

METHODS FOR PREDICTING RESPONSE TO A THERAPY FOR A DISORDER THROUGH CORE MICROBIOME GUILDS

CROSS REFERENCE TO RELATED APPLICATIONS

[0001] This application claims priority to U.S. Provisional Patent Application No. 63/595,189, filed November 1, 2023, the content of which is hereby incorporated by reference herein, in its entirety, for all purposes.

BRIEF DESCRIPTION OF THE SEQUENCE LISTING

[0002] This submission incorporates by reference the "Sequence Listing XML" file named ST26_126146_5002_PR.XML containing SEQ ID NOs: 1-99534, created on April 22, 2023, and having a size of 2,491,699 kilobytes, in accordance with 37 CFR §§ 1.831 through 1.835, submitted on a read only optical disc (DVD) as an XML file via mail for U.S. Provisional Application Serial No. 63/498,177. 37 CFR § 1.835(a)(1). The Sequence Listing XML is hereby incorporated by reference herein in its entirety.

[0003] This submission incorporates by reference the "Sequence Listing XML" file named ST26_126146_5001_WO.XML containing SEQ ID NOs: 1-99534, created on April 24, 2024, and having a size of 2,491,699 kilobytes, in accordance with 37 CFR §§ 1.831 through 1.835, submitted on a read only optical disc (DVD) as an XML file via mail on April 25, 2024, for PCT Application Serial No. PCT/US24/26282. 37 CFR § 1.835(a)(1). The Sequence Listing XML is hereby incorporated by reference herein in its entirety.

BACKGROUND

[0004] The human gut microbiome, emblematic of a complex adaptive system (CAS), hosts trillions of microorganisms, embodying a rich array of phylogenetic diversity. This sophisticated ecosystem not only sustains active interaction with its host environment but also showcases dynamic adaptability, thereby playing a pivotal role in the maintenance of health and modulation of disease susceptibility. Amidst the staggering microbial diversity resident within the human gut, the concept of a 'core microbiome' has gained considerable traction. This core is hypothesized to incorporate microbes that ubiquitously colonize healthy individuals, thus

contributing significantly to the preservation of homeostasis in nutrition, metabolism, immunity, and behavior. The integral role of this core microbiome is akin to that of an essential organ, underscoring its criticality in overall health management.

[0005] Historically, the demarcation of the core microbiome has predominantly rested upon the evaluation of the presence or absence, supplemented by the quantification of the abundance or prevalence of specific taxa or genes/pathways within a cohort of healthy individuals. While these methodologies have undoubtedly provided significant insights into the structural configuration and potential functional traits of the microbiome, they may inadequately represent the vital ecological interactions that underscore the stability and resilience of this intricate system. This oversight is particularly relevant when considering the critical role these interactions play in the inception, progression, and remission of various disease states.

[0006] As a CAS, the microbiome adheres to the modular design principle. Integral components of a CAS are organized into modules, which interconnect to establish a network. Within the gut ecosystem, individual microbes are integrated into a modular structure referred to as guilds. Each guild, despite comprising microorganisms of diverse taxonomic backgrounds, functions as a coherent functional unit or module within the microbiome's CAS. Members of a guild display cooperative behavior through co-abundance, and different guilds may engage in cooperative or competitive interactions to shape an ecological network. Consequently, the characterization of the core microbiome in terms of guilds emerges as a promising and intriguing approach.

SUMMARY

[0007] Throughout their co-evolution, gut microbiota has established a vital role in sustaining human health. Identifying core microbiome constituents that reliably confer essential health benefits, however, remains a significant challenge. It was posited that these core members should sustain their ecological interactions, cooperative or competitive, in spite of changing environmental conditions. Drawing from a high-fiber intervention trial in type 2 diabetes patients and 26 diverse case-control datasets, 284 high-quality metagenome-assembled genomes consistently forming stable pairs across individuals amidst dietary shifts or disease progression were identified. These genomes correspond to two guilds, encompassing the most resilient and highly interconnected bacteria, which collectively correlate with an expansive range of health

conditions. One guild's genomes were gene-rich for plant polysaccharide degradation and butyrate production, while the other was typified by a high prevalence of genes linked to virulence and antibiotic resistance. Utilizing these genomes as a reference, Random Forest models adeptly differentiated between cases and controls across 15 distinct diseases and forecasted patients' responses to immunotherapy. Therefore, this core microbiome signature has potential as a unifying therapeutic target for enhancing health.

[0008] Individual microbial cells are considered to be the fundamental components or agents of the CAS, representing the principal ecologically meaningful structural and functional units within the gut ecosystem. In some embodiments, the use of high-quality metagenome-assembled genomes (HQMAGs) are used as a surrogate to profile these microbial cells, thus providing a more comprehensive and realistic depiction of the microbiome compared to gene-, pathway-, or taxon-centric approaches. This perspective encompasses the complete genetic potential and ecological identity of a microbe, reinforcing the essential ecological axiom that organisms (or more accurately, cells) interact with each other and their environment, rather than genes/pathways or taxa.

[0009] To identify the core constituents of the gut microbiome, a genome-centric, reference-free approach that emphasizes the stability of ecological interactions was used. This methodology involved detecting stable relationships among HQMAGs across varying conditions, with environmental perturbations to the gut ecosystem being introduced via dietary interventions or disease progression. These stable relationships can unveil the core members of the microbiome. This aligns with a foundational principle of systems biology, whereby relationship stability often signifies pivotal system components. In the context of the gut microbiome, these core components are likely to execute essential functions contributing to system resilience and host health, demanding their persistent presence and predictable interaction patterns. Therefore, uncovering these stable relationships could disclose these critical microbial components, potentially exposing the backbone of the ecological network conserved within the gut microbiome, across individuals, populations, or health states.

[0010] A robust seesaw-like network comprising two competing bacterial guilds was identified. This network was discerned by searching for stable genome pairs across co-abundance networks among individuals pre- and post-high fiber intervention (the QD trial, FIG.

1 A), or between healthy and diseased cohorts. This seesaw-like network embodies both cooperative and competitive interactions, potentially indicating a key feature of a stable microbiome structure. The HQMAGs identified within this novel core microbiome demonstrated correlations with various clinical parameters in patients with type 2 diabetes mellitus (T2DM) undergoing a high fiber intervention. Moreover, a universal machine learning model, premised on these HQMAGs in the seesaw-networked core microbiome, successfully differentiated cases from controls in 26 independent datasets spanning 15 different diseases. Furthermore, these HQMAGs supported a machine learning model for predicting personalized treatment responses to immunotherapy in patients with cancer or autoimmune diseases. The disclosure introduces a novel conceptual and analytical paradigm for studying the core gut microbiome. This paradigm provides enhanced health maintenance strategies and disease management, enabling personalized interventions that accommodate the intricate interplay of microbial relationships within the gut ecosystem.

[0011] Accordingly, one aspect of the present disclosure provides methods, and systems for training a model for predicting a subject's response to a therapy. The method includes, at a computer system having at least one processor, and memory storing one or more programs for execution by the one or more processors, obtaining, in electronic form, for each respective training subject in a plurality of training subjects, wherein each respective training subject in the plurality of training subjects has received a therapy for a disorder: (i) a corresponding plurality of genomic abundance values for the respective training subject at a time prior to receiving the therapy, wherein the corresponding plurality of genomic abundance values comprising, for each respective gut microorganism in a plurality of gut microorganisms, a corresponding value for the abundance of the genome of the respective gut microorganism in a corresponding biological sample from the gut of the respective training subject, and (ii) an indication of the respective training subject's response to the therapy of the respective training subject. The method also includes inputting, for each respective training subject in the plurality of training subjects, information about the respective training subject into a model comprising a plurality of parameters, wherein the model applies the plurality of parameters to the information through at least 10,000 computations to obtain a corresponding output for the respective training subject from the model, wherein the corresponding output comprises a prediction of the respective training subject's response to the therapy, the information about the respective training subject

comprises the corresponding genomic abundance value for each respective gut microorganism in the plurality of gut microorganisms, and the plurality of gut microorganisms are selected from Table 1, Table 2, or Figure 13A-13XX. The method also includes adjusting the plurality of parameters based on, for each respective training subject in the first plurality of training subjects, one or more differences between (i) the corresponding output from the model and (ii) the corresponding indication of the respective training subject's response to the therapy.

[0012] Accordingly, another aspect of the present disclosure provides methods, and systems for using a model for predicting a subject's response to a therapy. The method includes, at a computer system having one or more processors, and memory storing one or more programs for execution by the one or more processors, obtaining, in electronic form, a plurality of genomic abundance values comprising, for each respective gut microorganism in a plurality of gut microorganisms selected from Table 1, Table 2, or Figure 13A-13XX, a corresponding abundance value for the genome of the respective species of gut microorganisms, in the plurality of gut microorganisms, in a biological sample from the subject. The method also includes inputting the plurality of genomic abundance values into a model comprising a plurality of parameters, wherein the model applies the plurality of parameters to the plurality of genomic abundance values through at least 10,000 computations to generate as output from the model a prediction of the subject's response to the therapy.

[0013] As disclosed herein, any embodiment disclosed herein when applicable can be applied to any other aspect.

[0014] Additional aspects and advantages of the present disclosure will become readily apparent to those skilled in this art from the following detailed description, where only illustrative embodiments of the present disclosure are shown and described. As will be realized, the present disclosure is capable of other and different embodiments, and its several details are capable of modifications in various obvious respects, all without departing from the disclosure. Accordingly, the drawings and description are to be regarded as illustrative in nature, and not as restrictive.

[0015] Accordingly, one aspect of the invention provides a method of training a model for predicting subject response to a therapy at a computer system having one or more processors, and memory storing one or more programs for execution by the one or more processors.

[0016] In some embodiments, the method includes obtaining, in electronic form, for each respective training subject in a plurality of training subjects, wherein each respective training subject in the plurality of training subjects has received a therapy for a disorder, (i) a corresponding plurality of genomic abundance values for the respective training subject at a time prior to receiving the therapy, wherein the corresponding plurality of genomic abundance values comprise, for each respective gut microorganism in a plurality of gut microorganisms, a corresponding value for the abundance of the genome of the respective gut microorganism in a corresponding biological sample from the gut of the respective training subject, (ii) an indication of the respective training subject's response to the therapy of the respective training subject.

[0017] In some such embodiments, the method includes sequencing, for each respective subject in the plurality of training subjects, genomic DNA from the corresponding biological sample from the gut of the respective training subject, thereby obtain the corresponding plurality of at least 100,000 nucleic acid sequences.

[0018] In some such embodiments, the method includes obtaining, for each respective training subject in the plurality of training subjects, in electronic form, a corresponding plurality of at least 100,000 nucleic acid sequences for genomic DNA from the corresponding biological sample from the gut of the respective training subject.

[0019] In some such embodiments, the method includes determining, for each respective gut microorganism in the plurality of gut microorganisms, the corresponding value for the abundance of the genome of the respective gut microorganism from the corresponding plurality of at least 100,000 nucleic acid sequences.

[0020] In some such embodiments, the method includes, for each respective training subject in the plurality of training subjects, assembling, in electronic form, a corresponding plurality of gut microorganism genomes by metagenomic de novo sequence assembly from the corresponding plurality of at least 100,000 nucleic acid sequences, and calculating, for each respective gut microorganism in the plurality of gut microorganisms, the corresponding value for the abundance of the genome of the respective gut microorganism based on the prevalence of respective nucleic acid sequences, in the plurality of at least 100,000 nucleic acid sequences, used to assemble a respective gut microorganism genome in the plurality of gut microorganism genomes corresponding to the respective gut microorganism.

[0021] In some such embodiments, the method includes, for each respective subject in the plurality of training subjects, assigning each respective nucleic acid sequence in the corresponding plurality of at least 100,000 sequences to a respective gut microorganism in the plurality of gut microorganisms, thereby generating, for each respective gut microorganism in the plurality of gut microorganism, a corresponding count of respective nucleic acid sequences in the corresponding plurality of nucleic acid sequences assigned to the respective gut microorganism, and determining, for each respective gut microorganism in the plurality of gut microorganisms, the corresponding genomic abundance value for the respective gut microorganism based on the corresponding count of respective nucleic acid sequences assigned to the respective gut microorganism.

[0022] In some such embodiments, the plurality of gut microorganisms comprises at least 20 gut microorganisms selected from Table 1, Table 2, or Figure 13A-13XX.

[0023] In some such embodiments, the plurality of gut microorganisms comprises at least 20 microorganisms selected from those microorganisms in Table 1, Table 2, or Figure 13A-13XX having a connectivity of at least 2.

[0024] In some such embodiments, the biological sample from the gut of the respective subject is a fecal sample from the respective training subject.

[0025] In some such embodiments, the therapy is a biological therapy, an immunotherapy, a chemotherapy, a radiation therapy, a gene therapy, a hormone therapy, a photodynamic therapy, a targeted therapy, small molecules, antibodies, polynucleotide, natural compound, immune modulator, bone marrow therapy, stem cell therapy, surgery therapy, induction therapy, maintenance therapy, or a combination thereof.

[0026] In some such embodiments, the disorder is selected from the group consisting of type-2 diabetes, hypertension, schizophrenia, atherosclerotic cardiovascular disease (ACVD), liver cirrhosis (LC), inflammatory bowel diseases (IBD), colorectal cancer (CRC), ankylosing spondylitis (AS), and Parkinson's disease (PD), inflammatory bowel disease (IBD), rheumatoid arthritis (RA), or advanced melanoma and B cell lymphoma.

[0027] In some such embodiments, the disorder is cancer.

[0028] In some embodiments, the method includes inputting, for each respective training subject in the plurality of training subjects, information about the respective training subject into a model comprising a plurality of parameters, wherein the model applies the plurality of parameters to the information through at least 10,000 computations to obtain a corresponding output for the respective training subject from the model, wherein the corresponding output comprises a prediction of the respective training subject's response to the therapy, and the information about the respective training subject comprises the corresponding genomic abundance value for each respective gut microorganism in the plurality of gut microorganisms, and the plurality of gut microorganisms are selected from Table 1, Table 2, or Figure 13A-13XX.

[0029] In some such embodiments, the prediction of the respective training subject's response is a class output of a respective response, in a plurality of possible responses, of the respective training subject.

[0030] In some such embodiments, the prediction of the respective training subject's response is a probability output for the respective training subject's response.

[0031] In some such embodiments, the model is a neural network algorithm, a support vector machine algorithm, a Naive Bayes algorithm, a nearest neighbor algorithm, a boosted trees algorithm, a Random Forest algorithm, a convolutional neural network algorithm, a decision tree algorithm, a regression algorithm, or a clustering algorithm.

[0032] In some such embodiments, the plurality of parameters is at least 1000, at least 10,000, at least 15,000, at least 50,000, at least 100,000, at least 250,000, at least 500,000, or at least 1,000,000 parameters.

[0033] In some such embodiments, the model applies the plurality of parameters to the information through at least 25,000, at least 50,000, at least 100,000, at least 250,000, at least 500,000, or at least 1,000,000 computations to obtain a corresponding output for the respective training subject from the model.

[0034] In some embodiments, the method includes adjusting the plurality of parameters based on, for each respective training subject in the plurality of training subjects, one or more differences between (i) the corresponding output from the model and (ii) the corresponding indication of the respective training subject's response to the therapy.

[0035] Another aspect of the present disclosure provides a method of using a model for predicting a subject's response to a therapy for a disorder at a computer system having one or more processors, and memory storing one or more programs for execution by the one or more processors.

[0036] In some embodiments, the method includes obtaining, in electronic form, a plurality of genomic abundance values comprising, for each respective gut microorganism in a plurality of gut microorganisms selected from Table 1, Table 2, or Figure 13A-13XX, a corresponding abundance value for the genome of the respective species of gut bacteria, in the plurality of gut microorganisms, in a biological sample from the subject.

[0037] In some such embodiments, the method includes sequencing genomic DNA from the biological sample from the gut of the subject, thereby obtaining the plurality of at least 100,000 nucleic acid sequences.

[0038] In some such embodiments, the method includes obtaining, in electronic form, a plurality of at least 100,000 nucleic acid sequences for genomic DNA from the biological sample from the gut of the subject.

[0039] In some such embodiments, the method includes determining, for each respective gut microorganism in the plurality of gut microorganisms, the corresponding value for the abundance of the genome of the respective gut microorganism from the plurality of at least 100,000 nucleic acid sequences.

[0040] In some such embodiments, the method includes assembling, in electronic form, a corresponding plurality of gut microorganism genomes by metagenomic de novo sequence assembly from the plurality of at least 100,000 nucleic acid sequences, and calculating, for each respective gut microorganism in the plurality of gut microorganisms, the corresponding value for the abundance of the genome of the respective gut microorganism based on the prevalence of respective nucleic acid sequences, in the plurality of at least 100,000 nucleic acid sequences, used to assemble a respective gut microorganism genome in the plurality of gut microorganism genomes corresponding to the respective gut microorganism.

[0041] In some such embodiments, the method includes assigning each respective nucleic acid sequence in the plurality of at least 100,000 sequences to a respective gut microorganism in

the plurality of gut microorganisms, thereby generating, for each respective gut microorganism in the plurality of gut microorganism, a corresponding count of respective nucleic acid sequences in the plurality of nucleic acid sequences assigned to the respective gut microorganism, and determining, for each respective gut microorganism in the plurality of gut microorganisms, the corresponding genomic abundance value for the respective gut microorganism based on the corresponding count of respective nucleic acid sequences assigned to the respective gut microorganism.

[0042] In some such embodiments, the plurality of gut microorganisms comprises at least 20 microorganisms selected from those microorganisms in Table 1, Table 2, or Figure 13A- 13XX having a connectivity of at least 2.

[0043] In some such embodiments, the biological sample from the gut of the subject is a fecal sample.

[0044] In some such embodiments, the therapy is a biological therapy, an immunotherapy, a chemotherapy, a radiation therapy, a gene therapy, a hormone therapy, a photodynamic therapy, a targeted therapy, small molecules, antibodies, polynucleotide, natural compound, immune modulator, bone marrow therapy, stem cell therapy, surgery therapy, induction therapy, maintenance therapy, or a combination thereof.

[0045] In some such embodiments, the disorder is selected from the group consisting of type-2 diabetes, hypertension, schizophrenia, atherosclerotic cardiovascular disease (ACVD), liver cirrhosis (LC), inflammatory bowel diseases (IBD), colorectal cancer (CRC), ankylosing spondylitis (AS), and Parkinson's disease (PD), inflammatory bowel disease (IBD), rheumatoid arthritis (RA), advanced melanoma and B cell lymphoma.

[0046] In some such embodiments, the disorder is cancer.

[0047] In some embodiments, the method includes inputting the plurality of genomic abundance values into a model comprising a plurality of parameters, wherein the model applies the plurality of parameters to the plurality of genomic abundance values through at least 10,000 computations to generate as output from the model a prediction of the subject's response to the therapy.

[0048] In some such embodiments, the prediction of the subject's response is a class output of a respective response, in a plurality of possible responses, of the respective subject.

[0049] In some such embodiments, the prediction of the subject's response of the subject is a probability output for the respective subject's response.

[0050] In some such embodiments, the model is a neural network algorithm, a support vector machine algorithm, a Naive Bayes algorithm, a nearest neighbor algorithm, a boosted trees algorithm, a Random Forest algorithm, a convolutional neural network algorithm, a decision tree algorithm, a regression algorithm, or a clustering algorithm.

[0051] In some such embodiments, the plurality of parameters is at least 1000, at least 10,000, at least 15,000, at least 50,000, at least 100,000, at least 250,000, at least 500,000, or at least 1,000,000 parameters.

[0052] In some such embodiments, the model applies the plurality of parameters to the information through at least 25,000, at least 50,000, at least 100,000, at least 250,000, at least 500,000, or at least 1,000,000 computations to obtain a corresponding output for the respective subject from the model.

[0053] In some embodiments, the method includes treating the subject by: i) when the prediction of the subject's response to the therapy satisfies a threshold likelihood that the subject will respond favorably to the therapy, administer the therapy to the subject; ii) when the prediction of the subject's response to the therapy does not satisfy the threshold likelihood that the subject will respond favorably to the therapy, administer one or more of the plurality of gut microorganisms to the subject.

[0054] Another aspect of the present disclosure provides a computer system. The computer system comprises one or more processors and a non-transitory computer-readable medium including computer-executable instructions that, when executed by the one or more processors, cause the processors to perform the method described herein.

[0055] Another aspect of the present disclosure provides a non-transitory computer readable storage medium. The non-transitory computer readable storage medium stores instructions, which when executed by a computer system, cause the computer system to perform any of the methods described herein.

BRIEF DESCRIPTION OF THE DRAWINGS

[0056] Figure 1 illustrates a block diagram of an example computing device, in accordance with some embodiments of the present disclosure.

[0057] Figures 2A, 2B, 2C, and 2D collectively provide a flow chart of processes and features for training a model for predicting a subject's response to a therapy for a disorder, in accordance with some embodiments of the present disclosure.

[0058] Figures 3A, 3B, and 3C collectively provide a flow chart of processes and features for predicting a subject's response to a therapy for a disorder, in accordance with some embodiments of the present disclosure.

[0059] Figures 4A, 4B, 4C, 4D, 4E, 4F, 4G, and 4H collectively illustrate reversible alterations in the gut microbiota induced by a high-fiber diet are associated with corresponding shifts in metabolic phenotypes in patients with Type 2 Diabetes Mellitus (T2DM). **(A)** Study design of the QD trial. During the Run-in period, written informed consent, questionnaire of personal information and HbA1c-based screening were conducted. After Run-in, medical checkup and sample collection were conducted at baseline (M0), three months (M3) after on the high fiber intervention (W) or usual diet (U) and one year (M15) after the high fiber intervention stopped. **(B)** Changes of fiber intake. **(C)** Global changes of the gut microbiome as shown by the principal coordinate analysis based on the Bray-Curtis distance with the 1845 HQMAGs and **(D)** Average Bray-Curtis distance between the groups. PERMANOVA test (9,999 permutations) was performed to compare the groups. * $P < 0.05$ and *** $P < 0.001$. The color intensity of the square showed the magnitude of average Bray-Curtis distance. **(E)** Change of HbA1c, **(F)** The percentage of participants who achieved adequate glycemic control, **(G)** Fasting blood glucose, and **(H)** The glucose area under the curve (AUC) in a meal tolerance test (MTT). For **(E)**, **(G)** and **(H)**, data shown as percent changes from baseline ($\pm$ S.E.M). Friedman test followed by Nemenyi post-hoc test was used for comparison in the same group, compact letters reflect significance ($P < 0.05$). $n = 67$ in W group and $n = 28$ in U group. Mann-Whitney test (two-sided) was used for comparison between W and U at the same time point, * $P < 0.05$, ** $P < 0.01$ and *** $P < 0.001$. $n = 74$ in W (M0) (For panel H, $n=72$), $n = 74$ in W (M3), $n = 67$ in W (M15), $n = 36$ in U (M0), $n = 36$ in U (M3) and $n = 28$ in U (M15).

[0060] Figures 5A and 5B collectively illustrate that despite substantial global changes in the gut microbiota induced by the high-fiber intervention, two competing bacterial guilds, which are associated with HbA1c levels, form a robust seesaw-like network within the ecosystem. **(A)** The distribution of different types of correlations of the genome pairs during the trial. The 3 letters show the correlations (N for negative, P for positive and U for un-correlated) of the genome pairs at M0, M3 and M15 subsequently. Stable correlations, NNN and PPP, were highlighted. **(B)** Correlations between genome clusters and HbA1c using linear mixed effect model by MaAslin2 package. Abundance was log transformed. Subject was used as random effect. N = 67. * BH adjusted P < 0.05, *** BH adjusted P < 0.001.

[0061] Figures 6A, 6B, 6C1, 6C2, 6D, 6E1, 6E2, 6E3, 6E4, 6E5, 6E6, 6E7, 6E8, 6E9, 6E10, and 6E11 collectively illustrate Genomes within the two competing guilds predict metabolic health outcomes in T2DM patients of the QD trial, and distinguish cases from controls across seven diseases in eleven independent case-control metagenomic datasets (Case-Control Dataset Collection I). **(A)** Change of the mean range-scaled robust centered-log-ratio (rcrlr) transformed abundances of Guild 1, Guild 2, and their ratio across the trial in the W group. Friedman test followed by Nemenyi test was used to analyze the difference between time points. Compact letters reflect the significance at P < 0.05. **(B)** Prediction of clinical parameters with the genomes in the two competing guilds. Linear mixed effect models were trained based on the mean range-scaled rcrlr abundances of Guild 1 and Guild 2, and clinical parameters at M0 and M3 with subject as random effect. The mean range-scaled rcrlr abundances of Guild 1 and Guild 2 at M15 were used to predict clinical parameters based on the trained model. The bar plot shows the Pearson's correlations coefficient between the predicted clinical parameters and measured values at M15. The asterisk before the parameter's name shows the significance of the Pearson's correlations. P values were adjusted by Benjamini & Hochberg's method. * adjusted P < 0.05, ** adjusted P < 0.01 and *** adjusted P < 0.001. BMI, body mass index; SBP, systolic blood pressure; DBP, diastolic blood pressure; WC, waist circumference; HP, hip circumference; TNF- α , tumor necrosis factor- α ; WBC, white blood cell count; CRP, C-reactive protein; LBP, lipopolysaccharide-binding protein; TC, total cholesterol; TG, triglyceride; Lpa, lipoprotein a; HDL, high-density lipoprotein; APOA, apolipoprotein A; LDL, low-density lipoprotein; APOB, apolipoprotein B; GFR (MDRR), glomerular filtration rate; CysC, Cystatin C; ACR, urinary microalbumin to creatinine ratio; IMT, intima-media thickness; DAN, diabetic autonomic

neuropathy score; MHR, mean heart rate; SDNN, standard deviation of NN intervals; SDANN, standard deviation of the average NN intervals calculated over 5 minutes; SDNNIndex, mean of standard deviation of NN intervals for 5-minute segments; rMSSD, root-mean-square of the differences of successive NN intervals; pNN50, percentage of the interval differences of successive NN intervals greater than 50 ms; TP, total power; VLF, very low frequency power; LF, low frequency power; HF, high frequency power; DPN, diabetic peripheral neuropathy score. (C) Differences in genetic capacity of carbohydrate substrate utilization (CAZy), short-chain fatty acid production (SCFA), antibiotic resistance genes (ARG) and virulence factor genes (VF). The heatmaps show the proportion (CAZy) or gene copy numbers (SCFA, ARG and VF) of each category in each genome. For carbohydrate substrate utilization, CAZy genes were predicted in each genome. The proportion of CAZy genes for a particular substrate was calculated as the number of the CAZy genes involved in its utilization divided by the total number of the CAZy genes. Arabinoxylan-related CAZy families: CE1, CE2, CE4, CE6, CE7, GH10, GH11, GH115, GH43, GH51, GH67, GH3 and GH5; cellulose-related: GH1, GH44, GH48, GH8, GH9, GH3 and GH5; inulin-related: GH32 and GH91; mucin-related families: GH1, GH2, GH3, GH4, GH18, GH19, GH20, GH29, GH33, GH38, GH58, GH79, GH84, GH85, GH88, GH89, GH92, GH95, GH98, GH99, GH101, GH105, GH109, GH110, GH113, PL6, PL8, PL12, PL13 and PL21; pectin-related: CE12, CE8, GH28, PL1 and PL9; starch-related: GH13, GH31 and GH97. For short chain fatty acid production, FTHFS: formate-tetrahydrofolate ligase for acetate production; ScpC: propionyl-CoA succinate-CoA transferase and Pct: propionate-CoA transferase for propionate production; But: Butyryl-coenzyme A (butyryl-CoA): acetate CoA transferase, Buk: butyrate kinase, 4Hbt: butyryl-CoA: 4-hydroxybutyrate CoA transferase, Ato: butyryl-CoA:acetoacetate CoA transferase (AtoA: alpha subunit, AtoD: beta subunit) for butyrate production. Mann-Whitney test (two-sided) was used to analyze the difference between Guild 1 and Guild 2. # $P < 0.1$, * $P < 0.05$, ** $P < 0.01$ and *** $P < 0.001$. Number of genomes in Guild 1 (green bar): $n = 50$, in Guild 2 (purple bar): $n = 91$. (D) Datasets from 11 different datasets on 7 diseases including type 2 diabetes (T2D), liver cirrhosis (LC), ankylosing spondylitis (AS), atherosclerotic cardiovascular disease (ACVD), schizophrenia (SCZ), colorectal cancer (CRC), inflammatory bowel disease (IBD) dataset were collected as Case-Control Dataset Collection I. Sample size of each case-control dataset is showed in red and green number in the left panel respectively. For each dataset in the collection, metagenomic reads were

recruited to the 141 genomes of the QD two competing guilds to estimate genome abundance in each sample (E) Based on the abundance matrix of the 141 genomes, a Random Forest classification model with leave-one-out cross validation was trained to classify case and control subjects in each dataset. ROC curves and the area under the curves (AUC) were shown here.

[0062] Figures 7A and 7B collectively illustrate genomes forming the two competing guilds, as identified from a case-control dataset specific to one disease, demonstrate significant effectiveness in classifying cases from controls across independent datasets on different diseases within the Case-Control Dataset Collection I. **(A)** Identification of the seesaw networked two competing guilds in Case-Control Dataset Collection I. Sample size of each case-control study is showed in red and green number in the left panel respectively. Case-Control Dataset Collection I has 11 published metagenomic case-control datasets on 7 diseases including type 2 diabetes (T2D), liver cirrhosis (LC), ankylosing spondylitis (AS), atherosclerotic cardiovascular disease (ACVD), schizophrenia (SCZ), colorectal cancer (CRC), inflammatory bowel disease (IBD) dataset. Datasets from 3 studies were combined to analyze CRC. Datasets from 2 studies were combined to analyze IBD. The percentage of correlations followed the pattern in the seesaw networked two competing guilds (i.e., positive edges within each guild, negative edges between the 2 guilds) was in yellow, and the ratio of correlations that were negative within each guild and positive between the guilds was in black of the 100% stacked bar. **(B)** The two competing guilds found in one dataset was used as predictors to classify case and control in the same dataset and in all the other 10 datasets. Random Forest classification model with leave-one-out cross validation was applied in each dataset on each set of the two competing guilds. The area under the ROC curve (AUC) values were showed in the heatmap.

[0063] Figures 8A, 8B1, 8B2, 8B3, 8B4, 8B5, 8B6, 8B7, 8B8, 8B9, 8B10, 8B11, 8B12, 8B13, 8B13, 8B14, 8B15, 8B16, 8C1, and 8C2 collectively illustrate the combined core genomes, drawn from all identified competing guilds, effectively differentiate cases from controls across a broader range of diseases, and predict treatment outcomes in independent datasets. **(A)** Identification of the combined genomes and the combined core genomes from the 8 sets of the two competing guilds, which were generated from QD dataset and Case-Control Dataset Collection I. All the HQMAGs in each set of the two competing guilds were dereplicated based on the cutoff of 99% average nucleotide identity (ANI) between two genomes. 788 non-redundant HQMAGs were obtained as the combined genomes of all the 8 sets of the two

competing guilds. Random forest classification model with leave-one-out cross validation was constructed based on the 788 HQMAGs in each dataset. The HQMAGs were ranked based on their importance across all the models. From the least important HQMAGs (biggest importance rank), subsequently removing one HQMAGs to do random forest classification model in each dataset. In each dataset, rank the HQMAG number based on the area under the ROC curve (AUC) values. The scatter plot shows the relationship between HQMAG number and model performance. The y axis is the sum of rank based on AUC values (the smaller the value, the better the performance). 302 HQMAG reached best performance. After excluding 18 HQMAGs that exhibited inconsistent C1A and C1B assignments across the datasets, a total of 284 HQMAG were kept from the 302 HQMAG as the Combined Core genomes of all the 8 sets of the two competing guilds. **(B)** The Combined Core were used as predictors in Case-Control Dataset Collection II which has 15 published metagenomic case-control datasets on 10 diseases including 1 dataset on ankylosing spondylitis (AS#2): Case n = 85, Control n = 55, 1 on autism spectrum disorder (ASD): Case n = 64, Control n = 64, 1 on Behcet's disease (BD): Case n = 24, Control n = 52, 1 on COVID-19: Case n = 47, Control n = 19, 3 on colorectal cancer (CRC): CRC#4 Case n = 40, Control n = 40; CRC#5 Case n = 61, Control n = 52; CRC#6 Case n = 52, Control n = 52; 1 on Graves' disease (GD), Case n = 88, Control n = 62, 2 on hypertension (HT): HT#1 Case n = 60, Control n = 56, HT#2 Case n = 99, Control n = 41; 1 on multiple sclerosis (MS): Case n = 24, Control n = 24, 3 on pancreatic cancer (PC): PC#1 Case n = 43, Control n = 235, PC#2 Case n = 57, Control n = 50, PC#3 Case n = 44, Control n = 32; and 1 on Parkinson's disease (PD): Case n = 39, Control n = 40. Random Forest classification model with leave-one-out cross validation was applied in each dataset. **(C)** The Combined Core genomes were used as predictors in Treatment Dataset Collection to predict responder (R) and non-responder (NR) under a treatment. For inflammatory bowel disease (IBD), 14-week remission was used to determine R and NR of patients with IBD to anti-cytokine or anti-antigen treatment. IBD_anti-cytokine: R n = 29, NR n = 18; IBD_anti-integrin#1, R n = 27, NR n = 40; IBD_anti-integrin#2: R n = 29, NR n = 53. For rheumatoid arthritis (RA), responder to methotrexate (MTX) was defined a priori as any patient with new-onset RA with an improvement in the Disease Activity Score in 28 joints (DAS28) of ≥ 1.8 by month 4 after initiation of MTX monotherapy. R n = 19, NR n = 28. For advanced melanoma, progression-free survival was used to determined R and NR to immune checkpoint inhibitor (ICI) treatment. AM_ICI#1: R n = 4, NR n = 7, AM_ICI#2: R n

= 10, NR n = 8, AM_ICI#3: R n = 12, NR n = 13, AM_ICI#4: R n = 25, NR n = 30, AM_ICI#5: R n = 26, NR n = 28. For B cell lymphoma, tumor response to CAR-T cell immunotherapy was classified as either complete remission or non-complete remission (partial remission, stable disease, progressive disease or death) at 180 days after CAR-T cell infusion by the treating physician. Model was trained on German cohort and validated by US cohort. Germany: R n = 21, NR n = 29; US: R n = 21, NR n = 24.

[0064] Figures 9A, 9B, and 9C collectively illustrate the discriminative power of the combined core genomes from all the 8 sets of the two competing guilds in classifying healthy individuals vs. patients across colorectal cancer (CRC), inflammatory bowel diseases (IBD), and Pancreatic Cancer (PC) datasets in the Case-Control Dataset Collection I and II. A prediction matrix was shown for the classification of cases and controls based on the combined core genomes from all eight sets of the two competing guilds within each dataset (diagonal values), across pairs of datasets (one dataset used for model training and the other for testing), and in a leave-one-dataset-out setting (training the model on all but one dataset and testing on the left-out dataset). Random Forest classification model with leave-one-out cross validation was applied. The area under the ROC curve (AUC) values were shown in the matrix. **(A)** CRC.#1: case n = 74, control n = 54;#2: case n = 46, control n = 63;#3: case n = 22, control n = 60;#4: case n = 40, control n = 40;#5: case n = 61, control n = 52;#6: case n = 52, control n = 52. **(B)** IBD. #1: case n = 80 control n = 26; #2: case n = 121, control n = 34;#3: case n = 43, control n = 22; **(C)** PC.#1: case n = 43 control n = 235;#2: case n = 57 control n = 50; #3: case n = 44 control n = 32.

[0065] Figures 10A1, 10A2, 10B1, 10B2, 10C1, 10C2, 10D1, and 10D2 collectively illustrate the combined core of the two competing guilds supports the prediction of therapeutic effects in the Treatment Dataset Collection for inflammatory bowel diseases, rheumatoid arthritis, advanced melanoma, and B cell lymphoma. The abundance of the combined core genomes (284 HQMAGs) in the pre-treatment samples were used as predictors in Random Forest classification models to predict responder (R) and non-responder (NR) under treatment. Area under the ROC curve (AUC) and AUC values were showed in the panels. **(A)** 14-week remission was used to determine R and NR. IBD_anti-cytokine, R n = 29, NR n = 18; IBD_anti-integrin#1, R n = 27, NR n = 40; IBD_anti-integrin#2, R n = 29, NR n = 53. **(B)** Responder to MTX was defined a priori as any patient with new-onset RA with an improvement in the Disease

Activity Score in 28 joints (DAS28) (25) of ≥ 1.8 by month 4 after initiation of MTX monotherapy. R n = 19, NR n = 28. **(C)** Overall response Rate (ORR, left matrix) and progression-free survival (PFS12, right matrix) was used to determined R and NR, respectively. Prediction matrix for microbiome-based prediction of response assessed via ORR (left matrix) and PFS12 (right matrix) within each cohort (values on the diagonal), across pairs of cohorts (one cohort used to train the model and the other for testing) and in the leave-one-cohort-out setting (training the model on all but one cohort and testing on the left-out cohort). ORR: R n = 94, NR n = 71; PFS12: R n = 77, NR n = 86. **(D)** Tumor response to CAR-T cell immunotherapy was classified as either complete remission or non-complete remission (partial remission, stable disease, progressive disease or death) at 180 days after CAR-T cell infusion by the treating physician. Model was trained on #1 (German cohort) and validated by #2 (US cohort). #1: R n = 21, NR n = 29; #2: R n = 21, NR n = 24.

[0066] Figures 11A, 11A2, 11B1, 11B2, 11C1, 11C2, 11D1, and 11D2 collectively illustrate the Combined Core genomes of the two competing guilds provide a universal model for distinguishing between cases and controls across a variety of diseases (Case-Control Dataset Collection I and II). **(A)** All control and case samples from Case-Control Dataset Collection I and II, encompassing a total of 26 datasets on 15 different diseases, were combined and randomly allocated, with 80% used for training a Random Forest classification model and 20% for testing. **(B)** ROC curve and the area under the ROC curve (AUC). **(C)** The density plot of the probability score of between case and control. The probability score was generated from the Random Forest classification model and showed the probability of one sample to be predicted as case. **(D)** Box plot of the probability score between control and case samples. Mann-Whitney test was applied. *** $P < 0.001$. Training: control n = 1,285, case n = 1424; testing: control n = 319, case n = 356.

[0067] Figures 12A, 12B, 12C, 12D, 12E, 12F, 12G, 12H, and 12I collectively illustrate the corresponding contigs, referenced by SEQ IDs, obtained for each of the 788 genomes.

[0068] Figures 13A, 13B, 13C, 13D, 13E, 13F, 13G, 13H, 13I, 13J, 13K, 13L, 13M, 13N, 13O, 13P, 13Q, 13R, 13S, 13T, 13U, 13V, 13W, 13X, 13Y, 13Z, 13AA, 13BB, 13CC, 13DD, 13EE, 13FF, 13GG, 13HH, 13II, 13JJ, 13KK, 13LL, 13MM, 13NN, 13OO, 13PP, 13QQ, 13RR,

13SS, 13TT, 13UU, 13VV, 13WW and 13XX collectively illustrate the Taxonomy Assignment of the 788 genomes in the combined microbiome.

[0069] Figures 14A and 14B collectively illustrate that the CC-TCGs predict treatment outcomes in independent datasets. CC-TCG was used as a predictor in the treatment dataset collection to predict whether respondents (R) and non-responders (NR) were under treatment. For inflammatory bowel disease (IBD), 14-week remission was used to determine the R and NR of patients with IBD to anti-cytokine or anti-antigen treatment. For rheumatoid arthritis (RA), responder to methotrexate (MTX) was defined a priori as any patient with new-onset RA with an improvement in the disease activity score in 28 joints (DAS28) of R1.8 by month four after initiation of MTX monotherapy. For advanced melanoma (AM), progression-free survival was used to determine R and NR to immune checkpoint inhibitor (ICI) treatment. For B cell lymphoma, the treating physician classified tumor response to CAR-T cell immunotherapy as complete or non-complete (partial remission, stable disease, progressive disease, or death) at 180 days after CAR-T cell infusion. The model was trained on the German cohort and validated by the US cohort. A sample size of R and NR is listed.

[0070] Figure 15A, 15B, 15C, and 15D collectively illustrate that the combined core of the TCGs supports the prediction of therapeutic effects in the treatment dataset collection for inflammatory bowel disease, rheumatoid arthritis, advanced melanoma, and B cell lymphoma. The abundance of the combined core genomes (284 HQMAGs) in the pre-treatment samples was used as predictors in random forest classification models to predict responder (R) and non-responder (NR) under treatment. Area under the ROC curve (AUC) and AUC values were shown in the panels. (15A) 14-week remission was used to determine R and NR. IBD_anti-cytokine, n = 29 for R, n = 18 for NR; IBD_anti-integrin#1, R n = 27, NR n = 40; and IBD_antiintegrin#2, R n = 29, NR n = 53. (15B) Responder to MTX was defined a priori as any patient with new-onset RA with an improvement in the disease activity score in 28 joints (DAS28) (25) of R1.8 by month 4 after initiation of MTX monotherapy. R n = 19, NR n = 28. (15C) Overall response rate (ORR, left matrix) and progression-free survival (PFS12, right matrix) were used to determine R and NR, respectively. Prediction matrix for microbiome-based prediction of response assessed via ORR (left matrix) and PFS12 (right matrix) within each cohort (values on the diagonal) across pairs of cohorts (one cohort was used to train the model and the other for testing) and in the leave-one-cohort-out setting (training the model on all but one cohort and testing on the left-

out cohort). ORR: R n = 94, NR n = 71; PFS12: R n = 77, NR n = 86. (15D) Tumor response to CAR-T cell immunotherapy was classified as either complete remission or non-complete remission (partial remission, stable disease, progressive disease, or death) at 180 days after CAR-T cell infusion by the treating physician. Model was trained on #1 (German cohort) and validated by #2 (US cohort). #,1: R n = 21, NR n = 29; #,2: R n = 21, NR n = 24.

[0071] Figures 16A and 16B collectively illustrate guild-level functional consistency among the 8 TCGs found in the QD trial and CCDC-I. (16A) PCoA plot based on the Euclidean distance showed significant separations between C1As and C1Bs from the eight sets of TCGs. (16B) Ward linkage based on Euclidean distance showed separations between C1As and C1Bs from the eight sets of TCGs. Euclidean distance was calculated based on the mean copies of KEGG Orthology in each guild.

[0072] Figures 17A, 17B, 17C, 17D, 17E, and 17F collectively illustrate differences in genes related to carbohydrate utilization and SCFA production between C1A and C1B. (17A, 17C, and 17E) Bar plot shows the proportion of CAZy genes for a particular substrate, calculated as the number of CAZy genes involved in its utilization divided by the total number of CAZy genes. Arabinoxylan-related CAZy families: CE1, CE2, CE4, CE6, CE7, GH10, GH11, GH115, GH43, GH51, GH67, GH3, and GH5; cellulose-related: GH1, GH44, GH48, GH8, GH9, GH3, and GH5; inulin-related: GH32 and GH91; mucin-related: GH1, GH2, GH3, GH4, GH18, GH19, GH20, GH29, GH33, GH38, GH58, GH79, GH84, GH85, GH88, GH89, GH92, GH95, GH98, GH99, GH101, GH105, GH109, GH110, GH113, PL6, PL8, PL12, PL13, and PL21; pectin-related: CE12, CE8, GH28, PL1, and PL9; and starch-related: GH13, GH31, and GH97. (17B, 17D, and 17F) Barplot showed the copy number of genes encoding FTHFS: formate-tetrahydrofolate ligase for acetate production; ScpC: propionyl-coenzyme A (CoA) succinate-CoA transferase and Pct: propionate-CoA transferase for propionate production; But, butyryl-coenzyme A (butyryl-CoA): acetate CoA transferase; Buk, butyrate kinase; 4Hbt, butyryl-CoA: 4-hydroxybutyrate CoA transferase; and Ato, butyryl-CoA:acetoacetate CoA transferase (AtoA: alpha subunit, AtoD: beta subunit) for butyrate production. Mann-Whitney test (two-sided) was used to analyze the difference between C1A 1 and C1B 2. # $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

[0073] Like reference numerals refer to corresponding parts throughout the several views of the drawings.

DETAILED DESCRIPTION

[0074] The methods and systems described herein facilitate prediction of a subject's response to a therapy for a disorder based on the constitution of the subject's microbiome.

[0075] *Definitions.*

[0076] The terminology used in the present disclosure is for the purpose of describing particular embodiments only and is not intended to be limiting of the invention. As used in the description of the invention and the appended claims, the singular forms "a", "an" and "the" are intended to include the plural forms as well, unless the context clearly indicates otherwise. It will also be understood that the term "and/or" as used herein refers to and encompasses any and all possible combinations of one or more of the associated listed items. It will be further understood that the terms "includes," "comprising," or any variation thereof, when used in this specification, specify the presence of stated features, integers, steps, operations, elements, and/or components, but do not preclude the presence or addition of one or more other features, integers, steps, operations, elements, components, and/or groups thereof. Furthermore, to the extent that the terms "including," "includes," "having," "has," "with," or variants thereof are used in either the detailed description and/or the claims, such terms are intended to be inclusive in a manner similar to the term "comprising."

[0077] As used herein, the term "if" may be construed to mean "when" or "upon" or "in response to determining" or "in response to detecting," depending on the context. Similarly, the phrase "if it is determined" or "if [a stated condition or event] is detected" may be construed to mean "upon determining" or "in response to determining" or "upon detecting [the stated condition or event]" or "in response to detecting [the stated condition or event]," depending on the context.

[0078] It will also be understood that, although the terms first, second, etc. may be used herein to describe various elements, these elements should not be limited by these terms. These terms are only used to distinguish one element from another. For example, a first subject could be termed a second subject, and, similarly, a second subject could be termed a first subject, without departing from the scope of the present disclosure. The first subject and the second

subject are both subjects, but they are not the same subject. Furthermore, the terms "subject," "user," and "patient" are used interchangeably herein.

[0079] As used herein, the term “measure of central tendency” refers to a central or representative value for a distribution of values. Non-limiting examples of measures of central tendency include an arithmetic mean, weighted mean, midrange, midhinge, trimean, geometric mean, geometric median, Winsorized mean, median, and mode of the distribution of values.

[0080] As used herein, the term “subject” refers to any living or non-living organism including, but not limited to, a human (*e.g.*, a male human, female human, fetus, pregnant female, child, or the like), a non-human mammal, or a non-human animal. Any human or non-human animal can serve as a subject, including but not limited to mammal, reptile, avian, amphibian, fish, ungulate, ruminant, bovine (*e.g.*, cattle), equine (*e.g.*, horse), caprine and ovine (*e.g.*, sheep, goat), swine (*e.g.*, pig), camelid (*e.g.*, camel, llama, alpaca), monkey, ape (*e.g.*, gorilla, chimpanzee), ursid (*e.g.*, bear), poultry, dog, cat, mouse, rat, fish, dolphin, whale and shark. In some embodiments, a subject is a male or female of any age (*e.g.*, a man, a woman, or a child).

[0081] As used herein, the term “administering” with respect to the methods of the invention, means a method for therapeutically or prophylactically preventing, treating or ameliorating a syndrome, disorder or disease as described herein. Such methods include administering an effective amount of said therapeutic agent at different times during the course of a therapy or concurrently in a combination form. The methods of the invention are to be understood as embracing all known therapeutic treatment regimens.

[0082] As used herein, the term “cancer,” “cancerous tissue,” or “tumor” refers to an abnormal mass of tissue in which the growth of the mass surpasses, and is not coordinated with, the growth of normal tissue, including both solid masses (*e.g.*, as in a solid tumor) or fluid masses (*e.g.*, as in a hematological cancer). A cancer or tumor can be defined as “benign” or “malignant” depending on the following characteristics: degree of cellular differentiation including morphology and functionality, rate of growth, local invasion and metastasis. A “benign” tumor can be well differentiated, have characteristically slower growth than a malignant tumor and remain localized to the site of origin. In addition, in some cases a benign tumor does not have the capacity to infiltrate, invade or metastasize to distant sites. A

“malignant” tumor can be a poorly differentiated (anaplasia), have characteristically rapid growth accompanied by progressive infiltration, invasion, and destruction of the surrounding tissue. Furthermore, a malignant tumor can have the capacity to metastasize to distant sites. Accordingly, a cancer cell is a cell found within the abnormal mass of tissue whose growth is not coordinated with the growth of normal tissue. Accordingly, a “tumor sample” refers to a biological sample obtained or derived from a tumor of a subject, as described herein.

[0083] Non-limiting examples of cancer types include ovarian cancer, cervical cancer, uveal melanoma, colorectal cancer, chromophobe renal cell carcinoma, liver cancer, endocrine tumor, oropharyngeal cancer, retinoblastoma, biliary cancer, adrenal cancer, neural cancer, neuroblastoma, basal cell carcinoma, brain cancer, breast cancer, non-clear cell renal cell carcinoma, glioblastoma, glioma, kidney cancer, gastrointestinal stromal tumor, medulloblastoma, bladder cancer, gastric cancer, bone cancer, non-small cell lung cancer, thymoma, prostate cancer, clear cell renal cell carcinoma, skin cancer, thyroid cancer, sarcoma, testicular cancer, head and neck cancer (*e.g.*, head and neck squamous cell carcinoma), meningioma, peritoneal cancer, endometrial cancer, pancreatic cancer, mesothelioma, esophageal cancer, small cell lung cancer, Her2 negative breast cancer, ovarian serous carcinoma, HR+ breast cancer, uterine serous carcinoma, uterine corpus endometrial carcinoma, gastroesophageal junction adenocarcinoma, gallbladder cancer, chordoma, and papillary renal cell carcinoma.

[0084] As used herein, the terms “cancer state” or “cancer condition” refer to a characteristic of a cancer patient's condition, *e.g.*, a diagnostic status, a type of cancer, a location of cancer, a primary origin of a cancer, a cancer stage, a cancer prognosis, and/or one or more additional characteristics of a cancer (*e.g.*, tumor characteristics such as morphology, heterogeneity, size, *etc.*). In some embodiments, one or more additional personal characteristics of the subject are used further describe the cancer state or cancer condition of the subject, *e.g.*, age, gender, weight, race, personal habits (*e.g.*, smoking, drinking, diet), other pertinent medical conditions (*e.g.*, high blood pressure, dry skin, other diseases), current medications, allergies, pertinent medical history, current side effects of cancer treatments and other medications, *etc.*

[0085] As used herein, the term “treat”, “treating”, “treatment”, or “therapy”, refers to both therapeutic treatment and prophylactic or preventative measures, wherein the object is to prevent

or slow down (lessen) the targeted pathologic condition or disorder. Those in need of treatment include those diagnosed with the disorder as well as those prone to have the disorder (e.g., a genetic predisposition) or those in whom the disorder is to be prevented. The terms “prevent,” “preventing,” and “prevention” refer to reducing the likelihood of the onset (or recurrence) of a disease, disorder, condition, or associated symptom(s). The term means obtaining beneficial or desired results, for example, clinical results. Beneficial or desired results can include, but are not limited to, alleviation of one or more symptoms. The term “alleviation”, for example in reference to a symptom of a condition, as used herein, refers to reducing at least one of the frequency and amplitude of a symptom of a condition in a patient.

[0086] As used herein, the "response" , as used herein, refers to the response to a biological drug, chemical drug, or physical therapy of the subject suffering from a pathology which is treatable with said biological drug, chemical drug, or physical therapy. Standard criteria may vary from disease to disease.

[0087] As used herein, “immunotherapies” are all therapies that either directly or indirectly modify the immune response or the immune system of a patient. For immunotherapeutic strategies, it has been found that the detection of a strong immune response at the tumor site was a reliable marker for a plurality of cancers, like colon cancers as well as rectum cancers, this association of a pre-existing immune response with a better therapeutic efficacy was assumed. Immune response encompasses any form of immune response of said patient through direct or indirect, or both, action towards said cancer or tumor sites. The immune response means the immune response of the host cancer patient in reaction to the tumor and encompasses the presence of, the number of, or alternatively the activity of, cells and related signaling molecules involved in the immune response of the host which includes: all cytokines, chemokines, growth factors, stem cell growth factors. In some embodiments, the immune response encompasses a multitude of different cellular subtypes, such as T cell lineage, the B cell lineage, the natural killer cells, macrophages, dendritic cells, myelo-derived suppressor cells, lytic dendritic cells, fibroblasts, endothelial cells, as well as an enormous number of signaling molecules (cytokines, chemokines, other signaling molecules).

[0088] As used herein, “immunotherapeutic agent,” refers to a compound, composition or treatment that indirectly or directly enhances, stimulates, or augments the body's immune

response against cancer cells and/or that lessens the side effects of other anticancer therapies. Immunotherapy is thus a therapy that directly or indirectly stimulates or enhances the immune system's responses to cancer cells and/or lessens the side effects that may have been caused by other anti-cancer agents. Immunotherapy is also referred to in the art as immunologic therapy, biological therapy biological response modifier therapy and biotherapy. Examples of common immunotherapeutic agents known in the art include, but are not limited to, cytokines, cancer vaccines, monoclonal antibodies, and non-cytokine adjuvants. Alternatively the immunotherapeutic treatment may consist of administering the patient with an amount of immune cells (T cells, NK, cells, dendritic cells, B cells...).

[0089] Immunotherapeutic agents can be non-specific, i.e. boost the immune system generally so that it becomes more effective in fighting the growth and/or spread of cancer cells, or they can be specific, i.e. targeted to the cancer cells themselves immunotherapy regimens may combine the use of non-specific and specific immunotherapeutic agents. Non-specific immunotherapeutic agents are substances that stimulate or indirectly augment the immune system. Non-specific immunotherapeutic agents have been used alone as the main therapy for the treatment of cancer, as well as in addition to a main therapy, in which case the non-specific immunotherapeutic agent functions as an adjuvant to enhance the effectiveness of other therapies (e.g. cancer vaccines). Non-specific immunotherapeutic agents can also function in this latter context to reduce the side effects of other therapies, for example, bone marrow suppression induced by certain chemotherapeutic agents. Non-specific immunotherapeutic agents can act on key immune system cells and cause secondary responses, such as increased production of cytokines and immunoglobulins. Alternatively, the agents can themselves comprise cytokines. Non-specific immunotherapeutic agents are generally classified as cytokines or non-cytokine adjuvants.

[0090] A number of cytokines have found application in the treatment of cancer either as general non-specific immunotherapies designed to boost the immune system, or as adjuvants provided with other therapies. Suitable cytokines include, but are not limited to, interferons, interleukins and colony-stimulating factors.

[0091] Interferons (IFNs) contemplated by the present invention include the common types of IFNs, IFN-alpha (IFN-a), IFN-beta (IFN-beta) and IFN-gamma (IFN-y). IFNs can act directly

on cancer cells, for example, by slowing their growth, promoting their development into cells with more normal behavior and/or increasing their production of antigens thus making the cancer cells easier for the immune system to recognize and destroy. IFNs can also act indirectly on cancer cells, for example, by slowing down angiogenesis, boosting the immune system and/or stimulating natural killer (NK) cells, T cells and macrophages. Recombinant IFN- α is available commercially as Roferon (Roche Pharmaceuticals) and Intron A (Schering Corporation). The use of IFN- α , alone or in combination with other immunotherapeutics or with chemotherapeutics, has shown efficacy in the treatment of various cancers including melanoma (including metastatic melanoma), renal cancer (including metastatic renal cancer), breast cancer, prostate cancer, and cervical cancer (including metastatic cervical cancer).

[0092] Interleukins contemplated by the present invention include IL-2, IL-4, IL-11 and IL-12. Examples of commercially available recombinant interleukins include Proleukin® (IL-2; Chiron Corporation) and Neumega® (IL-12; Wyeth Pharmaceuticals). Zymogenetics, Inc. (Seattle, Wash.) is currently testing a recombinant form of IL-21, which is also contemplated for use in the combinations of the present invention. Interleukins, alone or in combination with other immunotherapeutics or with chemotherapeutics, have shown efficacy in the treatment of various cancers including renal cancer (including metastatic renal cancer), melanoma (including metastatic melanoma), ovarian cancer (including recurrent ovarian cancer), cervical cancer (including metastatic cervical cancer), breast cancer, colorectal cancer, lung cancer, brain cancer, and prostate cancer. Interleukins have also shown good activity in combination with IFN- α in the treatment of various cancers (Negrier et al., *Ann Oncol.* 2002 13(9):1460-8; Tourani et al., *J Clin Oncol.* 2003 21(21):3987-94).

[0093] Colony-stimulating factors (CSFs) contemplated by the present invention include granulocyte colony stimulating factor (G-CSF or filgrastim), granulocyte-macrophage colony stimulating factor (GM-CSF or sargramostim) and erythropoietin (epoetin alfa, darbepoietin). Treatment with one or more growth factors can help to stimulate the generation of new blood cells in patients undergoing traditional chemotherapy. Accordingly, treatment with CSFs can be helpful in decreasing the side effects associated with chemotherapy and can allow for higher doses of chemotherapeutic agents to be used. Various recombinant colony stimulating factors are available commercially, for example, Neupogen® (G-CSF; Amgen), Neulasta (pelfilgrastim; Amgen), Leukine (GM-CSF; Berlex), Procrit (erythropoietin; Ortho Biotech), Epogen

(erythropoietin; Amgen), Arnesp (erythropoietin). Colony stimulating factors have shown efficacy in the treatment of cancer, including melanoma, colorectal cancer (including metastatic colorectal cancer), and lung cancer.

[0094] Non-cytokine adjuvants suitable for use in the combinations of the present invention include, but are not limited to, Levamisole, alum hydroxide (alum), bacillus Calmette-Guerin (ACG), incomplete Freund's Adjuvant (IFA), QS-21, DETOX, Keyhole limpet hemocyanin (KLH) and dinitrophenyl (DNP). Non-cytokine adjuvants in combination with other immuno- and/or chemotherapeutics have demonstrated efficacy against various cancers including, for example, colon cancer and colorectal cancer (Levamisole); melanoma (BCG and QS-21); renal cancer and bladder cancer (BCG).

[0095] In addition to having specific or non-specific targets, immunotherapeutic agents can be active, i.e. stimulate the body's own immune response, or they can be passive, i.e. comprise immune system components that were generated external to the body.

[0096] Passive specific immunotherapy typically involves the use of one or more monoclonal antibodies that are specific for a particular antigen found on the surface of a cancer cell or that are specific for a particular cell growth factor. Monoclonal antibodies may be used in the treatment of cancer in a number of ways, for example, to enhance a subject's immune response to a specific type of cancer, to interfere with the growth of cancer cells by targeting specific cell growth factors, such as those involved in angiogenesis, or by enhancing the delivery of other anticancer agents to cancer cells when linked or conjugated to agents such as chemotherapeutic agents, radioactive particles or toxins.

[0097] Monoclonal antibodies currently used as cancer immunotherapeutic agents that are suitable for inclusion in the combinations of the present invention include, but are not limited to, rituximab (Rituxan®), trastuzumab (Herceptin®), ibritumomab tiuxetan (Zevalin®), tositumomab (Bexxar®), cetuximab (C-225, Erbitux®), bevacizumab (Avastin®), gemtuzumab ozogamicin (Mylotarg®), alemtuzumab (Campath®), and BL22. Monoclonal antibodies are used in the treatment of a wide range of cancers including breast cancer (including advanced metastatic breast cancer), colorectal cancer (including advanced and/or metastatic colorectal cancer), ovarian cancer, lung cancer, prostate cancer, cervical cancer, melanoma and brain tumours.

[0098] Other examples include antibodies specific a co-stimulatory molecule. Co-stimulatory molecules include, for example B7-1/CD80, CD28, B7- 2/CD86, CTLA-4, B7-H1/PD-L1, Gi24/Dies 1/VISTA, B7-H2, ICOS, B7-H3 PD-1, B7-H4, PD-L2/B7-DC, B7-H6, PDCD6, BTLA, 4-1 BB/TNFRSF9/CD137, CD40 Ligand/TNFSF5, 4-1BB Ligand/TNFSF9 GITR/TNFRSF18, HVEM/TNFRSF14, CD27/TNFRSF7, LIGHT/TNFSF14, CD27 Ligand/TNFSF7, OX40/TNFRSF4, CD30/TNFRSF8, OX40 Ligand/TNFSF4, CD30 Ligand/TNFSF8, TACI/TNFRSF13B, CD40/TNFRSF5, 2B4/CD244/SLAMF4 CD84/SLAMF5, BLAME/SLAMF8, CD229/SLAMF3, CD2CRACC/SLAMF7, CD2F-10/SLAMF9 NTB-A/SLAMF6, CD48/SLAMF2, SLAM/CD 150, CD58/LFA-3, CD2 Ikaros, CD53 Integrin alpha 4/CD49d, CD82/Kai-1 Integrin alpha 4 beta 1, CD90/Thyl Integrin alpha 4 beta 7/LPAM-1, CD96 LAG-3, CD160 LMIR1/CD300A, CRTAM TCL1A, DAP 12 TCL1B, Dectin-1/CLEC7A TIM-1/KIM- 1/HAVCR, DPPIV/CD26 TIM-4, EphB6 TSLP, HLA Class I TSLP R, HLA-DR. In particular, the antibody is selected from the group consisting of anti-CTLA4 antibodies (e.g. Ipilimumab), anti-PD1 antibodies, anti-PDL1 antibodies, anti-TIMP3 antibodies, anti-LAG3 antibodies, anti-B7H3 antibodies, anti-B7H4 antibodies anti-TREM antibodies, anti-BTLA antibodies, anti-LIGHT antibodies or anti-B7H6 antibodies.

[0099] Monoclonal antibodies can be used alone or in combination with other immunotherapeutic agents or chemotherapeutic agents.

[00100] Active specific immunotherapy typically involves the use of cancer vaccines. Cancer vaccines have been developed that comprise whole cancer cells, parts of cancer cells or one or more antigens derived from cancer cells. Cancer vaccines, alone or in combination with one or more immuno- or chemotherapeutic agents are being investigated in the treatment of several types of cancer including melanoma, renal cancer, ovarian cancer, breast cancer, colorectal cancer, and lung cancer. Non-specific immunotherapeutics are useful in combination with cancer vaccines in order to enhance the body's immune response.

[00101] The immunotherapeutic treatment may consist of an adoptive immunotherapy as described by Nicholas P. Restifo, Mark E. Dudley and Steven A. Rosenberg "Adoptive immunotherapy for cancer: harnessing the T cell response, Nature Reviews Immunology, Volume 12, April 2012). In adoptive immunotherapy, the patient's circulating lymphocytes, or tumor infiltrated lymphocytes, are isolated in vitro, activated by lymphokines such as IL-2 or

transduced with genes for tumor necrosis, and readministered (Rosenberg et al., 1988; 1989). The activated lymphocytes are most preferably the patient's own cells that were earlier isolated from a blood or tumor sample and activated (or "expanded") in vitro. This form of immunotherapy has produced several cases of regression of melanoma and renal carcinoma.

[00102] As used herein, the term "genomic abundance value" refers to an absolute or relative amount of a microorganism's genome in a biological sample from the gut of a subject. A genomic abundance value can be expressed different units, including copy number, molarity, mass (e.g., normalized against the size of the genome), unique sequence reads (e.g., normalized against the size of the genome), a percentage of any of the former metrics relative to the total amount of the metric across all genomes in the sample, a percentage of any of the former metrics relative to the total amount of the metric across a plurality of genomes in the sample, etc. In some embodiments, a genomic abundance value is normalized against a total genomic abundance in the sample. In some embodiments, a genomic abundance value is normalized against a genomic abundance value for a control genome in the sample. In some embodiments, the values for a plurality of genomic abundance values in a sample are standardized, normalized, and/or scaled. Examples of methods for normalizing genomic abundance values are described, for example, in Lin, H., Peddada, S.D., Analysis of microbial compositions: a review of normalization and differential abundance analysis, *Biofilms Microbiomes*, 6(60) (2020) and Lutz K.C., et al., A Survey of Statistical Methods for Microbiome Data Analysis, *Frontiers in Applied Mathematics and Statistics*, 8 (2022) the contents of which are incorporated herein by reference in their entireties. Methods for measuring genomic abundance values are known in the art. For example, metagenomic sequencing can be used to largely reconstruct microbial genomes from next generation sequencing of genomic DNA in biological samples, such as biological samples from the gut of a subject. For a review of metagenomic sequence see, for example, Quince C, et al., Shotgun metagenomics, from sampling to analysis, *Nat Biotechnol*, 35(9):833-44 (2017), the content of which is incorporated herein by reference in its entirety. Genomic abundance may also be determined by quantification of the copy number of a ribosomal gene, for example the 16S rRNA gene. Examples of rRNA quantification are described in Manzari C., et al., Accurate quantification of bacterial abundance in metagenomic DNAs accounting for variable DNA integrity levels, *Microb Genom.*, 6(10):mgen000417 (2020) and Barlow, J.T., et al., A quantitative sequencing framework for absolute abundance measurements of mucosal and

luminal microbial communities, Nat Commun., 11:2590 (2020), the contents of which are incorporated herein by reference in their entireties.

[00103] As used herein, the term “relative abundance” refers to a ratio of a first amount of a compound measured in a sample, *e.g.*, a genome for a first microorganism, to a second amount of a compound measured in a second sample. In some embodiments, relative abundance refers to a ratio of an amount of a compound, *e.g.*, a genome for a first microorganism, to a total amount of compounds, *e.g.*, the total amount of microorganism genomes or the total amount of a plurality of genomes, in the same sample. In other embodiments, relative abundance refers to a ratio of an amount of a compound, *e.g.*, a genome for a first microorganism, in a first sample to an amount of the compound of the compound in a second sample. For instance, a ratio of a normalized amount of a genome for a first microorganism in a first sample to a normalized amount of the genome for the first microorganism in a second and/or reference sample.

[00104] As used herein, the terms “sequencing,” “sequence determination,” and the like refer to any biochemical processes that may be used to determine the order of biological macromolecules such as nucleic acids or proteins. For example, sequencing data can include all or a portion of the nucleotide bases in a nucleic acid molecule such as an mRNA transcript or a genomic locus.

[00105] As used herein, the term “sequence reads” or “reads” refers to nucleotide sequences produced by any nucleic acid sequencing process described herein or known in the art. Reads can be generated from one end of nucleic acid fragments (“single-end reads”) or from both ends of nucleic acid fragments (*e.g.*, paired-end reads, double-end reads). The length of the sequence read is often associated with the particular sequencing technology. High-throughput methods, for example, provide sequence reads that can vary in size from tens to hundreds of base pairs (bp). In some embodiments, the sequence reads are of a mean, median or average length of about 15 bp to 900 bp long (*e.g.*, about 20 bp, about 25 bp, about 30 bp, about 35 bp, about 40 bp, about 45 bp, about 50 bp, about 55 bp, about 60 bp, about 65 bp, about 70 bp, about 75 bp, about 80 bp, about 85 bp, about 90 bp, about 95 bp, about 100 bp, about 110 bp, about 120 bp, about 130, about 140 bp, about 150 bp, about 200 bp, about 250 bp, about 300 bp, about 350 bp, about 400 bp, about 450 bp, or about 500 bp. In some embodiments, the sequence reads are of a mean, median or average length of about 1000 bp, 2000 bp, 5000 bp, 10,000 bp, or 50,000 bp or more.

Nanopore® sequencing, for example, can provide sequence reads that can vary in size from tens to hundreds to thousands of base pairs. Illumina® parallel sequencing, for example, can provide sequence reads that do not vary as much, for example, most of the sequence reads can be smaller than 200 bp. A sequence read (or sequencing read) can refer to sequence information corresponding to a nucleic acid molecule (*e.g.*, a string of nucleotides). For example, a sequence read can correspond to a string of nucleotides (*e.g.*, about 20 to about 150) from part of a nucleic acid fragment, can correspond to a string of nucleotides at one or both ends of a nucleic acid fragment, or can correspond to nucleotides of the entire nucleic acid fragment. A sequence read can be obtained in a variety of ways, *e.g.*, using sequencing techniques or using probes, *e.g.*, in hybridization arrays or capture probes, or amplification techniques, such as the polymerase chain reaction (PCR) or linear amplification using a single primer or isothermal amplification.

[00106] As used herein, the term “read segment” refers to any form of nucleotide sequence read including the raw sequence reads obtained directly from a nucleic acid sequencing technique or from a sequence derived therefrom, *e.g.*, an aligned sequence read, a collapsed sequence read, or a stitched sequence read.

[00107] As used herein, the term “read count” refers to the total number of nucleic acid reads generated, which may or may not be equivalent to the number of nucleic acid molecules generated, during a nucleic acid sequencing reaction.

[00108] As used herein, the term “read-depth,” “sequencing depth,” or “depth” can refer to a total number of unique nucleic acid fragments encompassing a particular locus or region of the genome of a microorganism that are sequenced in a particular sequencing reaction. Sequencing depth can be expressed as “Yx”, *e.g.*, 50x, 100x, *etc.*, where “Y” refers to the number of unique nucleic acid fragments encompassing a particular locus that are sequenced in a sequencing reaction. In such a case, Y is necessarily an integer, because it represents the actual sequencing depth for a particular locus. Alternatively, read-depth, sequencing depth, or depth can refer to a measure of central tendency (*e.g.*, a mean or mode) of the number of unique nucleic acid fragments that encompass one of a plurality of loci or regions of the genome of a microorganism that are sequenced in a particular sequencing reaction. For example, in some embodiments, sequencing depth refers to the average depth of every locus across a targeted sequencing panel, an exome, or an entire genome for the microorganism. In such case, Y may be expressed as a

fraction or a decimal, because it refers to an average coverage across a plurality of loci. When a mean depth is recited, the actual depth for any particular locus may be different than the overall recited depth. Metrics can be determined that provide a range of sequencing depths in which a defined percentage of the total number of loci fall. For instance, a range of sequencing depths within which 90% or 95%, or 99% of the loci fall. As understood by the skilled artisan, different sequencing technologies provide different sequencing depths. For instance, low-pass whole genome sequencing can refer to technologies that provide a sequencing depth of less than 5x, less than 4x, less than 3x, or less than 2x, *e.g.*, from about 0.5x to about 3x.

[00109] As used herein, the term “sequencing breadth” refers to what fraction of a particular microorganism genome has been sequenced. Sequencing breadth can be expressed as a fraction, a decimal, or a percentage, and is generally calculated as (the number of loci analyzed / the total number of loci in the genome). The denominator of the fraction can be a repeat-masked genome, and thus 100% can correspond to all of the reference genome minus the masked parts. A repeat-masked genome can refer to a genome in which sequence repeats are masked (*e.g.*, sequence reads align to unmasked portions of the genome). In some embodiments, any part of a genome can be masked and, thus, sequencing breadth can be evaluated for any desired portion of a genome.

[00110] As used herein, the terms “sequence ratio” and “coverage ratio” interchangeably refer to any measurement of a number of units of a genomic sequence in a first one or more biological samples (*e.g.*, a test and/or tumor sample) compared to the number of units of the respective genomic sequence in a second one or more biological samples (*e.g.*, a reference and/or control sample). In some embodiments, a sequence ratio is a copy ratio, a log₂-transformed copy ratio (*e.g.*, log₂ copy ratio), a coverage ratio, a base fraction, an allele fraction (*e.g.*, a variant allele fraction), and/or a tumor ploidy. In some embodiments sequence ratio is a log_N-transformed copy ratio, where N is any real number greater than 1.

[00111] As used herein, the term “sequencing probe” refers to a molecule that binds to a nucleic acid with affinity that is based on the expected nucleotide sequence of the RNA or DNA present at that locus.

[00112] As used herein, the term “targeted panel” or “targeted gene panel” refers to a combination of probes for sequencing (*e.g.*, by next-generation sequencing) nucleic acids present

in a biological sample from a subject (*e.g.*, a tumor sample, liquid biopsy sample, germline tissue sample, white blood cell sample, or tumor or tissue organoid sample), selected to map to one or more loci of interest in a genome.

[00113] As used herein, the term “sensitivity” or “true positive rate” (TPR) refers to the number of true positives divided by the sum of the number of true positives and false negatives. Sensitivity can characterize the ability of an assay or method to correctly identify a proportion of the population that truly has a condition. For example, sensitivity can characterize the ability of a method to correctly identify the number of subjects within a population having a particular biological characteristic.

[00114] As used herein, the term “specificity” or “true negative rate” (TNR) refers to the number of true negatives divided by the sum of the number of true negatives and false positives. Specificity can characterize the ability of an assay or method to correctly identify a proportion of the population that truly does not have a condition. For example, specificity can characterize the ability of a method to correctly identify the number of subjects within a population not having a particular biological characteristic.

[00115] As used interchangeably herein, the term “classifier” or “model” refers to a machine learning model or algorithm.

[00116] In some embodiments, a model includes an unsupervised learning algorithm. One example of an unsupervised learning algorithm is cluster analysis. In some embodiments, a model includes supervised machine learning. Nonlimiting examples of supervised learning algorithms include, but are not limited to, logistic regression, neural networks, support vector machines, Naive Bayes algorithms, nearest neighbor algorithms, random forest algorithms, decision tree algorithms, boosted trees algorithms, multinomial logistic regression algorithms, linear models, linear regression, Gradient Boosting, mixture models, hidden Markov models, Gaussian NB algorithms, linear discriminant analysis, diffusion models, or any combinations thereof. In some embodiments, a model is a multinomial classifier algorithm. In some embodiments, a model is a 2-stage stochastic gradient descent (SGD) model. In some embodiments, a model is a deep neural network (*e.g.*, a deep-and-wide sample-level model).

[00117] *Neural networks.* In some embodiments, the model is a neural network (*e.g.*, a convolutional neural network and/or a residual neural network). Neural network algorithms, also

known as artificial neural networks (ANNs), include convolutional and/or residual neural network algorithms (deep learning algorithms). In some embodiments, neural networks are machine learning algorithms that are trained to map an input dataset to an output dataset, where the neural network includes an interconnected group of nodes organized into multiple layers of nodes. For example, in some embodiments, the neural network architecture includes at least an input layer, one or more hidden layers, and an output layer. In some embodiments, the neural network includes any total number of layers, and any number of hidden layers, where the hidden layers function as trainable feature extractors that allow mapping of a set of input data to an output value or set of output values. In some embodiments, a deep learning algorithm is a neural network including a plurality of hidden layers, *e.g.*, two or more hidden layers. In some instances, each layer of the neural network includes a number of nodes (or “neurons”). In some embodiments, a node receives input that comes either directly from the input data or the output of nodes in previous layers, and performs a specific operation, *e.g.*, a summation operation. In some embodiments, a connection from an input to a node is associated with a parameter (*e.g.*, a weight and/or weighting factor). In some embodiments, the node sums up the products of all pairs of inputs, x_i , and their associated parameters. In some embodiments, the weighted sum is offset with a bias, b . In some embodiments, the output of a node or neuron is gated using a threshold or activation function, f , which, in some instances, is a linear or non-linear function. In some embodiments, the activation function is, for example, a rectified linear unit (ReLU) activation function, a Leaky ReLU activation function, or other function such as a saturating hyperbolic tangent, identity, binary step, logistic, arcTan, softsign, parametric rectified linear unit, exponential linear unit, softPlus, bent identity, softExponential, Sinusoid, Sine, Gaussian, or sigmoid function, or any combination thereof.

[00118] In some implementations, the weighting factors, bias values, and threshold values, or other computational parameters of the neural network, are “taught” or “learned” in a training phase using one or more sets of training data. For example, in some implementations, the parameters are trained using the input data from a training dataset and a gradient descent or backward propagation method so that the output value(s) that the ANN computes are consistent with the examples included in the training dataset. In some embodiments, the parameters are obtained from a back propagation neural network training process.

[00119] Any of a variety of neural networks are suitable for use in accordance with the present disclosure. Examples include, but are not limited to, feedforward neural networks, radial basis function networks, recurrent neural networks, residual neural networks, convolutional neural networks, residual convolutional neural networks, and the like, or any combination thereof. In some embodiments, the machine learning makes use of a pre-trained and/or transfer-learned ANN or deep learning architecture. In some implementations, convolutional and/or residual neural networks are used, in accordance with the present disclosure.

[00120] For instance, a deep neural network model includes an input layer, a plurality of individually parameterized (*e.g.*, weighted) convolutional layers, and an output scorer. The parameters (*e.g.*, weights) of each of the convolutional layers as well as the input layer contribute to the plurality of parameters (*e.g.*, weights) associated with the deep neural network model. In some embodiments, at least 50 parameters, at least 100 parameters, at least 1000 parameters, at least 2000 parameters or at least 5000 parameters are associated with the deep neural network model. As such, deep neural network models require a computer to be used because they cannot be mentally solved. In other words, given an input to the model, the model output needs to be determined using a computer rather than mentally in such embodiments. *See*, for example, Krizhevsky *et al.*, 2012, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems 2*, Pereira, Burges, Bottou, Weinberger, eds., pp. 1097-1105, Curran Associates, Inc.; Zeiler, 2012 “ADADELTA: an adaptive learning rate method,” CoRR, vol. abs/1212.5701; and Rumelhart *et al.*, 1988, “Neurocomputing: Foundations of research,” ch. Learning Representations by Back-propagating Errors, pp. 696-699, Cambridge, MA, USA: MIT Press, each of which is hereby incorporated by reference.

[00121] Neural network algorithms, including convolutional neural network algorithms, suitable for use as models are disclosed in, for example, Vincent *et al.*, 2010, “Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion,” J Mach Learn Res 11, pp. 3371-3408; Larochelle *et al.*, 2009, “Exploring strategies for training deep neural networks,” J Mach Learn Res 10, pp. 1-40; and Hassoun, 1995, *Fundamentals of Artificial Neural Networks*, Massachusetts Institute of Technology, each of which is hereby incorporated by reference. Additional example neural networks suitable for use as models are disclosed in Duda *et al.*, 2001, *Pattern Classification*, Second Edition, John Wiley & Sons, Inc., New York; and Hastie *et al.*, 2001, *The Elements of Statistical Learning*, Springer-

Verlag, New York, each of which is hereby incorporated by reference in its entirety. Additional example neural networks suitable for use as models are also described in Draghici, 2003, *Data Analysis Tools for DNA Microarrays*, Chapman & Hall/CRC; and Mount, 2001, *Bioinformatics: sequence and genome analysis*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, each of which is hereby incorporated by reference in its entirety.

[00122] *Support vector machines.* In some embodiments, the model is a support vector machine (SVM). SVM algorithms suitable for use as models are described in, for example, Cristianini and Shawe-Taylor, 2000, "An Introduction to Support Vector Machines," Cambridge University Press, Cambridge; Boser *et al.*, 1992, "A training algorithm for optimal margin classifiers," in Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory, ACM Press, Pittsburgh, Pa., pp. 142-152; Vapnik, 1998, Statistical Learning Theory, Wiley, New York; Mount, 2001, Bioinformatics: sequence and genome analysis, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.; Duda, Pattern Classification, Second Edition, 2001, John Wiley & Sons, Inc., pp. 259, 262-265; and Hastie, 2001, The Elements of Statistical Learning, Springer, New York; and Furey *et al.*, 2000, Bioinformatics 16, 906-914, each of which is hereby incorporated by reference in its entirety. When used for classification, SVMs separate a given set of binary labeled data with a hyper-plane that is maximally distant from the labeled data. For certain cases in which no linear separation is possible, SVMs work in combination with the technique of 'kernels', which automatically realizes a non-linear mapping to a feature space. The hyper-plane found by the SVM in feature space corresponds, in some instances, to a non-linear decision boundary in the input space. In some embodiments, the plurality of parameters (*e.g.*, weights) associated with the SVM define the hyper-plane. In some embodiments, the hyper-plane is defined by at least 10, at least 20, at least 50, or at least 100 parameters and the SVM model requires a computer to calculate because it cannot be mentally solved.

[00123] *Naïve Bayes algorithms.* In some embodiments, the model is a Naïve Bayes algorithm. Naïve Bayes models suitable for use as models are disclosed, for example, in Ng *et al.*, 2002, "On discriminative vs. generative classifiers: A comparison of logistic regression and naïve Bayes," Advances in Neural Information Processing Systems, 14, which is hereby incorporated by reference. A Naïve Bayes model is any model in a family of "probabilistic models" based on applying Bayes' theorem with strong (naïve) independence assumptions

between the features. In some embodiments, they are coupled with Kernel density estimation. See, for example, Hastie *et al.*, 2001, *The elements of statistical learning: data mining, inference, and prediction*, eds. Tibshirani and Friedman, Springer, New York, which is hereby incorporated by reference.

[00124] *Nearest neighbor algorithms.* In some embodiments, a model is a nearest neighbor algorithm. In some implementations, nearest neighbor models are memory-based and include no model to be fit. For nearest neighbors, given a query point x_0 (a test subject), the k training points $x_{(1)}, x_{(2)}, \dots, x_{(k)}$ (here the training subjects) closest in distance to x_0 are identified and then the point x_0 is classified using the k nearest neighbors. In some embodiments, Euclidean distance in feature space is used to determine distance as $d_{(i)} = \|x_{(i)} - x_{(0)}\|$. Typically, when the nearest neighbor algorithm is used, the abundance data used to compute the linear discriminant is standardized to have mean zero and variance 1. In some embodiments, the nearest neighbor rule is refined to address issues of unequal class priors, differential misclassification costs, and feature selection. Many of these refinements involve some form of weighted voting for the neighbors. For more information on nearest neighbor analysis, see Duda, *Pattern Classification*, Second Edition, 2001, John Wiley & Sons, Inc; and Hastie, 2001, *The Elements of Statistical Learning*, Springer, New York, each of which is hereby incorporated by reference.

[00125] A k -nearest neighbor model is a non-parametric machine learning method in which the input consists of the k closest training examples in feature space. The output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If $k = 1$, then the object is simply assigned to the class of that single nearest neighbor. See, Duda *et al.*, 2001, *Pattern Classification*, Second Edition, John Wiley & Sons, which is hereby incorporated by reference. In some embodiments, the number of distance calculations needed to solve the k -nearest neighbor model is such that a computer is used to solve the model for a given input because it cannot be mentally performed.

[00126] *Random forest, decision tree, and boosted tree algorithms.* In some embodiments, the model is a decision tree. Decision trees suitable for use as models are described generally by Duda, 2001, *Pattern Classification*, John Wiley & Sons, Inc., New York, pp. 395-396, which is hereby incorporated by reference. Tree-based methods partition the feature space into a set of

rectangles, and then fit a model (like a constant) in each one. In some embodiments, the decision tree is random forest regression. For example, one specific algorithm is a classification and regression tree (CART). Other specific decision tree algorithms include, but are not limited to, ID3, C4.5, MART, and Random Forests. CART, ID3, and C4.5 are described in Duda, 2001, *Pattern Classification*, John Wiley & Sons, Inc., New York, pp. 396-408 and pp. 411-412, which is hereby incorporated by reference. CART, MART, and C4.5 are described in Hastie *et al.*, 2001, *The Elements of Statistical Learning*, Springer-Verlag, New York, Chapter 9, which is hereby incorporated by reference in its entirety. Random Forests are described in Breiman, 1999, "Random Forests--Random Features," Technical Report 567, Statistics Department, U.C. Berkeley, September 1999, which is hereby incorporated by reference in its entirety. In some embodiments, the decision tree model includes at least 10, at least 20, at least 50, or at least 100 parameters (*e.g.*, weights and/or decisions) and requires a computer to calculate because it cannot be mentally solved.

[00127] *Regression.* In some embodiments, the model uses a regression algorithm. In some embodiments, a regression algorithm is any type of regression. For example, in some embodiments, the regression algorithm is logistic regression. In some embodiments, the regression algorithm is logistic regression with lasso, L2 or elastic net regularization. In some embodiments, those extracted features that have a corresponding regression coefficient that fails to satisfy a threshold value are pruned (removed from) consideration. In some embodiments, a generalization of the logistic regression model that handles multcategory responses is used as the model. Logistic regression algorithms are disclosed in Agresti, *An Introduction to Categorical Data Analysis*, 1996, Chapter 5, pp. 103-144, John Wiley & Son, New York, which is hereby incorporated by reference. In some embodiments, the model makes use of a regression model disclosed in Hastie *et al.*, 2001, *The Elements of Statistical Learning*, Springer-Verlag, New York. In some embodiments, the logistic regression model includes at least 10, at least 20, at least 50, at least 100, or at least 1000 parameters (*e.g.*, weights) and requires a computer to calculate because it cannot be mentally solved.

[00128] *Linear discriminant analysis algorithms.* In some embodiments, linear discriminant analysis (LDA), normal discriminant analysis (NDA), or discriminant function analysis is a generalization of Fisher's linear discriminant, a method used in statistics, pattern recognition, and machine learning to find a linear combination of features that characterizes or separates two

or more classes of objects or events. In some embodiments, the resulting combination is used as the model (linear model) in some embodiments of the present disclosure.

[00129] *Mixture model and Hidden Markov model.* In some embodiments, the model is a mixture model, such as that described in McLachlan *et al.*, Bioinformatics 18(3):413-422, 2002. In some embodiments, in particular, those embodiments including a temporal component, the model is a hidden Markov model such as described by Schliep *et al.*, 2003, Bioinformatics 19(1):i255-i263.

[00130] *Clustering.* In some embodiments, the model is an unsupervised clustering model. In some embodiments, the model is a supervised clustering model. Clustering algorithms suitable for use as models are described, for example, at pages 211-256 of Duda and Hart, Pattern Classification and Scene Analysis, 1973, John Wiley & Sons, Inc., New York, (hereinafter "Duda 1973") which is hereby incorporated by reference in its entirety. As an illustrative example, in some embodiments, the clustering problem is described as one of finding natural groupings in a dataset. To identify natural groupings, two issues are addressed. First, a way to measure similarity (or dissimilarity) between two samples is determined. This metric (*e.g.*, similarity measure) is used to ensure that the samples in one cluster are more like one another than they are to samples in other clusters. Second, a mechanism for partitioning the data into clusters using the similarity measure is determined. One way to begin a clustering investigation is to define a distance function and to compute the matrix of distances between all pairs of samples in the training set. If distance is a good measure of similarity, then the distance between reference entities in the same cluster is significantly less than the distance between the reference entities in different clusters. However, in some implementations, clustering does not use a distance metric. For example, in some embodiments, a nonmetric similarity function $s(x, x')$ is used to compare two vectors x and x' . In some such embodiments, $s(x, x')$ is a symmetric function whose value is large when x and x' are somehow "similar." Once a method for measuring "similarity" or "dissimilarity" between points in a dataset has been selected, clustering uses a criterion function that measures the clustering quality of any partition of the data. Partitions of the dataset that extremize the criterion function are used to cluster the data. Particular exemplary clustering techniques contemplated for use in the present disclosure include, but are not limited to, hierarchical clustering (agglomerative clustering using a nearest-neighbor algorithm, farthest-neighbor algorithm, the average linkage algorithm, the centroid

algorithm, or the sum-of-squares algorithm), k-means clustering, fuzzy k-means clustering algorithm, and Jarvis-Patrick clustering. In some embodiments, the clustering includes unsupervised clustering (*e.g.*, with no preconceived number of clusters and/or no predetermination of cluster assignments).

[00131] *Ensembles of models and boosting.* In some embodiments, an ensemble (two or more) of models is used. In some embodiments, a boosting technique such as AdaBoost is used in conjunction with many other types of learning algorithms to improve the performance of the model. In this approach, the output of any of the models disclosed herein, or their equivalents, is combined into a weighted sum that represents the final output of the boosted model. In some embodiments, the plurality of outputs from the models is combined using any measure of central tendency known in the art, including but not limited to a mean, median, mode, a weighted mean, weighted median, weighted mode, *etc.* In some embodiments, the plurality of outputs is combined using a voting method. In some embodiments, a respective model in the ensemble of models is weighted or unweighted.

[00132] As used herein, the term “parameter” refers to any coefficient or, similarly, any value of an internal or external element (*e.g.*, a weight and/or a hyperparameter) in an algorithm, model, regressor, and/or classifier that can affect (*e.g.*, modify, tailor, and/or adjust) one or more inputs, outputs, and/or functions in the algorithm, model, regressor and/or classifier. For example, in some embodiments, a parameter refers to any coefficient, weight, and/or hyperparameter that can be used to control, modify, tailor, and/or adjust the behavior, learning, and/or performance of an algorithm, model, regressor, and/or classifier. In some instances, a parameter is used to increase or decrease the influence of an input (*e.g.*, a feature) to an algorithm, model, regressor, and/or classifier. As a nonlimiting example, in some embodiments, a parameter is used to increase or decrease the influence of a node (*e.g.*, of a neural network), where the node includes one or more activation functions. Assignment of parameters to specific inputs, outputs, and/or functions is not limited to any one paradigm for a given algorithm, model, regressor, and/or classifier but can be used in any suitable algorithm, model, regressor, and/or classifier architecture for a desired performance. In some embodiments, a parameter has a fixed value. In some embodiments, a value of a parameter is manually and/or automatically adjustable. In some embodiments, a value of a parameter is modified by a validation and/or training process for an algorithm, model, regressor, and/or classifier (*e.g.*, by error minimization and/or

backpropagation methods). In some embodiments, an algorithm, model, regressor, and/or classifier of the present disclosure includes a plurality of parameters. In some embodiments, the plurality of parameters is n parameters, where: $n \geq 2$; $n \geq 5$; $n \geq 10$; $n \geq 25$; $n \geq 40$; $n \geq 50$; $n \geq 75$; $n \geq 100$; $n \geq 125$; $n \geq 150$; $n \geq 200$; $n \geq 225$; $n \geq 250$; $n \geq 350$; $n \geq 500$; $n \geq 600$; $n \geq 750$; $n \geq 1,000$; $n \geq 2,000$; $n \geq 4,000$; $n \geq 5,000$; $n \geq 7,500$; $n \geq 10,000$; $n \geq 20,000$; $n \geq 40,000$; $n \geq 75,000$; $n \geq 100,000$; $n \geq 200,000$; $n \geq 500,000$, $n \geq 1 \times 10^6$, $n \geq 5 \times 10^6$, or $n \geq 1 \times 10^7$. As such, the algorithms, models, regressors, and/or classifiers of the present disclosure cannot be mentally performed. In some embodiments n is between 10,000 and 1×10^7 , between 100,000 and 5×10^6 , or between 500,000 and 1×10^6 . In some embodiments, the algorithms, models, regressors, and/or classifier of the present disclosure operate in a k -dimensional space, where k is a positive integer of 5 or greater (*e.g.*, 5, 6, 7, 8, 9, 10, *etc.*). As such, the algorithms, models, regressors, and/or classifiers of the present disclosure cannot be mentally performed.

[00133] As used herein, the term “untrained model” (*e.g.*, “untrained classifier” and/or “untrained neural network”) refers to a machine learning model or algorithm, such as a classifier or a neural network, that has not been trained on a target dataset. In some embodiments, “training a model” (*e.g.*, “training a neural network”) refers to the process of training an untrained or partially trained model (*e.g.*, “an untrained or partially trained neural network”). Moreover, it will be appreciated that the term “untrained model” does not exclude the possibility that transfer learning techniques are used in such training of the untrained or partially trained model. For instance, Fernandes *et al.*, 2017, “Transfer Learning with Partial Observability Applied to Cervical Cancer Screening,” Pattern Recognition and Image Analysis: 8th Iberian Conference Proceedings, 243-250, which is hereby incorporated by reference, provides non-limiting examples of such transfer learning. In instances where transfer learning is used, the untrained model described above is provided with additional data over and beyond that of the primary training dataset. Typically, this additional data is in the form of parameters (*e.g.*, coefficients, weights, and/or hyperparameters) that were learned from another, auxiliary training dataset. Moreover, while a description of a single auxiliary training dataset has been disclosed, it will be appreciated that there is no limit on the number of auxiliary training datasets that can be used to complement the primary training dataset in training the untrained model in the present disclosure. For instance, in some embodiments, two or more auxiliary training datasets, three or more auxiliary training datasets, four or more auxiliary training datasets or five or more auxiliary

training datasets are used to complement the primary training dataset through transfer learning, where each such auxiliary dataset is different than the primary training dataset. Any manner of transfer learning is used, in some such embodiments. For instance, consider the case where there is a first auxiliary training dataset and a second auxiliary training dataset in addition to the primary training dataset. In such a case, the parameters learned from the first auxiliary training dataset (by application of a first model to the first auxiliary training dataset) are applied to the second auxiliary training dataset using transfer learning techniques (*e.g.*, a second model that is the same or different from the first model), which in turn results in a trained intermediate model whose parameters are then applied to the primary training dataset and this, in conjunction with the primary training dataset itself, is applied to the untrained model. Alternatively, in another example embodiment, a first set of parameters learned from the first auxiliary training dataset (by application of a first model to the first auxiliary training dataset) and a second set of parameters learned from the second auxiliary training dataset (by application of a second model that is the same or different from the first model to the second auxiliary training dataset) are each individually applied to a separate instance of the primary training dataset (*e.g.*, by separate independent matrix multiplications) and both such applications of the parameters to separate instances of the primary training dataset in conjunction with the primary training dataset itself (or some reduced form of the primary training dataset such as principal components or regression coefficients learned from the primary training set) are then applied to the untrained model in order to train the untrained model.

[00134] As used herein, the term “AUC” refers to the Area Under the Curve, for example, of a ROC Curve. That value can assess the merit of a test on a given sample population with a value of 1 representing a good test ranging down to 0.5 which means the test is providing a random response in classifying test subjects. Since the range of the AUC is only 0.5 to 1.0, a small change in AUC has greater significance than a similar change in a metric that ranges for 0 to 1 or 0 to 100%. When the % change in the AUC is given, it will be calculated based on the fact that the full range of the metric is 0.5 to 1.0. A variety of statistics packages can calculate AUC for an ROC curve. AUC can be used to compare the accuracy of the classification algorithm across the complete data range. Classification algorithms with greater AUC have, by definition, a greater capacity to classify unknowns correctly between the two groups of interest (disease and no disease, responder and non-responder).

[00135] As used herein, the term “instruction” refers to an order given to a computer processor by a computer program. On a digital computer, each instruction is a sequence of 0s and 1s that describes a physical operation the computer is to perform. Such instructions can include data transfer instructions and data manipulation instructions. In some embodiments, each instruction is a type of instruction in an instruction set that is recognized by a particular processor type used to carry out the instructions. Examples of instruction sets include, but are not limited to, Reduced Instruction Set Computer (RISC), Complex Instruction Set Computer (CISC), Minimal instruction set computers (MISC), Very long instruction word (VLIW), Explicitly parallel instruction computing (EPIC), and One instruction set computer (OISC).

[00136] Several aspects are described below with reference to example applications for illustration. It should be understood that numerous specific details, relationships, and methods are set forth to provide a full understanding of the features described herein. One having ordinary skill in the relevant art, however, will readily recognize that the features described herein can be practiced without one or more of the specific details or with other methods. The features described herein are not limited by the illustrated ordering of acts or events, as some acts can occur in different orders and/or concurrently with other acts or events. Furthermore, not all illustrated acts or events are required to implement a methodology in accordance with the features described herein.

[00137] *Select Abbreviations ****

HQMAGs - High-quality metagenome-assembled genomes

CCDC-I – Case-Control Dataset Collection I

CCDC-II – Case-Control Dataset Collection II

TDC – Treatment Dataset Collection

TCG – Two Competing Guilds

QD-TCG – Two Competing Guilds identified from the QD study (141 HQMAGs)

C-TCG – Combined Two Competing Guilds (788 HQMAGs)

CC-TCG – Combined Core set of the Two Competing Guilds (284 HQMAGs)

ANI – Average Nucleotide Identity

[00138] Reference will now be made in detail to embodiments, examples of which are illustrated in the accompanying drawings. In the following detailed description, numerous specific details are set forth in order to provide a thorough understanding of the present disclosure. However, it will be apparent to one of ordinary skill in the art that the present disclosure may be practiced without these specific details. In other instances, well-known methods, procedures, components, circuits, and networks have not been described in detail so as not to unnecessarily obscure aspects of the embodiments.

[00139] *Example System Embodiments.*

[00140] Now that an overview of some aspects of the present disclosure and some definitions used in the present disclosure have been provided, details of an exemplary system are described in conjunction with Figure 1. Figure 1 is a block diagram illustrating a system 100 in accordance with some implementations. The system 100 in some implementations includes one or more processing units CPU(s) 102 (also referred to as processors), one or more network interfaces 104, a user interface 106 including (optionally) a display 108 and an input system 110, a non-persistent memory 111, a persistent memory 112, and one or more communication buses 114 for interconnecting these components. The one or more communication buses 114 optionally include circuitry (sometimes called a chipset) that interconnects and controls communications between system components. The non-persistent memory 111 typically includes high-speed random access memory, such as DRAM, SRAM, DDR RAM, ROM, EEPROM, flash memory, whereas the persistent memory 112 typically includes CD-ROM, digital versatile disks (DVD) or other optical storage, magnetic cassettes, magnetic tape, magnetic disk storage or other magnetic storage devices, magnetic disk storage devices, optical disk storage devices, flash memory devices, or other non-volatile solid state storage devices. The persistent memory 112 optionally includes one or more storage devices remotely located from the CPU(s) 102. The persistent memory 112, and the non-volatile memory device(s) within the non-persistent memory 112, comprise non-transitory computer readable storage medium. In some implementations, the non-persistent memory 111 or alternatively the non-transitory computer readable storage medium stores the following programs, modules and data structures, or a subset thereof, sometimes in conjunction with the persistent memory 112:

- an optional operating system 116, which includes procedures for handling various basic system services and for performing hardware dependent tasks;
- an optional network communication module (or instructions) 118 for connecting the system 100 with other devices and/or a communication network 104;
- a microbiome evaluation module 140 for determining a disease state, in a plurality of disease states, of a subject based on the constitution of the subject's microbiome; and
- a datastore of subject information 140 based on microbiome sequencing results 150, including abundance values 152 for microbes in each of guilds 152-A and 152-B as described herein.

[00141] In various implementations, one or more of the above identified elements are stored in one or more of the previously mentioned memory devices, and correspond to a set of instructions for performing a function described above. The above identified modules, data, or programs (e.g., sets of instructions) need not be implemented as separate software programs, procedures, datasets, or modules, and thus various subsets of these modules and data may be combined or otherwise re-arranged in various implementations. In some implementations, the non-persistent memory 111 optionally stores a subset of the modules and data structures identified above. Furthermore, in some embodiments, the memory stores additional modules and data structures not described above. In some embodiments, one or more of the above identified elements is stored in a computer system, other than that of visualization system 100, that is addressable by visualization system 100 so that visualization system 100 may retrieve all or a portion of such data when needed.

[00142] Although Figure 1 depicts a "system 100," the figure is intended more as a functional description of the various features that may be present in computer systems than as a structural schematic of the implementations described herein. In practice, and as recognized by those of ordinary skill in the art, items shown separately could be combined and some items could be separated. Moreover, although Figure 1 depicts certain data and modules in non-persistent memory 111, some or all of these data and modules instead may be stored in persistent memory 112.

[00143] **1. Methods of training a model for predicting subject response to a therapy for a disorder**

[00144] Figure 2 is a schematic diagram of a method of training a model for predicting a subject's response to a therapy for a disorder as discussed below. The method may be implemented using a computer system (*e.g.*, the computer system 100 shown and described above in reference to Figure 1).

[00145] Referring to block 200, in some embodiments, the methods including obtaining, in electronic form, for each respective training subject in a plurality of training subjects, (i) a corresponding plurality of genomic abundance values for the respective training subject at a time prior to receiving the therapy, wherein the corresponding plurality of genomic abundance values comprise, for each respective gut microorganism in a plurality of gut microorganisms, a corresponding value for the abundance of the genome of the respective gut microorganism in a corresponding biological sample from the gut of the respective training subject, (ii) an indication of the respective training subject's response to the therapy of the respective training subject. Each respective training subject in the plurality of training subjects has received a therapy for a disorder.

[00146] In some embodiments, the plurality of training subjects comprises at least 50, at least 100, at least 200, at least 500, at least 1000, at least 5000, at least 10,000, at least 20,000, at least 50,000, at least 100,000, at least 500,000, or at least 1,000,000 subjects. In some embodiments, the plurality of training subjects comprises no more than 1,000,000, no more than 500,000, no more than 100,000, no more than 50,000, no more than 20,000, no more than 10,000, no more than 1000 subjects, no more than 500 subjects, no more than 100 subjects, or no more than 50 subjects. In some embodiments, the plurality of training subjects consists of from 50 to 100, from 50 to 200, from 50 to 500, from 100 to 500, from 200 to 500, from 200 to 1000, from 500 to 1000, from 200 to 5,000, from 1000 to 10,000, from 5000 to 20,000, from 10,000 to 50,000, from 20,000 to 100,000, or from 500,000 to 1,000,000. In some embodiments, the plurality of training subjects falls within another range starting no lower than 50 subjects and ending no higher than 100,000,000 subjects. In some embodiments, the plurality of subjects shares similar health status (such as physical or mental conditions, medical history, gene carrier, or medication use).

[00147] In some embodiment, a corresponding biological sample from the gut of the respective training subject was taken prior to a treatment or a therapy. In some embodiments, the

biological sample is taken no more than 15 minutes, 30 minutes, 1 hour, 2 hours, 3 hours, 4 hours, 5 hours, 6 hours, 12 hours, or 24 hours prior to a treatment or a therapy. In some embodiments, the biological sample is taken 1 day, 2, days, 3 days, 4 days, 5 days, 6 days, 1 week, 2 weeks, 3 weeks, 4 weeks, or more prior to a treatment or a therapy. In some embodiments, the biological sample is taken about any of 1 month, 2 months, 3 months, 4 months, 5 months, 6 months, or more prior to a treatment or a therapy.

[00148] In some embodiments, sample data (including plasma, stool specimens) and corresponding clinical information (including gender/age/body fat count/underlying disease/histopathological characteristics, etc.) were collected for each training subject prior to receiving a therapy. Individual biological samples were subjected to full microbiome analysis. In some embodiments, the sample is a tissue biopsy, an intestinal, or mucosal sample. See, for example, Tang Q, Jet al., Current Sampling Methods for Gut Microbiota: A Call for More Precise Devices, *Front Cell Infect Microbiol.*, 10:151 (2020), the content of which is incorporated herein by reference in its entirety. In some embodiment, the biological sample from the gut of the respective subject is a fecal sample from the respective training subject.

[00149] In some of the embodiments, the corresponding value for the abundance of the genome is a value representative of the absolute abundance of a microorganism genome. In some of the embodiments, the corresponding value for the abundance of the genome is a value representative of a normalized abundance value, or a relative abundance value (e.g., an abundance of one microorganism normalized against the abundance of total microbiome of interest). In some of the embodiments, corresponding value for the abundance of the genome is a value representative of an averaged abundance value (e.g., average of abundances obtained at different time points or from different biological samples from the patients, or average of abundances obtained using different probes, etc.), or a combination of any of above. The corresponding value for the abundance of the genome is measured by any technique known in the art. In some embodiments, the genomic abundance value for the genome is measured by quantitative PCR(qPCR), such as bacterial 16S rRNA qPCR, RT-PCR, or qRT-PCR, for quantifying the abundance of region of interests in the genome, e.g., as described in U.S. Patent No. 11,427,865, the disclosure of which is hereby incorporated by reference in its entirety. In some embodiments, the genomic abundance value is measured by targeted sequencing (e.g., 16S rRNA sequencing, or any other suitable biomarker), partial genome sequencing or whole genome

sequencing, thereby quantifying the number of reads of the targeted regions in a microorganism genome to determine the abundance of the genome, e.g., as disclosed in U.S. Patent Application Publication No. 2021/0403986 or U.S. Patent No. 11,332,783, the disclosures of which are hereby incorporated by reference in their entireties. In some embodiments, deep sequencing is employed to determine the abundance of targeted sequences, e.g., as disclosed in U.S. Patent Application Publication No. 2018/0237863, the disclosure of which is incorporated herein by reference in its entirety. In some embodiments, the sequencing depth is at least 2X, at least 3X, at least 4X, at least 5X, at least 6X, at least 7X, at least 8X, at least 9X, at least 10X, at least 11X, at least 12X, at least 13X, at least 14X, at least 15X, at least 16X, at least 17X, at least 18X, at least 19X, at least 20X, at least 21X, at least 22X, at least 23X, at least 24X, at least 25X, at least 26X, at least 27X, at least 28X, at least 29X, at least 30X, at least 31X, at least 32X, at least 33X, at least 34X, at least 35X, at least 36X, at least 37X, at least 38X, at least 39X, at least 40X, at least 41X, at least 42X, at least 43X, at least 44X, at least 45X, at least 46X, at least 47X, at least 48X, at least 49X, at least 50X, at least 51X, at least 52X, at least 53X, at least 54X, at least 55 X, at least 56X, at least 57X, at least 58X, at least 59X, at least 60X, at least 70X, at least 80X, at least 90X, at least 100X, at least 110X, at least 120X, at least 130X, at least 150X, at least 200X, at least 300X, at least 400X, at least 500X, at least 750X, at least 1000X, or more. In some embodiments, shotgun metagenomic sequencing is employed to provide sequence reads for genomes in a sample, e.g., as described in U.S. Patent No. 11,028,449, the content of which is incorporated herein by reference in its entirety.

[00150] In some embodiments, the indication of subject's response is characterized by clinical outcome measures include, but are not limited to, complete remission, partial remission, non-remission, survival, development of adverse events, or any combination thereof. In some embodiments, one responder has complete remission in response to the treatment, and the non-responders has non-remission or partial remission in response to the treatment. In some embodiment, training subjects were subjected to routine clinical examinations, laboratory analyses, and computed tomography. Tumor responses were evaluated using RECIST criteria. In some embodiment, complete response was defined as complete radiographic disappearance of measurable or evaluable disease or stable, minimal radiographic findings; partial response was defined as reduction of the longest dimension of measurable disease by at least 50%; stable disease was defined as reduction of the longest dimension by less than 25%; Progressive disease

was defined as growth of the tumor by more than 25% in the longest dimension or development of new lesions. In some embodiment, overall response rate was defined as the sum of the complete and partial response rates and the tumor control rate was defined as the sum of overall response rates with stable disease rates.

[00151] In some embodiments, the indication of subject's response is characterized by the actual treatment efficacy of an therapy, including progression-free survival (PFS), the duration of the progression free survival under treatment, total Survival (OS), response to therapy (RT), overall response rate (ORR), sustained clinical effect (DCB), Disease Activity Score, or any combination thereof, or any other method for evaluating the progression or prognosis of a disease or disorder known in the art.

[00152] In some embodiments, "progression free survival" (PFS) has its art-understood meaning relating to the length of time during and after the treatment of a disease, such as cancer, that a patient lives with the disease but it does not get worse. In some embodiments, measuring the progression-free survival is utilized as an assessment of how well a new treatment works. In some embodiments, PFS is determined in a randomized clinical trial; in some such embodiments, PFS refers to time from randomization until objective tumor progression and/or death.

[00153] In some embodiments, ORR may be defined as the proportion of patients in whom partial (PR) or complete (CR) responses are identified as a best overall response (BOR) according to some metric, such as Response Evaluation Criteria in Solid Tumors (RECIST 1.1). Stable disease (SD) was categorized as non-response together with progressive disease (PD). In some embodiments, ORR has its art-understood meaning referring to the proportion of patients with tumor size reduction of a predefined amount and for a minimum time period. In some embodiments, response duration usually measured from the time of initial response until documented tumor progression. In some embodiments, ORR involves the sum of partial responses plus complete responses.

[00154] In some embodiments, "clinical effect" refers to a clinical benefit. In some embodiments, such a clinical benefit is or comprises reduction in tumor size, increase in progression free survival, increase in overall survival, decrease in overall tumor burden, decrease in the symptoms caused by tumor growth such as pain, organ failure, bleeding, damage to the skeletal system, and other related sequelae of metastatic cancer and combinations thereof. In

some embodiments, the clinical effect is a “sustained clinical effect” (DCB) that is maintained for a relevant period of time. In some embodiments, the relevant period of time is at least 1 month, 2 months, 3 months, 4 months, 5 months, 6 months, 7 months, 8 months, 9 months, 10 months, 11 months, 1 year, 2 years, 3 years, 4 years, 5 years, or longer.

[00155] In some embodiments, the subject’s response is measured by Disease Activity Score (DAS) (see, e.g., Van der Heijde D. M. et al., J Rheumatol, 1993, 20(3): 579-81; Prevoo M. L. et al, Arthritis Rheum, 1995, 38: 44-8). The DAS system represents both current state of disease activity and change. The DAS scoring system uses a weighted mathematical formula, derived from clinical trials in RA. For example, the DAS 28 is $0.56(T28)+0.28(SW28)+0.70(\ln ESR)+0.014 GH$ wherein T represents tender joint number, SW is swollen joint number, ESR is erythrocyte sedimentation rate, and GH is global health. Various values of the DAS represent high or low disease activity as well as remission, and the change and endpoint score result in a categorization of the patient by degree of response (none, moderate, good).

[00156] In some embodiments, the indication of the subject’s response is measured by the level of the immune response or immune parameters of a cancer-bearing patient resulting from an immunotherapy. In some embodiments, the immune response or immune parameters are characterized by expression level of various biological markers of the host immune response in conjunction with the occurrence of a cancer at a given stage of cancer development (i.e., treatment efficacy). In some embodiments, the expression level of a biological marker is compared with a reference value for the same biological marker, and when required with reference values. The reference value for the same biological marker is thus predetermined and is already known to be indicative of a reference value that is pertinent for discriminating between a low level and a high level of the immune response of a patient with cancer, for said biological marker. Said predetermined reference value for said biological marker is correlated with a responder to treatment in a cancer patient, or conversely is correlated with non-responder to treatment in a cancer patient.

[00157] In some embodiments, a change of a combination of biological markers are quantified. In some embodiments, a combination of 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, or more distinct biological markers are quantified.

[00158] In certain embodiments, biological markers are quantified with immunohistochemical techniques. Example biological markers include 18s, ACE, ACTB, AGTR1, AGTR2, APC, APOA1, ARF1, AXIN1, BAX, BCL2, BCL2L1, CXCR5, BMP2, BRCA1, BTLA, C3, CASP3, CASP9, CCL1, CCL11, CCL13, CCL16, CCL17, CCL18, CCL19, CCL2, CCL20, CCL21, CCL22, CCL23, CCL24, CCL25, CCL26, CCL27, CCL28, CCL3, CCL5, CCL7, CCL8, CCNB1, CCND1, CCNE1, CCR1, CCR10, CCR2, CCR3, CCR4, CCR5, CCR6, CCR7, CCR8, CCR9, CCRL2, CD154, CD19, CD1a, CD2, CD226, CD244, PDCD1LG1, CD28, CD34, CD36, CD38, CD3E, CD3G, CD3Z, CD4, CD40LG, CD5, CD54, CD6, CD68, CD69, CLIP, CD80, CD83, SLAMF5, CD86, CD8A, CDH1, CDH7, CDK2, CDK4, CDKN1A, CDKN1B, CDKN2A, CDKN2B, CEACAM1, COL4A5, CREBBP, CRLF2, CSF1, CSF2, CSF3, CTLA4, CTNNB1, CTSC, CX3CL1, CX3CR1, CXCL1, CXCL10, CXCL11, CXCL12, CXCL13, CXCL14, CXCL16, CXCL2, CXCL3, CXCL5, CXCL6, CXCL9, CXCR3, CXCR4, CXCR6, CYP1A2, CYP7A1, DCC, DCN, DEFA6, DICER1, DKK1, Dok-1, Dok-2, DOK6, DVL1, E2F4, EBI3, ECE1, ECGF1, EDN1, EGF, EGFR, EIF4E, CD105, ENPEP, ERBB2, EREG, FCGR3A, CGR3B, FN1, FOXP3, FYN, FZD1, GAPD, GLI2, GNLY, GOLPH4, GRB2, GSK3B, GSTP1, GUSB, GZMA, GZMH, GZMK, HLA-B, HLA-C, HLA-D, MA, HLA-DMB, HLA-DOA, HLA-DOB, HLA-DPA1, HLA-DQA2, HLA-DRA, HLX1, HMOX1, HRAS, HSPB3, HUWE1, ICAM1, ICAM-2, ICOS, ID1, ifna1, ifna17, ifna2, ifna5, ifna6, ifna8, IFNAR1, IFNAR2, IFNG, IFNGR1, IFNGR2, IGF1, IHH, IKBKB, IL10, IL12A, IL12B, IL12RB1, IL12RB2, IL13, IL13RA2, IL15, IL15RA, IL17, IL17R, IL17RB, IL18, IL1A, IL1B, IL1RI, IL2, IL21, IL21R, IL23A, IL23R, IL24, IL27, IL2RA, IL2RB, IL2RG, IL3, IL31RA, IL4, IL4RA, IL5, IL6, IL7, IL7RA, IL8, CXCR1, CXCR2, IL9, IL9R, IRF1, ISGF3G, ITGA4, ITGA7, integrin, alpha E (antigen CD103, human mucosal lymphocyte, antigen 1; alpha polypeptide), Gene hCG33203, ITGB3, JAK2, JAK3, KLRB1, KLRC4, KLRF1, KLRG1, KRAS, LAG3, LAIR2, LEF1, LGALS9, LILRB3, LRP2, LTA, SLAMF3, MADCAM1, MADH3, MADH7, MAF, MAP2K1, MDM2, MICA, MICB, MKI67, MMP12, MMP9, MTA1, MTSS1, MYC, MYD88, MYH6, NCAM1, NFATC1, NKG7, NLK, NOS2A, P2X7, PDCD1, PECAM-1, CXCL4, PGK1, PIAS1, PIAS2, PIAS3, PIAS4, PLAT, PML, PP1A, CXCL7, PPP2CA, PRF1, PROM1, PSMB5, PTCH, PTGS2, PTP4A3, PTPN6, PTPRC, RAB23, RAC/RHO, RAC2, RAF, RB1, RBL1, REN, Drosha, SELE, SELL, SELP, SERPINE1, SFRP1, SIRP beta 1, SKI, SLAMF1, SLAMF6, SLAMF7, SLAMF8, SMAD2, SMAD4, SMO, SMOH,

SMURF1, SOCS1, SOCS2, SOCS3, SOCS4, SOCS5, SOCS6, SOCS7, SOD1, SOD2, SOD3, SOS1, SOX17, CD43, ST14, STAM, STAT1, STAT2, STAT3, STAT4, STAT5A, STAT5B, STAT6, STK36, TAP1, TAP2, TBX21, TCF7, TERT, TFRC, TGFA, TGFB1, TGFB1, TGFB2, TIMP3, TLR1, TLR1, TLR2, TLR3, TLR4, TLR5, TLR6, TLR7, TLR8, TLR9, TNF, TNFRSF10A, TNFRSF11A, TNFRSF18, TNFRSF1A, TNFRSF1B, OX-40, TNFRSF5, TNFRSF6, TNFRSF7, TNFRSF8, TNFRSF9, TNFRSF10, TNFRSF6, TOB1, TP53, TSLP, VCAM1, VEGF, WIF1, WNT1, WNT4, XCL1, XCR1, ZAP70 and ZIC2.

[00159] In some embodiment, a treatment regime or a therapy can be administered via any common route so long as the target tissue or cell is available via that route. This includes, but is not limited to, intravenous, catheterization, orthotopic, intradermal, subcutaneous, intramuscular, intraperitoneal intertumoral, oral, nasal, buccal, rectal, vaginal, or topical administration. Selection of therapeutic agents and dosage regimes may depend on various factors, such as the drug combination employed, the particular disease being treated, and the condition and prior history of the patient.

[00160] Referring to block 202, in some embodiments, the methods include sequencing, for each respective subject in the plurality of training subjects, genomic DNA from the corresponding biological sample from the gut of the respective training subject, thereby obtain the corresponding plurality of nucleic acid sequences. In some embodiments, the corresponding plurality of nucleic acid sequences comprises at least 10,000, at least 100,000, at least 250,000, at least 500,000, at least 1,000,000, at least 2,500,000, at least 5,000,000, at least 10,000,000 or at least 50,000,000 nucleic acid sequences. In some embodiments, the corresponding plurality of nucleic acid sequences comprises no more than 250,000,000, no more than 100,000,000, no more than 50,000,000, no more than 25,000,000, no more than 10,000,000, no more than 5,000,000, no more than 1,000,000, no more than 100,000 nucleic acid sequences. In some embodiments, the corresponding plurality of nucleic acid sequences consists of from 100,000 to 1,000,000, from 200,000 to 5,000,000, from 500,000 from 10,000,000, from 1,000,000 to 20,000,000, from 5,000,000 to 50,000,000, from 10,000,000 to 100,000,000, or from 50,000,000 to 250,000,000 nucleic acid sequences. In some embodiments, the corresponding plurality of nucleic acid sequences falls within another range starting no lower than 1000 nucleic acid sequences and ending no higher than 250,000,000 nucleic acid sequences.

[00161] In some embodiments, the corresponding plurality of nucleic acid sequences are obtained through metagenomic sequencing, e.g., as disclosed in U.S. Patent Application Publication No. 2016/0239602 or U.S. Patent No. 11,495,326, the contents of which are incorporated herein by reference in their entireties. In some embodiments, metagenomes sequencing further comprise generating the plurality of metagenomic fragment reads. In some embodiments, metagenomic sequencing further comprise fragmenting microbial genomes into random fragments of targeted sizes. The resulting fragments can vary in size. In one embodiment, fragments of approximately 500 nucleotides can be obtained. In some embodiments, fragments of from 100-2000 nucleotides, e.g., 200-800, 100-900, 100-1000, 300-800, 400-900 nucleotides can be obtained. In some embodiments, the method may further comprise extracting the metagenomic fragments from the corresponding biological sample. In some embodiments, metagenomes sequencing further comprise sequencing the fragments using high throughput sequencing methods to generate a plurality of sequencing reads.

[00162] In some embodiments, the corresponding plurality of nucleic acid sequences are obtained through targeted panel sequencing. An example of targeted panel sequencing is described in U.S. Patent Application Publication No. 2019/0316209. In some embodiments, the targeted panel sequencing comprises hybridizing genomic DNA isolated from a biological sample from the gut of a subject with a panel of probes that include one or more probes that hybridize to a unique sequence in the genome of each microorganism being quantified, prior to sequencing recovered nucleic acids. In some embodiments, the microorganisms include a plurality of the microorganisms listed in Table 1, Table 2, and/or Figures 13A-13XX. In some embodiments, a combination of semi-unique sequences (e.g., sequences found in a small number of the microorganism genomes) can be used to deconvolute genomic abundance values using an algorithm, e.g., a system of equations. In some embodiments, the panel of probes includes at least 1 probe that hybridizes to a sequence unique to each microorganism genome being detected. In some embodiments, the panel of probes includes at least 2, at least 3, at least 4, at least 5, at least 10, at least 25, at least 50, or more probes that hybridize to a different sequence unique to each microorganism genome being detected. In some embodiments, the panel of probes includes at least 20, at least 30, at least 40, at least 50, at least 75, at least 100, at least 125, at least 150, at least 200, at least 150, at least 300, at least 400, at least 500, at least 750, at

least 1000, at least 1250, at least 1500, at least 2000, at least 2500, at least 3000, at least 4000, at least 5000, at least 7500, at least 10,000 or more unique probes.

[00163] In some embodiments, the sequencing genomic DNA from the corresponding biological sample comprises a partial or complete sequencing platform adapter sequence at their termini useful for sequencing using a sequencing platform of interest. Sequencing platforms of interest include, but are not limited to, the HiSeq™, MiSeq™ and Genome Analyzer™ sequencing systems from Illumina®; the Ion PGM™ and Ion Proton™ sequencing systems from Ion Torrent™; the PACBIO RS II Sequel system from Pacific Biosciences, the SOLiD sequencing systems from Life Technologies™, the 454 GS FLX+ and GS Junior sequencing systems from Roche, the MinION™ system from Oxford Nanopore, or any other sequencing platform of interest.

[00164] Referring to block 204, in some embodiments, the methods include obtaining, for each respective training subject in the plurality of training subjects, in electronic form, a corresponding plurality of nucleic acid sequences for genomic DNA from the corresponding biological sample from the gut of the respective training subject.

[00165] Referring to block 206, in some embodiments, the methods include determining, for each respective gut microorganism in the plurality of gut microorganisms, the corresponding value for the abundance of the genome of the respective gut microorganism from the corresponding plurality of nucleic acid sequences.

[00166] In some embodiments, the genomic abundance values determined for each respective subject in the plurality of training subjects comprise at least 20, at least 25, at least 50, at least 75, at least 100, at least 125, at least 150, at least 175, at least 200, at least 250, at least 300, at least 350, at least 400, at least 500, at least 600, at least 700, at least 700, at least 800, at least 900, at least 1000, at least 1500, at least 2000, at least 25000, at least 5,000 or at least 10,000 genome abundance values, where each genome abundance value corresponds to different gut microorganism. In some embodiments, the genomic abundance values determined for each respective subject in the plurality of training subjects comprise no more than 250,000, no more than 100,000, no more than 50,000, no more than 25,000, no more than 10,000, no more than 5000, no more than 2500, no more than 1000, no more than 750, no more than 500, or fewer genome abundance values. In some embodiments, the genomic abundance values determined for

each respective subject in the plurality of training subjects consist of from 10 to 40, from 20 to 50, from 30 to 80, from 40 to 100, from 50 to 150, from 60 to 200, from 80 to 300, from 90 to 500, from 100 to 1000, from 500 to 2,000, or from 1,000 to 5,000 genome abundance values. In some embodiments, the genomic abundance values determined for each respective subject in the plurality of training subjects fall within another range starting no lower than 10 genome abundance values and ending no higher than 250,000 genome abundance values.

[00167] Referring to block 208, in some embodiments, for each respective training subject in the plurality of training subjects, the methods include assembling, in electronic form, a corresponding plurality of gut microorganism genomes by metagenomic de novo sequence assembly from the corresponding plurality of nucleic acid sequences, and calculating, for each respective gut microorganism in the plurality of gut microorganisms, the corresponding value for the abundance of the genome of the respective gut microorganism based on the prevalence of respective nucleic acid sequences, in the plurality of nucleic acid sequences, used to assemble a respective gut microorganism genome in the plurality of gut microorganism genomes corresponding to the respective gut microorganism. In some embodiments, metagenomic de novo sequence assembly further comprise generating contigs based on the sequencing reads generated by a shotgun sequencing technique. Such a technique is described, for example, in U.S. Patent No. 10,529,443, the content of which is incorporated herein by reference in its entirety. In some embodiments, the first plurality of nucleic acid sequences is assembled into full genomes of the plurality of gut microorganisms. In some embodiments, the plurality of nucleic acid sequences is assembled into partial genomes of the plurality of gut microorganisms.

[00168] Referring to block 210, in some embodiments, for each respective subject in the plurality of training subjects, the methods including assigning each respective nucleic acid sequence in the corresponding plurality of nucleic acid sequences to a respective gut microorganism in the plurality of gut microorganisms, thereby generating, for each respective gut microorganism in the plurality of gut microorganism, a corresponding count of respective nucleic acid sequences in the corresponding plurality of nucleic acid sequences assigned to the respective gut microorganism, and determining, for each respective gut microorganism in the plurality of gut microorganisms, the corresponding genomic abundance value for the respective gut microorganism based on the corresponding count of respective nucleic acid sequences assigned to the respective gut microorganism. In some embodiments, the assigning each

respective nucleic acids to a respective gut microorganism includes mapping the nucleic acid to a reference nucleic acid (e.g., a contig listed in FIG.12) In some embodiments, the assigning each respective nucleic acids a respective gut microorganism includes annotating genome information based on existing databases. In some embodiments, nucleic acid sequences are analyzed, and annotations are to define taxonomic assignments using sequence similarity and phylogenetic placement methods or a combination of the two strategies.

[00169] Sequence similarity-based methods for assigning each nucleic acid sequence to a respective gut microorganism include those familiar to individuals skilled in the art including, but not limited to BLAST, BLASTx, tBLASTn, tBLASTx, RDP-classifier, DNAClust, and various implementations of these algorithms such as Qiime or Mothur. These methods rely on mapping a sequence read to a reference database and selecting the match with the best score and e-value. In some embodiments, phylogenetic methods are used in combination with sequence similarity methods to improve the calling accuracy of an annotation or taxonomic assignment. Common databases include, but not limited to, GT-DBTK, National Center for Biotechnology Information (NCBI) Genbank, European Bioinformatics Institute-European Nucleotide Archive (European Bioinformatics Institute-European Nucleotide Archive; EBI-ENA) , National Institute of Genetics, U.S. Department of ENERGY (USDOE) Integrated Microbial Genomes (Integrated Microbial Genomes) &Microbiomes; IMG/M) and other available databases in the art.

[00170] In some embodiments, the plurality of genomic abundance values is determined using a microarray comprising a probe sequence capable of detecting a unique genomic sequence of each respective genome for the plurality of gut microorganisms. In some embodiments, the panel of probes on a microarray includes at least 1 probe that hybridizes to a sequence unique to each microorganism genome being detected. In some embodiments, the panel of probes includes at least 2, at least 3, at least 4, at least 5, at least 10, at least 25, at least 50, or more probes that hybridize to a different sequence unique to each microorganism genome being detected. In some embodiments, the panel of probes includes at least 20, at least 30, at least 40, at least 50, at least 75, at least 100, at least 125, at least 150, at least 200, at least 150, at least 300, at least 400, at least 500, at least 750, at least 1000, at least 1250, at least 1500, at least 2000, at least 2500, at least 3000, at least 4000, at least 5000, at least 7500, at least 10,000 or more unique probes.

[00171] Referring to block 212, in some embodiments, the plurality of gut microorganisms comprises at least 20 gut microorganisms selected from Table 1, Table 2, or Figure 13A-13XX. In some embodiments, gut microorganisms of at least about 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100 or greater are selected from Table 1, Table 2 or Figure 13A-13XX. In some embodiments, the plurality of gut microorganisms comprises at least 25 gut microorganisms selected from Table 1, Table 2, or Figures 13A-13XX. In some embodiments, the plurality of gut microorganisms comprises at least 30 gut microorganisms selected from Table 1, Table 2, or Figures 13A-13XX. In some embodiments, the plurality of gut microorganisms comprises at least 40 gut microorganisms selected from Table 1, Table 2, or Figures 13A-13XX. In some embodiments, the plurality of gut microorganisms comprises at least 15, at least 20, at least 25, at least 30, at least 35, at least 40, at least 45, at least 50, at least 55, at least 60, at least 65, at least 70, at least 75, at least 80, at least 85, at least 90, at least 95, at least 100, at least 125, at least 150, at least 200, at least 250, at least 300, at least 400, at least 500, at least 600, at least 700, at least 800, or all of the gut microorganisms selected from Table 1, Table 2 or Figures 13A-13XX. In some embodiments, the plurality of gut microorganisms are at least 20, at least 25, at least 30, at least 35, at least 40, at least 45, at least 50, at least 55, at least 60, at least 65, at least 70, at least 75, at least 80, at least 85, at least 90, at least 95, at least 100, at least 125, at least 150, at least 200, at least 250, at least 300, at least 400, at least 500, at least 600, at least 700, at least 800, or all of the gut microorganisms selected from Table 1, Table 2 or Figures 13A-13XX. In some embodiments, the plurality of gut microorganisms are all of the gut microorganisms listed in Table 1. In some embodiments, the plurality of gut microorganisms are all of the gut microorganisms listed in Table 2. In some embodiments, the plurality of gut microorganisms are all of the gut microorganisms listed in Figures 13A-13XX.

[00172] **Table 1** - Taxonomy Assignment of 141 non-redundant genomes identified in two competing guilds (QD-TCG)

ID	Taxonomy Assignment (d=domain; p=phylum; c=class; o=order; f=family; g=genus; s = species)	Guilds
1U030.25	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Faecalibacterium;s__Faecalibacterium prausnitzii_G	Guild 1
1U035.49	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__CAG-103;s__	Guild 1

ID	Taxonomy Assignment (d=domain; p=phylum; c=class; o=order; f=family; g=genus; s = species)	Guilds
1U036.46	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__CAG-103;s	Guild 1
1W012.66	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__ER4;s__ER4 sp900317525	Guild 1
1W030.27	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__Oscillibacter;s__Oscillibacter sp001916835	Guild 1
1W037.25	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Faecalibacterium;s__Faecalibacterium prausnitzii K	Guild 1
1W059.36	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Faecalibacterium;s__Faecalibacterium prausnitzii K	Guild 1
1W065.34	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__Oscillibacter;s__Oscillibacter sp001916835	Guild 1
1W071.30	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Gemmiger;s	Guild 1
1W071.47	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Faecalibacterium;s__Faecalibacterium prausnitzii G	Guild 1
2U004.52	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__CAG-103;s	Guild 1
2U017.37	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Faecalibacterium;s__Faecalibacterium prausnitzii G	Guild 1
2U033.43	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Ruminococcus C;s	Guild 1
2W001.22	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Faecalibacterium;s__Faecalibacterium prausnitzii K	Guild 1
2W007.23	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__CAG-110;s__CAG-110 sp000434635	Guild 1
2W009.30	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Faecalibacterium;s__Faecalibacterium prausnitzii G	Guild 1
2W023.30	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Acetivibacteraceae;g__UBA1417;s	Guild 1
2W027.13	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Eisenbergiella;s__Eisenbergiella sp900066775	Guild 1
2W027.27	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Faecalibacterium;s__Faecalibacterium prausnitzii D	Guild 1
2W029.13	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Faecalibacterium;s__Faecalibacterium prausnitzii C	Guild 1

ID	Taxonomy Assignment (d=domain; p=phylum; c=class; o=order; f=family; g=genus; s = species)	Guilds
2W034.32	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__CAG-110;s__	Guild 1
2W062.21	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__Oscillibacter;s__Oscillibacter sp001916835	Guild 1
2W064.32	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Faecalibacterium;s__Faecalibacterium prausnitzii D	Guild 1
2W067.69	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__ER4;s__ER4 sp000765235	Guild 1
2W073.18	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Christensenellales;f__UBA1750;g__UBA7102;s__	Guild 1
2W073.55	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Butyricicoccaceae;g__Butyricoccus_A;s__Butyricoccus_A sp002395695	Guild 1
2W074.14	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Faecalibacterium;s__Faecalibacterium prausnitzii K	Guild 1
2W080.51	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Faecalibacterium;s__Faecalibacterium prausnitzii I	Guild 1
3U005.26	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Gemmiger;s__Gemmiger formicilis	Guild 1
3U010.20	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__CAG-110;s__	Guild 1
3U024.38	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Acetivibacteraceae;g__CAG-180;s__CAG-180 sp000432435	Guild 1
3U033.37	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__CAG-110;s__	Guild 1
3W015.22	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Faecalibacterium;s__Faecalibacterium prausnitzii J	Guild 1
3W048.18	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Gemmiger;s__	Guild 1
3W052.32	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Gemmiger;s__	Guild 1
3W062.32	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__CAG-353;s__CAG-353 sp900066885	Guild 1
3W072.38	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Faecalibacterium;s__Faecalibacterium prausnitzii G	Guild 1
3W073.49	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Christensenellales;f__CAG-138;g__Phil1;s__Phil1 sp001940855	Guild 1

ID	Taxonomy Assignment (d=domain; p=phylum; c=class; o=order; f=family; g=genus; s = species)	Guilds
1U009.16	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Faecalibacterium;s__Faecalibacterium prausnitzii G	Guild 1
1W011.28	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__Oscillibacter;s__Oscillibacter sp001916835	Guild 1
1W001.34	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Acetivibacteraceae;g__CAG-177;s__CAG-177 sp003514385	Guild 1
1U014.38	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Faecalibacterium;s__Faecalibacterium prausnitzii K	Guild 1
1W037.27	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Gemmiger;s__Gemmiger formicilis	Guild 1
1U002.48	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Ruminococcus C;s__Ruminococcus C callidus	Guild 1
2W066.63	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__ER4;s__ER4 sp900317525	Guild 1
2W043.36	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Gemmiger;s__Gemmiger formicilis	Guild 1
2W043.20	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Eisenbergiella;s__Eisenbergiella sp900066775	Guild 1
2W001.17	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Gemmiger;s__Gemmiger sp003476825	Guild 1
1U002.41	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Dorea;s__Dorea longicatena B	Guild 1
1W071.19	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Coprococcus A;s__Coprococcus A catus	Guild 1
1U017.16	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Blautia;s__	Guild 2
1U017.25	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Tyzzerella;s__Tyzzerella sp000209385	Guild 2
1U028.21	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Dorea;s__Dorea scindens	Guild 2
1U030.5	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Escherichia;s__Escherichia coli D	Guild 2
1U035.55	d__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;s__Streptococcus macedonicus	Guild 2
1U038.18	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__UBA7096;s__	Guild 2
1U038.6	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Escherichia;s__Escherichia coli D	Guild 2
1W006.13	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Faecalicatena;s__Faecalicatena faecis	Guild 2

ID	Taxonomy Assignment (d=domain; p=phylum; c=class; o=order; f=family; g=genus; s = species)	Guilds
1W032.17	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Paraprevotella;s__	Guild 2
1W032.19	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Eubacterium_G;s__Eubacterium_G sp000434315	Guild 2
1W033.3	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__	Guild 2
1W033.9	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__Flavonifractor;s__Flavonifractor plautii	Guild 2
1W036.28	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Tyzzerella;s__Tyzzerella nexilis	Guild 2
1W038.31	d__Bacteria;p__Fusobacteriota;c__Fusobacteriia;o__Fusobacteriales;f__Fusobacteriaceae;g__Fusobacterium_A;s__	Guild 2
1W042.8	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Clostridium_M;s__Clostridium_M sp001517625	Guild 2
1W047.42	d__Bacteria;p__Firmicutes_C;c__Negativicutes;o__Acidaminococcales;f__Acidaminococcaceae;g__Acidaminococcus;s__Acidaminococcus intestini	Guild 2
1W062.68	d__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;s__Streptococcus sp002300045	Guild 2
1W070.8	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Peptostreptococcales;f__Peptostreptococcaceae;g__Intestinibacter;s__Intestinibacter bartlettii	Guild 2
1W072.11	d__Bacteria;p__Firmicutes;c__Bacilli;o__Erysipelotrichales;f__Erysipelotrichaceae;g__Absiella;s__	Guild 2
1W075.10	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Blautia_A;s__Blautia_A wexlerae	Guild 2
1W075.14	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Marinifilaceae;g__Odoribacter;s__	Guild 2
1W078.21	d__Bacteria;p__Firmicutes_C;c__Negativicutes;o__Veillonellales;f__Veillonellaceae;g__Veillonella;s__Veillonella parvula_A	Guild 2
1W080.50	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__RF32;f__CAG-239;g__51-20;s__51-20 sp001917175	Guild 2
2U001.12	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Blautia_A;s__Blautia_A hydrogenotrophica	Guild 2
2U001.17	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Clostridium_M;s__Clostridium_M sp001517625	Guild 2
2U001.37	d__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Lactobacillaceae;g__Lactobacillus_H;s__Lactobacillus_H oris	Guild 2
2U008.16	d__Bacteria;p__Fusobacteriota;c__Fusobacteriia;o__Fusobacteriales;f__Fusobacteriaceae;g__Fusobacterium_A;s__Fusobacterium_A ulcerans	Guild 2
2U017.6	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Klebsiella;s__Klebsiella variicola	Guild 2

ID	Taxonomy Assignment (d=domain; p=phylum; c=class; o=order; f=family; g=genus; s = species)	Guilds
2U033.18	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides_A;s__Bacteroides_A sp000434735	Guild 2
2U037.14	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides clarus	Guild 2
2U037.18	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides_A;s__Bacteroides_A plebeius	Guild 2
2U037.6	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Escherichia;s__Escherichia flexneri	Guild 2
2U038.12	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Clostridium_Q;s__Clostridium_Q symbiosum	Guild 2
2U038.5	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Blautia;s__Blautia sp001304935	Guild 2
2W001.42	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__RF32;f__CAG-239;g__CAG-495;s__CAG-495 sp000432275	Guild 2
2W002.29	d__Bacteria;p__Actinobacteriota;c__Actinobacteria;o__Actinomycetales;f__Bifidobacteriaceae;g__Bifidobacterium;s__Bifidobacterium longum	Guild 2
2W003.7	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides_B;s__Bacteroides_B vulgatus	Guild 2
2W004.25	d__Bacteria;p__Firmicutes_C;c__Negativicutes;o__Veillonellales;f__Megasphaeraceae;g__Caecibacter;s__	Guild 2
2W005.9	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Tannerellaceae;g__Parabacteroides;s__Parabacteroides johnsonii	Guild 2
2W011.21	d__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus_B;s__Enterococcus B hirae	Guild 2
2W012.75	d__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;s__Streptococcus thermophilus	Guild 2
2W025.8	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Enterobacter;s__Enterobacter kobei	Guild 2
2W026.44	d__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;s__Streptococcus infantarius	Guild 2
2W037.10	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Tannerellaceae;g__Parabacteroides;s__Parabacteroides sp000436495	Guild 2
2W048.49	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Peptostreptococcales;f__Anaerovoracaceae;g__UBA1191;s__	Guild 2
2W050.17	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides stercoris	Guild 2
2W053.46	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Burkholderiales;f__Burkholderiaceae;g__Oxalobacter;s__Oxalobacter formigenes_A	Guild 2
2W054.20	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Blautia_A;s__Blautia_A sp900066165	Guild 2

ID	Taxonomy Assignment (d=domain; p=phylum; c=class; o=order; f=family; g=genus; s = species)	Guilds
2W065.60	d__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;s__Streptococcus salivarius	Guild 2
2W067.20	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides A;s__Bacteroides A coprocola	Guild 2
2W075.14	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Faecalicatena;s__Faecalicatena gnavus	Guild 2
2W080.58	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__RF32;f__CAG-239;g__CAG-495;s__	Guild 2
3U007.7	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Clostridium M;s__Clostridium M bolteae	Guild 2
3U020.14	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Eubacterium I;s__Eubacterium I ramulus	Guild 2
3U023.7	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Clostridium M;s__	Guild 2
3U038.19	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Blautia A;s__Blautia A sp000433815	Guild 2
3U040.11	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides stercoris	Guild 2
3W002.16	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Clostridium_Q;s__Clostridium_Q saccharolyticum	Guild 2
3W023.3	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides cellulosilyticus	Guild 2
3W028.33	d__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;s__Streptococcus pasteurianus	Guild 2
3W036.5	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides B;s__Bacteroides B dorei	Guild 2
3W036.6	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides fragilis	Guild 2
3W039.10	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides eggerthii	Guild 2
3W042.11	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides A;s__Bacteroides A plebeius A	Guild 2
3W042.9	d__Bacteria;p__Firmicutes;c__Bacilli;o__Erysipelotrichales;f__Erysipelatoclostridiaceae;g__Erysipelatoclostridium;s__Erysipelatoclostridium ramosum	Guild 2
3W049.14	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides stercoris	Guild 2
3W054.11	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Escherichia;s__Escherichia flexneri	Guild 2
3W055.19	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Barnesiella;s__Barnesiella intestinihominis	Guild 2

ID	Taxonomy Assignment (d=domain; p=phylum; c=class; o=order; f=family; g=genus; s = species)	Guilds
3W067.18	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Muribaculaceae;g__CAG-279;s__	Guild 2
3W074.2	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides thetaiotaomicron	Guild 2
3W079.9	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides A;s__Bacteroides A coprophilus	Guild 2
3W080.19	d__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus D;s__Enterococcus D sp002850555	Guild 2
3W080.64	d__Bacteria;p__Firmicutes;c__Bacilli;o__RFN20;f__CAG-288;g__UBA7642;s__	Guild 2
1U003.28	d__Bacteria;p__Actinobacteriota;c__Actinobacteria;o__Actinomycetales;f__Micrococcaceae;g__Rothia;s__Rothia mucilaginosa B	Guild 2
1U017.15	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Sellimonas;s__Sellimonas intestinalis	Guild 2
1U001.8	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Escherichia;s__Escherichia coli	Guild 2
1U017.10	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Dorea;s__Dorea sp000509125	Guild 2
1U027.33	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Acutalibacteraceae;g__Eubacterium_R;s__Eubacterium_R sp000431535	Guild 2
1U015.30	d__Bacteria;p__Firmicutes;c__Bacilli;o__Erysipelotrichales;f__Erysipelotrichaceae;g__Absiella;s__Absiella dolichum	Guild 2
1U036.9	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Paraprevotella;s__Paraprevotella clara	Guild 2
1U015.20	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Faecalicatena;s__Faecalicatena glycyrrhizinilyticum	Guild 2
1W023.6	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Klebsiella;s__Klebsiella pneumoniae	Guild 2
1W023.15	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides A;s__Bacteroides A plebeius	Guild 2
1U024.11	d__Bacteria;p__Actinobacteriota;c__Coriobacteriia;o__Coriobacteriales;f__Eggerthellaceae;g__Eggerthella;s__Eggerthella lenta	Guild 2
1W023.20	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__Lawsonibacter;s__	Guild 2
1U036.17	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides stercoris	Guild 2
1U021.14	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides A;s__Bacteroides A coprocola	Guild 2
1W026.10	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Blautia_A;s__	Guild 2

ID	Taxonomy Assignment (d=domain; p=phylum; c=class; o=order; f=family; g=genus; s = species)	Guilds
1U014.18	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Clostridium_M;s__Clostridium_M sp000431375	Guild 2
2W067.16	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Copro bacteraceae;g__Coprobacter;s__Coprobacter fastidiosus	Guild 2
1W079.48	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__UBA932;g__RC9;s__RC9 sp000431015	Guild 2

[00173] Table 2 - Taxonomy Assignment of 284 core microbiome (CC-TCG)

ID	Taxonomy Assignment (d=domain; p=phylum; c=class; o=order; f=family; g=genus; s = species)	Assignment
1U003.28	d__Bacteria;p__Actinobacteriota;c__Actinobacteria;o__Actinomycetales;f__Micrococcaceae;g__Rothia;s__Rothia mucilaginosa_B	Guild 2
1U015.20	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Faecalicatena;s__Faecalicatena glycyrrhizinilyticum	Guild 2
1U024.11	d__Bacteria;p__Actinobacteriota;c__Coriobacteriia;o__Coriobacteriales;f__Eggerthellaceae;g__Eggerthella;s__Eggerthella lenta	Guild 2
1U035.49	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__CAG-103;s__	Guild 1
1U035.55	d__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;s__Streptococcus macedonicus	Guild 2
1U038.18	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__UBA7096;s__	Guild 2
1W011.28	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__Oscillibacter;s__Oscillibacter sp001916835	Guild 1
1W023.20	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__Lawsonibacter;s__	Guild 2
1W030.27	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__Oscillibacter;s__Oscillibacter sp001916835	Guild 1
1W032.19	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Eubacterium_G;s__Eubacterium_G sp000434315	Guild 2
1W033.9	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__Flavonifractor;s__Flavonifractor plautii	Guild 2
1W042.8	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Clostridium_M;s__Clostridium_M sp001517625	Guild 2
1W065.34	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__Oscillibacter;s__Oscillibacter sp001916835	Guild 1
1W072.11	d__Bacteria;p__Firmicutes;c__Bacilli;o__Erysipelotrichales;f__Erysipelotrichaceae;g__Absiella;s__	Guild 2
1W075.14	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Marinifilaceae;g__Odoribacter;s__	Guild 2

ID	Taxonomy Assignment (d=domain; p=phylum; c=class; o=order; f=family; g=genus; s = species)	Assignment
1W078.2 1	d__Bacteria;p__Firmicutes_C;c__Negativicutes;o__Veillonellales;f__Veillonellaceae;g__Veillonella;s__Veillonella parvula A	Guild 2
1W080.5 0	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__RF32;f__CAG-239;g__51-20;s__51-20 sp001917175	Guild 2
2U001.1 7	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Clostridium M;s__Clostridium M sp001517625	Guild 2
2U008.1 6	d__Bacteria;p__Fusobacteriota;c__Fusobacteriia;o__Fusobacteriales;f__Fusobacteriaceae;g__Fusobacterium_A;s__Fusobacterium_A ulcerans	Guild 2
2U033.1 8	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides_A;s__Bacteroides_A sp000434735	Guild 2
2U037.1 4	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides clarus	Guild 2
2W001.4 2	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__RF32;f__CAG-239;g__CAG-495;s__CAG-495 sp000432275	Guild 2
2W011.2 1	d__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus B;s__Enterococcus B hirae	Guild 2
2W023.3 0	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Acetivibacteraceae;g__UBA1417;s__	Guild 1
2W027.2 7	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Faecalibacterium;s__Faecalibacterium prausnitzii D	Guild 1
2W048.4 9	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Peptostreptococcales;f__Anaerovoracaceae;g__UBA1191;s__	Guild 2
2W053.4 6	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Burkholderiales;f__Burkholderiaceae;g__Oxalobacter;s__Oxalobacter formigenes A	Guild 2
2W065.6 0	d__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;s__Streptococcus salivarius	Guild 2
2W073.5 5	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Butyrivibrioaceae;g__Butyrivibrio_A;s__Butyrivibrio_A sp002395695	Guild 1
3U020.1 4	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Eubacterium I;s__Eubacterium I ramulus	Guild 2
3U023.7	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Clostridium M;s__	Guild 2
3W002.1 6	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Clostridium Q;s__Clostridium Q saccharolyticum	Guild 2
3W015.2 2	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Faecalibacterium;s__Faecalibacterium prausnitzii J	Guild 1
3W039.1 0	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides eggerthii	Guild 2

ID	Taxonomy Assignment (d=domain; p=phylum; c=class; o=order; f=family; g=genus; s = species)	Assignment
3W055.19	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Barnesiellaceae;g__Barnesiella;s__Barnesiella intestinihominis	Guild 2
3W062.32	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__CAG-353;s__CAG-353 sp900066885	Guild 1
3W067.18	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Mucrobacteriaceae;g__CAG-279;s__	Guild 2
3W074.2	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides thetaiotaomicron	Guild 2
ERR1018189.10	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__CAG-110;s__CAG-110 sp900546075	Guild 2
ERR1018192.4	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Blautia A;s__Blautia A sp900066145	Guild 2
ERR1018196.11	d__Bacteria;p__Firmicutes_C;c__Negativicutes;o__Veillonellales;f__Dialisteraceae;g__Dialister;s__Dialister succinatiphilus	Guild 2
ERR1018197.5	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Acetivibacteraceae;g__UMGS263;s__UMGS263 sp900540535	Guild 1
ERR1018197.9	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Acetivibacteraceae;g__Clostridium A;s__Clostridium A leptum	Guild 1
ERR1018200.15	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Acetivibacteraceae;g__UMGS1071;s__UMGS1071 sp900542375	Guild 2
ERR1018200.4	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Eisenbergiella;s__Eisenbergiella tayi	Guild 2
ERR1018204.1	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Enterocloster;s__Enterocloster clostridioformis	Guild 2
ERR1018210.9	d__Bacteria;p__Actinobacteriota;c__Actinomycetia;o__Actinomycetales;f__Bifidobacteriaceae;g__Bifidobacterium;s__Bifidobacterium pseudocatenulatum	Guild 2
ERR1018213.10	d__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;s__Streptococcus salivarius	Guild 2
ERR1018218.4	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella hominis	Guild 2
ERR1018219.12	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Acetivibacteraceae;g__Eubacterium R;s__Eubacterium R sp000434995	Guild 2
ERR1018219.9	d__Bacteria;p__Desulfobacterota;c__Desulfovibrionia;o__Desulfovibrionales;f__Desulfovibrionaceae;g__Bilophila;s__Bilophila wadsworthia	Guild 2
ERR1018229.14	d__Bacteria;p__Firmicutes_C;c__Negativicutes;o__Veillonellales;f__Megasphaeraeae;g__Caecibacter;s__Caecibacter hominis	Guild 2
ERR1018232.13	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella copri	Guild 2
ERR1018237.11	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Eubacterium G;s__Eubacterium G ventriosum	Guild 2
ERR1018238.17	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Anaerostipes;s__Anaerostipes hadrus	Guild 2

ID	Taxonomy Assignment (d=domain; p=phylum; c=class; o=order; f=family; g=genus; s = species)	Assignment
ERR1018 239.7	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Blautia_A;s__Blautia_A sp003480185	Guild 2
ERR1018 250.9	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Enterocloster;s__Enterocloster sp900541315	Guild 2
ERR1018 258.4	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Tannerellaceae;g__Parabacteroides;s__Parabacteroides distasonis	Guild 2
ERR1018 262.5	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides uniformis	Guild 2
ERR1018 267.4	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Copro bacteraceae;g__Copro bacter;s__Copro bacter fastidiosus	Guild 2
ERR1018 278.5	d__Bacteria;p__Firmicutes_C;c__Negativicutes;o__Veillonellales;f__Dialisteraceae;g__Dialister;s__Dialister sp900541605	Guild 2
ERR1018 291.11	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__ER4;s__ER4 sp900317525	Guild 1
ERR1018 292.2	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides intestinalis	Guild 2
ERR1018 298.1	d__Bacteria;p__Firmicutes;c__Bacilli;o__Erysipelotrichales;f__Erysipelotrichaceae;g__Faecalitalea;s__Faecalitalea cylindroides	Guild 2
ERR1018 301.12	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Marinifilaceae;g__Butyricimonas;s__Butyricimonas virosa	Guild 2
ERR1018 308.3	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Agathobacter;s__Agathobacter sp900548765	Guild 2
ERR1018 309.23	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Monoglobales_A;f__UBA1381;g__CAG-41;s__CAG-41 sp900066215	Guild 2
ERR2017 433.5	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Lachnospira;s__Lachnospira sp000437735	Guild 1
ERR2017 443.11	d__Bacteria;p__Firmicutes_C;c__Negativicutes;o__Acidaminococcales;f__Acidaminococcaceae;g__Phascolarctobacterium;s__Phascolarctobacterium faecium	Guild 2
ERR2017 473.2	d__Bacteria;p__Fusobacteriota;c__Fusobacteriia;o__Fusobacteriales;f__Fusobacteriaceae;g__Fusobacterium_A;s__Fusobacterium_A mortiferum	Guild 2
ERR2017 500.4	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola massiliensis	Guild 2
ERR2017 517.1	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__RUG115;s__RUG115 sp900066395	Guild 1
ERR2017 522.17	d__Bacteria;p__Firmicutes;c__Bacilli;o__Erysipelotrichales;f__Erysipelotrichaceae;g__Longicatena;s__Longicatena caecimuris	Guild 2
ERR2017 539.14	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Ruminococcus_C;s__Ruminococcus_C sp000433635	Guild 1
ERR2017 540.1	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Rikenellaceae;g__Alistipes_A;s__Alistipes_A indistinctus	Guild 2

ID	Taxonomy Assignment (d=domain; p=phylum; c=class; o=order; f=family; g=genus; s = species)	Assignment
ERR2017 541.4	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__CAG-95;s__CAG-95 sp900066375	Guild 1
ERR2017 552.8	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Gemmiger;s__Gemmiger formicilis	Guild 1
ERR2017 555.15	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Faecalibacterium;s__Faecalibacterium sp900540455	Guild 1
ERR2017 555.23	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Ruthenibacterium;s__Ruthenibacterium lactatiformans	Guild 2
ERR2017 563.14	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Ruminococcus_D;s__Ruminococcus_D bicirculans	Guild 1
ERR2017 569.9	d__Bacteria;p__Firmicutes_C;c__Negativicutes;o__Acidaminococcales;f__Acidaminococcaceae;g__Phascolarctobacterium;s__Phascolarctobacterium sp900545535	Guild 2
ERR2017 573.11	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella pectinovora	Guild 1
ERR2017 577.6	d__Bacteria;p__Firmicutes;c__Bacilli;o__Erysipelotrichales;f__Erysipelotrichaceae;g__Coprobaecillus;s__Coprobaecillus cateniformis	Guild 2
ERR2017 581.15	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__ER4;s__ER4 sp000765235	Guild 1
ERR2017 584.6	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Anaerobutyricum;s__Anaerobutyricum hallii	Guild 1
ERR2017 586.7	d__Bacteria;p__Firmicutes;c__Bacilli;o__Erysipelotrichales;f__Erysipelotrichaceae;g__Longibaculum;s__Longibaculum muris	Guild 2
ERR2017 591.13	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaecicola;s__Phocaecicola sp900542985	Guild 2
ERR2017 596.4	d__Bacteria;p__Actinobacteriota;c__Actinomycetia;o__Actinomycetales;f__Bifidobacteriaceae;g__Bifidobacterium;s__Bifidobacterium longum	Guild 2
ERR2017 601.3	d__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Lactobacillaceae;g__Limosilactobacillus;s__Limosilactobacillus fermentum	Guild 2
ERR2017 608.6	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__UBA5446;s__UBA5446 sp900546485	Guild 1
ERR2017 610.5	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Klebsiella;s__Klebsiella aerogenes	Guild 2
ERR2017 617.4	d__Bacteria;p__Firmicutes_C;c__Negativicutes;o__Veillonellales;f__Veillonellaceae;g__Veillonella;s__Veillonella parvula A	Guild 2
ERR2017 618.4	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Faecalibacterium;s__Faecalibacterium prausnitzii_D	Guild 1

ID	Taxonomy Assignment (d=domain; p=phylum; c=class; o=order; f=family; g=genus; s = species)	Assignment
ERR2017 622.3	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Coprococcus;s__Coprococcus sp900066115	Guild 1
ERR2017 636.11	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Blautia_A;s__Blautia_A sp003471165	Guild 2
ERR2017 638.17	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Acetivibacteraceae;g__Ruminococcus_E;s__Ruminococcus_E bromii_B	Guild 1
ERR2017 641.10	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__COE1;s__COE1 sp001916965	Guild 1
ERR2017 647.13	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Acetivibacteraceae;g__CAG-217;s__CAG-217 sp000436335	Guild 1
ERR2017 650.13	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__CAG-81;s__CAG-81 sp000435795	Guild 2
ERR2017 651.6	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Clostridium_AP;s__Clostridium_AP sp000509125	Guild 2
ERR2017 654.11	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides uniformis	Guild 2
ERR2017 659.20	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Acetivibacteraceae;g__CAG-180;s__CAG-180 sp004556705	Guild 1
ERR2017 660.10	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Eubacterium_G;s__Eubacterium_G sp000434315	Guild 1
ERR2017 669.11	d__Bacteria;p__Firmicutes;c__Bacilli;o__Erysipelotrichales;f__Erysipelotrichaceae;g__Clostridium_AQ;s__Clostridium_AQ sp003481775	Guild 1
ERR2017 669.13	d__Bacteria;p__Actinobacteriota;c__Coriobacteriia;o__Coriobacteriales;f__Eggerthellaceae;g__Eggerthella;s__Eggerthella lenta	Guild 2
ERR2017 669.2	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Acetivibacteraceae;g__UMGS1071;s__UMGS1071 sp900541905	Guild 2
ERR2017 672.11	d__Bacteria;p__Actinobacteriota;c__Coriobacteriia;o__Coriobacteriales;f__Coriobacteriaceae;g__Collinsella;s__Collinsella sp002232035	Guild 2
ERR2017 680.3	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__UBA9502;s__UBA9502 sp003480315	Guild 1
ERR2017 696.13	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__UBA9502;s__UBA9502 sp900555625	Guild 2
ERR2017 699.1	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Eisenbergiella;s__Eisenbergiella tayi	Guild 1
ERR2017 700.13	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__CAG-81;s__CAG-81 sp900066535	Guild 1
ERR2017 701.13	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Ruminococcus_D;s__Ruminococcus_D bicirculans	Guild 1
ERR2017 705.14	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella copri_A	Guild 2
ERR2017 706.14	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__Lawsonibacter;s__Lawsonibacter sp000177015	Guild 1

ID	Taxonomy Assignment (d=domain; p=phylum; c=class; o=order; f=family; g=genus; s = species)	Assignment
ERR2017 714.15	d__Bacteria;p__Firmicutes_C;c__Negativicutes;o__Acidaminococcales;f__Acidaminococcaceae;g__Phascolarctobacterium;s__Phascolarctobacterium faecium	Guild 2
ERR2017 716.3	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Negativibacillus;s__Negativibacillus massiliensis	Guild 1
ERR2017 716.6	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__Flavonifractor;s__Flavonifractor plautii	Guild 2
ERR2017 717.14	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__CAG-127;s__CAG-127 sp900319515	Guild 1
ERR2017 729.15	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Agathobacter;s__Agathobacter sp900543445	Guild 1
ERR2017 731.2	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Anaerotignaceae;g__Anaerotignum;s__Anaerotignum faecicola	Guild 1
ERR2017 735.12	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Faecalibacterium;s__Faecalibacterium prausnitzii G	Guild 1
ERR2017 736.11	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__ER4;s__ER4 sp900550165	Guild 1
ERR2017 739.2	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__Lawsonibacter;s__Lawsonibacter sp000177015	Guild 1
ERR2017 740.16	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Monoglobales_A;f__UBA1381;g__CAG-41;s__CAG-41 sp900066215	Guild 1
ERR2017 740.2	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Blautia A;s__Blautia A sp900541345	Guild 1
ERR2017 741.15	d__Bacteria;p__Actinobacteriota;c__Actinomycetia;o__Actinomycetales;f__Bifidobacteriaceae;g__Bifidobacterium;s__Bifidobacterium dentium	Guild 2
ERR2017 742.15	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Mediterraneibacter;s__Mediterraneibacter faecis	Guild 1
ERR2017 746.14	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Dorea A;s__Dorea A longicatena B	Guild 2
ERR2017 750.5	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__Lawsonibacter;s__Lawsonibacter sp000177015	Guild 1
ERR2017 771.1	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Agathobacter;s__Agathobacter rectalis	Guild 1
ERR2017 781.12	d__Bacteria;p__Firmicutes;c__Bacilli;o__Erysipelotrichales;f__Erysipelotrichaceae;g__Holdemanella;s__Holdemanella biformis	Guild 1
ERR2017 814.13	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Blautia;s__Blautia stercoris	Guild 2
ERR2726 410.2	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Gemmiger;s__Gemmiger formicilis	Guild 1
ERR2726 411.1	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella nigrescens	Guild 2

ID	Taxonomy Assignment (d=domain; p=phylum; c=class; o=order; f=family; g=genus; s = species)	Assignment
ERR2726 419.10	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ac utalibacteraceae;g__Acutalibacter;s__Acutalibacter sp000435395	Guild 1
ERR2726 420.13	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__La chnospiraceae;g__Butyrivibrio_A;s__Butyrivibrio A crossotus	Guild 2
ERR2726 422.1	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__La chnospiraceae;g__Roseburia;s__Roseburia intestinalis	Guild 2
ERR2726 423.2	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Os cillospiraceae;g__CAG-170;s__CAG-170 sp900548625	Guild 1
ERR2726 433.15	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Ba cteroidaceae;g__Phocaeicola;s__Phocaeicola sp000434735	Guild 2
ERR2726 447.12	d__Bacteria;p__Firmicutes;c__Bacilli;o__RFN20;f__CAG- 288;g__CAG-568;s__CAG-568 sp000434395	Guild 1
ERR2726 447.19	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Ba cteroidaceae;g__Prevotella;s__Prevotella sp002265625	Guild 2
ERR2726 447.2	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__La chnospiraceae;g__KLE1615;s__KLE1615 sp900066985	Guild 1
ERR2726 451.2	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Ta nnerellaceae;g__Parabacteroides;s__Parabacteroides merdae	Guild 2
ERR2726 452.15	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Rik enellaceae;g__Alistipes;s__Alistipes putredinis	Guild 2
ERR2726 452.9	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ru minococcaceae;g__Negativibacillus;s__Negativibacillus sp900547455	Guild 1
ERR2726 458.13	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__La chnospiraceae;g__Mediterraneibacter;s__Mediterraneibacter lactaris	Guild 2
ERR2726 476.5	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Ba cteroidaceae;g__Phocaeicola;s__Phocaeicola sp000432735	Guild 2
ERR2726 486.19	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Rik enellaceae;g__Alistipes;s__Alistipes communis	Guild 2
ERR2726 493.4	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Os cillospiraceae;g__Dysosmobacter;s__Dysosmobacter welbionis	Guild 1
ERR2726 493.5	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ac utalibacteraceae;g__UMGS403;s__UMGS403 sp900540275	Guild 1
ERR2726 498.24	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Mu ribaculaceae;g__Muribaculum;s__Muribaculum sp003150235	Guild 2
ERR2726 502.1	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__La chnospiraceae;g__UBA11774;s__UBA11774 sp003507655	Guild 1
ERR2726 522.5	d__Bacteria;p__Firmicutes_A;c__Clostridia_A;o__Christensenellales; f__CAG-74;g__SFFH01;s__SFFH01 sp900542395	Guild 1
ERR2726 527.4	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Ba cteroidaceae;g__Bacteroides;s__Bacteroides uniformis	Guild 2
ERR2726 532.18	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ac utalibacteraceae;g__Pseudoruminococcus;s__Pseudoruminococcus massiliensis	Guild 1

ID	Taxonomy Assignment (d=domain; p=phylum; c=class; o=order; f=family; g=genus; s = species)	Assignment
ERR2855 787.12	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__CAG-115;s__CAG-115 sp003531585	Guild 1
ERR2855 790.1	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__Dysosmobacter;s__Dysosmobacter sp001916835	Guild 1
ERR2855 795.12	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Faecalibacterium;s__Faecalibacterium sp900539945	Guild 1
ERR2855 800.13	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Burkholderiales;f__Burkholderiaceae;g__Duodenibacillus;s__Duodenibacillus sp003472385	Guild 1
ERR2855 808.13	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__CAG-81;s__CAG-81 sp900066535	Guild 2
ERR2855 814.14	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__TF01-11;s__TF01-11 sp001414325	Guild 1
ERR2855 814.4	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__UBA9502;s__UBA9502 sp003480315	Guild 2
ERR2855 821.2	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__CAG-81;s__CAG-81 sp900066535	Guild 2
ERR2855 825.1	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaecicola;s__Phocaecicola massiliensis	Guild 2
ERR2855 828.7	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__UBA9502;s__UBA9502 sp003478505	Guild 2
ERR2855 831.6	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Burkholderiales;f__Burkholderiaceae;g__Duodenibacillus;s__Duodenibacillus massiliensis	Guild 2
ERR2855 838.18	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__TF01-11;s__TF01-11 sp003524945	Guild 2
ERR2855 840.6	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Gemmiger;s__Gemmiger formicilis	Guild 2
ERR2855 844.18	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Clostridium Q;s__Clostridium Q sp003024715	Guild 2
ERR2855 857.7	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Burkholderiales;f__Burkholderiaceae;g__Mesosutterella;s__Mesosutterella multiformis	Guild 2
ERR2855 864.8	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella stercorea	Guild 2
ERR2855 871.6	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Lachnospira;s__Lachnospira sp900545725	Guild 2
ERR2855 880.6	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Faecalibacterium;s__Faecalibacterium prausnitzii G	Guild 2
ERR2855 886.18	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Acetivibacteraceae;g__Ruminococcus_E;s__Ruminococcus_E sp003526955	Guild 1

ID	Taxonomy Assignment (d=domain; p=phylum; c=class; o=order; f=family; g=genus; s = species)	Assignment
ERR2855 908.3	d__Bacteria;p__Firmicutes_C;c__Negativicutes;o__Veillonellales;f__Dialisteraceae;g__Dialister;s__Dialister sp900538805	Guild 2
ERR2855 923.15	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Ruminococcus_C;s__Ruminococcus_C sp000433635	Guild 1
ERR2855 933.19	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Eubacterium_G;s__Eubacterium_G sp000435815	Guild 2
ERR2855 935.6	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Acetivibacteraceae;g__CAG-177;s__CAG-177 sp003514385	Guild 2
ERR2855 935.9	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Rikenellaceae;g__Alistipes_A;s__Alistipes_A sp900539755	Guild 2
ERR2855 947.28	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__CAG-83;s__CAG-83 sp900552475	Guild 1
ERR2855 947.3	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__Dysosmobacter;s__Dysosmobacter sp001916835	Guild 1
ERR6885 06.3	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Faecalibacterium;s__Faecalibacterium prausnitzii E	Guild 1
ERR6885 07.14	d__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;s__Streptococcus thermophilus	Guild 2
ERR6885 13.3	d__Bacteria;p__Actinobacteriota;c__Coriobacteriia;o__Coriobacteriales;f__Coriobacteriaceae;g__Collinsella;s__Collinsella sp900541745	Guild 2
ERR6885 18.6	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Anaerostipes;s__Anaerostipes sp900066705	Guild 2
ERR6885 20.2	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Acetivibacteraceae;g__Eubacterium_R;s__Eubacterium_R sp000436835	Guild 2
ERR6885 20.8	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__CAG-317;s__CAG-317 sp900543415	Guild 2
ERR6885 21.4	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Ruminococcus_A;s__Ruminococcus_A sp003011855	Guild 1
ERR6885 22.8	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Faecalibacterium;s__Faecalibacterium prausnitzii D	Guild 1
ERR6885 22.9	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Dorea;s__Dorea formicigenerans	Guild 2
ERR6885 24.6	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Enterocloster;s__Enterocloster sp000155435	Guild 2
ERR6885 28.23	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Bacteroides_F;s__Bacteroides_F pectinophilus	Guild 2
ERR6885 29.17	d__Bacteria;p__Actinobacteriota;c__Coriobacteriia;o__Coriobacteriales;f__Coriobacteriaceae;g__Collinsella;s__Collinsella sp900758375	Guild 2

ID	Taxonomy Assignment (d=domain; p=phylum; c=class; o=order; f=family; g=genus; s = species)	Assignment
ERR6885 36.9	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Ruminococcus_A;s__Ruminococcus_A sp003011855	Guild 1
ERR6885 37.6	d__Bacteria;p__Actinobacteriota;c__Actinomycetia;o__Actinomycetales;f__Bifidobacteriaceae;g__Bifidobacterium;s__Bifidobacterium pullorum	Guild 2
ERR6885 39.8	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Blautia;s__Blautia stercoris	Guild 1
ERR6885 58.20	d__Bacteria;p__Actinobacteriota;c__Coriobacteriia;o__Coriobacteriales;f__Eggerthellaceae;g__CAG-1427;s__CAG-1427 sp000436075	Guild 2
ERR6885 58.6	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Gemmiger;s__Gemmiger quicibialis	Guild 1
ERR6885 59.18	d__Bacteria;p__Actinobacteriota;c__Coriobacteriia;o__Coriobacteriales;f__Eggerthellaceae;g__CAG-1427;s__CAG-1427 sp000431675	Guild 2
ERR6885 63.10	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Acetivibacteraceae;g__UBA1417;s__UBA1417 sp003531055	Guild 2
ERR6885 64.13	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Fusicatenibacter;s__Fusicatenibacter saccharivorans	Guild 1
ERR6885 67.8	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__NK3B98;s__NK3B98 sp900545815	Guild 1
ERR6885 77.19	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Escherichia;s__Escherichia dysenteriae	Guild 2
ERR6885 89.10	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__CAG-110;s__CAG-110 sp000434635	Guild 1
ERR6885 98.5	d__Bacteria;p__Actinobacteriota;c__Actinomycetia;o__Actinomycetales;f__Bifidobacteriaceae;g__Bifidobacterium;s__Bifidobacterium breve	Guild 1
ERR6886 04.10	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Eubacterium_G;s__Eubacterium_G sp000434315	Guild 1
ERR6886 08.9	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Anaerobutyricum;s__Anaerobutyricum soehngenii	Guild 1
ERR6886 10.11	d__Bacteria;p__Firmicutes_A;c__Clostridia_A;o__Christensenellales;f__CAG-138;g__PeH17;s__PeH17 sp900542285	Guild 1
ERR6886 10.9	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Roseburia;s__Roseburia hominis	Guild 2
ERR6886 12.12	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Fusicatenibacter;s__Fusicatenibacter saccharivorans	Guild 1
ERR6886 15.3	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Blautia_A;s__Blautia_A obeum	Guild 2
ERR6886 15.8	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__CAG-274;g__CAG-274;s__CAG-274 sp000432155	Guild 1

ID	Taxonomy Assignment (d=domain; p=phylum; c=class; o=order; f=family; g=genus; s = species)	Assignment
ERR6886 17.7	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Gemmiger;s__Gemmiger sp900540775	Guild 1
ERR6886 23.13	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Butyrivibrionaceae;g__Agathobaculum;s__Agathobaculum butyrivibrionaceae	Guild 1
ERR6886 27.12	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Blautia_A;s__Blautia_A wexlerae	Guild 2
ERR6886 29.1	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides uniformis	Guild 2
ERR6886 29.15	d__Bacteria;p__Firmicutes;c__Bacilli;o__Erysipelotrichales;f__Erysipelotrichaceae;g__Bulleidia;s__Bulleidia sp900539965	Guild 2
ERR6886 31.6	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides stercoris	Guild 2
ERR6886 40.13	d__Bacteria;p__Firmicutes;c__Bacilli;o__Erysipelotrichales;f__Erysipelotrichaceae;g__Clostridium_AQ;s__Clostridium_AQ sp900552125	Guild 2
ERR6886 42.6	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Gemmiger;s__Gemmiger quicibialis	Guild 1
ERR6886 45.6	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides fragilis	Guild 2
HD-20.2	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Faecalibacterium;s__Faecalibacterium sp900539945	Guild 2
HD-63.13	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__	Guild 2
HD-76.2	d__Bacteria;p__Fusobacteriia;c__Fusobacteriia;o__Fusobacteriales;f__Fusobacteriaceae;g__Fusobacterium_A;s__Fusobacterium_A sp900543175	Guild 2
HD-8.10	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Eubacterium_F;s__Eubacterium_F sp003491505	Guild 2
LD-26.6	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__TF01-11;s__TF01-11 sp000436755	Guild 1
LD-33.1	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides ndongoniae	Guild 2
LD-35.30	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Anaerotruncaceae;g__Anaerotruncum;s__	Guild 2
LD-39.9	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Tannerellaceae;g__Parabacteroides;s__Parabacteroides distasonis	Guild 2
LD-7.2	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Agathobacter;s__Agathobacter faecis	Guild 1
LD-75.3	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Tannerellaceae;g__Parabacteroides;s__Parabacteroides gordonii	Guild 2
MSM79 H8B.5	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola dorei	Guild 2

ID	Taxonomy Assignment (d=domain; p=phylum; c=class; o=order; f=family; g=genus; s = species)	Assignment
PSMA26 40.3	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Faecalibacterium;s__Faecalibacterium prausnitzii D	Guild 2
SRR3416 54.1	d__Bacteria;p__Verrucomicrobiota;c__Verrucomicrobiae;o__Verrucomicrobiales;f__Akkermansiaceae;g__Akkermansia;s__Akkermansia muciniphila	Guild 2
SRR3416 80.2	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Roseburia;s__Roseburia sp003470905	Guild 2
SRR3417 24.5	d__Bacteria;p__Actinobacteriota;c__Actinomycetia;o__Actinomycetales;f__Bifidobacteriaceae;g__Bifidobacterium;s__Bifidobacterium pseudocatenulatum	Guild 1
SRR4135 59.10	d__Bacteria;p__Firmicutes_A;c__Clostridia_A;o__Christensenellales;f__CAG-74;g__UBA11524;s__UBA11524 sp000437595	Guild 2
SRR4135 67.10	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Tannerellaceae;g__Parabacteroides;s__Parabacteroides goldsteinii	Guild 2
SRR4135 84.9	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__CAG-603;s__CAG-603 sp900066105	Guild 2
SRR4135 88.5	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900551275	Guild 2
SRR4136 35.13	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella buccae	Guild 1
SRR4136 35.5	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__CAG-45;s__CAG-45 sp000438375	Guild 1
SRR4136 41.2	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides ovatus	Guild 2
SRR4136 42.4	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Lachnospira;s__Lachnospira sp000437735	Guild 2
SRR4136 42.5	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Agathobacter;s__Agathobacter faecis	Guild 2
SRR4136 55.3	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Lachnospira;s__Lachnospira rogosae A	Guild 2
SRR4136 77.1	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Marinifilaceae;g__Odoribacter;s__Odoribacter splanchnicus	Guild 2
SRR4136 91.1	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900313215	Guild 2
SRR4136 94.1	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Lachnospira;s__Lachnospira eligens A	Guild 2
SRR4137 29.3	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides caccae	Guild 2
SRR4137 36.4	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Marinifilaceae;g__Butyrivibrio;s__Butyrivibrio virosa	Guild 2
SRR4137 53.22	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Rikenellaceae;g__Alistipes;s__Alistipes communis	Guild 2

ID	Taxonomy Assignment (d=domain; p=phylum; c=class; o=order; f=family; g=genus; s = species)	Assignment
SRR4137 55.1	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Enterocloster;s__Enterocloster aldenensis	Guild 1
SRR4137 61.12	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Anaerotignaceae;g__Anaerotignum;s__Anaerotignum faecicola	Guild 2
SRR4137 62.3	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__CAG-95;s__CAG-95 sp000438155	Guild 2
SRR4137 62.6	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Marinifilaceae;g__Odoribacter;s__Odoribacter massiliensis	Guild 2
SRR5279 219.7	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Acetivibacteraceae;g__Ruminococcus_E;s__Ruminococcus_E sp900314705	Guild 2
SRR5279 222.3	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Blautia_A;s__Blautia_A sp003480185	Guild 2
SRR5279 239.4	d__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Lactobacillaceae;g__Limosilactobacillus;s__Limosilactobacillus mucosae	Guild 2
SRR5279 253.7	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Burkholderiales;f__Burkholderiaceae;g__Sutterella;s__Sutterella	Guild 2
SRR5279 269.27	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Eubacterium_G;s__Eubacterium_G ventriosum	Guild 2
SRR5279 278.2	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotellamassilia;s__Prevotellamassilia sp900539625	Guild 2
SRR5279 285.2	d__Bacteria;p__Firmicutes;c__Bacilli;o__Erysipelotrichales;f__Erysipelotrichaceae;g__Faecalibacillus;s__Faecalibacillus intestinalis	Guild 2
SRR5279 303.4	d__Bacteria;p__Firmicutes_C;c__Negativicutes;o__Veillonellales;f__Veillonellaceae;g__Veillonella;s__Veillonella parvula_A	Guild 2
SRR6468 507.8	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Blautia_A;s__Blautia_A wexlerae	Guild 1
SRR6468 510.13	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__UBA932;g__RC9;s__RC9 sp000434935	Guild 2
SRR6468 523.9	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__CAG-170;s__CAG-170 sp000432135	Guild 1
SRR6468 543.7	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Acetivibacteraceae;g__Eubacterium_R;s__Eubacterium_R sp900540305	Guild 1
SRR6468 590.5	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Eubacterium_G;s__Eubacterium_G sp000432355	Guild 1
SRR6468 594.3	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__V9D3004;s__V9D3004 sp900760345	Guild 2
SRR6468 602.3	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Burkholderiales;f__Burkholderiaceae;g__Parasutterella;s__Parasutterella sp000980495	Guild 2
SRR6468 612.10	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Acetivibacteraceae;g__Acetivibacter;s__Acetivibacter timonensis	Guild 2

ID	Taxonomy Assignment (d=domain; p=phylum; c=class; o=order; f=family; g=genus; s = species)	Assignment
SRR6468618.9	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Sellimonas;s__Sellimonas intestinalis	Guild 2
SRR6468619.8	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Enterocloster;s__Enterocloster bolteae	Guild 2
SRR6468633.4	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Gemmiger;s__Gemmiger sp900539695	Guild 1
SRR6468637.9	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides uniformis	Guild 2
SRR6468639.4	d__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;s__Streptococcus lutetiensis	Guild 1
SRR6468651.2	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Faecalimonas;s__	Guild 2
SRR6468652.5	d__Bacteria;p__Firmicutes_C;c__Negativicutes;o__Veillonellales;f__Megasphaeraceae;g__Anaeroglobus;s__Anaeroglobus micronuciformis	Guild 2
SRR6468672.4	d__Bacteria;p__Firmicutes_C;c__Negativicutes;o__Acidaminococcales;f__Acidaminococcaceae;g__Acidaminococcus;s__Acidaminococcus intestini	Guild 2
SRR6468684.3	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Lachnospira;s__Lachnospira sp000436475	Guild 1
SRR6468684.5	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__KLE1615;s__KLE1615 sp900066985	Guild 1
SRR6468688.21	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Acetivibacteraceae;g__Eubacterium_R;s__Eubacterium_R sp000431535	Guild 1
SRR6468699.2	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__Gemmiger;s__Gemmiger formicilis	Guild 1

[00174] The bacterial species listed in Table 1, Table 2, and Figures 13A-13XX were identified by metagenomic sequencing of genomic DNA isolated from human fecal samples and determined to be part of two competing microbiota guilds relative to at least one biological characteristic, as described in the Examples. Briefly, genomic DNA was isolated from each fecal sample was sequenced by next generation sequencing and contigs for microorganism genome sequences were constructed de novo. Generally, the contigs identified for each microorganism are predicted to represent greater than 95% of the entire genome for the microorganism. Genomic constructs having less than 1% sequence divergence from each other were combined and defined to be from the same microorganism. Genomic contigs for each microorganism listed in Table 1, Table 2, and Figures 13A-13XX are provided in the sequence listing filed with the application. The taxonomic assignment of each microorganism is given in

Table 1, Table 2, or Figures 13A-13XX. Correspondence between the sequence identifier assigned to each contig and the microorganism to which it belongs is provided in FIG.12. For example, the contigs provided as SEQ ID NOS:1-68 correspond to the genomic sequence of microorganism 1U001.8 (as indicated in FIG.12A), which is a microorganism classified as domain Bacteria, phylum Proteobacteria, class Gammaproteobacteria, order Enterobacterales, family Enterobacteria, genus Escherichia, and species Escherichia coli and is in Guild 2 of the 141 core microorganisms identified in Table 1.

[00175] Accordingly, in some embodiments of the methods described herein, a genome identified in a metagenomic analysis is classified as corresponding to a microorganism listed in Table 1, Table 2, and/or Figures 13A-13XX if the identified genomic constructs have at least 97% sequence identity when compared to the contigs for the microorganism provided in the sequence listing, as denoted in FIG.12. In some embodiments, a genome identified in a metagenomic analysis is classified as corresponding to a microorganism listed in Table 1, Table 2, and/or Figures 13A-13XX if the identified genomic constructs have at least 98% sequence identity when compared to the contigs for the microorganism provided in the sequence listing, as denoted in FIG.12. In some embodiments, a genome identified in a metagenomic analysis is classified as corresponding to a microorganism listed in Table 1, Table 2, and/or Figures 13A-13XX if the identified genomic constructs have at least 99% sequence identity when compared to the contigs for the microorganism provided in the sequence listing, as denoted in FIG.12. In some embodiments, a genome identified in a metagenomic analysis is classified as corresponding to a microorganism listed in Table 1, Table 2, and/or Figures 13A-13XX if the identified genomic constructs have at least 99.5% sequence identity when compared to the contigs for the microorganism provided in the sequence listing, as denoted in FIG.12. In some embodiments, a genome identified in a metagenomic analysis is classified as corresponding to a microorganism listed in Table 1, Table 2, and/or Figures 13A-13XX if the identified genomic constructs have at least 97%, at least 97.5%, at least 98%, at least 98.5%, at least 99%, at least 99.1%, at least 99.2%, at least 99.3%, at least 99.4%, at least 99.5%, at least 99.6%, at least 99.7%, at least 99.8%, at least 99.9%, or more sequence identity when compared to the contigs for the microorganism provided in the sequence listing, as denoted in FIG.12.

[00176] Referring to block 214, in some embodiments, the plurality of gut microorganisms comprises at least 20 microorganisms selected from those microorganisms in Table 1, Table 2, or

Figure 13A-13XX having a connectivity of at least 2. In some embodiments, the set of identified gut microorganisms are selected from those microorganisms having a connectivity of at least 2, at least 3, at least 4, at least 5, at least 10, at least 15, at least 20, at least 25, or more.

[00177] Referring to block 216, in some embodiments, the biological sample from the gut of the respective subject is a fecal sample from the respective training subject. In some embodiments, said biological sample is a sample obtained from the small or large intestine, preferably colon or rectum, more preferably obtained in the form of a fecal sample or rectal swab or in the form of a biopsy specimen of gastrointestinal mucosa.

[00178] Referring to block 218, in some embodiments, the therapy is a biological therapy, an immunotherapy, a chemotherapy, a radiation therapy, a gene therapy, a hormone therapy, a photodynamic therapy, a targeted therapy, small molecules, antibodies, polynucleotide, natural compound, immune modulator, bone marrow therapy, stem cell therapy, surgery therapy, induction therapy, maintenance therapy, or a combination thereof.

[00179] Referring to block 220, in some embodiments, the disorder is selected from the group consisting of type-2 diabetes, hypertension, schizophrenia, atherosclerotic cardiovascular disease (ACVD), liver cirrhosis (LC), inflammatory bowel diseases (IBD), colorectal cancer (CRC), ankylosing spondylitis (AS), and Parkinson's disease (PD), inflammatory bowel disease (IBD), rheumatoid arthritis (RA), or advanced melanoma and B cell lymphoma. In some embodiments, the disorder is, e.g., hypertension (HT), schizophrenia (SCZ), Multiple Sclerosis (MS), Gaucher disease type II (GDII), COVID-19 (COV), Behcet's disease (BD), autism spectrum disorder (ASD), or pancreatic cancer (PC). In some embodiments, the disorder is cancer, Alzheimer diseases, a cardiovascular disease, an autoimmune disease, a mental health disease, an infectious disease, or a genetic disorder.

[00180] In some embodiments, the disorder is categorized by any indicator of a biological state, function, structure, process, response, or condition in a patient. Such indicators include any of the numerous variables (parameters) that are commonly measured in medicine to evaluate a patient for purposes such as diagnosis, prognosis, and/or treatment. Typically, indicators of interest herein are those whose values (which may be quantitative or qualitative) reflect, characterize, or are related to the function or structure of organs and organ systems and/or whose values reflect, characterize, or are related to the presence or severity of conditions. In some

embodiments, the disease is categorized by the progression or prognosis of a disease or disorder, e.g., different stages of cancer, type, frequency, or degree of severity of the conditions that can be objectively measured or experienced by a subject. In certain embodiments, the disorder may be acquired by a medical device, which may be used to analyze the status of a body part, wound, or lesion, or of a physical substrate such as a test strip, dipstick, filter, or other substrate that provides means for detecting or measuring the presence of a substance in a biological sample obtained from a subject. In some embodiments, it is contemplated to detect antibodies against pathogens (e.g., viruses, bacteria, fungi), abnormal tissues (e.g., tumor site), or biomarkers in a biological sample and/or to detect the presence in a biological sample from a patient for purposes such as diagnosing the presence of a disorder or a disease.

[00181] Referring to block 222, in some embodiments, the disorder is cancer.

[00182] Referring to block 224, in some embodiments, the methods include inputting, for each respective training subject in the plurality of training subjects, information about the respective training subject into a model comprising a plurality of parameters. The model applies the plurality of parameters to the information, e.g., through at least 10,000 computations, to obtain a corresponding output for the respective training subject from the model. The corresponding output comprises a prediction of the respective training subject's response to the therapy, and the information about the respective training subject comprises the corresponding genomic abundance value for each respective gut microorganism in the plurality of gut microorganisms, and the plurality of gut microorganisms are selected from Table 1, Table 2, or Figure 13A-13XX.

[00183] In some embodiments, the model is trained against datasets collected across a plurality of therapies to disorders and the model is trained to distinguish between a responsive state and a non-responsive state for the therapy. In some embodiments, the model comprises a learning statistical classifier system. In some embodiments, the learning statistical classifier system is random forest, classification and regression tree, boosted tree, or neural network. For example, as described in Example 3, a random forest classifier was trained against datasets from 11 different studies collectively looking at microbiomes in 4 different disorders. As shown in Figure 8C, the resulting model was powered to predict responder or non-responder to anti-cytokine or anti-integrin therapy, methotrexate treatment in new-onset Rheumatoid Arthritis,

immune checkpoint inhibitor (ICI) treatment on advanced melanoma, and CD19-CAR-T immunotherapy on B cell lymphoma.

[00184] Referring to block 226, in some embodiments, the prediction of the respective training subject's response is a class output of a respective response, in a plurality of possible responses, of the respective training subject. The method allows the setting of a single "cut-off" value permitting discrimination between responder or non-responder to a treatment. In some embodiments, the prediction of the respective training subject's response includes a prediction of an objective response rate of the human subject to the treatment or therapy, and wherein the prediction of the objective response rate includes an indication or classification of a complete response or an amount of a partial response to the treatment.

[00185] Referring to block 228, in some embodiments, the prediction of the respective training subject's response is a probability output for the respective training subject's response. As it is disclosed above, the method allows the setting of a single "cut-off" value permitting discrimination between responder or non-responder to a treatment. In some embodiments, the methods comprise utilizing the model to calculate a probability value for a subject; compare the probability value to a threshold value derived from a cohort of responders/non-responders to determine whether or not the probability value is above/below the threshold value; classify the subject as responder/non-responder if the probability value is above/below the threshold. In embodiments, the threshold value may be about a probability value of at least 50%, 55%, 50%, 65%, 70%, 75% or about 80% or more. In other embodiments, the probability value is a positive predictive value as measured by area under the curve (AUC) of receiver operating characteristic (ROC) curves. In certain embodiments, the probability value is calculated using a multivariate logistic regression model, a neural network model, a random forest model or a decision tree model.

[00186] Referring to block 230, in some embodiments, the model is a neural network algorithm, a support vector machine algorithm, a Naive Bayes algorithm, a nearest neighbor algorithm, a boosted trees algorithm, a Random Forest algorithm, a convolutional neural network algorithm, a decision tree algorithm, a regression algorithm, or a clustering algorithm.

[00187] Referring to block 232, in some embodiments, the plurality of parameters is at least 1000, at least 10,000, at least 15,000, at least 50,000, at least 100,000, at least 250,000, at least

500,000, or at least 1,000,000 parameters, at least 2,500,000 parameters, at least 5,000,000 parameters, at least 10,000,000 parameters, or more.

[00188] Referring to block 234, in some embodiments, the model applies the plurality of parameters to the information through at least 1000 computations, at least 5000 computations, at least 10,000 computations, at least 25,000 computations, at least 50,000 computations, at least 100,000 computations, at least 250,000 computations, at least 500,000 computations, at least 1,000,000 computations, at least 2,500,000 computations, at least 5,000,000 computations, at least 10,000,000 computations, or more to obtain a corresponding output for the respective training subject from the model.

[00189] Referring to block 236, in some embodiments, the methods include adjusting the plurality of parameters based on, for each respective training subject in the plurality of training subjects, one or more differences between (i) the corresponding output from the model and (ii) the corresponding indication of the respective training subject's response to the therapy.

[00190] In some embodiments where deep learning techniques utilize a neural network as described above, the training of the neural network to improve the accuracy of its prediction involves modifying one or more parameters, including, but not limited to, weights in the filters in convolutional layers as well as biases in network layers. In some embodiments, the weights and biases are further constrained with various forms of regularization such as L1, L2, weight decay, and dropout.

[00191] For instance, in some embodiments, the neural network or any of the models disclosed herein optionally, where training data is labeled (*e.g.*, with an indication of the state of the biological characteristic), have their parameters (*e.g.*, weights) tuned (adjusted to potentially minimize the error between the system's predicted indications and the training data's measured indications). Various methods used to minimize error function, such as gradient descent methods, include, but are not limited to, log-loss, sum of squares error, hinge-loss methods. In some embodiments, these methods further include second-order methods or approximations such as momentum, Hessian-free estimation, Nesterov's accelerated gradient, adagrad, *etc.* In some embodiments, the methods also combine unlabeled generative pretraining and labeled discriminative training.

[00192] Accordingly, in some embodiments, the training of the neural network comprises adjusting one or more parameters in the plurality of parameters by back-propagation through a loss function. In some embodiments, the loss function is a regression task and/or a classification task. Non-limiting examples of loss functions suitable for the regression task include, but are not limited to, a mean squared error loss function, a mean absolute error loss function, a Huber loss function, a Log-Cosh loss function, or a quantile loss function. *See*, Wang *et al.*, 2020, “A Comprehensive Survey of Loss Functions in Machine Learning,” *Annals of Data Science*, doi.org/10.1007/s40745-020-00253-5, last accessed September 15, 2021, which is hereby incorporated by reference in its entirety. Non-limiting examples of loss functions suitable for the classification task include, but are not limited to, a binary cross entropy loss function, a hinge loss function, or a squared hinged loss function. In some embodiments, the loss function is any suitable regression task loss function or classification task loss function.

[00193] Other suitable methods for training the neural network that are contemplated for use in the present disclosure are further described herein (*see, e.g.*, Definitions: Untrained model, above).

[00194] In some embodiments, the parameters of the neural network are randomly initialized prior to training.

[00195] In some embodiments, the neural network comprises a dropout regularization parameter. For example, in some embodiments, a regularization is performed by adding a penalty to the loss function, where the penalty is proportional to the values of the parameters in the trained or untrained model. Generally, regularization reduces the complexity of the model by adding a penalty to one or more parameters to decrease the importance of the respective hidden neurons associated with those parameters. Such practice can result in a more generalized model and reduce overfitting of the data. In some embodiments, the regularization includes an L1 or L2 penalty.

[00196] In some embodiments, the training the neural network comprises an optimizer. In some embodiments, the optimizer may employ the loss function to update the parameters of the neural network or other model via back-propagation. In some embodiments, the training the neural network comprises a learning rate.

[00197] In some embodiments, the learning rate is at least 0.0001, at least 0.0005, at least 0.001, at least 0.005, at least 0.01, at least 0.05, at least 0.1, at least 0.2, at least 0.3, at least 0.4, at least 0.5, at least 0.6, at least 0.7, at least 0.8, at least 0.9, or at least 1. In some embodiments, the learning rate is no more than 1, no more than 0.9, no more than 0.8, no more than 0.7, no more than 0.6, no more than 0.5, no more than 0.4, no more than 0.3, no more than 0.2, no more than 0.1 no more than 0.05, no more than 0.01, or less. In some embodiments, the learning rate is from 0.0001 to 0.01, from 0.001 to 0.5, from 0.001 to 0.01, from 0.005 to 0.8, or from 0.005 to 1. In some embodiments, the learning rate falls within another range starting no lower than 0.0001 and ending no higher than 1.

[00198] In some embodiments, the learning rate further comprises a learning rate decay (*e.g.*, a reduction in the learning rate over one or more epochs). For example, a learning decay rate can be a reduction in the learning rate of 0.5 or 0.1. In some embodiments, the learning rate is a differential learning rate. In some embodiments, the training the neural network further uses a scheduler that conditionally applies the learning rate decay based on an evaluation of a performance metric over a threshold number of training epochs (*e.g.*, the learning rate decay is applied when the performance metric fails to satisfy a threshold performance value for at least a threshold number of training epochs).

[00199] In some embodiments, the performance of the neural network is measured at one or more time points using a performance metric, including, but not limited to, a training loss metric, a validation loss metric, and/or a mean absolute error. In some embodiments, the performance metric is an area under receiving operating characteristic (AUROC) and/or an area under precision-recall curve (AUPRC).

[00200] For instance, in some embodiments, the performance of the neural network is measured by validating the model using a validation (*e.g.*, development) dataset. In some such embodiments, the training the neural network forms a trained neural network when the neural network satisfies a minimum performance requirement based on a validation.

[00201] In some embodiments, any suitable method for validation can be used, including but not limited to K-fold cross-validation, advanced cross-validation, random cross-validation, grouped cross-validation (*e.g.*, K-fold grouped cross-validation), bootstrap bias corrected cross-validation, random search, and/or Bayesian hyperparameter optimization.

[00202] In some embodiments, a method is provided for training a model comprising a plurality of parameters by a procedure comprising (i) inputting corresponding genomic abundance value for each respective gut microorganism in a plurality of gut microorganisms for each respective training subject in a plurality of training subjects, thereby obtaining as output from the model, for each respective training subject in the plurality of training subjects, a corresponding prediction of a training subject's response to a therapy, and (ii) refining the plurality of model parameters based on a differential between the corresponding actual response to a therapy of the training subject and the corresponding predicted response to a therapy of the training subject.

[00203] 2. Methods of applying a model for predicting a subject's response to a therapy for a disorder

[00204] Figure 3 is a schematic diagram of a method for applying a model for predicting a subject's response to a therapy for a disorder as discussed below. The method 300 may be implemented using a computer system (*e.g.*, the computer system 100 shown and described above in reference to Figure 1).

[00205] Referring to block 300, in some embodiments, the methods include obtaining, in electronic form, a plurality of genomic abundance values comprising, for each respective gut microorganism in a plurality of gut microorganisms selected from Table 1, Table 2, or Figure 13A-13XX, a corresponding abundance value for the genome of the respective species of gut bacteria, in the plurality of gut microorganisms, in a biological sample from the subject.

[00206] In some embodiment, a corresponding biological sample from the gut of the respective subject was taken prior to a treatment or a therapy. In some embodiments, the biological sample is taken no more than 15 minutes, 30 minutes, 1 hour, 2 hours, 3 hours, 4 hours, 5 hours, 6 hours, 12 hours, or 24 hours prior to a treatment or a therapy. In some embodiments, the biological sample is taken 1 day, 2, days, 3 days, 4 days, 5 days, 6 days, 1 week, 2 weeks, 3 weeks, 4 weeks, or more prior to a treatment or a therapy. . In some embodiments, the biological sample is taken about any of 1 month, 2 months, 3 months, 4 months, 5 months, 6 months, or more prior to a treatment or a therapy.

[00207] In some embodiments, sample data (including plasma, stool specimens) and corresponding clinical information (including gender/age/body fat count/underlying

disease/histopathological characteristics, etc.) were collected for each subject prior to receiving a therapy. Individual biological samples were subjected to full microbiome analysis. In some embodiments, the sample is a tissue biopsy, an intestinal, or mucosal sample. See, for example, Tang Q, Jet al., Current Sampling Methods for Gut Microbiota: A Call for More Precise Devices, Front Cell Infect Microbiol., 10:151 (2020), the content of which is incorporated herein by reference in its entirety. In some embodiment, the biological sample from the gut of the respective subject is a fecal sample from the respective subject.

[00208] In some embodiments, the plurality of gut microorganisms comprises at least 25 gut microorganisms selected from Table 1, Table 2, or Figures 13A-13XX. In some embodiments, the plurality of gut microorganisms comprises at least 30 gut microorganisms selected from Table 1, Table 2, or Figures 13A-13XX. In some embodiments, the plurality of gut microorganisms comprises at least 40 gut microorganisms selected from Table 1, Table 2, or Figures 13A-13XX. In some embodiments, the plurality of gut microorganisms comprises at least 25 gut microorganisms selected from Table 1, Table 2, or Figures 13A-13XX. In some embodiments, the plurality of gut microorganisms comprises at least 15, at least 20, at least 25, at least 30, at least 35, at least 40, at least 45, at least 50, at least 55, at least 60, at least 65, at least 70, at least 75, at least 80, at least 85, at least 90, at least 95, at least 100, at least 125, at least 150, at least 200, at least 250, at least 300, at least 400, at least 500, at least 600, at least 700, at least 800, or all of the gut microorganisms selected from Table 1, Table 2 or Figures 13A-13XX. In some embodiments, the plurality of gut microorganisms are at least 20, at least 25, at least 30, at least 35, at least 40, at least 45, at least 50, at least 55, at least 60, at least 65, at least 70, at least 75, at least 80, at least 85, at least 90, at least 95, at least 100, at least 125, at least 150, at least 200, at least 250, at least 300, at least 400, at least 500, at least 600, at least 700, at least 800, or all of the gut microorganisms selected from Table 1, Table 2 or Figures 13A-13XX. In some embodiments, the plurality of gut microorganisms are all of the gut microorganisms listed in Table 1. In some embodiments, the plurality of gut microorganisms are all of the gut microorganisms listed in Table 2. In some embodiments, the plurality of gut microorganisms are all of the gut microorganisms listed in Figures 13A-13XX.

[00209] In some of the embodiments, the corresponding value for the abundance of the genome is a value representative of the absolute abundance of a microorganism genome. In some of the embodiments, the corresponding value for the abundance of the genome is a value

representative of a normalized abundance value, or a relative abundance value (e.g., an abundance of one microorganism normalized against the abundance of total microbiome of interest). In some of the embodiments, corresponding value for the abundance of the genome is a value representative of an averaged abundance value (e.g., average of abundances obtained at different time points or from different biological samples from the patients, or average of abundances obtained using different probes, etc.), or a combination of any of above. The corresponding value for the abundance of the genome is measured by any technique known in the art. In some embodiments, the genomic abundance value for the genome is measured by quantitative PCR(qPCR), such as bacterial 16S rRNA qPCR, RT-PCR, or qRT-PCR, for quantifying the abundance of region of interests in the genome, e.g., as described in U.S. Patent No. 11,427,865, the disclosure of which is hereby incorporated by reference in its entirety. In some embodiments, the genomic abundance value is measured by targeted sequencing (e.g., 16S rRNA sequencing, or any other suitable biomarker), partial genome sequencing or whole genome sequencing, thereby quantifying the number of reads of the targeted regions in a microorganism genome to determine the abundance of the genome, e.g., as disclosed in U.S. Patent Application Publication No. 2021/0403986 or U.S. Patent No. 11,332,783, the disclosures of which are hereby incorporated by reference in their entireties. In some embodiments, deep sequencing is employed to determine the abundance of targeted sequences, e.g., as disclosed in U.S. Patent Application Publication No. 2018/0237863, the disclosure of which is incorporated herein by reference in its entirety. In some embodiments, the sequencing depth is at least 2X, at least 3X, at least 4X, at least 5X, at least 6X, at least 7X, at least 8X, at least 9X, at least 10X, at least 11X, at least 12X, at least 13X, at least 14X, at least 15X, at least 16X, at least 17X, at least 18X, at least 19X, at least 20X, at least 21X, at least 22X, at least 23X, at least 24X, at least 25X, at least 26X, at least 27X, at least 28X, at least 29X, at least 30X, at least 31X, at least 32X, at least 33X, at least 34X, at least 35X, at least 36X, at least 37X, at least 38X, at least 39X, at least 40X, at least 41X, at least 42X, at least 43X, at least 44X, at least 45X, at least 46X, at least 47X, at least 48X, at least 49X, at least 50X, at least 51X, at least 52X, at least 53X, at least 54X, at least 55 X, at least 56X, at least 57X, at least 58X, at least 59X, at least 60X, at least 70X, at least 80X, at least 90X, at least 100X, at least 110X, at least 120X, at least 130X, at least 150X, at least 200X, at least 300X, at least 400X, at least 500X, at least 750X, at least 1000X, or more. In some embodiments, shotgun metagenomic sequencing is employed to provide sequence reads for

genomes in a sample, e.g., as described in U.S. Patent No. 11,028,449, the content of which is incorporated herein by reference in its entirety.

[00210] In some embodiments of the methods described herein, a genome identified in a metagenomic analysis is classified as corresponding to a microorganism listed in Table 1, Table 2, and/or Figures 13A-13XX if the identified genomic constructs have at least 97% sequence identity when compared to the contigs for the microorganism provided in the sequence listing, as denoted in FIG. 12. In some embodiments, a genome identified in a metagenomic analysis is classified as corresponding to a microorganism listed in Table 1, Table 2, and/or Figures 13A-13XX if the identified genomic constructs have at least 98% sequence identity when compared to the contigs for the microorganism provided in the sequence listing, as denoted in FIG. 12. In some embodiments, a genome identified in a metagenomic analysis is classified as corresponding to a microorganism listed in Table 1, Table 2, and/or Figures 13A-13XX if the identified genomic constructs have at least 99% sequence identity when compared to the contigs for the microorganism provided in the sequence listing, as denoted in FIG. 12. In some embodiments, a genome identified in a metagenomic analysis is classified as corresponding to a microorganism listed in Table 1, Table 2, and/or Figures AXX if the identified genomic constructs have at least 99.5% sequence identity when compared to the contigs for the microorganism provided in the sequence listing, as denoted in FIG. 12. In some embodiments, a genome identified in a metagenomic analysis is classified as corresponding to a microorganism listed in Table 1, Table 2, and/or Figures 13A-13XX if the identified genomic constructs have at least 97%, at least 97.5%, at least 98%, at least 98.5%, at least 99%, at least 99.1%, at least 99.2%, at least 99.3%, at least 99.4%, at least 99.5%, at least 99.6%, at least 99.7%, at least 99.8%, at least 99.9%, or more sequence identity when compared to the contigs for the microorganism provided in the sequence listing, as denoted in FIG. 12.

[00211] Referring to block 302, in some embodiments, the methods include sequencing genomic DNA from the biological sample from the gut of the subject, thereby obtaining the plurality of nucleic acid sequences. In some embodiments, the plurality of nucleic acid sequences comprises at least 10,000, at least 100,000, at least 250,000, at least 500,000, at least 1,000,000, at least 2,500,000, at least 5,000,000, at least 10,000,000 or at least 50,000,000 nucleic acid sequences. In some embodiments, the plurality of nucleic acid sequences comprises no more than 250,000,000, no more than 100,000,000, no more than 50,000,000, no more than

25,000,000, no more than 10,000,000, no more than 5,000,000, no more than 1,000,000, no more than 100,000 nucleic acid sequences. In some embodiments, the plurality of nucleic acid sequences consists of from 100,000 to 1,000,000, from 200,000 to 5,000,000, from 500,000 from 10,000,000, from 1,000,000 to 20,000,000, from 5,000,000 to 50,000,000, from 10,000,000 to 100,000,000, or from 50,000,000 to 250,000,000 nucleic acid sequences. In some embodiments, the plurality of nucleic acid sequences falls within another range starting no lower than 1000 nucleic acid sequences and ending no higher than 250,000,000 nucleic acid sequences.

[00212] In some embodiments, the corresponding plurality of nucleic acid sequences are obtained through metagenomic sequencing, e.g., as disclosed in U.S. Patent Application Publication No. 2016/0239602 or U.S. Patent No. 11,495,326, the contents of which are incorporated herein by reference in their entireties. In some embodiments, metagenomes sequencing further comprise generating the plurality of metagenomic fragment reads. In some embodiments, metagenomic sequencing further comprise fragmenting microbial genomes into random fragments of targeted sizes. The resulting fragments can vary in size. In one embodiment, fragments of approximately 500 nucleotides can be obtained. In some embodiments, fragments of from 100-2000 nucleotides, e.g., 200-800, 100-900, 100-1000, 300-800, 400-900 nucleotides can be obtained. In some embodiments, the method may further comprise extracting the metagenomic fragments from the corresponding biological sample. In some embodiments, metagenomes sequencing further comprise sequencing the fragments using high throughput sequencing methods to generate a plurality of sequencing reads.

[00213] In some embodiments, the corresponding plurality of nucleic acid sequences are obtained through targeted panel sequencing. An example of targeted panel sequencing is described in U.S. Patent Application Publication No. 2019/0316209. In some embodiments, the targeted panel sequencing comprises hybridizing genomic DNA isolated from a biological sample from the gut of a subject with a panel of probes that include one or more probes that hybridize to a unique sequence in the genome of each microorganism being quantified, prior to sequencing recovered nucleic acids. In some embodiments, the microorganisms include a plurality of the microorganisms listed in Table 1, Table 2, and/or Figures 13A-13XX. In some embodiments, a combination of semi-unique sequences (e.g., sequences found in a small number of the microorganism genomes) can be used to deconvolute genomic abundance values using an algorithm, e.g., a system of equations. In some embodiments, the panel of probes includes at

least 1 probe that hybridizes to a sequence unique to each microorganism genome being detected. In some embodiments, the panel of probes includes at least 2, at least 3, at least 4, at least 5, at least 10, at least 25, at least 50, or more probes that hybridize to a different sequence unique to each microorganism genome being detected. In some embodiments, the panel of probes includes at least 20, at least 30, at least 40, at least 50, at least 75, at least 100, at least 125, at least 150, at least 200, at least 300, at least 400, at least 500, at least 750, at least 1000, at least 1250, at least 1500, at least 2000, at least 2500, at least 3000, at least 4000, at least 5000, at least 7500, at least 10,000 or more unique probes.

[00214] In some embodiments, the sequencing genomic DNA from the corresponding biological sample comprise a partial or complete sequencing platform adapter sequence at their termini useful for sequencing using a sequencing platform of interest. Sequencing platforms of interest include, but are not limited to, the HiSeq™, MiSeq™ and Genome Analyzer™ sequencing systems from Illumina®; the Ion PGM™ and Ion Proton™ sequencing systems from Ion Torrent™; the PACBIO RS II Sequel system from Pacific Biosciences, the SOLiD sequencing systems from Life Technologies™, the 454 GS FLX+ and GS Junior sequencing systems from Roche, the MinION™ system from Oxford Nanopore, or any other sequencing platform of interest.

[00215] Referring to block 304, in some embodiments, the methods include obtaining, in electronic form, a plurality of nucleic acid sequences for genomic DNA from the biological sample from the gut of the subject.

[00216] Referring to block 306, in some embodiments, the methods include determining, for each respective gut microorganism in the plurality of gut microorganisms, the corresponding value for the abundance of the genome of the respective gut microorganism from the plurality of nucleic acid sequences. In some embodiments, the genomic abundance values determined for the subject comprise at least 20, at least 25, at least 50, at least 75, at least 100, at least 125, at least 150, at least 175, at least 200, at least 250, at least 300, at least 350, at least 400, at least 500, at least 600, at least 700, at least 800, at least 900, at least 1000, at least 1500, at least 2000, at least 25000, at least 5,000 or at least 10,000 genome abundance values, where each genome abundance value corresponds to different gut microorganism. In some embodiments, the genomic abundance values comprise no more than 250,000, no more than 100,000, no more than

50,000, no more than 25,000, no more than 10,000, no more than 5000, no more than 2500, no more than 1000, no more than 750, no more than 500, or fewer genome abundance values. In some embodiments, the genomic abundance values consist of from 10 to 40, from 20 to 50, from 30 to 80, from 40 to 100, from 50 to 150, from 60 to 200, from 80 to 300, from 90 to 500, from 100 to 1000, from 500 to 2,000, or from 1,000 to 5,000 genome abundance values. In some embodiments, the number of genomic abundance values fall within another range starting no lower than 10 genome abundance values and ending no higher than 250,000 genome abundance values.

[00217] Referring to block 308, in some embodiments, the methods include assembling, in electronic form, a corresponding plurality of gut microorganism genomes by metagenomic de novo sequence assembly from the plurality of nucleic acid sequences, and calculating, for each respective gut microorganism in the plurality of gut microorganisms, the corresponding value for the abundance of the genome of the respective gut microorganism based on the prevalence of respective nucleic acid sequences, in the plurality of nucleic acid sequences, used to assemble a respective gut microorganism genome in the plurality of gut microorganism genomes corresponding to the respective gut microorganism. In some embodiments, metagenomic de novo sequence assembly further comprise generating contigs based on the sequencing reads generated by a shotgun sequencing technique. Such a technique is described, for example, in U.S. Patent No. 10,529,443, the content of which is incorporated herein by reference in its entirety. In some embodiments, the plurality of nucleic acid sequences can be assembled into full genomes of the plurality of gut microorganisms. In some embodiments, the plurality of nucleic acid sequences can be assembled into partial genomes of the plurality of gut microorganisms.

[00218] Referring to block 310, in some embodiments, the methods include assigning each respective nucleic acid sequence in the plurality of nucleic acid sequences to a respective gut microorganism in the plurality of gut microorganisms, thereby generating, for each respective gut microorganism in the plurality of gut microorganism, a corresponding count of respective nucleic acid sequences in the plurality of nucleic acid sequences assigned to the respective gut microorganism, and determining, for each respective gut microorganism in the plurality of gut microorganisms, the corresponding genomic abundance value for the respective gut microorganism based on the corresponding count of respective nucleic acid sequences assigned to the respective gut microorganism. In some embodiments, the assigning each respective nucleic

acids to a respective gut microorganism includes mapping the nucleic acid to a reference nucleic acid. In some embodiments, the assigning each respective nucleic acids a respective gut microorganism includes annotating genome information based on existing databases. In some embodiments, nucleic acid sequences are analyzed, and annotations are to define taxonomic assignments using sequence similarity and phylogenetic placement methods or a combination of the two strategies.

[00219] Sequence similarity based methods for assigning each respective nucleic acid sequence in a respective gut microorganism include those familiar to individuals skilled in the art including, but not limited to BLAST, BLASTx, tBLASTn, tBLASTx, RDP-classifier, DNAClust, and various implementations of these algorithms such as Qiime or Mothur. These methods rely on mapping a sequence read to a reference database and selecting the match with the best score and e-value. In some embodiments, phylogenetic methods are used in combination with sequence similarity methods to improve the calling accuracy of an annotation or taxonomic assignment. Common databases include, but not limited to, GT-DBTK, National Center for Biotechnology Information (NCBI) Genbank, European Bioinformatics Institute-European Nucleotide Archive (European Bioinformatics Institute-European Nucleotide Archive; EBI-ENA), National Institute of Genetics, U.S. Department of ENERGY (USDOE) Integrated Microbial Genomes (Integrated Microbial Genomes) & Microbiomes; IMG/M) and other available databases in the art.

[00220] Referring to block 312, in some embodiments, the plurality of gut microorganisms comprises at least 20 microorganisms selected from those microorganisms in Table 1, Table 2, or Figure 13A-13XX having a connectivity of at least 2, at least 3, at least 4, at least 5, at least 10, at least 15, at least 20, at least 25, or more. In some embodiments, the plurality of gut microorganisms comprises at least 20 microorganisms selected from those microorganisms listed in Table 1, Table 2, or Figure 13A- 13XX as having a connectivity of at least 2. In some embodiments, the plurality of gut microorganisms comprises at least 20 microorganisms selected from those microorganisms listed in Table 1, Table 2, or Figure 13A- 13XX as having a connectivity of at least 3, at least 4, at least 5, at least 10, at least 15, at least 20, at least 25, or more.

[00221] Referring to block 314, in some embodiments, the biological sample from the gut of the subject is a fecal sample. In some embodiments, the sample is a tissue biopsy, an intestinal, or mucosal sample. In some embodiments, said biological sample is a sample obtained from the small or large intestine, preferably colon or rectum, more preferably obtained in the form of a fecal sample or rectal swab or in the form of a biopsy specimen of gastrointestinal mucosa.

[00222] Referring to block 316, in some embodiments, the therapy is a biological therapy, an immunotherapy, a chemotherapy, a radiation therapy, a gene therapy, a hormone therapy, a photodynamic therapy, a targeted therapy, small molecules, antibodies, polynucleotide, natural compound, immune modulator, bone marrow therapy, stem cell therapy, surgery therapy, induction therapy, maintenance therapy, or a combination thereof.

[00223] Referring to block 318, in some embodiments, the disorder is selected from the group consisting of type-2 diabetes, hypertension, schizophrenia, atherosclerotic cardiovascular disease (ACVD), liver cirrhosis (LC), inflammatory bowel diseases (IBD), colorectal cancer (CRC), ankylosing spondylitis (AS), and Parkinson's disease (PD), inflammatory bowel disease (IBD), rheumatoid arthritis (RA), advanced melanoma and B cell lymphoma. In some embodiments, the disorder is, e.g., hypertension (HT), schizophrenia (SCZ), Multiple Sclerosis (MS), Gaucher disease type II (GDII), COVID-19 (COV), Behcet's disease (BD), autism spectrum disorder (ASD), or pancreatic cancer (PC). In some embodiments, the disorder is cancer, Alzheimer diseases, a cardiovascular disease, an autoimmune disease, a mental health disease, an infectious disease, or a genetic disorder.

[00224] In some embodiments, the disorder is categorized by any indicator of a biological state, function, structure, process, response, or condition in a patient. Such indicators include any of the numerous variables (parameters) that are commonly measured in medicine to evaluate a patient for purposes such as diagnosis, prognosis, and/or treatment. Typically, indicators of interest herein are those whose values (which may be quantitative or qualitative) reflect, characterize, or are related to the function or structure of organs and organ systems and/or whose values reflect, characterize, or are related to the presence or severity of conditions. In some embodiments, the disease is categorized by the progression or prognosis of a disease or disorder, e.g., different stages of cancer, type, frequency, or degree of severity of the conditions that can be objectively measured or experienced by a subject. In certain embodiments, the disorder may

be acquired by a medical device, which may be used to analyze the status of a body part, wound, or lesion, or of a physical substrate such as a test strip, dipstick, filter, or other substrate that provides means for detecting or measuring the presence of a substance in a biological sample obtained from a subject. In some embodiments, it is contemplated to detect antibodies against pathogens (e.g., viruses, bacteria, fungi), abnormal tissues (e.g., tumor site), or biomarkers in a biological sample and/or to detect the presence in a biological sample from a patient for purposes such as diagnosing the presence of a disorder or a disease.

[00225] Referring to block 320, in some embodiments, the disorder is cancer.

[00226] In some embodiments, the disorder is inflammatory bowel disease and the therapy comprises anti-cytokine or anti-integrin treatment. For example, as reported in Example 5, for each of three data sets with microbiome metagenomic data collected before and after treatment with anti-cytokine or anti-integrin therapy, a random forest classifier was built based on the abundance of the 284 HQMAGs in CC-TCG to predict therapeutic responses. The models predicted therapeutic response with ROC AUC values of from 0.64 to 0.69.

[00227] The method according to any one of claims 19-29, wherein the disorder is rheumatoid arthritis and the therapy comprises methotrexate treatment. For example, as reported in Example 5, for a data set with microbiome metagenomic data collected before and after treatment with methotrexate, a random forest classifier was built based on the abundance of the 284 HQMAGs in CC-TCG to predict therapeutic responses. The model predicted therapeutic response with an ROC AUC value of 0.69.

[00228] The method according to any one of claims 19-29, wherein the disorder is B cell lymphoma and the therapy comprises CAR-T cell immunotherapy. In some embodiments, the CAR-T cell immunotherapy is CD19-CAR-T immunotherapy. For example, as reported in Example 5, for each of five data sets with microbiome metagenomic data collected before and after treatment with CD19-CAR-T immunotherapy, a random forest classifier was built based on the abundance of the 284 HQMAGs in CC-TCG to predict therapeutic responses. The models predicted therapeutic response with ROC AUC values of from 0.64 to 0.69.

[00229] The method according to any one of claims 19-29, wherein the disorder is melanoma and the therapy comprises immune checkpoint inhibitor treatment. For example, as reported in Example 5, one of two data sets with microbiome metagenomic data collected before and after

treatment with immune checkpoint inhibitor (ICI) treatment, a random forest classifier was built based on the abundance of the 284 HQMAGs in CC-TCG to predict therapeutic responses. The model predicted therapeutic outcomes on the training data set with an AUC of 0.66 and on the other data set, not used in training, with an AUC of 0.64.

[00230] Referring to block 322, in some embodiments, the methods include inputting the plurality of genomic abundance values into a model comprising a plurality of parameters. The model applies the plurality of parameters to the plurality of genomic abundance values through, e.g., at least 10,000 computations, to generate as output from the model a prediction of the subject's response to the therapy.

[00231] In some embodiments, the model is trained against datasets collected across a plurality of therapies to disorders and the model is trained to distinguish between a responsive state and a non-responsive state for the therapy. In some embodiments, the model comprises a learning statistical classifier system. In some embodiments, the learning statistical classifier system is random forest classification and regression tree, boosted tree, neural network. For example, as described in Example 3, a random forest classifier was trained against datasets from 11 different studies collectively looking at microbiomes in 4 different disorders. As shown in Figure 8C, the resulting model was powered to predict responder or non-responder to anti-cytokine or anti-integrin therapy, methotrexate treatment in new-onset Rheumatoid Arthritis, immune checkpoint inhibitor (ICI) treatment on advanced melanoma, and CD19-CAR-T immunotherapy on B cell lymphoma.

[00232] In some embodiments, the indication of subject's response is characterized by clinical outcome measures include, but are not limited to, complete remission, partial remission, non-remission, survival, development of adverse events, or any combination thereof. In some embodiments, one responder has complete remission in response to the treatment, and the non-responders has non-remission or partial remission in response to the treatment. In some embodiment, patients were subjected to routine clinical examinations, laboratory analyses, and computed tomography. Tumor responses were evaluated using RECIST criteria. In some embodiment, complete response was defined as complete radiographic disappearance of measurable or evaluable disease or stable, minimal radiographic findings; partial response was defined as reduction of the longest dimension of measurable disease by at least 50%; stable

disease was defined as reduction of the longest dimension by less than 25%; Progressive disease was defined as growth of the tumor by more than 25% in the longest dimension or development of new lesions. In some embodiment, overall response rate was defined as the sum of the complete and partial response rates and the tumor control rate was defined as the sum of overall response rates with stable disease rates.

[00233] In some embodiments, the indication of subject's response is characterized by the actual treatment efficacy of an therapy, including progression-free survival (PFS), the duration of the progression free survival under treatment, total Survival (OS), response to therapy (RT), overall response rate (ORR), sustained clinical effect (DCB), Disease Activity Score, or any combination thereof, or any other methods for evaluating the progression or prognosis of a disease or disorder known in the art.

[00234] In some embodiments, "progression free survival" (PFS) has its art-understood meaning relating to the length of time during and after the treatment of a disease, such as cancer, that a patient lives with the disease but it does not get worse. In some embodiments, measuring the progression-free survival is utilized as an assessment of how well a new treatment works. In some embodiments, PFS is determined in a randomized clinical trial; in some such embodiments, PFS refers to time from randomization until objective tumor progression and/or death.

[00235] In some embodiments, ORR may be defined as the proportion of patients in whom partial (PR) or complete (CR) responses are identified as a best overall response (BOR) according to some metric, such as Response Evaluation Criteria in Solid Tumors (RECIST 1.1). Stable disease (SD) was categorized as non-response together with progressive disease (PD). In some embodiments, ORR has its art-understood meaning referring to the proportion of patients with tumor size reduction of a predefined amount and for a minimum time period. In some embodiments, response duration usually measured from the time of initial response until documented tumor progression. In some embodiments, ORR involves the sum of partial responses plus complete responses.

[00236] In some embodiments, "clinical effect" refers to a clinical benefit. In some embodiments, such a clinical benefit is or comprises reduction in tumor size, increase in progression free survival, increase in overall survival, decrease in overall tumor burden, decrease in the symptoms caused by tumor growth such as pain, organ failure, bleeding, damage to the

skeletal system, and other related sequelae of metastatic cancer and combinations thereof. In some embodiments, the clinical effect is a “sustained clinical effect” (DCB) that is maintained for a relevant period of time. In some embodiments, the relevant period of time is at least 1 month, 2 months, 3 months, 4 months, 5 months, 6 months, 7 months, 8 months, 9 months, 10 months, 11 months, 1 year, 2 years, 3 years, 4 years, 5 years, or longer.

[00237] In some embodiments, the subject’s response is measured by Disease Activity Score (DAS) (see, e.g., Van der Heijde D. M. et al., *J Rheumatol*, 1993, 20(3): 579-81; Prevoo M. L. et al, *Arthritis Rheum*, 1995, 38: 44-8). The DAS system represents both current state of disease activity and change. The DAS scoring system uses a weighted mathematical formula, derived from clinical trials in RA. For example, the DAS 28 is $0.56(T28)+0.28(SW28)+0.70(\ln ESR)+0.014 GH$ wherein T represents tender joint number, SW is swollen joint number, ESR is erythrocyte sedimentation rate, and GH is global health. Various values of the DAS represent high or low disease activity as well as remission, and the change and endpoint score result in a categorization of the patient by degree of response (none, moderate, good).

[00238] In some embodiments, the indication of the subject’s response is measured by the level of the immune response or immune parameters of a cancer-bearing patient resulting from an immunotherapy. In some embodiments, the immune response or immune parameters are characterized by expression level of various biological markers of the host immune response in conjunction with the occurrence of a cancer at a given stage of cancer development (i.e. treatment efficacy). In some embodiments, the expression level of a biological marker is compared with a reference value for the same biological marker, and when required with reference values. The reference value for the same biological marker is thus predetermined and is already known to be indicative of a reference value that is pertinent for discriminating between a low level and a high level of the immune response of a patient with cancer, for said biological marker. Said predetermined reference value for said biological marker is correlated with a responder to treatment in a cancer patient, or conversely is correlated with non-responder to treatment in a cancer patient.

[00239] In some embodiments, a change of a combination of biological markers are quantified. In some embodiments, a combination of 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, or more distinct biological markers are quantified.

[00240] In certain embodiments, biological markers are quantified with immunohistochemical techniques. Example biological markers include 18s, ACE, ACTB, AGTR1, AGTR2, APC, APOA1, ARF1, AXIN1, BAX, BCL2, BCL2L1, CXCR5, BMP2, BRCA1, BTLA, C3, CASP3, CASP9, CCL1, CCL11, CCL13, CCL16, CCL17, CCL18, CCL19, CCL2, CCL20, CCL21, CCL22, CCL23, CCL24, CCL25, CCL26, CCL27, CCL28, CCL3, CCL5, CCL7, CCL8, CCNB1, CCND1, CCNE1, CCR1, CCR10, CCR2, CCR3, CCR4, CCR5, CCR6, CCR7, CCR8, CCR9, CCRL2, CD154, CD19, CD1a, CD2, CD226, CD244, PDCD1LG1, CD28, CD34, CD36, CD38, CD3E, CD3G, CD3Z, CD4, CD40LG, CD5, CD54, CD6, CD68, CD69, CLIP, CD80, CD83, SLAMF5, CD86, CD8A, CDH1, CDH7, CDK2, CDK4, CDKN1A, CDKN1B, CDKN2A, CDKN2B, CEACAM1, COL4A5, CREBBP, CRLF2, CSF1, CSF2, CSF3, CTLA4, CTNNB1, CTSC, CX3CL1, CX3CRI, CXCL1, CXCL10, CXCL11, CXCL12, CXCL13, CXCL14, CXCL16, CXCL2, CXCL3, CXCL5, CXCL6, CXCL9, CXCR3, CXCR4, CXCR6, CYP1A2, CYP7A1, DCC, DCN, DEFA6, DICER1, DKK1, Dok-1, Dok-2, DOK6, DVL1, E2F4, EBI3, ECE1, ECGF1, EDN1, EGF, EGFR, EIF4E, CD105, ENPEP, ERBB2, EREG, FCGR3A, CGR3B, FN1, FOXP3, FYN, FZD1, GAPD, GLI2, GNLY, GOLPH4, GRB2, GSK3B, GSTP1, GUSB, GZMA, GZMH, GZMK, HLA-B, HLA-C, HLA-D, MA, HLA-DMB, HLA-DOA, HLA-DOB, HLA-DPA1, HLA-DQA2, HLA-DRA, HLX1, HMOX1, HRAS, HSPB3, HUWE1, ICAM1, ICAM-2, ICOS, ID1, ifna1, ifna17, ifna2, ifna5, ifna6, ifna8, IFNAR1, IFNAR2, IFNG, IFNGR1, IFNGR2, IGF1, IHH, IKBKB, IL10, IL12A, IL12B, IL12RB1, IL12RB2, IL13, IL13RA2, IL15, IL15RA, IL17, IL17R, IL17RB, IL18, IL1A, IL1B, IL1RI, IL2, IL21, IL21R, IL23A, IL23R, IL24, IL27, IL2RA, IL2RB, IL2RG, IL3, IL31RA, IL4, IL4RA, IL5, IL6, IL7, IL7RA, IL8, CXCR1, CXCR2, IL9, IL9R, IRF1, ISGF3G, ITGA4, ITGA7, integrin, alpha E (antigen CD103, human mucosal lymphocyte, antigen 1; alpha polypeptide), Gene hCG33203, ITGB3, JAK2, JAK3, KLRB1, KLRC4, KLRF1, KLRG1, KRAS, LAG3, LAIR2, LEF1, LGALS9, LILRB3, LRP2, LTA, SLAMF3, MADCAM1, MADH3, MADH7, MAF, MAP2K1, MDM2, MICA, MICB, MKI67, MMP12, MMP9, MTA1, MTSS1, MYC, MYD88, MYH6, NCAM1, NFATC1, NKG7, NLK, NOS2A, P2X7, PDCD1, PECAM-1, CXCL4, PGK1, PIAS1, PIAS2, PIAS3, PIAS4, PLAT, PML, PP1A, CXCL7, PPP2CA, PRF1, PROM1, PSMB5, PTCH, PTGS2, PTP4A3, PTPN6, PTPRC, RAB23,

RAC/RHO, RAC2, RAF, RB1, RBL1, REN, Drosha, SELE, SELL, SELP, SERPINE1, SFRP1, SIRP beta 1, SKI, SLAMF1, SLAMF6, SLAMF7, SLAMF8, SMAD2, SMAD4, SMO, SMOH, SMURF1, SOCS1, SOCS2, SOCS3, SOCS4, SOCS5, SOCS6, SOCS7, SOD1, SOD2, SOD3, SOS1, SOX17, CD43, ST14, STAM, STAT1, STAT2, STAT3, STAT4, STAT5A, STAT5B, STAT6, STK36, TAP1, TAP2, TBX21, TCF7, TERT, TFRC, TGFA, TGFB1, TGFB1, TGFB2, TIMP3, TLR1, TLR1, TLR2, TLR3, TLR4, TLR5, TLR6, TLR7, TLR8, TLR9, TNF, TNFRSF10A, TNFRSF11A, TNFRSF18, TNFRSF1A, TNFRSF1B, OX-40, TNFRSF5, TNFRSF6, TNFRSF7, TNFRSF8, TNFRSF9, TNFRSF10, TNFRSF6, TOB1, TP53, TSLP, VCAM1, VEGF, WIF1, WNT1, WNT4, XCL1, XCR1, ZAP70 and ZIC2.

[00241] Referring to block 324, in some embodiments, the prediction of the subject's response is a class output of a respective response, in a plurality of possible responses, of the respective subject. The method allows the setting of a single "cut-off" value permitting discrimination between responder or non-responder to a treatment. In some embodiments, the prediction of the respective subject's response includes a prediction of an objective response rate of the human subject to the treatment or therapy, and wherein the prediction of the objective response rate includes an indication or classification of a complete response or an amount of a partial response to the treatment.

[00242] Referring to block 326, in some embodiments, the prediction of the subject's response of the subject is a probability output for the respective subject's response. As it is disclosed above, the method allows the setting of a single "cut-off" value permitting discrimination between responder or non-responder to a treatment. In some embodiments, the methods comprise utilizing the model to calculate a probability value for a subject; compare the probability value to a threshold value derived from a cohort of responders/non-responders to determine whether or not the probability value is above or below the threshold value; classify the subject as responder/non-responder if the probability value is above or below the threshold. In embodiments, the threshold value may be about a probability value of at least 50%, 55%, 50%, 65%, 70%, 75% or about 80% or more. In other embodiments, the probability value is a positive predictive value as measured by area under the curve (AUC) of receiver operating characteristic (ROC) curves. In certain embodiments, the probability value is calculated using a multivariate logistic regression model, a neural network model, a random forest model or a decision tree model.

[00243] Referring to block 328, in some embodiments, the model is a neural network algorithm, a support vector machine algorithm, a Naive Bayes algorithm, a nearest neighbor algorithm, a boosted trees algorithm, a Random Forest algorithm, a convolutional neural network algorithm, a decision tree algorithm, a regression algorithm, or a clustering algorithm.

[00244] Referring to block 330, in some embodiments, the plurality of parameters is at least 1000, at least 10,000, at least 15,000, at least 50,000, at least 100,000, at least 250,000, at least 500,000, at least 1,000,000 parameters, at least 2,500,000 parameters, at least 5,000,000 parameters, at least 10,000,000 parameters, or more.

[00245] Referring to block 332, in some embodiments, the model applies the plurality of parameters to the information through at least 1000 computation, at least 5000 computations, at least 10,000 computations, at least 25,000 computations, at least 50,000 computations, at least 100,000 computations, at least 250,000 computations, at least 500,000 computations, at least 1,000,000 computations, at least 2,500,000 computations, at least 5,000,000 computations, at least 10,000,000 computations, or more to obtain a corresponding output for the respective subject from the model.

[00246] Referring to block 334, in some embodiments, the method further comprises treating the subject by: i) when the prediction of the subject's response to the therapy satisfies a threshold likelihood that the subject will respond favorably to the therapy, administering the therapy to the subject; ii) when the prediction of the subject's response to the therapy does not satisfy the threshold likelihood that the subject will respond favorably to the therapy, administering one or more of the plurality of gut microorganisms to the subject.

[00247] In some embodiments, the administering comprises identifying one or more of the plurality of gut microorganisms that is underrepresented in the subject, e.g., as determined based on the corresponding genomic abundance value for the microorganism, and administering the identified one or more gut microorganism to the subject. In some embodiments, the identifying includes determining whether the abundance of a gut microorganism, e.g., as determined based on the corresponding genomic abundance value for the microorganism, satisfies a corresponding threshold amount. When the abundance of the microorganism does not satisfy the corresponding threshold amount, identifying that microorganism for administration. In some embodiments, the corresponding threshold amount is a relative abundance. In some embodiments, the

corresponding threshold amount is an amount relative to the abundance of one or more different gut microorganisms in the subject. In some embodiments, the corresponding threshold amount is an amount relative to the total abundance of the plurality of gut microorganisms in the subject.

[00248] In some embodiments, the administering comprises administering a pre-defined set of microorganisms. In some embodiments, the predefined set of microorganisms includes at least 5 gut microorganisms selected from Table 1, Table 2, or Figures 13A-13XX. In some embodiments, the predefined set of microorganisms includes at least 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, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 110, 120, 130, 140, 150, 160, 170, 175, 180, 190, 200, 225, 250, 275, 300, 350, 400, 450, 500, 600, 700, or more gut microorganisms selected from Table 1, Table 2, or Figures 13A-13XX.

[00249] In some embodiments, the predefined set of microorganisms only includes gut microorganisms selected from Table 1, Table 2, or Figures 13A-13XX that are assigned to Guild 1. That is, the predefined set of microorganisms does not include microorganisms selected from Table 1, Table 2, or Figures 13A-13XX that are assigned to Guild 2. In some embodiments, the predefined set of microorganisms includes at least 5 gut microorganisms selected from Table 1, Table 2, or Figures 13A-13XX that are assigned to Guild 1. In some embodiments, the predefined set of microorganisms includes at least 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, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 110, 120, 130, 140, 150, 160, 170, 175, 180, 190, 200, 225, 250, 275, 300, 350, 400, or more gut microorganisms selected from Table 1, Table 2, or Figures 13A-13XX that are assigned to Guild 1.

[00250] In some embodiments, the predefined set of microorganisms only includes gut microorganisms selected from Table 1, Table 2, or Figures 13A-13XX that are assigned to Guild 2. That is, the predefined set of microorganisms does not include microorganisms selected from Table 1, Table 2, or Figures 13A-13XX that are assigned to Guild 1. In some embodiments, the predefined set of microorganisms includes at least 5 gut microorganisms selected from Table 1, Table 2, or Figures 13A-13XX that are assigned to Guild 2. In some embodiments, the predefined set of microorganisms includes at least 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, 35, 40, 45, 50, 55, 60, 65,

70, 75, 80, 85, 90, 95, 100, 110, 120, 130, 140, 150, 160, 170, 175, 180, 190, 200, 225, 250, 275, 300, 350, 400, or more gut microorganisms selected from Table 1, Table 2, or Figures 13A-13XX that are assigned to Guild 2.

[00251] In some embodiments, when the prediction of the subject's response to the therapy does not satisfy the threshold likelihood that the subject will respond favorably to the therapy, the method further comprises administering the therapy to the subject. In some embodiments, the therapy is administered to the subject around the same time as the one or more of the plurality of gut microorganisms are administered. In some embodiments, the therapy is administered to the subject after the one or more of the plurality of gut microorganisms are administered. In some embodiments, the therapy is administered to the subject at least 1 day, at least 2 days, at least 3 days, at least 4 days, at least 5 days, at least 6 days, at least 7 days, at least 1 week, at least 2 weeks, at least 3 weeks, at least 4 weeks, at least 5 weeks, at least 6 weeks, at least 7 weeks, at least 8 weeks, or more after the one or more of the plurality of gut microorganisms are administered. In some embodiments, the therapy is administered to the subject no more than 3 months, no more than 2 months, no more than one month, no more than 4 weeks, no more than 3 weeks, no more than 2 weeks, no more than 1 week, no more than 6 days, no more than 5 days, no more than 4 days, no more than 3 days, or no more than 2 days after the one or more of the plurality of gut microorganisms are administered. In some embodiments, the therapy is administered to the subject from 1 day to 2 months, from 1 day to 1 month, from 1 day to 3 weeks, from 1 day to 2 weeks, from 1 day to 1 week, from 1 day to 3 days, from 2 days to 2 months, from 2 days to 1 month, from 2 days to 3 weeks, from 2 days to 2 weeks, from 2 days to 1 week, from 2 days to 3 days, from 3 days to 2 months, from 3 days to 1 month, from 3 days to 3 weeks, from 3 days to 2 weeks, from 3 days to 1 week, from 1 week to 2 months, from 1 week to 1 month, from 1 week to 3 weeks, or from 1 week to 2 weeks after the one or more of the plurality of gut microorganisms are administered.

[00252] In some embodiments, if a subject is classified ahead of treatment as a predicted non-responder, then a clinician may treat that subject differently to a subject classified as a predicted responder. Classifying the subject as a predicted non-responder or as a predicted responder may allow the adoption of a particular, or an alternative, treatment regime more suited to the patient.

[00253] In some embodiments, a treatment regime or a therapy can be administered via any common route so long as the target tissue or cell is available via that route. In some embodiment, a treatment regime or a therapy can be administered via any common route so long as the target tissue or cell is available via that route. This includes, but is not limited to, intravenous, catheterization, orthotopic, intradermal, subcutaneous, intramuscular, intraperitoneal, intertumoral, oral, nasal, buccal, rectal, vaginal, or topical administration. Selection of therapeutic agents and dosage regimes may depend on various factors, such as the drug combination employed, the particular disease being treated, and the condition and prior history of the patient.

[00254] In some embodiments, a non-responder is administered with one or more of the pluralities of gut microorganisms via, but is not limited to, oral administration or by colonoscopy. A gut microorganism therapeutic composition for use as described herein can be prepared and administered using methods known in the art. In general, compositions are formulated for oral, colonoscopic, or nasogastric delivery although any appropriate method can be used.

[00255] In some embodiments, a non-responder receives fecal microbiota transplantation from a responder population through methods as disclosed in e.g., US 20230109343, US20200147151, or US 2021036172. In some embodiments, a non-responder receives an effective amount of pre-selected isolated population of gut microorganisms from fecal matters of a responder. In some embodiments, a non-responder receives an effective amount of pre-selected isolated population of gut microorganisms from Table 1, Table 2 or Figure 13A-13XX. In some embodiments, the one or more of the pluralities of gut microorganisms administered to a non-responder comprise a therapeutically effective or sufficient amount of at least 1, at least 5, at least 10, at least 15, at least 20, at least 25, at least 30, at least 35, at least 40, at least 45, at least 50, at least 55, at least 60, at least 65, at least 70, at least 75, at least 80, at least 85, at least 90, at least 95, at least 100, at least 125, at least 150, at least 200, at least 250, at least 300, at least 400, at least 500, at least 600, at least 700, at least 800, or all of the gut microorganisms isolated or purified populations of gut microorganisms selected from Table 1, Table 2 or Figures 13A-13XX. In some embodiments, the one or more of the pluralities of gut microorganisms administered to a non-responder comprise at least about 1×10^3 viable colony forming units (CFU) of bacteria or at least about 1×10^4 , 1×10^5 , 1×10^6 , 1×10^7 , 1×10^8 , 1×10^9 , 1×10^{10} , 1×10^{11} , 1×10^{12} , 1×10^{13} , 1×10^{14} , 1×10^{15}

viable CFU (or any derivable range therein). In some embodiments, a single dose will contain an amount of gut microorganisms (such as a specific bacteria or species, genus, or family described herein) of at least, at most, or exactly 1×10^4 , 1×10^5 , 1×10^6 , 1×10^7 , 1×10^8 , 1×10^9 , 1×10^{10} , 1×10^{11} , 1×10^{12} , 1×10^{13} , 1×10^{14} , 1×10^{15} or greater than 1×10^{15} viable CFU (or any derivable range therein) of a specified bacteria. In some embodiments, a single dose will contain at least, at most, or exactly 1×10^4 , 1×10^5 , 1×10^6 , 1×10^7 , 1×10^8 , 1×10^9 , 1×10^{10} , 1×10^{11} , 1×10^{12} , 1×10^{13} , 1×10^{14} , 1×10^{15} or greater than 1×10^{15} viable CFU (or any derivable range therein) of total gut microorganisms.

[00256] In some embodiments, the pluralities of gut microorganisms are administered concomitantly or sequentially with one or more therapies to a disease or a disorder. In some embodiments, some, most, or substantially all of the subject's colon, gut or intestinal microbiota are removed prior to the administering of the composition.

[00257] In some embodiments, the pluralities of gut microorganisms are administered more than once. In certain aspects, the composition is administered daily, weekly, or monthly. In some embodiments, the pluralities of gut microorganisms are administered for two, three, or four months to induce and/or maintain an appropriate microbiome in the non-responder's GI tract.

[00258] *Examples*

[00259] *Example 1 – The two competing guilds identified in the QD trial (QD-TCG) distinguish cases from controls in 10 independent case-control metagenomic datasets of 6 different diseases.*

[00260] **1.1 Reversible changes in the gut microbiota associate with reversible changes of host metabolic phenotypes**

[00261] QD is an open label, randomized, and controlled interventional trial, in which T2DM patients were randomized at baseline (M0) to receive either 3 months (M3) of a high fiber intervention (W group; n = 74) or standard care (U group; n= 36) followed by a one-year follow-up (M15) (FIG. 4A). Dietary fiber intake in the U group remained unchanged throughout the trial, whereas the W group had a significant increase in the intake of dietary fibers from M0 to M3 and a decrease from M3 to M15 (FIG. 4B). Compared with the U group, fiber intake was significantly higher in the W group at both M3 and M15 (FIG. 4B), but energy and protein

consumption were similar between the two groups across the study period. Fat intake showed no significant difference between the two groups though a trend of decrease in the W group was observed. Compared with the U group, the W group had similar carbohydrate consumption at M0 and M3, but a higher consumption at M15.

[00262] To investigate the gut microbial responses to the introduction and withdrawal of the high fiber intervention, we performed shotgun metagenomic sequencing on 315 fecal samples collected from 110 patients of the W and U group, among whom 95 patients provided samples at all 3 time points and 15 provided samples at M0 and M3 only. To achieve genome-level resolution, we reconstructed 1,845 non-redundant high-quality draft genomes (HQMAGs, two HQMAGs were collapsed into one if the average nucleotide identity, ANI, between them was > 99%) from the metagenomic datasets. These HQMAGs accounted for more than 70% of the total reads. In the context of beta-diversity based on Bray-Curtis distance, the overall structure of the gut microbiota in the W group significantly changed from M0 to M3 (PERMANOVA test, $P < 0.001$) and returned to that of M0 at M15; there was no difference in the U group across the 3 timepoints (FIG. 4C-D). Similar changes in abundance-weighted alpha-diversity based on Shannon and Simpson indices were also observed. Regarding to abundance-unweighted indices, richness and Chao1 were increased from M0 to M3 and remained at M15 in both W and U groups. These results showed that the high fiber intervention induced significant structural and abundance changes of the gut microbiota, however the gut microbiota structure reverted to baseline after the intervention was withdrawn indicating a high resilience in community structure, though there may be lingering in richness increase.

[00263] To determine if host metabolic phenotypes would show similar reversible changes as the gut microbiota, we examined 43 bio-clinical parameters of 9 categories across the 3 time points. Hemoglobin A1c (HbA1c) in the U group showed no significant changes throughout the trial. The high fiber intervention significantly reduced the level of HbA1c in the W group from M0 to M3 by $15.22 \pm 9.82\%$ (mean $\pm$ SD). At one-year follow-up of the W group, HbA1c was significantly increased from M3 but remained lower than that at M0 (FIG. 4E). The proportion of patients who achieved adequate glycemic control ($\text{HbA1c} \leq 7\%$) was significantly higher in the W group (61.6 % versus 33.3% in the U group) at M3, but showed no difference between the two groups at M15 (FIG. 4F). The level of fasting blood glucose and postprandial glucose in

meal tolerance test followed a similar trend as HbA1c (FIG. 4G, H). Among the rest 40 biochemical parameters, 14 also showed an alleviation from M0 to M3 but rebounded at one-year follow-up in the W group. These results indicate that changes of the host metabolic phenotypes were associated with the reversible changes of the gut microbiota in response to the introduction or withdrawal of the high fiber intervention.

[00264] 1.2 Genome pairs with stable interactions form a seesaw-like network of two competing guilds.

[00265] To facilitate the identification of genome pairs that keep their ecological interactions stable during the trial, particularly in the W group with profound microbiota and host phenotypic changes, we constructed a co-abundance network for each time point based on the abundance matrix of the HQMAGs representing the prevalent microbes. Co-abundance network is a data-driven way to investigate ecological interactions between microbes across habitats. A total of 477 HQMAGs were selected for network construction because they were detectable in more than 75% of the samples at each time point in the W group. These 477 HQMAGs also accounted for ~60% of the total abundance of the 1,845 HQMAGs. In the W group, we calculated pairwise correlations of all 113,526 possible genome pairs among these 477 prevalent HQMAGs based on their abundance across the patients at each time point and constructed 3 co-abundance networks (G_{M0} , G_{M3} and G_{M15}) by using Fastspar, a rapid and scalable correlation estimation tool for compositional data. The three networks were of similar order S , i.e., the total number of nodes (HQMAGs), $S_{M0}(442)$, $S_{M3}(421)$, and $S_{M15}(429)$, but they varied considerably in their size L , i.e., the total number of edges (correlations), $L_{M0}(4,231)$, $L_{M3}(2,587)$ and $L_{M15}(4,592)$. L in G_{M3} decreased to 61.14% of that in G_{M0} and rebounded back in G_{M15} to 108.53% of that in G_{M0} . This pattern was confirmed by changes in connectance, which is defined as the proportion of realized ecological interactions among the potential ones (in undirected network, $\text{connectance} = \frac{L}{S(S-1)/2}$, range: [0,1]). Connectance decreased from 0.043 in G_{M0} to 0.029 in G_{M3} and rebounded to 0.050 in G_{M15} . Changes in L and connectance showed that the high fiber intervention dramatically reduced the correlations among the prevalent genomes in the network. In addition, we found that the distributions of degree, i.e. the number of edges a node has, fit well with a power-law model (R^2 values G_{M0} : 0.79, G_{M3} : 0.82, G_{M15} : 0.79), indicating the presence of network hubs²¹. Defining hubs as nodes that connect with more than one-fifth of the total nodes in the network,

we found 24 hubs, in which 10 were in G_{M0} , 20 were in G_{M15} but none were in G_{M3} . These results indicate that the overall structure of the gut microbiome undergone profound changes during the trial, particularly, the high fiber intervention resulted in the loss of interactions between genome pairs.

[00266] We considered two genomes are connected with robust and stable ecological relationship if they keep the same type of correlation across all the three timepoints. Out of the 113,526 possible genome pairs, 92.39% showed no correlations at any of the three time points, suggesting that it is uncommon for two genomes to establish an ecological relationship even transiently (FIG. 5A). Interestingly, 517 genome pairs showed positive correlations and 118 negative correlations at all the three time points. All the 635 stable correlations involved 184 HQMAGs. These HQMAGs were grouped into 17 unconnected clusters based on the Connected Components Clustering analysis. Cluster C2-C16 had only between 2-9 HQMAGs which had only positive correlations. While cluster C1 was more complex as it had 141 HQMAGs which had both negative and positive correlations. To identify if sub-clusters existed in C1, we further built a clustering tree based on the negative and positive correlations with average linkage method and applied WGCNA analysis. The 141 genomes in C1 were further grouped into 2 subclusters, C1A and C1B. Interestingly, among C1A, C1B and C2-C16, only C1A and C1B were significantly correlated with our primary outcome HbA1c (FIG. 5B). Thus, we focused C1A and C1B for further analysis.

[00267] C1A and C1B can be considered as guilds as HQMAGs in each cluster were highly interconnected with only positive correlations no matter which were robust or transient (FIG. 5B). The two guilds were connected by negative edges only, indicating a competitive relationship that structures a seesaw-like network. Such a network feature was termed as two competing guilds (TCG). The members of the TCG had significantly higher degree, betweenness centrality, eigenvector centrality, closeness centrality and stress centrality than the rest of the genomes in the networks. This finding indicates that the two guilds exerted a relatively large amount of control over the interaction of other nodes (reflected by betweenness centrality and eigenvector centrality) and the information flow in the network (reflected by closeness centrality and stress centrality). Removing the two guilds would lead to the collapse of the networks since on average 86.08% of the total edges would have been lost. These suggest that the genomes of the two guilds can be considered as the core nodes of the three large networks (G_{M0} , G_{M3} , and G_{M15}) as

they were not only the most stably connected but also the most highly connected nodes in the gut ecosystem network.

[00268] The members of the two guilds were also highly prevalent among participants, as 137 of them were in > 90%, and 95 were in 100% of the individuals in the W and U groups. In addition, most of these 141 HQMAGs were also predominant members of the gut microbiota as the abundance of 111 of them was higher than the median of the 1,845 HQMAGs and accounted for 20.78% of the total sequencing reads. Based on Bray-Curtis distance, beta-diversity analysis showed significant correlations between the profiles of the two guilds' members and all the 1,845 HQMAGs, as evidenced in the Mantel test ($R^2 = 0.62$, $P = 0.001$) and Procrustes analysis ($P = 0.001$) (FIG. 4C, D). These indicate that the variations of the two guilds contributed to the major variations of the whole gut microbial community across the 3 time points.

[00269] Our data showed that the TCG of the 141 HQMAGs existed in all three ecological networks G_{M0} , G_{M3} , and G_{M15} in the W group. Furthermore, the finding of the TCG in the W group at M0 suggests that such microbial organization exists irrespective of the high fiber intervention in our study. Given the similar overall gut microbiota structure between the W and U groups at M0 and in the U group across 3 timepoints (FIG. 4C, D), we speculated that the TCG can also be observed in the U group across the trial. Thus, we constructed the co-abundance networks based on the abundance of the 141 HQMAGs across the individuals in the U group at each time point. In the co-abundance networks, 99.8%, 99.51% and 99.74% of the total edges agreed with our TCG, i.e., positive correlations inside the guilds and negative between guilds. This showed that the detection of the TCG was independent of the high fiber intervention, indicating that such a pattern may be an inherent structure of the gut microbiome in our study.

[00270] 1.3 Dynamics of the two competing guilds associate with changes of host metabolic phenotypes.

[00271] We sought to determine whether the balance of the TCG could be modulated by dietary fiber and describe how the TCG affects the host metabolic phenotypes. In the W group, the abundance of Guild 1 increased and Guild 2 decreased significantly from M0 to M3. Then at M15, Guild 1 decreased to a level similar to that at M0, and Guild 2 increased but were not significantly different from that at M3. Subsequently, from M0 to M3, high fiber intervention significantly increased the Guild 1-to-Guild 2 ratio. At one-year follow-up, the ratio significantly

decreased (FIG. 6A). Neither the abundances of the 2 guilds nor their ratio was changed in the U group across the trial. These results showed that the changes of the balance between the two guilds were concomitant with changes in dietary fiber intake, overall gut microbiota and host phenotypes. To further explore the importance of the two guilds to host health, we applied linear mixed effect model to identify the associations between the two guilds and each host bio-clinical parameter. The models were trained using the abundances of the two guilds and bio-clinical parameters at M0 and M3. Subsequently, the models were used to generate predicted values for each bio-clinical parameter based on the guilds' abundance at M15. These predicted values were then correlated with the measured bio-clinical parameters at M15. Forty-two out of the 43 bio-clinical parameters had significant Pearson's correlation coefficient ranged from 0.14 to 0.88 (adjusted P value < 0.05) between the predicted and measured values (FIG. 6B). These results showed the TCG constitutes an important microbiome signature for T2DM and the related metabolic phenotypes.

[00272] Next, we performed genome-centric analysis of the 141 HQMAGs in the TCG to explore the genetic basis underlying the association between the dynamic changes of the seesaw networked microbiome signature and relationships with the host's metabolic phenotypes. As the balance between the two guilds can be shifted by dietary fibers, we first sought to identify carbohydrate-active enzyme (CAZy)-encoding genes and genes encoding key enzymes in short-chain fatty acid (SCFA) production to compare the genetic capacity for carbohydrate utilization between the two guilds. Compared with genomes in Guild 2 (C1B), those in Guild 1 (C1A) enriched CAZy genes for arabinoxylan ($P < 0.001$), cellulose ($P < 0.01$) and had lower proportion of CAZy genes for inulin utilization ($P < 0.01$) (FIG. 6C). There was no difference in genes for starch, pectin, and mucin utilization between the two guilds. Our previous study showed that gut microbiota benefited patients with T2DM via acetic and butyric acid production from carbohydrate fermentation¹¹. Among the terminal genes for the butyrate biosynthetic pathways from both carbohydrates (i.e., *but* and *buk*) and proteins (i.e., *atoA/D* and *4Hbt*), the copy number of *but* was significantly higher in Guild 1 and there was no difference in the other terminal genes between the two guilds (FIG. 6C). More than one-third of the genomes in Guild 1 harbored the *but* gene while less than 5% of the genomes in Guild 2 had this gene (Fisher's exact test $P < 0.001$). Compared with Guild 2, Guild 1 also trended higher in its genetic capacity for acetate production ($P = 0.06$) but had a lower genetic capacity for propionate production ($P < 0.05$)

(FIG. 6C). These results showed that compared to Guild 2, Guild 1 had significantly higher genetic capacity for utilizing complex plant polysaccharides and producing acetate and butyrate.

[00273] From the perspective of pathogenicity, 21 out of all the 1,845 HQMAGs encoded 750 virulence factor (VF) genes. Among the 21 VF-encoding HQMAGs, 3 were in Guild 1 while 18 were in Guild 2. Three out of the 50 HQMAGs in Guild 1 had one VF gene involved in antiphagocytosis. In Guild 2, 18 out of the 91 HQMAGs encoded 747 VF genes across 15 different VF classes i.e., acid resistance, adherence, antiphagocytosis, biofilm formation, efflux pump, endotoxin, invasion, iron uptake, manganese uptake, motility, nutritional factor, protease, regulation, secretion system, and toxin (FIG. 6C). Notably, 98.53% of all the VF genes in Guild 2 were harbored in 8 HQMAGs (1 in *Enterobacter kobei*, 2 in *Escherichia flexneri*, 3 in *Escherichia coli* and 2 in *Klebsiella*). The highly enriched genes for virulence factors in HQMAGs of Guild 2 (Fisher's Exact test, $P < 2.2 \times 10^{-16}$,) indicates that this guild may play an important role in aggravating the metabolic disease phenotypes.

[00274] In terms of antibiotic resistance genes (ARG), in Guild 1, only 1 HQMAG (2.00% of the genomes in this guild) harbored a copy of an ARG related to phenicol (FIG. 6C). In Guild 2, 17 HQMAGs (18.68% of the genomes in this guild) encode 40 ARGs for resistance to 7 different antibiotic classes i.e., aminoglycosides, beta-lactam, fosfomycin, glycopeptide, quinolone, macrolide, and tetracycline. Thus, Guild 2 may serve as a reservoir of ARGs for horizontal transfer to opportunistic pathogens. Taken together, our data showed that the two competing guilds had distinct genetic capacity with Guild 1 being potentially beneficial and Guild 2 detrimental.

[00275] **1.4 The two competing guilds identified in the QD trial (QD-TCG) distinguish cases from controls in 10 independent case-control metagenomic datasets of 6 different diseases.**

[00276] As the two competing guilds identified in the QD trial (QD-TCG) were found to be significant microbiome signatures for responding to dietary intervention in T2DM, we investigated whether they could function as biomarkers to distinguish T2DM from controls. We addressed this question in an independent T2DM metagenomic dataset, comprising 136 T2DM cases and 136 controls, using all the 141 HQMAGs from QD-TCG as reference genomes. These

were used in a read recruitment analysis, a widely utilized method to estimate the abundance of reference genomes in metagenomes. On average, 35.28% and 32.92% reads were recruited in the cases and controls samples, respectively. Following this, we developed a machine learning classifier based on a Random Forest algorithm, employing the abundance of the 141 HQMAGs to determine if we could distinguish cases from controls. Receiver operating characteristic curve analysis demonstrated a moderate diagnostic capability, with an area under the curve (AUC) of 0.70, ascertained through leave-one-out cross-validation.

[00277] Subsequently, we hypothesized that the QD-TCG might represent an intrinsic pattern in the human microbiome, irrespective of disease types. To assess this, we collected 10 independent metagenomic datasets of cases and controls spanning six different diseases: liver cirrhosis (LC), ankylosing spondylitis (AS), atherosclerotic cardiovascular disease (ACVD), schizophrenia (SCZ), colorectal cancer (CRC), and inflammatory bowel disease (IBD). We refer to the collection of these 10 datasets, together with the aforementioned T2DM dataset, as the Case-Control Dataset Collection I (CCDC-I). On average, from these 10 metagenomic datasets, 32.12% and 31.34% reads were recruited to the two guilds in the cases and control samples, respectively (FIG 6D). Within each dataset, a Random Forest classifier using the 141 HQMAGs demonstrated diagnostic power in distinguishing cases from controls, with the AUC ranging from 0.68 for SCZ to 0.98 for AS#1 (FIG 6E).

[00278] These findings underscore the discriminative power of the QD-TCG in distinguishing cases versus controls across various disease types. This suggests a common microbiome signature that is associated with a broad spectrum of human diseases.

[00279] **Materials and Methods**

[00280] **Clinical Experiment**

[00281] Study design: QD trial was conducted at the Qidong People's Hospital (Jiangsu, China), examined the effect of a high fiber diet in free-living conditions in a cohort of individuals clinically diagnosed T2DM. The study protocol was approved by Ethics Committee of Shanghai General Hospital (2014KY104), and the study was conducted in accordance with the principles of the Declaration of Helsinki. All participants provided written informed consent. The trial was registered in the Chinese Clinical Trial Registry (ChiCTR-IPC-14005346).

[00282] T2DM patients of the Chinese Han ethnicity were recruited for the study (age: 37 - 70 years; HbA1c: 6.5% - 12.0%). More detailed description of inclusion and exclusion criteria were shown in Chinese Clinical Trial registry, found at the URL www.chictr.org.cn.

[00283] Patients received either a high-fiber diet (WTP diet) as the treatment group (W group) or the usual care (Usual diet) as the control group (U group) for 3 months. Total caloric and macronutrients prescriptions were based on age-specific Chinese Dietary Reference Intakes (Chinese Nutrition Society, 2013). The WTP diet, based on wholegrains, traditional Chinese medicinal foods and prebiotics, included three ready-to-consume pre-prepared foods. The usual care included standard dietary and exercise advice that was made according to the Chinese Diabetes Society guidelines for T2DM. Patients in W group were provided with the WTP diet to perform a self-administered intervention at home for three months, while patients in U group accepted the usual care. W group stopped WTP diet intervention at the end of the third month (at M3). Then W and U continued a one-year follow-up (M15). A meal-based food frequency questionnaire and 24-h dietary recall were used to calculate nutrient intake based on the China Food Composition 2009. Patients in both groups continued with their antidiabetic medications according to their physician prescriptions .

[00284] Before a 2-week run-in period, all participants attended a lecture on diabetes intervention and improvements and received diabetes education and metabolic assessments. 119 eligible individuals were enrolled based on the inclusion and exclusion criteria and assigned into two groups in a 2:1 ratio (n = 79 in W group, n = 40 in U group) determined by SAS software.

[00285] Physical examinations were carried out at M0, M3, and M15 in Qidong People's Hospital (Jiangsu, China). Sample collection instructions were provided to the participants at the day before. The participants provided the feces and first early morning urine as requested. After collecting fasting venous blood sample, a 3-h meal tolerance test (Chinese buns containing 75 g of available carbohydrates; MTT test) was conducted and the postprandial venous blood samples at 30, 60, 120, and 180 min were collected. All the blood samples were centrifuged at 3000 rpm for 20 min at 4°C after standing at room temperature for 30 min to obtain serum. The fasting blood serum were divided into two parts, one used for hospital tests and the other used for lab tests. The feces, urine, and serum samples were stored in dry ice immediately then transported to lab and frozen at -80°C. Subsequently, anthropometric markers and diabetic complication

indexes were measured. Ewing test and 24-h dynamic electrocardiogram were conducted to estimate diabetic autonomic neuropathy (DAN). B-mode carotid ultrasound was conducted to estimate atherosclerosis. Michigan Neuropathy Screening Instrument was conducted to estimate diabetic peripheral neuropathy (DPN). In addition, A meal-based food frequency questionnaire and the 24-h dietary review were recorded for nutrient intake calculation.

[00286] The fasting venous blood was used to measure HbA1c, fasting blood glucose, fasting insulin, fasting C-Peptide, C-reactive protein (CRP), blood routine examination, blood biochemical examination and five analytes of thyroid. The venous blood samples at 30, 60, 120, and 180 min of MTT were used to measure the postprandial blood glucose, insulin, and C-Peptide. The fasting early morning urine was used to measure the routine urine examination and urinary microalbumin creatinine ratio. The measurements above were completed at Qidong People's Hospital. Fasting venous blood was used to quantify TNF- α (R&D Systems, MN, USA), lipopolysaccharide-binding protein (Hycult Biotech, PA, USA), leptin (P&C, PCDBH0287, China) and adiponectin (P&C, PCDBH0016, China) by enzyme-linked immunosorbent assays (ELISAs) at Shanghai Jiao Tong University.

[00287] The homeostatic model assessments of insulin resistance (HOMA-IR) and islet β -cell function (HOMA- β) were calculated based on fasting blood glucose (mmol/L) and fasting C-Peptide (pmol/L) : $\text{HOMA-IR} = 1.5 + \text{FBG} * \text{Fasting-C-Peptide} / 2800$; $\text{HOMA-}\beta = 0.27 * \text{Fasting-C-Peptide} / (\text{FBG} - 3.5)$. Glomerular Filtration Rate was estimated by formula $\text{GFR} (\text{ml/min per } 1.73 \text{ m}^2) = 186 * \text{Scr}^{-1.154} * \text{age}^{-0.203} * 0.742 (\text{if female}) * 1.233 (\text{if Chinese})$, where Scr (serum creatinine) is in mg/dl and age is in years.

[00288] *Gut microbiome analysis*

[00289] **Metagenomic sequencing.** DNA was extracted from fecal samples using the methods as previously described. Metagenomic sequencing was performed using Illumina Hiseq 3000 at GENEWIZ Co. (Beijing, China). Cluster generation, template hybridization, isothermal amplification, linearization, and blocking denaturing and hybridization of the sequencing primers were performed according to the workflow specified by the service provider. Libraries were constructed with an insert size of approximately 500 bp followed by high-throughput sequencing to obtain paired-end reads with 150 bp in the forward and reverse directions.

[00290] Data quality control. Prinseq was used to: 1) trim the reads from the 3' end until reaching the first nucleotide with a quality threshold of 20; 2) remove read pairs when either read was < 60 bp or contained "N" bases; and 3) de-duplicate the reads. Reads that could be aligned to the human genome (H. sapiens, UCSC hg19) were removed (aligned with Bowtie2 using --reorder --no-hd --no-contain --dovetail).

[00291] De novo assembly, abundance calculation, and taxonomic assignment of genomes. De novo assembly was performed for each sample by using IDBA_UD (--step 20 --mink 20 --maxk 100 --min_contig 500 --pre_correction). The assembled contigs were further binned using MetaBAT (--minContig 1500 --superspecific -B 20). The quality of the bins was assessed using CheckM. Bins had completeness > 95%, contamination < 5% and strain heterogeneity < 5% were retained as high-quality draft genomes . The assembled high-quality draft genomes were further dereplicated by using dRep. DiTASiC, which applied kallisto for pseudo-alignment and a generalized linear model for resolving shared reads among genomes, was used to calculate the abundance of the genomes in each sample, estimated counts with P-value > 0.05 were removed, and all samples were downsized to 36 million reads (One sample with read mapping ratio < 25%, which could not be well represented by the high quality genomes, were removed in downstream analysis). Taxonomic assignment of the genomes was performed by using GTDB-Tk with default parameters .

[00292] Gut microbiome network construction and analysis. In W group, prevalent genomes shared by more than 75% of the samples at every timepoint were used to construct the co-abundance network at each timepoint. Fastspar, a rapid and scalable correlation estimation tool for microbiome study, was used to calculate the correlations between the genomes with 1,000 permutations at each time point based on the abundances of the genomes across the patients and the correlations with $P \leq 0.001$ were retained for further analysis. The networks were visualized with Cytoscape v3.8.1. The layout of the nodes and edges was determined by Edge-weighted Spring Embedded Layout using the correlation coefficient as weights. The links between the nodes are treated as metal springs attached to the pair of nodes. The correlation coefficient was used to determine the repulsion and attraction of the spring. The layout algorithm sets the position of the nodes to minimize the sum of forces in the network. We defined robust stable edges as the unchanged positive/negative correlations between the same

two genomes across all the 3 networks at M0, M3, and M15. Nodes in the stable network were clustered by Connected Components Clustering analysis in Cytoscope. To identify if sub-clusters existed in Cluster C1, a clustering tree based on the negative (set as -1) and positive correlations (set as 1) with average linkage method followed by WGCNA analysis.

[00293] Case-Control Dataset Collection I (CCDC-I). Eleven independent metagenomic datasets on T2DM, LC, AS, ACVD, SCZ, CRC and IBD were downloaded from SRA or ENA database. The group information, i.e., case or control, was collected from the corresponding papers or from curatedMetagenomicData. Quality control of raw reads was conducted by KneadData (see the URL bitbucket.org/biobakery/kneaddata). DiTASiC was used to recruit reads and estimate the abundance of the 141 HQMAGs in QD-CTG in each sample, estimated counts with P-value > 0.05 were removed and further converted to relative abundance divided by the total number of reads. A random forest classification model to classify case and control was constructed based on the estimated abundances of the HQMAGs in each dataset with leave-one-out cross-validation. High quality reads were de novo assembled for each sample using MEGAHIT (-min-contig-len 500, -presets meta-large). The assembled contigs were binned using MetaBAT 2 and MaxBin 2. The quality of the bins was assessed using CheckM. Bins had completeness > 95%, contamination < 5% and strain heterogeneity < 5% were retained as HQMAGs, which were further dereplicated by using dRep. In the case and control groups, Fastpar with 1,000 permutations was used to calculate the co-abundance correlations based on the HQMAGs shared by more than 75% of the samples in both groups. Correlations with $P \leq 0.001$ were retained for further analysis. Robust stable edges were defined as the unchanged positive/negative correlations between the same two genomes in both case and control groups. Same clustering analysis performed in the QD trial was conducted on stable network. DiTASiC was used to estimate the abundance of HQMAGs in each TCG in each sample. A random forest classification model to classify case and control was constructed with leave-one-out cross-validation to test each TCG.

[00294] Case-Control Dataset Collection II (CCDC-II). Fifteen independent metagenomic datasets on AS, ASD, BD, COVID-19, CRC, GD, HT, MS, PC and PD were downloaded from SRA or ENA database. The group information, i.e. case or control, was collected from the corresponding papers or from curatedMetagenomicData. Quality control of raw reads was

conducted by KneadData. DiTASiC was used to recruit reads and estimate the abundance of HQMAGs in C-TCG and CC-TCG. A random forest classification model to classify case and control was constructed based on the estimated abundances of the HQMAGs in each dataset with leave-one-out cross-validation.

[00295] Treatment Dataset Collection (TDC). Eleven independent metagenomic datasets on pre-treatment samples related to four diseases, including IBD, rheumatoid arthritis (RA), advanced melanoma and B cell lymphoma, were download from SRA or ENA database. The responder and non-responder categories of each sample were collected from the corresponding paper. Quality control of raw reads was conducted by KneadData. DiTASiC was used to recruit reads and estimate the abundance of HQMAGs in CC-TCG. A random forest classification model to predict responder and non-responder was constructed based on the estimated abundances of the HQMAGs in each dataset with leave-one-out cross-validation.

[00296] Gut microbiome functional analysis. Prokka was used to annotate the HQMAGs. KEGG Orthologue (KO) IDs were assigned to the predicted protein sequences in each HQMAG by HMMSEARCH against KOfam using KofamKOALA with default parameters. KOs were further assigned to KEGG modules. Antibiotic resistance genes were predicted using ResFinder with default parameters. The identification of virulence factors was based on the core set of Virulence Factors of Pathogenic Bacteria Database (VFDB). The predicted proteins sequences were aligned to the reference sequence in VFDB using BLASTP (best hist with E-value < 1e-5, identity > 80% and query coverage > 70%). Genes encoding carbohydrate-active enzymes (CAZys) were identified using dbCAN (release 6.0), and the best-hit alignment was retained. Genes encoding formate-tetrahydrofolate ligase, propionyl-CoA: succinate-CoA transferase, propionate CoA-transferase, 4Hbt, AtoA, AtoD, Buk and But were identified as described previously.

[00297] Statistical Analysis.

[00298] Statistical analysis was performed in the R environment (R version 3.6.1). Friedman test followed by Nemenyi post-hoc test was used for intra-group comparisons for repeat measurements. Mann-Whitney test (two-sided) was used for comparisons between W and U at the same time point. Pearson Chi-square tests was performed to compare the differences of

categorical data between groups or timepoints. PERMANOVA test (9,999 permutations) was used to compare the groups of gut microbiota structure.

[00299] Mann-Whitney test (two-sided) and Fisher's exact test (two sided) were used to compare the target functions between Guild 1 and Guild 2. Hierarchical clustering analysis based on Jaccard distance on the KO profiles was conducted to compare HQMAGs in CC-TCG.

[00300] Linear mixed effect model with subject id as random effect was applied to explore the associates between the abundance of guilds in QD-TCG and clinical parameters. For each HQMAG belonging to a guild, the robust clr-transformed abundance across each sample was first range-scaled. Subsequently, the guild abundance was then obtained as the mean of the range-scaled abundances of HQMAGs belonging to that guild. The M0 and M3 timepoints was used to train the linear mixed effect model, and M15 was used as testing.

[00301] The Random Forest with leave-one-out cross-validation was used to perform classification analysis based on HQMAGs of TCGs in each dataset of CCDC-I, CCDC-II and TDC. The Random Forest with 10-fold cross-validation was used to perform classification analysis based on HQMAGs in CC-TCG for universal model.

[00302] *Example 2 – The combined core genomes from all the two competing guilds (CC-TCG) shows better performances in classifying case vs control across diseases*

[00303] **2.1 Disease as the ecosystem perturbation leads to identification of more sets of two competing guilds with stable correlations in independent case-control datasets.**

[00304] To extend our understanding of the two competing guilds i.e., QD-TCG which were identified based on unveiling stable genome pairs despite dietary interventions, we focused our research on pinpointing stable genome pairs across both cases and controls in cross-sectional datasets, with disease progression acting as a perturbation to the gut ecosystem. For this purpose, we first reconstructed non-redundant HQMAGs from the metagenomic datasets in CCDC-I (FIG. 7A). In total, we obtained 682, 384, 564, 1308, 812, 1644 and 774 HQMAGs in the case and control datasets on T2D, LC, AS, ACVD, SCZ, CRC and IBD. For each disease, we proceeded to construct a co-abundance network for both case and control groups, predicated on the prevalence of the HQMAGs that were shared by over 75% of subjects in both groups. We

identified stable HQMAG pairs, which demonstrated positive/negative correlations in both case and control groups, and further grouped them using Components Clustering analysis. Notably, prevalent HQMAGs with stable correlations were classified into 2, 6, 10, 4, 10, 8 and 7 clusters in the studies on T2D, LC, AS, ACVD, SCZ, CRC and IBD.

[00305] Within each disease type, we observed one primary cluster which held the majority of genome numbers of the stable network and displayed intricate interactions with both positive and negative correlations. To ascertain if C1s contained sub-clusters, we built a clustering tree using the negative and positive correlations with the average linkage method and applied WGCNA analysis. Each Cluster C1 was subsequently segmented into two sub-clusters, C1A and C1B. Interestingly, every pair of C1A and C1B followed the pattern of two competing guilds that we observed in the QD-TCG. The majority of stable correlations between the HQMAGs belonging to C1A and C1B exhibited positive correlations within each cluster and negative correlations between the two clusters. These correlations accounted for 76.88%, 95.31%, 100%, 96.23%, 96.43%, 96.72%, and 97.10% of the stable correlations within or between the C1A and C1B in the studies on T2D, LC, AS, ACVD, SCZ, CRC and IBD respectively. These results highlight the possibility of detecting TCGs from case-control datasets and demonstrate the existence of TCGs in the human gut microbiome across various diseases, ethnicities and geographical regions.

[00306] Next, we investigated whether the TCGs identified from CCDC-I could function as biomarkers to differentiate cases and controls. For this, we used the corresponding HQMAGs as reference genomes to perform read recruitment analysis in the metagenomic datasets of CCDC-I for each TCG. On average, 54.58%, 28.32%, 26.93%, 63.62%, 44.54%, 55.72% and 34.21% of the reads in CCDC-I were recruited to the TCGs discovered in the studies on T2D, LC, AS, ACVD, SCZ, CRC and IBD respectively. For each CCDC-I dataset, we then employed a Random Forest classifier to differentiate cases from controls using the abundance of the HQMAGs in each TCG. Out of the 11 datasets in CCDC-I, the TCGs from the studies on T2D, LC, AS, ACVD, SCZ, CRC and IBD achieved moderate to excellent diagnostic power to classify cases and controls in 8, 9, 9, 9, 8, 7, 7 datasets, with an average AUC of 0.79, 0.78, 0.78, 0.77, 0.77, 0.77 and 0.74 (FIG. 7B). These findings underscore the versatility of the TCGs in classifying cases versus controls across various disease types, suggesting that the TCGs with

stable correlations constitute a shared microbiome signature associated with a broad spectrum of human diseases.

[00307] 2.2 The combined core genomes from all the two competing guilds (CC-TCG) shows better performances in classifying case vs control across diseases.

[00308] In total, we amassed 925 HQMAGs pertaining to 8 distinct TCGs from the QD trial and CCDC-I. These were consolidated into a pool of 788 non-redundant HQMAGs after a de-redundancy analysis based on a genomic ANI cutoff > 99%. The collective of these 788 non-redundant HQMAGs is hereby referred to as the combined genomes of the two competing guilds (C-TCG), a representation of the confluence of the 8 TCG sets. Within the C-TCG, 701 HQMAGs were unique to one of the 8 sets and 87 shared across multiple sets. Among the unique ones, 301 belonged to C1A and 400 to C1B. Among the shared ones, 10 and 40 consistently belonged to C1A and C1B respectively, while 37 exhibited inconsistent assignments across different TCGs (FIG. 8A). To ascertain if C-TCG would improve case-control classification across diseases, we conducted a read recruitment analysis on the metagenomic datasets of CCDC-I, using corresponding HQMAGs as reference genomes. C-TCG accounted for 84.54% of total abundance on average. Subsequently, a Random Forest classifier was trained on each CCDC-I dataset using the abundance of HQMAGs in C-TCG. Overall, C-TCG demonstrated superior case-control classification capacity in CCDC-I compared to individual TCGs, with significantly higher AUC values than classifiers trained on TCGs from the T2D, AS, IBD, SCZ, and LC studies.

[00309] Next, we aimed to identify C-TCG members most relevant to classification performance as the core genomes. We ranked the 788 HQMAGs in C-TCG based on their feature importance in the Random Forest models built for each CCDC-I dataset. Starting from the least significant HQMAGs, we sequentially removed one HQMAG and trained new Random Forest models in each dataset, assessing AUC values for classifying cases and controls. This resulted in 788 different classifiers, each employing a variable number of HQMAGs ranging from 1 to 788 in each CCDC-I dataset. We assigned ranks to HQMAG numbers based on the AUC values of their corresponding models, with lower ranks indicating higher AUC values. Notably, classifiers trained on the top 302 HQMAGs showed the best classification performance in CCDC-I as demonstrated by the smallest cumulative rank (FIG. 8A). From these 302

HQMAGs, 103 were unique to C1A, 181 unique to C1B, and 18 showed inconsistent C1A and C1B assignment across different TCGs. After discarding the 18 inconsistent HQMAGs, we obtained a set of 284 HQMAGs that were not only most relevant to classification performance but also consistently assigned to the two competing guilds. We referred to these HQMAGs as the combined core set of the two competing guilds (CC-TCG). Overall, the Random Forest classifier built on CC-TCG demonstrated superior performance in classifying cases and controls compared to both C-TCG and individual TCGs from the QD trial and CCDC-I, with significantly higher AUC values than classifiers trained on TCGs from the CRC, T2D, AS, IBD, SCZ, and LC studies.

[00310] To decipher the genetic underpinnings of the associations between these genomes and host health, we conducted a genome-centric analysis of the HQMAGs belonging to CC-TCG. We first performed targeted functional analysis and compared HQMAGs assigned to C1A and C1B. Similar to QD-TCG findings, C1A had a significantly higher gene copy number for butyrate biosynthesis and lower for propionate production. In relation to carbohydrate-degrading genes, HQMAGs in C1A were rich in CAZy genes for arabinoxylan and cellulose utilization. These findings suggest that compared to C1B, HQMAGs from C1A have a higher genetic capacity for utilizing complex plant polysaccharides and producing butyrate. From the perspective of antibiotic resistance and pathogenicity, C1A had fewer ARGs and VFs than C1B.

[00311] Furthermore, we conducted an untargeted functional analysis based on the assignment of KEGG Orthology (KO) to all predicted genes from the 284 core HQMAGs. In total, we found 3,553 and 5,495 KOs in C1A and C1B respectively. Hierarchical clustering analysis based on KO profiles demonstrated significant functional differences between C1A and C1B (PERMANOVA, $P = 0.001$). KOs from C1A and C1B were further mapped to 253 and 291 KEGG modules. There were 250 shared modules between the two groups. C1A had 3 unique modules for acarbose biosynthesis, benzoate degradation and staphyloferrin B biosynthesis. C1B had 41 unique modules including those for multidrug resistance, KDO2-lipid A modification, pathogenicity signature and gamma-aminobutyrate production. In conclusion, these results show that the CC-TCG has distinct genetic capacities, with C1A being potentially beneficial and C1B detrimental.

[00312] 2.3 The combined core genomes in the two competing guilds (CC-TCG) differentiate cases from controls for additional datasets.

[00313] In order to further substantiate the capability of CC-TCG in distinguishing cases from controls across a spectrum of diseases, we compiled an additional 15 independent metagenomic datasets. These comprised case and control data from 10 different diseases, including Ankylosing Spondylitis (AS), Autism Spectrum Disorder (ASD), Behcet's Disease (BD), COVID-19, Colorectal Cancer (CRC), Graves' Disease (GD), Hypertension (HT), Multiple Sclerosis (MS), Pancreatic Cancer (PC), and Parkinson's Disease (PD). These 15 datasets are collectively referred to as the Case-Control Dataset Collection II (CCDC-II). Employing the 284 HQMAGs that belong to CC-TCG as reference genomes, we conducted a read recruitment analysis on the metagenomic datasets of the CCDC-II. On average, we found that 34.41% and 34.76% of reads were recruited in the case and control samples respectively. For each dataset in the CCDC-II, we then utilized the abundance of the 284 HQMAGs in CC-TCG to train a Random Forest classifier to discern cases from controls. The CC-TCG showed moderate to excellent diagnostic power in 10 of the 15 datasets, specifically those related to AS, ASD, COVID-19, CRC, GD, HT, MS, and PC, although it only achieved an AUC value of 0.58 for HT#2, and AUC values between 0.6-0.7 for BD, PD, CRC#4 and CRC#5 datasets (FIG.8B).

[00314] In terms of diseases like CRC, IBD, and PC, which had more than two different datasets, we also employed cross-dataset analysis (with one dataset used for model training and the other for testing) and leave-one-dataset-out (LODO) analysis to assess the universal applicability of the diagnostic power of CC-TCG for these diseases. For CRC, we found the model's transportability from one individual dataset to another to be insufficient, similar to the microbiome signature report in CRC by Thomas et al. However, in the case of IBD and PC, the classification model trained on CC-TCG in one dataset exhibited moderate to outstanding performance when classifying cases and controls in the other datasets. Pooling the training datasets, as done in the LODO analysis, improved the performance of CC-TCG-based models in classifying cases versus controls in CRC, with AUC values ranging from 0.66 to 0.79 (FIG. 9A) LODO analysis produced AUC values ranging from 0.75 to 0.91 for IBD, and 0.71-0.72 for PC (FIG. 9B-C). These results attest to the classification power of CC-TCG in completely independent datasets.

[00315] Materials and Methods**[00316] *Gut microbiome analysis***

[00317] Metagenomic sequencing. DNA was extracted from fecal samples using the methods as previously described. Metagenomic sequencing was performed using Illumina HiSeq 3000 at GENEWIZ Co. (Beijing, China). Cluster generation, template hybridization, isothermal amplification, linearization, and blocking denaturing and hybridization of the sequencing primers were performed according to the workflow specified by the service provider. Libraries were constructed with an insert size of approximately 500 bp followed by high-throughput sequencing to obtain paired-end reads with 150 bp in the forward and reverse directions.

[00318] Data quality control. Prinseq was used to: 1) trim the reads from the 3' end until reaching the first nucleotide with a quality threshold of 20; 2) remove read pairs when either read was < 60 bp or contained "N" bases; and 3) de-duplicate the reads. Reads that could be aligned to the human genome (H. sapiens, UCSC hg19) were removed (aligned with Bowtie2 using --reorder --no-hd --no-contain --dovetail).

[00319] De novo assembly, abundance calculation, and taxonomic assignment of genomes. De novo assembly was performed for each sample by using IDBA_UD (--step 20 --mink 20 --maxk 100 --min_contig 500 --pre_correction). The assembled contigs were further binned using MetaBAT (--minContig 1500 --superspecific -B 20). The quality of the bins was assessed using CheckM. Bins had completeness > 95%, contamination < 5% and strain heterogeneity < 5% were retained as high-quality draft genomes. The assembled high-quality draft genomes were further dereplicated by using dRep. DiTASiC, which applied kallisto for pseudo-alignment and a generalized linear model for resolving shared reads among genomes, was used to calculate the abundance of the genomes in each sample, estimated counts with P-value > 0.05 were removed, and all samples were downsized to 36 million reads (One sample with read mapping ratio < 25%, which could not be well represented by the high quality genomes, were removed in downstream analysis). Taxonomic assignment of the genomes was performed by using GTDB-Tk with default parameters.

[00320] Gut microbiome network construction and analysis. In W group, prevalent genomes shared by more than 75% of the samples at every timepoint were used to construct the co-abundance network at each timepoint. Fastspar, a rapid and scalable correlation estimation

tool for microbiome study, was used to calculate the correlations between the genomes with 1,000 permutations at each time point based on the abundances of the genomes across the patients and the correlations with $P \leq 0.001$ were retained for further analysis. The networks were visualized with Cytoscape v3.8.1. The layout of the nodes and edges was determined by Edge-weighted Spring Embedded Layout using the correlation coefficient as weights. The links between the nodes are treated as metal springs attached to the pair of nodes. The correlation coefficient was used to determine the repulsion and attraction of the spring. The layout algorithm sets the position of the nodes to minimize the sum of forces in the network. We defined robust stable edges as the unchanged positive/negative correlations between the same two genomes across all the 3 networks at M0, M3, and M15. Nodes in the stable network were clustered by Connected Components Clustering analysis in Cytoscape. To identify if sub-clusters existed in Cluster C1, a clustering tree based on the negative (set as -1) and positive correlations (set as 1) with average linkage method followed by WGCNA analysis.

[00321] Case-Control Dataset Collection I (CCDC-I). Eleven independent metagenomic datasets on T2DM, LC, AS, ACVD, SCZ, CRC and IBD were downloaded from SRA or ENA database. The group information, i.e. case or control, was collected from the corresponding papers or from curatedMetagenomicData. Quality control of raw reads was conducted by KneadData (<https://bitbucket.org/biobakery/kneaddata>). DiTASiC was used to recruit reads and estimate the abundance of the 141 HQMAGs in QD-CTG in each sample, estimated counts with $P\text{-value} > 0.05$ were removed and further converted to relative abundance divided by the total number of reads. A random forest classification model to classify case and control was constructed based on the estimated abundances of the HQMAGs in each dataset with leave-one-out cross-validation. High quality reads were de novo assembled for each sample using MEGAHIT ((-min-contig-len 500, -presets meta-large). The assembled contigs were binned using MetaBAT 2 and MaxBin 2. The quality of the bins was assessed using CheckM. Bins had completeness $> 95\%$, contamination $< 5\%$ and strain heterogeneity $< 5\%$ were retained as HQMAGs, which were further dereplicated by using dRep (Table 3). In the case and control groups, Fastpar with 1,000 permutations was used to calculate the co-abundance correlations based on the HQMAGs shared by more than 75% of the samples in both groups. Correlations with $P \leq 0.001$ were retained for further analysis Robust stable edges were defined as the

unchanged positive/negative correlations between the same two genomes in both case and control groups. Same clustering analysis performed in the QD trial was conducted on stable network. DiTASiC was used estimate the abundance of HQMAGs in each TCG in each sample. A random forest classification model to classify case and control was constructed with leave-one-out cross-validation to test each TCG.

[00322] Case-Control Dataset Collection II (CCDC-II). Fifteen independent metagenomic datasets on AS, ASD, BD, COVID-19, CRC, GD, HT, MS, PC and PD were downloaded from SRA or ENA database. The group information, i.e., case or control, was collected from the corresponding papers or from curated metagenomic data. Quality control of raw reads was conducted by KneadData. DiTASiC was used to recruit reads and estimate the abundance of HQMAGs in C-TCG and CC-TCG. A random forest classification model to classify case and control was constructed based on the estimated abundances of the HQMAGs in each dataset with leave-one-out cross-validation.

[00323] Treatment Dataset Collection (TDC). Eleven independent metagenomic datasets on pre-treatment samples related to four diseases, including IBD, rheumatoid arthritis (RA), advanced melanoma and B cell lymphoma, were download from SRA or ENA database. The responder and non-responder categories of each sample were collected from the corresponding paper. Quality control of raw reads was conducted by KneadData. DiTASiC was used to recruit reads and estimate the abundance of HQMAGs in CC-TCG. A random forest classification model to predict responder and non-responder was constructed based on the estimated abundances of the HQMAGs in each dataset with leave-one-out cross-validation.

[00324] Gut microbiome functional analysis. Prokka was used to annotate the HQMAGs. KEGG Orthologue (KO) IDs were assigned to the predicted protein sequences in each HQMAG by HMMSEARCH against Kofam using KofamKOALA with default parameters. KOs were further assigned to KEGG modules. Antibiotic resistance genes were predicted using ResFinder with default parameters. The identification of virulence factors was based on the core set of Virulence Factors of Pathogenic Bacteria Database (VFDB). The predicted proteins sequences were aligned to the reference sequence in VFDB using BLASTP (best hist with E-value < 1e-5, identity > 80% and query coverage > 70%). Genes encoding carbohydrate-active enzymes (CAZys) were identified using dbCAN (release 6.0), and the best-hit alignment was retained.

Genes encoding formate-tetrahydrofolate ligase, propionyl-CoA: succinate-CoA transferase, propionate CoA-transferase, 4Hbt, AtoA, AtoD, Buk and But were identified as described previously.

[00325] *Statistical Analysis.*

[00326] Statistical analysis was performed in the R environment (R version3.6.1). Friedman test followed by Nemenyi post-hoc test was used for intra-group comparisons for repeat measurements. Mann-Whitney test (two-sided) was used for comparisons between W and U at the same time point. Pearson Chi-square tests was performed to compare the differences of categorical data between groups or timepoints. PERMANOVA test (9,999 permutations) was used to compare the groups of gut microbiota structure.

[00327] Mann-Whitney test (two-sided) and Fisher's exact test (two sided) were used to compare the target functions between Guild 1 and Guild 2. Hierarchical clustering analysis based on Jaccard distance on the KO profiles was conducted to compare HQMAGs in CC-TCG.

[00328] Linear mixed effect model with subject id as random effect was applied to explore the associates between the abundance of guilds in QD-TCG and clinical parameters. For each HQMAG belonging to a guild, the robust clr-transformed abundance across each sample was first range-scaled. Subsequently, the guild abundance was then obtained as the mean of the range-scaled abundances of HQMAGs belonging to that guild. The M0 and M3 timepoints was used to train the linear mixed effect model, and M15 was used as testing.

[00329] The Random Forest with leave-one-out cross-validation was used to perform classification analysis based on HQMAGs of TCGs in each dataset of CCDC-I, CCDC-II and TDC. The Random Forest with 10-fold cross-validation was used to perform classification analysis based on HQMAGs in CC-TCG for universal model.

[00330] *Example 3 – The combined core genomes in the two competing guilds (CC-TCG) predict immunotherapy outcomes across various independent datasets spanning a diverse range of diseases.*

[00331] There have been reports linking gut microbiota to the efficacy of biological therapies in many diseases. We hypothesized that pre-treatment variations in CC-TCG might be predictive

of clinical success in diseases. To test our hypothesis, we compiled 11 pre-treatment metagenomic datasets related to four diseases, including IBD, rheumatoid arthritis (RA), advanced melanoma and B cell lymphoma, along with their corresponding categories indicating responder and non-responder to therapies. These 11 datasets are referred to as the Treatment Dataset Collection (TDC). We used HQMAGs from the CC-TCG as reference genomes to perform read recruitment analysis on the metagenomic datasets of TDC. On average, we found that 16.22 % and 13.06% of reads were recruited in the responder and non-responder samples respectively. For each dataset in the TDC, we built a Random Forest classifier based on the abundance of the HQMAGs in CC-TCG to predict therapeutic responses (FIG. 8C).

[00332] In the 3 datasets of IBD treatment, 14-week remission was used to classify patients as responder or non-responder to anti-cytokine or anti-integrin therapy. CC-TCG predicted remission at week 14 with AUC of 0.68 for IBD_anti-cytokine, 0.64 for IBD_anti-integrin#1 and 0.69 for IBD_anti-integrin#2 (FIG. 8C and FIG. 10A). |In the dataset of methotrexate treatment in new-onset RA, a responder to methotrexate was defined as any patient with new-onset RA with an improvement in the Disease Activity Score in 28 joints. CC-TCG predicted improvement under methotrexate with AUC of 0.69 (FIG. 8C and FIG. 10B). |The cross-cohort datasets of immune checkpoint inhibitor (ICI) treatment on advanced melanoma, conducted by Lee et al, were collected in TDC. The cohorts were from Barcelona (AM_ICI#1), Leeds (AM_ICI #2), Manchester (AM_ICI #3), PRIMM-UK (AM_ICI #4) and PRIMM-NL (AM_ICI #5). Responders and responders were defined based on overall response rate (ORR) or progression-free survival at 12 months (PFS12). Averagely, CC-TCG predicted ORR and PFS12 with AUC of 0.61 and 0.7 within each cohort (FIG. 8C and FIG. 10C). |The transportability of the prediction model from one single cohort to another was found to be insufficient. However, when pooling training datasets, as done in the LODO analysis, the prediction performance improved, reaching average AUC values of 0.66 and 0.67 for ORR and FPS12, respectively. |In the datasets of CD19-CAR-T immunotherapy on B cell lymphoma, responses to therapy were classified as either complete remission or non-complete remission at 180 days after CAR-T cell infusion. We trained a prediction model in the cohort from Germany (BCL_CD19-CAR-T#1) and validated it in the cohort from the United States (BCL_CD19-CAR-T#2). CC-TCG predicted responses to CD19-CAR-T immunotherapy with an AUC of 0.66 in the German cohort, and the model was sufficiently transportable to predict for the US cohort with an AUC of 0.64. These results

confirm the associations between pre-treatment gut microbiome and therapeutic effect and highlight the potential of using CC-TCG to predict treatment effects across various diseases.

[00333] Materials and Methods

[00334] Gut microbiome analysis

[00335] Metagenomic sequencing. DNA was extracted from fecal samples using the methods as previously described. Metagenomic sequencing was performed using Illumina Hiseq 3000 at GENEWIZ Co. (Beijing, China). Cluster generation, template hybridization, isothermal amplification, linearization, and blocking denaturing and hybridization of the sequencing primers were performed according to the workflow specified by the service provider. Libraries were constructed with an insert size of approximately 500 bp followed by high-throughput sequencing to obtain paired-end reads with 150 bp in the forward and reverse directions.

[00336] Data quality control. Prinseq was used to: 1) trim the reads from the 3' end until reaching the first nucleotide with a quality threshold of 20; 2) remove read pairs when either read was < 60 bp or contained "N" bases; and 3) de-duplicate the reads. Reads that could be aligned to the human genome (H. sapiens, UCSC hg19) were removed (aligned with Bowtie2 using --reorder --no-hd --no-contain --dovetail).

[00337] De novo assembly, abundance calculation, and taxonomic assignment of genomes. De novo assembly was performed for each sample by using IDBA_UD (--step 20 --mink 20 --maxk 100 --min_contig 500 --pre_correction). The assembled contigs were further binned using MetaBAT (--minContig 1500 --superspecific -B 20). The quality of the bins was assessed using CheckM. Bins had completeness > 95%, contamination < 5% and strain heterogeneity < 5% were retained as high-quality draft genomes. The assembled high-quality draft genomes were further dereplicated by using dRep. DiTASiC, which applied kallisto for pseudo-alignment and a generalized linear model for resolving shared reads among genomes, was used to calculate the abundance of the genomes in each sample, estimated counts with P-value > 0.05 were removed, and all samples were downsized to 36 million reads (One sample with read mapping ratio < 25%, which could not be well represented by the high quality genomes, were removed in downstream analysis). Taxonomic assignment of the genomes was performed by using GTDB-Tk with default parameters.

[00338] Gut microbiome network construction and analysis. In W group, prevalent genomes shared by more than 75% of the samples at every timepoint were used to construct the co-abundance network at each timepoint. Fastspar, a rapid and scalable correlation estimation tool for microbiome study, was used to calculate the correlations between the genomes with 1,000 permutations at each time point based on the abundances of the genomes across the patients and the correlations with $P \leq 0.001$ were retained for further analysis. The networks were visualized with Cytoscape v3.8.1. The layout of the nodes and edges was determined by Edge-weighted Spring Embedded Layout using the correlation coefficient as weights. The links between the nodes are treated as metal springs attached to the pair of nodes. The correlation coefficient was used to determine the repulsion and attraction of the spring. The layout algorithm sets the position of the nodes to minimize the sum of forces in the network. We defined robust stable edges as the unchanged positive/negative correlations between the same two genomes across all the 3 networks at M0, M3, and M15. Nodes in the stable network were clustered by Connected Components Clustering analysis in Cytoscape. To identify if sub-clusters existed in Cluster C1, a clustering tree based on the negative (set as -1) and positive correlations (set as 1) with average linkage method followed by WGCNA analysis.

[00339] Case-Control Dataset Collection I (CCDC-I). Eleven independent metagenomic datasets on T2DM, LC, AS, ACVD, SCZ, CRC and IBD were downloaded from SRA or ENA database. The group information, i.e. case or control, was collected from the corresponding papers or from curatedMetagenomicData. Quality control of raw reads was conducted by KneadData. DiTASiC was used to recruit reads and estimate the abundance of the 141 HQMAGs in QD-CTG in each sample, estimated counts with P-value > 0.05 were removed and further converted to relative abundance divided by the total number of reads. A random forest classification model to classify case and control was constructed based on the estimated abundances of the HQMAGs in each dataset with leave-one-out cross-validation. High quality reads were de novo assembled for each sample using MEGAHIT ((-min-contig-len 500, -presets meta-large). The assembled contigs were binned using MetaBAT 2 and MaxBin 2. The quality of the bins was assessed using CheckM. Bins had completeness > 95%, contamination < 5% and strain heterogeneity < 5% were retained as HQMAGs, which were further dereplicated by using dRep (Table 3). In the case and control groups, Fastspar with 1,000 permutations was used to calculate the co-abundance correlations based on the HQMAGs shared by more than 75% of the

samples in both groups. Correlations with $P \leq 0.001$ were retained for further analysis. Robust stable edges were defined as the unchanged positive/negative correlations between the same two genomes in both case and control groups. Same clustering analysis performed in the QD trial was conducted on stable network. DiTASiC was used to estimate the abundance of HQMAGs in each TCG in each sample. A random forest classification model to classify case and control was constructed with leave-one-out cross-validation to test each TCG.

[00340] Case-Control Dataset Collection II (CCDC-II). Fifteen independent metagenomic datasets on AS, ASD, BD, COVID-19, CRC, GD, HT, MS, PC and PD were downloaded from SRA or ENA database (Table 4). The group information, i.e. case or control, was collected from the corresponding papers or from curated MetagenomicData. Quality control of raw reads was conducted by KneadData. DiTASiC was used to recruit reads and estimate the abundance of HQMAGs in C-TCG and CC-TCG. A random forest classification model to classify case and control was constructed based on the estimated abundances of the HQMAGs in each dataset with leave-one-out cross-validation.

[00341] Treatment Dataset Collection (TDC). Eleven independent metagenomic datasets on pre-treatment samples related to four diseases, including IBD, rheumatoid arthritis (RA), advanced melanoma and B cell lymphoma, were downloaded from SRA or ENA database. The responder and non-responder categories of each sample were collected from the corresponding paper. Quality control of raw reads was conducted by KneadData. DiTASiC was used to recruit reads and estimate the abundance of HQMAGs in CC-TCG. A random forest classification model to predict responder and non-responder was constructed based on the estimated abundances of the HQMAGs in each dataset with leave-one-out cross-validation.

[00342] Gut microbiome functional analysis. Prokka was used to annotate the HQMAGs. KEGG Orthologue (KO) IDs were assigned to the predicted protein sequences in each HQMAG by HMMSEARCH against Kofam using KofamKOALA with default parameters. KOs were further assigned to KEGG modules. Antibiotic resistance genes were predicted using ResFinder with default parameters. The identification of virulence factors was based on the core set of Virulence Factors of Pathogenic Bacteria Database (VFDB). The predicted protein sequences were aligned to the reference sequence in VFDB using BLASTP (best hit with E-value $< 1e-5$, identity $> 80\%$ and query coverage $> 70\%$). Genes encoding carbohydrate-active enzymes

(CAZys) were identified using dbCAN (release 6.0), and the best-hit alignment was retained. Genes encoding formate-tetrahydrofolate ligase, propionyl-CoA: succinate-CoA transferase, propionate CoA-transferase, 4Hbt, AtoA, AtoD, Buk and But were identified as described previously.

[00343] *Statistical Analysis.*

[00344] Statistical analysis was performed in the R environment (R version 3.6.1). Friedman test followed by Nemenyi post-hoc test was used for intra-group comparisons for repeat measurements. Mann-Whitney test (two-sided) was used for comparisons between W and U at the same time point. Pearson Chi-square tests were performed to compare the differences of categorical data between groups or timepoints. PERMANOVA test (9,999 permutations) was used to compare the groups of gut microbiota structure.

[00345] Mann-Whitney test (two-sided) and Fisher's exact test (two-sided) were used to compare the target functions between Guild 1 and Guild 2. Hierarchical clustering analysis based on Jaccard distance on the KO profiles was conducted to compare HQMAGs in CC-TCG.

[00346] Linear mixed effect model with subject id as random effect was applied to explore the associations between the abundance of guilds in QD-TCG and clinical parameters. For each HQMAG belonging to a guild, the robust clr-transformed abundance across each sample was first range-scaled. Subsequently, the guild abundance was then obtained as the mean of the range-scaled abundances of HQMAGs belonging to that guild. The M0 and M3 timepoints were used to train the linear mixed effect model, and M15 was used as testing.

[00347] The Random Forest with leave-one-out cross-validation was used to perform classification analysis based on HQMAGs of TCGs in each dataset of CCDC-I, CCDC-II and TDC. The Random Forest with 10-fold cross-validation was used to perform classification analysis based on HQMAGs in CC-TCG for universal model.

[00348] *Example 4 – A universal model based on the combined core genomes of the two competing guilds distinguish cases from controls across diseases.*

[00349] We demonstrated the efficacy of CC-TCG in distinguishing cases from controls and predicting therapeutic responses. This is substantiated by the moderate to outstanding Random Forest models built for each dataset in CCDC-I, CCDC-II, and TDC. Further validation of the

robustness and transportability of CC-TCG trained models was achieved through cross-dataset and LODO analysis for diseases with multiple datasets.

[00350] Our findings prompted us to probe whether a universal model based on CC-TCG could discern cases from controls, regardless of the disease types. To this end, we amalgamated all the 26 datasets from CCDC-I and CCDC-II, encompassing 1,780 cases spanning 15 disease types and 1,604 controls. The HQMAGs in CC-TCG were highly prevalent, with 210 of them present in over 80%, and 101 in over 90% of the samples .

[00351] Consistent trends of CC-TCG were observed across 20 of the 26 datasets, where control samples exhibited greater diversity in CC-TCG and a higher C1A-to-C1B ratio compared to case samples. We allocated all the case and control samples randomly, with 80% serving to train a Random Forest classifier based on CC-TCG, and the remaining 20% set aside for testing (Fig. 11A).

[00352] By employing 10-fold cross-validation, we found that the universal classifier displayed a moderate ability, with an AUC of 0.73, to distinguish cases from controls (Figure 11B). The model generated a probability score, indicating the likelihood of a sample being classified as a case, which exhibited a unimodal distribution in both case and control samples, and clear separation between peaks (Fig. 11C). Probability scores were significantly lower in controls compared to cases (Fig 11D). When applying the universal classifier to the testing data, an AUC of 0.76 was achieved in distinguishing cases from controls (Figure 11B). Parallel to the training data, similar distributions of probability scores for cases and controls were observed in the testing data (Fig 11C). Similarly, in the testing data, controls also had significantly lower probability scores than those of cases. For both the training and testing data, a probability score of 0.52 provided a classification specificity of 0.7 and sensitivity of 0.7 for distinguishing cases from controls. These results indicate the feasibility of using CC-TCG to differentiate cases and controls with a universal model, which is agnostic to disease types. In other words, the variation in CC-TCG may serve as a generalized indicator of healthy recovery and maintenance.

[00353] Discussion

[00354] Our study unveils the existence of a core gut microbiome structure consisting of two adversarial bacterial guilds, manifested as a seesaw-like network in humans. This discovery was facilitated by a unique methodological combination: genome-centric, reference-free, and

interaction-focused approaches. These revealed robust associations between this core microbiome signature and a diverse range of host phenotypes, notably in individuals with Type 2 Diabetes Mellitus (T2DM). Furthermore, our random forest models suggest that these bacterial genomes could help distinguish between cases and controls for multiple diseases and predict immunotherapy outcomes, emphasizing the relevance of this core microbiome signature across different ethnicities, geographical locations, and disease states.

[00355] Several lines of evidence support the characterization of these two interconnected guilds as a core gut microbiome signature. Firstly, their presence is consistent across populations, regardless of ethnicity and geography. Secondly, they demonstrate remarkable temporal stability in terms of their members and interaction patterns. Thirdly, despite comprising approximately 10% of the gut microbiome, they exercise substantial influence over the ecological community due to their highly interconnected and stable roles in the gut ecosystem. Lastly, the guilds' structure could be a product of natural selection over a long co-evolutionary history between the microbiomes and their hosts, potentially modulated by dietary fibers, a direct external energy source for the gut ecosystem.

[00356] Historical dietary trends in humans, favoring high fiber intake until about 150 years ago, bolster this notion. The high dietary fiber potentially provided an evolutionary advantage to the beneficial bacteria in Guild 1, owing to their superior capacity to degrade plant polysaccharides, thereby granting them an upper hand over the pathobionts in Guild 2. It is important to note that this seesaw-like network isn't solely linked to T2DM or entirely reliant on a high fiber diet. It can be detected in other independent metagenomic datasets, shows correlations with a variety of diseases, and could be a fundamental attribute of the human microbiome.

[00357] Guild 1 members, equipped with an exceptional ability to degrade complex plant polysaccharides, produce beneficial metabolites like short-chain fatty acids (SCFAs). These metabolites could potentially restrain the overgrowth of pathobionts in Guild 2, whose unregulated proliferation might harm host health through mechanisms such as inflammation. However, maintaining a certain population of these pathobionts is necessary as they play an essential role in priming our immune system early in life. Guild 1, akin to tall trees acting as foundational species in a dense forest, can be considered the "foundation guild" that structures

and stabilizes a gut environment unfavorable to Guild 2, the pathobiont guild. Therefore, maintaining a delicate balance between the Foundation and Pathobiont Guilds is critical to determine whether the gut microbiome is health-promoting or disease-inducing.

[00358] The identified network is marked by cooperative and competitive interactions. While cooperation improves overall metabolic efficiency, it can induce dependencies leading to destabilization, which is mitigated by infusing competition into the network. Interestingly, the seesaw-like network maintains stability, but the relative abundances of the foundation guild and the pathobiont guild can be swayed, suggesting the role of external energy inputs, specifically dietary fiber. Understanding how external energy input impacts the balance between order and chaos in a complex adaptive system is vital to comprehend the ecological dynamics within the gut microbiome. With the competing guilds acting as the gut ecosystem's backbone, 'order' (stable and predictable interactions) can be achieved with sufficient dietary fiber input, bolstering the foundation guild's dominance over the pathobiont guild. Conversely, 'chaos' (instability and potential system disruption) may arise if the energy input to the gut ecosystem shifts from external dietary fibers to host-produced mucin. These dynamics are perpetually counterbalanced by energy inputs, as evidenced by the identified seesaw-like network.

[00359] Our analysis is founded on a relationship-centric approach, arguing that stable relationships within the microbiome aid in identifying its core components. This approach involves studying interaction and co-occurrence patterns among various microbial members and interpreting these relationships as ecological role indicators. Using this method, the seesaw-like network of the two competing guilds emerged as a robust signature of the gut microbiome, displaying stability across different ethnicities, geographical locations, and disease states. This finding highlights the significance of relationship-based analyses in microbiome research, showcasing the importance of cooperation and competition dynamics within the gut microbiome. It also depicts how interaction dynamics can be modulated by external energy inputs, like dietary fiber, to maintain a balance between order (stability) and chaos (instability) within the microbiome.

[00360] In conclusion, our findings suggest that the seesaw-like network may be a foundational characteristic of the human gut microbiome. We underscore the pivotal role of dietary fibers as the external energy input required to maintain order in the gut microbiota for the

benefit of host health. This characteristic, potentially established through natural selection over a lengthy co-evolutionary history between microbiomes and their hosts under a high-fiber dietary context, emphasizes the intricate interplay of these principles within the human gut microbiome.

[00361] Enhancing our comprehension of gut microbiome dynamics through our relationship-centric approach could guide the development of interventions targeting this core signature. The ultimate goal is to restore and uphold the dominance of the foundation guild over the pathobiont guild, thereby promoting and preserving human health. Our study underlines the importance of acknowledging stable relationships as key indicators of microbial components and their ecological roles, providing a promising framework for future microbiome research. Further investigation is needed to fully comprehend these complex dynamics and to unlock the enormous potential of microbiome-centered therapeutics.

[00362] **Example 5 – *Integrated TCGs predict immunotherapy responses in independent treatment datasets.***

[00363] Previous reports have linked the gut microbiota to the efficacy of biological therapies in various diseases. It was hypothesized that pre-treatment variations in CC-TCG might predict clinical success in these treatments. To test this hypothesis, 11 pre-treatment metagenomic datasets related to four diseases were compiled: IBD, (Lee, J.W.J., et al., Cell Host Microbe 29:1294-1304.e4 (2021) and Ananthakrishnan, A.N., et al., Cell Host Microbe 21:603-10.e3 (2017)) rheumatoid arthritis (RA) (Artacho, A., et al., Arthritis Rheumatol. 73:931-42 (2021)), advanced melanoma (Lee, K.A., et al., Nat. Med. 28, 535-44 (2022)), and B cell lymphoma (Stein-Thoeringer, C.K., et al., Nat. Med. 29:906-16 (2023)) These datasets included responder and non-responder categories and are collectively referred to as the treatment dataset collection (TDC) (Figure 14). Using the 284 HQMAGs from the CC-TCG as reference genomes, read recruitment analysis was performed on the TDC metagenomic datasets. On average, 16.22% and 13.06% of reads were recruited in the responder and non-responder samples, respectively. For each dataset in the TDC, a random forest classifier was built based on the abundance of the 284 HQMAGs in CC-TCG to predict therapeutic responses (Figure 14).

[00364] In the three IBD treatment datasets, CC-TCG predicted remission at week 14 with AUC values of 0.68 for IBD_anti-cytokine, 0.64 for IBD_anti-integrin#1, and 0.69 for IBD_anti-

integrin#2 (Figures 14B and 15A). In the methotrexate treatment dataset for new-onset RA, CC-TCG predicted improvement with an AUC of 0.69 (Figures 14B and 15B).

[00365] The immune checkpoint inhibitor (ICI) treatment datasets for advanced melanoma included multiple cohorts, including Barcelona (AM_ICI#1), Leeds (AM_ICI#2), Manchester (AM_ICI#3), PRIMM-UK (AM_ICI#4), and PRIMM-NL (AM_ICI#5). On average, CC-TCG predicted overall response rate (ORR) and progression-free survival at 12 months (FPS12) with AUC values of 0.61 and 0.7 within each cohort, respectively (Figures 14B and 15C). The transportability of the prediction model from one cohort to another was insufficient. However, when pooling training datasets, as done in the LODO analysis, the prediction performance improved, reaching average AUC values of 0.66 and 0.67 for ORR and FPS12, respectively. In the datasets for CD19-CAR-T immunotherapy on B cell lymphoma, a prediction model was trained against the cohort from Germany (BCL_CD19-CAR-T#1) and validated it against the cohort from the United States (BCL_CD19-CAR-T#2). CC-TCG predicted responses to CD19-CAR-T immunotherapy with an AUC of 0.66 in the German cohort, and the model was sufficiently transportable to predict responses in the US cohort with an AUC of 0.64 (Figures 14A and 15D). These results highlight the potential of using CC-TCG to predict treatment effects across various diseases and therapies.

[00366] Example 6 – *The dichotomy of conserved health-relevant functionalities of the genomes in the TCGs.*

[00367] Using HQMAGs allows annotation and understanding of the health-relevant functionalities of the two guilds in the TCGs. With a 99% genomic ANI cutoff, the HQMAGs from the eight different TCGs found in QD and CCDC-I displayed limited overlap, with 701 of the 788 genomes in C-TCG being specific to just one TCG. At the same time, only 87 were present in more than one TCG. While this revealed the varying genomic composition of the HQMAGs from different TCGs derived from datasets of different studies, we observed similar functional profiles within C1As and C1Bs and the functional dichotomy between the two guilds across the 8 TCGs through non-targeted functional annotation of KEGG Ortholog (KO) in each guild. Pearson correlation analysis showed a strong correlation of KO profiles within C1As ($r = 0.89 \pm 0.071$, mean $\pm$ SD) and within C1Bs ($r = 0.96 \pm 0.023$). A lower correlation was observed between the C1As and C1Bs ($r = 0.76 \pm 0.093$). Principal coordinate analysis and clustering

analysis, performed on Euclidean distance calculated from the KO profiles, further illustrated clear distinctions between the C1As and C1Bs from different TCGs (Figure 16), with a significant PERMANOVA test ($p = 0.002$). Among all the 5,870 identified KOs, 2,473 had significant differences in copy numbers between the two guilds, 354 were higher in C1As, and 2,119 were higher in C1Bs. Thus, though the HQMAGs from different TCGs had few overlaps, the functions of each guild in different TCGs were highly consistent.

[00368] To understand why C1A in QD-TCG was promoted by the fibers and negatively correlated with HbA1c, the carbohydrate-active enzyme (CAZy)-encoding genes and genes encoding key enzymes in short-chain fatty acid (SCFA) production were identified to compare the genetic capacity for carbohydrate utilization between the two guilds. Within QD-TCG, compared with genomes in C1B, those in C1A were enriched with CAZy genes for arabinoxylan ($p < 0.001$), cellulose ($p < 0.001$) and had a lower proportion of CAZy genes for inulin utilization ($p < 0.01$) (Figure 17A). The two guilds were not different in genes for starch, pectin, and mucin utilization. Previous work showed that gut microbiota benefited patients with T2D via acetic and butyric acid production from carbohydrate fermentation. Within QD-TCG, among the terminal genes for the butyrate biosynthetic pathways from both carbohydrates (e.g., *but* and *buk*) and proteins (e.g., *atoA/D* and *4Hbt*), the copy number of *but* was significantly higher in C1A, and there was no difference in the other terminal genes between the two guilds (Figure 17B). More than one-third of the genomes in C1A harbored *but* gene, while less than 5% of the genomes in C1B had this gene (Fisher's exact test $p < 0.001$). Compared with C1B, C1A also trended higher in its genetic capacity for acetate production ($p = 0.06$) but had a lower genetic capacity for propionate production ($p < 0.05$) (Figure 17B). Thus, C1A in QD-TCG had a significantly higher genetic capacity for utilizing complex plant polysaccharides and producing butyrate and a lower genetic capacity for producing propionate. When CAZy and SCFA-related genes were explored in all eight different TCGs, this functional genetic distinction between C1A and C1B was further validated (Figures 17C and 17D). Moreover, CC-TCG also inherited this functional dichotomy (Figures 17E and 17F).

[00369] Next, to investigate the functional basis leading to the potential detrimental effect of C1B on the host, antibiotic resistance genes (ARGs) and virulence factor genes (VFs) were identified in the genomes of each guild. Within QD-TCG, C1A had 8 VF genes involved in 1 VF class and 1 ARG related to 1 antibiotic class, while C1B had 753 VF genes from 11 classes and

25 ARGs related to 7 antibiotic classes. In addition to QD-TCG, across all TCGs from CCDC-I, C1As consistently carried fewer VF genes and ARGs than C1Bs. CC-TCG inherited such a genetically functional difference between the two guilds, as evidenced by its C1A harboring 30 VF genes involved in 6 VF classes and 42 ARGs related to 5 antibiotic classes. At the same time, its C1B had 193 VF genes from 11 VF classes and 122 ARGs related to 7 antibiotic classes. In conclusion, these results show that the two guilds in TCGs have distinct genetic capacities, with C1A potentially beneficial and C1B detrimental.

[00370] Discussion

[00371] The studies described above have uncovered a core microbiome signature marked by a dynamic, seesaw-like interplay between two competing bacterial guilds. This microbiome signature correlates with various host phenotypes, notably in individuals with T2D, emerging from our genome-specific, database-independent, and interaction-focused analytical method. Furthermore, this distinctive microbiome signature underpins our machine learning models, successfully differentiating between cases and controls across various conditions and predicting immunotherapy outcomes, underscoring its broad applicability. The articulation of the TCG model as a core microbiome signature represents a novel insight into our understanding of the microbiome-health nexus.

[00372] Based on the three distinct methodological pillars for our microbiome analysis, namely genome-specific analysis, database independence, and a focus on microbial interactions, this study concentrated on stable interactions within the gut microbiome. Such a stable interaction-focused approach has enhanced the understanding of the fundamental structure of the gut microbiome and devised a method to streamline the complex data inherent to this field. This approach allows maintenance of the integrity of crucial biological patterns while managing the high dimensionality and sparsity challenges of extensive microbiome datasets.

[00373] This research encompassed approximately 4,000 metagenomic samples from 38 studies, covering 15 diseases across three continents. 284 HQMAGs within the CC-TCG were identified from a pool of 8,000 by specifically targeting stably correlated genomes. Intriguingly, the refined predictive prowess of our TCG-informed models, when benchmarked against species-level, entire-genome-list, differential-feature, or high prevalent-feature models, substantiates the merit of the targeted approach. Notably, although high prevalence has been widely used as a

criterion for core microbiome, the results herein indicate that the prevalent gut microbiota members are not equally important to the host. In addition to high prevalence, stable interaction is critical in defining core health-relevant microbiomes. Despite selecting a narrower array of microbial actors, the TCG models maintain, if not enhance, their predictive accuracy, highlighting the TCGs as parsimonious yet potent representatives of the microbiome's functional core.

[00374] In addition, the classification and prediction models based on TCGs, as described herein, showed consistent efficacy across different datasets and conditions, even without batch effect adjustments, demonstrating the robustness of this approach. Using stably correlated genomes has emerged as a powerful strategy to mitigate confounding factors like batch effects and technological discrepancies across studies. These results underscore the potential of stably correlated HQMAGs as a more precise biomarker in disease classification and prediction, setting a new standard for microbiome research.

[00375] The identified TCGs embody a functional dichotomy directly linked to host health, which holds promise but requires cautious interpretation across various datasets. In this binary, the beneficial guild (C1As) comprises genomes rich in CAZy genes essential for the digestion of dietary fiber and genes instrumental in the production of SCFAs, such as butyrate, which play vital roles in nutrition, metabolism, immune function, and overall physiology. By contrast, what might be considered the detrimental counterpart (C1Bs) is characterized by genomes abundant with genes that confer antibiotic resistance and express VFs that may predispose to a pathogenic interaction with the host. It is thought that these safe pathogens or pathobionts might play a pivotal role in the developmental shaping of the immune system in early life. However, the dominance of this guild over the beneficial one could contribute to a pro-inflammatory state, potentially underpinning various chronic conditions and the aging process. This observed functional dichotomy offers an intriguing lens to view and redefine eubiosis and dysbiosis, pointing to a delicate equilibrium within TCGs and their broader health implications.

[00376] Despite the stability of the TCG framework, the balance between these two guilds appears responsive to external influences like diet, as evidenced in the QD trial with individuals with T2D. In the QD trial, modifications in fiber intake were shown to alter the guild dynamics within the TCG framework and selectively foster the beneficial guild. These findings indicate

that the proportional representation of the TCGs is not merely correlative but may be causally connected to metabolic health. The beneficial guild members ferment dietary fibers to extract energy to support their growth. This guild shows greater resilience in the more acidic gut environment caused by SCFA accumulation, a byproduct of their fiber digestion, indicating that dietary fibers serve not just as an energy source but also promote the growth of these beneficial microbes. They may progressively outcompete pathobionts due to the acidified environment, SCFA's antimicrobial properties, and competitive exclusion. The beneficial guild thus may function similarly to foundation species in ecology, forming a supportive microbiome structure that underpins host health while making the gut environment inhibitive to pathobionts. The beneficial guild (C1As) can be designated as the foundation guild, postulating that their role may be analogous to towering trees within a forest ecosystem. While drawing parallels between this foundation guild and ecological concepts offers an alluring analogy, this a working hypothesis that merits further empirical exploration to solidify its role in the human gut microbiome.

[00377] The pervasive presence of the TCG structure across the case-control datasets reinforces the hypothesis that dietary fiber plays a pivotal role in shaping the backbone of the human gut microbiome throughout evolutionary trajectories. Indeed, studies on coprolites, the fossilized feces of ancient humans, have estimated a daily fiber intake of over 100 grams. This substantial fiber intake experienced a dramatic decline only in the past century or so, and during this transition, the rampant emergence of pathobionts and inflammation is associated with nearly all major chronic conditions. This might reflect the dwindling foundation guild within the TCGs, a repercussion of the steep reduction in dietary fiber intake to less than a tenth of what our ancestors consumed. Thus, dietary fiber may act as a constant energy infusion, nurturing an orderly TCG structure predominated by the foundation guild and bolstering host health within the intricate ecosystem of the gut microbiome. A broad spectrum of epidemiological and interventional studies has validated that dietary fibers can potentially deter and alleviate various disease conditions. The TCG model holds promise in unraveling the mechanisms underpinning this phenomenon.

[00378] Members of the TCG model can be viewed as housekeeping microbes in the human gut microbiome, analogous to housekeeping genes in genomes. Just as housekeeping genes are essential for the maintenance of basic cellular functions and are ubiquitously expressed in all cells of an organism, the core members identified in our TCG model represent foundational

components of the gut microbiome that are critical for maintaining its overall stability and functionality. These housekeeping microbes could be involved in fundamental processes such as nutrient cycling, metabolic regulation, and immune modulation, similar to how housekeeping genes are involved in basic cellular processes such as transcription, translation, and replication.

[00379] By focusing on these stable and core interactions, the TCG model aims to identify the most crucial microbial relationships that persist across different conditions and host environments. This approach allows pinpointing of the key players in the gut ecosystem that support the broader microbiome structure and functionality, much like housekeeping genes support the basic infrastructure of the genome. Understanding these core microbial interactions provides valuable insights into the essential functions of the gut microbiome and its role in health and disease. It also offers a robust framework for developing targeted microbiome-based therapies, including a new wave of products like prebiotics, probiotics, synbiotics, postbiotics, microbiome transplantation, and live biotherapeutic products (LBPs) focused on restoring and nurturing the foundation guild. Ultimately, the consistent core structure of TCGs that the research described herein highlights may propel forward the fields of proactive and preventive healthcare, marking a significant stride in microbiome studies and their clinical ramifications.

[00380] Materials and Methods

[00381] Gut microbiome analysis

[00382] Metagenomic sequencing. DNA was extracted from fecal samples using the methods as previously described¹⁰. Metagenomic sequencing was performed using Illumina HiSeq 3000 at GENEWIZ Co. (Beijing, China). Cluster generation, template hybridization, isothermal amplification, linearization, and blocking denaturing and hybridization of the sequencing primers were performed according to the workflow specified by the service provider. Libraries were constructed with an insert size of approximately 500 bp followed by high-throughput sequencing to obtain paired-end reads with 150 bp in the forward and reverse directions.

[00383] Data quality control. Prinseq⁶⁰ was used to: 1) trim the reads from the 3' end until reaching the first nucleotide with a quality threshold of 20; 2) remove read pairs when either read was < 60 bp or contained "N" bases; and 3) de-duplicate the reads. Reads that could be aligned

to the human genome (H. sapiens, UCSC hg19) were removed (aligned with Bowtie2⁶¹ using --reorder --no-hd --no-contain --dovetail).

[00384] De novo assembly, abundance calculation, and taxonomic assignment of genomes. De novo assembly was performed for each sample by using IDBA_UD⁶² (--step 20 --mink 20 --maxk 100 --min_contig 500 --pre_correction). The assembled contigs were further binned using MetaBAT⁶³ (--minContig 1500 --superspecific -B 20). The quality of the bins was assessed using CheckM⁶⁴. Bins had completeness > 95%, contamination < 5% and strain heterogeneity < 5% were retained as high-quality draft genomes. The assembled high-quality draft genomes were further dereplicated by using dRep⁶⁵. DiTASiC⁶⁶, which applied kallisto for pseudo-alignment⁶⁷ and a generalized linear model for resolving shared reads among genomes, was used to calculate the abundance of the genomes in each sample, estimated counts with P-value > 0.05 were removed, and all samples were downsized to 36 million reads (One sample with read mapping ratio < 25%, which could not be well represented by the high quality genomes, were removed in downstream analysis). Taxonomic assignment of the genomes was performed by using GTDB-Tk⁶⁸ with default parameters.

[00385] Gut microbiome network construction and analysis. In W group, prevalent genomes shared by more than 75% of the samples at every timepoint were used to construct the co-abundance network at each timepoint. Fastspar⁷⁴, a rapid and scalable correlation estimation tool for microbiome study, was used to calculate the correlations between the genomes with 1,000 permutations at each time point based on the abundances of the genomes across the patients and the correlations with $P \leq 0.001$ were retained for further analysis. The networks were visualized with Cytoscape v3.8.1⁷⁵. The layout of the nodes and edges was determined by Edge-weighted Spring Embedded Layout using the correlation coefficient as weights. The links between the nodes are treated as metal springs attached to the pair of nodes. The correlation coefficient was used to determine the repulsion and attraction of the spring⁷⁵. The layout algorithm sets the position of the nodes to minimize the sum of forces in the network. We defined robust stable edges as the unchanged positive/negative correlations between the same two genomes across all the 3 networks at M0, M3, and M15. Nodes in the stable network were clustered by Connected Components Clustering analysis in Cytoscape. To identify if sub-

clusters existed in Cluster C1, a clustering tree based on the negative (set as -1) and positive correlations (set as 1) with average linkage method followed by WGCNA analysis.

[00386] Case-Control Dataset Collection I (CCDC-I). Eleven independent metagenomic datasets on T2DM, LC, AS, ACVD, SCZ, CRC and IBD were downloaded from SRA or ENA database. The group information, i.e. case or control, was collected from the corresponding papers or from curatedMetagenomicData. Quality control of raw reads was conducted by KneadData (<https://bitbucket.org/biobakery/kneaddata>). DiTASiC was used to recruit reads and estimate the abundance of the 141 HQMAGs in QD-CTG in each sample, estimated counts with P-value > 0.05 were removed and further converted to relative abundance divided by the total number of reads. A random forest classification model to classify case and control was constructed based on the estimated abundances of the HQMAGs in each dataset with leave-one-out cross-validation. High quality reads were de novo assembled for each sample using MEGAHIT ((-min-contig-len 500, -presets meta-large). The assembled contigs were binned using MetaBAT 2 and MaxBin 2. The quality of the bins was assessed using CheckM. Bins had completeness > 95%, contamination < 5% and strain heterogeneity < 5% were retained as HQMAGs, which were further dereplicated by using dRep (Table 3). In the case and control groups, Fastspar with 1,000 permutations was used to calculate the co-abundance correlations based on the HQMAGs shared by more than 75% of the samples in both groups. Correlations with $P \leq 0.001$ were retained for further analysis. Robust stable edges were defined as the unchanged positive/negative correlations between the same two genomes in both case and control groups. Same clustering analysis performed in the QD trial was conducted on stable network. DiTASiC was used to estimate the abundance of HQMAGs in each TCG in each sample. A random forest classification model to classify case and control was constructed with leave-one-out cross-validation to test each TCG.

[00387] Case-Control Dataset Collection II (CCDC-II). Fifteen independent metagenomic datasets on AS, ASD, BD, COVID-19, CRC, GD, HT, MS, PC and PD were downloaded from SRA or ENA database (Table 4). The group information, i.e. case or control, was collected from the corresponding papers or from curatedMetagenomicData. Quality control of raw reads was conducted by KneadData. DiTASiC was used to recruit reads and estimate the abundance of HQMAGs in C-TCG and CC-TCG. A random forest classification model to classify case and

control was constructed based on the estimated abundances of the HQMAGs in each dataset with leave-one-out cross-validation.

[00388] Treatment Dataset Collection (TDC). Eleven independent metagenomic datasets on pre-treatment samples related to four diseases, including IBD, rheumatoid arthritis (RA), advanced melanoma and B cell lymphoma, were download from SRA or ENA database (Table 5). The responder and non-responder categories of each sample were collected from the corresponding paper. Quality control of raw reads was conducted by KneadData. DiTASiC was used to recruit reads and estimate the abundance of HQMAGs in CC-TCG. A random forest classification model to predict responder and non-responder was constructed based on the estimated abundances of the HQMAGs in each dataset with leave-one-out cross-validation.

[00389] Gut microbiome functional analysis. Prokka⁶⁹ was used to annotate the HQMAGs. KEGG Orthologue (KO) IDs were assigned to the predicted protein sequences in each HQMAG by HMMSEARCH against KOfam using KofamKOALA with default parameters⁷⁰. KOs were further assigned to KEGG modules. Antibiotic resistance genes were predicted using ResFinder⁷¹ with default parameters. The identification of virulence factors was based on the core set of Virulence Factors of Pathogenic Bacteria Database (VFDB⁷²). The predicted proteins sequences were aligned to the reference sequence in VFDB using BLASTP (best hist with E-value < 1e-5, identity > 80% and query coverage > 70%). Genes encoding carbohydrate-active enzymes (CAZys) were identified using dbCAN (release 6.0)⁷³, and the best-hit alignment was retained. Genes encoding formate-tetrahydrofolate ligase, propionyl-CoA: succinate-CoA transferase, propionate CoA-transferase, 4Hbt, AtoA, AtoD, Buk and But were identified as described previously¹¹.

[00390] Statistical Analysis.

[00391] Statistical analysis was performed in the R environment (R version 3.6.1). Friedman test followed by Nemenyi post-hoc test was used for intra-group comparisons for repeat measurements. Mann-Whitney test (two-sided) was used for comparisons between W and U at the same time point. Pearson Chi-square tests was performed to compare the differences of categorical data between groups or timepoints. PERMANOVA test (9,999 permutations) was used to compare the groups of gut microbiota structure.

[00392] Mann-Whitney test (two-sided) and Fisher's exact test (two sided) were used to compare the target functions between Guild 1 and Guild 2. Hierarchical clustering analysis based on Jaccard distance on the KO profiles was conducted to compare HQMAGs in CC-TCG.

[00393] Linear mixed effect model with subject id as random effect was applied to explore the associates between the abundance of guilds in QD-TCG and clinical parameters. For each HQMAG belonging to a guild, the robust clr-transformed abundance across each sample was first range-scaled. Subsequently, the guild abundance was then obtained as the mean of the range-scaled abundances of HQMAGs belonging to that guild. The M0 and M3 timepoints was used to train the linear mixed effect model, and M15 was used as testing.

[00394] The Random Forest with leave-one-out cross-validation was used to perform classification analysis based on HQMAGs of TCGs in each dataset of CCDC-I, CCDC-II and TDC. The Random Forest with 10-fold cross-validation was used to perform classification analysis based on HQMAGs in CC-TCG for universal model.

REFERENCES CITED AND ALTERNATIVE EMBODIMENTS

[00395] All references cited herein are incorporated herein by reference in their entirety and for all purposes to the same extent as if each individual publication or patent or patent application was specifically and individually indicated to be incorporated by reference in its entirety for all purposes.

[00396] The present invention can be implemented as a computer program product that comprises a computer program mechanism embedded in a non-transitory computer readable storage medium. For instance, the computer program product could contain the program modules shown in Figure 1, and/or as described in Figure 2. These program modules can be stored on a CD-ROM, DVD, magnetic disk storage product, USB key, or any other non-transitory computer readable data or program storage product.

[00397] Many modifications and variations of this invention can be made without departing from its spirit and scope, as will be apparent to those skilled in the art. The specific embodiments described herein are offered by way of example only. The embodiments were chosen and described in order to best explain the principles of the invention and its practical applications, to thereby enable others skilled in the art to best utilize the invention and various

embodiments with various modifications as are suited to the particular use contemplated. The invention is to be limited only by the terms of the appended claims, along with the full scope of equivalents to which such claims are entitled.

WHAT IS CLAIMED IS:

1. A method of training a model for predicting a subject's response to a therapy for a disorder, comprising:
at a computer system having one or more processors, and memory storing one or more programs for execution by the one or more processors:

A) obtaining, in electronic form, for each respective training subject in a plurality of training subjects, wherein each respective training subject in the plurality of training subjects has received a therapy for a disorder:

(i) a corresponding plurality of genomic abundance values for the respective training subject at a time prior to receiving the therapy, wherein the corresponding plurality of genomic abundance values comprising, for each respective gut microorganism in a plurality of gut microorganisms, a corresponding value for the abundance of the genome of the respective gut microorganism in a corresponding biological sample from the gut of the respective training subject, and

(ii) an indication of the respective training subject's response to the therapy of the respective training subject;

B) inputting, for each respective training subject in the plurality of training subjects, information about the respective training subject into a model comprising a plurality of parameters, wherein the model applies the plurality of parameters to the information through at least 10,000 computations to obtain a corresponding output for the respective training subject from the model, wherein:

the corresponding output comprises a prediction of the respective training subject's response to the therapy,

the information about the respective training subject comprises the corresponding genomic abundance value for each respective gut microorganism in the plurality of gut microorganisms, and

the plurality of gut microorganisms are selected from Table 1, Table 2, or Figure 13A-13XX; and

C) adjusting the plurality of parameters based on, for each respective training subject in the plurality of training subjects, one or more differences between (i) the corresponding output

from the model and (ii) the corresponding indication of the respective training subject's response to the therapy.

2. The method of claim 1, wherein the obtaining A) comprises, for each respective training subject in the plurality of training subjects:

(i) obtaining, in electronic form, a corresponding plurality of at least 100,000 nucleic acid sequences for genomic DNA from the corresponding biological sample from the gut of the respective training subject; and

(ii) determining, for each respective gut microorganism in the plurality of gut microorganisms, the corresponding value for the abundance of the genome of the respective gut microorganism from the corresponding plurality of at least 100,000 nucleic acid sequences.

3. The method of claim 2, wherein the determining A) (ii) comprises, for each respective training subject in the plurality of training subjects:

assembling, in electronic form, a corresponding plurality of gut microorganism genomes by metagenomic de novo sequence assembly from the corresponding plurality of at least 100,000 nucleic acid sequences, and

calculating, for each respective gut microorganism in the plurality of gut microorganisms, the corresponding value for the abundance of the genome of the respective gut microorganism based on the prevalence of respective nucleic acid sequences, in the plurality of at least 100,000 nucleic acid sequences, used to assemble a respective gut microorganism genome in the plurality of gut microorganism genomes corresponding to the respective gut microorganism.

4. The method of claim 2, wherein the determining A) (ii) comprises, for each respective subject in the plurality of training subjects:

assigning each respective nucleic acid sequence in the corresponding plurality of at least 100,000 sequences to a respective gut microorganism in the plurality of gut microorganisms, thereby generating, for each respective gut microorganism in the plurality of gut microorganism, a corresponding count of respective nucleic acid sequences in the corresponding plurality of nucleic acid sequences assigned to the respective gut microorganism, and

determining, for each respective gut microorganism in the plurality of gut microorganisms, the corresponding genomic abundance value for the respective gut microorganism based on the corresponding count of respective nucleic acid sequences assigned to the respective gut microorganism.

5. The method according to any one of claims 2-4, further comprising sequencing, for each respective subject in the plurality of training subjects, genomic DNA from the corresponding biological sample from the gut of the respective training subject, thereby obtaining the corresponding plurality of at least 100,000 nucleic acid sequences.

6. The method according to any one of claims 1-5, wherein the plurality of gut microorganisms comprises at least 20 gut microorganisms selected from Table 1, Table 2, or Figure 13A-13XX.

7. The method according to any one of claims 1-5, wherein the plurality of gut microorganisms comprises at least 50, at least 100, at least 150, at least 200, or at least 250 gut microorganisms selected from Table 2.

8. The method according to any one of claims 1-5, wherein the plurality of gut microorganisms comprises all of the gut microorganisms from Table 2.

9. The method of any one of claims 1-8, wherein the plurality of gut microorganisms comprises at least 20 microorganisms selected from those microorganisms in Table 1, Table 2, or Figure 13A-13XX having a connectivity of at least 2.

10. The method according to any one of claims 1-9, wherein for each respective subject in the plurality of training subjects, the biological sample from the gut of the respective subject is a fecal sample from the respective training subject.

11. The method according to any one of claims 1-10, wherein the therapy is a biological therapy, an immunotherapy, a chemotherapy, a radiation therapy, a gene therapy, a hormone therapy, a photodynamic therapy, a targeted therapy, small molecules, antibodies,

polynucleotide, natural compound, immune modulator, bone marrow therapy, stem cell therapy, surgery therapy, induction therapy, maintenance therapy, or a combination thereof.

12. The method of claim 1, wherein the disorder is selected from the group consisting of type-2 diabetes, hypertension, schizophrenia, atherosclerotic cardiovascular disease (ACVD), liver cirrhosis (LC), inflammatory bowel diseases (IBD), colorectal cancer (CRC), ankylosing spondylitis (AS), and Parkinson's disease (PD), inflammatory bowel disease (IBD), rheumatoid arthritis (RA), or advanced melanoma and B cell lymphoma.

13. The method of claim 1, wherein the disorder is cancer.

14. The method of any one of claims 1-13, wherein the prediction of the respective training subject's response is a class output of a respective response, in a plurality of possible responses, of the respective training subject.

15. The method of any one of claims 1-13, wherein the prediction of the respective training subject's response is a probability output for the respective training subject's response.

16. The method of any one of claims 1-15, wherein the model is a neural network algorithm, a support vector machine algorithm, a Naive Bayes algorithm, a nearest neighbor algorithm, a boosted trees algorithm, a Random Forest algorithm, a convolutional neural network algorithm, a decision tree algorithm, a regression algorithm, or a clustering algorithm.

17. The method of any one of claims 1-16, wherein the plurality of parameters is at least 1000, at least 10,000, at least 15,000, at least 50,000, at least 100,000, at least 250,000, at least 500,000, or at least 1,000,000 parameters.

18. The method of any one of claims 1-17, wherein the model applies the plurality of parameters to the information through at least 25,000, at least 50,000, at least 100,000, at least 250,000, at least 500,000, or at least 1,000,000 computations to obtain a corresponding output for the respective training subject from the model.

19. A method for predicting a subject's response to a therapy for a disorder, comprising:
at a computer system having one or more processors, and memory storing one or more programs for execution by the one or more processors:

A) obtaining, in electronic form, a plurality of genomic abundance values comprising, for each respective gut microorganism in a plurality of gut microorganisms selected from Table 1, Table 2, or Figure 13A-13XX, a corresponding abundance value for the genome of the respective species of gut bacteria, in the plurality of gut microorganisms, in a biological sample from the subject; and

B) inputting the plurality of genomic abundance values into a model comprising a plurality of parameters, wherein the model applies the plurality of parameters to the plurality of genomic abundance values through at least 10,000 computations to generate as output from the model a prediction of the subject's response to the therapy.

20. The method of claim 19, wherein the obtaining A) comprises:

(i) obtaining, in electronic form, a plurality of at least 100,000 nucleic acid sequences for genomic DNA from the biological sample from the gut of the subject; and

(ii) determining, for each respective gut microorganism in the plurality of gut microorganisms, the corresponding value for the abundance of the genome of the respective gut microorganism from the plurality of at least 100,000 nucleic acid sequences.

21. The method of claim 20, wherein the determining A) (ii) comprises:

assembling, in electronic form, a corresponding plurality of gut microorganism genomes by metagenomic de novo sequence assembly from the plurality of at least 100,000 nucleic acid sequences, and

calculating, for each respective gut microorganism in the plurality of gut microorganisms, the corresponding value for the abundance of the genome of the respective gut microorganism based on the prevalence of respective nucleic acid sequences, in the plurality of at least 100,000 nucleic acid sequences, used to assemble a respective gut microorganism genome in the plurality of gut microorganism genomes corresponding to the respective gut microorganism.

22. The method of claim 20, wherein the determining A) (ii) comprises:

assigning each respective nucleic acid sequence in the plurality of at least 100,000 sequences to a respective gut microorganism in the plurality of gut microorganisms, thereby generating, for each respective gut microorganism in the plurality of gut microorganism, a corresponding count of respective nucleic acid sequences in the plurality of nucleic acid sequences assigned to the respective gut microorganism, and

determining, for each respective gut microorganism in the plurality of gut microorganisms, the corresponding genomic abundance value for the respective gut microorganism based on the corresponding count of respective nucleic acid sequences assigned to the respective gut microorganism.

23. The method according to any one of claims 20-22, further comprising sequencing genomic DNA from the biological sample from the gut of the subject, thereby obtaining the plurality of at least 100,000 nucleic acid sequences.

24. The method according to any one of claims 19-23, the method further comprising treating the subject by:

when the prediction of the subject's response to the therapy satisfies a threshold likelihood that the subject will respond favorably to the therapy, administering the therapy to the subject; and

when the prediction of the subject's response to the therapy does not satisfy the threshold likelihood that the subject will respond favorably to the therapy, administering one or more of the plurality of gut microorganisms to the subject.

25. The method of any one of claims 19-24, wherein the plurality of gut microorganisms comprises at least 20 gut microorganisms selected from Table 1, Table 2, or Figure 13A-13XX.

26. The method according to any one of claims 19-24, wherein the plurality of gut microorganisms comprises at least 50, at least 100, at least 150, at least 200, or at least 250 gut microorganisms selected from Table 2.

27. The method according to any one of claims 19-24, wherein the plurality of gut microorganisms comprises all of the gut microorganisms from Table 2.

28. The method according to any one of claims 19-27, wherein the plurality of gut microorganisms comprises at least 20 microorganisms selected from those microorganisms in Table 1, Table 2, or Figure 13A-13XX having a connectivity of at least 2.

29. The method according to any one of claims 19-28, wherein the biological sample from the gut of the subject is a fecal sample.

30. The method according to any one of claims 19-29, wherein the therapy is a biological therapy, an immunotherapy, a chemotherapy, a radiation therapy, a gene therapy, a hormone therapy, a photodynamic therapy, a targeted therapy, small molecules, antibodies, polynucleotide, natural compound, immune modulator, bone marrow therapy, stem cell therapy, surgery therapy, induction therapy, maintenance therapy, or a combination thereof.

31. The method of claim 19-30, wherein the disorder is selected from the group consisting of type-2 diabetes, hypertension, schizophrenia, atherosclerotic cardiovascular disease (ACVD), liver cirrhosis (LC), inflammatory bowel diseases (IBD), colorectal cancer (CRC), ankylosing spondylitis (AS), and Parkinson's disease (PD), inflammatory bowel disease (IBD), rheumatoid arthritis (RA), advanced melanoma and B cell lymphoma.

32. The method of claim 19-30, wherein the disorder is cancer.

33. The method according to any one of claims 19-29, wherein the disorder is inflammatory bowel disease and the therapy comprises anti-cytokine or anti-integrin treatment.

34. The method according to any one of claims 19-29, wherein the disorder is rheumatoid arthritis and the therapy comprises methotrexate treatment.

35. The method according to any one of claims 19-29, wherein the disorder is B cell lymphoma and the therapy comprises CAR-T cell immunotherapy.

36. The method according to any one of claims 19-29, wherein the disorder is melanoma and the therapy comprises immune checkpoint inhibitor treatment.

37. The method of any one of claims 19-36, wherein the prediction of the subject's response is a class output of a respective response, in a plurality of possible responses, of the respective subject.

38. The method of any one of claims 19-36, wherein the prediction of the subject's response of the subject is a probability output for the respective subject's response.

39. The method of any one of claims 19-38, wherein the model is a neural network algorithm, a support vector machine algorithm, a Naive Bayes algorithm, a nearest neighbor algorithm, a boosted trees algorithm, a Random Forest algorithm, a convolutional neural network algorithm, a decision tree algorithm, a regression algorithm, or a clustering algorithm.

40. The method of any one of claims 19-39, wherein the plurality of parameters is at least 1000, at least 10,000, at least 15,000, at least 50,000, at least 100,000, at least 250,000, at least 500,000, or at least 1,000,000 parameters.

41. The method of any one of claims 19-40, wherein the model applies the plurality of parameters to the information through at least 25,000, at least 50,000, at least 100,000, at least 250,000, at least 500,000, or at least 1,000,000 computations to obtain a corresponding output for the respective subject from the model.

42. A computer system, comprising:
one or more processors; and
a non-transitory computer-readable medium including computer-executable instructions that, when executed by the one or more processors, cause the processors to perform the method according to any one of claims 1-41.

43. A non-transitory computer-readable storage medium having stored thereon program code instructions that, when executed by a processor, cause the processor to perform the method according to any one of claims 1-41.