(19) Евразийское патентное ведомство

(21) 202391791 (13) A1

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ЕВРАЗИЙСКОЙ ЗАЯВКЕ

(43) Дата публикации заявки
2023.10.02

(22) Дата подачи заявки
2021.12.16

(51) Int. Cl. ***C12Q 1/6881*** (2018.01)
C12Q 1/6886 (2018.01)
G16B 30/00 (2019.01)

(54) СПОСОБЫ КЛАССИФИКАЦИИ ОБРАЗЦА ПО КЛИНИЧЕСКИ ЗНАЧИМЫМ КАТЕГОРИЯМ

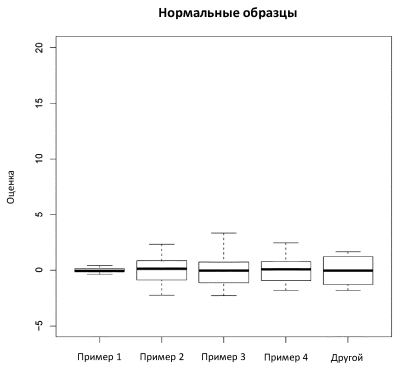
(31) **20215730.1**
(32) **2020.12.18**
(33) **EP**
(86) **PCT/EP2021/086243**
(87) **WO 2022/129360 2022.06.23**

(71) Заявитель:
МЕДИКАВЕР БИОТЕК ЛТД (CY)

(72) Изобретатель:
Коумбарис Джордж, Ахиллеос Ахиллеас, Элиадес Алексия, Лойзидес Хараламбос, Тсангарас Кириакос, Кипри Елена, Йоаннидес Мариос, Патсалис Филиппос (CY)

(74) Представитель:
Нилова М.И. (RU)

(57) В настоящем изобретении предложены способы и наборы для классификации биологических образцов по клинически значимым категориям. Способ включает этапы: (i) определения в образце, содержащем множество фрагментов внеклеточной ДНК (вкДНК), координат на последовательности для начала и/или конца по меньшей мере у 100000 фрагментов вкДНК путем выравнивания с референсной последовательностью; и (ii) определения в референсной последовательности всех мотивов нуклеиновой кислоты, состоящих из тринуклеотидов, тетрануклеотидов и пентануклеотидов. Указанные нуклеотиды: (a) находятся в диапазоне от 1 до 5 пар нуклеотидов по направлению внутрь, но рядом с координатой на последовательности для каждого начала и/или конца, определенной на этапе (i), и/или (b) находятся в диапазоне от 1 до 5 пар нуклеотидов по направлению наружу, но рядом с координатой на последовательности для каждого начала и/или конца, определенной на этапе (i). Затем на третьем этапе (iii) определяют частоту (a) каждой координаты на последовательности плюс и/или минус 1 пара нуклеотидов, определенной на этапе (i), во множестве фрагментов вкДНК, содержащихся в образце, и (b) каждого из мотивов нуклеиновой кислоты, определенных на этапе (ii) a) и b), во множестве фрагментов вкДНК, содержащихся в образце. Применяя эти частоты, вычисляют отношения каждой из частот, определенных на этапе (iii) a) и b), к соответствующей референсной частоте. Для каждого упомянутого выше отношения отдельно рассчитывают диагностическую оценку. На заключительном этапе вычисляют комбинированную диагностическую оценку по меньшей мере из двух или более определенных диагностических оценок. Диагностические оценки и комбинированные диагностические оценки позволяют классифицировать образец на клинически значимые категории, а комбинированная диагностическая оценка дополнительно позволяет классифицировать внеклеточную и/или циркулирующую опухолевую ДНК в образце как содержащуюся в низком, умеренном и высоком количестве.



202391791 A1

202391791 A1

СПОСОБЫ КЛАССИФИКАЦИИ ОБРАЗЦА ПО КЛИНИЧЕСКИ ЗНАЧИМЫМ КАТЕГОРИЯМ

ОБЛАСТЬ ТЕХНИКИ, К КОТОРОЙ ОТНОСИТСЯ ИЗОБРЕТЕНИЕ

Настоящее изобретение относится к области биологии, медицины и химии, в частности, к области молекулярной биологии и, более конкретно, к области молекулярной диагностики.

УРОВЕНЬ ТЕХНИКИ

Эукариотические геномы организованы в хроматин, который не только обеспечивает возможность компактизации ДНК, но также регулирует метаболизм ДНК (репликацию, транскрипцию, репарацию, рекомбинацию). Было показано, что сигнатуры структуры хроматина у эукариотических организмов, в частности, расположение нуклеосом, можно применять для идентификации редких фрагментов нуклеиновой кислоты в сложных смесях, которые присутствуют в эукариотических организмах (Heitzer E. *et al., Nat. Rev. Genet*., 2019, 20(2):71-88).

Предполагается, что защита ДНК нуклеосомой отвечает за наличие горячих точек неслучайной фрагментации (hot spots of non-random fragmentation — HSNRF), которые определяются как области в геноме, где, как было обнаружено, концы фрагментов нуклеиновых кислот со специфическим распределением по размеру встречаются с более высокой частотой, чем ожидалось, по сравнению с соседними локациями в геноме.

Рак часто обнаруживают в труднодоступных местах человеческого тела. Инвазивная хирургическая биопсия, являющаяся «золотым стандартом» для диагностики рака, сопряжена со значительными клиническими рисками, включая кровотечение и инфекцию. Одним из недостатков таких инвазивных процедур является тот факт, что образец, взятый из опухолевой ткани, является лишь пространственно ограниченным отображением по состоянию на момент проведения процедуры. Рак, однако, не остается статичным, а претерпевает постоянные изменения, которые приводят к генетической гетерогенности внутри опухоли и между первичным и метастатическими очагами рака. Были предприняты значительные усилия по разработке неинвазивных/минимально инвазивных методов диагностики, мониторинга и руководства терапией рака. Успешная технологическая разработка неинвазивного пренатального тестирования на многочисленные аномалии с

использованием внеклеточной ДНК из материнской плазмы также может быть применена для обнаружения биомаркеров с целью диагностики рака. Обнаружение циркулирующей опухолевой ДНК в плазме предоставило возможность использовать ее в качестве биомаркера и применять анализ методом жидкой биопсии для обнаружения рака, оценки прогноза и предсказания ответа на лечение рака без необходимости учитывать риски, связанные с инвазивными хирургическими процедурами. Эта технология приносит пользу онкологическим пациентам, позволяя обнаруживать рак на ранних стадиях и тем самым увеличивая возможность успешного выздоровления, помогая выбрать наиболее подходящую терапию, а также дает возможность выявлять минимальную остаточную болезнь после курса лечения, тем самым помогая клиницистам в проведении необходимых медицинских вмешательств. В отличие от современных инвазивных методов тестирования, которые имеют риск осложнений, жидкая биопсия по своей сути безопасна для пациента, поскольку в ней используются такие образцы, как кровь, моча или мокрота.

На сегодняшний день описано лишь ограниченное количество способов, с помощью которых пытаются дать оценку вклада опухоли в общее количество внеклеточной ДНК (вкДНК), обнаруженной в плазме, причем указанная внеклеточная опухолевая ДНК (вкоДНК) должна применяться в качестве прогностического биомаркера, индикатора ответа и/или резистентности к терапии и рецидива заболевания (Smith C.G. *et al., Genome Med.,* 2020, 12(1): 23; Peiyong Jiang *et al*., *PNAS,* 2018, 115(46): E10925-E10933; Cristiano S. *et al. Nature,* 2019, 570: 385-389; Mouliere et al., *Sci. Transl. Med*., 2018, 10(466): eaat4921; Newman A. *et al*., *Nat. Med*., 2014, 20(5): 548-554).

Современные тесты на основе жидкой биопсии не удовлетворяют потребностям прецизионной онкологии из-за их сложности, а также их ограниченной чувствительности и специфичности (De Rubis G. *et al*., *Trends Pharmacol Sci*., 2019, 40(3): 172-186; Peiyong Jiang *et al*., *Cancer Discov*., 2020, CD-19-0622). Таким образом, точность таких способов недостаточно высока и может привести к ошибочным результатам.

Настоящее изобретение направлено на преодоление ограничений, с которыми сталкиваются современные подходы к жидкой биопсии, путем расширения диапазона информации, извлекаемой из секвенирования циркулирующей опухолевой ДНК (цоДНК), и внедрения новых многопараметрических стратегий для создания надежного, чувствительного и специфичного анализа методом жидкой биопсии для классификации (распределения) образцов по клинически значимым категориям.

КРАТКОЕ ОПИСАНИЕ ИЗОБРЕТЕНИЯ

Настоящее изобретение направлено на решение проблемы ограничений точности, с которыми в настоящее время сталкиваются другие подходы к жидкой биопсии. Настоящее изобретение преодолевает указанные ограничения точности путем расширения диапазона информации, извлекаемой из секвенирования внеклеточной опухолевой ДНК или цоДНК, и внедрения новых многопараметрических стратегий для создания надежного, чувствительного и специфичного анализа методом жидкой биопсии для классификации образцов по клинически значимым категориям.

В одном из вариантов реализации настоящее изобретение относится к способу классификации образца как содержащего внеклеточную опухолевую ДНК, включающему следующие этапы:

(i) определение в образце, содержащем множество фрагментов внеклеточной ДНК (вкДНК), координат на последовательности для начала и/или конца по меньшей мере у 100 000 фрагментов вкДНК путем выравнивания с референсной последовательностью,

(ii) определение в референсной последовательности всех мотивов нуклеиновой кислоты, состоящих из тринуклеотидов, тетрануклеотидов и пентануклеотидов:

a) в диапазоне от 1 до 5 пар нуклеотидов по направлению внутрь, но рядом с координатой на последовательности для каждого начала и/или конца, определенной на этапе (i), и/или

b) в диапазоне от 1 до 5 пар нуклеотидов по направлению наружу, но рядом с координатой на последовательности для каждого начала и/или конца, определенной на этапе (i),

(iii) определение частоты:

a) каждой координаты на последовательности плюс и/или минус 1 пара нуклеотидов, определенной на этапе (i), во множестве фрагментов вкДНК, содержащихся в образце,

b) каждого из мотивов нуклеиновой кислоты, определенных на этапе (ii) a) и b), во множестве фрагментов вкДНК, содержащихся в образце,

(iv) вычисление отношения каждой из частот, определенных на этапе (iii) a) и b), к соответствующей референсной частоте,

(v) вычисление диагностической оценки отдельно для каждого отношения, определенного на этапе (iv), причем указанная оценка представляет собой соответствующую взвешенную сумму всех соответствующих отношений частот,

полученных на этапе (iv),

(vi) вычисление комбинированной диагностической оценки из по меньшей мере двух или более диагностических оценок, определенных на этапе (v), причем указанная оценка представляет собой взвешенную сумму указанных двух или более диагностических оценок, определенных на этапе (v), и

(vii) определение классификации образца путем сравнения комбинированной диагностической оценки с референсной оценкой,

при этом образец классифицируют как содержащий опухолевую вкДНК, если значение комбинированной диагностической оценки выше среднего значения референсной оценки по меньшей мере на одно стандартное отклонение референсной оценки, причем референсную оценку рассчитывают из одного или более референсных значений.

В одном варианте реализации комбинированную диагностическую оценку рассчитывают из всех диагностических оценок, рассчитанных для каждого отношения, рассчитанного на этапе (v) по способу, описанному выше.

В одном из вариантов реализации настоящее изобретение относится к способу классификации образца как содержащего внеклеточную опухолевую ДНК, включающему следующие этапы:

(i) определение в образце, содержащем множество фрагментов внеклеточной ДНК (вкДНК), координат на последовательности для начала и/или конца и для начала и/или конца плюс и/или минус 1 пара нуклеотидов по меньшей мере у 100 000 фрагментов вкДНК путем выравнивания с референсной последовательностью,

(ii) определение частоты каждой координаты, определенной на этапе (i), во множестве фрагментов вкДНК, содержащихся в образце,

(iii) вычисление отношения частоты каждой координаты, определенной на этапе (ii), к соответствующей референсной частоте,

(iv) вычисление диагностической оценки из всех отношений, определенных на этапе (iii), причем указанная оценка представляет собой взвешенную сумму всех отношений частоты, определенных на этапе (iii), и

(v) определение классификации образца путем сравнения диагностической оценки с референсной оценкой,

при этом образец классифицируют как содержащий опухолевую вкДНК, если значение

диагностической оценки выше среднего значения референсной оценки по меньшей мере на одно стандартное отклонение референсной оценки, причем референсную оценку рассчитывают из одного или более референсных значений.

В одном из вариантов реализации настоящее изобретение относится к способу классификации образца как содержащего внеклеточную опухолевую ДНК, включающему следующие этапы:

(i) определение в образце, содержащем множество фрагментов внеклеточной ДНК (вкДНК), координат на последовательности для начала и/или конца по меньшей мере у 100 000 фрагментов вкДНК путем выравнивания с референсной последовательностью,

(ii) определение в референсной последовательности всех мотивов нуклеиновой кислоты, состоящих из тринуклеотидов, тетрануклеотидов и пентануклеотидов, в диапазоне от 1 до 5 пар нуклеотидов по направлению внутрь, но рядом с координатой на последовательности для каждого начала и/или конца, определенной на этапе (i),

(iii) определение частоты каждого из мотивов нуклеиновой кислоты, определенных на этапе (ii), во множестве фрагментов вкДНК, содержащихся в образце,

(iv) вычисление отношения каждой из частот, определенных на этапе (iii), к соответствующей референсной частоте,

(v) вычисление диагностической оценки из всех отношений, определенных на этапе (iv), причем указанная оценка представляет собой взвешенную сумму всех отношений частоты, определенных на этапе (iv), и

(vi) определение классификации образца путем сравнения диагностической оценки с референсной оценкой,

при этом образец классифицируют как содержащий опухолевую вкДНК, если значение диагностической оценки выше среднего значения референсной оценки по меньшей мере на одно стандартное отклонение референсной оценки, причем референсную оценку рассчитывают из одного или более референсных значений.

В другом варианте реализации настоящее изобретение относится к способу классификации образца как содержащего внеклеточную опухолевую ДНК, включающему следующие этапы:

(i) определение в образце, содержащем множество фрагментов внеклеточной ДНК (вкДНК), координат на последовательности для начала и/или конца по меньшей мере у 100 000 фрагментов вкДНК путем выравнивания с референсной последовательностью,

(ii) определение в референсной последовательности всех мотивов нуклеиновой кислоты, состоящих из тринуклеотидов, тетрануклеотидов и пентануклеотидов, в диапазоне от 1 до 5 пар нуклеотидов по направлению наружу, но рядом с координатой на последовательности для каждого начала и/или конца, определенной на этапе (i),

(iii) определение частоты каждого из мотивов нуклеиновой кислоты, определенных на этапе (ii), во множестве фрагментов вкДНК, содержащихся в образце,

(iv) вычисление отношения каждой из частот, определенных на этапе (iii), к соответствующей референсной частоте,

(v) вычисление диагностической оценки из всех отношений, определенных на этапе (iv), причем указанная оценка представляет собой взвешенную сумму всех отношений частоты, определенных на этапе (iv), и

(vi) определение классификации образца путем сравнения диагностической оценки с референсной оценкой,

при этом образец классифицируют как содержащий опухолевую вкДНК, если значение диагностической оценки выше среднего значения референсной оценки по меньшей мере на одно стандартное отклонение референсной оценки, причем референсную оценку рассчитывают из одного или более референсных значений.

В одном из вариантов реализации диапазон пар нуклеотидов по направлению внутрь, но рядом с координатой на последовательности для каждого начала и/или конца, может составлять от 2 п. н. до 6 п. н., или от 3 п. н. до 7 п. н., или от 4 п. н. до 8 п. н., или от 5 п. н. до 9 п. н., или от 6 п. н. до 10 п. н. от координаты каждого начала и/или конца.

В одном из вариантов реализации минимальное количество фрагментов вкДНК, содержащихся в анализируемом образце, составляет от 100 тысяч до 500 тысяч, от 500 тысяч до 1 миллиона, от 1 миллиона до 2 миллионов, от 2 миллионов до 5 миллионов, или от 5 миллионов до 10 миллионов, или от 10 миллионов до 20 миллионов, или от 20 миллионов до 50 миллионов, или от 50 миллионов до 500 миллионов.

В одном из вариантов реализации количество опухолевой вкДНК в образце может быть классифицировано как низкое, если комбинированная диагностическая оценка составляет от 2 до 4 стандартных отклонений референсной оценки, как среднее, если комбинированная оценка составляет от 4 до 6,5 стандартных отклонений референсной оценки, и как высокое, если комбинированная оценка составляет более 6,5 стандартных отклонения референсной оценки.

В одном из вариантов реализации референсные образцы могут представлять собой образцы от пациентов без рака, или от пациентов без рецидивов, или от успешно вылеченных пациентов с раком.

В одном из вариантов реализации этап (i) любого из способов, описанных выше, заключающийся в определении в образце, содержащем множество фрагментов внеклеточной ДНК (вкДНК), координат на последовательности для начала и/или конца по меньшей мере у 100 000 фрагментов вкДНК путем выравнивания с референсной последовательностью, включает определение последовательности нуклеиновой кислоты по меньшей мере у части из множества фрагментов вкДНК в образце перед выравниванием с референсной последовательностью.

В одном из вариантов реализации этап (i) любого из способов, описанных выше, заключающийся в определении в образце, содержащем множество фрагментов внеклеточной ДНК (вкДНК), координат на последовательности для начала и/или конца по меньшей мере у 100 000 фрагментов вкДНК путем выравнивания с референсной последовательностью, дополнительно включает обогащение фрагментов вкДНК перед определением последовательности нуклеиновой кислоты у фрагментов вкДНК.

В одном из вариантов реализации образец классифицируют как содержащий опухолевую вкДНК, происходящую из опухоли, выбранной из группы, включающей рак крови, рак печени, рак легкого, рак поджелудочной железы, рак предстательной железы, рак молочной железы, рак желудка, глиобластому, колоректальный рак, рак головы и шеи, солидную опухоль, доброкачественную опухоль, злокачественную опухоль, рак на поздней стадии, метастаз или предраковую ткань.

В еще одном варианте реализации настоящее изобретение относится к набору, включающему:

(i) компоненты для осуществления любого из описанных выше способов, при этом компоненты содержат:

a) один или более компонентов для выделения внеклеточной ДНК из биологического образца,

b) один или более компонентов для приготовления и обогащения библиотеки для секвенирования, и/или

c) один или более компонентов для амплификации и/или секвенирования обогащенной

библиотеки,

(ii) программное обеспечение для проведения статистического анализа.

КРАТКОЕ ОПИСАНИЕ ФИГУР

Были проанализированы двадцать нормальных образцов от пациентов, не болеющих раком, и 27 не соответствующих норме образцов от пациентов с диагностированным немелкоклеточным раком легкого (НМРЛ) на поздней стадии или раком толстой кишки. Десять случайно выбранных нормальных образцов и десять случайно выбранных не соответствующих норме образцов применили на этапе обучения для оценки неизвестных параметров в Примерах 1–4.

Фигура 1. На Фигуре показано распределение оценок, полученных в Примерах 1–4 для «нормальных» образцов (контрольные образцы от здоровых, не болеющих раком индивидуумов, не включенные в этап обучения), по сравнению с оценками, полученными с помощью способа, описанного специалистами в данной области, который называется в этом документе «другим» способом (Peiyong Jiang *et al*., Cancer *Discov., 2020*, CD-19-0622). Указанный другой способ измерения количества концевых мотивов последовательности фрагментов вкДНК, содержащихся в анализируемых образцах, также учитывает и включает координаты для начала и/или конца указанных фрагментов, в отличие от настоящего раскрытия, которое исключает указанные координаты начала и/или конца. Соответствующий незначимому различию критерий суммы рангов Краскела — Уоллиса (*p-значение* = 0,9966) указывает на то, что ни один из способов не демонстрирует стохастического доминирования над другим подходом в случае нормальных образцов. Среднее значение расчетных оценок устанавливается для каждого примера равным нулю.

Фигура 2. На Фигуре проиллюстрированы значения оценок и их соответствующее распределение, полученные с помощью способа по настоящему изобретению в Примерах 1–4 и с применением описанного специалистами в данной области способа (далее называемого в этом документе «другим» способом) для образцов, содержащих внеклеточную опухолевую («не соответствующую норме ») ДНК (указанные образцы не включены в этап обучения). Когда эти оценки сравнивают с оценками, полученными для нормальных образцов (Фигура 1), самая высокая степень установления различий достигается с помощью способов по настоящему изобретению из Примеров 1–4, что явным образом иллюстрирует улучшение (увеличение) чувствительности настоящего способа (Примеры 1–4) по сравнению с

описанным специалистами в данной области способом при установлении различий между не соответствующими норме образцами и нормальными образцами.

Фигура 3. На Фигуре проиллюстрировано сравнение характеристик чувствительности между способами, описанными в Примерах 1–4, и описанным специалистами в данной области способом (далее называемым в этом документе «другим» способом). Из эмпирических распределений каждой из оценок нормальных и не соответствующих норме образцов вычислили расчетную чувствительность для всех способов в Примерах 1–4 и описанного специалистами в данной области («другого») способа. Специфичность для всех способов (т. е. уровень значимости при проверке статистической гипотезы) установлена на уровне 99,9%, при этом расчетные чувствительности для этого набора данных равны 96,8%, 99,94%, 99,48%, 99,9997% для способов из Примеров 1–4 соответственно. Все способы по настоящему изобретению значимо превосходят описанный специалистами в данной области способ, который обеспечивает чувствительность всего 84,3%, а также другие способы, доступные в настоящее время в литературе, в которых применяют информацию о размере фрагментов и изменении количества копий для классификации образцов на клинически информативные категории и достижения чувствительности в диапазоне от 60% до 90% (Mouliere et al. 2018 and Adalsteinsson et al. 2017) (данные не показаны).

Фигура 4. Таблица 1. В Таблице приведены оценки, полученные способом по настоящему изобретению в Примере 4, для четырех дополнительных нормальных образцов и трех дополнительных не соответствующих норме образцов, причем не соответствующие норме образцы получены от онкологических пациентов с диагнозом НМРЛ (I стадия). В Таблице выделена классификация количества цоДНК на низкую, среднюю и высокую. Количество цоДНК в образце классифицируется как низкое, если комбинированная диагностическая оценка составляет от 2 до 4,5, как среднее, если комбинированная диагностическая оценка составляет от 4,5 до 6, и как высокое, если комбинированная диагностическая оценка составляет более 6.

ПОДРОБНОЕ ОПИСАНИЕ ИЗОБРЕТЕНИЯ

В настоящем изобретении описан способ жидкой биопсии, в котором применяется новый биоинформационный анализ, основанный на расширенном диапазоне информации, извлекаемой из секвенирования цоДНК, и внедряются новые многопараметрические стратегии для создания надежного, чувствительного и специфичного анализа жидкой биопсии для классификации образцов по клинически значимым категориям.

Один из вариантов реализации настоящего изобретения относится к способу классификации образца как содержащего внеклеточную опухолевую ДНК, причем указанный способ включает определение координат на последовательности для концов, или «начала и/или конца», и необязательно для начала и/или конца плюс и/или минус 1 пара нуклеотидов, у множества фрагментов вкДНК, содержащихся в образце. «Начало и/или конец» фрагмента вкДНК в настоящей заявке относится к концам, границам или самым крайним парам нуклеотидов или нуклеотидам фрагмента вкДНК. Определение координат на последовательности для фрагментов вкДНК может быть осуществлено путем выравнивания с референсной последовательностью, причем референсная последовательность может представлять собой последовательность ДНК организма, предпочтительно последовательность ДНК человека, такую как последовательность генома человека hg19 или hg38, или последовательность генома субъекта-человека, который может быть, согласно одному из вариантов реализации, здоровым субъектом-человеком или субъектом-человеком, не имеющим рак.

В одном из вариантов реализации настоящего изобретения определение координат на последовательности может включать анализ и/или определение последовательности нуклеиновой кислоты у множества фрагментов вкДНК, например, путем анализа секвенирования. В одном из вариантов реализации определение координат на последовательности может дополнительно включать выделение или очистку нуклеиновых кислот и/или, в частности, фрагментов вкДНК из образца, и/или обогащение фрагментов вкДНК из образца, и/или получение библиотеки для секвенирования из выделенной ДНК, РНК или вкДНК перед анализом секвенирования.

Анализ данных секвенирования может включать выравнивание полученной информации о последовательности нуклеиновой кислоты вкДНК с референсной последовательностью генома. Это выравнивание позволяет картировать координаты на последовательности для «начала и/или конца» или концов анализируемых фрагментов вкДНК на референсной последовательности генома. В предпочтительном варианте реализации настоящего изобретения в дополнение к координатам начала и/или конца секвенированного фрагмента вкДНК на референсной последовательности генома также определяют координаты на последовательности для положений +1 п. н. и -1 п. н. от начала и/или конца.

Впоследствии может быть определена частота каждой определенной координаты на последовательности для начала и/или конца у множества фрагментов вкДНК, содержащихся в образце. Все координаты, обнаруженные для одного и того же фрагмента вкДНК

(технический дубликат) или для двух разных фрагментов вкДНК (биологические дубликаты), учитывают при вычислении частоты (распространенности) каждой координаты на последовательности для начала и/или конца, обнаруженной у множества фрагментов вкДНК. В предпочтительном варианте реализации настоящего изобретения в дополнение к частотам каждой координаты для начала и/или конца также определяют частоту каждой координаты на последовательности для положений +1 п. н. и -1 п. н. от начала и/или конца в пределах множества фрагментов вкДНК в образце.

В одном из вариантов реализации настоящего изобретения определяют отношение частоты каждой определенной координаты на референсном геноме к соответствующей референсной частоте. В предпочтительном варианте реализации это отношение частоты координат в образце к референсной частоте также вычисляют для каждой частоты координат на последовательности для положений +1 п. н. и -1 п. н. от начала и/или конца.

Затем из всех отношений частот может быть вычислена диагностическая оценка в соответствии со способом по настоящему изобретению, причем указанная диагностическая оценка определяется как взвешенная сумма всех отношений частот, полученных в соответствии с Примером 1, при этом анализируемый образец классифицируют как содержащий опухолевую вкДНК, если значение диагностической оценки выше среднего значения референсной оценки по меньшей мере на одно стандартное отклонение референсной оценки, причем референсную оценку рассчитывают из одного или более референсных значений.

В одном из вариантов реализации настоящего изобретения после определения координат начала и/или конца у множества фрагментов вкДНК, содержащихся в образце, могут быть определены все мотивы нуклеиновой кислоты в референсной последовательности, состоящие, например, из тринуклеотидов (трех последовательных нуклеотидов), тетрануклеотидов (четырех последовательных нуклеотидов) и/или пентануклеотидов (пяти последовательных нуклеотидов), в пределах определенного диапазона пар нуклеотидов по направлению внутрь, но расположенных смежно на расстоянии 1 или более п. н. от каждой координаты на последовательности для начала и/или конца. В одном из вариантов реализации настоящего изобретения определенный диапазон пар нуклеотидов по направлению внутрь, но расположенных смежно на расстоянии 1 или более п. н. от каждой координаты на последовательности для начала и/или конца, может составлять от 1 п. н. до 5 п. н., от 2 п. н. до 6 п. н., от 3 п. н. до 7 п. н., от 4 п. н. до 8 п. н., от 5 п. н. до 9 п. н. или от 6 п. н. до 10 п. н. В предпочтительном варианте реализации диапазон может составлять от

1 п. н. до 5 п. н. по направлению внутрь от каждой координаты на последовательности для начала и/или конца, определенной у множества фрагментов вкДНК в образце. Мотивы берут из референсной последовательности генома, чтобы избежать межиндивидуальной вариабельности (т. е. однонуклеотидных полиморфизмов).

Мотивы нуклеиновых кислот могут быть определены на основе каждого обнаруженного положения начала и/или конца на референсной последовательности, с которой выравнивали фрагмент вкДНК, а не на основе фактической последовательности фрагмента.

Затем может быть определена частота (распространенность) каждого обнаруженного мотива нуклеиновой кислоты во множестве фрагментов вкДНК в образце. Все мотивы, обнаруженные для одного и того же фрагмента вкДНК или для двух разных фрагментов вкДНК, учитывают при вычислении частоты (распространенности) каждого мотива, обнаруженного во множестве фрагментов вкДНК. После этого вычисляют отношение каждой из частот мотивов нуклеиновой кислоты во множестве фрагментов вкДНК к соответствующей референсной частоте. Затем из всех отношений частот вычисляют диагностическую оценку в соответствии со способом по настоящему изобретению, причем указанная диагностическая оценка определяется как взвешенная сумма всех отношений частот в соответствии с Примером 2, при этом анализируемый образец классифицируют как содержащий опухолевую вкДНК, если значение диагностической оценки выше среднего значения референсной оценки по меньшей мере на одно стандартное отклонение референсной оценки, причем референсную оценку рассчитывают из одного или более референсных значений.

В одном из вариантов реализации настоящего изобретения после определения координат начала и/или конца у множества фрагментов вкДНК, содержащихся в образце, могут быть определены все мотивы нуклеиновой кислоты в референсной последовательности, состоящие, например, из тринуклеотидов (трех последовательных нуклеотидов), тетрануклеотидов (четырех последовательных нуклеотидов) и/или пентануклеотидов (пяти последовательных нуклеотидов), в пределах определенного диапазона пар нуклеотидов по направлению наружу, но расположенных смежно на расстоянии 1 или более п. н. от каждой координаты на последовательности для начала и/или конца.

В одном из вариантов реализации настоящего изобретения определенный диапазон пар нуклеотидов по направлению наружу, но расположенных смежно на расстоянии 1 или более п. н. от каждой координаты на последовательности для начала и/или конца, может

составлять от 1 п. н. до 5 п. н., от 2 п. н. до 6 п. н., от 3 п. н. до 7 п. н., от 4 п. н. до 8 п. н., от 5 п. н. до 9 п. н. или от 6 п. н. до 10 п. н. В предпочтительном варианте реализации диапазон может составлять от 1 п. н. до 5 п. н. по направлению наружу от каждой координаты на последовательности для начала и/или конца, определенной у множества фрагментов вкДНК в образце. Мотивы нуклеиновых кислот могут быть определены на основе каждого обнаруженного положения начала и/или конца на референсной последовательности, с которой выравнивали фрагмент вкДНК. Такие мотивы нуклеиновой кислоты могут содержать только последовательность нуклеиновой кислоты референсной последовательности, расположенную смежно на расстоянии 1 или более п. н. от места выравнивания фрагмента вкДНК. Такие мотивы не содержат последовательность нуклеиновой кислоты фрагмента вкДНК, но содержат последовательность, начинающуюся непосредственно за пределами координаты начала или конца на референсной последовательности, например, начинающуюся с координаты на расстоянии от 1 п. н. до 5 п. н. по направлению наружу, но рядом с началом/концом.

Затем может быть определена частота каждого обнаруженного мотива нуклеиновой кислоты во множестве фрагментов вкДНК в образце. Все мотивы, обнаруженные для одного и того же фрагмента вкДНК или для двух разных фрагментов вкДНК, учитывают при вычислении частоты (распространенности) каждого мотива, обнаруженного во множестве фрагментов вкДНК. После этого может быть рассчитано отношение каждой из частот мотивов нуклеиновой кислоты во множестве фрагментов вкДНК к соответствующей референсной частоте. Наконец, из всех отношений частот может быть рассчитана диагностическая оценка в соответствии со способом по настоящему изобретению, причем указанная диагностическая оценка определяется как взвешенная сумма всех отношений частот в соответствии с Примером 3, при этом анализируемый образец классифицируют как содержащий опухолевую вкДНК, если значение диагностической оценки выше среднего значения референсной оценки по меньшей мере на одно стандартное отклонение референсной оценки, причем референсную оценку рассчитывают из одного или более референсных значений.

В одном варианте реализации настоящего изобретения все описанные в настоящей заявке выше этапы способа вычисления оценки из отношений: (a) частот координат на последовательности для начала и/или конца (необязательно положений -1 п. н. и/или +1 п. н.), (b) частот всех мотивов нуклеиновых кислот, расположенных по направлению внутрь, но смежно на расстоянии одной или более п. н. от координат начала и/или конца

фрагментов вкДНК, и (c) частот всех мотивов нуклеиновых кислот, расположенных по направлению наружу, но смежно на расстоянии 1 или более п. н. от координат начала и/или конца фрагментов вкДНК, без включения последовательности вкДНК, в сравнении с референсными частотами, могут проводиться параллельно или в определенном порядке, причем впоследствии значения диагностических оценок двух или всех этапов (a), (b) и (c) могут быть применены для вычисления значения комбинированной диагностической оценки в соответствии со способами по настоящему изобретению, как описано в Примере 4. В соответствии с этим значением комбинированной диагностической оценки анализируемый образец классифицируют как содержащий опухолевую вкДНК или циркулирующую опухолевую ДНК (цоДНК), если значение комбинированной диагностической оценки выше среднего значения референсной оценки по меньшей мере на одно стандартное отклонение референсной оценки, причем референсную оценку рассчитывают из одного или более референсных значений.

В одном из вариантов реализации путем сравнения значения комбинированной диагностической оценки, полученной для каждого не соответствующего норме образца, с референсной оценкой количество опухолевой вкДНК или цоДНК в образце может быть классифицировано как (a) низкое, если комбинированная диагностическая оценка составляет от 2 до 4 стандартных отклонений референсной оценки, как (b) среднее, если комбинированная оценка составляет от 4 до 6,5 стандартных отклонений референсной оценки, и как (c) высокое, если комбинированная оценка составляет более 6,5 стандартных отклонений референсной оценки (Таблица 1).

Внеклеточные нуклеиновые кислоты

В настоящей заявке смесь фрагментов нуклеиновой кислоты предпочтительно выделяют из образца, полученного от эукариотического организма, предпочтительно от примата, более предпочтительно от человека. Образец может содержать клетки или нуклеиновые кислоты из разных типов тканей. По существу, образец в силу своей природы может содержать смесь фрагментов нуклеиновой кислоты.

В настоящей заявке термины «нуклеиновая кислота» или «последовательность нуклеиновой кислоты» могут быть использованы как взаимозаменяемые для обозначения, без ограничений, ДНК, РНК, геномной ДНК, внеклеточной ДНК и/или РНК и тРНК, матричной РНК (мРНК), синтетической ДНК или РНК.

В контексте настоящего изобретения термины «фрагменты нуклеиновой кислоты» и «фрагментированные нуклеиновые кислоты» могут использоваться как взаимозаменяемые. В предпочтительном варианте реализации способа по настоящему изобретению фрагменты нуклеиновой кислоты представляют собой циркулирующую внеклеточную ДНК или РНК.

В одном из вариантов реализации настоящего изобретения может быть проанализировано не менее 100 000 фрагментов вкДНК, содержащихся в образце. В другом варианте реализации количество фрагментов вкДНК, содержащихся в анализируемом образце, может находиться в диапазоне от 100 тысяч до 500 тысяч, от 500 тысяч до 1 миллиона, от 1 миллиона до 2 миллионов, от 2 миллионов до 5 миллионов, от 5 миллионов до 10 миллионов, от 10 миллионов до 20 миллионов, от 20 миллионов до 50 миллионов или от 50 миллионов до 500 миллионов.

В одном варианте реализации изобретения «образец» представляет собой образец крови, образец сыворотки, образец плазмы, жидкий биопсийный образец или образец ДНК (например, смесь фрагментов нуклеиновой кислоты), содержащий внеклеточную ДНК (вкДНК), внеклеточную опухолевую ДНК (вкоДНК), циркулирующую опухолевую ДНК (цоДНК) или циркулирующую вкоДНК. В контексте настоящего изобретения термины «вкДНК», «вкоДНК», «цоДНК» или «циркулирующая вкоДНК» могут использоваться как взаимозаменяемые.

В одном из вариантов реализации образец выбран из группы, состоящей из образца плазмы, образца крови, образца мочи, образца мокроты, образца спинномозговой жидкости, образца асцитической жидкости и образца плевральной жидкости от субъекта, имеющего опухоль, или субъекта с подозрением на опухоль. В одном из вариантов реализации образец или образец ДНК происходит из образца ткани субъекта, имеющего опухоль, или субъекта с подозрением на опухоль, либо из набора злокачественных клеток.

В контексте настоящего изобретения термины «опухоль», «рак» или «не соответствующий норме/патологический» могут использоваться как взаимозаменяемые. В настоящей заявке термины «рак» или «опухоль» также могут включать раннюю стадию рака или позднюю стадию рака, метастазы или предраковые ткани или клетки. В настоящей заявке образец опухоли или не соответствующий норме образец могут относиться к образцу, содержащему (внеклеточную) ДНК или РНК, происходящую из первичной опухоли или метастатической опухоли. В настоящей заявке нормальный образец или референсный образец может относиться к образцу, содержащему только (внеклеточную) ДНК или РНК, происходящую из

нераковой (-ых), здоровой (-ых) или «нормальной (-ых)» ткани (-ей) или клетки (-ок). В контексте настоящего изобретения термины «нормальный», «контрольный» или «референсный» могут использоваться как взаимозаменяемые.

Способы по настоящему изобретению можно применять с разнообразными биологическими образцами. По существу, любой биологический образец, содержащий генетический материал, например, РНК или ДНК, и в частности внеклеточную ДНК (вкДНК) или внеклеточную РНК, можно применять в качестве образца в способах, обеспечивающих возможность генетического анализа РНК или ДНК в нем. Например, в одном варианте реализации образец ДНК представляет собой образец плазмы или образец крови, содержащий внеклеточную ДНК (вкДНК).

Согласно еще одному варианту реализации, предназначенному для онкологических целей, образец представляет собой биологический образец, полученный от субъекта, имеющего опухоль или рак, или субъекта с подозрением на опухоль или рак. В одном из вариантов реализации образец содержит циркулирующую внеклеточную опухолевую ДНК (вкоДНК). В другом варианте реализации образец представляет собой мочу, мокроту, асцитическую, спинномозговую жидкость или плевральный выпот субъекта. В другом варианте реализации онкологический образец представляет собой образец плазмы субъекта, полученный из периферической крови субъекта. Таким образом, образец может представлять собой жидкий биопсийный образец, который получают неинвазивным способом из образца крови субъекта, тем самым обеспечивая потенциальную возможность раннего выявления рака до развития поддающейся выявлению или пальпируемой опухоли, или возможность мониторинга прогрессирования заболевания, лечения заболевания или рецидива заболевания.

В настоящей заявке внеклеточная ДНК (вкДНК) относится к ДНК, которая не содержится в клетке. Образец может содержать вкДНК из нормальных или здоровых клеток и/или из раковых клеток. Внеклеточная ДНК может высвобождаться в кровь или сыворотку посредством секреции, в результате апоптоза или некроза. Если вкДНК высвобождается из опухоли или раковой клетки, ее можно назвать внеклеточной опухолевой ДНК (вкоДНК).

Термин «субъект» в контексте настоящего изобретения относится к животным, предпочтительно млекопитающим и еще более предпочтительно к людям или пациентам-людям. При использовании в настоящей заявке термин «субъект» может относиться к субъекту, страдающему от опухоли, или субъекту, у которого подозревают наличие опухоли.

Термин «опухоль» в настоящей заявке относится к раку в целом, включая, без ограничений, солидную опухоль, аденому, рак крови, рак печени, рак легкого, рак поджелудочной железы, рак предстательной железы, рак молочной железы, рак желудка, глиобластому, колоректальный рак, рак головы и шеи, опухоль на поздней стадии рака, доброкачественную или злокачественную опухоль, метастаз или предраковую ткань.

В настоящей заявке «концы» фрагментов вкДНК определяют как самые крайние нуклеотиды на 3'- и 5'-концах фрагмента нуклеиновой кислоты, и они могут также называться в настоящей заявке «началом и/или концом (положениями начала и/или конца)» или «точками разрыва» или «границами» фрагмента вкДНК. При выравнивании с референсной последовательностью «координаты (начала и/или конца)» или «координаты на последовательности» фрагмента вкДНК определяются самыми крайними положениями последовательности нуклеиновой кислоты, с которыми выравниваются концы фрагментов вкДНК в референсной последовательности. Например, если фрагмент вкДНК комплементарен или выравнивается с референсной последовательностью нуклеиновой кислоты, охватывающей положение на последовательности от 1500 п. н. до 1700 п. н., координаты последовательности будут составлять 1500 и 1700 п. н., определяя длину фрагмента вкДНК 200 п. н.

Профиль размеров вкДНК, демонстрирующий основной пик в области 166 п. н. и меньшие пики с интервалами 10 п. н., позволяет предположить, что биология вкДНК может быть связана с нуклеосомной организацией. Аналогичные закономерности также наблюдали в ДНК в плазме крови у пациентов с раком. Неслучайный характер фрагментации вкДНК, связанный с тканями происхождения, также может быть связан с состоянием здоровья пациента. Следовательно, координаты концов или начала и/или конца и частота внеклеточных фрагментов ДНК указывают на прогрессирование заболевания. Они варьируют в зависимости от происхождения опухоли и массы опухоли, что отражает выраженность заболевания и, следовательно, его ответ на данную терапию.

При использовании в настоящей заявке термин «по направлению внутрь» от координаты начала и/или конца относится к направлению от координаты «начала и/или конца» фрагмента нуклеиновой кислоты на референсной последовательности, в котором продолжается последовательность или мотив. Термин «по направлению внутрь» может относиться к последовательности нуклеиновой кислоты или мотиву, содержащимся в последовательности фрагмента нуклеиновой кислоты, или референсной последовательности, с которой осуществляют выравнивание. Термин «по направлению внутрь» может означать +1, +2, +3, +4, +5 и т. д. пар нуклеотидов от координаты начала и/или

-1, -2, -3, -4, -5 пар нуклеотидов от координаты конца фрагмента нуклеиновой кислоты. В одном из вариантов реализации диапазон пар нуклеотидов по направлению внутрь, но рядом с координатой на последовательности для каждого начала и/или конца, может составлять от 1 п. н. до 5 п. н., или от 2 п. н. до 6 п. н., или от 3 п. н. до 7 п. н., или от 4 п. н. до 8 п. н., или от 5 п. н. до 9 п. н., или от 6 п. н. до 10 п. н. от координаты каждого начала и/или конца.

При использовании в настоящей заявке термин «по направлению наружу» от координаты начала и/или конца относится к направлению от координаты «начала и/или конца» фрагмента нуклеиновой кислоты на референсной последовательности, в котором продолжается последовательность. Термин «по направлению наружу» может относиться к последовательности нуклеиновой кислоты или мотиву, не содержащимся в последовательности фрагмента нуклеиновой кислоты, или референсной последовательности, с которой его выравнивают. Термин «по направлению наружу» может означать +1, +2, +3, +4, +5 и т. д. пар нуклеотидов от координаты конца и/или -1, -2, -3, -4, -5 пар нуклеотидов от координаты начала фрагмента нуклеиновой кислоты. В одном из вариантов реализации диапазон пар нуклеотидов по направлению наружу, но рядом с координатой на последовательности для каждого начала и/или конца, может составлять от 1 п. н. до 5 п. н., или от 2 п. н. до 6 п. н., или от 3 п. н. до 7 п. н., или от 4 п. н. до 8 п. н., или от 5 п. н. до 9 п. н., или от 6 п. н. до 10 п. н. от координаты каждого начала и/или конца.

В настоящем способе анализируются частоты и/или мотивы последовательности координат начала и/или конца плюс и минус 1 п. н., поскольку наблюдаемые конечные сайты фрагментов не обязательно могут быть истинными сайтами разрезания/расщепления (Peiyong Jiang *et al.*, *Genome Res.*, 2020, doi: 10.1101/gr.261396.120). Следовательно, принимая во внимание вероятность того, что близлежащие геномные основания являются истинным сайтом расщепления, настоящее изобретение приводит к повышению точности по сравнению с современным уровнем техники при классификации биологических образцов по клинически значимым категориям.

В настоящей заявке термины «мотив нуклеиновой кислоты», «мотив последовательности» или «мотив» относятся к отрезку последовательных нуклеотидов в последовательности нуклеиновой кислоты, состоящему из 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 100 и т. д. последовательных нуклеотидов. Этот отрезок последовательных нуклеотидов также может быть назван «тринуклеотидами», «тетрануклеотидами», «пентануклеотидами», «гексануклеотидами» и т. д. Указанные мотивы представляют собой подмножество

местоположений в геноме человека, предпочтительно расщепляемых, например, специфическими нуклеазами, когда генерируются и высвобождаются в плазму крови внеклеточные и/или циркулирующие молекулы ДНК. Такие концевые мотивы ДНК в плазме, образующиеся под действием нуклеаз, которые расщепляют нуклеиновые кислоты, такие как ДНК, во время апоптоза, демонстрируют отчетливую сигнатуру, которая может содержать HSNRF или быть специфичной для HSNRF. В предпочтительном варианте реализации «мотив» относится к отрезку из 3, 4 или 5 последовательных нуклеотидов из референсной последовательности генома.

В одном из вариантов реализации мотив нуклеиновой кислоты может быть расположен на конце или в точке разрыва фрагмента вкДНК, при этом указанный мотив может содержаться в последовательности нуклеиновой кислоты фрагмента вкДНК или находиться за пределами границ последовательности фрагмента вкДНК и в пределах референсной последовательности нуклеиновой кислоты, например, рядом с местом выравнивания фрагмента вкДНК.

Анализ вкДНК

В настоящей заявке «референсная последовательность» может представлять собой любую последовательность нуклеиновой кислоты, геномную последовательность, геномную последовательность организма или субъекта, предпочтительно последовательность генома человека (например, hg19 или hg38) или здорового индивидуума или субъекта.

В настоящей заявке «референсная частота» для частоты координат на последовательности для начала и/или конца может представлять собой частоту соответствующей координаты на последовательности для начала и/или конца в одном или более референсных геномах, референсных последовательностях или в одном или более геномах или последовательностях одного или более здоровых или «нормальных» контрольных образцов, субъектов или пациентов. В настоящей заявке «референсная частота» для мотива нуклеиновой кислоты может представлять собой частоту соответствующего мотива нуклеиновой кислоты в одном или более референсных геномах, референсных последовательностях или в одном или более геномах или последовательностях одного или более здоровых или «нормальных» контрольных образцов, субъектов или пациентов.

В настоящей заявке термин «частота» может использоваться как взаимозаменяемый с терминами «распространенность» и «встречаемость». В одном из вариантов реализации изобретения «частота» описывает распространенность и встречаемость или количество,

например, мотивов последовательности нуклеиновой кислоты, фрагментов нуклеиновой кислоты (вкДНК) или координат на последовательности для начала и/или конца, которые были обнаружены или подсчитаны во множестве нуклеиновых кислот или фрагментов вкДНК, содержащихся в образце.

В настоящей заявке термин «отношение» может относиться к математическому соотношению или пропорции частоты, например, мотива последовательности нуклеиновой кислоты, обнаруженного во множестве фрагментов нуклеиновой кислоты в образце, и частоты того же мотива последовательности нуклеиновой кислоты в референсном образце. В настоящей заявке отношение может быть рассчитано путем деления частоты каждой координаты или мотива на соответствующую референсную частоту соответствующей координаты или мотива.

Для приготовления образца нуклеиновые кислоты, такие как ДНК и/или РНК, выделяют из образца с применением стандартных способов, известных в данной области техники, неограничивающими примерами которых являются протокол QIAsymphony (QIAGEN), протокол для циркулирующей нуклеиновой кислоты QIAamp Circulating Nucleic Acid (QIAGEN), протокол KingFisher (Thermofisher), набор для внеклеточной ДНК MagMAX™ Cell-Free DNA (Thermofisher) или любой другой ручной или автоматизированный способ выделения, подходящий для выделения внеклеточной ДНК.

После выделения внеклеточную ДНК образца можно применять для получения библиотеки для секвенирования, чтобы сделать образец совместимым с применяемым далее способом секвенирования, таким как секвенирование следующего поколения (Next Generation Sequencing — NGS). Обычно это включает лигирование адаптеров с концами фрагментов внеклеточной ДНК. Наборы для приготовления библиотеки для секвенирования коммерчески доступны или могут быть разработаны.

Целевое обогащение вкДНК осуществляют с применением последовательностей захвата мишени (ПЗМ), которые связываются с представляющими интерес областями в геноме человека, и при этом: длина каждой последовательности в пуле составляет 125–260 пар нуклеотидов и/или 125–300 п. н., и/или 125–350 п. н., причем каждая последовательность имеет 5'-конец и 3' -конец; каждая последовательность в пуле связывается с представляющей интерес областью на расстоянии по меньшей мере 10 пар нуклеотидов как с 5'-конца, так и с 3' -конца от областей, несущих вариации числа копий, сегментные дупликации или

повторяющиеся элементы ДНК; и содержание GC в ПЗМ составляет от 20% до 50%, и/или от 20% до 60%, и/или от 20% до 70%, и/или от 20% до 80%.

В настоящей заявке термин «последовательности захвата мишени» или «ПЗМ» относится к последовательностям ДНК, которые комплементарны представляющей (-им) интерес области (-ям) на представляющей (-их) интерес геномной (-ых) последовательности (-ях) и которые применяются в качестве «приманки» для захвата и обогащения представляющей интерес области из большой библиотеки последовательностей, такой как библиотека полного генома для секвенирования, полученная из биологического образца. В контексте настоящего изобретения термины «последовательности захвата мишени», или «ПЗМ», или «зонды» могут использоваться как взаимозаменяемые.

В другом варианте реализации пул ПЗМ связывается со множеством представляющих интерес последовательностей опухолевых биомаркеров, выбранных из группы, включающей, без ограничений, AKT1, ALK, APC, AR, ARAF, ATM, BAP1, BARD1, BMPR1A, BRAF, BRCA1, BRCA2, BRIP1, CDH1, CDK4, CDKN2A (pl4ARF), CDKN2A (pl6INK4a), CHEK2, CTNNB1, DDB2, DDR2, DICERl, EGFR, EPCAM, ERBB2, ERBB3, ERBB4, ERCC1, ERCC2, ERCC3, ERCC4, ERCC5, ESR1, FANCA, FANCB, FANCC, FANCD2, FANCE, FANCF, FANCG, FANCI, FANCL, FANCM, FBXW7, FGFR1, FGFR2, FLT3, FOXA1, FOXL2, GATA3, GNA11, GNAQ, GNAS, GREM1, HOXB13, IDH1, IDH2, JAK2, KEAP1, KIT, KRAS, MAP2K1, MAP3K1, MEN1, MET, MLH1, MPL, MRE11A, MSH2, MSH6, MTOR, MUTYH, MYC, MYCN, NBN, NPM1, NRAS, NTRK1, PALB2, PDGFRA, PIK3CA, PIK3CB, PMS2, POLD1, POLE, POLH, PTEN, RAD50, RAD51C, RAD51D, RAF1, RBI, RET, ROS1, RUNX1, SDHA, SDHAF2, SDHB, SDHC, SDHD, SLX4, SMAD4, SMARCA4, SPOP, STAT, STK11, TMPRSS2, TP53, VHL, XPA, XPC и их комбинации. В одном из вариантов реализации пул ПЗМ связывается со множеством представляющих интерес последовательностей опухолевых биомаркеров, выбранных из группы, содержащей EGFR_6240, KRAS_521, EGFR_6225, NRAS_578, NRAS_580, PIK3CA_763, EGFR_13553, EGFR_18430, BRAF_476, KIT_1314, NRAS_584, EGFR_12378 и их комбинации.

В другом варианте реализации пул ПЗМ связывается со множеством представляющих интерес последовательностей опухолевых биомаркеров, выбранных из группы, содержащей, без ограничений, COSM6240 (EGFR_6240), COSM521 (KRAS_521), COSM6225 (EGFR_6225), COSM578 (NRAS_578), COSM580 (NRAS_580), COSM763 (PIK3CA_763), COSM13553 (EGFR_13553), COSM18430 (EGFR_18430), COSM476 (BRAF_476), COSM1314 (KIT_1314), COSM584 (NRAS_584), COSM12378 (EGFR_12378) и их комбинации, при этом идентификаторы относятся к идентификационному номеру биомаркера в базе данных COSMIC. В целом, этап

гибридизации с зондом или обогащения может быть осуществлен перед созданием библиотеки для секвенирования или после того, как библиотека была создана.

В одном из вариантов реализации настоящего изобретения библиотека для секвенирования может быть обогащена представляющими интерес областями последовательности путем гибридизации библиотеки с одним или более зондами, охватывающими, например, горячие точки неслучайной фрагментации (HSNRF). Такие области HSNRF представляют собой области, с высокой вероятностью содержащие на небольшом отрезке многочисленные вариации последовательности нуклеиновой кислоты, облегчающие идентификацию типов тканей различного происхождения (например, раковых и нормальных), которые присутствуют в смеси вкДНК.

Представляющую (-ие) интерес область (-и) на представляющей (-их) интерес хромосоме (-ах), где располагаются HSNRF, обогащают путем гибридизации пула зондов для захвата HSNRF с библиотекой для секвенирования с последующим выделением тех последовательностей в пределах библиотеки для секвенирования, которые связываются с зондами. В одном из вариантов реализации зонд охватывает сайт HSNRF таким образом, чтобы только 5'-конец фрагментированных внеклеточных нуклеиновых кислот был захвачен зондом. В другом варианте реализации зонд охватывает сайт HSNRF таким образом, чтобы только 3'-конец фрагментированных внеклеточных нуклеиновых кислот, возникающих из HSNRF, мог связываться с зондом. В другом предпочтительном варианте реализации зонд охватывает оба сайта HSNRF, ассоциированных с фрагментированной нуклеиновой кислотой, таким образом, чтобы как 5'-конец, так и 3'-конец внеклеточной нуклеиновой кислоты, ассоциированной с данным сайтом HSNRF, были захвачены зондом.

Чтобы облегчить выделение желаемых обогащенных последовательностей (HSNRF), обычно последовательности зондов модифицируют таким образом, чтобы последовательности, которые гибридизуются с зондами, можно было отделить от последовательностей, которые не гибридизуются с зондами. Обычно этого достигают путем фиксирования зондов на подложке. Это обеспечивает возможность физического отделения тех последовательностей, которые связываются с зондами, от тех последовательностей, которые не связываются с зондами. Например, каждую последовательность в пределах пула зондов можно пометить при помощи биотина, а затем связать пул с шариками, покрытыми субстанцией, связывающей биотин, такой как стрептавидин или авидин. В предпочтительном варианте реализации зонды метят биотином и связывают с магнитными шариками, покрытыми стрептавидином, тем самым обеспечивая возможность разделения за счет использования

магнитных свойств шариков. При этом, однако, обычному специалисту в данной области техники будет понятно, что в данной области техники известны и другие системы аффинного связывания, и их можно применять вместо системы «биотин — стрептавидин/авидин». Например, можно применять систему на основе антител, в которой зонды метят антигеном и затем связывают с шариками, покрытыми антителами. Более того, в зонды можно внедрить на одном конце последовательность метки и связывать их с подложкой через комплементарную последовательность на подложке, которая гибридизуется с последовательностью метки. Помимо этого, наряду с магнитными шариками можно применять другие типы подложек, такие как полимерные шарики, стекло и тому подобное.

В определенных вариантах реализации члены библиотеки для секвенирования, которые связываются с пулом зондов, являются полностью комплементарными зонду. В других вариантах реализации члены библиотеки для секвенирования, которые связываются с пулом зондов, являются частично комплементарными зонду. Например, при определенных обстоятельствах может быть желательно применять и анализировать данные, полученные от фрагментов ДНК, которые являются продуктами процесса обогащения, но не обязательно принадлежат к представляющим интерес геномным областям (т. е. такие фрагменты ДНК могут связываться с зондом по причине частичных гомологий) и при секвенировании производили бы очень малое покрытие всего генома по не связанным с зондами координатам.

После обогащения представляющей (-их) интерес последовательности (-ей) с применением зондов, с формированием вследствие этого обогащенной библиотеки ДНК с сайтами HSNRF, членов обогащенной библиотеки HSNRF подвергают элюированию, амплификации и секвенированию с применением стандартных способов, известных в данной области техники. В другом варианте реализации зонды предлагаются вместе с подложкой, такие как, например, биотинилированные зонды, предлагаемые вместе с магнитными шариками, покрытыми стрептавидином.

Зонды для выявления опухолевых биомаркеров сконструированы на основе критериев конструирования, описанных в настоящей заявке, а также известных последовательностей генов опухолевых биомаркеров и генетических мутаций в них, ассоциированных с раком. В одном из вариантов реализации множество зондов, применяемых в данном способе, связывается со множеством последовательностей представляющих интерес опухолевых биомаркеров. В данной ситуации зонд может располагаться в горячих точках неслучайной фрагментации, примыкающих к сайту мутации.

В настоящей заявке для анализа последовательности нуклеиновой кислоты можно применять секвенирование следующего поколения (NGS), хотя также можно задействовать и другие технологии секвенирования, которые обеспечивают очень точный подсчет наряду с информацией о последовательности. Таким образом, вместо NGS можно также применять и другие способы точного подсчета, такие как, без ограничений, цифровая ПЦР, одномолекулярное секвенирование, нанопоровое секвенирование, секвенирование наношариков ДНК, секвенирование путем лигирования, ионное полупроводниковое секвенирование, секвенирование путем синтеза и анализ на микрочипах.

В одном из вариантов реализации настоящее изобретение относится к способу, в котором фрагменты нуклеиновой кислоты, подлежащие выявлению, или происхождение которых предстоит определить, присутствуют в смеси в более низкой концентрации, чем фрагмент нуклеиновой кислоты из того же генетического локуса, но другого происхождения.

Настоящий способ в частности подходит для того, чтобы анализировать такие низкие концентрации целевой вкДНК. В способе по настоящему изобретению фрагмент нуклеиновой кислоты, подлежащий выявлению, или происхождение которого предстоит определить, и фрагмент нуклеиновой кислоты из того же генетического локуса, но другого происхождения, присутствуют в смеси в соотношении, которое выбрано из группы 1 : 2, 1 : 4, 1 : 10, 1 : 20, 1 : 50, 1 : 100, 1 : 200, 1 : 500, 1 : 1000, 1 : 2000 и 1 : 5000. Данные соотношения следует понимать как приблизительные соотношения, что означает плюс/минус 30%, 20% или 10%. Специалисту в данной области техники известно, что такие соотношения не будут возникать в точных численных значениях, приведенных выше. Соотношения означают отношение количества локус-специфических молекул редкого типа к количеству локус-специфических молекул часто встречающегося типа.

Анализ данных.

Информация, полученная в результате секвенирования обогащенной библиотеки, анализируется с применением инновационной цепочки процессов анализа биоматематических/биостатистических данных. В настоящем способе применяют признаки фрагментов вкДНК, включая комбинацию всех возможных мотивов, расположенных рядом на расстоянии 1 или более п. н. от координат концов, с применением референсной последовательности генома и исключая наблюдаемые сайты концов вкДНК, поскольку они могут не представлять истинные сайты расщепления. Кроме того, путем объединения анализа различных признаков вкДНК, включая местоположения и мотивы, настоящее

изобретение обеспечивает неожиданный технический эффект повышенной точности, т. е. повышенную чувствительность при тех же уровнях специфичности.

В соответствии с предпочтительным вариантом реализации настоящего изобретения проводят целевое секвенирование спаренных концов следующего поколения. Мультиплексированные данные для всех образцов демультиплексируют с применением инструмента компании Illumina bcltofastq. Данные секвенирования указанного образца обрабатывают для удаления адаптерных последовательностей и некачественных чтений (Q-показатель <25) с применением программного обеспечения cutadapt (Martin, M. et al. 2011 EMB.netJournal 17.1).

Обработанные чтения, длина которых составляла по меньшей мере 25 оснований, выравнивали по референсной сборке генома человека GRCh37 (hg19) (UCSC Genome Bioinformatics) с применением алгоритма выравнивания Burrows-Wheel (Li, H. and Durbin, R. (2009) Bioinformatics 25:1754-1760). Удаляли парные чтения с размером вставки, превышающим пороговое значение, причем указанное пороговое значение находилось в диапазоне 100–600. Если применимо, идентифицируют повторяющиеся чтения, группируют их по семейству уникальных молекулярных идентификаторов (Unique Molecular Identifier — UMI) и применяют для получения консенсусных чтений для каждого семейства UMI, после выравнивания.

Если применимо, выходные данные секвенирования, относящиеся к одному и тому же образцу, но обработанные на разных полосах секвенирования, объединяли в один файл выходных данных секвенирования. Применение дубликатов и процедуры объединения были выполнены с использованием программного пакета fgbio, picard (Broad Institute) и программного пакета Sambamba (ссылка на Sambamba: Tarasov, Artem, et al. Sambamba: fast processing of NGS alignment formats. Bioinformatics 31.12 (2015): 2032-2034). Информация с точки зрения позиций картирования (самые внешние и близлежащие координаты), глубины чтений на одно основание в интересующих локусах и размера фрагментов была получена с применением опции mpileup программного пакета SAMtools, далее называемого файлом mpileup, и обработана с применением пользовательских программных интерфейсов приложения (API), написанных на языках программирования Python и R (Python Software Foundation (2015) Python; The R Foundation (2015) The R Project for Statistical Computing).

Координата конца фрагмента определяется как самая крайняя координата в референсном геноме, которая охватывается фрагментом, т. е. каждый выровненный фрагмент имеет две

координаты концов (координата начального/крайнего левого положения (5'-конец) и конечного/крайнего правого положения (3'-конец) относительно референсного генома).

В различных вариантах реализации настоящего изобретения целевая панель состояла как минимум из 500 целевых геномных оснований. Минимальное количество фрагментов, необходимых на один образец, составляет 100 000.

В настоящей заявке «значение диагностической оценки» рассчитывают как взвешенную сумму всех отношений частот, как описано в Примерах 1, 2 и 3 в разделе «Примеры».

В настоящей заявке «значение комбинированной диагностической оценки» вычисляют как взвешенную сумму по меньшей мере двух или более отношений частот на всех этапах, описанных в настоящем изобретении, как описано в Примере 4.

В одном из вариантов реализации настоящего изобретения «референсная оценка» может быть рассчитана из одного или более «референсных значений».

В одном из вариантов реализации референсное значение или референсная оценка могут быть рассчитаны из данных, полученных из одного или более нормальных или референсных образцов. В одном из вариантов реализации референсное значение или референсная оценка и значение анализируемого образца (например, частоты мотивов нуклеиновых кислот или частоты координат начала и/или конца) или диагностическая оценка для анализируемого образца, с которым его сравнивают, вычисляют в соответствии с тем же способом расчета, который описан в настоящей заявке.

Классификация образца

В настоящей заявке классификация образца включает бинарную классификацию (т. е. рак, отсутствие рака; хороший прогноз, плохой прогноз; рецидивирующий, не рецидивирующий) и классификацию количества вкоДНК как низкого, среднего и высокого количества.

Клинически значимыми категориями для классификации образца могут быть наличие или отсутствие рака, ремиссия заболевания или рака, рецидив заболевания или рака, ранние стадии рака и прогноз.

В одном из вариантов реализации количество, наличие или распространенность опухолевой вкДНК в образце может быть классифицировано как низкое, если комбинированная диагностическая оценка составляет от 2 до 4 стандартных отклонений референсной оценки, как среднее, если комбинированная оценка составляет от 4 до 6,5 стандартного отклонения

референсной оценки, и как высокое, если комбинированная оценка составляет более 6,5 стандартного отклонения референсной оценки.

Применение в онкологии

Настоящее изобретение можно применять при лечении рака или для оценки опухолевой нагрузки, выявления минимальной остаточной болезни, мониторинга исходов лечения, долгосрочного мониторинга исхода у пациента. Настоящее изобретение можно дополнительно применять для идентификации мутаций, подходящих для таргетной терапии, и для выявления раковых соматических и герминативных мутаций. Настоящий способ облегчает раннее выявление небольших опухолей, которые не обнаруживаются другими способами, и обеспечивает более целенаправленный индивидуальный подход к лечению.

Наборы

В другом аспекте в изобретении предложены наборы для выполнения способа по изобретению. В одном из вариантов реализации набор содержит контейнер, состоящий из пула зондов, и программное обеспечение и инструкции для выполнения способа.

Наряду с пулом зондов набор может содержать одно или более из следующего: (i) один или более компонентов для выделения внеклеточной ДНК из биологического образца, (ii) один или более компонентов для приготовления и обогащения библиотеки для секвенирования (например, праймеры, адаптеры, буферы, линкеры, модифицирующие ДНК ферменты, ферменты лигирования, полимеразные ферменты, зонды и т. п.), (iii) один или более компонентов для амплификации и/или секвенирования обогащенной библиотеки и/или (iv) программное обеспечение для выполнения статистического анализа. Компоненты, подходящие для выполнения этапов, указанных в (i), (ii) и (iii), хорошо известны специалисту в данной области техники.

В одном из вариантов реализации зонды предлагаются в форме, которая обеспечивает возможность их связывания с твердой подложкой, например, в форме биотинилированных зондов. В другом варианте реализации зонды предлагаются вместе с твердой подложкой, такие как, например, биотинилированные зонды, предлагаемые вместе с магнитными шариками, покрытыми стрептавидином.

В различных других вариантах реализации набор может содержать дополнительные компоненты для реализации других аспектов способа. Например, помимо пула зондов, набор может содержать одно или более из следующего: (i) один или более компонентов для

выделения внеклеточной ДНК из образца материнской плазмы; (ii) один или более компонентов для приготовления библиотеки для секвенирования (например, праймеры, адаптеры, линкеры, ферменты рестрикции, ферменты лигирования, полимеразные ферменты); (iii) один или более компонентов для амплификации и/или секвенирования обогащенной библиотеки; и/или (iv) программное обеспечение для выполнения статистического анализа. Компоненты, подходящие для выполнения этапов, указанных в (i), (ii) и (iii), хорошо известны специалисту в данной области техники.

ПРИМЕРЫ

Пример 1

Определение начала и/или конца (плюс и/или минус 1 пара нуклеотидов) множества фрагментов вкДНК, содержащихся в образце, осуществили путем выравнивания с референсной последовательностью. Затем определили частоту каждой определенной координаты на последовательности для начала и/или конца во множестве фрагментов вкДНК, содержащихся в образце. Определили отношение частоты каждой определенной координаты в референсном геноме к соответствующей референсной частоте и рассчитали взвешенную сумму (называемую в настоящей заявке «диагностической оценкой») всех полученных отношений частот.

В соответствии с одним вариантом реализации настоящего изобретения для каждого основания i, при $i = 1, \ldots, B$, где B равно общему количеству целевых оснований в указанной панели, случайная переменная X_i определялась как общее количество картированных чтений, удовлетворяющих по меньшей мере одному из следующих условий:

(A1) координата начального положения соответствует основанию i, или

(A2) координата конечного положения соответствует основанию i, или

(A3) координата начального положения минус одно основание соответствует основанию i, или

(A4) координата начального положения плюс одно основание соответствует основанию i, или

(A5) координата конечного положения минус одно основание соответствует основанию i, или

(A6) координата конечного положения плюс одно основание соответствует основанию i.

В соответствии с нулевой гипотезой (т. е. базовой моделью) ожидается, что будет наблюдаться различное, но стабильное количество чтений, удовлетворяющих по меньшей мере одному из условий A1–A6 по различным основаниям генома, причем указанная базовая модель распределения вероятности по основаниям оценивается с помощью набора нормальных образцов. Из приведенного выше определения X_i получаем, что $X_i \sim Bin(x_i; n_i, p_i)$, где n_i равно общему количеству чтений, охватывающих основание i, и p_i, оцененный для всех i, скажем $\widehat{p_i}$, вычисляется следующим образом:

$$\widehat{p_i} = \frac{\sum_{j=1}^{N} z_{i,j}}{\sum_{j=1}^{N} n_{i,j}},$$

где $z_{i,j}$ — наблюдаемое количество чтений, удовлетворяющих по меньшей мере одному из условий A1–A6 по основанию i, для нормального образца j и $n_{i,j}$ — общее количество чтений, охватывающих основание i, для нормального образца j из N нормальных образцов в целом. Биномиальное распределение с очень малым p и большим n может быть аппроксимировано распределением Пуассона с параметром частоты, равным np. Следовательно, базовая модель по основаниям определяется следующей математической формулой: $X_i \sim Po(\widehat{p_i} n_i)$, где n_i равно общему количеству чтений, охватывающих основание i. В другом варианте реализации настоящего изобретения к модели применяют распределение Вейбулла или бета-распределение по каждому основанию i, при этом случайная переменная определяется как $z_{i,j}/n_{i,j}$ для всех j.

После обучения базовой модели по каждому основанию далее действовали следующим образом. Для каждого образца k, в одном из вариантов реализации настоящего изобретения, выполняют следующее: для каждой X_i наблюдаемое значение, скажем x_i, сравнивают с расчетной базовой моделью по основаниям. Если p-значение, т. е. $P(X_i > x_i) = 1 - P(X_i \leq x_i)$, было меньше 0,001, то наблюдаемое значение X_i делили на общее количество чтений, охватывающих основание i, т. е. $Y_i = X_i/n_i$; в противном случае $Y_i = 0$. Затем вычисляли конкретную оценку образца следующим образом: $S_{0,k} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$, где n_2 — общее количество оснований с $Y_i > 0$. Потом $S_{0,k}$ нормируют для получения нормированной оценки $S_{1,k}$ с использованием следующей математической формулы:

$$S_{1,k} = \frac{S_{0,k} - m}{s},$$

где m и s — среднее значение и стандартное отклонение всех значений S_0, полученных для нормальных референсных образцов. (Фигуры 1, 2 и 3.)

Пример 2

После определения координат на последовательности для начала и/или конца (плюс и/или минус 1 пара нуклеотидов) фрагмента вкДНК определяли все мотивы нуклеиновой кислоты в референсной последовательности из референсного генома. Указанные мотивы состояли из тринуклеотидов, тетрануклеотидов и/или пентануклеотидов и находились в пределах

определенного диапазона пар нуклеотидов по направлению внутрь, но на расстоянии 1 или более пар нуклеотидов от координат начала и/или конца. Определили отношение частоты каждого из мотивов нуклеиновой кислоты во множестве фрагментов вкДНК к соответствующей референсной частоте и рассчитали взвешенную сумму (называемую в настоящей заявке «диагностической оценкой») всех полученных отношений частот.

В одном варианте реализации изобретения для каждого образца, скажем k, определили две последовательности для каждого фрагмента вкДНК, выровненного относительно референсного генома hg19, причем указанные последовательности содержали последовательность генома hg19 в диапазоне от 1 до 5 пар нуклеотидов по направлению внутрь от двух концов выровненных фрагментов вкДНК (исключая последовательность нуклеиновой кислоты, охватываемую фрагментом), и рассчитали абсолютную частоту всех мотивов тринуклеотидных (например, ACC, GGT и т. д.), тетрануклеотидных и пентануклеотидных последовательностей в пределах указанных последовательностей, скажем T_{ij} для $i = 1, \dots, n_j$, где $j = 3{,}4{,}5$ представляет собой количество нуклеотидов и n_j представляет собой количество всех возможных j-нуклеотидных мотивов ($n_3 = 64, n_4 = 256,$ $n_5 = 1024$). Конкретную оценку образца $S_{2,k}$ рассчитывают следующим образом:

$$S_{2,k} = \sum_{j=3}^{5} S_{2,jk} b_j$$

где: $$S_{2,jk} = \frac{1}{n_j} \sum_{i=1}^{n_j} \chi_{ij}^2 w_{ij}$$

$$\chi_{ij}^2 = \left(\frac{\widetilde{f_{ij}} - m_{ij}}{s_{ij}} \right)^2 \quad , \quad \widetilde{f_{ij}} = \frac{f_{ij}}{r_{ij}} \quad , \quad f_{ij} = \frac{T_{ij}}{D_k}$$

В приведенных выше формулах D_k представляет собой общее количество консенсусных фрагментов в образце k, r_{ij} — референсное значение f_{ij}, рассчитанное по обучающему набору данных от не содержащих цоДНК образцов, m_{ij} и s_{ij} — референсные среднее значение и стандартное отклонение $\widetilde{f_{ij}}$, рассчитанные по обучающему набору данных от не содержащих цоДНК образцов, w_{ij} являются весами $\left(\sum_{i=1}^{n_j} w_{ij} = 1\right)$, которые оптимизированы из обучающего набора для обеспечения оптимального разделения между нормальными и не соответствующими норме образцами. Веса b_j могут варьировать в различных вариантах

реализации настоящего изобретения: $b_3 = 1/12$ или $1/6$ или $1/3$ или $1/2$ или $b_4 = 1/12$ или $1/6$ или $1/3$ или $1/2$ и $b_5 = 1 - b_3 - b_4$. (Фигуры 1, 2 и 3.)

Пример 3

После определения координат на последовательности для начала и/или конца (плюс и/или минус 1 пара нуклеотидов) фрагмента вкДНК определяли все мотивы нуклеиновой кислоты в референсной последовательности из референсного генома. Указанные мотивы состояли из тринуклеотидов, тетрануклеотидов и/или пентануклеотидов и находились в пределах определенного диапазона пар нуклеотидов по направлению наружу, но на расстоянии 1 или более пар нуклеотидов от координат начала и/или конца. Определили отношение частоты каждого из мотивов нуклеиновой кислоты во множестве фрагментов вкДНК к соответствующей референсной частоте и рассчитали взвешенную сумму (называемую в настоящей заявке «диагностической оценкой») всех полученных отношений частот.

В одном варианте реализации способа для каждого образца, скажем k, определили две последовательности для каждого фрагмента вкДНК, выровненного относительно референсного генома hg19, причем указанные последовательности содержали последовательность генома hg19 в диапазоне от 1 до 5 пар нуклеотидов по направлению наружу от двух концов выровненных фрагментов вкДНК (исключая последовательность нуклеиновой кислоты, охватываемую фрагментом), и рассчитали абсолютную частоту всех мотивов тринуклеотидных (например, ACC, GGT и т. д.), тетрануклеотидных и пентануклеотидных последовательностей в пределах указанных последовательностей, скажем T_{ij} для $i = 1, \ldots, n_j$, где $j = 3{,}4{,}5$ представляет собой количество нуклеотидов и n_j представляет собой количество всех возможных j-нуклеотидных мотивов ($n_3 = 64, n_4 = 256$, $n_5 = 1024$). Конкретную оценку образца $S_{3,k}$ рассчитывают следующим образом:

$$S_{3,k} = \sum_{j=3}^{5} S_{3,jk} b_j$$

где: $$S_{3,jk} = \frac{1}{n_j} \sum_{i=1}^{n_j} \chi_{ij}^2 w_{ij}$$

$$\chi_{ij}^2 = \left(\frac{\widetilde{f_{ij}} - m_{ij}}{s_{ij}} \right)^2 \quad , \quad \widetilde{f_{ij}} = \frac{f_{ij}}{r_{ij}} \quad , \quad f_{ij} = \frac{T_{ij}}{D_k}$$

В приведенных выше формулах D_k представляет собой общее количество консенсусных фрагментов в образце k, r_{ij} — референсное значение f_{ij}, рассчитанное по обучающему набору данных от не содержащих цоДНК образцов, m_{ij} и s_{ij} — референсные среднее значение и стандартное отклонение $\widetilde{f_{ij}}$, рассчитанные по обучающему набору данных от не содержащих цоДНК образцов, w_{ij} являются весами $\left(\sum_{i=1}^{n_j} w_{ij} = 1\right)$, которые оптимизированы из обучающего набора для обеспечения оптимального разделения между нормальными и не соответствующими норме образцами. Веса b_j могут варьировать в различных вариантах реализации настоящего изобретения: $b_3 = 1/12$ или $1/6$ или $1/3$ или $1/2$ или $b_4 = 1/12$ или $1/6$ или $1/3$ или $1/2$ и $b_5 = 1 - b_3 - b_4$. (Фигуры 1, 2 и 3.)

Пример 4

В одном варианте реализации способа взвешенную сумму по меньшей мере двух оценок, рассчитанных в Примерах 1, 2 и 3, вычислили для каждого образца; далее указанная взвешенная сумма называется «комбинированной диагностической оценкой». Диагностическая оценка для образца k, скажем DS_k, определяется как средневзвешенное значение по меньшей мере двух из оценок, описанных в Примерах 1, 2 и 3 выше, т. е.:

$$\mathrm{DS}_k = \sum_{i=1}^{3} w_i S_{i,k},$$

где S_1, S_2 и S_3 рассчитаны в Примерах 1, 2 и 3 соответственно, и в различных вариантах реализации изобретения $w_1 = 0{,}5$ или $0{,}4$ или $0{,}3$ или $0{,}2$ или 0 с округлением до одного знака после запятой, $w_2 = 0{,}5$ или $0{,}4$ или $0{,}3$ или $0{,}2$ или 0 с округлением до одного знака после запятой, и $w_3 = 1 - w_1 - w_2$. В другом варианте реализации способа средневзвешенное значение максимума и минимума $\{S_1, S_2, S_3\}$ применяют для вычисления оценки DS для образца k, то есть $DS_k = z\mathrm{MAX}\left(S_{1,k}, S_{2,k}, S_{3,k}\right) + (1 - z)\mathrm{MIN}\left(S_{1,k}, S_{2,k}, S_{3,k}\right)$, при $0{,}5 < z < 1$.

ФОРМУЛА ИЗОБРЕТЕНИЯ

1. Способ классификации образца как содержащего внеклеточную опухолевую ДНК, включающий следующие этапы:

   (i) определение в образце, содержащем множество фрагментов внеклеточной ДНК (вкДНК), координат на последовательности для начала и/или конца и для начала и/или конца плюс и/или минус 1 пара нуклеотидов по меньшей мере у 100 000 фрагментов вкДНК путем выравнивания с референсной последовательностью,

   (ii) определение частоты каждой координаты, определенной на этапе (i), в указанном множестве фрагментов вкДНК, содержащихся в указанном образце,

   (iii) вычисление отношения частоты каждой координаты, определенной на этапе (ii), к соответствующей референсной частоте,

   (iv) вычисление диагностической оценки на основе всех отношений, определенных на этапе (iii), причем указанная оценка представляет собой взвешенную сумму всех отношений частоты, определенных на этапе (iii), и

   (v) определение классификации указанного образца путем сравнения диагностической оценки с референсной оценкой,

   при этом указанный образец классифицируют как содержащий опухолевую вкДНК, если значение диагностической оценки выше среднего значения референсной оценки по меньшей мере на одно стандартное отклонение референсной оценки, причем указанную референсную оценку рассчитывают на основе одного или более референсных значений.

2. Способ классификации образца как содержащего внеклеточную опухолевую ДНК, включающий следующие этапы:

   (i) определение в образце, содержащем множество фрагментов внеклеточной ДНК (вкДНК), координат на последовательности для начала и/или конца по меньшей мере у 100 000 фрагментов вкДНК путем выравнивания с референсной последовательностью,

(ii) определение в референсной последовательности всех мотивов нуклеиновой кислоты, состоящих из тринуклеотидов, тетрануклеотидов и пентануклеотидов, в диапазоне от 1 до 5 пар нуклеотидов по направлению внутрь, но рядом с координатой на последовательности для каждого начала и/или конца, определенной на этапе (i),

(iii) определение частоты каждого из мотивов нуклеиновой кислоты, определенных на этапе (ii), в указанном множестве фрагментов вкДНК, содержащихся в указанном образце,

(iv) вычисление отношения каждой из частот, определенных на этапе (iii), к соответствующей референсной частоте,

(v) вычисление диагностической оценки на основе всех отношений, определенных на этапе (iv), причем указанная оценка представляет собой взвешенную сумму всех отношений частоты, определенных на этапе (iv), и

(vi) определение классификации указанного образца путем сравнения диагностической оценки с референсной оценкой,

при этом указанный образец классифицируют как содержащий опухолевую вкДНК, если значение диагностической оценки выше среднего значения референсной оценки по меньшей мере на одно стандартное отклонение референсной оценки, причем указанную референсную оценку рассчитывают на основе одного или более референсных значений.

3. Способ классификации образца как содержащего внеклеточную опухолевую ДНК, включающий следующие этапы:

(i) определение в образце, содержащем множество фрагментов внеклеточной ДНК (вкДНК), координат на последовательности для начала и/или конца по меньшей мере у 100 000 фрагментов вкДНК путем выравнивания с референсной последовательностью,

(ii) определение в референсной последовательности всех мотивов нуклеиновой кислоты, состоящих из тринуклеотидов, тетрануклеотидов и пентануклеотидов, в диапазоне от 1 до 5 пар нуклеотидов по направлению наружу, но рядом с координатой на последовательности для каждого начала и/или конца, определенной на этапе (i),

(iii) определение частоты каждого из мотивов нуклеиновой кислоты, определенных на этапе (ii), в указанном множестве фрагментов вкДНК, содержащихся в указанном образце,

(iv) вычисление отношения каждой из частот, определенных на этапе (iii), к соответствующей референсной частоте,

(v) вычисление диагностической оценки на основе всех отношений, определенных на этапе (iv), причем указанная оценка представляет собой взвешенную сумму всех отношений частоты, определенных на этапе (iv), и

(vi) определение классификации указанного образца путем сравнения диагностической оценки с референсной оценкой,

при этом указанный образец классифицируют как содержащий опухолевую вкДНК, если значение диагностической оценки выше среднего значения референсной оценки по меньшей мере на одно стандартное отклонение референсной оценки, причем указанную референсную оценку рассчитывают на основе одного или более референсных значений.

4. Способ классификации образца как содержащего внеклеточную опухолевую ДНК, включающий следующие этапы:

(i) определение в образце, содержащем множество фрагментов внеклеточной ДНК (вкДНК), координат на последовательности для начала и/или конца по меньшей мере у 100 000 фрагментов вкДНК путем выравнивания с референсной последовательностью,

(ii) определение в референсной последовательности всех мотивов нуклеиновой кислоты, состоящих из тринуклеотидов, тетрануклеотидов и пентануклеотидов:

a) в диапазоне от 1 до 5 пар нуклеотидов по направлению внутрь, но рядом с координатой на последовательности для каждого начала и/или конца, определенной на этапе (i), и/или

b) в диапазоне от 1 до 5 пар нуклеотидов по направлению наружу, но рядом с координатой на последовательности для каждого начала и/или конца, определенной на этапе (i),

(iii) определение частоты:

a) каждой координаты на последовательности плюс и/или минус 1 пара нуклеотидов, определенной на этапе (i), во множестве фрагментов вкДНК, содержащихся в указанном образце,

b) каждого из мотивов нуклеиновой кислоты, определенных на этапе (ii) a) и b), во множестве фрагментов вкДНК, содержащихся в указанном образце,

(iv) вычисление отношения каждой из частот, определенных на этапе (iii) a) и b), к соответствующей референсной частоте,

(v) вычисление диагностической оценки отдельно для каждого отношения, определенного на этапе (iv), причем указанная оценка представляет собой соответствующую взвешенную сумму всех соответствующих отношений частот, полученных на этапе (iv),

(vi) вычисление комбинированной диагностической оценки на основе по меньшей мере двух или более диагностических оценок, определенных на этапе (v), причем указанная оценка представляет собой взвешенную сумму указанных двух или более диагностических оценок, определенных на этапе (v), и

(vii) определение классификации указанного образца путем сравнения комбинированной диагностической оценки с референсной оценкой,

при этом указанный образец классифицируют как содержащий опухолевую вкДНК, если значение комбинированной диагностической оценки выше среднего значения референсной оценки по меньшей мере на одно стандартное отклонение референсной оценки, причем указанную референсную оценку рассчитывают на основе одного или более референсных значений.

5. Способ по п. 4, отличающийся тем, что комбинированную диагностическую оценку рассчитывают на основе всех диагностических оценок, вычисленных на этапе (v) в п. 4.

6. Способ по пп. 2–5, отличающийся тем, что диапазон пар нуклеотидов по направлению внутрь, но рядом с координатой на последовательности для каждого начала и/или конца, может составлять от 2 п. н. до 6 п. н., или от 3 п. н. до 7 п. н., или от 4 п. н. до 8 п. н., или от 5 п. н. до 9 п. н., или от 6 п. н. до 10 п. н. от координаты каждого начала и/или конца.

7. Способ по любому из пп. 1–6, отличающийся тем, что минимальное количество фрагментов вкДНК, содержащихся в образце для анализа, составляет от 100 тысяч до 500 тысяч, от

500 тысяч до 1 миллиона, от 1 миллиона до 2 миллионов, от 2 миллионов до 5 миллионов, или от 5 миллионов до 10 миллионов, или от 10 миллионов до 20 миллионов, или от 20 миллионов до 50 миллионов, или от 50 миллионов до 500 миллионов.

8. Способ по пп. 4–7, отличающийся тем, что количество опухолевой вкДНК в образце может быть классифицировано как низкое, если указанная комбинированная диагностическая оценка составляет от 2 до 4 стандартных отклонений референсной оценки, как среднее, если указанная комбинированная оценка составляет от 4 до 6,5 стандартных отклонений референсной оценки, и как высокое, если указанная комбинированная оценка составляет более 6,5 стандартных отклонений референсной оценки.

9. Способ по любому из пп. 1–8, отличающийся тем, что референсные образцы могут представлять собой образцы от пациентов без рака, или от пациентов без рецидива, или от успешно вылеченных пациентов с раком.

10. Способ по любому из пп. 1–8, отличающийся тем, что этап (i) включает определение последовательности нуклеиновой кислоты по меньшей мере у части из множества фрагментов вкДНК в указанном образце перед выравниванием с референсной последовательностью.

11. Способ по пп. 1–10, отличающийся тем, что этап (i) дополнительно включает обогащение фрагментов вкДНК перед определением последовательности нуклеиновой кислоты фрагментов вкДНК.

12. Способ по любому из предшествующих пунктов, отличающийся тем, что указанный образец классифицируют как содержащий опухолевую вкДНК, происходящую из опухоли, выбранной из группы, включающей рак крови, рак печени, рак легкого, рак поджелудочной железы, рак предстательной железы, рак молочной железы, рак желудка, глиобластому, колоректальный рак, рак головы и шеи, солидную опухоль, доброкачественную опухоль, злокачественную опухоль, рак на поздней стадии, метастаз или предраковую ткань.

13. Набор, содержащий:

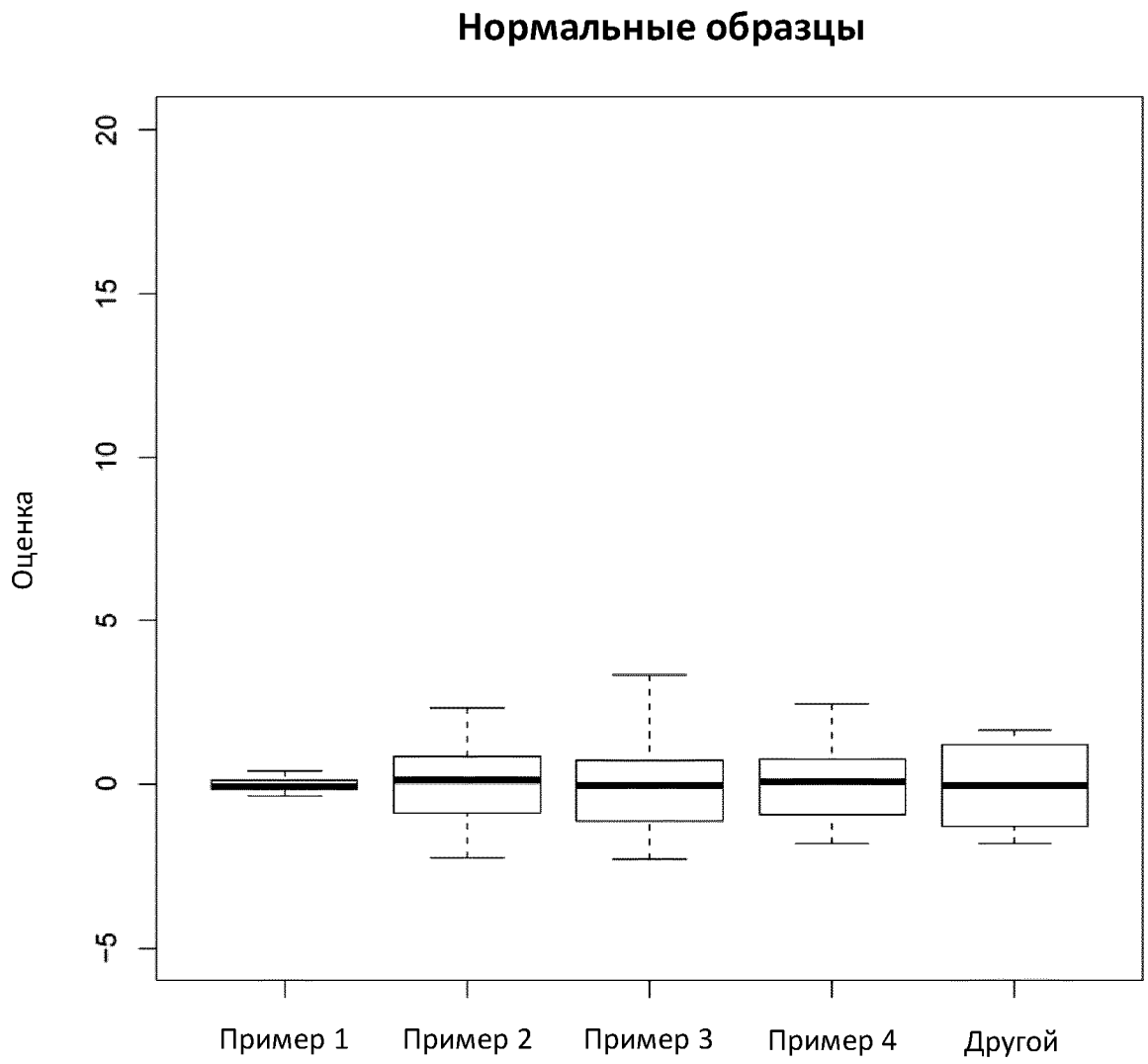
    (i) компоненты для осуществления способа по любому из пп. 1–12,
    причем компоненты содержат:
    a) один или более компонентов для выделения внеклеточной ДНК из биологического образца,

b) один или более компонентов для приготовления и обогащения библиотеки для секвенирования, и/или

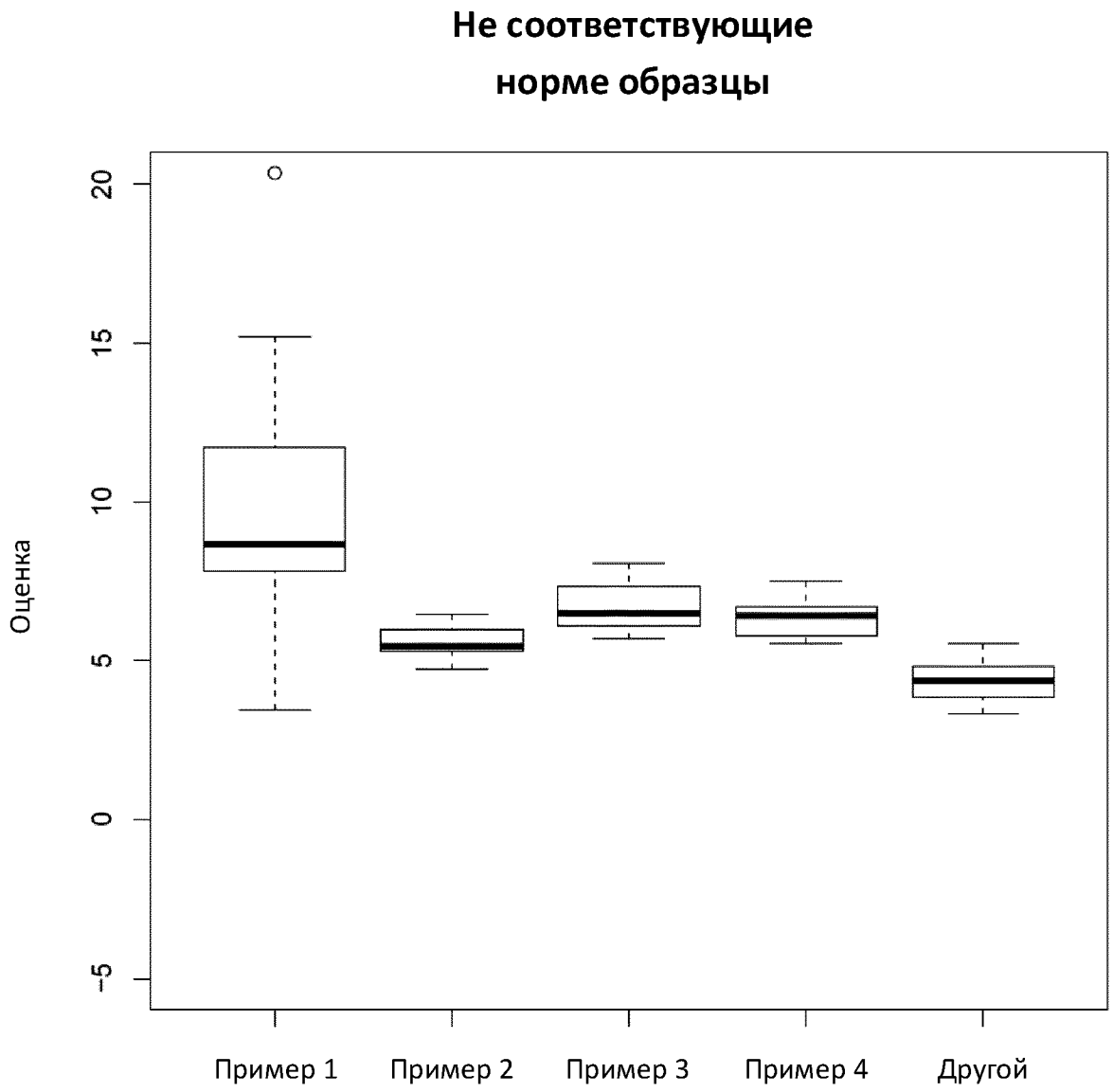
c) один или более компонентов для амплификации и/или секвенирования указанной обогащенной библиотеки,

(ii) программное обеспечение для проведения статистического анализа.

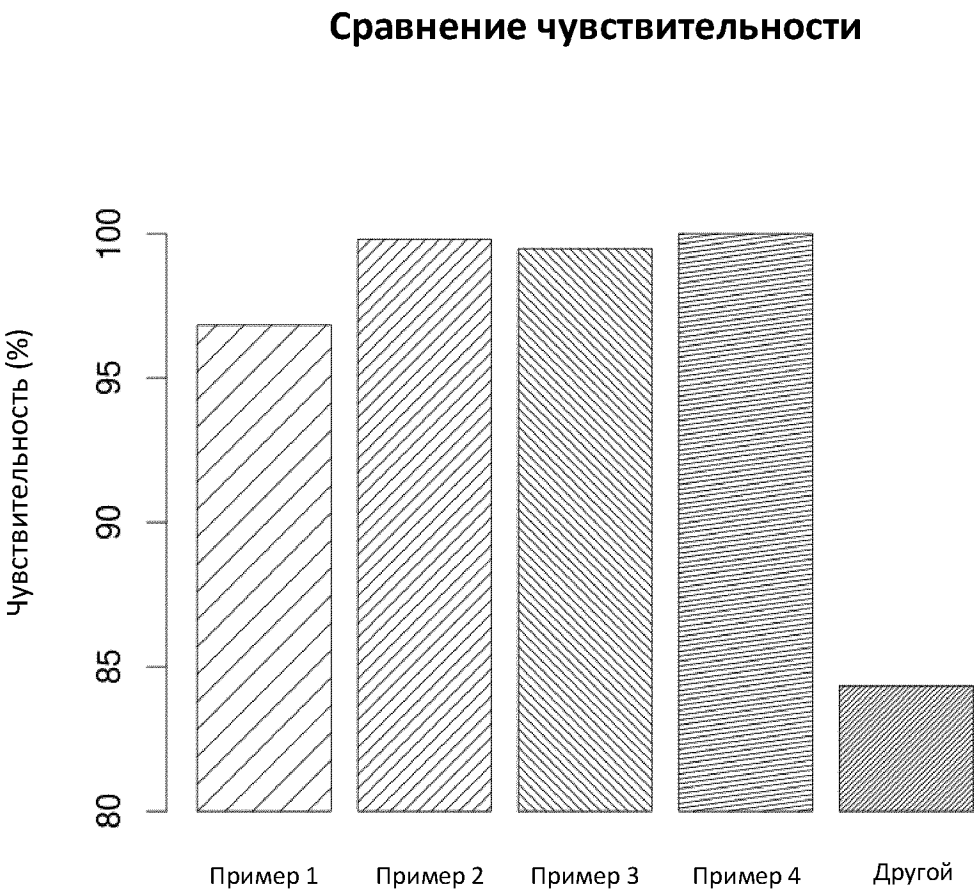
Фигура 1


Нормальные образцы
Оценка
20
15
10
5
0
−5
Пример 1
Пример 2
Пример 3
Пример 4
Другой

Фигура 2


Не соответствующие
норме образцы
Оценка
20
15
10
5
0
−5
Пример 1
Пример 2
Пример 3
Пример 4
Другой

Фигура 3


Сравнение чувствительности
Чувствительность (%)
80
85
90
95
100
Пример 1
Пример 2
Пример 3
Пример 4
Другой

Фигура 4

Таблица 1

Образец	DS	Истинный статус	Прогнозируемый класс
1	**1,3**	**Нормальный**	**Нормальный**
2	**0,9**	**Нормальный**	**Нормальный**
3	**-0,6**	**Нормальный**	**Нормальный**
4	**0,8**	**Нормальный**	**Нормальный**
5	**4,6**	**НМРЛ, I стадия**	**Умеренное количество цоДНК**
6	**3,3**	**НМРЛ, I стадия**	**Низкое количество цоДНК**
7	**3,8**	**НМРЛ, I стадия**	**Низкое количество цоДНК**