(12) МЕЖДУНАРОДНАЯ ЗАЯВКА, ОПУБЛИКОВАННАЯ В СООТВЕТСТВИИ С ДОГОВОРОМ О ПАТЕНТНОЙ КООПЕРАЦИИ (PCT)

(19) Всемирная Организация Интеллектуальной Собственности
Международное бюро

(43) Дата международной публикации
11 апреля 2019 (11.04.2019)

WIPO | PCT

(10) Номер международной публикации
WO 2019/070143 A1

(51) Международная патентная классификация:
G16H 10/40 (2018.01) *G06F 19/22* (2011.01)
G16H 50/20 (2018.01)

(21) Номер международной заявки: PCT/RU2017/000734

(22) Дата международной подачи:
03 октября 2017 (03.10.2017)

(25) Язык подачи: Русский

(26) Язык публикации: Русский

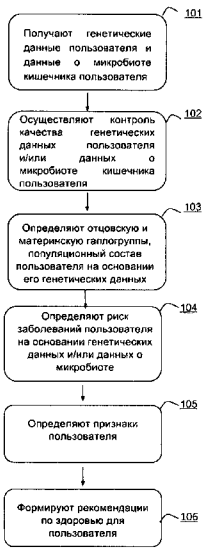
(30) Данные о приоритете:
2017134097 03 октября 2017 (03.10.2017) RU

(71) Заявитель: **ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "АТЛАС" (OBSHCHESTVO S OGRANICHENNOJ OTVETSTVENNOSTYU "ATLAS")** [RU/RU]; Инновационный центр "Сколково", ул. Нобеля, 7, Москва, 121205, Moscow (RU).

(72) Изобретатели: **МУСИЕНКО, Сергей Владимирович (MUSIENKO, Sergej Vladimirovich)**; ул. Пушкинская, 7, кв.33, мкрн. Юбилейный, Королев, Московская обл., 141092, Korolev Moskovskaya obl. (RU). **ПЕРФИЛЬЕВ, Андрей Валентинович (PERFIL'EV, Andrej Valentinovich)**; 1-й Нагатинский проезд, 11, к.1, кв.475, Москва, 115533, Moscow (RU). **ОСИПЕНКО, Дмитрий Александрович (OSIPENKO, Dmitrij Aleksandrovich)**; корп.445, кв.87, Зеленоград, Москва, , 124498, Zelenograd, Moscow (RU). **НИКОГОСОВ, Дмитрий Аркадьевич (NIKOGOSOV, Dmitrij Arkad'evich)**; Комсомольский пр-т, 38/16, кв.12, Москва, 119270, Moscow (RU). **АЛЕКСЕЕВ, Дмитрий Глебович (ALEKSEEV, Dmitrij Glebovich)**;

(54) **Title:** SYSTEM AND METHOD FOR INTERPRETING DATA AND PROVIDING RECOMMENDATIONS TO A USER BASED ON HIS/HER GENETIC DATA AND ON DATA RELATED TO THE COMPOSITION OF HIS/HER INTESTINAL MICROBIOTA

(54) **Название изобретения:** СИСТЕМА И СПОСОБ ИНТЕРПРЕТАЦИИ ДАННЫХ И ПРЕДОСТАВЛЕНИЯ РЕКОМЕНДАЦИЙ ПОЛЬЗОВАТЕЛЮ НА ОСНОВЕ ЕГО ГЕНЕТИЧЕСКИХ ДАННЫХ И ДАННЫХ О СОСТАВЕ МИКРОБИОТЫ КИШЕЧНИКА



Фиг. 1

101 The user's genetic data and data related to the user's intestinal microbiota are obtained
102 The quality of the user's genetic data and/or data related to the user's intestinal microbiota is checked
103 The user's paternal and maternal haplogroups and population structure are determined based on his/her genetic data
104 The user's disease risk is determined based on genetic data and/or microbiota-related data
105 The user's signs are determined
106 Health recommendations are generated for the user

(57) **Abstract:** This technical solution relates to the sphere of computer technology in genetics and microbiology, and more specifically, to a new system and method, within the sphere of microbiology, for studying and interpreting genetic data and data related to the composition of a person's intestinal microbiota. The system for providing recommendations to a user, based on genetic data and/or data related to the composition of intestinal microbiota, comprises: a unit for obtaining primary data, designed with the ability to obtain the user's genetic data and/or data related to the user's intestinal microbiota; a quality checking unit, designed with the ability to check the quality of the user's genetic data and data related to the user's intestinal microbiota, said data having been obtained by the unit for obtaining primary data, wherein the genetic data includes single nucleotide polymorphisms, while the data related to the microbiota includes readings; a unit for determining the user's lineage, said unit being designed with the ability to determine the paternal and maternal haplogroup, the population structure of a user based on his/her genetic data; a unit for the taxonomic analysis of microbiota-related data, said unit being designed with the ability to classify metagenomic risks; a unit for determining disease risk, said unit being designed with the ability to determine the degree of immunity to diseases, as well to check mutations for the presence of pathogenic alleles and for assessing the status of diseases; a unit for determining signs, said unit being designed with the ability to determine the states of the user's signs by means of reduction of the dependency graph of signs; a unit for generating recommendations for the user, said unit being designed with the ability to generate recommendations for the user based on data from the unit for determining disease risk and the unit for determining the user's signs. The technical result is an increase in the accuracy of recommendations provided to a user based on his/her genetic data and on data related to the composition of his/her intestinal microbiota.

(57) **Реферат:** Данное техническое решение относится к области вычислительной техники в генетике и микробиологии, а более конкретно, к новой системе и способу для исследования и интерпретации генетических данных и данных о составе микробиоты кишечника

[продолжение на следующей странице]

WO 2019/070143 A1

ул. Мурановская, 13Б, кв.109, Москва, 127349, Moscow (RU). **ТЯХТ, Александр Викторович (TYAKHT, Aleksandr Viktorovich)**; ул. Лизы Чайкиной, 4, корп.1, кв.28, Москва, 125315, Moscow (RU).

(74) **Агент: ГЕРАСИН, Борис Валерьевич (GERASIN, Boris Valer'evich)**; ул. Героев Панфиловцев, 18, стр.1, кв.96, Москва, 125480, Moscow (RU).

(81) **Указанные государства** *(если не указано иначе, для каждого вида национальной охраны)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) **Указанные государства** *(если не указано иначе, для каждого вида региональной охраны)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), евразийский (AM, AZ, BY, KG, KZ, RU, TJ, TM), европейский патент (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Опубликована:

— *с отчётом о международном поиске (статья 21.3)*

человека в области микробиологии. Система рекомендаций пользователю на основе генетических данных и/или данных о составе микробиоты кишечника содержит блок получения первичных данных, выполненный с возможностью получения генетических данных пользователя и/или данных о микробиоте кишечника пользователя; блок контроля качества, выполненный с возможностью контроля качества генетических данных пользователя и данных о микробиоте кишечника пользователя, полученных блоком получения первичных данных, причем генетические данные включают в себя однонуклеотидные полиморфизмы, а данные о микробиоте включают в себя чтения; блок определения происхождения пользователя, выполненный с возможностью определения отцовской и материнской гаплогруппы, популяционного состава пользователя на основании его генетических данных; блок таксономического анализа данных о микробиоте, выполненный с возможностью классификации метагеномных рисков; блок определения риска заболеваний, выполненный с возможностью определения защищенности от заболеваний, а также проверки мутаций на наличие патогенных аллелей и оценки статуса заболеваний, блок определения признаков, выполненный с возможностью определения состояний признаков пользователя путем редукции графа зависимостей признаков; блок формирования рекомендаций для пользователя, выполненный с возможностью формирования рекомендации пользователю на основании данных блока определения риска заболеваний и блока определения признаков пользователя. Технический результат - повышение точности рекомендаций пользователю на основании его генетических данных и данных о составе микробиоты кишечника.

СИСТЕМА И СПОСОБ ИНТЕРПРЕТАЦИИ ДАННЫХ И ПРЕДОСТАВЛЕНИЯ РЕКОМЕНДАЦИЙ ПОЛЬЗОВАТЕЛЮ НА ОСНОВЕ ЕГО ГЕНЕТИЧЕСКИХ ДАННЫХ И ДАННЫХ О СОСТАВЕ МИКРОБИОТЫ КИШЕЧНИКА

ОБЛАСТЬ ТЕХНИКИ

Данное техническое решение относится в общем к области вычислительной техники в генетике и микробиологии, а более конкретно, к новой системе и способу для исследования и интерпретации генетических данных и/или данных о составе микробиоты кишечника человека в области микробиологии, с целью выработки рекомендаций пользователю.

УРОВЕНЬ ТЕХНИКИ

Организм человека является одной из наиболее плотно населенных сред обитания на Земле. Число микроорганизмов, обитающих в такой «биологической системе», насчитывает порядка 100 триллионов бактерий, что значительно превышает общее число эукариотических клеток всех тканей и органов человека. Только 10% клеток организма являются нашими собственными, остальные 90% принадлежат бактериям. Совокупность всех микроорганизмов человека называется микрофлорой или микробиотой, а совокупность их генов — метагеномом. При этом метагеном человека в 100–150 раз больше генома самого человека. Большая часть микроорганизмов приходится на желудочно-кишечный тракт, поэтому его исследование и интерпретация этих данных является очень важной технической задачей. Фактически в настоящий момент формируется представление о микробиоте кишечника как об отдельном органе человеческого организма, что не противоречит исторически сложившемуся определению органа как части организма, представляющей собой эволюционно сложившийся комплекс тканей, объединенный общей функцией, структурной организацией и развитием. При этом человека можно рассматривать как «сверхорганизм», обмен веществ которых обеспечивается четко организованной работой ферментов, кодируемых не только геномом собственно Homo sapiens, но и геномами всех микроорганизмов.

Генетика человека – это врожденные особенности человека, передаваемые через гены, которые представляют собой участки ДНК, несущие информацию о наследственности. Генетика человека часто способствует возникновению наиболее распространенных заболеваний. Нельзя не учитывать наследственные особенности человека в определении образа жизни и режима питания, выборе профессии, занятий каким-то видом спорта и др. Многофакторные заболевания развиваются под действием нескольких факторов, например, таких как экология, образ жизни, физическая активность

и наследственность. Соответственно корректируя модифицируемые факторы, мы можем снижать риски. Таким образом, знание о генетических рисках важно для формирования индивидуальных профилактических мер. Многие факторы вызывают нарушения всех процессов в организме и несут за собой развитие различных заболеваний, которые можно предотвратить исследуя генетические данные человека и формируя рекомендации по факторам: здоровью, питанию, спорту, и в целом образу жизни.

Учитывая большое влияние микробиоты и генетики на здоровье человека, следует продолжать усилия, связанные с их исследованием и интерпретацией.

СУЩНОСТЬ ИЗОБРЕТЕНИЯ

Данное техническое решение направлено на устранение недостатков, свойственных решениям, известным из уровня техники.

Технической задачей или проблемой, решаемой в данном техническом решении, является формирование рекомендаций по образу жизни, профилактике заболеваний, питанию и физической активности пользователю на основании генетических данных и/или данных о составе микробиоты кишечника.

Техническим результатом, достигаемым при решении вышеуказанной технической проблемы, является повышение точности рекомендаций пользователю на основании учета генетических данных и данных о составе микробиоты кишечника.

Указанный технический результат достигается благодаря реализации системы формирования рекомендаций пользователю на основе генетических данных и/или данных о составе микробиоты кишечника, которая содержит блок получения первичных данных, выполненный с возможностью использования генетических данных пользователя и/или данных о микробиоте кишечника пользователя; блок контроля качества, выполненный с возможностью контроля качества генетических данных пользователя и данных о микробиоте кишечника пользователя, полученных блоком получения первичных данных, причем генетические данные включают в себя однонуклеотидные полиморфизмы, а данные о микробиоте включают в себя чтения; блок популяционного анализа генетических данных, выполненный с возможностью определения отцовской и материнской гаплогруппы, популяционного состава генетических данных пользователя; блок таксономического анализа данных о микробиоте, выполненный с возможностью картирования метагеномных чтений на каталог, состоящий из набора последовательностей генов микробов микробиоты кишечника; блок определения риска заболеваний, выполненный с возможностью определения защищенности от заболеваний, а также проверки мутаций на наличие патогенных аллелей и оценки статуса заболеваний; блок определения признаков, выполненный с возможностью определения состояний признаков пользователя путем редукции графа зависимостей признаков; блок

формирования рекомендаций для пользователя, выполненный с возможностью формирования рекомендаций пользователю на основании данных блока определения риска заболеваний и блока определения признаков пользователя.

В некоторых вариантах осуществления технического решения блок получения первичных данных получает файлы секвенирования в формате FASTQ или FASTA, полученные с секвенатора.

В некоторых вариантах осуществления технического решения блок контроля качества получает генетические данные пользователя из кремниевого биочипа посредством сканера биочипов.

В некоторых вариантах осуществления технического решения генетические данные содержат данные о генотипах однонуклеотидных полиморфизмов пользователя, содержащие полиморфизмы X- и Y-хромосом.

В некоторых вариантах осуществления технического решения блок контроля качества дополнительно определяет пол пользователя посредством подсчета количества однонуклеотидных полиморфизмов по X- и по Y-хромосомам.

В некоторых вариантах осуществления технического решения в случае мужского пола блок контроля качества конвертирует однонуклеотидные полиморфизмы в гомозиготном состоянии с X- и Y-хромосомами в однонуклеотидные полиморфизмы в гемизиготном состоянии.

В некоторых вариантах осуществления технического решения блок контроля качества осуществляет отсеивание чтений со средним баллом качества, полученного с ДНК-секвенатора, ниже заранее заданного порогового.

В некоторых вариантах осуществления технического решения блок контроля качества с концов чтений удаляет позиции, имеющие низкий балл качества.

В некоторых вариантах осуществления технического решения блок контроля качества отсеивает из чтений постороннюю генетическую информацию, не относящуюся к микробиоте кишечника, имеющих как биологическое происхождение, так и техническое, возникающую из-за прочтения артефактных генетических последовательностей.

В некоторых вариантах осуществления технического решения блок популяционного анализа генетических данных определяет отцовскую гаплогруппу на основании дерева мутаций для Y-хромосомы и генетических данных пользователя.

В некоторых вариантах осуществления технического решения блок популяционного анализа генетических данных определяет материнскую гаплогруппу на основании дерева мутаций для митохондрии и генетических данных пользователя.

В некоторых вариантах осуществления технического решения блок популяционного анализа генетических данных определяет популяционный состав на основании данных о генотипах людей из разных популяций и генетических данных пользователя.

В некоторых вариантах осуществления технического решения блок популяционного анализа генетических данных на основании генетических данных пользователя и набора унаследованных от неандертальца аллелей в определенных полиморфизмах определяет общее количество неандертальских аллелей.

В некоторых вариантах осуществления технического решения при картировании метагеномных чтений на каталог блоком таксономического анализа данных, данный каталог содержит геномные последовательности бактерий и/или архей и/или эукариот, встречающихся в кишечнике пользователя.

В некоторых вариантах осуществления технического решения блок таксономического анализа данных о микробиоте определяет относительную представленность генома или микробного вида.

В некоторых вариантах осуществления технического решения блок таксономического анализа данных о микробиоте формирует прореженные таблицы представленности по другим таксономическим уровням помимо таксономического уровня рода или вида.

В некоторых вариантах осуществления технического решения блок определения риска заболеваний оценивает аномальность образца посредством проверки суммарного процента чтений, относящихся к одному из таксонов из списка оппортунистических патогенов или микробов, заведомо не относящихся к кишечной микробиоте.

В некоторых вариантах осуществления технического решения блок определения риска заболеваний определяет защищенность от заболеваний пользователя по данным о микробиоте на основании референсных данных.

В некоторых вариантах осуществления технического решения блок определения признаков осуществляет проверку на наличие циклов в графе зависимостей, и при наличии циклов блок не дает запустить редукцию графа.

КРАТКОЕ ОПИСАНИЕ ЧЕРТЕЖЕЙ

Признаки и преимущества настоящего технического решения станут очевидными из приведенного ниже подробного описания и прилагаемых чертежей, на которых:

Фиг. 1 - приведена блок-схема способа формирования рекомендаций пользователю на основе генетических данных и/или данных о составе микробиоты кишечника;

Фиг. 2 - приведена блок-схема системы формирования рекомендаций пользователю на основе генетических данных и/или данных о составе микробиоты кишечника;

Фиг. 3 - показан процесс работы системы формирования рекомендаций пользователю на основе генетических данных и/или данных о составе микробиоты кишечника.

Фиг. 4 - показан вариант реализации, когда образцы для одних и тех же пользователей имеют разные генотипы.

Фиг. 5 - показан вариант реализации, когда в зависимости от количества образцов генотип одного и того же пользователя может различаться.

ПОДРОБНОЕ ОПИСАНИЕ ИЗОБРЕТЕНИЯ

Ниже будут подробно рассмотрены термины и их определения, используемые в описании технического решения.

В данном изобретении под системой подразумевается компьютерная система, ЭВМ (электронно-вычислительная машина), ЧПУ (числовое программное управление), ПЛК (программируемый логический контроллер), компьютеризированные системы управления и любые другие устройства, способные выполнять заданную, четко определенную последовательность операций (действий, инструкций).

Под устройством обработки команд подразумевается электронный блок либо интегральная схема (микропроцессор), исполняющая машинные инструкции (программы). Устройство обработки команд считывает и выполняет машинные инструкции (программы) с одного или более устройства хранения данных. В роли устройства хранения данных могут выступать, но, не ограничиваясь, жесткие диски (HDD), флеш-память, ПЗУ (постоянное запоминающее устройство), твердотельные накопители (SSD), оптические приводы.

Программа - последовательность инструкций, предназначенных для исполнения устройством управления вычислительной машины или устройством обработки команд.

Микробиота (нормальная микрофлора, нормофлора) человека – это совокупность всех микроорганизмов в теле человека.

Генетические данные - это информация о структуре ДНК, последовательности нуклеотидов ДНК, одно- и олигонуклеотидных изменений в последовательности ДНК, включая все хромосомы конкретного организма. Генетическая информация частично определяет морфологическое строение, рост, развитие, обмен веществ, психический склад, предрасположенность к заболеваниям и генетические пороки организма.

Однонуклеотидный полиморфизм (ОНП, англ. single nucleotide polymorphism, SNP, произносится как снип) — отличия последовательности ДНК размером в один нуклеотид (A, T, G или C) в геноме (или в другой сравниваемой последовательности) представителей одного вида или между гомологичными участками гомологичных хромосом.

Гаплогруппа — группа схожих гаплотипов, имеющих общего предка, у которого произошла мутация, унаследованная всеми потомками (обычно — однонуклеотидный полиморфизм). Термин «гаплогруппа» широко применяется в генетической генеалогии - науке, изучающей генетическую историю человечества, с помощью исследования гаплогрупп Y-хромосомы (Y-ДНК), митохондриальной ДНК (мтДНК) и ГКГ-гаплогруппы.

Аллели — различные формы (значения) одного и того же гена, расположенные в одинаковых участках (локусах) гомологичных хромосом.

Секвенирование ДНК — определение последовательности нуклеотидов в молекуле ДНК. Под этим может подразумеваться как амликонное секвенирование (прочтение последовательностей выделенных фрагментов ДНК, полученных в результате ПЦР реакции - таких, как ген 16S рРНК или его фрагменты) и полногеномное секвенирование (прочтение последовательностей всей ДНК, присутствующей в образце).

Гомозиготное состояние — это такое состояние локуса, при котором аллели в данном локусе идентичны друг другу на гомологичных хромосомах.

Гетерозиготное состояние — это такое состояние локуса, при котором аллели в данном локусе отличаются друг от друга на гомологичных хромосомах.

Гемизиготное состояние — это такое состояние локуса, при котором у него отсутствует гомологичный аллель, то есть хромосома, в которой находится данный локус, не имеет гомологичной пары.

RsID — это обозначение идентификатора индивидуального однонуклеотидного полиморфизма.

Чтения (риды, reads) – данные, представляющие собой нуклеотидные последовательности фрагментов ДНК, полученные с помощью ДНК-секвенатора.

FASTA – формат записи последовательностей ДНК.

Филогенетика или филогенетическая систематика — область биологической систематики, которая занимается выявлением и прояснением эволюционных взаимоотношений среди разных видов жизни на Земле, как современных, так и вымерших.

α-разнообразие - числовая величина, характеризующая разнообразие микробного сообщества внутри одной ниши. α-разнообразие вычисляется с помощью того или иного алгоритма на основании данных о видовом составе микробиоты.

β-разнообразие - числовая величина, характеризующая меру различия между 2 микробными сообществами. Это разнообразие между сообществами, показатель степени дифференцированности распределения видов или скорости изменения видового состава, видовой структуры вдоль градиентов среды. Возможный путь определения β-разнообразия – сравнение видового состава различных сообществ. Чем меньше общих видов в сообществах или в разных точках градиента, тем выше β-разнообразие.

Картирование коротких прочтений — биоинформатический метод анализа результатов секвенирования нового поколения, состоящий в определении позиций в референсной базе геномов или генов, откуда с наибольшей вероятностью могло быть получено каждое конкретное короткое прочтение.

В результате секвенирования ДНК создается набор **чтений**. Длина чтения у современных секвенаторов составляет от нескольких сотен до нескольких тысяч нуклеотидов.

«Золотой стандарт» (эталон, **референс**) генома — последовательность ДНК в цифровом виде, составленная учеными как общий репрезентативный пример генетического кода того или иного вида живых организмов. В случае человеческого генома, это может быть, например, версия сборки GRChg37 (Genome Reference Consortium human genome 37), которая представляет собой гаплоидный геном с перемежающимися локусами (т. е. изначально сведенные в одну последовательность аллельные варианты могут располагаться на разных хромосомах).

Таксономия — учение о принципах и практике классификации и систематизации сложноорганизованных иерархически соотносящихся сущностей.

В некоторых вариантах осуществления способ **100** реализован в системе **200**, которая представляет из себя набор блоков, как показано на Фиг. 2. Однако способ **100** может быть альтернативно реализован с использованием любой другой подходящей системы (систем), сконфигурированной для приема и обработки генетических данных пользователей и данных о микробиоте кишечника этих пользователей, в совокупности с другой информацией для создания и обмена данными, полученными из проведенных микробиологических анализов.

Блок **201** получения первичных данных получает пробы образца по меньшей мере одного пользователя. Вышеуказанные данные получают от пользователя посредством *использования* набора для отбора проб, включающий контейнер **301** для образцов, как показано на Фиг. 3, имеющий компонент технологического реагента и сконфигурированный для приема образца из места сбора пользователем. Пользователь, находящийся в удаленном месте от блока получения первичных данных, может предоставлять образцы надежным образом. Доставка набора для отбора проб предпочтительно выполняется с использованием службы доставки посылок (например, почтовой службы, службы доставки и т. д.). Дополнительно или альтернативно набор для отбора проб может быть предоставлен непосредственно через устройство, установленное в помещении или на улице, которое предназначено для облегчения приема пробы от пользователя. В других вариантах осуществления набор для отбора проб может быть сдан в клинику или другое медицинское заведение медицинскому лабораторному технику. Однако предоставление набора (-ов) для отбора проб пользователя в блоке **201** получения первичных данных может дополнительно или альтернативно выполняться любым другим подходящим способом.

Набор(ы) для отбора проб, предусмотренный в блоке **201** получения первичных данных, предпочтительно выполнен с возможностью облегчения приема образцов от пользователей неинвазивным образом. В некоторых вариантах осуществления изобретения неинвазивные способы получения образца от человека могут использовать любой или несколько: проницаемый субстрат (например, тампон, выполненный с возможностью протирать область тела человека, туалетную бумага, губка и т. д.),

контейнер (например, флакон, трубка, мешок и т. д.), сконфигурированный для приема образца из области тела пользователя и любого другого подходящего элемента приема (слюна, кал, моча и т.д.). В конкретном примере образцы могут быть собраны неинвазивным образом из одного органа или нескольких, например, таких как нос, кожа, половой орган человека, полость рта и кишечник (например, с использованием тампона и флакона). Однако набор для отбора пробы, предусмотренный в блоке **201** получения первичных данных, может дополнительно или альтернативно быть использован для облегчения приема образцов полуинвазивным способом или инвазивным способом. В некоторых вариантах осуществления инвазивные способы приема образца могут использовать следующие объекты: игла, шприц, биопсийные щипцы, трепан и любой другой подходящий инструмент для сбора образца полуинвазивным или инвазивным способом. В конкретных примерах пробы пользователей могут содержать один или несколько образцов крови, образцов плазмы / сыворотки (например, для экстракции бесклеточной ДНК) и образцов тканей.

Входные образцы могут представлять из себя образцы (слюна, моча, кал, кровь), которые могут быть обработаны, например, в лаборатории, и из которых получают генетические данные и данные о составе микробиоты кишечника путем секвенирования или генотипирования.

В некоторых вариантах реализации блок **201** получения первичных данных может получать дополнительные данные, которые будут учитываться при формировании рекомендаций для пользователя, из датчиков, связанных с пользователем (пользователями) (например, датчиков носимых вычислительных устройств, датчиков мобильных устройств, биометрических датчиков, связанных с пользователем и т. д.). Таким образом, блок **201** получения первичных данных может включать в себя получение данных о физической активности пользователя или физическом воздействии на него (например, данные акселерометра и гироскопа с мобильного устройства или носимого вычислительного устройства пользователя), данные об окружающей среде (например, данные о температуре, данные о высоте, данные о климате, данные о параметрах света и т. д.), данные о питании пользователя или данные о диете (например, данные из регистрационных записей о принимаемых пищевых продуктах, данные спектрофотометрического анализа и т. д.), биометрические данные (например, данные, регистрируемые с помощью датчиков в мобильном вычислительном устройстве пользователя), данные о местоположении (например, с использованием датчиков GPS), диагностические данные или любые другие подходящие данные. Дополнительно или альтернативно, дополнительный набор данных может быть получен из данных медицинской записи и / или клинических данных пользователя (ей). В некоторых вариантах реализации дополнительный набор данных может быть получен из одной или нескольких электронных медицинских записей (EHR) пользователя (ей).

Блок **202** контроля качества на основании проб образца пользователя, полученных в блоке **201** получения первичных данных, получает однонуклеотидные полиморфизмы и чтения пользователя.

В уровне техники при получении генетических данных бывают несколько типов ошибок. Например, образцы для одних и тех же пользователей имеют разные генотипы, как показано на Фиг. 4. Или, например, в зависимости от количества образцов генотип одного и того же пользователя может различаться (Фиг. 5).

Чтобы предотвратить вывод неправильных интерпретаций, раньше в уровне техники генетики вручную проверяли, является ли образец правильным, исходя из числа патогенных аллелей для одного однонуклеотидного полиморфизма и показателя интенсивности, что является очень неэффективным.

При получении однонуклеотидных полиморфизмов, блок **202** контроля качества осуществляет их контроль качества (QC - Quality Control). Данные могут получать из кремниевого биочипа посредством сканера биочипов, на который нанесены небольшие отрезки ДНК - зонды, которые специфически связываются с ДНК пользователя. При успешном связывании к этим данным может присоединяться флюоресцентная метка. Биочипы для генотипирования позволяют проводить SNP-типирование и анализ вариаций числа копий генов, генотипирование образцов для биобанков, таргетное генотипирование. В результате работы сканера биочипа получают информацию о генотипах однонуклеотидных полиморфизмов для конкретного пользователя, содержащую также полиморфизмы с X- и Y-хромосом. Вышеуказанная информация может в себя включать идентификатор генетического полиморфизма (rsID) и один или два аллеля. Аллель в данном случае - это строка из символов A, T, G, C, -. Например, данные могут быть представлены в следующем виде:

# rsid	chromosome	position	genotype
rs4477212	1	72017	A/A
rs3094315	1	742429	A/A
rs10195681	2	8674	C/C
rs11901199	2	8856	G/G
rs7885198	X	67869221	G
rs2066847	16	50729867	-/C

На первом шаге определяют пол пользователя посредством подсчета количества однонуклеотидных полиморфизмов по X- и по Y-хромосомам. В частности осуществляют подсчет доли однонуклеотидных полиморфизмов на X-хромосоме в гомозиготном состоянии и доли однонуклеотидных полиморфизмов, для которых не удалось осуществить генотипирование, на Y-хромосоме. Для подсчета однонуклеотидных полиморфизмов на X-хромосоме определяют число однонуклеотидных полиморфизмов на X-хромосоме в гомозиготном состоянии, общее число однонуклеотидных полиморфизмов на X-хромосоме, после чего определяют отношение первого числа ко второму. Для подсчета однонуклеотидных полиморфизмов на Y-хромосоме определяют число однонуклеотидных полиморфизмов с неопределенным генотипом, общее число однонуклеотидных полиморфизмов на У-хромосоме, после чего находят отношение первого числа ко второму.

В случае совпадения определения пола по X- и по Y-хромосомам итоговый пол однозначно определён. В случае, если по X определился мужчина, а по Y – женщина, то результатом является X0 – признак синдрома Тернера; если наоборот – то признак синдрома Клайнфельтера. В некоторых вариантах осуществления в случае несовпадения определения пола по X- и Y-хромосомам производится дополнительная проверка образца на брак, так как с высокой долей вероятности речь идет именно о браке, а не об указанных двух синдромах.

После определения пола блоком **202** контроля качества на этапе контроля качества в случае мужского пола однонуклеотидные полиморфизмы в гомозиготном состоянии с X- и Y-хромосомами конвертируются в однонуклеотидные полиморфизмы в гемизиготном состоянии; при этом гетерозиготные однонуклеотидные полиморфизмы с X и Y-хромосомами отфильтровываются и не попадают в итоговую выборку генетических данных. В случае женского пола все однонуклеотидные полиморфизмы с Y-хромосомой отфильтровываются и не попадают в итоговую выборку генетических данных. Под конвертацией в данном техническом решении понимается удаление одного аллеля из пары.

Также блок **201** получения первичных данных получает данные в результате секвенирования последовательностей микробных генов 16S рРНК микробиоты кишечника. В некоторых вариантах осуществления блок **201** получения первичных данных получает файлы секвенирования в формате FASTQ или FASTA, полученные с секвенатора, по одному файлу на каждый образец. Предпочтительно может применяться ампликонное секвенирование, но также может применяться полногеномное секвенирование (WGS).

В процессе секвенирования, заключительным этапом запуска секвенатора является нахождение нуклеотидов (англ. base calling) - преобразование промежуточных “сырых” (внутренних) сигналов прибора (изображений, спектров, карт интенсивности) во множество чтений, сопровождаемых баллами качества (по одному баллу для каждой нуклеотидной

позиции). Чтения состоят из четырех символов нуклеотидов (A, C, G и T), а также служебного символа N или ".", или "?", обозначающего полную неопределенность относительно значения в данной позиции (секвенатор не может определить нуклеотид), например в виде "GCAAAAAACTTACCCCGGAACAGGCCGAGCAGATCAAAACGCTACTGCAATACAGACCAT CAAGCACCAACTCCCNNNCGTAGNNNNNNTATGTTNNNNG". С точки зрения биоинформатического анализа важнейшими являются следующие характеристики чтений две вещи: во-первых, какой длины получатся чтения, во-вторых, какие в них могут быть ошибки и как часто. Приборный балл качества (англ. quality value) - величина, характеризующая вероятность отсутствия ошибки в данной позиции, вычисляемая секвенатором исходя из качества сигнала:

$$Q = -10log_{10}P$$

где P - вероятность ошибки в данной позиции. В разных вариантах осуществления чтения и их баллы качества могут генерироваться в виде двух файлов для каждого образца (формат FASTA) либо объединены в единый файл (формат FASTQ); при этом с целью экономии дискового пространства данные текстовые представления могут быть переведены в двоичный формат.

Для ускорения расчетов, файлы размером, например, более 500 МБ формата FASTQ прореживаются до, например, 89951 чтений (это количество чтений в среднем соответствует размеру файла в 500 МБ при длине ридов 250 нуклеотидов). Начиная с некоторого значения увеличение глубины секвенирования слабо влияет на получаемый видовой состав микробиоты.

Блок **202** контроля качества осуществляет отсеивание чтений со средним баллом качества ниже заранее заданного порогового. В других вариантах осуществления с концов чтений могут адаптивно удаляться позиции, имеющие низкий балл качества (например, последовательно удаляются все нуклеотиды с 5' к 3' концу до тех пор, пока не встретится позиция с баллом качества, превышающим фиксированный порог). Дополнительно блок **202** контроля качества отсеивает постороннюю генетическую информацию в чтениях, имеющих небиологическое происхождение, возникающую из-за прочтения артефактных последовательностей, возникающих в ходе неправильной химической модификации исходной ДНК.

При выполнении процесса контроля качества блок **202** контроля качества может использовать вычислительные методы (например, статистические методы, методы машинного обучения, методы искусственного интеллекта, методы биоинформатики и т. д.).

Затем блок **202** контроля качества передает список однонуклеотидных полиморфизмов пользователя с координатами (хромосома и ее позиция) и генотипом пользователя в блок **203** популяционного анализа генетических данных.

Гаплогруппы бывают двух типов: материнская гаплогруппа и отцовская гаплогруппа.

В блоке **203** популяционного анализа генетических данных сначала определяют отцовскую гаплогруппу на основании дерева мутаций для Y-хромосомы и генетических данных пользователя. Дерево мутаций может быть представлено, например, в формате XML. Генетические данные пользователя содержат список однонуклеотидных полиморфизмов с координатами (хромосома и позиция) и генотипом пользователя. Дерево мутаций для Y-хромосомы содержит мутации, характерные для каждой гаплогруппы (позиция - полиморфизм).

Структура данных и способ расчета для материнской гаплогруппы совпадает с отцовской, за исключением того, что материнская гаплогруппа рассчитывается по SNP (однонуклеотидным полиморфизмам) с MT-хромосомы, а отцовская гаплогруппа рассчитывается по SNP (однонуклеотидным полиморфизмам) с Y-хромосомы. В результате для мужчин вычисляется и отцовская и материнская гаплогруппы, а для женщин только материнская гаплогруппа.

Каждая гаплогруппа, кроме изначальной, имеет одну родительскую гаплогруппу, и одну или несколько дочерних. Каждая гаплогруппа имеет конечный список определяющих мутаций. Таким образом, образуется дерево гаплогрупп, где ребра определяются наборами мутаций.

Блок **203** популяционного анализа генетических данных при определении отцовской гаплогруппы использует дерево мутаций, генетические данные пользователя и работает следующим образом:

определяет количество вхождений каждого полиморфизма в дерево мутаций (например, A123G встречается 3 раза в дереве, T456C встречается в дереве 22 раза);

70. сохраняет отдельно максимально возможное количество вхождений полиморфизма в дерево (получается число, например 30);

оценивает каждый полиморфизм по формуле: максимальное число вхождений какого-либо полиморфизма (определенное на предыдущем шаге) минус количество вхождений данного полиморфизма в дерево. Это значение является весом полиморфизма;

находит совпадения полиморфизмов между образцом (пользовательские данные) и каждой гаплогруппой;

находит несовпадения полиморфизмов между образцом (пользовательские данные) и каждой гаплогруппой. В контексте данного технического решения несовпадающий полиморфизм - это такой полиморфизм, при котором мутация является обратной указанной. Например, если в дереве мутаций находится мутация A12345C, а у пользователя генотип A, то блок **203** популяционного анализа генетических данных определяет, что это не совпадающий полиморфизм.

если в дереве мутаций находится мутация A12345C, а у пользователя генотип не C и не A, то блок **203** популяционного анализа генетических данных отображает мутацию на комплементарную цепочку и получается T12345G. На данном шаге меняется обозначение

аллелей на комплементарные, то есть меняются аллели так, как будто они были на FWD цепочке, а стали на REV.

определяет количество совпавших и не совпавших полиморфизмов для каждой гаплогруппы;

оценивает гаплогруппу (которая является элементом дерева мутаций) по формуле: сумма весов совпавших полиморфизмов минус сумма весов не совпавших полиморфизмов;

находит такой путь по дереву мутаций, чтобы сумма оценок гаплогрупп была максимальна. Конечная гаплогруппа в таком пути и будет являться искомой отцовской гаплогруппой.

Аналогично, блок **203** популяционного анализа генетических данных определяет материнскую гаплогруппу, однако на основании дерева мутаций для митохондрии и генетических данных пользователя. мтДНК хранит дерево мутаций — устойчивые генетические маркеры (гаплогруппы), которые повторяются у всех потомков. Дерево образуется так: маркеры возникают во время мутаций и накапливаются в мтДНК. Есть возможность проследить отношение родства различных популяций по количеству совпадающих маркеров — чем больше маркеров совпадает, тем ближе родство. Если маркеры после определенной мутации не совпадают, можно сказать, когда популяции разошлись.

Далее блок **203** популяционного анализа генетических данных определяет популяционный состав пользователя на основании данных о генотипах людей из разных популяций, списка однонуклеотидных полиморфизмов с координатами (хромосома и позиция) и генотипом пользователя.

Блок **203** популяционного анализа генетических данных определяет популяционный состав посредством применения метода главных компонент. Каждый генетический образец из базы геномов для популяций разбивается на отрезки, состоящие из определенного количества однонуклеотидных полиморфизмов, следующих в геноме друг за другом последовательно. Для каждого отрезка образца определяется вектор методом главных компонент.

Аналогично, определяется вектор методом главных компонент для каждого отрезка входного образца.

Каждый отрезок образца относится в итоге к определенной популяции путем сравнения с векторами, определенными ранее.

Доля популяции рассчитывается как количество отрезков образца, отнесенных к этой популяции, поделенное на общее число отрезков образца.

В некоторых вариантах осуществления изобретения может использоваться метод главных компонент для разложения образца на вектор из 12 популяционных компонент, причем на вход подается образец целиком.

В некоторых вариантах осуществления блок **203** популяционного анализа генетических данных на основании списка однонуклеотидных полиморфизмов с координатами (хромосома и позиция) и генотипа пользователя, набора унаследованных от неандертальца аллелей в определенных полиморфизмах определяет общее количество неандертальских аллелей в образце следующим образом: если неандертальская аллель в гомозиготном состоянии, тогда +2 к результату, если неандертальская аллель в гетерозиготном состоянии, тогда +1 к результату, иначе +0. Изначально набор унаследованных от неандертальца аллелей может быть разделен на три части по популяциям: ASN, EUR и EURASN и в итоге слит в один набор. Далее, позиции на хромосоме переводятся из 37 в 38 сборок генома.

В некоторых вариантах осуществления при полногеномном (WGS) профилировании состава микробиоты блок **204** таксономического анализа данных о микробиоте картирует метагеномные чтения на неизбыточный каталог, состоящий из представительного набора геномов микробов кишечника. В данный каталог могут входить геномы бактерий, а также архей, встречающихся в кишечнике пользователя. Данный каталог может быть выработан на основании общедоступных крупных баз данных, а также автоматического анализа публикаций, имеющихся в уровне техники. В некоторых вариантах реализации расширяют набор референсных геномов, что позволяет регулярно добавлять новые опубликованные геномы. Результат картирования может быть сохранен в файле формата BAM. В некоторых вариантах осуществления для каждого генома определяют суммарную длину картировавшихся на него чтений (глубина покрытия).

При полногеномном анализе микробиоты относительная представленность генома далее может определяться блоком **204** таксономического анализа данных о микробиоте путем нормализации покрытия на длину генома и общую длину картировавшихся чтений:

$$\text{относительная представленность гена} = 10^{12} \times \left(\frac{\sum \text{длин картировавшихся ридов}/\text{длина гена}}{\text{общая длина картировавшихся ридов образца}}\right)$$

При анализе микробиоты с помощью 16S рРНК секвенирования после предварительной обработки, блок **204** таксономического анализа данных о микробиоте осуществляет количественный таксономический анализ данных путем определения, к какой известной бактерии принадлежит каждое чтение 16S рРНК (или его фрагмента) и как можно охарактеризовать чтения от неизвестных бактерий. Поиск осуществляется с применением стратегии поиска на основании шаблона (англ. reference-based). Таксономическая классификация опирается на базовое понятие операционной таксономической единицы (OTE, англ. operational taxonomic unit, OTU) - определение бактериального вида на основании одной лишь последовательности 16S рРНК. Набор чтений гена 16S рРНК (или его региона) сопоставляется с представительной базой последовательностей данного гена. Каждое чтение относится к той таксономической единице, с которой он обладает высокой степенью сходства. В случае нескольких

совпадений возможно случайное назначение чтения одной из этих OTE. В базе каждая запись является представительной последовательностью соответствующего OTE, полученнного ранее в результате кластерного анализа. В то время как порог сходства можно варьировать, традиционно в метагеномных исследованиях используется значение 97% сходства как эвристическая оценка степени сходства 16S рРНК внутри одного бактериального вида. Однако данное значение не является абсолютным: с одной стороны, и в пределах одного бактериального вида могут встречаться бактерии с сильно различающимися последовательностями этого гена, с другой стороны, у двух разных видов могут быть идентичные последовательности (например, Escherichia и Shigella).

В данном техническом решении могут применяться в некоторых вариантах осуществления две другие основные стратегии идентификации OTE, известные из уровня техники: поиск de novo и гибридный подход (сочетающий элементы поиска на основании шаблона и поиска de novo).

Накопленные последовательности по 16S рРНК секвенированию микробиоты сводятся в объединенные базы данных и филогенетически аннотируются. Среди наиболее используемых баз данных в уровне техники Greengenes (курируемая база полных последовательностей гена 16S рРНК), SILVA (включает последовательности не только 16S, но и 18S, 23S/28S для эукариот), RDP (аннотация менее унифицирована, но объем выше, чем у Greengenes).

В результате обработки набора метагеномов в формате 16S рРНК получается таблица относительной представленности, которая отражает количество чтений, отнесенных к каждой таксономической единице (OTE) из базы данных для каждого образца. Прореженная таблица относительной представленности может определяться по следующему принципу:

a. Если суммарное число чтений для образца по каждой таксономической единице меньше порогового значения (например, 5000), такой образец исключается из дальнейшего анализа как не подходящий по качеству и подлежащий повторному секвенированию.

b. Если суммарное число чтений для образца по каждой таксономической единице больше или равно пороговому значению (например, 5000), то число чтений для каждой таксономической единицы пропорционально нормируется таким образом, чтобы суммарное число чтений для образца стало равно пороговому значению (например, 5000).

В некоторых вариантах осуществления относительная представленность нормируется. Для этого для каждого образца суммируется количество его чтений, которые успешно откартировались на референсную базу. Нормированная представленность для каждого таксона рассчитывается как количество чтений, отнесенных к этому таксону для данного образца, деленное на общую сумму откартированных чтений для этого образца и помноженное на 100%. Из полученных значений нормированной представленности

составляется нормированная таблица представленности, содержащая процент чтений, отнесенных к каждому таксону из базы данных для каждого образца.

Из непрореженных таблиц относительной представленности по OTU блок **204** таксономического анализа формирует прореженные таблицы представленности по другим таксономическим уровням (родам, семействам и т.п.). Для каждого таксономического уровня применяется следующий способ:

a. Количество чтений в образце для всех OTU, которые относятся к данному таксономическому уровню, суммируются;
b. Из полученных сумм составляется таблица представленности для данного таксономического уровня.

Далее на основании прореженной таблицы представленности (таблица, которая отражает количество чтений, отнесенных к каждому таксону одного из таксономических уровней для каждого образца) оценивается относительная представленность групп микробных генов.

Для этого прореженную таблицу представленности нормализуют на число копий 16S рРНК. Для этого количество чтений, отнесенных к каждому из таксонов для каждого образца, делят на оценочное количество копий гена 16S рРНК, характерное для данного таксона.

Затем для каждого гена определяют его представленность для каждого образца следующим образом: используя существующую таблицы наличия тех или иных метаболических путей и/или групп генов в них входящих в различных микроорганизмах, для каждого образца составляется таблица представленности групп генов (EC) и метаболических путей, которая пропорциональная представленности микроорганизмам, в которых эти гены/метаболические пути содержатся.

В итоге из получившихся сумм составляется таблица представленности генов в каждом образце.

Таксономический профиль популяции микробных сообществ по 16S рРНК, полученный блоком **204** таксономического анализа, применяется для оценки важных характеристик популяции пользователя: альфа- и бета-разнообразия. Они являются численными величинами, характеризующими разнообразие одиночного микробного сообщества и различие между двумя сообществами, соответственно. Чем больше чтений на образец будет секвенировано, тем больше различных видов будет найдено, причем с увеличением числа чтений происходит насыщение; для сообщества низкой сложности оно будет происходить быстрее, чем для сложного; поэтому при вычислении альфа-разнообразия учитывается число чтений на образец Среди наиболее широко применяемых оценивателей альфа-разнообразия может применяться в данном техническом решении филогенетическое разнообразие (пропорционально доле дерева жизни, которую покрывает собой сообщество), а также индексы Chao 1 и ACE.

Предварительно осуществляется фильтрация малопредставленных таксонов - например, по следующему принципу: оставляются таксоны, представленность которых превышает 0,2% от общей микробной представленности не менее чем в 10% образцах.

Далее на основании нормированной таблицы представленности, содержащейся в блоке **204** таксономического анализа, блок **205** определения риска заболеваний предварительно осуществляет предобработку и оценивает аномальность состава микробиоты в образце. Для каждого образца проверяется суммарный процент чтений, относящихся к каждому из таксонов из списка оппортунистических патогенов. Образец, в котором суммарный процент превышает фиксированный - например, 20% - считается аномальным. В некоторых вариантах осуществления учитывается процентная доля отдельно взятых таксонов из списка, в том числе с возможностью взвешивания их вкладов в оценку аномальности. В некоторых вариантах осуществления дополнительно для каждого образца проверяется процент чтений, относящихся к роду бифидобактерий. Образец, в котором этот процент превышает фиксированный - например, 50% - считается аномальным. В некоторых вариантах осуществления относительные представленности таксонов для каждого образца могут просматриваться экспертом на предмет обнаружения нетипичной представленности ряда выделенных таксонов, в том числе условно патогенных. На основании суждения эксперта и/или результатов работы алгоритмов машинного обучения образец также может считаться аномальным. Образцы, признанные аномальными, исключаются из дальнейшего анализа. Пользователям, которым принадлежали эти образцы, высылается уведомление о нетипичном составе микробиоты.

Затем на основании нормированной таблицы представленности и базы данных связей бактерий и болезней, блок **205** определения риска заболеваний определяет защищенность от заболеваний пользователя по данным о микробиоте.

Предварительно из микробиотных образцов популяционной выборки создается так называемый контекст — референсные данные для сравнения - следующим образом.

Для каждой бактерии (рода или иного таксона) рассчитывается набор фиксированных перцентилей по представленности - например, 33%- и 67%-перцентили. Иначе говоря, получают два порога представленности: треть образцов из популяционной выборки имеет меньшую представленность по данной бактерии, чем меньший порог; а треть образцов из популяционной выборки имеет большую представленность по данной бактерии, чем больший порог.

В некоторых вариантах реализации пороговые значения для перцентилей могут быть заранее вычислены на основании результатов статистического анализа относительной представленности таксона у пациентов с данным заболеванием (или лиц с повышенным риском данного заболевания) по сравнению со здоровыми индивидами.

Для каждого образца блок **205** определения риска заболеваний определяет его защищенность от каждой болезни. Каждому заболеванию заранее сопоставлен список прямо или обратно ассоциированных с ним микробных таксонов (биомаркеров). Далее образцу по данной болезни выставляется величина защищенности, которая может быть рассчитана по следующим правилам:

Для данного образца каждому микроорганизму (таксону) из числа биомаркеров данного заболевания присваивается значение 0, N(k) или M(k) (где k - номер биомаркера, а N(k) и M(k) - специфичные для данного биомаркера данного заболевания константы) по следующим правилам:

i. Если данная бактерия не содержится в данном образце, этой бактерии присваивается число 0.

ii. Если представленность данной бактерии в данном образце ниже верхнего и выше нижнего перцентиля, этой бактерии присваивается число 0.

iii. Если данная бактерия по данным о связи бактерий и болезней не влияет на данную болезнь, этой бактерии присваивается число 0.

iv. Если представленность данной бактерии в данном образце превышает верхний перцентиль и, по данным таблицы связей бактерий и болезней, положительно ассоциирована с этим заболеванием, этой бактерии присваивается число -M(k).

v. Если представленность данной бактерии в данном образце ниже нижнего перцентиля и, по данным о связи бактерий и болезней, положительно ассоциирована с этим заболеванием, этой бактерии присваивается число N(k).

vi. Если представленность данной бактерии в данном образце выше верхнего перцентиля и, по данным о связи бактерий и болезней, отрицательно ассоциирована с этим заболеванием, этой бактерии присваивается число 1.

vii. Если представленность данной бактерии в данном образце ниже нижнего перцентиля и, по данным о связи бактерий и болезней, отрицательно ассоциирована с этим заболеванием, этой бактерии присваивается число -1.

В некоторых примерных вариантах реализации для всех биомаркеров (k=1,...) N(k)=M(k)=1.

Данному образцу присваивается величина защищенности от данной болезни, равное сумме величин, присвоенных бактериям-биомаркерам на предыдущем шаге.

Для каждой болезни рассчитываются фиксированные перцентили по защищенности, например, 33%- и 67%-перцентили. Иначе говоря, получаются два порога защищенности: треть образцов из популяционной выборки имеет меньшую защищенность по данной болезни, чем меньший порог; а треть образцов из популяционной выборки имеет большую защищенность по данной болезни, чем больший порог.

Затем шкалированная величина защищенности для пользователя определяется блоком **205** определения риска заболеваний так:

Для каждой болезни рассчитывается величина защищенности микробиотой по способу, описанному выше при анализе контекста.

Затем защищенность для пользователя шкалируется по следующему правилу:

a. Нижний перцентиль защищенности от данного заболевания, вычисленный по контексту, принимается в новой шкале за 0;

b. Верхний перцентиль защищенности от данного заболевания, вычисленный по *контексту*, принимается в новой шкале за 10;

c. Значение защищенности пользователя пропорционально нормализуется по новой шкале.

Если значение защищенности по новой шкале меньше 4, ему присваивается значение 4. Полученное значение является уровнем защищенности образца от данной болезни.

В других вариантах осуществления технического решения могут быть использованы иные перцентили. Также каждый таксон может иметь свой индивидуальный вес, складывающийся из оценки его влияния на признак и его представленности в конкретном образце, отличный от 1, -1 или 0.

В рекомендациях для пользователя предлагается повысить относительную представленность бактерий, которые отрицательно ассоциированы с болезнью и имеют низкую (ненулевую) и/или нормальную представленность (лежат между верхним и нижним перцентилем), а также если они не являются положительно ассоциированными с другими болезнями.

В некоторых вариантах осуществления блок **205** определения риска заболеваний определяет состав наследственных моногенных заболеваний. Для этого может быть использован список мутаций и патогенных аллелей для наследственных заболеваний. Эти данные содержат информацию только о патогенных мутациях. Образец пользователя содержит идентификатор мутации и генотипы.

Блок **205** определения риска заболеваний проверяет каждую мутацию на наличие патогенной аллели и оценивает статус заболевания, например следующим образом:

a. 0 — нет патогенной аллели;

b. 1 — одна и только одна мутация с одной патогенной аллелью;

c. 2 — одна и более мутаций с обеими патогенными аллелями;

d. 3 — две и более мутаций с одной патогенной аллелью (составная гетерозигота);

Для одного заболевания, в одном образце могут быть первые три случая одновременно, порядок назначения: 2 > 3 > 1.

В уровне техники существует следующие типы наследования мутаций: аутосомно-рецессивный (AR), аутосомно-доминантный (AD), X-сцепленный рецессивный (XR), X-сцепленный доминантный (XD), Y-сцепленный (Y), митохондриальный (MT).

В случае оценки статуса заболевания как 2 (одна и более мутаций с обеими патогенными аллелями) или 3 (две и более мутаций с одной патогенной аллелью), порядок назначения итогового типа наследования при комбинации AD и AR — AD следующий AR — AD>AR; при комбинации XD и XR — XD > XR. В итоге на выходе блок 205 определения риска заболеваний выдает статус заболевания с типом наследования.

В некоторых вариантах осуществления блок 205 определения риска заболеваний может ранжировать пользователей на основании полученных данных (индивидуальные данные, полученные в результате расчета рисков, а также данные метагеномного анализа). Для каждого заболевания блок 205 определения риска заболеваний ранжирует всех пользователей по величине показателя относительного риска и делит их, например, на пять групп таким образом, чтобы первая группа включала 10% пользователей, вторая - 20%, третья - 40%, четвертая - 20%, пятая - 10%.

Далее блок 205 определения риска заболеваний формирует например следующее распределение пользователей по группам риска:

1. Высокий риск — от 0-го до 10-го перцентиля
2. Повышенный риск — от 10-го до 30-го перцентиля
3. Средний риск — от 30-го до 70-го перцентиля
4. Умеренный риск — от 70-го до 90-го перцентиля
5. Низкий риск — от 90-го до 100-го перцентиля

По результатам метагеномного анализа, как было показано выше, блок 205 определения риска заболеваний определяет степень защищенности организма от развития некоторых заболеваний. Уровень защищенности может выражаться в целых числах по шкале от 0 до 10. Блок 205 определения риска заболеваний использует следующие принципы включения данных о степени защищенности микробиотой в ранжирование рисков по генетике:

- 0–5 баллов — пользователь перемещается на группу риска выше, чем группа, определившаяся по результатам расчета рисков (но не выше первой);
- 6–7 баллов — группа риска остается неизменной;
- 8–10 баллов — пользователь перемещается на группу риска ниже, чем группа, определившаяся по результатам расчета рисков (но не ниже пятой).

В случае, если пользователь сдавал только тест микробиоты (генетические данные не учитываются), распределение рисков может выглядеть следующим образом:

1. Высокий риск — от 0 до 3 баллов;
2. Повышенный риск — от 4 до 5 баллов;

3. Средний риск — от 6 до 7 баллов;
4. Умеренный риск — от 8 до 9 баллов;
5. Низкий риск — 10 баллов.

Специалисту в данной области техники очевидно, что способ ранжирования и количество баллов является примерным, а не ограничивающим и не влияет на сущность технического решения.

В некоторых вариантах осуществления при расчете риска может быть сделано допущение, что все факторы (внешние и генетические) независимы друг от друга. Для определения риска заболевания может использоваться логистическая модель, отправной точкой в которой считается средняя встречаемость заболевания в популяции и учитываются вклады внешних и генетических факторов риска.

Для генетических факторов риска численные значения вклада могут быть извлечены из исследований типа широкогеномный поиск ассоциаций (GWAS) по данному заболеванию. Например, для такого заболевания как "сахарный диабет 2 типа" это исследование Morris, A.P. et al., 2012. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nature Genetics, 44(9), pp.981–990.

Для внешних факторов риска используются источники информации, где показана связь между тем или иным фактором риска и риском развития данного заболевания. Например, для сахарного диабета это могут быть следующие факторы и статьи:

riskFactor	OR	PMID

BMI у женщин	BMI < 22 - 1 RR BMI 22-23 - 2.9 RR BMI 23-24 - 4.3 RR BMI 24-25 - 5.0 RR BMI 25-27 - 8.1 RR BMI 27-29 - 15.8 RR BMI 29-31 - 27.6 RR BMI 31-33 - 40.3 RR BMI 33-35 - 54.0 RR BMI ≥ 35 - 93.2 RR	www.ncbi.nlm.nih.gov/pubmed/7872581
BMI у мужчин	BMI < 24 - 1 RR BMI 24-25 - 1.6 RR BMI 25-27 - 2.3 RR BMI 27-29 - 4.8 RR BMI 29-31 - 8.1 RR	www.ncbi.nlm.nih.gov/pubmed/7988316

	BMI 31-33 - 13.8 RR BMI 33-35 - 26.9 RR BMI ≥ 35 - 50.7 RR	
Умеренное употребление алкоголя	0,87 RR (Male) 0,60 RR (Female)	www.ncbi.nlm.nih.gov/pubmed/19875607
Курение	1,44 RR	www.ncbi.nlm.nih.gov/pubmed/18073361
СД у родственников	2,44 HR	www.ncbi.nlm.nih.gov/pubmed/23052052
Присутствие в рационе свежих фруктов	0,91 RR	www.ncbi.nlm.nih.gov/pubmed/26816602
Присутствие в рационе свежих овощей	0,87 RR	www.ncbi.nlm.nih.gov/pubmed/26816602
Синдром поликистозных яичников	8,8 OR	www.ncbi.nlm.nih.gov/pubmed/24081730
Присутствие в рационе злаковых	0,68 RR	www.ncbi.nlm.nih.gov/pubmed/24158434
Употребление кофе	0,7 RR	www.ncbi.nlm.nih.gov/pubmed/24459154

Употребление сладких напитков	1,26 RR	www.ncbi.nlm.nih.gov/pubmed/20693348
Употребление мяса 50 г/день	1,51 RR	www.ncbi.nlm.nih.gov/pubmed/21831992
Низкий уровень ЛПВП	< 1.0 mmol/L — 5,74 HR 1,0-1,4 mmol/L — 2,72 HR 1,4-1,7 mmol/L — 1,44 HR ≥ 1,7 mmol/L — 1,0 HR	www.ncbi.nlm.nih.gov/pubmed/25972569
Высокий уровень адипонектина	0,72 OR	www.ncbi.nlm.nih.gov/pubmed/19584347
Высокий уровень гормон-связывающего глобулина у мужчин	0,12 OR	www.ncbi.nlm.nih.gov/pubmed/19657112
Высокий уровень гормон-связывающего глобулина у женщин	0,11 OR	www.ncbi.nlm.nih.gov/pubmed/19657112
Высокий СРБ (выше верхней границы референса)	1,28 OR	www.ncbi.nlm.nih.gov/pubmed/23264288
Сон больше 9 часов или меньше 6 часов в сутки	1,28 RR (< 6 ч/сутки)	www.ncbi.nlm.nih.gov/pubmed/19910503

1,48 RR (> 9
ч/сутки)

В некоторых вариантах осуществления определяют по составу образца микробиоты относительную представленность групп генов согласно номенклатуре ЕС (Enzyme commission number), входящих в состав метаболического пути синтеза масляной кислоты. Их представленность соотносится с контекстными данными, и каждой группе генов назначается балл по способу, аналогичному описанному выше для расчета защищенности от заболеваний. Контекстные данные представленности микроорганизмов содержат распределение представленности прокариотических микроорганизмов, значения для 33% и 67% перцентилей. Далее определяется определяется балл от 4 до 10 и это будет являться баллом синтеза масляной кислоты. В том случае, если этот балл менее порогового значения, находятся те таксоны, которые потенциально несут в своем геноме те группы генов, которые на первом шаге оказались не представлены (попали ниже 33%-перцентиля), и проверяется их представленность по контекстным данным. Если эти таксоны также попали ниже 33%-перцентиля, то они используются в дальнейшем при формировании рекомендаций пользователю.

В других вариантах осуществления происходит определение представленности групп генов ЕС в образце, входящих в состав пути синтеза витаминов, для каждого в отдельности из числа витаминов B1, B2, B3, B5, B6, B7, B9, K. Их представленность соотносится с контекстными данными, каждому ЕС назначается балл по способу, аналогичному в пункте 116. Далее считается средний балл по всем витаминам и целое от него будет являться баллом синтеза витамина. В том случае если этот балл менее порогового значения, находятся те микроорганизмы, которые потенциально обладают в своем геноме теми ЕС, которые на первом шаге оказались недопредставленны (попали в 33%), и проверяется их представленность по контекстным данным. Если эти микроорганизмы также попали в 33%, то они используются в дальнейшем способе рекомендаций для пользователя.

В некоторых вариантах реализации может быть использована другая номенклатура микробных функциональных групп генов - например, KEGG Orthology groups или группы генов из базы MetaCyc.

В других вариантах осуществления происходит определение потенциала расщепления для каждого типа пищевого волокна из заранее заданного набора. Производится оценка суммы представленности относительно контекстных данных тех микроорганизмов, для которых из базы данных ассоциаций известно, что они способны их расщеплять. Если их суммарная представленность попадает в 33%, алгоритм решает,

что потенциал расщепления этого волокна низок. Для каждого волокна рассчитывается балл от 4 до 10 в зависимости от значения их суммарной представленности. Суммарный потенциал расщепления волокон считается как целое среднее значение баллов по всем пищевым волокнам.

В некоторых вариантах осуществления блок 202 контроля качества передает генетические данные в блок 206 определения признаков пользователя. Признак (trait) в генетической терминологии - это измеряемая характеристика пользователя. Признак может быть получен из заполненного пользователем опросника, генетического теста, носимых гаджетов, медкарты и т. д.

Примеры признаков пользователя:

- непереносимость лактозы (дискретные состояния: есть предрасположенность, нет предрасположенности, неизвестно);
- возраст (непрерывное состояние: 30 лет, 49 лет и т. д.);
- активность CYP2D6 (дискретные состояния: ультрабыстрый метаболизатор, нормальный метаболизатор, плохой метаболизатор);
- риск ожирения (непрерывное состояние: риск 50%, риск 43,4% и т. д.).

В некоторых вариантах осуществления признаки могут быть объединены в группы по наследственным заболеваниям, реакции на лекарства, признакам по питанию, признакам по спорту, гаплотипам.

В зависимости от типа признака пользователя, признак может иметь два или более возможных состояний (states). В некоторых вариантах осуществления состояния могут быть дискретными или непрерывными, но не одновременно для одного признака. Пока признак не рассчитан для пользователя он имеет неопределенное состояние. В некоторых вариантах осуществления признак пользователя зависит от состояний других признаков. Все возможные комбинации состояний зависимостей составляют область определения признака.

Признаки могут быть изменяемыми (потребление кофе), неизменяемыми (активность CYP2D6, статус фенилкетонурии) и условно изменяемыми (некоторые риски), которые зависят от изменяемых признаков.

Признак может иметь срок давности, спустя который он будет инвалидирован, то есть перейдет в неопределенное состояние. Например, концентрация холестерина в крови по результатам анализа будет действительна в течение года, а затем признак вернется в состояние неопределенного признака пользователя.

Изменяемые и условно изменяемые признаки хранят историю изменений своего состояния, в том числе при инвалидации по истечении срока давности.

Так как признаки могут ссылаться на другие признаки в своей интерпретации, при наполнении системы признаками блок 206 определения признаков образует направленный граф зависимостей между признаками. Узлы графа, которые ни на что не

ссылаются, — это узлы исходных данных (мутаций, ответов на вопросы, микробиоты). Все остальные узлы напрямую или опосредованно зависят от узлов исходных данных.

Блок 206 определения признаков пользователя осуществляет определение состояний признаков для конкретного пользователя путем редукции графа, начиная с узлов исходных данных.

Если какой-то из признаков изменил свое состояние, например, пользователь по-другому ответил на вопрос в опроснике, все зависимые от опросника признаки будут пересчитаны, т.е. обновлены. Пересчет одних признаков вызовет пересчет других, пока не будет достигнут конец графа зависимостей.

Перед определением состояния признака блок 206 определения признаков осуществляет проверку на наличие циклов в графе зависимостей, и при наличии циклов блок 206 не дает запустить редукцию графа.

На основании показателя по меньшей мере одного признака, состояния данного признака и генетических данных (однонуклеотидный полиморфизм, пол и т.д.), блок 206 определения признаков может сформулировать интерпретацию для пользователя (по спорту, по питанию, по личным качествам и т.д., не ограничиваясь) например в следующем виде:

TRAIT	STATES	PREDICATE	TEXT
femaleHormoneBindingGlobulinLevels	high	and (gender.female, "rs727428 FWD C/C")	У вас предрасположенность к высокому уровню гормон-связывающих глобулинов. За это отвечает вариант гена SHBG. В норме, уровень свободного эстрадиола может быть повышен.
femaleHormoneBindingGlobulinLevels	mean	and (gender.female, "rs727428 FWD T/C"	У вас предрасположенность к среднему уровню гормон-связывающих

		)	глобулинов. За это отвечает вариант гена SHBG.
femaleHormoneBindingGlobulinLevels	low	and (gender.female, "rs727428 FWD T/T")	У вас предрасположенность к низкому уровню гормон-связывающих глобулинов. За это отвечает вариант гена SHBG. В норме, уровень свободного эстрадиола может быть понижен.

В некоторых вариантах осуществления блок 206 определения признаков определяет признак пользователя по питанию на основании данных микробиоты. Для этого используются результаты расчета защищенности от заболеваний, потенциала к расщеплению пищевых волокон, синтеза короткоцепочечных жирных кислот, синтеза витаминов, а также база данных ассоциаций между продуктами питания и представителями кишечной микробиоты. Данная база данных формируется с использованием алгоритмов машинного анализа текста совместно с ручным наполнением на основании фактов о продуктах, прием которых положительно ассоциирован с теми или иными микроорганизмами, обитающих в кишечнике человека.

Если итоговый балл по одному из данных (например, защищенность от заболеваний) меньше заранее заданного порогового значения, то из базы данных ассоциаций берутся те продукты питания, которые ассоциированы с ростом тех микроорганизмов, представленность которых оказалась недостаточной. Чем чаще рекомендуется тот или иной продукт для данного пользователя по результатам различных алгоритмов, тем выше его ранг и вероятность его рекомендации для пользователя.

Блок 207 формирования рекомендаций для пользователя выполнен с возможностью формирования рекомендации пользователю на основании данных блока 205 определения риска заболеваний и блока 206 определения признаков пользователя.

На вход данному блоку подаются индивидуальные данные, полученные в результате определения признаков, рисков и статусов носительства заболеваний, а также данные метагеномного анализа с других блоков системы.

Работа блока 207 формирования рекомендаций для пользователя основана на выполнении условия, приводящего к выдаче результата. Условие является сочетанием простых логических операций над входными данными. Результатом является текст рекомендации, направленный на то, чтобы побудить пользователя к выполнению определенного набора действий. Рекомендации в некоторых вариантах осуществления делятся на следующие группы:

- рекомендации по нежелательным видам нагрузки;
- рекомендации по изменению образа жизни;
- рекомендации по изменению приема в пищу тех или иных продуктов питания или их групп;
- рекомендации по посещению врача.

Группа рекомендаций по продуктам питания выдается с учетом как данных генотипирования, так и данных состава микробиоты кишечника или одного варианта.

В некоторых вариантах осуществления блок 207 формирования рекомендаций для пользователя формирует рекомендации по снижению риска, рекомендации по самодиагностике, рекомендации по обращению к врачу, рекомендации по признакам.

При формировании рекомендаций по снижению риска заболевания необходимым условием отображения рекомендаций является повышенный риск заболевания.

Рекомендация побуждает перейти из одного состояния признака в другое. То есть рекомендация относится к собственно признаку. Признак может иметь массив рекомендаций, размер которого равен количеству заданных переходов между разными состояниями. Сам переход может произойти только тогда, когда меняются исходные данные пользователя, затрагивающие данный признак, и выполняется переинтерпретация.

Переход может иметь дополнительные условия, при которых он будет осуществлен. Например, пол пользователя может влиять на выдачу рекомендации.

Наличие определенного состояния какого-либо признака, определенного блоком 206 определения признаков может требовать определенное состояние другого признака, то есть, существует требователь и требуемое состояние. У каждого из состояний, среди которых предстоит выбрать целевое, есть вес, который складывается из весов всех требователей. В том случае, если требуемое состояние отличается от текущего, начинает работать переход и выдается рекомендация, побуждающая пользователя совершить этот переход. Выбор рекомендации, которую нужно дать пользователю, зависит от того, какое из требуемых состояний признака перевешивает.

Если у пользователя повышенный риск заболевания, ему выдаются рекомендации по коррекции изменяемых внешних факторов риска, которые у него находятся в состоянии, повышающем риск. Например, для сахарного диабета рекомендации могут выглядеть так:

"Пейте кофе каждый день.

Следует ежедневно включать в свой рацион кофе, не превышая, однако, допустимую для вас норму.

Включите в свой ежедневный рацион фрукты

Рекомендуется употреблять фрукты каждый день. Они богаты клетчаткой и полезными витаминами и микроэлементами.

Рекомендуется употреблять продукты, богатые витамином Е

Следует увеличить поступление в организм токоферола с пищей. Витамин Е является мощным антиоксидантом, необходим для мышечной ткани и иммунной системы".

Более подробно, сформированные рекомендации в блоке 207 формирования рекомендаций для пользователя могут включать в себя предоставление уведомлений пользователю относительно рекомендуемых терапевтических мер и / или других вариантов действий в отношении целей, связанных со здоровьем. Уведомления о рекомендациях могут предоставляться отдельному лицу посредством электронного устройства (например, персонального компьютера, мобильного устройства, планшета, смарт-часов и т. д.), причем с отображением в графическом интерфейсе пользователя (GUI). Рекомендации могут отображаться в приложении, веб-интерфейс в личном кабинете пользователя, в смс-сообщении или PUSH-уведомлении. В одном примере реализации веб-интерфейс персонального компьютера или ноутбука, связанный с пользователем, может предоставлять пользователю доступ к учетной записи пользователя, в котором учетная запись пользователя включает информацию о данных пользователя, подробную информацию о генетических данных и данных о составе микробиоты кишечника, а также уведомления о рекомендациях, сгенерированных в блоке **207** формирования рекомендаций. В другом варианте реализации приложение, выполняемое на персональном электронном устройстве (например, смартфон, смарт-часы, интеллектуальное устройство на голове), может быть сконфигурировано для предоставления уведомлений (например, на дисплее или звуковым образом и т. д.), в отношении рекомендаций, полученных с помощью блока **207** формирования рекомендаций. Уведомления могут дополнительно или альтернативно предоставляться непосредственно через человека, связанного с пользователем системы (например, смотритель, супруг, медицинский работник и т. д.). В некоторых дополнительных вариантах реализации уведомления могут дополнительно или альтернативно быть предоставлены человеку (например, специалисту в области здравоохранения), связанному с пользователем, в котором человек может влиять на осуществление рекомендаций (например, посредством рецепта, путем проведения терапевтического сеанса и т. д.). Однако рекомендации и уведомления о них могут быть предоставлены пользователю системы любым другим подходящим способом.

Хотя варианты осуществления описаны в связи с иллюстративной вычислительной системной средой, они могут быть реализованы с помощью многочисленных других вычислительных системных сред, конфигураций и устройств общего и специального назначения.

Примеры известных вычислительных систем, сред и/или конфигураций, которые могут являться подходящими для использования с аспектами изобретения, включают в себя, но без ограничения, мобильные вычислительные устройства, персональные компьютеры, серверные компьютеры, карманные устройства или ноутбуки, многопроцессорные системы, игровые консоли, системы на основе микропроцессоров, телеприставки, программируемую бытовую электронику, мобильные телефоны, сетевые персональные компьютеры, миникомпьютеры, суперкомпьютеры, распределенные вычислительные среды, которые включают в себя любые из упомянутых выше систем или устройств (например, фитнес-браслеты), и т.п. Такие системы или устройства могут принимать данные от пользователя в любом виде, в том числе от устройств ввода данных, таких как клавиатура или указательное устройство, через жестовый ввод и/или через речевой ввод.

Варианты осуществления изобретения могут быть описаны в общем контексте исполняемых с помощью компьютера команд, таких как программные модули или блоки, исполняемые одним или несколькими компьютерами или другими устройствами. Исполняемые с помощью компьютера команды могут быть организованы в один или несколько исполняемых с помощью компьютера компонентов или модулей. Обычно программные модули включают в себя, но без ограничения, подпрограммы, программы, объекты, компоненты и структуры данных, которые выполняют конкретные задачи или реализуют конкретные абстрактные типы данных. Аспекты изобретения могут быть реализованы с помощью любого количества и любой организации таких компонентов или модулей. Например, аспекты изобретения не ограничены конкретными исполняемыми с помощью компьютера командами или конкретными компонентами или модулями, проиллюстрированными на фигурах и описанными здесь. Другие варианты осуществления изобретения могут включать в себя другие исполняемые с помощью компьютера команды или компоненты, имеющие большую или меньшую функциональность, чем проиллюстрированная и описанная здесь.

Аспекты изобретения преобразовывают компьютер общего назначения в вычислительную систему специального назначения, выполненную с возможностью интерпретировать генетические данные пользователя и данные о составе микробиоты кишечника.

Следует понимать, что различные способы, описанные в данном документе, могут быть реализованы вместе с аппаратным или программным обеспечением или, при необходимости, с их комбинацией. Поэтому, способы и система данного предмета

изобретения, или некоторые аспекты или части такового, могут содержать программный код (т.е., инструкции), реализованные в материальном носителе, таком как гибкие дискеты, CD-ROM, накопители на жестких магнитных дисках, облачные хранилища или любые другие носители информации, при этом, когда программный код загружается и выполняется машиной, такой как компьютер, машина становится устройством для применения предмета изобретения. В случае выполнения программного кода на программируемых компьютерах, вычислительное устройство, в основном, содержит процессор, носитель информации доступный для чтения процессором (включая энергозависимую и энергонезависимую память и/или элементы памяти), по меньшей мере одно устройство ввода, и по меньшей мере одно устройство вывода. Одна или более программ может реализовывать или использовать процессы, описанные вместе с настоящим раскрытым предметом изобретения, например, посредством использования прикладного программного интерфейса (API), повторно используемых элементов управления, и тому подобного. Такие программы могут быть реализованы при помощи высокоуровневого процедурного или объектно-ориентированного языка программирования для обмена данными с компьютерной системой. Однако, при необходимости, программа(ы) могут быть реализованы на ассемблере, или машинном языке программирования. В любом случае, язык программирования может быть компилируемым или интерпретируемым языком, и это может быть комбинироваться с аппаратными реализациями.

Несмотря на то, что предмет изобретения был описан специфическим языком структурных особенностей и/или методологических функций, является понятным, что предмет изобретения получил определение в прилагаемых пунктах формулы патента, и нет необходимости ограничивать описанные выше характерные особенности или функции. В значительной степени, описанные выше характерные особенности и функции раскрываются как типовые разновидности реализации формулы изобретения.

ФОРМУЛА

1. Система рекомендаций для пользователя на основе его генетических данных и данных о составе микробиоты кишечника, содержащая:
   - блок получения первичных данных, выполненный с возможностью получения генетических данных пользователя и/или данных о микробиоте кишечника пользователя;
   - блок контроля качества, выполненный с возможностью проверки качества генетических данных пользователя и/или данных о микробиоте кишечника пользователя, полученных блоком получения первичных данных, причем генетические данные включают в себя однонуклеотидные полиморфизмы, а данные о микробиоте включают в себя чтения;
   - блок популяционного анализа генетических данных, выполненный с возможностью определения отцовской и материнской гаплогруппы, популяционного состава генетических данных;
   - блок таксономического анализа данных по микробиоте, выполненный с возможностью классификации чтений по базе данных последовательностей генов 16S рРНК;
   - блок определения риска заболеваний, выполненный с возможностью определения риска заболеваний, защищенности от заболеваний, а также проверки генотипа на наличие патогенных аллелей и оценки статуса носительства наследственных заболеваний,
   - блок определения признаков, выполненный с возможностью определения состояний признаков пользователя путем редукции графа зависимостей признаков;
   - блок формирования рекомендаций для пользователя, выполненный с возможностью формирования рекомендации пользователю на основании данных блока определения риска заболеваний и блока определения признаков пользователя.

2. Способ по п.1, характеризующийся тем, что блок получения первичных данных получает файлы секвенирования в формате FASTQ или FASTA, полученные с секвенатора.

3. Способ по п.1, характеризующийся тем, что блок контроля качества получает генетические данные пользователя из кремниевого биочипа посредством сканера биочипов.

4. Способ по п.3, характеризующийся тем, что генетические данные содержат данные о генотипах однонуклеотидных полиморфизмов пользователя, содержащие полиморфизмы X- и Y-хромосом.

5. Способ по п.1, характеризующийся тем, что блок контроля качества дополнительно определяет пол пользователя посредством подсчета количества однонуклеотидных полиморфизмов по X- и по Y-хромосомам.

6. Способ по п.5, характеризующийся тем, что в случае мужского пола блок контроля качества конвертирует однонуклеотидные полиморфизмы в гомозиготном состоянии с X- и Y-хромосом в однонуклеотидные полиморфизмы в гемизиготном состоянии. Однонуклеотидные полиморфизмы в гетерозиготном состоянии с X- и Y-хромосом отфильтровываются.

7. Способ по п.5, характеризующийся тем, что в случае женского пола все однонуклеотидные полиморфизмы с Y-хромосомой отфильтровываются и не попадают в итоговую выборку генетических данных.

8. Способ по п.1, характеризующийся тем, что блок контроля качества осуществляет отсеивание чтений со средним баллом качества ниже заранее заданного порогового.

9. Способ по п. 1, характеризующийся тем, что блок контроля качества с концов чтений удаляет позиции, имеющие низкий балл качества.

10. Способ по п. 1, характеризующийся тем, что блок контроля качества отсеивает постороннюю генетическую информацию в чтениях, имеющих биологическое или небиологическое происхождение, возникающую из-за прочтения артефактных последовательностей.

11. Способ по п. 1, характеризующийся тем, что блок популяционного анализа генетических данных определяет отцовскую гаплогруппу на основании дерева мутаций для Y хромосомы и генетических данных пользователя.

12. Способ по п. 1, характеризующийся тем, что блок популяционного анализа генетических данных определяет материнскую гаплогруппу на основании дерева мутаций для митохондрии и генетических данных пользователя.

13. Способ по п. 1, характеризующийся тем, что блок популяционного анализа генетических данных определяет популяционный состав на основании данных о генотипах людей из разных популяций и генетических данных пользователя.

14. Способ по п. 1, характеризующийся тем, что блок популяционного анализа генетических данных на основании генетических данных пользователя, набора унаследованных от неандертальца аллелей в определенных полиморфизмах определяет общее количество неандертальских аллелей.

15. Способ по п. 1, характеризующийся тем, что при классификации по базе данных последовательностей генов 16S рРНК, данная база данных содержит набор геномов бактерий и/или архей, встречающихся в кишечнике пользователя.

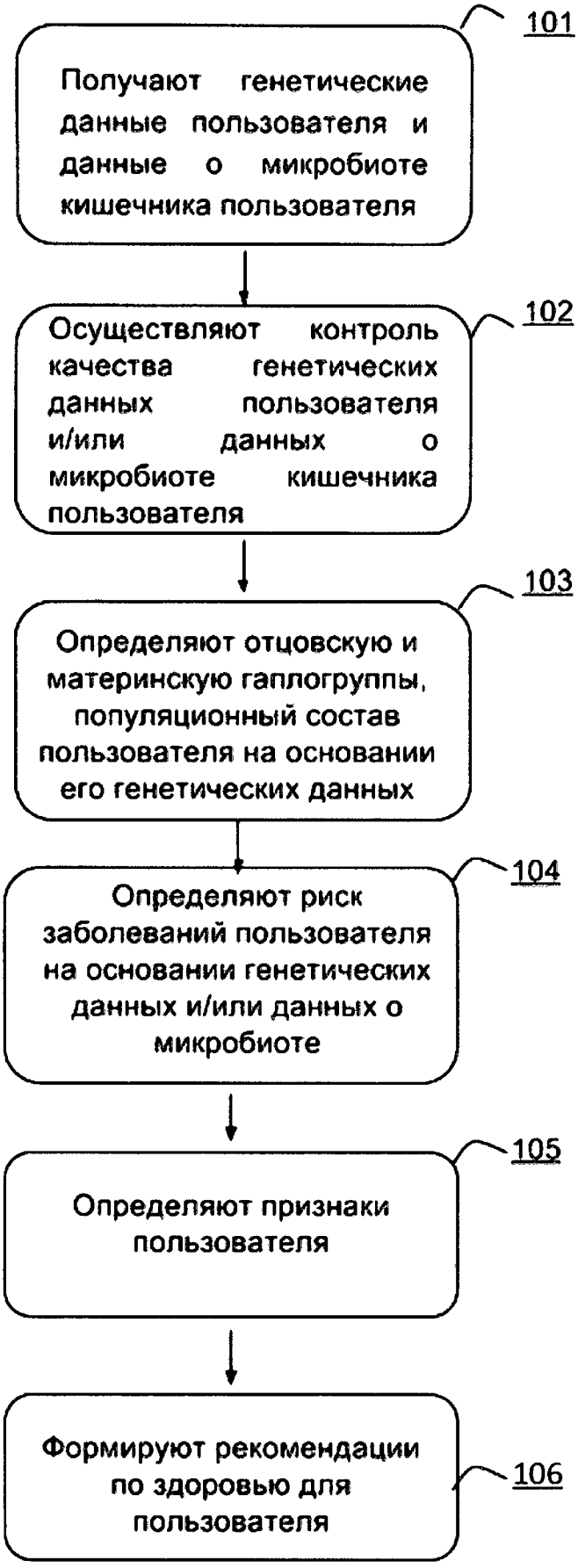
16. Способ по п. 1, характеризующийся тем, что блок таксономического анализа данных о микробиоте определяет относительную представленность микробного генома или вида.

17. Способ по п. 1, характеризующийся тем, что блок таксономического анализа данных о микробиоте формирует прореженные таблицы представленности по другим таксономическим уровням.

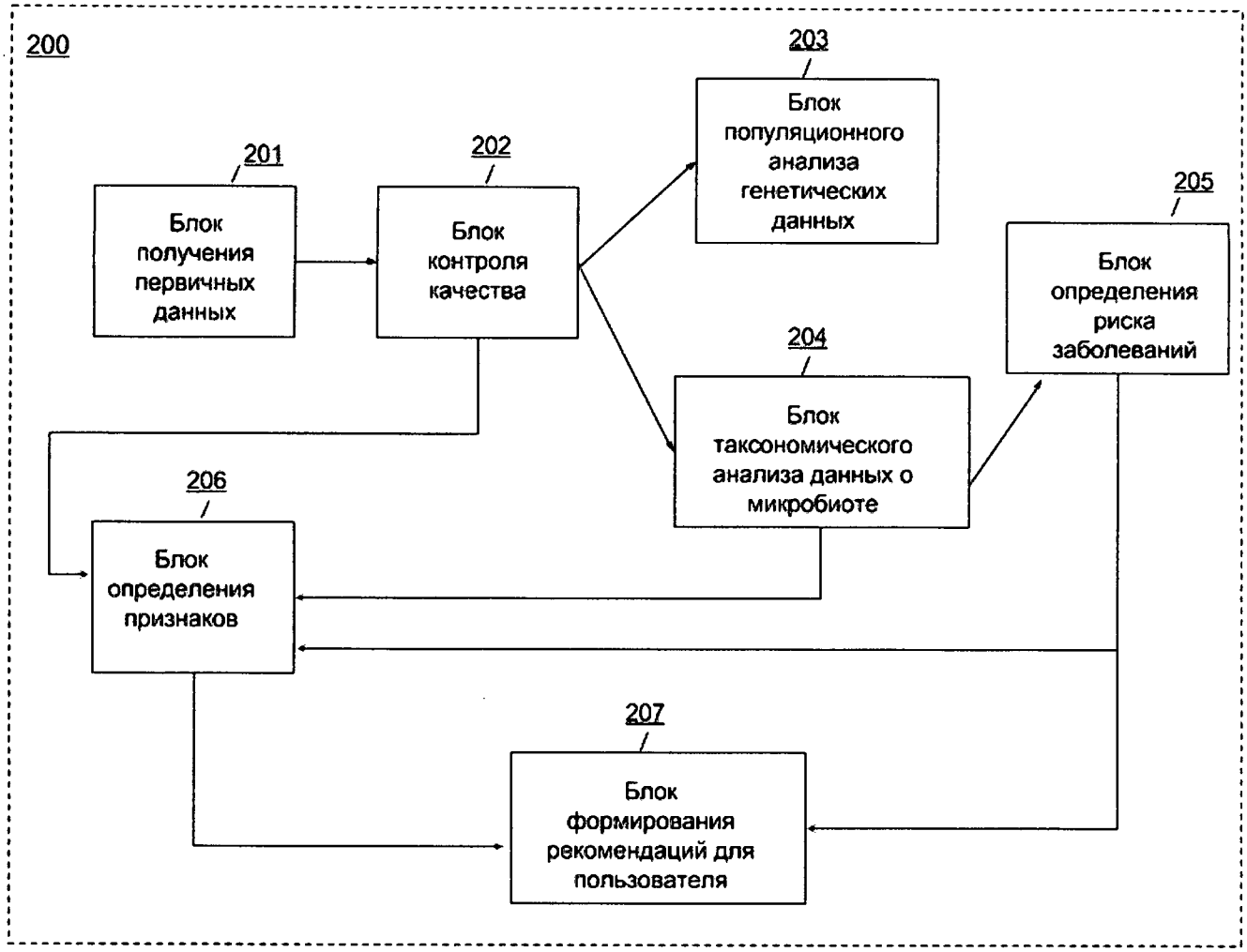
18. Способ по п. 1, характеризующийся тем, что блок определения риска заболеваний оценивает аномальность генетических данных посредством проверки суммарного процента чтений, относящихся к одному из таксонов из списка оппортунистических патогенов.

19. Способ по п. 1, характеризующийся тем, что блок определения риска заболеваний определяет защищенность от заболеваний пользователя по данным о микробиоте на основании референсных данных.

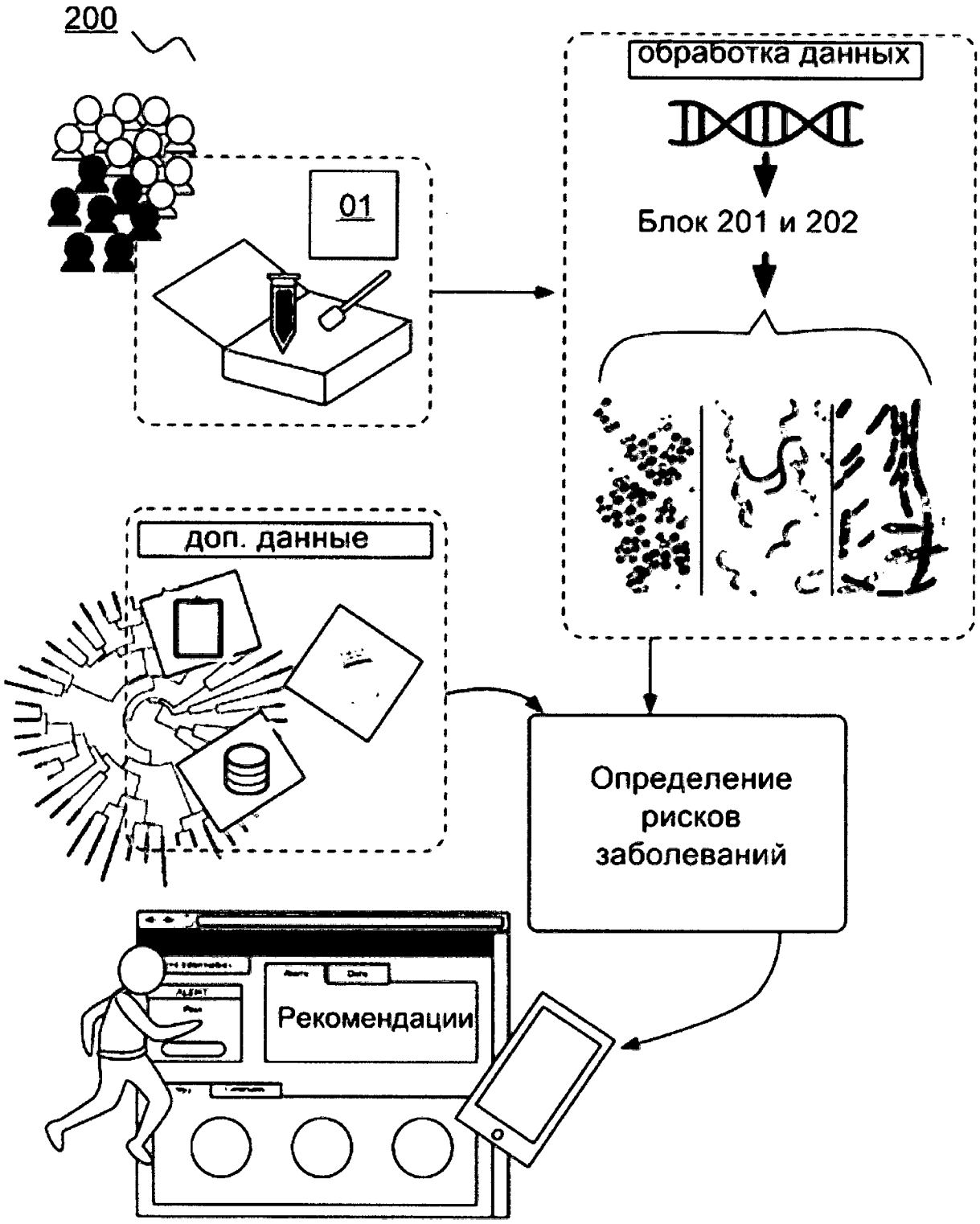
20. Способ по п. 1, характеризующийся тем, что блок определения признака осуществляет проверку на наличие циклов в графе зависимостей, и при наличии циклов блок не дает запустить редукцию графа.


101
Получают генетические данные пользователя и данные о микробиоте кишечника пользователя
102
Осуществляют контроль качества генетических данных пользователя и/или данных о микробиоте кишечника пользователя
103
Определяют отцовскую и материнскую гаплогруппы, популяционный состав пользователя на основании его генетических данных
104
Определяют риск заболеваний пользователя на основании генетических данных и/или данных о микробиоте
105
Определяют признаки пользователя
Формируют рекомендации по здоровью для пользователя
106

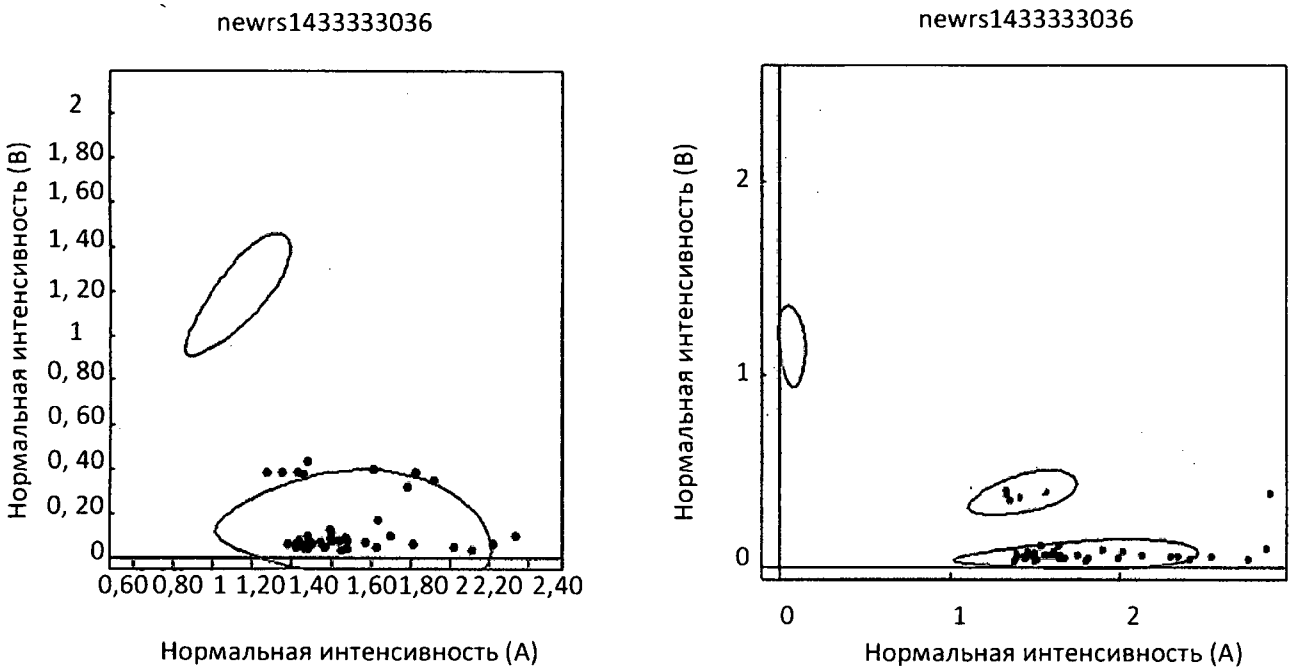
Фиг. 1


200
201
Блок получения первичных данных
202
Блок контроля качества
203
Блок популяционного анализа генетических данных
204
Блок таксономического анализа данных о микробиоте
205
Блок определения риска заболеваний
206
Блок определения признаков
207
Блок формирования рекомендаций для пользователя

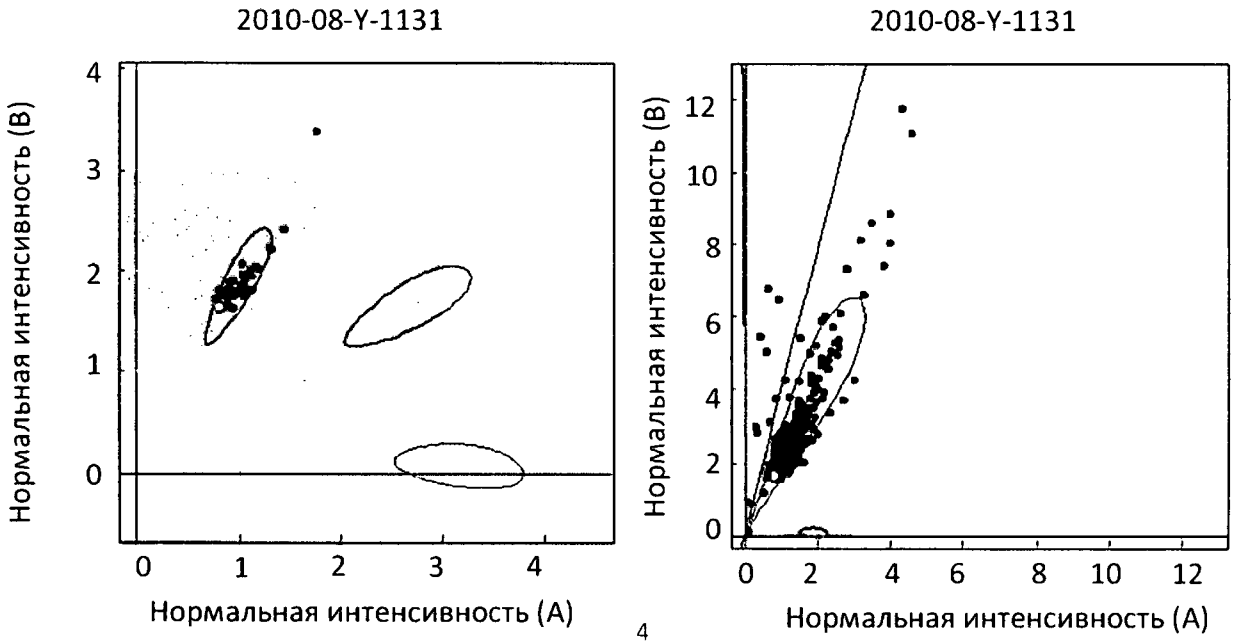
Фиг. 2


200
01
обработка данных
Блок 201 и 202
доп. данные
Определение рисков заболеваний
Рекомендации

Фиг. 3


newrs1433333036
Нормальная интенсивность (B)
2
1, 80
1, 60
1, 40
1, 20
1
0, 80
0, 60
0, 40
0, 20
0
0,60 0,80 1 1,20 1,40 1,60 1,80 2 2,20 2,40
Нормальная интенсивность (A)
newrs1433333036
Нормальная интенсивность (B)
2
1
0
0 1 2
Нормальная интенсивность (A)

Фиг. 4


2010-08-Y-1131
Нормальная интенсивность (B)
4
3
2
1
0
0 1 2 3 4
Нормальная интенсивность (A)
2010-08-Y-1131
Нормальная интенсивность (B)
12
10
8
6
4
2
0
0 2 4 6 8 10 12
Нормальная интенсивность (A)

4

Фиг. 5

INTERNATIONAL SEARCH REPORT

International application No.
PCT/RU 2017/000734

A. CLASSIFICATION OF SUBJECT MATTER

G16H 10/40 *(2018.01)*
G16H 50/20 *(2018.01)*
G06F 19/22 *(2011.01)*

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G16H 10/40, 50/20, G06F 19/00, 19/10, 19/22

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

DB WIPO (Patentscope), DB Espacenet, PatFT & AppFT USPTO, RUPTO, EAPATIS, CIPO, J-PlatPat, KIPRIS, K-PION, SIPO, DWPI, DEPATISnet, PSS; MedLine, PubMed Central, Scopus, EMBASE, ScienceDirect, РИНЦ (e-Library)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2016/0281166 Al (BHATTACHARJEE ARINDAM et al.) 29.09.2016, para. [0005] -[0006], [0011], [0021]-[0023], [0026], [0034], [0067], [0078], [0080], [0107]-[0122], [0125], [0132]- [0142], examples 1-3	1-20
Y	BODILY P.M. et al. "ScaffoldScaffolder: solving contig orientation via bidirected to directed graph reduction." Bioinformatics, 2016, 32(1), p.17-24, in particular p.20-23	1-20
Y	US 8 877442 B2 (THE BOARD OF TRUSTEES OF THE LELAND STANFORD JUNIOR UNIVERSITY) 04.11.2014, the claims	4-7, 11, 12
A	DEURENBERG RUUD H. et al. "Application of next generation sequencing in clinical microbiology and infection prevention." Journal of Biotechnology, 2017 (Available online 29 December 2016), 243, pp. 16-24, in particular abstract	1-20
A	KOSTRJUKOVA E. S. et al. "Variabelnost otnositelnogo soderzhania genomnoi DNK cheloveka pri metagenomnom analize mikrobioty kishechnika." Biomeditsinskaia khimiia, 2014, 60(6)? P./ 695-701, in particular abstract	1-20

☐ Further documents are listed in the continuation of Box C.

☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
29 May 2018 (29.05.2018)	21 June 2018 (21.06.2018)
Name and mailing address of the ISA/ RU	Authorized officer
Facsimile No.	Telephone No.

ОТЧЕТ О МЕЖДУНАРОДНОМ ПОИСКЕ

Номер международной заявки

PCT/RU 2017/000734

A. КЛАССИФИКАЦИЯ ПРЕДМЕТА ИЗОБРЕТЕНИЯ

G16H 10/40 *(2018.01)*
G16H 50/20 *(2018.01)*
G06F 19/22 *(2011.01)*

Согласно Международной патентной классификации МПК

B. ОБЛАСТЬ ПОИСКА

Проверенный минимум документации (система классификации с индексами классификации)

G16H 10/40, 50/20, G06F 19/00, 19/10, 19/22

Другая проверенная документация в той мере, в какой она включена в поисковые подборки

Электронная база данных, использовавшаяся при поиске (название базы и, если, возможно, используемые поисковые термины)

DB WIPO (Patentscope), DB Espacenet, PatFT & AppFT USPTO, RUPTO, EAPATIS, CIPO, J-PlatPat, KIPRIS, K-PION, SIPO, DWPI, DEPATISnet, PSS; MedLine, PubMed Central, Scopus, EMBASE, ScienceDirect, РИНЦ (e-Library)

C. ДОКУМЕНТЫ, СЧИТАЮЩИЕСЯ РЕЛЕВАНТНЫМИ:

Категория*	Цитируемые документы с указанием, где это возможно, релевантных частей	Относится к пункту №
Y	US 2016/0281166 A1 (BHATTACHARJEE ARINDAM et al.) 29.09.2016, параграфы [0005]-[0006], [0011], [0021]-[0023], [0026], [0034], [0067], [0078], [0080], [0107]-[0122], [0125], [0132]- [0142], примеры 1-3	1-20
Y	BODILY P.M. et al. "ScaffoldScaffolder: solving contig orientation via bidirected to directed graph reduction." Bioinformatics, 2016, 32(1), cc.17-24, особенно cc.20-23	1-20
Y	US 8 877442 B2 (THE BOARD OF TRUSTEES OF THE LELAND STANFORD JUNIOR UNIVERSITY) 04.11.2014, формула	4-7, 11, 12
A	DEURENBERG RUUD H. et al. "Application of next generation sequencing in clinical microbiology and infection prevention." Journal of Biotechnology, 2017 (Available online 29 December 2016), 243, pp. 16-24, особенно реферат	1-20
A	КОСТРЮКОВА Е. С. и др. "Вариабельность относительного содержания геномной ДНК человека при метагеномном анализе микробиоты кишечника." Биомедицинская химия, 2014, 60(6)? Сс/ 695-701, особенно реферат	1-20

☐ последующие документы указаны в продолжении графы С.

☐ данные о патентах-аналогах указаны в приложении

* Особые категории ссылочных документов:

"A" документ, определяющий общий уровень техники и не считающийся особо релевантным

"E" более ранняя заявка или патент, но опубликованная на дату международной подачи или после нее

"L" документ, подвергающий сомнению притязание(я) на приоритет, или который приводится с целью установления даты публикации другого ссылочного документа, а также в других целях (как указано)

"O" документ, относящийся к устному раскрытию, использованию, экспонированию и т.д.

"P" документ, опубликованный до даты международной подачи, но после даты испрашиваемого приоритета

"T" более поздний документ, опубликованный после даты международной подачи или приоритета, но приведенный для понимания принципа или теории, на которых основывается изобретение

"X" документ, имеющий наиболее близкое отношение к предмету поиска; заявленное изобретение не обладает новизной или изобретательским уровнем, в сравнении с документом, взятым в отдельности

"Y" документ, имеющий наиболее близкое отношение к предмету поиска; заявленное изобретение не обладает изобретательским уровнем, когда документ взят в сочетании с одним или несколькими документами той же категории, такая комбинация документов очевидна для специалиста

"&" документ, являющийся патентом-аналогом

Дата действительного завершения международного поиска	Дата отправки настоящего отчета о международном поиске
29 мая 2018 (29.05.2018)	21 июня 2018 (21.06.2018)
Наименование и адрес ISA/RU: Федеральный институт промышленной собственности, Бережковская наб., 30-1, Москва, Г-59, ГСП-3, Россия, 125993 Факс: (8-495) 531-63-18, (8-499) 243-33-37	Уполномоченное лицо: Т.Бабакова Телефон № 495 531 65 15