

overlapping taxa, the different environments presented by the guts of other hosts are expected to support different microbiomes. In situations where contributions from multiple sub-environments are combined, such as a water source potentially contaminated by a variety of sources, just identifying the thousands of taxa is a significant challenge to current methods of detection.

[0005] Since the study of microbiomes can offer new insight into origins of environmental change, disease, immunological functions, and physiological functions, improved methods for designing nucleic acids, proteins, or other probes that can recognize specific organisms, or taxa are needed. Similarly, improved methods for data analysis that allow detection and quantification of the members of a microbial community at high confidence levels are also needed.

SUMMARY OF THE INVENTION

[0005a] In one aspect, the present invention refers to a system comprising a plurality of probes capable of determining the presence, absence, relative abundance, and/or quantity of at least 10,000 different Operational Taxon Units (OTUs) in a single assay, the system comprising:

(a) a probe set for each of the at least 10,000 OTUs, each probe set comprising interrogation probes having complementarity to a target polynucleotide specific to a corresponding OTU; and

(b) a computer readable storage medium having stored therein instructions executable by a computer processor to:

(i) calculate a probability that an OTU is present based on hybridization signal intensity of interrogation probes corresponding to the OTU; and

(ii) penalize likelihood that an OTU is present in the sample based on potential cross hybridization of probes from said OTU with sequences from other OTUs.

[0006] Some ~~embodiments~~provide examples disclose a system comprising a plurality of probes capable of determining the presence, absence, relative abundance, and/or quantity of at least 10,000 different OTUs in a single assay. In some ~~embodiments~~examples, the system is configured to produce a biosignature that is indicative of fecal contamination. In some ~~embodiments~~examples, the probes selectively hybridize to one or more highly conserved polynucleotides, which can include 16S rRNA gene, 23S rRNA gene, 5S rRNA gene, 5.8S rRNA gene, 12S rRNA gene, 18S rRNA gene, 28S rRNA gene, gyrB gene, rpoB gene, fusA gene, recA gene, coxI gene, nifH gene, RNA molecules derived therefrom, or a combination thereof. In some ~~embodiments~~examples, the conserved polynucleotides are amplicons. In some ~~embodiments~~examples, the probes can be attached to a substrate. In some ~~embodiments~~examples, the probes can form an array. In some ~~embodiments~~examples, the substrate comprises a bead, microsphere, glass, plastic, or silicon. In some ~~embodiments~~examples, the system is capable of performing sequencing reactions on the same highly conserved region of each of the OTUs. In some ~~embodiments~~examples, the system further comprises one or more species-specific probes. In some ~~embodiments~~examples, each of the OTUs is bacterial, archaeal, or fungal.

[0007] In some ~~embodiments~~examples, the system further comprises a plurality of positive control probes. In some ~~embodiments~~examples, the system further comprises a plurality of negative control probes. In some ~~embodiments~~examples, the negative control probes comprise sequences that are not complementary to sequence found in the highly conserved polynucleotide. In some ~~embodiments~~examples, the positive control probes comprise sequences that are complementary to a polynucleotide selected from SEQ ID NOs:51-100. In some ~~embodiments~~examples, the positive control probes comprise one or more sequences selected from SEQ ID NOs: 51-100.

[0008] In some ~~embodiments~~examples, the system removes data from at least a subset of said interrogation probes before making a final call on the presence, absence, relative abundance, and/or quantity of said OTUs. In some ~~embodiments~~examples, the data is removed based on interrogation probe cross-hybridization potential.

[0009] Some ~~embodiments provide examples disclose~~ a system capable of detecting one or more first nucleic acid sequences comprising 1×10^{-3} or less of the total nucleic acids present in a single assay with a confidence level greater than 95% and sensitivity level greater than 95%, wherein the one or more first nucleic acid sequences and set of remaining target nucleic acids are at least 95% homologous. In some ~~embodiments examples~~, the system is configured to produce a biosignature that is indicative of fecal contamination. In some ~~embodiment examples~~, one or more of the nucleic acid sequences are 16S rRNA gene, 23S rRNA gene, 5S rRNA gene, 5.8S rRNA gene, 12S rRNA gene, 18S rRNA gene, 28S rRNA gene, gyrB gene, rpoB gene, fusA gene, recA gene, coxI gene, nifH gene, RNA molecules derived therefrom, or a combination thereof. In some ~~embodiments examples~~, the nucleic acids comprise amplicons.

[0010] Some ~~embodiments provide examples disclose~~ a system for determining the presence, absence, relative abundance, and/or quantity of a plurality of different OTUs in a single assay, said system comprising a plurality of polynucleotide interrogation probes, a plurality of polynucleotide positive control probes, and a plurality of polynucleotide negative control probes. In some ~~embodiments examples~~, the system is configured to produce a biosignature that is indicative of fecal contamination. In some ~~embodiments examples~~, the system removes data from at least a subset of said interrogation probes before making a final call on the presence, absence, relative abundance, and/or quantity of said microorganisms. In some ~~embodiments examples~~, data is removed based on interrogation probe cross hybridization potential.

[0011] Some ~~embodiments provide examples disclose~~ a system capable of detecting the presence, absence, relative abundance, and/or quantity of more than 10,000 different OTUs of a single domain (e.g. bacterial, archaeal, or fungal) in a single assay with confidence greater than 95%. In some ~~embodiments examples~~, the system is configured to produce a biosignature that is indicative of fecal contamination. In some ~~embodiments examples~~, the system comprises a plurality of probes that selectively hybridize to the same highly conserved region in each of said OTUs. In some ~~embodiments examples~~, the system is capable of performing sequencing reactions on the same highly conserved region of each of said OTUs. In some ~~embodiments examples~~, the system further comprises species-specific probes, wherein the probes do not hybridize to said highly conserved sequence. In some ~~embodiments examples~~, the system comprises 100 species-specific probes.

[0012] Some ~~embodiments provide examples disclose~~ a system for determining the presence, absence, relative abundance, and/or quantity of one or more microorganisms from a sample, said system comprising a plurality of OTUs, wherein the median number of probes per OTU is less than 26. Some ~~embodiments provide examples disclose~~ a system for determining the presence, absence, relative abundance, and/or quantity of one or more microorganisms from a sample, said system comprising a plurality of OTUs, wherein the median number of cross-hybridizations per probe is less than 20. In some ~~embodiments examples~~, the system is configured to produce a biosignature that is indicative of fecal contamination.

[0012a] In another aspect, the present invention refers to a method for selecting probes for assaying a condition in a sample comprising:

- a) applying one or more test samples having said condition to the system as defined above;
- b) applying one or more control samples not having said condition to said system to determine the probability of the presence or absence of said OTUs in said control samples;
- c) determining a pattern of OTU's associated with the test samples that is not associated with the control samples;
- d) identifying probes that selectively detect the OTU's associated with the test sample for use in a low density probe system.

[0013] Some embodiments provide examples disclose a method for determining a condition of a sample comprising: a) contacting said sample with a plurality of different probes; b) determining hybridization

signal strength for each of said probes, wherein said determination establishes a biosignature for said sample; and c) comparing the biosignature of said sample to a biosignature for fecal contamination.

[0013a] In yet another aspect, the present invention refers to a method for determining the probability of the presence or quantity of a unique sequence or microorganism in a sample comprising:

a) contacting said sample with a plurality of different probes comprising probe sets for each of the at least 10,000 Operational Taxon Units (OTUs), each probe set comprising interrogation probes having complementarity to a target polynucleotide specific to a corresponding OTU;

b) determining hybridization signal strength for sample sequences to each of said probes;

c) removing or attenuating from analysis an Operational Taxon Unit (OTU) based on hybridization signal strength data, thereby increasing the confidence in the remaining hybridization signal strength data;

wherein the removing or attenuating is performed by penalizing likelihood that an OTU is present in the sample based on potential cross hybridization of probes from said OTU with sequences from other OTUs.

[0014] In one ~~example~~aspect of the invention, a method is ~~provided~~disclosed for determining the probability of the presence, relative abundance, and/or quantity of a microorganism in a sample comprising a) determining hybridization signal strength distributions of negative control probes that do not specifically hybridize to a highly conserved polynucleotide in the microorganism; b) determining hybridization signal strength distributions of positive control probes; c) determining hybridization signal strengths for a plurality of different interrogation probes, each of which is complementary to a section within the highly conserved polynucleotide; and d) using the hybridization signal strengths of the negative and positive probes to determine the probability that the hybridization signal for the different interrogation probes represents the presence, relative abundance, and/or quantity of the microorganism. In some ~~embodiments~~examples, the hybridization signal strengths of the negative and positive probes are used to normalize or fit the interrogation probes hybridization data. In further ~~embodiments~~examples, the normalization or fitting of interrogation probes hybridization data utilizes A+T content or normal and gamma distributions of the negative and positive control probes. In other ~~embodiments~~examples, the negative control probes and/or the positive control probes comprise perfect match and mismatch probes. In further ~~embodiments~~examples, the normal and gamma distribution of the negative and positive control probes involves calculating a pair difference score for said probes. In other ~~embodiments~~examples, the hybridization signal strengths for the plurality of different interrogation probes are attenuated based on the G+C content of each probe.

[0015] In one ~~example~~ aspect, a method is ~~provided~~ disclosed for determining the probability of the presence or quantity of a unique polynucleotide or microorganism in a sample comprising a) contacting the sample with a plurality of different probes; b) determining hybridization signal strength for sample polynucleotides to each of the probes; c) removing or attenuating from analysis an OTU/taxa from the possible list based on hybridization signal strength data, thereby increasing the confidence level of the remaining hybridization signal strength data. In some ~~embodiments~~ examples, the removing or attenuating is performed only on OTUs having a percentage of probes that pass a certain threshold intensity within such OTU. In some ~~embodiments~~ examples, only OTUs that pass a certain threshold are further analyzed. In still further ~~embodiments~~ examples, the removing or attenuating is performed by penalizing OTUs present in the sample based on potential cross hybridization of probes from the OTU with polynucleotides from other OTUs. In some ~~embodiments~~ examples, the penalization positively correlates with potential for cross hybridization with other OTUs. In other ~~embodiments~~ examples, penalization based on cross hybridization is performed at each level of a phylogenic tree starting with the lowest level. In further ~~embodiments~~ examples, only penalized OTUs scoring above a hybridization signal strength threshold are further analyzed. In still other ~~embodiments~~ examples, only parts of phylogenic tree that include an OTU are analyzed.

[0016] In ~~another example~~ further aspect of the invention, a method is ~~provided~~ disclosed for determining presence or quantity of a plurality of different organisms in a sample comprising determining GC content of each

probe and comparing each probe intensity to a positive control probe intensity and negative probe intensity to determine quantity of said probes.

[0017] In another ~~aspect of the invention~~example, computer executable logic is ~~provided~~disclosed for determining a probability that one or more organisms from a set of different organisms are present in a sample said logic comprising: a) an algorithm for determining likelihood that individual interrogation probe intensities are accurate based on comparison with intensities of negative control probes and positive control probes; b) an algorithm for determining likelihood that an individual OTU is present based on intensities of interrogation probes from said OTU passing a first quantile threshold; and c) an algorithm for penalizing one or more OTUs that have passed the first quantile threshold based on potential for cross-hybridization of probes analyzing said OTUs sequences with sequences from other OTUs.

[0018] In a further ~~aspect~~example, computer executable logic is ~~provided~~disclosed for determining the presence and optionally quantity of one or more microorganisms in a sample comprising: logic for analyzing intensities from a set of probes that selectively binds each of at least 10,000, 20,000, 30,000, 40,000, 50,000, 60,000, 70,000, 80,000, 90,000 or 100,000 highly conserved polynucleotides, and determining the presence of at least 90%, 95%, 97%, or more of all species present in said sample. The determination can be made with at least a 90%, 95%, 98%, 99%, or 99.5% confidence level.

[0019] In another ~~aspect~~example, computer executable logic is ~~provided~~disclosed for determining the presence of one or more microorganisms in a sample comprising: logic for analyzing a set of at least 1000 different interrogation perfect probes, and logic for discarding information from at least 10% of said interrogation perfect probes in the process of making said determination.

[0020] In one ~~aspect~~example, a method is ~~provided~~disclosed for probe selection comprising: a) selecting a set of highly conserved polynucleotides; b) comparing said plurality of polynucleotides against a plurality of standard polynucleotides to identify chimeric sequences; c) removing chimeric sequences identified in the comparison step; and d) selecting probes that are complementary to the remaining polynucleotides. In some ~~embodiments~~examples, at least 500,000 highly conserved polynucleotides are selected. In other ~~embodiments~~examples, a member of the plurality of polynucleotides is considered not a chimeric sequence if it shares greater than 95% similarity with a member of the plurality of standard polynucleotides. In still other ~~embodiments~~examples, the plurality of polynucleotides are compared against themselves to identify chimeric sequences. In other ~~embodiments~~examples, highly conserved polynucleotides comprise sequences from a 16S RNA gene, 23S RNA gene, 5S RNA gene, 5.8S rRNA gene, 12S rRNA gene, 18S rRNA gene, 28S rRNA gene, gyrB gene, rpoB gene, fusA gene, recA gene, coxI gene, nifD gene, or combinations thereof.

[0021] In another ~~aspect~~example, a method of probe selection is ~~provided~~disclosed comprising: a) selecting a plurality of nucleic acid sequences; b) aligning the plurality of nucleic acid sequences with a plurality of standard nucleic acid sequences to identify insertion points in each of the plurality of nucleic acid sequences; c) removing sequences with at least 10, 20, 30, 40, 50, or more insertion points or with

insertions that are at least 100 nucleic acids in length; and d) selecting probes that are complementary to the remaining nucleic acids.

[0022] In a further ~~aspect~~example, a method of probe selection is ~~provided~~disclosed comprising: a) selecting a plurality of nucleic acid sequences; b) filtering the plurality of nucleic acid sequences; c) performing hierarchical clustering on remaining nucleic acid sequences to generate a guide tree; and d) selecting probes that are complementary to each node in said guide tree. In some ~~embedments~~examples, filtering the plurality of nucleic acid sequences comprises removing sequences that are identified to comprise PCR primer artifacts, removing sequences that are identified to comprise insertions, removing sequences that are identified as chimeric, or any combination thereof.

[0023] In one ~~aspect~~example, a method is ~~provided~~disclosed for identifying a microbiome signature indicative of a condition, the method comprising a) comparing the presence and optionally abundance of at least 1,000 different OTUs in a control sample without said condition and a reference sample with said condition; and b) identifying one or more OTUs that associate with said condition. In some ~~embedments~~examples, the condition is an oil spill. In some ~~embedments~~examples, an increase in the similarity in the presence and optionally abundance of said OTUs in said reference sample with respect to said control sample is indicative of remediation of said condition. In some ~~embedments~~examples, changes in the degree of similarity in the presence and optionally abundance of said OTUs in said reference sample with respect to said control sample are provided as a measure of remediation of said condition. In some ~~embedments~~examples, the method further comprises projecting a time to reaching a predetermined level of remediation of said condition.

[0024] In one ~~aspect~~example, a method is ~~provided~~disclosed for selecting probes for assaying a condition in a sample comprising: a) applying one or more test samples having said condition to a detection system that simultaneously assays for the probability of the presence or absence of at least 10,000 OTUs of a single domain, such as bacteria, archaea, fungus, or each known OTU of a single domain; b) applying one or more control samples not having said condition to said detection system to determine the probability of the presence or absence of said OTUs in said control samples; c) determining a pattern of OTUs associated with the test samples that is not associated with the control samples; and d) identifying probes that selectively detect the OTUs associated with the test sample. In some ~~embedments~~examples, one or more of the identified probes are selected for use in a low-density probe system. In some ~~embedments~~examples, the pattern consists of up to 200 different OTUs. In other ~~embedments~~examples, the sample is a water sample and the condition is fecal contamination, toxic alga-bloom contamination, presence of fish pathogens, a point source contamination, a non-point source contamination, or a combination thereof. In some ~~embedments~~examples, a unique biosignature of a type of contamination is used to determine the source of the contamination. In some ~~embedments~~examples, the sample is a human or animal sample. In some ~~embedments~~examples, the sample is obtained from the gut, respiratory system, oral cavity, sinuses, nares, urogenital tract, skin, feces, udders, or a combination thereof. In some ~~embedments~~examples, the condition being characterized (e.g., diagnosed or prognosed) in that sample is Crohn's Disease, irritable bowel syndrome, cancer, rhinitis,

stomach ulcers, colitis, atopy, asthma, neonatal necrotizing enterocolitis, acne, food allergy, Gastroesophageal reflux disease, obesity or periodontal disease. In some embodimentsexamples, the sample is a food, water, soil, or air sample. In some embodimentsexamples, the sample is from a forest, industrial crop, or other plant.

[0025] In one aspectexample, a method is provideddisclosed to identify at least one new indicator species for a condition comprising: a) assaying in a single experiment a control sample without said condition to determine the presence or absence of each OTU of all known bacteria, archaea, or fungi; b) assaying in a single experiment a test sample with said condition to determine the presence or absence of each OTU of all known bacteria, archaea, or fungi; c) comparing results from (a) and (b) to identify at least one microorganism whose abundance changes by a predetermined measure in response to the change in the condition, wherein the identified microorganism species represents said new indicator species for said condition. In some embodimentsexamples, the identified microorganism decreases in abundance in the presence of the condition while in others the identified microorganism increases in abundance. In some embodimentsexamples, the predetermined measure is at least a 2-fold change in abundance. In some embodimentsexamples, the predetermined measure is a statistically significant change in abundance.

[0026] In another aspectexample, a system is provideddisclosed for determining the probability that a microorganism or a select group of microorganisms are present in a sample, the system comprising two or more probes identified by the disclosed algorithms. In some embodimentsexamples, the system determines the probability with a confidence level greater than 95%, 99% or 99.5%. In other embodimentsexamples, the determination is performed simultaneously or using a single assay.

[0027] In one aspectexample, a system is provideddisclosed that is capable in a single assay of distinguishing between two OTUs on a phylogenetic tree with an accuracy/confidence of greater than or equal to 95%, 99% or 99.5% based on the selective hybridization of a plurality of probes to highly conserved nucleic acids isolated from each organism to be distinguished.

[0028] In another aspectexample, a system is provideddisclosed that is capable of generating a microbiome signature comprising at least 10,000 OTUs from an environment in a single assay with an accuracy and/or confidence level greater than 95%. In some embodimentsexamples the probes selectively hybridize to nucleic acids from each organism being detected.

[0029] In one aspectexample a method is provideddisclosed for detecting a source of microorganism contamination, the method comprising in a single assay, determining the present and quantity of at least 20, 50, 100, or more microorganism OTUs not naturally occurring in said sample and identifying the source of the contamination using a pattern of the presence and quantity of the OTUs.

[0030] In another aspectexample, a system is provideddisclosed that is capable of detecting the presence and quantity of at least 50 different fecal taxa in a single assay. In some embodimentsexamples, the detection is based on the selective hybridization of a plurality of probes to highly conserved nucleic acids isolated from each organism to be detected. In some embodimentsexamples, detection is based on the selective hybridization of

a plurality of probes that identify the organisms or taxa listed in Table 4. In some ~~embodiment~~examples, detection comprises detecting hybridization of one or more probes that selectively hybridize to nucleic acids indicative of clean water taxa, wherein said probes are selected from that a plurality of probes that identify the organisms or taxa listed in Table 11.

[0031] In a further ~~aspect~~example, a method is ~~provided~~disclosed for detecting fecal contamination in water comprising: detecting the presence or absence in the water sample of one or more polynucleotides which detect the taxa listed in Table 4. In some ~~embodiment~~examples, the method further comprises detecting hybridization of one or more probes that selectively hybridize to polynucleotides indicative of clean water taxa listed in Table 11.

[0032] In another ~~aspect~~example, a method is ~~provided~~disclosed for testing a water sample, the method comprising calculating a ratio of Bacilli, Bacteroidetes and Clostridia (BBC) species and α -proteobacteria (A) species in said water, wherein a value greater than 1.0 is indicative of fecal contamination. In some ~~embodiment~~examples, calculating the ratio does not rely on culturing, directly counting, PCR cloning, sequencing or use of a gene expression array. In some ~~embodiment~~examples, the Bacilli, Bacteroidetes, Clostridia and α -proteobacteria species comprise the species listed in Table 4. In some ~~embodiment~~examples, calculating the ratio of BBC species to A species comprises contacting the water sample with a plurality of probes. In some ~~embodiment~~examples, the plurality of probes are complimentary to a highly conserved gene.

[0033] In a further ~~aspect~~example, a method is ~~provided~~disclosed for predicting the likelihood of a toxic alga bloom, the method comprising: a) contacting a water sample with a plurality of probes that selectively bind to nucleic acids derived from cyanobacteria selected from Table 6; b) using hybridization data derived to determine the quantity and composition of cyanobacteria in the water sample; c) measuring environmental conditions; and d) predicting the likelihood of a toxic alga bloom based on cyanobacteria quantity and composition and environmental conditions. In some ~~embodiment~~examples, the probes to cyanobacteria nucleic acids are selected using the present methods and detect the genera listed in Table 6. In some ~~embodiment~~examples, the environmental conditions comprise water temperature, turbidity, nitrogen concentration, oxygen concentration, carbon concentration, phosphate concentration and/or sunlight level. In further ~~embodiment~~examples, a water management decision is made based on the likelihood of a toxic alga bloom.

[0034] In one ~~aspect~~example, a method is ~~provided~~disclosed for determining a condition of a subject or a therapy for a subject, the method comprising performing a single nucleic acid assay on a sample from said subject to determine the presence and/or amount of at least 1000 OTUs.

[0035] In another ~~aspect~~example, a method is ~~provided~~disclosed for predicting a condition of a sample, the method comprising a) determining microorganism population data as the probability of the presence or absence of at least 1,000 OTUs of microorganisms in the sample; b) determining gene expression data of one or more genes of said microorganisms in the sample; and c) using the expression data and population data to predict the condition of the sample. In some ~~embodiment~~examples, the sample is a water or soil sample.

[0036] In another aspect, ~~the invention provides example, there is disclosed~~ a method for assessing damage caused by an oil spill. In some ~~embodiment examples~~, the method comprises (a) determining the presence, absence, and/or abundance of at least 1,000 OTUs in one or more samples from one or more locations unaffected by said oil spill, thereby establishing an unaffected biosignature; (b) determining the presence, absence, and/or abundance of at least 1,000 OTUs in one or more samples from a location affected by said oil spill, thereby establishing an oil-spill-affected biosignature; and (c) comparing said unaffected biosignature to said oil-spill-affected biosignature, wherein differences in said biosignatures are indicative of affects on the microbiome of said location affected by said oil spill. In some ~~embodiment examples~~, step (b) is performed at a first time and a second time. In some ~~embodiment examples~~, a change in said differences in said biosignatures between said first time and said second time are used to track the progress of remediation of oil spill damage.

INCORPORATION BY REFERENCE

[0037] All publications, patents, and patent applications mentioned in this specification are herein incorporated by reference to the same extent as if each individual publication, patent, or patent application was specifically and individually indicated to be incorporated by reference.

BRIEF DESCRIPTION OF THE DRAWINGS

[0038] The novel features of the invention are set forth with particularity in the appended claims. A better understanding of the features and advantages of the present invention will be obtained by reference to the following detailed description that sets forth illustrative embodiments, in which the principles of the invention are utilized, and the accompanying drawings of which:

[0039] Figure 1 illustrates an example of a suitable computer system environment.

[0040] Figure 2 illustrates a networked system for the remote acquisition or analysis of data obtained through a method of the invention.

[0041] Figure 3 illustrates a flow chart of the probe selection process.

[0042] Figures 4A-B demonstrate the distribution of observed pair difference score, d , from quantitative standards (QS) probes and negative controls (NC) probes.

[0043] Figure 5 is a graph showing variations of gamma scale across 79 arrays.

[0044] Figure 6 illustrates the pre-partition process for computational load balancing.

[0045] Figure 7 is a vector plot comparing the microbial community composition in polluted water samples compared to three potential pollution sources: sewage, septage and cattle waste to determine the source of the pollution.

[0046] Figure 8 is a logarithmic bar graph showing the number of OTUs detected (y-axis) by the PhyloChip for each pooled clean room sample (x-axis). The number of spores detected by the spore count are shown.