(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
16 March 2017 (16.03.2017)

WIPO | PCT

(10) International Publication Number
WO 2017/044885 A1

(51) International Patent Classification:
A61B 5/00 (2006.01) *G06F 19/12* (2011.01)
C12Q 1/68 (2006.01) *G06F 19/22* (2011.01)
G06F 19/00 (2011.01) *G06F 19/28* (2011.01)

(21) International Application Number:
PCT/US2016/051155

(22) International Filing Date:
9 September 2016 (09.09.2016)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

62/216,016	9 September 2015 (09.09.2015)	US
62/216,028	9 September 2015 (09.09.2015)	US
62/215,939	9 September 2015 (09.09.2015)	US
62/216,035	9 September 2015 (09.09.2015)	US
62/216,042	9 September 2015 (09.09.2015)	US

(71) Applicant: **UBIOME, INC.** [US/US]; 360 Langton Street, Suite 301, San Francisco, CA 94103 (US).

(72) Inventors: **APTE, Zachary**; c/o uBiome, Inc., 360 Langton Street, Suite 301, San Francisco, CA 94103 (US). **RICHMAN, Jessica**; c/o uBiome, Inc., 360 Langton Street, Suite 301, San Francisco, CA 94103 (US). **ALMONACID, Daniel**; c/o uBiome, Inc., 360 Langton Street, Suite 301, San Francisco, CA 94103 (US). **BEHBAHANI, Siavosh, Rezvan**; c/o uBiome, Inc., 360 Langton Street, Suite 301, San Francisco, CA 94103 (US).

(74) Agents: **RACZKOWSKI, David** et al.; Kilpatrick Townsend & Stockton LLP, Mailstop: IP Docketing - 22, 1100 Peachtree Street, Suite 2800, Atlanta, GA 30309 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: 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, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, 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) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (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).

[Continued on next page]

(54) Title: METHOD AND SYSTEM FOR MICROBIOME-DERIVED DIAGNOSTICS AND THERAPEUTICS FOR CONDITIONS ASSOCIATED WITH CEREBRO-CRANIOFACIAL HEALTH

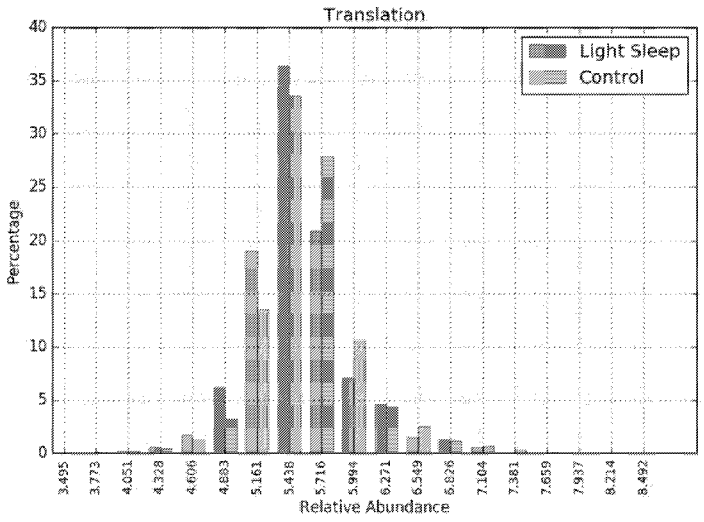

FIG. 10

(57) Abstract: Methods, compositions, and systems are provided for detecting one or more a cognition health issues by characterizing the microbiome of an individual, monitoring such effects, and/or determining, displaying, or promoting a therapy for the cognition health issue. Methods, compositions, and systems are also provided for generating and comparing microbiome composition and/or functional diversity datasets. Methods, compositions, and systems are also provided for generating a characterization model and/or therapy model for insomnia issues, light sleep issues, headache issues, sinusitis issues, and poor concentration issues.

WO 2017/044885 A1

Published:

— *with international search report (Art. 21(3))*

METHOD AND SYSTEM FOR MICROBIOME-DERIVED DIAGNOSTICS AND THERAPEUTICS FOR CONDITIONS ASSOCIATED WITH CEREBRO-CRANIOFACIAL HEALTH

CROSS-REFERENCE TO RELATED PATENT APPLICATIONS

[0001] The present patent application claims benefit of priority to U.S. Provisional Application No. 62/216,016, filed September 9, 2015; U.S. Provisional Application No. 62/216,028, filed September 9, 2015; U.S. Provisional Application No. 62/215,939, filed September 9, 2015; U.S. Provisional Application No. 62/216,035, filed September 9, 2015; and U.S. Provisional Application No. 62/216,042, filed September 9, 2015, the disclosures of each which are incorporated herein in the entirety and for all purposes.

BACKGROUND

[0002] A microbiome is an ecological community of commensal, symbiotic, and pathogenic microorganisms that are associated with an organism. The human microbiome comprises more microbial cells than human cells, but characterization of the human microbiome is still in nascent stages due to limitations in sample processing techniques, genetic analysis techniques, and resources for processing large amounts of data. Nonetheless, the microbiome is suspected to play at least a partial role in a number of health/disease-related states (e.g., preparation for childbirth, diabetes, auto-immune disorders, gastrointestinal disorders, rheumatoid disorders, neurological disorders, etc.).

[0003] Given the profound implications of the microbiome in affecting a subject's health, efforts related to the characterization of the microbiome, the generation of insights from the characterization, and the generation of therapeutics configured to rectify states of dysbiosis should be pursued. Current methods and systems for analyzing the microbiomes of humans and providing therapeutic measures based on gained insights have, however, left many questions unanswered. In particular, methods for characterizing certain health conditions and therapies (e.g., probiotic therapies) tailored to specific subjects based upon microbiome compositional or functional diversity features have not been viable due to limitations in current technologies.

[0004] As such, there is a need in the field of
and system for characterizing health conditions in an individualized and population-wide manner. This invention creates such a new and useful method and system.

BRIEF SUMMARY

[0005] A method for identification and classification of occurrence of a microbiome associated with a cerebro-craniofacial health issue or screening for the presence or absence of a microbiome associated with a cerebro-craniofacial health issue in an individual and/or determining a course of treatment for an individual human having a microbiome composition associated with a cerebro-craniofacial health issue, the method comprising:

providing a sample comprising microorganisms from the individual human;

determining an amount(s) of one or more of the following in the sample:

(a) bacteria and/or archaeal taxon or gene sequence corresponding to gene functionality as set forth in Tables A, B, C, D, or E;

(b) unicellular eukaryotic taxon or gene sequence corresponding to gene functionality,

comparing the determined amount(s) to a condition pattern or signature having cut-off or probability values for amounts of the microorganisms taxon and/or gene sequence for an individual having a microbiome composition associated with a cerebro-craniofacial health issue or an individual not having a microbiome composition associated with a cerebro-craniofacial health issue or both; and

identifying a classification of the presence or absence of the microbiome composition associated with a cerebro-craniofacial health issue and/or determining the course of treatment for the individual human having the microbiome composition associated with a cerebro-craniofacial health issue based on the comparing.

[0006] In embodiments described herein, reference is made to "bacteria" and "bacterial material" (e.g., DNA). Additionally or alternatively, other microorganisms and their material (e.g., DNA) can be detected, classified, and used in the methods and compositions described herein and thus every occurrence of "bacterial" or "bacterial material" or equivalents thereof apply equally to other microorganisms, including but not limited to archaea, unicellular eukaryotic organisms, viruses, or the combinations thereof.

[0007] In some embodiments, a method of determining a classification of occurrence of a microbiome indicative of a cerebro-craniofacial health issue or screening for the presence or

absence of a microbiome indicative of a cerebr

and/or determining a course of treatment for an individual human having a microbiome indicative of a cerebro-craniofacial health issue , the method comprising,

providing a sample comprising microorganisms including bacteria (or at least one of the following microorganisms including: bacteria, archaea, unicellular eukaryotic organisms and viruses, or the combinations thereof) from the individual human;

determining an amount(s) of one or more of the following in the sample:

bacteria taxon or gene sequence corresponding to gene functionality as set forth in Tables A, B, C, D, or E;

comparing the determined amount(s) to a disease signature having cut-off or probability values for amounts of the bacteria taxon and/or gene sequence for an individual having a microbiome indicative of a cerebro-craniofacial health issue or an individual not having a microbiome indicative of a cerebro-craniofacial health issue or both; and

determining a classification of the presence or absence of the microbiome indicative of a cerebro-craniofacial health issue and/or determining the course of treatment for the individual human having the microbiome indicative of a cerebro-craniofacial health issue based on the comparing.

[0008] In some embodiments, the determining comprises preparing DNA from the sample and performing nucleotide sequencing of the DNA.

[0009] In some embodiments, the determining comprises deep sequencing bacterial DNA from the sample to generate sequencing reads, receiving at a computer system the sequencing reads; and mapping, with the computer system, the reads to bacterial genomes to determine whether the reads map to a sequence from the bacterial taxon or a gene sequence from Tables A, B, C, D, or E; and determining a relative amount of different sequences in the sample that correspond to a sequence from the bacteria taxon or gene sequence corresponding to gene functionality from Tables A, B, C, D, or E.

[0010] In some embodiments, the deep sequencing is random deep sequencing.

[0011] In some embodiments, the deep sequencing comprises deep sequencing of 16S rRNA coding sequences.

[0012] In some embodiments, the method further comprises obtaining physiological, demographic or behavioral information from the individual human, wherein the disease

signature comprises physiological, demographi

determining comprises comparing the obtained physiological, demographic or behavioral information to corresponding information in the disease signature.

[0013] In some embodiments, the sample is a fecal, blood, saliva, cheek swab, urine or bodily fluid from the individual human.

[0014] In some embodiments, comprising determining that the individual human likely has a microbiome indicative of a cerebro-craniofacial health issue ; and treating the individual human to ameliorate at least one symptom of the microbiome indicative of a cerebro-craniofacial health issue . In some embodiments, the treating comprises administering a dose of one of more of the bacteria taxon listed in Tables A, B, C, D, or E to the individual human for which the individual human is deficient.

[0015] Also provided is method for determining a classification of the presence or absence of a microbiome indicative of a cerebro-craniofacial health issue and/or determine a course of treatment for an individual human having a microbiome indicative of a cerebro-craniofacial health issue. In some embodiments, the method comprises performing, by a computer system:

receiving sequence reads of bacterial DNA obtained from analyzing a test sample from the individual human;

mapping the sequence reads to a bacterial sequence database to obtain a plurality of mapped sequence reads, the bacterial sequence database including a plurality of reference sequences of a plurality of bacteria;

assigning the mapped sequence reads to sequence groups based on the mapping to obtain assigned sequence reads assigned to at least one sequence group, wherein a sequence group includes one or more of the plurality of reference sequences;

determining a total number of assigned sequence reads;

for each sequence group of a disease signature set of one or more sequence groups selected from Tables A, B, C, D, or E:

determining a relative abundance value of assigned sequence reads assigned to the sequence group relative to the total number of assigned sequence reads, the relative abundance values forming a test feature vector;

comparing the test feature vector to calibration
abundance values of calibration samples having a known status of a cerebro-craniofacial health issue ; and

determining the classification of the presence or absence of the microbiome indicative of a cerebro-craniofacial health issue and/or determining the course of treatment for the individual human having the microbiome indicative of a cerebro-craniofacial health issue based on the comparing.

[0016] In some embodiments, the comparing includes:

clustering the calibration feature vectors into a control cluster not having the microbiome indicative of a cerebro-craniofacial health issue and a disease cluster having the microbiome indicative of a cerebro-craniofacial health issue; and

determining which cluster the test feature vector belongs.

In some embodiments, the clustering includes using a Bray–Curtis dissimilarity.

In some embodiments, the comparing includes comparing each of the relative abundance values of the test feature vector to a respective cutoff value determined from the calibration feature vectors generated from the calibration samples.

[0017] In some embodiments, the comparing includes:

comparing a first relative abundance value of the test feature vector to a disease probability distribution to obtain a disease probability for the individual human having a microbiome indicative of a cerebro-craniofacial health issue, the disease probability distribution determined from a plurality of samples having the microbiome indicative of a cerebro-craniofacial health issue and exhibiting the sequence group;

comparing the first relative abundance value to a control probability distribution to obtain a control probability for the individual human not having a microbiome indicative of a cerebro-craniofacial health issue , wherein the disease probabilities and the control probabilities are used to determine the classification of the presence or absence of the microbiome indicative of a cerebro-craniofacial health issue and/or determining the course of treatment for the individual human having the microbiome indicative of a cerebro-craniofacial health issue .

[0018] In some embodiments, the sequence reads are mapped to one or more predetermined regions of the reference sequences.

[0019] In some embodiments, the disease sig
group and at least one functional group.

[0020] In some embodiments, the analyzing comprises deep sequencing.

[0021] In some embodiments, the deep sequencing reads are random deep sequencing reads.

[0022] In some embodiments, the deep sequencing reads comprise 16S rRNA deep sequencing reads.

[0023] In some embodiments, further comprising:

receiving physiological, demographic or behavioral information from the individual human; and

using the physiological, demographic or behavioral information in combination with the classification with the comparing of the test feature vector to the calibration feature vectors to determine the classification of the presence or absence of the microbiome indicative of a cerebro-craniofacial health issue and/or determining the course of treatment for the individual human having the microbiome indicative of a cerebro-craniofacial health issue .

[0024] In some embodiments, comprising preparing DNA from the sample and performing nucleotide sequencing of the DNA.

[0025] Also provided is a non-transitory computer readable medium storing a plurality of instructions that when executed, by the computer system, perform the method of any of those above.

BRIEF DESCRIPTION OF THE DRAWINGS

[0026] FIG. 1A is a flowchart of an embodiment of a method for determining a classification of the presence or absence of a cerebro-craniofacial health issue and/or determining the course of treatment for the individual human having a cerebro-craniofacial health issue.

[0027] FIG. 1B is a flowchart of an embodiment of a method for determining a classification of the presence or absence of a cerebro-craniofacial health issue and/or determining the course of treatment for an individual human having a cerebro-craniofacial health issue.

[0028] FIG. 1C is a flowchart of an embodim
abundances of a plurality of taxa from a sample and outputting the estimates to a database.

[0029] FIG. 1D is a flowchart of an embodiment of a method for generating features derived from composition and/or functional components of a biological sample or an aggregate of biological samples.

[0030] FIG. 1E is a flowchart of an embodiment of a method for characterizing a microbiome-associated condition and identifying therapeutic measures.

[0031] FIG. 1F is a flow chart of an embodiment of a method for generating microbiome-derived diagnostics.

[0032] FIG. 2 depicts an embodiment of a method and system for generating microbiome-derived diagnostics and therapeutics.

[0033] FIG. 3 depicts variations of a portion of an embodiment of a method for generating microbiome-derived diagnostics and therapeutics.

[0034] FIG. 4 depicts a variation of a process for generation of a model in an embodiment of a method and system for generating microbiome-derived diagnostics and therapeutics.

[0035] FIG. 5 depicts variations of mechanisms by which therapies (*e.g.*, probiotic-based or prebiotic-based therapies) operate in an embodiment of a method for characterizing a health condition.

[0036] FIG. 6 depicts examples of therapy-related notification provision in an example of a method for generating microbiome-derived diagnostics and therapeutics.

[0037] FIG. 7 shows a plot illustrating the control distribution and the disease distribution for insomnia where the sequence group is Moryella for the Genus taxonomic group according to embodiments of the present invention.

[0038] FIG. 8 shows a plot illustrating the control distribution and the disease distribution for insomnia where the sequence group is Selenocompound metabolism for the function taxonomic group according to embodiments of the present invention.

[0039] FIG. 9 shows a plot illustrating the control distribution and the disease distribution for light sleep where the sequence group is Lactobacillaceae for the Family taxonomic group according to embodiments of the present invention.

[0040] FIG. 10 shows a plot illustrating the c
for light sleep where the sequence group is Translation for the function taxonomic group according to embodiments of the present invention.

[0041] FIG. 11 shows a plot illustrating the control distribution and the disease distribution for headache where the sequence group is Marvinbryantia for the Genus taxonomic group according to embodiments of the present invention.

[0042] FIG. 12 shows a plot illustrating the control distribution and the disease distribution for headache where the sequence group is Selenocompound metabolism for the function taxonomic group according to embodiments of the present invention.

[0043] FIG. 13 shows a plot illustrating the control distribution and the disease distribution for sinusitis where the sequence group is Clostridiales for the Genus taxonomic group according to embodiments of the present invention.

[0044] FIG. 14 shows a plot illustrating the control distribution and the disease distribution for poor concentration where the sequence group is Moryella for the Genus taxonomic group according to embodiments of the present invention.

[0045] FIG. 15 shows a plot illustrating the control distribution and the disease distribution for poor concentration where the sequence group is Propanoate metabolism for the function taxonomic group according to embodiments of the present invention.

DETAILED DESCRIPTION

[0046] The inventors have discovered that characterization of the microbiome of individuals is useful for detecting a microbiome indicative of insomnia, light sleep, headache, sinusitis, or poor concentration. For example, an individual having symptoms indicative of insomnia, light sleep, headache, sinusitis, or poor concentration, or in whom insomnia, light sleep, headache, sinusitis, or poor concentration is suspected, can be tested to confirm or provide further evidence to support or refute a diagnosis of the subject. As another example, an individual can be assayed to determine whether they have a microbiome that is likely to increase the risk of insomnia, light sleep, headache, sinusitis, or poor concentration. As another example, an individual having, or suspected of having, or having a history of, insomnia, light sleep, headache, sinusitis, or poor concentration can be assayed to determine whether the microbiome is likely to be a causative agent, or contribute to the frequency or severity of the insomnia, light sleep, headache, sinusitis, or poor concentration.

[0047] An individual having symptoms of ins
poor concentration, or has insomnia, light sleep, headache, sinusitis, or poor concentration, or has a microbiome (e.g., a gut or stool microbiome) that causes or contributes to the frequency or severity of insomnia, light sleep, headache, sinusitis, or poor concentration is referred to herein as having a "cerebro-craniofacial health issue." Similarly, an individual having symptoms of insomnia, or has insomnia, or has a microbiome (e.g., a gut or stool microbiome) that causes or contributes to the frequency or severity of insomnia is referred to herein as having a "insomnia issue." Likewise, an individual having symptoms of light sleep, or has light sleep, or has a microbiome (e.g., a gut or stool microbiome) that causes or contributes to the frequency or severity of light sleep is referred to herein as having a "light sleep issue." An individual having symptoms of a headache, or has a headache, or has a microbiome (e.g., a gut or stool microbiome) that causes or contributes to the frequency or severity of a headache is referred to herein as having a "headache issue." An individual having symptoms of sinusitis, or has sinusitis, or has a microbiome (e.g., a gut or stool microbiome) that causes or contributes to the frequency or severity of sinusitis is referred to herein as having a "sinusitis issue." An individual having symptoms of poor concentration, or has poor concentration, or has a microbiome (e.g., a gut or stool microbiome) that causes or contributes to the frequency or severity of light sleep is referred to herein as having a "poor concentration issue."

[0048] Such characterizations are also useful for screening individuals for and/or determining a course of treatment for an individual that has a cerebro-craniofacial health issue. For example, by deep sequencing bacterial DNAs from control (healthy, or at least not having a cerebro-craniofacial health issue) individuals and diseased individuals (having a cerebro-craniofacial health issue), the inventors have discovered that the amount of certain bacteria and/or bacterial sequences corresponding to certain genetic pathways can be used to predict the presence or absence of a cerebro-craniofacial health issue. The bacteria and genetic pathways in some cases are present in a certain abundance in individuals having a cerebro-craniofacial health issue, or having a specific cerebro-craniofacial health issue, as discussed in more detail below whereas the bacteria and genetic pathways are at a statistically different abundance in control individuals that do not have a cerebro-craniofacial health issue, or do not have a specific cerebro-craniofacial health issue.

I. BACTERIA GROUPS

[0049] Details of these associations for the specific cerebro-craniofacial health issue of insomnia can be found in TABLE A for bacteria groups (also called taxonomic groups) and

or genetic pathways (also called functional grou

functional groups are referred to as features, or as sequence groups in the context of determining an amount of sequence reads corresponding to a particular group (feature). Scoring of a particular bacteria or genetic pathway can be determined according to a comparison of an abundance value to one or more reference (calibration) abundance values for known samples, e.g., where a detected abundance value less than a certain value is associated with a insomnia issue and above the certain value is scored as associated with a lack of a insomnia issue, depending on the particular criterion. Similarly, depending on the particular criterion, a detected abundance value greater than a certain value can be associated with a insomnia issue and below the certain value can be scored as associated with a lack of a insomnia issue or a microbiome that is not indicative of a insomnia issue. The scoring for various bacteria or genetic pathways can be combined to provide a classification for a subject.

TABLE A

Insomnia (901) vs control (4865)	**p-value**	**# disease subjects detected**	**# control subjects detected**	**Mean % abundance for disease**	**Mean % abundance for control**
Taxa (microbiome composition):					
Species:					
Parabacteroides distasonis_823	2.49E-11	535	3456	1.280	1.140
Flavonifractor plautii_292800	6.44E-09	518	2401	0.369	0.278
Collinsella aerofaciens_74426	3.84E-06	446	2827	0.575	0.591
Blautia sp. YHC-4_1157314	8.23E-06	268	1024	1.525	0.840
Genus:					
Moryella_437755	9.34E-09	373	1577	0.502	0.375
Roseburia_841	4.53E-07	888	4785	6.844	7.742
Faecalibacterium_216851	8.10E-07	840	4681	11.076	12.470
Bacteroides_816	6.84E-06	888	4801	26.326	23.923
Family:					
Oscillospiraceae_216572	1.99E-10	704	3646	0.415	0.285
Ruminococcaceae_541000	1.02E-07	885	4782	14.986	17.134
Lactobacillaceae_33958	2.03E-07	584	3114	0.850	0.591
Bacteroidaceae_815	5.01E-06	888	4801	26.395	23.972
Porphyromonadaceae_171551	3.24E-05	854	4573	3.227	2.888

Order:					
Clostridiales_186802	1.51E-05	901	4865	53.201	55.076
Class:					
Clostridia_186801	1.24E-05	901	4865	53.268	55.143
Bacteroidia_200643	3.95E-05	896	4851	34.385	32.252
Phylum:					
Bacteroidetes_976	3.23E-05	897	4851	34.586	32.541
Function (microbiome functionality):					
KEGG L2:					
Translation	1.24E-09	897	4853	5.651	5.743
Cell Growth and Death	2.22E-07	899	4855	0.517	0.524
Cellular Processes and Signaling	2.66E-07	897	4852	4.239	4.198
Carbohydrate Metabolism	4.78E-07	897	4852	11.113	11.005
Metabolism of Cofactors and Vitamins	9.26E-07	897	4853	4.416	4.453
Metabolism of Terpenoids and Polyketides	1.35E-06	897	4852	1.654	1.671
Nucleotide Metabolism	2.27E-06	899	4855	4.015	4.060
Genetic Information Processing	5.28E-06	897	4852	2.594	2.612
Signal Transduction	6.45E-06	897	4853	1.442	1.418
Biosynthesis of Other Secondary Metabolites	8.12E-06	899	4855	0.983	0.968
Energy Metabolism	8.99E-06	896	4852	6.120	6.163
Replication and Repair	1.64E-05	897	4852	8.883	8.963
Metabolism	2.23E-04	898	4854	2.469	2.451
Transcription	2.66E-04	897	4853	2.924	2.902
KEGG L3:					
Selenocompound metabolism	1.99E-13	899	4851	0.368	0.373
Amino acid related enzymes	2.32E-13	899	4851	1.497	1.516
Ribosome biogenesis in eukaryotes	5.02E-12	899	4851	0.047	0.048
D-Alanine metabolism	2.81E-10	899	4851	0.100	0.103
Ribosome Biogenesis	6.06E-10	899	4851	1.401	1.419

Phenylpropanoid biosynthesis	7.15E-10	89			
Starch and sucrose metabolism	1.55E-09	899	4851	1.138	1.113
Pentose and glucuronate interconversions	1.96E-09	899	4851	0.589	0.569
Galactose metabolism	2.30E-09	899	4851	0.864	0.841
Inorganic ion transport and metabolism	3.83E-09	899	4851	0.189	0.181
Ribosome	4.34E-09	899	4851	2.345	2.391
Sphingolipid metabolism	6.39E-09	899	4851	0.277	0.262
Cyanoamino acid metabolism	1.20E-08	899	4851	0.314	0.304
Translation factors	1.83E-08	899	4851	0.533	0.541
DNA repair and recombination proteins	3.00E-08	899	4851	2.825	2.855
Translation proteins	3.04E-08	899	4851	0.890	0.900
Propanoate metabolism	1.04E-07	899	4851	0.503	0.511
Aminoacyl-tRNA biosynthesis	1.30E-07	899	4851	1.175	1.195
Carbohydrate metabolism	1.38E-07	899	4851	0.199	0.194
Cell cycle - Caulobacter	1.41E-07	899	4851	0.512	0.520
Others	2.48E-07	899	4851	0.917	0.903
Fructose and mannose metabolism	2.62E-07	899	4851	1.073	1.053
Peptidoglycan biosynthesis	4.41E-07	899	4851	0.831	0.843
Glycosphingolipid biosynthesis - globo series	5.33E-07	899	4851	0.133	0.126
Phosphonate and phosphinate metabolism	5.95E-07	899	4851	0.056	0.054
Terpenoid backbone biosynthesis	7.89E-07	899	4851	0.578	0.587
RNA polymerase	1.43E-06	899	4851	0.160	0.163
Homologous recombination	1.89E-06	899	4851	0.934	0.945
Pentose phosphate pathway	1.94E-06	899	4851	0.931	0.921
Amino sugar and nucleotide sugar metabolism	1.98E-06	899	4851	1.487	1.469
Pyrimidine metabolism	2.90E-06	899	4851	1.825	1.849
Carbon fixation pathways in prokaryotes	3.55E-06	899	4851	1.012	1.026
Nicotinate and nicotinamide metabolism	5.51E-06	899	4851	0.431	0.437
Lipid biosynthesis proteins	7.38E-06	899	4851	0.587	0.593
Purine metabolism	8.98E-06	899	4851	2.190	2.212
Two-component system	1.03E-05	899	4851	1.307	1.284
Other glycan degradation	1.44E-05	899	4851	0.373	0.355
Oxidative phosphorylation	1.87E-05	899	4851	1.194	1.210

Nucleotide excision repair	2.91E-05	89			
Amino acid metabolism	3.59E-05	899	4851	0.204	0.200
Riboflavin metabolism	4.22E-05	899	4851	0.233	0.238
Drug metabolism - other enzymes	4.75E-05	899	4851	0.323	0.328
Fatty acid biosynthesis	4.84E-05	899	4851	0.498	0.504
One carbon pool by folate	7.14E-05	899	4851	0.630	0.640
Glycerolipid metabolism	8.34E-05	899	4851	0.401	0.393
General function prediction only	1.18E-04	899	4851	3.645	3.659
Benzoate degradation	1.19E-04	899	4851	0.200	0.205
Bacterial toxins	1.52E-04	899	4851	0.121	0.119
Folate biosynthesis	2.02E-04	899	4851	0.393	0.399
Other transporters	2.22E-04	899	4851	0.272	0.269
Butirosin and neomycin biosynthesis	2.24E-04	899	4851	0.075	0.074
Pantothenate and CoA biosynthesis	2.50E-04	899	4851	0.660	0.665
Fatty acid metabolism	2.50E-04	899	4851	0.218	0.223
Protein kinases	2.63E-04	899	4851	0.296	0.291
Function unknown	2.85E-04	899	4851	1.188	1.176
Chloroalkane and chloroalkene degradation	3.49E-04	899	4851	0.187	0.184

[0050] Details of these associations for the specific cerebro-craniofacial health issue of light sleep can be found in TABLE B for bacteria groups (also called taxonomic groups) and or genetic pathways (also called functional groups). Scoring of a particular bacteria or genetic pathway can be determined according to a comparison of an abundance value to one or more reference (calibration) abundance values for known samples, e.g., where a detected abundance value less than a certain value is associated with a light sleep issue and above the certain value is scored as associated with a lack of a light sleep issue, depending on the particular criterion. Similarly, depending on the particular criterion, a detected abundance value greater than a certain value can be associated with a light sleep issue and below the certain value can be scored as associated with a lack of a light sleep issue or a microbiome that is not indicative of a light sleep issue. The scoring for various bacteria or genetic pathways can be combined to provide a classification for a subject.

TABLE B

Lightsleep (627) vs control (4471)	**p-value**	**# disease subjects detected**	**# control subjects detected**	**Mean % abundance for disease**	**Mean % abundance for control**

Taxa (microbiome composition):					
Family:					
Lactobacillaceae_33958	4.72E-08	451	2928	0.961	0.592
Function (microbiome functionality):					
Translation	1.13E-07	647	4544	5.646	5.741
Cellular Processes and Signaling	2.79E-07	647	4543	4.247	4.199
Metabolism	1.28E-05	647	4545	2.479	2.452
Replication and Repair	8.28E-05	647	4543	8.877	8.961
Cell Growth and Death	2.00E-04	647	4546	0.518	0.524

[0051] Details of these associations for the specific cerebro-craniofacial health issue of a headache can be found in TABLE C for bacteria groups (also called taxonomic groups) and or genetic pathways (also called functional groups). Collectively, the taxonomic groups and functional groups are referred to as features, or as sequence groups in the context of determining an amount of sequence reads corresponding to a particular group (feature). Scoring of a particular bacteria or genetic pathway can be determined according to a comparison of an abundance value to one or more reference (calibration) abundance values for known samples, e.g., where a detected abundance value less than a certain value is associated with a headache issue and above the certain value is scored as associated with a lack of a headache issue, depending on the particular criterion. Similarly, depending on the particular criterion, a detected abundance value greater than a certain value can be associated with a headache issue and below the certain value can be scored as associated with a lack of a headache issue or a microbiome that is not indicative of a headache issue. The scoring for various bacteria or genetic pathways can be combined to provide a classification for a subject.

TABLE C

	p-value	# disease subjects detected	# control subjects detected	Mean % abundance for disease	Mean % abundance for control
Headaches & Migraine (795) vs Control (4349)					
Taxa (microbiome composition):					
Species:					
Parabacteroides	8.16E-11	456	3070	1.247	1.188

distasonis_823					
Flavonifractor plautii_292800	7.45E-09	456	2175	0.420	0.284
Bacteroides acidifaciens_85831	1.92E-06	516	2690	1.364	0.984
Genus:					
Marvinbryantia_248744	1.31E-06	379	2512	0.255	0.276
Family:					
Oscillospiraceae_216572	2.64E-09	628	3265	0.420	0.291
Lactobacillaceae_33958	1.70E-05	529	2724	0.812	0.568
Function (microbiome functionality):					
KEGG L2					
Energy Metabolism	3.79E-15	795	4348	6.086	6.169
Metabolism of Terpenoids and Polyketides	4.20E-13	795	4348	1.645	1.671
Cell Motility	1.60E-10	795	4349	1.741	1.612
Cell Growth and Death	2.34E-10	795	4349	0.513	0.524
Signal Transduction	8.73E-10	795	4348	1.457	1.422
Metabolism of Cofactors and Vitamins	2.19E-09	795	4348	4.400	4.452
Nucleotide Metabolism	3.02E-09	795	4349	3.998	4.055
Transcription	1.27E-08	795	4348	2.936	2.899
Folding, Sorting and Degradation	5.26E-07	795	4348	2.470	2.496
Environmental Adaptation	8.40E-07	795	4348	0.163	0.160
Membrane Transport	1.15E-06	795	4349	11.883	11.640
Replication and Repair	1.72E-06	795	4348	8.860	8.951
Digestive System	2.64E-06	795	4348	0.040	0.045
Translation	1.13E-05	795	4348	5.659	5.731
Metabolism of Other Amino Acids	6.70E-05	795	4348	1.455	1.468
KEGG L3:					
Selenocompound metabolism	1.71E-17	795	4346	0.367	0.373
Oxidative phosphorylation	4.93E-17	795	4346	1.178	1.213
Carbon fixation pathways in prokaryotes	4.40E-16	795	4346	1.001	1.027
Folate biosynthesis	9.62E-15	795	4346	0.385	0.400

Amino acid related enzymes	1.49E-12	795			
Starch and sucrose metabolism	9.77E-12	795	4346	1.144	1.113
General function prediction only	1.11E-11	795	4346	3.634	3.658
RNA transport	2.30E-11	795	4346	0.154	0.149
Protein kinases	2.31E-11	795	4346	0.299	0.291
Phenylpropanoid biosynthesis	3.63E-11	795	4346	0.191	0.180
Cell cycle - Caulobacter	3.65E-11	795	4346	0.509	0.519
Two-component system	9.68E-11	795	4346	1.325	1.287
Glycerolipid metabolism	1.89E-10	795	4346	0.404	0.393
Insulin signaling pathway	3.39E-10	795	4346	0.095	0.092
Translation factors	6.08E-10	795	4346	0.531	0.541
One carbon pool by folate	6.53E-10	795	4346	0.625	0.639
Drug metabolism - other enzymes	7.41E-10	795	4346	0.320	0.327
Purine metabolism	3.05E-09	795	4346	2.181	2.210
DNA repair and recombination proteins	3.40E-09	795	4346	2.817	2.852
Glycerophospholipid metabolism	6.32E-09	795	4346	0.576	0.569
Chaperones and folding catalysts	1.40E-08	795	4346	1.027	1.044
Pyrimidine metabolism	1.59E-08	795	4346	1.818	1.847
Riboflavin metabolism	1.77E-08	795	4346	0.232	0.239
Citrate cycle (TCA cycle)	1.95E-08	795	4346	0.581	0.602
Glycine, serine and threonine metabolism	3.18E-08	795	4346	0.827	0.835
Ribosome	3.28E-08	795	4346	2.340	2.386
Bacterial chemotaxis	3.62E-08	795	4346	0.383	0.353
Pentose phosphate pathway	3.70E-08	795	4346	0.932	0.921
Carbohydrate metabolism	4.66E-08	795	4346	0.200	0.194
Transporters	4.99E-08	795	4346	6.656	6.488
Cyanoamino acid metabolism	1.59E-07	795	4346	0.314	0.305
Glutathione metabolism	2.12E-07	795	4346	0.171	0.178
Biosynthesis of ansamycins	2.35E-07	795	4346	0.122	0.119
Transcription factors	2.41E-07	795	4346	1.708	1.667
Prenyltransferases	4.25E-07	795	4346	0.308	0.316
Propanoate metabolism	4.41E-07	795	4346	0.503	0.512
Butirosin and neomycin biosynthesis	6.70E-07	795	4346	0.076	0.074
Ribosome biogenesis in	6.75E-07	795	4346	0.047	0.048

eukaryotes					
Chloroalkane and chloroalkene degradation	7.98E-07	795	4346	0.188	0.184
Homologous recombination	1.05E-06	795	4346	0.932	0.944
Pentose and glucuronate interconversions	1.11E-06	795	4346	0.586	0.571
Nucleotide excision repair	1.17E-06	795	4346	0.391	0.397
Energy metabolism	2.57E-06	795	4346	0.893	0.905
Terpenoid backbone biosynthesis	2.60E-06	795	4346	0.576	0.585
beta-Alanine metabolism	2.65E-06	795	4346	0.191	0.197
Plant-pathogen interaction	3.00E-06	795	4346	0.163	0.160
D-Glutamine and D-glutamate metabolism	3.43E-06	795	4346	0.147	0.149
Arginine and proline metabolism	3.62E-06	795	4346	1.285	1.270
Peptidases	3.85E-06	795	4346	1.882	1.899
DNA replication proteins	4.45E-06	795	4346	1.233	1.248
Geraniol degradation	4.95E-06	795	4346	0.032	0.036
Cellular antigens	7.15E-06	795	4346	0.037	0.042
DNA replication	1.17E-05	795	4346	0.646	0.654
Galactose metabolism	1.45E-05	795	4346	0.861	0.844
Nitrogen metabolism	1.61E-05	795	4346	0.699	0.708
Zeatin biosynthesis	1.98E-05	795	4346	0.054	0.056
ABC transporters	2.10E-05	795	4346	3.198	3.136
Others	2.64E-05	795	4346	0.917	0.905
RNA polymerase	3.81E-05	795	4346	0.160	0.163
Secretion system	6.45E-05	795	4346	1.039	1.021
Peptidoglycan biosynthesis	6.73E-05	795	4346	0.830	0.841
Translation proteins	6.79E-05	795	4346	0.890	0.898
Protein export	1.11E-04	795	4346	0.592	0.598
D-Alanine metabolism	1.24E-04	795	4346	0.101	0.103
Taurine and hypotaurine metabolism	1.35E-04	795	4346	0.106	0.108
Phosphonate and phosphinate metabolism	1.42E-04	795	4346	0.056	0.054
Mismatch repair	2.15E-04	795	4346	0.826	0.833
Vitamin B6 metabolism	2.51E-04	795	4346	0.200	0.203
RNA degradation	2.87E-04	795	4346	0.463	0.468
Lipid metabolism	3.05E-04	795	4346	0.148	0.146
Protein folding and associated processing	3.32E-04	795	4346	0.605	0.613

Carbon fixation in photosynthetic organisms	3.42E-04	795	4346	0.682	0.688

[0052] Details of these associations for the specific cerebro-craniofacial health issue of sinusitis can be found in TABLE D for bacteria groups (also called taxonomic groups) and or genetic pathways (also called functional groups). Collectively, the taxonomic groups and functional groups are referred to as features, or as sequence groups in the context of determining an amount of sequence reads corresponding to a particular group (feature). Scoring of a particular bacteria or genetic pathway can be determined according to a comparison of an abundance value to one or more reference (calibration) abundance values for known samples, e.g., where a detected abundance value less than a certain value is associated with a sinusitis issue and above the certain value is scored as associated with a lack of a sinusitis issue, depending on the particular criterion. Similarly, depending on the particular criterion, a detected abundance value greater than a certain value can be associated with a sinusitis issue and below the certain value can be scored as associated with a lack of a sinusitis issue or a microbiome that is not indicative of a sinusitis issue. The scoring for various bacteria or genetic pathways can be combined to provide a classification for a subject.

TABLE D

Sinusitis (218) vs control (1049)	**p-value**	**# disease subjects detected**	**# control subjects detected**	**Mean % abundance for disease**	**Mean % abundance for control**
Taxa (microbiome composition):					
Order:					
Clostridiales_186802	5.22E-05	172	896	5.832	7.606
Class:					
Clostridia_186801	5.07E-05	172	897	5.839	7.616

[0053] Details of these associations for the specific cerebro-craniofacial health issue of poor concentration can be found in TABLE E for bacteria groups (also called taxonomic groups) and or genetic pathways (also called functional groups). Collectively, the taxonomic groups and functional groups are referred to as features, or as sequence groups in the context of determining an amount of sequence reads corresponding to a particular group (feature). Scoring of a particular bacteria or genetic pathway can be determined according to a comparison of an abundance value to one or more reference (calibration) abundance values

for known samples, e.g., where a detected abun

associated with a poor concentration issue and above the certain value is scored as associated with a lack of a poor concentration issue, depending on the particular criterion. Similarly, depending on the particular criterion, a detected abundance value greater than a certain value can be associated with a poor concentration issue and below the certain value can be scored as associated with a lack of a poor concentration issue or a microbiome that is not indicative of a poor concentration issue. The scoring for various bacteria or genetic pathways can be combined to provide a classification for a subject.

TABLE E

Poor Concentration (1396) vs control (6276)	**p-value**	**# disease subjects detected**	**# control subjects detected**	**Mean % abundance for disease**	**Mean % abundance for control**
Taxa (microbiome composition):					
Species:					
Parabacteroides distasonis_823	5.31E-12	831	4324	1.279	1.181
Flavonifractor plautii_292800	1.56E-10	795	3018	0.357	0.283
Blautia sp. YHC-4_1157314	1.49E-07	419	1349	1.250	0.930
Odoribacter splanchnicus_28118	3.44E-07	742	3102	0.300	0.255
Blautia luti_89014	2.92E-06	1021	4664	1.348	1.548
Genus:					
Moryella_437755	5.06E-07	563	2027	0.420	0.404
Sarcina_1266	2.44E-06	1189	5391	1.767	2.009
Bacteroides_816	4.03E-06	1372	6204	26.304	24.115
Bilophila_35832	1.82E-05	823	3496	0.320	0.249
Dorea_189330	1.86E-05	1304	5856	1.304	1.376
Odoribacter_283168	1.99E-05	822	3525	0.341	0.311
Terrisporobacter_1505652	3.04E-05	656	3303	0.285	0.262
Family:					
Oscillospiraceae_216572	2.63E-09	1086	4606	0.387	0.293
Sutterellaceae_995019	1.24E-07	1124	4967	1.462	1.254
Flavobacteriaceae_49546	3.47E-07	662	3394	0.464	0.468
Bacteroidaceae_815	4.51E-06	1372	6207	26.331	24.150
Order:					

Flavobacteriales_200644	2.71E-07	662			
Burkholderiales_80840	7.76E-07	1138	5055	1.463	1.258
Bacteroidales_171549	3.65E-05	1390	6276	33.992	32.296
Class:					
Flavobacteriia_117743	2.40E-07	662	3400	0.464	0.468
Betaproteobacteria_28216	1.47E-06	1146	5104	1.463	1.273
Bacteroidia_200643	8.46E-05	1390	6276	34.036	32.387
Function (microbiome functionality):					
KEGG L2					
Signal Transduction	3.18E-08	1395	6268	1.450	1.423
Energy Metabolism	3.70E-07	1394	6265	6.119	6.161
Metabolism of Terpenoids and Polyketides	1.03E-06	1395	6266	1.654	1.669
Translation	1.81E-06	1395	6268	5.675	5.736
Nucleotide Metabolism	1.30E-05	1396	6272	4.019	4.053
Replication and Repair	2.82E-05	1395	6266	8.883	8.949
Metabolism of Cofactors and Vitamins	3.16E-05	1395	6268	4.422	4.449
Genetic Information Processing	1.26E-04	1395	6266	2.596	2.609
KEGG L3:					
Propanoate metabolism	1.63E-08	1396	6269	0.504	0.511
Selenocompound metabolism	1.89E-08	1396	6269	0.369	0.372
Two-component system	8.05E-08	1396	6269	1.314	1.289
Translation factors	1.02E-07	1396	6269	0.534	0.541
Carbon fixation pathways in prokaryotes	1.64E-07	1396	6269	1.011	1.025
Folate biosynthesis	1.83E-07	1396	6269	0.391	0.398
DNA repair and recombination proteins	2.45E-07	1396	6269	2.826	2.851
Ribosome Biogenesis	1.06E-06	1396	6269	1.406	1.418
Oxidative phosphorylation	3.14E-06	1396	6269	1.195	1.210
Nucleotide excision repair	7.50E-06	1396	6269	0.392	0.397
Purine metabolism	9.38E-06	1396	6269	2.190	2.208
Pyrimidine metabolism	1.74E-05	1396	6269	1.828	1.846
Galactose metabolism	2.91E-05	1396	6269	0.857	0.843

Aminoacyl-tRNA biosynthesis	4.17E-05	1396			
Terpenoid backbone biosynthesis	4.55E-05	1396	6269	0.580	0.586
Peptidoglycan biosynthesis	5.99E-05	1396	6269	0.833	0.841
Drug metabolism - other enzymes	6.08E-05	1396	6269	0.323	0.327
Pantothenate and CoA biosynthesis	7.54E-05	1396	6269	0.661	0.666
Starch and sucrose metabolism	8.29E-05	1396	6269	1.131	1.116
One carbon pool by folate	9.52E-05	1396	6269	0.631	0.638
Others	9.63E-05	1396	6269	0.914	0.905
Amino acid metabolism	2.03E-04	1396	6269	0.203	0.200
Mismatch repair	2.62E-04	1396	6269	0.827	0.833

[0054] The comparison of an abundance value to one or more reference abundance values can involve a comparison to a cutoff value determined from the one or more reference values. Such cutoff value(s) can be part of a decision tree or a clustering technique (where a cutoff value is used to determine which cluster the abundance value(s) belong) that are determined using the reference abundance values. The comparison can include intermediate determination of other values, e.g., probability values. The comparison can also include a comparison of an abundance value to a probability distribution of the reference abundance values, and thus a comparison to probability values.

[0055] The inventors have identified the specific bacteria taxa and genetic pathways listed in TABLE A by deep sequencing of bacterial DNA associated with samples from test individuals having a insomnia issue and control individuals that do not have a insomnia issue and determining those criteria that readily distinguish test individuals from control individuals. Similarly, the inventors have identified the specific bacteria taxa and genetic pathways listed in TABLE B by deep sequencing of bacterial DNA associated with samples from test individuals having a light sleep issue and control individuals that do not have a light sleep issue and determining those criteria that readily distinguish test individuals from control individuals. Similarly, the inventors have identified the specific bacteria taxa and genetic pathways listed in TABLE C by deep sequencing of bacterial DNA associated with samples from test individuals having a headache issue and control individuals that do not have a headache issue and determining those criteria that readily distinguish test individuals from control individuals. Similarly, the inventors have identified the specific bacteria taxa and genetic pathways listed in TABLE D by deep sequencing of bacterial DNA associated with

samples from test individuals having a sinusitis

a sinusitis issue and determining those criteria that readily distinguish test individuals from control individuals. Similarly, the inventors have identified the specific bacteria taxa and genetic pathways listed in TABLE E by deep sequencing of bacterial DNA associated with samples from test individuals having a poor concentration issue and control individuals that do not have a poor concentration issue and determining those criteria that readily distinguish test individuals from control individuals.

[0056] Deep sequencing allows for determination of a sufficient number of copies of DNA sequences to determine relative amount of corresponding bacteria or genetic pathways in the sample. Having identified the criteria in TABLEs A, B, C, D, and E, one can now detect an individual that has a cerebro-craniofacial health issue by detecting one or more (e.g., 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, or more) of the options in TABLEs A, B, C, D, or E by any quantitative detection method. In some cases, one can now detect an individual that has a cerebro-craniofacial health issue by detecting from about 1 to about 20, from about 2 to about 15, from about 3 to about 10, from about 1 to about 10, from about 1 to about 15, from about 1 to about 5, or from about 5 to about 30 of the options in TABLEs A, B, C, D, or E by any quantitative detection method. For example, while deep sequencing can be used to detect the presence, absence or amount of one or more option in TABLEs A, B, C, D, or E , one can also use other detection methods, including but not limited to protein detection methods. For example, without intending to limit the scope of the invention, one could use protein-based diagnostics such as immunoassays to detect bacterial taxons by detecting taxon-specific protein markers.

[0057] As a result of these discoveries (e.g., as set forth in TABLEs A, B, C, D, and E), one can design treatments to ameliorate one or more symptoms of a cerebro-craniofacial health issue and/or alleviate or reduce the frequency and/or severity of insomnia, light sleep, headache, sinusitis, or poor concentration. As a non-limiting example, one can determine whether an individual having a insomnia issue lacks, or has a reduced abundance of, one or more type of bacteria as listed in TABLE A and if so, that one or more type of bacteria can be administered to the individual. Additionally, or alternatively, one can determine whether an individual having a insomnia issue lacks, or has a reduced abundance of, one or more type of bacteria as listed in TABLE A and if so, a prebiotic that promotes the growth of that one or more type of bacteria can be administered to the individual. Additionally, or alternatively, one can determine whether an individual having a insomnia issue has an increased abundance of one or more type of bacteria as listed in TABLE A and if so, a targeted therapy that

reduces the abundance of such bacteria (*e.g.*, ba
therapy) can be administered to the individual.

[0058] As another non-limiting example, one can determine whether an individual having a light sleep issue lacks, or has a reduced abundance of, one or more type of bacteria as listed in TABLE B and if so, that one or more type of bacteria can be administered to the individual. Additionally, or alternatively, one can determine whether an individual having a light sleep issue lacks, or has a reduced abundance of, one or more type of bacteria as listed in TABLE B and if so, a pre-biotic that promotes the growth of that one or more type of bacteria can be administered to the individual. Additionally, or alternatively, one can determine whether an individual having a light sleep issue has an increased abundance of one or more type of bacteria as listed in TABLE B and if so, a targeted therapy that reduces the abundance of such bacteria (*e.g.*, bacteriophage therapy or selective antibiotic therapy) can be administered to the individual.

[0059] As another non-limiting example, one can determine whether an individual having a headache issue lacks, or has a reduced abundance of, one or more type of bacteria as listed in TABLE C and if so, that one or more type of bacteria can be administered to the individual. Additionally, or alternatively, one can determine whether an individual having a headache issue lacks, or has a reduced abundance of, one or more type of bacteria as listed in TABLE C and if so, a pre-biotic that promotes the growth of that one or more type of bacteria can be administered to the individual. Additionally, or alternatively, one can determine whether an individual having a headache issue has an increased abundance of one or more type of bacteria as listed in TABLE C and if so, a targeted therapy that reduces the abundance of such bacteria (*e.g.*, bacteriophage therapy or selective antibiotic therapy) can be administered to the individual.

[0060] As another non-limiting example, one can determine whether an individual having a sinusitis issue lacks, or has a reduced abundance of, one or more type of bacteria as listed in TABLE D and if so, that one or more type of bacteria can be administered to the individual. Additionally, or alternatively, one can determine whether an individual having a sinusitis issue lacks, or has a reduced abundance of, one or more type of bacteria as listed in TABLE D and if so, a pre-biotic that promotes the growth of that one or more type of bacteria can be administered to the individual. Additionally, or alternatively, one can determine whether an individual having a sinusitis issue has an increased abundance of one or more type of bacteria as listed in TABLE D and if so, a targeted therapy that reduces the abundance of such

bacteria (*e.g.*, bacteriophage therapy or selectiv
the individual.

[0061] As another non-limiting example, one can determine whether an individual having a poor concentration issue lacks, or has a reduced abundance of, one or more type of bacteria as listed in TABLE E and if so, that one or more type of bacteria can be administered to the individual. Additionally, or alternatively, one can determine whether an individual having a poor concentration issue lacks, or has a reduced abundance of, one or more type of bacteria as listed in TABLE E and if so, a pre-biotic that promotes the growth of that one or more type of bacteria can be administered to the individual. Additionally, or alternatively, one can determine whether an individual having a poor concentration issue has an increased abundance of one or more type of bacteria as listed in TABLE E and if so, a targeted therapy that reduces the abundance of such bacteria (*e.g.*, bacteriophage therapy or selective antibiotic therapy) can be administered to the individual.

II. DETERMINING LIKELIHOOD OF a cerebro-craniofacial health ISSUE

[0062] In some embodiments, a method of determining whether, or the likelihood whether, an individual has a cerebro-craniofacial health issue is provided. As described herein, an individual having a cerebro-craniofacial health issue can exhibit an increase in one or more taxonomic groups in the microbiome, a decrease in one or more taxonomic groups in the microbiome, an increase in one or more functional groups in the microbiome, a decrease in one or more functional groups in the microbiome, or a combination thereof (e.g., relative to a control/healthy individual or population of control or healthy individuals).

[0063] The method can include one or more of the following steps:

obtaining a sample from the individual;

purifying nucleic acids (e.g., DNA) from the sample;

deep sequencing nucleic acids from the sample so as to determine the amount of one or more (e.g., 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, or more, e.g., 1-20, 2-15, 3-10, 1-10, 1-15, 1-5, or 5-30) of the features listed in TABLEs A, B, C, D, or E; and

comparing the resulting amount of each feature to one or more reference amounts of the one or more of the features listed in TABLEs A, B, C, D, or E as occurs in an average individual having a cerebro-craniofacial health issue or an individual not having a cerebro-craniofacial health issue or both. The compilation of features can sometimes be referred to as a "disease signature" for a specific disease (*i.e.*, a cerebro-craniofacial health issue such as insomnia,

light sleep, headache, sinusitis, or poor concent

specific condition. The disease signature can act as a characterization model, and may include probability distributions for control population (no cerebro-craniofacial health issue) or disease populations having the disease (a cerebro-craniofacial health issue) or both. The disease signature can include one or more of the features (e.g., bacterial taxa or genetic pathways) in TABLEs A, B, C, D, or E and can optionally include criteria determined from abundance values of the control and/or disease populations. Example criteria can include cutoff or probability values for amounts of those features associated with average control individuals (no cerebro-craniofacial health issue) or individuals having the disease (a cerebro-craniofacial health issue).

[0064] The likelihood of an individual having a microbiome indicative of a cerebro-craniofacial health issue (e.g., as listed in TABLEs A, B, C, D, or E) refers to the chance (degree of confidence) that the results from the individual's sample can be correlated with a cerebro-craniofacial health issue. Alternatively, one can simply screen for a cerebro-craniofacial health issue, i.e., one can generate a yes or no indication for the presence or absence of a microbiome indicative of insomnia, light sleep, headache, sinusitis, or poor concentration. In some embodiments, the individual will not yet have been diagnosed with insomnia, light sleep, headache, sinusitis, or poor concentration or a insomnia issue, light sleep issue, headache issue, sinusitis issue, or poor concentration issue. In other examples, the individual can have been initially diagnosed by other methods and the methods described herein can be used to provide better (or worse) confidence of the initial diagnosis.

[0065] Any type of sample containing bacteria can be used from the individual. Exemplary sample types include, for example, a fecal sample, blood sample, saliva sample, throat swab, cheek swab, gum swab, urine or other bodily fluid from the individual. Nucleic acids (e.g., DNA and/or RNA) can be purified from the sample. Basic texts disclosing the general molecular biology methods include Sambrook and Russell, *Molecular Cloning, A Laboratory Manual* (3rd ed. 2001); Kriegler, *Gene Transfer and Expression: A Laboratory Manual* (1990); and *Current Protocols in Molecular Biology* (Ausubel et al., eds., 1994-1999). Such nucleic acids may also be obtained through in vitro amplification methods such as those described herein and in Berger, Sambrook, and Ausubel, as well as Mullis *et al.*, (1987) U.S. Pat. No. 4,683,202; *PCR Protocols A Guide to Methods and Applications* (Innis *et al.*, eds) Academic Press Inc. San Diego, Calif. (1990) (Innis); Arnheim & Levinson (Oct. 1, 1990) C&EN 36-47; *The Journal Of NIH Research* (1991) 3: 81-94; Kwoh *et al.* (1989) *Proc. Natl. Acad. Sci. USA* 86: 1173; Guatelli *et al.* (1990) *Proc. Natl. Acad. Sci. USA* 87, 1874; Lomell

et al. (1989) *J. Clin. Chem.*, 35: 1826; Landegr
Brunt (1990) *Biotechnology* 8: 291-294; Wu and Wallace (1989) *Gene* 4: 560; and Barringer et al. (1990) *Gene* 89: 117, each of which is incorporated by reference in its entirety for all purposes and in particular for all teachings related to amplification methods. In some embodiments, the nucleic acids will not be amplified before they are quantified.

[0066] Any of a variety of detection methods can be used to screen an individual's sample for one or more of the features listed in TABLEs A, B, C, D, or E . For example, in some embodiments, nucleic acid hybridization and/or amplification methods are used to detect and quantify one or more of the features. In some embodiments, an immunoassay or other assay to detect and quantify one or more specific proteins determinative of one or more of the criteria can be used. For example, solid-phase ELISA immunoassays, Western blots, or immunohistochemistry are routinely used to specifically detect a protein. *See*, Harlow and Lane *Antibodies, A Laboratory Manual*, Cold Spring Harbor Publications, NY (1988) for a description of immunoassay formats and conditions that can be used to determine specific immunoreactivity. In some preferred embodiments, nucleotide sequencing is used to identify and quantify one or more of the criteria.

[0067] DNA sequencing can be performed as desired. Such sequencing can be performed using known sequencing methodologies, *e.g.*, Illumina, Life Technologies, and Roche 454 sequencing systems. In typical embodiments, a sample is sequenced using a large-scale sequencing method that provides the ability to obtain sequence information from many reads. Such sequencing platforms include those commercialized by Roche 454 Life Sciences (GS systems), Illumina (e.g., HiSeq, MiSeq) and Life Technologies (e.g., SOLiD systems).

[0068] The Roche 454 Life Sciences sequencing platform involves using emulsion PCR and immobilizing DNA fragments onto bead. Incorporation of nucleotides during synthesis is detected by measuring light that is generated when a nucleotide is incorporated.

[0069] The Illumina technology involves the attachment of genomic DNA to a planar, optically transparent surface. Attached DNA fragments are extended and bridge amplified to create an ultra-high density sequencing flow cell with clusters containing copies of the same template. These templates are sequenced using a sequencing-by-synthesis technology that employs reversible terminators with removable fluorescent dyes.

[0070] Methods that employ sequencing by hybridization may also be used. Such methods, *e.g.*, used in the Life Technologies SOLiD4+ technology uses a pool of all possible oligonucleotides of a fixed length, labeled according to the sequence. Oligonucleotides are

annealed and ligated; the preferential ligation b
in a signal informative of the nucleotide at that position.

[0071] The sequence can be determined using any other DNA sequencing method including, *e.g.,* methods that use semiconductor technology to detect nucleotides that are incorporated into an extended primer by measuring changes in current that occur when a nucleotide is incorporated (see, *e.g.*, U.S. Patent Application Publication Nos. 20090127589 and 20100035252). Other techniques include direct label-free exonuclease sequencing in which nucleotides cleaved from the nucleic acid are detected by passing through a nanopore (Oxford Nanopore) (Clark e*t al., Nature Nanotechnology* 4: 265 – 270, 2009); and Single Molecule Real Time (SMRT™) DNA sequencing technology (Pacific Biosciences), which is a sequencing-by synthesis technique.

[0072] Deep sequencing can be used to quantify the number of copies of a particular sequence in a sample and then also be used to determine the relative abundance of different sequences in a sample. Deep sequencing refers to highly redundant sequencing of a nucleic acid sequence, for example such that the original number of copies of a sequence in a sample can be determined or estimated. The redundancy (i.e., depth) of the sequencing is determined by the length of the sequence to be determined (X), the number of sequencing reads (N), and the average read length (L). The redundancy is then NxL/X. The sequencing depth can be, or be at least about 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55 ,56, 57, 58, 59, 60, 70, 80, 90, 100, 110, 120, 130, 150, 200, 300, 500, 500, 700, 1000, 2000, 3000, 4000, 5000 or more. *See*, *e.g.*, Mirebrahim, Hamid *et al.*, *Bioinformatics* 31 (12): i9–i16 (2015).

[0073] In some embodiments, specific sequences in the sample can be targeted for amplification and/or sequencing. For example, specific primers can be used to detect and sequence bacterial sequences of interest. Exemplary target sequences can include, but are not limited to, the 16S rRNA coding sequence (e.g., gene families mentioned in the discussion of Block S120), as well as gene sequences involved in one or more genetic pathway as shown in TABLEs A, B, C, D, or E. In addition, or alternatively, whole genome sequencing methods that randomly sequence DNA fragments in a sample can be used.

[0074] Once sequencing raw data is generated, the resulting sequence reads can be "mapped" to known sequences in a genomic database. Exemplary algorithms that are suitable for determining percent sequence identity and sequence similarity and thus aligning

and identifying sequence reads are the BLAST described in Altschul *et al.* (1990) *J. Mol. Biol.* 215: 403-410 and Altschul *et al.* (1977) *Nucleic Acids Res.* 25: 3389-3402, respectively. Software for performing BLAST analyses is publicly available through the National Center for Biotechnology Information (NCBI) web site. Accordingly, for the sequence reads generated, a subset of these reads will be aligned to one or more bacterial genomes of the bacterial taxa in TABLEs A, B, C, D, or E or can be aligned to a gene sequence in any genome that has a genetic function as set forth in TABLEs A, B, C, D, or E. For example, one can align a read with a database of bacterial sequences and the read can be designated as from a particular bacteria if that read has the best alignment to a DNA sequence from that bacteria in the database.

[0075] Similarly, one can align a read with a database of bacterial sequences and the read can be designated as from a genetic pathway if that read has the best alignment to a DNA sequence from that genetic pathway in the database. For example, one can assign the read to a sequence from a particular Kyoto Encyclopedia of Genes and Genomes (KEGG) category or Clusters of Orthologous Groups (COG) categories. KEGGs are described more at genome.jp/kegg/. COGs are described in, e.g., Tatusov, *et al.*, *Nucleic Acids Res.* 2000 Jan 1; 28(1): 33–36. The TABLEs provided herein lists various KEGG and COG categories that are correlated with the presence or absence of a microbiome indicative of a cerebro-craniofacial health issue. Different levels of KEGG or COG categories are provided in TABLEs A, B, C, D, or E. Values in TABLEs A, B, C, D, and E for particular criteria are proportional values compared to totals at that taxonomic or functional designation level.

[0076] Assuming sequencing has occurred at a sufficient depth, one can quantify the number of reads for sequences indicative of the presence of a feature of TABLEs A, B, C, D, or E, thereby allowing one to set a value for an estimated amount of one of the criterion. The number of reads or other measures of amount of one of the features can be provided as an absolute or relative value. An example of an absolute value is the number of reads of 16S rRNA coding sequence reads that map to the genus of Bacteroides. Alternatively, relative amounts can be determined. An exemplary relative amount calculation is to determine the amount of 16S rRNA coding sequence reads for a particular bacterial taxon (e.g., genus , family, order, class, or phylum) relative to the total number of 16S rRNA coding sequence reads assigned to the bacterial domain. A value indicative of amount of a feature in the sample can then be compared to a cut-off value or a probability distribution in a disease signature for a microbiome indicative of a cerebro-craniofacial health issue. For example, if the signature indicates that a relative amount of feature #1 of 50% or more of all features

possible at that level indicates the likelihood of craniofacial health issue, then quantification of gene sequences associated with feature #1 less than 50% in a sample would indicate a higher likelihood of a microbiome that is not indicative of a cerebro-craniofacial health issue and alternatively, quantification of gene sequences associated with feature #1 more than 50% in a sample would indicate a higher likelihood of a microbiome indicative of a cerebro-craniofacial health issue.

[0077] Once amounts of various features from TABLEs A, B, C, D, or E have been determined and compared to a cut-off or probability value for the corresponding criteria in a disease signature for a cerebro-craniofacial health issue, one can determine the likelihood of a microbiome indicative of a cerebro-craniofacial health issue in the individual.

[0078] Disease signatures can include criteria corresponding to one or at least one of the features set forth in TABLEs A, B, C, D, or E . In some embodiments, 2, 3, or 4 of the criteria of TABLE A can be used in a disease signature for a microbiome indicative of a insomnia issue. In some embodiments, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 or more (e.g., all) of the criteria of TABLE B can be used in a disease signature for a microbiome indicative of a light sleep issue. In some embodiments, various numbers of the criteria of TABLE C can be used in a disease signature for a microbiome indicative of a headache issue. In some embodiments, various numbers of the criteria of TABLE D can be used in a disease signature for a microbiome indicative of a sinusitis issue. In some embodiments, various numbers of the criteria of TABLE E can be used in a disease signature for a microbiome indicative of a poor concentration issue.

[0079] In some embodiments, supplementary information about the individual can also be used in the disease signature and thus also for determining the likelihood of occurrence of a microbiome indicative of a cerebro-craniofacial health issue in the individual. Supplementary information can include, for example, different demographics (e.g., genders, ages, marital statuses, ethnicities, nationalities, socioeconomic statuses, sexual orientations, etc.), different health conditions (e.g., health and disease states), different living situations (e.g., living alone, living with pets, living with a significant other, living with children, etc.), different dietary habits (e.g., omnivorous, vegetarian, vegan, sugar consumption, acid consumption, etc.), different behavioral tendencies (e.g., levels of physical activity, drug use, alcohol use, etc.), different levels of mobility (e.g., related to distance traveled within a given time period), biomarker states (e.g., cholesterol levels, lipid levels, etc.), weight, height, body mass index, genotypic factors, and any other suitable trait that has an effect on microbiome composition.

[0080] FIG. 1A is a flowchart of an embodim
classification of the presence or absence of a microbiome indicative of a cerebro-craniofacial health issue, such as insomnia, light sleep, headache, sinusitis, or poor concentration and/or determining the course of treatment for the individual human having the microbiome indicative of a cerebro-craniofacial health issue, such as insomnia, light sleep, headache, sinusitis, or poor concentration.

[0081] At block 10, a sample comprising bacteria from the individual human is provided. In specific examples, samples can comprise stool samples, blood samples, saliva samples, plasma/serum samples (e.g., to enable extraction of cell-free DNA), cerebrospinal fluid, and tissue samples. In some cases, the sample is an oral sample (e.g., a throat, tongue, or gum swab, or saliva), or a sample (*e.g.*, a nucleic acid sample, such as a DNA sample) extracted from an oral sample.

[0082] At block 11, an amount(s) of bacteria taxon and/or gene sequence corresponding to gene functionality as set forth in TABLEs A, B, C, D, or E is determined. As various examples, an amount of one bacteria taxon can be determined; an amount of one gene sequence corresponding to gene functionality can be determined; an amount of one bacteria taxon and an amount one gene sequence corresponding to gene functionality can be determined; multiple amounts (e.g., 2-4) of bacteria taxa can be determined; multiple amounts (e.g., 2-6) of gene sequences corresponding to gene functionalities can be determined; and multiple amounts of both can be determined.

[0083] The amount can be determined in various ways, e.g., by sequencing nucleic acids in the sample, using a hybridization array, and PCR. As examples, the amounts can correspond to levels of a signal or a count of numbers of nucleic acids corresponding to each taxa. The amount can be a relative abundance value.

[0084] At block 12, the determined amount(s) are compared to a condition signature having cut-off or probability values for amounts of the bacteria taxon and/or gene sequence for an individual having a microbiome indicative of a cerebro-craniofacial health issue or an individual not having a microbiome indicative of a cerebro-craniofacial health issue or both. In various embodiments, each amount can be compared to a separate value, and a number of taxa exceeding that value can be compared to a threshold for determining whether a sufficient number of the taxa provide the condition signature. Other examples are provider herein. Before a comparison to a probability value, the amount can be transformed (e.g., via a probability distribution). As another example, the amounts can be used to determine a

measure probability, which can be compared to
among classifications.

[0085] At block 13, a classification of the presence or absence of the microbiome indicative of a cerebro-craniofacial health issue is determined based on the comparing, and/or the course of treatment for the individual human having the microbiome indicative of a cerebro-craniofacial health issue is determined based on the comparing. As described herein, the classification can be binary or includes more levels, e.g., corresponding to a probability.

III. TREATMENT OF ISSUES RELATED TO THE DISEASE

[0086] Also provided are methods of determining a course of treatment, and/or optionally of treating, an individual having a microbiome indicative of a cerebro-craniofacial health issue. For example, by detecting the presence, absence, or quantity of one or more of the criteria set forth in TABLEs A, B, C, D, or E , one can determine treatments to increase those criteria that are reduced in individuals having a condition/disease (*i.e.*, individuals having a microbiome indicative of a cerebro-craniofacial health issue) or decrease these criteria that are increased in individuals having the disease (a cerebro-craniofacial health issue) compared to healthy individuals (*i.e.*, individuals having a microbiome that is not indicative of a cerebro-craniofacial health issue). In some embodiments, the individual will have been diagnosed, optionally by other methods, of having a microbiome associated with a cerebro-craniofacial health issue, or symptoms thereof, and the methods described herein (e.g., comparison to the disease signature) will reveal excessive amounts and/or deficient amounts of one or more of the features that can then be used to guide treatment.

[0087] For example, in embodiments in which the amount of a particular bacteria type is lower in individuals having a microbiome indicative of a cerebro-craniofacial health issue than in individuals having a microbiome that is not indicative of a cerebro-craniofacial health issue, a possible treatment is providing a probiotic or prebiotic treatment that provides or stimulates growth of the particular bacteria type.

[0088] In embodiments in which the higher amount of bacteria is in the individual having a microbiome indicative of a cerebro-craniofacial health issue, one can administer treatments that reduce the relative amount of that particular bacteria. In some embodiments, antibiotics can be administered to reduce the target bacterial population. Alternatively, other treatments can be administered including promoting (by administration of probiotics or prebiotics) bacteria that compete with the target bacteria. In yet another embodiment, bacteriophage targeting the particular bacteria can be administered to the individual.

[0089] Similarly, where a particular function
one can increase or reduce that function by selectively promoting or reducing growth of bacterial populations that have that particular function.

[0090] Additional mechanisms of treatment are listed, for example, in FIG. 5.

[0091] Further, one can monitor treatment of an individual having a microbiome indicative of a cerebro-craniofacial health issue by obtaining samples from the individual before, during, and/or after treatment of the cerebro-craniofacial health issue, or before, during, and/or after treatment to mitigate the symptoms of a cerebro-craniofacial health issue (*e.g.*, prebiotic, probiotic, or bacteriophage therapy), or the combination thereof, to monitor progression of the cerebro-craniofacial health issue (e.g., monitor progression of insomnia, light sleep, headache, sinusitis, or poor concentration). For example, in some embodiments, levels of one or more of the criteria in TABLEs A, B, C, D, or E are determined one or more (e.g., 2 or more, 3, 4, 5 or more) times and the dosage of a pre-biotic and/or pro-biotic treatment can be adjusted up or down depending on how the criteria respond to the treatment.

IV. ANALYSIS OF SEQUENCE INFORMATION

[0092] In some embodiments, sequence information can be received. The sequence information can correspond to one or more sequence reads per nucleic acid molecule (e.g., a DNA fragment). The sequence reads can be obtained in a variety of ways. For example, a hybridization array, PCR, or sequencing techniques can be used.

[0093] When sequencing is performed, a sequence read can be aligned (mapped) to a plurality of reference bacterial genomes (also called reference genomes) to determine which reference bacterial genome the sequence read aligns and where on that reference genome the sequence read aligns. The alignment can be to a particular region (e.g., 16S region) of a reference genome, and thus to a reference sequence, which can be all or part of the reference genome. For paired-end sequencing, both sequence reads can be aligned as a pair, with an expected length of the nucleic acid molecule being used to aid in the alignment.

[0094] Accordingly, it can be determined that a particular DNA fragment is derived from a particular gene of a particular bacterial taxonomic group (also called taxon) based on the aligned location of a sequence read to the particular gene of the particular bacterial taxonomic group. The same determination may be made by various hybridization probes using a variety of techniques, as will be known by one skilled in the art. Thus, the mapping can be performed in a variety of ways.

[0095] In this manner, a count of the number more genes of different bacterial taxonomic groups can be determined. The count for each gene and for each taxonomic group can be used to determine relative abundances. For example, a relative abundance value (RAV) of a particular taxonomic group can be determined based on a fraction (proportion) of sequence reads aligning to that taxonomic group relative to other taxonomic groups. The RAV can correspond to the proportion of reads assigned to a particular taxonomic or functional group. The proportion can be relative to various denominator values, e.g., relative to all of the sequence reads, relative to all assigned to at least one group (taxonomic or functional), or all assigned to for a given level in the hierarchy. The alignment can be implemented in any manner that can assign a sequence read to a particular taxonomic or functional group. For example, based on the mappings to the reference sequence(s) in the 16S region, a taxonomic group with the best match for the alignment can be identified. The RAV can then be determined for that taxonomic group using the number of sequence reads (or votes of sequence reads) for a particular sequence group divided by the number of sequence reads identified as being bacterial, which may be for a specific region or even for a given level of a hierarchy.

[0096] A taxonomic group can include one or more bacteria and their corresponding reference sequences. A taxonomic group can correspond to any set of one or more reference sequences for one or more loci (e.g., genes) that represent the taxonomic group. Any given level of a taxonomic hierarchy would include a plurality of taxonomic groups. For instance, a reference sequence in the one group at the genus level can be in another group at the family level. A sequence read can be assigned based on the alignment to a taxonomic group when the sequence read aligns to a reference sequence of the taxonomic group. A functional group can correspond to one or more genes labeled as having a similar function. Thus, a functional group can be represented by reference sequences of the genes in the functional group, where the reference sequences of a particular gene can correspond to various bacteria. The taxonomic and functional groups can collectively be referred to as sequence groups, as each group includes one or more reference sequences that represent the group. A taxonomic group of multiple bacteria can be represented by multiple reference sequence, e.g., one reference sequence per bacteria species in the taxonomic group. Embodiments can use the degree of alignment of a sequence read to multiple reference sequences to determine which sequence group to assign the sequence read based on the alignment.

[0097] As mentioned above, a particular genomic region (e.g., gene 16S) can be analyzed. For example, the region can be amplified, and a portion of the amplified DNA fragments can

be sequenced. The amplification can be to such
the amplified region. Other example regions can be smaller than a gene, e.g., variable regions within a gene. The longer the region, more resolution can be obtained to determine voting to assign a sequence read to a group. Multiple non-contiguous regions can be analyzed, e.g., by amplifying multiple regions.

A. Example determination of relative abundance of a sequence group (feature)

[0098] As mentioned above, a relative abundance value can correspond to a proportion of sequence reads that align to at least one reference sequence of a sequence group, also referred to as a feature herein. A sequence read can be assigned to one or more sequence groups based on the alignment to the reference sequence(s) for each sequence group. A sequence read can be assigned to more than one sequence group if the assigned groups are in different categories (e.g., taxonomic or functional) or in different levels of a hierarchy (e.g., genus and family). And, a sequence group can include multiple sequences for different regions or a same region, e.g., a sequence group can include more than one base at a particular position, e.g., if the group encompasses various polymorphisms at a genomic position. A sequence group is an example of a feature that can be used to characterize a sample, e.g., when the sequence group has a statistically significant separation between the control population and the disease population.

1. Assignment to a sequence group

[0099] In some embodiments, sequence reads can be obtained for two ends of a nucleic acid molecule, e.g., via paired-end sequencing. Embodiments can identify whether each sequence read of a pair of sequence reads corresponds to a particular sequence group. Each sequence read can effectively have a vote, and the nucleic acid molecule can be identified as corresponding to a particular sequence group only if both sequence reads are aligned to that sequence group (alignment may allow mismatches when less than 100% sequence identity is used). In such embodiments, molecules that do not have both sequence reads aligning to the same sequence group can be discarded. The alignment to a reference sequence may be required to be perfect (i.e., no mismatches), while other embodiments can allow mismatches. Further, the alignment can be required to be unique, or else the read is discarded.

[0100] In other embodiments, a partial vote can be attributed to each sequence group to which a sequence read aligns. In one implementation, a weight of the partial vote based on the degree of alignment, e.g., whether there are any mismatches. In other implementations,

each sequence read can get a vote when it does
weighted by the probability of its existence in humans. A total weight for a read being assigned to a particular reference sequence can be determined by various factors, each providing a weight. The total votes to the reference sequence of a group can be determined and compared to the total votes for other groups in the same level. For each read, the sequence group at a given level with the highest percentage for assignment to the read can be assigned the read. Various techniques of partial assignment can be used, e.g., Dirichlet partial assignment.

[0101] Sequencing can be advantageous for assigning sequence reads to a group, as sequencing provides the actual sequence of at least a portion of a nucleic acid molecule. The sequence might be slightly different than what has already been known for a particular taxonomic group, but it may be similar enough to assign to a particular taxonomic group. If predetermined probes were used, then that nucleic acid molecule might not be identified. Thus, one can identify unknown bacteria, but whose sequence is similar enough to an existing taxonomic group, or even assigned to an unknown group.

[0102] In some embodiments, the proportion can be the total of sequence reads, even if some are not assigned, or equivalently assigned to an unknown group. As an example, the 16S gene can be analyzed, and a read can be determined to align to one or more reference sequences in the region, e.g., with a certain number of mismatches below a threshold, but with a high enough variations to not correspond to any known taxonomic group (or functional group as discussed below). Thus, embodiments can include unassigned reads that contribute to the denominator for determining the proportion of reads of a certain sequence group relative to the sequence reads identified, e.g., as being bacterial. Thus, a proportion of the bacterial population of sequence reads can be determined. Using predetermined probes would generally not allow one to identify unknown bacterial sequences.

2. Sequence group corresponds to a particular taxonomic group

[0103] A taxonomic group can correspond to any set of one or more reference sequences for one or more loci (e.g., genes) that represent the taxonomic group. Any given level of a taxonomic hierarchy would include a plurality of taxonomic groups. The taxonomic groups of a given level of the taxonomic hierarchy would typically be mutually exclusive. Thus, a reference sequence of one taxonomic group would not be included in another taxonomic group in the same level. For example, a reference sequence in one group at the genus level

would not be included in another group at the g
one group at the genus level can be in another group at the family level.

[0104] The RAV can correspond to the proportion of reads assigned to a particular taxonomic group. The proportion can be relative to various denominator values, e.g., relative to all of the sequence reads, relative to all assigned to at least one group (taxonomic or functional), or all assigned to for a given level in the hierarchy. The alignment can be implemented in any manner that can assign a sequence read to a particular taxonomic group.

[0105] For example, based on the mappings to the reference sequence(s) in the 16S region, a taxonomic group with the best match for the alignment can be identified. The RAV can then be determined for that taxonomic group using the number of sequence reads (or votes of sequence reads) for a particular sequence group divided by the number of sequence reads identified, e.g., as being bacterial, which may be for a specific region or even for a given level of a hierarchy.

3. Sequence group corresponds to a particular gene or functional group

[0106] Instead of or in addition to determining a count of the sequence reads that correspond to a particular taxonomic group, embodiments can use a count of a number of sequence reads that correspond to a particular gene or a collection of genes having an annotation of a particular function, where the collection is called a functional group. The RAV can be determined in a similar manner as for a taxonomic group. For example, functional group can include a plurality of reference sequences corresponding to one or more genes of the functional group. Reference sequences of multiple bacteria for a same gene can correspond to a same functional group. Then, to determine the RAV, the number of sequence reads assigned to the functional group can be used to determine a proportion for the functional group.

[0107] The use of a function group, which may include a single gene, can help to identify situations where there is a small change (e.g., increase) in many taxonomic groups such that the change is too small to be statistically significant. But, the changes may all be for a same gene or set of genes of a same functional group, and thus the change for that functional group can be statistically significant, even though the changes for the taxonomic groups may not be significant. The reverse can be true of a taxonomic group being more predictive than a particular functional group, e.g., when a single taxonomic group includes many genes that have changed by a relatively small amount.

[0108] As an example, if 10 taxonomic group
discriminate between the two groups may be low when each taxonomic group is analyzed individually. But, if the increase is all for genes(s) of a same functional group, then the increase would be 100%, or a doubling of the proportion for that taxonomic group. This large increase would have a much larger statistical power for discriminating between the two groups. Thus, the functional group can act to provide a sum of small changes for various taxonomic groups. And, small changes for various functional groups, which happen to all be on a same taxonomic group, can sum to provide high statistical power for that particular taxonomic group.

[0109] The taxonomic groups and functional groups can supplement each other as the information can be orthogonal, or at least partially orthogonal as there still may be some relationship between the RAVs of each group. For example, the RAVs of one or more taxonomic groups and functional groups can be used together as multiple features of a feature vector, which is analyzed to provide a diagnosis, as is described herein. For instance, the feature vector can be compared to a disease signature as part of a characterization model.

B. Example determination of statistically significant separation of abundance of a sequence group between control and disease populations

[0110] Embodiments can use the relative abundance values (RAVs) for populations of subjects that have a disease (condition population; *i.e.*, individuals having a microbiome indicative of a cerebro-craniofacial health issue) and that do not have the disease (control population; *i.e.*, individuals having a microbiome that is not indicative of a cerebro-craniofacial health issue). If the distribution of RAVs of a particular sequence group for the disease population is statistically different than the distribution of RAVs for the control population, then the particular sequence group can be identified for including in a disease signature. Since the two populations have different distributions, the RAV for a new sample for a sequence group in the disease signature can be used to classify (e.g., determine a probability) of whether the sample does or does not have the disease. The classification can also be used to determine a treatment, as is described herein. A discrimination level can be used to identify sequence groups that have a high predictive value. Thus, embodiment can filter out taxonomic groups that are not very accurate for providing a diagnosis.

1. Discrimination level of sequence group

[0111] Once RAVs of a sequence group have been determined for the control and condition populations, various statistical tests can be used to determine the statistical power of the

sequence group for discriminating between a ce
and no cerebro-craniofacial health issue (control). In one embodiment, the Kolmogorov-Smirnov (KS) test can be used to provide a probability value (p-value) that the two distributions are actually identical. The smaller the p-value the greater the probability to correctly identify which population a sample belongs. The larger the separation in the mean values between the two populations generally results in a smaller p-value (an example of a discrimination level). Other tests for comparing distributions can be used. The Welch's t-test presumes that the distributions are Gaussian, which is not necessarily true for a particular sequence group. The KS test, as it is a non-parametric test, is well suited for comparing distributions of taxa or functions for which the probability distributions are unknown.

[0112] The distribution of the RAVs for the control and condition populations can be analyzed to identify sequence groups with a large separation between the two distributions. The separation can be measured as a p-value (See example section). For example, the relative abundance values for the control population may have a distribution peaked at a first value with a certain width and decay for the distribution. And, the disease population can have another distribution that is peaked a second value that is statistically different than the first value. In such an instance, an abundance value of a control sample has a lower probability to be within the distribution of abundance values encountered for the disease samples. The larger the separation between the two distributions, the more accurate the discrimination is for determining whether a given sample belongs to the control population or the disease population. As is discussed later, the distributions can be used to determine a probability for an RAV as being in the control population and determine a probability for the RAV being in the disease population.

[0113] FIG. 7 shows a plot illustrating the control distribution and the disease distribution for insomnia where the sequence group is Moryella for the Genus taxonomic group according to embodiments of the present invention. As one can see, the RAVs for the disease group having a microbiome indicative of insomnia tend to have higher values than the control distribution. Thus, if Moryella is present, a higher RAV would have a higher probability of being in the insomnia population. The p-value in this instance is 9.34×10^{-9}, as indicated in TABLE A.

[0114] One of skill in the art will appreciate that, in some cases, the RAVs for the disease having a microbiome indicative of a cerebro-craniofacial health issue can have lower values than the control distribution. For example, the RAVs of the genus taxonomic group Roseburia for the insomnia condition group tend to have lower values than the control group.

Thus, if Roseburia is present, a lower RAV wo
insomnia population. The p-value in this instance is 4.53×10^{-7}, as indicated in TABLE A.

[0115] FIG. 8 shows a plot illustrating the control distribution and the disease distribution for insomnia where the sequence group is Selenocompound metabolism for the function taxonomic group according to embodiments of the present invention. As one can see, the RAVs for the disease group having a microbiome indicative of insomnia tend to have lower values than the control distribution. Thus, if sequences associated with Selenocompound metabolism is present, a lower RAV would have a higher probability of being in the insomnia population. The p-value in this instance is 1.99×10^{-13}, as indicated in TABLE A.

[0116] FIG. 9 shows a plot illustrating the control distribution and the disease distribution for light sleep where the sequence group is Lactobacillaceae for the Family taxonomic group according to embodiments of the present invention. As one can see, the RAVs for the disease group having a microbiome indicative of light sleep tend to have higher values than the control distribution. Thus, if Lactobacillaceae is present, a higher RAV would have a higher probability of being in the light sleep population. The p-value in this instance is 4.72×10^{-8}, as indicated in TABLE B.

[0117] FIG. 10 shows a plot illustrating the control distribution and the disease distribution for light sleep where the sequence group is Translation for the function taxonomic group according to embodiments of the present invention. As one can see, the RAVs for the disease group having a microbiome indicative of light sleep tend to have lower values than the control distribution. Thus, if sequences associated with Translation is present, a lower RAV would have a higher probability of being in the light sleep population. The p-value in this instance is 1.13×10^{-7}, as indicated in TABLE B.

[0118] FIG. 11 shows a plot illustrating the control distribution and the disease distribution for headache where the sequence group is Marvinbryantia for the Genus taxonomic group according to embodiments of the present invention. As one can see, the RAVs for the disease group having a microbiome indicative of headache tend to have lower values than the control distribution. Thus, if Marvinbryantia is present, a lower RAV would have a higher probability of being in the headache population. The p-value in this instance is 1.31×10^{-6}, as indicated in TABLE C.

[0119] FIG. 12 shows a plot illustrating the control distribution and the disease distribution for headache where the sequence group is Selenocompound metabolism for the function taxonomic group according to embodiments of the present invention. As one can see, the

RAVs for the disease group having a microbior
values than the control distribution. Thus, if sequences associated with Selenocompound metabolism is present, a lower RAV would have a higher probability of being in the headache population. The p-value in this instance is 1.71×10^{-17}, as indicated in TABLE C.

[0120] FIG. 13 shows a plot illustrating the control distribution and the disease distribution for sinusitis where the sequence group is Clostridiales for the Genus taxonomic group according to embodiments of the present invention. As one can see, the RAVs for the disease group having a microbiome indicative of sinusitis tend to have lower values than the control distribution. Thus, if Clostridiales is present, a lower RAV would have a higher probability of being in the sinusitis population. The p-value in this instance is 5.22×10^{-5}, as indicated in TABLE D.

[0121] FIG. 14 shows a plot illustrating the control distribution and the disease distribution for poor concentration where the sequence group is Moryella for the Genus taxonomic group according to embodiments of the present invention. As one can see, the RAVs for the disease group having a microbiome indicative of poor concentration tend to have higher values than the control distribution. Thus, if Moryella is present, a higher RAV would have a higher probability of being in the poor concentration population. The p-value in this instance is 5.06×10^{-7}, as indicated in TABLE E.

[0122] FIG. 15 shows a plot illustrating the control distribution and the disease distribution for poor concentration where the sequence group is Propanoate metabolism for the function taxonomic group according to embodiments of the present invention. As one can see, the RAVs for the disease group having a microbiome indicative of poor concentration tend to have lower values than the control distribution. Thus, if sequences associated with Propanoate metabolism is present, a lower RAV would have a higher probability of being in the poor concentration population. The p-value in this instance is 1.63×10^{-8}, as indicated in TABLE E.

2. Prevalence of sequence group in population

[0123] In some embodiments, certain samples may not have any presence of a particular taxonomic group, or at least not a presence above a relatively low threshold (i.e., a threshold below either of the two distributions for the control and condition population). Thus, a particular sequence group may be prevalent in the population, e.g., more than 30% of the population may have the taxonomic group. Another sequence group may be less prevalent in the population, e.g., showing up in only 5% of the population. The prevalence (e.g.,

percentage of population) of a certain sequence
likely the sequence group may be used to determine a diagnosis.

[0124] In such an example, the sequence group can be used to determine a status of the disease (e.g., diagnose for the disease) when the subject falls within the 30%. But, when the subject does not fall within the 30%, such that the taxonomic group is simply not present, the particular taxonomic group may not be helpful in determining a diagnosis of the subject. Thus, whether a particular taxonomic group or functional group is useful in diagnosing a particular subject can be dependent on whether nucleic acid molecules corresponding to the sequence group are actually sequenced.

[0125] Accordingly, the disease signature can include more sequence groups that are used for a given subject. As an example, the disease signature can include 100 sequence groups, but only 60 of sequence groups may be detected in a sample. The classification of the subject (including any probability for being in the application) would be determined based on the 60 sequence groups.

C. Example generation of characterization model

[0126] The sequence groups with high discrimination levels (e.g., low p-values) for a given condition (*e.g.*, a cerebro-craniofacial health issue) can be identified and used as part of a characterization model, e.g., which uses a disease signature to determine a probability of a subject having the disease. The disease signature can include a set of sequence groups as well as discriminating criteria (e.g., cutoff values and/or probability distributions) used to provide a classification of the subject. The classification can be binary (e.g., indicative of a cerebro-craniofacial health issue or not indicative of a cerebro-craniofacial health issue) or have more classifications (e.g., probability of being indicative of a cerebro-craniofacial health issue or not being indicative of a cerebro-craniofacial health issue). Which sequence groups of the disease signature that are used in making a classification be dependent on the specific sequence reads obtained, e.g., a sequence group would not be used if no sequence reads were assigned to that sequence group. In some embodiments, a separate characterization model can be determined for different populations, e.g., by geography where the subject is currently residing (e.g., country, region, or continent), the generic history of the subject (e.g., ethnicity), or other factors.

1. Selection of sequence g

[0127] As mentioned above, sequence groups having at least a specified discrimination level can be selected for inclusion in the characterization model. In various embodiments, the specified discrimination level can be an absolute level (e.g., having a p-value below a specified value), a percentage (e.g., being in the top 10% of discriminating levels), or a specified number of the top discrimination levels (e.g., the top 100 discriminating levels). In some embodiments, the characterization model can include a network graph, where each node in a graph corresponds to a sequence group having at least a specified discrimination level.

[0128] The sequence groups used in a disease signature of a characterization model can also be selected based on other factors. For example, a particular sequence group may only be detected in a certain percentage of the population, referred to as a coverage percentage. An ideal sequence group would be detected in a high percentage of the population and have a high discriminating level (e.g., a low p-value). A minimum percentage may be required before adding the sequence group to the characterization model for a particular disease (*e.g.*, a cerebro-craniofacial health issue). The minimum percentage can vary based on the accompanying discriminating level. For instance, a lower coverage percentage may be tolerated if the discriminating level is higher. As a further example, 95% of the patients with a disease may be classified with one or a combination of a few sequence groups, and the 5% remaining can be explained based on one sequence group, which relates to the orthogonality or overlap between the coverage of sequence groups. Thus, a sequence group that provides discriminating power for 5% of the individuals having the disease (*e.g.*, a cerebro-craniofacial health issue) may be valuable.

[0129] Another factor for determining which sequence to include in a disease signature of the characterization model is the overlap in the subjects exhibiting the sequence groups of a disease signature. For example, to sequence groups can both have a high coverage percentage, but sequence groups may cover the exact same subjects. Thus, adding one of the sequence groups does increase the overall coverage of the disease signature. In such a situation, the two sequence groups can be considered parallel to each other. Another sequence group can be selected to add to the characterization model based on the sequence group covering different subjects than other sequence groups already in the characterization model. Such a sequence group can be considered orthogonal to the already existing sequence groups in the characterization model.

[0130] As examples, selecting a sequence gr
may appear in 100% of control individuals and in 100% of individuals having a specified disease (*e.g.*, a cerebro-craniofacial health issue), but where the distributions are so close in both groups, that knowing the relative abundance of that taxa only allows to catalogue a few individuals as having the disease or lacking the disease (i.e. it has a low discriminating level). Whereas, a taxa that appears in only 20% of individuals not having the disease and 30% of individuals having the disease can have distributions of relative abundance that are so different from one another, it allows to catalogue 20% of individuals not having the disease and 30% of individuals having the disease (i.e. it has a high discriminating level).

[0131] In some embodiments, machine learning techniques can allow the automatic identification of the best combination of features (e.g., sequence groups). For instance, a Principal Component Analysis can reduce the number of features used for classification to only those that are the most orthogonal to each other and can explain most of the variance in the data. The same is true for a network theory approach, where one can create multiple distance metrics based on different features and evaluate which distance metric is the one that best separates individuals having the disease (a cerebro-craniofacial health issue) from individuals that do not have the disease.

2. Discrimination criteria sequence groups

[0132] The discrimination criteria for the sequence groups included in the disease signature of a characterization model can be determined based on the disease distributions and the control distributions for the disease. For example, a discrimination criterion for a sequence group can be a cutoff value that is between the mean values for the two distributions. As another example, discrimination criteria for a sequence group can include probability distributions for the control and disease populations. The probability distributions can be determined in a separate manner from the process of determining the discrimination level.

[0133] The probability distributions can be determined based on the distribution of RAVs for the two populations. The mean values (or other average or median) for the two populations can be used to center the peaks of the two probability distributions. For example, if the mean RAV of the disease population is 20% (or 0.2), then the probability distribution for the disease population can have its peak at 20%. The width or other shape parameters (e.g., the decay) can also be determined based on the distribution of RAVs for the disease population. The same can be done for the control population.

D. Use of sequence groups

[0134] The sequence groups included in the disease signature of the characterization can be used to classify a new subject. The sequence groups can be considered features of the feature vector, or the RAVs of the sequence groups considered as features of a feature vector, where the feature vector can be compared to the discriminating criteria of the disease signature. For instance, the RAVs of the sequence groups for the new subject can be compared to the probability distributions for each sequence group of the disease signature. If an RAV is zero or nearly zero, then the sequence group may be skipped and not used in the classification.

[0135] The RAVs for sequence groups that are exhibited in the new subject can be used to determine the classification. For example, the result (e.g., a probability value) for each exhibited sequence group can be combined to arrive at the final classification. As another example, clustering of the RAVs can be performed, and the clusters can be used to determine a classification of a disease.

1. Classification of disease using sequence groups

[0136] Embodiments can provide a method for determining a classification of the presence or absence for a disease and/or determine a course of treatment for an individual human having the disease (a cerebro-craniofacial health issue such as insomnia, light sleep, headache, sinusitis, or poor concentration). The method can be performed by a computer system, as described herein. FIG. 1B is a flowchart of an embodiment of a method for determining a classification of the presence or absence of a microbiome indicative of a cerebro-craniofacial health issue and/or determining the course of treatment for an individual human having the microbiome indicative of a cerebro-craniofacial health issue.

[0137] In block 20, sequence reads of bacterial DNA obtained from analyzing a test sample from the individual human are received. The analysis can be done with various techniques, e.g., as described herein, such as sequencing or hybridization arrays. The sequence reads can be received at a computer system, e.g., from a detection apparatus, such as a sequencing machine that provides data to a storage device (which can be loaded into the computer system) or across a network to the computer system.

[0138] In block 21, the sequence reads are mapped to a bacterial sequence database to obtain a plurality of mapped sequence reads. The bacterial sequence database includes a plurality of reference sequences of a plurality of bacteria. The reference sequences can be for predetermined region(s) of the bacteria, e.g., the 16S region.

[0139] In block 22, the mapped sequence rea
the mapping to obtain assigned sequence reads assigned to at least one sequence group. A sequence group includes one or more of the plurality of reference sequences. The mapping can involve the sequence reads being mapped to one or more predetermined regions of the reference sequences. For example, the sequence reads can be mapped to the 16S gene. Thus, the sequence reads do not have to be mapped to the whole genome, but only to the region(s) covered by the reference sequences of a sequence group.

[0140] In block 23, a total number of assigned sequence reads is determined. In some embodiments, the total number of assigned reads can include reads identified as being, e.g., bacterial, but not assigned to a known sequence group. In other embodiments, the total number can be a sum of sequence reads assigned to known sequence groups, where the sum may include any sequence read assigned to at least one sequence group.

[0141] In block 24, relative abundance value(s) can be determined. For example, for each sequence group of a disease signature set of one or more sequence groups selected from TABLEs A, B, C, D, or E, a relative abundance value of assigned sequence reads assigned to the sequence group relative to the total number of assigned sequence reads can be determined. The relative abundance values can form a test feature vector, where each values of the test feature vector is an RAV of a different sequence group.

[0142] In block 25, the test feature vector is compared to calibration feature vectors generated from relative abundance values of calibration samples having a known status of the disease. The calibration samples may be samples of a disease population and samples of a control population. In some embodiments, the comparison can involve various machine learning techniques, such as supervised machine learning (e.g. decision trees, nearest neighbor, support vector machines, neural networks, naïve Bayes classifier, etc…) and unsupervised machine learning (e.g., clustering, principal component analysis, etc…).

[0143] In one embodiment, clustering can use a network approach, where the distance between each pair of samples in the network is computed based on the relative abundance of the sequence groups that are relevant for each disease. Then, a new sample can be compared to all samples in the network, using the same metric based on relative abundance, and it can be decided to which cluster it should belong. A meaningful distance metric would allow all individuals having the disease (a cerebro-craniofacial health issue) to form one or a few clusters and all individuals lacking the disease to form one or a few clusters. One distance

metric is the Bray-Curtis dissimilarity, or equiv
is 1 – Bray-Curtis dissimilarity. Another example distance metric is the Tanimoto coefficient.

[0144] In some embodiments, the feature vectors may be compared by transforming the RAVs into probability values, thereby forming probability vectors. Similar processing for the feature vectors can be performed for the probability, with such a process still involving a comparison of the feature vectors since the probability vectors are generated from the feature vectors.

[0145] Block 26 can determine a classification of the presence or absence of the disease (*e.g.*, a cerebro-craniofacial health issue) and/or determine a course of treatment for an individual human having the disease based on the comparing. For example, the cluster to which the test feature vector is assigned may be a disease cluster, and the classification can be made that the individual human has the disease or a certain probability for having the disease.

[0146] In one embodiment involving clustering, the calibration feature vectors can be clustered into a control cluster not having the disease and a disease cluster having the disease. Then, which cluster the test feature vector belongs can be determined. The identified cluster can be used to determine the classification or select a course of treatment. In one implementation, the clustering can use a Bray–Curtis dissimilarity.

[0147] In one embodiment involving a decision tree, the comparison may be performed to by comparing the test feature vector to one or more cutoff values (e.g., as a corresponding cutoff vector), where the one or more cutoff values are determined from the calibration feature vectors, thereby providing the comparison. Thus, the comparison can include comparing each of the relative abundance values of the test feature vector to a respective cutoff value determined from the calibration feature vectors generated from the calibration samples. The respective cutoff values can be determined to provide an optimal discrimination for each sequence group.

2. Use of probability values

[0148] A new sample can be measured to detect the RAVs for the sequence groups in the disease signature. The RAV for each sequence group can be compared to the probability distributions for the control and disease populations for the particular sequence group. For example, the probability distribution for the disease population can provide an output of a probability (e.g., a conditional probability) of having the disease (condition) for a given input

of the RAV. Similarly, the probability distribut
output of a probability (control probability) of not having the disease for a given input of the RAV. Thus, the value of the probability distribution at the RAV can provide the probability of the sample being in each of the populations. Thus, it can be determined which population the sample is more likely to belong to, by taking the maximum probability.

[0149] In some embodiments, just the maximum probability is used in further steps of a characterization process. In other embodiments, both the disease probability and the control probability are used. As noted above, the probability distributions used here for classification may be different than the statistical test used to determine whether the distribution of RAV values are separated, e.g., the KS test.

[0150] A total probability across sequence groups of a disease signature can be used. For all of the sequence groups that are measured, a disease probability can be determined for whether the sample is in the disease group and a control probability can be determined for whether the sample is in the control population. In other embodiments, just the disease probabilities or just the control probabilities can be determined.

[0151] The probabilities across the sequence groups can be used to determine a total probability. For example, an average of the conditional probabilities can be determined, thereby obtaining a final disease probability of the subject having the disease based on the disease signature. An average of the control probabilities can be determined, thereby obtaining a final control probability of the subject not having the disease based on the disease signature.

[0152] In one embodiment, the final disease probability and final control probability can be compared to each other to determine the final classification. For instance, a difference between the two final probabilities can be determined, and a final classification probability determined from the difference. A large positive difference with final disease probability being higher would result in a higher final classification probability of the subject having the disease.

[0153] In other embodiments, only the final disease probability can be used to determine the final classification probability. For example, the final classification probability can be the final disease probability. Alternatively, the final classification probability can be one minus the final control probability, or 100% minus the final control probability depending on the formatting of the probabilities.

[0154] In some embodiments, a final classific
can be combined with other final classification probabilities of other disease of the same class. The aggregated probability can then be used to determine whether the subject has at least one of the class of diseases. Thus, embodiments can determine whether a subject has a health issue that may include a plurality of diseases associated with that health issue.

[0155] The classification can be one of the final probabilities. In other examples, embodiments can compare a final probability to a threshold value to make a determination of whether the disease exists. For example, the respective conditional probabilities can be averaged, and an average can be compared to a threshold value to determine whether the disease exists. As another example, the comparison of the average to the threshold value can provide a treatment for treating the subject.

V. ADDITIONAL EMBODIMENTS

[0156] Described herein, and with reference to the FIGs, are additional illustrative embodiments of the methods, compositions, and systems provided herein. It will be appreciated that one of ordinary skill in the art can readily determine where and when any one or more of the methods, compositions, and/or systems described above can be utilized additionally, or alternatively, in the embodiments described below.

[0157] As shown in FIG. 1E, a first method 100 for diagnosing and treating an individual having a microbiome indicative of a cerebro-craniofacial health issue can comprise: receiving an aggregate set of samples from a population of subjects S110; characterizing a microbiome composition and/or functional features for each of the aggregate set of samples associated with the population of subjects, thereby generating at least one microbiome composition dataset, at least one microbiome functional diversity dataset, or a combination thereof, for the population of subjects S120. In some cases, the method can further comprise: receiving a supplementary dataset, associated with at least a subset of the population of subjects, wherein the supplementary dataset is informative of characteristics associated with a cerebro-craniofacial health issue S130. Typically, the method further comprises: and transforming the features extracted from the at least one microbiome composition dataset, microbiome functional diversity dataset, or the combination thereof, into a characterization model of a cerebro-craniofacial health issue S140. In some cases, the transforming includes transforming the supplementary dataset, if received. In some variations, the first method 100 can further include: based upon the characterization,

generating a therapy model configured to in
having a cerebro-craniofacial health issue S150.

[0158] The first method 100 functions to generate models that can be used to characterize and/or diagnose subjects according to at least one of their microbiome composition and functional features (e.g., as a clinical diagnostic, as a companion diagnostic, etc.), and provide therapeutic measures (e.g., probiotic-based therapeutic measures, phage-based therapeutic measures, small-molecule-based therapeutic measures, prebiotic-based therapeutic measures, clinical measures, etc.) to subjects based upon microbiome analysis for a population of subjects. As such, data from the population of subjects can be used to characterize subjects according to their microbiome composition and/or functional features, indicate states of health and areas of improvement based upon the characterization(s), and promote one or more therapies that can modulate the composition of a subject's microbiome toward one or more of a set of desired equilibrium states.

[0159] In variations, the method 100 can be used to promote targeted therapies to subjects having a microbiome indicative of a cerebro-craniofacial health issue. In some cases, the targeted therapies are promoted when the cerebro-craniofacial health issue produces observed differences in insomnia, light sleep, headache, sinusitis, or poor concentration or at least one of social behavior, motor behavior, and energy levels, gastrointestinal heath, etc. In these variations, diagnostics associated with a cerebro-craniofacial health issue can be typically assessed using one or more of: a survey instrument or study, such as a sleep study, and any other standard tool. As such, the method 100 can be used to characterize the effects of a cerebro-craniofacial health issue, including disorders, and/or adverse states in an entirely non-typical method. In particular, the inventors propose that characterization of the microbiome of individuals can be useful for predicting the likelihood of a cerebro-craniofacial health issue in subjects. Such characterizations can also be useful for screening for symptoms related to a cerebro-craniofacial health issue and/or determining a course of treatment for an individual human having a microbiome indicative of a cerebro-craniofacial health issue. For example, by deep sequencing bacterial DNAs from subjects having a cerebro-craniofacial health issue and control subjects, the inventors propose that features associated with certain microbiome compositional and/or functional features (e.g., the amount of certain bacteria and/or bacterial sequences corresponding to certain genetic pathways) can be used to predict the presence or absence of a microbiome indicative of a cerebro-craniofacial health issue. The bacteria and genetic pathways in some cases are present in a certain abundance in individuals having a microbiome indicative of a cerebro-

craniofacial health issue as discussed in more

pathways are at a statistically different abundance in individuals not having a microbiome indicative of a cerebro-craniofacial health issue.

[0160] As such, in some embodiments, outputs of the first method 100 can be used to generate diagnostics and/or provide therapeutic measures for a subject based upon an analysis of the subject's microbiome composition and/or functional features of the subject's microbiome. Thus, as shown in FIG. 1F, a second method 200 derived from at least one output of the first method 100 can include: receiving a biological sample from a subject S210; characterizing the subject as having or not having a microbiome indicative of a cerebro-craniofacial health issue based upon processing a microbiome dataset derived from the biological sample S220; and promoting a therapy to the subject with the microbiome indicative of a cerebro-craniofacial health issue based upon the characterization and the therapy model S230. Variations of the method 200 can further facilitate monitoring and/or adjusting of therapies provided to a subject, for instance, through reception, processing, and analysis of additional samples from a subject throughout the course of therapy. Embodiments, variations, and examples of the second method 200 are described in more detail below.

[0161] Thus, methods 100 and/or 200 can function to generate models that can be used to classify individuals and/or provide therapeutic measures (e.g., therapy recommendations, therapies, therapy regimens, etc.) to individuals based upon microbiome analysis for a population of individuals. As such, data from the population of individuals can be used to generate models that can classify individuals according to their microbiome compositions (e.g., as a diagnostic measure), indicate states of health and areas of improvement based upon the classification(s), and/or provide therapeutic measures that can push the composition of an individual's microbiome toward one or more of a set of improved equilibrium states. Variations of the second method 200 can further facilitate monitoring and/or adjusting of therapies provided to an individual, for instance, through reception, processing, and analysis of additional samples from an individual throughout the course of therapy.

[0162] In one application, at least one of the methods 100, 200 is implemented, at least in part, at a system 300, as shown in FIG. 2, that receives a biological sample derived from the subject (or an environment associated with the subject) by way of a sample reception kit, and processes the biological sample at a processing system implementing a characterization process and a therapy model configured to positively influence a microorganism

distribution in the subject (e.g., human, non-hu

In variations of the application, the processing system can be configured to generate and/or improve the characterization process and the therapy model based upon sample data received from a population of subjects. The method 100 can, however, alternatively be implemented using any other suitable system(s) configured to receive and process microbiome-related data of subjects, in aggregation with other information, in order to generate models for microbiome-derived diagnostics and associated therapeutics. Thus, the method 100 can be implemented for a population of subjects (e.g., including the subject, excluding the subject), wherein the population of subjects can include patients dissimilar to and/or similar to the subject (e.g., in health condition, in dietary needs, in demographic features, etc.). Thus, information derived from the population of subjects can be used to provide additional insight into connections between behaviors of a subject and effects on the subject's microbiome, due to aggregation of data from a population of subjects.

[0163] Thus, the methods 100, 200 can be implemented for a population of subjects (e.g., including the subject, excluding the subject), wherein the population of subjects can include subjects dissimilar to and/or similar to the subject (e.g., health condition, in dietary needs, in demographic features, etc.). Thus, information derived from the population of subjects can be used to provide additional insight into connections between behaviors of a subject and effects on the subject's microbiome, due to aggregation of data from a population of subjects.

A. Sample Handling

[0164] Block S110 recites: receiving an aggregate set of biological samples from a population of subjects, which functions to enable generation of data from which models for characterizing subjects and/or providing therapeutic measures to subjects can be generated. In Block S110, biological samples are preferably received from subjects of the population of subjects in a non-invasive manner. In variations, non-invasive manners of sample reception can use any one or more of: a permeable substrate (e.g., a swab configured to wipe a region of a subject's body, toilet paper, a sponge, etc.), a non-permeable substrate (e.g., a slide, tape, etc.), a container (e.g., vial, tube, bag, etc.) configured to receive a sample from a region of a subject's body, and any other suitable sample-reception element. In a specific example, samples can be collected from one or more of a subject's nose, skin, genitals, mouth, and gut in a non-invasive manner (e.g., using a swab and a vial). However, one or more biological samples of the set of biological samples can additionally or alternatively be received in a semi-invasive manner or an invasive manner. In variations,

invasive manners of sample reception can use biopsy element, a lance, and any other suitable instrument for collection of a sample in a semi-invasive or invasive manner. In specific examples, samples can comprise blood samples, plasma/serum samples (e.g., to enable extraction of cell-free DNA), cerebrospinal fluid, and tissue samples. In some cases, the sample is a stool sample, or a sample (*e.g.*, a nucleic acid sample, such as a DNA sample) extracted from a stool sample.

[0165] In the above variations and examples, samples can be taken from the bodies of subjects without facilitation by another entity (e.g., a caretaker associated with an individual, a health care professional, an automated or semi-automated sample collection apparatus, etc.), or can alternatively be taken from bodies of individuals with the assistance of another entity. In one example, wherein samples are taken from the bodies of subjects without facilitation by another entity in the sample extraction process, a sample-provision kit can be provided to a subject. In the example, the kit can include one or more swabs or sample vials for sample acquisition, one or more containers configured to receive the swab(s) or sample vials for storage, instructions for sample provision and setup of a user account, elements configured to associate the sample(s) with the subject (e.g., barcode identifiers, tags, etc.), and a receptacle that allows the sample(s) from the individual to be delivered to a sample processing operation (e.g., by a mail delivery system). In another example, wherein samples are extracted from the user with the help of another entity, one or more samples can be collected in a clinical or research setting from a subject (e.g., during a clinical appointment).

[0166] In Block S110, the aggregate set of biological samples is preferably received from a wide variety of subjects, and can involve samples from human subjects and/or non- human subjects. In relation to human subjects, Block S110 can include receiving samples from a wide variety of human subjects, collectively including subjects of one or more of: different demographics (e.g., genders, ages, marital statuses, ethnicities, nationalities, socioeconomic statuses, sexual orientations, etc.), different health conditions (e.g., health and disease states), different living situations (e.g., living alone, living with pets, living with a significant other, living with children, etc.), different dietary habits (e.g., omnivorous, vegetarian, vegan, sugar consumption, acid consumption, etc.), different behavioral tendencies (e.g., levels of physical activity, drug use, alcohol use, etc.), different levels of mobility (e.g., related to distance traveled within a given time period), biomarker states (e.g., cholesterol levels, lipid levels, etc.), weight, height, body mass index, genotypic factors, and any other suitable trait that has an effect on microbiome composition. As such, as the number of subjects increases,

the predictive power of feature-based models g
100 increases, in relation to characterizing a variety of subjects based upon their microbiomes. Additionally or alternatively, the aggregate set of biological samples received in Block S110 can include receiving biological samples from a targeted group of similar subjects in one or more of: demographic traits, health conditions, living situations, dietary habits, behavior tendencies, levels of mobility, age range (e.g., pediatric, adulthood, geriatric), and any other suitable trait that has an effect on microbiome composition. Additionally or alternatively, the methods 100, and/or 200 can be adapted to characterize diseases typically detected by way of lab tests (e.g., polymerase chain reaction based tests, cell culture based tests, blood tests, biopsies, chemical tests, etc.), physical detection methods (e.g., manometry), medical history based assessments, behavioral assessments, and imagenology based assessments. Additionally or alternatively, the methods 100, 200 can be adapted to characterization of acute conditions, chronic conditions, conditions with difference in prevalence for different demographics, conditions having characteristic disease areas (e.g., the head, the gut, endocrine system diseases, the heart, nervous system diseases, respiratory diseases, immune system diseases, circulatory system diseases, renal system diseases, locomotor system diseases, etc.), and comorbid conditions.

[0167] In some embodiments, receiving the aggregate set of biological samples in Block S110 can be performed according to embodiments, variations, and examples of sample reception as described in U.S. App. No. 14/593,424 filed on 09-JAN-2015 and entitled "Method and System for Microbiome Analysis", which is incorporated herein in its entirety by this reference. However, receiving the aggregate set of biological samples in Block S110 can additionally or alternatively be performed in any other suitable manner. Furthermore, some alternative variations of the first method 100 can omit Block S110, with processing of data derived from a set of biological samples performed as described below in subsequent blocks of the method 100.

B. Sample Analysis

[0168] Block S120 recites: characterizing a microbiome composition and/or functional features for each of the aggregate set of biological samples associated with a population of subjects, thereby generating at least one of a microbiome composition dataset and a microbiome functional diversity dataset for the population of subjects. Block S120 functions to process each of the aggregate set of biological samples, in order to determine compositional and/or functional aspects associated with the microbiome of each of a population of subjects. Compositional and functional aspects can include compositional

aspects at the microorganism level, including p
microorganisms across different groups of kingdoms, phyla, classes, orders, families, genera, species, subspecies, strains, infraspecies taxon (e.g., as measured in total abundance of each group, relative abundance of each group, total number of groups represented, etc.), and/or any other suitable taxa. Compositional and functional aspects can also be represented in terms of operational taxonomic units (OTUs). Compositional and functional aspects can additionally or alternatively include compositional aspects at the genetic level (e.g., regions determined by multilocus sequence typing, 16S sequences, 18S sequences, ITS sequences, other genetic markers, other phylogenetic markers, etc.). Compositional and functional aspects can include the presence or absence or the quantity of genes associated with specific functions (e.g., enzyme activities, transport functions, immune activities, etc.). Outputs of Block S120 can thus be used to provide features of interest for the characterization process of Block S140, wherein the features can be microorganism-based (e.g., presence of a genus of bacteria), genetic-based (e.g., based upon representation of specific genetic regions and/or sequences) and/or functional-based (e.g., presence of a specific catalytic activity, presence of metabolic pathways, etc.).

[0169] In one variation, Block S120 can include characterization of features based upon identification of phylogenetic markers derived from bacteria and/or archaea in relation to gene families associated with one or more of: ribosomal protein S2, ribosomal protein S3, ribosomal protein S5, ribosomal protein S7, ribosomal protein S8, ribosomal protein S9, ribosomal protein S10, ribosomal protein S11, ribosomal protein S12/S23, ribosomal protein S13, ribosomal protein S15P/S13e, ribosomal protein S17, ribosomal protein S19, ribosomal protein L1, ribosomal protein L2, ribosomal protein L3, ribosomal protein L4/L1e, ribosomal protein L5, ribosomal protein L6, ribosomal protein L10, ribosomal protein L11, ribosomal protein L13, ribosomal protein L14b/L23e, ribosomal protein L15, ribosomal protein L16/L10E, ribosomal protein L18P/L5E, ribosomal protein L22, ribosomal protein L24, ribosomal protein L25/L23, ribosomal protein L29, translation elongation factor EF-2, translation initiation factor IF-2, metalloendopeptidase, ffh signal recognition particle protein, phenylalanyl-tRNA synthetase alpha subunit, phenylalanyl- tRNA synthetase beta subunit, tRNA pseudouridine synthase B, porphobilinogen deaminase, phosphoribosylformylglycinamidine cyclo-ligase, and ribonuclease HII. However, the markers can include any other suitable marker(s).

[0170] Characterizing the microbiome composition and/or functional features for each of the aggregate set of biological samples in Block S120 thus can include a combination of

sample processing techniques (e.g., wet laborat
(e.g., utilizing tools of bioinformatics) to quantitatively and/or qualitatively characterize the microbiome and functional features associated with each biological sample from a subject or population of subjects.

[0171] In variations, sample processing in Block S120 can include any one or more of: lysing a biological sample, disrupting membranes in cells of a biological sample, separation of undesired elements (e.g., RNA, proteins) from the biological sample, purification of nucleic acids (e.g., DNA) in a biological sample, amplification of nucleic acids from the biological sample, further purification of amplified nucleic acids of the biological sample, and sequencing of amplified nucleic acids of the biological sample. Thus, portions of Block S120 can be implemented using embodiments, variations, and examples of the sample handling network and/or computing system as described in U.S. App. No. 14/593,424 filed on 09-JAN-2015 and entitled "Method and System for microbiome Analysis", which is incorporated herein in its entirety by this reference. Thus the computing system implementing one or more portions of the method 100 can be implemented in one or more computing systems, wherein the computing system(s) can be implemented at least in part in the cloud and/or as a machine (e.g., computing machine, server, mobile computing device, etc.) configured to receive a computer-readable medium storing computer-readable instructions. However, Block S120 can be performed using any other suitable system(s).

[0172] In variations, lysing a biological sample and/or disrupting membranes in cells of a biological sample preferably includes physical methods (e.g., bead beating, nitrogen decompression, homogenization, sonication), which omit certain reagents that produce bias in representation of certain bacterial groups upon sequencing. Additionally or alternatively, lysing or disrupting in Block S120 can involve chemical methods (e.g., using a detergent, using a solvent, using a surfactant, etc.). Additionally or alternatively, lysing or disrupting in Block S120 can involve biological methods. In variations, separation of undesired elements can include removal of RNA using RNases and/or removal of proteins using proteases. In variations, purification of nucleic acids can include one or more of: precipitation of nucleic acids from the biological samples (e.g., using alcohol-based precipitation methods), liquid-liquid based purification techniques (e.g., phenol-chloroform extraction), chromatography-based purification techniques (e.g., column adsorption), purification techniques involving use of binding moiety-bound particles (e.g., magnetic beads, buoyant beads, beads with size distributions, ultrasonically responsive beads, etc.) configured to bind nucleic acids and configured to release nucleic acids in the presence of an elution environment (e.g., having an

elution solution, providing a pH shift, providing suitable purification techniques.

[0173] In variations, performing an amplification operation S123 on purified nucleic acids can include performing one or more of: polymerase chain reaction (PCR)-based techniques (e.g., solid-phase PCR, RT-PCR, qPCR, multiplex PCR, touchdown PCR, nanoPCR, nested PCR, hot start PCR, etc.), helicase-dependent amplification (HDA), loop mediated isothermal amplification (LAMP), self-sustained sequence replication (3SR), nucleic acid sequence based amplification (NASBA), strand displacement amplification (SDA), rolling circle amplification (RCA), ligase chain reaction (LCR), and any other suitable amplification technique. In amplification of purified nucleic acids, the primers used are preferably selected to prevent or minimize amplification bias, as well as configured to amplify nucleic acid regions/sequences (e.g., of the 16S region, the 18S region, the ITS region, etc.) that are informative taxonomically, phylogenetically, for diagnostics, for formulations (e.g., for probiotic formulations), and/or for any other suitable purpose. Thus, universal primers (e.g., a F27-R338 primer set for 16S rRNA, a F515-R806 primer set for 16S RNA, etc.) configured to avoid amplification bias can be used in amplification. Primers used in variations of Block S120 (*e.g.*, S123 and/or S124) can additionally or alternatively include incorporated barcode sequences specific to each biological sample, which can facilitate identification of biological samples post-amplification. Primers used in variations of Block S120 (*e.g.*, S123 and/or S124) can additionally or alternatively include adaptor regions configured to cooperate with sequencing techniques involving complementary adaptors (e.g., according to protocols for Illumina Sequencing).

[0174] Identification of a primer set for a multiplexed amplification operation can be performed according to embodiments, variations, and examples of methods described in U.S. App. No. 62/206,654 filed 18-AUG-2015 and entitled "Method and System for Multiplex Primer Design", which is herein incorporated in its entirety by this reference. Performing a multiplexed amplification operation using a set of primers in Block S123 can additionally or alternatively be performed in any other suitable manner.

[0175] Additionally or alternatively, as shown in FIG. 3, Block S120 can implement any other step configured to facilitate processing (e.g., using a Nextera kit) for performance of a fragmentation operation S122 (e.g., fragmentation and tagging with sequencing adaptors) in cooperation with the amplification operation S123 (e.g., S122 can be performed after S123, S122 can be performed before S123, S122 can be performed substantially contemporaneously with S123, etc.). Furthermore, Blocks S122 and/or S123

can be performed with or without a nucleic a

can be performed prior to amplification of nucleic acids, followed by fragmentation, and then amplification of fragments. Alternatively, extraction can be performed, followed by fragmentation and then amplification of fragments. As such, in some embodiments, performing an amplification operation in Block S123 can be performed according to embodiments, variations, and examples of amplification as described in U.S. App. No. 14/593,424 filed on 09-JAN-2015 and entitled "Method and System for microbiome Analysis". Furthermore, amplification in Block S123 can additionally or alternatively be performed in any other suitable manner.

[0176] In a specific example, amplification and sequencing of nucleic acids from biological samples of the set of biological samples includes: solid-phase PCR involving bridge amplification of DNA fragments of the biological samples on a substrate with oligo adapters, wherein amplification involves primers having a forward index sequence (e.g., corresponding to an illumina forward index for miSeq/NextSeq/HiSeq platforms) and/or a reverse index sequence (e.g., corresponding to an Illumina reverse index for MiSeq/NextSeq/HiSeq platforms), a forward barcode sequence and/or a reverse barcode sequence, optionally a transposase sequence (e.g., corresponding to a transposase binding site for MiSeq/NextSeq/HiSeq platforms), optionally a linker (e.g., a zero, one, or two-base fragment configured to reduce homogeneity and improve sequence results), optionally an additional random base, and optionally a sequence for targeting a specific target region (e.g., 16S region, 18S region, ITS region). In some cases, amplification involves one or both primers having any combination of the foregoing elements, or all of the foregoing elements. Amplification and sequencing can further be performed on any suitable amplicon, as indicated throughout the disclosure. In the specific example, sequencing comprises Illumina sequencing (e.g., with a HiSeq platform, with a MiSeq platform, with a NextSeq platform, etc.) using a sequencing-by-synthesis technique. Additionally or alternatively, any other suitable next generation sequencing technology (e.g., PacBio platform, MinION platform, Oxford Nanopore platform, etc.) can be used. Additionally or alternatively, any other suitable sequencing platform or method can be used (e.g., a Roche 454 Life Sciences platform, a Life Technologies SOLiD platform, etc.). In examples, sequencing can include deep sequencing to quantify the number of copies of a particular sequence in a sample and then also be used to determine the relative abundance of different sequences in a sample. The sequencing depth can be, or be at least about 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53,

54, 55 ,56, 57, 58, 59, 60, 70, 80, 90, 100, 110
1000, 2000, 3000, 4000, 5000 or more.

[0177] Some variations of sample processing in Block S120 can include further purification of amplified nucleic acids (e.g., PCR products) prior to sequencing, which functions to remove excess amplification elements (e.g., primers, dNTPs, enzymes, salts, etc.). In examples, additional purification can be facilitated using any one or more of: purification kits, buffers, alcohols, pH indicators, chaotropic salts, nucleic acid binding filters, centrifugation, and any other suitable purification technique.

[0178] In variations, computational processing in Block S120 can include any one or more of: performing a sequencing analysis operation S124 including identification of microbiome-derived sequences (e.g., as opposed to subject sequences and contaminants), performing an alignment and/or mapping operation S125 of microbiome-derived sequences (e.g., alignment of fragmented sequences using one or more of single-ended alignment, ungapped alignment, gapped alignment, pairing), and generating features S126 derived from compositional and/or functional aspects of the microbiome associated with a biological sample.

[0179] Performing the sequencing analysis operation S124 with identification of microbiome-derived sequences can include mapping of sequence data from sample processing to a subject reference genome (e.g., provided by the Genome Reference Consortium), in order to remove subject genome-derived sequences. Unidentified sequences remaining after mapping of sequence data to the subject reference genome can then be further clustered into operational taxonomic units (OTUs) based upon sequence similarity and/or reference-based approaches (e.g., using VAMPS, using MG-RAST, and/or using QIIME databases), aligned (e.g., using a genome hashing approach, using a Needleman- Wunsch algorithm, using a Smith-Waterman algorithm), and mapped to reference bacterial genomes (e.g., provided by the National Center for Biotechnology Information), using an alignment algorithm (e.g., Basic Local Alignment Search Tool, FPGA accelerated alignment tool, BWT-indexing with BWA, BWT-indexing with SOAP, BWT-indexing with Bowtie, etc.). Mapping of unidentified sequences can additionally or alternatively include mapping to reference archaeal genomes, viral genomes and/or eukaryotic genomes. Furthermore, mapping of taxa can be performed in relation to existing databases, and/or in relation to custom-generated databases.

[0180] Additionally or alternatively, in relati
diversity dataset, Block S120 can include extracting candidate features associated with functional aspects of one or more microbiome components of the aggregate set of biological samples S127, as indicated in the microbiome composition dataset. Extracting candidate functional features can include identifying functional features associated with one or more of: prokaryotic clusters of orthologous groups of proteins (COGs); eukaryotic clusters of orthologous groups of proteins (KOGs); any other suitable type of gene product; an RNA processing and modification functional classification; a chromatin structure and dynamics functional classification; an energy production and conversion functional classification; a cell cycle control and mitosis functional classification; an amino acid metabolism and transport functional classification; a nucleotide metabolism and transport functional classification; a carbohydrate metabolism and transport functional classification; a coenzyme metabolism functional classification; a lipid metabolism functional classification; a translation functional classification; a transcription functional classification; a replication and repair functional classification; a cell wall/membrane/envelop biogenesis functional classification; a cell motility functional classification; a post-translational modification, protein turnover, and chaperone functions functional classification; an inorganic ion transport and metabolism functional classification; a secondary metabolites biosynthesis, transport and catabolism functional classification; a signal transduction functional classification; an intracellular trafficking and secretion functional classification; a nuclear structure functional classification; a cytoskeleton functional classification; a general functional prediction only functional classification; and a function unknown functional classification; and any other suitable functional classification.

[0181] Additionally or alternatively, extracting candidate functional features in Block S127 can include identifying functional features associated with one or more of: systems information (e.g., pathway maps for cellular and organismal functions, modules or functional units of genes, hierarchical classifications of biological entities); genomic information (e.g., complete genomes, genes and proteins in the complete genomes, orthologous groups of genes in the complete genomes); chemical information (e.g., chemical compounds and glycans, chemical reactions, enzyme nomenclature); health information (e.g., human diseases, approved drugs, crude drugs and health-related substances); metabolism pathway maps; genetic information processing (e.g., transcription, translation, replication and repair, etc.) pathway maps; environmental information processing (e.g., membrane transport, signal transduction, etc.) pathway maps; cellular processes (e.g., cell growth, cell death, cell membrane functions, etc.) pathway maps; organismal systems (e.g., immune system,

endocrine system, nervous system, etc.) pathwa
development pathway maps; and any other suitable pathway map.

[0182] In extracting candidate functional features, Block S127 can comprise performing a search of one or more databases, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) and/or the Clusters of Orthologous Groups (COGs) database managed by the National Center for Biotechnology Information (NCBI). Searching can be performed based upon results of generation of the microbiome composition dataset from one or more of the set of aggregate biological samples and/or sequencing of material from the set of samples. In more detail, Block S127 can include implementation of a data-oriented entry point to a KEGG database including one or more of a KEGG pathway tool, a KEGG BRITE tool, a KEGG module tool, a KEGG ORTHOLOGY (KO) tool, a KEGG genome tool, a KEGG genes tool, a KEGG compound tool, a KEGG glycan tool, a KEGG reaction tool, a KEGG disease tool, a KEGG drug tool, or a KEGG medicus tool. Searching can additionally or alternatively be performed according to any other suitable filters. Additionally or alternatively, Block S127 can include implementation of an organism-specific entry point to a KEGG database including a KEGG organisms tool. Additionally or alternatively, Block S127 can include implementation of an analysis tool including one or more of: a KEGG mapper tool that maps KEGG pathway, BRITE, or module data; a KEGG atlas tool for exploring KEGG global maps, a BlastKOALA tool for genome annotation and KEGG mapping, a BLAST/FASTA sequence similarity search tool, a SIMCOMP chemical structure similarity search tool, and a SUBCOMP chemical substructure search tool. In specific examples, Block S127 can include extracting candidate functional features, based on the microbiome composition dataset, from a KEGG database resource and a COG database resource; moreover, Block S127 can comprise extracting candidate functional features in any other suitable manner. For instance, Block S127 can include extracting candidate functional features, including functional features derived from a Gene Ontology functional classification, and/or any other suitable features.

[0183] In one example, a taxonomic group can include one or more bacteria and their corresponding reference sequences. A sequence read can be assigned based on the alignment to a taxonomic group when the sequence read aligns to a reference sequence of the taxonomic group. A functional group can correspond to one or more genes labeled as having a similar function. Thus, a functional group can be represented by reference sequences of the genes in the functional group, where the reference sequences of a particular gene can correspond to various bacteria. The taxonomic and functional groups can collectively be referred to as

sequence groups, as each group includes one or
group. A taxonomic group of multiple bacteria can be represented by multiple reference sequence, e.g., one reference sequence per bacteria species in the taxonomic group. Embodiments can use the degree of alignment of a sequence read to multiple reference sequences to determine which sequence group to assign the sequence read based on the alignment.

1. Analysis of Sequence Groups

[0184] Instead of or in addition to determining a count of the sequence reads that correspond to a particular taxonomic group, embodiments can use a count of a number of sequence reads that correspond to a particular gene or a collection of genes having an annotation of a particular function, where the collection is called a functional group. The RAV can be determined in a similar manner as for a taxonomic group. For example, functional group can include a plurality of reference sequences corresponding to one or more genes of the functional group. Reference sequences of multiple bacteria for a same gene can correspond to a same functional group. Then, to determine the RAV, the number of sequence reads assigned to the functional group can be used to determine a proportion for the functional group. In exemplary embodiment, the functional group is a KEGG or COG group.

[0185] The use of a functional group, which may include a single gene, can help to identify situations where there is a small change (e.g., increase) in many taxonomic groups such that the individual changes are too small to be statistically significant. In such cases, the changes may all be for a same gene or set of genes of a same functional group, and thus the change for that functional group can be statistically significant, even though the changes for the taxonomic groups may not be statistically significant for a given sequence dataset. The reverse can be true of a taxonomic group being more predictive than a particular functional group, e.g., when a single taxonomic group includes many genes that have changed by a relatively small amount.

[0186] As an example, if 10 taxonomic groups increase by approximately 10%, the statistical power to discriminate between the two groups may be low when each taxonomic group is analyzed individually. But, if the increase is similar all for genes(s) of a shared functional group, then the increase would be 100%, or a doubling of the proportion for that taxonomic group. This large increase would have a much larger statistical power for discriminating between the two groups. Thus, the functional group can act to provide a sum of small changes for various taxonomic groups. And, small changes for various functional

groups, which happen to all be on a same taxon
statistical power for that particular taxonomic group.

2. Exemplary Pipeline for Detecting and Analyzing Taxonomic Groups

[0187] Embodiments can provide a bioinformatics pipeline that taxonomically annotates the microorganisms present in a sample. The example clinical annotation pipeline can comprise the following procedures described herein. FIG. 1C is a flowchart of an embodiment of a method for estimating the relative abundances of a plurality of taxa from a sample and outputting the estimates to a database..

[0188] In block 30, the samples can be identified and the sequence data can be loaded. For example, the pipeline can begin with demultiplexed fastq files (or other suitable files) that are the product of pair-end sequencing of amplicons (e.g., of the V4 region of the 16S gene). All samples can be identified for a given input sequencing file, and the corresponding fastq files can be obtained from the fastq repository server and loaded into the pipeline.

[0189] In block 31, the reads can be filtered. For example, a global quality filtering of reads in the fastq files can accept reads with a global Q-score > 30. In one implementation, for each read, the per-position Q-scores are averaged, and if the average is equal or higher than 30, then the read is accepted, else the read is discarded, as is its paired read.

[0190] In block 32, primers can be identified and removed. In one embodiment, only forward reads that contain the forward primer and reverse reads that contain the reverse primer (allowing annealing of primers with up to 5 mismatches or other number of mismatches) are further considered. Primers and any sequences 5' to them are removed from the reads. The 125 bp (or other suitable number) towards the 3' of the forward primer are considered from the forward reads, and only 124 bp (or other suitable number) towards the 3' of the reverse primer are considered for the reverse reads. All processed forward reads that are < 125bp and reverse reads that are < 124bp are eliminated from further processing as are their paired reads.

[0191] In block 33, the forward and reverse reads can be written to files (e.g., FASTA files). For example, the forward and reverse reads that remained paired can be used to generate files that contain 125bp from the forward read, concatenated to 124bp from the reverse read (in the reverse complement direction).

[0192] In block 34, the sequence reads can be
sequences or determine a consensus sequence for a bacterium. For example, the sequences in the files can be subjected to clustering using the Swarm algorithm [Mahé, F. et al. 2014] with a distance of 1. This treatment allows the generation of cluster composed of a central biological entity, surrounded by sequences which are 1 mutation away from the biological entity, which are less abundant and the result of the normal base calling error associated to high throughput sequencing. Singletons are removed from further analyses. In the remaining clusters, the most abundant sequence per cluster is then used as the representative and assigned the counts of all members in the cluster.

[0193] In block 35, chimeric sequences can be removed. For example, amplification of gene superfamilies can produce the formation of chimeric DNA sequences. These result from a partial PCR product from one member of the superfamily that anneals and extends over a different member of the superfamily in a subsequent cycle of PCR. In order to remove chimeric DNA sequences, some embodiments can use the VSEARCH chimera detection algorithm with the *de novo* option and standard parameters [Rognes, T. et al. 2016]. This algorithm uses abundance of PCR products to identify reference "real" sequences as those most abundant, and chimeric products as those less abundant and displaying local similarity to two or more of the reference sequences. All chimeric sequences can be removed from further analysis.

[0194] In block 36, taxonomy annotation can be assigned to sequences using sequence identity searches. To assign taxonomy to the sequences that have passed all filters above, some embodiments can perform identity searches against a database that contains bacterial strains (e.g., reference sequences) annotated to phylum, class, order, family, genus and species level, at least to a subsection of those taxonomic levels, or any other taxonomic levels. The most specific level of taxonomic annotation for a sequence can be kept, given that higher order taxonomy designations for a lower level taxonomy level can be inferred. The sequence identity search can be performed using the algorithm VSEARCH [Rognes, T. et al. 2016] with parameters (maxaccepts=0, maxrejects=0, id=1) that allow an exhaustive exploration of the reference database used. Decreasing values of sequence identity can be used to assign sequences to different taxonomic groups: > 97% sequence identity for assigning to a species, > 95% sequence identity for assigning to a genus, > 90% for assigning to family, > 85% for assigning to order, > 80% for assigning to class, and > 77% for assigning to phylum.

[0195] In block 37, relative abundances of ea

database. For example, once all sequences have been used to identify identical sequences in the reference database, relative abundance per taxa can be determined by dividing the count of all sequences that are assigned to the same taxonomic group by the total number of reads that passed filters, e.g., were assigned. Results can be uploaded to database tables that are used as repository for the taxonomic annotation data.

3. Exemplary Pipeline for Detecting and Analyzing Functional Groups

[0196] For functional groups, the process can proceed as follows. FIG. 1D is a flowchart of an embodiment of a method for generating features derived from composition and/or functional components of a biological sample or an aggregate of biological samples.

[0197] In block 40, sample OTUs (Operational Taxonomic Units) can be found. This may occur, e.g., after the sixth block described above in section V.B.2. After sample OTUs are found, sequences can be clustered, e.g., based on sequence identity (e.g., 97% sequence identity).

[0198] In block 41, a taxonomy can be assigned, e.g., by comparing OTUs with reference sequences of known taxonomy. The comparison can be based on sequence identity (e.g., 97%).

[0199] In block 42, taxonomic abundance can be adjusted for 16S copy number, or whatever genomic regions may be analyzed. Different species may have different number of copies of the 16S gene, so those possessing a higher number of copies will have more 16S material for PCR amplification at same number of cells than other species. Therefore, abundance can be normalized by adjusting the number of 16S copies.

[0200] In block 43, a pre-computed genomic lookup table can be used to relate taxonomy to functions, and amount of function. For example, a pre-computed genomic lookup table that shows the number of genes for important KEGG or COG functional categories per taxonomic group can be used to estimate the abundance of those functional categories based on the normalized 16S abundance data.

[0201] Upon identification of represented groups of microorganisms of the microbiome associated with a biological sample and/or identification of candidate functional aspects (e.g., functions associated with the microbiome components of the biological samples),

generating features derived from compositional

associated with the aggregate set of biological samples can be performed.

[0202] In one variation, generating features can include generating features derived from multilocus sequence typing (MLST), which can be performed experimentally at any stage in relation to implementation of the methods 100, 200, in order to identify markers useful for characterization in subsequent blocks of the method 100. Additionally or alternatively, generating features can include generating features that describe the presence or absence of certain taxonomic groups of microorganisms, and/or ratios between exhibited taxonomic groups of microorganisms. Additionally or alternatively, generating features can include generating features describing one or more of: quantities of represented taxonomic groups, networks of represented taxonomic groups, correlations in representation of different taxonomic groups, interactions between different taxonomic groups, products produced by different taxonomic groups, interactions between products produced by different taxonomic groups, ratios between dead and alive microorganisms (e.g., for different represented taxonomic groups, e.g., based upon analysis of RNAs), phylogenetic distance (e.g., in terms of Kantorovich-Rubinstein distances, Wasserstein distances etc.), any other suitable taxonomic group-related feature(s), or any other suitable genetic or functional feature(s).

[0203] Additionally or alternatively, generating features can include generating features describing relative abundance of different microorganism groups, for instance, using a sparCC approach, using Genome Relative Abundance and Average size (GAAS) approach and/or using a genome Relative Abundance using Mixture Model theory (GRAMM) approach that uses sequence-similarity data to perform a maximum likelihood estimation of the relative abundance of one or more groups of microorganisms. Additionally or alternatively, generating features can include generating statistical measures of taxonomic variation, as derived from abundance metrics. Additionally or alternatively, generating features can include generating features derived from relative abundance factors (e.g., in relation to changes in abundance of a taxon, which affects abundance of other taxa). Additionally or alternatively, generating features can include generation of qualitative features describing presence of one or more taxonomic groups, in isolation and/or in combination. Additionally or alternatively, generating features can include generation of features related to genetic markers (e.g., representative 16S, 18S, and/or ITS sequences) characterizing microorganisms of the microbiome associated with a biological sample. Additionally or alternatively, generating features can include generation of features related to functional associations of specific genes and/or organisms having the specific genes.

Additionally or alternatively, generating feature
pathogenicity of a taxon and/or products attributed to a taxon. Block S120 can, however, include generation of any other suitable feature(s) derived from sequencing and mapping of nucleic acids of a biological sample. For instance, the feature(s) can be combinatory (e.g., involving pairs, triplets), correlative (e.g., related to correlations between different features), and/or related to changes in features (i.e., temporal changes, changes across sample sites, spatial changes, etc.). Features can, however, be generated in any other suitable manner in Block S120.

4. Use of Supplementary Data

[0204] Block S130 recites: receiving a supplementary dataset, associated with at least a subset of the population of subjects, wherein the supplementary dataset is informative of characteristics associated with the disease or condition. The supplementary dataset can thus be informative of presence of the disease within the population of subjects. Block S130 functions to acquire additional data associated with one or more subjects of the set of subjects, which can be used to train and/or validate the characterization processes performed in block S140. In Block S130, the supplementary dataset can include survey-derived data, and can additionally or alternatively include any one or more of: contextual data derived from sensors, medical data (e.g., current and historical medical data associated with a cerebro-craniofacial health issue or health conditions associated with a cerebro-craniofacial health issue, brain scan data (e.g., imaging or electrocardiogram, EKG), behavioral instrument data, data derived from a tool derived from the Diagnostic and Statistical Manual of Mental Disorders, etc.), and any other suitable type of data.

[0205] In variations of Block S130 including reception of survey-derived data, the survey-derived data preferably provides physiological, demographic, and behavioral information in association with a subject. Physiological information can include information related to physiological features (e.g., height, weight, body mass index, body fat percent, body hair level, etc.). Demographic information can include information related to demographic features (e.g., gender, age, ethnicity, marital status, number of siblings, socioeconomic status, sexual orientation, etc.). Behavioral information can include information related to one or more of: health conditions (e.g., health and disease states), living situations (e.g., living alone, living with pets, living with a significant other, living with children, etc.), dietary habits (e.g., omnivorous, vegetarian, vegan, sugar consumption, acid consumption, etc.), behavioral tendencies (e.g., levels of physical activity, drug use, alcohol use, etc.), different levels of mobility (e.g., related to distance traveled within a given time period), different levels of

sexual activity (e.g., related to numbers of partı
suitable behavioral information. Survey-derived data can include quantitative data and/or qualitative data that can be converted to quantitative data (e.g., using scales of severity, mapping of qualitative responses to quantified scores, etc.).

[0206] In facilitating reception of survey-derived data, Block S130 can include providing one or more surveys to a subject of the population of subjects, or to an entity associated with a subject of the population of subjects. Surveys can be provided in person (e.g., in coordination with sample provision and/or reception from a subject), electronically (e.g., during account setup by a subject, at an application executing at an electronic device of a subject, at a web application accessible through an internet connection, etc.), and/or in any other suitable manner.

[0207] Additionally or alternatively, portions of the supplementary dataset received in Block S130 can be derived from sensors associated with the subject(s) (e.g., sensors of wearable computing devices, sensors of mobile devices, biometric sensors associated with the user, etc.). As such, Block S130 can include receiving one or more of: physical activity- or physical action-related data (e.g., accelerometer and gyroscope data from a mobile device or wearable electronic device of a subject), environmental data (e.g., temperature data, elevation data, climate data, light parameter data, etc.), patient nutrition or diet-related data (e.g., data from food establishment check-ins, data from spectrophotometric analysis, etc.), biometric data (e.g., data recorded through sensors within the patient's mobile computing device, data recorded through a wearable or other peripheral device in communication with the patient's mobile computing device), location data (e.g., using GPS elements), and any other suitable data. Additionally or alternatively, portions of the supplementary dataset can be derived from medical record data and/or clinical data of the subject(s). As such, portions of the supplementary dataset can be derived from one or more electronic health records (EHRs) of the subject(s).

[0208] Additionally or alternatively, the supplementary dataset of Block S130 can include any other suitable diagnostic information (e.g., clinical diagnosis information), which can be combined with analyses derived from features to support characterization of subjects in subsequent blocks of the method 100. For instance, information derived from a colonoscopy, biopsy, blood test, diagnostic imaging, survey-related information, and any other suitable test can be used to supplement Block S130.

5. Characterization of cer

[0209] Block S140 recites: transforming the supplementary dataset and features extracted from at least one of the microbiome composition dataset and the microbiome functional diversity dataset into a characterization model of the disease or condition. Block S140 functions to perform a characterization process for identifying features and/or feature combinations that can be used to characterize subjects or groups with a cerebro-craniofacial health issue based upon their microbiome composition and/or functional features. Additionally or alternatively, the characterization process can be used as a diagnostic tool that can characterize a subject (e.g., in terms of behavioral traits, in terms of medical conditions, in terms of demographic traits, etc.) based upon their microbiome composition and/or functional features, in relation to other health condition states, behavioral traits, medical conditions, demographic traits, and/or any other suitable traits. Such characterization can then be used to suggest or provide personalized therapies by way of the therapy model of Block S150.

[0210] In performing the characterization process, Block S140 can use computational methods (e.g., statistical methods, machine learning methods, artificial intelligence methods, bioinformatics methods, etc.) to characterize a subject as exhibiting features characteristic of a group of subjects with a cerebro-craniofacial health issue.

[0211] In one variation, characterization can be based upon features derived from a statistical analysis (e.g., an analysis of probability distributions) of similarities and/or differences between a first group of subjects exhibiting a target state (e.g., a health condition state) associated with the cerebro-craniofacial health issue, and a second group of subjects not exhibiting the target state (e.g., a "normal" state) associated with absence of a cerebro-craniofacial health issue, or the absence of a microbiome indicative of a cerebro-craniofacial health issue, or the absence of a microbiome indicative of a health and/or quality of life issue caused by a cerebro-craniofacial health issue. In implementing this variation, one or more of a Kolmogorov-Smirnov (KS) test, a permutation test, a Cramér-von Mises test, and any other statistical test (e.g., t-test, Welch's t-test, z-test, chi-squared test, test associated with distributions, etc.) can be used. In particular, one or more such statistical hypothesis tests can be used to assess a set of features having varying degrees of abundance in (or variations across) a first group of subjects exhibiting a target state (e.g., an adverse state) associated with the a cerebro-craniofacial health issue and a second group of subjects not exhibiting the target state (e.g., having a normal state) associated with cerebro-craniofacial health issue. In more detail, the set of features assessed can be constrained based upon percent abundance

and/or any other suitable parameter pertaining t
of subjects and the second group of subjects, in order to increase or decrease confidence in the characterization. In a specific implementation of this example, a feature can be derived from a taxon of microorganism and/or presence of a functional feature that is abundant in a certain percentage of subjects of the first group and subjects of the second group, wherein a relative abundance of the taxon between the first group of subjects and the second group of subjects can be determined from one or more of a KS test or a Welch's t-test (e.g., a t-test with a log normal transformation), with an indication of significance (e.g., in terms of p-value). Thus, an output of Block S140 can comprise a normalized relative abundance value (e.g., 25% greater abundance of a taxon-derived feature and/or a functional feature in cerebro-craniofacial health issue subjects vs. control subjects) with an indication of significance (e.g., a p-value of 0.0013). Variations of feature generation can additionally or alternatively implement or be derived from functional features or metadata features (e.g., non-bacterial markers).

[0212] In variations and examples, characterization can use the relative abundance values (RAVs) for populations of subjects that have the disease (a cerebro-craniofacial health issue) and that do not have the disease (control population). If the distribution of RAVs of a particular sequence group for the disease population is statistically different than the distribution of RAVs for the control population, then the particular sequence group can be identified for including in a disease signature. Since the two populations have different distributions, the RAV for a new sample for a sequence group in the disease signature can be used to classify (e.g., determine a probability) of whether the sample does or does not have, or is indicative of, the disease. The classification can also be used to determine a treatment, as is described herein. A discrimination level can be used to identify sequence groups that have a high predictive value. Thus, embodiment can filter out taxonomic groups and/or functional groups that are not very accurate for providing a diagnosis.

[0213] Once RAVs of a sequence group have been determined for the control and disease populations, various statistical tests can be used to determine the statistical power of the sequence group for discriminating between disease (a cerebro-craniofacial health issue) and the absence of the disease (control). In one embodiment, the Kolmogorov-Smirnov (KS) test can be used to provide a probability value (p-value) that the two distributions are actually identical. The smaller the p-value the greater the probability to correctly identify which population a sample belongs. The larger the separation in the mean values between the two populations generally results in a smaller p-value (an example of a discrimination level).

Other tests for comparing distributions can be u
distributions are Gaussian, which is not necessarily true for a particular sequence group. The KS test, as it is a non- parametric test, is well suited for comparing distributions of taxa or functions for which the probability distributions are unknown.

[0214] The distribution of the RAVs for the control and disease populations can be analyzed to identify sequence groups with a large separation between the two distributions. The separation can be measured as a p-value (See example section). For example, the RAVs for the control population may have a distribution peaked at a first value with a certain width and decay for the distribution. And, the disease population can have another distribution that is peaked a second value that is statistically different than the first value. In such an instance, an abundance value of a control sample has a lower probability to be within the distribution of abundance values encountered for the disease samples. The larger the separation between the two distributions, the more accurate the discrimination is for determining whether a given sample belongs to the control population or the disease population. As is described herein, the distributions can be used to determine a probability for an RAV as being in the control population and determine a probability for the RAV being in the disease population, where sequence groups associated with the largest percentage difference between two means have the smallest p-value, signifying a greater separation between the two populations.

[0215] In performing the characterization process, Block S140 can additionally or alternatively transform input data from at least one of the microbiome composition datasets and/or microbiome functional diversity datasets into feature vectors that can be tested for efficacy in predicting characterizations of the population of subjects. Data from the supplementary dataset can be used to inform characterizations of the cerebro-craniofacial health issue, wherein the characterization process is trained with a training dataset of candidate features and candidate classifications to identify features and/or feature combinations that have high degrees (or low degrees) of predictive power in accurately predicting a classification. As such, refinement of the characterization process with the training dataset identifies feature sets (e.g., of subject features, of combinations of features) having high correlation with a cerebro-craniofacial health issue or a health issue (e.g., symptom) associated with a cerebro-craniofacial health issue.

[0216] In some embodiments, feature vectors effective in predicting classifications of the characterization process can include features related to one or more of: microbiome diversity metrics (e.g., in relation to distribution across taxonomic groups, in relation to distribution across archaeal, bacterial, viral, and/or eukaryotic groups), presence of taxonomic groups in

one's microbiome, representation of specific ge
one's microbiome, relative abundance of taxonomic groups in one's microbiome, microbiome resilience metrics (e.g., in response to a perturbation determined from the supplementary dataset), abundance of genes that encode proteins or RNAs with given functions (enzymes, transporters, proteins from the immune system, hormones, interference RNAs, etc.) and any other suitable features derived from the microbiome composition dataset, the microbiome functional diversity dataset (e.g., COG-derived features, KEGG derived features, other functional features, etc.), and/or the supplementary dataset. Additionally, combinations of features can be used in a feature vector, wherein features can be grouped and/or weighted in providing a combined feature as part of a feature set. For example, one feature or feature set can include a weighted composite of the number of represented classes of bacteria in one's microbiome, presence of a specific genus of bacteria in one's microbiome, representation of a specific 16S sequence in one's microbiome, and relative abundance of a first phylum over a second phylum of bacteria. However, the feature vectors can additionally or alternatively be determined in any other suitable manner.

[0217] In examples of Block S140, assuming sequencing has occurred at a sufficient depth, one can quantify the number of reads for sequences indicative of the presence of a feature, thereby allowing one to set a value for an estimated amount of one of the criteria. The number of reads or other measures of amount of one of the features can be provided as an absolute or relative value. An example of an absolute value is the number of reads of 16S rRNA coding sequence reads that map to the genus of Lachnospira. Alternatively, relative amounts can be determined. An exemplary relative amount calculation is to determine the amount of 16S rRNA coding sequence reads for a particular bacterial taxon (e.g., genus , family, order, class, or phylum) relative to the total number of 16S rRNA coding sequence reads assigned to the bacterial domain. A value indicative of amount of a feature in the sample can then be compared to a cut-off value or a probability distribution in a disease signature for a cerebro-craniofacial health issue. For example, if the disease signature indicates that a relative amount of feature #1 of 50% or more of all features possible at that level indicates the likelihood of a cerebro-craniofacial health issue or a health or quality of life issue attributable to, indicative of, or caused by a cerebro-craniofacial health issue, then quantification of gene sequences associated with feature #1 less than 50% in a sample would indicate a higher likelihood of being from a healthy subject (or at least from a subject that does not have a cerebro-craniofacial health , or does not have a specific a cerebro-craniofacial health issue) and alternatively, quantification of gene sequences associated with feature #1 of more than 50% in a sample would indicate a higher likelihood of the disease.

[0218] In some cases, the taxonomic groups a
features, or as sequence groups in the context of determining an amount of sequence reads corresponding to a particular group (feature). In some cases, scoring of a particular bacteria or genetic pathway can be determined according to a comparison of an abundance value to one or more reference (calibration) abundance values for known samples, e.g., where a detected abundance value less than a certain value is associated with the cerebro-craniofacial health issue in question and above the certain value is scored as associated with healthy, or vice versa depending on the particular criterion. The scoring for various bacteria or genetic pathways can be combined to provide a classification for a subject. Furthermore, in the examples, the comparison of an abundance value to one or more reference abundance values can include a comparison to a cutoff value determined from the one or more reference values. Such cutoff value(s) can be part of a decision tree or a clustering technique (where a cutoff value is used to determine which cluster the abundance value(s) belong) that are determined using the reference abundance values. The comparison can include intermediate determination of other values, (e.g., probability values). The comparison can also include a comparison of an abundance value to a probability distribution of the reference abundance values, and thus a comparison to probability values.

[0219] A disease signature can include more sequence groups than are used for a given subject. As an example, the disease signature can include 100 sequence groups, but only 60 of sequence groups may be detected in a sample, or detected above a threshold cutoff. The classification of the subject (including any probability for having or lacking a disease such as a cerebro-craniofacial health issue) can be determined based on the 60 sequence groups.

[0220] In relation to generation of the characterization model, the sequence groups with high discrimination levels (e.g., low p-values) for a given disease can be identified and used as part of a characterization model, e.g., which uses a disease signature to determine a probability of a subject having a cerebro-craniofacial health issue. The disease signature can include a set of sequence groups as well as discriminating criteria (e.g., cutoff values and/or probability distributions) used to provide a classification of the subject. The classification can be binary (e.g., disease or control) or have more classifications (e.g., probability values for having the disease of a cerebro-craniofacial health issue, or not having the disease). Which sequence groups of the disease signature that are used in making a classification be dependent on the specific sequence reads obtained, e.g., a sequence group would not be used if no sequence reads were assigned to that sequence group. In some embodiments, a separate characterization model can be determined for different populations, e.g., by geography where

the subject is currently residing (e.g., country, r
subject (e.g., ethnicity), or other factors.

6. Selection of Sequence Groups, Discrimination Criteria for Sequence Groups, and Use of Sequence Groups

[0221] As shown in FIG. 4, in one embodiment of Block S140, the characterization process can be generated and trained according to a random forest predictor (RFP) algorithm that combines bagging (i.e., bootstrap aggregation) and selection of random sets of features from a training dataset to construct a set of decision trees, T, associated with the random sets of features. In using a random forest algorithm, N cases from the set of decision trees are sampled at random with replacement to create a subset of decision trees, and for each node, m prediction features are selected from all of the prediction features for assessment. The prediction feature that provides the best split at the node (e.g., according to an objective function) is used to perform the split (e.g., as a bifurcation at the node, as a trifurcation at the node). By sampling many times from a large dataset, the strength of the characterization process, in identifying features that are strong in predicting classifications can be increased substantially. In this variation, measures to prevent bias (e.g., sampling bias) and/or account for an amount of bias can be included during processing to increase robustness of the model.

[0222] In one implementation, a characterization process of Block S140 based upon statistical analyses can identify the sets of features that have the highest correlations with a cerebro-craniofacial health issue, for which one or more therapies would have a positive effect, based upon an algorithm trained and validated with a validation dataset derived from a subset of the population of subjects. In particular, a cerebro-craniofacial health issue in this first variation is characterized by an alteration of the microbiome that is predictive of the presence or absence of insomnia, light sleep, headache, sinusitis, or poor concentration.

[0223] In one variation, a set of features useful for diagnostics associated with cerebro-craniofacial disorders includes features derived from one or more of the taxa of TABLEs A, B, C, D, or E (e.g., one or more of the family, order, class, and/or phylum of TABLE A, or the species of TABLE B) and/or one or more of the functional groups of TABLE B (e.g., one or more of the KEGG level 2 (KEGG L2) functional groups and/or one or more of the KEGG level 3 (KEGG L3) functional groups of TABLE B). One skilled in the art will appreciate other combinations of sequence groups from various tables.

7. Therapy Models

[0224] In some embodiments, as noted above, outputs of the first method 100 can be used to generate diagnostics and/or provide therapeutic measures for an individual based upon an analysis of the individual's microbiome. As such, a second method 200 derived from at least one output of the first method 100 can include: receiving a biological sample from a subject S210; characterizing the subject with a form of a cerebro-craniofacial health issue based upon the characterization and the therapy model S230.

[0225] Block S210 recites: receiving a biological sample from the subject, which functions to facilitate generation of a microbiome composition dataset and/or a microbiome functional diversity dataset for the subject. As such, processing and analyzing the biological sample preferably facilitates generation of a microbiome composition dataset and/or a microbiome functional diversity dataset for the subject, which can be used to provide inputs that can be used to characterize the individual in relation to diagnosis of the cerebro-craniofacial health issue, as in Block S220. Receiving a biological sample from the subject is preferably performed in a manner similar to that of one of the embodiments, variations, and/or examples of sample reception described in relation to Block S110 above. As such, reception and processing of the biological sample in Block S210 can be performed for the subject using similar processes as those for receiving and processing biological samples used to generate the characterization(s) and/or the therapy provision model of the first method 100, in order to provide consistency of process. However, biological sample reception and processing in Block S210 can alternatively be performed in any other suitable manner.

[0226] Block S220 recites: characterizing the subject characterizing the subject with a form of a disease or condition based upon processing a microbiome dataset derived from the biological sample. Block S220 functions to extract features from microbiome-derived data of the subject, and use the features to positively or negatively characterize the individual as having a form of the cerebro-craniofacial health issue. Characterizing the subject in Block S220 thus preferably includes identifying features and/or combinations of features associated with the microbiome composition and/or functional features of the microbiome of the subject, and comparing such features with features characteristic of subjects with the cerebro-craniofacial health issue. Block S220 can further include generation of and/or output of a confidence metric associated with the characterization for the individual. For instance, a confidence metric can be derived from the number of features used to generate the classification, relative weights or rankings of features used to generate the characterization,

measures of bias in the models used in Block S

parameter associated with aspects of the characterization operation of Block S140.

[0227] In some variations, features extracted from the microbiome dataset can be supplemented with survey-derived and/or medical history-derived features from the individual, which can be used to further refine the characterization operation(s) of Block S220. However, the microbiome composition dataset and/or the microbiome functional diversity dataset of the individual can additionally or alternatively be used in any other suitable manner to enhance the first method 100 and/or the second method 200.

[0228] Block S230 recites: promoting a therapy to the subject with the disease or condition based upon the characterization and the therapy model. Block S230 functions to recommend or provide a personalized therapeutic measure to the subject, in order to shift the microbiome composition of the individual toward a desired equilibrium state. As such, Block S230 can include correcting the cerebro-craniofacial health issue, or otherwise positively affecting the user's health in relation to the cerebro-craniofacial health issue. Block S230 can thus include promoting one or more therapeutic measures to the subject based upon their characterization in relation to the cerebro-craniofacial health issue, as described herein, wherein the therapy is configured to modulate taxonomic makeup of the subject's microbiome and/or modulate functional feature aspects of the subject in a desired manner toward a "normal" or "control" state in relation to the characterizations described above.

[0229] In Block S230, providing the therapeutic measure to the subject can include recommendation of available therapeutic measures configured to modulate microbiome composition of the subject toward a desired state (e.g., having a microbiome that is not indicative of (e.g., altered by) a cerebro-craniofacial health issue). Additionally or alternatively, Block S230 can include provision of customized therapy to the subject according to their characterization (e.g., in relation to a specific type of a cerebro-craniofacial health issue, such as insomnia, light sleep, headache, sinusitis, or poor concentration). In variations, therapeutic measures for adjusting a microbiome composition of the subject, in order to improve a state of the cerebro-craniofacial health issue can include one or more of: probiotics, prebiotics, bacteriophage-based therapies, consumables, suggested activities, topical therapies, adjustments to hygienic product usage, adjustments to diet, adjustments to sleep behavior, living arrangement, adjustments to level of sexual activity, nutritional supplements, medications, antibiotics, and any other suitable therapeutic measure. Therapy provision in Block S230 can include provision of notifications by way of an electronic device, through an entity associated with the individual, and/or in any other suitable manner.

[0230] In more detail, therapy provision in B
notifications to the subject regarding recommended therapeutic measures and/or other courses of action, in relation to health-related goals, as shown in FIG. 6. Notifications can be provided to an individual by way of an electronic device (e.g., personal computer, mobile device, tablet, head-mounted wearable computing device, wrist-mounted wearable computing device, etc.) that executes an application, web interface, and/or messaging client configured for notification provision. In one example, a web interface of a personal computer or laptop associated with a subject can provide access, by the subject, to a user account of the subject, wherein the user account includes information regarding the subject's characterization, detailed characterization of aspects of the subject's microbiome composition and/or functional features, and notifications regarding suggested therapeutic measures generated in Block S150. In another example, an application executing at a personal electronic device (e.g., smart phone, smart watch, head-mounted smart device) can be configured to provide notifications (e.g., at a display, haptically, in an auditory manner, etc.) regarding therapeutic suggestions generated by the therapy model of Block S150. Notifications can additionally or alternatively be provided directly through an entity associated with a subject (e.g., a caretaker, a spouse, a significant other, a healthcare professional, etc.). In some further variations, notifications can additionally or alternatively be provided to an entity (e.g., healthcare professional) associated with the subject, wherein the entity is able to administer the therapeutic measure (e.g., by way of prescription, by way of conducting a therapeutic session, etc.). Notifications can, however, be provided for therapy administration to the subject in any other suitable manner.

[0231] Furthermore, in an extension of Block S230, monitoring of the subject during the course of a therapeutic regimen (e.g., by receiving and analyzing biological samples from the subject throughout therapy, by receiving survey-derived data from the subject throughout therapy) can be used to generate a therapy-effectiveness model for each recommended therapeutic measure provided according to the model generated in Block S150.

[0232] As shown in FIG. 1E, in some variations, the first method 100, or any of the methods described herein (*e.g.*, as in any one or more of FIGs 1A-1F) can further include Block S150, which recites: based upon the characterization model, generating a therapy model configured to correct or otherwise improve a state of the disease or condition. Block S150 functions to identify or predict therapies (e.g., probiotic-based therapies, prebiotic-based therapies, phage-based therapies, small molecule-based therapies (*e.g.*, selective, pan-selective, or non-selective antibiotics), etc.) that can shift a subject's microbiome composition

and/or functional features toward a desired equ
health (*e.g.*, toward a microbiome that is not indicative of a cerebro-craniofacial health issue, or to correct or otherwise improve a state or symptom of a cerebro-craniofacial health issue). In Block S150, the therapies can be selected from therapies including one or more of: probiotic therapies, phage-based therapies, prebiotic therapies, small molecule-based therapies, cognitive/behavioral therapies, physical rehabilitation therapies, clinical therapies, medication-based therapies, diet-related therapies, and/or any other suitable therapy designed to operate in any other suitable manner in promoting a user's health. In a specific example of a bacteriophage-based therapy, one or more populations (e.g., in terms of colony forming units) of bacteriophages specific to a certain bacteria (or other microorganism) represented in a subject with the cerebro-craniofacial health issue can be used to down-regulate or otherwise eliminate populations of the certain bacteria. As such, bacteriophage-based therapies can be used to reduce the size(s) of the undesired population(s) of bacteria represented in the subject. Complementarily, bacteriophage-based therapies can be used to increase the relative abundances of bacterial populations not targeted by the bacteriophage(s) used.

[0233] For instance, in relation to the variations of cerebro-craniofacial health issues described herein, therapies (e.g., probiotic therapies, bacteriophage-based therapies, prebiotic therapies, etc.) can be configured to downregulate and/or upregulate microorganism populations or subpopulations (and/or functions thereof) associated with features characteristic of the cerebro-craniofacial health issue.

[0234] In one such variation, the Block S150 can include one or more of the following steps: obtaining a sample from the subject; purifying nucleic acids (e.g., DNA) from the sample; deep sequencing nucleic acids from the sample so as to determine the amount of one or more of the features of TABLEs A, B, C, D, or E ; and comparing the resulting amount of each feature to one or more reference amounts of the one or more of the features listed in one or more of TABLEs A, B, C, D, or E as occurs in an average individual having a cerebro-craniofacial health issue or an individual not having the cerebro-craniofacial health issue or both. The compilation of features can sometimes be referred to as a "disease signature" for a specific condition related to a cerebro-craniofacial health issue. The disease signature can act as a characterization model, and may include probability distributions for control population (no cerebro-craniofacial health issue) or disease populations having the condition or both. The disease signature can include one or more of the features (e.g., bacterial taxa or genetic pathways) listed and can optionally include criteria determined from abundance

values of the control and/or disease populations

probability values for amounts of those features associated with average control or disease (e.g., insomnia, light sleep, headache, sinusitis, or poor concentration) individuals.

[0235] In a specific example of probiotic therapies, as shown in FIG. 5, candidate therapies of the therapy model can perform one or more of: blocking pathogen entry into an epithelial cell by providing a physical barrier (e.g., by way of colonization resistance), inducing formation of a mucous barrier by stimulation of goblet cells, enhance integrity of apical tight junctions between epithelial cells of a subject (e.g., by stimulating up regulation of zona-occludens 1, by preventing tight junction protein redistribution), producing antimicrobial factors, stimulating production of anti-inflammatory cytokines (e.g., by signaling of dendritic cells and induction of regulatory T-cells), triggering an immune response, and performing any other suitable function that adjusts a subject's microbiome away from a state of dysbiosis.

[0236] In variations, the therapy model is preferably based upon data from a large population of subjects, which can comprise the population of subjects from which the microbiome-related datasets are derived in Block S110, wherein microbiome composition and/or functional features or states of health, prior exposure to and post exposure to a variety of therapeutic measures, are well characterized. Such data can be used to train and validate the therapy provision model, in identifying therapeutic measures that provide desired outcomes for subjects based upon different microbiome characterizations. In variations, support vector machines, as a supervised machine learning algorithm, can be used to generate the therapy provision model. However, any other suitable machine learning algorithm described above can facilitate generation of the therapy provision model.

[0237] While some methods of statistical analyses and machine learning are described in relation to performance of the Blocks above, variations of the method 100, or any one of Figs 1A-1F, can additionally or alternatively utilize any other suitable algorithms in performing the characterization process. In variations, the algorithm(s) can be characterized by a learning style including any one or more of: supervised learning (e.g., using logistic regression, using back propagation neural networks), unsupervised learning (e.g., using an Apriori algorithm, using K-means clustering), semi-supervised learning, reinforcement learning (e.g., using a Q-learning algorithm, using temporal difference learning), and any other suitable learning style. Furthermore, the algorithm(s) can implement any one or more of: a regression algorithm (e.g., ordinary least squares, logistic regression, stepwise regression, multivariate adaptive regression splines, locally estimated scatterplot smoothing,

etc.), an instance-based method (e.g., k-nearest organizing map, etc.), a regularization method (e.g., ridge regression, least absolute shrinkage and selection operator, elastic net, etc.), a decision tree learning method (e.g., classification and regression tree, iterative dichotomiser 3, C4.5, chi-squared automatic interaction detection, decision stump, random forest, multivariate adaptive regression splines, gradient boosting machines, etc.), a Bayesian method (e.g., naïve Bayes, averaged one-dependence estimators, Bayesian belief network, etc.), a kernel method (e.g., a support vector machine, a radial basis function, a linear discriminant analysis, etc.), a clustering method (e.g., k-means clustering, expectation maximization, etc.), an associated rule learning algorithm (e.g., an Apriori algorithm, an Eclat algorithm, etc.), an artificial neural network model (e.g., a Perceptron method, a back-propagation method, a Hopfield network method, a self-organizing map method, a learning vector quantization method, etc.), a deep learning algorithm (e.g., a restricted Boltzmann machine, a deep belief network method, a convolutional network method, a stacked autoencoder method, etc.), a dimensionality reduction method (e.g., principal component analysis, partial least squares regression, Sammon mapping, multidimensional scaling, projection pursuit, etc.), an ensemble method (e.g., boosting, bootstrapped aggregation, AdaBoost, stacked generalization, gradient boosting machine method, random forest method, etc.), and any suitable form of algorithm.

[0238] Additionally or alternatively, the therapy model can be derived in relation to identification of a "normal" or baseline microbiome composition and/or functional features, as assessed from subjects of a population of subjects who are identified to be in good health. Upon identification of a subset of subjects of the population of subjects who are characterized to be in good health (e.g., characterized as not having an altered microbiome caused by, or indicative of, a cerebro-craniofacial health issue, e.g., using features of the characterization process), therapies that modulate microbiome compositions and/or functional features toward those of subjects in good health can be generated in Block S150. Block S150 can thus include identification of one or more baseline microbiome compositions and/or functional features (e.g., one baseline microbiome for each of a set of demographics), and potential therapy formulations and therapy regimens that can shift microbiomes of subjects who are in a state of dysbiosis toward one of the identified baseline microbiome compositions and/or functional features. The therapy model can, however, be generated and/or refined in any other suitable manner.

[0239] Microorganism compositions associated with probiotic therapies associated with the therapy model preferably include microorganisms that are culturable (e.g., able to be

expanded to provide a scalable therapy) and no therapeutic dosages). Furthermore, microorganism compositions can comprise a single type of microorganism that has an acute or moderated effect upon a subject's microbiome. Additionally or alternatively, microorganism compositions can comprise balanced combinations of multiple types of microorganisms that are configured to cooperate with each other in driving a subject's microbiome toward a desired state. For instance, a combination of multiple types of bacteria in a probiotic therapy can comprise a first bacteria type that generates products that are used by a second bacteria type that has a strong effect in positively affecting a subject's microbiome. Additionally or alternatively, a combination of multiple types of bacteria in a probiotic therapy, e.g., can comprise several bacteria types that produce proteins with the same functions that positively affect a subject's microbiome.

[0240] In examples of probiotic therapies, probiotic compositions can comprise components of one or more of the identified taxa of microorganisms (e.g., as described in TABLEs A, B, C, D, or E) provided at dosages of 1 million to 10 billion CFUs, as determined from a therapy model that predicts positive adjustment of a subject's microbiome in response to the therapy. Additionally or alternatively, the therapy can comprise dosages of proteins resulting from functional presence in the microbiome compositions of subjects without the cerebro-craniofacial health issue. In the examples, a subject can be instructed to ingest capsules comprising the probiotic formulation according to a regimen tailored to one or more of his/her: physiology (e.g., body mass index, weight, height), demographics (e.g., gender, age), severity of dysbiosis, sensitivity to medications, and any other suitable factor.

[0241] Furthermore, probiotic compositions of probiotic-based therapies can be naturally or synthetically derived. For instance, in one application, a probiotic composition can be naturally derived from fecal matter or other biological matter (e.g., of one or more subjects having a baseline microbiome composition and/or functional features, as identified using the characterization process and the therapy model). Additionally or alternatively, probiotic compositions can be synthetically derived (e.g., derived using a benchtop method) based upon a baseline microbiome composition and/or functional features, as identified using the characterization process and the therapy model. In one embodiment, the probiotic composition is or is derived from the subject's own fecal matter that has been stored or "banked" from a period during which the subject is in a healthy state for use when the microbiome is imbalanced (e.g., due to antibiotic usage, or due to a cerebro-craniofacial health issue).

[0242] In variations, microorganism agents th include one or more of: yeast (e.g., *Saccharomyces boulardii*), gram-negative bacteria (e.g., *E. coli Nissle, Akkermansia muciniphila, Prevotella bryantii, etc.*), gram-positive bacteria (e.g., *Bifidobacterium animalis* (including subspecies *lactis*), *Bifidobacterium longum* (including subspecies *infantis*), *Bifidobacterium bifidum, Bifidobacterium pseudolongum, Bifidobacterium thermophilum, Bifidobacterium breve, Lactobacillus rhamnosus, Lactobacillus acidophilus, Lactobacillus casei, Lactobacillus helveticus, Lactobacillus plantarum, Lactobacillus fermentum, Lactobacillus salivarius, Lactobacillus delbrueckii* (including subspecies *bulgaricus*), *Lactobacillus johnsonii, Lactobacillus reuteri, Lactobacillus gasseri, Lactobacillus brevis* (including subspecies *coagulans*), *Bacillus cereus, Bacillus subtilis* (including var. *Natto*), *Bacillus polyfermenticus, Bacillus clausii, Bacillus licheniformis, Bacillus coagulans, Bacillus pumilus, Faecalibacterium prausnitzii, Streptococcus thermophilus, Brevibacillus brevis, Lactococcus lactis, Leuconostoc mesenteroides, Enterococcus faecium, Enterococcus faecalis, Enterococcus durans, Clostridium butyricum, Sporolactobacillus inulinus, Sporolactobacillus vineae, Pediococcus acidilactici, Pediococcus pentosaceus,* etc.), and any other suitable type of microorganism agent.

[0243] Additionally or alternatively, therapies promoted by the therapy model of Block S150 can include one or more of: consumables (e.g., food items, beverage items, nutritional supplements), suggested activities (e.g., exercise regimens, adjustments to alcohol consumption, adjustments to cigarette usage, adjustments to drug usage), topical therapies (e.g., lotions, ointments, antiseptics, etc.), adjustments to hygienic product usage (e.g., use of shampoo products, use of conditioner products, use of soaps, use of makeup products, etc.), adjustments to diet (e.g., sugar consumption, fat consumption, salt consumption, acid consumption, etc.), adjustments to sleep behavior, living arrangement adjustments (e.g., adjustments to living with pets, adjustments to living with plants in one's home environment, adjustments to light and temperature in one's home environment, etc.), nutritional supplements (e.g., vitamins, minerals, fiber, fatty acids, amino acids, prebiotics, probiotics, etc.), medications, antibiotics, and any other suitable therapeutic measure. Among the prebiotics suitable for treatment, as either part of any food or as supplement, are included the following components: 1,4-dihydroxy-2-naphthoic acid (DHNA), Inulin, trans-Galactooligosaccharides (GOS), Lactulose, Mannan oligosaccharides (MOS), Fructooligosaccharides (FOS), Neoagaro-oligosaccharides (NAOS), Pyrodextrins, Xylo-oligosaccharides (XOS), Isomalto-oligosaccharides (IMOS), Amylose-resistant starch, Soybean oligosaccharides (SBOS), Lactitol, Lactosucrose (LS), Isomaltulose (including

Palatinose), Arabinoxylooligosaccharides (AXO
Arabinoxylans (AX), Polyphenols or any other compound capable of changing the microbiota composition with a desirable effect.

[0244] Additionally or alternatively, therapies promoted by the therapy model of Block S150 can include one or more of: different forms of therapy having different therapy orientations (e.g., motivational, increase energy level, reduce weight gain, improve diet, psychoeducational, cognitive behavioral, biological, physical, mindfulness-related, relaxation-related, dialectical behavioral, acceptance-related, commitment-related, etc.) configured to address a variety of factors contributing to an adverse states due to a microbiome that is altered by a cerebro-craniofacial health issue or a microbiome that is caused by or indicative of a cerebro-craniofacial health issue; weight management interventions (e.g., to prevent adverse weight-related (*e.g.*, weight gain or loss) side effects due to insomnia, light sleep, headache, sinusitis, or poor concentration, or a therapy to prevent, mitigate, or reduce the frequency or severity of insomnia, light sleep, headache, sinusitis, or poor concentration); physical therapy; rehabilitation measures; and any other suitable therapeutic measure.

[0245] The first method 100 can, however, include any other suitable blocks or steps configured to facilitate reception of biological samples from individuals, processing of biological samples from individuals, analyzing data derived from biological samples, and generating models that can be used to provide customized diagnostics and/or therapeutics according to specific microbiome compositions of individuals.

[0246] The methods 100, 200 and/or system of the embodiments can be embodied and/or implemented at least in part as a machine configured to receive a computer-readable medium storing computer-readable instructions. The instructions can be executed by computer-executable components integrated with the application, applet, host, server, network, website, communication service, communication interface, hardware/firmware/software elements of a patient computer or mobile device, or any suitable combination thereof. Other systems and methods of the embodiments can be embodied and/or implemented at least in part as a machine configured to receive a computer-readable medium storing computer-readable instructions. The instructions can be executed by computer-executable components integrated with apparatuses and networks of the type described above. The computer-readable medium can be stored on any suitable computer readable media such as RAMs, ROMs, flash memory, EEPROMs, optical devices (CD or DVD), hard drives, floppy drives, or any suitable device.

The computer-executable component can be a p
hardware device can (alternatively or additionally) execute the instructions.

[0247] The FIGs illustrate the architecture, functionality and operation of possible implementations of systems, methods and computer program products according to preferred embodiments, example configurations, and variations thereof. In this regard, each block in the flowchart or block diagrams may represent a module, segment, step, or portion of code, which comprises one or more executable instructions for implementing the specified logical function(s). It should also be noted that, in some alternative implementations, the functions noted in the block can occur out of the order noted in the Figs. For example, two blocks shown in succession may, in fact, be executed substantially concurrently, or the blocks may sometimes be executed in the reverse order, depending upon the functionality involved. It will also be noted that each block of the block diagrams and/or flowchart illustration, and combinations of blocks in the block diagrams and/or flowchart illustration, can be implemented by special purpose hardware-based systems that perform the specified functions or acts, or combinations of special purpose hardware and computer instructions.

VI. EXAMPLES FOR CEREBRO-CRANIOFACIAL HEALTH

A. Example for Insomnia

[0248] Some examples of sequence groups, discriminating levels, coverage percentages, and discriminating criteria are provided in TABLE A.

[0249] TABLE A shows data for insomnia. The data was obtained from 901 subjects in the condition population and 4865 subjects in the control population. TABLE A shows taxonomic groups for Species, Genus, and Family all in the first column of TABLE A. Each of the rows containing data corresponds to a different sequence group. For example, Parabacteroides distasonis corresponds to a sequence group in the Species level of the taxonomic hierarchy.

[0250] A level can have many sequence groups. The number "823" after "Parabacteroides distasonis" is the NCBI taxonomy ID for that taxonomic group. The IDs correspond to those at www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=200643. The p-values are determined via either the Kolmogorov-Smirnov test, or the Welch's t-test.

[0251] Sequence groups having a p-value less than 0.01 are shown in the second column. Other sequence groups may exist, but likely would not be selected for inclusion into a disease signature. The third column ("# disease subjects detected") shows the number of samples

tested that had the condition of insomnia and w

sequence group. The fourth column ("# control subjects detected") shows the number of samples tested that did not have the disease (control) and where the sample exhibited bacteria in the sequence group. The coverage percentage of the sequence group can be determined from the values in the third and fourth columns.

[0252] The fifth column shows the mean percentage for the abundance for the subjects having the disease and where the sample exhibited bacteria in the sequence group. The sixth column shows the mean percentage for the abundance for the subjects not having the disease and where the sample exhibited bacteria in the sequence group. As one can see, the sequence groups with the largest percentage difference between the two means have the smallest p-value, signifying a greater separation between the two populations.

[0253] A set of sequence groups (taxonomic and/or functional) can be selected from TABLE A for forming a disease signature that can be used to classify a sample regarding a presence or absence of a microbiome indicative of a insomnia issue. For example, all taxonomic sequence groups can be selected, or just the 2, 3, 4, 5, or 6 ones with the smallest p-value, as may include the function groups as well. The sequence groups for the disease signature can be selected to optimize accuracy for discriminating between the two groups and coverage of the population such that a likelihood of being able to provide a classification is higher (e.g., if a sequence group is not present then that sequence group cannot be used to determine the classification). The total coverage can dependent on the individual coverage percentages and based on the overlap in the coverages among the sequence groups, as described above.

B. Example for Light sleep

[0254] Some examples of sequence groups, discriminating levels, coverage percentages, and discriminating criteria are provided in TABLE B.

[0255] TABLE B shows data for light sleep. 627 subjects are in the condition population and 4471 subjects are in the control population. TABLE B shows the taxonomic group for Species, Genus, and Family and shows functional groups all in the first column of TABLE B. As mentioned above, the functional groups correspond to one or more genes with the function. Each of the rows containing data corresponds to a different sequence group.

[0256] A set of sequence groups (taxonomic and/or functional) can be selected from TABLE B for forming a disease signature that can be used to classify a sample regarding a

presence or absence of a microbiome indicative
other number) sequence groups can be selected, e.g., with the smallest p-value. The sequence groups for the disease signature can be selected to optimize accuracy for discriminating between the two groups and coverage of the population such that a likelihood of being able to provide a classification is higher (e.g., if a sequence group is not present then that sequence group cannot be used to determine the classification). The total coverage can dependent on the individual coverage percentages and based on the overlap in the coverages among the sequence groups, as described above.

C. Example for Headache

[0257] Some examples of sequence groups, discriminating levels, coverage percentages, and discriminating criteria are provided in TABLE C.

[0258] TABLE C shows data for headache. 795 subjects are in the condition population and 4349 subjects are in the control population. TABLE C shows the taxonomic group for Species, Genus, and Family and shows functional groups all in the first column of TABLE C. As mentioned above, the functional groups correspond to one or more genes with the function. Each of the rows containing data corresponds to a different sequence group.

[0259] A set of sequence groups (taxonomic and/or functional) can be selected from TABLE C for forming a disease signature that can be used to classify a sample regarding a presence or absence of a microbiome indicative of a headache issue. For example, 6 (or other number) sequence groups can be selected, e.g., with the smallest p-value. The sequence groups for the disease signature can be selected to optimize accuracy for discriminating between the two groups and coverage of the population such that a likelihood of being able to provide a classification is higher (e.g., if a sequence group is not present then that sequence group cannot be used to determine the classification). The total coverage can dependent on the individual coverage percentages and based on the overlap in the coverages among the sequence groups, as described above.

D. Example for Sinusitis

[0260] Some examples of sequence groups, discriminating levels, coverage percentages, and discriminating criteria are provided in TABLE D.

[0261] TABLE D shows data for sinusitis. 218 subjects are in the condition population and 1049 subjects are in the control population. TABLE D shows the taxonomic group for Species, Genus, and Family and shows functional groups all in the first column of TABLE D.

As mentioned above, the functional groups corı
function. Each of the rows containing data corresponds to a different sequence group.

[0262] A set of sequence groups (taxonomic and/or functional) can be selected from TABLE D for forming a disease signature that can be used to classify a sample regarding a presence or absence of a microbiome indicative of a sinusitis issue. For example, 6 (or other number) sequence groups can be selected, e.g., with the smallest p-value. The sequence groups for the disease signature can be selected to optimize accuracy for discriminating between the two groups and coverage of the population such that a likelihood of being able to provide a classification is higher (e.g., if a sequence group is not present then that sequence group cannot be used to determine the classification). The total coverage can dependent on the individual coverage percentages and based on the overlap in the coverages among the sequence groups, as described above.

E. Example for Poor Concentration

[0263] Some examples of sequence groups, discriminating levels, coverage percentages, and discriminating criteria are provided in TABLE E.

[0264] TABLE E shows data for poor concentration. 1396 subjects are in the condition population and 6276 subjects are in the control population. TABLE E shows the taxonomic group for Species, Genus, and Family and shows functional groups all in the first column of TABLE E. As mentioned above, the functional groups correspond to one or more genes with the function. Each of the rows containing data corresponds to a different sequence group.

[0265] A set of sequence groups (taxonomic and/or functional) can be selected from TABLE E for forming a disease signature that can be used to classify a sample regarding a presence or absence of a microbiome indicative of a light sleep issue. For example, 6 (or other number) sequence groups can be selected, e.g., with the smallest p-value. The sequence groups for the disease signature can be selected to optimize accuracy for discriminating between the two groups and coverage of the population such that a likelihood of being able to provide a classification is higher (e.g., if a sequence group is not present then that sequence group cannot be used to determine the classification). The total coverage can dependent on the individual coverage percentages and based on the overlap in the coverages among the sequence groups, as described above.

[0266] Although the foregoing invention has been described in some detail by way of illustration and example for purposes of clarity of understanding, one of skill in the art will

appreciate that certain changes and modificatio

appended claims. In addition, each reference provided herein is incorporated by reference in its entirety to the same extent as if each reference was individually incorporated by reference. Where a conflict exists between the instant application and a reference provided herein, the instant application shall dominate.

WHAT IS CLAIMED IS:

1. A method of determining a classification of occurrence of a microbiome indicative of a cerebro-craniofacial health issue or screening for the presence or absence of a microbiome indicative of a cerebro-craniofacial health issue in an individual and/or determining a course of treatment for an individual human having a microbiome indicative of a cerebro-craniofacial health issue, the method comprising,

providing a sample comprising bacteria (or at least one of the following microorganisms including: bacteria, archaea, unicellular eukaryotic organisms and viruses, or the combinations thereof) from the individual human;

determining an amount(s) of one or more of the following in the sample:

bacteria taxon or gene sequence corresponding to gene functionality as set forth in TABLEs A, B, C, D, or E ;

comparing the determined amount(s) to a disease signature having cut-off or probability values for amounts of the bacteria taxon and/or gene sequence for an individual having a microbiome indicative of a cerebro-craniofacial health issue or an individual not having a microbiome indicative of a cerebro-craniofacial health issue or both; and

determining a classification of the presence or absence of the microbiome indicative of a cerebro-craniofacial health issue and/or determining the course of treatment for the individual human having the microbiome indicative of a cerebro-craniofacial health issue based on the comparing.

2. The method of claim 1, wherein the cerebro-craniofacial health issue is:

(i) insomnia and the bacteria taxa or gene sequences are selected from those in TABLE A;

(ii) light sleep and the bacteria taxa or gene sequences are selected from those in TABLE B;

(iii) headache and the bacteria taxa or gene sequences are selected from those in TABLE C;

(iv) sinusitis and the bacteria taxa or gene sequences are selected from those in TABLE D; or

(v) poor concentration and the bacteria taxa or gene sequences are selected from those in TABLE E.

3. The method of claim 1, wherein the determining comprises preparing DNA from the sample and performing nucleotide sequencing of the DNA.

4. The method of any of claims 1-3, wherein the determining comprises deep sequencing bacterial DNA from the sample to generate sequencing reads,

receiving at a computer system the sequencing reads; and

mapping, with the computer system, the reads to bacterial genomes to determine whether the reads map to a sequence from the bacterial taxon or a gene sequence from TABLEs A, B, C, D, or E ; and

determining a relative amount of different sequences in the sample that correspond to a sequence from the bacteria taxon or gene sequence corresponding to gene functionality from TABLEs A, B, C, D, or E .

5. The method of claim 4, wherein the deep sequencing is random deep sequencing.

6. The method of claim 4, wherein the deep sequencing comprises deep sequencing of 16S rRNA coding sequences.

7. The method of any of claims 1-6, wherein the method further comprises obtaining physiological, demographic or behavioral information from the individual human, wherein the disease signature comprises physiological, demographic or behavioral information; and

the determining comprises comparing the obtained physiological, demographic or behavioral information to corresponding information in the disease signature.

8. The method of any of claims 1-7, wherein the sample includes at least one of the following: a fecal, blood, saliva, cheek swab, urine, or bodily fluid from the individual human

9. The method of any of claims 1-8, further comprising determining that the individual human likely has a microbiome indicative of a cerebro-craniofacial health issue; and

treating the individual human to ameliorate at least one symptom of the microbiome indicative of the cerebro-craniofacial health issue.

10. The method of claim 9, wherein the treating comprises administering a dose of one or more of the bacteria taxon listed in TABLEs A, B, C, D, or E to the individual human for which the individual human is deficient.

11. A method for determining a classification of the presence or absence of a microbiome indicative of a cerebro-craniofacial health issue and/or determine a course of treatment for an individual human having a microbiome indicative of a cerebro-craniofacial health issue, the method comprising performing, by a computer system:

receiving sequence reads of bacterial DNA obtained from analyzing a test sample from the individual human;

mapping the sequence reads to a bacterial sequence database to obtain a plurality of mapped sequence reads, the bacterial sequence database including a plurality of reference sequences of a plurality of bacteria;

assigning the mapped sequence reads to sequence groups based on the mapping to obtain assigned sequence reads assigned to at least one sequence group, wherein a sequence group includes one or more of the plurality of reference sequences;

determining a total number of assigned sequence reads;

for each sequence group of a disease signature set of one or more sequence groups selected from TABLEs A, B, C, D, or E :

determining a relative abundance value of assigned sequence reads assigned to the sequence group relative to the total number of assigned sequence reads, the relative abundance values forming a test feature vector;

comparing the test feature vector to calibration feature vectors generated from relative abundance values of calibration samples having a known status of cerebro-craniofacial health ; and

determining the classification of the presence or absence of the microbiome indicative of a cerebro-craniofacial health issue and/or determining the course of treatment for the individual human having the microbiome indicative of a cerebro-craniofacial health issue based on the comparing.

12. The method of claim 11, wherein the comparing includes:

clustering the calibration feature vectors into a control cluster not having the microbiome indicative of a cerebro-craniofacial health issue and a disease cluster having the microbiome indicative of a cerebro-craniofacial health issue; and

determining which cluster the test feature vector belongs.

13. The method of claim 12, wherein the clustering includes using a Bray–Curtis dissimilarity.

14. The method of claim 11, wherein the comparing includes comparing each of the relative abundance values of the test feature vector to a respective cutoff value determined from the calibration feature vectors generated from the calibration samples.

15. The method of claim 11, wherein the comparing includes:

comparing a first relative abundance value of the test feature vector to a disease probability distribution to obtain a disease probability for the individual human having a microbiome indicative of a cerebro-craniofacial health issue, the disease probability distribution determined from a plurality of samples having the microbiome indicative of the cerebro-craniofacial health issue and exhibiting the sequence group;

comparing the first relative abundance value to a control probability distribution to obtain a control probability for the individual human not having a microbiome indicative of a cerebro-craniofacial health issue, wherein the disease probabilities and the control probabilities are used to determine the classification of the presence or absence of the microbiome indicative of a cerebro-craniofacial health issue and/or determining the course of treatment for the individual human having the microbiome indicative of a cerebro-craniofacial health issue.

16. The method of claim 11, wherein the sequence reads are mapped to one or more predetermined regions of the reference sequences.

17. The method of claim 11, wherein the disease signature set includes at least one taxonomic group and at least one functional group.

18. The method of claim 11, wherein the cerebro-craniofacial health issue is:

(i) insomnia and the sequence groups are selected from those in TABLE A;

(ii) light sleep and the sequence groups are selected from those in TABLE B;

(iii) headache and the sequence groups are selected from those in TABLE C;

(iv) sinusitis and the sequence groups are selected from those in TABLE D;

(v) poor concentration and tl

TABLE E.

19. The method of claim 11, wherein the analyzing comprises deep sequencing.

20. The method of claim 19, wherein the deep sequencing reads are random deep sequencing reads.

21. The method of claim 19, wherein the deep sequencing reads comprise 16S rRNA deep sequencing reads.

22. The method of any of claims 11-21, further comprising:

receiving physiological, demographic or behavioral information from the individual human; and

using the physiological, demographic or behavioral information in combination with the classification with the comparing of the test feature vector to the calibration feature vectors to determine the classification of the presence or absence of the microbiome indicative of a cerebro-craniofacial health issue and/or determining the course of treatment for the individual human having the microbiome indicative of a cerebro-craniofacial health issue.

23. The method of claim 11, further comprising preparing DNA from the sample and performing nucleotide sequencing of the DNA.

24. A non-transitory computer readable medium storing a plurality of instructions that when executed, by the computer system, perform the method of any one of claims 11-22.

25. A method for at least one of characterizing, diagnosing, and treating a cerebro-craniofacial health issue in at least a subject, the method comprising:

- at a sample handling network, receiving an aggregate set of samples from a population of subjects;
- at a computing system in communication with the sample handling network, generating a microbiome composition dataset and a microbiome functional diversity dataset for the population of subjects upon processing nucleic acid content of each of the aggregate set of samples with a fragmentation operation, a multiplexed amplification operation using a set of primers, a sequencing analysis operation, and an alignment operation;

- at the computing system, with at least a subset of the population of subjects, wherein the supplementary dataset is informative of characteristics associated with the cerebro-craniofacial health issue;
- at the computing system, transforming the supplementary dataset and features extracted from at least one of the microbiome composition dataset and the microbiome functional diversity dataset into a characterization model of the cerebro-craniofacial health issue;
- based upon the characterization model, generating a therapy model configured to correct the cerebro-craniofacial health issue; and
- at an output device associated with the subject and in communication with the computing system, promoting a therapy to the subject with the cerebro-craniofacial health issue, upon processing a sample from the subject with the characterization model, in accordance with the therapy model.

26 The method of claim 25, wherein generating the characterization model comprises performing a statistical analysis to assess a set of microbiome composition features and microbiome functional features having variations across a first subset of the population of subjects exhibiting the cerebro-craniofacial health issue and a second subset of the population of subjects not exhibiting the cerebro-craniofacial health issue.

27. The method of claim 26, wherein generating the characterization model comprises:

- extracting candidate features associated with a set of functional aspects of microbiome components indicated in the microbiome composition dataset to generate the microbiome functional diversity dataset; and
- characterizing the mental health issue in association with a subset of the set of functional aspects, the subset derived from at least one of clusters of orthologous groups of proteins features, genomic functional features from the Kyoto Encyclopedia of Genes and Genomes (KEGG), chemical functional features, and systemic functional features.

28. The method of claim 27, wherein generating the characterization model of the cerebro-craniofacial health issue comprises generating a characterization that is diagnostic of at least one symptom of insomnia, light sleep, headache, sinusitis, or poor concentration.

29. The method of claim 28, wherein the generating the characterization model of the cerebro-craniofacial health issue comprises generating a characterization that is diagnostic of at least one symptom of insomnia, and generating a characterization that is diagnostic of at least one symptom of insomnia comprises generating the characterization upon processing the aggregate set of samples and determining presence of features derived from 1) a set of taxa of TABLE A, and 2) a set of one or more functional groups of TABLE A.

30. The method of claim 28, wherein the generating the characterization model of the cerebro-craniofacial health issue comprises generating a characterization that is diagnostic of at least one symptom of light sleep, and generating a characterization that is diagnostic of at least one symptom of light sleep comprises generating the characterization upon processing the aggregate set of samples and determining presence of features derived from 1) a set of taxa of TABLE B, and 2) a set of one or more functional groups of TABLE B.

31. The method of claim 28, wherein the generating the characterization model of the cerebro-craniofacial health issue comprises generating a characterization that is diagnostic of at least one symptom of headache, and generating a characterization that is diagnostic of at least one symptom of headache comprises generating the characterization upon processing the aggregate set of samples and determining presence of features derived from 1) a set of taxa of TABLE C, and 2) a set of one or more functional groups of TABLE C.

32. The method of claim 28, wherein the generating the characterization model of the cerebro-craniofacial health issue comprises generating a characterization that is diagnostic of at least one symptom of sinusitis, and generating a characterization that is diagnostic of at least one symptom of sinusitis comprises generating the characterization upon processing the aggregate set of samples and determining presence of features derived from a set of taxa of TABLE D.

33. The method of claim 28, wherein the generating the characterization model of the cerebro-craniofacial health issue comprises generating a characterization that is diagnostic of at least one symptom of poor concentration, and generating a characterization that is diagnostic of at least one symptom of poor concentration comprises generating the characterization upon processing the aggregate set of samples and determining presence of

features derived from 1) a set of taxa of TABLI
groups of TABLE E.

34. A method for characterizing a cerebro-craniofacial health issue, the method comprising:

- upon processing an aggregate set of samples from a population of subjects, generating at least one of a microbiome composition dataset and a microbiome functional diversity dataset for the population of subjects, the microbiome functional diversity dataset indicative of systemic functions present in the microbiome components of the aggregate set of samples;
- at the computing system, transforming at least one of the microbiome composition dataset and the microbiome functional diversity dataset into a characterization model of the cerebro-craniofacial health issue, wherein the characterization model is diagnostic of the cerebro-craniofacial health issue producing observed changes in dental and/or gingival health; and
- based upon the characterization model, generating a therapy model configured to improve a state of the cerebro-craniofacial health issue.

35. The method of claim 34, wherein generating the characterization comprises analyzing a set of features from the microbiome composition dataset with a statistical analysis, wherein the set of features includes features associated with: relative abundance of different taxonomic groups represented in the microbiome composition dataset, interactions between different taxonomic groups represented in the microbiome composition dataset, and phylogenetic distance between taxonomic groups represented in the microbiome composition dataset.

36. The method of claim 34, wherein generating the characterization comprises performing a statistical analysis with at least one of a Kolmogorov-Smirnov test and a t-test to assess a set of microbiome composition features and microbiome functional features having varying degrees of abundance in a first subset of the population of subjects exhibiting the cerebro-craniofacial health issue and a second subset of the population of subjects not exhibiting the cerebro-craniofacial health issue, wherein generating the characterization further includes clustering using a Bray–Curtis dissimilarity.

37. The method of claim 34, wherein generating the characterization model comprises generating a characterization that is diagnostic of at least one symptom of a

insomnia issue, upon processing the aggregate set of samples and determining presence of features derived from 1) a set of taxa of TABLE A, and 2) a set of one or more functional groups of TABLE A.

38. The method of claim 34, wherein generating the characterization model comprises generating a characterization that is diagnostic of at least one symptom of a light sleep issue, upon processing the aggregate set of samples and determining presence of features derived from 1) a set of taxa of TABLE B, and 2) a set of one or more functional groups of TABLE B.

39. The method of claim 34, wherein generating the characterization model comprises generating a characterization that is diagnostic of at least one symptom of a headache issue, upon processing the aggregate set of samples and determining presence of features derived from 1) a set of taxa of TABLE C, and 2) a set of one or more functional groups of TABLE C.

40. The method of claim 34, wherein generating the characterization model comprises generating a characterization that is diagnostic of at least one symptom of a sinusitis issue, upon processing the aggregate set of samples and determining presence of features derived from a set of taxa of TABLE D.

41. The method of claim 34, wherein generating the characterization model comprises generating a characterization that is diagnostic of at least one symptom of a poor concentration issue, upon processing the aggregate set of samples and determining presence of features derived from 1) a set of taxa of TABLE E, and 2) a set of one or more functional groups of TABLE E.

42. The method of claim 34, further including diagnosing a subject with the cerebro-craniofacial health issue upon processing a sample from the subject with the characterization model; and at an output device associated with the subject, promoting a therapy to the subject with the cerebro-craniofacial health issue based upon the characterization model and the therapy model.

43. The method of claim 42, wherein promoting the therapy comprises promoting a bacteriophage-based therapy to the subject, the bacteriophage-based therapy providing a bacteriophage component that selectively downregulates a population size of an undesired taxon associated with the cerebro-craniofacial health issue.

44. The method of claim 42, wherein promoting the therapy comprises promoting a prebiotic therapy to the subject, the prebiotic therapy affecting a microorganism component that selectively supports a population size increase of a desired taxon associated with correction of the cerebro-craniofacial health issue, based on the therapy model.

45. The method of claim 42, wherein promoting the therapy comprises promoting a probiotic therapy to the subject, the probiotic therapy affecting a microorganism component of the subject, in promoting correction of the cerebro-craniofacial health issue, based on the therapy model.

46. The method of claim 42, wherein promoting the therapy comprises promoting a microbiome modifying therapy to the subject in order to improve a state of the cerebro-craniofacial health associated symptom.

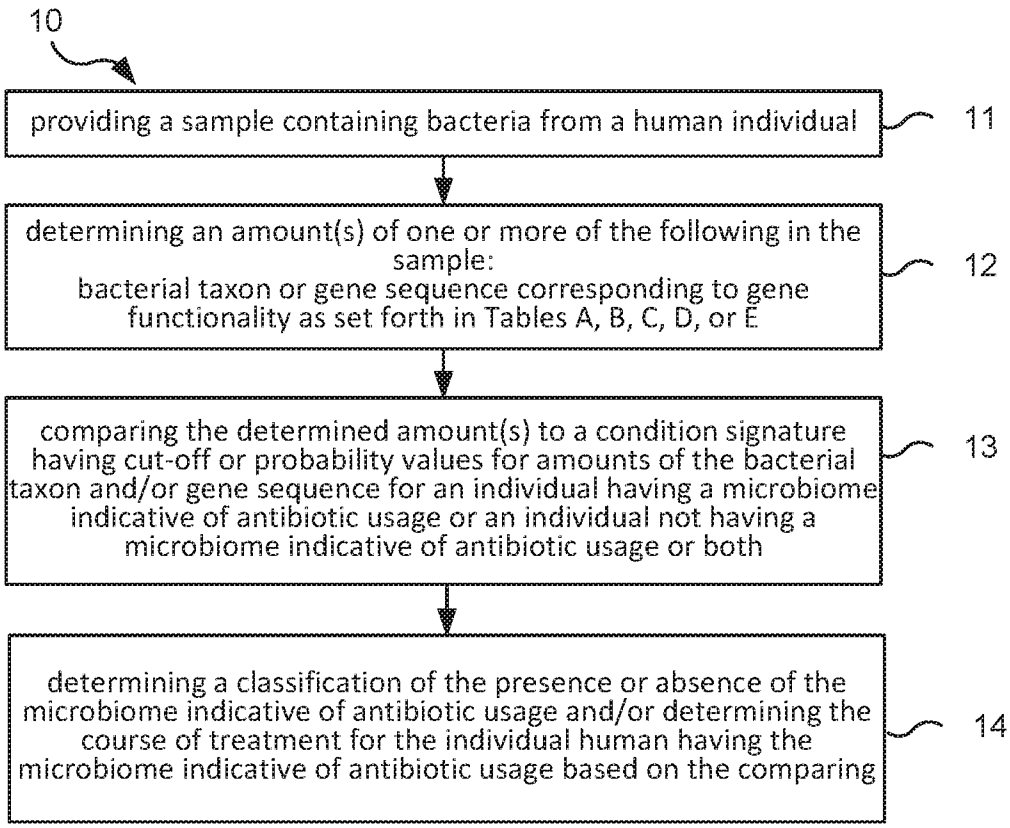
10
providing a sample containing bacteria from a human individual
11
determining an amount(s) of one or more of the following in the sample:
bacterial taxon or gene sequence corresponding to gene functionality as set forth in Tables A, B, C, D, or E
12
comparing the determined amount(s) to a condition signature having cut-off or probability values for amounts of the bacterial taxon and/or gene sequence for an individual having a microbiome indicative of antibiotic usage or an individual not having a microbiome indicative of antibiotic usage or both
13
determining a classification of the presence or absence of the microbiome indicative of antibiotic usage and/or determining the course of treatment for the individual human having the microbiome indicative of antibiotic usage based on the comparing
14

FIG. 1A

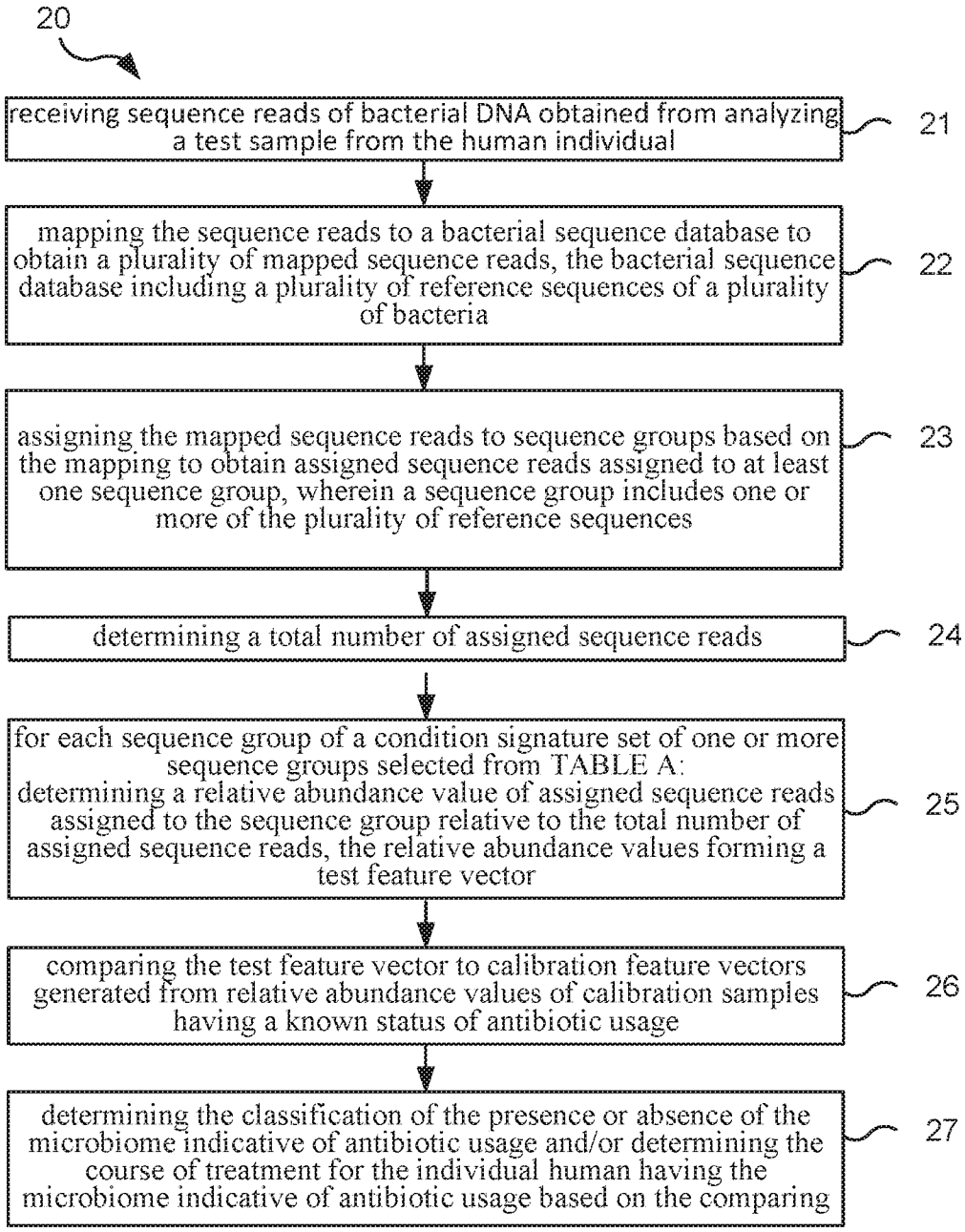
20
receiving sequence reads of bacterial DNA obtained from analyzing a test sample from the human individual
21
mapping the sequence reads to a bacterial sequence database to obtain a plurality of mapped sequence reads, the bacterial sequence database including a plurality of reference sequences of a plurality of bacteria
22
assigning the mapped sequence reads to sequence groups based on the mapping to obtain assigned sequence reads assigned to at least one sequence group, wherein a sequence group includes one or more of the plurality of reference sequences
23
determining a total number of assigned sequence reads
24
for each sequence group of a condition signature set of one or more sequence groups selected from TABLE A:
determining a relative abundance value of assigned sequence reads assigned to the sequence group relative to the total number of assigned sequence reads, the relative abundance values forming a test feature vector
25
comparing the test feature vector to calibration feature vectors generated from relative abundance values of calibration samples having a known status of antibiotic usage
26
determining the classification of the presence or absence of the microbiome indicative of antibiotic usage and/or determining the course of treatment for the individual human having the microbiome indicative of antibiotic usage based on the comparing
27

FIG. 1B

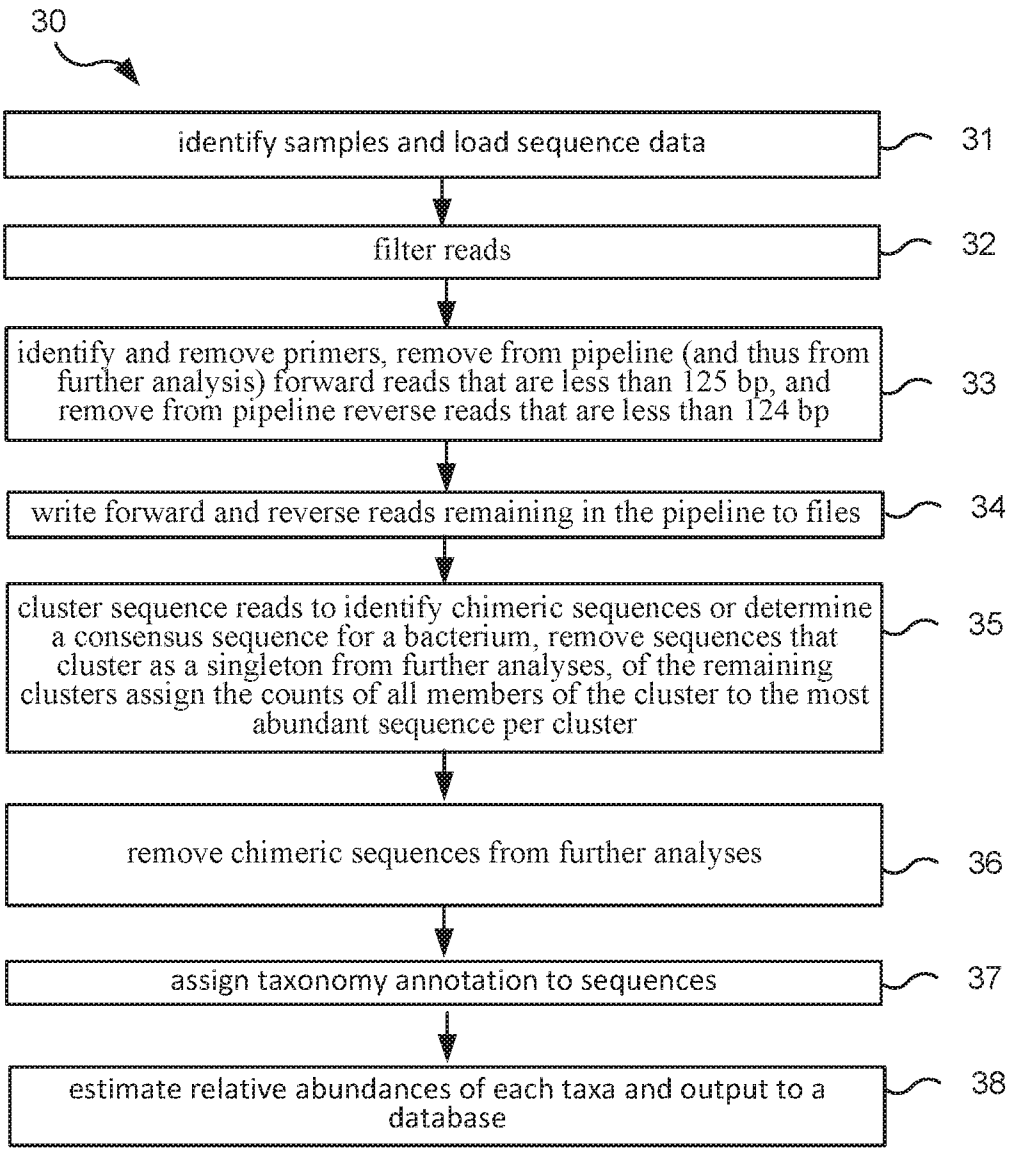
30
identify samples and load sequence data
31
filter reads
32
identify and remove primers, remove from pipeline (and thus from further analysis) forward reads that are less than 125 bp, and remove from pipeline reverse reads that are less than 124 bp
33
write forward and reverse reads remaining in the pipeline to files
34
cluster sequence reads to identify chimeric sequences or determine a consensus sequence for a bacterium, remove sequences that cluster as a singleton from further analyses, of the remaining clusters assign the counts of all members of the cluster to the most abundant sequence per cluster
35
remove chimeric sequences from further analyses
36
assign taxonomy annotation to sequences
37
estimate relative abundances of each taxa and output to a database
38

FIG. 1C

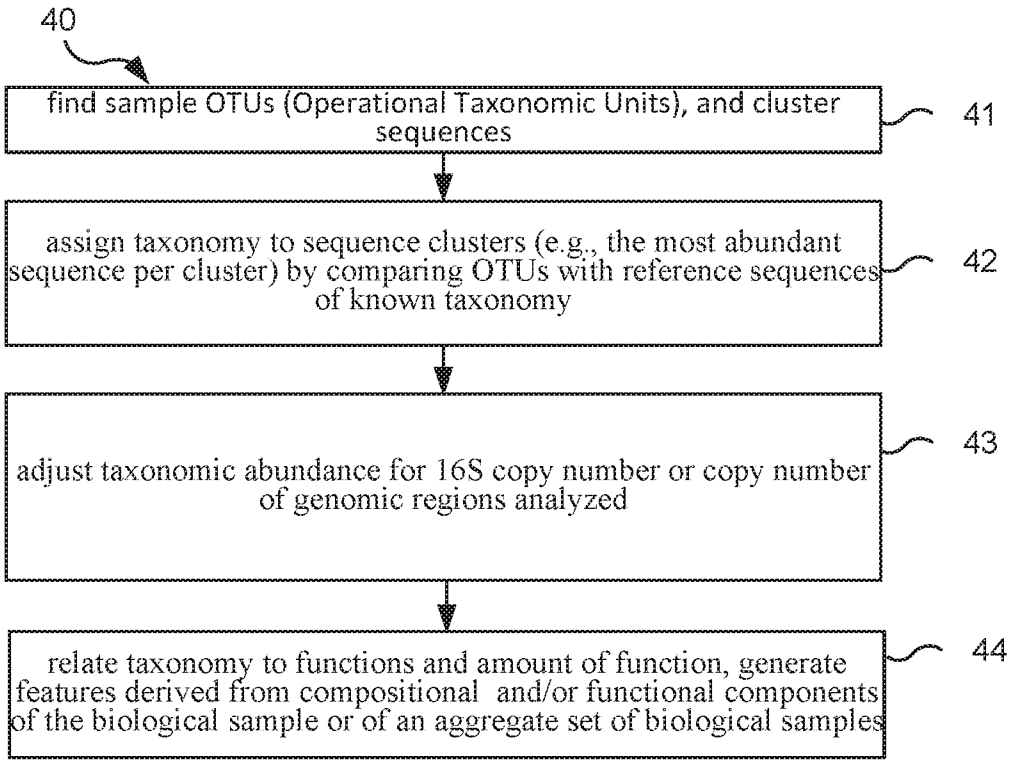
40
find sample OTUs (Operational Taxonomic Units), and cluster sequences
41
assign taxonomy to sequence clusters (e.g., the most abundant sequence per cluster) by comparing OTUs with reference sequences of known taxonomy
42
adjust taxonomic abundance for 16S copy number or copy number of genomic regions analyzed
43
relate taxonomy to functions and amount of function, generate features derived from compositional and/or functional components of the biological sample or of an aggregate set of biological samples
44

FIG. 1D

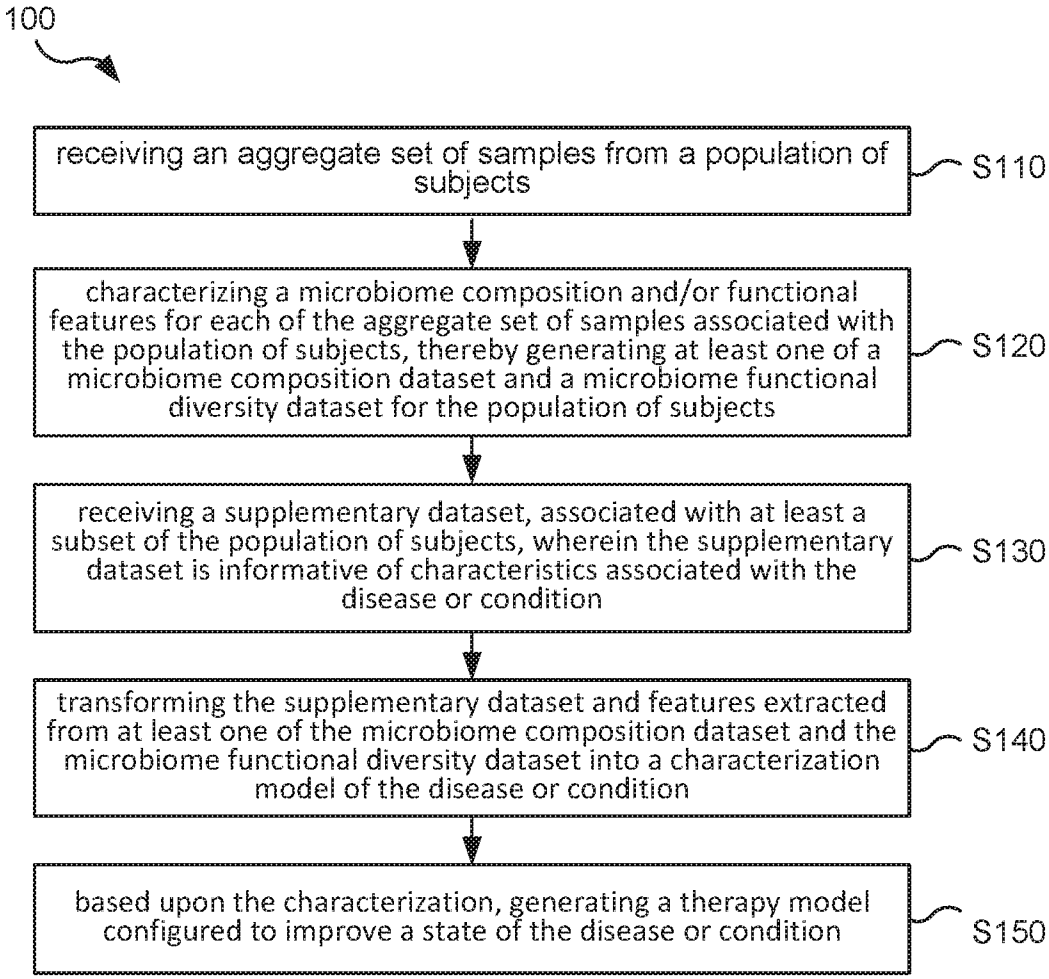
100
receiving an aggregate set of samples from a population of subjects
S110
characterizing a microbiome composition and/or functional features for each of the aggregate set of samples associated with the population of subjects, thereby generating at least one of a microbiome composition dataset and a microbiome functional diversity dataset for the population of subjects
S120
receiving a supplementary dataset, associated with at least a subset of the population of subjects, wherein the supplementary dataset is informative of characteristics associated with the disease or condition
S130
transforming the supplementary dataset and features extracted from at least one of the microbiome composition dataset and the microbiome functional diversity dataset into a characterization model of the disease or condition
S140
based upon the characterization, generating a therapy model configured to improve a state of the disease or condition
S150

FIG. 1E

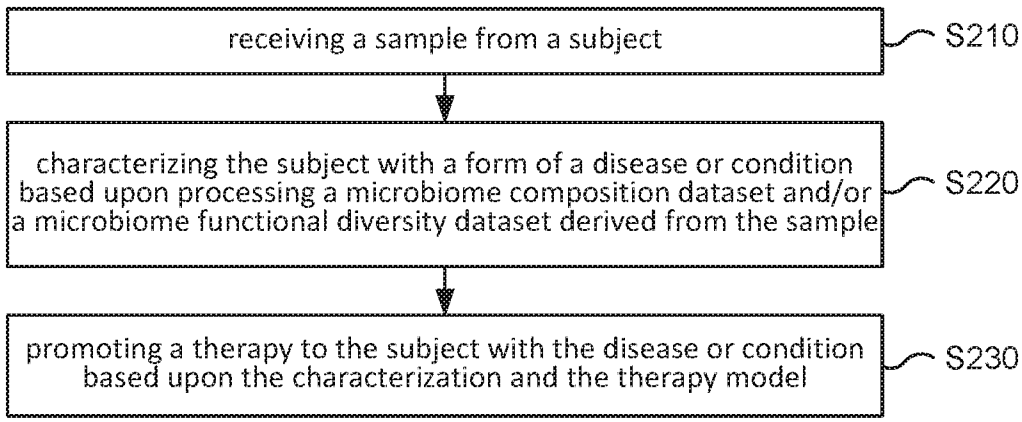
receiving a sample from a subject
S210
characterizing the subject with a form of a disease or condition based upon processing a microbiome composition dataset and/or a microbiome functional diversity dataset derived from the sample
S220
promoting a therapy to the subject with the disease or condition based upon the characterization and the therapy model
S230

FIG. 1F

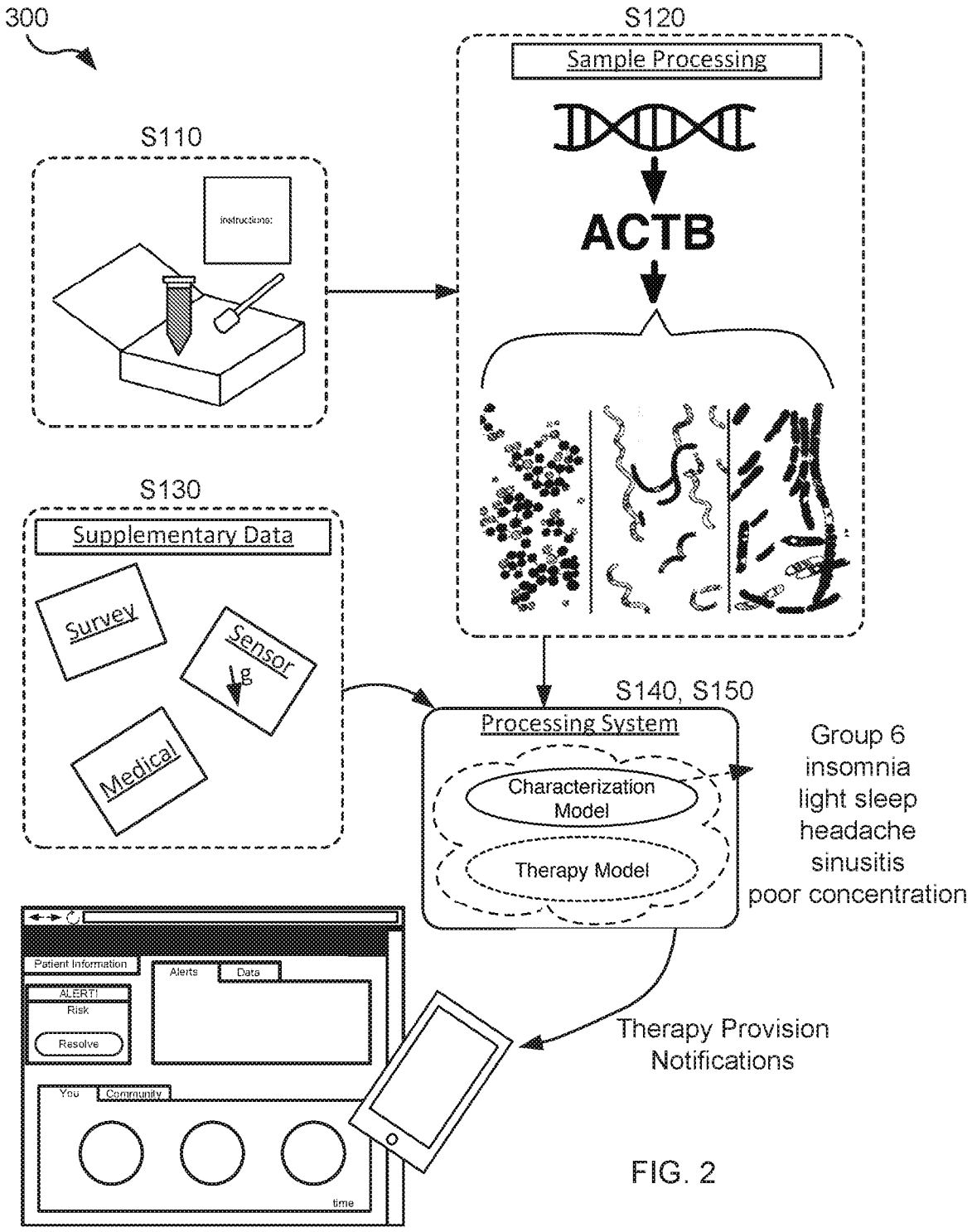
300
S120
Sample Processing
ACTB
S110
Instructions:
S130
Supplementary Data
Survey
Sensor
Medical
S140, S150
Processing System
Characterization Model
Therapy Model
Group 6
insomnia
light sleep
headache
sinusitis
poor concentration
Patient Information
Alerts
Data
ALERT!
Risk
Resolve
You
Community
time
Therapy Provision
Notifications

FIG. 2

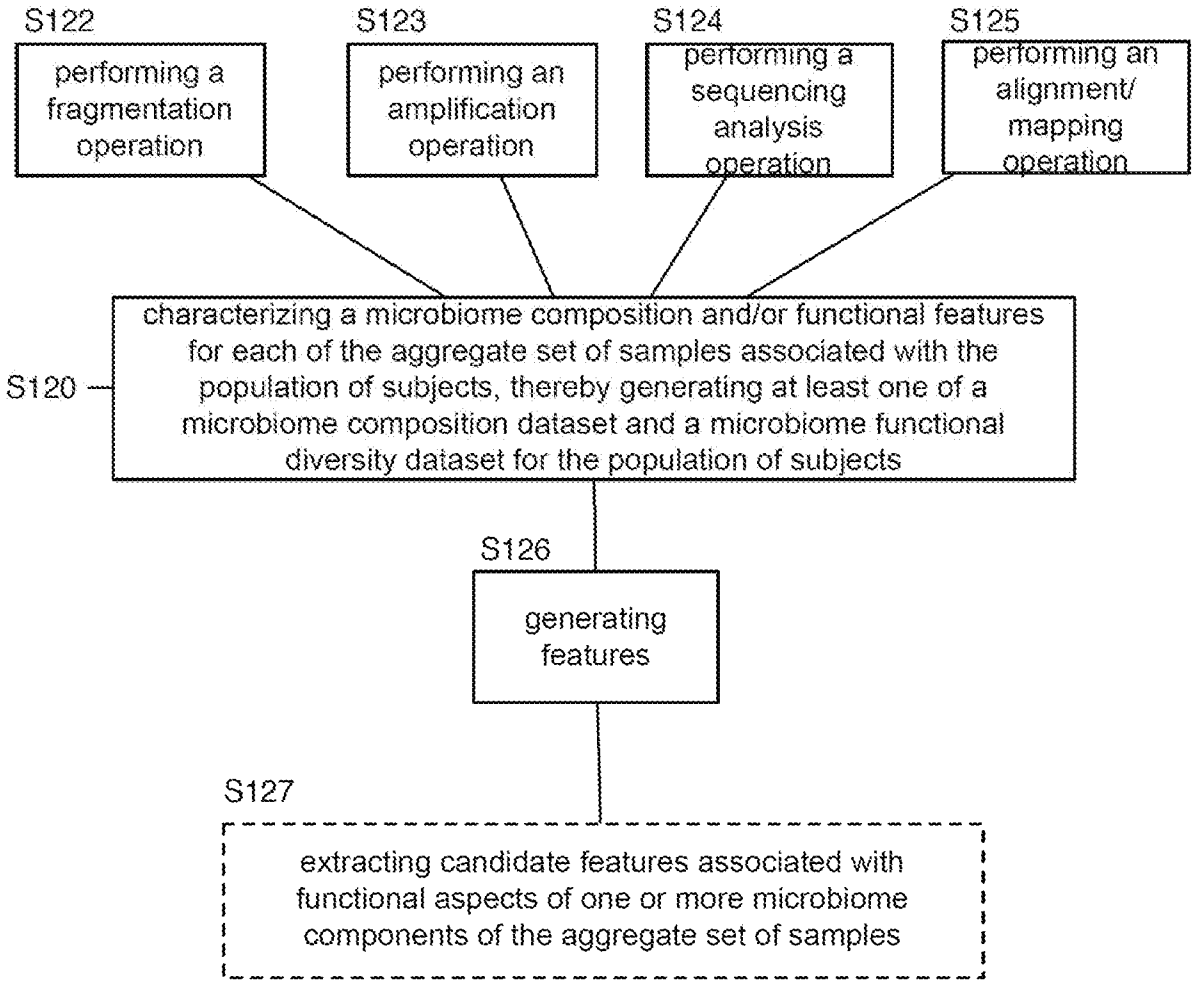
S122
performing a fragmentation operation
S123
performing an amplification operation
S124
performing a sequencing analysis operation
S125
performing an alignment/ mapping operation
S120
characterizing a microbiome composition and/or functional features for each of the aggregate set of samples associated with the population of subjects, thereby generating at least one of a microbiome composition dataset and a microbiome functional diversity dataset for the population of subjects
S126
generating features
S127
extracting candidate features associated with functional aspects of one or more microbiome components of the aggregate set of samples

FIG. 3

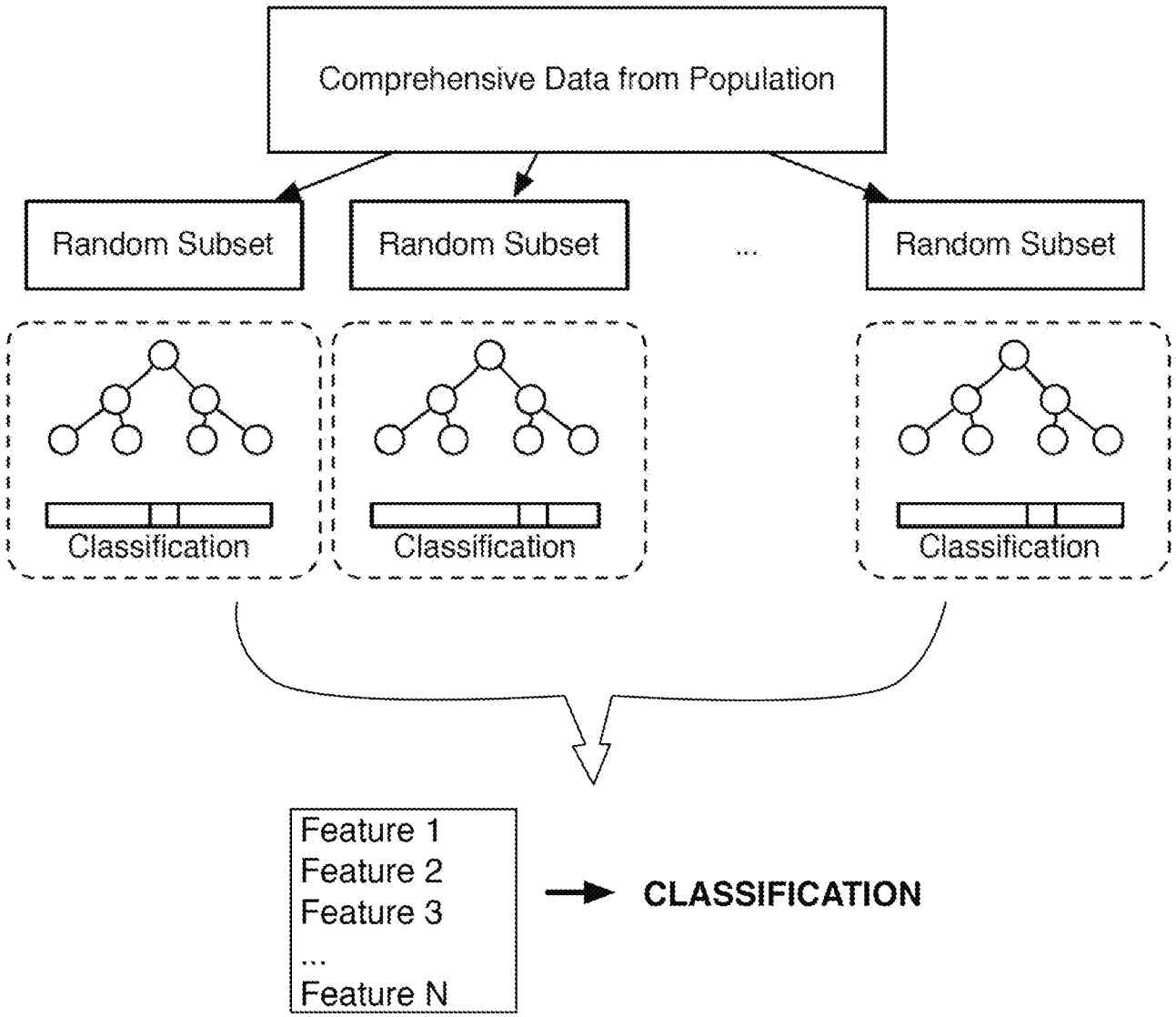
Comprehensive Data from Population
Random Subset
Random Subset
...
Random Subset
Classification
Classification
Classification
Feature 1
Feature 2
Feature 3
...
Feature N
CLASSIFICATION

FIG. 4

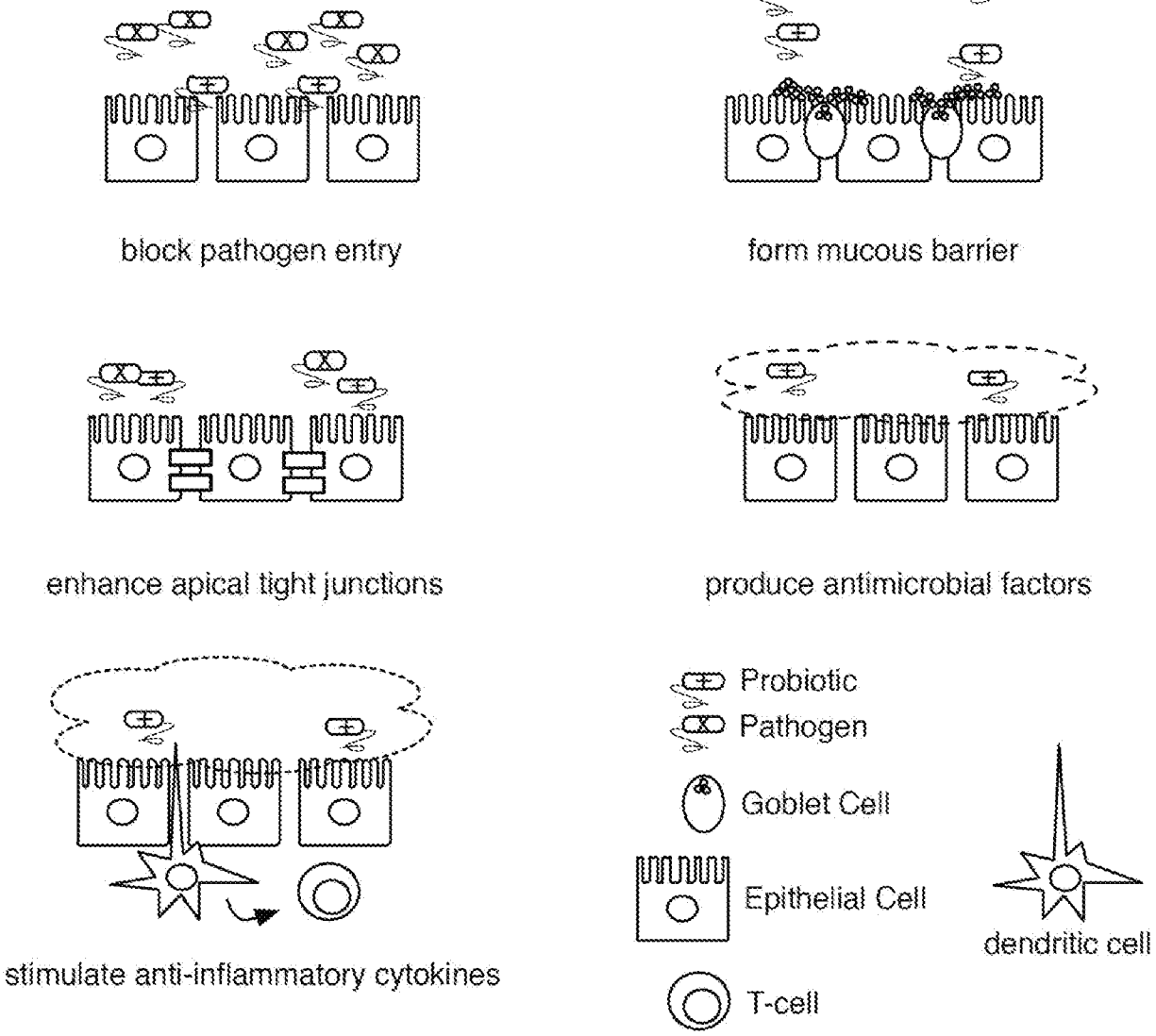
block pathogen entry
form mucous barrier
enhance apical tight junctions
produce antimicrobial factors
stimulate anti-inflammatory cytokines
Probiotic
Pathogen
Goblet Cell
Epithelial Cell
T-cell
dendritic cell

FIG. 5

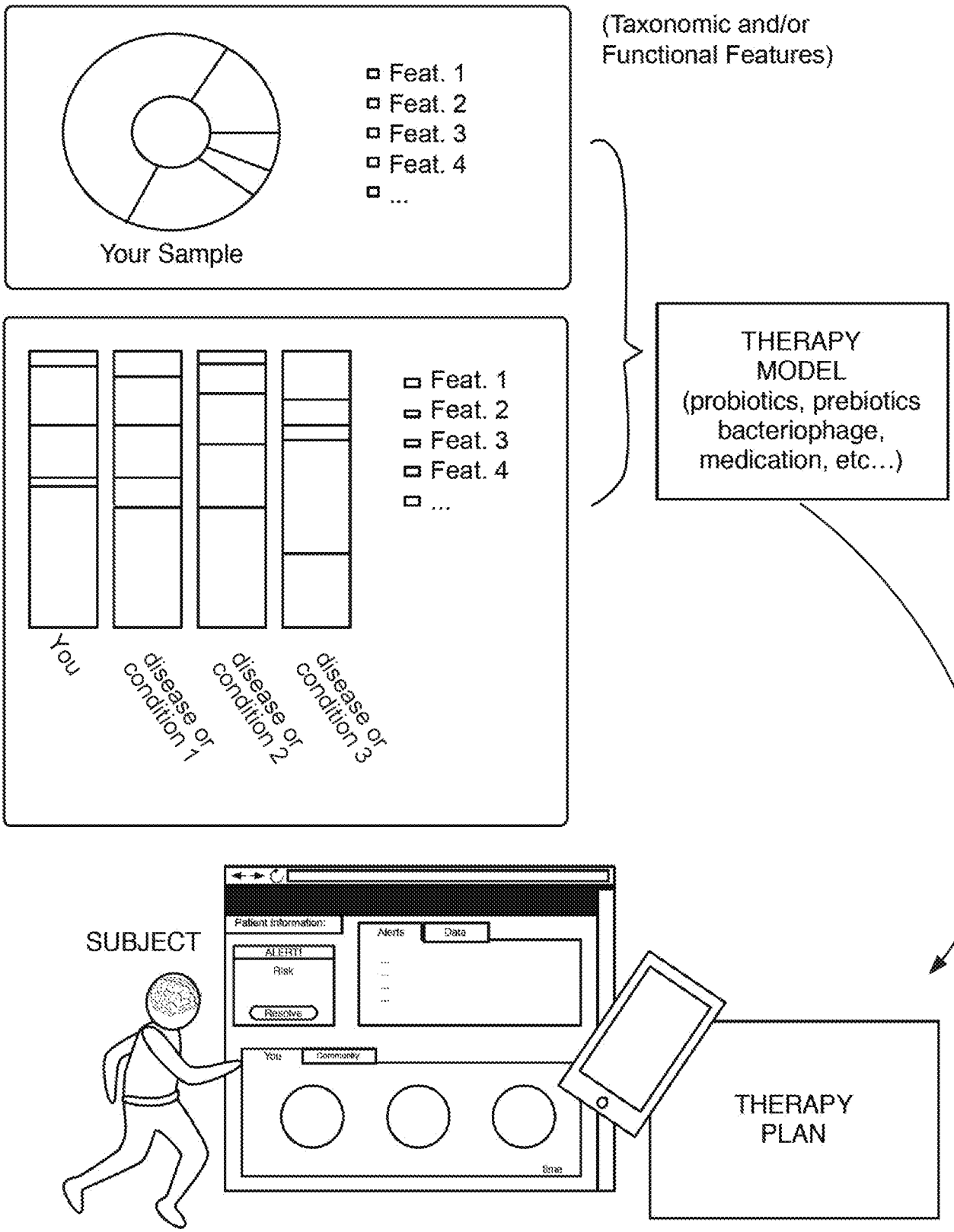
(Taxonomic and/or Functional Features)
Feat. 1
Feat. 2
Feat. 3
Feat. 4
...
Your Sample
Feat. 1
Feat. 2
Feat. 3
Feat. 4
...
You
disease or condition 1
disease or condition 2
disease or condition 3
THERAPY MODEL (probiotics, prebiotics bacteriophage, medication, etc...)
SUBJECT
Patient Information:
Alerts
Data
ALERT!
Risk
Resolve
You
Community
Time
THERAPY PLAN

FIG. 6

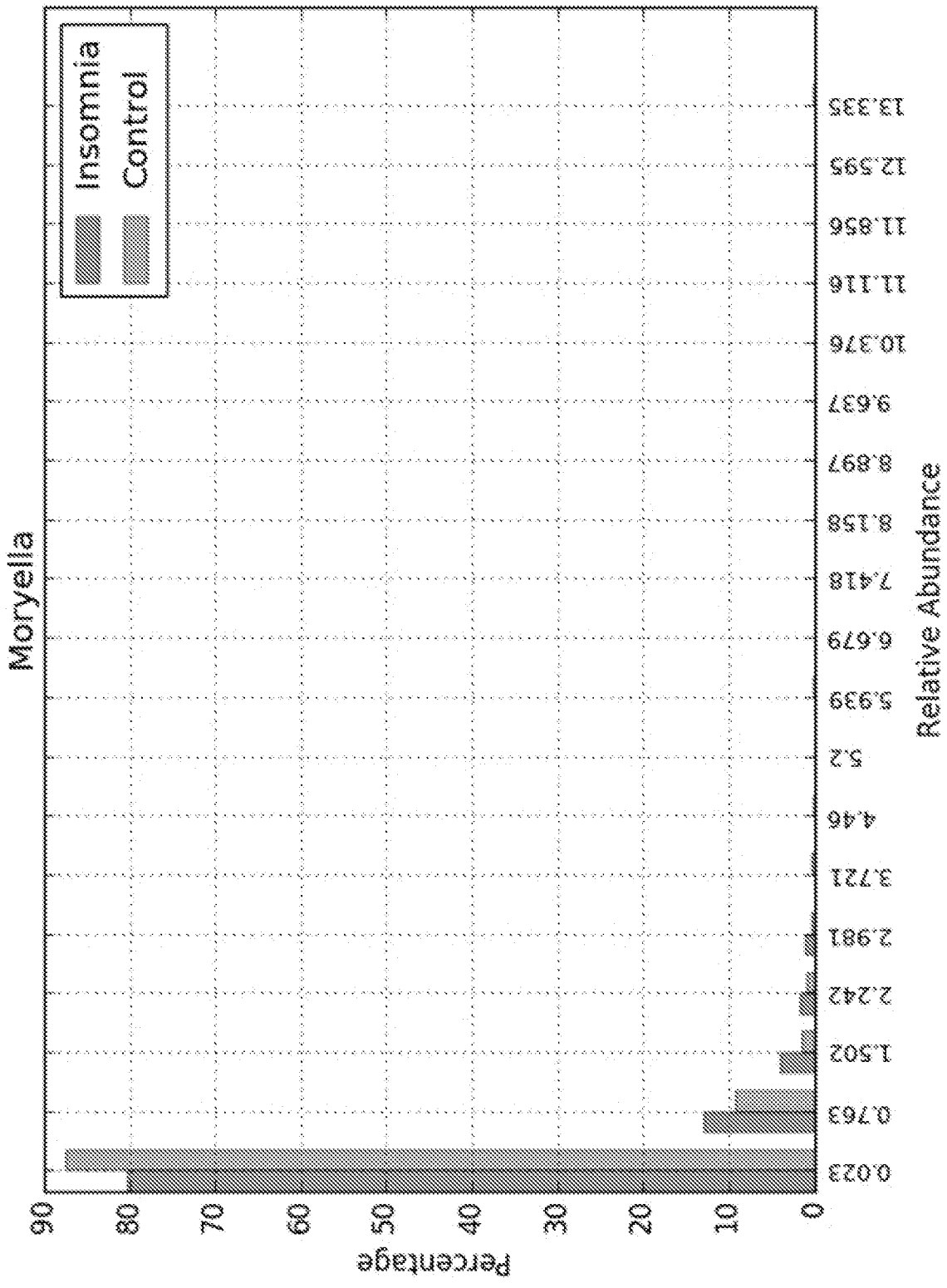
Moryella
Insomnia
Control
Percentage
0
10
20
30
40
50
60
70
80
90
Relative Abundance
0.023
0.763
1.502
2.242
2.981
3.721
4.46
5.2
5.939
6.679
7.418
8.158
8.897
9.637
10.376
11.116
11.856
12.595
13.335

FIG. 7

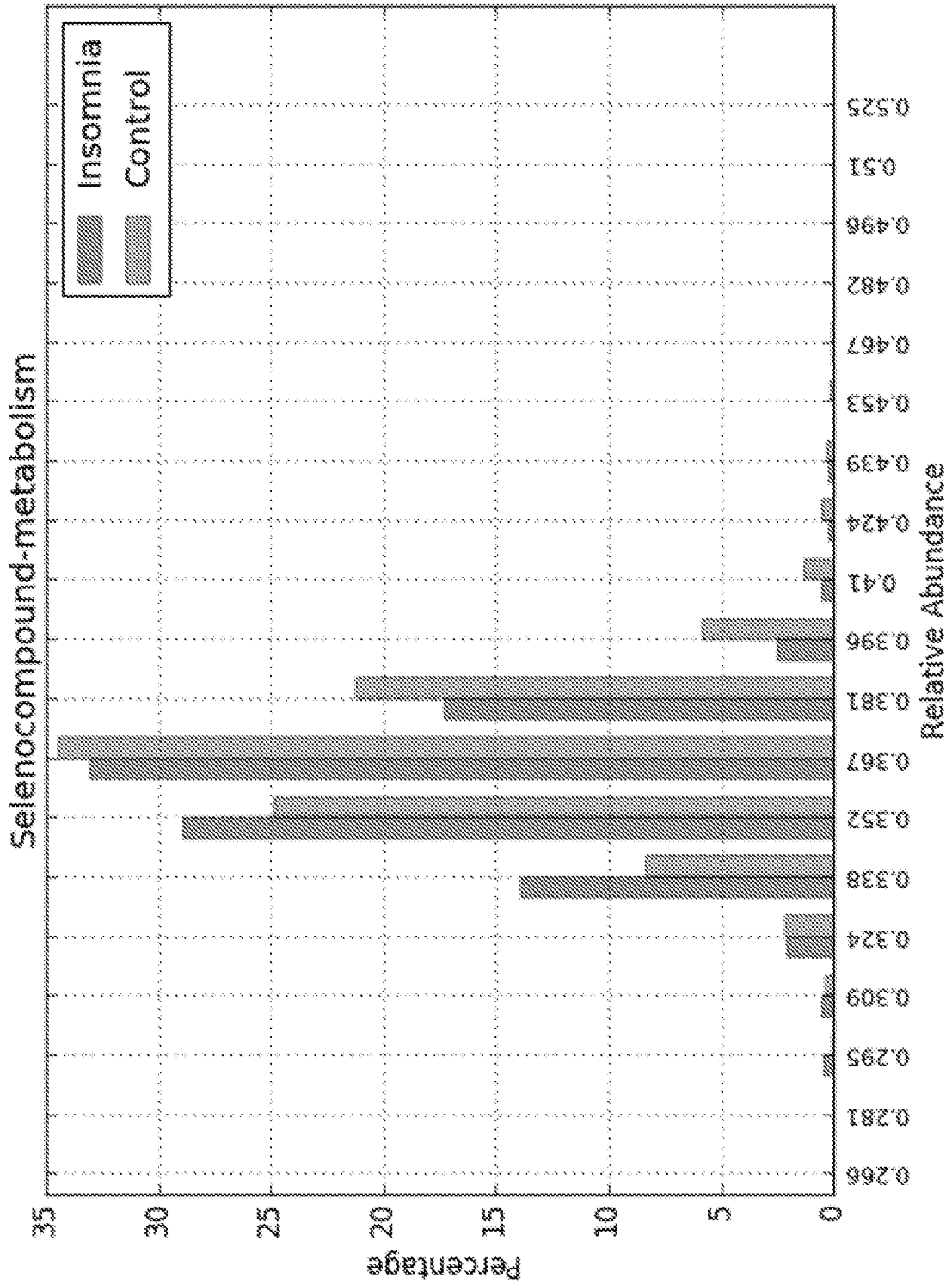
Selenocompound-metabolism
Insomnia
Control
Percentage
0
5
10
15
20
25
30
35
Relative Abundance
0.266
0.281
0.295
0.309
0.324
0.338
0.352
0.367
0.381
0.396
0.41
0.424
0.439
0.453
0.467
0.482
0.496
0.51
0.525

FIG. 8

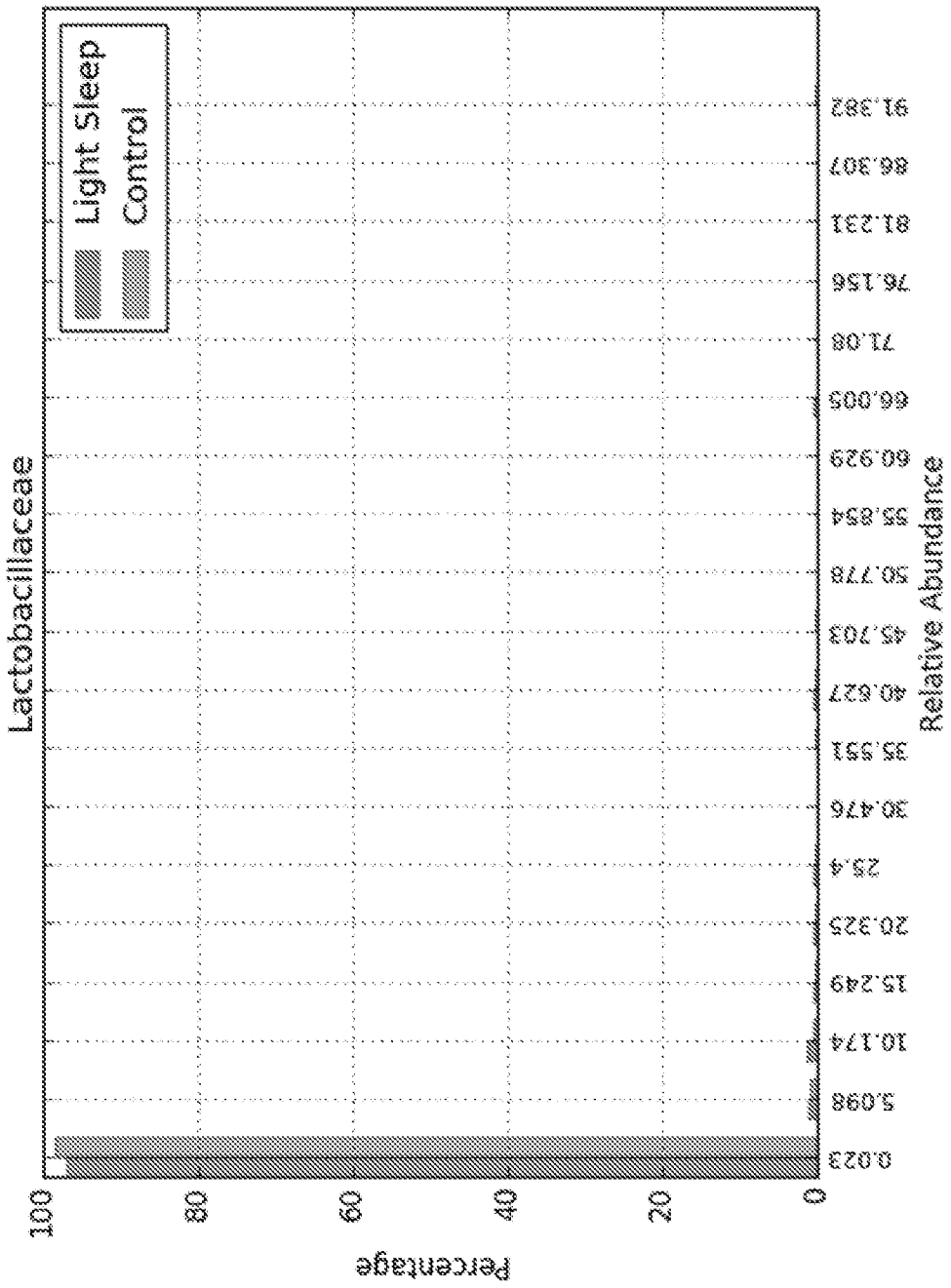
Lactobacillaceae
Light Sleep
Control
Percentage
100
80
60
40
20
0
Relative Abundance
0.023
5.098
10.174
15.249
20.325
25.4
30.476
35.551
40.627
45.703
50.778
55.854
60.929
66.005
71.08
76.156
81.231
86.307
91.382

FIG. 9

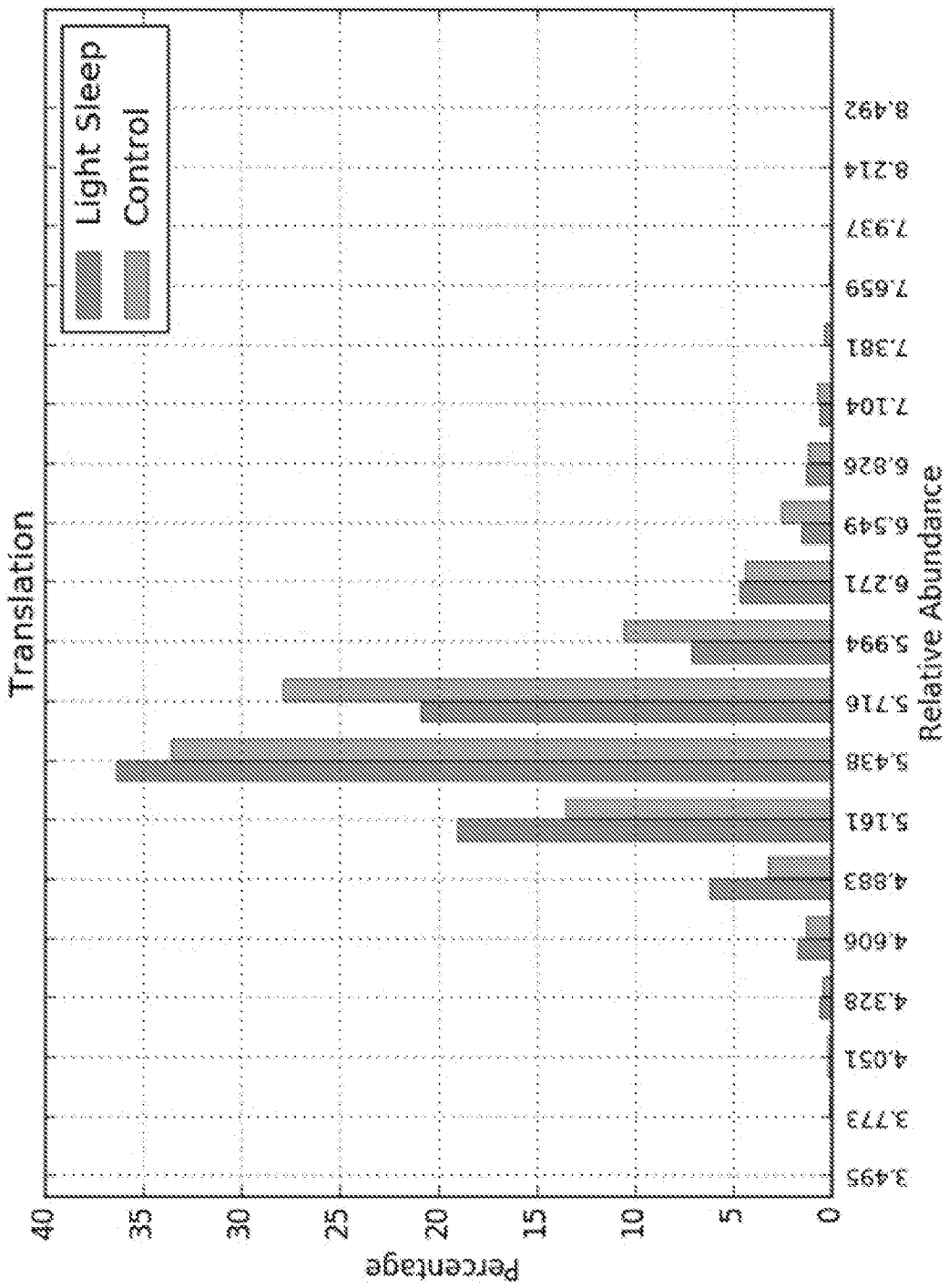
Translation
Light Sleep
Control
Percentage
0
5
10
15
20
25
30
35
40
Relative Abundance
3.495
3.773
4.051
4.328
4.606
4.883
5.161
5.438
5.716
5.994
6.271
6.549
6.826
7.104
7.381
7.659
7.937
8.214
8.492

FIG. 10

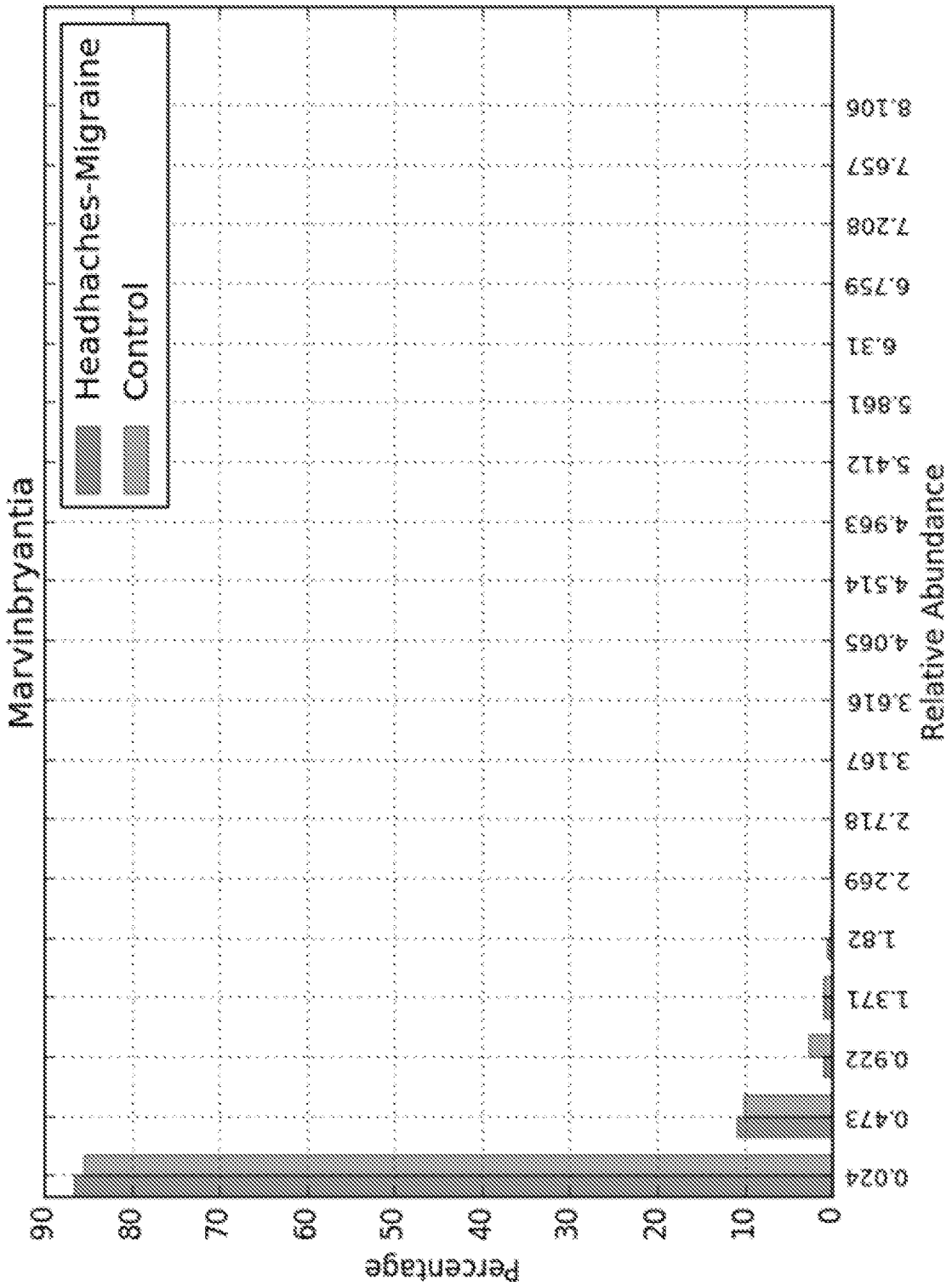
Marvinbryantia
Headaches-Migraine
Control
Percentage
0
10
20
30
40
50
60
70
80
90
Relative Abundance
0.024
0.473
0.922
1.371
1.82
2.269
2.718
3.167
3.616
4.065
4.514
4.963
5.412
5.861
6.31
6.759
7.208
7.657
8.106

FIG. 11

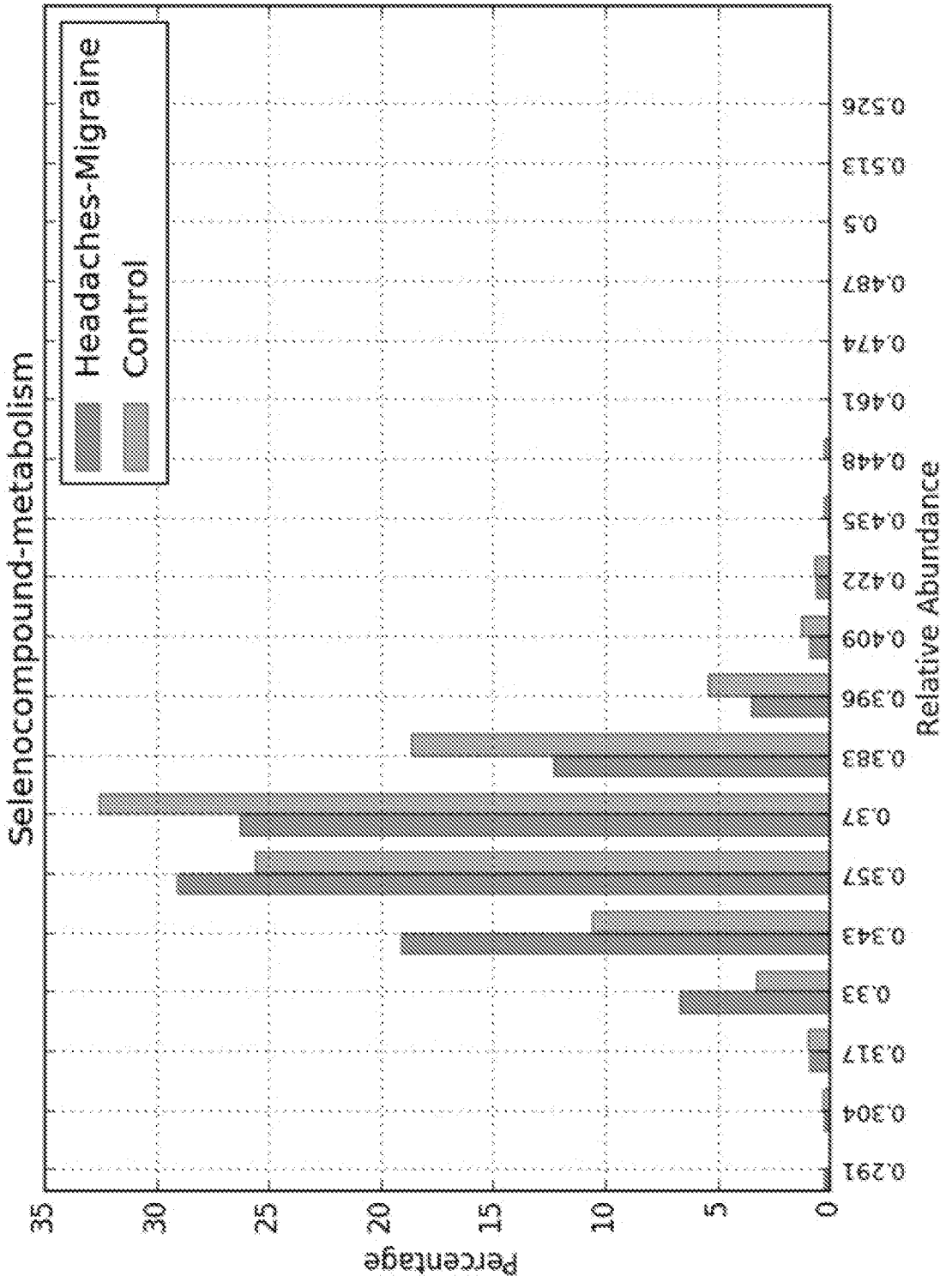
Selenocompound-metabolism
Headaches-Migraine
Control
Percentage
Relative Abundance
0
5
10
15
20
25
30
35
0.291
0.304
0.317
0.33
0.343
0.357
0.37
0.383
0.396
0.409
0.422
0.435
0.448
0.461
0.474
0.487
0.5
0.513
0.526

FIG. 12

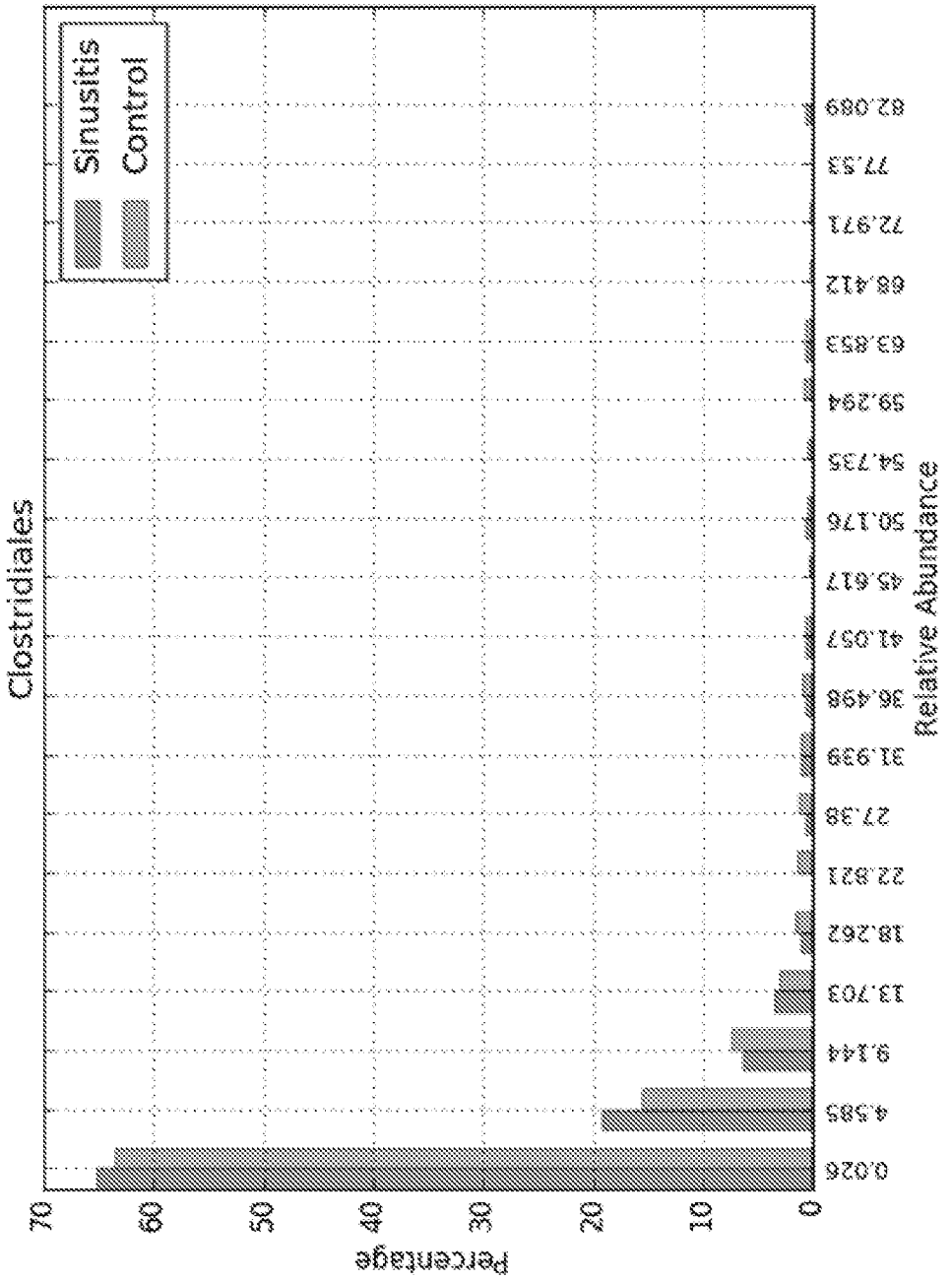
Clostridiales
Sinusitis
Control
Percentage
0
10
20
30
40
50
60
70
Relative Abundance
0.026
4.585
9.144
13.703
18.262
22.821
27.38
31.939
36.498
41.057
45.617
50.176
54.735
59.294
63.853
68.412
72.971
77.53
82.089

FIG. 13

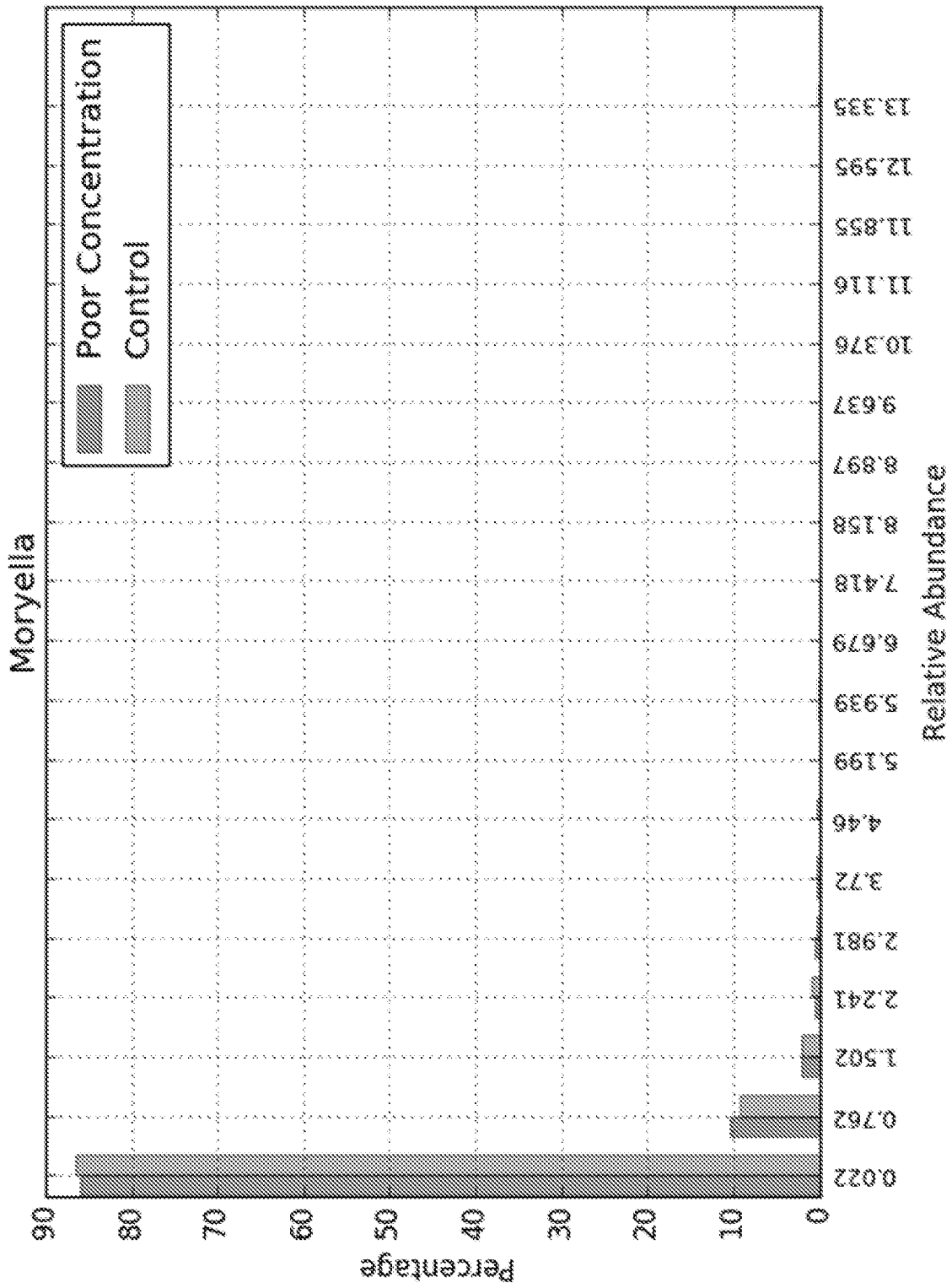
Moryella
Poor Concentration
Control
Percentage
0
10
20
30
40
50
60
70
80
90
Relative Abundance
0.022
0.762
1.502
2.241
2.981
3.72
4.46
5.199
5.939
6.679
7.418
8.158
8.897
9.637
10.376
11.116
11.855
12.595
13.335

FIG. 14

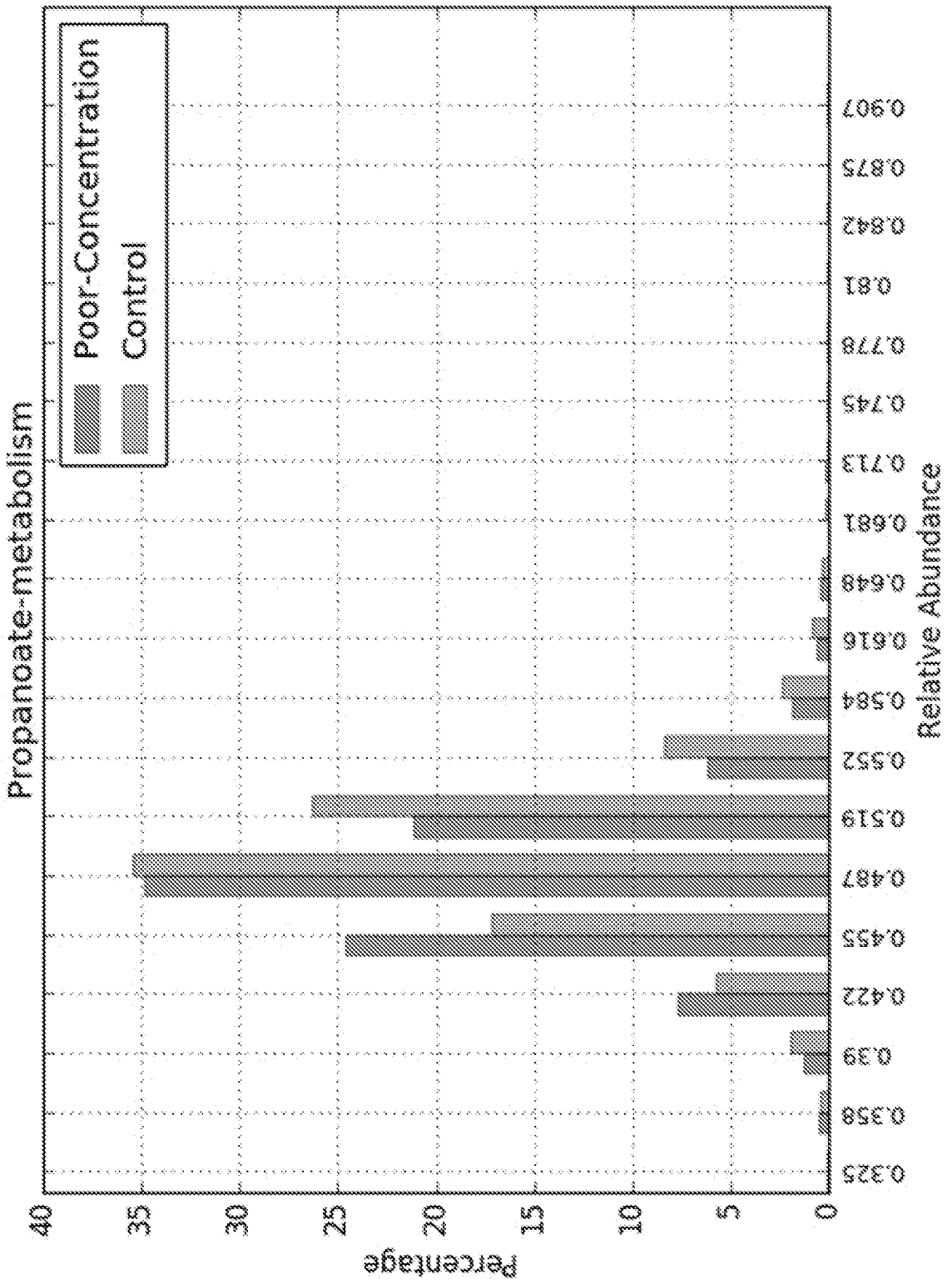
Propanoate-metabolism
Poor-Concentration
Control
Percentage
0
5
10
15
20
25
30
35
40
Relative Abundance
0.325
0.358
0.39
0.422
0.455
0.487
0.519
0.552
0.584
0.616
0.648
0.681
0.713
0.745
0.778
0.81
0.842
0.875
0.907

FIG. 15

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2016/051155

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - A61B 5/00; C12Q 1/68; G06F 19/00; G06F 19/12; G06F 19/22; G06F 19/28 (2016.01)
CPC - A61B 5/4233; A61B 5/4836; C12Q 1/689; C12Q 2600/112; G06F 19/12; G06F 19/22 (2016.11)
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)
IPC - A61B 5/00; C12Q 1/68; G06F 19/00; G06F 19/12; G06F 19/22; G06F 19/28
CPC - A61B 5/4233; A61B 5/4836; C12Q 1/689; C12Q 2600/112; G06F 19/12; G06F 19/22; G06F 19/24; G06F 19/28; G06F 19/3418; G06F 19/3443; G06F 19/345

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
USPC - 435/6.12; 435/287.2; 506/2; 506/36; 702/19 (keyword delimited)

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
Orbit, Google Patents, Google Scholar
Search terms used: microbiome condition classification inassignee:ubiome (Excema OR antibiotic OR mental health OR sleep OR infectious OR craniofacial) 16s ribosomal "Bray-Curtis dissimilarity" ("Kyoto Encyclopedia of Genes and Genomes" OR KEGG) (Kolmogorov-Smirnov OR t-test)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2015/0211078 A1 (UBIOME INC.) 30 July 2015 (30.07.2015) entire document	1-6, 11-23, 25-46
Y	WO 2014/121298 A2 (SERES HEALTH, INC.) 07 August 2014 (07.08.2014) entire document	1-6, 11-23, 25-46
Y	WO 2015/074054 A1 (THE TRUSTEES OF COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK) 21 May 2015 (21.05.2015) entire document	4-6, 36
Y	US 2013/0121968 A1 (ATOSSA GENETICS, INC. et al) 16 May 2013 (16.05.2013) entire document	13, 36
Y	US 2015/0086581 A1 (THE REGENTS OF THE UNIVERSITY OF CALIFORNIA) 26 March 2015 (26.03.2015) entire document	27-33, 43-46
A	FAUST et al. "Microbial Co-Occurrence Relationships in the Human Microbiome," PLoS Comput Biol, 12 July 2012 (12.07.2012), Vol. 8, Pgs. 1-17. entire document	1-6, 11-23, 25-46
A	HOLLISTER et al. "Structure and Function of the Healthy Pre-Adolescent Pediatric Gut Microbiome," Microbiome, 26 August 2015 (26.08.2015), Vol. 3, Pgs. 1-13. entire document	1-6, 11-23, 25-46
A	MORGAN et al. "Human Microbiome Analysis," PLoS Comput Biol, 27 December 2012 (27.12.2012), Vol. 8, Pgs. 1-14. entire document	1-6, 11-23, 25-46
A	US 2012/0149584 A1 (OLLE et al) 14 June 2012 (14.06.2012) entire document	1-6, 11-23, 25-46
P, X	US 2016/0224749 A1 (UBIOME INC.) 04 August 2016 (04.08.2016) entire document	1-6, 11-23, 25-46

☐ 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
01 November 2016	12 DEC 2016
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, VA 22313-1450 Facsimile No. 571-273-8300	Authorized officer Blaine R. Copenheaver PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2016/051155

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☒ Claims Nos.: 7-10, 24
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.