путем измерения импульсов оптического сигнала, соответствующего нуклеотидам. Окно может представлять собой любое окно, описанное в этапе 1022 на фиг. 102, и последовательность может представлять собой любую последовательность, описанную в этапе 1022 на фиг. 102. Входная структура данных может включать в себя значения тех же свойств, которые описаны в этапе 1022 на фиг. 102. Способ 1030 может включать в себя секвенирование молекулы нуклеиновой кислоты образца.

Нуклеотиды в пределах окна могут быть или не быть выровнены с эталонным геномом. Нуклеотиды в пределах окна могут быть определены с использованием кольцевой консенсусной последовательности (ККП) без выравнивания секвенированных нуклеотидов с эталонным геномом. Нуклеотиды в каждом окне можно идентифицировать с помощью ККП, вместо выравнивания с эталонным геномом. В некоторых вариантах осуществления, окно может быть определено без ККП и без выравнивания секвенированных нуклеотидов с эталонным геномом.

Нуклеотиды в пределах окна могут быть обогащены или отфильтрованы. Обогащение может осуществляться с помощью подхода с участием Cas9. Подход Cas9 может включать в себя разрезание двухцепочечной молекулы ДНК с использованием комплекса Cas9 для формирования разрезанной двухцепочечной молекулы ДНК и лигирование адаптера в виде шпильки к концу разрезанной двухцепочечной молекулы ДНК, аналогично фиг. 91. Фильтрация может осуществляться путем отбора двухцепочечных молекул ДНК, имеющих размер в пределах диапазона размеров. Нуклеотиды могут происходить из этих двухцепочечных молекул ДНК. Могут быть использованы другие способы, которые сохраняют статус метилирования молекул (например, метил-связывающие белки).

На этапе 1034 входную структуру данных вводят в модель. Модель может быть обучена с помощью способа 1020 на фиг. 102.

В некоторых вариантах осуществления, для проверки модели можно использовать молекулы химерной нуклеиновой кислоты. По меньшей мере, некоторые из множества первых молекул нуклеиновой кислоты каждая содержат первую часть, соответствующую первой эталонной последовательности, и вторую часть, соответствующую второй эталонной последовательности, которая не пересекается с первой эталонной последовательностью. Первая эталонная последовательность может происходить из другой хромосомы, ткани (например, опухолевой или неопухолевой), органеллы (например, митохондрии, ядра, хлоропласта), организма (млекопитающего, вируса, бактерии, т.д.), или вида, нежели вторая эталонная последовательность. Первая эталонная последовательность может быть человеческой, а вторая эталонная последовательность может быть из другого животного. Каждая химерная молекула нуклеиновой кислоты может содержать первую часть, соответствующую первой эталонной последовательности, и вторую часть, соответствующую второй эталонной последовательности. Первая часть может иметь первый паттерн метилирования, а вторая часть может иметь второй паттерн метилирования. Первая часть может быть обработана метилазой. Вторая часть может быть не обработана метилазой и может соответствовать неметилированной части второй эталонной последовательности.

На этапе 1036 с помощью модели определяют, присутствует ли модификация на нуклеотиде в целевой позиции в пределах окна в входной структуре данных.

Входная структура данных может быть одной входной структурой данных из множества входных структур данных. Каждая входная структура данных может соотноситься с соответствующим окном нуклеотидов, секвенированных в соответствующей молекуле нуклеиновой кислоты образца из множества молекул нуклеиновой кислоты образца. Множество молекул нуклеиновой кислоты образца может быть получено из биологического образца субъекта. Биологический образец может быть любым биологическим образцом, описанным в данном документе. Способ 1030 может повторяться для каждой входной структуры данных. Способ может включать в себя прием множества входных структур данных. В модель может быть введено множество входных структур данных. С помощью модели можно опреде-

лить присутствует ли модификация на нуклеотиде в целевой позиции в соответствующем окне каждой входной структуры данных.

Каждая молекула нуклеиновой кислоты образца из множества молекул нуклеиновой кислоты образца может иметь размер, превышающий пороговое значение размера. Например, пороговое значение размера может составлять 100, 200, 300, 400, 500, 600, 700, 800, 900 п.о., 1, 2, 3, 4, 5, 6, 7, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 500 т.п.о., или 1 млн.п.о. Наличие порогового значения размера может дать более высокое значение глубины субпрочтений, что может повысить точность обнаружения модификации. В некоторых вариантах осуществления, способ может включать в себя фракционирование молекул ДНК до определенных размеров перед секвенированием молекул ДНК.

Множество молекул нуклеиновой кислоты образца могут выравниваться с множеством геномных областей. Для каждой области генома из множества областей генома, с областью генома может быть выровнен ряд молекул нуклеиновой кислоты образца. Количество молекул нуклеиновой кислоты образца может быть больше порогового значения. Величина порогового значения может являться пороговым значением глубины субпрочтений. Величина порогового значения глубины субпрочтений может составлять 1x, 10x, 30x, 40x, 50x, 60x, 70x, 80x, 900x, 100x, 200x, 300x, 400x, 500x, 600x, 700x или 800x. Для улучшения или оптимизации точности может быть определена величина порогового значения глубины субпрочтений. Величина порогового значения глубины субпрочтений может быть связана с величиной множества геномных областей. Например, чем больше величина порогового значения глубины субпрочтений, тем меньше величина множества геномных областей.

Можно определить, присутствует ли модификация на одном или большем количестве нуклеотидов. Классификация нарушения может быть выполнена по наличию модификации на одном или большем количестве нуклеотидов. Классификация нарушения может включать в себя использование ряда модификаций. Количество модификаций можно сравнить с пороговым значением. В альтернативном или дополнительном варианте, классификация может включать в себя позицию одной или большего количества модификаций. Позиция одной или большего количества модификаций может быть определена путем выравнивания прочтений последовательности молекулы нуклеиновой кислоты с эталонным геномом. Может быть определено нарушение, если показано, что определенные позиции, которые, как известно, коррелируют с нарушением, имеют модификацию. Например, паттерн метилированных сайтов можно сравнить с контрольным паттерном для нарушения, и определение нарушения может быть основано на сравнении. Совпадение с эталонным паттерном или существенное совпадение (например, 80, 90 или 95% или больше) с эталонным паттерном может указывать на нарушение или высокую вероятность нарушения. Нарушение может представлять собой рак или любое нарушение (например, нарушение, связанное с беременностью, аутоиммунное заболевание), описанное в данном документе.

Может быть проанализировано статистически значимое количество молекул нуклеиновой кислоты, чтобы обеспечить точное определение нарушения, тканевого происхождения или клинически значимой фракции ДНК. В некоторых вариантах осуществления, анализируют по меньшей мере 1000 молекул нуклеиновой кислоты. В других вариантах осуществления, может быть проанализировано по меньшей мере 10000, или 50000, или 100000, или 500000, или 1000000, или 5000000 молекул нуклеиновой кислоты, или большее количество. В качестве дополнительного примера может быть сгенерировано по меньшей мере 10000, 50000, 100000, или 500000, или 1000000, или 5000000 прочтений последовательности.

Способ может включать в себя определение того, что классификация нарушения соответствует тому, что субъект имеет нарушение. Классификация может включать в себя степень нарушения с использованием количества модификаций и/или сайтов модификаций.

Клинически значимая фракция ДНК, профиль метилирования плода, профиль метилирования матери, наличие подверженной импринтингу области гена или тканевое происхождения (например, из образца, содержащего смесь различных типов клеток) могут быть определены за счет наличия модификации одного или большего количества нуклеотидов. Клинически значимая фракция ДНК включает в себя, но не ограничивается лишь этими: фракцию ДНК плода, фракцию опухолевой ДНК (например, из образца, содержащего смесь опухолевых клеток и неопухолевых клеток) и фракцию ДНК трансплантата (например, из образца, содержащего смесь клеток донора и клеток реципиента).

Способ может дополнительно включать в себя лечение нарушения. Лечение может быть предоставлено согласно определенной степени нарушения, идентифицированным модификациями и/или тканевому происхождению (например, опухолевые клетки, выделенные из кровотока больного раком). Например, на выявленную модификацию можно нацелить конкретное лекарство или химиотерапию. Тканевое происхождение может быть использовано для проведения операции или любой другой формы лечения. Кроме того, степень нарушения можно использовать для определения того, насколько агрессивным должно быть лечение.

Варианты осуществления могут включать в себя лечение нарушения у пациента после определения степени нарушения у пациента. Лечение может включать в себя любую подходящую терапию, лекарство, химиотерапию, облучение или хирургическое вмешательство, включая любое лечение, описанное в источнике, упомянутом в данном документе. Информация о лечении из источников включена в данный документ посредством ссылки.

VI. Анализ гаплотипа.

Различия в профилях метилирования между двумя гаплотипами были обнаружены в образцах опухолевой ткани. Таким образом, дисбаланс метилирования между гаплотипами может быть использован для определения классификации степени рака или другого заболевания. Дисбаланс по гаплотипах также может использоваться для определения наследования гаплотипа плодом. Заболевания плода также могут быть идентифицированы путем анализа дисбаланса метилирования между гаплотипами. Клеточная ДНК может использоваться для анализа уровней метилирования гаплотипов.

A. Анализ метилирования, связанного с гаплотипом.

Технология секвенирования отдельной молекулы в реальном времени позволяет идентифицировать отдельные ОНП. Длинные прочтения, полученные секвенированием отдельной молекулы в реальном времени (например, вплоть до нескольких т.п.о.) позволят группировать варианты в геномах, используя информацию о гаплотипах, присутствующую в каждом консенсусном прочтении (Edge et al. Genome Res. 2017;27:801-812; Wenger et al. Nat Biotechnol. 2019;37:1155-1162). Профиль метилирования гаплотипа может быть проанализирован по уровням метилирования сайтов CpG, связанных посредством ККП с аллелями соответствующих гаплотипов, как показано на фиг. 77. Этот анализ оценки гаплотипа метилирования может быть использован для решения вопроса о том, имеют ли две копии гомологичных хромосом одинаковые или разные паттерны метилирования при различных клинически значимых патологиях, таких как рак. В одном варианте осуществления, метилирование гаплотипа будет представлять собой агрегированные уровни метилирования, вклад в которые вносит ряд фрагментов ДНК, отнесенных к этому гаплотипу. Гаплотип может представлять собой блоки разных размеров, включающие в себя, но не ограничивающиеся лишь этими: 50, 100, 200, 300, 400, 500 нт, 1, 2, 3, 4, 5, 10, 20, 30, 40, 50, 100, 200, 300, 400, 500 кнт, 1, 2 и 3 Мнт.

B. Анализ относительного дисбаланса метилирования на основе гаплотипов.

Фиг. 104 иллюстрирует анализ относительного дисбаланса метилирования на основе гаплотипов. Гаплотипы (т.е. Гапл I и Гапл II) определяли путем анализа результатов секвенирования отдельной молекулы в реальном времени. Паттерны метилирования, связанные с каждым гаплотипом, могут быть определены с использованием тех связанных с гаплотипом фрагментов, профили метилирования которых были определены в соответствии с подходом, описанным на фиг. 77. Таким образом, можно было сравнить паттерны метилирования между Гапл I и Гапл II.

Чтобы количественно оценить разницу в метилировании между Гапл I и Гапл II, была рассчитана разница уровней метилирования (ΔF) между Гапл I и Гапл II. Разницу ΔF рассчитывали как:

$$\Delta F = M_{\text{Гапл}I} - M_{\text{Гапл}II}$$

где ΔF представляет собой разницу в уровне метилирования между Гапл I и Гапл II, и $M_{\text{ГаплI}}$ и $M_{\text{ГаплII}}$ представляют уровни метилирования Гапл I и Гапл II, соответственно.

Положительное значение ΔF свидетельствует о более высоком уровне метилирования ДНК для Гапл I по сравнению с Гапл II.

C. Анализ относительного дисбаланса метилирования на основе гаплотипов для ДНК опухоли ГЦК.

В одном варианте осуществления, анализ метилирования гаплотипа может быть полезен для обнаружения аберраций метилирования в раковых геномах. Например, будет проанализировано изменение метилирования между двумя гаплотипами в пределах геномной области. Гаплотип в пределах геномной области определяется как гаплотипный блок. Гаплотипный блок можно рассматривать как набор аллелей на хромосоме, которые были гаплотипированы. В некоторых вариантах осуществления, гаплотипный блок может быть расширен как можно дальше согласно набору информации о последовательности, которая обосновывает физическое сцепление двух аллелей на хромосоме. Для случая 3033 мы получили 97475 гаплотипных блоков из результатов секвенирования ДНК прилегающих нормальных тканей. Средний размер гаплотипных блоков составил 2,8 т.п.о. 25% гаплотипных блоков имели размер больше чем 8,2 т.п.н. Максимальный размер гаплотипных блоков составил 282,2 т.п.н. Набор данных был сгенерирован из ДНК, полученной с помощью Sequel II Sequencing Kit 1.0.

В целях иллюстрации мы использовали ряд критериев для идентификации потенциальных гаплотипных блоков, которые демонстрируют дифференциальное метилирование между Гапл I и Гапл II в ДНК опухоли по сравнению с ДНК прилегающей неопухолевой ткани. Критерии представляли собой: (1) анализируемый гаплотипный блок содержал по меньшей мере 3 последовательности ККП, которые были получены из трех ячеек для секвенирования, соответственно; (2) абсолютная разница в уровне метилирования между Гапл I и Гапл II в ДНК прилегающей неопухолевой ткани составляла меньше чем 5%; (3) абсолютная разница в уровне метилирования между Гапл I и Гапл II в ДНК опухолевой ткани составляла больше чем 30%. Мы идентифицировали 73 гаплотипных блока, удовлетворяющих указанным выше критериям.

Фиг. 105A и 105B представляют собой табл. 73 гаплотипных блоков, демонстрирующие дифференциальные уровни метилирования между Гапл I и Гапл II в ДНК опухоли ГЦК по сравнению с ДНК прилегающей неопухолевой ткани для случая TBR3033. Первый столбец показывает хромосому, связанную с гаплотипным блоком. Второй столбец показывает начальную координату гаплотипного блока в хромо-

соме. Третий столбец показывает конечную координату гаплотипного блока. Четвертый столбец показывает длину гаплотипного блока. В четвертом столбце указан идентификатор гаплотипного блока. Пятый столбец показывает уровень метилирования Гапл I в неопухолевой ткани, прилегающей к опухолевой ткани. Шестой столбец показывает уровень метилирования Гапл II в неопухолевой ткани. Седьмой столбец показывает уровень метилирования Гапл I в опухолевой ткани. Восьмой столбец показывает уровень метилирования Гапл II в опухолевой ткани.

В отличие от 73 гаплотипных блоков, показывающих разницу в уровне метилирования между гаплотипами ДНК опухолевой ткани больше чем 30%, только один гаплотипный блок показал разницу больше чем 30% для ДНК неопухолевой ткани, но меньше чем 5% отличие в ДНК опухолевой ткани ДНК. В некоторых вариантах осуществления, можно использовать другой набор критериев для идентификации гаплотипных блоков, демонстрирующих неодинаковое метилирование. Могут быть использованы другие максимальные и минимальные пороговые значения разницы. Например, минимальные пороговые значения разницы могут составлять 10, 15, 20, 25, 30, 35, 40, 45, 50% или более. Максимальные пороговые значения разницы могут составлять, например, 1, 5, 10, 15, 20 или 30%. Эти результаты предполагают, что различие в метилировании между гаплотипами может служить новым биомаркером для диагностики, обнаружения, мониторинга, прогнозирования и протокола для лечения рака.

В некоторых вариантах осуществления, длинный гаплотипный блок in silico может быть разделен на меньшие блоки при изучении паттернов метилирования.

Для случая 3032 мы получили 61958 гаплотипных блоков из результатов секвенирования ДНК прилегающей неопухолевой ткани. Средний размер гаплотипных блоков составил 9,3 т.п.о. 25% гаплотипных блоков имели размер больше чем 27,6 т.п.н. Максимальный размер гаплотипных блоков составил 717,8 т.п.н. В целях иллюстрации мы использовали те же три критерия, описанные выше, для идентификации потенциальных гаплотипных блоков, которые демонстрируют дифференциальное метилирование между Гапл I и Гапл II в ДНК опухоли по сравнению с ДНК прилегающей нормальной ткани. Мы идентифицировали 20 гаплотипных блоков, удовлетворяющих указанным выше критериям. Набор данных был сгенерирован из ДНК, полученной с помощью Sequel II Sequencing Kit 1.0.

Фиг. 106 представляет собой таблицу 20 гаплотипных блоков, демонстрирующую отличающиеся уровни метилирования между Гапл I и Гапл II в ДНК опухоли по сравнению с ДНК прилегающей нормальной ткани для случая TBR3032. Первый столбец показывает хромосому, связанную с гаплотипным блоком. Второй столбец показывает начальную координату гаплотипного блока в хромосоме. Третий столбец показывает конечную координату гаплотипного блока. Четвертый столбец показывает длину гаплотипного блока. В четвертом столбце указан идентификатор гаплотипного блока. Пятый столбец показывает уровень метилирования Гапл I в неопухолевой ткани, прилегающей к опухолевой ткани. Шестой столбец показывает уровень метилирования Гапл II в неопухолевой ткани. Седьмой столбец показывает уровень метилирования Гапл I в опухолевой ткани. Восьмой столбец показывает уровень метилирования Гапл II в опухолевой ткани.

В отличие от 20 гаплотипных блоков, показывающих разницу в опухолевой ткани ГЦК на фиг. 106, только один гаплотипный блок показал разницу больше чем 30% в неопухолевой ткани, но меньше чем 5% разницу в опухолевой ткани. Эти результаты дополнительно предполагают, что различие в метилировании между гаплотипами может служить новым биомаркером для диагностики, обнаружения, мониторинга, прогнозирования и протокола лечения рака. Для других вариантов осуществления, могут быть использованы критерии для идентификации гаплотипных блоков, демонстрирующих дифференциальное метилирование.

D. Анализ относительного дисбаланса метилирования на основе гаплотипов для ДНК из других типов опухолей.

Как указано выше, анализ уровней метилирования между гаплотипами показал, что опухолевые ткани ГЦК содержат больше гаплотипных блоков, демонстрирующих дисбаланс метилирования, по сравнению с поставленным к ним в пару прилегающими неопухолевыми тканями. В качестве одного примера, критерии для гаплотипного блока, демонстрирующего дисбаланс метилирования в опухолевой ткани, представляли собой: (1) анализируемый гаплотипный блок содержал по меньшей мере три последовательности ККП, которые были получены из трех ячеек для секвенирования; (2) абсолютная разница в уровне метилирования между Гапл I и Гапл II в ДНК прилегающей неопухолевой ткани или ДНК нормальной ткани, исходя из предыдущих данных, составляла меньше чем 5%; (3) абсолютная разница в уровне метилирования между Гапл I и Гапл II в ДНК опухолевой ткани составляла больше чем 30%. Был включен критерий (2), поскольку неопухолевые/нормальные ткани, демонстрирующие дисбаланс гаплотипа по уровнях метилирования, могут указывать на области импринтинга, а не на опухолевые области. Критерии для гаплотипного блока, демонстрирующего дисбаланс метилирования в неопухолевой ткани, представляли собой: (1) анализируемый гаплотипный блок содержал по меньшей мере три последовательности ККП, которые были получены из трех ячеек для секвенирования; (2) абсолютная разница в уровне метилирования между Гапл I и Гапл II в ДНК прилегающей неопухолевой ткани или ДНК нормальной ткани, исходя из предыдущих данных, составляла больше чем 30%; (3) абсолютная разница в уровне метилирования между Гапл I и Гапл II в ДНК опухолевой ткани составляла меньше чем 5%.

В других вариантах осуществления могут использоваться другие критерии. Например, для идентификации дисбаланса гаплотипа I генома рака, разница в уровне метилирования между Гапл I и Гапл II может составлять меньшей чем 1, 5, 10, 20, 40, 50 или 60% и т.д., в неопухолевых тканях, тогда как разница в уровне метилирования между Гапл I и Гапл II может составлять больше чем 1, 5, 10, 20, 40, 50 или 60% и т.д. в опухолевых тканях. Для идентификации дисбаланса гаплотипа I неракового генома, разница в уровне метилирования между Гапл I и Гапл II может составлять больше чем 1, 5, 10, 20, 40, 50 или 60% и т.д., в неопухолевых тканях, тогда как разница в уровне метилирования между Гапл I и Гапл II может составлять меньше чем 1, 5, 10, 20, 40, 50 или 60% и т.д. в опухолевых тканях.

Фиг. 107A представляет собой таблицу, резюмирующую количество гаплотипных блоков, демонстрирующих дисбаланс метилирования между двумя гаплотипами, между опухолевой и прилегающей неопухолевой тканями на основе данных, полученных с помощью Sequel II Sequencing Kit 2.0. В первом столбце приведен тип ткани. Во втором столбце приведено количество гаплотипных блоков, демонстрирующих дисбаланс метилирования между двумя гаплотипами в опухолевых тканях. В третьем столбце приведено количество гаплотипных блоков, демонстрирующих дисбаланс метилирования между двумя гаплотипами в прилегающих неопухолевых тканях в паре с ними. Строки показывают опухолевую ткань с большим количеством гаплотипных блоков, демонстрирующих дисбаланс метилирования между двумя гаплотипами, чем у поставленной в пару, прилегающей неопухолевой ткани.

Медианная длина гаплотипных блоков, вовлеченных в этот анализ, составляла 15,7 т.п.н. (МКД: 10,3-26,1 т.п.н.). Включая результаты ГНК для печени, данные показывают 7 типов тканей, для которых опухолевая ткань содержала больше гаплотипных блоков с дисбалансом метилирования. Помимо печени, к другим тканям относятся ткани толстой кишки, молочной железы, почек, легких, предстательной железы и желудка. Таким образом, в некоторых вариантах осуществления, для определения наличия у пациента опухоли или рака можно использовать ряд гаплотипных блоков, имеющих дисбаланс метилирования.

Фиг. 107B представляет собой таблицу, резюмирующую ряд гаплотипных блоков, демонстрирующих дисбаланс метилирования между двумя гаплотипами в опухолевых тканях для разных стадий опухоли на основе данных, полученных с помощью Sequel II Sequencing Kit 2.0. В первом столбце показан тип ткани с опухолью. Во втором столбце показан ряд гаплотипных блоков с дисбалансом метилирования между двумя гаплотипами в опухолевых тканях. В третьем столбце приведена информация о стадиях опухоли с использованием классификации TNM злокачественных опухолей. T3 и T3a - это опухоль большего размера, чем T2.

В таблице показано больше гаплотипных блоков, демонстрирущих дисбаланс метилирования для более крупных опухолей как для груди, так и для почек. Например, для ткани молочной железы, ткань, классифицированная как опухоль стадии T3 (классификация TNM), ER-положительная и демонстрирующая амплификацию ERBB2, имела больше гаплотипных блоков (57), демонстрирующих дисбаланс метилирования, чем ткань, имеющая такие гаплотипные блоки (18), классифицированная как опухоль стадии T2 (классификация TNM), PR (рецептор прогестерона)/ER (рецептор эстрогена) положительная и без амплификации ERBB2. Для почечной ткани, ткань, классифицированная как опухоль стадии T3a, имела больше гаплотипных блоков (68), демонстрирующих дисбаланс метилирования, чем ткань, имеющая такие гаплотипные блоки (0), классифицированная как опухоль стадии T2.

В некоторых вариантах осуществления, можно использовать гаплотипные блоки, демонстрирующие дисбаланс метилирования, для классификации опухолей и для установления связи с их клиническим поведением (например, прогрессированием, прогнозом или ответом на лечение). Эти данные предполагают, что степень дисбаланса метилирования на основе гаплотипов может служить классификатором опухолей и может быть включена в клинические исследования, или испытания, или возможные клинические услуги. Классификация опухолей может включать в себя размер и степень тяжести.

E. Анализ метилирования внеклеточной ДНК из материнской плазмы на основе гаплотипов.

Могут быть определены гаплотипы обеих родителей или одного из родителей. Способы гаплотипирования могут включать в себя секвенирование отдельной молекулы с длинным прочтением, секвенирование с соединенными короткими прочтениями (например, 10x Genomics), ПЦР длинных фрагментов отдельной молекулы, вывод на основе популяции. Если отцовские гаплотипы известны, метилом внеклеточной эмбриональной ДНК может быть собран путем соединения профилей метилирования множества внеклеточных молекул ДНК, каждая из которых содержит по меньшей мере один отцовский специфический аллель ОНП, которые представлены в отцовском гаплотипе. Другими словами, отцовский гаплотип используется как каркас для соединения специфичных для плода прочтений последовательностей.

Фиг. 108 иллюстрирует анализ гаплотипов на предмет относительного дисбаланса метилирования. Если материнские гаплотипы известны, дисбаланс метилирования между двумя гаплотипами (т.е. Гапл I и Гапл II) может быть использован для определения материнского гаплотипа, унаследованного от плода. Как показано на фиг. 108, молекулы ДНК из плазмы беременной женщины секвенируют с использованием технологии секвенирования отдельной молекулы в реальном времени. Информация о метилировании и аллелях может быть получена согласно раскрытию изобретения в данном документе. В одном варианте осуществления, ОНП, связанные с вызывающим заболевание геном, обозначены как Гапл I. Если плод

унаследовал Гапл I, в материнской плазме будет присутствовать больше фрагментов, несущих аллели Гапл I, по сравнению с фрагментами, несущими аллели Гапл II. Гипометилирование фрагментов ДНК, полученных из плода, будет снижает уровень метилирования Гапл I по сравнению с Гапл II. В результате, если метилирование Гапл I показывает более низкий уровень метилирования, чем Гапл II, плод с большей вероятностью наследует материнский Гапл I. В противном случае плод с большей вероятностью наследует Гапл II от матери. В клинической практике анализ дисбаланса метилирования на основе гаплотипов может использоваться для определения того, унаследовал ли плод материнский гаплотип, связанный с генетическими нарушениями, например, но без ограничения, моногенными нарушениями, включающими в себя синдром ломкой Х-хромосомы, мышечную дистрофию, болезнь Хантингтона, или бета-талассемию.

F. Пример способа классификации нарушений.

Фиг. 109 демонстрирует иллюстративный способ 1090 классификации нарушения в организме, обладающем первым гаплотипом и вторым гаплотипом. Способ 1090 включает в себя сравнение относительных уровней метилирования между двумя гаплотипами.

На этапе 1091 молекулы ДНК из биологического образца анализируют для определения их местоположения в эталонном геноме, соответствующем организму. Молекулы ДНК могут быть молекулами клеточной ДНК. Например, молекулы ДНК могут быть секвенированы для получения прочтений последовательности, и прочтения последовательности могут быть картированы (выровнены с) на эталонном геноме. Если бы организм представлял собой человека, то эталонный геном был бы эталонным геномом человека, возможно, из определенной субпопуляции. В качестве другого примера, молекулы ДНК могут быть проанализированы с помощью различных зондов (например, после ПЦР или других способов амплификации), где каждый зонд соответствует геномному местоположению, которое может охватывать гетерозиготный, и один или большее количество сайтов CpG, как описано ниже.

Дополнительно, молекулы ДНК могут быть проанализированы для определения соответствующего аллеля молекулы ДНК. Например, аллель молекулы ДНК может быть определен из последовательности прочтения, полученного при секвенировании или для конкретного зонда, который гибридизируется с молекулой ДНК, при этом оба метода могут предоставить последовательность прочтения (например, зонд можно рассматривать как последовательность прочтения при наличии гибридизации). Состояние метилирования в каждом одном или большем количестве сайтов (например, сайтов CpG) может быть определено для молекул ДНК.

На этапе 1092 идентифицируют один или большее количество гетерозиготных локусов первой части первой хромосомной области. Каждый гетерозиготный локус может содержать соответствующий первый аллель в первом гаплотипе и соответствующий второй аллель во втором гаплотипе. Один или большее количество гетерозиготных локусов могут быть первым множеством гетерозиготных локусов, где второе множество гетерозиготных локусов может соответствовать другой хромосомной области.

На этапе 1093 идентифицируют первый набор из множества молекул ДНК. Каждая из множества молекул ДНК размещается в любом из гетерозиготных локусов из этапа 1096 и содержит соответствующий первый аллель, так что молекула ДНК может быть идентифицирована как соответствующая первому гаплотипу. Молекула ДНК может быть расположена больше чем в одном гетерозиготном локусе, но обычно прочтение будет включать в себя только один гетерозиготный локус. Каждая из первого набора молекул ДНК также содержать по меньшей мере один из N геномных сайтов, при этом геномные сайты используются для измерения уровней метилирования. N - целое число, например, большее или равное 1, 2, 3, 4, 5, 10, 20, 50, 100, 200, 500, 1000, 2000 или 5000. Таким образом, прочтение молекулы ДНК может указывать на покрытие 1 сайта, 2 сайтов и т.д. 1 геномный сайт может включать в себя сайт, в котором присутствует нуклеотид CpG.

На этапе 1094 определяют первый уровень метилирования первой части первого гаплотипа с использованием первого набора из множества молекул ДНК. Первый уровень метилирования можно определить любым способом, описанным в данном документе. Первая часть может соответствовать одному сайту или включать в себя множество сайтов. Первая часть первого гаплотипа может быть длиннее чем или равной 1 т.п.н. Например, первая часть первого гаплотипа может быть длиннее чем или равной 1 т.п.н., 5 т.п.н., 10 т.п.н., 15 т.п.н. или 20 т.п.н. Данные по метилированию могут быть данными по клеточной ДНК.

В некоторых вариантах осуществления, множество первых уровней метилирования может быть определено для множества частей первого гаплотипа. Каждая часть может иметь длину большую чем или равную 5 т.п.н., или любой размер, раскрытый в данном документе для первой части первого гаплотипа.

На этапе 1095 идентифицируеют второй набор из множества молекул ДНК. Каждая из множества молекул ДНК расположена в любом из гетерозиготных локусов из этапа 1096 и содержит соответствующий второй аллель, так что молекула ДНК может быть идентифицирована как соответствующая второму гаплотипу. Каждая из второго набора молекул ДНК также содержит по меньшей мере один из N геномных сайтов, при этом геномные сайты используются для определения уровней метилирования.

На этапе 1096 определяют второй уровень метилирования первой части второго гаплотипа с использованием второго набора из множества молекул ДНК. Второй уровень метилирования можно опре-

делить любым способом, описанным в данном документе. Первая часть второго гаплотипа может быть длиннее чем или равной 1 т.п.н., или любого размера для первой части первого гаплотипа. Первая часть первого гаплотипа может быть комплементарной первой части второго гаплотипа. Первая часть первого гаплотипа и первая часть второго гаплотипа могут формировать кольцевую молекулу ДНК. Первый уровень метилирования первой части первого гаплотипа может быть определен с использованием данных из кольцевой молекулы ДНК. Например, анализ кольцевой ДНК может включать в себя анализ, описанный на фиг. 1, фиг. 2, фиг. 4, фиг. 5, фиг. 6, фиг. 7, фиг. 8, фиг. 50 или фиг. 61.

Кольцевая молекула ДНК может быть образована путем разрезания двухцепочечной молекулы ДНК с использованием комплекса Cas9 для формирования разрезанной двухцепочечной молекулы ДНК. Адаптер в виде шпильки может быть лигирован к концу разрезанной двухцепочечной молекулы ДНК. В вариантах осуществления, оба конца двухцепочечной молекулы ДНК могут быть разрезаны и лигированы. Например, разрезание, лигирование и последующий анализ могут выполняться, как описано на фиг. 91.

В некоторых вариантах осуществления, множество вторых уровней метилирования может быть определено для множества частей второго гаплотипа. Каждая часть из множества частей второго гаплотипа может быть комплементарной части из множества частей первого гаплотипа.

На этапе 1097 значение параметра вычисляется с использованием первого уровня метилирования и второго уровня метилирования. Параметр может представлять собой значение разграничения. Значение разграничения может представлять собой разницу между двумя уровнями метилирования или соотношение двух уровней метилирования.

Если используется множество частей второго гаплотипа, то для каждой части из множества частей второго гаплотипа значение разграничения может быть рассчитано с использованием второго уровня метилирования части второго гаплотипа и первого уровня метилирования с использованием комплементарной части первого гаплотипа. Значение разграничения можно сравнить с пороговым значением.

Пороговое значение может быть определено для тканей, не подверженных заболеванию. Параметр может представлять собой количество частей второго гаплотипа, при этом значение разграничения превышает пороговое значение. Например, количество частей второго гаплотипа, для которых значение разграничения превышает пороговое значение, может быть аналогично количеству областей, для которых показано, что они имеют отличие, составляющее больше чем 30% на фиг. 105A, 105B, и 106. На фиг. 105A, 105B и 106, значение разграничения представляет собой соотношение, а пороговое значение составляет 30%. В некоторых вариантах осуществления, пороговое значение может быть определено для тканей, подверженных заболеванию.

В другом примере, значение разграничения для каждой части может быть агрегированным, например, суммированным, что может быть выполнено с помощью взвешенной суммы или суммы функций соответствующих значений разграничения. Такое агрегирование может предоставить значение параметра.

На этапе 1098 значение параметра сравнивают с эталонным значением. Контрольное значение может быть определено с использованием контрольной ткани без нарушения. Контрольное значение может представлять собой значение разграничения. Например, эталонное значение может отображать то, что не должно быть значительной разницы между уровнями метилирования двух гаплотипов. Например, эталонное значение может быть статистической разницей, составляющей 0, или соотношением около 1. Когда используется множество частей, эталонное значение может быть количеством частей в здоровом организме, где два гаплотипа демонстрируют значение разграничения, превышающее пороговое значение. В некоторых вариантах осуществления, эталонное значение может быть определено с использованием контрольной ткани, подверженной болезни.

На этапе 1099 классификацию нарушения в организме выполняют с использованием сравнения значения параметра с эталонным значением. Нарушение может быть определено как присутствующее или имеющее большую вероятность появления, если значение параметра превышает эталонное значение. Нарушение может включать в себя рак. Рак может представлять собой любой тип рака, описанный в данном документе.

Классификация нарушения может представлять собой вероятность возникновения нарушения. Классификация нарушения может включать в себя степень тяжести нарушения. Например, большее значение параметра, указывающее на большее количество частей с дисбалансом гаплотипов, может указывать на более тяжелую форму рака.

Хотя способ, описанный на фиг. 109 включает в себя классификацию нарушения, аналогичные способы могут использоваться для определения любой патологии или признака, которые могут возникнуть в результате дисбаланса уровней метилирования между гаплотипами. Например, уровень метилирования гаплотипа из ДНК плода может быть ниже, чем метилирование гаплотипа из ДНК матери. Уровни метилирования можно использовать для классификации нуклеиновых кислот на материнские или эмбриональне.

Когда нарушение представляет собой рак, разные хромосомные области опухоли могут иметь такие различия в метилировании. В зависимости от того, какие области затрагиваются, может быть предостав-

лено различное лечение. Дополнительно, субъекты, имеющие разные области, демонстрирующие такие различия в метилировании, могут иметь разные прогнозы.

Хромосомные области (части), которые имеют достаточное разграничение (например, больше порогового значения), могут быть идентифицированы как аберрантные (или имеющие аберрантное разграничение). Паттерн аберрантной области (возможно учитывая для какого гаплотипа выше, чем для другого) можно сравнить с эталонным паттреном (например, как определено для субъекта, страдающего раком, возможно конкретного типа рака, или для здорового субъекта). Если два паттерна идентичны в пределах порогового значения (например, меньше указанного числа областей/частей, которые отличаются), когда эталонный паттерн конкретно классифицирован, то субъект может быть идентифицирован как имеющий этот тип нарушения. Такая классификация может включать в себя нарушение импринтинга, например, как описано в данном документе.

VII. Анализ метилирования отдельных молекул для гибридных молекул.

Чтобы дополнительно оценить эффективность и полезность раскрытых в данном документе вариантов осуществления в отношении определения модификаций оснований нуклеиновых кислот, мы искусственно создали гибридные фрагменты ДНК человека и мыши, для которых часть человека была метилирована, а часть мыши - неметилирована, или наоборот. Определение областей соединения гибридных или химерных молекул ДНК может позволить обнаруживать слияния генов для различных нарушений или заболеваний, включая рак.

A. Способы создания гибридных фрагментов ДНК человека и мыши.

В этом разделе описывается создание гибридных фрагментов ДНК, а затем процедура определения профилей метилирования фрагментов.

В одном варианте осуществления, человеческая ДНК была амплифицирована посредством амплификации всего генома, так что исходная сигнатура метилирования в геноме человека была удалена, поскольку амплификация всего генома не сохраняет состояния метилирования. Амплификация всего генома может быть выполнена с использованием устойчивых к экзонуклеазам тиофосфат-модифицированных вырожденных гексамеров в качестве праймеров, которые могут случайным образом связываться по всему геному, что позволяет полимеразе (например, ДНК-полимеразе Phi29) амплифицировать ДНК без термоциклирования. Амплифицированный продукт ДНК будет неметилированным. Амплифицированные молекулы ДНК человека дополнительно обрабатывали M.SssI, метилтрансферазой CpG, которая теоретически полностью метилирует все цитозины в контексте CpG в двухцепочечной, неметилированной или полуметилированной ДНК. Таким образом, такая амплифицированная ДНК человека, обработанная M.SssI, станет метилированными молекулами ДНК.

Напротив, ДНК мыши подвергали полногеномной амплификации, чтобы получить неметилированные фрагменты ДНК мыши.

Фиг. 110 иллюстрирует создание гибридных фрагментов ДНК человек-мышь, у которых человеческая часть метилирована, а мышиная часть неметилирована. Закрашенные кружочки на палочке представляют собой метилированные сайты CpG. Незакрашенные кружки на палочке представляют неметилированные сайты CpG. Толстая полоса 11010 с диагональными полосами представляет метилированную человеческую часть. Толстая полоса 11020 с вертикальными полосками представляет неметилированную мышиную часть.

Для создания гибридных молекул ДНК человек-мышь, в одном варианте осуществления, молекулы ДНК после амплифиции всего генома и обработки M.SssI были дополнительно расщеплены HindIII и NcoI для генерации липких концов для облегчения последующего лигирования. В одном варианте осуществления, метилированные фрагменты ДНК человека дополнительно смешивали с неметилированными фрагментами ДНК мыши в эквимолярном соотношении. Такая смесь ДНК человек-мишь была подвергнута процессу лигирования, который в одном варианте осуществления был опосредован ДНК-лигазой при 20°C в течение 15 мин. Как показано на фиг. 110, эта реакция лигирования будет давать 3 типа образующихся молекул, включающих в себя гибридные молекулы ДНК человек-мышь (a: гибридные фрагменты человек-мышь); молекулы ДНК только человека (b: лигирование человек-человек, и c: ДНК человека без лигирования); и молекулы ДНК только мыши (d: лигирование мышь-мышь, и e: ДНК мыши без лигирования). Продукт ДНК после лигирования подвергали секвенированию отдельной молекулы в реальном времени. Результаты секвенирования анализировали в соответствии с приведенным в данном документе раскрытием изобретения для определения состояний метилирования.

Фиг. 111 иллюстрирует создание гибридных фрагментов ДНК человек-мышь, у которых человеческая часть неметилирована, а мышиная часть метилирована. Закрашенные кружочки на палочке представляют собой метилированные сайты CpG. Незакрашенные кружки на палочке представляют неметилированные сайты CpG. Толстая полоса 11110 с диагональными полосами представляет метилированную мышиную часть. Толстая полоса 11120 с вертикальными полосами представляет неметилированную человеческую часть.

Для варианта осуществления, показанного на фиг. 111, молекулы ДНК мыши были амплифицированы посредством амплификации всего генома, так что исходное метилирование в геноме мыши было устранено. Амплифицированный продукт ДНК будет неметилированным. Амплифицированная ДНК

мыши будет дополнительно обработана M.SssI. Таким образом, такая амплифицированная ДНК мыши, обработанная M.SssI, превращается в метилированные молекулы ДНК. Напротив, фрагменты ДНК человека подвергали полногеномной амплификации, чтобы получить неметилированные фрагменты ДНК. В одном варианте осуществления, метилированные фрагменты человека дополнительно смешивали с неметилированными фрагментами в эквимолярном соотношении. Такую смесь ДНК человек-мышь подвергали процессу лигирования, опосредуемому ДНК-лигазой. Как показано на фиг. 111, эта реакция лигирования будет давать 3 типа образующихся молекул, включающих в себя гибридные молекулы ДНК человек-мышь (a: гибридные фрагменты человек-мышь); молекулы ДНК только человека (b: лигирование человек-человек, и c: ДНК человека без лигирования); и молекулы ДНК только мыши (d: лигирование мышь-мышь, и e: ДНК мыши без лигирования). Продукт ДНК после лигирования подвергали секвенированию отдельной молекулы в реальном времени. Результаты секвенирования анализировали в соответствии с приведенным в данном документе раскрытием изобретения для определения состояний метилирования.

Согласно варианту осуществления, показанному на фиг. 110, мы приготовили искусственную смесь ДНК (названную образцом MIX01), содержащую гибридные молекулы ДНК человек-мышь, ДНК только человека и ДНК только мыши, для которых молекулы ДНК, ассоциированные с человеком, были метилированы, тогда как молекулы ДНК мыши были неметилированы. Для образца MIX01 мы получили 166 миллионов субпрочтений, которые можно было выровнять либо с эталонным геномом человека или мыши, либо частично с геномом человека и частично с геномом мыши. Эти субпрочтения были получены из примерно 5 миллионов ячеек секвенирования отдельной молекулы в реальном времени (SMRT) Pacific Biosciences. Каждая молекула в ячейке секвенирования отдельной молекулы в реальном времени секвенировалась в среднем 32 раза (диапазон: 1-881 раз).

Чтобы определить человеческую часть ДНК и мышиную часть ДНК в гибридном фрагменте, мы сначала сконструировали консенсусные последовательности, объединив нуклеотидную информацию из всех соответствующих субпрочтений в ячейке. Всего мы получили 3435657 консенсусных последовательностей для образца MIX01. Набор данных был сгенерирован из ДНК, полученной с помощью Sequel II Sequencing Kit 1.0.

Консенсусные последовательности выровняли с эталонными геномами, включающие в себя эталонные геномы как человека, так и мыши. Мы получили 3,2 миллиона выровненных консенсусных последовательностей. Среди них, 39,6% из них были отнесены к типу ДНК только человека; 26,5% из них были отнесены к типу ДНК только мыши, а 30,2% из них были классифицированы как гибридная ДНК человек-мышь.

Фиг. 112 демонстрирует распределение длин молекул ДНК в смеси ДНК после лигирования (образец MIX01). Ось абсцисс показывает длину молекулы ДНК. Ось ординат показывает частоту встречаемости, связанную с длиной молекулы ДНК. Как показано на фиг. 112, гибридные молекулы ДНК человек-мышь имеют более длинное распределение длинны, что согласуется с тем фактом, что они представляют собой комбинацию по меньшей мере двух типов молекул.

Фиг. 113 иллюстрирует область соединения, с помощью которой первую ДНК (A) и вторую ДНК (B) соединяют вместе. ДНК (A) и ДНК (B) могут быть расщеплены ферментом рестрикции. В одном варианте осуществления, для повышения эффективности лигирования с применением липких концов мы использовали ферменты рестрикции HindIII и NcoI, распознающие сайты A^AGCTT и C^CATGG соответственно, для расщепления ДНК человека и мыши перед стадией лигирования. Затем можно лигировать ДНК (A) и ДНК (B). Среди 698492 гибридных молекул ДНК человек-мышь, несущих области соединения, мы обнаружили, что 88% молекул гибридной ДНК человек-мышь, несущих сайт распознавания фермента - A^AGCTT и C^CATGG, дополнительно предполагая наличие лигирования между фрагментами ДНК человек-мышь. Указанную область соединения определяют, как область или сайт, посредством которой(ого) первый фрагмент ДНК и второй фрагмент ДНК были физически соединены вместе. Поскольку область соединения включает в себя последовательности, общие как для ДНК (A), так и для ДНК (B), часть одной цепи, соответствующая области соединения, не может быть определена как являющаяся частью либо ДНК (A), либо ДНК (B), только по последовательности. Анализ паттерна или плотности метилирования части одной цепи, соответствующей месту соединения, может быть использован для определения того, происходит ли часть из ДНК (A) или ДНК (B). Например, ДНК (A) может быть вирусной ДНК, а ДНК (B) может быть ДНК человека. Определение полной области соединения может дать информацию о том, нарушает ли такая интегрированная ДНК белковые структуры и каким образом.

Фиг. 114 иллюстрирует анализ метилирования смеси ДНК. Полоса 11410 с диагональными полосками обозначает область соединения, наблюдаемую в анализе выравнивания, которая может быть внесена обработкой ферментом рестрикции перед лигированием. "Сайт RE" обозначает сайт распознавания фермента рестрикции (RE).

Как показано на фиг. 114, в одном варианте осуществления, выровненные консенсусные последовательности были сгруппированы в три категории следующим образом.

(1) Секвенированную ДНК выравнивали только с эталонным геномом человека, но не выравнивали

с эталонным геномом мыши, в отношении одного или нескольких критериев выравнивания. В одном варианте осуществления, один критерий выравнивания может быть определен как, но без ограничения, 100, 95, 90, 80, 70, 60, 50, 40, 30 или 20% смежных нуклеотидов секвенированной ДНК можно сопоставить с человеческим эталоном. В одном варианте осуществления, одним из критериев выравнивания было бы то, что оставшаяся часть секвенированного фрагмента, которая не выровнялась с человеческим эталоном, не могла быть выровнена с эталонным геномом мыши. В одном варианте осуществления, одним из критериев выравнивания было то, что секвенированная ДНК могла быть выровнена с одной областью в эталонном геноме человека. В одном варианте осуществления, выравнивание может быть абсолютным. Тем не менее, в другом варианте осуществления, выравнивание может учитывать расхождения нуклеотидов, включая вставки, несовпадения и делеции, при условии, что такие расхождения были меньше определенных пороговых значений, таких как, но без ограничения, 1, 2, 3, 4, 5, 10, 20 или 30% длины выровненных последовательностей. В другом варианте осуществления, выравнивание может происходить с более чем одним местом в эталонном геноме. В еще других вариантах осуществления, выравнивание с одним или большим количеством сайтов в эталонном геноме может быть установлено вероятностным способом (например, указанием вероятности ошибочного выравнивания), и измерение вероятностей может быть использовано в последующей обработке.

(2) Секвенированную ДНК выравнивали только с эталонным геномом мыши, но не выравнивали с эталонным геномом человека в отношении одного или нескольких критериев выравнивания. В одном варианте осуществления, один критерий выравнивания может быть определен как, но без ограничения, 100, 95, 90, 80, 70, 60, 50, 40, 30 или 20% смежных нуклеотидов секвенированной ДНК можно выровнять с мышиным эталоном. В одном варианте осуществления, одним из критериев выравнивания было бы то, что оставшаяся часть, не могла быть выровнена с эталонным геномом человека. В одном варианте осуществления, одним из критериев выравнивания было то, что секвенированная ДНК могла быть выровнена с одной областью в эталонном геноме мыши. В одном варианте осуществления, выравнивание может быть абсолютным. В еще других вариантах осуществления, выравнивание может учитывать расхождения нуклеотидов, включая вставки, несовпадения и делеции, при условии, что такие расхождения были меньше определенных пороговых значений, таких как, но без ограничения, 1, 2, 3, 4, 5, 10, 20 или 30% длины выровненных последовательностей. В другом варианте осуществления, выравнивание может происходить с более чем одним местом в эталонном геноме. В еще других вариантах осуществления, выравнивание с одним или большим количеством сайтов в эталонном геноме может быть установлено вероятностным способом (например, указанием вероятности ошибочного выравнивания), и измерение вероятностей может быть использовано в последующей обработке.

(3) Одна часть секвенированной ДНК была однозначно выровнена с эталонным геномом человека, тогда как другая часть была однозначно выровнена с эталонным геномом мыши. В одном варианте осуществления, если фермент рестрикции был использован до лигирования, при анализе выравнивания будет наблюдаться область соединения, соответствующая сайту разрезания фермента рестрикции. В некоторых вариантах осуществления, области соединения между частями ДНК человека и мыши можно было определить только примерно в пределах определенной области из-за ошибок секвенирования и выравнивания. В некоторых вариантах осуществления, сайты распознавания фермента рестрикции не будут наблюдаться в областях соединения гибридных фрагментов ДНК человек-мышь, если лигирование затрагивает молекулы, не разрезанные ферментами рестрикции (например, если имело место лигирование тупых концов).

Межимпульсные периоды (МИП), значения ширины импульсов (ШИ) и контекст последовательности, в котором находятся сайты CpG, были получены из тех субпрочтений, которые соответствуют консенсусным последовательностям. Таким образом, метилирование для каждой молекулы ДНК, включая ДНК только человека, только мыши, и гибридную ДНК человек-мышь, может быть определено согласно вариантам осуществления, представленным в данном раскрытии изобретения.

B. Результаты метилирования.

В этом разделе описаны результаты метилирования гибридных фрагментов ДНК. Плотность метилирования может быть использована для определения происхождения различных частей гибридных фрагментов ДНК.

Фиг. 115 демонстрирует диаграмму вероятностей метилирования для сайтов CpG в образце MIX01. На оси абсцисс показаны три различные молекулы, присутствующие в образце MIX01: ДНК только человека, ДНК только мыши, и гибридная ДНК человек-мышь (включает в себя как человеческую часть, так и мышиную часть). По оси ординат показана вероятность метилирования сайта CpG конкретной отдельной молекулы ДНК. Этот анализ проводился таким образом, чтобы ДНК человека была более метилированной, а ДНК мыши - более неметилированной.

Как показано на фиг. 115, вероятность метилирования сайта CpG в ДНК только человека (медиана: 0,66; диапазон: 0-1) была значительно выше, чем таковая для ДНК только мыши (медиана: 0,06; диапазон: 0-1) (p-значение <0,0001). Эти результаты соответствовали дизайну анализа, в котором ДНК человека была более метилированной из-за обработки CpG-метилтрансферазой M.SssI, тогда как ДНК мыши была более неметилированной, поскольку метилирование не могло быть сохранено в процессе амплифи-

кации всего генома. Более того, сайты CpG в человеческой части ДНК в гибридной молекуле ДНК человек-мышь показали более высокую вероятность метилирования (медиана: 0,69; диапазон: 0-1) по сравнению с сайтами в мышиной части ДНК (медиана: 0,06; диапазон: 0-1) (p-значение <0,0001). Эти данные показывают, что раскрытый способ может точно определять статус метилирования молекул ДНК, а также сегментов в пределах молекулы ДНК.

Вероятность метилирования относится к оцененной вероятности для конкретного сайта CpG в пределах отдельной молекулы на основе используемой статистической модели. Вероятность, составляющая 1, указывает на то, что на основе статистической модели 100% сайтов CpG, с использованием определенных параметров (включая МИП, ШИ и контекст последовательности), будут метилированы. Вероятность, составляющая 0, указывает на то, что на основе статистической модели 0% сайтов CpG, с использованием определенных параметров (включая МИП, ШИ и контекст последовательности), будут метилированы. Другими словами, все сайты CpG, с использованием определенных параметров, будут неметилированными. Фиг. 115 демонстрирует распределение вероятностей метилирования, с более широким распределением только для ДНК человека и человеческой части, нежели для соответствующих частей мыши. Бисульфитное секвенирование используется для определения метилирования аналогичных образцов, чтобы подтвердить, что метилирование не было полным, и результаты продемонстрированы ниже. Фиг. 115 демонстрирует значительную разницу между метилированием ДНК человека и мыши.

Согласно варианту осуществления, показанному на фиг. 111, мы приготовили искусственную смесь ДНК (названную образцом MIX02), содержащую гибридные молекулы ДНК человек-мышь, ДНК только человека и ДНК только мыши, для которых человеческая часть являлась неметилированной, а мышиная часть - метилированной. Для образца MIX02 мы получили 140 миллионов субпрочтений, которые можно выровнять либо с эталонным геномом человека, либо мыши, либо частично с геномом человека и частично с геномом мыши. Эти субпрочтения были получены из примерно 5 миллионов ячеек секвенирования отдельной молекулы в реальном времени (SMRT) Pacific Biosciences. Каждая молекула в ячейке секвенирования отдельной молекулы в реальном времени секвенировалась в среднем 27 раз (диапазон: 1-1028 раз).

Мы также сконструировали консенсусные последовательности, объединив нуклеотидную информацию из всех соответствующих субпрочтений в ячейке. Всего мы получили 3265487 консенсусных последовательностей для образца MIX02. Консенсусные последовательности выровняли с эталонными геномами, включающие в себя эталонные геномы как человека, так и мыши, с использованием BWA (Li H et al., Bioinformatics. 2010;26(5):589-595). Мы получили 3,0 миллиона выровненных консенсусных последовательностей. Среди них, 30,5% было отнесено к типу ДНК только человека; 32,2% было отнесено к типу ДНК только мыши, а 33,8% было классифицировано как гибридная ДНК человек-мышь. Набор данных был сгенерирован из ДНК, полученной с помощью Sequel II Sequencing Kit 1.0.

Фиг. 116 демонстрирует распределение длин молекул ДНК в смеси ДНК после перекрестного лигирования образца MIX02. Ось абсцисс показывает длину молекулы ДНК. Ось ординат показывает частоту встречаемости, связанную с длиной молекулы ДНК. Как показано на фиг. 116, гибридные молекулы ДНК человек-мышь имеют более длинное распределение длинны, что согласуется с тем фактом, что они были произведены путем лигирования больше чем одной молекулы.

Фиг. 117 демонстрирует диаграмму вероятностей метилирования для сайтов CpG в образце MIX02. Статус метилирования определяли согласно описанными в данном документе способам. На оси абсцисс показаны три различные молекулы, присутствующие в образце MIX01: ДНК только человека, ДНК только мыши, и гибридная ДНК человек-мышь (включает в себя как человеческую часть, так и мышиную часть). По оси ординат показана вероятность метилирования сайта CpG. Этот анализ проводился таким образом, чтобы ДНК человека была неметилированной, а ДНК мыши - метилированной.

Как показано на фиг. 117, вероятность метилирования сайтов CpG в ДНК только человека (медиана: 0,06; диапазон: 0-1) была значительно ниже, чем таковая для ДНК только мыши (медиана: 0,93; диапазон: 0-1) (p-значение <0,0001). Эти результаты соответствовали дизайну анализа, в котором ДНК человека была более неметилированной, поскольку метилирование не могло быть сохранено в процессе амплификации всего генома, тогда как ДНК мыши была более метилированной, из-за обработки CpG-метилтрансферазой M.SssI. Более того, сайты CpG в человеческой части ДНК в гибридной молекуле ДНК человек-мышь показали более низкую вероятность метилирования (медиана: 0,07; диапазон: 0-1) по сравнению с таковыми в мышиной части ДНК (медиана: 0,93; диапазон: 0-1) (p-значение <0,0001). Эти данные показывают, что раскрытый способ может точно определять статус метилирования молекул ДНК, а также сегментов в пределах молекулы ДНК.

Бисульфитное секвенирование использовали для измерения метилирования гибридных фрагментов человек-мышь, паттерны метилирования которых были определены путем секвенирования отдельной молекулы в реальном времени согласно вариантам осуществления в данном раскрытии изобретения. Образец MIX01 (ДНК человека была метилирована, а ДНК мыши не метилирована) и MIX02 (ДНК человека была неметилирована, а ДНК мыши метилирована) были фрагментированы посредством обработки ультразвуком, что дало смесь с медианным размером фрагмента ДНК, составляющим 196 п.о. (межквартильный диапазон: 161-268). Затем было выполнено бисульфитное секвенирование одинаковых концов

(БС-секв.) на платформе MiSeq (Illumina) с длиной считывания 300 п.о. x2. Мы получили 3,7 миллиона и 2,9 миллиона секвенированных фрагментов для MIX01 и MIX02, соответственно, которые были выровнены с эталонным геномом человека или мыши, или частично с геномом человека и частично с геномом мыши. Для MIX01 41,6% выровненных фрагментов были классифицированы как ДНК только человека, 56,6% как ДНК только мыши, и 1,8% как гибридная ДНК человек-мышь. Для MIX02 61,8% выровненных фрагментов были классифицированы как ДНК только человека, 36,3% как ДНК только мыши, и 1,9% как гибридная ДНК человек-мышь. Процент секвенированных фрагментов, определенных как гибридная ДНК человек-мышь при БС-секв. (<2%), был намного ниже, чем наблюдаемый в результатах секвенирования Pacific Biosciences (>30%). Примечательно, что длинные фрагменты (медианное значение ~2 т.п.о.) были секвенированы с помощью секвенирования Pacific Biosciences, в то время как длинные фрагменты были разделены на короткие фрагменты (в среднем ~196 п.о.), которые подходили для MiSeq. Такой процесс фрагментирования сильно разбавил бы гибридные фрагменты человек-мышь.

Фиг. 118 демонстрирует таблицу, в которой сравнивается метилирование, определенное с помощью бисульфитного секвенирования и секвенирования Pacific Biosciences для MIX01. В самой левой части таблицы показан тип ДНК: 1) только человека; 2) только мышь; и 3) гибрид человек-мышь, разделенный на человеческую и мышиную части. В средней части таблицы показаны детали бисульфитного секвенирования, включая количество сайтов CG и плотность метилирования. В самой правой части таблицы показаны детали секвенирования Pacific Biosciences, включая количество сайтов CG и плотность метилирования.

Как показано на фиг. 118, ДНК только человека постоянно демонстрировала более высокую плотность метилирования, чем ДНК только мыши для MIX01, в результатах как бисульфитного секвенирования, так и секвенирования Pacific Biosciences. Для гибридных фрагментов человек-мышь уровни метилирования человеческой части и мышиной части были определены как 46,8 и 2,3%, соответственно, в результатах бисульфитного секвенирования. Эти результаты подтвердили более высокие плотности метилирования для человеческой части по сравнению с мышиной, как определено с помощью секвенирования Pacific Biosciences согласно данному изобретению. При секвенировании Pacific Biosciences наблюдали плотность метилирования 57,4% в человеческой части, и наблюдали более низкую плотность метилирования 12,1% в мышиной части. Эти результаты предполагают, что метилирование, определенное с помощью секвенирования Pacific Biosciences согласно данному раскрытию изобретения, может быть обоснованным. В частности, секвенирование Pacific Biosciences может быть использовано для определения отличающихся плотностей метилирования, в том числе в ДНК, имеющей участок с более высокой плотностью метилирования, чем в другом участке. Мы заметили, что плотность метилирования, определенная с помощью секвенирования Pacific Biosciences согласно данному изобретению, была выше по сравнению с бисульфитным секвенированием. Такая оценка может быть скорректирована с использованием разницы между результатами, полученными с помощью этих двух технологий, для сравнения результатов по технологиям.

Фиг. 119 демонстрирует таблицу, в которой сравнивается метилирование, определенное с помощью бисульфитного секвенирования и секвенирования Pacific Biosciences для MIX02. В самой левой части таблицы показан тип ДНК: 1) только человека; 2) только мышь; и 3) гибрид человек-мышь, разделенный на человеческую и мышиную части. В средней части таблицы показаны детали бисульфитного секвенирования, включая количество сайтов CG и плотность метилирования. В самой правой части таблицы показаны детали секвенирования Pacific Biosciences, включая количество сайтов CG и плотность метилирования.

Как показано на фиг. 119, ДНК только человека постоянно демонстрировала более низкую плотность метилирования, чем ДНК только мыши для MIX02, в результатах как бисульфитного секвенирования, так и секвенирования Pacific Biosciences. Для гибридных фрагментов человек-мышь уровни метилирования человеческой части и мышиной части были определены как 1,8 и 67,4%, соответственно, в результатах бисульфитного секвенирования. Эти результаты дополнительно подтвердили более низкие плотности метилирования для человеческой части по сравнению с мышиной, как определено с помощью секвенирования Pacific Biosciences согласно данному изобретению. При секвенировании Pacific Biosciences наблюдали плотность метилирования, составляющую 13,1% в человеческой части, и наблюдали более высокую плотность метилирования, составляющую 72,2%, в мышиной части, как определено с помощью секвенирования Pacific Biosciences согласно данному раскрытию изобретения. Также предполагается, что было осуществимо определение метилирования с помощью секвенирования Pacific Biosciences согласно данному изобретению. В частности, секвенирование Pacific Biosciences может быть использовано для определения неодинаковой плотности метилирования, в том числе в ДНК, имеющей участок с более низкой плотностью метилирования, чем в другом участке. Мы также наблюдали, что плотность метилирования, определенная с помощью секвенирования Pacific Biosciences согласно данному изобретению, была выше по сравнению с бисульфитным секвенированием. Такая оценка может быть скорректирована с использованием разницы между результатами, полученными с помощью этих двух технологий, для сравнения результатов по технологиям.

Фиг. 120A демонстрирует уровни метилирования в группах 5 млн.п.о. для ДНК отдельно человека и

отдельно мыши для MIX01. Фиг. 120B демонстрирует уровни метилирования в группах 5 млн.п.о. для ДНК только человека и только мыши для MIX02. Уровень метилирования в процентах на обеих фигурах отложен по оси ординат. Бисульфитное секвенирование и секвенирование Pacific Biosciences для каждой - ДНК только человека и ДНК только мыши показаны на оси абсцисс.

Результаты на фиг. 120A и фиг. 120B, определенные с помощью секвенирования Pacific Biosciences согласно данному изобретению, оказались систематически выше для групп как в образце MIX01, так и в MIX02.

Фиг. 121A демонстрирует уровни метилирования в группах 5 млн.п.о. для человеческой части и мышиной части гибридных фрагментов ДНК человек-мышь для MIX01. Фиг. 121B демонстрирует уровни метилирования в группах 5 млн.п.о. для человеческой части и мышиной части гибридных фрагментов ДНК человек-мышь для MIX02. Уровень метилирования в процентах на обеих фигурах отложен по оси ординат. Бисульфитное секвенирование и секвенирование Pacific Biosciences для каждой - человеческой части ДНК и мышиной части ДНК показаны на оси абсцисс.

Фиг. 121A и фиг. 121B обе продемонстрировали увеличение уровня метилирования при использовании секвенирования Pacific Biosciences по сравнению с бисульфитным секвенированием. Это увеличение аналогично увеличению уровней метилирования при секвенировании Pacific Biosciences, наблюдаемому с ДНК только человека и ДНК только мыши на фиг. 120A и 120B. Повышенная вариабельность уровней метилирования среди групп 5 млн.п.о., представленных в результатах бисульфитного секвенирования для гибридных фрагментов, вероятно, связана с меньшим количеством сайтов CpG, используемых для анализа.

Фиг. 122A и 122B представляют собой иллюстративные графики, демонстрирующие состояния метилирования в отдельной гибридной молекуле человек-мышь. Фиг. 122A демонстрирует гибридный фрагмент человек-мышь в образце MIX01. Фиг. 122B демонстрирует гибридный фрагмент человек-мышь в образце MIX02. Закрашенный кружок обозначает метилированный сайт, а не закрашенный кружок обозначает неметилированный сайт. Состояния метилирования в этих фрагментах определяли согласно вариантам осуществления, описанным в данном документе.

Как показано на фиг. 122A, человеческая часть гибридной молекулы из образца MIX01 была определена как более метилированная. Напротив, было установлено, что мышиная часть ДНК более гипометилирована. Напротив, фиг. 122B демонстрирует, что человеческая часть гибридной молекулы из образца MIX02 была определена как более гипометилированная, тогда как мышиная часть ДНК была определена как более метилированная.

Эти результаты продемонстрировали, что варианты осуществления, представленные в данном раскрытии изобретения, позволили определить изменения метилирования в отдельной молекуле ДНК с различными паттернами метилирования в разных частях молекулы. В одном варианте осуществления, может быть определен статус метилирования гена или других областей генома, в которых разные части гена или области генома будут демонстрировать разный статус метилирования (например, промотора по сравнению с телом гена). В другом варианте осуществления способы, представленные в данном документе, могут обнаруживать гибридные фрагменты человек-мышь, предлагая типичный подход для обнаружения молекул ДНК, содержащих несмежные фрагменты (т.е. химерные молекулы), относительно эталонного генома, и для анализа их состояний метилирования. Например, мы могли бы использовать этот подход для анализа, но без ограничения, слияния генов, геномных перестроек, трансляций, инверсий, дупликаций, вариаций структуры, интеграций вирусной ДНК, мейотических рекомбинаций и т.д.

В некоторых вариантах осуществления, эти гибридные фрагменты могут быть обогащены перед секвенированием с использованием способов гибридизации на основе зондов или систем CRISPR-Cas, или вариантов этих подходов для обогащения целевой ДНК. Недавно сообщалось, что связанная с CRISPR транспозаза цианобактерии Scytonema hofmanni способна вставлять сегменты ДНК в область рядом с целевым сайтом интереса (Strecker et al. Science. 2019;365:48-53). CRISPR-ассоциированная транспозаза может осуществлять подобное Th7-опосредованной транспозиции. В одном варианте осуществления, мы смогли бы приспособить эту CRISPR-ассоциированную транспозазу для вставки ссылочных последовательностей, меченных, например, биотином, в одну или большее количество представляющих интерес геномных областей, с нацеливанием с помощью нРНК. Мы могли бы использовать магнитные гранулы, покрытые, например, стрептавидином, для захвата ссылочных последовательностей, тем самым одновременно извлекая целевые последовательности ДНК для секвенирования и анализа метилирования согласно вариантам осуществления, представленным в данном раскрытии изобретения.

В некоторых вариантах осуществления, фрагменты могут быть обогащены с помощью ферментов рестрикции, которые могут включать в себя любой фермент рестрикции, описанный в данном документе.

C. Пример способа обнаружения химерных молекул.

Фиг. 123 демонстрирует способ 1230 обнаружения химерных молекул в биологическом образце. Химерные молекулы могут включать в себя последовательности двух разных генов, хромосом, органелл (например, митохондрий, ядер, хлоропластов), организмов (млекопитающих, бактерий, вирусов и т.д.) и/или видов. Способ 1230 может применяться к каждой из множества молекул ДНК из биологического образца. В некоторых вариантах осуществления, множество молекул ДНК может представлять собой

клеточную ДНК. В других вариантах осуществления, множество молекул ДНК может представлять собой внеклеточные молекулы ДНК из плазмы беременной женщины.

На этапе 1232 может быть выполнено одномолекулярное секвенирование молекулы ДНК для получения прочтения последовательности, которое дает статус метилирования в каждом из N сайтов. N может составлять 5 или больше, включая 5-10, 10-15, 15-20 или больше 20. Статусы метилирования прочтения последовательности могут формировать паттерн метилирования. Молекула ДНК может представлять собой одну молекулу ДНК из множества молекул ДНК, и способ 1230 может быть выполнен для множества молекул ДНК. Паттерн метилирования может принимать различные формы. Например, паттерн может составлять N (например, 2, 3, 4 и т.д.) метилированных сайтов, за которыми следуют N неметилированных сайтов, или наоборот. Такое изменение метилирования может указывать на область соединения. Количество смежных сайтов, которые метилированы, может отличаться от количества смежных сайтов, которые не метилированы.

На этапе 1234 паттерн метилирования может быть сопоставлен с одним или большим количеством эталонных паттернов, которые соответствуют химерным молекулам, которые имеют две части из двух участков эталонного генома человека. Контрольный паттерн может функционировать в качестве фильтра для идентификации совпадающего паттерна, указывающего на область соединения. Количество сайтов, которые совпадают с эталонным паттерном, можно отслеживать так, чтобы позиция совпадения соответствовала максимальному количеству совпадающих сайтов (т.е. количеству, где статус метилирования соответствует эталонному паттерну). Две части эталонного генома человека могут быть перемежающимися частями эталонного генома человека. Две части эталонного генома человека могут быть разделены больше чем 1 т.п.о., 5 т.п.о., 10 т.п.о., 100 т.п.о., 1 млн.п.о., 5 млн.п.о. или 10 млн.п.о. Эти две части могут быть из двух разных хромосомных плеч или хромосом. Один или большее количество эталонных паттернов могут включать в себя смену между метилированными состояниями и неметилированными состояниями.

На этапе 1236 позиция совпадения может быть идентифицирована между паттерном метилирования и первым эталонным паттерном из одного или большего количества эталонных паттернов. Позиция совпадения может идентифицировать область соединения между двумя частями эталонного генома человека в прочтении последовательности. Данная позиция совпадения может соответствовать максимуму функции перекрытия между эталонным паттерном и паттерном метилирования. Функция перекрытия может использовать множество эталонных паттернов, при этом выходные данные, возможно, представляют собой максимум агрегатной функции (т.е. каждый эталонный паттерн вносит вклад в выходное значение) или единый максимум, который определяется по эталонным паттернам.

На этапе 1238 область соединения может быть получен как местоположение слияния гена в химерной молекуле. Местоположение слияния гена можно сравнить с эталонными местоположениями слияния генов для различных нарушений или заболеваний, включая рак. Организм, из которого получают биологический образец, можно лечить от нарушения или заболевания.

Положение совпадения может выводиться в функцию выравнивания. Местоположение слияния генов может быть уточнено. Уточнение местоположения слияния генов может включать в себя выравнивание первой части прочтения последовательности с первой частью эталонного генома человека. Первая часть может быть перед областью соединения. Уточнение местоположения слияния генов может включать в себя выравнивание второй части прочтения последовательности с второй частью эталонного генома человека. Вторая часть может быть после области соединения. Первая часть эталонного генома человека может располагаться по меньшей мере в 1 т.п.н. от второй части эталонного генома человека. Например, первая часть эталонного генома человека и вторая часть эталонного генома человека могут находиться на расстоянии друг от друга, составляющем от 1,0 до 1,5 т.п.о., от 1,5 до 2,0 т.п.о., от 2,0 до 2,5 т.п.о., от 2,5 до 3,0 т.п.о., от 3 до 5 т.п.о. или больше чем 5 т.п.о.

Области соединения множества химерных молекул можно сравнивать друг с другом для подтверждения местоположения областей слияния генов.

VIII. Вывод.

Мы разработали эффективный подход для прогнозирования уровней модификации оснований (например, метилирования) нуклеиновых кислот при разрешении, составляющем одно основание. Этот новый подход реализует новую схему для одновременного получения информации о кинетике полимеразы вблизи исследуемого основания, контексте последовательности и информации о цепи. Такое новое преобразование кинетики позволило идентифицировать и смоделировать тонкие прерывания, возникающие в кинетических импульсах. По сравнению с предыдущими способами, использовавшими только МИП, новый подход, представленный в данной патентной заявке, значительно улучшил разрешение и точность анализа метилирования. Эта новая схема может быть легко расширена для других целей, например, для обнаружения 5hmC (5-гидроксиметилцитозина), 5fC (5-формилцитозина), 5caC (5-карбоксилцитозина), 4mC (4-метилцитозина), 6mA (N6-метиладенина), 8oxoG (7,8-дигидро-8-оксогуанина), 8oxoA (7,8-дигидро-8-оксоаденина) и других форм модификаций оснований, а также повреждений ДНК. В другом варианте осуществления, эта новая схема (например, преобразование кинетики, аналогичное 2-мерной цифровой матрице, представленной в данной заявке) может использоваться для анализа модификаций

оснований с использованием системы нанопорового секвенирования.

Эта реализация обнаружения метилирования может быть использована для образцов нуклеиновых кислот из различных источников, например, для клеточных нуклеиновых кислот, нуклеиновых кислот из образцов окружающей среды (например, клеточных загрязнений), нуклеиновых кислот из патогенов (например, бактерий и грибов) и вкДНК в плазмы беременных. Это откроет много новых возможностей для геномных исследований и молекулярной диагностики, таких как неинвазивное пренатальное тестирование, обнаружение рака и мониторинг трансплантации. Для неинвазивной пренатальной диагностики на основе вкДНК это новое изобретение сделало возможным одновременное использование аберраций числа копий, размеров, мутаций, концов фрагментов и модификации оснований для каждой молекулы в диагностике без ПНР и экспериментального преобразования до секвенирования, тем самым повышая чувствительность. Дисбалансы уровней метилирования между гаплотипами могут быть обнаружены с использованием описанных в данном документе способов. Такие дисбалансы могут указывать на происхождение молекулы ДНК (например, выделенной из нарушения, например, раковой клетки, выделенной из крови больного раком) или нарушение.

IX. Примеры систем.

Фиг. 124 иллюстрирует систему измерения 12400 согласно варианту осуществления данного изобретения. Показанная система включает в себя образец 12405, такой как молекулы ДНК, в держателе образца 12410, где образец 12405 может контактировать с аналитической системой 12408 для получения сигнала физической характеристики 12415. Примером держателя образца может быть проточная кювета, которая содержит зонды и/или праймеры для анализа, или пробирка, через которую движется капля (с каплей, содержащей аналитическую систему). Физическую характеристику 12415 (например, интенсивность флуоресценции, напряжение или ток) образца определяют с помощью детектора 12420. Детектор 12402 может выполнять измерения с интервалами (например, периодическими интервалами), чтобы получать точки данных, которые составляют сигнал данных. В одном варианте осуществления, аналого-цифровой преобразователь несколько раз преобразует аналоговый сигнал от детектора в цифровую форму. Держатель 12401 образца и детектор 12402 могут формировать устройство анализа, например, устройство секвенирования, которое выполняет секвенирование согласно вариантам осуществления, описанным в данном документе. Сигнал данных 12425 отправляется из детектора 12402 к логической системе 12403. Сигнал данных 12425 может храниться в локальной памяти 12435, внешней памяти 12404 или запоминающем устройстве 12445.

Логическая система 12403 может представлять собой или может включать в себя компьютерную систему, ASIC, микропроцессор и т.д. Она также может включать в себя или быть соединена с дисплеем (например, монитором, светодиодным дисплеем и т.д.) и устройством ввода пользователя (например, мышью, клавиатурой, кнопками и т.д.). Логическая система 12403 и другие компоненты могут быть частью автономной или подключенной к сети компьютерной системы, или они могут быть непосредственно присоединены к или включены в устройство (например, устройство секвенирования), которое включает в себя детектор 12402 и/или держатель образца 12401. Логическая система 12403 также может включать в себя программное обеспечение, которое выполняется в процессоре 12405. Логическая система 12403 может включать в себя машиночитаемый носитель, хранящий инструкции для управления системой 12400, для осуществления любого из способов, описанных в данном документе. Например, логическая система 12403 может подавать команды в систему, которая включает в себя держатель образца 12401, 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 могут выполняться робототехнической системой, например, включая роботизированную руку, которая может использоваться для получения образца и выполнения анализа.

Любая из упомянутых в данном документе компьютерных систем может использовать любое подходящее количество подсистем. Примеры таких подсистем показаны на фиг. 125 в компьютерной системе 10. В некоторых вариантах осуществления, компьютерная система включает в себя одно компьютерное устройство, где подсистемы могут быть компонентами компьютерного устройства. В других вариантах осуществления, компьютерная система может включать в себя множество компьютерных устройств, каждое из которых является подсистемой, с внутренними компонентами. Компьютерная система может включать в себя настольные и портативные компьютеры, планшеты, мобильные телефоны, другие мобильные устройства и облачные системы.

Подсистемы, показанные на фиг. 125 соединены 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, несъёмного 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 данном документе данных могут выводиться из одного компонента в другой компонент, и могут выводиться пользователю.

Компьютерная система может включать в себя множество одинаковых компонентов или подсистем, например, связанных друг с другом с помощью внешнего интерфейса 81, внутреннего интерфейса или через съемные запоминающие устройства, которые могут быть подключены и удалены от одного компонента к другому компоненту. В некоторых вариантах осуществления, 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 в себя множество систем, подсистем или компонентов.

Аспекты вариантов осуществления могут быть реализованы 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 данного изобретения, процессор может включать в себя 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 раскрытия изобретения и идей, представленных в данном документе, специалист в данной области техники будет осведомлен о и хорошо разбираться в других путях и/или способах реализации вариантов осуществления данного изобретения с использованием аппаратных средств, и комбинации аппаратных средств и программного обеспечения.

Любой из программных компонентов или любая из функций, описанных в данной заявке, может быть реализован(а) в виде программного кода, который должен выполняться процессором с использованием любого подходящего языка программирования, такого как, например, Java, C, C++, C#, Objective-C, Swift или скриптового языка программирования, такого как Perl или Python, с использованием, например, стандартных или объектно-ориентированных методов. Программный код может храниться в виде серии инструкций или команд на машиночитаемом носителе для хранения и/или передачи. Подходящий энергонезависимый машиночитаемый носитель может включать в себя оперативную память (ОЗУ), постоянное запоминающее устройство (ПЗУ), магнитный носитель, такой как жесткий диск или гибкий диск, или оптический носитель, такой как 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 могут кодироваться и передаваться с использованием несущих сигналов, адаптированных для передачи через проводные, оптические и/или беспроводные сети, соответствующих множеству протоколов, включая Интернет. По существу, машиночитаемый носитель может быть создан с использованием сигнала данных, закодированного с помощью таких программ. Машиночитаемый носитель, закодированный программным кодом, может быть упакован с совместимым устройством или предоставлен отдельно от других устройств (например, путем загрузки через Интернет). Любой такой машиночитаемый носитель может поставляться с или в составе компьютерного продукта (например, жесткий диск, компакт-диске, или вся компьютерная система), и может поставляться с или в составе различных компьютерных продуктов в пределах системы или сети. Компьютерная система может включать в себя монитор, принтер или другой подходящий дисплей для предоставления пользователю любых из упомянутых в данном документе результатов.

Любой из описанных в данном документе способов может быть полностью или частично выполнен компьютерной системой, содержащей один или большее количество процессоров, которые могут быть сконфигурированы для выполнения этапов. Таким образом, варианты осуществления могут быть направлены на компьютерные системы, сконфигурированные для выполнения этапов любого из способов, описанных в данном документе, возможно, с различными компонентами, выполняющими соответствующий этап или соответствующую группу этапов. Хотя этапы представлены в виде пронумерованных, этапы описанных в данном документе способов могут выполняться в одно и то же время, в разное время, или в другом порядке. Кроме того, фазы этих этапов могут использоваться с фазами других этапов из других способов. Кроме того, весь этап или его фазы могут быть необязательными. Кроме того, любой из этапов любого из способов может выполняться при наличии модулей, блоков, схем или других средств системы для выполнения этих этапов.

Конкретные детали конкретных вариантов осуществления могут быть скомбинированы любым подходящим способом без отступления от сущности и объема вариантов осуществления данного изобре-

тения. Однако другие варианты осуществления изобретения могут указывать на конкретные варианты осуществления, относящиеся к каждому отдельному аспекту, или к конкретным комбинациям этих отдельных аспектов.

Приведенное выше описание иллюстративных вариантов осуществления данного раскрытия изобретения было представлено в целях иллюстрации и описания. Оно не предназначено для того, чтобы быть исчерпывающим или ограничивать раскрытие изобретения описанной точной формой, и в свете изложенного выше возможны многочисленные модификации и варианты.

Указание существительных в единственном и множественном числе служит для обозначения "одного или большего количества" до тех пор, пока специально не указанно инное. При использовании "или" имеется ввиду "включающее или", а не "исключающее или", если специально не указано иное. Ссылка на "первый" компонент не обязательно требует предоставления второго компонента. Более того, ссылка на "первый" или "второй" компонент не ограничивает упомянутый компонент конкретным местоположением, если явно неуказано иное. Под термином "основанный на" имеется в виду "основанный, по меньшей мере, частично на".

Все патенты, заявки на патенты, публикации и описания, упомянутые в данном документе, полностью включены посредством ссылки для всех целей. Ни один из них не рассматривается как предшествующий уровень техники.

Список литературы.

Albert, T.J. *et al.* (2007) Direct selection of human genomic loci by microarray hybridization. *Nat. Methods*, **4**, 903–905.

Beckmann *et al.* (2014) Detecting epigenetic motifs in low coverage and metagenomics settings. *BMC Bioinformatics*, 15(Suppl 9): S16.

Beaulaurier, J. *et al.* (2019) Deciphering bacterial epigenomes using modern sequencing technologies. *Nature Reviews Genetics*, 20:157–172.

Blow, M.J. *et al.* (2016) The Epigenomic Landscape of Prokaryotes. *PLOS Genet.*, **12**, e1005854.

Breiman, L. (2001) Random Forests. *Mach. Learn.*, **45**, 5–32.

Chan, K.C.A. *et al.* (2013) Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, 18761–8.

Clark, T.A. *et al.* (2013) Enhanced 5-methylcytosine detection in single-molecule, real-time sequencing via Tet1 oxidation. *BMC Biol.*, **11**, 4.

Clark, T.A. *et al.* (2012) Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. *Nucleic Acids Res.*, 40:e29.

Eid, J. *et al.* (2009) Real-Time DNA Sequencing from Single Polymerase Molecules. *Science* **323**, 133–138.

Feinberg, A.P. and Irizarry, R.A. (2010) Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proc. Natl. Acad. Sci.*, **107**, 1757–1764.

Feng, Z. *et al.* (2013) Detecting DNA modifications from SMRT sequencing data by modeling sequence context dependence of polymerase kinetic. *PLoS Comput Biol.*, 9:e1002935.

Flusberg, B.A. *et al.* (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods*, **7**, 461–465.

Frommer, M. *et al.* (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci.*, **89**, 1827–1831.

Gai, W. *et al.* (2018) Liver- and colon-specific DNA methylation markers in plasma for investigation of colorectal cancers with or without liver metastases. *Clin. Chem.*, **64**, 1239–1249.

Gouil, Q. *et al.* (2019) Latest techniques to study DNA methylation. *Essays Biochem.* 63(6):639-648.

Grunau, C. (2001) Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. *Nucleic Acids Res.*, **29**, 65e – 65.

Herman, J.G. *et al.* (1996) Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. *Proc. Natl. Acad. Sci. U. S. A.*, **93**, 9821–9826.

Jiang, P. *et al.* (2014) Methy-Pipe: An Integrated Bioinformatics Pipeline for Whole Genome Bisulfite Sequencing Data Analysis. *PLoS One*, **9**, e100360.

LeCun, Y. *et al.* (1989) Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Comput.*, **1**, 541–551.

Lee, E.-J. *et al.* (2011) Targeted bisulfite sequencing by solution hybrid selection and

massively parallel sequencing. *Nucleic Acids Res.*, **39**, e127–e127.

Lehmann-Werman, R. *et al.* (2016) Identification of tissue-specific cell death using methylation patterns of circulating DNA. *Proc. Natl. Acad. Sci.*, **113**, E1826–E1834.

Lister, R. *et al.* (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**, 315–322.

Liu, Q. *et al.* (2019) Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. *Nature Commun.*, **10**, 2449.

Liu, Y. *et al.* (2019) Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. *Nat. Biotechnol.*, **37**, 424–429.

Lun, F.M.F. *et al.* (2013) Noninvasive prenatal methylomic analysis by genomewide bisulfite sequencing of maternal plasma DNA. *Clin. Chem.*, **59**, 1583–1594.

Nattestad, M. *et al.* (2018) Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. *Genome Res.*, **28**, 1126–1135.

Ng, A.Y. (2004) Feature selection, L_1 vs. L_2 regularization, and rotational invariance. In, *Twenty-first International Conference on Machine Learning - ICML '04*. ACM Press, New York, New York, USA, p. 78.

Ni, P. *et al.* (2019) DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning. *Bioinformatics*,**35**, 4586-4595

Okou, D.T. *et al.* (2007) Microarray-based genomic selection for high-throughput resequencing. *Nat. Methods*, **4**, 907–909.

Olova, N. *et al.* (2018) Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. *Genome Biol.*, **19**, 33.

Robertson, K.D. (2005) DNA methylation and human disease. *Nat. Rev. Genet.*, **6**, 597–610.

Smith, Z.D. and Meissner, A. (2013) DNA methylation: roles in mammalian development. *Nat. Rev. Genet.*, **14**, 204–20.

Schadt, E.E. *et al.* (2013) Modeling kinetic rate variation in third generation DNA sequencing data to detect putative modifications to DNA bases. *Genome Res.*, 23(1):129-41.

Sun, K. *et al.* (2015) Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proc. Natl. Acad. Sci.*, **112**, E5503–E5512.

Suzuki, Y. *et al.* (2016) AgIn: measuring the landscape of CpG methylation of individual repetitive elements. *Bioinformatics*, **32**, 2911–2919.

Watson, C.M. *et al.* (2019) Cas9-based enrichment and single-molecule sequencing for precise characterization of genomic duplications. *Lab. Investig*, **100**, 135-146.

Zhang, W. *et al.* (2015) Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. *Genome Biol.*, **16**, 14.

ФОРМУЛА ИЗОБРЕТЕНИЯ

1. Способ обнаружения модификации нуклеотида в молекуле нуклеиновой кислоты, включающий в себя:

получение входной структуры данных, причем входная структура данных соответствует окну секвенированных нуклеотидов в молекуле нуклеиновой кислоты образца, при этом молекулу нуклеиновой кислоты образца секвенируют путем измерения импульсов оптического сигнала, соответствующего нуклеотидам, при этом входная структура данных содержит значения для следующих свойств:

для каждого нуклеотида:

тип нуклеотида;

позицию нуклеотида в пределах нуклеиновой кислоты образца;

ширину импульса, соответствующего нуклеотиду; и

межимпульсный период, представляющий время между импульсом, соответствующим нуклеотиду, и импульсом, соответствующим соседнему нуклеотиду;

введение входной структуры данных в модель, при этом модель тренировали путем:

получения первого множества первых структур данных, причем каждая первая структура данных из первого множества первых структур данных соотноситься с соответствующим окном секвенированных нуклеотидов в соответствующей молекуле нуклеиновой кислоты из множества первых молекул нуклеиновой кислоты, при этом каждую из первых молекул нуклеиновой кислоты секвенируют путем измерения импульсов в сигнале соответствующих нуклеотидов, при этом модификация имеет известное первое состояние на нуклеотиде в целевой позиции в каждом окне каждой первой молекулы нуклеиновой кислоты, при этом каждая первая структура данных содержит значения для тех же свойств, что и входная структура данных;

сохранения множества первых обучающих выборок, каждая из которых включает в себя одну из первого множества первых структур данных и первую метку, обозначающую первое состояние нуклеотида в целевой позиции; и

оптимизации с использованием множества первых обучающих выборок, параметров модели на основе выходных данных модели, совпадающих или не совпадающих с соответствующими метками из первых меток, когда первое множество первых структур данных вводится в модель, при этом выходные данные модели указывают на то, имеет ли модификацию нуклеотид в целевой позиции в соответствующем окне; и

определение с использованием модели того, присутствует ли модификация на нуклеотиде в целевой позиции в пределах окна в входной структуре данных.

2. Способ по п.1, отличающийся тем, что:

входная структура данных представляет собой одну входную структуру данных из множества входных структур данных;

молекула нуклеиновой кислоты образца представляет собой одну молекулу нуклеиновой кислоты образца из множества молекул нуклеиновой кислоты образца;

множество молекул нуклеиновой кислоты образца получают из биологического образца субъекта; и

каждая входная структура данных соотноситься с соответствующим окном нуклеотидов, секвенированных в соответствующей молекуле нуклеиновой кислоты образца из множества молекул нуклеиновой кислоты образца; и

способ дополнительно включает в себя:

получение множества входных структур данных;

ввод множества входных структур данных в модель; и

определение с использованием модели, присутствует ли модификация в нуклеотиде в целевом местоположении в соответствующем окне каждой входной структуры данных.

3. Способ по п.2, дополнительно включающий в себя:

определение того, присутствует ли модификация на одном или большем количестве нуклеотидов; и

определение классификации нарушения у субъекта с использованием наличия модификации на одном или большем количестве нуклеотидов, где молекула нуклеиновой кислоты образца получена от субъекта.

4. Способ по п.3, где нарушение включает в себя рак.

5. Способ по п.3, где способ дополнительно включает определение того, что классификация нарушения соответствует тому, что субъект имеет нарушение.

6. Способ по п.3, где определение классификации нарушения использует количество модификаций или сайтов модификации.

7. Способ по любому из пп.1-6, где модификация представляет собой метилирование.

8. Способ по п.7, где метилирование включает в себя 5mC (5-метилцитозин).

9. Способ по п.7, где метилирование включает в себя 6mA (N6-метиладенин).

10. Способ по п.2, дополнительно включающий в себя:

определение статуса метилирования как оценки того, присутствует ли модификация на одном или большем количестве нуклеотидов; и

определение клинически значимой фракции ДНК, профиля метилирования плода, профиля метилирования матери, наличия подверженной импринтингу области гена, или тканевого происхождения, используя статус метилирования в одном или большем количестве нуклеотидов.

11. Способ по п.10, где:

способ включает определение происхождения ткани; и

определение происхождения ткани включает определение происходит ли нуклеиновая кислота образца от плода или матери.

12. Способ по п.11, где определение происходит ли нуклеиновая кислота образца от плода или матери включает:

определение уровня метилирования нуклеиновой кислоты образца используя статусы метилирования одного или нескольких нуклеотидов; и

сравнение уровня метилирования молекулы нуклеиновой кислоты образца с эталонным значением.

13. Способ по п.11 или 12, где эталонное значение определяется из уровня метилирования одной или более молекул нуклеиновых кислот матери.

14. Способ по п.11 или 12, где:

сравнение уровня метилирования молекулы нуклеиновой кислоты образца с эталоном включает определение, что уровень метилирования молекулы нуклеиновой кислоты образца ниже чем эталонное значение; и

определение, происходит ли нуклеиновая кислота образца от плода или матери включает определение, что молекула нуклеиновой кислоты произошла от плода в данном сравнении.

15. Способ по п.2, где модификация представляет собой метилирование и где способ дополнительно включает:

идентификацию каждой молекулы нуклеиновой кислоты образца из множества молекул нуклеиновых кислот образца как выравнивающуюся с областью в геноме;

определение с использованием данной модели статуса метилирования как оценки того, присутствует ли модификация на одном или большем количестве нуклеотидов для каждой молекулы нуклеиновой кислоты образца из множества молекул нуклеиновых кислот образца;

определение уровня метилирования для данной области генома с использованием множества статусов метилирования одного или нескольких нуклеотидов из множества молекул нуклеиновых кислот образца; и

определение присутствует ли аберрация числа копий в данной области генома с использованием уровня метилирования.

16. Способ по п.15, дополнительно включающий сравнение уровня метилирования данной области с эталонным значением, где определение присутствует ли аберрация числа копий в данной области включает использование данного сравнения.

17. Способ по п.16, где эталонное значение определяют с использованием области не имеющей такого типа аберрации числа копий.

18. Способ по любому из пп.15-17, где областью является хромосома, а субъект является женщиной, беременной плодом, а данный способ дополнительно включает:

определение наличия аберрации числа копий, и

определение, что у плода есть анеуплоидия хромосом.

19. Способ по любому из пп.2-18, отличающийся тем, что каждая молекула нуклеиновой кислоты образца из множества молекул нуклеиновой кислоты образца может иметь размер, больший чем пороговое значение размера.

20. Способ по любому из пп.1-11, где нуклеотиды в пределах окна определяются с помощью кольцевой консенсусной последовательности и без выравнивания последовательности нуклеотидов с эталонным геномом.

21. Способ по любому из пп.1-11, где нуклеотиды в пределах окна определяются без помощи кольцевой консенсусной последовательности и без выравнивания последовательности нуклеотидов с эталонным геномом.

22. Способ по любому из пп.2-11, отличающийся тем, что:

множество молекул нуклеиновой кислоты образца выравнивается с множеством областей генома;

для каждой области генома из множества областей генома:

ряд молекул нуклеиновой кислоты образца выравнивают с геномной областью, и

количество молекул нуклеиновой кислоты образца является большим, чем пороговое значение.

23. Способ по любому из пп.1-22, отличающийся тем, что модель включает в себя модель машинного обучения, анализ главных компонентов, сверточную нейронную сеть или логистическую регрессию.

24. Способ по любому из пп.1-23, отличающийся тем, что:

окно нуклеотидов, соответствующее входной структуре данных, содержит нуклеотиды первой цепи молекулы нуклеиновой кислоты образца и нуклеотиды второй цепи молекулы нуклеиновой кислоты об-

разца; и

входная структура данных дополнительно содержит для каждого нуклеотида, в пределах окна, значение свойства цепи, при этом свойство цепи указывает на то, что нуклеотид находиться либо в первой цепи, либо во второй цепи.

25. Способ по п.24, где молекула нуклеиновой кислоты образца представляет собой кольцевую молекулу ДНК, сформированную путем:

разрезания двухцепочечной молекулы ДНК с использованием комплекса Cas9 для формирования разрезанной двухцепочечной молекулы ДНК; и

лигирования адаптера в виде шпильки к концу разрезанной двухцепочечной молекулы ДНК.

26. Способ по любому из пп.1-25, отличающийся тем, что каждый нуклеотид в пределах окна обогащают или отфильтровывают.

27. Способ по п.26, где каждый нуклеотид в пределах окна обогащают путем:

разрезания двухцепочечной молекулы ДНК с использованием комплекса Cas9 для формирования разрезанной двухцепочечной молекулы ДНК, и лигирования адаптера в виде шпильки к концу разрезанной двухцепочечной молекулы ДНК; или

фильтруют путем:

отбора двухцепочечных молекул ДНК, имеющих размер в диапазоне размеров.

28. Способ по п.1, далее включающий использование наличия метилирования для идентификации нарушений у субъекта, для выявления и прогнозирования опухолей или злокачественных новообразований субъекта, для определения ткани, из которой происходит молекула нуклеиновой кислоты образца, или для идентификации химерной или гибридной ДНК, где молекула нуклеиновой кислоты образца получена от субъекта.

29. Способ по любому из пп.1-28, где по меньшей мере, некоторые из множества первых молекул нуклеиновой кислоты каждая содержат первую часть, соответствующую первой эталонной последовательности, и вторую часть, соответствующую второй эталонной последовательности, которая не пересекается с первой эталонной последовательностью.

30. Способ по любому из пп.1-29, далее содержащий:

проверку модели с использованием множества химерных молекул нуклеиновых кислот, каждая из которых содержит первую часть, соответствующую первой эталонной последовательности, и вторую часть, соответствующую второй эталонной последовательности, при этом первая часть имеет первый паттерн метилирования, а вторая часть имеет второй паттерн метилирования.

31. Способ по п.29 или 30, где первую часть обрабатывают метилазой.

32. Способ по п.31, где вторая часть соответствует неметилированной части второй эталонной последовательности.

33. Способ по п.29 или 30, где первая эталонная последовательность является человеческой, а вторая эталонная последовательность принадлежит другому животному.

34. Способ по любому из пп.1-33, где окно включает по меньшей мере три нуклеотида выше целевого положения внутри окна.

35. Способ по любому из пп.1-33, где окно, соответствующее структуре входных данных имеет иное число последовательных нуклеотидов выше нуклеотида в целевом положении чем число последовательных нуклеотидов ниже нуклеотида в целевом положении.

36. Способ по любому из пп.1-33, где окно, соответствующее структуре входных данных содержит по меньшей мере 6 последовательных нуклеотидов выше нуклеотида в целевом положении и по меньшей мере 6 последовательных нуклеотидов ниже нуклеотида в целевом положении.

37. Способ по любому из пп.1-33, где окно, соответствующее структуре входных данных содержит по меньшей мере 10 последовательных нуклеотидов выше нуклеотида в целевом положении и по меньшей мере 10 последовательных нуклеотидов ниже нуклеотида в целевом положении

38. Способ по любому из пп.1-33, где окно, соответствующее структуре входных данных содержит по меньшей мере 21 последовательных нуклеотида выше нуклеотида в целевом положении и по меньшей мере 21 последовательных нуклеотида ниже нуклеотида в целевом положении.

39. Способ по любому из пп.1-34 далее включающий секвенироввание нуклеиновой кислоты образца.

40. Способ по п.39, где секвенирование нуклеиновой кислоты образца включает измерение импульсов в оптическом сигнале соответствующих нуклеотидов в нуклеиновой кислоте образца.

41. Способ обнаружения модификации нуклеотида в молекуле нуклеиновой кислоты, включающий в себя:

получение первого множества первых структур данных, причем каждая первая структура данных из первого множества структур данных соотноситься с соответствующим окном секвенированных нуклеотидов в соответствующей молекуле нуклеиновой кислоты из множества первых молекул нуклеиновой кислоты, при этом каждую из первых молекул нуклеиновой кислоты секвенируют путем измерения импульсов в оптическом сигнале соответствующих нуклеотидов, при этом модификация имеет известное первое состояние на нуклеотиде в целевой позиции в каждом окне каждой первой молекулы нуклеино-

вой кислоты, при этом каждая первая структура данных содержит значения для следующих свойств:

для каждого нуклеотида:

тип нуклеотида;

позицию нуклеотида по отношению в пределах первой последовательности нуклеиновой кислоты,

ширину импульса, соответствующего нуклеотиду; и

межимпульсный период, представляющий время между импульсом, соответствующим нуклеотиду, и импульсом, соответствующим соседнему нуклеотиду;

сохранение множества первых обучающих выборок, каждая из которых включает в себя одну из первого множества первых структур данных и первую метку, обозначающую первое состояние для модификации нуклеотида в целевой позиции; и

обучение модели с использованием множества первых обучающих выборок путем оптимизации параметров модели на основе выходных данных модели, совпадающих или не совпадающих с соответствующими метками из первых меток, когда первое множество первых структур данных вводится в модель, при этом выходные данные модели указывают на то, имеет ли модификацию нуклеотид в целевой позиции в соответствующем окне.

42. Способ по п.41, дополнительно включающий в себя:

получение второго множества вторых структур данных, причем каждая из вторых структура данных из второго множества вторых структур данных соотноситься с соответствующим окном секвенированных нуклеотидов в соответствующей молекуле нуклеиновой кислоты из множества вторых молекул нуклеиновой кислоты, при этом модификация имеет второе известное состояние на нуклеотиде в целевой позиции в каждом окне каждой из вторых молекул нуклеиновой кислоты, при этом каждая из вторых структур данных содержит значения для тех же свойств, что и первое множество первых структур данных;

сохранение множества вторых обучающих выборок, каждая из которых включает в себя одну из второго множества вторых структур данных и вторую метку, обозначающую второе состояние нуклеотида в целевой позиции; и

в обучении:

первое состояние или второе состояние представляет собой то, что модификация присутствует, а другое состояние - то, что модификация отсутствует, и

модель дополнительно включает в себя использование множества вторых обучающих выборок путем оптимизации параметров модели на основе выходных данных модели, совпадающих или не совпадающих с соответствующими метками из вторых меток, когда второе множество вторых структур данных представляет собой входные данные для модели.

43. Способ по п.42, где множество первых молекул нуклеиновой кислоты является тем же, что и множество вторых молекул нуклеиновой кислоты.

44. Способ по п.42, где модификация включает в себя метилирование, причем множество первых молекул нуклеиновой кислоты генерируют с использованием амплификации множественного вытеснения с метилированными нуклеотидами первого типа, и при этом множество вторых молекул нуклеиновой кислоты генерируют с использованием амплификации множественного вытеснения с неметилированными нуклеотидами первого типа.

45. Способ по п.1 или 41, отличающийся тем, что оптический сигнал представляет собой сигнал флуоресценции нуклеотида, меченного красителем.

46. Способ по п.1 или 41, отличающийся тем, что каждое окно, связанное с первым множеством первых структур данных, содержит по меньшей мере 4 последовательных нуклеотида из первой цепи каждой из первых молекул нуклеиновой кислоты.

47. Способ по п.1 или 41, отличающийся тем, что каждое окно, связанное с первым множеством первых структур данных, содержит по меньшей мере 6 последовательных нуклеотида из первой цепи каждой из первых молекул нуклеиновой кислоты.

48. Способ по п.1 или 41, отличающийся тем, что каждое окно, связанное с первым множеством первых структур данных, содержит по меньшей мере 13 последовательных нуклеотида из первой цепи каждой из первых молекул нуклеиновой кислоты.

49. Способ по п.46, отличающийся тем, что окна, связанные с первым множеством структур данных, содержат одинаковое количество последовательных нуклеотидов.

50. Способ по п.41, отличающийся тем, что:

каждое окно, связанное с первым множество структур данных, содержит нуклеотиды из первой цепи первой молекулы нуклеиновой кислоты и нуклеотиды из второй цепи первой молекулы нуклеиновой кислоты; и

каждая первая структура данных дополнительно содержит для каждого нуклеотида в пределах окна значение свойства цепи, при этом свойство цепи указывает на то, что нуклеотид находиться либо в первой цепи, либо во второй цепи.

51. Способ по п.41, отличающийся тем, что соседний нуклеотид представляет собой прилегающий нуклеотид.

52. Способ по п.41, отличающийся тем, что ширина импульса - это ширина импульса на половине максимального значения импульса.

53. Способ по п.41, где межимпульсный период может представлять собой время между максимальным значением импульса, связанного с нуклеотидом, и максимальным значением импульса, связанного с соседним нуклеотидом.

54. Способ по п.41, отличающийся тем, что модель включает в себя сверточную нейронную сеть, содержащую:

набор сверточных фильтров, сконфигурированных для фильтрации первого множества структур данных;

входной слой, сконфигурированный для приема отфильтрованного первого множества структур данных;

множество скрытых слоев, включая множество узлов, первый слой множества скрытых слоев, связанный с входным слоем; и

выходной слой, соединенный с последним слоем множества скрытых слоев, и сконфигурированный для вывода выходной структуры данных, причем выходная структура данных содержит свойства.

55. Способ по п.1 или 41, отличающийся тем, что модификация включает в себя метилирование нуклеотида в целевой позиции.

56. Способ по п.55, где известные первые состояния могут включать в себя метилированное состояние для первой части первых структур данных и неметилированное состояние для второй части первых структур данных.

57. Способ по п.55, где метилирование включает в себя 4mC (N3-метилцитозин), 5hmC (5-гидроксиметилцитозин), 5fC (5-формилцитозин), 5caC (5-карбоксилцитозин), 1mA (N1-метиладенин), 3mA (N3-метиладенин), 7mA (N7-метиладенин), 3mC (N3-метилцитозин), 2mG (N2-метилгуанин), 6mG (O6-метилгуанин), 7mG (N7-метилгуанин), 3mT (N3-метилтимин) и 4mT (O4-метилтимин).

58. Способ по п.55, где метилирование включает в себя 5mC (5-метилцитозин).

59. Способ по п.55, где метилирование включает в себя 6mA (N6-метиладенин)

60. Способ по п.41, отличающийся тем, что модификация включает в себя изменение окисления.

61. Способ по п.41, отличающийся тем, что каждая структура данных дополнительно содержит значение высоты импульса, соответствующего каждому нуклеотиду в пределах окна.

62. Способ по п.41, отличающийся тем, что оптический сигнал, соответствующий нуклеотидам, происходит от нуклеотидов или тэга, соединенного с нуклеотидами.

63. Способ по п.41, отличающийся тем, что каждая целевая позиция является центром соответствующего окна.

64. Способ по п.41, отличающийся тем, что модификация отсутствует в каждом окне каждой первой молекулы нуклеиновой кислоты.

65. Способ по п.41, отличающийся тем, что:

каждая первая структура данных из множества первых структур данных исключает первые молекулы нуклеиновой кислоты с межимпульсным периодом или шириной импульса ниже порогового значения.

66. Способ по п.41, отличающийся тем, что:

модификация включает в себя метилирование, и

множество первых обучающих выборок генерируют путем:

амплификации множества молекул нуклеиновой кислоты с использованием набора нуклеотидов, при этом набор нуклеотидов включает в себя 6mA в заданном соотношении.

67. Способ по п.66, где метилирование включает в себя 6mA (N3-метиладенин).

68. Способ по п.1 или 41, отличающийся тем, что, по меньшей мере, некоторые из множества первых молекул нуклеиновой кислоты каждая содержат первую часть, соответствующую первой эталонной последовательности, и вторую часть, соответствующую второй эталонной последовательности, которая не пересекается с первой эталонной последовательностью.

69. Способ по п.1 или 41, дополнительно включающий в себя:

проверку модели с использованием множества химерных молекул нуклеиновых кислот, каждая из которых содержит первую часть, соответствующую первой эталонной последовательности, и вторую часть, соответствующую второй эталонной последовательности, при этом первая часть имеет первый паттерн метилирования, а вторая часть имеет второй паттерн метилирования.

70. Способ по п.68 или 69, отличающийся тем, что первую часть обрабатывают метилазой.

71. Способ по п.70, где вторая часть соответствует неметилированной части второй эталонной последовательности.

72. Способ по п.68 или 69, отличающийся тем, что первая эталонная последовательность является человеческой, а вторая эталонная последовательность принадлежит другому животному.

73. Способ анализа биологического образца из организма, имеющего первый гаплотип и второй гаплотип в первой хромосомной области, причем биологический образец содержит молекулы ДНК, при этом способ включает в себя:

анализ множества молекул ДНК из биологического образца, причем анализ молекулы ДНК включает в себя:

определение местоположения молекулы ДНК в эталонном геноме человека;

определение соответствующего аллеля молекулы ДНК; и

определение того, метилирована ли молекула ДНК в одном или большем количестве геномных сайтов;

идентификацию одного или большего количества гетерозиготных локусов первой части первой хромосомной области, каждый гетерозиготный локус содержит соответствующий первый аллель в первом гаплотипе и соответствующий второй аллель во втором гаплотипе;

идентификацию первого набора из множества молекул ДНК, каждая из которых:

расположена в любом одном или большем количестве гетерозиготных локусов;

содержит соответствующий первый аллель гетерозиготного локуса; и

содержит по меньшей мере один из N геномных сайтов, где N является целым числом, которое больше чем или равное единице;

определение первого уровня метилирования первой части первого гаплотипа с использованием первого набора множества молекул ДНК;

идентификацию второго набора из множества молекул ДНК, каждая из которых:

расположена в любом одном или большем количестве гетерозиготных локусов;

содержит соответствующий второй аллель гетерозиготного локуса; и

содержит по меньшей мере один из N геномных сайтов;

определение второго уровня метилирования первой части второго гаплотипа с использованием второго набора множества молекул ДНК;

вычисление значения параметра с использованием первого уровня метилирования и второго уровня метилирования;

сравнение значения параметра с эталонным значением; и

определение классификации нарушения в организме с использованием сравнения значения параметра с эталонным значением.

74. Способ по п.73, отличающийся тем, что первый уровень метилирования определяют с использованием уровней метилирования одной цепи первого набора множества молекул ДНК, и при этом второй уровень метилирования определяют с использованием уровней метилирования одной цепи второго набора множества молекул ДНК.

75. Способ по п.73, отличающийся тем, что первый уровень метилирования определяют с использованием уровней метилирования одной двухцепочечной молекулы ДНК первого набора из множества молекул ДНК, и при этом второй уровень метилирования определяют с использованием уровней метилирования одной двухцепочечной молекулы ДНК второго набора из множества молекул ДНК.

76. Способ по п.73, отличающийся тем, что заболевание представляет собой рак.

77. Способ по п.73, отличающийся тем, что параметр представляет собой значение разграничения.

78. Способ по п.73, дополнительно включающий в себя:

определение множества первых уровней метилирования для множества частей первого гаплотипа;

определение множества вторых уровней метилирования для множества частей второго гаплотипа, при этом каждая часть из множества частей второго гаплотипа является комплементарной части из множества частей первого гаплотипа;

для каждой части из множества частей второго гаплотипа:

вычисление значения разграничения с использованием второго уровня метилирования части второго гаплотипа и первого уровня метилирования комплементарной части первого гаплотипа; и

сравнение значения разграничения с пороговым значением;

при этом:

первая часть первого гаплотипа является комплементарной первой части второго гаплотипа; и

параметр включает в себя ряд частей второго гаплотипа, где значение разграничения превышает пороговое значение.

79. Способ по п.78, отличающийся тем, что пороговое значение определяют по тканям, не подверженных нарушению.

80. Способ по п.78, отличающийся тем, что каждая часть из множества частей первого гаплотипа имеет длину больше чем или равную 5 т.п.н.

81. Способ по п.73, дополнительно включающий в себя:

определение множества первых уровней метилирования для множества частей первого гаплотипа;

определение множества вторых уровней метилирования для множества частей второго гаплотипа, при этом каждая часть из множества частей второго гаплотипа является комплементарной части из множества частей первого гаплотипа;

для каждой части из множества частей второго гаплотипа:

вычисление значения разграничения с использованием второго уровня метилирования части второго гаплотипа и первого уровня метилирования комплементарной части первого гаплотипа;

при этом:

первая часть первого гаплотипа является комплементарной первой части второго гаплотипа; и

параметр включает в себя сумму значений разграничения.

82. Способ по п.73, дополнительно включающий в себя:

определение множества первых уровней метилирования для множества частей первого гаплотипа;

определение множества вторых уровней метилирования для множества частей второго гаплотипа, при этом каждая часть из множества частей второго гаплотипа является комплементарной части из множества частей первого гаплотипа;

для каждой части из множества частей второго гаплотипа:

вычисление значения разграничения с использованием второго уровня метилирования части второго гаплотипа и первого уровня метилирования комплементарной части первого гаплотипа; и

сравнение значения разграничения с пороговым значением, чтобы определить, имеет ли часть аберрантное разграничение между первым уровнем метилирования и вторым уровнем метилирования;

при этом определение классификации нарушения в организме включает в себя сравнение паттерна частей, имеющих аберрантное разграничение, с эталонным паттерном.

83. Способ по п.73, отличающийся тем, что классификация нарушения представляет собой вероятность нарушения.

84. Способ по п.73, отличающийся тем, что:

первая часть первого гаплотипа и первая часть второго гаплотипа формируют кольцевую молекулу ДНК; и

определение первого уровня метилирования первой части первого гаплотипа включает в себя использование данных из кольцевой молекулы ДНК.

85. Способ по п.84, где кольцевую молекулу ДНК формируют путем:

разрезания двухцепочечной молекулы ДНК с использованием комплекса Cas9 для формирования разрезанной двухцепочечной молекулы ДНК; и

лигирования адаптера в виде шпильки к концу разрезанной двухцепочечной молекулы ДНК.

86. Способ по п.73, отличающийся тем, что первая часть первого гаплотипа длиннее чем или равная 1 т.п.н.

87. Способ по п.73, отличающийся тем, что эталонное значение определяют с использованием эталонной ткани без нарушения.

88. Способ по п.73, отличающийся тем, что нарушение является нарушением импринтинга.

89. Способ обнаружения химерных молекул в биологическом образце, включающий в себя:

для каждой из множества молекул ДНК из биологического образца:

выполнение одномолекулярного секвенирования молекулы ДНК для получения прочтения последовательности, которое дает статус метилирования в каждом из N сайтов, причем N составляет 5 или больше, при этом статусы метилирования прочтения последовательности формируют паттерн метилирования;

сопоставление паттерна метилирования с одним или большим количеством эталонных паттернов, которые соответствуют химерным молекулам, которые имеют две части из двух участков эталонного генома человека, при этом один или большее количество эталонных паттернов включают в себя смену между метилированными состояниями и неметилированными состояниями; и

идентификацию позиции совпадения между паттерном метилирования и первым эталонным паттерном одного или большего количества эталонных паттернов, при этом позиция совпадения определяет область соединения между двумя частями из эталонного генома человека в прочтении последовательности; и

вывод области соединения как места слияния генов в химерной молекуле.

90. Способ по п.89, отличающийся тем, что позиция совпадения выводится в функцию выравнивания, при этом способ дополнительно включает в себя:

уточнение местоположения слияния генов путем:

выравнивания первой части прочтения последовательности с первой частью эталонного генома человека, причем первая часть находится перед областью соединения; и

выравнивание второй части прочтения последовательности со второй частью эталонного генома человека, при этом вторая часть находится после области соединения, при этом первая часть эталонного генома человека находится на расстоянии в 1 т.п.о. от второй части эталонного генома.

91. Способ по п.89, дополнительно включающий в себя сравнение областей соединения химерных молекул друг с другом для подтверждения местоположения слияния генов.

92. Машиночитаемый носитель, хранящий множество инструкций, которые при выполнении контролируют компьютерную систему для выполнения способа по любому пп.1-91.

93. Система для осуществления способа по любому из пп.1-91, включающая в себя:

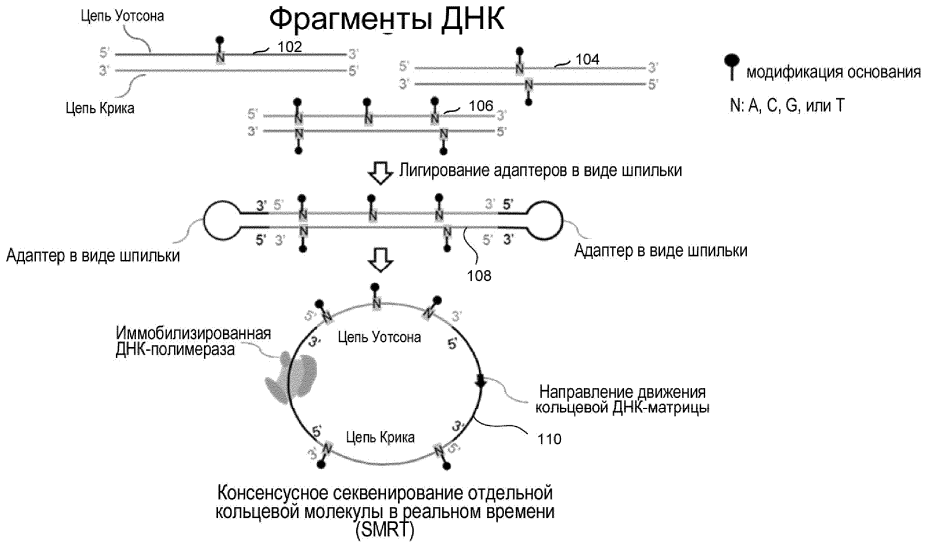
машиночитаемый носитель по п.92; и

один или большее количество процессоров сконфигурированных для выполнения инструкций, хранящихся на машиночитаемом носителе, для осуществления способа по любому из пп.1-91.

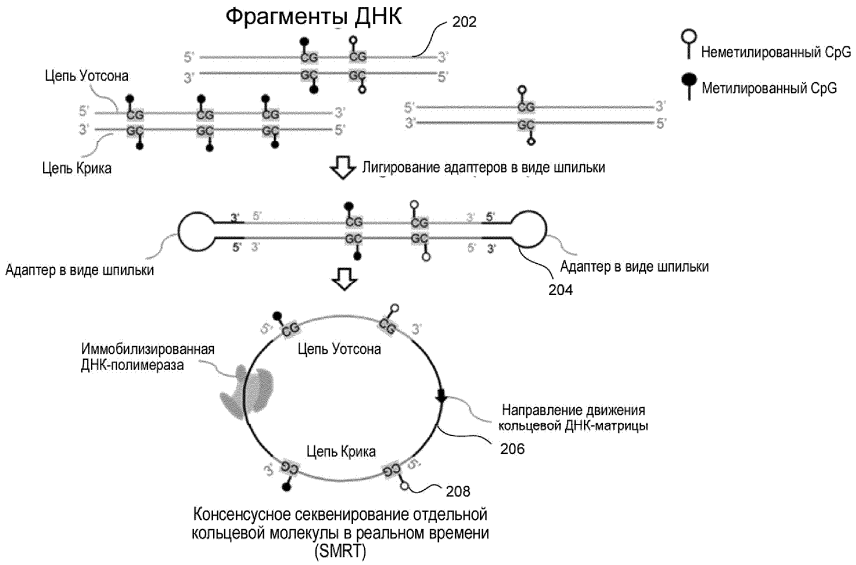
94. Система для осуществления способа по любому из пп.1-91, содержащая средства для выполнения любого способов по любому из пп.1-91.

95. Система для осуществления способа по любому из пп.1-91, содержащая один или большее количество процессоров, сконфигурированных для выполнения способов по любому из пп.1-91.

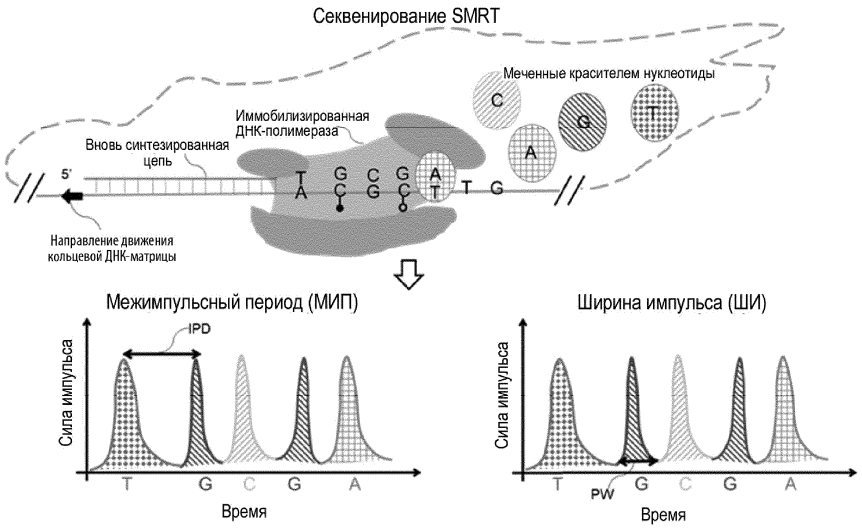
96. Система для осуществления способа по любому из пп.1-91, содержащая модули, которые соответствующим образом выполняют шаги любого из способов по любому из пп.1-91.



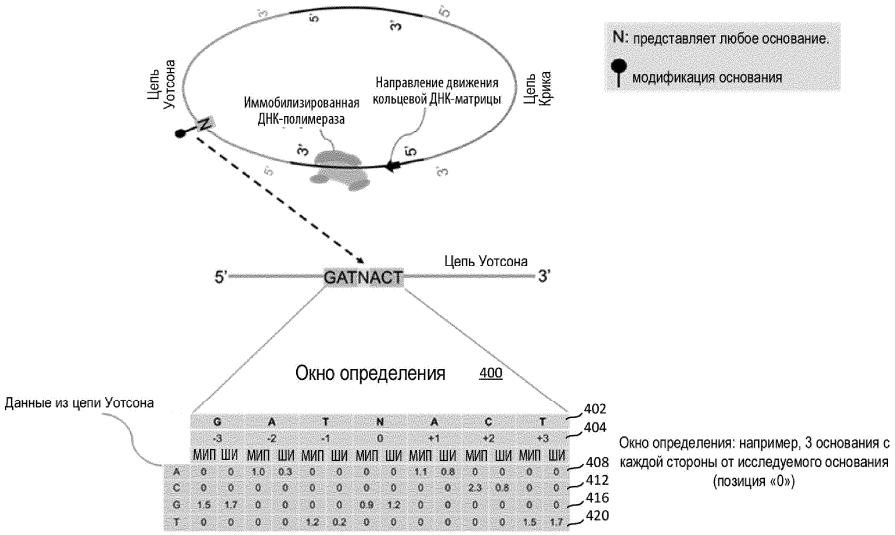
Фиг. 1



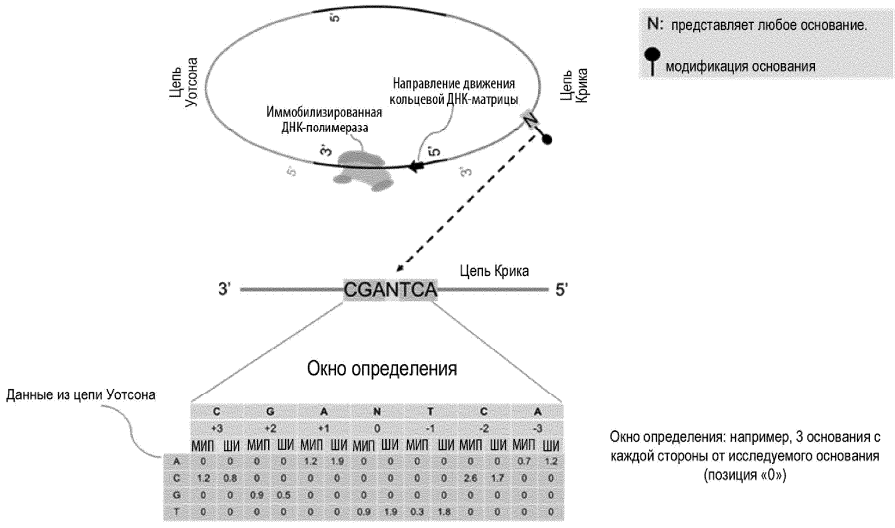
Фиг. 2



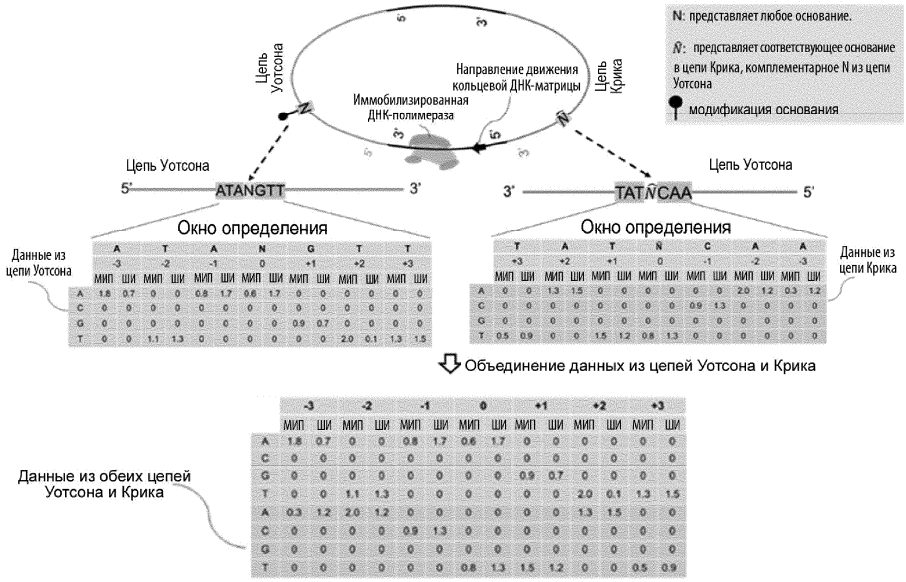
Фиг. 3



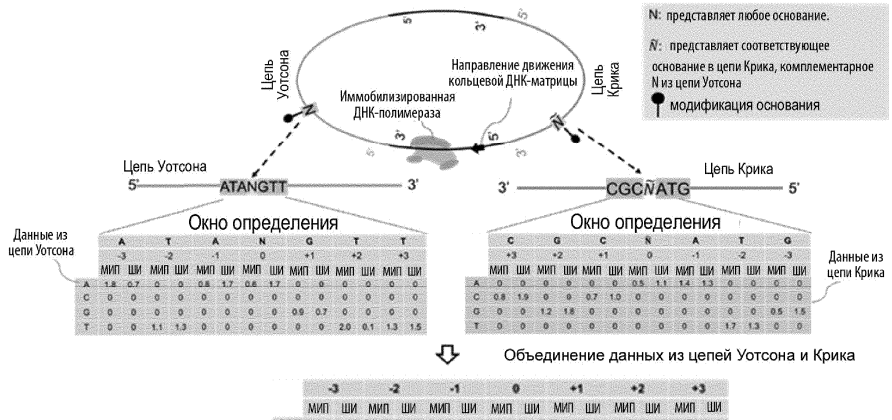
Фиг. 4



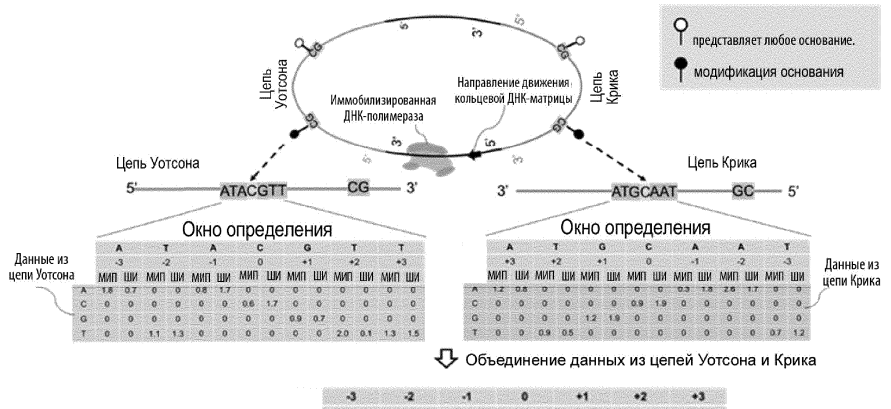
Фиг. 5



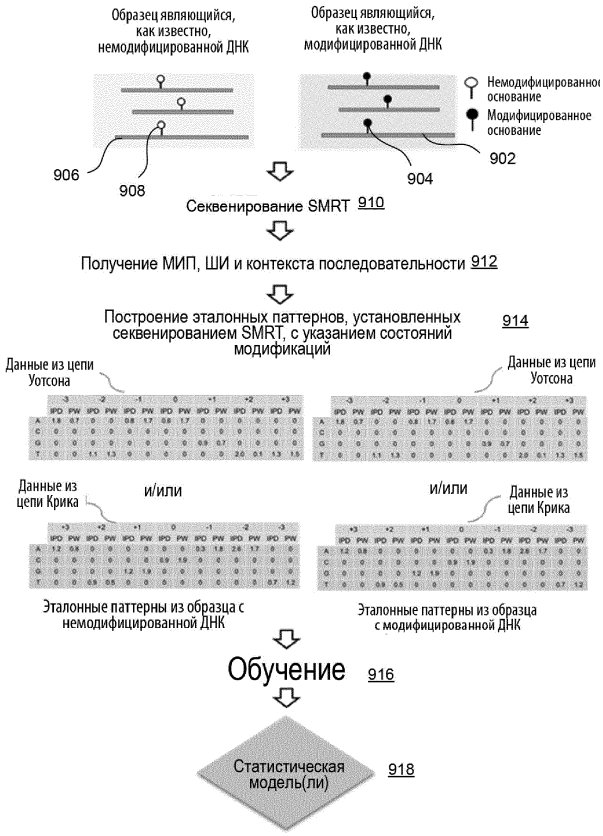
Фиг. 6



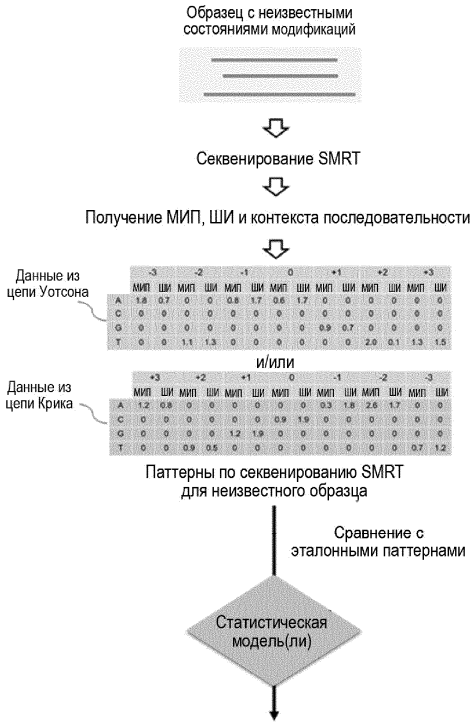
Фиг. 7



Фиг. 8

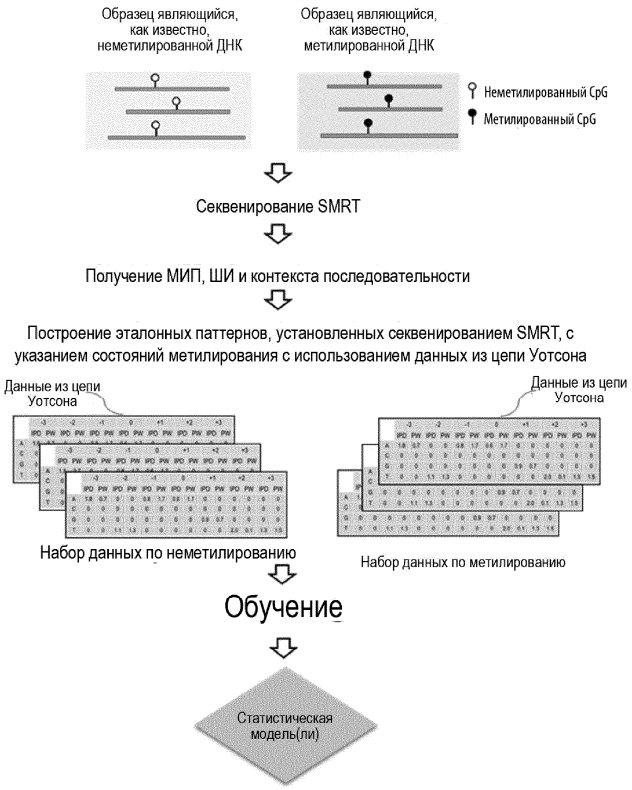


Фиг. 9

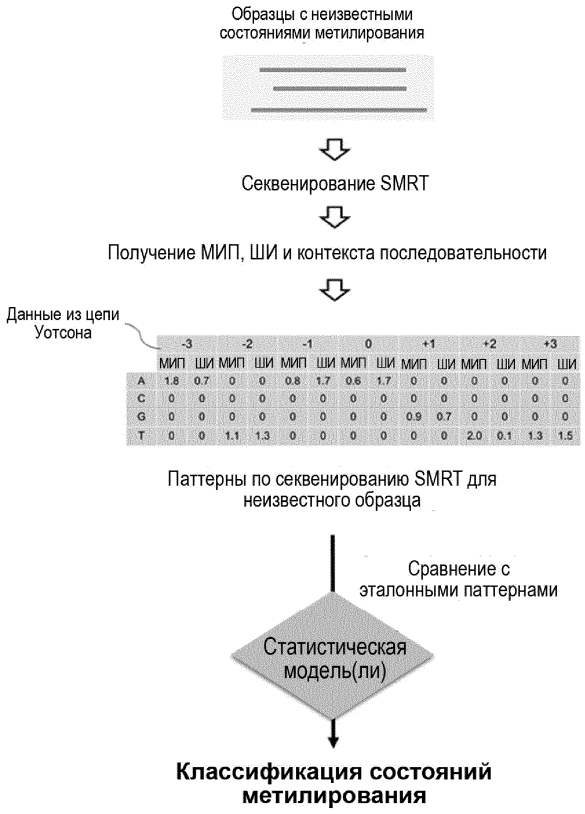


Классификация модификаций оснований

Фиг. 10

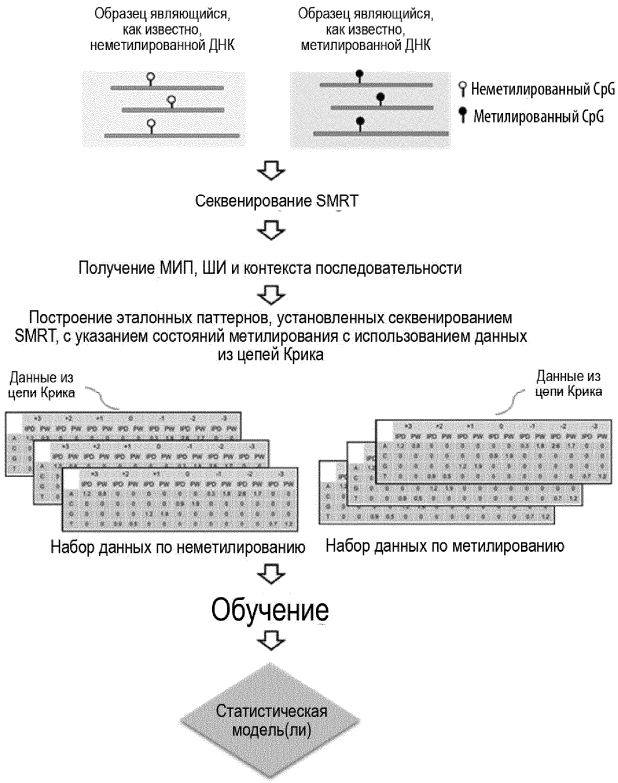


Фиг. 11

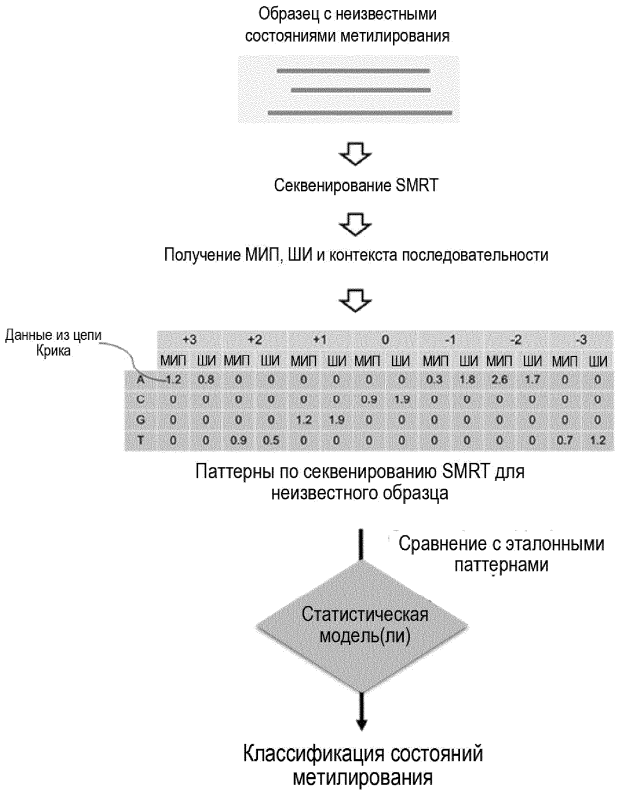


	-3		-2		-1		0		+1		+2		+3	
	МИП	ШИ	МИП	ШИ	МИП	ШИ	МИП	ШИ	МИП	ШИ	МИП	ШИ	МИП	ШИ
A	1.8	0.7	0	0	0.8	1.7	0.6	1.7	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0.9	0.7	0	0	0	0
T	0	0	1.1	1.3	0	0	0	0	0	0	2.0	0.1	1.3	1.5

Фиг. 12

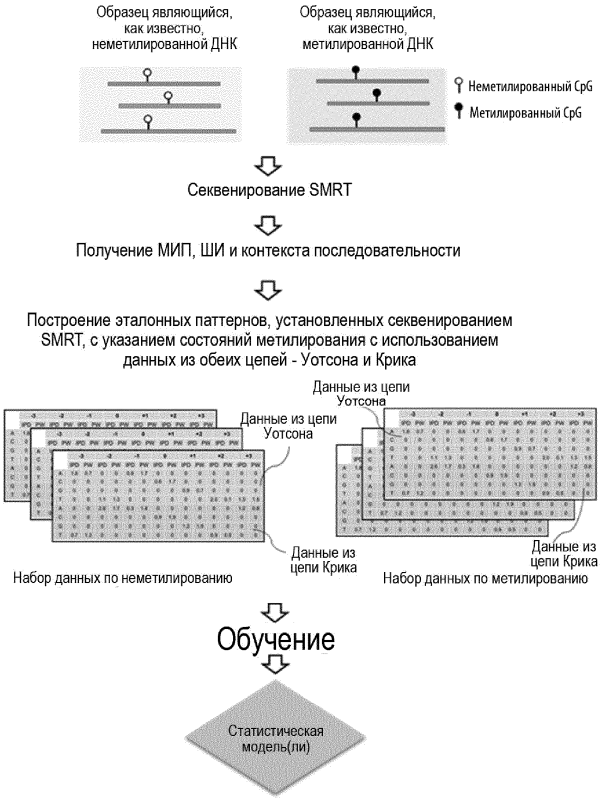


Фиг. 13

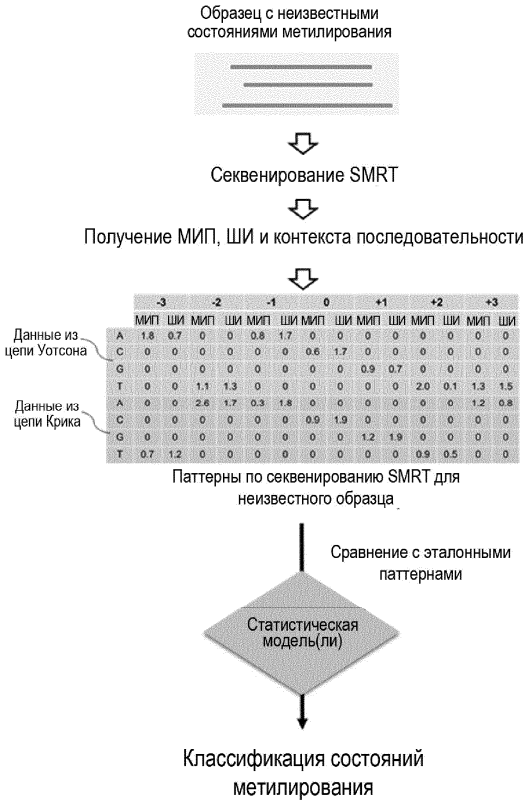


	+3		+2		+1		0		-1		-2		-3	
	МИП	ШИ	МИП	ШИ	МИП	ШИ	МИП	ШИ	МИП	ШИ	МИП	ШИ	МИП	ШИ
A	1.2	0.8	0	0	0	0	0	0	0.3	1.8	2.6	1.7	0	0
C	0	0	0	0	0	0	0.9	1.9	0	0	0	0	0	0
G	0	0	0	0	1.2	1.9	0	0	0	0	0	0	0	0
T	0	0	0.9	0.5	0	0	0	0	0	0	0	0	0.7	1.2

Фиг. 14

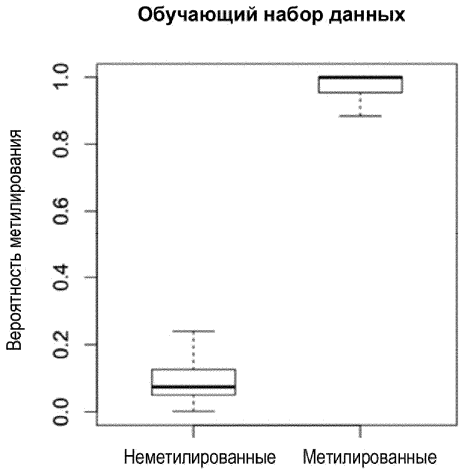


Фиг. 15

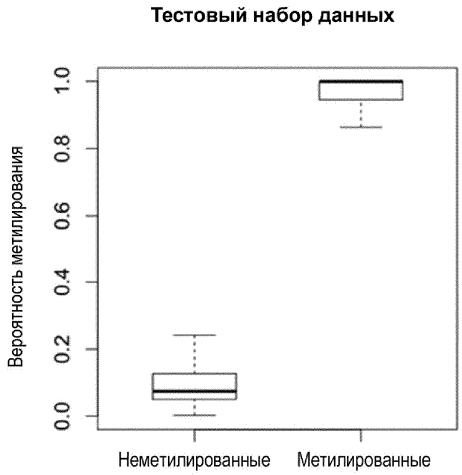


	-3		-2		-1		0		+1		+2		+3	
	МИП	ШИ	МИП	ШИ	МИП	ШИ	МИП	ШИ	МИП	ШИ	МИП	ШИ	МИП	ШИ
A	1.8	0.7	0	0	0.8	1.7	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0.6	1.7	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0.9	0.7	0	0	0	0
T	0	0	1.1	1.3	0	0	0	0	0	0	2.0	0.1	1.3	1.5
A	0	0	2.6	1.7	0.3	1.8	0	0	0	0	0	0	1.2	0.8
C	0	0	0	0	0	0	0.9	1.9	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	1.2	1.9	0	0	0	0
T	0.7	1.2	0	0	0	0	0	0	0	0	0.9	0.5	0	0

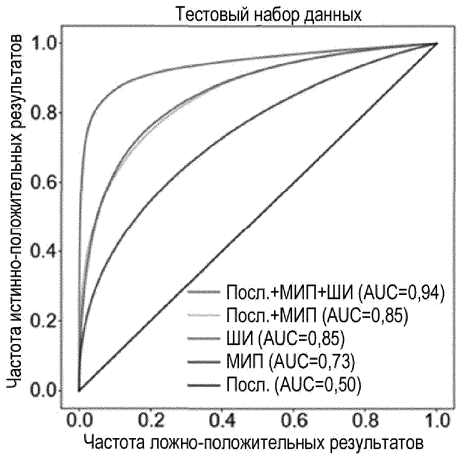
Фиг. 16



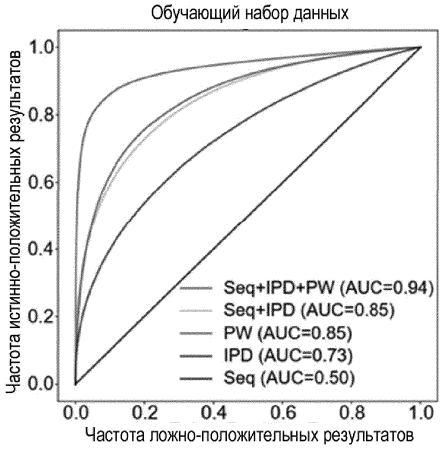
Фиг. 17A



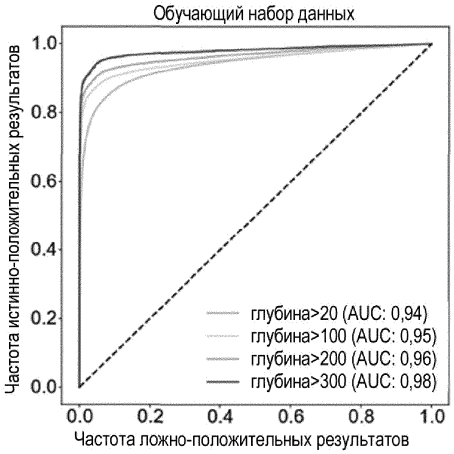
Фиг. 17B



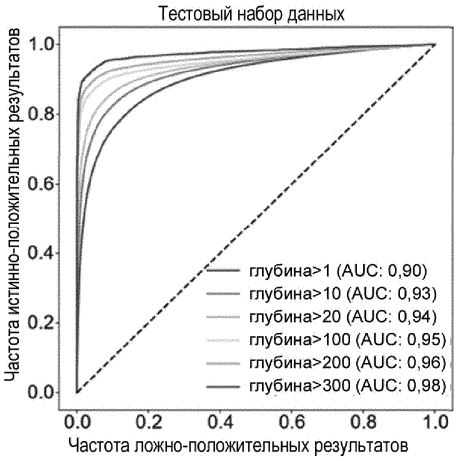
Фиг. 18A



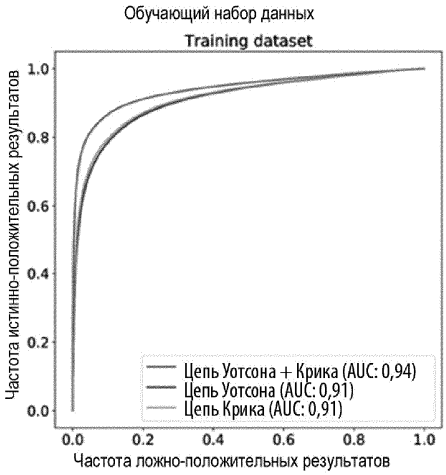
Фиг. 18B



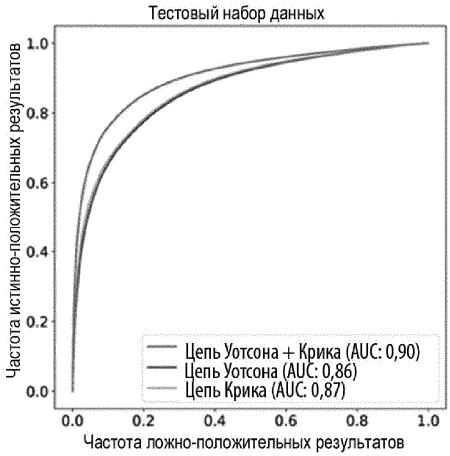
Фиг. 19A



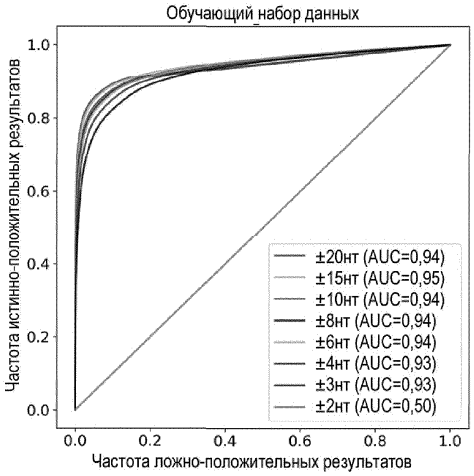
Фиг. 19B



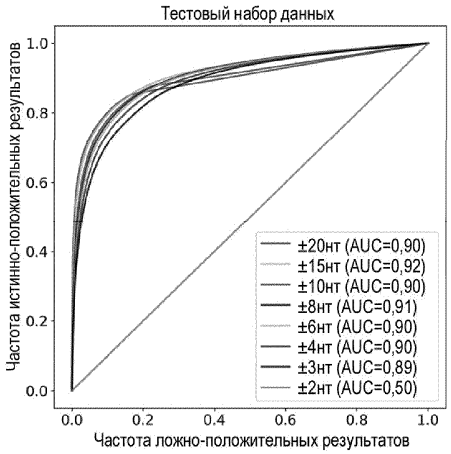
Фиг. 20A



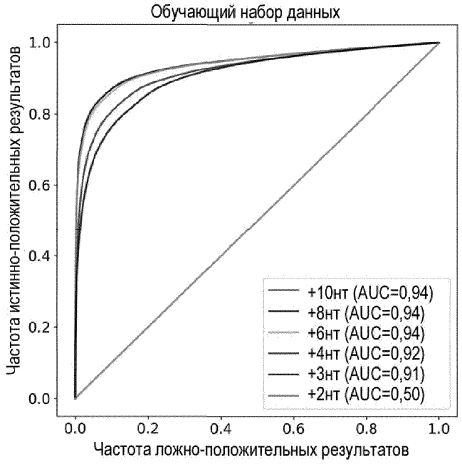
Фиг. 20B



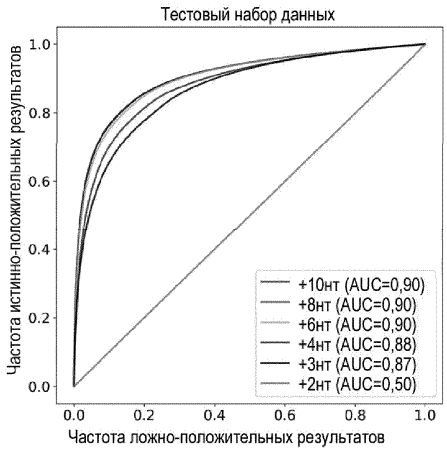
Фиг. 21A



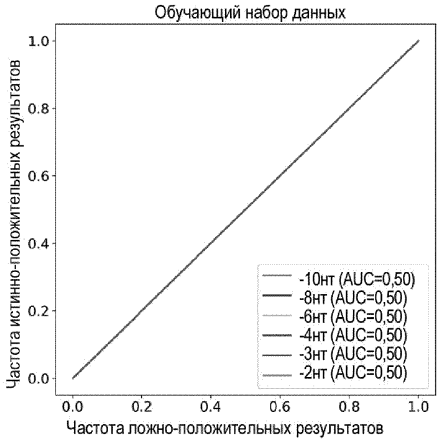
Фиг. 21B



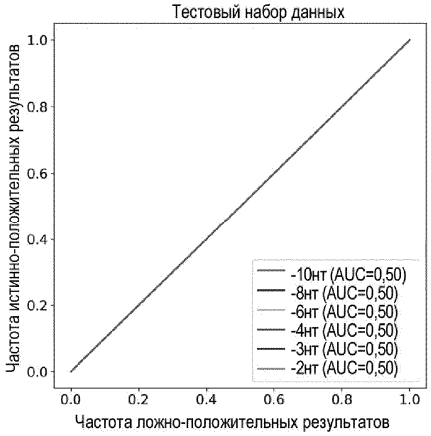
Фиг. 22A



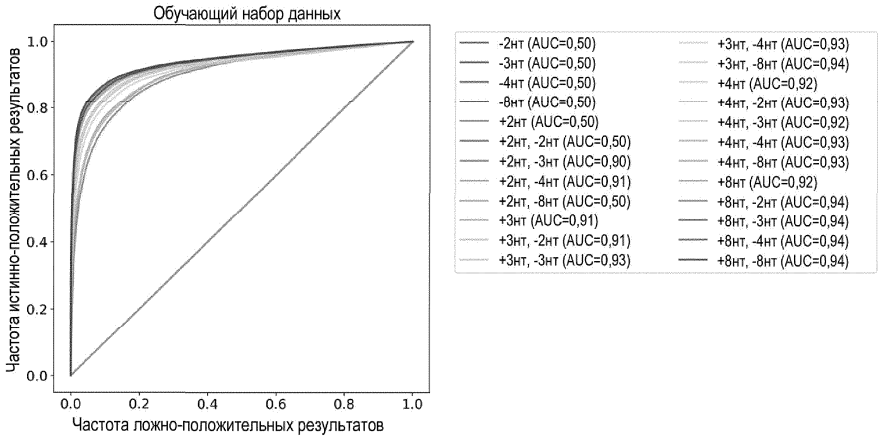
Фиг. 22B



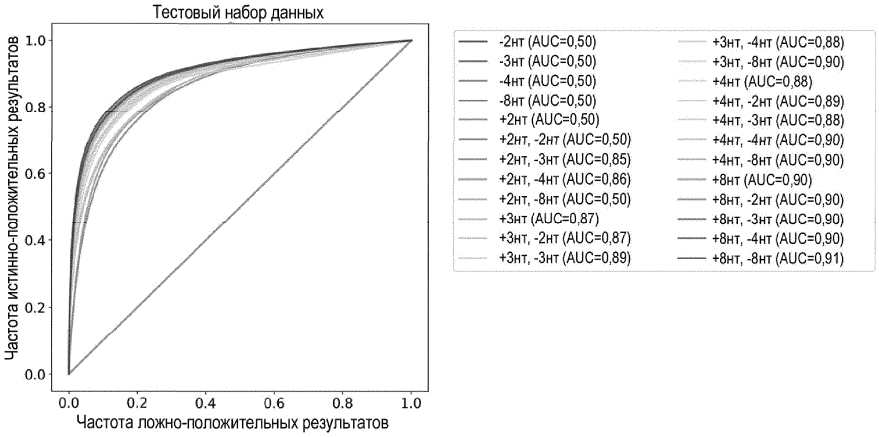
Фиг. 23A



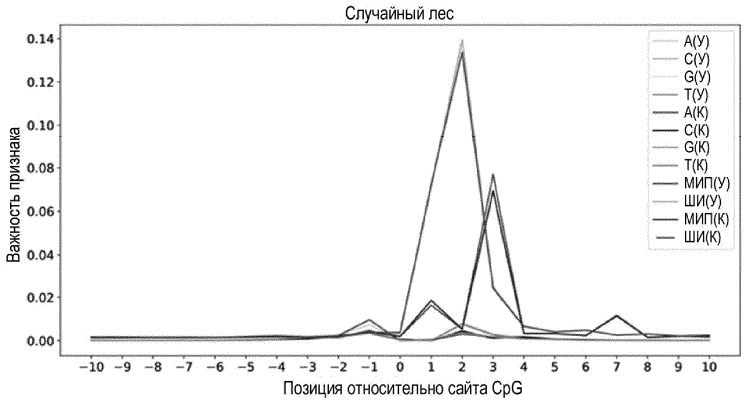
Фиг. 23B



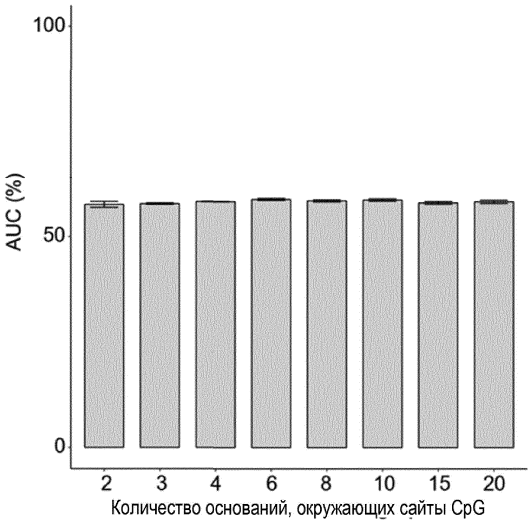
Фиг. 24



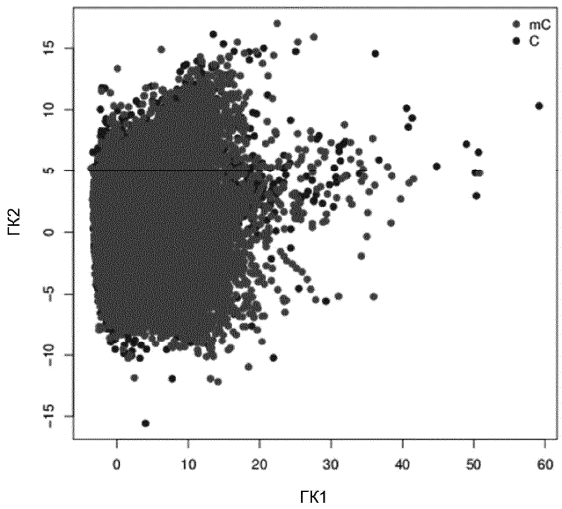
Фиг. 25



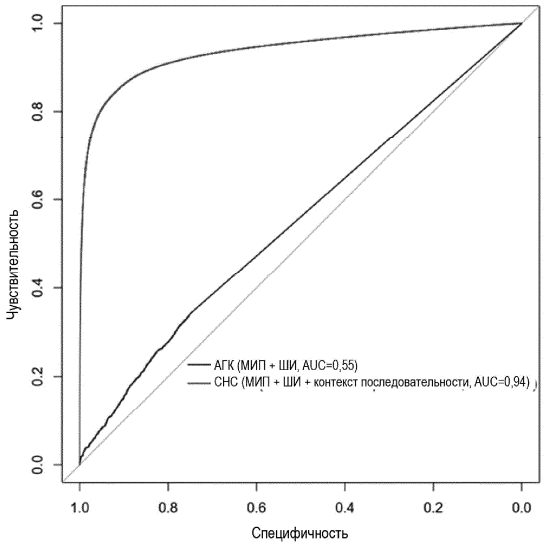
Фиг. 26



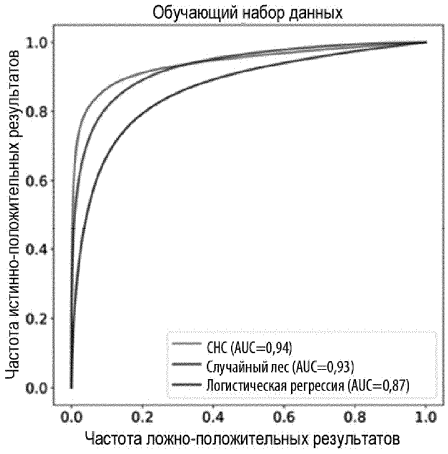
Фиг. 27



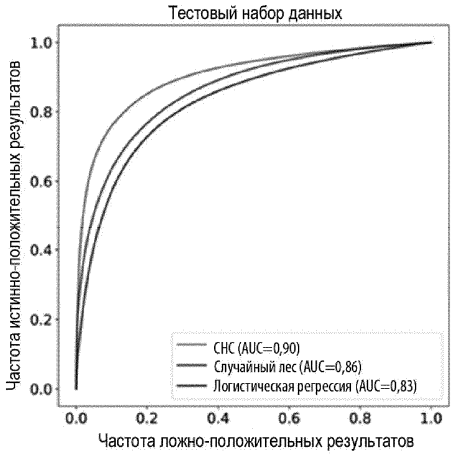
Фиг. 28



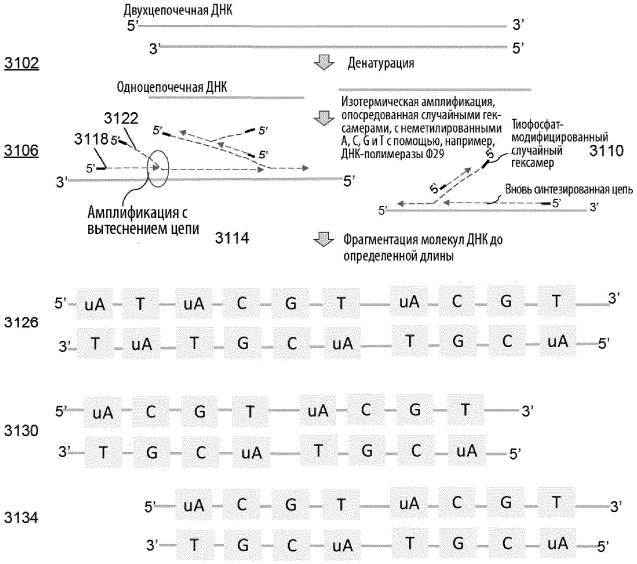
Фиг. 29



Фиг. 30А

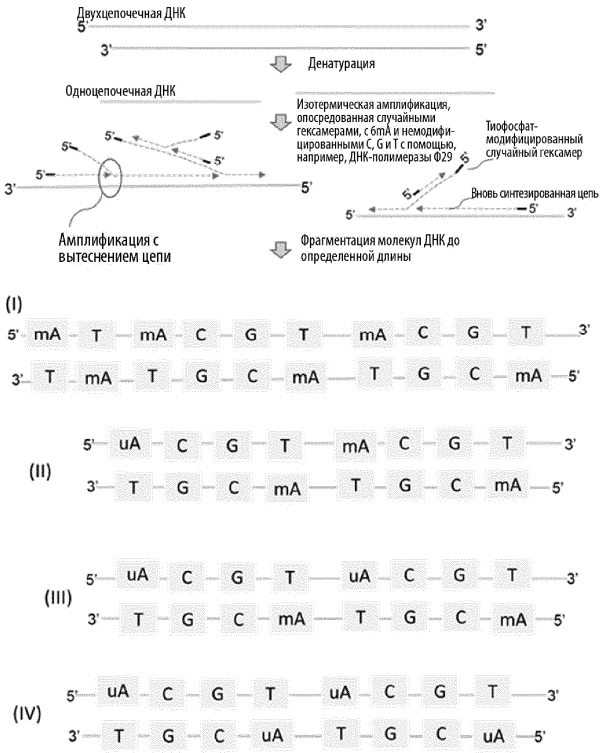


Фиг. 30B



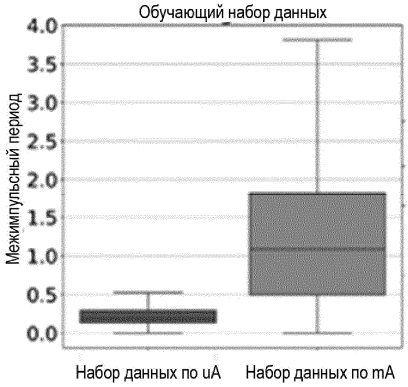
Продукты ДНК полногеномной амплификации

Фиг. 31A

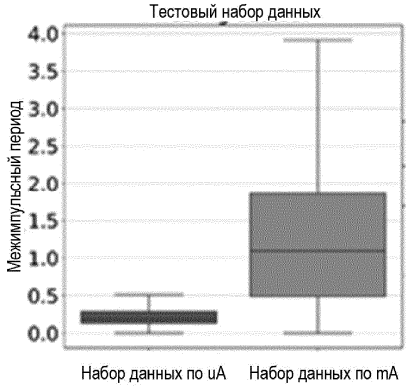


Продукты ДНК полногеномной амплификации

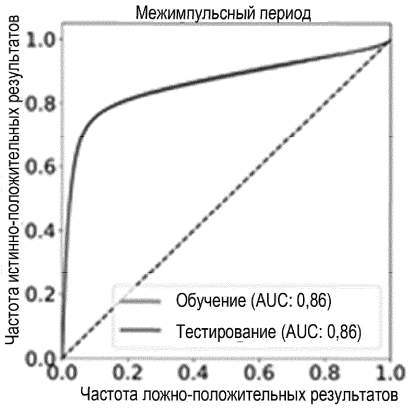
Фиг. 31B



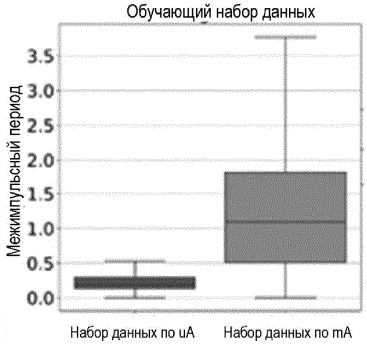
Фиг. 32A



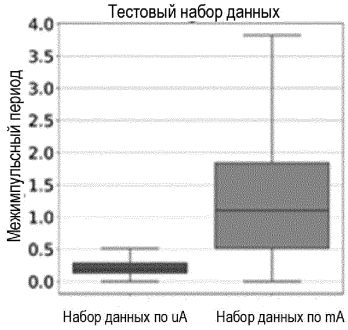
Фиг. 32B



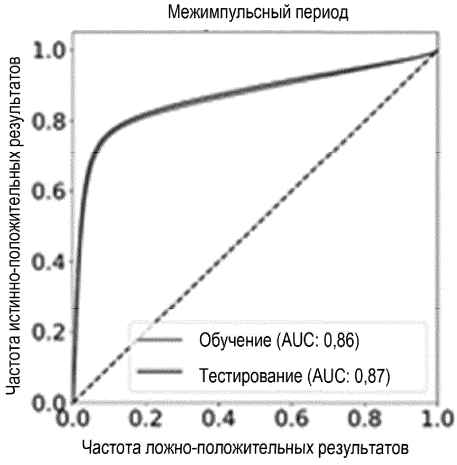
Фиг. 32C



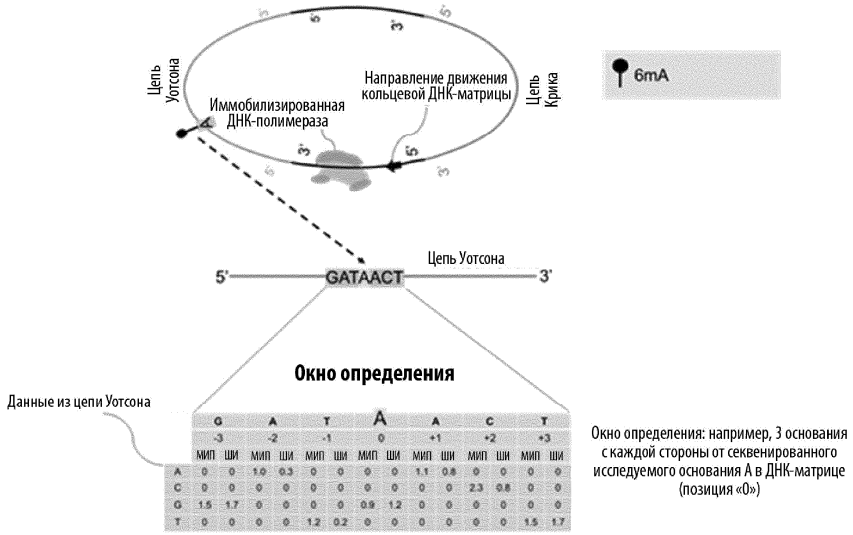
Фиг. 33A



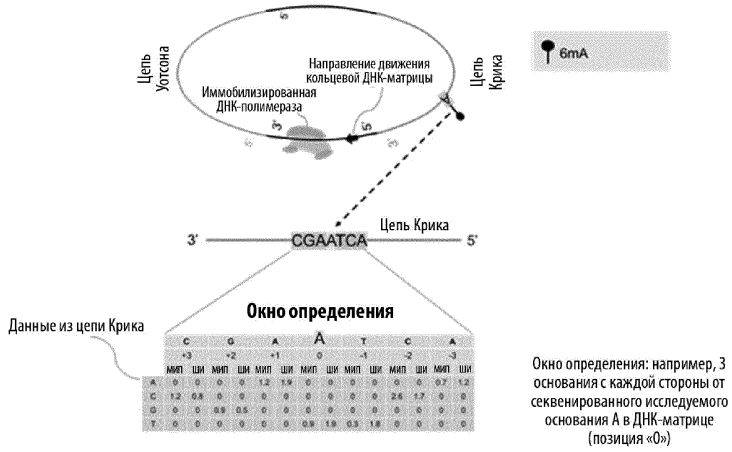
Фиг. 33B



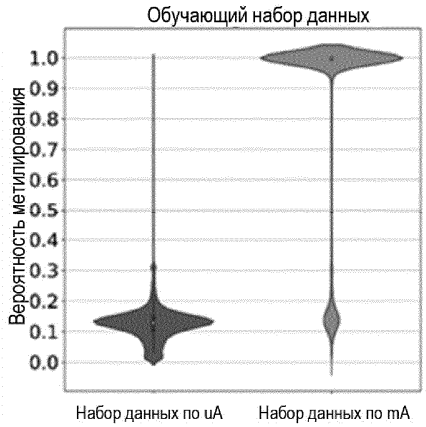
Фиг. 33C



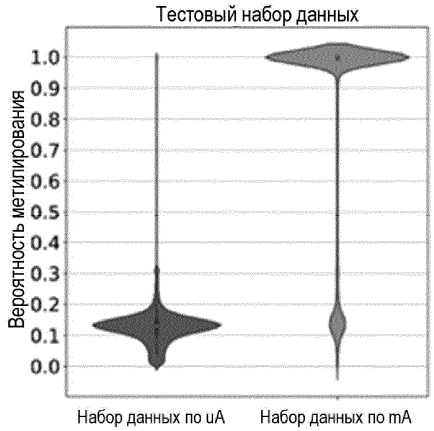
Фиг. 34



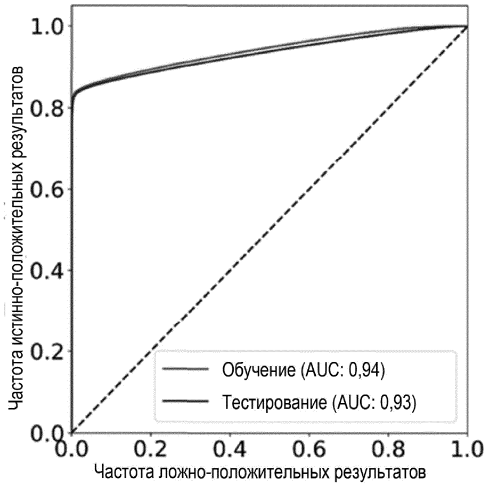
Фиг. 35



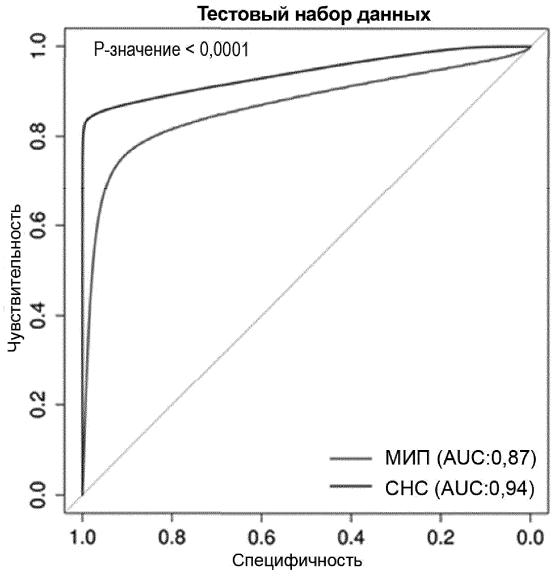
Фиг. 36A



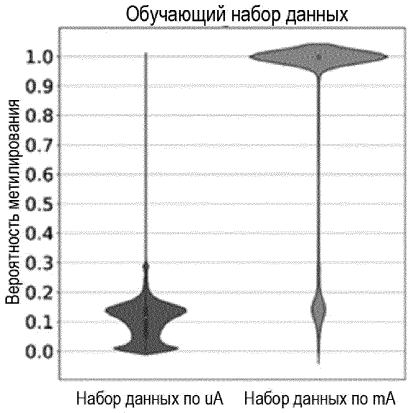
Фиг. 36B



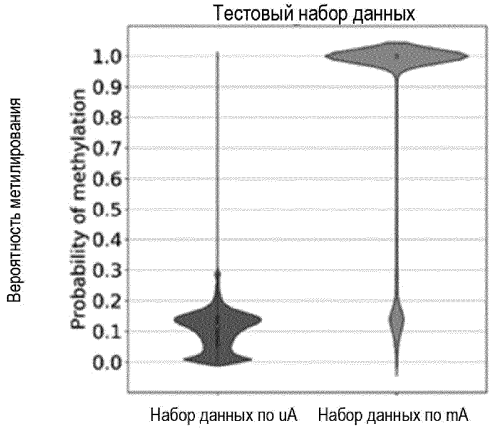
Фиг. 37



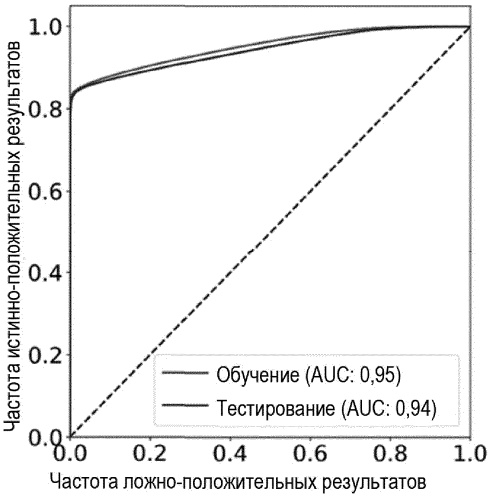
Фиг. 38



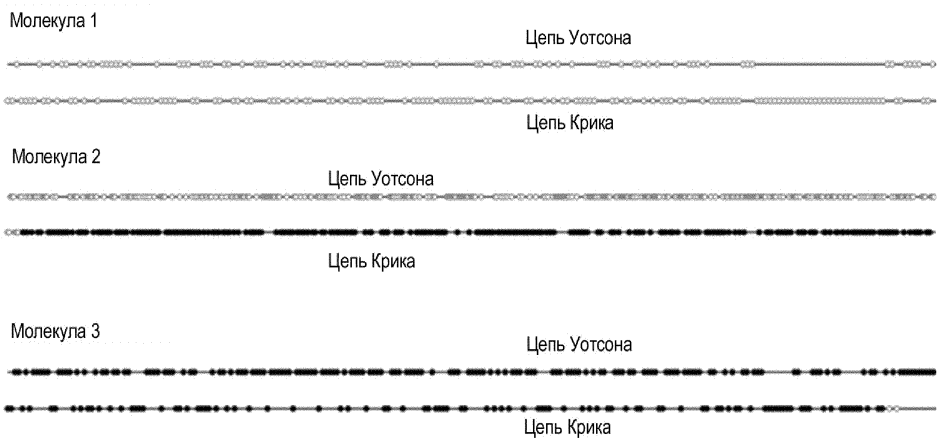
Фиг. 39A



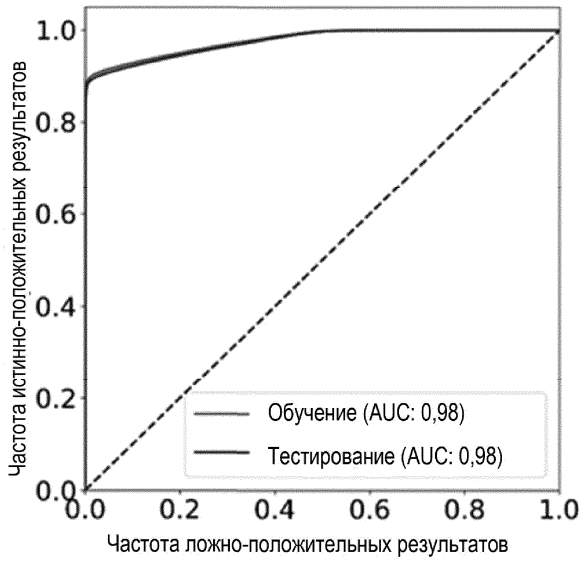
Фиг. 39B



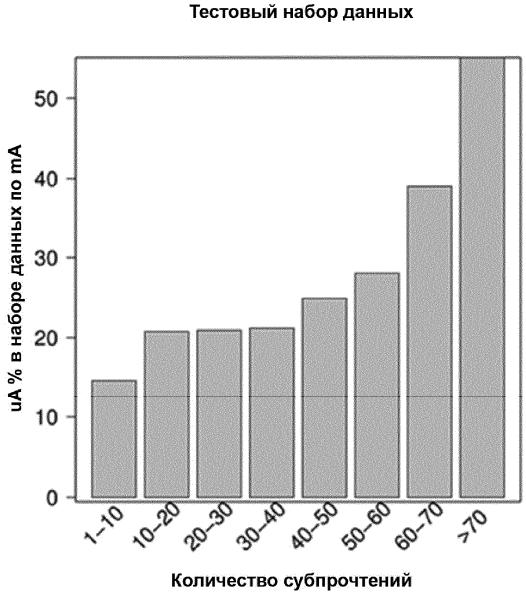
Фиг. 40



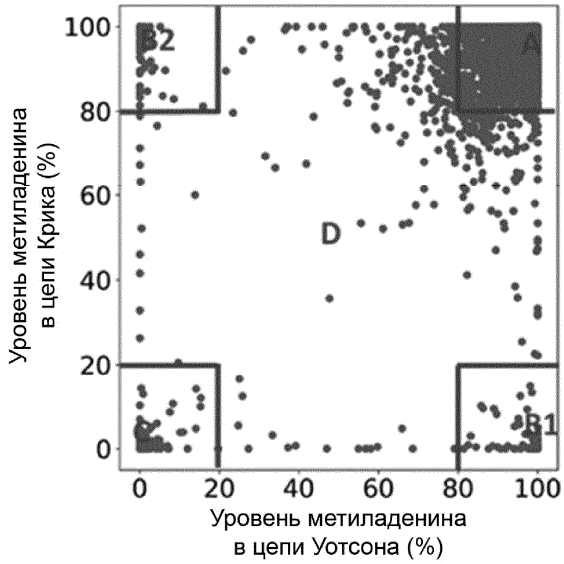
Фиг. 41



Фиг. 42



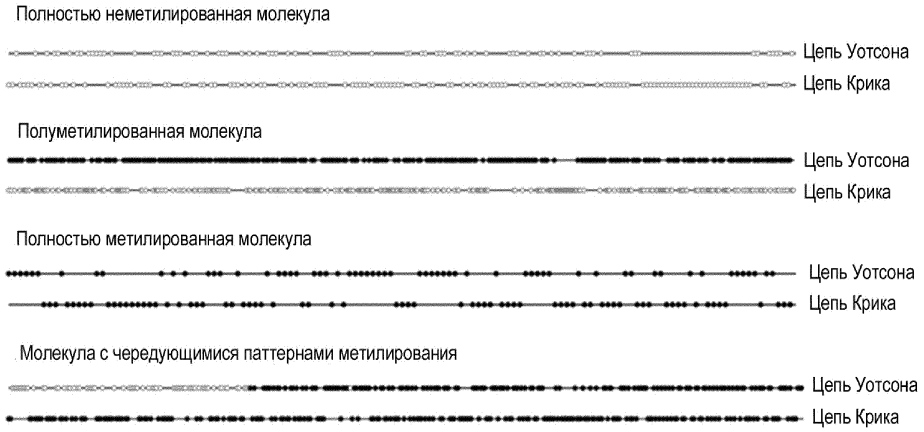
Фиг. 43



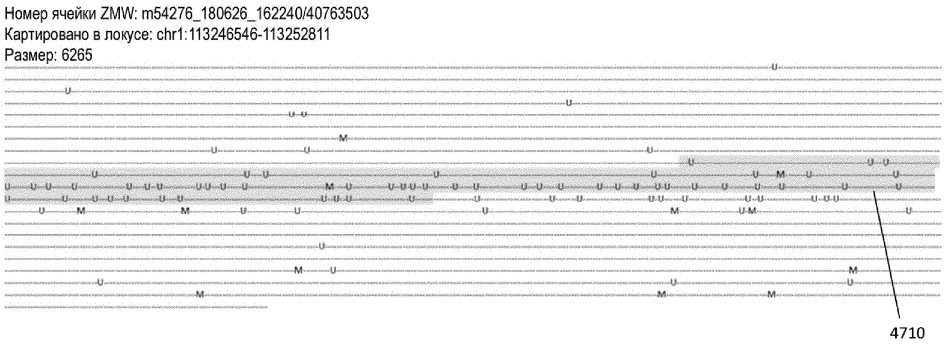
Фиг. 44

Категории	Обучающий набор данных	Тестовый набор данных
Полностью неметилированные	283 (7.0%)	276 (7.0%)
Полуметилированные	401 (10.0%)	389 (9.8%)
Полностью метилированные	3194 (79.4%)	3142 (79.4%)
Чередующееся паттерны метилирования	145 (3.6%)	148 (3.7%)

Фиг. 45



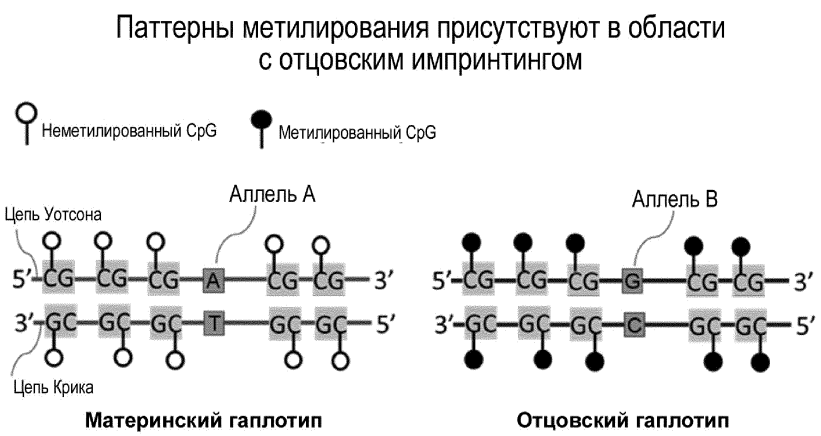
Фиг. 46



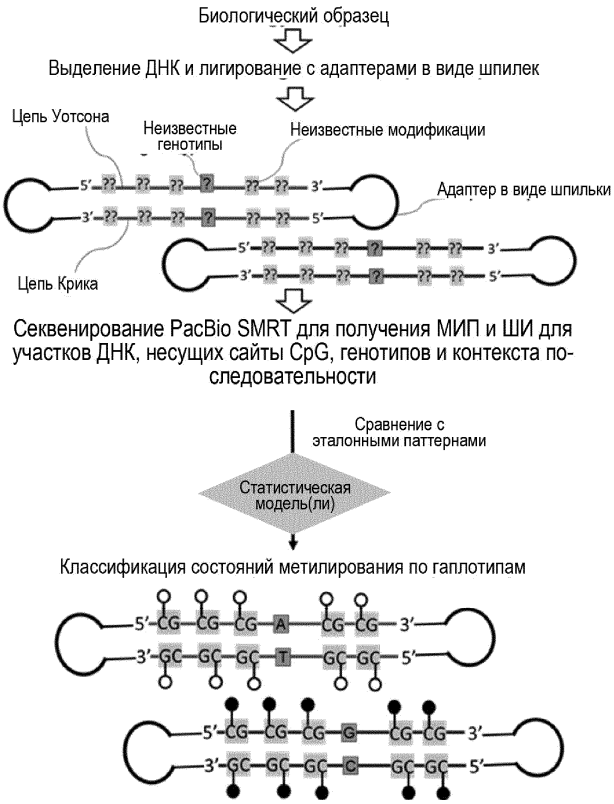
Фиг. 47

Хромосомы	Начало	Конец	Названия генов, подверженных импринтингу	Длина CpG-островков	Молекулы, секвенированные с помощью секвенирования PacBio SMRT, и состояния метилирования, определенные согласно вариантам осуществления, представленным в данном изобретении	Сигнал метилирования молекулы
chr11	2013333	2013617	H19	284	-U----------M----------M-M---------U----U----------M----------M----- --U----------M----------M----- M----------M-M------U---------[T]----------M----------M----- --M---	Метилированные
chr11	2019565	2019863	H19	298	-M-M----------M----------M----------[C]-----M-M------M---------- ----------M----------M---------M-M----------M---------M---------- M-----M-M---M----------M------M----------M-M----------M----- ----------	Метилированные
chr11	32460586	32461004	WT1-AS/WT1	418	-U--U----------U-M-------[C]----------U----------U---U-------- U----------U---U-U----------U----M---M---------U---------- -----U-U-U------U------U----------U-----U----------U-U----------U-- ----M	Неметилированные
chr14	101192851	101193499	DLK1	648	-U--------U----------U----------U----------M---M----------U----------U---------- -U--------U----------U-U--U---U--U----------U----------U- --U-U----------U----------U-----U-U---U----------U--U---U------U----- M-----	Неметилированные
chr14	101201559	101201763	DLK1	204	-M--M----------U----------M--------M---M----------[T]---------- ----------M----------M---M----------M----M----------M---------- ----------M----------M----------M-	Метилированные
chr14	101292863	101293101	MEG3	238	----------M----------U----------U----------M- -----M----------M-M----------M-	Метилированные
chr15	25981176	25981392	ATP10A	216	-----*----------M----------M----------M----------M----------M---------- M---M----------M----------[T]-----M-M----------U-	Метилированные
chr2	80531367	80531719	LRRTM1	352	---*----[G]----------U---U-------M-------M----------M----------U----U-- ----U------U----------U----------U--------M----------U---- --U-U----------M----------U----------U--U----------U-U--- --	Неметилированные
chr7	79082174	79082427	MAGI2	253	----------*-U----------[A]-M-U----------U----U---U---------U--- U----------U----------*----------	Неметилированные

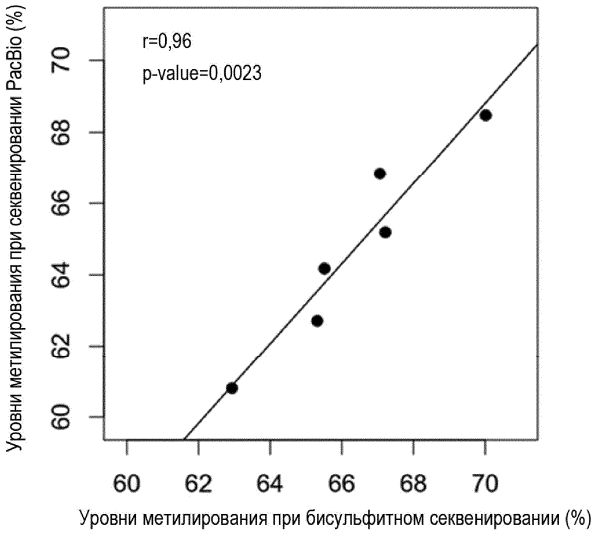
Фиг. 48



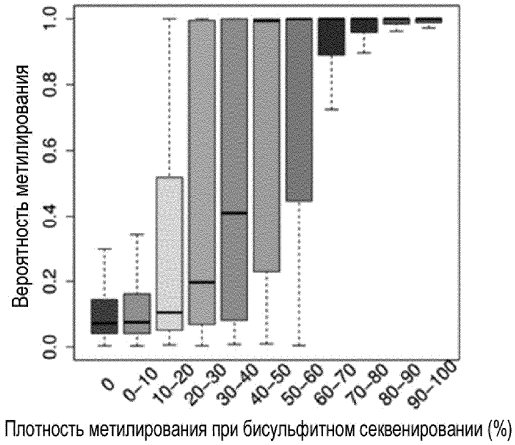
Фиг. 49


Биологический образец
Выделение ДНК и лигирование с адаптерами в виде шпилек
Цепь Уотсона
Неизвестные генотипы
Неизвестные модификации
Адаптер в виде шпильки
Цепь Крика
Секвенирование PacBio SMRT для получения МИП и ШИ для участков ДНК, несущих сайты CpG, генотипов и контекста последовательности
Сравнение с эталонными паттернами
Статистическая модель(ли)
Классификация состояний метилирования по гаплотипам

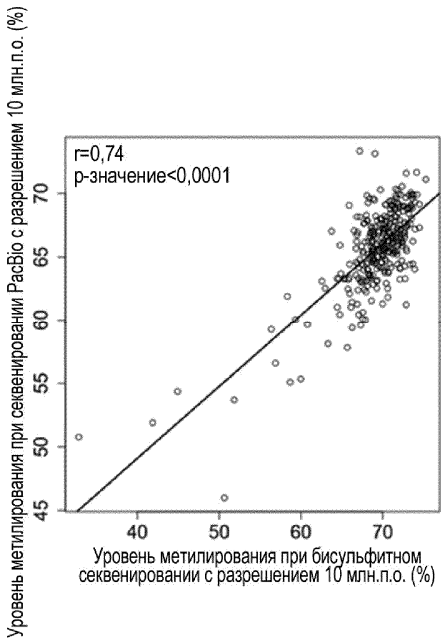
Фиг. 50


r=0,96
p-value=0,0023
Уровни метилирования при секвенировании PacBio (%)
Уровни метилирования при бисульфитном секвенировании (%)
60
62
64
66
68
70

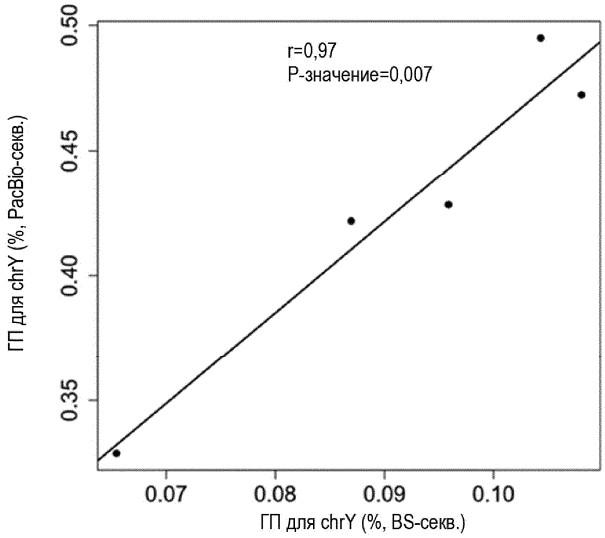
Фиг. 51



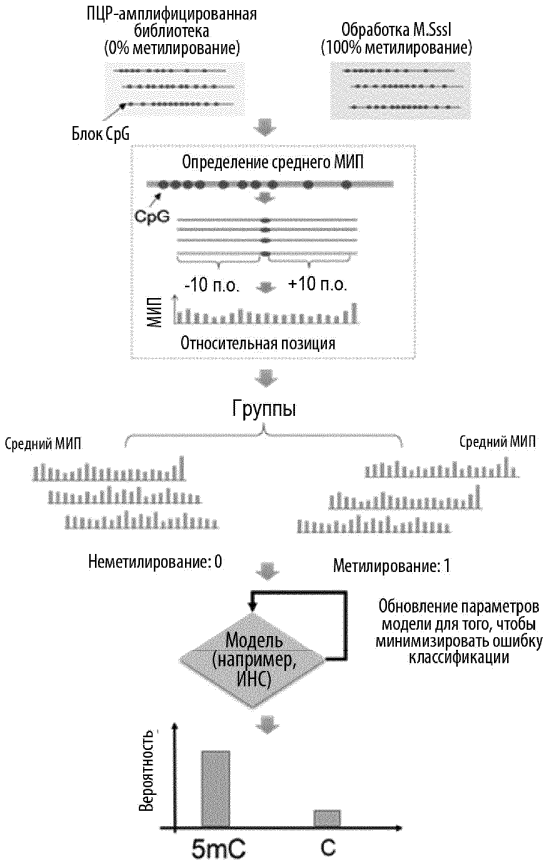
Фиг. 52A



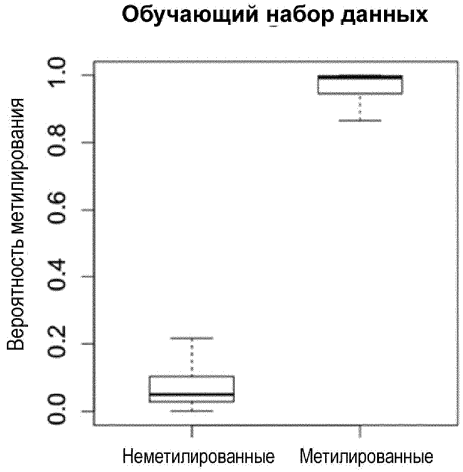
Фиг. 52B



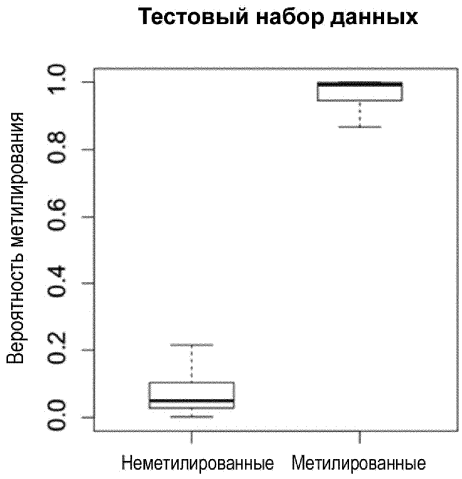
Фиг. 53



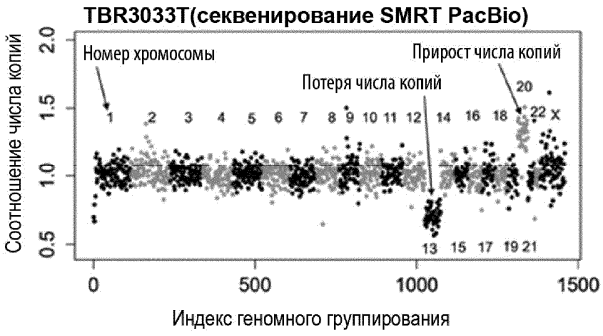
Фиг. 54



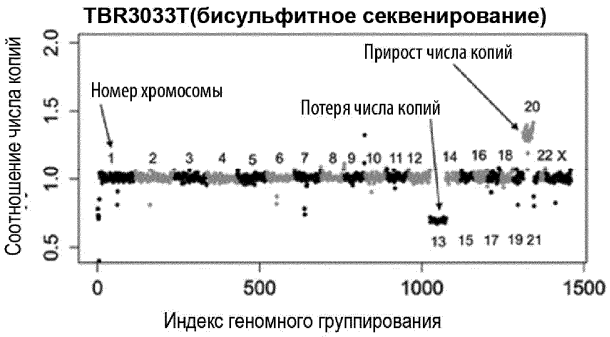
Фиг. 55A



Фиг. 55B



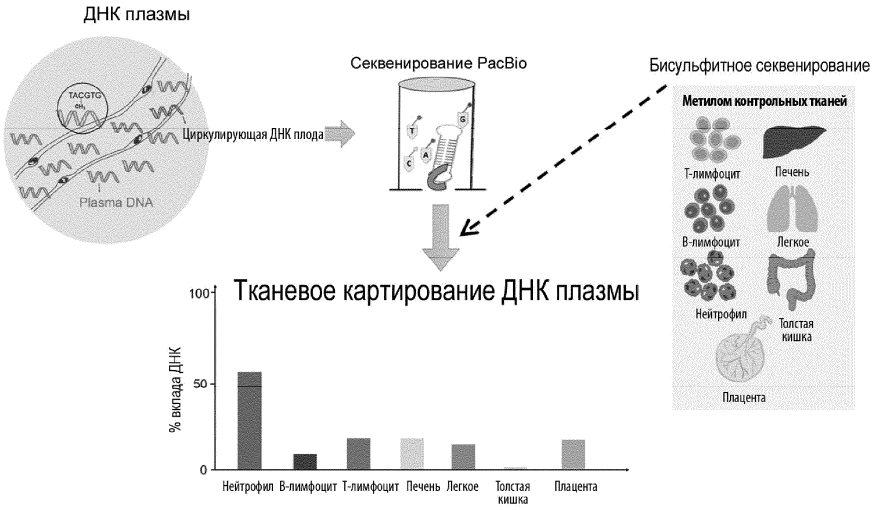
Фиг. 56A



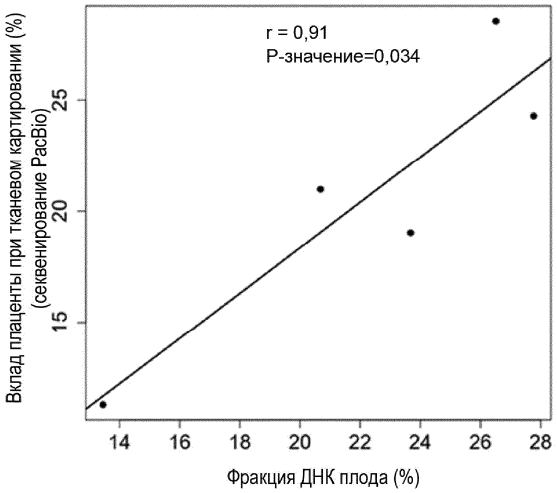
Фиг. 56B

Фиг. 57A

Фиг. 57B



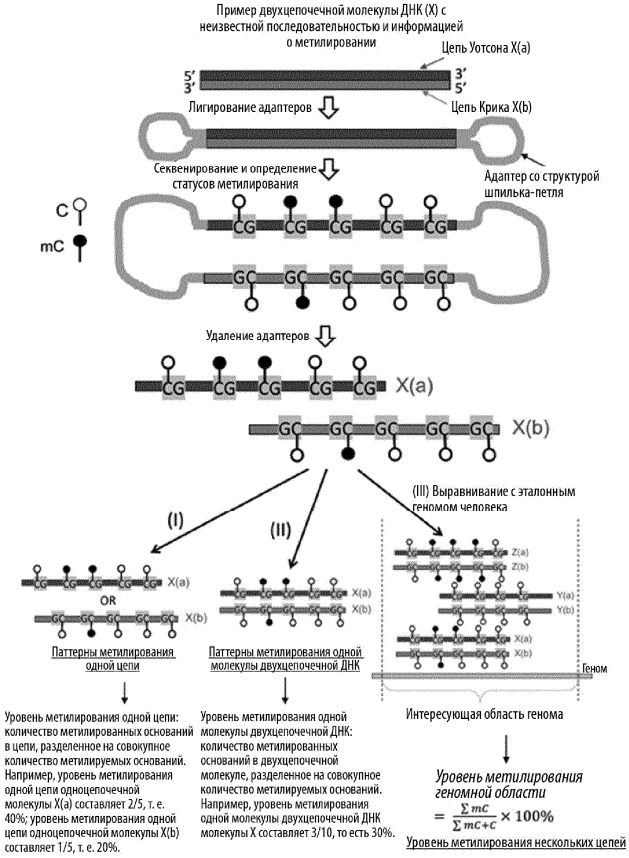
Фиг. 58



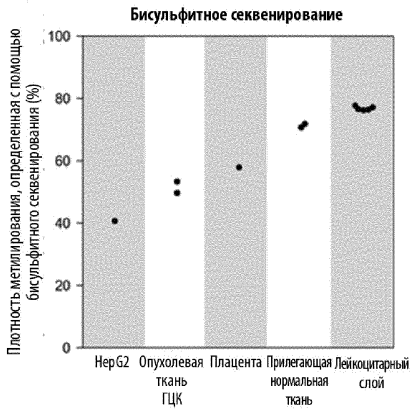
Фиг. 59

Группы	Образцы	Совокупно субпрочтений	Картированные субпрочтения	Картированность субпрочтений (%)	Средняя глубина субпрочтения на ячейку SMRT (x)	Количество ячеек SMRT	Картируемые ячейки	Доля картируемых ячеек (%)
Материнский образец лейкоцитарного слоя	M13153w	39,006,460	30,673,525	78.6	13.4	3,157,310	2,295,002	72.7
Плацента	N13153	23,013,428	16,374,758	71.2	10.4	2,393,400	1,573,540	65.7
Ткани ГЦК	TBR3032T	20,164,513	15,232,744	75.5	13.1	1,742,990	1,147,985	64.8
	TBR3033T	22,639,692	17,479,024	77.2	8.1	2,832,627	2,157,196	76.2
Прилегающие здоровые ткани	TBR3033N	73,118,110	56,446,202	77.2	12.6	6,881,142	4,471,370	65.0
	TBR3032N	76,852,680	60,145,452	78.3	12.8	6,000,227	4,702,130	78.4
Образец лейкоцитарного слоя (здоровые контрольные субъекты)	M1	44,777,423	28,325,587	63.3	7.7	7,316,000	3,659,996	50.0
	F2	49,840,758	32,994,645	66.2	8.6	7,215,112	3,823,329	53.0
	F1	40,012,804	24,717,289	61.8	6.5	7,301,768	3,800,392	52.0
	M2	152,530,411	88,596,520	58.1	7.7	21,794,606	11,563,500	53.1
Клеточная линия ГЦК	HepG2	47,308,982	34,581,721	73.1	7.3	6,220,000	4,750,581	76.4

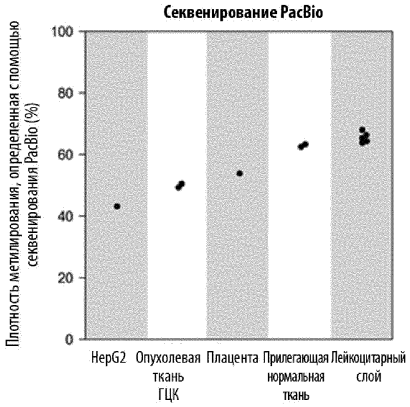
Фиг. 60



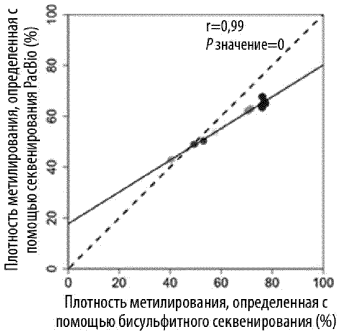
Фиг. 61



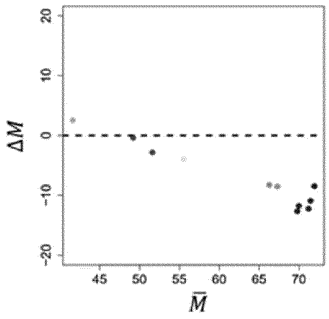
Фиг. 62A



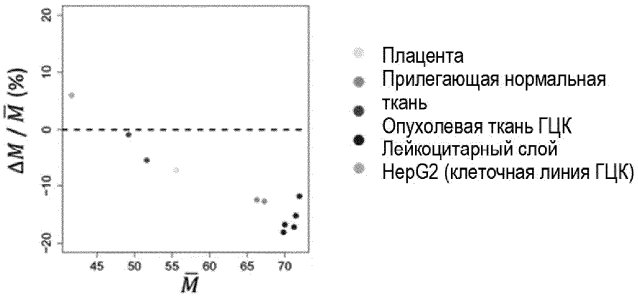
Фиг. 62B



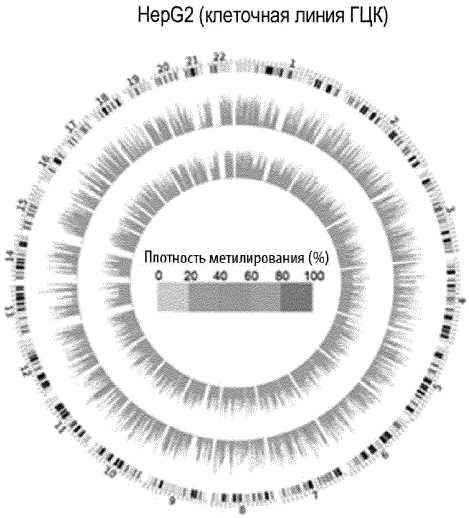
Фиг. 63A



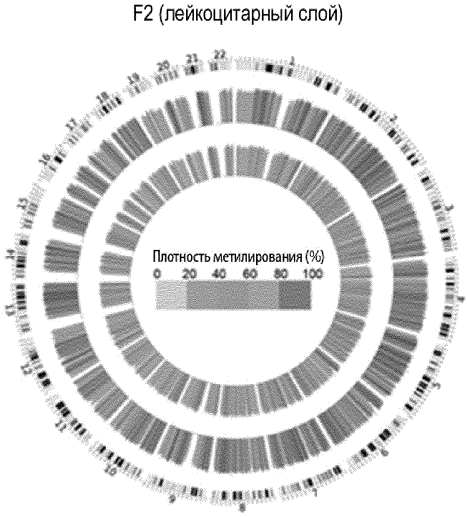
Фиг. 63B



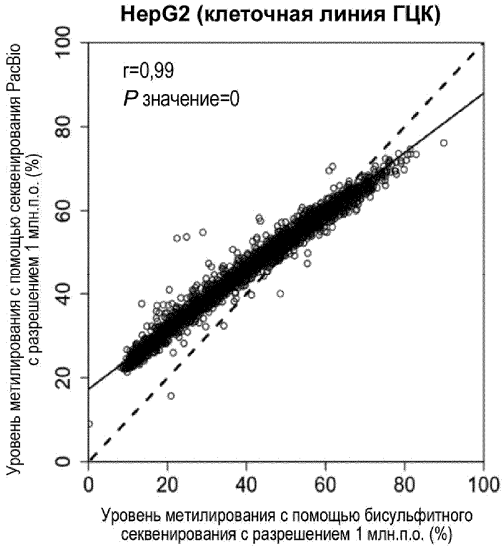
Фиг. 63C



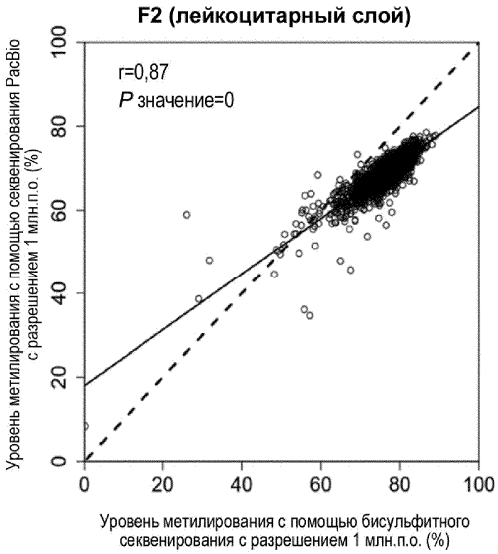
Фиг. 64A



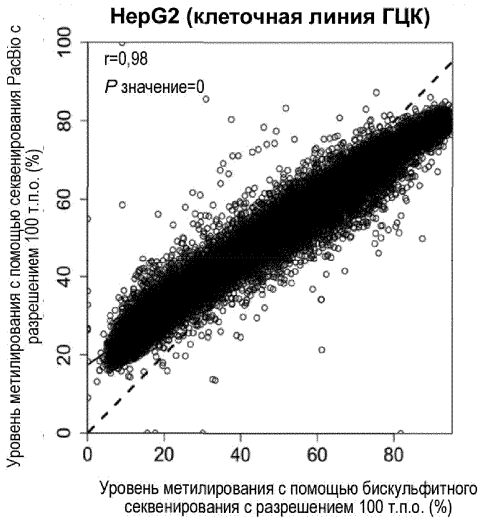
Фиг. 64B



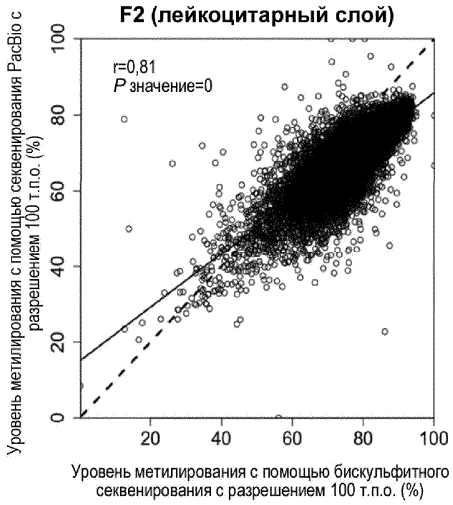
Фиг. 65A



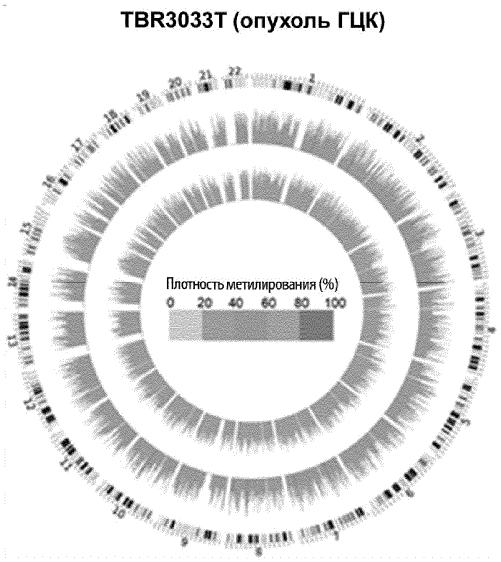
Фиг. 65B



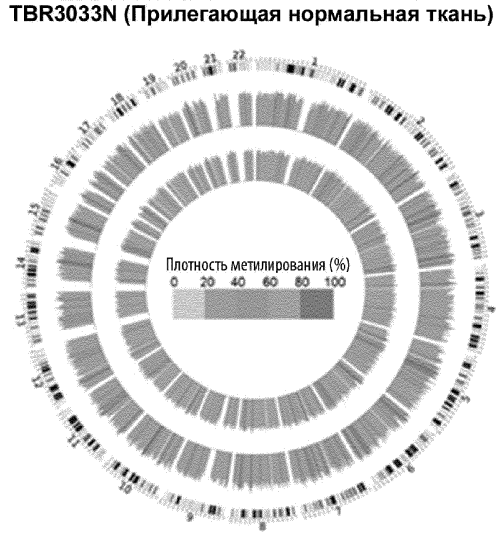
Фиг. 66A



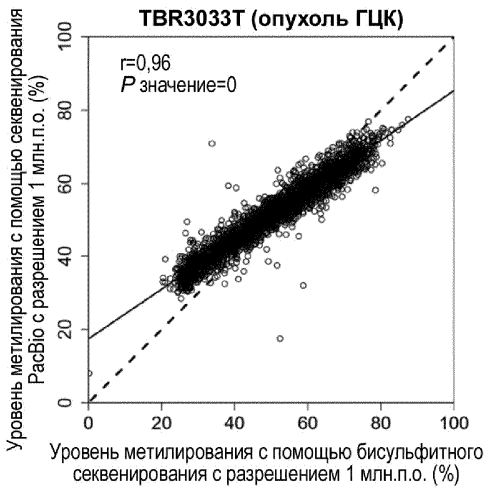
Фиг. 66B



Фиг. 67A

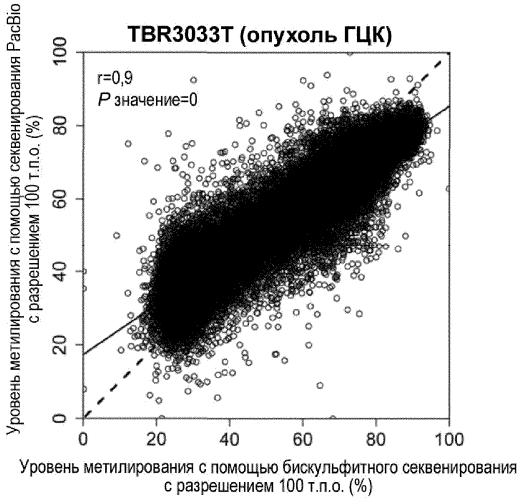


Фиг. 67B

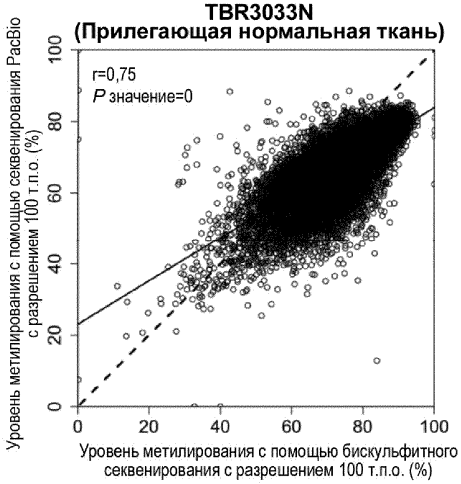


Фиг. 68A

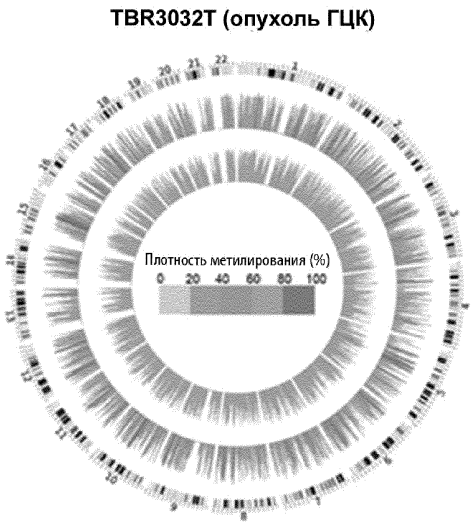
Фиг. 68B



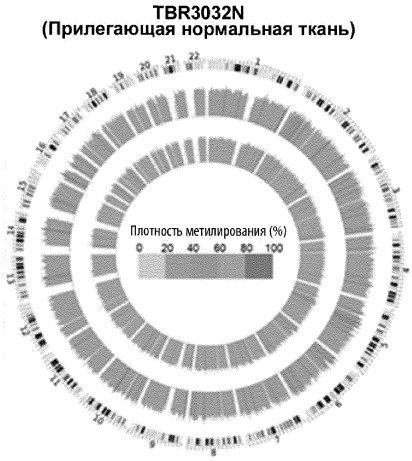
Фиг. 69A



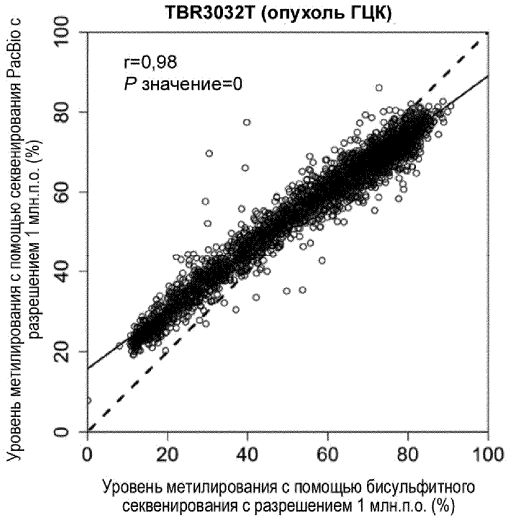
Фиг. 69B



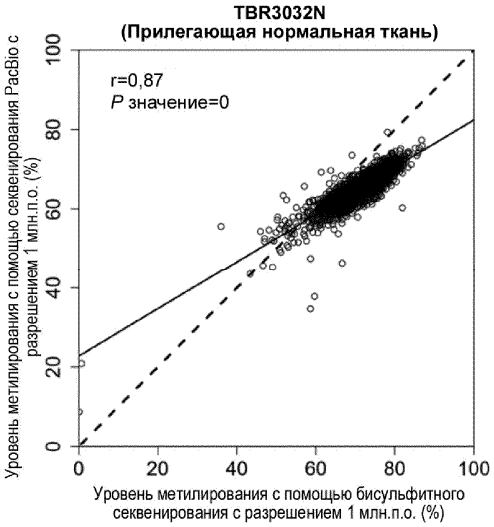
Фиг. 70A



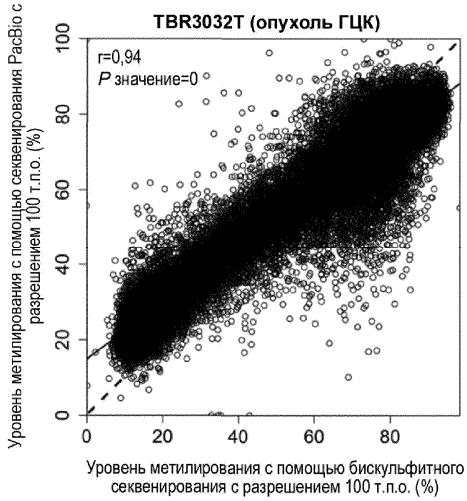
Фиг. 70B



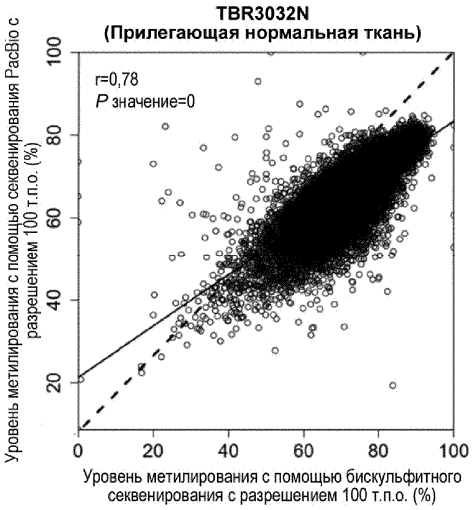
Фиг. 71A



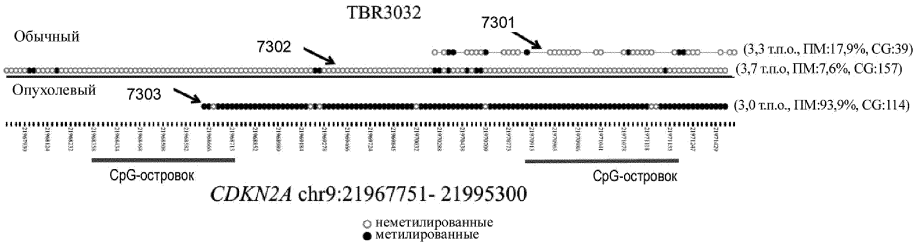
Фиг. 71B



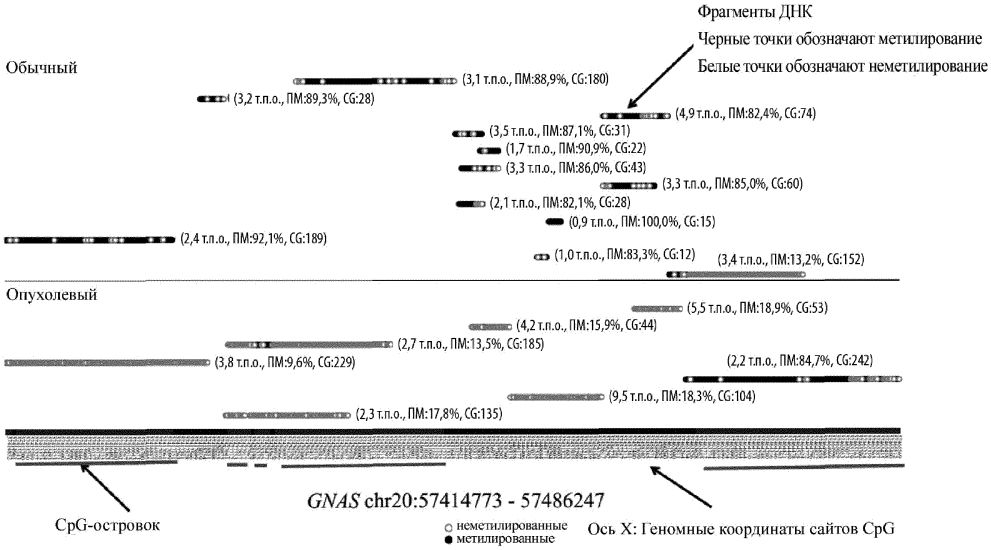
Фиг. 72A



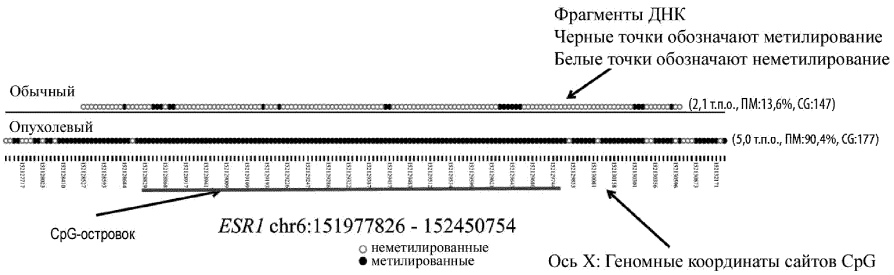
Фиг. 72B



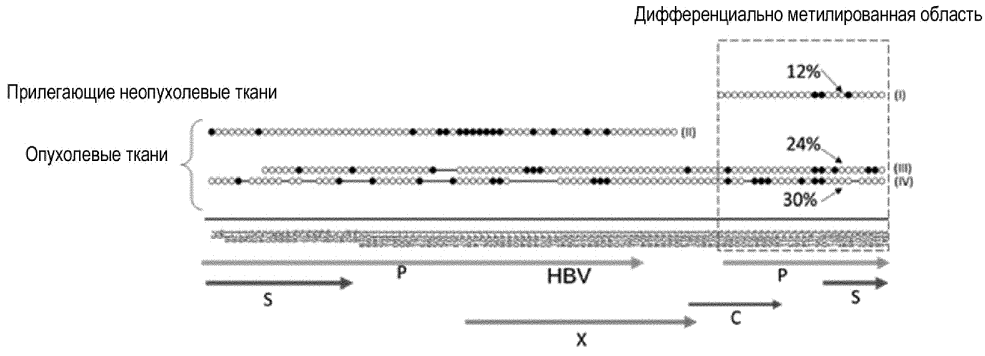
Фиг. 73



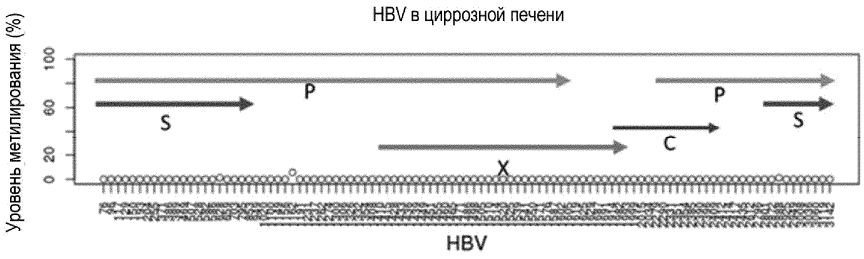
Фиг. 74A



Фиг. 74B

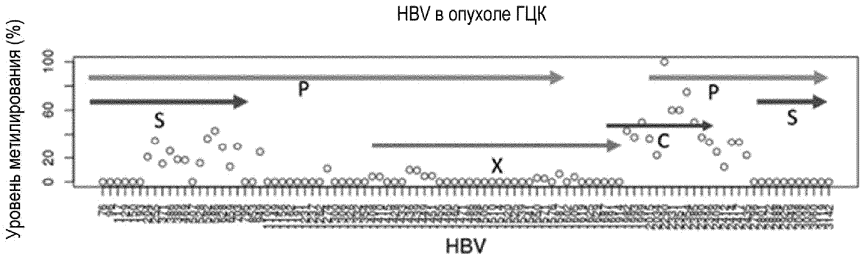


Фиг. 75



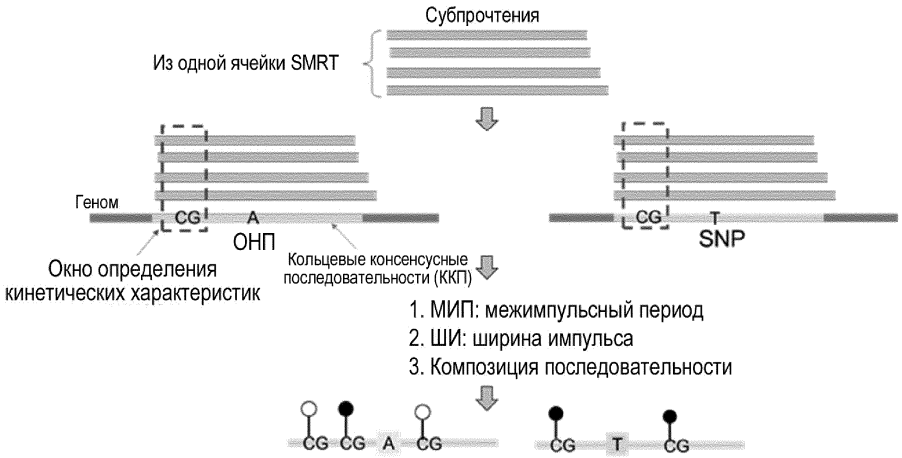
HBV в 4 образцах тканей циррозной печени: медианна глубины - 25 X

Фиг. 76A

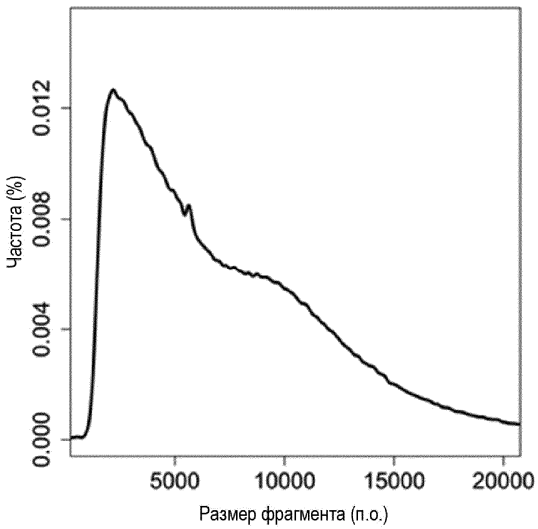


HBV в 15 образцах тканей ГЦК: медианна глубины - 14 X

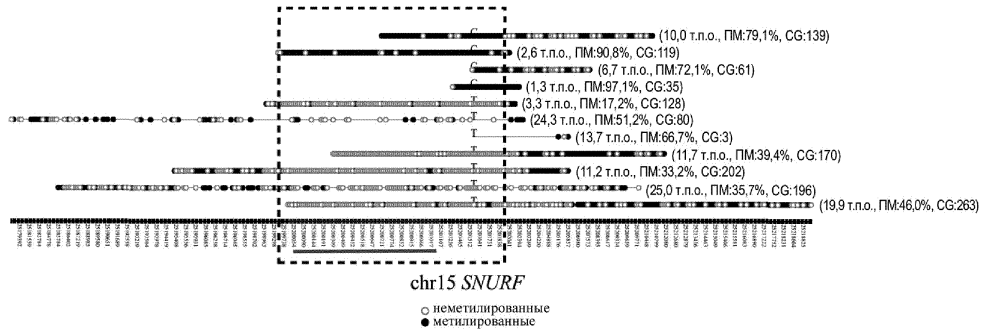
Фиг. 76B



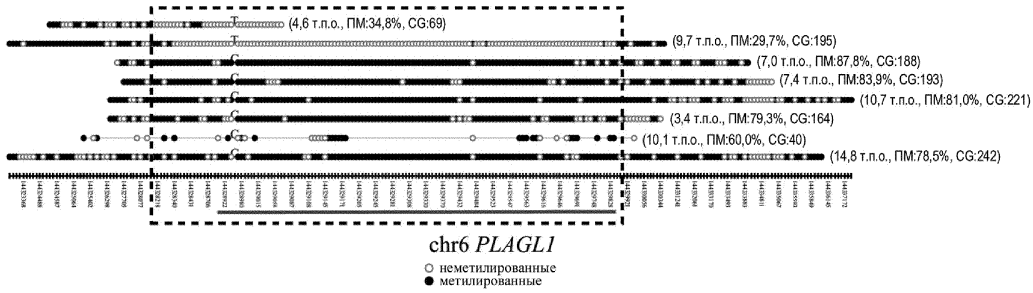
Фиг. 77



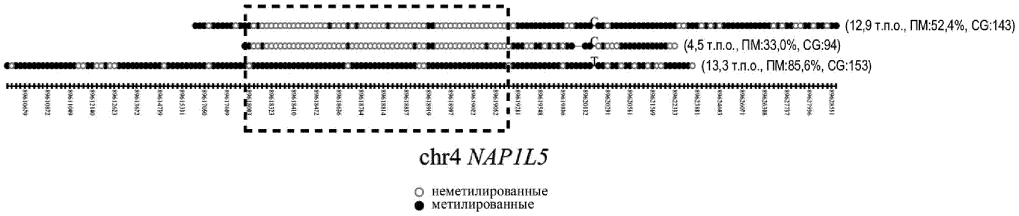
Фиг. 78



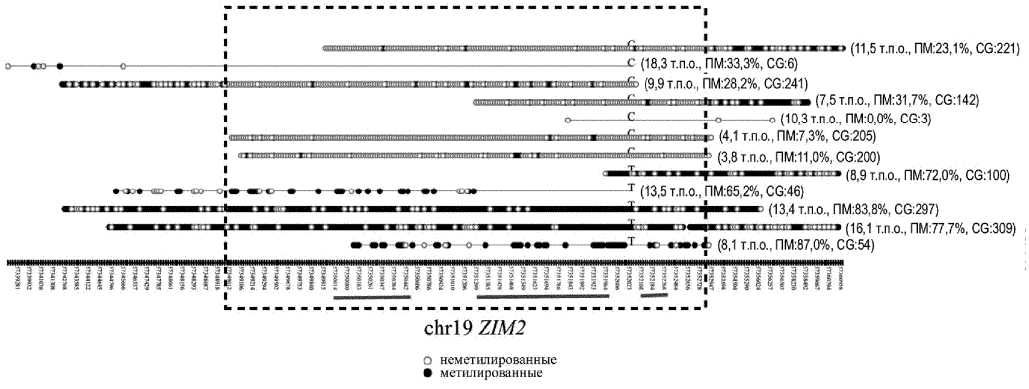
Фиг. 79A



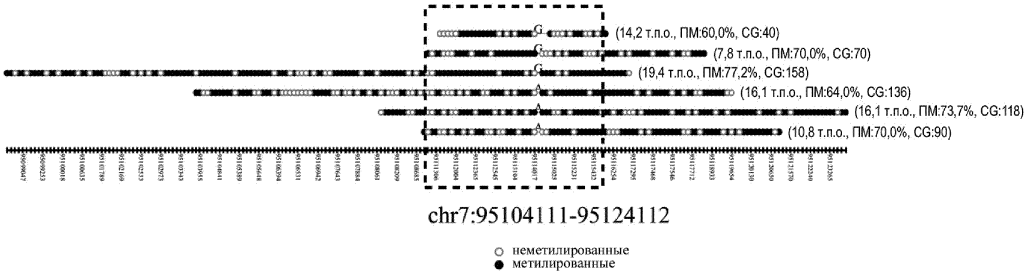
Фиг. 79B



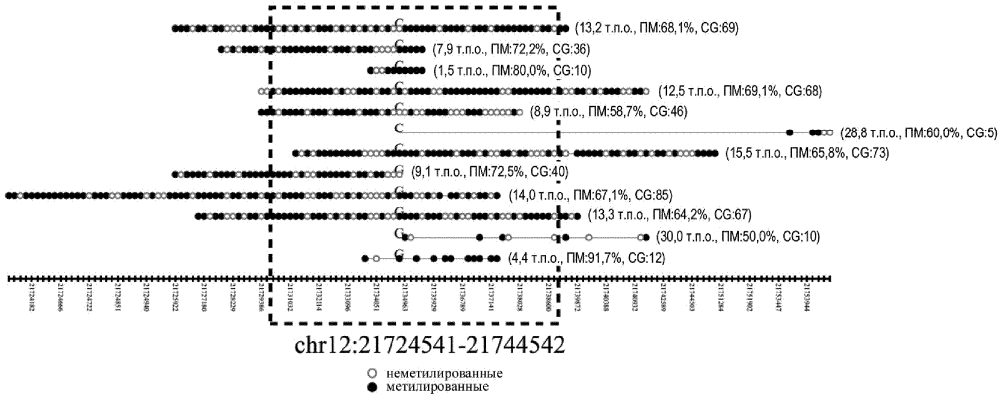
Фиг. 79C



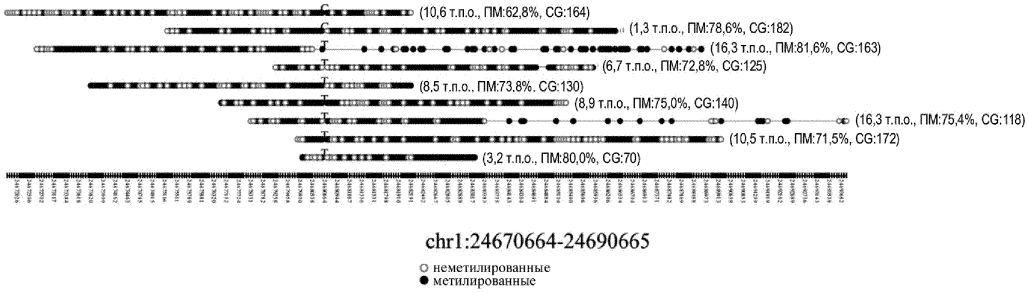
Фиг. 79D



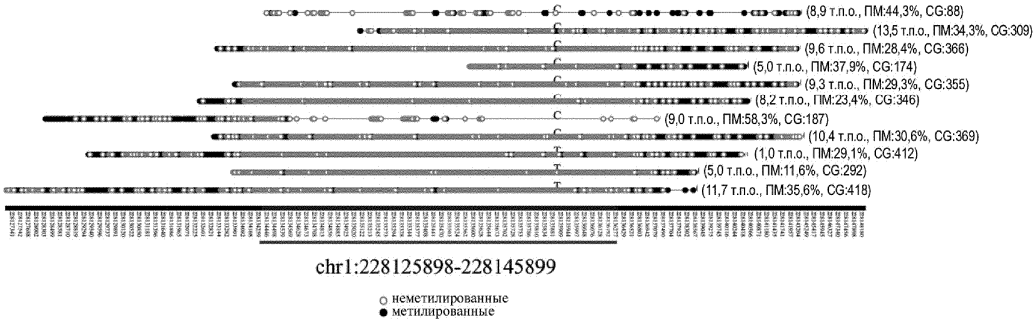
Фиг. 80A



Фиг. 80B



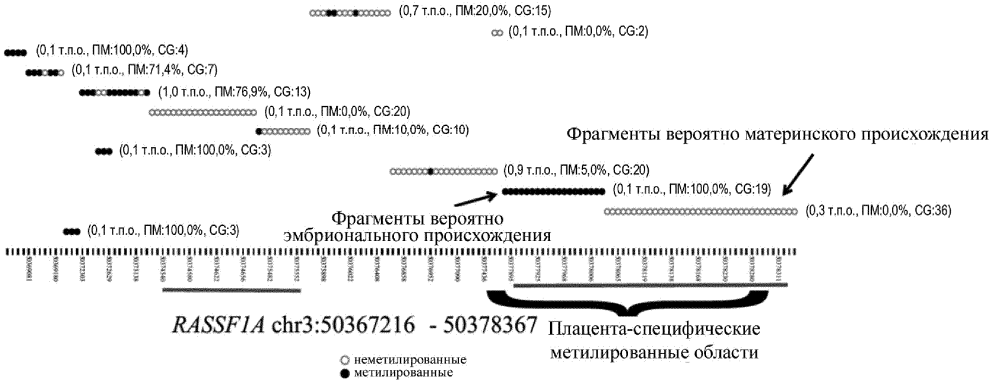
Фиг. 80C



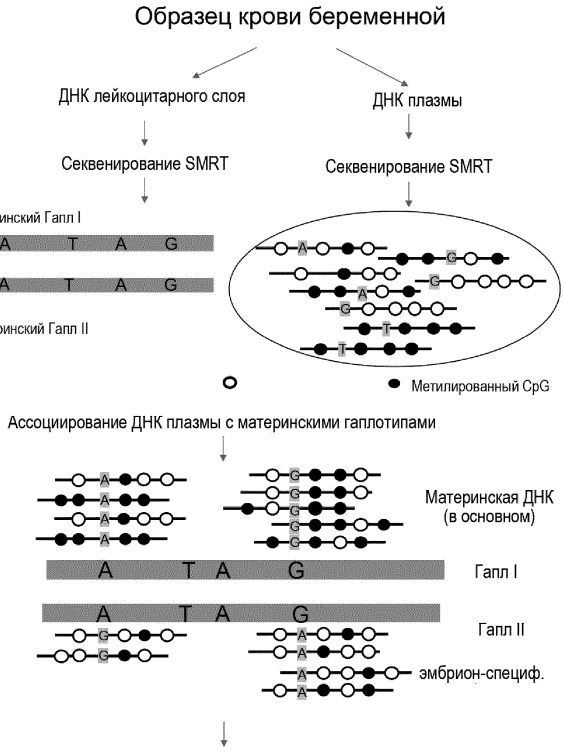
Фиг. 80D

	Ген	Аллель 1	Аллель 2	Уровень метилирования (%)	
				Аллель 1	Аллель 2
Подверженные импринтингу гены	*SNURF*	T	C	15.73	89.37
	PLAGL1	T	C	7.56	89.41
	NAP1L5	C	T	12.5	91.07
	ZIM2	C	T	13	84.64
Случайно выбранные области	Область 01	G	A	71.79	69.17
	Область 02	T	G	63.22	65.22
	Область 03	C	T	73.33	74.9
	Область 04	C	T	10.83	8.51

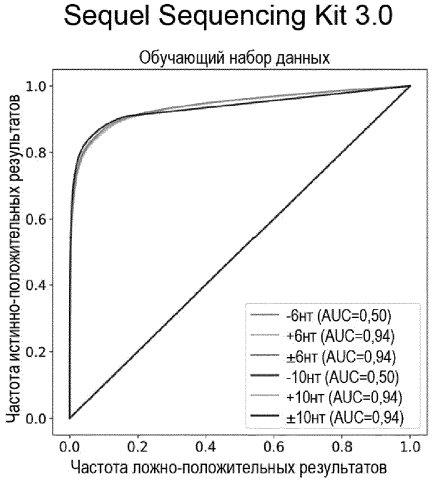
Фиг. 81



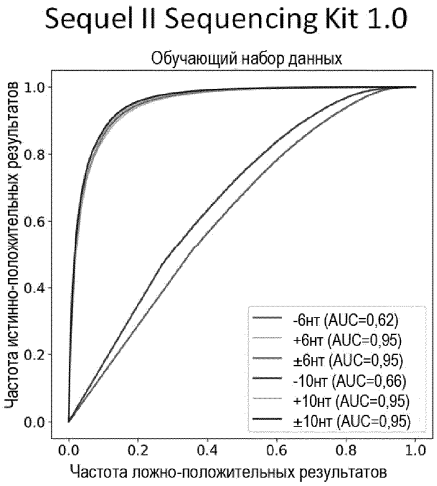
Фиг. 82



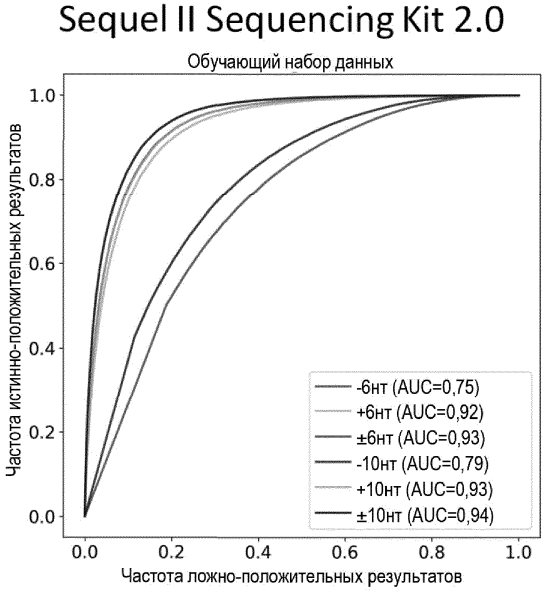
Фиг. 83



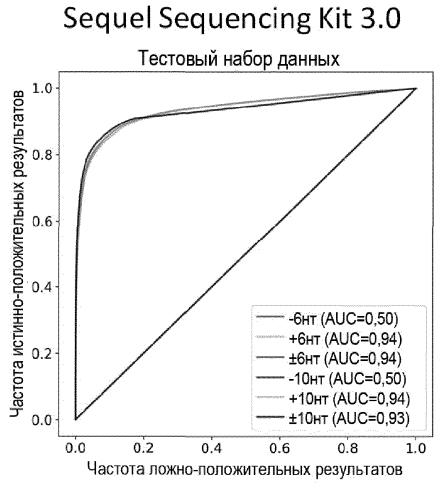
Фиг. 84A



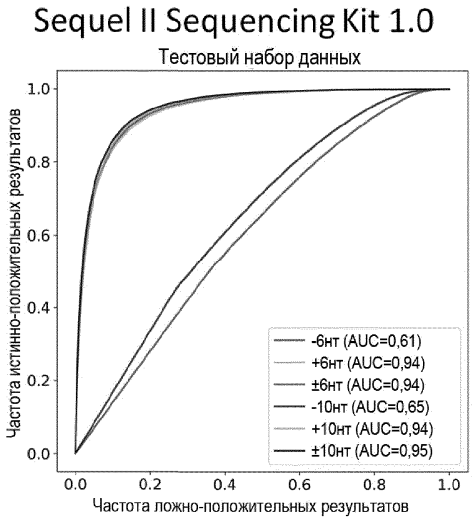
Фиг. 84B



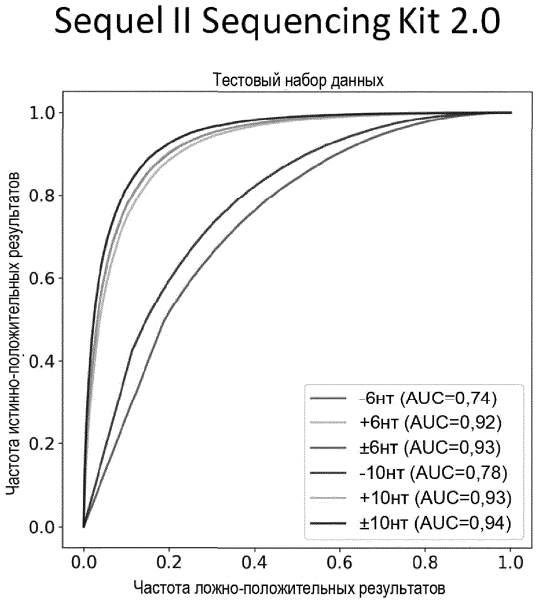
Фиг. 84C



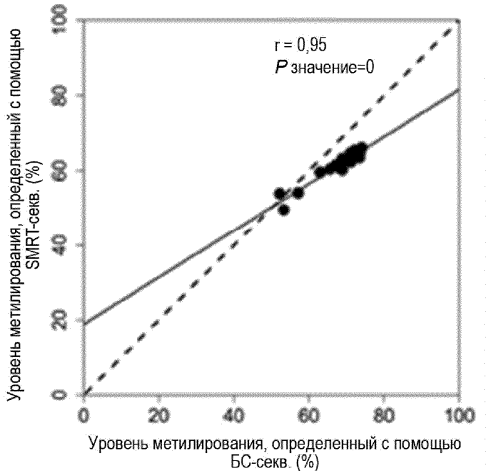
Фиг. 85A



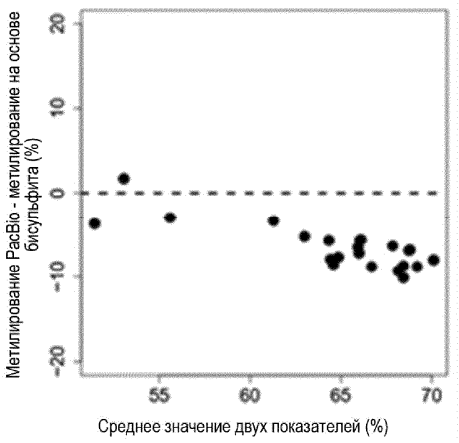
Фиг. 85B



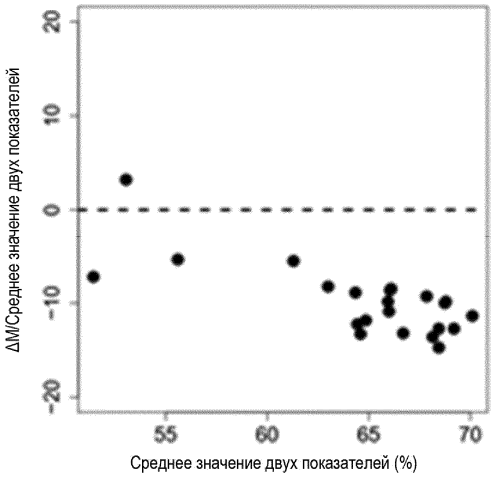
Фиг. 85C



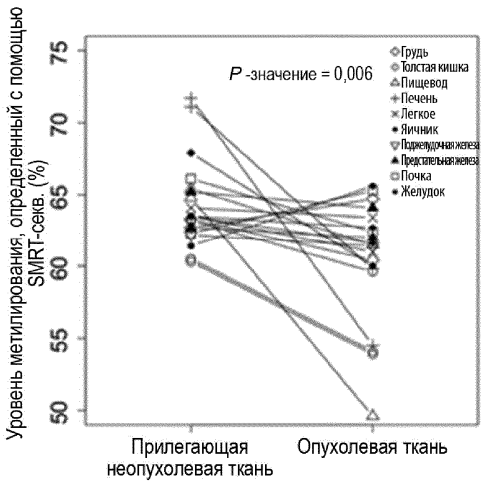
Фиг. 86A



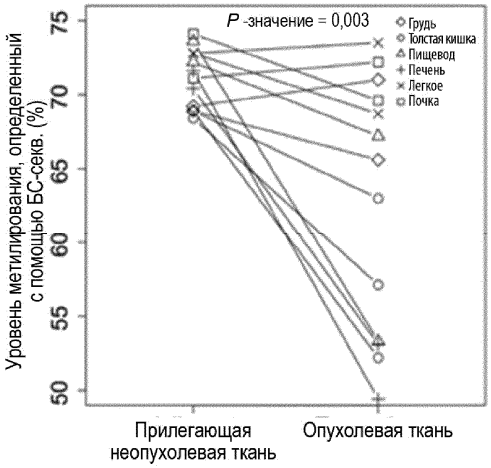
Фиг. 86B



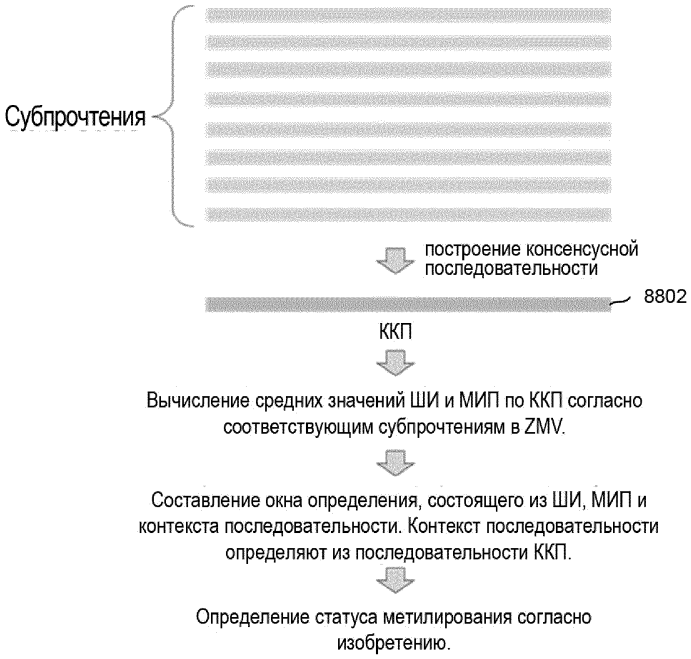
Фиг. 86C



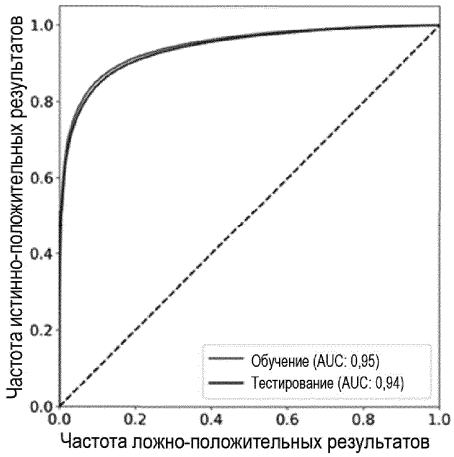
Фиг. 87A



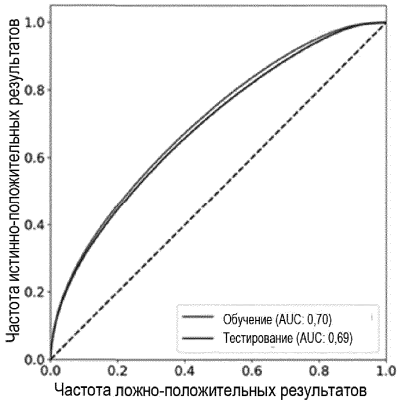
Фиг. 87B



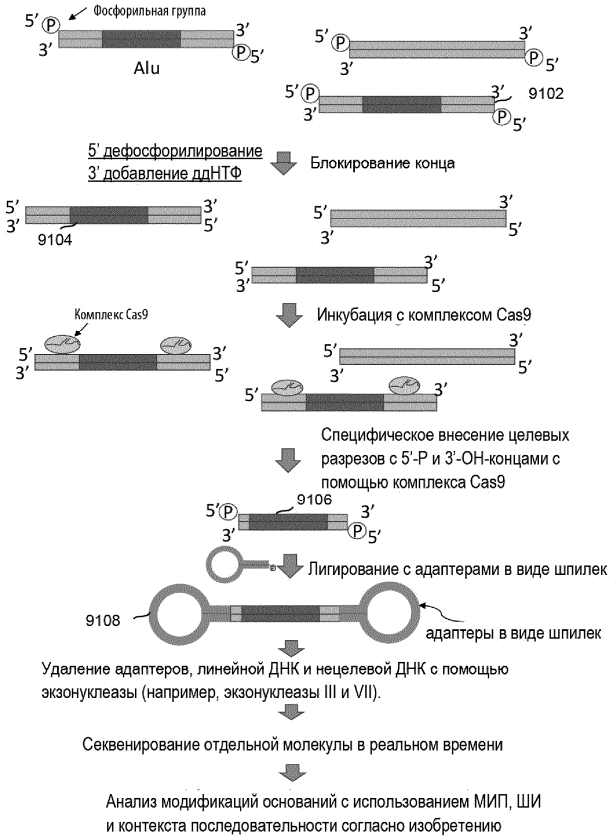
Фиг. 88



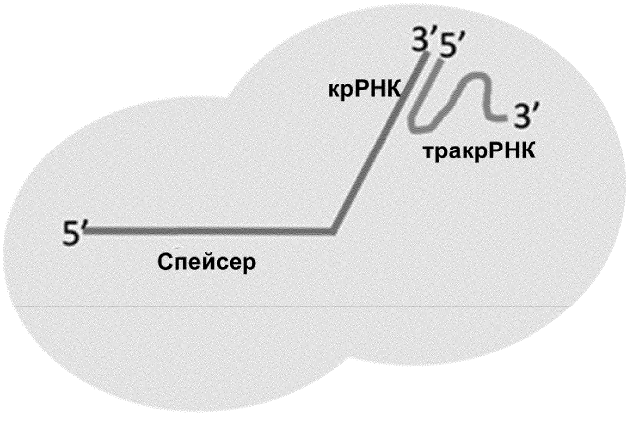
Фиг. 89



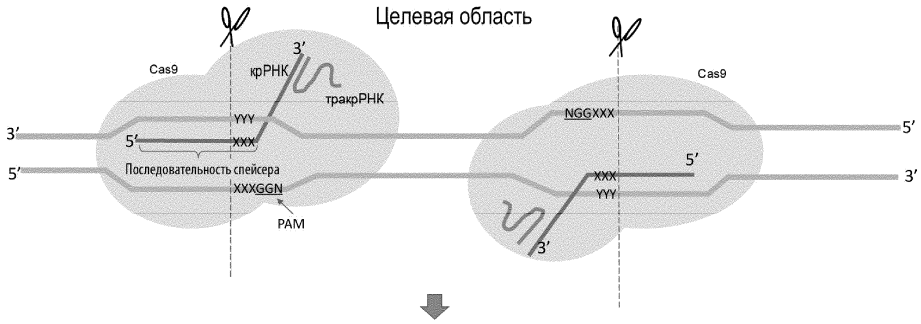
Фиг. 90



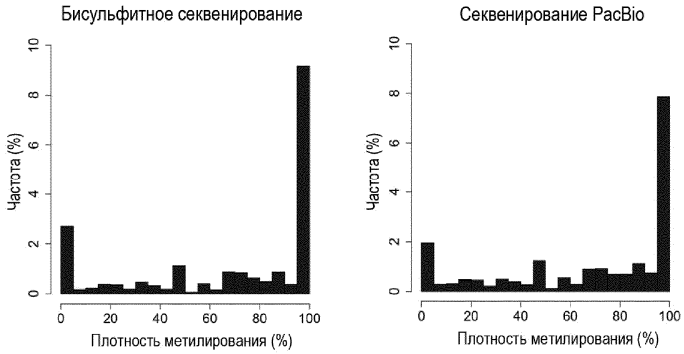
Фиг. 91



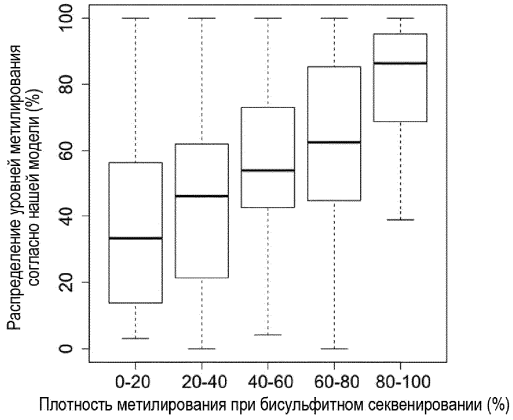
Фиг. 92



Фиг. 93



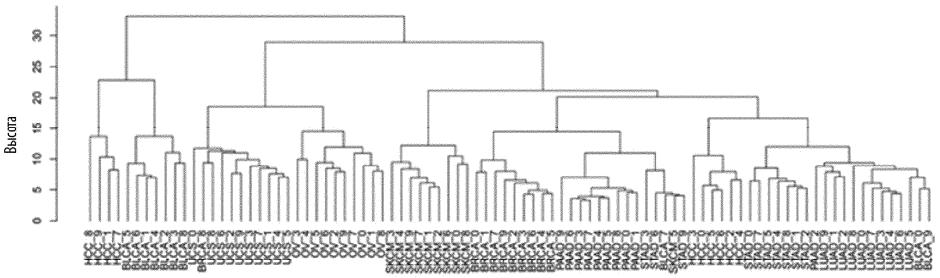
Фиг. 94



Фиг. 95

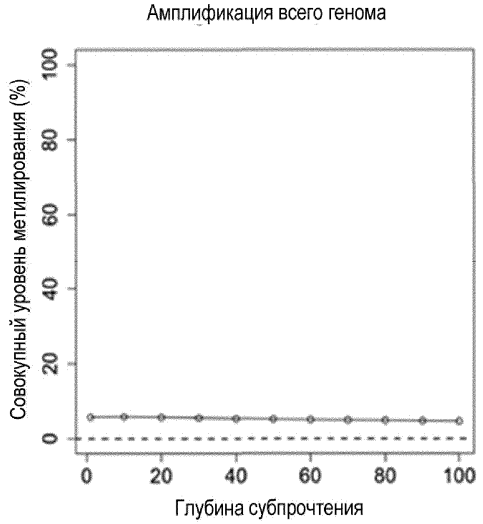
Ткани	Уровень метилирования Alu (%)
Лейкоцитарный слой	89.54
Печень	88.18
Толстая кишка	89.56
Легкое	91.52
Тонкий кишечник	86.56
Надпочечник	89.07
Жировая ткань	91.44
Поджелудочная железа	85.82
Мозг	91.79
ГЦК	76.74
Плацента	73.04

Фиг. 96

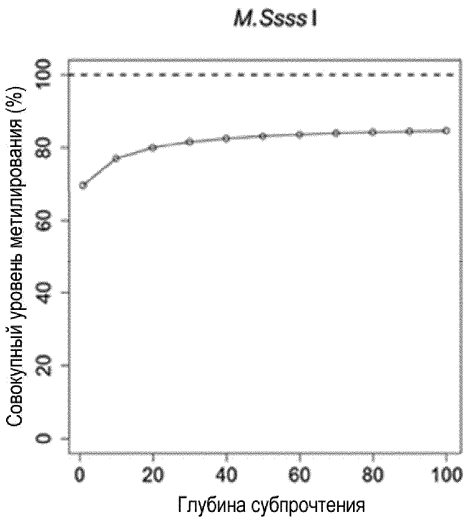


Типы рака
BLCA: Уротелиальная карцинома мочевого пузыря
BRCA: Инвазивная карцинома молочной железы
OV: Серозная аденокарцинома яичников
PAAD: Аденокарцинома поджелудочной железы
HCC: Гепатоцеллюлярная карцинома печени
LUAD: Аденокарцинома легкого
STAD: Аденокарцинома желудка
SKCM: Накожная меланома кожи
UCS: Карциносакрома матки;

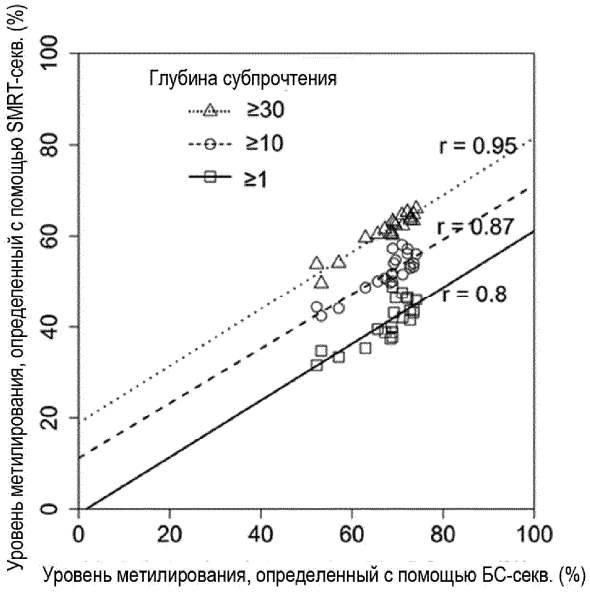
Фиг. 97



Фиг. 98A



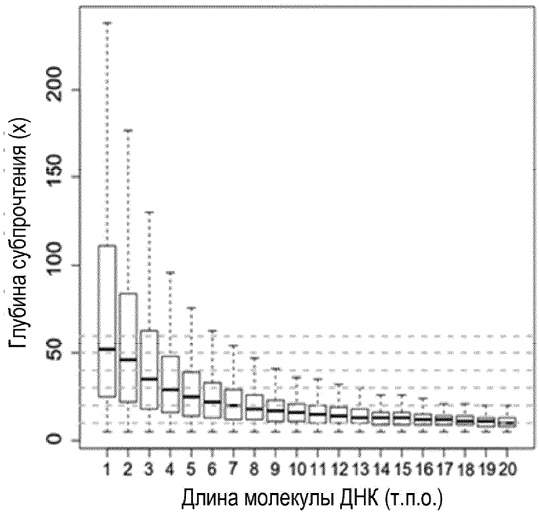
Фиг. 98B



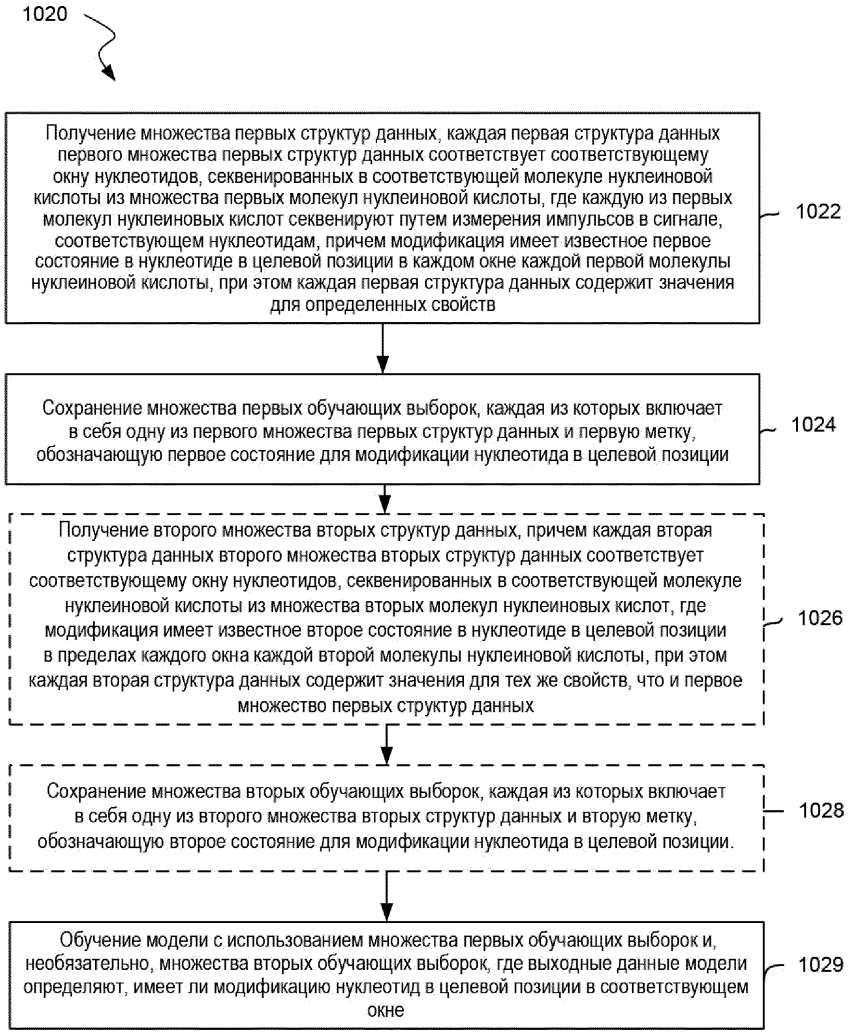
Фиг. 99

Пороговые значение глубины субпрочтений ≥	Коэффициент корреляции Пирсона (SMRT-секв. в сравнении с БС-секв.)	Количество сайтов CpG
1	0.797	25,606,068 (23,949,832-27,008,582)
10	0.873	21,668,418 (18,263,886-23,515,147)
20	0.933	14,276,212 (10,526,406-16,736,887)
30	0.952	6,736,890 (4,255,452-10,449,814)
40	0.948	3,420,790 (2,232,511-5,792,825)
50	0.941	1,684,871 (1,278,475-3,055,876)
60	0.929	911,961 (707,295-1,581,313)
70	0.917	532,422 (350,001-866,045)
80	0.907	284,375 (177,698-534,540)
90	0.906	150,974 (98,000-333,933)
100	0.875	89,788 (58,552-182,861)

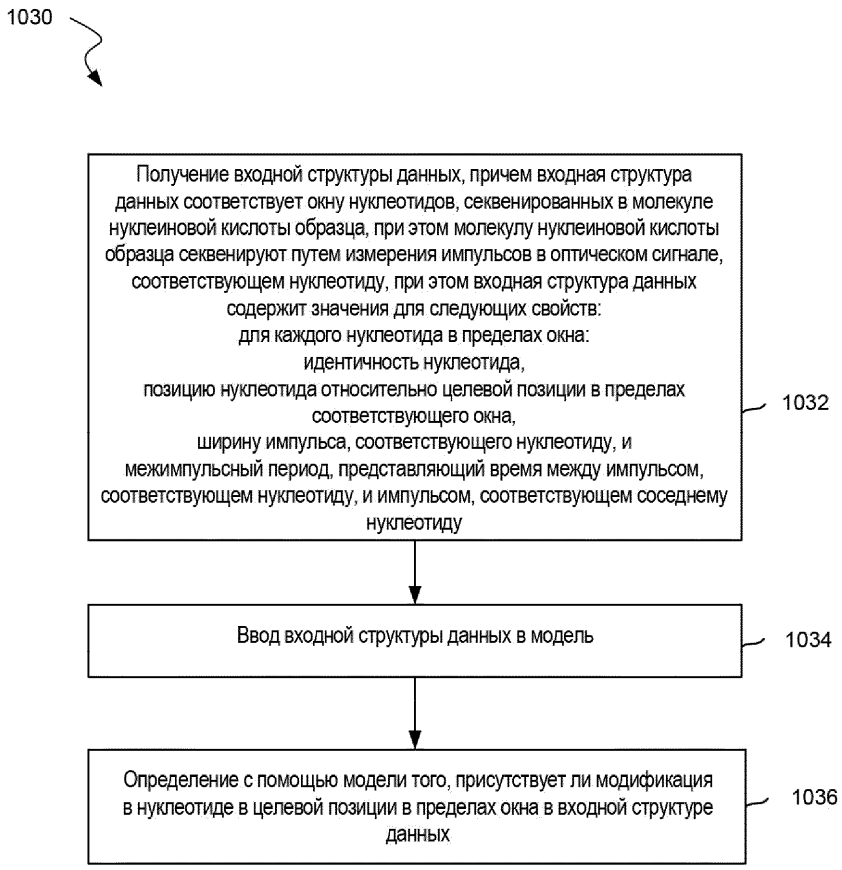
Фиг. 100



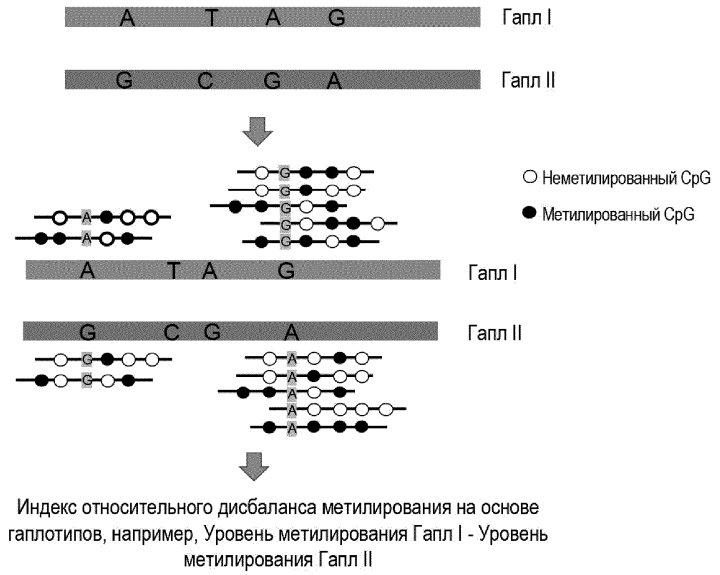
Фиг. 101



Фиг. 102



Фиг. 103



Фиг. 104

Chr	начало	конец	Длинна	Идентификатор гаплотипного блока	Секвенирование PacBio			
					Уровень метилирования в прилегающей неопухолевой ткани		Уровень метилирования в опухолевой ткани	
					Гапл I	Гапл II	Гапл I	Гапл II
chr1	56312395	56347696	35301	hap1927	68.2	67.4	60.3	23.5
chr1	194413819	194424806	10987	hap5953	52.8	49.5	48.8	9.3
chr1	220674478	220699011	24533	hap6863	63.0	64.5	50.4	17.3
chr10	113088792	113124248	35456	hap11838	62.7	63.4	38.1	5.7
chr11	5482746	5498801	16055	hap12904	70.3	75.0	16.3	51.7
chr11	42819351	42852772	33421	hap14385	54.6	54.9	65.3	17.8
chr11	57983961	58051078	67117	hap14930	67.3	66.4	58.2	18.6
chr11	60174708	60204209	29501	hap14990	58.4	59.8	49.6	10.8
chr12	128079419	128114656	35237	hap22249	60.0	58.3	12.1	45.2
chr15	20480575	20533464	52889	hap29631	64.7	69.1	27.7	59.3
chr15	94902853	94946231	43378	hap32161	74.1	74.5	74.9	15.8
chr15	96526684	96549225	22541	hap32221	70.8	68.8	28.9	64.4
chr16	31595372	31613277	17905	hap33499	55.9	59.3	46.3	14.4
chr16	80151778	80182097	30319	hap34821	71.1	71.0	11.5	51.8
chr16	82519715	82554191	34476	hap34920	71.3	66.5	47.4	13.0
chr17	21668593	21685572	16979	hap36049	50.3	47.8	67.4	19.6
chr17	44999177	45012087	12910	hap36640	47.1	45.2	81.6	35.1
chr17	69911623	69926625	15002	hap37435	67.3	63.0	37.8	5.2
chr18	11441122	11458521	17399	hap38335	65.5	66.8	65.9	22.4
chr18	23405569	23423387	17818	hap38673	66.3	61.7	3.3	48.1
chr18	68887284	68925031	37747	hap40390	63.0	61.0	22.0	53.4
chr18	69487809	69505470	17661	hap40414	74.5	74.1	33.3	72.2
chr2	41480394	41514135	33741	hap43972	54.0	54.0	14.9	77.8
chr2	114171214	114182880	11666	hap46226	72.4	68.8	79.7	16.7
chr2	123762541	123797629	35088	hap46589	66.7	68.1	24.0	54.5
chr2	125236882	125241950	5068	hap46673	58.9	59.2	10.7	46.4
chr2	130016110	130040331	24221	hap46835	54.6	50.8	5.6	41.6
chr2	137757638	137783716	26078	hap47090	61.8	61.4	13.5	69.2
chr2	144128597	144160845	32248	hap47343	65.8	66.6	9.3	50.3
chr20	15736792	15753459	16667	hap51505	78.9	74.3	45.8	77.3
chr20	26167979	26177235	9256	hap51868	55.0	52.2	38.5	68.6
chr20	44255808	44264190	8382	hap52246	57.4	56.1	9.7	50.6
chr20	59518410	59559273	40863	hap52761	61.0	62.4	30.0	72.8
chr21	21402034	21424129	22095	hap53197	63.5	67.3	25.0	75.5
chr21	24750027	24768793	18766	hap53333	68.2	64.6	3.4	38.9
chr21	26666833	26701575	34742	hap53418	62.1	66.5	47.6	16.7
chr3	2364024	2387896	23872	hap55539	67.4	67.8	54.9	10.9
chr3	21036965	21049451	12486	hap56223	54.8	51.4	53.1	21.1
chr3	56011690	56046642	34952	hap57346	64.2	61.2	71.2	22.6

Фиг. 105A

chr3	73330942	73371216	40274	hap57939	60.9	62.9	9.4	42.9
chr3	106372440	106401301	28861	hap59077	67.8	67.9	13.8	53.2
chr3	107772994	107807482	34488	hap59122	69.6	73.5	30.4	66.4
chr3	116742501	116776747	34246	hap59493	64.3	69.1	14.1	51.6
chr3	171076306	171100102	23796	hap61495	68.0	66.0	80.6	48.8
chr3	193058272	193080344	22072	hap62231	65.5	64.7	54.6	20.0
chr4	30411613	30432317	20704	hap63589	59.3	60.6	53.4	14.6
chr4	31304718	31338193	33475	hap63633	60.2	60.0	7.2	55.0
chr4	92003467	92030505	27038	hap65794	65.3	65.1	54.1	21.7
chr4	155224697	155250915	26218	hap68104	60.5	57.5	57.3	25.0
chr5	2281802	2299281	17479	hap69632	71.5	66.9	69.9	6.6
chr5	4624948	4664704	39756	hap69739	62.8	61.0	14.0	52.0
chr5	89593236	89606080	12844	hap72628	76.6	74.0	20.3	78.4
chr5	119214026	119233058	19032	hap73698	62.8	61.2	57.6	13.1
chr5	119940397	119972658	32261	hap73720	59.1	54.7	53.8	12.2
chr5	132859668	132877415	17747	hap74150	62.5	66.6	59.5	28.3
chr6	26914610	26936918	22308	hap76887	41.9	40.9	71.9	32.6
chr6	66879106	66957243	78137	hap78266	61.6	59.6	25.4	62.0
chr6	77349083	77377529	28446	hap78674	64.5	66.4	27.0	62.9
chr6	159738794	159751033	12239	hap81616	79.6	79.0	21.2	59.8
chr7	26585255	26641907	56652	hap83161	66.2	64.7	49.4	13.3
chr7	48214640	48248036	33396	hap84003	76.0	76.7	78.0	32.3
chr7	88558182	88575482	17300	hap85335	63.8	59.6	63.8	22.9
chr7	96588562	96607580	19018	hap85620	60.4	63.1	19.7	50.0
chr7	122942180	122956897	14717	hap86454	42.3	39.0	19.2	50.0
chr7	132321970	132344802	22832	hap86807	61.4	60.7	52.5	11.5
chr7	153296219	153302441	6222	hap87487	48.7	53.7	64.4	19.3
chr7	156356247	156371897	15650	hap87631	74.9	71.6	87.5	56.6
chr7	159091986	159119486	27500	hap87738	54.0	49.1	52.0	13.2
chr8	51530582	51550889	20307	hap89477	66.4	65.7	68.0	19.9
chr8	63513932	63537543	23611	hap89942	62.0	63.3	11.6	48.4
chr8	72373321	72398122	24801	hap90226	58.0	54.9	71.6	32.0
chr8	94100451	94141855	41404	hap90991	65.2	65.7	36.2	68.7
chr8	109300499	109326404	25905	hap91510	63.6	67.7	29.5	65.8

Фиг. 105B

Chr	начало	конец	Длинна	Идентификатор гаплотипного блока	Секвенирование PacBio			
					Уровень метилирования в прилегающей неопухолевой ткани		Уровень метилирования в опухолевой ткани	
					Гапл I	Гапл II	Гапл I	Гапл II
chr9	27803548	27888202	84654	hap58508	64.2	60.9	20.6	75.4
chr6	242149	386636	144487	hap47880	62.3	63.3	77.4	32.2
chr5	28219159	28302858	83699	hap44666	59.3	58.0	16.8	58.2
chr5	18119943	18153743	33800	hap44475	61.6	65.0	53.2	21.7
chr7	24906307	25046195	139888	hap52069	69.3	68.7	44.0	76.2
chr15	27689897	27752573	62676	hap18337	65.9	61.9	64.8	20.5
chr12	42183870	42212433	28563	hap12045	63.5	68.4	19.4	51.2
chr21	9825597	9935752	110155	hap34175	54.3	53.5	60.9	29.1
chr2	118813055	118893366	80311	hap30060	62.6	62.3	77.0	38.6
chr6	90307702	90344869	37167	hap49779	69.1	66.4	84.7	53.9
chr7	107932914	108049376	116462	hap53838	67.2	62.9	43.8	76.4
chr7	137039327	137160933	121606	hap54447	59.5	60.9	22.9	72.0
chr17	21193754	21254930	61176	hap22633	59.2	54.3	69.7	31.6
chr12	11473697	11644714	171017	hap11451	62.8	66.4	35.5	75.9
chr5	129212299	129353349	141050	hap46632	50.9	54.5	45.5	14.0
chr11	93910738	94028887	118149	hap10288	67.6	63.6	36.6	74.2
chr3	131707434	132003636	296202	hap38642	57.8	55.9	17.9	60.2
chr3	43024004	43161785	137781	hap36769	69.1	66.5	46.1	80.2
chr3	190403156	190606658	203502	hap39947	60.9	61.6	36.9	72.7
chr15	40218970	40279780	60810	hap18606	53.4	57.5	79.1	47.4

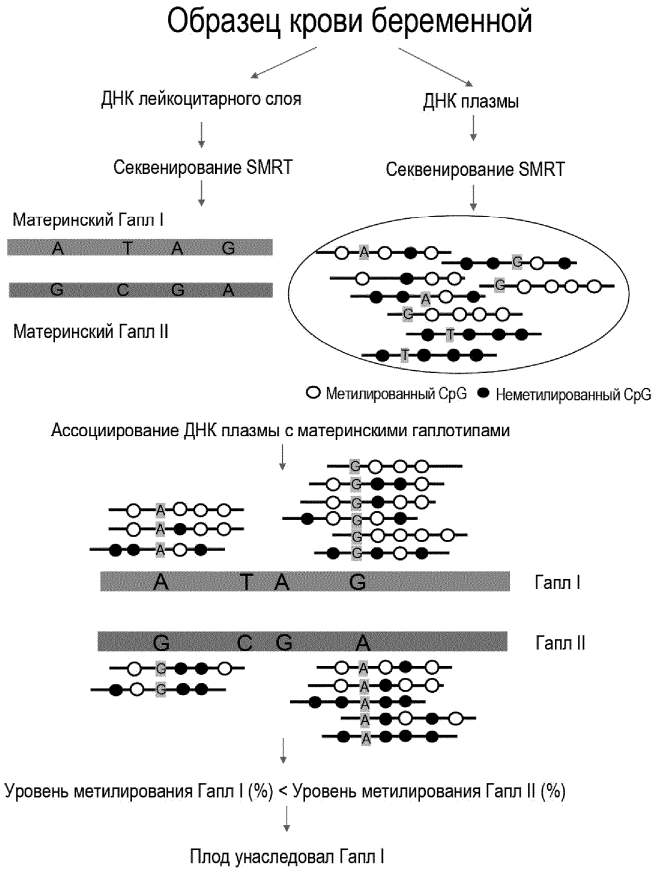
Фиг. 106

Типы тканей	Количество гаплотипных блоков, демонстрирующих дисбаланс метилирования между двумя гаплотипами в опухолевых тканях	Количество гаплотипных блоков, демонстрирующих дисбаланс метилирования между двумя гаплотипами в парных прилегающих неопухолевых тканях
Толстая кишка	92	47
Грудь	57	13
Почка	68	18
Легкое	31	21
Предстательная железа	26	19
Желудок	2	0

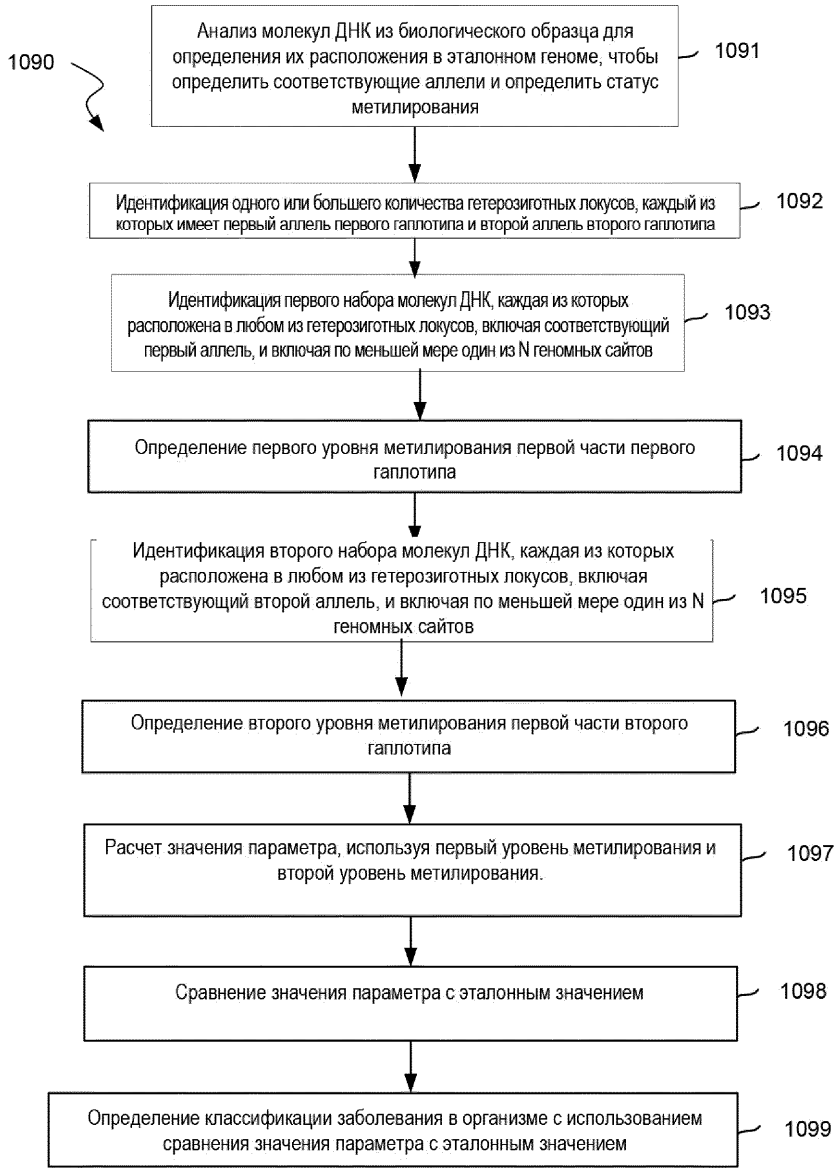
Фиг. 107A

Типы тканей	Количество гаплотипных блоков, демонстрирующих дисбаланс метилирования между двумя гаплотипами в опухолевых тканях	Доступна информация о стадии опухоли (TNM)
Грудь	18	T2
	57	T3
Почка	68	T3a
	0	T2

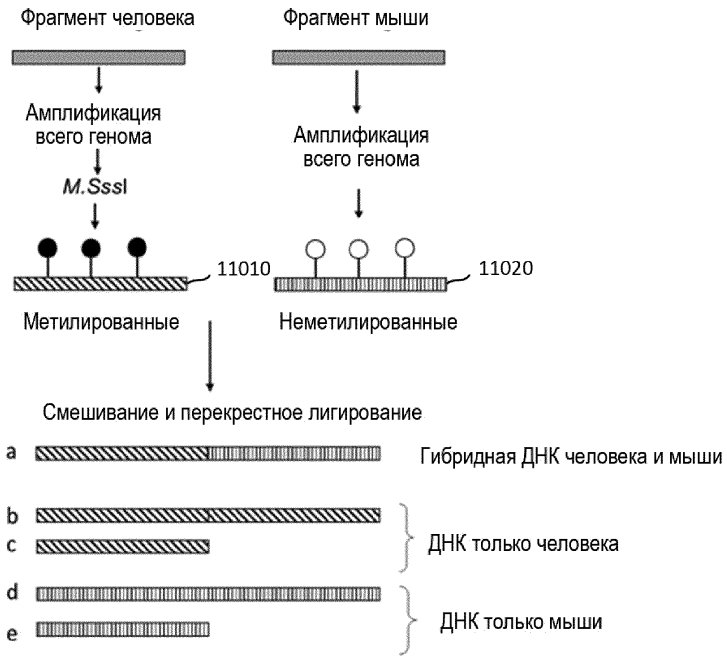
Фиг. 107B



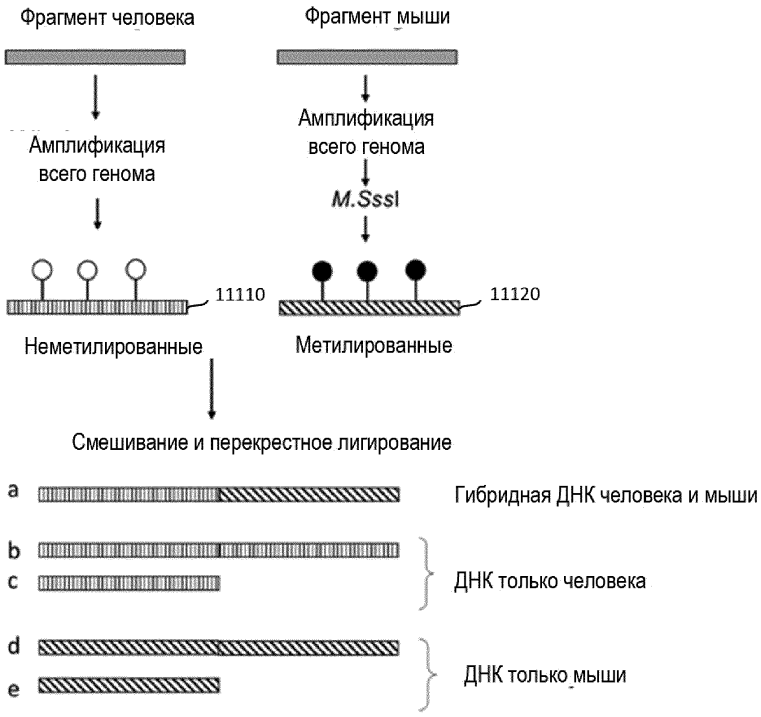
Фиг. 108


1090
Анализ молекул ДНК из биологического образца для определения их расположения в эталонном геноме, чтобы определить соответствующие аллели и определить статус метилирования
1091
Идентификация одного или большего количества гетерозиготных локусов, каждый из которых имеет первый аллель первого гаплотипа и второй аллель второго гаплотипа
1092
Идентификация первого набора молекул ДНК, каждая из которых расположена в любом из гетерозиготных локусов, включая соответствующий первый аллель, и включая по меньшей мере один из N геномных сайтов
1093
Определение первого уровня метилирования первой части первого гаплотипа
1094
Идентификация второго набора молекул ДНК, каждая из которых расположена в любом из гетерозиготных локусов, включая соответствующий второй аллель, и включая по меньшей мере один из N геномных сайтов
1095
Определение второго уровня метилирования первой части второго гаплотипа
1096
Расчет значения параметра, используя первый уровень метилирования и второй уровень метилирования.
1097
Сравнение значения параметра с эталонным значением
1098
Определение классификации заболевания в организме с использованием сравнения значения параметра с эталонным значением
1099

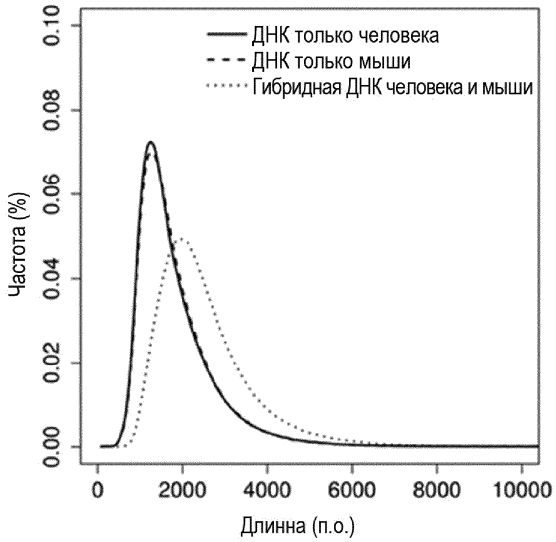
Фиг. 109



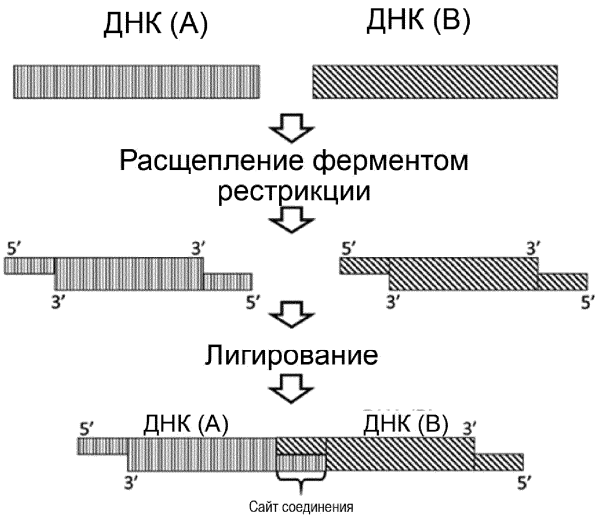
Фиг. 110



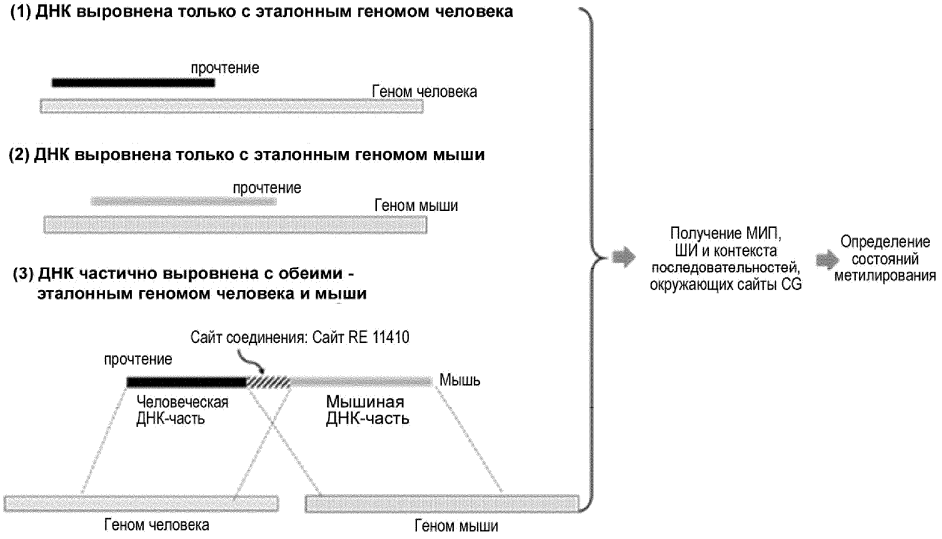
Фиг. 111



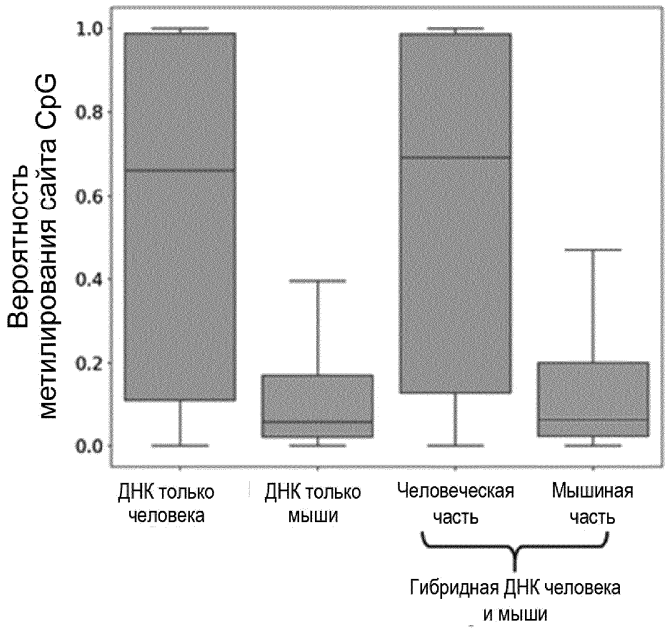
Фиг. 112



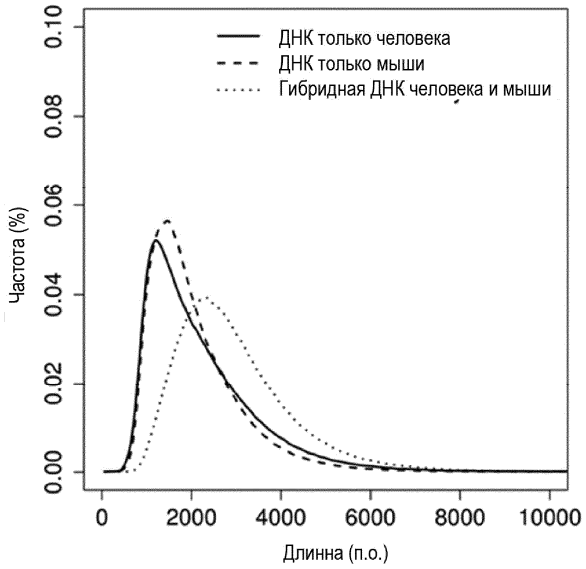
Фиг. 113



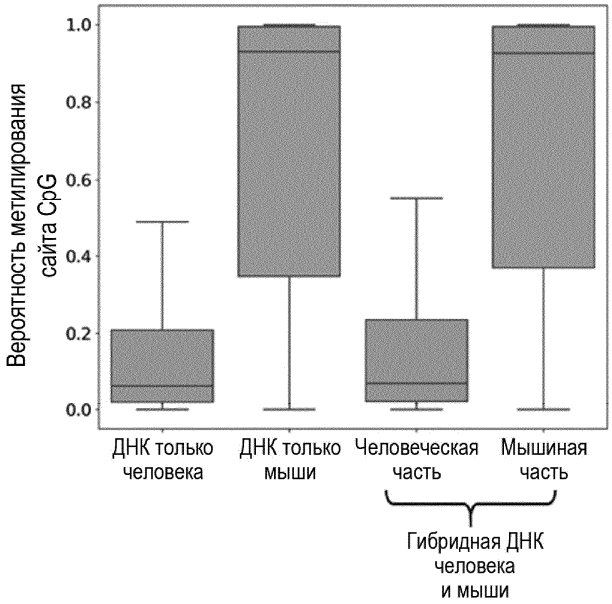
Фиг. 114



Фиг. 115



Фиг. 116



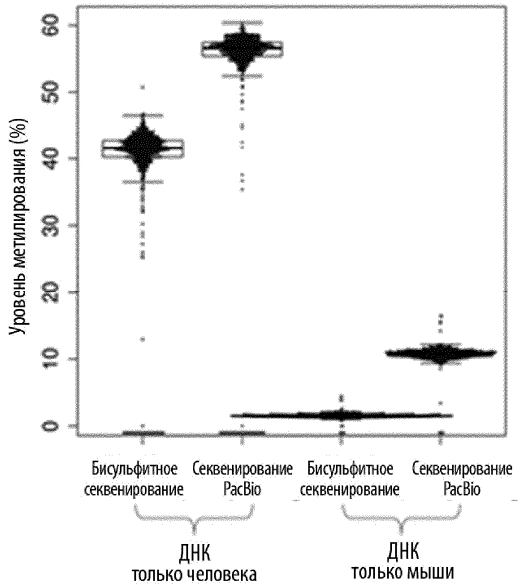
Фиг. 117

		Бисульфитное секвенирование		Секвенирование PacBio	
		Количество сайтов CpG	Плотность метилирования (%)	Количество сайтов CpG	Плотность метилирования (%)
1) Только человеческая		2,230,407	41.4	16,226,014	56.0
2) Только мышиная		2,726,499	1.6	9,398,340	10.7
3) Гибридная ДНК человека и мыши	Человеческая часть	73,780	46.8	4,838,454	57.4
	Мышиная часть	76,312	2.3	4,385,046	12.1

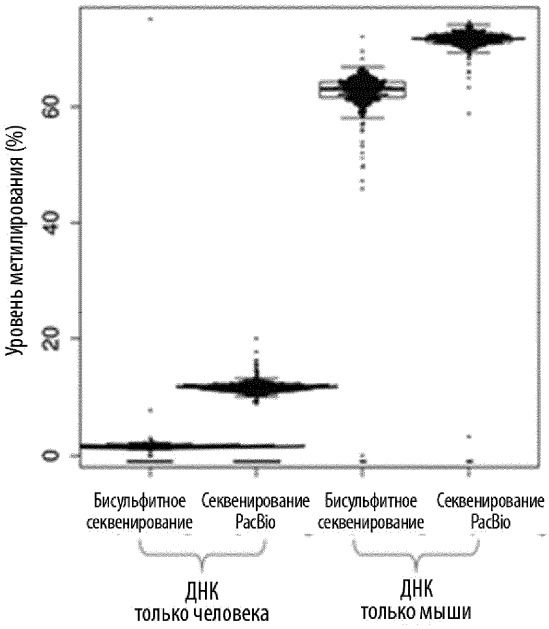
Фиг. 118

		Бисульфитное секвенирование		Секвенирование PacBio	
		Количество сайтов CpG	Плотность метилирования (%)	Количество сайтов CpG	Плотность метилирования (%)
1) Только человеческая		2,938,088	1.6	14,503,548	11.6
2) Только мышиная		1,513,971	62.4	11,348,555	71.5
3) Гибридная ДНК человека и мыши	Человеческая часть	67,371	1.8	5,824,379	13.1
	Мышиная часть	58,242	67.4	5,093,097	72.2

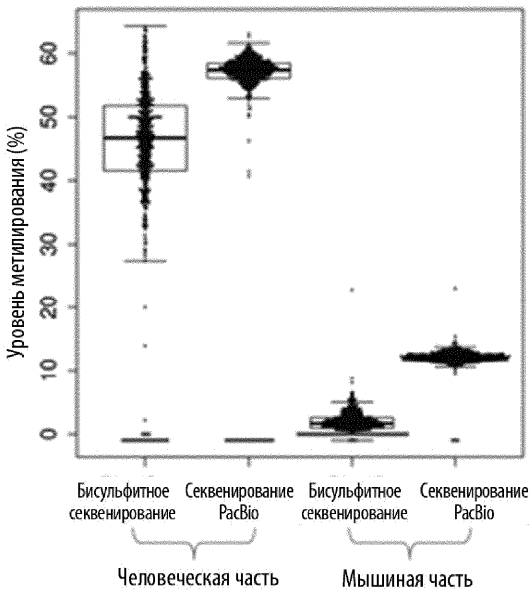
Фиг. 119



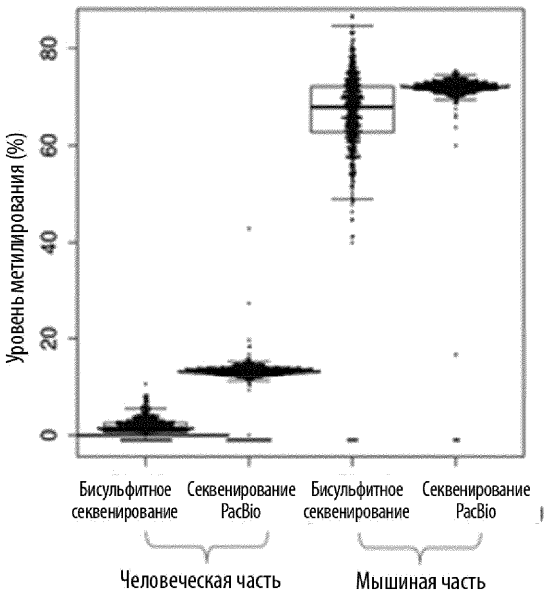
Фиг. 120А



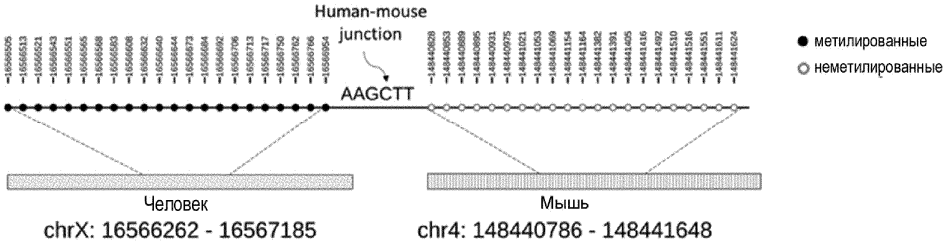
Фиг. 120B



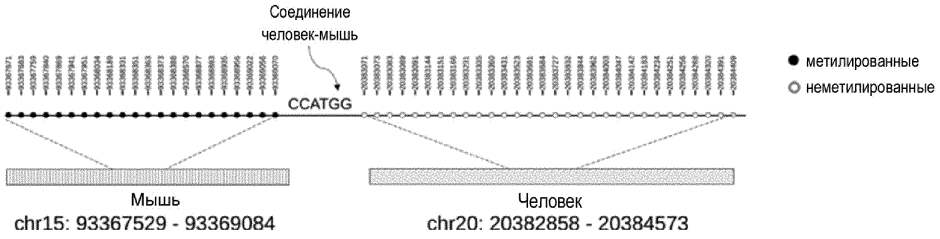
Фиг. 121A



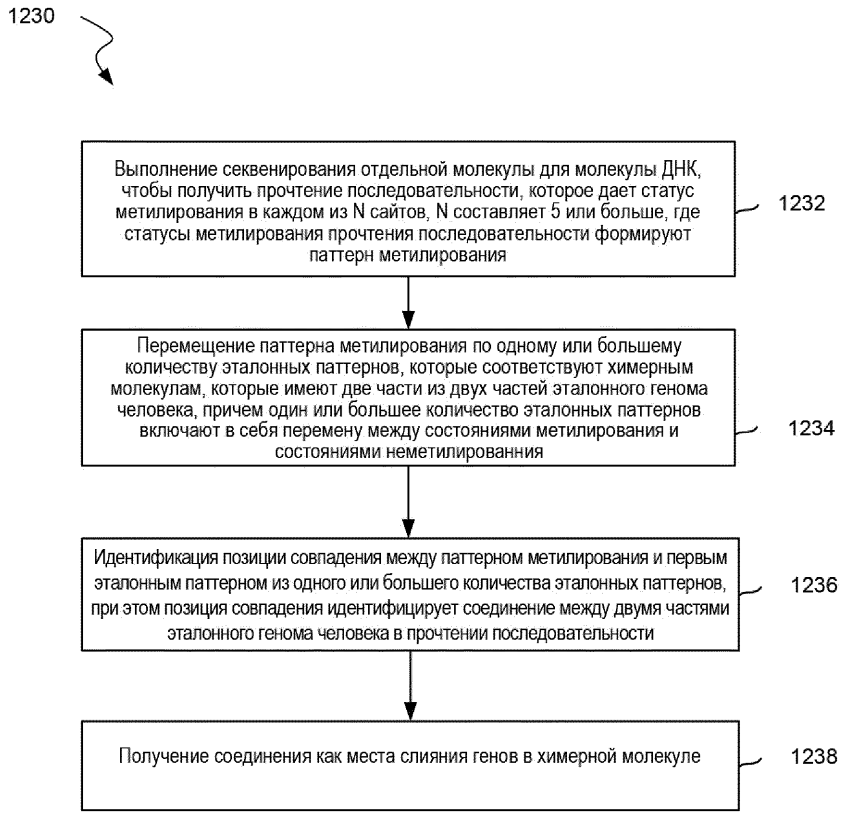
Фиг. 121B



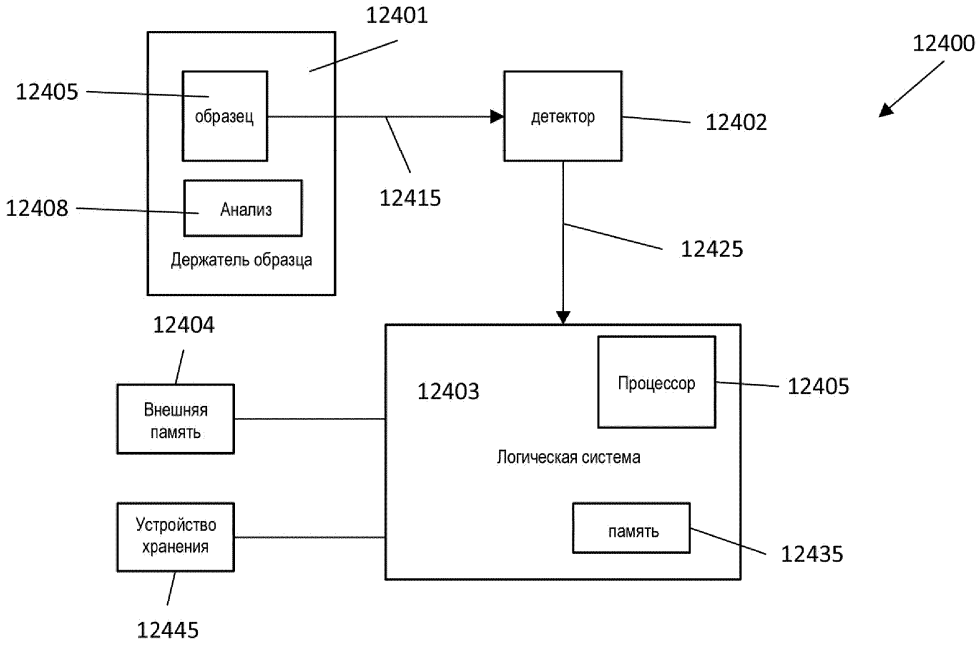
Фиг. 122 A



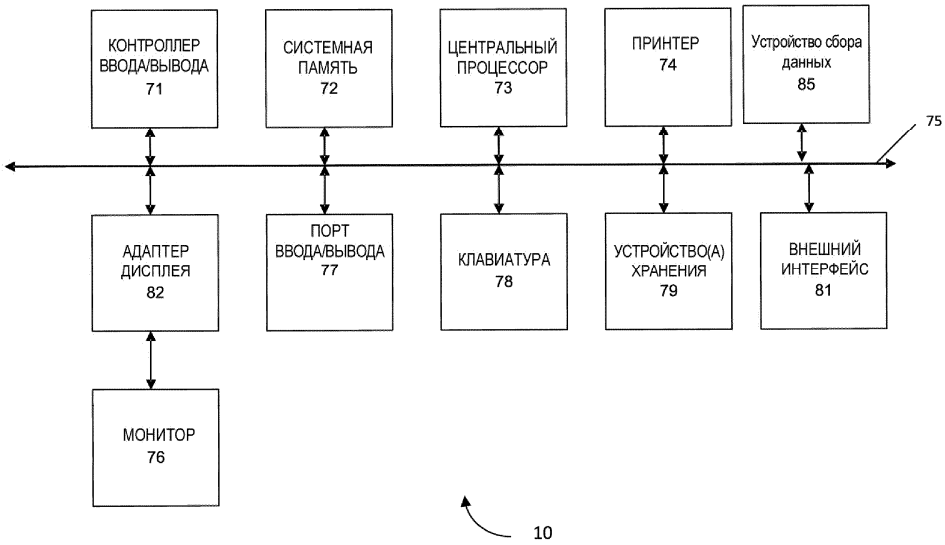
Фиг. 122B



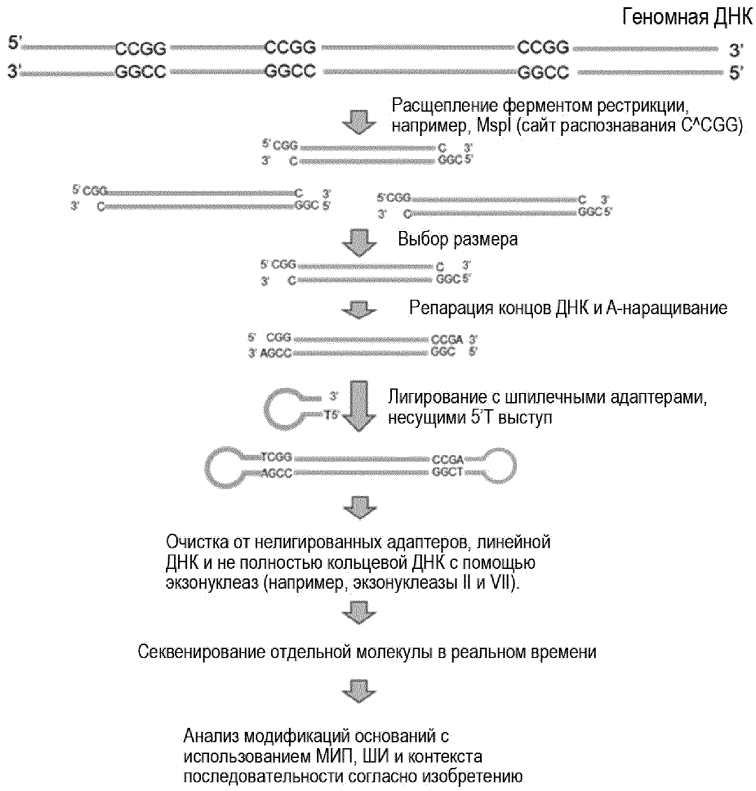
Фиг. 123



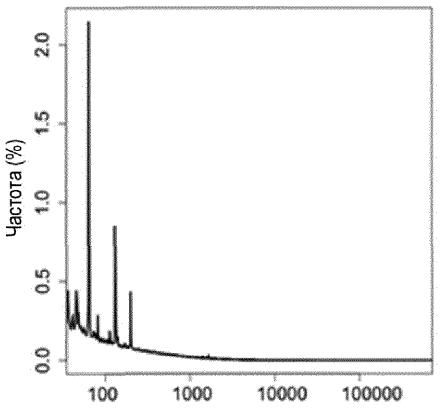
Фиг. 124



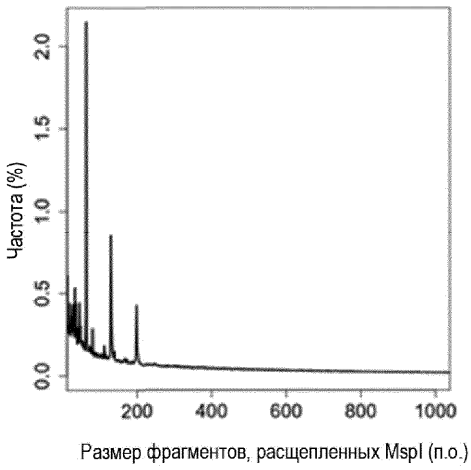
Фиг. 125



Фиг. 126



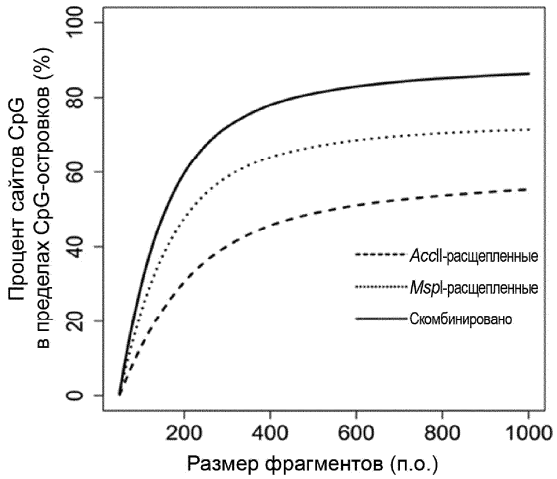
Фиг. 127A



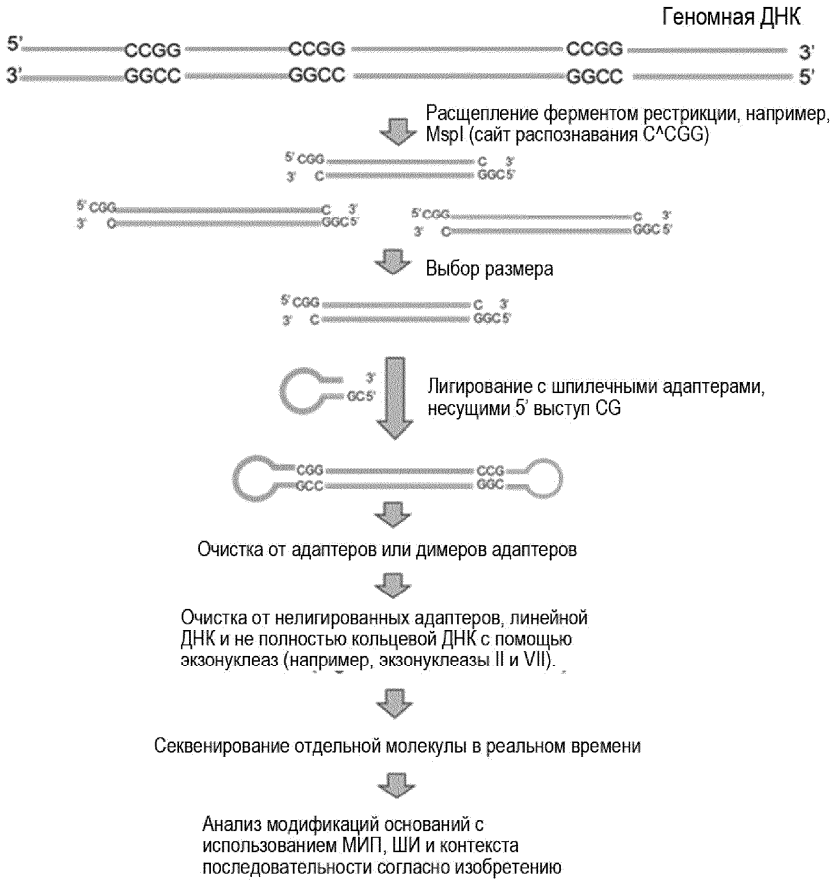
Фиг. 127B

Диапазон размеров (п.о.)	Количество молекул	Процент молекул в пределах диапазона размеров по отношению к совокупному количеству фрагментов (%)	Количество молекул в диапазоне размеров, перекрывающихся с островками CpG	Процент молекул в диапазоне размеров, перекрывающихся с островками CpG (%)	Количество секвенируемых сайтов CpG	Количество сайтов CpG, попадающих в пределы островков CpG	Процент сайтов CpG, на которые распространяется выбор размера и которые попадают в пределы островков CpG (%)
50-200	526,543	23.03	104,059	19.76	2,358,020	885,041	37.53
200-400	269,562	11.79	23,927	8.88	1,781,556	353,087	19.82
400-600	177,776	7.77	7,369	4.15	1,468,561	107,130	7.29
600-800	133,927	5.86	3,673	2.74	1,326,544	48,851	3.68
800-1000	104,976	4.59	2,168	2.07	1,193,233	25,821	2.16
1000-2000	311,596	13.63	4,596	1.47	4,610,504	58,288	1.26
2000-3000	149,468	6.54	1,771	1.18	3,036,951	25,106	0.83
3000-4000	86,760	3.79	809	0.93	2,165,171	10,785	0.50
5000-6000	36,931	1.62	266	0.72	1,242,712	3,412	0.27
6000-7000	25,027	1.09	202	0.81	947,874	3,354	0.35
7000-8000	17,597	0.77	86	0.49	736,830	791	0.11
8000-9000	12,658	0.55	76	0.60	583,680	993	0.17
9000-10000	9,184	0.40	48	0.52	461,935	591	0.13
10000-15000	20,790	0.91	97	0.47	1,255,731	2,003	0.16
15000-20000	5,111	0.22	16	0.31	414,400	163	0.04
20000-25000	1,441	0.06	6	0.42	147,731	34	0.02

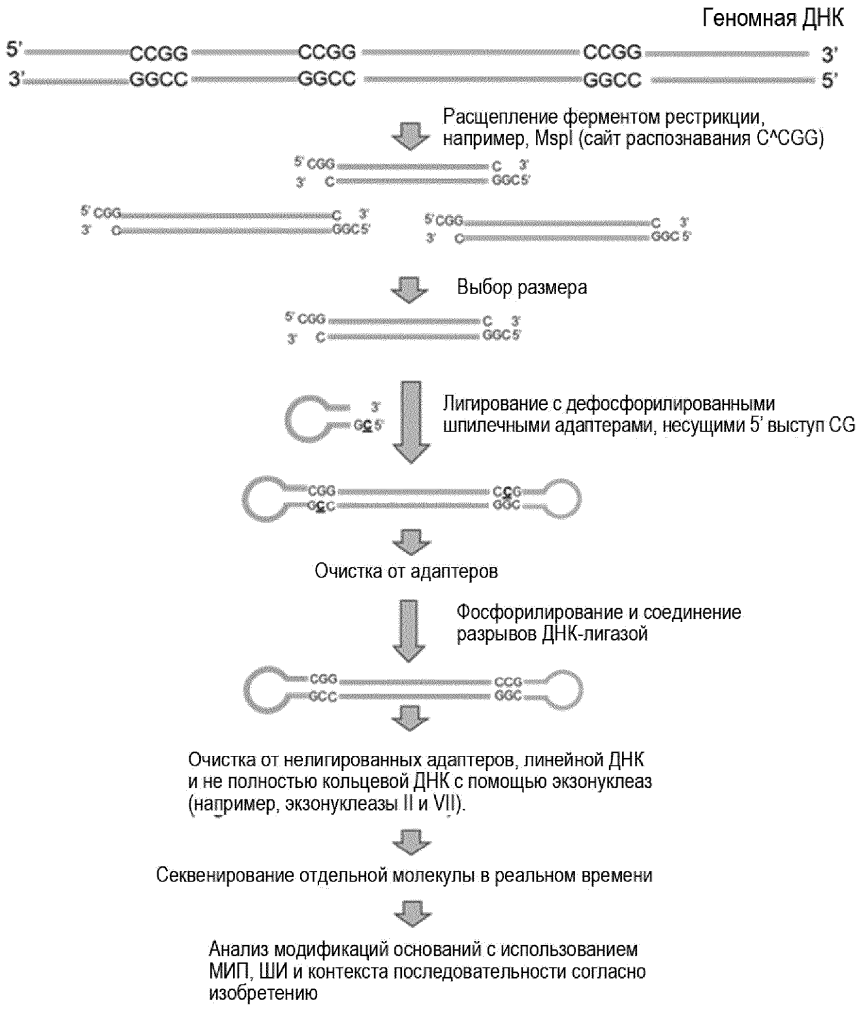
Фиг. 128



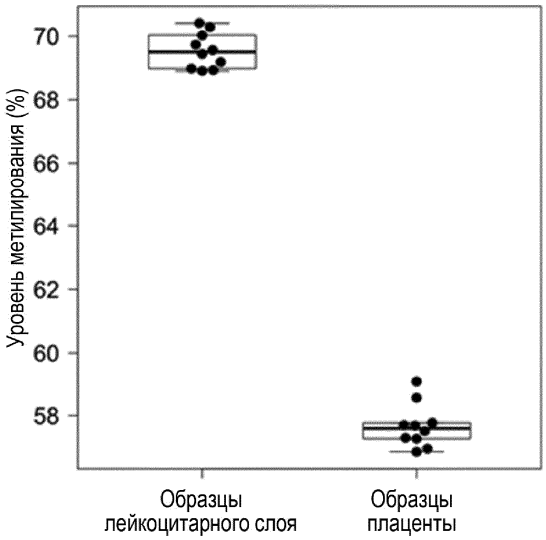
Фиг. 129



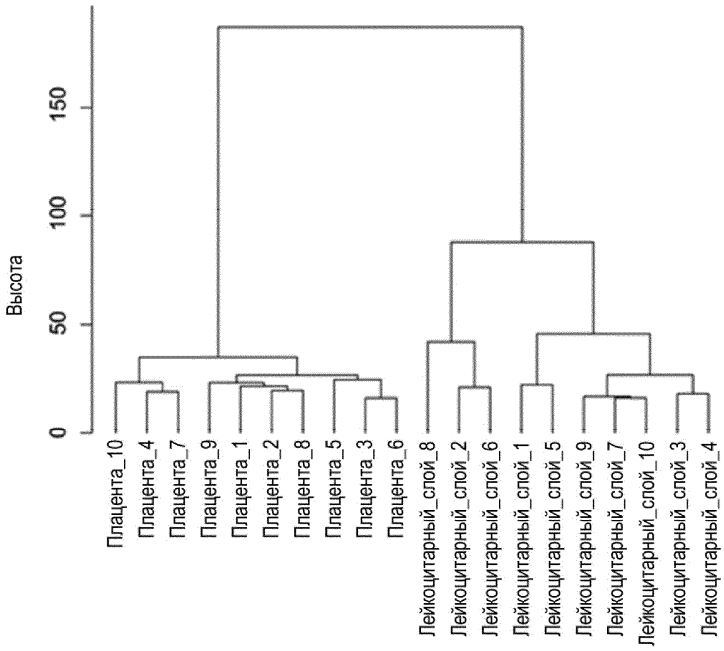
Фиг. 130



Фиг. 131



Фиг. 132


Высота
0
50
100
150
Плацента_10
Плацента_4
Плацента_7
Плацента_9
Плацента_1
Плацента_2
Плацента_8
Плацента_5
Плацента_3
Плацента_6
Лейкоцитарный_слой_8
Лейкоцитарный_слой_2
Лейкоцитарный_слой_6
Лейкоцитарный_слой_1
Лейкоцитарный_слой_5
Лейкоцитарный_слой_9
Лейкоцитарный_слой_7
Лейкоцитарный_слой_10
Лейкоцитарный_слой_3
Лейкоцитарный_слой_4

Фиг. 133