CLAIMS

1. A method of identifying a set of gut microorganisms, 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 subject in a first plurality of subjects having a first state of a biological characteristic, a 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 biological sample from the gut of the respective subject;

B) obtaining, in electronic form, for each respective subject in a second plurality of subjects having a second state of a biological characteristic, a corresponding plurality of genomic abundance values comprising, for each respective gut microorganism in the plurality of gut microorganisms, a corresponding value for the abundance of the genome of the respective gut microorganism in a biological sample from the gut of the respective subject;

C) computing a first plurality of similarity metrics from the corresponding pluralities of genomic abundance values across the first plurality of subjects, wherein:

the first plurality of similarity metrics comprises a first corresponding similarity metric for each unique pair of gut microorganisms in the plurality of gut microorganisms, and

the first corresponding similarity metric quantifies a similarity between (i) a corresponding first vector formed by the corresponding genomic abundance values of the first microorganism in the unique pair of gut microorganisms across the first plurality of subjects and (ii) a corresponding second vector formed by the corresponding genomic abundance values of the second microorganism in the unique pair of gut microorganisms across the first plurality of subjects;

D) computing a second plurality of similarity metrics using the corresponding genomic abundance values for the second plurality of subjects, wherein:

the second plurality of similarity metrics comprises a second corresponding similarity metric for each unique pair of gut microorganisms in the plurality of gut microorganisms, and

the second corresponding similarity metric quantifies a similarity between (i) a corresponding second vector formed by the corresponding genomic abundance values of the first

microorganism in the unique pair of gut microorganisms across the second plurality of subjects and (ii) a corresponding second vector formed by the corresponding genomic abundance values of the second microorganism in the unique pair of gut microorganisms across the second plurality of subjects;

E) determining a set of unique pairs of gut microorganisms in the plurality of gut microorganisms based on the first plurality of similarity metrics and the second plurality of similarity metrics, wherein, for each respective unique pair of gut microorganisms in the set of unique pairs of gut microorganisms:

the first corresponding similarity metric and the second corresponding similarity metric both indicate a statistically significant positive correlation between the abundance of the first gut microorganism and the abundance of the second gut microorganism in the respective unique pair of gut microorganisms, or

the first corresponding similarity metric and the second corresponding similarity metric both indicate a statistically significant negative correlation between the abundance of the first gut microorganism and the abundance of the second gut microorganism in the respective unique pair of gut microorganisms; and

F) identifying a set of gut microorganisms comprising respective gut microorganisms represented in the set of unique pairs of gut microorganisms.

2. The method according to claim 1, wherein the first corresponding similarity metric and the second similarity metric are both a Pearson correlation coefficient, an intraclass correlation coefficient, or a rank correlation coefficient.

3. A method of training a model for evaluating human health, 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:

(i) a 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) a corresponding state of a biological characteristic 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 an indication of the corresponding state of the biological characteristic of the respective training subject, 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 42 A-42XX; and

C) 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 state of the biological characteristic of the respective training subject.

4. A method for evaluating the health of a subject, 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 at least 20 gut microorganisms selected from Table 1, Table 2, or Figure 42 A-42XX, a corresponding abundance value for the genome of the respective species of gut bacteria, in the plurality of at least 20 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 an indication {e.g., a class output or probability output} of the health of the subject.

5. The method of claim 4, wherein the obtaining A) comprises:

(i) obtaining, in electronic form, a plurality of at least 100,000 nucleic acid sequences {Include support in the spec for minimum numbers, maximum numbers, and ranges of NA 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.

6. The method of claim 5, 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.

7. The method of claim 5, 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.

8. The method according to any one of claims 5-7, 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.

9. The method of any one of claims 4-8, wherein the plurality of gut microorganisms comprises at least 20 microorganisms selected from those microorganisms in Table 1, Table 2, or Figure 42 A-42XX having a connectivity of at least 2.

10. The method according to any one of claims 4-9, wherein the biological sample from the gut of the subject is a fecal sample.

11. The method according to any one of claims 4-10, wherein the indication of the health of the subject is an indication of a biological characteristic, wherein the biological characteristic is a disease or disorder, a therapy administered to the subject, or a diet of the subject.

12. The method of claim 11, wherein the disease or 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).

13. The method of claim 11, wherein the disease or disorder is cancer.

14. The method of any one of claims 4-13, wherein the indication of the health of the subject is a class output of a respective state, in a plurality of possible states, of the health of the subject.

15. The method of any one of claims 4-13, wherein the indication of the health of the subject is a probability output for the corresponding state of the health of the subject.

16. The method of any one of claims 4-15, 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 4-16, 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 4-17, 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 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, 9, 14, or 4-18.

20. 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, 9, 14, or 4-18.