(19) Евразийское патентное ведомство

(21) 202490768 (13) A1

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ЕВРАЗИЙСКОЙ ЗАЯВКЕ

(43) Дата публикации заявки
2024.05.22

(22) Дата подачи заявки
2022.09.19

(51) Int. Cl. ***C12Q 1/6855*** (2018.01)
C12Q 1/6869 (2018.01)
G16B 30/10 (2019.01)
G16B 20/20 (2019.01)

(54) СПОСОБ АНАЛИЗА СТЕПЕНИ СХОДСТВА ПО МЕНЬШЕЙ МЕРЕ ДВУХ ОБРАЗЦОВ С ИСПОЛЬЗОВАНИЕМ ПОЛНОГЕНОМНОЙ АМПЛИФИКАЦИИ С ДЕТЕРМИНИСТИЧЕСКИМ УЧАСТКОМ РЕСТРИКЦИИ (DRS-WGA)

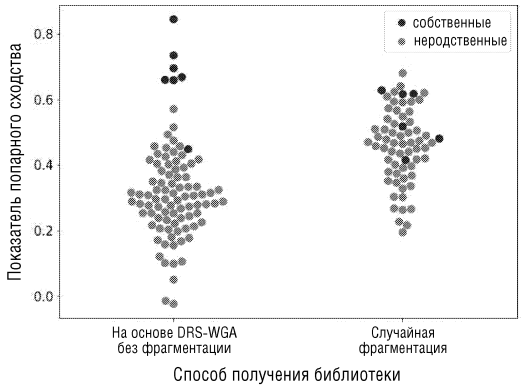
(31) **102021000024101**
(32) **2021.09.20**
(33) **IT**
(86) **PCT/IB2022/058833**
(87) **WO 2023/042173 2023.03.23**

(71) Заявитель:
МЕНАРИНИ СИЛИКОН БАЙОСИСТЕМЗ С.П.А. (IT)

(72) Изобретатель:
Манарези Николо, Форкато Клаудио, Феррарини Альберто (IT)

(74) Представитель:
Медведев В.Н. (RU)

(57) Настоящее изобретение относится к способу анализа степени сходства по меньшей мере двух образцов в множестве образцов, содержащих геномную ДНК. Способ включает следующие стадии. a) Получение множества образцов, содержащих геномную ДНК. b) Проведение, отдельно для каждого образца, полногеномной амплификации с детерминистическим участком рестрикции (DRS-WGA) указанной геномной ДНК. c) Получение библиотеки для массивного параллельного секвенирования с использованием реакции ПЦР с адаптером для секвенирования/праймером для слияния WGA, без фрагментации, из каждого продукта указанной DRS-WGA. d) Проведение полногеномного секвенирования с низким покрытием при средней глубине покрытия < 1x для указанной библиотеки для массивного параллельного секвенирования. e) Выравнивание для каждого образца прочтений, полученных на стадии d), на эталонном геноме. f) Извлечение для каждого образца аллельного содержания в множестве полиморфных локусов. g) Расчет показателя попарного сходства для по меньшей мере двух образцов как функции от аллельного содержания, измеренного в указанном множестве локусов. h) Определение степени сходства по меньшей мере двух образцов на основании показателя сходства.



202490768 A1

202490768 A1

ОПИСАНИЕ ИЗОБРЕТЕНИЯ

2420-580781EA/085

СПОСОБ АНАЛИЗА СТЕПЕНИ СХОДСТВА ПО МЕНЬШЕЙ МЕРЕ ДВУХ ОБРАЗЦОВ С ИСПОЛЬЗОВАНИЕМ ПОЛНОГЕНОМНОЙ АМПЛИФИКАЦИИ С ДЕТЕРМИНИСТИЧЕСКИМ УЧАСТКОМ РЕСТРИКЦИИ (DRS-WGA)

Перекрестная ссылка на родственные заявки

По настоящей патентной заявке испрашивается приоритет Итальянской патентной заявки no. 102021000024101, поданной 20 сентября 2021 г., полное содержание которой приведено в настоящем описании в качестве ссылки.

Область техники

Настоящее изобретение относится к способу спаривания образцов, присвоения идентичности каждого из множества образцов классу или индивидууму, посредством анализа данных, полученных посредством полногеномного секвенирования с низким покрытием, проведенного для указанного множества образцов, достигающего разрешения отдельных клеток, с использованием или без использования эталонов.

В дополнение к спариванию образцов, способ предоставляет унифицированный анализ, позволяющий одновременную идентификацию и характеризацию тестируемого образца среди образцов.

Способ в соответствии с настоящим изобретением можно использовать в нескольких областях применения, включая, но без ограничения:

- криминалистическую идентификацию человека по отдельной клетке
- идентификацию образца в ходе анализа циркулирующих клеток опухолей
- идентификацию фетальных клеток или фетальной внеклеточной ДНК (вкДНК) в материнских жидкостях организма для неинвазивного пренатального тестирования
- идентификацию клеток эмбриона или вкДНК при инвазивном преимплантационном генетическом тестировании (PGT) и неинвазивном PGT в использованной среде для эмбриона
- идентификацию фетального компонента при пренатальной диагностике в инвазивно полученных образцах и продуктах зачатия (например: оценка материнской или экзогенной контаминации)
- молярную беременность, многоплодную беременность (включая исчезновение/химеру), однородительскую дисомию (изодисомию или гетеродисомию), ROH и идентификацию кровного родства, классификацию ошибки нерасхождения в материале, происходящем из концептуса
- микрохимеризм

аутентификацию линии клеток (например: стволовых клеток).

Предшествующий уровень техники

Предшествующий уровень техники для идентификации образца и спаривания образцов

Наиболее широко распространенный способ идентификации образца основан на

анализе высокополиморфных локусов коротких тандемных повторов (STR) (также называемых микросателлиты). Этот способ включает проведение нацеленной ПЦР для множества локусов и детекцию ампликонов с использованием капиллярного электрофореза. При идентификации человека, поскольку для каждого локуса каждый аллель (из-за материнского и отцовского происхождения) может иметь множество различных значений, образуется большое разнообразие при относительно низком количестве амплифицированных генетических локусов, такое, что размеры аллелей индивидуума, измеренные среди 10 или 20 локусов, могут идентифицировать с высокой вероятностью индивидуума в большой когорте. Применение этого способа для отдельных клеток может являться проблематичным, особенно, если качество ДНК является низким, или она деградирована (например, деградирована из-за фиксации или условий окружающей среды для хранения, или других биологических процессов), поскольку выпадение аллелей может нарушать извлечение достаточной информации для присвоения идентичности образца. Это справедливо, независимо от того факта, проводят ли мультиплексную ПЦР напрямую для образца отдельной клетки (таким образом, расходуя этот образец) или для аликвоты продукта полногеномной амплификации из отдельной клетки, таким образом, позволяя повторяющееся тестирование для различных аликвот одного и того же продукта WGA.

Выпадение аллелей может значительно уменьшать количество аллелей, детектированных на электрофореграмме анализа STR, вплоть до 80%, 70%, 60%, 50%, 40%, 30%, 20%, 10% или ниже. Кроме того, может происходить вброс аллеля, приводящий к дополнительным пикам, искажающих интерпретацию, особенно для высоко деградированных образцов и матрицы с низким вводом, например, с использованием отдельных клеток. Поэтому полученная информация является недостаточной для присвоения идентичности образца с доверием.

Требования к минимальному количеству аллелей из локусов STR зависят от нескольких факторов, но в общем справедливо и известно специалисту в данной области, что при установлении соответствия профиля большой популяции необходимы намного более информативные локусы, в то время как установление соответствия образца меньшей когорте потенциальных участников представляет более простую проблему, которую можно разрешить с использованием более низкого количества детектированных аллелей.

Например, в криминалистических экспертизах, таких как после изнасилования, могут присутствовать ДНК и клетки от одного или нескольких виновных и потерпевшего, с количеством участников, которое может составлять 1 потерпевший и 1, 2, 3, 4, 5 или более виновных. В случае множества виновных мужского пола, проблема может усугубляться тем фактом, что клетки-мишени для анализа представляют собой клетки спермы, которые, являясь гаплоидными, имеют только один аллель на локус. При анализе отдельных клеток для экспертизы, может, таким образом, становиться невозможным использовать информацию отдельной клетки для надежного вывода количества участников и сборки реконструированного полного профиля для данного участника при ограниченных данных

для отдельной клетки.

В качестве примера, отдельные клетки спермы можно выделять с использованием DEPArray (Fontana et. al, «Isolation and genetic analysis of pure cells from forensic biological mixtures: The precision of a digital approach», Forensic Sciences International: Genetics 2007, http://dx.doi.org/10.1016/j.fsigen.2017.04.023), который позволяет сбор вплоть до 48 отдельных сперматозоидов из одного цикла DEPArray, с использованием проверенного криминалистического приложения, или вплоть до 96 отдельных клеток с использованием различных прикладных программ, доступных в системе DEPArray.

Криминалистическая идентификация профилей различных участников по отдельным клеткам в крови из смешанных улик крови, с использованием системы DEPArray для выделения индивидуальных клеток, была показана в K. Anslinger, B. Bayer, «Whose blood is it? Application of DEPArray™ technology for the identification of individual/s who contributed blood to a mixed stain» Int J Legal Med. 2019 Mar;133(2):419-426. doi: 10.1007/s00414-018-1912-7. Epub 2018 Aug 18.

Как правило, проблему реконструкции полного профиля и/или определения генетической информации посредством реконструкции *in silico* полного профиля из множества неполных профилей отдельных клеток, труднее разрешить, чем

(i) ниже количество анализируемых отдельных клеток,

(ii) ниже количество аллелей, детектированных на клетку,

(iii) выше количество участников,

(iv) ниже уровень представленности минорных участников среди анализируемых клеток.

Помимо криминалистического анализа отдельных клеток, полученных посредством прямого выделения индивидуальных клеток, другие способы, такие как взятие подвыборок (K. Huffman, E. Hanson and J Ballantyne, «Recovery of single source DNA profiles from mixtures by direct single cell subsampling and simplified micromanipulation», Science & Justice Volume 61, Issue 1, January 2021, Pages 13-25), подразумевают анализ множества образцов, состоящих из коллекций небольших пулов клеток, например, из 2 или 3 клеток на пул. Также, в этом случае, может обеспечивать преимущество наличие системы для идентификации того, состоит ли пул из клеток от одного и того же участника или множества участников, и, возможно, идентификации общего количества участников среди всех пулов, так же как обеспечения возможности дальнейшего генетического анализа гомогенных пулов, например, для дополнительных исследовательский целей, таких как определение происхождения или физических признаков, связанных с геномными характеристиками.

В качестве следующего примера, аутентификацию линий клеток обычно проводят с использованием анализа STR. Большинство наборов для STR требуют секвенаторов с капиллярным электрофорезом для анализа длины фрагментов флуоресцентных продуктов амплификации. С распространением секвенаторов для массивного параллельного секвенирования, доступность капиллярного электрофореза уменьшилась, и многие

лаборатории находят затруднительным анализ своими силами профилей STR с использованием капиллярного электрофореза. В настоящее время являются доступными панели для нацеленной ПЦР для анализа STR с использованием секвенаторов для массивного параллельного секвенирования. Однако, это подразумевает приобретение дополнительных реагентов, часто ранее не присутствовавших в лаборатории.

В качестве следующего примера, существует потребность в идентификации и/или спаривании образцов в способах неинвазивной пренатальной диагностики на основе выделения фетальных клеток из материнской физиологической жидкости. Они могут представлять собой, например, фетальные клетки (такие как фетальные ядросодержащие эритроциты или трофобласты), выделенные из материнской крови. Принимая во внимание, что клетки являются настолько редкими, существует значительный риск того, что индивидуальные клетки, выделенные в результате процесса обогащения, могут представлять собой материнские клетки, в отличие от фетальных клеток, из-за нескольких причин, таких как ограниченная специфичность при иммунофлуоресцентном окрашивании или неоднозначный морфологический отбор, технические несовершенства и ошибки в оборудовании для сортировки, используемом при их выделении. Независимо от способа и критериев, используемых для выделения этих клеток, принимая во внимание важность гарантии того, что диагностику проводят для фактической фетальной клетки, являются необходимыми подтверждение, только ли фетальный генетический материал является вводом генетического анализа, и детекция возможной материнской контаминации (смешанных клеток), или полной замены образца (отдельная клетка является материнской), или даже контаминации, например, от оператора. В то время как смешанный образец (например, 1 фетальная клетка, 1 материнская клетка, т.е., 50% контаминация) может все еще являться приемлемым для некоторых анализов хромосомной анеуплоидии, более низкая чистота может нарушать детекцию меньших аберраций, подобных микроделециям, в зависимости от используемого анализа.

Таким образом современной практикой на текущем уровне техники является проведение анализа STR в качестве дополнительного подтверждающего теста фетального происхождения для клетки, выделенной в ходе NIPD на основе клеток (Vossaert L, Wang Q, Salman R et al. «Validation Studies for Single Circulating Trophoblast Genetic Testing as a Form of Noninvasive Prenatal Diagnosis» American Journal of Human Genetics (2019) 105(6) 1262-1273; L.D. Jeppesen et al., «Cell-based non-invasive prenatal diagnosis in a pregnancy at risk of cystic fibrosis» Prenatal Diagnosis. 2020;1-7.; Manaresi et al., EP2152859B1).

В недавней статье (Zhuo X, Wang Q, Vossaert L, Salman R, Kim A, Van den Veyver I, et al. (2021) «Use of amplicon-based sequencing for testing fetal identity and monogenic traits with Single Circulating Trophoblast (SCT) as one form of cell-based NIPT» PLoS ONE 16(4): e0249695. https://doi.org/10.1371/journal.pone.0249695) выявили, что «полногеномное секвенирование (WGS) способом дробовика с низким покрытием (5-10 миллионов прочтений на клетку) предоставляет хорошие данные количества копий, но не различает легко фетальные и материнские клетки, если плод женского пола». В этой работе,

генотипирование с использованием панели из 90 высокополиморфных SNP с использованием нацеленной амплификации на основе ПЦР (40 ампликонов) и массивное параллельное секвенирование предложено в качестве альтернативы анализу STR, для подтверждения фетального происхождения клетки, выделенной для диагностики. В этом способе используют небольшую аликвоту ДНК из продукта WGA отдельной клетки, однако это все еще имеет недостаток требования дополнительного исследования образцов и ассоциированных затрат, относительно технологического маршрута для оценки анеуплоидии на основе WGS с низким покрытием.

Неинвазивная оценка случаев молярной беременности и гестационной трофобластической болезни показана для циркулирующих трофобластов (Sunde L et al., «Hydatidiform mole diagnostics using circulating gestational trophoblasts isolated from maternal blood» Mol Genet Genomic Med. 2020;00:e1565. https://doi.org/10.1002/mgg3.1565), однако, анализ STR снова сочли необходимым для определения происхождения редких трофобластов, выделенных из материнской крови. Пузырные заносы (HM) могут представлять собой «полные заносы», которые являются, как правило, диплоидными, с обоими геномными наборами, происходящими от отца (родительский тип: PP), из-за оплодотворения яйцеклетки, утратившей материнское ядро, за которым следует, в большинстве случаев, дупликация хромосом сперматозоида, или - в меньшинстве случаев - оплодотворение посредством двух сперматозоидов. Для большинства HM с родительским типом PP показана гомозиготность по всем локусам (P1P1), в то время как для приблизительно 15% показана гетерозиготность по некоторым локусам (P1P2). Частичные заносы представляют собой HM, как правило, триплоидные, с двумя геномными наборами от отца и одним от матери (родительский тип: PPM). Полные заносы приводят к увеличенному риску хориокарциномы (15%, относительно 0,5% при частичных заносах). Таким образом, представляет интерес понимание, несут ли HM копию материнского генома или она отсутствует.

В качестве следующего примера потребности в способах спаривания образцов, присутствует идентификация для отслеживание образца на лабораторном технологическом маршруте. При секвенировании множества образцов для полногеномного секвенирования с низким покрытием, для получения полногеномного профиля количества копий, может обеспечивать преимущество подтверждение, что не присутствует путаницы образцов, и что присвоение кода образца от пациента в лабораторной информационно-управляющей системе (LIMS) согласуется с присвоением пациента, полученным по данным секвенирования.

Другим примером потребности в способах спаривания образцов является оценка происхождения эндотелиальных клеток (хозяина или донора) у пациентов при аллогенной трансплантации гематопоэтических клеток (алло-HSCT). Детекция происходящих от донора эндотелиальных клеток представляет интерес в исследовании физио-патологических взаимосвязей между эндотелием и реакцией трансплантат против хозяина (GVHD), по потенциальной роли эндотелия сосудов в качестве мишени в ранней фазе

GVHD и потенциальной толерогенной роли происходящих от донора эндотелиальных клеток, так же как реакции трансплантат против опухоли (обзор приведен в Penack O. et al., «The importance of neovascularization and its inhibition for allogeneic hematopoietic stem cell transplantation» Blood, Volume 117, Issue 16, 21 April 2011, Pages 4181-4189). Не совпадающие по полу образцы часто используют для обеспечения возможности такого анализа, но было бы желательно иметь способ для анализа образцов, когда хозяин и донор имеют одинаковый пол. Анализ STR после выделения отдельных клеток посредством DEPArray опубликован для анализа циркулирующих эндотелиальных клеток, обогащенных из периферической крови. Однако, анализ STR отдельных клеток в архивных образцах, таких как FFPE, является трудно достижимым, из-за деградации ДНК, мешающей анализу STR отдельных клеток.

Неинвазивный пренатальный скрининг на основе циркулирующей вкДНК по случаям фетального хромосомного дисбаланса можно оценивать для достаточной фетальной фракции (FF) ДНК, поскольку низкие уровни могут приводить к получению ложноотрицательньных результатов. Таким образом, может являться важным точно оценить фетальную фракцию ДНК, убедиться, что она преодолевает порог QC для обеспечения достаточного количества фетальной ДНК, присутствующего в образце для тестирования, и обеспечить возможность достижения надлежащей интерпретации результата секвенирования. В некоторых лабораториях могут не оценивать FF или оценивать не с использованием оптимальных способов детекции, и это может потенциально приводить к получению ложноотрицательньных результатов для пациентов. Современные способы, разработанные для оценки фетальной фракции ДНК с использованием секвенирования нового поколения, включают:

- непрямое выведение ее оценки посредством оценки характеристик фетальной/плацентарной вкДНК, отличающихся от характеристик вкДНК материнского происхождения (способ на основе размера внеклеточной ДНК, способ на основе отслеживания нуклеосом внеклеточной ДНК, способ на основе фетальных маркеров метилирования, способ на основе данных секвенирования малой глубины ДНК материнской плазмы)
- прямые оценку и количественное определение генетических вариантов, не присутствующих в материнском фоне (способ на основе хромосомы Y, способ на основе данных секвенирования ДНК материнской плазмы с использованием родительского генотипа, способ на основе данных секвенирования высокой глубины ДНК материнской плазмы, способ на основе данных секвенирования малой глубины ДНК материнской плазмы с использованием материнского генотипа) (Peng XL, Jiang P Bioinformatics Approaches for Fetal DNA Fraction Estimation in Noninvasive Prenatal Testing. Int J Mol Sci. 2017 Feb 20;18(2):453).

С использованием способа на основе данных секвенирования ДНК материнской плазмы с родительским генотипом (в основном, посредством анализа SNP), специфические фетальные аллели в материнской плазме можно легко идентифицировать из прочтений

последовательностей. Даже несмотря на то, что этот способ представляет собой прямой и точный способ оценки фетальной фракции ДНК и в основном рассматривается как золотой стандарт, осуществимости этого способа иногда препятствует необходимость родительских генотипов, поскольку i) только материнские образцы крови собирают, и ДНК материнской плазмы является объектом секвенирования для NIPT в большинстве клинических условий; и ii) не является редким, что генотип биологического отца может не являться доступным на практике.

Чтобы обойти необходимость информации о родительских генотипах, разработан способ измерения фетальной фракции ДНК посредством анализа данных секвенирования ДНК материнской плазмы при высокой глубине с использованием нацеленного массивного параллельного секвенирования. В этом способе, модель смеси биномиальных распределений использовали для установления соответствия наблюдаемых аллельных частот с использованием лежащих в основе четырех типов комбинаций материнских-фетальных генотипов, и фетальную фракцию определяли посредством оценки максимального правдоподобия. Ограничением этого способа может являться то, что необходимо, чтобы глубина секвенирования являлась настолько высокой, как ~120X посредством нацеленного секвенирования для надежного определения фетальных аллелей, что влияет на стоимость тестирования.

Расширенный вариант этого способа недавно был разработан на основе данных секвенирования малой глубины в сочетании с информацией только о материнском генотипе (способ на основе данных секвенирования малой глубины ДНК материнской плазмы с использованием материнского генотипа). Обоснованием этого способа является получение преимущества от того факта, что любой альтернативный аллель (не материнские аллели), присутствующий в локусе SNP, по которому мать является гомозиготной, может теоретически предполагать специфический для фетальной ДНК аллель. Таким образом, выдвинута гипотеза, что фракции таких не материнских аллелей коррелируют с фетальными фракциями ДНК, исходя из предположения, что частоты ошибок, обусловленные платформами секвенирования и генотипирования, являются относительно постоянными среди различных случаев. Однако, параметры в этой модели могут изменяться, в соответствии с платформами секвенирования и генотипирования, поскольку различные платформы характеризуются различными свойствами в отношении ошибок, которые могут вносить вклад в измеренные не материнские аллели. Таким образом, понятно, что с использованием секвенирования малой глубины ДНК материнской плазмы и с использованием только гомозиготных материнских локусов (полученных посредством генотипирования на основе SNParray материнской лейкоцитарной пленки) является проблематичным надежное измерение FF одновременно с детекцией изменчивости фетального количества копий.

Среди наиболее близких документов предшествующего уровня техники, можно процитировать следующие: Sejoon Lee et al., «NGSCheckMate: software for validating sample identity in next-generation sequencing studies within and across data types», Nucleic Acids

Research, 2017, Vol. 45, No. 11, в котором объясняют способ, чтобы убедиться, что наборы данных NGS от одного и того же субъекта являются надлежащим образом парными. В способе NGSCheckMate подтверждают идентичность образцов из файлов FASTQ, BAM или VCF с использованием способа на основе модели для сравнения фракций прочтения аллелей в приблизительно 12000 или 21000 локусах однонуклеотидных полиморфизмов (SNP), принимая во внимание зависимое от глубины поведение метрических показателей сходства для идентичных и неродственных образцов. NGSCheckMate является эффективным для разнообразного множества типов данных, включая экзомное секвенирование, полногеномное секвенирование, РНК-seq, ChIP-seq, нацеленное секвенирование и полногеномное секвенирование отдельных клеток, но задает требование глубины секвенирования >0,5X. Это требование является даже более высоким (>3x) в случае образцов с кровным родством или родительской взаимосвязью. Фактически, когда Sejoon Lee et al. тестировали свой способ на наборе данных, состоящем из 89 профилей WGS отдельных клеток злокачественных опухолей от двух неродственных пациентов с глиобластомой (39 и 50 клеток от каждого пациента), секвенированных с глубиной (0,01-0,3X) для характеризации CNV на уровне отдельных клеток, они достигали только 87,8% точности в группировке клеток, где все ошибки неправильной классификации были обусловлены небольшим количеством клеток с особенно малой глубиной секвенирования (<0,15X).

Полногеномная амплификация из отдельных клеток, и полногеномное секвенирование с низким покрытием

Полногеномная амплификация (WGA) геномной ДНК отдельной клетки часто требуется для получения большего количества ДНК, для упрощения и/или обеспечения возможности различных типов генетических анализов, включая секвенирование, детекцию SNP и т.д. WGA с использованием LM-ПЦР на основе детерминистического участка рестрикции (в дальнейшем DRS-WGA) известна из WO2000/017390.

Показано, что DRS-WGA является лучшим в своем классе способом WGA с многих точек зрения, в частности, в отношении более низкого выпадения аллелей из отдельных клеток (Borgstrom et al., 2017; Normand et al., 2016; Babayan et al., 2016; Binder et al., 2014).

Коммерческий набор для DRS-WGA на основе LM-ПЦР (Ampli1™ WGA, Silicon Biosystems) использовали в Hodgkinson C.L. et al., Nature Medicine 20, 897-903 (2014). В этой работе, проводили анализ количества копий посредством полногеномного секвенирования с низким покрытием на материале WGA отдельной клетки, проводя расщепление адаптеров WGA и фрагментацию до лигирования снабженного штрих-кодом адаптера Illumina для секвенирования.

В WO2017/178655 и WO2019/016401A1 объяснен упрощенный способ получения библиотек для массивного параллельного секвенирования после DRS-WGA (например, Ampli1 WGA) для полногеномного секвенирования с низким покрытием и получения профиля количества копий. В Ferrarini et al., PLoSONE 13(3):e0193689 https://doi.org/10.1371/journal.pone.0193689, эффективность способа из WO2017/178655 с

использованием платформы Ion Torrent подробно описана в отношении получения профиля количества копий.

Показано, что DRS-WGA лучше, чем DOP-ПЦР, для анализа профилей количества копий из незначительных количеств микронарезанного FFPE материала (Stoecklein et al., Am J Pathol. 2002 Jul; 161(1):43-51; Arneson et al., ISRN Oncol. 2012;2012:710692. doi: 10.5402/2012/710692. Epub 2012 Mar 14.), при использовании массива CGH, метафазной CGH, так же как для другого генетического анализа, такого как анализ потери гетерозиготности с использованием нацеленных праймеров и ПЦР для анализа избранных микросателлитов, однако, было показано, что в зависимости от качества FFPE ДНК, LP-WGS отдельных FFPE клеток является возможным, но может становиться непрактичным для более низких показателей качества ДНК (Mangano, C., Ferrarini, A., Forcato, C. *et al.* «Precise detection of genomic imbalances at single-cell resolution reveals intra-patient heterogeneity in Hodgkin's lymphoma». *Blood Cancer J.* **9,** 92 (2019). https://doi.org/10.1038/s41408-019-0256-y).

В общем, существует потребность в предоставлении способа, позволяющего делать заключения об идентичности образца и/или анализировать степень сходства с разрешением вплоть до отдельных клеток, с использованием данных секвенирования с низким покрытием (< 0,15x), преодолевающего одно или несколько из следующих ограничений, присущих современному уровню техники:

- необходимость отдельного исследования для анализа микросателлитов;
- необходимость отдельного исследования для генотипирования SNP;
- покрытие полногеномного секвенирования > 0,5x;
- невозможность надежного повторного анализа отдельной клетки для подтверждения или дополнения целевой геномной информации.

Для криминалистической идентификации отдельных клеток, может являться желательным наличие эффективного способа, для присвоения идентичности каждого из множества образцов отдельных клеток, даже при плохом качестве, и дальнейшего исследования генетических характеристик индивидуума, которому принадлежат указанные образцы.

Для получения полногеномного профиля количества копий для образцов опухолей, включая анализ отдельной клетки, такой как анализ отдельной CTC или отдельных FFPE клеток, может являться желательным предоставление встроенного алгоритма отслеживания образца, чтобы избегать замены образцов при полногеномном секвенировании с низким покрытием, и/или детектировать смешивание различных образцов.

Для неинвазивного пренатального тестирования или диагностики по циркулирующим фетальным клеткам, собранным из материнской крови, может являться желательным наличие эффективного способа анализа, объединяющего в одном анализе (i) получение фетального полногеномного профиля (например, получение полногеномного профиля количества копий) с (ii) способностью подтверждать фетальное происхождение образца.

Для неинвазивного пренатального тестирования на основе циркулирующей фетальной внеклеточной ДНК, смешанной с ДНК материнского происхождения, с использованием полногеномного массивного параллельного секвенирования с низким покрытием, может являться желательным наличие эффективного способа анализа, позволяющего i) идентификацию фетального компонента и оценку его количества, относительно материнского компонента (например: фетальной фракции, FF) и ii) получение полногеномного профиля количества копий в образце по тем же данным секвенирования с низким покрытием.

Для преимплантационного генетического скрининга (PGS; также обозначенного как преимплантационное генетическое тестирование или «PGT») например, на бластоцистах, использованной культуральной среде для эмбриона, может являться желательным наличие способа с использованием одного анализа для детекции и/или количественной оценки контаминации материнскими клетками или экзогенной контаминации, чтобы избегать ложноотрицательных или несогласованных по полу сигналов из анализа, объединяющего возможность (i) получения полногеномного профиля генома эмбриона (например, получения полногеномного профиля количества копий), который можно использовать, например, для подтверждения присутствия или отсутствия анеуплоидии в образце и (ii) количественной оценки и/или определения отсутствия материнской контаминации, по тем же данным секвенирования с низким покрытием.

Для пренатальных образцов (например: ворсин хориона, амниотической жидкости, продуктов зачатия) может являться желательным наличие способа с использованием одного анализа для детекции и/или количественной оценки контаминации материнскими клетками или экзогенной контаминации, чтобы избегать ложноотрицательных или несогласованных по полу сигналов из анализа, объединяющего возможность i) получения фетального полногеномного профиля и (ii) количественной оценки и/или определения отсутствия материнской контаминации, по тем же данным секвенирования с низким покрытием.

В дополнение к этому, может являться желательным наличие способа с использованием одного анализа для детекции в генетическом материале, происходящем из концептуса, в любой фазе эмбриофетального развития, таких состояний, как молярная беременность, многоплодная беременность (включая исчезновение/химеру), однородительская дисомия (изодисомия или гетеродисомия) и ROH (Патент n. WO2021019459A1), классификации кровного родства и ошибки нерасхождения.

Для аутентификации линии клеток, может являться желательным наличие способа с использованием одного анализа для одновременных

(i) идентификации линии клеток с использованием широко доступных секвенаторов для массивного параллельного секвенирования, без необходимости проведения анализа STR в менее доступных устройствах для капиллярного электрофореза, и

(ii) получения полногеномного профиля (например, получения полногеномного профиля количества копий) линии клеток для возможной детекции дрейфов, связанных с

геномной нестабильностью, или артефактов, обусловленных высоким количеством культуральных пассажей.

Для архивных FFPE образцов, где желательна характеризация по отдельной клетке являющегося источником индивидуума, например, в анализе, применительно к эндотелиальным клеткам при аллогенной трансплантации гематопоэтических стволовых клеток, может являться желательным наличие способа, посредством которого можно получать надежные результаты по отдельным клеткам, выделенным из FFPE материалов (сортированных или микронарезанных).

Сущность изобретения

Таким образом, целью настоящего изобретения является предоставление способа, преодолевающего недостатки способов предшествующего уровня техники.

В частности, целью настоящего изобретения является предоставление способа анализа степени сходства по меньшей мере двух образцов в множестве образцов, содержащих геномную ДНК, совместимого с несколькими клетками, вплоть до отдельной клетки, так же как с количествами ДНК, сравнимыми с одним геномным эквивалентом или ниже.

Эта цель достигнута посредством способа, как определено в пункте формулы изобретения 1.

Краткое описание чертежей

На фиг. 1 показано более высокое разрешение между собственными и неродственными образцами с использованием способа по настоящему изобретению, включающего DRS-WGA с последующей реакцией ПЦР с адаптером для секвенирования/праймером для слияния WGA, без фрагментации, относительно получения библиотеки со случайной фрагментацией, известного в данной области.

На фиг. 2 показан эффект увеличения количества локусов до 300000 полиморфных локусов, на основании наиболее высокой гетерозиготности - в соответствии с настоящим изобретением - против выбора NGScheckMate из 21000 SNP: дискриминационная мощность увеличивается.

На фиг. 3A и 3B показано распределение показателей сходства парных образцов, принадлежащих одним и тем же (собственные) или различным (неродственные) индивидуумам (с использованием линий клеток), рассчитанное с использованием различных способов в соответствии с настоящим изобретением. На фиг. 3A корреляцию используют в качестве способа определения расстояний (стандартный способ NGSCheckMate). На фиг. 3B согласованность используют для оценки сходства образцов. Подробно: - если запрашиваемые аллели являются одинаковыми, добавляют 1 к оценке в баллах; - если запрашиваемые аллели перекрываются частично (например, если один образец имеет 2 аллеля, а другой только 1), добавляют 0,5; - если запрашиваемые аллели являются различными, добавляют 0 к оценке в баллах. Оценку в баллах затем делят на количество аллелей, покрытых в обоих сравниваемых образцах.

На фиг. 4A - 4C и 4D - 4F показана взаимосвязь между параметрами, такими как

минимальная средняя гетерозиготность, количество прочтений, и полученным в результате разделением между собственными и неродственными образцами.

На фиг. 5A - 5D показана эффективность классификации родственных образцов, в отношении собственных образцов родителя женского пола и неродственных образцов, для количества прочтений, равного 500000 на образец.

На фиг. 6 показано распределение показателей попарного сходства, рассчитанных как согласованность, в отношении образцов от родителя женского пола, для собственных (родитель женского пола), родственных и неродственных образцов, как функция от минимальной средней гетерозиготности (диапазон=0,2-0,498).

На фиг. 7 показано схематическое изображение способа детекции двуплодной беременности. Все попарные прогнозы фетальных клеток, описанных посредством «родственной» связи с материнским контролем, используют в качестве ввода в алгоритм кластеризации графов для нахождения «сообществ» фетальных клеток.

На фиг. 8 показано распределение средних показателей попарного сходства, рассчитанных в отношении образцов от родителя женского пола, в выделенных препаратах клеток эритробластов, выделенных из периферической крови из двух отдельных материнских образцов.

На фиг. 9A - 9C показана классификация на основе кластеризации выделенных препаратов клеток из образца BO1368. Коэффициент силуэта 2 смешанных клеток является намного более низким, чем коэффициент силуэта фетальных клеток, и может быть использован для установления их отличий от фетальных и создания нового кластера со смешанными образцами.

На фиг. 10A - 10C показана классификация на основе кластеризации выделенных препаратов клеток из образца BO1383.

На фиг. 11A и 11B показана эффективность классификации индивидуальных образцов, в отношении неродственных образцов с максимум 50% компонента собственных образцов. Фиг. 11A представляет собой график «типа ROC» с TPR и 1-PPV для родственного класса как функциями от порогового значения «согласия». На фиг. 11B показаны TPR и PPV при различной средн. гетерозиг. Порог (серый) был установлен, чтобы иметь по меньшей мере ppv 99,9%. Порог показан серым на вторичной оси y.

На фиг. 12 показано распределение показателей попарного сходства (согласованность), рассчитанных для парных образцов с различной степенью контаминации от различных индивидуумов.

На фиг. 13A - 13C показана классификация отдельных выделенных препаратов клеток из FFPE образцов, в соответствии с идентичностью индивидуумов. FFPE образцы (лимфома) от 4 пациентов. Подвыборка 500000 прочтений. Согласие основано на согласованности. Сравнения отмечены как высок. DLRS (x-ось), если один или оба члена имели DLRS > 0,4, и низк. DLRS, если оба члена имели DLRS≤0,4. На фиг. 13C показано, что кластеризация правильно приписывает все FFPE образцы 4 различным кластерам, соответствующим 4 индивидуумам.

На фиг. 14 показано моделирование *in silico* свободных от клеток использованных культуральных сред с различной степенью контаминации материнской ДНК от 0 (100% фетальные) до 90% (10% фетальные), и связанный показатель сходства. В частности, на фигуре показана эмуляция, проводимая посредством смешивания *in silico* различных долей последовательностей ДНК из отдельных фетальных клеток с последовательностями из материнских клеток. Сплошная линия соответствует среднему показателю попарного сходства при различных процентах ввода фетальных. Закрашенная область соответствует 95% доверительному интервалу. Пунктирной линией показан пример смешанного образца с известным % материнского компонента (80%) и показателем попарного сходства с материнским эталоном=0,807, который, в соответствии с моделью, имеет средний прогнозируемый фетальный компонент=27,7% (C.I.=25,4%-30,7%), соответствующий оцененной контаминации материнской ДНК ≈ 75%.

На фиг. 15A и 15B показан эффект компенсации для контаминации в полногеномном анализе количества копий смешанного образца. В частности, фигура представляет полногеномный анализ количества копий смешанного образца, полученного посредством смешивания *in silico* различных долей последовательностей ДНК из отдельных фетальных клеток (20%) с последовательностями из материнских клеток (80%). На фиг. 15A показан полногеномный профиль количества копий; каждая точка соответствует геномному интервалу 10 млн.п.о. На фиг. 15B показано полногеномное количество копий после применения поправочного коэффициента=0,75, на основе оцененного процента контаминации материнской ДНК, на основе показателя попарного сходства с материнским эталоном. Статистически значимые отклонения показаны как сплошные черные линии.

Определения

Если не определено иное, все технические и научные термины, используемые в настоящем описании, имеют такое же значение, которое является общепринятым для специалиста в области, к которой относится настоящее изобретение. Несмотря на то, что многие способы и материалы, сходные или эквивалентные описанным в настоящем описании, можно использовать в практическом осуществлении или тестировании настоящего изобретения, предпочтительные способы и материалы описаны ниже. Если не упомянуто иное, способы, описанные в настоящем описании для применения по настоящему изобретению, представляют собой стандартные способы, хорошо известные специалисту в данной области.

Под выражением «массивное параллельное секвенирование нового поколения (NGS или MPS)» в настоящем описании понимают способ секвенирования ДНК, включающий получение библиотеки молекул ДНК, разделенных в пространстве и/или во времени, клонально секвенированных (в присутствии или в отсутствие предшествующей клональной амплификации). Примеры включают платформу Illumina (Illumina Inc), платформу Ion Torrent (Thermo Fisher Scientific Inc), платформу Pacific Biosciences, MinIon (Oxford Nanopore Technologies Ltd).

Под выражением «полногеномное секвенирование с низким покрытием» в

настоящем описании понимают полногеномное секвенирование при средней глубине секвенирования ниже чем 1x, в отношении полного эталонного генома, библиотеки для массивного параллельного секвенирования, которая не была обогащена по специфическим для последовательности фрагментам. Это определение явно исключает случай основанного на ПЦР нацеленного обогащения или специфического для последовательности обогащения с использованием ловли-наживок по набору локусов, например, таких как локусы однонуклеотидных полиморфизмов (SNP) и/или короткие тандемные повторы (STR).

Под выражением «средняя глубина секвенирования» в настоящем описании понимают, на основании образцов, общее количество секвенированных оснований, картированных на эталонном геноме, деленное на общий размер эталонного генома. Общее количество секвенированных и картированных оснований можно аппроксимировать по количеству раз картированных прочтений средней длины прочтения.

Под выражением «эталонный геном» в настоящем описании понимают эталонную последовательность ДНК для специфического вида.

Под термином «локус» (множественное «локусы») в настоящем описании понимают фиксированное положение на хромосоме (в отношении эталонного генома).

Под выражением «полиморфный локус» в настоящем описании понимают локус, имеющий 2 или более аллелей с наблюдаемой частотой, большей, чем 1%, в популяции.

Под выражением «гетерозиготный локус» в настоящем описании понимают локус, имеющий 2 или более аллелей, наблюдаемых в конкретном образце.

Под выражением «средняя гетерозиготность» для локуса в настоящем описании понимают значение 1 минус сумма квадратов частот аллелей. В частности, произведение 2pq, где p и q=(1-p) представляют собой частоты аллелей для локуса, в случае локусов с двумя аллелями в популяции, или сумма произведений 2pq+2pr+2qr, где p, q и r (с p+q+r=1) представляют собой три частоты аллелей для локуса с тремя возможными аллелями.

Под выражением «покрытый геном» в настоящем описании понимают часть эталонного генома, покрытую по меньшей мере одним прочтением.

Под термином «прочтение» в настоящем описании понимают фрагмент ДНК, секвенированный («прочитанный») посредством секвенатора.

Под выражением «кратность уменьшения» в настоящем описании понимают общее количество оснований фрагментов, полученных посредством расщепления *in silico* эталонного генома, в соответствии с рестрикционным ферментом, используемым в DRS-WGA, содержащихся в указанном диапазоне пар оснований, деленное на общее количество оснований в эталонном геноме.

Под выражением «аллельное содержание» в настоящем описании понимают состав, в отношении аллелей, детектированных в локусе.

Под выражением получение библиотеки для массивного параллельного секвенирования «и реакция ПЦР с адаптером для секвенирования/праймером для слияния WGA, без фрагментации», в настоящем описании понимают получение библиотеки для массивного параллельного секвенирования из продуктов DRS-WGA, без стадий

фрагментации ДНК, в результате чего адаптеры для секвенирования добавляют к продукту WGA посредством праймеров для слияния, например, в соответствии с патентными заявками (WO2017/178655) или (WO2019/016401A1).

Под выражением «показатель попарного сходства», в настоящем описании понимают функцию от множества парных вводов с конечным кодоменом. Кодомен, предпочтительно, нормализуют на стандартное значение, например, [-1;1] или [0;1], независимо от количества парных вводов.

Под выражением «кластеризация образцов», в настоящем описании понимают алгоритм для разбиения образцов таким образом, что образцы, принадлежащие к одному и тому же разделу (называемому также «кластер»), разделяют общее свойство, выбранное из группы, состоящей из идентичности одного индивидуума (или более индивидуумов), вносящих значительный вклад ДНК в образцы из этого раздела, свойство содержания недостаточных количеств ДНК и свойство содержания высоко деградированной ДНК или ДНК непонятного происхождения.

В данной области известны несколько метрических показателей оценки эффективности алгоритмов кластеризации, когда реальная ситуация является неизвестной, такие как «коэффициент силуэта», «индекс Калинского-Харабаша», «индекс Дэвиса-Булдина», которые можно использовать для определения «оптимального» количества кластеров для разбиения множества образцов на гомогенные, хорошо определенные кластеры.

Под выражением «кластер идентичности» в настоящем описании понимают группу, состоящую из образцов, содержащих, с высокой вероятностью, ДНК только от одного и того же индивидуума. Значение высокой вероятности (далее в настоящем описании Вер.[Отд.-ID]) зависит от применения, как понятно специалисту в данной области, и его определяют, применительно к специфике применения и его требований эффективности. Например, в случае анализа фетальных клеток, принимают, что диагноз выдают, только когда по меньшей мере три отдельных «предположительных» фетальных клетки (т.е., принадлежащих к кластеру идентичности клеток, которые находятся в родственной взаимосвязи с материнским эталоном) индивидуально анализируют и регистрируют. Диагностика, например, анеуплоидии с использованием выведенного из WGS с низким покрытием профиля количества копий, может быть нарушена, если ни одна из клеток не происходит из пораженного плода, и все анализируемые клетки представляют собой материнские клетки, по ошибке принятые за фетальные. Кроме того, устанавливают в качестве приемлемой минимальную чувствительность (Чувств._мин.) для детекции анеуплоидного плода. Последующая вероятность названия нормальным анеуплоидного плода, вызванная ошибочным наименованием идентичности каждой из отдельных клеток, может требовать, чтобы все клетки, на которых основана диагностика, были названы фетальными вместо материнских. Как правило, является целесообразным принимать, что эти события (попарное сравнение с материнским эталоном) являются независимыми среди предположительных фетальных клеток, таким образом Вер.[Ложн._ID из анализируемых N

клеток]=Вер.[Ложн._ID]$^{N\ клеток}$, где N клеток представляет собой количество клеток, индивидуально анализируемых, где Вер.[Ложн._ID]=1-Вер.[Отд._ID] представляет собой вероятность ошибки в наименовании образца, как принадлежащего к кластеру идентичности одного и того же индивидуума (более конкретно, кластеру образцов в родственной взаимосвязи с материнским эталоном, как указано выше). Желательно, чтобы

(1-Вер.[Отд._ID])$^{N\ клеток}$ ≤(1-Чувств._мин.), т.е.

Вер.[Отд._ID]≥1-(1-Чувств._мин.)$^{1/N\ клеток}$

Например, при Чувств._мин.=99,9%, N клеток=5 требует Вер.[Отд._ID]≥75%,

В то время как принятие N клеток=3 требует

Вер.[Отд._ID]≥90%

В обоих случаях исключая, с целью упрощения, другие источники ошибки, подобные вероятности того, что истинно фетальную клетку действительно анализируют, но терпят неудачу в детекции анеуплоидии.

В случае криминалистических экспертизы и исследования не служащих уликами образцов, значение высокой вероятности может быть различным. Например, способ в соответствии с настоящим изобретением можно использовать для реконструкции профиля STR из количества N клеток из индивидуальных клеток. В зависимости от допустимой строгости поиска в базе данных ДНК, количества анализируемых отдельных клеток, среднего уровня определения STR для каждого индивидуального образца из экспертизы, различные требования могут возникать для точного значения высокой вероятности (Вер.[Отд._ID]) для достижения целей. Это требование более сложно моделировать аналитически, и его можно выводить, например, путем моделирований способом Монте-Карло посредством использования доступных баз данных и моделирования *in silico* различной степени выпадений аллелей, количества отдельных клеток, фактически анализируемых, и алгоритмических выборов в реконструкции профиля.

Под выражением «образец WGA-ДНК отдельного индивидуума», в настоящем описании понимают образец, содержащий смесь продуктов DRS-WGA, полученных из образцов, содержащих ДНК от отдельного индивидуума.

Под выражением «неинвазивное пренатальное тестирование» в настоящем описании понимают проведение генетических анализов для оценки фетальной внеклеточной ДНК или интактных фетальных клеток, циркулирующих в материнской крови.

Под выражением «преимплантационные генетическое тестирование/скрининг» в настоящем описании понимают проведение генетических анализов для оценки эмбрионов до переноса в матку посредством полногеномного анализа, например, отклонений количества копий для определения присутствия анеуплоидии (либо слишком большого, либо слишком малого количества хромосом), в развивающемся эмбрионе.

Под выражением «преимплантационная генетическая диагностика» в настоящем описании понимают преимплантационное генетическое тестирование посредством нацеленного секвенирования для анализа присутствия вариантов последовательности в развивающемся эмбрионе, например, таких как мутации, сцепленные с моногенными

нарушениями (например, болезнью Хантингтона, кистозным фиброзом, синдромом ломкой Х-хромосомы), включая те, которые являются аутосомно-доминантными и рецессивными или Х-сцепленными, или наследственными раковыми синдромами (например, наследственным раком молочной железы и яичника, синдромом Линча). Кроме того, этот термин предназначен для секвенирования для идентификации совместимых по человеческому лейкоцитарному антигену, неповрежденных эмбрионов, вынашиваемых с целью обеспечения больных членов семьи получением совместимых трансплантатов костного мозга или трансфузий пуповинной крови.

Под выражением «эмбриональный образец», в настоящем описании понимают образец, содержащий ДНК из эмбриона, например, такой как бластоциста, использованная культуральная среда для эмбриона, полярное тельце.

Под выражением «данные WGA-ДНК отдельного индивидуума» в настоящем описании понимают данные, полученные посредством слияния данных секвенирования, полученных от образцов, содержащих DRS-WGA ДНК от отдельного индивидуума.

С целью упрощения описания применений способа в соответствии с настоящим изобретением в пренатальной и репродуктивной медицине, термин «материнский» можно использовать для распространения его значения на «принадлежащий женщине» или «принадлежащий родителю женского пола», и «мать» для распространения на «женщину» или «родителя женского пола», применительно к индивидууму женского пола, предоставившему яйцеклетку для эмбриона, плода из текущей беременности, хотя эта женщина могла еще не стать матерью в результате рождения потомства, соответствующего указанному эмбриону или плоду, и т.д.

Подобным образом, термин «отцовский» можно использовать для распространения его значения на «принадлежащий мужчине» или «принадлежащий родителю мужского пола», и «отец» для распространения на «мужчину» или «родителя мужского пола», применительно к индивидууму мужского пола, предоставившему сперму для эмбриона, плода из текущей беременности, пузырного заноса, хотя этот мужчина мог еще не стать отцом в результате рождения женщиной потомства, соответствующего указанному эмбриону или плоду, и т.д.

Подробное описание

Способ в соответствии с настоящим изобретением применяют к анализу множества образцов, содержащих геномную ДНК. В частности, способ предназначен для анализа степени сходства по меньшей мере двух образцов в множестве образцов, содержащих геномную ДНК. В конкретных вариантах осуществления, вид образцов представляет собой *Homo Sapiens*, и если не отмечено иное, на этот вид ссылаются в остальной части описания, без ограничения применимости для других видов, когда это применимо.

Способ включает следующие стадии.

На стадии a), получают множество образцов, содержащих геномную ДНК.

На стадии b), полногеномную амплификацию с детерминистическим участком рестрикции (DRS-WGA) указанной геномной ДНК проводят отдельно для каждого образца.

На стадии c), библиотеку для массивного параллельного секвенирования получают из каждого продукта указанной DRS-WGA с использованием реакции ПЦР с адаптером для секвенирования/праймером для слияния WGA, без фрагментации.

На стадии d), полногеномное секвенирование с низким покрытием проводят при средней глубине покрытия < 1x для указанной библиотеки для массивного параллельного секвенирования. Среднее покрытие составляет, предпочтительно, 0,01x, предпочтительно, при покрытии < 0,05x, более предпочтительно, при покрытии < 0,1x, даже более предпочтительно, при покрытии < 0,5x. Это позволяет уменьшение стоимости секвенирования, с сохранением в то же время хороших результатов анализа, в отношении применения.

На стадии e), прочтения, полученные на стадии d), выравнивают на эталонном геноме.

На стадии f), аллельное содержание в множестве полиморфных локусов извлекают для каждого образца, т.е., получают из выровненных прочтений. Указанное множество локусов содержит полиморфные локусы для рассматриваемого вида.

Указанное множество полиморфных локусов предпочтительно содержит полиморфные локусы со средней гетерозиготностью > 0,499, более предпочтительно, со средней гетерозиготностью > 0,49, даже более предпочтительно, со средней гетерозиготностью > 0,4, даже более предпочтительно, со средней гетерозиготностью > 0,3, наиболее предпочтительно, со средней гетерозиготностью > 0,2.

Указанное множество полиморфных локусов предпочтительно содержит > 200000 локусов, более предпочтительно, > 300000 локусов, даже более предпочтительно, > 500000 локусов, наиболее предпочтительно, > 1000000 локусов.

На стадии g), показатель попарного сходства для по меньшей мере двух образцов рассчитывают, как функцию от аллельного содержания, измеренного в указанном множестве локусов.

На стадии h), степень сходства по меньшей мере двух образцов определяют на основании показателя сходства.

Как правило, сходство можно измерять на основании согласованности аллельного содержания в общих полиморфных локусах, где слово «общие» означает, что локусы покрыты по меньшей мере одним прочтением ДНК образцов в паре или наборе из по меньшей мере двух образцов. Например, показатель попарного сходства, предпочтительно, рассчитывают посредством вычисления корреляции частоты B-аллеля среди локусов, покрытых по меньшей мере одним прочтением в по меньшей мере двух образцах.

В качестве альтернативы, показатель попарного сходства, предпочтительно, рассчитывают посредством вычисления среднего показателя согласованности среди локусов, покрытых по меньшей мере одним прочтением в обоих парных образцах, где показателю согласованности для каждого локуса приписывают одно из следующих значений:

a) 1, если запрашиваемые аллели являются идентичными;

b) 0, если запрашиваемые аллели являются различными или полностью различными;

c) 0,5, если запрашиваемые аллели являются частично перекрывающимися.

Например, в некоторых вариантах осуществления, показателю согласованности для каждого локуса можно приписать:

A1) 1, если запрашиваемые аллели являются идентичными; и

B1) 0, если запрашиваемые аллели являются различными. Альтернативно, в некоторых вариантах осуществления, показателю согласованности для каждого локуса можно приписать:

A2) 1, если запрашиваемые аллели являются идентичными;

B2) 0, если запрашиваемые аллели являются полностью различными; и

C2) 0,5, если запрашиваемые аллели являются частично перекрывающимися.

Для целей по настоящему изобретению, способы, описанные в настоящем описании, можно использовать для спаривания образцов (например, образцов отдельных клеток, образцов внеклеточной ДНК и т.д.) для измерения степени «сходства» между образцами. Включение в набор образцов (т.е., «по меньшей мере два образца») контрольного образца, такого как материнский/отцовский образец, в случае анализа NIPT или установления отцовства, соответственно, может позволять улучшенное установление отличий между образцами, такими как материнские/отцовские и фетальные клетки.

Способ в соответствии с настоящим изобретением, предпочтительно, дополнительно включает стадию определения группы кластеров образцов, разделяющих общее свойство, такое как идентичность одного индивидуума (или более индивидуумов), вносящих значительный вклад ДНК в образцы из кластера, или свойство содержания недостаточных количеств ДНК, и/или свойство содержания высоко деградированной ДНК или ДНК непонятного происхождения.

В другом предпочтительном варианте осуществления, алгоритм кластеризации (например, иерархической кластеризации) можно осуществлять для нахождения указанных кластеров с использованием индивидуальных образцов (например, отдельных клеток). Этот тип анализа может являться наиболее подходящим для установления отличий групп образцов, где один из образцов представляет собой эталонный образец, используемый для идентификации эталонного кластера. Например, в анализах NIPT, пулы материнских клеток можно использовать в качестве эталона для установления отличий других групп клеток, таких как фетальные клетки, у беременных женщин, с использованием показателя сходства, как описано в настоящем описании. Способы кластеризации в общем, и НС конкретно, можно осуществлять, включая итерационный способ для нахождения наиболее правильного количества кластеров, показатель качества (например, коэффициент силуэта) для выбора наилучшего кластерного разбиения и способ идентификации смешанных выделенных препаратов (например, образцов, принадлежащих к большему количеству кластеров) и, в случае анализа NIPT, множества плодов.

Предпочтительно, по меньшей мере два образца приписывают по меньшей мере одному кластеру посредством классификатора с использованием, в качестве ввода,

указанного показателя попарного сходства. Как более подробно описано ниже, классификатор можно использовать независимо от анализа кластеризации.

В предпочтительном варианте осуществления, определение количества указанных кластеров проводят посредством проведения агломеративной кластеризации показателя попарного сходства. В предпочтительном варианте осуществления, такую агломеративную кластеризацию проводят с использованием евклидова расстояния и связи Варда. В предпочтительном варианте осуществления, такую кластеризацию проводят с использованием диапазона количества кластеров, приводящего к различным альтернативным выходам кластеризации.

В предпочтительном варианте осуществления, такие альтернативные выходы кластеризации оценивают посредством расчета коэффициента силуэта и выбирают кластеризацию с наиболее высоким усредненным коэффициентом силуэта среди всех подкластеров.

Предпочтительно, в указанном классификаторе используют, в качестве дополнительного ввода, по меньшей мере одно значение, измеренное по указанным данным полногеномного секвенирования с низким покрытием, выбранное из группы, содержащей:

a) DLRS: производное логарифма отношения разброса;

b) R50: процент фрагментов WGA, покрытых посредством 50% секвенированных прочтений, среди всех фрагментов WGA, покрытых по меньшей мере одним прочтением;

c) YFRAC: доля прочтений, картированных на хромосоме Y;

d) Аберрантный: процент генома, соответствующий добавлениям или потерям, относительно медианной плоидности клеток;

e) Хр. 13: плоидность хромосомы 13;

f) Хр. 18: плоидность хромосомы 18;

g) Хр. 21: плоидность хромосомы 21;

h) RSUM: среднее абсолютное отклонение от ближайшего целочисленного уровня количества копий, рассчитанное для события аберрации количества копий с наивысшим абсолютным отклонением от медианной плоидности клеток;

i) Смеш._показатель: z-показатель по RSUM, рассчитанный для события аберрации количества копий с наивысшим абсолютным отклонением от медианной плоидности клеток; и

j) Дегр._показатель: количество событий небольшой потери (< 10 млн.п.о., которая является распространенной в деградированных образцах).

Количество указанных кластеров, предпочтительно, рассчитывают посредством

a) выбора количества кластеров после первой итерации, максимизирующего средний коэффициент силуэта;

b) для каждого из указанных кластеров после первой итерации, вычисления коэффициента силуэта каждого из указанных образцов, принадлежащих к кластеру после первой итерации, где образцы, принадлежащие к кластеру, имеющие коэффициент силуэта ниже, чем фиксированный порог, лежащий в диапазоне 0,19-0,21, приписывают новому

кластеру.

В предпочтительном варианте осуществления, указанная группа кластеров, предпочтительно, содержит один или несколько кластеров идентичности, содержащих образцы, содержащие, с высоким уровнем доверия, ДНК только от одного и того же индивидуума.

В присутствии большего количества кластеров идентичности, кардинальность указанного множества кластеров идентичности, предпочтительно, соответствует количеству индивидуальных вносящих вклад в ДНК участников в указанном множестве образцов.

Предпочтительно, способ дополнительно включает определение группы кластеров смешанной идентичности, где каждый из указанных кластеров смешанной идентичности содержит образцы, содержащие ДНК от по меньшей мере двух индивидуумов.

Предпочтительно, способ дополнительно включает определение по меньшей мере одного кластера без распознавания, содержащего образцы, содержащие ДНК непонятного происхождения.

Обеспечивающим преимущество образом, этот кластер включает образцы, где количество локусов, оцененное для расчета показателя сходства, ниже, чем порог. Обеспечивающим преимущество образом, указанный порог устанавливают, принимая во внимание один или несколько элементов, выбранных из группы, содержащей:

1. количество прочтений образца,

2. минимальную среднюю гетерозиготность в локусах, используемых для сравнения.

Множество образцов, предпочтительно, содержит по меньшей мере один эталонный образец, и указанная группа кластеров идентичности включает по меньшей мере один эталонный кластер, содержащий указанный эталонный образец.

Предпочтительно, классификатор можно использовать независимо от анализа кластеризации для присвоения образца, в паре, правильному классу, используя, в качестве основного ввода, указанный показатель попарного сходства, и принимая, что по меньшей мере один из двух парных образцов представляет собой эталонный образец. Кроме того, машинообучаемый классификатор может использовать дополнительные признаки для получения наивысшего возможного уровня доверия. Для целей по настоящему изобретению, понятно, что классификатор не обязательно приписывает образец кластеру, но вместо этого приписывает образец одному из нескольких предопределенных классов. Таким образом, является возможным классифицировать образец без его кластеризации. И наоборот, способами неконтролируемой кластеризации можно находить сходство между образцами, без предварительных определений классов.

В предпочтительном варианте осуществления машинообучаемый классификатор (например, типа случайного леса) можно осуществлять и обучать с использованием подходящего обучающего набора для установления отличий образцов. Такой классификатор может использовать, среди других признаков, указанный показатель попарного сходства. Этот способ может наилучшим образом подходить для попарных

сравнений, где отдельный тестируемый образец необходимо оценивать против эталонного образца. Пример может представлять собой способ, целью которого является классификация отдельной клетки с использованием пула клеток известного происхождения в качестве контроля (например, пула материнских клеток в качестве контроля). В случае неинвазивного пренатального тестирования на основе клеток, при установлении отличий между материнскими и фетальными клетками, ожидаемые классы могут представлять собой (i) «собственные» для материнских клеток, (ii) «родственные» для фетальных клеток, (iii) «смешанные» для выделенных препаратов, содержащих смесь фетальных и материнских клеток, (iv) «неродственные» для образцов, не родственных для матери или плода (т.е. экзогенной контаминации, яйцеклетки донора при беременности с ЭКО и т.д.), и «без распознавания» для ненадежных образцов с плохими метрическими показателями. Классификатор, такой как классификатор случайного леса, может устанавливать отличия образцов с использованием, в дополнение к указанному показателю попарного сходства, ввода из по меньшей мере одного признака, измеренного с использованием данных полногеномного секвенирования с низким покрытием, включая, но без ограничения,:

a) DLRS: производное логарифма отношения разброса;

b) R50: процент фрагментов WGA, покрытых посредством 50% секвенированных прочтений, среди всех фрагментов WGA, покрытых по меньшей мере одним прочтением;

c) YFRAC: доля прочтений, картированных на хромосоме Y;

d) Аберрантный: процент генома, соответствующий добавлениям или потерям, относительно медианной плоидности клеток;

e) Хр. 13: плоидность хромосомы 13;

f) Хр. 18: плоидность хромосомы 18;

g) Хр. 21: плоидность хромосомы 21;

h) RSUM: среднее абсолютное отклонение от ближайшего целочисленного уровня количества копий, рассчитанное для события аберрации количества копий с наивысшим абсолютным отклонением от медианной плоидности клеток;

i) Смеш._-показатель: z-показатель по RSUM, рассчитанный для события аберрации количества копий с наивысшим абсолютным отклонением от медианной плоидности клеток; и

j) Дегр._-показатель: количество событий небольшой потери (< 10 млн.п.о., которая является распространенной в деградированных образцах).

Кроме того, другие типы классификаторов, которые являются подходящими для описанных способов, могут быть основаны, например, на предопределенных фиксированных порогах указанного показателя попарного сходства, описывающих «родственные», «собственные» или «неродственные» взаимосвязи (т.е., пример 6).

В некоторых вариантах осуществления, способы кластеризации (например, иерархическую кластеризацию) и способы с использованием классификатора (например, классификатора RF) можно использовать взаимозаменяемо для установления отличий образцов на основании данных прочтения последовательности, принимая во внимание, что

способом с использованием классификатора сравнивают тестируемый образец против эталонного образца, в то время как целью способов кластеризации является нахождение групп/кластеров образцов, в которых один из них идентифицирует эталонный кластер.

В предпочтительном варианте осуществления, указанный по меньшей мере один эталонный образец представляет собой образец от беременного индивидуума - родителя женского пола.

Указанная группа кластеров идентичности, предпочтительно, дополнительно содержит по меньшей мере один родственный кластер, состоящий из образцов из по меньшей мере одного плода из текущей беременности указанного индивидуума - родителя женского пола.

Предпочтительно, указанный родственный кластер разбивают на множество фетальных кластеров, состоящих из образцов, содержащих ДНК только от одного и того же плода.

В альтернативном предпочтительном варианте осуществления, указанный по меньшей мере один эталонный кластер, предпочтительно, состоит из образцов, содержащих ДНК только от одного и того же индивидуума, соответствующего потерпевшему в криминалистической экспертизе, дополнительно включающей определение по меньшей мере одного кластера виновного, содержащего образцы, содержащие ДНК только от одного и того же индивидуума, отличного от потерпевшего.

В этом случае, способ в соответствии с настоящим изобретением, предпочтительно, включает смешивание по кластерам аликвот DRS-WGA из множества образцов, принадлежащих к каждому из указанных по меньшей мере одного из кластеров виновных, с получением для каждого кластера соответствующего образца WGA-ДНК отдельного индивидуума, и проведением дополнительного анализа ДНК для по меньшей мере одного из указанных образцов WGA-ДНК отдельного индивидуума.

Способ, предпочтительно, включает слияние по кластерам данных генетического анализа из по меньшей мере одного типа анализа, из множества образцов, принадлежащих к каждому из указанных по меньшей мере одного из кластеров виновных, с получением для каждого из указанных по меньшей мере одного из кластеров виновных соответствующих данных WGA-ДНК отдельного индивидуума.

Тип анализа выбран из группы, состоящей из анализа микросателлитов, анализа однонуклеотидного полиморфизма, массивного параллельного нацеленного секвенирования и полногеномного секвенирования.

В одном предпочтительном варианте осуществления способа по настоящему изобретению, множество образцов содержит образцы опухоли и/или нормальные образцы.

В другом предпочтительном варианте осуществления, множество образцов содержит по меньшей мере эталонный образец, содержащий ДНК от индивидуума - родителя женского пола, и по меньшей мере один другой эмбриональный образец из указанного множества образцов выбран из группы, состоящей из:

а) образцов, содержащих ДНК из эмбриона, происходящего от указанного

индивидуума - родителя женского пола; и

b) образцов, содержащих ДНК из использованной культуральной среды для эмбриона, полученной от эмбриона от указанного индивидуума - родителя женского пола.

В последнем варианте осуществления, способ, предпочтительно, дополнительно включает проведение преимплантационного генетического скрининга для указанного эмбриона посредством полногеномного анализа хромосомных аберраций по данным указанного полногеномного секвенирования с низким покрытием для указанного по меньшей мере одного другого эмбрионального образца с использованием коэффициента контаминации, соответствующего материнской контаминации, измеренного для указанного по меньшей мере одного другого эмбрионального образца как функция от указанного попарного сходства указанного по меньшей мере одного другого эмбрионального образца и указанного образца от индивидуума - родителя женского пола.

В другом предпочтительном варианте осуществления, множество образцов содержит по меньшей мере эталонный образец, содержащий ДНК от индивидуума - родителя женского пола, и по меньшей мере один другой образец, содержащий ДНК из образца внеклеточной ДНК. В некоторых вариантах осуществления, способ, предпочтительно, дополнительно включает проведение неинвазивного пренатального тестирования для указанного образца внеклеточной ДНК посредством полногеномного анализа хромосомных аберраций по данным указанного полногеномного секвенирования с низким покрытием для указанного по меньшей мере одного образца внеклеточной ДНК с использованием поправочного коэффициента, соответствующего фетальной фракции, измеренного для указанного по меньшей мере одного образца внеклеточной ДНК как функция от указанного попарного сходства.

В другом предпочтительном варианте осуществления, множество образцов содержит по меньшей мере эталонный образец, содержащий ДНК от индивидуума - родителя женского пола, и по меньшей мере один другой пренатальный образец, содержащий ДНК из ворсин хориона, амниотической жидкости или продуктов зачатия. В некоторых вариантах осуществления, способ, предпочтительно, дополнительно включает проведение анализа пренатального тестирования для указанных пренатальных образцов посредством полногеномного анализа хромосомных аберраций по данным указанного полногеномного секвенирования с низким покрытием для указанного по меньшей мере одного пренатального образца, с использованием поправочного коэффициента, соответствующего материнской или экзогенной контаминации, измеренного для указанного по меньшей мере одного пренатального образца как функция от указанного попарного сходства.

В частности, для аутентификации линии клеток, предпочтительно, множество эталонных кластеров получают из множества образцов ДНК из линий клеток, и указанная группа кластеров идентичности дополнительно содержит по меньшей мере один из образцов из линии клеток, подлежащей аутентификации.

В частности, для исследования аллотрансплантатов, предпочтительно, указанный по

меньшей мере один эталонный кластер состоит из образцов, содержащих ДНК зародышевой линии от подвергаемого трансплантации пациента, и указанная группа кластеров идентичности дополнительно содержит один кластер донора, состоящий из образцов от аллогенного донора для указанного подвергаемого трансплантации пациента.

В частности, для неинвазивного установления отцовства, предпочтительно, указанный по меньшей мере один эталонный образец содержит эталонный образец от родителя мужского пола, содержащий ДНК только от указанного родителя мужского пола, и указанный по меньшей мере один эталонный кластер дополнительно содержит кластер идентичности родителя мужского пола, включающий указанный образец от родителя мужского пола, и:

(i) если показатель сходства родственного образца, в отношении образца от родителя мужского пола, согласуется с кровным родством, отцовство подтверждают;

(ii) если показатель сходства родственного образца, в отношении образца от родителя мужского пола, согласуется с неродственным индивидуумом, отцовство не подтверждают.

В частности, для неинвазивной оценки молярной беременности, предпочтительно, указанный по меньшей мере один образец содержит по меньшей мере один образец циркулирующих трофобластных клеток и, если показатель сходства указанного образца трофобластных клеток, в отношении образцов от родителя женского пола, согласуется с неродственными образцами, подтверждают полный занос.

В последнем варианте осуществления, указанный по меньшей мере один образец, предпочтительно, содержит множество образцов трофобластных клеток и:

(i) если показатель сходства среди указанных образцов трофобластных клеток превышает ожидаемый 99-й процентиль ожидаемого показателя сходства для собственных образцов, подтверждают гомозиготный отцовский занос P1P1.

(ii) если показатель сходства среди указанных образцов трофобластных клеток согласуется с ожидаемым показателем сходства для собственных образцов, подтверждают гетерозиготный отцовский занос P1P2.

Предпочтительно, указанный по меньшей мере один образец дополнительно содержит образец от родителя мужского пола, и показатель сходства среди указанных образцов трофобластных клеток согласуется с ожидаемым показателем сходства для собственных образцов, и:

(i) если показатель сходства указанных образцов трофобластных клеток, в отношении образца от родителя мужского пола, согласуется с ожидаемым показателем сходства для собственных образцов, подтверждают гетерозиготный отцовский занос P1P2.

(ii) если показатель сходства указанных образцов трофобластных клеток, в отношении образца от родителя мужского пола, ниже чем 1-й процентиль ожидаемого показателя сходства для собственных образцов, не подтверждают гетерозиготный отцовский занос P1P2.

В отличие от предшествующего уровня техники, авторы настоящего изобретения

неожиданно обнаружили, что комбинирование DRS-WGA с получением библиотеки для массивного параллельного секвенирования с использованием реакции ПЦР с адаптером для секвенирования/праймером для слияния WGA, без фрагментации, для полногеномного секвенирования с низким покрытием, улучшает возможность установления отличий образцов ДНК, даже по полногеномному секвенированию с низким покрытием при очень малой глубине, ниже чем 1x, для собственных и родственных образцов, и кроме того, также разделения смешанных собственных и родственных образцов с относительно хорошей точностью. Кроме того, для неродственных индивидуумов, даже полногеномное секвенирование с чрезвычайно низким покрытием, таким как < 0,15x, является достаточным.

Чтобы доказать вышеуказанное, проводили следующие эксперименты.

Примеры

Пример 1

Данные секвенирования первоначально получали с использованием 7 линий клеток. На фиг. 1 показан эффект способа получения полногеномной библиотеки на корреляцию частот аллелей SNP между собственными и неродственными образцами. На оси X показан способ получения библиотеки. Библиотеки без фрагментации получали посредством проведения полногеномной амплификации с детерминистическим участком рестрикции (DRS-WGA) геномной ДНК 2 отдельных клеток из 7 линий клеток опухолей (NCI-H1650, NCI-H23, NCI-H661, NCI-H1563, NCI-H1573, NCI-H441, OE19) с последующей реакцией ПЦР с адаптером для секвенирования/праймером для слияния WGA, без фрагментации; библиотеки со случайной фрагментацией получали из геномной ДНК 6 линий клеток опухолей (NCI-H1650, NCI-H23, NCI-H661, NCI-H1563, NCI-H1573, NCI-H441) с использованием набора для получения библиотек Ion Xpress™ Plus gDNA Fragment Library preparation kit (Thermo Fisher Scientific). На оси Y показан показатель попарного сходства, рассчитанный как корреляция частоты B-аллеля среди локусов, покрытых по меньшей мере одним прочтением в парных образцах, как зарегистрировано посредством NGSCheckMate (фиксация 8ea2c0438). NGSCheckMate выполняли для 500000 прочтений (≈0,025x покрытие), выровненных с эталонным геномом (hg19), с параметрами по умолчанию и набором полиморфных локусов по умолчанию (21067 SNP). Черными точками (собственные) показаны показатели попарного сходства парных образцов, принадлежащих к одной и той же линии клеток. Серыми точками (неродственные) показаны показатели попарного сходства парных образцов, принадлежащих к различным линиям клеток. График показывает явное преимущество получения библиотеки на основе DRS-WGA без фрагментации над способом со случайной фрагментацией, с более сильным разделением между значениями показателя попарного сходства собственных и неродственных.

Пример 2

Полиморфные локусы для сравнений, в соответствии с настоящим изобретением, предпочтительно, выбирают на основании их средней гетерозиготности. Предпочтительно, полиморфные локусы выбирают на основании свойства наличия средней гетерозиготности,

более высокой, чем конкретный минимальный порог.

На фиг. 2 показан эффект выбора набора полиморфных локусов на показатели попарного сходства парных образцов, принадлежащих к одной и той же (собственные) или различным линиям клеток (неродственные). Библиотеки получали посредством проведения полногеномной амплификации с детерминистическим участком рестрикции (DRS-WGA) геномной ДНК 2 отдельных клеток из 7 линий клеток опухолей (NCI-H1650, NCI-H23, NCI-H661, NCI-H1563, NCI-H1573, NCI-H441, OE19) с последующей реакцией ПЦР с адаптером для секвенирования/праймером для слияния WGA, без фрагментации. На оси X показан набор полиморфных локусов, используемый для анализа: набор 21000 соответствует набору SNP, предоставленному по умолчанию посредством NGSCheckMate и выбранному на основании частот аллелей полиморфных локусов в dbSNP в наборе из 40 профилей WGS зародышевой линии для пациентов с раком желудка из TCGA; набор 300000 состоит из 312458 полиморфных локусов, выбранных из dbSNP (построение 150) на основании минимальной средней гетерозиготности 0,498. На оси Y показан показатель попарного сходства, рассчитанный как корреляция частоты B-аллеля среди локусов, покрытых по меньшей мере одним прочтением в по меньшей мере двух образцах, степень сходства которых анализируют. NGSCheckMate выполняли для 500000 прочтений (≈0,025X покрытие), выровненных с эталонным геномом (hg19), с параметрами по умолчанию и либо с набором полиморфных локусов по умолчанию (21000), либо с набором 300000. График показывает, что посредством использования выбора полиморфных локусов на основании средней гетерозиготности, различие между показателями попарного сходства парных образцов, принадлежащих к одной и той же линии клеток (собственные) и показателями попарного сходства парных образцов, принадлежащих к различным линиям клеток (неродственные) увеличивается, приводя к четкому разделению между двумя типами сравнения.

Различные способы расчета показателей сходства можно использовать на стадии g) в соответствии с настоящим изобретением.

Как упомянуто в предшествующем описании, в предпочтительном варианте осуществления, показатель попарного сходства со стадии g) рассчитывают посредством вычисления корреляции частоты B-аллеля среди локусов, покрытых по меньшей мере одним прочтением в по меньшей мере двух образцах, степень сходства которых анализируют.

В другом предпочтительном варианте осуществления, показатель попарного сходства со стадии g) рассчитывают посредством вычисления среднего показателя согласованности среди локусов, покрытых по меньшей мере одним прочтением в обоих парных образцах, где показателю согласованности для каждого локуса приписывают одно из следующих значений:

a) 1, если запрашиваемые аллели являются идентичными;

b) 0, если запрашиваемые аллели являются полностью различными;

c) 0,5, если запрашиваемые аллели являются частично перекрывающимися.

Пример 3

На фиг. 3A и 3B показано распределение показателей попарного сходства, вычисленное среди образцов, происходящих от одного и того же индивидуума («собственные») или различных неродственных индивидуумов («неродственные»), для 500000 прочтений и минимальной средней гетерозиготности=0,46 или 5000000 прочтений и минимальной средней гетерозиготности=0,49, с использованием способов корреляции (фиг. 3A) или согласованности (фиг.3B).

Обоими способами получают сходные результаты в отношении разделения и разброса образцов из одного и того же класса, однако, абсолютное значение показателя попарного сходства (y-ось) должно явно изменяться, в соответствии с конкретным используемым способом. Показатель попарного сходства на основании согласованности имеет преимущество более простого вычисления, по сравнению с корреляцией, предоставляя лучшую производительность вычислений, особенно в случае больших наборов полиморфных локусов. Для обеих глубин прочтения графики не показывают явных различий в отношении распределений показателя попарного сходства, при разделении собственных и неродственных парных образцов, между двумя используемыми показателями сходства, однако абсолютное значение показателя сходства необходимо корректировать для конкретной функции, используемой в расчете.

Пример 4 - Средняя гетерозиготность и количество полиморфных локусов

Минимальная средняя гетерозиготность, предпочтительно, лежит в диапазоне [0,2;0,499]. Количество рассматриваемых полиморфных локусов уменьшается монотонно с увеличением минимальной средней гетерозиготности.

Количество локусов, покрытых посредством парных образцов, увеличивается монотонно с количеством прочтений на образец. Как правило, существует оптимальная минимальная средняя гетерозиготность для увеличения разделения между совпадающими (тот же индивидуум) и неродственными образцами, для определенного количества прочтений. Дальнейшее увеличение минимальной средней гетерозиготности за этот оптимум может сначала постепенно, и затем резко, уменьшать количество локусов, покрытых в парных образцах, которые доступны для сравнения, таким образом, уменьшая общее разделение между совпадающими и неродственными образцами по показателю попарного сходства.

На фиг. 4A - 4C показана взаимосвязь между параметрами. На фиг. 4A показана взаимосвязь между порогом средней гетерозиготности (X-ось; диапазон=0,2-0,5), используемым для выбора набора полиморфных локусов, и количеством полиморфных локусов (Y-ось). На фиг. 4B показана взаимосвязь между количеством полиморфных локусов в наборе (Y-ось) и средним количеством локусов, покрытым в обоих парных образцах по меньшей мере одним прочтением (X-ось) при различных глубинах прочтения. На фиг. 4C показана взаимосвязь между средним количеством локусов, покрытым в обоих парных образцах (X-ось), и расстоянием между распределением показателя попарного сходства (согласованности) парных образцов, принадлежащих к одной и той же линии

клеток (собственные), против распределения показателя парных образцов, принадлежащих к различным линиям клеток (неродственные), рассчитанного как 5-й процентиль распределения показателя попарного сходства для собственных минус 95-й процентиль распределения показателя попарного сходства для неродственных, при различных глубинах прочтения, лежащих в диапазоне от 500000 прочтений до 4000000 прочтений.

Фиг. 4D - 4F представляют собой крупный план такого же типа анализа для более узкого диапазона минимальной средней гетерозиготности.

Пример 5 - Анализ кровного родства

Даже более сложная проблема при идентификации образца возникает в случаях родства, такого как родственные отношения, как, например, когда половина генома является общей между матерью и ее дочерью.

Для оценки эффективности способа в соответствии с настоящим изобретением в этом случае применения, авторы настоящего изобретения моделировали этот случай путем получения, *in silico*, родственных образцов посредством смешивания (50%/50%) данных полногеномного секвенирования с низким покрытием, полученных, в соответствии со способом, для отдельных лейкоцитов, полученных от нескольких (N=3) различных неродственных индивидуумов, в результате чего для каждого индивидуума, полиморфные локусы были редактированы в данных таким образом, чтобы регистрировать только один из детектированных аллелей для этого индивидуума, таким образом, моделируя вклад гаплоидного генома от этого индивидуума в «родственные» данные. Из периферической крови, собранной в пробирки для сбора крови CellSave (Menarini Silicon Biosystems), после иммуно-магнитного обогащения с использованием CELLSEARCH AutoPrep, клетки окрашивали с использованием коктейля флуоресцентных антител и DAPI, затем CD45+, DAPI+ отдельные клетки выделяли посредством DEPArray (Menarini Silicon Biosystems), и проводили полногеномную амплификацию с использованием DRS-WGA (Ampli1 WGA, Menarini Silicon Biosystems). Аликвоту продукта WGA использовали для получения библиотеки для массивного параллельного секвенирования из каждого продукта этой DRS-WGA с использованием реакции ПЦР с адаптером для секвенирования/праймером для слияния WGA, без фрагментации (набор Ampli1 LowPass для Illumina, Menarini Silicon Biosystems).

Чтобы избежать смещений, данные секвенирования для каждой отдельной клетки использовали только один раз (для получения либо собственного, либо родственного типа данных).

На фиг. 5A - 5D показана эффективность классификации родственных образцов, в отношении собственных (родитель женского пола) и неродственных образцов. Два переменных порога для показателя сходства, рассчитанных в отношении образцов от родителя женского пола, используют в качестве классификаторов для установления отличий родственных образцов от собственных и неродственных образцов. Порог родственного-собственного устанавливают на значения, лежащие в диапазоне от медианы распределения показателя сходства для родственных до медианы распределения показателя

сходства для собственных. Порог родственного-неродственного устанавливают на значения, лежащие в диапазоне от медианы распределения показателя сходства для родственных до медианы распределения показателя сходства для неродственных. Количество прочтений сохраняют постоянным при 500000 прочтений. На фиг. 5A показаны значения TPR и 1-PPV для классификации родственных образцов, в отношении собственных образцов родителя женского пола, по мере изменения порога, при различной минимальной средней гетерозиготности (порог средн. гетерозиг.). На фиг. 5B показаны значения TPR и 1-PPV для классификации родственных образцов, в отношении неродственных образцов, по мере изменения порога, при различной минимальной средней гетерозиготности (порог средн. гетерозиг.). На фиг. 5C показан порог показателя сходства родственного-собственного (сплошная серая линия; вторичная Y-ось), необходимый для получения PPV по меньшей мере 0,999 и соответствующей TPR (первичная Y-ось), по мере изменения значения минимальной средней гетерозиготности (X-ось). На фиг. 5D показан порог показателя сходства родственного-собственного (сплошная серая линия; вторичная Y-ось), необходимый для получения PPV по меньшей мере 0,999 и соответствующей TPR (первичная Y-ось), по мере изменения значения минимальной средней гетерозиготности (X-ось). Графики показывают, что высокую чувствительность (TPR≥0,99) получают с наборами SNP, выбранными с использованием порога средней гетерозиготности от 0,2 вплоть до 0,495 для классификации родственного-собственного и вплоть до 0,48 для классификации родственного-неродственного, с быстрым уменьшением значений чувствительности после этих значений.

Пример 6

На фиг. 6 показано распределение показателей попарного сходства, рассчитанных как согласованность, в отношении образцов от родителя женского пола, для собственных (родитель женского пола), родственных и неродственных образцов, как функция от минимальной средней гетерозиготности (диапазон=0,2-0,498). Количество прочтений сохраняют постоянным при 500000 прочтений. Пороги показателя сходства, используемые для классификации родственных образцов против собственных образцов от родителя женского пола и неродственных образцов с PPV по меньшей мере 0,999, показаны как пунктирная и штрих-пунктирная линии, соответственно.

Соответственно, в предпочтительном варианте осуществления, данные LPWGS прореживают до 500000 отдельных прочтений, минимальную среднюю гетерозиготность для полиморфных локусов выбирают в диапазоне [0,2;0,49], и пороги показателя сходства выбирают в диапазоне [0,73;0,79] для родственных-собственных и [0,62;0,7] для родственных-неродственных, с использованием, в качестве показателя сходства, «согласованности», рассчитанной, как объяснено выше. Множество полиморфных локусов, предпочтительно, содержит локусы, полученные из базы данных, такой как dbSNP. Предпочтительно, указанное множество полиморфных локусов содержит > 200000, 300000, 500000 или 1000000 локусов с наиболее высокой средней гетерозиготностью.

Кластеризация

В предпочтительном варианте осуществления, способ в соответствии с настоящим изобретением дополнительно включает стадию определения группы кластеров образцов, разделяющих общее свойство, такое как идентичность одного индивидуума (или более индивидуумов), вносящих значительный вклад ДНК в образцы из кластера, или свойство содержания недостаточных количеств ДНК, и/или свойство содержания высоко деградированной ДНК или ДНК непонятного происхождения. По меньшей мере два образца, предпочтительно, приписывают по меньшей мере одному кластеру посредством классификатора, с использованием указанного показателя сходства и других метрических показателей качества.

Пример 7 - Применение для неинвазивной пренатальной диагностики на основании фетальных циркулирующих клеток.

В предпочтительном варианте осуществления, по меньшей мере один эталонный кластер составляют из образцов от беременного индивидуума - родителя женского пола. Указанные «эталонные образцы» можно собирать, выделяя материнские клетки из той же обогащенной физиологической жидкости, что использовали для выделения фетальных клеток, или альтернативно, посредством другого источника материнской ДНК. В случае, когда материнская физиологическая жидкость состоит из периферической крови, ядросодержащие клетки, положительные по материнским маркерам и отрицательные по фетальным маркерам, можно собирать в качестве эталона.

Предпочтительно, указанная группа кластеров идентичности может дополнительно содержать по меньшей мере один родственный кластер, состоящий из образцов из по меньшей мере одного плода из текущей беременности указанного индивидуума - родителя женского пола. Указанные образцы идентифицируют, предпочтительно, как имеющие показатель попарного сходства согласующийся с родственной взаимосвязью с эталонным родителем женского пола.

Указанный родственный кластер, предпочтительно, далее разбивают на множество фетальных кластеров, состоящих из образцов, которые содержат ДНК только от одного и того же плода.

Образцы, принадлежащие одному и тому же плоду, распознают как имеющие показатель попарного расстояния, согласующийся с классификацией как собственные, в отношении друг друга. Другие родственные клетки, имеющие показатель попарного расстояния, согласующийся с родственной взаимосвязью в отношении других родственных клеток, помещают в отличный раздел, как принадлежащие отличному плоду.

Фиг. 7 представляет способ детекции двуплодной беременности. Все попарные прогнозы фетальных клеток, описанных посредством «родственной» связи с материнским контролем, используют в качестве ввода в алгоритм кластеризации графов для нахождения «сообществ» фетальных клеток.

В другом варианте осуществления, который можно использовать в контексте неинвазивной пренатальной диагностики, циркулирующие фетальные клетки, смешанные с материнскими клетками, детектируют посредством наблюдения показателя попарного

сходства, промежуточного, в отношении того, что ожидают для «собственного» типа ДНК и «родственного» типа ДНК. Фактически, совместное выделение материнской клетки вместе с целевой фетальной клеткой может происходить случайно в результате погрешности в способе сортировки (либо из-за выбора клеток для выделения, либо из-за способа выделения, либо из-за того и другого). Совместное выделение материнской клетки вместе с целевой фетальной клеткой может также происходить неслучайно, поскольку может обеспечивать преимущество анализ каким-либо образом дополнительного смешанного образца, вместо его отбрасывания, если слишком мало несмешанных и чистых образцов фетальных клеток являются доступными.

В зависимости от типа анализа, смесь двух клеток, одной фетальной и одной материнской, все еще может являться приемлемой, если чувствительность анализа значительно не нарушена. Это может представлять собой, например, случай при анализе анеуплоидий целых хромосом, с использованием адекватных количеств прочтений. Контаминацию можно обеспечивающим преимущества образом учитывать в ходе анализа посредством применения специфического коэффициента контаминации, как это доступно в определенных биоинформатических конвейерах, таких как ControlFreec (Boeva, V. et al, Bioinformatics 2012 Feb 1;28(3):423-5), таким образом, поддерживая адекватную чувствительность.

В предпочтительном варианте осуществления, указанные фетальные клетки, циркулирующие в материнской крови, представляют собой (i) трофобласты, (ii) эритробласты или (iii) оба типа.

Пример 8 - Идентификация циркулирующих фетальных эритробластов из материнской крови.

Ядросодержащие клетки сначала выделяли из материнской крови с использованием градиента фиколла (плотность 1,107 г/мл), и фетальные эритробласты (ядросодержащие эритроциты) обогащали посредством иммуно-магнитного истощения по CD45/CD15/CD14 нежелательных материнских клеток с использованием магнитной сортировки активированных клеток (MACS) от Miltenyi.

Обогащенные клетки фиксировали, с использованием либо

(A) Параформальдегида (PFA) 4% в течение 30’ при комнатной температуре, либо

(B) PFA 4% 60’, 37° с последующим 0,05% глутаральдегидом в течение 30” при комнатной температуре

Второй тип фиксации образует более сильное перекрестное сшивание и может способствовать фиксации целевого гемоглобина внутри клетки, однако, препятствует амплификации ДНК.

После фиксации, клетки окрашивали посредством антител с FITC против гамма-гемоглобина (в качестве маркера фетальных клеток) и DAPI для окрашивания ДНК в ядрах.

Предположительные фетальные клетки сортировали посредством DEPArray™ как отдельные клетки, или совместно с дополнительными материнскими контаминирующими клетками, которые случайно оказывались локализованными совместно в той же самой

диэлектрофоретической ячейке. Выделенные препараты клеток (независимо от того, отдельные или контаминированные) амплифицировали с использованием набора Ampli1 WGA, Menarini Silicon Biosystems S.p.A., осуществляющего способ DRS-WGA в соответствии с настоящим изобретением.

Аликвоту (1 мкл) продукта первичной ПЦР с Ampli1 WGA использовали для анализа микросателлитов, с использованием мультиплексной ПЦР для амплификации следующих локусов: D21S1435, D21S11, HPRT, SRY, D21S1413, D21S1411, D18S535, D13S317, D21S2039, D13S631, D21S1442, с последующим анализом фрагментов с использованием капиллярного электрофореза в ABI Prism 310 (Applied Biosystems). С использованием протокола «более слабой» фиксации - варианта (A) выше - 56% ожидаемых аллелей выделили в среднем (диапазон 30%-90%). В среднем, обнаружили 3,2 информативных аллеля, определенных как аллели, не общие между материнским и фетальным эталонным профилем, полученным посредством анализа образца ворсин хориона (CVS).

С использованием протокола «более сильной» фиксации - варианта (B) выше - только 28% ожидаемых аллелей выделили в среднем (диапазон 6%-68%), т.е., приблизительно половину от выделенных с использованием более слабой фиксации. Иными словами, с использованием *более сильной* фиксации (B), получили среднее выпадение аллелей 72%. Соответственно, в среднем обнаружили только 1,7 информативных аллелей, включая также смешанные образцы (BO1368B_4, BO1368B_6), имеющие как материнские, так и фетальные информативные аллели, таким образом, имеющие две клетки и двойное количество исходной ДНК-матрицы. Действительно, 4 образца отдельных клеток (BO1368B_3, BO1368B_5, BO1368B_9, BO1368B_12) имели 0 информативных аллелей в вышеуказанном мультиплексном анализе STR. Первые три из них разделили только с помощью дополнительного анализа с использованием дополнительных локусов STR, анализа, в котором не удалось получить информацию для классификации образца BO1368B_12, который оставался «неизвестного» происхождения.

Таким образом, ясно, что, в то время как она предоставляет большее количество фетальных эритробластов, более сильная фиксация (такая как PFA 4% 60’ 37° с последующим 0,05% глутаральдегидом в течение 30” при комнатной температуре), увеличивает выпадение аллелей и уменьшает уровень определения STR, таким образом, подвергая серьезному риску классификацию образца как материнского, фетального или смешанного.

И наоборот, при получении из другой аликвоты продукта WGA библиотеки для массивного секвенирования с использованием набора Ampli1 LowPass, и анализа данных с использованием способа в соответствии с настоящим изобретением, является возможным с доверием приписать каждый образец, как дополнительно более подробно описано ниже, даже для таких образцов с очень высоким выпадением аллелей.

На фиг. 8 показано распределение средних показателей попарного сходства, рассчитанных в отношении образцов от родителя женского пола, в выделенных препаратах клеток эритробластов из 2 образцов. График показывает, что классификатор порог

родственного-собственного устанавливает отличия выделенных препаратов родственных (серые точки) от выделенных препаратов клеток беременного индивидуума - родителя женского пола (светло-серые точки). Однако, классификатор не может устанавливать отличия выделенных препаратов родственных от выделенных препаратов смешанных клеток (черные точки).

В предпочтительном варианте осуществления, кластеризация образцов включает вычисление коэффициента силуэта, на основании сходства, для определения количества кластеров. Обеспечивающим преимущество образом, кластер, где для показателей попарного сходства показывают два отдельных уровня сходства, можно далее фракционировать посредством использования фиксированного порога, предпочтительно, 0,205, на основании распределения коэффициентов силуэта в наборе образцов, содержащих материнские клетки и фетальные клетки, для установления отличий смешанных фетальных-материнских образцов (от фетальных или материнских образцов). В предпочтительном варианте осуществления, указанный фиксированный порог лежит в диапазоне [0,19-0,21].

Таким образом, смешанные материнские-фетальные клетки можно идентифицировать как отдельный кластер от собственной (материнской) и родственной (фетальной) субпопуляции.

Пример 9

На фиг. 9A - 9C показана классификация на основе кластеризации выделенных препаратов клеток из образца BO1368. Образец материнских клеток (BO1368_MC) и образец ворсин хориона (BO1368_CVS) включены в качестве эталона. На фиг. 9A показаны средние коэффициенты силуэта для различных количеств кластеров, используемых в качестве ввода для кластеризации показателей попарного сходства, имеющих наиболее высокий показатель для 2 кластеров. На фиг. 9B показан анализ индивидуального коэффициента силуэта для каждого выделения в двух кластерах, показывающий, что 2 выделенных препарата в кластере #0, соответствующие выделенным препаратам смешанных клеток, имеют показатель, близкий к 0, показывающий, что они находятся очень близко к границе принятия решений между двумя соседними кластерами; посредством установки фиксированного минимального порога коэффициента силуэта (0,205), является возможным установление отличий 2 выделенных препаратов смешанных фетальных-материнских клеток, которые, таким образом, приписывают к третьему независимому кластеру. На фиг. 9C показана тепловая карта, показывающая показатели сходства между всеми 17 выделенными препаратами клеток в оттенках серого, где более темные цвета показывают более высокое сходство; кластеры отмечены цветным метками строк и столбцов.

Пример 10

На фиг. 10A - 10C показана классификация на основе кластеризации выделенных препаратов клеток из образца BO1383. Образец материнских клеток (BO1383_MC) включен в качестве эталона. На фиг. 10A показаны средние коэффициенты силуэта для различных

количеств кластеров, используемых в качестве ввода для кластеризации показателей попарного сходства, имеющих наиболее высокий показатель для 2 кластеров. На фиг. 10B показан анализ индивидуального коэффициента силуэта для каждого выделения в двух кластерах, показывающий, что 2 выделенных препарата в кластере #0, соответствующие выделенным препаратам смешанных клеток, имеют показатель, близкий к 0, показывающий, что они находятся очень близко к границе принятия решений между двумя соседними кластерами; посредством установки фиксированного минимального порога коэффициента силуэта (0,205), является возможным установление отличий 2 выделенных препаратов смешанных фетальных-материнских клеток, которые, таким образом, приписывают к третьему независимому кластеру. На фиг. 10C показана тепловая карта, показывающая показатели сходства между всеми 8 выделенными препаратами клеток в оттенках серого, где более темные цвета показывают более высокое сходство; кластеры отмечены цветным метками строк и столбцов.

Пример 11 - Применение для неинвазивного пренатального установления отцовства на основании фетальных циркулирующих клеток.

В другом варианте осуществления настоящего изобретения, образец от родителя мужского пола (отцовский образец) является доступным, в дополнение к материнскому образцу, и анализ кровного родства можно применять с использованием, в свою очередь, в качестве эталона также отцовского образца. Показатель попарного сходства, согласующийся с «родственным» типом ДНК, в отношении отцовского эталонного образца, подтверждает отцовство для плода. Альтернативно, если показатель попарного сходства фетального образца (т.е., подтвержденного фетального, поскольку классифицированного как родственный в отношении эталонного образца от родителя женского пола) согласуется с «неродственным» типом ДНК при использовании образцов от родителя мужского пола, результат опровергает отцовство.

Пример 12 - Применение для молярной беременности.

В другом варианте осуществления настоящего изобретения, по меньшей мере одну предположительную циркулирующую фетальную трофобластную клетку обогащают из материнской крови. Образец трофобластных клеток сравнивают с материнским эталонным образцом, и показатель попарного сходства, согласующийся с «неродственным» типом ДНК, показывает возможный полный занос (или лабораторную контаминацию/замену образца). Если выделяют более одного образца циркулирующих трофобластных клеток, сравнение показателя попарного сходства среди этих образцов можно использовать для исследования генотипа заноса. Если попарное расстояние сильно превышает ожидаемое значение для парных образцов типа «собственных», подтверждают гомозиготный отцовский занос P1P1, поскольку все сравнение полиморфных локусов может являться идентичным, за исключением редких ошибок секвенирования (или даже более редких ошибок амплификации WGA), которые могут иногда возникать в таких же геномных положениях, соответствующих проверяемым полиморфным локусам. Альтернативно, в присутствии заноса P1P2 с гетерозиготностью в некоторых из полиморфных локусов,

значение попарного сходства, наблюдаемое среди различных образцов трофобластов, лежит в диапазоне, ожидаемом для парных образцов типа «собственных». В этом последнем случае заноса P1P2, если отцовский образец ДНК является доступным, показатель попарного расстояния образцов трофобластов, согласующийся с «собственным» типом ДНК в отношении отцовского эталонного образца, можно использовать для установления отличий молярной беременности от лабораторной контаминации или замены образца.

Пример 13 - Применение для криминалистической идентификации и идентификации человека по отдельным клеткам.

В предпочтительном варианте осуществления, указанный по меньшей мере один эталонный кластер состоит из образцов, содержащих ДНК только от одного и того же индивидуума, соответствующего потерпевшему в криминалистической экспертизе, дополнительно включающей определение по меньшей мере одного кластера виновного, содержащего образцы, содержащие ДНК только от одного и того же индивидуума, отличного от потерпевшего.

Образцы приписывают кластеру виновного, если они имеют показатель попарного расстояния, согласующийся с «неродственной» взаимосвязью с образцами от потерпевшего, и «собственной» взаимосвязью с другими образцами, принадлежащими к одному и тому же кластеру виновного. Во всех случаях, когда новый образец согласуется с «неродственным» как с потерпевшим, так и с виновными, уже принадлежащими к другим кластерам виновных, определяют новый кластер виновного.

Альтернативно, применение алгоритма кластеризации на основе коэффициента силуэта, как подробно описано для случая применения для неинвазивной пренатальной диагностики, можно использовать для присвоения каждого индивидуального образца гомогенному кластеру.

Обеспечивающим преимущество образом, в случае криминалистической идентификации, образцы с показателем попарного расстояния, согласующимся с «родственной» взаимосвязью (как получено для неинвазивной пренатальной диагностики - NIPD-типа анализа), можно интерпретировать как «смешанные образцы», поскольку они, вероятно, содержат ДНК от двух неродственных индивидуумов (потерпевшего и виновного или различных виновных), подобно случаю «родственных» образцов при применении NIPD, содержащих ДНК от одного родителя женского пола и одного неродственного родителя мужского пола.

Обеспечивающим преимущество образом, информацию о количестве копий для половых хромосом, полученную по данным того же самого полногеномного секвенирования с низким покрытием, можно использовать для дальнейшего уточнения и/или подтверждения классификации на основе уточненного показателя попарного расстояния.

В случае несовпадения по полу между потерпевшим и виновными, как является обычным при доказательстве изнасилования, информация о количестве копий для

хромосомы X и Y может способствовать получению информации для классификации образца как потерпевшего или виновного.

В другом предпочтительном варианте осуществления, указанный по меньшей мере один эталонный кластер состоит из образцов, содержащих ДНК только от одного и того же индивидуума, соответствующего подозреваемому виновному в криминалистической экспертизе, дополнительно включающей определение по меньшей мере одного кластера виновного, содержащего образцы, содержащие ДНК только от одного и того же индивидуума.

В другом предпочтительном варианте осуществления, множество образцов, полученных из смешанных криминалистических улик с вкладом ДНК множества участников, каждый образец, содержащий одну или несколько клеток, анализируют в соответствии со способом, дополнительно включающим определение по меньшей мере одного кластера виновного, содержащего образцы, содержащие ДНК только от одного и того же индивидуума.

В предпочтительном варианте осуществления, аликвоты после DRS-WGA из множества образцов, каждый из которых принадлежит к одному и тому же из указанных по меньшей мере одного из кластеров виновных, смешивают вместе, таким образом, получая для каждого кластера соответствующий образец WGA-ДНК отдельного индивидуума, таким образом, обеспечивая возможность проводить дальнейший анализ ДНК для указанного образца WGA-ДНК отдельного индивидуума. Преимуществом этого способа является то, что потенциальные случайные выпадения аллелей, возникающие в образце отдельной клетки, дополняют сигналом от других индивидуальных клеток, таким образом, получая более полный профиль. Этот способ обеспечивает особенное преимущество, когда ДНК из образца каждой отдельной клетки от индивидуума является сильно деградированной. Это может происходить, в частности, для нераскрытых дел, особенно когда улику хранили при комнатной температуре, или в случаях, когда образец ткани от потерпевшего был фиксирован в формалине и погружен в парафин для позднейшего использования.

Другой предпочтительный вариант осуществления относится к слиянию по кластерам данных генетического анализа из по меньшей мере одного типа анализа, из множества образцов, принадлежащих к каждому из указанных по меньшей мере одного из кластеров виновных, с получением для каждого из указанных по меньшей мере одного из кластеров виновных соответствующих данных WGA-ДНК отдельного индивидуума.

В предпочтительном варианте осуществления, указанный по меньшей мере один тип анализа выбран из группы, состоящей из:

a) анализа микросателлитов;

b) анализа однонуклеотидного полиморфизма;

c) массивного параллельного нацеленного секвенирования;

d) полногеномного секвенирования.

На фиг. 11A и 11B показана эффективность классификации индивидуальных

образцов, в отношении неродственных образцов с максимум 50% компонента собственных образцов. Классификатор на основе переменного порога показателя попарного сходства используют для установления отличий образцов от индивидуума от смешанных образцов. Порог устанавливают на значения, лежащие в диапазоне от медианы распределения «собственного» показателя сходства до медианы распределения «смешанного» показателя сходства. Количество прочтений сохраняют постоянным при 500000 прочтений. A) Значения TPR и 1-PPV для классификатора по мере изменения порога, при различной средней гетерозиготности (порог средн. гетерозиг.). B) Порог показателя попарного сходства (сплошная серая линия; вторичная Y-ось), необходимый для получения PPV по меньшей мере 0,999 и соответствующей TPR (первичная Y-ось) как функции от средней гетерозиготности (X-ось). Графики показывают, что высокую чувствительность (TPR≥0,99) получают с набором SNP, выбранным с использованием порога среднее гетерозиготности от 0,2 вплоть до 0,495 для классификации родственного-собственного и вплоть до 0,48 для классификации собственного-смешанного, с значениями чувствительности, быстро уменьшающимися после этих значений.

На фиг. 12 показано распределение показателей попарного сходства (согласованность), рассчитанное для парных образцов от одного и того же индивидуума (собственные), для парных образцов, где один из образцов содержит 50% компонента от того же индивидуума, что и другой образец (смешанные_1/2), для парных образцов, где один из образцов содержит 1/3 (33%) от того же индивидуума, что и «собственные», и 66% компонента от того же индивидуума, что и другой образец (смешанные_1/3), для парных образцов, принадлежащих различным индивидуумам (неродственные), как функция от средней гетерозиготности (диапазон=0,2-0,499). Количество прочтений сохраняют постоянным при 500000 прочтений. Классификатор на основе показателя попарного сходства показан как пунктирная линия.

Термин виновный и потерпевший, использованные выше, предназначены только в качестве ориентира и чтобы способствовать пониманию. Специалисту в данной области понятно, что вышеуказанный способ является применимым, без отклонения от настоящего изобретения, также к другим условиям идентификации человека, таким как идентификация индивидуальных жертв катастроф, где значение кластера всего лишь изменяют с виновного на отличное произвольное наименование.

Пример 14 - Применение идентификации образца в технологическом маршруте в онкологической лаборатории.

В другом предпочтительном варианте осуществления, способ в соответствии с настоящим изобретением используют для установления соответствия образцов, принадлежащих одному и тому же пациенту, и детекции возможных замен образцов или возможной перекрестной контаминации от образцов, принадлежащих различным пациентам. Например, это может обеспечивать особенные преимущества при работе с FFPE образцами отдельных клеток. Фактически, является чрезвычайно сложным получение исчерпывающей геномной информации из отдельной клетки (или ядра), экстрагированных

из FFPE, из-за повреждения ДНК, вызванного фиксацией. STR или даже нацеленное секвенирование для SNP могут являться непрактичными. Однако с использованием способа в соответствии с настоящим изобретением, все еще является возможным установление отличий образцов.

На фиг. 13A - 13C показана классификация отдельных выделенных препаратов клеток из FFPE образцов, в соответствии с идентичностью индивидуумов. Продукты WGA отдельных клеток получали, как подробно описано в Mangano C. et al., «Precise detection of genomic imbalances at single-cell resolution reveals intra-patient heterogeneity in Hodgkin's lymphoma», Blood Cancer Journal volume 9, Article number: 92 (2019). На фиг. 13A показан роевый график, показывающий показатели попарного сходства парных образцов, принадлежащих одному и тому же индивидууму (собственные) или к различным индивидуумам (неродственные). Данные группируют, в соответствии с сигналом DLRS полногеномного количества копий (X-ось), где низк. DLRS соответствует парным образцам с DLRS < 0,4, показательным для низкого шума сигнала, и высок. DLRS соответствует парным образцам, где по меньшей мере для одного из образцов в паре показано DLRS≥0,4, показательное для высокого шума сигнала. Для обеих групп графики показывают четкое разделение, в отношении показателя попарного сходства, между собственными и неродственными образцами. На фиг. 13B показаны средние коэффициенты силуэта для различных количеств кластеров, используемых в качестве ввода для кластеризации KMeans показателей попарного сходства, показывающие наивысшую оценку в баллах для 4 кластеров. На фиг. 13C показана тепловая карта, показывающая показатели попарного сходства между всеми 17 выделенными препаратами клеток в оттенках серого, где более темные цвета показывают более высокое сходство; кластеры отмечены цветным метками строк и столбцов; для целей визуализации, строки и столбцы упорядочены посредством иерархической кластеризации на основе евклидова расстояния.

Пример 15 - Применение идентификации образца в преимплантационном генетическом скрининге (PGS).

В другом предпочтительном варианте осуществления, способ в соответствии с настоящим изобретением используют для анализа образцов, происходящих из свободной от клеток использованной культуральной среды для эмбриона. Как известно в данной области, обеспечивает преимущества оценка эмбрионов для определения приоритета для имплантации, для увеличения частоты прикрепления и успешности процедуры. Способы на основе свободной от клеток использованной культуральной среды для эмбриона являются привлекательными, поскольку они упрощают технологический маршрут и могут являться менее инвазивными для развивающегося эмбриона. Однако, опубликована контаминация материнской ДНК культуральной среды и показано, что она ухудшает разрешение PGS при детекции анеуплоидий у плода.

В одном варианте осуществления настоящего изобретения, в контексте этого применения, материнский эталон используют в качестве эталона для «собственного» (родителя женского пола). Показатель попарного сходства с образцом свободной от клеток

использованной культуральной среды для эмбриона вычисляют в соответствии с настоящим изобретением. Указанный показатель попарного сходства используют для оценки контаминации материнской ДНК в отношении ДНК эмбриона. Показатель попарного сходства, более низкий или равный ожидаемому медианному значению для «родственного» типа ДНК, в отношении материнского эталона, используют для принятия 100% чистоты эмбриональной ДНК. Показатель попарного сходства, равный или более высокий, чем ожидаемое медианное значение для «собственного» типа ДНК, в отношении материнского эталона, используют для принятия 0% чистоты эмбриональной ДНК (полностью материнской ДНК) в свободном от клеток образце. Промежуточное значение попарного сходства показывает степень контаминации материнской ДНК. Это значение контаминации можно использовать в качестве ввода в анализ получения полногеномного профиля количества копий на основе тех же самых данных полногеномного секвенирования с низким покрытием, чтобы компенсировать потенциальное разбавление - из-за смешанного сигнала, происходящего из нормального диплоидного материнского генома - сигнала количества копий, обусловленного потенциальной анеуплоидией или субхромосомными изменениями количества копий эмбриона. Таким образом, благодаря компенсации, чувствительность операции, обозначающей количество копий, менее ухудшается из-за разбавления сигнала. Кроме того, значение контаминации можно использовать для оценки пригодности образца для надежной детекции изменений количества копий данного размера, поскольку степень диплоидного материнского фона может нарушать детекцию субхромосомных CNV, например, микроделеций.

На фиг. 14 показано моделирование, проводимое посредством смешивания, *in silico*, различных долей последовательностей ДНК из отдельных фетальных клеток с последовательностями из материнских клеток. Сплошная линия соответствует среднему показателю попарного сходства при различных процентах ввода фетальных. Закрашенная область соответствует 95% доверительному интервалу. Пунктирной линией показан пример смешанного образца с известным % материнского компонента (80%) и показателем попарного сходства с материнским эталоном=0,807, который, в соответствии с моделью, имеет средний прогнозируемый фетальный компонент=27,7% (C.I.= 25,4%-30,7%), соответствующий оцененной контаминации материнской ДНК ≈ 75%.

На фиг. 15A и 15B показан полногеномный анализ количества копий смешанного образца, полученного посредством смешивания *in silico* различных долей последовательностей ДНК из отдельных фетальных клеток (20%) с последовательностями из материнских клеток (80%). На фиг. 15A показан полногеномный профиль количества копий; каждая точка соответствует геномному интервалу 10 млн.п.о. На фиг. 15B показано полногеномное количество копий после применения поправочного коэффициента=0,75, на основе оцененного процента контаминации материнской ДНК, на основе показателя попарного сходства с материнским эталоном. Статистически значимые отклонения показаны как сплошные черные линии.

Сходный способ можно использовать также для внеклеточной ДНК или инвазивных

пренатальных образцов для определения фетальной фракции и контаминации, соответственно, с использованием эталона, содержащего лейкоциты плазмы, для внеклеточной ДНК, материнскую децидуальную оболочку, буккальный мазок или кровь.

Пример 16 - Применение для идентификации образца в аутентификации линии клеток.

В другом предпочтительном варианте осуществления, способ в соответствии с настоящим изобретением используют для аутентификации линий клеток, используемых в исследовательских лабораториях.

В этом варианте осуществления, сначала организуют *эталонную* базу данных, собирающую - из всех *эталонных* типов линий клеток - - исходные данные WGS с низким покрытием, в соответствии со способом, так что данные из этой эталонной базы данных используют для аутентификации *тестируемой* линии клеток. В предпочтительном варианте осуществления для этого применения, исходные образцы предпочтительно выбирают из группы, состоящей из (i) пула клеток или (ii) ДНК, выделенной из пула клеток.

Таким образом:

- для *эталонного* образца чистых линий клеток получают средний исчерпывающий профиль линии клеток, наилучшим образом обобщающий разнообразие, связанное с гетерогенностью клеток;

- для *тестируемого* образца, кроме того, можно наблюдать потенциальную контаминацию другой линией клеток. Порог на основе распределения показателей сходства среди повторений анализа можно использовать для обозначения контаминации, с определенной степенью доверия, если показатель сходства ниже, чем этот минимальный порог. Кроме того, с использованием способа, сходного с приведенным выше для применения для преимплантационного генетического скрининга, можно получать непрямой показатель уровня контаминации, сравнивая наблюдаемый показатель сходства тестируемого образца с калибровочной кривой, представляющей ожидаемый показатель сходства как функцию от контаминации чистого «собственного» другим характерным «неродственным» образцом.

Количество клеток в указанном пуле, предпочтительно, лежит в диапазоне [50-1500]. Нижний предел 50 обеспечивает минимум разнообразия, репрезентативного для геномной гетерогенности (если она присутствует). Кроме того, этот нижний предел можно использовать - в тестируемом образце - для детекции потенциальной контаминации другой линией клеток с более высокой чувствительностью, поскольку низкий уровень контаминации -например, 10% - может вообще не оказаться представленным в пуле клеток с более низким количеством клеток, или иным образом привести к получению образца, где минорный контаминант является недостаточно представленным, относительно реального % в популяции, таким образом, потенциально уменьшая общую чувствительность при детекции указанной контаминации. Верхний предел 1500 (т.е., эквивалент 10 нг) является предпочтительным для обеспечения хорошей амплификации WGA без ингибирования, которое может возникать при перегрузке реакционной смеси для WGA вводом ДНК, или

ингибирующего эффекта лизата цельной клетки при начале напрямую с клеток без очистки ДНК.

Пример 17 - Применение для аллогенной трансплантации гематопоэтических клеток.

В другом предпочтительном варианте осуществления, способ в соответствии с настоящим изобретением используют для оценки происхождения эндотелиальных клеток у пациентов при аллогенной трансплантации гематопоэтических клеток (алло-HSCT).

В предпочтительном варианте осуществления настоящего изобретения, выделение индивидуальных эндотелиальных клеток проводят из

1. FFPE срезов, после дезагрегации, окрашивания с использованием маркеров эндотелиальных клеток, таких как CD146, и сортировки отдельных клеток, например, такой как с использованием DEPArray™.

2. периферической крови, после обогащения и окрашивания циркулирующих эндотелиальных клеток (CEC) с использованием CELLSEARCH® AutoPrep и набора CEC, и сортировки отдельных клеток, например, такой как с использованием DEPArray™.

Получают первый эталонный образец, содержащий ДНК зародышевой линии от хозяина. Отдельные эндотелиальные клетки выделяют от пациентов, и оценивают их показатель сходства с эталонным образцом от хозяина. Если тестируемую клетку классифицируют как собственную, это означает, что для нее подтверждают происхождение от хозяина, в то время как если ее классифицируют как неродственную, ее классифицируют как принадлежащую неродственному донору.

Способ можно применять с использованием также анализа кровного родства для идентификации клеток донора, в случае, если донор связан с хозяином родственными отношениями.

Если, кроме того, образец ДНК зародышевой линии донора является доступным, можно получать второй эталонный образец в качестве подтверждения классификации.

Дополнительные общие детали и соображения, применимые среди различных применений

Однозначная взаимосвязь локуса с длиной фрагментов в DRS-WGA

Более подробно, в способе в соответствии с настоящим изобретением используют тот факт, что в DRS-WGA, такой как Ampli1™ WGA, каждый локус в геноме представлен в библиотеке WGA только фрагментами, имеющими специфическую длину в парах оснований. Это свойство может быть обозначено «однозначная взаимосвязь локуса с длиной фрагментов» (L2FLUR). Рассматривая общий нормальный локус, например, локус для полиморфного SNP, указанный локус может быть представлен только фрагментом данной длины, равной размеру соответствующего фрагмента (измеренному по любой из одиночных цепей) после расщепления рестрикционным ферментом, плюс двойная длина универсальных адаптеров WGA (длина праймера LIB1 в случае Ampli1 WGA). Когда продукты WGA секвенируют после получения библиотеки, в соответствии с наборами Ampli1 LowPass, вводят предсказуемую дополнительную длину, связанную с длинами

адаптеров и штрих-кодов для секвенирования, которые известны.

Воспроизводимость и уменьшенное представление генома

В способе в соответствии с настоящим изобретением, свойство DRS-WGA в комбинации со случайным получением библиотеки без фрагментации используют для получения уменьшенного представления генома (относительно исходного размера эталонного генома образцов), в результате чего для данных секвенирования с низким покрытием, для данного количества прочтений, увеличивается вероятность покрытия одинаковых фрагментов среди различных образцов, относительно случайного процесса, присущего WGA (например, как для способов WGA с использованием амплификации с множественным вытеснением цепи или DOP-ПЦР) и/или получению библиотеки для секвенирования (например, посредством случайной фрагментации или тагментации).

Иными словами, происходит детерминистическое взятие подвыборок из эталонного генома. Термин «детерминистический» является принципиальным, в том смысле, что - для любого данного количества прочтений - перекрывание в геномных локусах, покрытых среди двух парных образцов, является более высоким, таким образом, увеличивая количество высокополиморфных локусов, доступных для измерения сходства ДНК этих образцов.

Следует отметить, что способ является гибким в том смысле, что различные детерминистические ферменты могут являться подходящими, в зависимости от желательного разрешения и/или используемых платформы для секвенирования и протокола секвенирования. Например, можно использовать различные частощепящие ферменты. В примерах Ampli1 WGA, мотив TTAA представляет собой участок рестрикции. Другие разрезающие по четырем основаниям ферменты можно использовать для разрезания различных участков рестрикции, таких как GTAC, CTAG, с получением различного распределения фрагментов, позволяющего настраивать количество локусов, общих среди различных образцов для данного количества прочтений.

Когда продукты DRS-WGA впервые очищают после первичной ПЦР, происходит первый отбор по размеру, в результате чего более короткие фрагменты после WGA удаляют вместе со свободными праймерами. Обеспечивающим преимущество образом, в способе используют дополнительную стадию отбора. Эту дополнительную стадию отбора можно осуществлять посредством отбора по размеру определенных фрагментов после первичной WGA и/или получения библиотеки для массивного параллельного секвенирования посредством способа, ограничивающего поддающиеся секвенированию фрагменты. Например, для наборов Ampli1 LowPass включают присущую им стадию отбора по размеру, которая является достаточной, чтобы положительно влиять на способ. В WO2017/178655, проводят отбор по размеру в геле. В WO2019/016401, посредством последовательных стадий очистки с использованием бусин SPRI эффективно проводят первый отбор по размеру, в результате которого длину в парах оснований ограничивают до диапазона, по существу зависимого от концентрации бусин SPRI. Кроме того, в сам секвенатор может быть также введен отбор по размеру, поскольку для более длинных фрагментов получают

данные о последовательности с меньшей и меньшей эффективностью (например, из-за эффективности эмульсионной ПЦР в Ion Torrent, или мостиковой ПЦР для формирования кластеров на платформах Illumina).

В DRS-WGA также существует детерминистическая взаимосвязь между средним размером библиотеки для секвенирования и соотношением взятия подвыборок эталонного генома.

В анализе *in silico*, проведенном с расщеплением TTAA для эталонного генома человека hg19, получили всего приблизительно 19000000 фрагментов, включающих последовательности всех хромосом, которые можно перевести в 38000000 фрагментов в нормальном диплоидном геноме человека. В качестве примера, при отборе *in silico*, фрагменты в диапазоне 175-225 п.о. составляют только 1252559, покрывая приблизительно всего 248000000 оснований из 3090000000 оснований, т.е., 8,02% эталонного генома человека. См. таблицу 1 ниже, в которой количество фрагментов, общее количество пар оснований и кратность уменьшения (%) перечислены для различных диапазонах отбора по размеру. Это взятие подвыборок может быть обозначено кратность уменьшения (RR).

Таблица 1

Кратность уменьшения, в зависимости от отбора по размеру фрагментов

Диапазон	**Кол. фрагментов**	**Всего п.о.**	**Кратность уменьшения**
75-125	3057163	298483600	9,64
175-225	1252559	248367191	8,02
275-325	703011	210389610	6,80
375-425	340419	155603924	5,03
475-525	217861	108653407	3,51
725-775	68581	51428399	1,66
975-1025	24091	24070638	0,78

В предпочтительном варианте осуществления настоящего изобретения, целью является получение хорошего разрешения для показателя попарного сходства среди образцов. Для увеличения разрешения для данного количества прочтений, которое может быть доступно для каждого образца (связанного со стоимостью секвенирования на образец), перекрывание покрытых пар оснований между любыми двумя образцами является важным, поскольку сравнивают только области, покрытые в обоих образцах. Таким образом, увеличение диапазона пар оснований секвенируемых фрагментов может способствовать уменьшению разнообразия фрагментов, увеличивая перекрывание между различными образцами.

Однако, существуют компромиссы, в зависимости от применения. В конкретных вариантах осуществления настоящего изобретения, помимо идентификации происхождения ДНК образца, данные полногеномного секвенирования с низким

покрытием служат также двойной цели получения полногеномного профиля количества копий собственно образцов, как в случае применения для NIPD или для свободной от клеток использованной культуральной среды для эмбрионов.

В этом случае, диапазон фрагментов сходной ширины, но центрированный на более коротких фрагментах, увеличивает разнообразие и может приводить к получению лучших результатов и разрешения для операции, обозначающей количество копий, поскольку присутствует более высокое количество фрагментов, вносящих вклад в счет прочтений в данном геномном окне.

Отбор по размеру фрагментов

Различные способы отбора по размеру можно также использовать для достижения желательной кратности уменьшения, в зависимости от избранных количества прочтений секвенирования на образец и/или разрешения. Для данной средней длины фрагментов - меньшее или большее общее количество фрагментов можно получать, выбирая, соответственно, меньшую или большую полосу, центрированную на этой средней длине фрагментов.

Устройства, подобные Pipping prep (Sage Science), можно использовать для получения более строгого контроля распределения длины фрагментов и, с использованием аналогии с полосовыми фильтрами, также для получения более высокого фактора Q, определенного как

Q=Fцентр./ДельтаF=[(Fмин.+FМАКС.)/2]/(FМАКС.-Fмин.),

где

Fцентр.=(Fмин.+FМАКС.)/2 представляет собой средний размер фрагментов

ДельтаF=FМАКС.-Fмин. представляет собой ширину диапазона размеров фрагментов

Fмин. представляет собой размер фрагментов, ниже которого фрагменты представлены на общепринятом относительном уровне (например, 1/10=10%) или менее, в отношении нормализованного, внутриполосного, пикового количества фрагментов на интервал.

FМАКС. представляет собой размер фрагментов, выше которого фрагменты представлены на таком же общепринятом относительном уровне или менее, в отношении нормализованного, внутриполосного пикового количества фрагментов на интервал.

С использованием секвенирования Illumina, способ секвенирования представляет собой, предпочтительно, секвенирование спаренных концов, поскольку покрытие генома увеличивается и таким образом, количество локусов на миллион прочитанных пар увеличивается, увеличивая разрешение. Однако, когда размер, выбранный для секвенирования, опускается ниже определенного размера, секвенирование спаренных концов не может увеличивать покрытие, поскольку два парных прочтения перекрываются полностью.

Для секвенирования Ion Torrent, более высокие длины прочтений могут пропорционально увеличивать покрытый геном и таким образом, количество локусов на

миллион прочтений увеличивается, увеличивая разрешение. Для набора Ampli1 LowPass IonTorrent (Menarini Silicon Biosystems), снабженные штрих-кодом пулированные образцы отбирают по размеру, в геле или с использованием других способов, подобных Pippin Prep. Выбор различных фактора Q и средней длины фрагментов могут обеспечивать различные разрешения на основе миллиона прочтений.

Одним из преимуществ пулирования образцов и впоследствии отбора по размеру библиотеки для секвенирования является то, что все образцы имеют одинаковое распределение длин фрагментов, и в свою очередь, это может максимизировать перекрывание покрытого генома среди различных образцов, как необходимо для обеспечения более высокого количества высокополиморфных локусов для сравнения.

С другой стороны, при использовании набора Ampli1 LowPass для Illumina, различные библиотеки LowPass сначала отбирают по размеру и затем пулируют, получая немного различные отборы по размеру среди различных образцов, таким образом, уменьшая покрытый геном среди различных образцов.

Отбор по размеру после пулирования библиотеки, хотя и не является обязательным по стандартному протоколу, может быть использован для увеличения перекрывания среди образцов, которое может обеспечивать преимущества в анализе на основе контроля.

Однако, является важным, что присутствует перекрывание между распределением секвенированных фрагментов DRS-WGA среди различных образцов, поскольку уменьшение перекрывания распределения фрагментов может уменьшать количество общих полиморфных локусов для оценки показателя попарного сходства, в свою очередь, уменьшая разрешение способа.

В соответствии с настоящим изобретением, комбинирование DRS-WGA и LPWGS приводит к уменьшенному представлению из ввода образцов. Посредством секвенирования с использованием NGS, это уменьшает представление библиотек эталонного генома, в свою очередь, сокращает покрытый геном в избранном (или по иным причинам поддающемся секвенированию) диапазоне пар оснований, и получают эффективно более высокое перекрывание покрытого генома среди различных образцов, на основании прочтений.

Этот эффект можно использовать, в соответствии с настоящим изобретением, различными способами, в зависимости от ситуации.

Предпочтительно, получение библиотеки после DRS-WGA представляет собой один из способов, описанных в WO2017/178655 или WO2019/016401.

Установление порога показателя сходства и обозначение идентичности

Необязательно, можно устанавливать порог показателя сходства, полученного на предыдущих стадиях, для определения классов образцов. В большинстве случаев, количество полиморфных локусов, доступных для сравнения среди двух образцов, увеличивается при более высокой глубине прочтения. Чтобы позволять установление порога показателя сходства с использованием предварительно вычисленного значения, количество картированных прочтений в каждом образце, предпочтительно, нормализуют на фиксированное количество прочтений. Такую нормализацию проводят посредством

случайного отбора образцов прочтений, картирования на эталонном геноме, пока не будет достигнуто желательное количество (предпочтительно, лежащее в диапазоне, простирающемся от 100000 картированных прочтений до 10000000 картированных прочтений).

В предпочтительном варианте осуществления настоящего изобретения, «собственной» взаимосвязь между двумя образцами называют, если показатель сходства является более высоким, чем первый выбранный порог.

В предпочтительном варианте осуществления настоящего изобретения, «неродственной» взаимосвязь между двумя образцами называют, если показатель сходства является более низким, чем второй выбранный порог.

При применении для неинвазивной пренатальной диагностики, «родственной» взаимосвязь между двумя образцами называют, если показатель сходства содержится между третьим порогом, равным или более низким, чем указанный первый порог, и четвертым порогом, равным или более высоким, чем указанный второй порог.

При применении для криминалистической идентификации человека, «смешанной» взаимосвязь между двумя образцами называют, если показатель сходства содержится между третьим порогом, равным или более низким, чем указанный первый порог, и четвертым порогом, равным или более высоким, чем указанный второй порог.

Заявление о соответствии Статье 170bis(2) Итальянского кодекса интеллектуальной собственности

Биологический материал человеческого происхождения, используемый в настоящем описании, был получен в соответствии с применимыми правовыми положениями.

ФОРМУЛА ИЗОБРЕТЕНИЯ

1. Способ анализа степени сходства по меньшей мере двух образцов в множестве образцов, содержащих геномную ДНК, включающий стадии:

a) получения множества образцов, содержащих геномную ДНК;

b) проведения, отдельно для каждого образца, полногеномной амплификации с детерминистическим участком рестрикции (DRS-WGA) указанной геномной ДНК;

c) получения библиотеки для массивного параллельного секвенирования с использованием реакции ПЦР с адаптером для секвенирования/праймером для слияния WGA, без фрагментации, из каждого продукта указанной DRS-WGA;

d) проведения полногеномного секвенирования с низким покрытием при средней глубине покрытия < 1x для указанной библиотеки для массивного параллельного секвенирования;

e) выравнивания для каждого образца прочтений, полученных на стадии d), на эталонном геноме;

f) извлечения для каждого образца аллельного содержания в множестве полиморфных локусов;

g) расчета показателя попарного сходства для по меньшей мере двух образцов как функции от аллельного содержания, измеренного в указанном множестве локусов;

h) определения степени сходства по меньшей мере двух образцов на основании показателя сходства.

2. Способ по п.1, где указанное полногеномное секвенирование с низким покрытием проводят при покрытии < 0,01x, предпочтительно, при покрытии < 0,05x, более предпочтительно, при покрытии < 0,1x, даже более предпочтительно, при покрытии < 0,5x.

3. Способ по п.1 или 2, где указанное множество полиморфных локусов содержит полиморфные локусы со средней гетерозиготностью > 0,499, предпочтительно, со средней гетерозиготностью > 0,49, более предпочтительно, со средней гетерозиготностью > 0,4, даже более предпочтительно, со средней гетерозиготностью > 0,3, наиболее предпочтительно, со средней гетерозиготностью > 0,2.

4. Способ по любому из пунктов 1-3, где указанное множество полиморфных локусов содержит > 200000 локусов, предпочтительно, > 300000 локусов, более предпочтительно, > 500000 локусов, даже более предпочтительно, > 1000000 локусов.

5. Способ по любому из пунктов 1-4, где указанный показатель попарного сходства рассчитывают посредством вычисления корреляции частоты B-аллеля среди локусов, покрытых по меньшей мере одним прочтением в по меньшей мере двух образцах.

6. Способ по любому из пунктов 1-4, где указанный показатель попарного сходства рассчитывают посредством вычисления среднего показателя согласованности среди локусов, покрытых по меньшей мере одним прочтением в обоих парных образцах, где показателю согласованности для каждого локуса приписывают одно из следующих значений:

A1) 1, если запрашиваемые аллели являются идентичными; и

B1) 0, если запрашиваемые аллели являются различными; или

A2) 1, если запрашиваемые аллели являются идентичными;

B2) 0, если запрашиваемые аллели являются полностью различными; и

C2) 0,5, если запрашиваемые аллели являются частично перекрывающимися.

7. Способ по любому из предшествующих пунктов, дополнительно включающий определение группы кластеров образцов, разделяющих общее свойство, выбранное из группы, состоящей из идентичности одного индивидуума (или более индивидуумов), вносящих значительный вклад ДНК в образцы из кластера, или свойства содержания недостаточных количеств ДНК, и/или свойства содержания высоко деградированной ДНК или ДНК непонятного происхождения.

8. Способ по п.7, где по меньшей мере два образца приписывают по меньшей мере одному кластеру посредством алгоритма с использованием, в качестве ввода, указанного показателя попарного сходства.

9. Способ по п.8, где алгоритм представляет собой алгоритм иерархической кластеризации.

10. Способ по п.8, где количество указанных кластеров рассчитывают посредством

a) выбора количества кластеров после первой итерации, максимизирующего средний коэффициент силуэта;

b) для каждого из указанных кластеров после первой итерации, вычисления коэффициента силуэта для каждого из указанных образцов, принадлежащих к кластеру после первой итерации, где образцы, принадлежащие к кластеру, имеющие коэффициент силуэта ниже, чем фиксированный порог, лежащий в диапазоне 0,19-0,21, приписывают новому кластеру.

11. Способ по п.10, где указанная группа кластеров содержит один или несколько кластеров идентичности, содержащих образцы, содержащие ДНК только от одного и того же индивидуума.

12. Способ по п.11, где, в присутствии большего количества кластеров идентичности, кардинальность указанного множества кластеров идентичности соответствует количеству индивидуальных вносящих вклад в ДНК участников в указанном множестве образцов.

13. Способ по любому из пунктов от 8 до 12, дополнительно включающий определение группы кластеров смешанной идентичности, где каждый из указанных кластеров смешанной идентичности содержит образцы, содержащие ДНК от по меньшей мере двух индивидуумов.

14. Способ по п.13, дополнительно включающий определение по меньшей мере одного кластера без распознавания, содержащего образцы, содержащие ДНК непонятного происхождения.

15. Способ по любому из пунктов от 8 до 14, где указанное множество образцов содержит по меньшей мере один эталонный образец, и указанная группа кластеров идентичности включает по меньшей мере один эталонный кластер, содержащий указанный

эталонный образец.

16. Способ по п.15, где указанный по меньшей мере один эталонный образец представляет собой образец от беременного индивидуума - родителя женского пола.

17. Способ по п.16, где указанная группа кластеров идентичности дополнительно содержит по меньшей мере один родственный кластер, состоящий из образцов из по меньшей мере одного плода из текущей беременности указанного индивидуума - родителя женского пола.

18. Способ по п.17, где указанный родственный кластер разбивают на множество фетальных кластеров, состоящих из образцов, содержащих ДНК только от одного и того же плода.

19. Способ по п.15, где указанный по меньшей мере один эталонный кластер состоит из образцов, содержащих ДНК только от одного и того же индивидуума, соответствующего потерпевшему в криминалистической экспертизе, дополнительно включающей определение по меньшей мере одного кластера виновного, содержащего образцы, содержащие ДНК только от одного и того же индивидуума, отличного от потерпевшего.

20. Способ по п.19, включающий смешивание по кластерам аликвот DRS-WGA из множества образцов, принадлежащих к каждому из указанных по меньшей мере одного из кластеров виновных, с получением для каждого кластера соответствующего образца WGA-ДНК отдельного индивидуума, и проведение дополнительного анализа ДНК для по меньшей мере одного из указанных образцов WGA-ДНК отдельного индивидуума.

21. Способ по п.19, включающий слияние по кластерам данных генетического анализа из по меньшей мере одного типа анализа, из множества образцов, принадлежащих к каждому из указанных по меньшей мере одного из кластеров виновных, с получением для каждого из указанных по меньшей мере одного из кластеров виновных соответствующих данных WGA-ДНК отдельного индивидуума.

22. Способ по п.21, где указанный тип анализа выбран из группы, состоящей из:

a) анализа микросателлитов;

b) анализа однонуклеотидного полиморфизма;

c) массивного параллельного нацеленного секвенирования; и

d) полногеномного секвенирования.

23. Способ по любому из пунктов 1-15, где указанное множество образцов содержит образцы опухоли и/или нормальные образцы.

24. Способ по п.1 или 15, где указанное множество образцов содержит по меньшей мере эталонный образец, содержащий ДНК от индивидуума - родителя женского пола, и по меньшей мере один другой эмбриональный образец из указанного множества образцов выбран из группы, состоящей из:

a) образцов, содержащих ДНК из эмбриона, происходящего от указанного индивидуума - родителя женского пола; и

b) образцов, содержащих ДНК из использованной культуральной среды для эмбриона, полученной от эмбриона от указанного индивидуума - родителя женского пола.

25. Способ по п.24, дополнительно включающий проведение преимплантационного генетического скрининга для указанного эмбриона посредством полногеномного анализа хромосомных аберраций по данным указанного полногеномного секвенирования с низким покрытием для указанного по меньшей мере одного другого эмбрионального образца с использованием коэффициента контаминации, соответствующего материнской контаминации, измеренного для указанного по меньшей мере одного другого эмбрионального образца как функция от указанного попарного сходства указанного по меньшей мере одного другого эмбрионального образца и указанного образца от индивидуума - родителя женского пола.

26. Способ по п.15, где указанное множество образцов содержит по меньшей мере эталонный образец, содержащий ДНК от индивидуума - родителя женского пола, и по меньшей мере один другой образец, содержащий ДНК из образца внеклеточной ДНК.

27. Способ по п.26, дополнительно включающий проведение неинвазивного пренатального тестирования для указанного образца внеклеточной ДНК посредством полногеномного анализа хромосомных аберраций по данным указанного полногеномного секвенирования с низким покрытием для указанного по меньшей мере одного образца внеклеточной ДНК с использованием поправочного коэффициента, соответствующего фетальной фракции, измеренного для указанного по меньшей мере одного образца внеклеточной ДНК как функция от указанного попарного сходства с эталонным образцом от родителя женского пола.

28. Способ по п.15, где указанное множество образцов содержит по меньшей мере эталонный образец, содержащий ДНК от индивидуума - родителя женского пола, и по меньшей мере один другой пренатальный образец, содержащий ДНК из ворсин хориона, амниотической жидкости или продуктов зачатия.

29. Способ по п.28, дополнительно включающий проведение пренатального тестирования для указанных пренатальных образцов посредством полногеномного анализа хромосомных аберраций по данным указанного полногеномного секвенирования с низким покрытием для указанного по меньшей мере одного пренатального образца с использованием поправочного коэффициента, соответствующего материнской или экзогенной контаминации, измеренного для указанного по меньшей мере одного пренатального образца как функция от указанного попарного сходства с эталонным образцом от родителя женского пола.

30. Способ по п.15, в частности, для аутентификации линии клеток, где множество эталонных кластеров получают из множества образцов ДНК из линий клеток, и указанная группа кластеров идентичности дополнительно содержит по меньшей мере один из образцов из линии клеток, подлежащей аутентификации.

31. Способ по п.15, в частности, для исследования аллотрансплантатов, где указанный по меньшей мере один эталонный кластер состоит из образцов, содержащих ДНК зародышевой линии от подвергаемого трансплантации пациента, и указанная группа кластеров идентичности дополнительно содержит один кластер донора, состоящий из

образцов от аллогенного донора для указанного подвергаемого трансплантации пациента.

32. Способ по п.17, в частности для неинвазивного установления отцовства, где указанный по меньшей мере один эталонный образец содержит эталонный образец от родителя мужского пола, содержащий ДНК только от указанного родителя мужского пола, и указанный по меньшей мере один эталонный кластер дополнительно содержит кластер идентичности родителя мужского пола, включающий указанный образец от родителя мужского пола, где:

(i) если показатель сходства родственного образца, в отношении образца от родителя мужского пола, согласуется с кровным родством, отцовство подтверждают

(ii) если показатель сходства родственного образца, в отношении образца от родителя мужского пола, согласуется с неродственным индивидуумом, отцовство не подтверждают.

33. Способ по п.17, в частности, для неинвазивной оценки молярной беременности, где указанный по меньшей мере один образец содержит по меньшей мере один образец циркулирующих трофобластных клеток и где, если показатель сходства указанного образца трофобластных клеток, в отношении образцов от родителя женского пола, согласуется с неродственными образцами, подтверждают полный занос.

34. Способ по п.33, где указанный по меньшей мере один образец содержит множество образцов трофобластных клеток и где:

(i) если показатель сходства среди указанных образцов трофобластных клеток превышает ожидаемый 99-й процентиль ожидаемого показателя сходства для собственных образцов, подтверждают гомозиготный отцовский занос P1P1.

(ii) если показатель сходства среди указанных образцов трофобластных клеток согласуется с ожидаемым показателем сходства для собственных образцов, подтверждают гетерозиготный отцовский занос P1P2.

35. Способ по п.30, где указанный по меньшей мере один образец дополнительно содержит образец от родителя мужского пола, и показатель сходства среди указанных образцов трофобластных клеток согласуется с ожидаемым показателем сходства для собственных образцов, где:

(i) если показатель сходства указанных образцов трофобластных клеток, в отношении образца от родителя мужского пола, согласуется с ожидаемым показателем сходства для собственных образцов, подтверждают гетерозиготный отцовский занос P1P2.

(ii) если показатель сходства указанных образцов трофобластных клеток, в отношении образца от родителя мужского пола, ниже чем 1-й процентиль ожидаемого показателя сходства для собственных образцов, не подтверждают гетерозиготный отцовский занос P1P2.

36. Способ по любому из пунктов 1-6, дополнительно включающий классификацию образцов, выбранных из множества образцов, на основе предопределенных классов, с использованием машинообучаемого классификатора с использованием, в качестве ввода, указанного показателя попарного сходства.

37. Способ по п.36, где машинообучаемый классификатор представляет собой классификатор типа случайного леса.

38. Способ по п.36 или 37, где машинообучаемый классификатор использует, в качестве дополнительного ввода, по меньшей мере одно значение, измеренное по указанным данным полногеномного секвенирования с низким покрытием, выбранное из группы, содержащей:

a) DLRS: производное логарифма отношения разброса;

b) R50: процент фрагментов WGA, покрытых посредством 50% секвенированных прочтений, среди всех фрагментов WGA, покрытых по меньшей мере одним прочтением;

c) YFRAC: доля прочтений, картированных на хромосоме Y;

a) Аберрантный: процент генома, соответствующий добавлениям или потерям, относительно медианной плоидности клеток;

b) Xp. 13: плоидность хромосомы 13;

c) Xp. 18: плоидность хромосомы 18;

d) Xp. 21: плоидность хромосомы 21;

e) RSUM: среднее абсолютное отклонение от ближайшего целочисленного уровня количества копий, рассчитанное для события аберрации количества копий с наивысшим абсолютным отклонением от медианной плоидности клеток;

f) Смеш._-показатель: z-показатель по RSUM, рассчитанный для события аберрации количества копий с наивысшим абсолютным отклонением от медианной плоидности клеток; и

g) Дегр._-показатель: количество событий небольшой потери (< 10 млн.п.о., которая является распространенной в деградированных образцах).

39. Способ по любому из пунктов от 36 до 38, где по меньшей мере один из образцов представляет собой эталонный образец.

40. Способ по п.39, где указанный по меньшей мере один эталонный образец содержит образец от беременного индивидуума - родителя женского пола.

41. Способ по п.40, где указанное множество образцов содержит по меньшей мере один образец, классифицированный как «родственный», в отношении эталона родителя женского пола, представляющий образец из плода из текущей беременности указанного индивидуума - родителя женского пола.

42. Способ по п.39, где указанный по меньшей мере один эталонный образец представляет собой образец, содержащий ДНК только от одного и того же индивидуума, соответствующего потерпевшему в криминалистической экспертизе, дополнительно включающей определение по меньшей мере одной группы виновного, представленной всеми образцами, классифицированными как «не собственные», в отношении эталонных образцов, и классифицированными как «собственные» в отношении друг друга, содержащей образцы, содержащие ДНК только от одного и того же индивидуума, отличного от потерпевшего.

43. Способ по п.42, включающий смешивание по группам аликвот DRS-WGA из

множества образцов, принадлежащих к каждой из указанных по меньшей мере одной из групп одного виновного, с получением для каждой группы одного виновного соответствующего образца WGA-ДНК отдельного индивидуума, и проведением дополнительного анализа ДНК для по меньшей мере одного из указанных образцов WGA-ДНК отдельного индивидуума.

44. Способ по п.42, включающий слияние по группам данных генетического анализа из по меньшей мере одного типа анализа, из множества образцов, принадлежащих к каждой из указанных по меньшей мере одной из групп одного виновного, с получением для каждой из указанной по меньшей мере одной группы одного виновного соответствующих данных WGA-ДНК отдельного индивидуума.

45. Способ по любому из пунктов 36-39, где указанное множество образцов содержит образцы опухоли и/или нормальные образцы.

46. Способ по любому из пунктов 36-39, где указанное множество образцов содержит по меньшей мере эталонный образец, содержащий ДНК от индивидуума - родителя женского пола, и по меньшей мере один другой эмбриональный образец, классифицированный как «не собственный» в отношении эталона - родителя женского пола, из указанного множества образцов выбран из группы, состоящей из:

a) образцов, содержащих ДНК из эмбриона, происходящего от указанного индивидуума - родителя женского пола; и

b) образцов, содержащих ДНК из использованной культуральной среды для эмбриона, полученной от эмбриона от указанного индивидуума - родителя женского пола.

47. Способ по п.46, дополнительно включающий проведение преимплантационного генетического скрининга для указанного эмбриона посредством полногеномного анализа хромосомных аберраций по данным указанного полногеномного секвенирования с низким покрытием для указанного по меньшей мере одного другого эмбрионального образца с использованием коэффициента контаминации, соответствующего материнской контаминации, измеренного для указанного по меньшей мере одного другого эмбрионального образца как функция от указанного попарного сходства указанного по меньшей мере одного другого эмбрионального образца и указанного образца от индивидуума - родителя женского пола.

48. Способ по п.39, где множество эталонных групп получают из множества образцов ДНК из линий клеток, и указанное множество образцов дополнительно содержит по меньшей мере один образец из линии клеток, подлежащей аутентификации.

49. Способ по п.39, где указанная по меньшей мере одна эталонная группа содержит образцы, содержащие ДНК зародышевой линии от подвергаемого трансплантации пациента, и указанное множество образцов дополнительно содержит один образец от донора, представляющий собой по меньшей мере один образец от аллогенного донора для указанного подвергаемого трансплантации пациента.

50. Способ по п.41, в частности, для неинвазивного установления отцовства, где указанный по меньшей мере один эталонный образец дополнительно содержит эталонный

образец от родителя мужского пола, содержащий ДНК только от указанного родителя мужского пола, и указанное множество образцов дополнительно содержит образцы, где: (i) отцовство подтверждают, если их классифицируют как «собственные», в отношении эталонного образца от родителя мужского пола

(ii) отцовство не подтверждают, если их классифицируют как «неродственные», в отношении эталонного образца от родителя мужского пола.

51. Способ по п.40, в частности, для неинвазивной оценки молярной беременности, где указанный по меньшей мере один образец содержит по меньшей мере один образец циркулирующих трофобластных клеток и где, если указанный образец трофобластных клеток классифицируют как «неродственный», в отношении эталона - родителя женского пола, подтверждают полный пузырный занос отцовского происхождения.

52. Способ по п.51, где указанный по меньшей мере один образец содержит множество образцов трофобластных клеток, которые классифицируют как «собственные» в отношении друг друга, и где:

(i) если их показатель сходства превышает ожидаемый 99-й процентиль ожидаемого показателя сходства для «собственных» образцов, подтверждают гомозиготный пузырный занос отцовского происхождения P1P1.

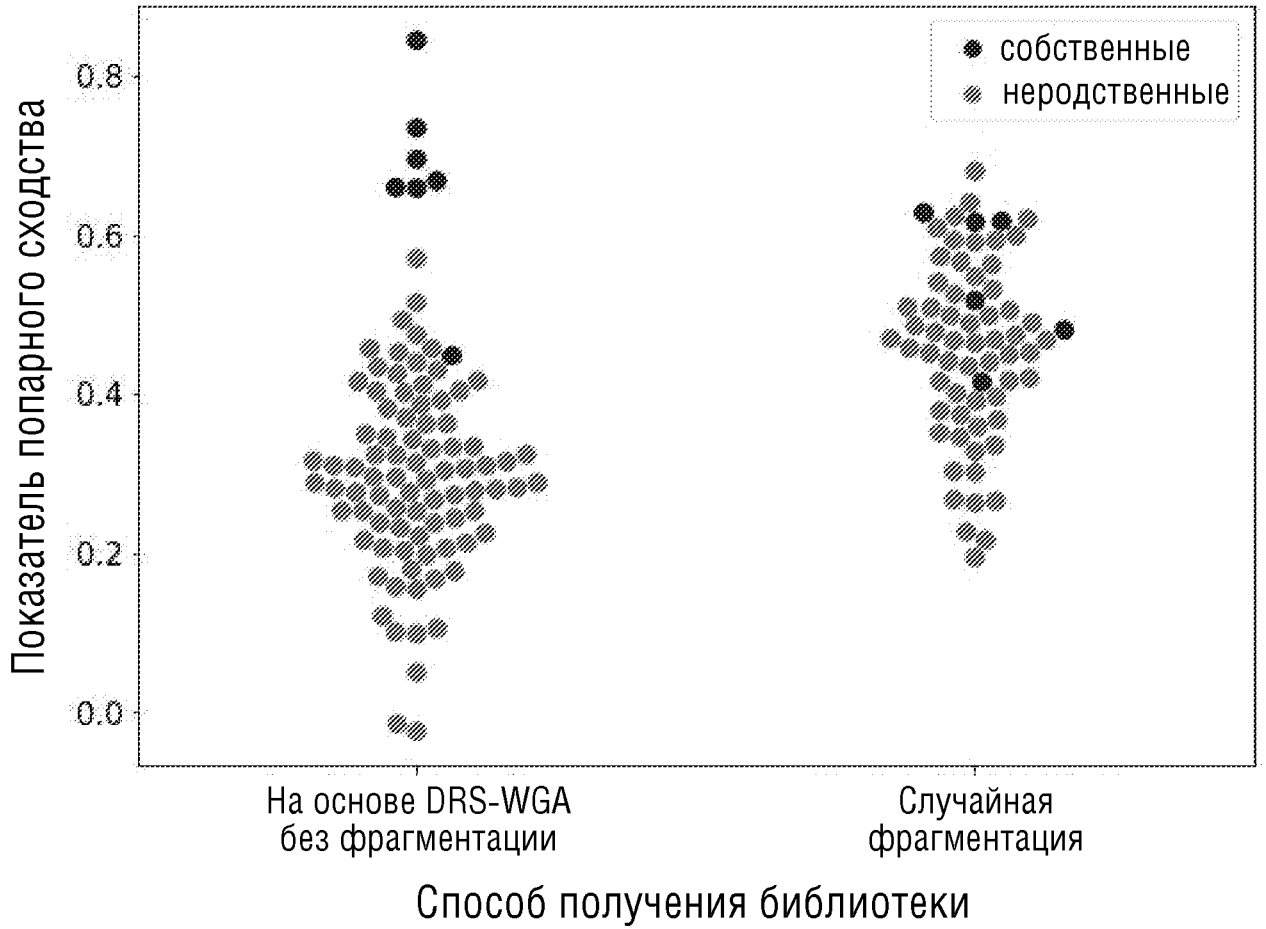
(ii) если их показатель сходства согласуется с ожидаемым показателем сходства для «собственных» образцов, подтверждают гетерозиготный пузырный занос отцовского происхождения P1P2.

53. Способ по п.52, где указанный по меньшей мере один образец дополнительно содержит образец от родителя мужского пола, где указанный образец от родителя мужского пола классифицируют как «собственный», в отношении по меньшей мере одного образца из указанного множества образцов трофобластных клеток, и где:

(i) если показатель сходства указанных образцов трофобластных клеток, в отношении образца от родителя мужского пола, согласуется с ожидаемым показателем сходства для «собственных» образцов, подтверждают гетерозиготный пузырный занос отцовского происхождения P1P2.

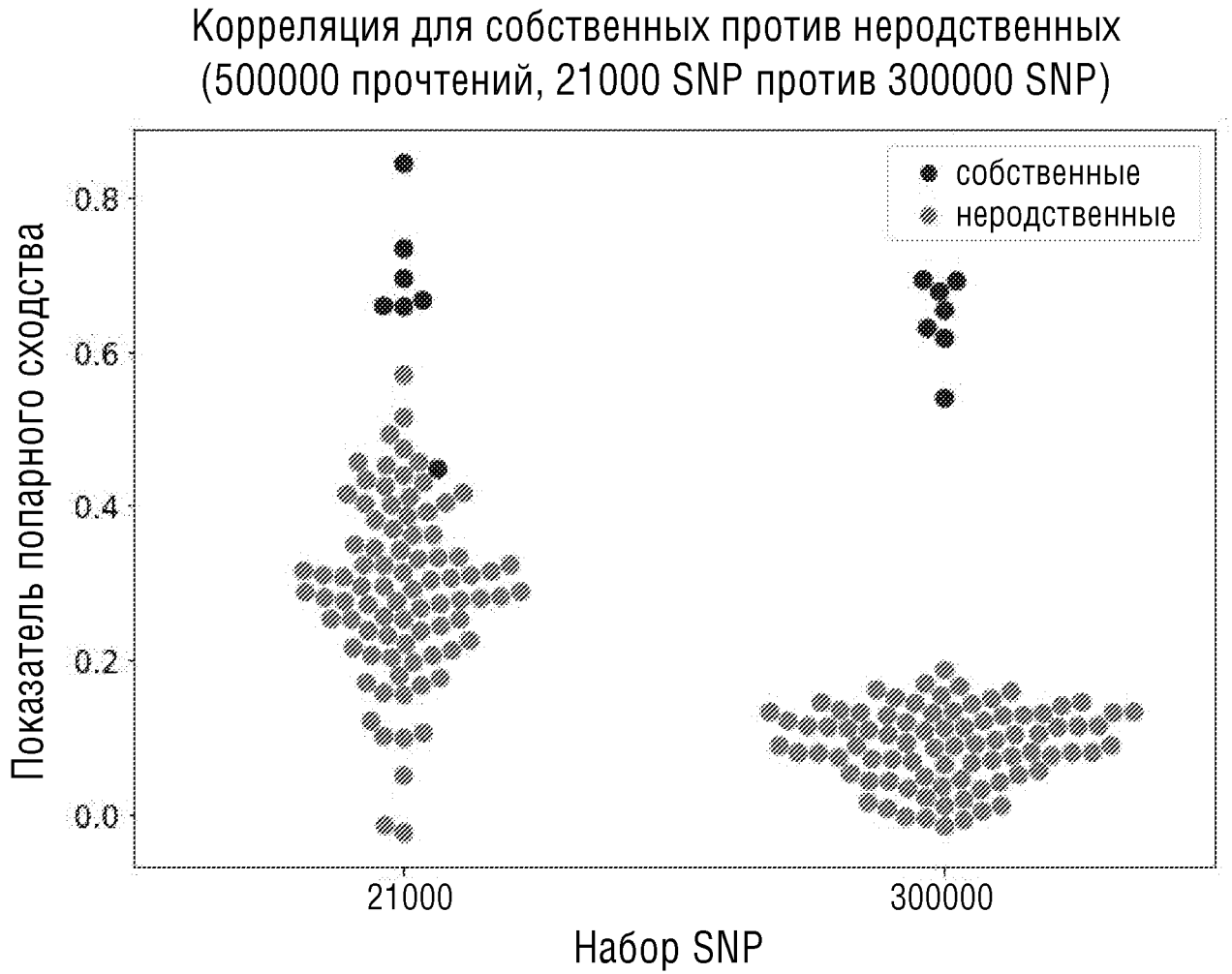
(ii) если показатель сходства указанных образцов трофобластных клеток, в отношении образца от родителя мужского пола, ниже чем 1-й процентиль ожидаемого показателя сходства для «собственных» образцов, не подтверждают гетерозиготный пузырный занос отцовского происхождения P1P2.

ФИГ.1

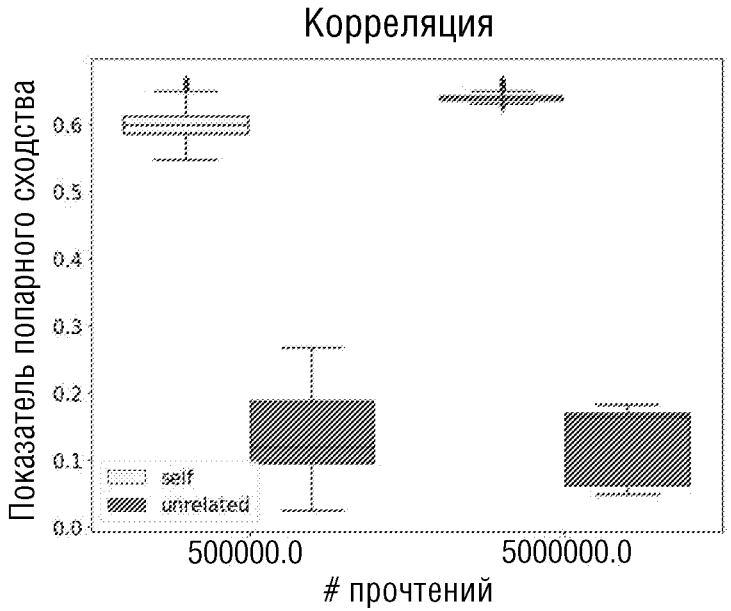

собственные
неродственные
Показатель попарного сходства
0.0
0.2
0.4
0.6
0.8
На основе DRS-WGA
без фрагментации
Случайная
фрагментация
Способ получения библиотеки

580781

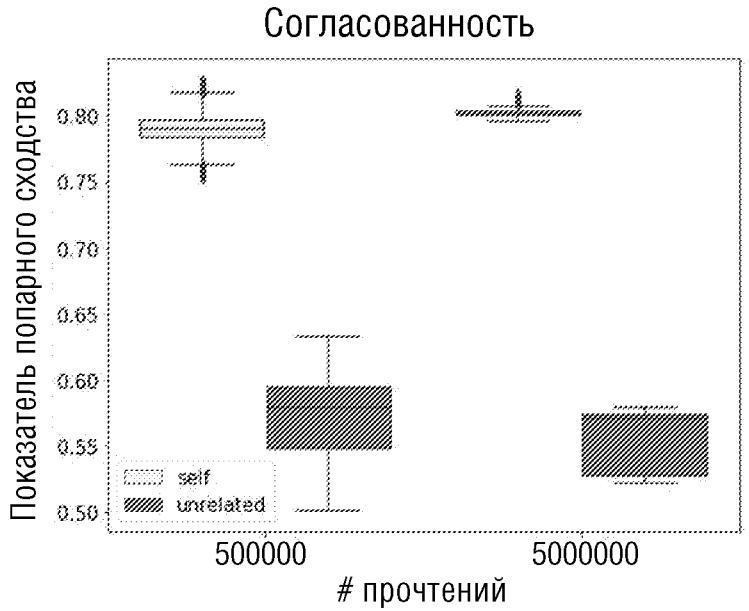
ФИГ.2


Корреляция для собственных против неродственных
(500000 прочтений, 21000 SNP против 300000 SNP)
собственные
неродственные
Показатель попарного сходства
0.0
0.2
0.4
0.6
0.8
21000
300000
Набор SNP

ФИГ.3А

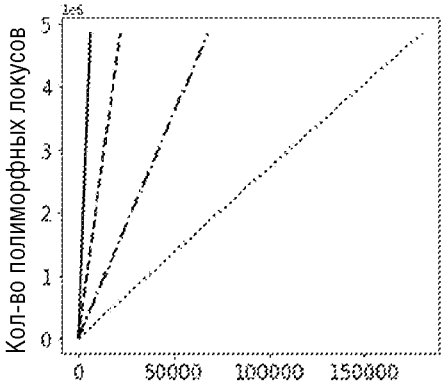


ФИГ.3В

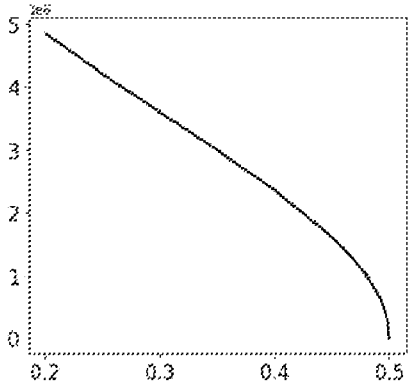


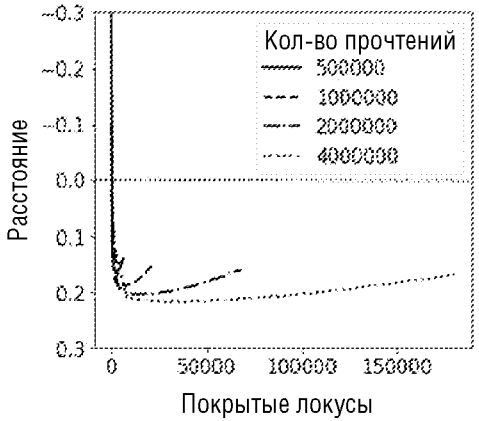
Пороги средн. гетерозиг.:
500000 прочтений: 0,46
5000000 прочтений: 0,49

ФИГ.4B



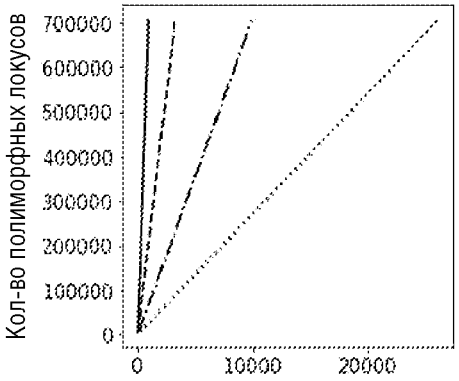
ФИГ.4A



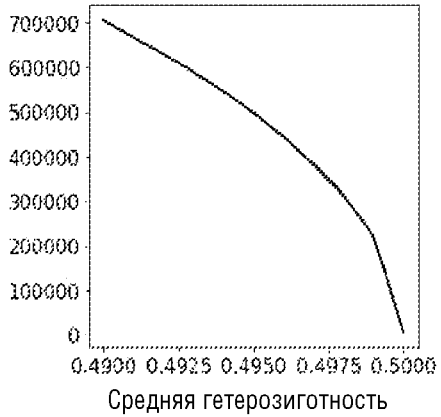


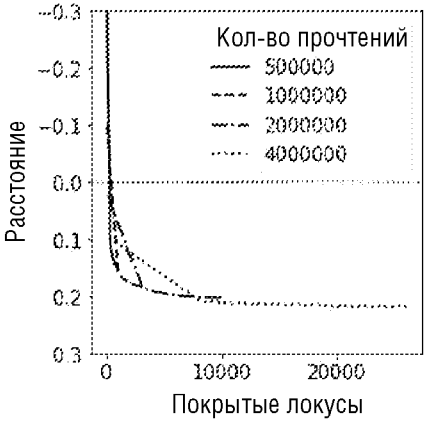
ФИГ.4C

ФИГ.4E



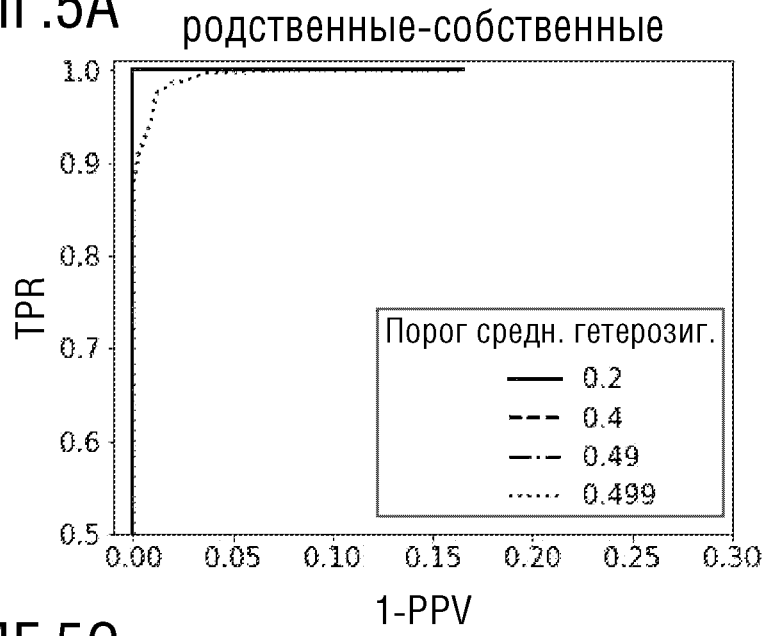
ФИГ.4D



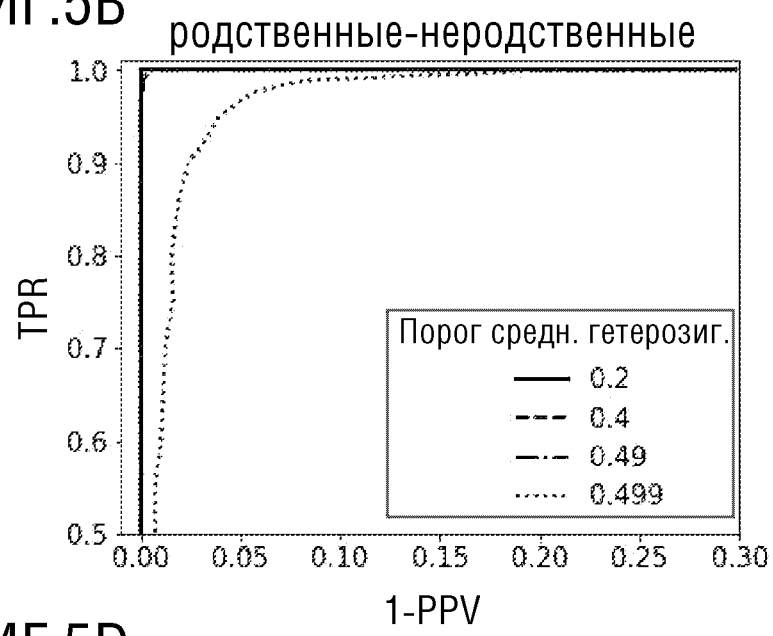


ФИГ.4F

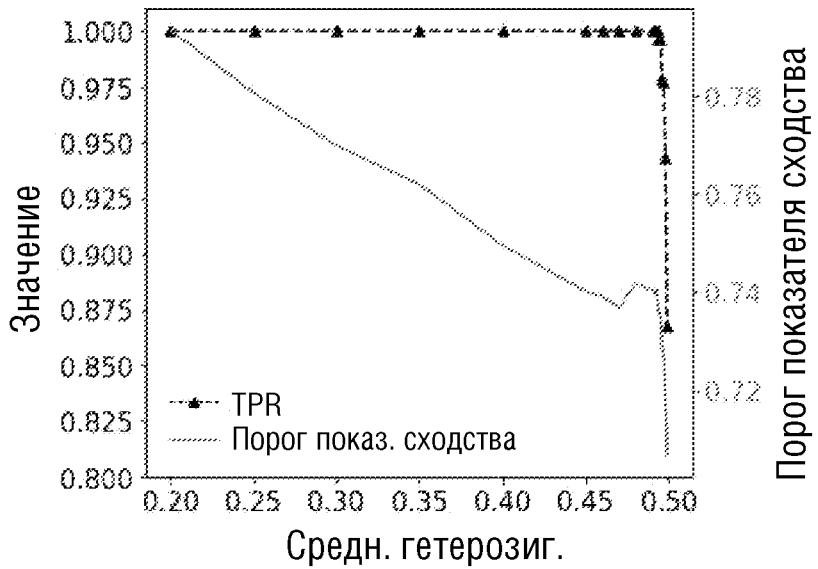
ФИГ.5A



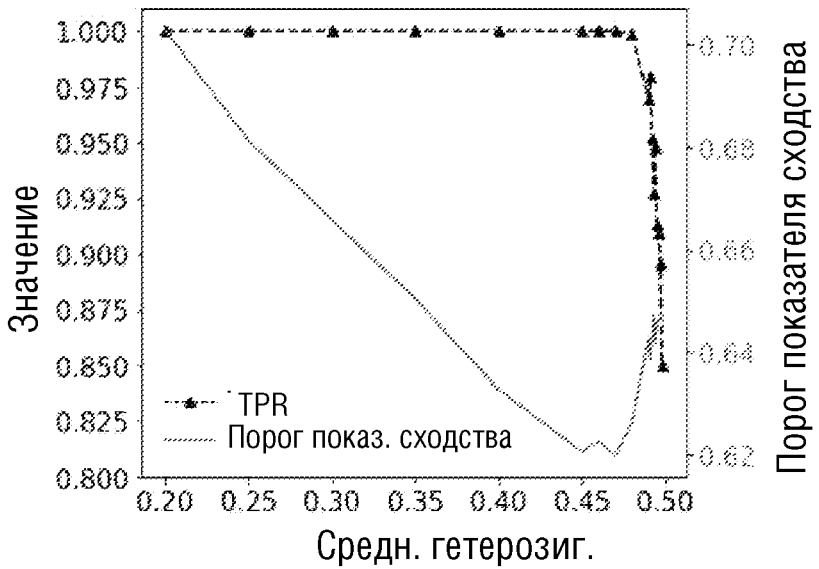
ФИГ.5B



ФИГ.5C



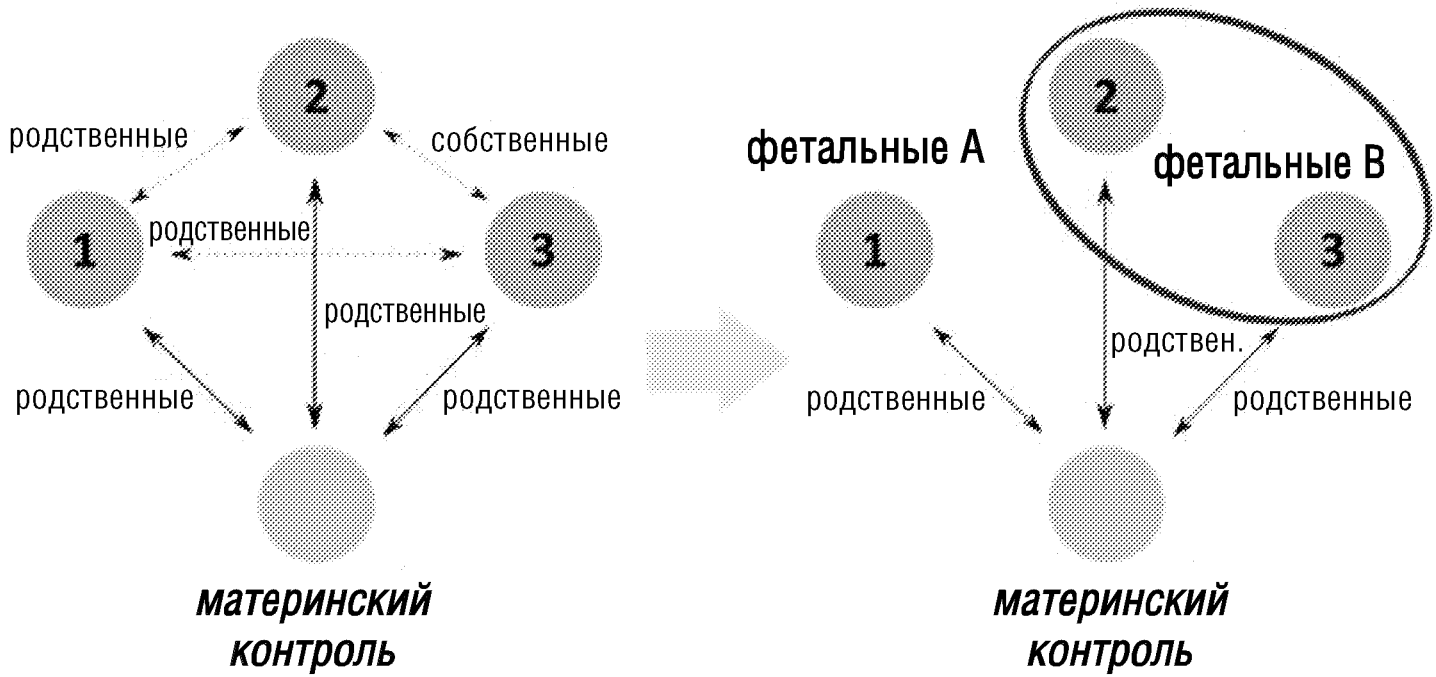
ФИГ.5D



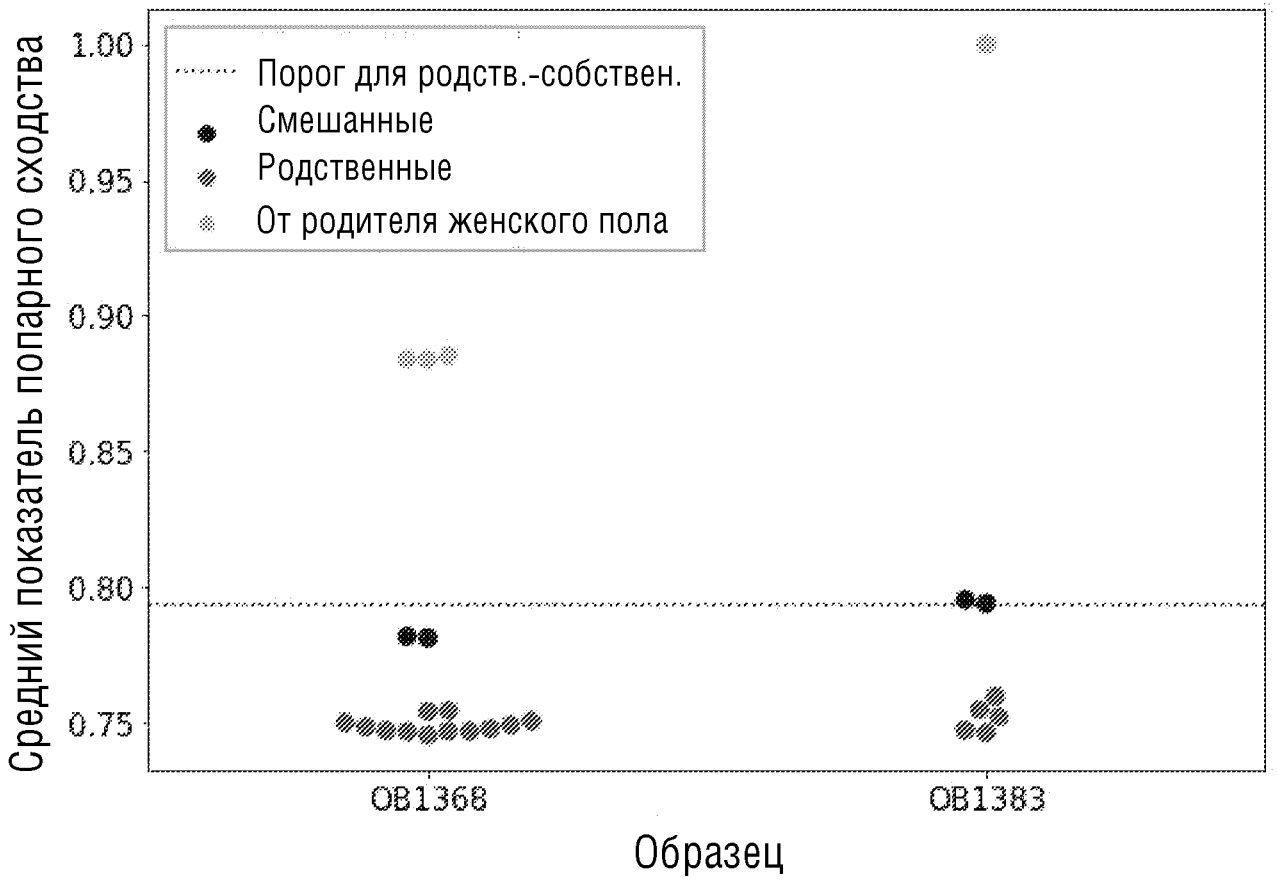
ФИГ.6

Пороги при различной средн. гетерозиг., для которой PPV > 0.999

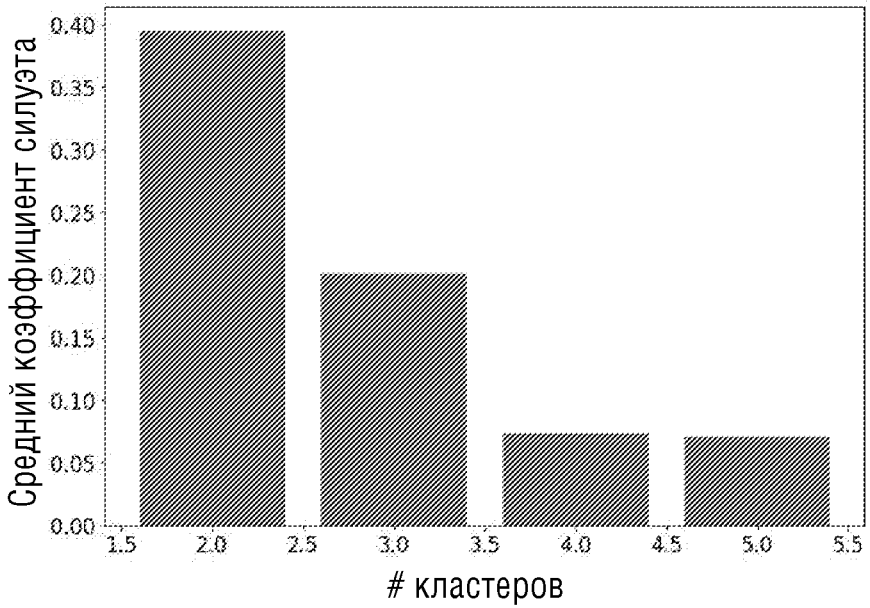
ФИГ.7


2
родственные
собственные
1
родственные
3
родственные
родственные
родственные
материнский контроль
фетальные А
2
фетальные В
1
3
родствен.
родственные
родственные
материнский контроль

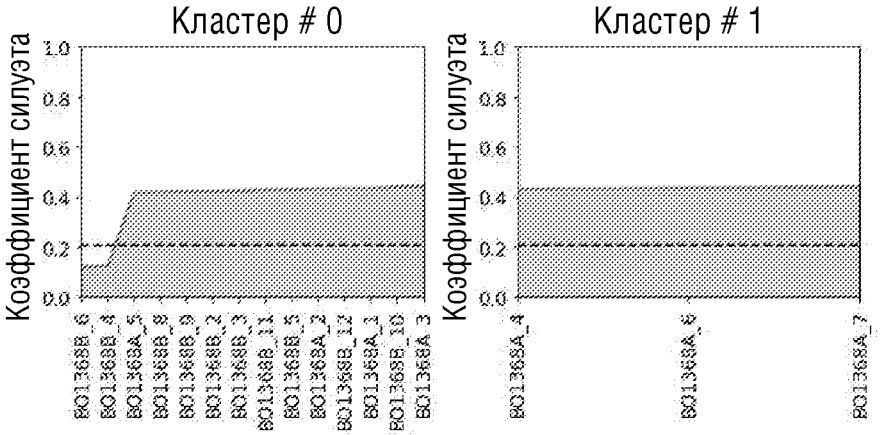
ФИГ.8


Средний показатель попарного сходства
1.00
0.95
0.90
0.85
0.80
0.75
Порог для родств.-собствен.
Смешанные
Родственные
От родителя женского пола
OB1368
OB1383
Образец

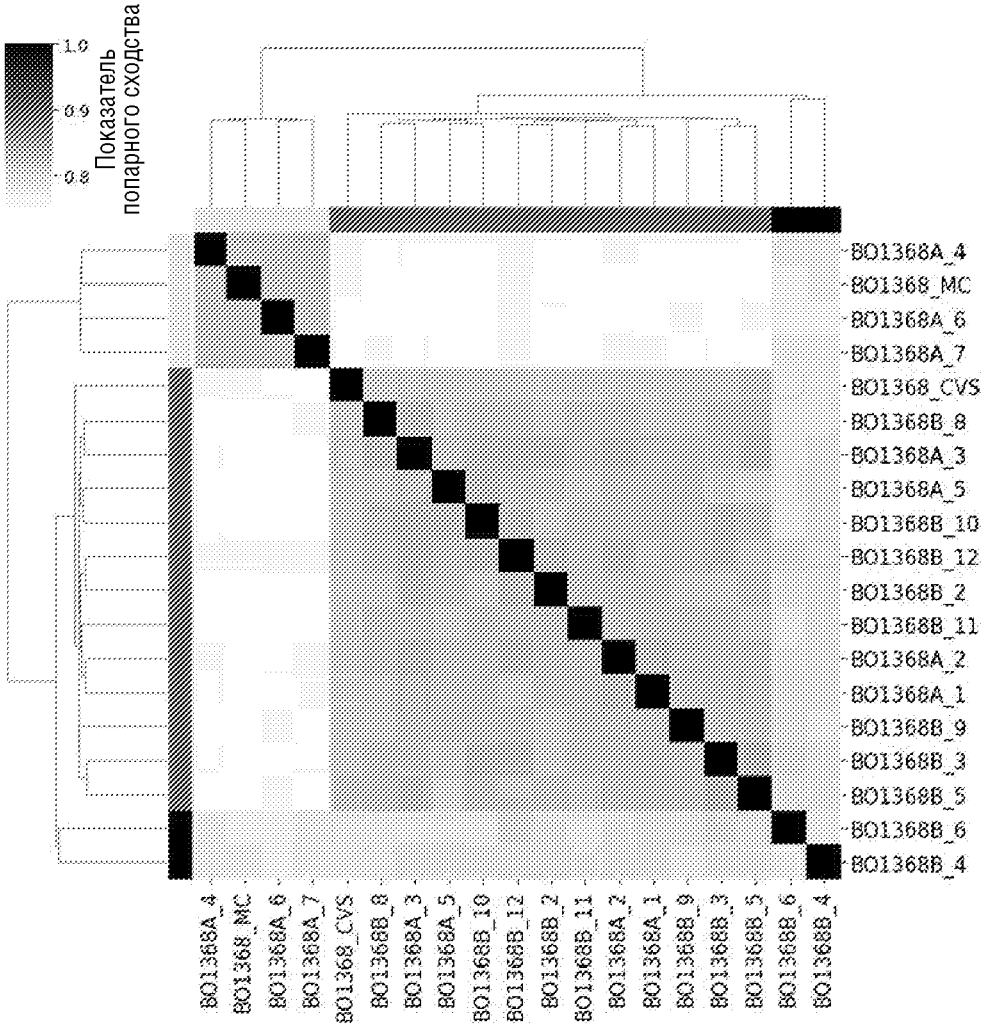
ФИГ.9А



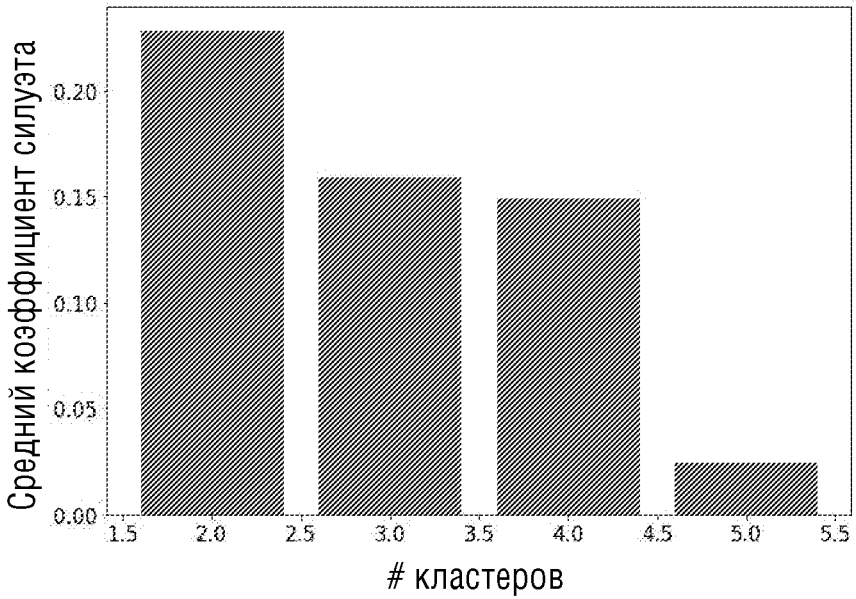
ФИГ.9В



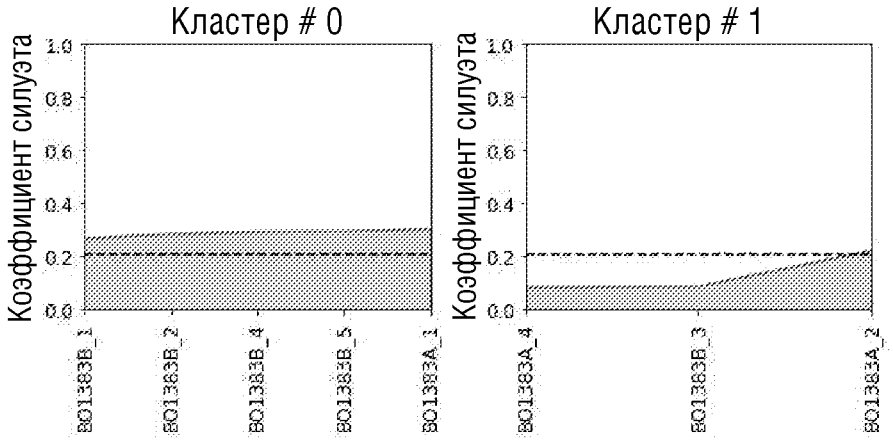
ФИГ.9С



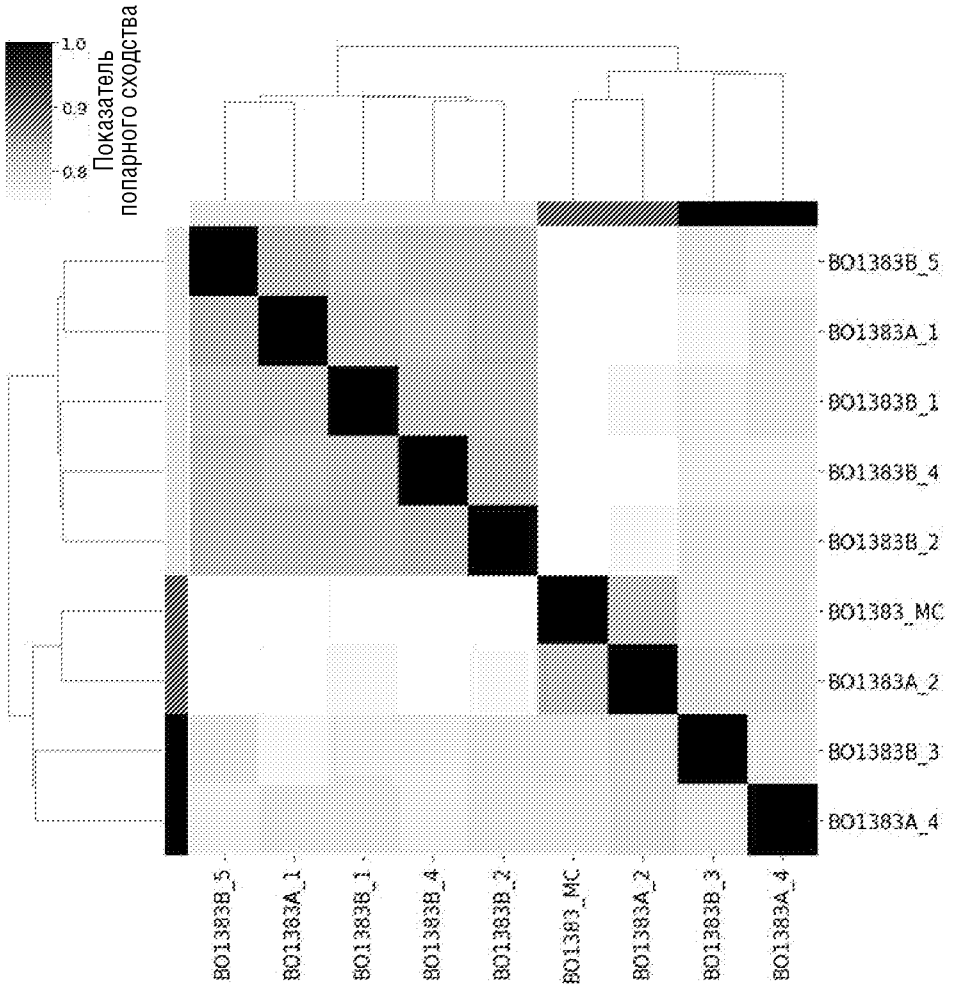
ФИГ.10A



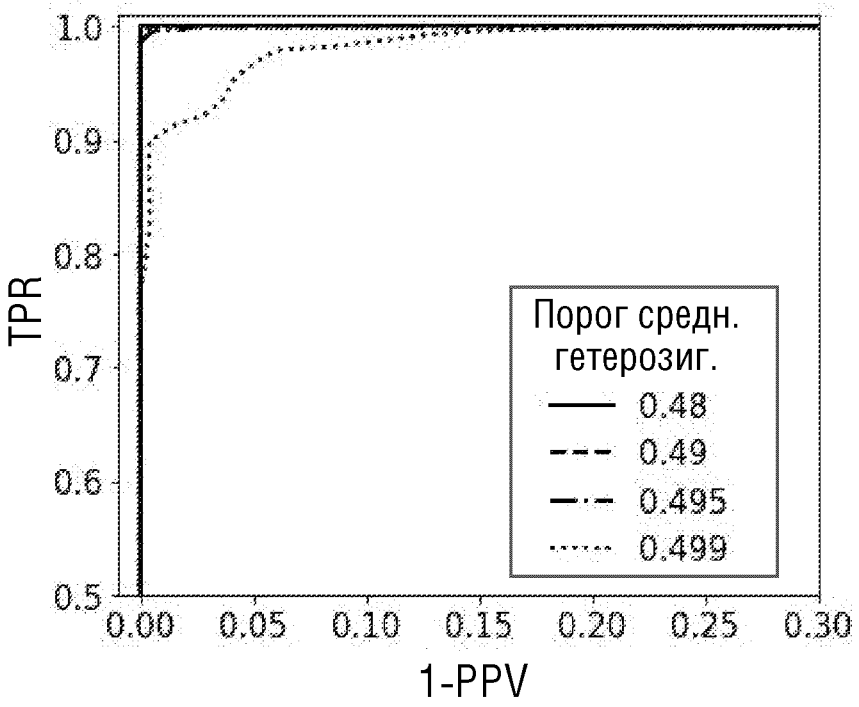
ФИГ.10B



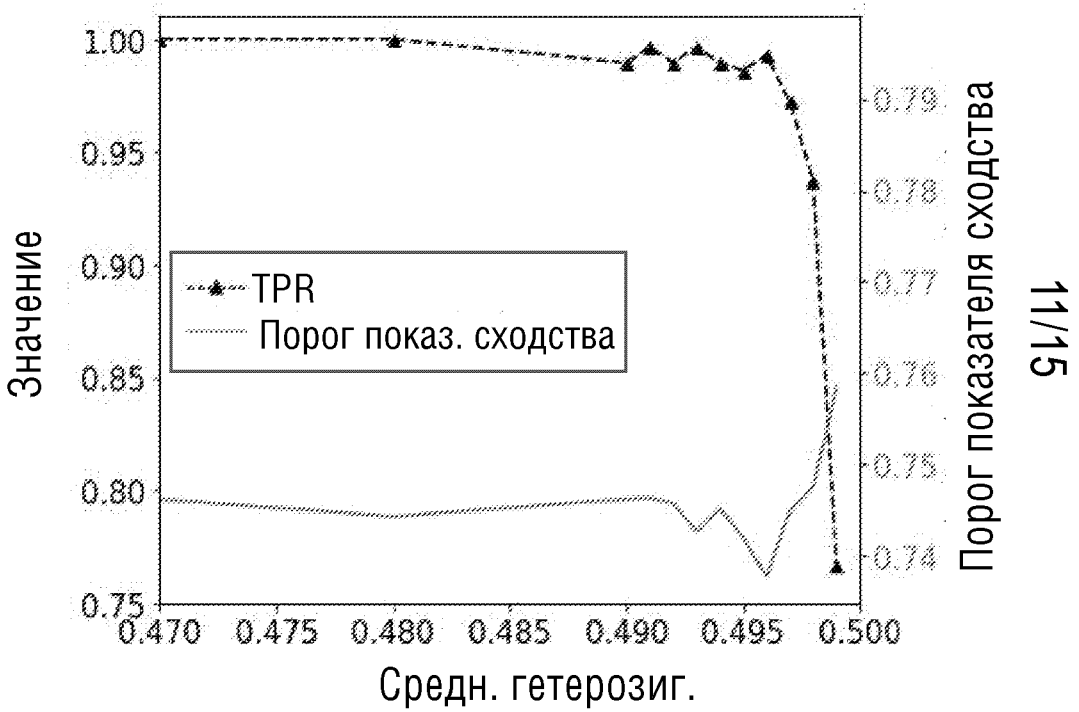
ФИГ.10C



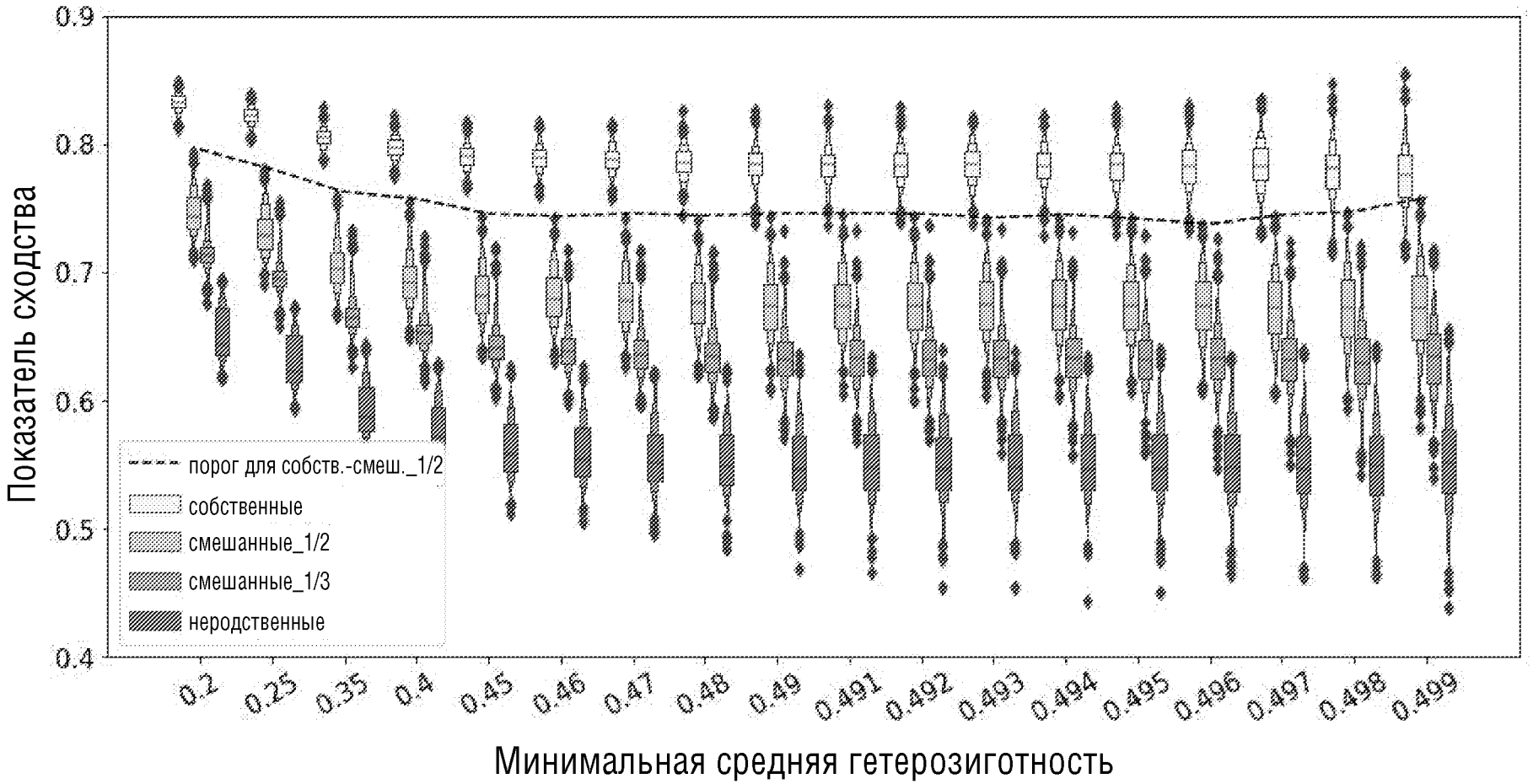
ФИГ.11A



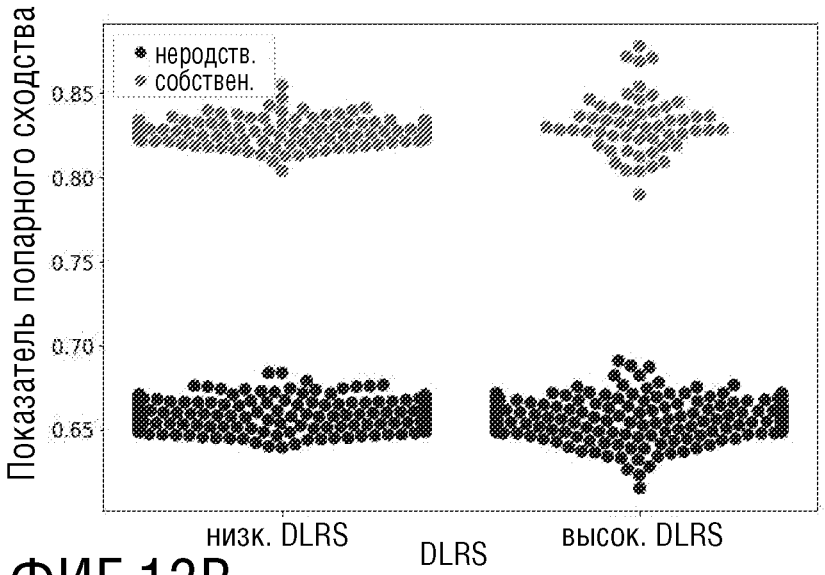
ФИГ.11B



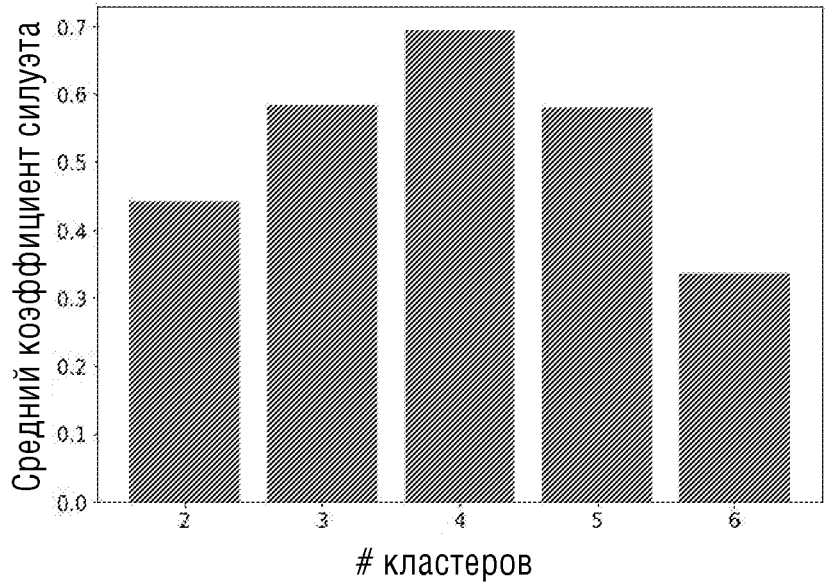
ФИГ.12


0.9
0.8
0.7
0.6
0.5
0.4
Показатель сходства
порог для собств.-смеш._1/2
собственные
смешанные_1/2
смешанные_1/3
неродственные
0.2
0.25
0.35
0.4
0.45
0.46
0.47
0.48
0.49
0.491
0.492
0.493
0.494
0.495
0.496
0.497
0.498
0.499
Минимальная средняя гетерозиготность

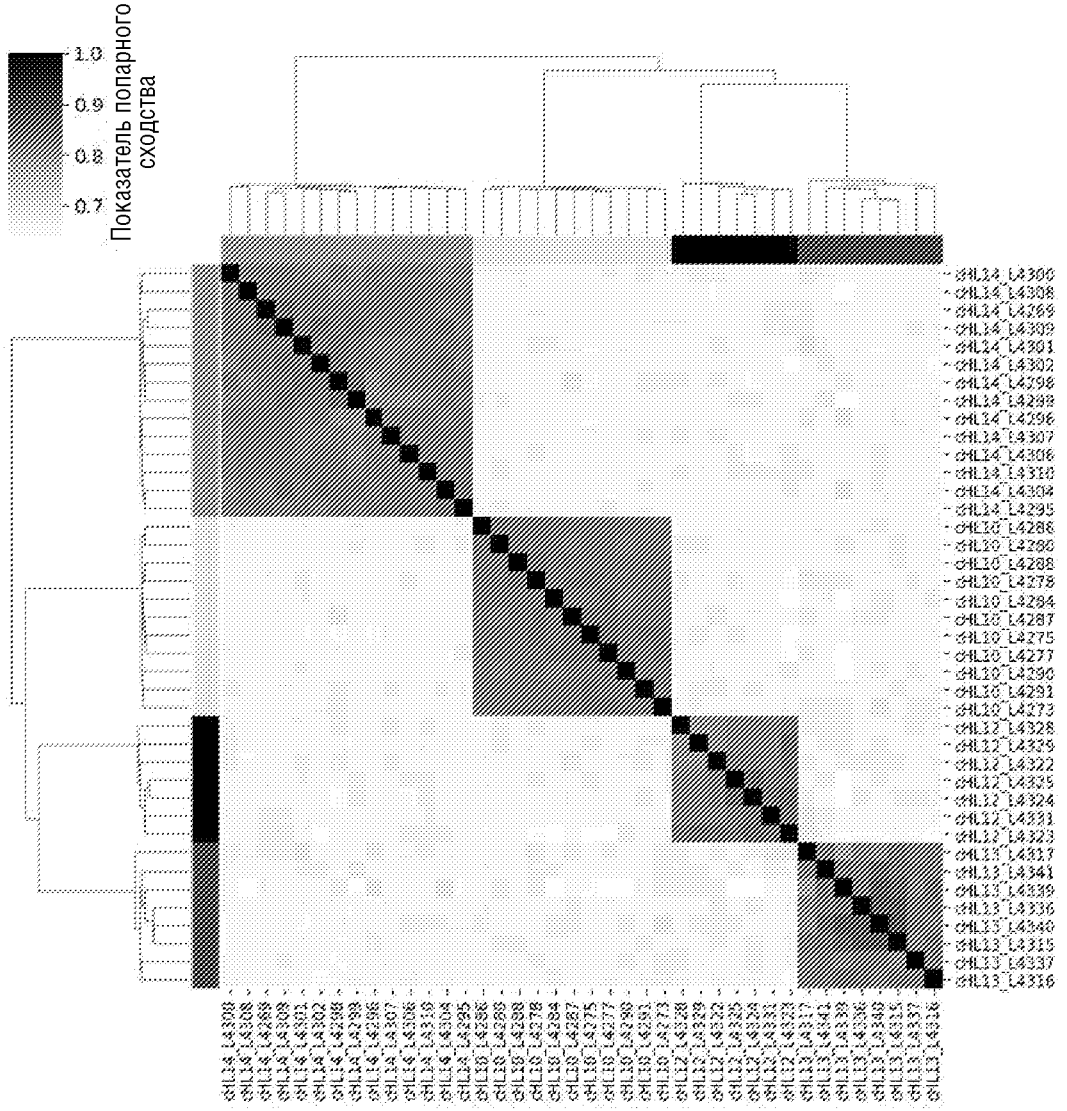
ФИГ.13А



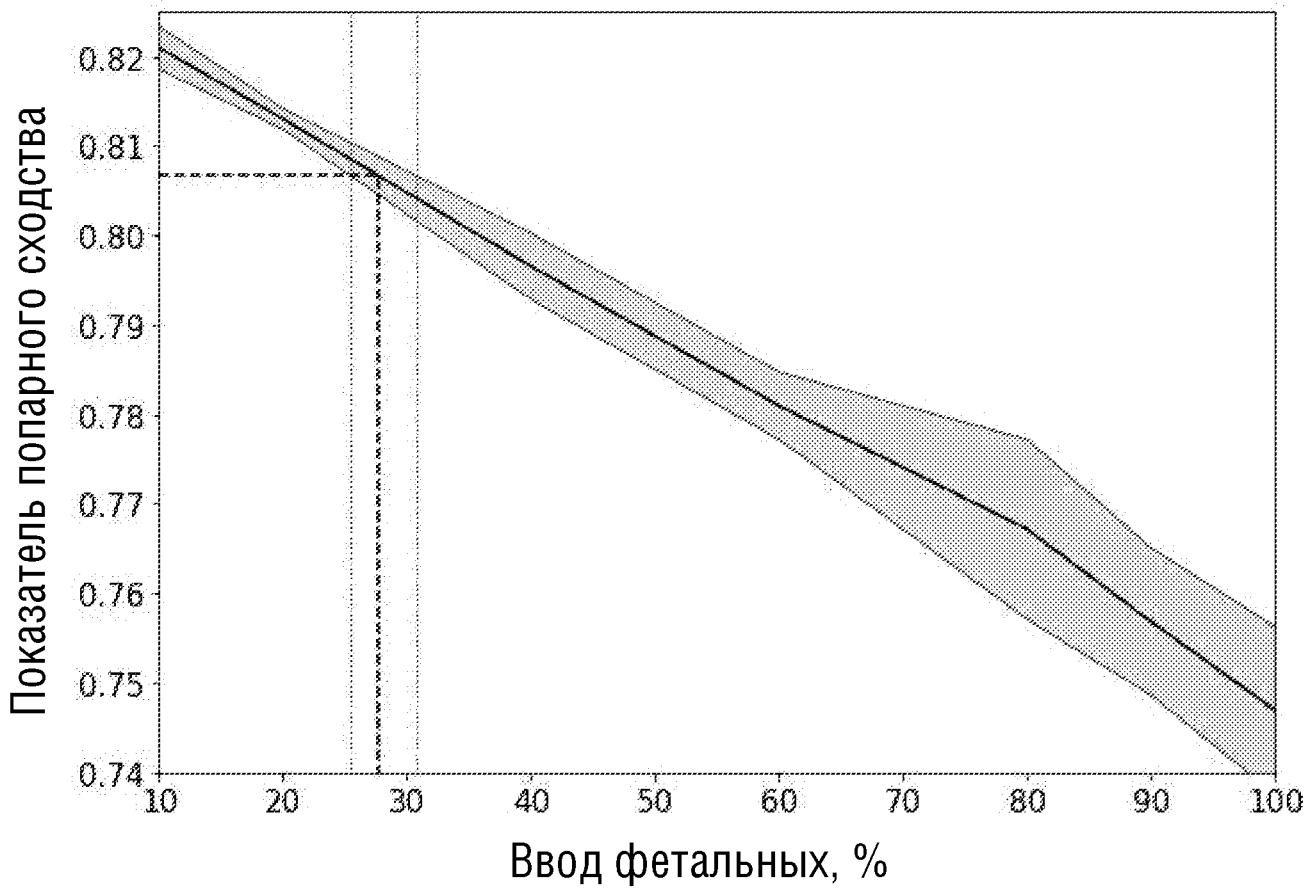
ФИГ.13В



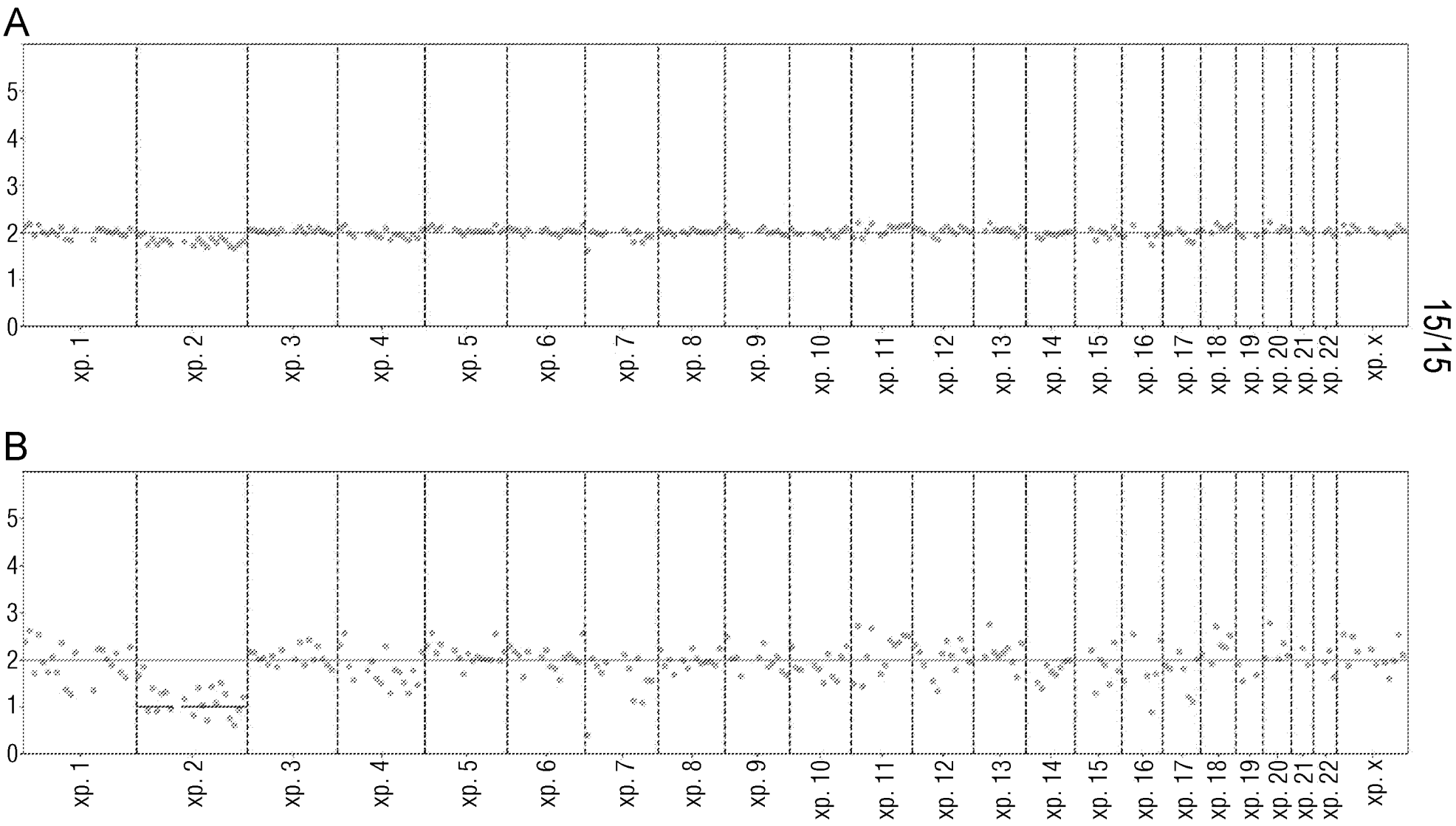
ФИГ.13С



ФИГ.14


Показатель попарного сходства
0.82
0.81
0.80
0.79
0.78
0.77
0.76
0.75
0.74
10
20
30
40
50
60
70
80
90
100
Ввод фетальных, %

ФИГ.15


A
B
0
1
2
3
4
5
хр. 1
хр. 2
хр. 3
хр. 4
хр. 5
хр. 6
хр. 7
хр. 8
хр. 9
хр. 10
хр. 11
хр. 12
хр. 13
хр. 14
хр. 15
хр. 16
хр. 17
хр. 18
хр. 19
хр. 20
хр. 21
хр. 22
хр. x