

(19) **DANMARK**



Patent- og
Varemærkestyrelsen

(12)

Oversættelse af europæisk patentskrift

(10) **DK/EP 2278030 T3**

-
- (51) Int.Cl.: **C 12 Q 1/68 (2006.01)**
- (45) Oversættelsen bekendtgjort den: **2017-07-24**
- (80) Dato for Den Europæiske Patentmyndigheds bekendtgørelse om meddelelse af patentet: **2017-05-24**
- (86) Europæisk ansøgning nr.: **10012384.3**
- (86) Europæisk indleveringsdag: **2001-04-10**
- (87) Den europæiske ansøgnings publiceringsdag: **2011-01-26**
- (30) Prioritet: **2000-04-10 US 196063 P** **2000-04-11 US 196258 P**
- (62) Stamansøgningsnr: **10152284.5**
- (84) Designerede stater: **AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE TR**
- (73) Patenthaver: **Taxon Biosciences, Inc., 3150 Paradise Drive, Tiburon, CA 94920, USA**
- (72) Opfinder: **Ashby, Matthew, 91, Longfellow Road, Mill Valley, CA 94941, USA**
- (74) Fuldmægtig i Danmark: **Zacco Denmark A/S, Arne Jacobsens Allé 15, 2300 København S, Danmark**
- (54) Benævnelse: **FREMGANGSMÅDER TIL UNDERSØGELSE OG GENETISK ANALYSE AF POPULATIONER**
- (56) Fremdragne publikationer:
VOORDOUW GERRIT ET AL: "Characterization of 16S rRNA genes from oil field microbial communities indicates the presence of a variety of sulfate-reducing, fermentative, and sulfide-oxidizing bacteria", APPLIED AND ENVIRONMENTAL MICROBIOLOGY, AMERICAN SOCIETY FOR MICROBIOLOGY, US, vol. 62, no. 5, 1 May 1996 (1996-05-01), pages 1623-1629, XP002209959, ISSN: 0099-2240
LARSEN N ET AL: "The ribosomal database project.", NUCLEIC ACIDS RESEARCH 1 JUL 1993 LNKD-PUBMED:8332524, vol. 21, no. 13, 1 July 1993 (1993-07-01), pages 3021-3023, XP002616478, ISSN: 0305-1048
VON WINTZINGERODE F ET AL: "Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis.", FEMS MICROBIOLOGY REVIEWS NOV 1997 LNKD- PUBMED:9451814, vol. 21, no. 3, November 1997 (1997-11), pages 213-229, XP002587456, ISSN: 0168-6445
FELSKE A ET AL: "Phylogeny of the main bacterial 16S rRNA sequences in Drentse A grassland soils (The Netherlands).", APPLIED AND ENVIRONMENTAL MICROBIOLOGY MAR 1998 LNKD- PUBMED:9501427, vol. 64, no. 3, March 1998 (1998-03), pages 871-879, XP002587457, ISSN: 0099-2240
WANG G C ET AL: "Frequency of formation of chimeric molecules as a consequence of PCR coamplification of 16S rRNA genes from mixed bacterial genomes.", APPLIED AND ENVIRONMENTAL MICROBIOLOGY DEC 1997 LNKD- PUBMED:9406382, vol. 63, no. 12, December 1997 (1997-12), pages 4645-4650, XP002587458, ISSN: 0099-2240
VELCULESCU V E ET AL: "SERIAL ANALYSIS OF GENE EXPRESSION", SCIENCE, AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE, WASHINGTON, DC; US LNKD- DOI:10.1126/SCIENCE.270.5235.484, vol. 270, no. 5235, 20 October 1995 (1995-10-20), pages 484-487, XP001024449, ISSN: 0036-8075

DESCRIPTION

TECHNICAL FIELD OF THE INVENTION

[0001] The present invention relates to methods for performing surveys of the genetic diversity of a population. The invention also relates to methods for performing genetic analyses of a population. The invention further relates to methods for the creation of databases comprising the survey information and the databases created by these methods. The invention also relates to methods for analyzing the information to correlate the presence of nucleic acid markers with desired parameters in a sample. These methods have application in the fields of geochemical exploration, agriculture, bioremediation, environmental analysis, clinical microbiology, forensic science and medicine.

BACKGROUND OF THE INVENTION

[0002] Microbes have been used previously as biosensors to identify chemicals in the environment. For instance, microbes have been utilized as biosensors for the presence of nitrates (Larsen, L. H. et al, 1997, A microscale NO₃ biosensor for environmental applications. Anal. Chem. 69:3527-3531), metals (Virta M. et al, 1998, Bioluminescence-based metal detectors. Methods Mol. Biol. 102:219-229), and a variety of hydrocarbons (Sticher P. et al., 1997, Development and characterization of a whole-cell bioluminescent sensor for bioavailable middle-chain alkanes in contaminated groundwater samples. Appl. Environ. Microbial. 63(10):4053-4060). In these examples however, the indicator microbes are not native species, but rather, the product of recombinant manipulations designed for specific applications. These modifications involve coupling the nutrient sensing machinery of well-characterized bacterial strains with reporter genes to aid identification. This approach is limited, however, by the metabolic diversity of a few well-characterized bacterial strains. In contrast, the large and diverse pool of microbes in the environment represents a source of biosensors for a much larger range of applications than currently exists. Thus, there is a need to identify and use other microbes, especially those found *in situ*, as biosensors.

[0003] Microbes also have an important impact on health and medicine. Estimates have been made there may be ten times the number of microbial cells associated with the human body as there are human cells. Many microbial cell populations that are associated with the human body play a beneficial role in maintaining health. For instance, gut microflora is important for proper digestion and absorption of nutrients and for production of certain factors, including some vitamins. In general, the human immune system is able to keep the bacterial populations of the human body in check and prevent the overgrowth of beneficial microbial populations and infection by detrimental microbial populations. Nevertheless, the list of human diseases that are now attributed to microbial pathogens is growing. However, nearly all of the information regarding the relationships between microbes and human disease have been gained from

approaches that require culture of microbial species.

[0004] Two examples of diseases where the causative agents were identified through molecular methods include bacillary angiomatosis (Relman, D.A. et al., 1990, New Engl. J. Med. 323: 1573) and Whipple's disease (Wilson, K.H. et al., 1991, Lancet 338: 474). Further, the central aspects of atherosclerosis are consistent with the inflammation that results from infection. DNA sequences from *Chlamydia* have been identified from atherosclerotic lesions and has led to suggestions that this organism plays a role in the disease.

[0005] In addition, bacterial infections have become an increasing health problem because of the advent of antibiotic-resistant strains of bacteria. Further, microbial infections caused by bacteria or fungi that do not usually infect humans may be a problem in immunocompromised individuals. Further, individuals in developing countries who may be malnourished or lack adequate sanitary facilities may also support a large load of opportunistic bacteria, many of which may cause sickness and disease. In veterinary medicine, livestock living in close quarters also may be prey to infections caused by a variety of different types of microbes. Thus, there is a need to develop sensitive methods of identifying many different types of microbes without having to cultivate them first in order to treat or prevent microbial infections in humans and other animals.

[0006] Assays for microbial contamination is an important component of food testing as well. A large number of different types of microbes may contaminate food for humans or animals. Thus, an ability to test food for contamination quickly and effectively is critical for maintaining food safety. However, many of the microbes responsible for causing sickness in humans and animals are difficult to isolate or identify.

[0007] Assays for microbial populations also has use in fields such as forensic science. Over the past ten to twenty years, scientists have determined that microbial populations change when bodies begin to decay, and have begun to identify certain microbial species that are indicative of decomposition (Lawrence Osborne, Crime-Scene Forensics; Dead Men Talking, New York Times, December 3, 2000). However, only a few microbial species that may be useful in these analyses has been identified.

[0008] The problem of determining genetic diversity is not confined to microbial populations. Antibody diversity is critical for a proper immune response. During B cell differentiation, antibody diversity is generated in the heavy and light chains of the immunoglobulin by mechanisms including multiple germ line variable (V) genes, recombination of V gene segments with joining (J) gene segments (V-J recombination) and recombination of V gene segments with D gene segments and J gene segments (V-D-J recombination) as well as recombinational inaccuracies. Furthermore, somatic point mutations that occur during the lifetime of the individual also lead to antibody diversity. Thus, a huge number of different antibody genes coding for antibodies with exquisite specificity can be generated. T cell receptor (TCR) diversity is generated in a similar fashion through recombination between numerous V, D and J segments and recombinational inaccuracies. It has been estimated that

10^{14} γ chains, more than 10^{13} β chains and more than 10^{12} forms of α chains can be made (Roitt, I. et al., Immunology, 3rd Ed., 1993, pages 5.1-5.14). A knowledge of the antibody or TCR diversity in a particular individual would be useful for diagnosis of disease, such as autoimmune disease, or for potential treatment.

[0009] The identification of microbes, especially soil microbes, has traditionally relied upon culture-dependent methods, whereby the detection of a microbial species depends upon the ability to find laboratory conditions that support its growth. To this end, 96-well plates have been commercially developed to identify microbes with different metabolic requirements. For instance, BioLog plates incorporate 96 different media formulations into the wells of a 96-well plate. Despite these efforts, it is now accepted that far fewer than 1% of microbes can propagate under laboratory conditions (Amann, R. I. et al., 1995. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol. Rev. 59:143-169).

[0010] The widespread interest in genomics has created many exciting new technologies for the parallel quantitation of thousands of distinct nucleic acid sequences simultaneously. While still in their infancy, these technologies have provided unprecedented insight into biology. To date, these technologies have predominately been utilized in pharmaceutical and agricultural applications. Genome expression profiling has gained general acceptance in biology and is likely to become commonplace in all academic, biotechnology and pharmaceutical institutions in the 21st century. For instance, Serial Analysis of Gene Expression (SAGE) is a hybridization-independent method designed to quantitate changes in gene expression (Velculescu, V. E. et al., 1995, Serial analysis of gene expression. Science 270:484-487 and United States Patent 5,866,330). However, SAGE only measures RNA levels from tissues or organisms, and is not suitable for examining genetic diversity.

[0011] The widespread interest in genomics has also led to the development of many technologies for the rapid analysis of tens of thousands of nucleic acid sequences. One such technology is the DNA chip. Although this approach had been used as a diagnostic for distinguishing between several species of the genus *Mycobacterium* (Troesch, A., et al., 1999, *Mycobacterium* species identification and rifampin resistance testing with high-density DNA probe arrays. J. Clin. Microbiol. 37:49-55), it has limited utility for an environmental microbial survey for two reasons. First, the sequence of the target DNAs to be analyzed must be known in order to synthesize the complementary probes on the chip. However, the vast majority of environmental microbes have not been characterized. Second, DNA chips rely on hybridization of nucleic acids which is subject to cross hybridization from DNA molecules with similar sequence. However, the resolving power of a hybridization-based approach is limited because one must identify regions of DNA that do not cross-hybridize, which may be difficult for related microbial species.

[0012] Genomic technologies and bioinformatics hold much untapped potential for application in other areas of biology, especially in the field of microbiology. However, to date there has not been a method to rapidly and easily determine the genomic diversity of a population, such as a

microbial or viral population. Further, there has not been a method to easily determine the antibody or TCR diversity of a population of B or T cells, respectively. Thus, there remains a need to develop such methods in these areas.

[0013] Characterization of 16S rRNA Genes from Oil Field Microbial Communities Indicates the Presence of a Variety of Sulfate-Reducing, Fermentative, and Sulfide-Oxidizing Bacteria by Gerrit Voordouw et al. discloses characterizing oil field bacteria by cloning and sequencing PCR-amplified 16S rRNA genes to detect a certain bacteria.

[0014] The ribosomal database project by Niels Larsen et al. discloses a curated database that offers ribosome data along with related programs and services. The offerings include phylogenetically ordered alignments of ribosomal RNA (rRNA) sequences, derived phylogenetic trees, rRNA secondary structure diagrams and various software packages for handling, analyzing and displaying alignments and trees.

[0015] Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis by Friedrich v. Wintzingerode et al. discloses specific aspects of sample collection, cell lysis, nucleic acid extraction, PCR amplification, separation of amplified DNA, application of nucleic probes and data analysis.

[0016] Phylogeny of the Main Bacterial 16S rRNA Sequences in Drentse A Grassland Soils (The Netherlands) by Andreas Felske et al. discloses investigating the main bacteria in peaty, acid grassland soils in the Netherlands by ribosome isolation, temperature gradient gel electrophoresis, hybridization, cloning, and sequencing using rRNA to classify and quantify the most active bacteria.

[0017] Frequency of Formation of Chimeric Molecules as a Consequence of PCT Coamplification of 16S rRNA Genes from Mixed Bacterial Genomes by Grace C. Y. Wang and Yue Wang discloses a method to quantify the frequency of chimeric molecule formation, wherein mixed genomic DNAs from eight actinomycete species whose 16S rRNA sequences had been determined were used for PCR coamplification of 16S rRNA genes. A large number of cloned 16S ribosomal DNAs were examined by sequence analysis, and chimeric molecules were identified by multiple-sequence alignment with reference species.

[0018] Serial Analysis of Gene Expression by Victor E. Velculescu et al. discloses a method called serial analysis of gene expression (SAGE), that allows the quantitative and simultaneous analysis of a large number of transcripts. SAGE should provide a broadly applicable means for the quantitative cataloguing and comparison of expressed genes in a variety of normal, developmental, and disease states.

BRIEF SUMMARY OF THE INVENTION

[0019] The present invention solves this problem by providing methods for rapidly determining

the diversity of a microbial population.

[0020] The present invention also provides a method for constructing a marker diversity profile (MDP) database comprising the steps of: a) sequencing a plurality of rRNA markers from a sample, wherein said sample comprises a microbial population, and wherein each rRNA marker comprises a microbial rRNA gene polymorphic sequence, and wherein the plurality of rRNA markers provides an accurate representation of the microbial population of said sample; b) determining the abundance, in said sample, of each of said plurality of rRNA markers; c) transducing the abundance of each rRNA marker into an electrical output signal; d) storing said electrical output signals produced in step c) in a matrix data structure and associating in said structure each output signal with the corresponding sequence of the rRNA marker from whose abundance the output signal was transduced; e) providing the abundances of one or more sample parameters that are associated with said sample; f) transducing the abundances of each sample parameter into an electrical output signal; g) storing said electrical output signals produced in step f) in a matrix data structure and associating in said structure each output signal with the sample parameter from whose abundance the output signal was transduced; h) designating the electrical output signals produced in step c) corresponding to the plurality of rRNA marker abundances and the electrical output signals produced in step f) corresponding to the one or more sample parameter abundances, as an MDP; and i) repeating steps a-h for a plurality of other samples, and designating the plurality of MDPs stored in said data structure as an MDP database.

[0021] In an embodiment, said rRNA marker abundance is an abundance relative to the total abundance of the plurality of rRNA markers in the sample.

[0022] In an embodiment, said microbial rRNA gene is a 16S rRNA gene.

[0023] In an embodiment, said microbial rRNA gene polymorphic sequence is from the intergenic region between a 16S rRNA gene and a 23S rRNA gene.

[0024] In an embodiment, the sample is selected from the group consisting of: a soil sample, a rock sample, a water sample, an air sample, a hydrocarbon sample, a petroleum sample and a biofilm sample.

[0025] In an embodiment, the sample is obtained from the group consisting of: petroleum, hydrocarbons, an oil reservoir, a gas reservoir, soil, rock, water, air, a building, and a roadway. In an embodiment, the sample parameter is selected from the group consisting of: petroleum, hydrocarbons, oil, gas, an inorganic compound, a mineral, pH, temperature, and salinity.

[0026] In an embodiment, the sample is obtained from the group consisting of: a human, a plant, an animal, a foodstuff a cell culture, and a tissue culture.

[0027] In an embodiment, the sample parameter is selected from the group consisting of: an acute disease state, a chronic disease state, a physiological state, and a development stage.

[0028] In an embodiment, the sample parameter is time, and wherein the plurality of samples comprise a time-course.

[0029] In an embodiment, the plurality of samples comprise a time-course useful to monitor environmental remediation efforts or oilfield production.

[0030] In an embodiment, a MDP database constructed according to the method of anyone of claims 1 to 10.

BRIEF DESCRIPTION OF THE DRAWINGS

[0031]

Figure 1 shows a representation of a 16S rDNA gene from Bacteria illustrating polymorphic regions (shown as dark bands) and constant regions (shown as light bands).

Figure 2 shows a schematic representation of the Serial Analysis Ribosomal DNA (SARD) method for isolating polymorphic sequence tags from Bacterial rDNA.

Figure 3 shows a number of representative members of the domain Bacteria with their taxonomic relationships.

Figure 4 shows a number of representative members of the domain Archaea with their taxonomic relationships.

Figure 5 shows examples of Marker-Marker correlation scatter plots. Each point represents a single sample population.

Figure 6 shows examples of various Marker-Parameter scatter plots. Each point represents a single sample population.

Figures 7A, 7B and 7C shows scatter plots comparing Marker Diversity Profile (MDP) profiles. Each point represents one marker. Fig. 7A: Significant correlation with all markers within the MDP. Fig. 7B: No correlation found using all the markers within the MDP. Fig. 7C: Same plot as B except that significant correlation is found using a subset of markers.

Figure 8 shows a schematic for generating a marker diversity profile matrix database. The first step 100 involves assigning N an integer value of 1 corresponding to the first MDP data structure. The second step 105 involves assembling polymorphic markers from a sample. Examples of methods for assembling such markers are described in this application and include, but are not limited to, SARD (Figure 2, Tables I and II) or mass tag compilation (Table III). The next step 110 involves detecting the polymorphic markers. Examples of methods to detect polymorphic markers described above include DNA sequence analysis of SARD tags and MALDI-TOF analysis of mass tags, respectively. This detection step includes detecting the presence and abundance of each marker in the sample. The next step **115** involves the

conversion or transduction of the MDP into an electrical signal output. Generally, this process is a linear electronic conversion of the data into a digital signal. Step **120** involves detecting parameters that are associated with the sample N. Step **125** involves transducing the sample parameter data, which may include, without limitation, such parameters as pH, grain size, elemental analysis and/or organic analysis, into an electrical signal output in the form of a digital signal. Step **130** involves storing into the memory of a computer the output signal from each MDP into a matrix data structure and associating it with sample parameters. The next step is a decision block **135** where if all the data structures have not been completed the routine advances to step **140** where n is incremented and the data structure generating steps **105-130** are repeated for the n +1th marker diversity profile. Once all the data structures are completed the routine proceed to step **145** to form the MDP database from n data structures. Each marker is assigned a unique identifier along with its relative abundance in the population. This information is also optionally indexed with other known parameters that are associated with the sample including, for instance, the time, date, elevation and geographical location. These signals are digitized and stored in the memory of a computer.

Figure 9 shows a schematic of the steps involved in determining parameters associated with an MDP. A marker diversity profile **200** is created for a sample. The marker diversity profile is subject to a comparison function **205** which compares the profile with resident marker diversity profiles in the database. Step **210** is a decision block where a query is made whether the new marker diversity profile equals a resident marker diversity profile in the database. If the query returns a "Yes" the new marker diversity profile is deduced **215** to share the same parameters as the resident marker diversity profile. Since the parameters associated with the resident marker diversity profile are characterized, the parameters associate with the new marker diversity profile are identified.

If the query in decision block **210** returns a "No" the routine proceeds to decision block **220** which queries whether the new marker diversity profile is a subset of a resident profile in the database. If the query returns a "No" the parameters remain undefined **225**. If the query returns a "Yes" the routine proceeds to step **230**. Step **230** is an optional step to determine the correlation between members of the common subset of markers and may either be performed for each new profile or may be queried from a matrix table of pre-calculated values from existing profiles. Such values generally would be maintained in a relational database. If this step is not performed all common markers are parsed into groups of individual markers and treated as correlated groups **255**. If the marker-marker correlation is performed between the common subset of markers, the routine proceeds to decision block **235** which queries whether all of the common markers are correlated. If the query returns a "Yes" the markers are correlated with the parameters **240** resident in the database. If none of the markers are correlated with a parameter, the parameter(s) remain undefined **245** whereas if the markers are correlated with a parameter, the parameter is deduced to be associated with the marker diversity profile **250**. If the decision block query **235** returns a "No", the common markers are sorted into groups of correlated markers **255**. The first correlated marker group N **260** is subject to a decision block **265** that queries whether the markers in this group are correlated with a parameter. A "No" determines that the parameters remain undefined. If the marker(s) are correlated with a parameter, the parameter is deduced to be associated with the marker

diversity profile. Steps **260-275** are repeated in steps **280-295** for each correlated group of markers. The groups of correlated markers may be comprised of a single or multiple markers. The confidence level in deducing that a parameter is associated with a marker diversity profile is determined by the level of correlation between the marker(s) and the parameter. Therefore, sets of correlated markers are expected to be more robust indicators of any given parameter.

Figure 10 shows a schematic for generating a marker diversity profile matrix database. The first step **300** involves assigning N an integer value of 1 corresponding to the first MDP data structure. The second step **305** involves assembling polymorphic markers from a sample. Examples of methods for assembling such markers are described in this application and include, but are not limited to, SARD (Figure 2, Tables I and II) or mass tag compilation (Table III). The next step **310** involves detecting the polymorphic markers. Examples of methods to detect polymorphic markers described above include DNA sequence analysis of SARD tags and MALDI-TOF analysis of mass tags, respectively. The next step **315** involves the conversion or transduction of the MDP into an electrical signal output. Generally, this process is a linear electronic conversion of the data into a digital signal. Step **320** involves storing into the memory of a computer the output signal from each MDP into a matrix data structure associating each MDP with a geographic coordinates such as longitude and latitude. The next step is a decision block **325** where if all the data structures have not been completed the routine advances to step **330** where n is incremented and the data structure generating steps **305-320** are repeated for the n +1th marker diversity profile. Once all the data structures are completed the routine proceed to step **335** to form the MDP database from n data structures. Each marker is assigned a unique identifier and indexed with its relative abundance in the population. These signals are digitized and stored in the memory of a computer.

Figure 11 shows a schematic for mapping applications using marker diversity profiles. Marker diversity profile data **400** can be processed in several ways to create maps that provides significant environmental information. In one example **405**, each marker diversity profile in a database can be correlated with every other marker diversity profile in a pairwise manner to create a correlation matrix. By appending this data to the geographical coordinates of each sample **410**, a map can be constructed that reflects the correlation values of physically neighboring sample sites. Preferably the correlation values will be color coded to reflect the level of correlation. The color is chosen from a reference color spectrum that is indexed to correlation values between 0-1.

Marker diversity profiles **400** can also be processed into maps at the individual marker or correlated marker group level. This approach is preferable since subsets of markers are likely to correlate to fewer number of sample associated parameters. Each marker in a marker diversity profile database is correlated with every other marker in the database in a pairwise manner to create a correlation matrix **415**. The source database can either be composed of marker diversity profiles from a single geographic area or several distinct areas. In step **420**, the markers from one geographic area are sorted into groups based upon their level of correlation. In step **425**, the relative representation of the correlated marker group N is determined along with its geographical coordinates for each marker diversity profile in a geographic area. A map is constructed **430** where the relative abundance of each correlated

marker group is color-coded with its geographical coordinates. Steps **425** and **430** are repeated as in **435** and **440** for each correlated group of markers.

Fig. 12 shows oligonucleotides useful for amplifying nucleic acid molecules for SARD.

Fig. 13 shows the use of the SARD strategy for Eubacteria. The double-underlined sequence and the wavy-underlined sequence represent the sequence tags for the two pools and the single-underlined sequence delineates the *Bpml* recognition site.

Fig. 14 is a graphical representation of a SARD analysis of a defined population.

Fig. 15 shows the sequence of SARD tags identified from Wy-1 sample. The number in parentheses indicates the number of tags having that sequence.

Fig. 16 shows SARD tags identified from Wy-2 sample. The number in parentheses indicates the number of tags having that sequence.

Fig. 17 is a graphical representation of the number and abundance of SARD tags. The upper panel shows the Wy-1 SARD Tag Diversity Profile and the lower panel shows the Wy-2 SARD Tag Diversity Profile.

DETAILED DESCRIPTION OF THE INVENTION

[0032] The extent of the diversity of microbes in our environment has only recently been recognized. With the advent of the polymerase chain reaction (PCR) and small subunit ribosomal DNA (rDNA) sequence analysis, researchers have been able to detect and perform phylogenetic analyses on individual microbes without first cultivating the microbes of interest. This molecular phylogenetic approach has significantly changed our view of microbial evolution and diversity (Woese, C. R, 1987, Bacterial evolution. Microbiol Rev. 51(2):221-71; Pace, N. R, 1997, A molecular view of microbial diversity and the biosphere. Science. 276(5313):734-40). For instance, the earliest life forms are now thought to have utilized inorganic compounds for nutrition rather than compounds based upon organic carbon. In addition, the vast proportion of biological diversity is now known to be due to microbial species. Estimates have been made that there may be more than ten thousand *distinct* species of microbes in a single gram of soil. Figures 3 and 4 show some of the representative members of the domains Bacteria and Archaea, respectively, that may be found in environmental samples.

[0033] Microbes inhabit virtually all niches including extreme environments with temperatures between 20° and 250° F. Microbes have even been isolated from deep petroleum reservoirs more than a mile beneath the earth's surface (Jeanthon, C. et al., 1995, Thermotoga subterranea sp. nov., a new thermophilic bacterium isolated from a continental oil reservoir. Arch. Microbiol. 164:91-97). In order to prevail under such diverse conditions, microbes have made remarkable adaptations and have attained the ability to utilize unusual carbon and

mineral resources that are immediately available. These physiological and metabolic adaptations that enable some microbes to inhabit a particular niche may also restrict their distribution to such areas. Numerous examples of environmental parameters that lead to restrictions of microbial distribution are well known and are usually dictated by a species' specific metabolic program (e.g. obligate nature of the carbon, nitrogen and energy source).

[0034] Microbes that have highly defined nutrient requirements are likely to have a restricted distribution in the environment. Thus, the microbes' dependence on the presence of a particular resource to proliferate can serve as the basis for an assay to identify the presence, and characterize the distribution, of various features in the environment, such as biological, chemical and geochemical features. In other words, microbes can function as environmental biosensors.

[0035] In one aspect of this invention, the ability of microbes to function as environmental biosensors is used to identify particular environmental states. In a preferred embodiment, a profile of a microbial population is used to identify one or more parameters of a particular environmental state. In a more preferred embodiment, a microbial population profile is used to identify areas that are likely to have mineral deposits and/or petroleum reserves. In another preferred embodiment, a microbial population profile is used in forensic science to identify decomposition of a body or to associate an individual to another individual, to an object or to a location. In yet another preferred embodiment, a microbial population profile is used to identify microbial contamination of human and animal foodstocks. In yet another preferred embodiment, the profile is used to diagnose human or animal disease.

Definitions

[0036] Unless otherwise defined, all technical and scientific terms used herein have the meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. The practice of the present invention employs, unless otherwise indicated, conventional techniques of chemistry, molecular biology, microbiology, recombinant DNA, genetics and immunology. See, e.g., Maniatis et al., 1982; Sambrook et al., 1989; Ausubel et al., 1992; Glover, 1985; Anand, 1992; Guthrie and Fink, 1991.

[0037] A microbe is defined as any microorganism that is of the domains *Bacterial*, *Eukarya* or *Archaea*. Microbes include, without limitation, bacteria, fungi, nematodes, protozoans, archaebacteria, algae, dinoflagellates, molds, bacteriophages, mycoplasma, viruses and viroids.

[0038] A marker is a DNA sequence that can be used to distinguish or identify a particular gene, genome or organism from another. In one embodiment, a marker may be generated by one of the methods described herein. A marker represents one or a limited number of taxonomic species or genes. In a preferred embodiment, a marker represents a single taxonomic species or gene. In one embodiment, the marker represents a single microbial

species. In another embodiment, the marker represents a single viral species or type. In another embodiment, the marker represents a single immunoglobulin or TCR variable domain.

[0039] A marker diversity profile (MDP) is a data set that is obtained from each population sample and that contains a collection of markers. In a preferred embodiment, the MDP also comprises other information, including all known parameters associated with a particular population sample. Such parameters that relate to environmental samples may include inorganic components (obtained through atomic adsorption analysis), organic components (obtained through GC-MS or LC-MS), grain size analysis, pH, and salinity. Parameters that relate to medical samples would include, but are not limited to, a complete medical history of the donor. In a preferred embodiment, the markers are obtained by SARD.

Methods for the Genetic Analysis of Populations

[0040] Ribosomes, which are comprised of numerous ribosomal proteins and three ribosomal RNA (rRNA) molecules, are a key component of protein synthesis. The 16S subunit rRNA, which is encoded by the 16S rDNA gene, has been the focus of much attention in microbial phylogenetic studies. The 16S rDNA sequence is highly conserved between taxonomic groups, yet also possesses regions that are highly polymorphic (Figure 1). Moreover, the rate of change in the RNA sequence is thought to have been relatively constant over evolutionary time, enabling scientists to determine the relative relatedness of different organisms.

[0041] Typical molecular microbial analyses involve utilizing the highly conserved regions of the 16S rDNA to amplify the roughly 1,500 bp gene. The sequence of the PCR-amplified product is determined and compared with other known rDNA sequences. Although this approach is highly informative, it is not amenable to a rapid survey of an environmental microbial community.

[0042] The instant invention provides methods for quickly and easily calculating the genetic diversity of a population. The method uses hybridization-independent genomic technologies to overcome the previously-identified problems of determining genetic diversity. This method may be used for any population of cells, viruses or organisms which comprise at least one DNA molecule that comprises regions of high sequence conservation interspersed with polymorphic sequences, wherein the polymorphic sequences can be used to distinguish different members of the population of interest. One aspect of the present invention describes a method (SARD) that can capture a designated polymorphic region from a given DNA molecule present in the members of a microbial community. In a preferred embodiment, the DNA molecule is a 16S rDNA molecule. In another embodiment, the DNA molecule is the intergenic region between the 16S and 23S rDNA genes.

[0043] The method may be performed as follows (see Figure 2):

Step 1. Sample Preparation and DNA Amplification by PCR

[0044] Samples may be obtained from any organism or region desired. For environmental microbial analyses, samples may be obtained from, without limitation, buildings, roadways, soil, rock, plants, animals, cell or tissue culture, organic debris, air or water. For medical microbial analyses, samples may be obtained from, without limitation, humans, animals, parasites, water, soil, air and foodstuffs. For viral analyses, samples may be obtained from, without limitation, viral culture stocks, humans, animals, plants, cell or tissue culture and microbes. For immunoglobulin or TCR analyses, samples may be obtained from, without limitation, humans, animals or cell or tissue cultures. DNA molecules from the sample of interest may be isolated by any method known in the art. See, e.g., Sambrook et al., 1989 and Ausubel et al., 1992. In a preferred embodiment, DNA is obtained as described by Yeates et al., "Methods for Microbiological DNA Extraction from Soil for PCR Amplification," Biological Procedures Online, Volume 1, May 14, 1998, available at www.science.uwaterloo.ca/bpo; Liu et al., Applied and Environmental Microbiology (1997) 63: 4516-4522; and Tsai et al., Applied and Environmental Microbiology (1992) 58: 2292-2295. The DNA molecules do not have to be completely purified but only need be isolated to the point at which PCR may be performed.

[0045] Environmental microbes often exist in biofilms (Costerton, JW., et al., 1999, Bacterial biofilms: a common cause of persistent infections. Science 284(5418):1318-1322) or in tight association with solid surfaces. Microbial DNA from a sample of interest is isolated by one of several methods that are widely known to those skilled in the art and are described in the literature (Gillan, D.C. et al., 1998, Genetic diversity of the biofilm covering Montacuta ferruginosa (Mollusca, bivalvia) as evaluated by denaturing gradient gel electrophoresis analysis and cloning of PCR-amplified gene fragments coding for 16S rRNA Appl. Environ. Microbial. 64(9):3464-72).

[0046] The samples may be selectively enriched before they are isolated by any method known in the art. In one embodiment, for a population of microbes that are known or suspected to feed on a hydrocarbon source such as propane, the hydrocarbon may be added to the environment in which the microbes live for a period of time before the microbes are harvested. In another embodiment, a viral population may be cultivated in cells before they are isolated. In a further embodiment, B and T cells may be expanded in culture before isolation. It may be easier to obtain sufficient sample amounts if a population is expanded before isolation. However, this must be weighed against the possibility that expansion will alter the ratio of different members of the population to each other.

[0047] In general, the primers used for amplification are designed to hybridize to a region of the DNA that is highly conserved between members of the population. Further, the primers should flank a polymorphic region the partial sequence of which should provide diagnostic information regarding the genetic diversity of the population. For instance, for the 16S rDNA gene, primers are designed to hybridize to a highly conserved region of the 16S rDNA gene flanking a polymorphic region (see Figure 2). In an immunoglobulin gene, the primers are designed to hybridize to a region of the B cell DNA flanking the V-J recombination site.

Alternatively, primers may be designed that bind to the relatively constant regions within certain regions of the V-J gene that flank a polymorphic region. See Roitt et al., Immunology, 3rd Ed., 1993, pp. 5.2-5.14, which shows regions of variability and conservation within immunoglobulin and TCR genes. One having ordinary skill in the art following the teachings of the specification will recognize that other genes that have regions that are highly conserved between members of the population and that flank polymorphic regions may be used in the design of primers.

[0048] The primers should also be designed to flank a region of DNA that comprises a restriction site for a restriction enzyme. In a preferred embodiment, the restriction enzyme is a cuts at a four-basepair recognition site, such as *AluI* (see Figure 2). Furthermore, the restriction site should be near but not in the polymorphic region of the gene of interest. In a preferred embodiment, the restriction site should be one that is present in the gene of interest in a majority of the known species or genes.

[0049] A single set of primers may be used or multiple sets of primers may be used. A single set of primers may be used if it is known that the region of DNA to which the primers will bind is very highly conserved. Alternatively, if it is known that there is some variation in the conserved region, multiple sets of primers may be used to bind to the conserved DNA region. Using multiple sets of primers may be useful to identify more members of a population, especially those members of the population that exhibit less sequence identity in the conserved areas of a nucleic acid sequence. In one embodiment, four to ten sets of primers may be used to identify members of a population. Alternatively, the primers used may be degenerate, such that different molecules within a primer population will include a different base at one or more specific sites in the primer. For instance, a primer may have a site that has either cytosine or thymidine. The purpose of making primers degenerate is to increase the number of different DNA molecules that will hybridize to a particular primer. Methods of making degenerate primers are well-known in the art.

[0050] The primers used in this and in subsequent steps are generally of a length sufficient to promote specific hybridization to a DNA sequence. The primers generally have a length of at least 12 bases, more preferably at least 15 bases, even more preferably at least 18 bases. The primers may have a length of up to 60 bases, although usually most are under 40 bases in length. Primers may include both bases that are naturally found in DNA, such as adenine, guanine, cytosine and thymidine, and may also include nucleotides that are not usually found in DNA, such as inosine.

[0051] One of the primers (the "upstream" primer) should be modified to incorporate a moiety that can be used to bind the PCR product to a solid support. The upstream primer is defined as the primer that is located on the opposite side of the polymorphic region of interest relative to the flanking four-base restriction site. A number of different binding moieties are known in the art. In a preferred embodiment, the moiety is biotin. In another preferred embodiment, the moiety is digoxigenin or six histidines.

[0052] PCR is performed using the primers to amplify a subregion that contains a polymorphic

site of interest. Methods for performing PCR are well-known in the art. In one embodiment, the PCR products are normalized or subtracted by methods known in the art to lower the representation of the dominant sequences. Exemplary methods are described in Sambrook et al., 1989, Ausubel et al., 1992 Glover, 1985; Anand, 1992.

Step II. Digestion of the Amplified Fragment and Binding to Solid Support

[0053] The amplified fragment is cut with the restriction enzyme as discussed in Step I. Any restriction enzyme may be used in this step so long as it cuts at a site immediately adjacent to the polymorphic sequence. In a preferred embodiment, the enzyme is a four-base restriction enzyme. Examples of four-base restriction enzymes are well-known in the art and include many that are commercially available. See, e.g., New England Biolabs Catalog 2000. Examples of four-base restriction enzymes include, without limitation, *AluI*, *Bsh1236I*, *DpnI*, *HpaII*, *MboI*, *MspI*, *PaiI* (an isoschizomer of *HaeIII*), *RsaI*, *Sau3AI* and *TaqI*. After restriction, the DNA fragment is bound to a solid support. Numerous solid supports to immobilize DNA are known in the art. Examples include, without limitation, streptavidin beads, which would bind to a PCR product labeled with biotin, and anti-digoxigenin beads, which would bind to a PCR product labeled with digoxigenin, and beads conjugated to nickel, which would bind to a six-histidine labeled product. In a preferred embodiment, streptavidin beads are used (Figure 2).

[0054] Since the SARD tag position is dictated by the first restriction enzyme recognition site distal to the biotinylated primer used in the initial PCR reaction, there may be cases in which the first restriction enzyme recognition site is located within a conserved region of the gene of interest. In general, this will not be a problem because even though the tags from the conserved region may not be informative, most tags derived by SARD will be from a polymorphic region and will be informative. However, if one desires to decrease the number of tags that contain information from a conserved region of a gene rather than from a polymorphic region, one may purify the desired PCR products after restriction. In a preferred embodiment, one may do this by gel purifying those PCR products that have the expected size.

Step III. Digestion of the Amplification Product and Ligation to Linkers

[0055] The immobilized products are split into two pools and linkers are attached to the immobilized products of each pool. Each linker is a double-stranded synthetic DNA molecule comprising a specific DNA sequence. Both linkers incorporate a Type IIS restriction enzyme site. In a preferred embodiment, the two linkers incorporate the same Type IIS restriction enzyme site. Each of the two linkers also comprises a DNA sequence that specifically hybridizes to a primer. In one embodiment, the linkers are identical to one another and hybridize to the same primer. In a preferred embodiment, the linkers are different from each other such that each hybridizes to a different primer.

[0056] The double-stranded linker is ligated to the immobilized PCR product. The linker may incorporate the Type IIS restriction enzyme site or it may incorporate only a portion of the site. In this case, the linker will be designed such that ligation of the linker to the restricted DNA will reconstitute the Type IIS restriction site. In a preferred embodiment, the linker incorporates a *BpmI* site. Linker ligation is well-known and may be accomplished by any method known in the art. After ligation, the immobilized PCR product is isolated from the free linkers by any method known in the art. See, e.g., Velcelescu et al., *Science* 270: 484-487, 1995; Powell, *Nucleic Acids Research* 14: 3445-3446, 1998; Sambrook et al., pp. F.8-F.10, 1989.

[0057] Type IIS restriction enzymes cleave at a defined distance up to 20 basepairs away from their asymmetric recognition sites. Type IIS restriction enzymes that are commercially available include enzymes that leave 5' overhangs and those that leave 3' overhangs as double-stranded DNA products. Some enzymes of the former class include: *BsmFI* (10/14), *Bst71I* (8/12), and *FokI* (9/13), where the number in parentheses indicate the cleavage position on the same DNA strand as the recognition sequence/cleavage position on the complementary DNA strand. Enzymes of the latter class include: *BpmI* (16/14), *BsgI* (16/14), *Eco57I* (16/14) and *GsuI* (16/14). The 3' overhang left by these enzymes must be removed for a blunt ligation (*Step IV*). Therefore, enzymes that cleave at positions 16/14 result in a 14 base-pair tag. Other enzymes that cut at a more distal position could create a larger tag. For instance, *MmeI* (20/18) leaves a 3' overhang, but is not commercially available (Tucholski, J. et al., 1995, *MmeI*, a class-IIS restriction endonuclease: purification and characterization. *Gene* 157: 87-92).

Step IV. Digestion of the Product with Type IIS Restriction Enzyme

[0058] The product is digested with the appropriate Type IIS restriction enzyme to release a DNA fragment from the anchoring bead and produce a short hybrid DNA fragment containing a portion of the polymorphic region of the DNA of interest (the tag) and the linker DNA. After digestion, the DNA must be either filled in or digested to create blunt ends. If the Type IIS restriction enzyme produces a 3' overhang, the fragment is digested with T4 DNA polymerase to remove the 3' overhang. If the Type IIS restriction enzyme produces a 5' overhang, the overhang must be filled in using the appropriate deoxynucleotides and the Klenow fragment of DNA polymerase I. The DNA fragment is separated from the rest of the immobilized PCR product. In a preferred embodiment, the two pools of immobilized PCR products are digested with *BpmI* to release the polymorphic markers and digested with T4 DNA polymerase to create blunt ends (Figure 2).

Step V. Ligation of Tags and PCR Amplification of Resulting Ditags

[0059] The tags are blunt-end ligated to one another using methods well-known in the art to form ditags. See, e.g., Sambrook et al., 1989; Velcelescu et al., *Science* 270: 484-487, 1995.

The ditags are subsequently amplified by PCR using primers that are unique to the linkers used in Step III. In a preferred embodiment, the primers are different from one another if the linkers used in Step III were different from one another. Alternatively, the primers may be the same if the linkers used in Step III were identical to one another. The number of PCR amplification reactions will vary depending upon the amount of DNA present in the starting material. If there is a large amount of DNA, then only one PCR amplification reaction, wherein each reaction comprises from approximately 15-30 cycles, will be required at this step. If the starting amount of DNA is low, then more than one PCR amplification reaction may be required at this step.

Step VI. Cleavage of the Ditags and Ligation to Form Ditag Concatemers

[0060] The ditags are cleaved with the four-base restriction enzyme used in Step II. The products are then ligated to create ditag concatemers. In one embodiment, the ditag concatemers range from 2 to 200 ditags. In a more preferred embodiment, the concatemer comprises 20-50 polymorphic tags. The concatemer may be sequenced directly, or may be cloned into a sequencing vector. Using a 96-channel capillary DNA sequencer, about 12,000 tags could be easily analyzed in one day. Alternatively, the concatemers may be sequenced manually.

Methods to analyze Marker Data

[0061] The invention is directed toward methods of analyzing the genetic diversity of a population in a sample. Each population that is analyzed will have its own unique set of different organisms or genes. The data set that is captured from each sample should recapitulate the genetic structure in a survey format to include a marker for each gene or organism and the relative abundance of each gene or organism in the population as a whole. The markers for a particular population form a marker diversity profiles (MDPs), that may be entered into a database. See, e.g., Fig. 8 which shows one schematic for generating such a database. The method by which the data are captured is not critical as long as it produces an accurate representation of each population.

[0062] In one aspect of the method, MDPs are entered into a database. In a preferred embodiment, the database is kept in a computer-readable form, such as on a diskette, on a non-removable disk, on a network server, or on the World Wide Web. However, the method by which the data are captured is not critical as long as it produces an accurate representation of each microbial community.

[0063] Artificial intelligence (AI) systems can perform many data management and analysis functions. Examples of AI systems include expert systems and neural networks. Expert systems analyze data according to a knowledge base in conjunction with a resident database.

Neural networks are comprised of interconnected processing units that can interpret multiple input signals and generate a single output signal. AI systems are ideally suited for analyzing complex biological systems including populations through the use of deduction protocols.

[0064] A marker may be correlated with a particular condition or with another marker. See, e.g., Fig. 9 for a schematic of the steps involved in determining particular parameters associated with an MDP and Fig. 10, which shows a schematic for generating a marker diversity matrix database. A condition or state may be an environmental condition such as pH, temperature, salinity, or the presence or absence of an organic or inorganic compound such as hydrocarbons, nitrates or mineral deposits. A condition may be a physiological or medical condition such as an acute or chronic disease state, physiological state, developmental stage or associated with a particular body tissue or fluid. Information regarding all known parameters associated with the samples will also be saved together with the MDPs.

[0065] Each MDP is composed of markers which represent a small number, more preferably one, species or gene. For instance, in the case of Example 1, each marker would be comprised of a 12 base-pair polymorphic 16S rDNA sequence. Such parameters that relate to environmental samples may include inorganic components (obtained through atomic adsorption analysis), organic components (obtained through GC-MS or LC-MS), grain size analysis, pH, and salinity. Parameters that relate to medical samples would include, but are not limited to, a complete medical history of the donor. See, e.g., Fig. 11, which shows a schematic for mapping applications using marker diversity profiles.

[0066] In another aspect of the invention, MDPs are collected for a time course, and each time point is one of the parameters included. Time courses may be useful for tracking changes over time for a wide variety of indication. For example, time courses may be useful for tracking the progression of a disease, during environmental remediation efforts, and during oilfield production.

[0067] In another aspect of the invention, MDPs are collected in various distinct locations, such as in various geographical locations or in various tissues of the body. Comparison of MDPs compiled from various distinct locations are useful for distinguishing changes between these various locations, which may be indicative of particular environmental conditions or disease states.

[0068] Comparison of marker diversity profiles can reveal trends in populations either relative to time or to geographical location. In the latter case, comparisons of microbial populations can resolve spacial information about the environment that would otherwise be undetected. Examples of such information include migration patterns of water, organic compounds and minerals. For instance, placer deposits of minerals are caused by the action of both water and wind causing the minerals to migrate from a lode deposit at one location only to deposited at another location. The migration of such minerals may leave a detectable trace upon the microbial populations in the path of migration. Physical attributes of the environment could also be detected such as structures, formations and fault lines. It is commonly understood that

faults offer a significant vertical migration route for gases such as methane which is known to be differentially utilized by microbes. By combining MDP data with geographical coordinates such as elevation, longitude and latitude that can easily be obtained with global positioning system devices, it is possible to create maps delineating the distribution of various microbes in the environment.

[0069] Correlation analyses between one marker and all the other markers in the database will reveal pairs of markers that have a propensity to coincide. This process can be repeated in an iterative manner for all markers to produce a matrix of correlation coefficients between all observed markers. Figure 5 shows a scatter plot for two pairs of markers with one of the pairs exhibiting a high degree of correlation. This approach can also be used to create a dendrogram that reflects the relative level of correlation between each marker. Therefore, at any chosen level of correlation, all of the observed markers can be divided into groups where the markers in each group share the same level of correlation with each other member of the group. If a high correlation coefficient value is chosen (e.g. 0.8), the markers of each group would, more often than not, be found in the same sample. Thus, this exercise will divide a given population into groups of genes or organisms that have a propensity to co-localize with each other. In one preferred embodiment, the exercise will divide a microbial community into groups of microbes that have a propensity to co-localize.

[0070] Correlation analysis between a marker (variable 1) and a sample parameter (variable 2) will identify markers whose presence often, or invariably, coincides with a component present in the samples. Some types of relationships between markers and sample components (or parameters) are shown in Figure 6. A strong correlation value between a marker and sample parameter would allow predictions to be made about the abundance of either variable (marker or sample parameter) as long as one of the variables is known.

[0071] In some cases, a marker will not be specific for a single species or gene. For example, the tag sequence that would be identified by the approach depicted in Example 1 would be identical for *Denitrobacter permanens* and *Legionella anisa*. In the cases where a significant correlation is found between a marker and sample parameter of interest, the preferred action is to use the tag sequence information to identify the complete gene sequence. The sequence can then be used to identify the species and to identify species-specific probes to verify the correlation. One may do this using methods known in the art, such as by PCR or by hybridization to DNA molecules isolated from the sample of interest, followed by sequencing or other method of analysis.

[0072] Species-specific probes that are identified from markers with a robust correlation to a sample parameter of interest can then be utilized as a diagnostic, or to prospect for the parameter of interest. Such assays would preferably be PCR-based and would be highly sensitive, rapid and inexpensive. In a preferred embodiment, a marker identified by these methods may be used as a hybridization probe to identify a larger piece of the DNA from which the marker is derived. The sequence of the larger DNA molecule can then be used to design primers that will specifically hybridize to the DNA molecule of interest and which can be used to

specifically amplify the DNA molecule by PCR. Alternatively, one may use a hybridization-based assay using a probe that binds specifically to the DNA molecule of interest. Using specific primers or probes are especially useful for quickly determining whether a large number of samples contains the DNA molecule that correlates to the parameter of interest.

[0073] In a preferred embodiment, a marker that correlates with a desired parameter is identified. The marker may be identified using SARD, or may be identified using another method, such as restriction fragment length polymorphisms (RFLP) or terminal restriction fragment length polymorphisms (T-RFLP; Liu et al., Applied and Environmental Microbiology 63: 4516-4522, 1997). A method such as denaturing gradient gel electrophoresis (DGGE) may be used to identify size differences. In a preferred embodiment, SARD is used to identify the marker. Other samples are screened to determine whether they have the marker of interest. In a preferred embodiment, the screen used is PCR or hybridization, more preferably PCR. In an even more preferred embodiment, the marker correlates with the presence of mineral deposits or petroleum reserves.

[0074] Correlation analysis between MDPs (MDP_n, variable 1; MDP_{n+1}, variable 2) can reveal the relative similarities between samples. Samples taken from the same individual or from proximal environmental sites that have similar composition, are expected to show a robust correlation coefficient (Figure 7A). However, samples that share only one or a few parameters in common are expected not to show a significant correlation when all of the markers are considered (Figure 7B). By incorporating the knowledge learned from the Marker-Marker and Marker-Parameter correlations, MDPs can be compared using either individual markers or preferably, subsets of correlated markers in the analysis (Figure 7C). This approach can eliminate much of the noise and enable one to identify hidden relationships.

[0075] Correlation analyses may be performed by any method or calculation known in the art. Correlation analyses for r and r^2 may be performed as described by M. J. Schmidt in Understanding and Using Statistics, 1975 (D. C. Heath and Company), pages 131-147. The degree of correlation for r may be defined as follows:

1.0	Perfect
0.8-0.99	High
0.5-0.7	Moderate
0.3-0.4	Low
0.1-0.2	Negligible

[0076] In one embodiment, the correlation between two markers or between a marker and a parameter is at least low (r is 0.3-0.4). In a preferred embodiment, the correlation is at least moderate (r is 0.5 to 0.7). In a more preferred embodiment, the correlation is high (r is 0.8 to 0.99).

[0077] With the development of numerous genomic technologies for analyzing complex sets

of nucleic acids, we have the opportunity to begin to catalog the reservoir of microbial, and hence, metabolic diversity. Since the proliferation of a microbe in a given location will depend upon the presence of the requisite metabolic nutrients, information as to the abundance of that microbe can serve as a biosensor for a given set of parameters. When viewed as a whole, the microbial community structure in a given location will hold intrinsic biosensor potential for a wide range of parameters. The predictive reliability of the data from a complete microbial community will also be significantly increased. For example, if a given microbe were present in 50% of soil samples taken above petroleum reservoirs and were found nowhere else, then the presence of ten such microbes would create a predictive value with 99.9% accuracy.

Applications of the Invention

Geochemical and Mineral Exploration

[0078] The methods described in this invention have several benefits over existing technologies. For instance, in the area of geochemical exploration, genomic rDNA-based assays potentially will be able to resolve an extensive set of geochemical parameters of interest to the petroleum and mining industries. Currently, many different technologies are required to measure these parameters. Because this invention is based upon a universal measure, nucleic acid detection, it can greatly reduce instrumentation and sample outsourcing costs.

[0079] Oil and gas reservoirs are located well beneath the earth's surface at depths from a few hundred feet to more than 10,000 feet. When oil is formed, it undergoes a migration in which one of two things take place. The oil may continue to migrate until it ultimately reaches the surface, where it evaporates over time. Alternatively, its migration may be blocked by an impermeable structure, a so-called "trap". Geophysical methods (such as three-dimensional seismic methods) for petroleum exploration relies on finding these trap structures with the hope that they contain oil.

[0080] Crude oil is made up of a variety of hydrocarbon chain lengths. The lightest hydrocarbons (namely methane, ethane, propane and butane) are often able to diffuse through the trap structures and, as a result of pressure gradients, undergo a vertical migration to the surface. Certain microbes present at the surface or in the surface layer are able to utilize these migrating hydrocarbons, which occasionally results in mineralogical changes that are detectable at the surface. Thus, these migrating hydrocarbons would be expected to affect microbial populations, such that the ability to determine the genetic diversity of a microbial population may reveal microbial signatures that are indicative of the presence of oil.

[0081] Recent advances in microfluidics in the genomics industry have resulted in the development of instruments that can detect specific nucleic acids within a few minutes. Utilizing

such instruments will enable measurements to be made in the field for a variety of parameters. In contrast, conventional chemical assays require laboratory analysis and interpretation.

[0082] Biosensors have been created that are able to detect hydrocarbons present at, or below, the level of detection of sophisticated GC-MS analytical instrumentation (Sticher, P. et al., 1997, Development and characterization of a whole-cell bioluminescent sensor for bioavailable middle-chain alkanes in contaminated groundwater samples. Appl Environ Microbiol. 63:4053-4060). Sticher et al. demonstrated that by using single reporter gene in a genetically engineered microbe comprising a reporter gene was able to sense extremely small changes in their environment in response to an acute treatment with a particular hydrocarbon. The instant invention could document the effect upon a population comprising thousands of microbes over geologic time and thus, has the potential of being more sensitive than current analytical instruments.

[0083] This invention may also be used to create a survey of biological entities that is limited only by the prerequisite that these entities contain nucleic acids that are arranged in regions that are conserved and regions that are polymorphic when compared to sequences from related organisms. Some additional examples of the application of this invention are described below.

Oil and Gas Reservoir Development

[0084] In addition to the application of this invention in petroleum exploration, this invention could also be useful in the development of oil and gas reservoirs. Several properties of oil reservoirs that directly affect the commercial viability of the reservoir are modulated at some level by microbes. Hydrogen sulfide is sometimes present in crude oil and can render otherwise 'sweet' oil into 'sour' oil. In addition to its corrosive effect on oilfield equipment, H₂S also poses risk to the workers and significantly reduces the value of an oil reservoir since a washing plant must be installed to remove the gas. The levels of H₂S can change during the development of a reservoir and is now thought to be the result of sulfate-reducing bacteria (Leu, J.-Y. et al., 1999, The same species of sulphate-reducing Desulfomicrobium occur in different oil field environments in the north sea. Lett. Appl. Microbiol. 29(4):246-252). By identifying the presence of microbes that could lead to H₂S production, the valuation of new reservoirs and the resulting developmental strategies could be made more effective.

[0085] Crude oil and natural gas are composed of a complex mixture of hydrocarbons including straight chain hydrocarbons of lengths generally between 2-40 carbon atoms. The shorter chain-length hydrocarbons are more valuable (e.g. gasoline, C₄-C₁₀). In some oil reservoirs, the lighter hydrocarbons are selectively removed either during or prior to development of the reservoir. Microbes have been suspected to play a role in this process since the shorter chain-length hydrocarbons are more bioavailable. This invention could identify microbes that are involved in this process and therefore make predictions as to the

susceptibility of certain reservoirs to the depletion of short chain hydrocarbons. This invention may also be able to identify microbes capable of shortening long chain hydrocarbons thereby increasing the value of existing reservoirs.

Insect and Parasite Detection

[0086] The significant negative impact insects can have on agriculture is widely known. Insects can also serve as vectors for the transmission of many disease causing microbes. Numerous microbe-insect relationships have been described. For example, the bacterial genus *Wolbachia* is found associated with many species of ants and has been shown to alter sex determination and fecundity in the host (Wenseleers, T. et al., 1998, Widespread occurrence of the micro-organism *Wolbachia* in ants. Proc. R Soc. Lond. B. Biol. Sci. 265(1404):1447-52). In addition, many intracellular endosymbiotic bacterial species have been identified in ants (Schroder, D. et al., 1996, Intracellular endosymbiotic bacteria of *Camponotus* species (carpenter ants): systematics, evolution and ultrastructural characterization. Mol. Microbiol. 21:479-89). Most, if not all, insects probably have species-specific intimate relationships with microbes which could represent an Achilles heel for the control of insect populations. The invention described in this application could provide a means to identify microbes that modulate the well-being of a given insect species.

[0087] The identification of microbes that are specifically associated with a given insect could also potentially serve as the basis for a highly sensitive test for the presence of the insect. For instance, current methods to identify the presence of termites in wooden structures is based on visual inspection and is largely inadequate. A test for the presence of a termite-associated microbe that is based on PCR-amplification would be both non-invasive and highly sensitive.

[0088] Further, the ability to create comprehensive inventories of microbial diversity has several applications that relate to microbial ecology that would have utility in the agricultural industry. For instance, the agriculture industry utilizes enormous amounts of pesticides prophylactically to prevent loss of crops. This invention provides the ability to perform comprehensive surveys of microbial populations and could lead to predictions as to the susceptibility of a given field to particular plant pathogens. This knowledge could lead to a better strategy of pesticide applications.

[0089] Further, surveying microbial diversity in an environment such as a agricultural field at various times of the year would reveal the dynamic changes in microbial populations that occur as a result of seasonal fluctuations (temperature and moisture), pesticide application and the proliferation of certain organisms.

[0090] Comparison of the diversity profiles taken from the same or similar site at different times would reveal interactions between species in a population. These productive interactions may manifest themselves either in the increase or decrease in the representation of one

marker relative to the decrease or increase in the representation of a second marker, respectively.

[0091] Such information could provide an early warning of the proliferation of a pest species as well as the identification of species that are pathogenic to a pest species. These pathogenic organisms may have value either as a biological agent to control proliferation of pest species and/or as a source for genes or compounds that would act as pesticides. This phenomenon is not be limited to microbe-microbe interactions. The eggs as well as larval stages of insects likely interact with soil microbes and create a detectable impact upon microbial populations. An example of such an organism is *Bacillus thuringensis*, which itself is commonly used in organic farming as an insecticidal agent. The gene responsible for this insecticidal activity (Bt toxin) has been widely used to create transgenic plants with resistance to insect attack.

Bioremediation

[0092] A considerable amount of effort has gone into the development of methods for microbe-based removal of chemicals from the environment. Such chemicals include heavy metals, polynuclear aromatics (PNAs), halogenated aromatics, crude oil, and a variety of other organic compounds such as MTBE. Regulatory considerations for the release of micro-organisms into the environment have re-directed efforts towards identifying and augmenting the growth of endemic organisms that have the capability to metabolize or remove compounds of interest from the environment.

[0093] The present invention can facilitate bioremediation efforts in two ways. First, organisms can be identified at a given site that have either been previously shown to be capable of removing compounds of interest, or that have significant likelihood of having the capacity metabolize the relevant compounds based upon its coincidence in the environment with the chemical in numerous geological settings. Secondly, this invention can identify trends of soil types and particular microbial species. In this case, the correlations that are drawn from the database between microbial distribution and soil types can be coupled with the existing knowledge base of geochemistry. For instance, the USGS provides many publically available maps describing numerous geochemical and geophysical parameters. Extrapolations of the distribution of microbial species can be made to the regional, and possibly worldwide, level. These extrapolated microbial distributions could serve as the basis for site-specific treatment regimens to augment the growth of certain relevant species without first performing a microbial survey.

Forensic Science

[0094] This invention, particularly SARD analysis, can be used in forensic applications as well. Studies over the past ten to twenty years have focused upon the changes that occur in a

person's or animal's body after death. Many of these changes involve changes in microbial populations that occur during decomposition of the body. Changes in microbial populations have been correlated with length of time that a person has been dead and the conditions that the body experienced after death, e.g., heat, sun exposure, partial or complete burial, rain, etc. SARD analyses would permit forensic scientists to quickly and accurately determine the size and type of microbial populations, which in turn may be used to determine more accurate times of death as well as conditions that the body may have been exposed to.

Other Applications

[0095] This approach can be utilized for any polymorphic region in a microbial genome. This method also need not be restricted to genes. The DNA sequence of intergenic regions of genomes are not under as high a level of selective pressure and thus, represent highly polymorphic DNA sequence. One example of such a region is the intergenic region between the large (23S) and small (16S) rDNA subunits coding regions. The above-described methods may be used to distinguish members of a population based upon size differences of the intergenic region between the 16S-23S or 23S-5S rDNA genes. The spacer region between these genes has been found to be hypervariable in microbial populations. In the case of the 16S-23S intergenic region, the spacer size ranges between about 200-1500 base-pairs depending the presence or absence of various tRNA genes. (Nour M. 1998, 16S-23S and 23S-5S intergenic spacer regions of lactobacilli: nucleotide sequence, secondary structure and comparative analysis, Res. Microbial. 149(6):433-448; Berthier F. et al., 1998, Rapid species identification within two groups of closely related lactobacilli using PCR primers that target the 16S/23S rRNA spacer region, FEMSMicrobiol Lett. 161(1):97-106; Tilsala-Timisjarvi A. et al., 1997, Development of oligonucleotide primers from the 16S-23S rRNA intergenic sequences for identifying different dairy and probiotic lactic acid bacteria by PCR Int. J. Food Microbial. 35(1):49-56). Both of these rDNA genes are transcribed on the same operon. Therefore, the conserved regions of the rDNA coding sequence of these subunits can be utilized to amplify the intergenic regions.

[0096] In order that this invention may be better understood, the following examples are set forth. These examples are for purposes of illustration only and are not to be construed as limiting the scope of the invention in any manner.

EXAMPLE 1

Serial Analysis of rDNA Polymorphic Tags from the Domain *Eubacteria*

[0097] A sample comprising environmental bacteria was obtained and total DNA was extracted from the sample. To amplify the DNA, PCR was performed by mixing 5 μ L 10X

Advantage2 reaction buffer (Clontech), 2.5 μ L dNTPs, 5 μ L 8 μ M TX9/TX16 primers, 50 ng of sample DNA, 0.5 μ L Advantage2 Taq polymerase (Clontech) and water to 50 μ L. Primer TX9 was biotinylated. The primers are shown in Figure 12 and a general strategy of this example is shown in Figure 13. The reaction mixture was then subjected to PCR under the following conditions:

1. a) 94°C for 5 minutes;
2. b) 94°C for 1 minute, 10 seconds;
3. c) 55°C for 50 seconds;
4. d) 68°C for 1 minute;
5. e) repeat steps (b) through (d) 20 times; and
6. f) 68°C for 2 minutes.

[0098] The approximately 600 basepair PCR product was gel purified using 1% agarose and a Qiaquick kit (Qiagen) according to manufacturer's instructions. The DNA was eluted with 50 μ L Tris-EDTA (TE; 10 mM Tris pH 8; 1 mM EDTA). To restrict the amplified DNA, 10X New England Biolabs Buffer #2 (NEB#2) was added to bring the concentration of the buffer to 1X and 25 units of *AluI* was added. The reaction mixture was incubated for 2 hours at 37°C; 25 additional units of *AluI* was added and the reaction mixture was incubated for a further one hour followed by inactivation of the enzyme at 65°C for 20 minutes.

[0099] In order to immobilize the restricted DNA fragment to a bead, an equal volume of 2X BW buffer (2X BW = 10 mM Tris pH 7.5; 1 mM EDTA; 2 M NaCl) was added to the reaction mixture, followed by addition of 50 μ L washed M-280 Streptavidin (SA) beads (Dyna). The reaction mixture was incubated for 20 minutes at room temperature. The beads were then washed twice in 1X BW and twice in wash buffer (10 mM Tris pH 8; 10 mM MgSO₄; 50 mM NaCl). During the last wash step, the beads were split into two pools.

[0100] To add the linkers to the immobilized DNA, one pool of beads were resuspended in 4 μ L of 10 μ M linker TX-12/13 (a double-stranded DNA molecule comprising primers TX-012 and TX-013) to form pool A, while the other pool (pool B) were resuspended in 4 μ L of 10 μ M linker TX-14/15 (a double-stranded DNA molecule comprising primers TX-014 and TX-015) was added to the other pool. 36 μ L of T4 ligase mix (4 μ L 10X ligase buffer; 32 μ L water; 0.2 μ L T4 ligase [400 U]) were added to the linker/bead mixture and the mixture was incubated overnight at 16°C.

[0101] The DNA molecule attached to the streptavidin beads was then incubated with *BpmI*, which recognizes its specific restriction sequence (GAGGTC) in the DNA molecule. *BpmI* cleaves the DNA such that it releases the streptavidin bead from the DNA molecule and incorporates a part of the polymorphic region of the DNA molecule first amplified. The beads were washed twice with 0.5 mL each 1X BW and twice with wash buffer. The beads were resuspended in 10 μ L *BpmI* mix (1 mL 10X NEB#3 [New England Biolabs]; 1 μ L 1 μ g/ μ L bovine serum albumin; 8 μ L water and 1 μ L *BpmI* [New England Biolabs]). The reaction mixture was

incubated at 37°C for 2 hours and then at 65°C to inactivate the enzyme. The supernatant containing the DNA tags was then isolated.

[0102] In order to remove the 3' overhang on the DNA tags and make them blunt ended, to the supernatant (10 µL) on ice was added 10 µL T4 polymerase mix (1 µL 10X NEB#2; 0.5 µL 4 mM dNTPs; 8.5 µL water and 0.33 µL T4 polymerase [1 U; New England Biolabs]). The reaction was incubated at 12°C for 20 minutes and at 65°C for 20 minutes to inactivate the polymerase. In order to form ditags, pool A and pool B were recombined to give a total volume of 40 µL. Four µL 10 mM rATP and 0.2 µL T4 DNA ligase (400 U) was added and the reaction was incubated 4 hours to overnight at 16°C.

[0103] As an intermediate amplification step, the ditags were then amplified in a 300 µL PCR reaction (30 µL 10X Advantage2 Taq buffer; 15 µL 4 mM dNTPs; 30 µL 8 µM TX111/TX121 primer mix; 3 µL Advantage2 Taq polymerase; 3 µL ditag template; 219 µL water). The 300 µL reaction mix was split into three 100 µL reactions and amplified using the following conditions:

1. a) 94°C for 5 minutes;
2. b) 94°C for 30 seconds;
3. c) 56°C for 30 seconds;
4. d) 68°C for 40 seconds;
5. e) repeat steps (b) through (d) for 15 cycles; and
6. f) 68° for 2 minutes.

After amplification, three volumes (900 µL) QG buffer (Qiagen) and one volume (300 µL) isopropanol was added, the mixture was bound to a Gel Extraction Spin Column (Qiagen) and the DNA was eluted with 50 µL TE.

[0104] The amplified DNA was then subjected to non-denaturing polyacrylamide gel electrophoresis (PAGE; Novex 1 mm 10% TBS-PAGE gel). The gel was stained with 5 µg/mL ethidium bromide, the approximately 106 basepair ditags were excised from the gel, the excised gel was fragmented and 0.3 mL of TE was added to the gel and incubated at 65°C for 30 minutes. The gel/TE mixture was transferred to a miniprep spin column (Qiagen) and the eluate containing the amplified ditags was collected (the ditag template). This amplified DNA was called the P300 ditag template.

[0105] In order to determine the optimal number of PCR cycles for large-scale amplification (PCR cycle titration), 6.4 µL P300 ditag template was mixed with 8 µL 10X Advantage2 Taq buffer, 4 µL 4 mM dNTPS, 8 µL TX111/TX121 primers, 0.8 µL Advantage2 Taq polymerase and 52.8 µL water. The reaction mixture was split into three 25 µL reactions and amplified for 6, 8 or 10 cycles using the PCR conditions described above for the intermediate amplification step. After amplification, the DNA products were desalted and purified using a Qiagen Gel Extraction Spin Column as described above except that only 75 µL QG buffer and 25 µL isopropanol were used, and the DNA products were eluted with 20 µL TE. The cycle number that produced the largest 106 basepair ditag yield without any detectable vertical smearing by 10% TBS-PAGE analysis was used for large-scale amplification.

[0106] For large-scale ditag amplification, 180 μ L 10X Advantage2 Taq buffer, 90 μ L 4mM dNTPs, 180 μ L 8 μ M TX111/TX121 mixture, 18 μ L Advantage2 Taq polymerase, 144 μ L P300 ditag template and 11881 μ L water was mixed together and then split into 18 100 μ L reactions. PCR was performed as described immediately above using the number of cycles that were determined from the previous titration step. The PCR reactions were desalted by adding three volumes QG buffer and one volume isopropanol to the reactions and passed through a total of four Qiagen Gel Extraction Spin Columns. Each column was eluted with 50 μ L TE. The samples were subjected to non-denaturing preparative PAGE (Novex 1.5 mm 10% TBS-PAGE). The DNA bands were stained and excised as above and split into three tubes. The polyacrylamide was fragmented and eluted using TE as described above and the DNA was purified by passing the contents of the gel/TE mixture through a Qiagen Plasmid Spin Miniprep column as described above.

[0107] To precipitate the DNA, 1 μ L glycogen, 150 μ L 10M ammonium acetate and 1125 μ L ethanol was added to each eluate of approximately 300 μ L. The mixture was incubated at -80°C for 20 minutes and microfuged at 13,000 rpm at 4°C for 15 minutes. The pellet was washed with 1 mL 70% ethanol and dried at 37°C for 5 minutes.

[0108] The DNA pellets were resuspended in 150 μ L *AluI* mix (15 μ L NEB#2, 135 μ L water and 15 μ L *AluI* [150 U]). The reaction mixture was incubated at 37°C for 2 hours. 150 μ L of 2X BW buffer was added and the sample was transferred to a new tube containing 150 packed SA beads. The beads were incubated with the DNA mixture for 30 minutes at room temperature and the beads were pelleted magnetically. The supernatant was removed, extracted with 250 μ L phenol/chloroform/isoamyl alcohol, and the aqueous phase was precipitated with 150 μ L 10M ammonium acetate and 1125 μ L ethanol overnight at -80°C . The DNA was centrifuged at 13,000 rpm at 4°C for 15 minutes, the pellet washed with 70% ethanol and dried. This step removes any free linkers that were present as well as any incompletely digested DNA products that were still bound to the biotinylated linkers.

[0109] In order to concatenate the ditags, the DNA pellet was resuspended in 10 μ L ligase mix (1 μ L 10X ligase buffer; 9 μ L water and 1 μ L T4 ligase [100 U]) and incubated at 16°C for 30 minutes. The DNA was precipitated by added 40 μ L TE, 25 μ L 10M ammonium acetate and 180 μ L ethanol. The DNA was precipitated for 15 minutes at -80°C and centrifuged, washed and dried as described above.

[0110] The concatenated ditags were purified by resuspending the pellet in 1X TBE loading buffer and subjecting the sample to non-denaturing PAGE (8% TBE-PAGE). The area between 0.5 to 1.2 kb was excised and the DNA was eluted from the gel into 300 μ L TE as described above. The DNA was precipitated using 1 μ L glycogen, 150 μ L 10M ammonium acetate and 1125 μ L ethanol. The DNA was precipitated, centrifuged, washed and dried as before.

[0111] The concatenated ditags were cloned in pUC19 by resuspending the DNA pellet in 3 μ L *SmaI*-cut pUC 19 (approximately 100 ng plasmid DNA) and adding 7 μ L T4 ligase mix (1 μ L

10X ligase buffer, 6 μ L water and 0.2 μ L T4 ligase [400 U]). The plasmid/ditag mixture was incubated for 2 hours at 16°C and 1 μ L of the mixture was used to transform 100 μ L chemically-competent DH10B. Amp-resistant transformants were screened by PCR using pUC/m13 forward/reverse 17mers as PCR primers. Transformants containing concatemerized ditags were then sequenced.

EXAMPLE 2

SARD Analysis of a Defined Population

[0112] SARD was performed essentially as described in Example 1. In this example, commercially available bacterial genomic DNA samples were mixed at an equal concentration (weight/volume). The bacterial DNA samples used included *Bacillus subtilis*, *Clostridium perfringens*, *Escherchia coli*, *Lactococcus lactis* and *Streptomyces coelicolor*. Equal volumes of the DNA samples (total genomic DNA) were mixed at a concentration of 50 nD/ μ L, each. Chart I shows the size of the genomes of each bacterial species, the number of 16S rDNA copies per genome and the molar percentage of 16S copies for each bacterial species in the total DNA sample.

Chart I

Bacteria	Genome (Mb)	16S Copies/Genome	Molar %
<i>B. subtilis</i>	4.2	10	17.2
<i>C. perfringens</i>	4.4	10	16.4
<i>E. coli</i>	4.6	7	11.0
<i>L. lactis</i>	2.4	10	30.7
<i>S. coelicolor</i>	8.0	6	5.4

[0113] After SARD, 120 tags were sequenced. Chart II shows the expected number and percentage of tags compared to the observed number and percentage of tags for the population. See also Fig. 14.

Chart II

Bacteria	Expected Tag Number	Observed Tag Number	Expected Tag Percentage	Observed Tag Percentage
<i>B. subtilis</i>	21	44	17.2%	36.7%
<i>C. perfringens</i>	20	18	16.4%	15.0%
<i>E. coli</i>	13	8	11.0%	6.7%
<i>L. lactis</i>	37	35	30.7%	29.2%
<i>S.</i>	6	6	5.4%	5.0%

Bacteria	Expected Tag Number	Observed Tag Number	Expected Tag Percentage	Observed Tag Percentage
<i>coelicolor</i>				

[0114] Each of the rDNA genes from these species produced a SARD tag that was distinguishable from the other members of the set. As can be observed, approximately twice as many tags that corresponded to 16S rDNA from *B. subtilis* were found than was expected based upon the molar percentage of 16S rDNA from *B. subtilis*. The observation that *B. subtilis* appeared to be twice as abundant as was expected has been reported previously. Farrelly et al. (Farrelly, V. F. et al., 1995, Effect of genome size and *rrn* gene copy number on PCR amplification of 16S rRNA genes from a mixture of bacterial species. Appl. Environ. Microbiol. 61(7):2798-2801) described similar results when PCR amplifying the 16S rDNA gene from mixed populations of genomic DNA. They concluded that this phenomenon was the result of the tandem organization of the *rrn* operons in the *B. subtilis* genome where multiple rDNA genes may be amplified as a single product. The remaining tags were found at abundances close (<40% deviation) to their expected values.

EXAMPLE 3

SHARD Analysis of Environmental Bacterial Diversity

[0115] In order to demonstrate that the SARD method could be used to survey environmental bacterial diversity, total DNA was extracted from two soil samples (Wy-1 and Wy-2) taken from the Rocky Mountain Oilfield Testing Center (RMOTC, Casper, Wyoming) in October, 2000. The samples were collected about 0.5 miles apart and from a depth of 14-18 inches. The environmental DNA samples were subjected to SARD analysis as described in Example 1. In a preliminary analysis, 148 tags were identified from Wy-1 and 234 tags were identified from Wy-2 (Figures 15 and 16, respectively).

[0116] In the Wy-1 sample, 58 distinct tags were identified and the abundance of each tag varied. The most abundant tag (ATGGCTGTCGTCAGCT) made up about 34% of the population. This tag sequence is identical to many bacterial sequences in GenBank and its position within the 16S rDNA gene indicates that it is located in a conserved region located distal to the targeted *AluI* restriction site. In other words, the contributing 16S gene(s) for this tag did not contain the conserved *AZuI* site. Since the SARD tag position is dictated by the first *AluI* site distal to the biotinylated primer used in the initial PCR reaction, it is likely that the first *AluI* site in the contributing 16S gene(s) was located downstream within a conserved region. In order to decrease the number of tags that do not contain the conserved *AluI* site next to the polymorphic region, one may gel purify the approximately 100 basepair PCR products after the first *AluI* restriction step. However, this may result in losing some information. Nevertheless,

39% of the tags (58/148) in this set were different from each other. See Figures 15 and 17.

[0117] The Wy-2 sample was found to contain 79 different tags out of a total of 234 tags that were examined. Thus, 34% of the tags (79/234) in this set were different from each other. See Figures 16 and 17. As in the case with Wy-1, the tag ATGGCTGTCGTCAGCT, which represents a conserved sequence in a 16S rDNA gene, was most abundant and made up about 30% of the population.

[0118] Combining the tags from the two sets reveals a total of 105 different tags from a total of 382 tags. Thus, 26 of 58 different tags (45%) in Wy-1 were not present in Wy-2. Likewise, 47 of 79 different tags (59%) in Wy-2 were not present in Wy-1. The tags that were only found in one of the samples are candidates for bacteria with indicator value for various parameters associated with each sample. However, there was no attempt in this preliminary analysis to obtain all of the tags present in the two samples, so it cannot be concluded that some or most of the tags found in one sample are not present in the other sample. Thus, one cannot conclude that there are tags in these two samples that are indicators for various parameters associated with each sample. Nonetheless, a full-fledged analysis of these samples may provide such indicators.

EXAMPLE 4

Serial Analysis of rDNA Polymorphic Tags From the Domain *Archaea*

[0119] The method described in Example 1 can also be applied to the domain *Archaea*. The domain *Archaea* is made up of two known kingdoms, Euryarchaeota and Crenarchaeota (Pace, N. R. 1997, A molecular view of microbial diversity and the biosphere. Science 276:734-740). One set of oligonucleotides and restriction enzymes can be used to survey both of these domains.

[0120] In this example, the following oligonucleotides are designed: 5' biotin-TA(CT)T(CT)CCCA(GA)GCGG(CT)(GCT)(GC)(GA)CTT(AGCT)-3' corresponding to position 817-838 of the *Methanococcus jannaschii* 16S rDNA gene (GenBank Accession number M59126), and (5'-GGTG(TGC)CA(GC)C(CA) GCCGCGGTAA(TC)ACC(AGCT)-3' corresponding to position 457-481 of the *Methanococcus jannaschii* 16S rDNA gene.

[0121] In this example, a *Bfal* site (CTAG) is utilized that corresponds to position 768-771 of the *Methanococcus jannaschii* 16S rDNA gene. This site is immediately flanking a polymorphic region.

[0122] The SARD method was tested *in silico* using 17 representatives from the domain Archaea (Figure 4). 15 of the SARD tags that would be identified were all unique to this set

(Table II). Two species of the genus *Methanobacterium* did not possess any *B*gl sites in the region that would be amplified and therefore, would not produce any tags.

EXAMPLE 5

Surveying PCR-Amplified Polymorphic rDNA Regions

[0123] Oligonucleotide primers that are complementary to conserved regions immediately flanking a polymorphic region could be used to amplify the polymorphic DNA sequences. The oligonucleotide primer sequences could include existing or introduced restriction sites to enable subsequence cloning into a bacterial vector. By utilizing or introducing different restriction sites in each primer, the restriction digested PCR products could be concatemerized in a unidirectional fashion prior to cloning. This step would allow a serial sequence analysis of multiple Polymorphic regions from a single recombinant product.

EXAMPLE 6

Surveying Translation Products of 16S rDNA Polymorphic Regions

[0124] The polymorphic regions of 16S rDNA could also be surveyed by the parallel identification of the translation products of a given polymorphic region. Oligonucleotides could be designed such that they can serve to amplify a polymorphic region. The 5' primer would also include a T7 polymerase binding site or other polymerase binding site, a Kozak consensus sequence, an initiator ATG codon and an epitope to facilitate purification. Examples of epitopes include hemagglutinin (HA), *myc*, Flag or polyhistidine. Following amplification, the products are subjected to *in vitro* transcription/translation to produce the peptide products. These peptides are purified from the cell extract and analyzed by mass spectrometry. This type of an approach has been applied to the identification of mutations in the BRCA1 gene (Garvin, A M. et al., 2000 MALDI-TOF based mutation detection using tagged in vitro synthesized peptides. Nature Biotechnol. 18:95-97).

[0125] Although this process would not readily provide DNA sequence information from which to deduce taxonomy, it would allow for the creation of microbial diversity profiles comprised of 'mass tags'. These mass tags could be used to identify correlations between specific tags and various sample parameters. To test the information content of this approach, a polymorphic region of the 16S rDNA genes from the species in Figure 3 was translated *in silico*. Of the 34 polymorphic regions examined, 32 produced a tag with a unique mass in this set (Table III).

[0126] Another approach to translating the amplified polymorphic regions would be to clone

and express the sequences in whole cells. For instance, oligonucleotides could be designed to amplify polymorphic regions that include sequences at the 5' ends that would allow for cloning by homologous recombination in yeast. These sequences could be cloned into an expression cassette to create a fusion between a secreted protein, such as alpha factor or invertase, and an epitope to facilitate purification and the translated amino acid sequence of the rDNA polymorphic region. Homologous recombination in yeast is quite robust and could easily enable the isolation of 10^3 - 10^4 independent recombinants on a single transformation plate. The secreted products could be isolated from the medium and identified by mass spectrometry.

EXAMPLE 7

Hybridization of Microbial rDNA to Immobilized Oligonucleotides

[0127] A complication with using short oligonucleotide probes in DNA microarrays is the instability of short oligonucleotides duplexes. A possible solution to this problem is to synthesize probes that include degenerate sequences to accommodate unknown sequences. The length of the oligonucleotide is dictated by the number of degeneracies, or sequence permutations, that are to be accommodated. For instance, a fully degenerate 9mer oligonucleotide requires 4^9 or 262,144 different oligonucleotide sequences. The efficient hybridization of 9mer oligonucleotides is not possible using standard conditions. One solution has been to incorporate a ligation step in the hybridization of a target sequence to a 9mer oligonucleotide probe (Gunderson, K. L. et al., 1998, Mutation detection by ligation to complete n-mer DNA arrays. *Genome Res.* 8:1142-1153).

[0128] Another solution could be to effectively increase the length of the oligonucleotide by including nucleotides in the primer that are not degenerate. This approach could be applied to a survey of a microbial community by constructing oligonucleotide probes that are composed of a constant region that corresponds to a well conserved region of a 16S rDNA gene together with a degenerate sequence that corresponds to the flanking polymorphic region.

[0129] An example of such a collection of degenerate oligonucleotides for the domain Bacteria could include permutations of the following primer: 5'-AACGAGCGCAACNNNNNNNNNN-3', where N indicates any nucleotide at that position. This sequence corresponds to position 1101-1122 of the *E. coli* 16S rDNA gene (GenBank Accession number E05133). Alternatively, the primers could be designed such that they are composed of a mixture of constant sequence, semi-degenerate positions (e.g. A or G) and degenerate positions (e.g. A, G, C or T).

Position	814-829	810-825	1021-1036	834-849	828-843	788-803	765-780	824-839	826-841	782-797	783-798	837-852	789-804	861-876	854-869	790-805	1029-1044	848-863	813-828	521-536	521-536	521-536	521-536	319-334	779-794	687-702	798-813	779-794	316-331	226-241	830-845
-----------------	---------	---------	-----------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	-----------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------

Table 1*

	Species	GenBank Acc#	Tag Sequence
804-819	<i>Desulfurobacterium thermolithotrophum</i>	AJ001049	GTCCAGTGCCTGAGCT
796-811	Uncultured Aquificales OPS132	AF027104	GTCCGTCGCCGTAAGCT
796-811	<i>Bacterioides caccae</i>	X83951	ATGCTGTGCTGCTGAGCT
317-332	<i>Actinomyces bovis</i>	X81061	TTTCCGCGCCGTAAGCT
796-811	<i>Actinomyces neyeri</i>	X82451	TTTCTGCGCCGTAAGCT
799-814	<i>Denitrobacterium detoxificans</i>	AF079507	CCTCCGCGCCGTAAGCT
778-793	Uncultured GNS bacteria BPC110	AF154084	CCCGTAGTCCCTAGCT
	Uncultured GNS bacteria GCA004	AF154104	CATCGGTGCCGTAAGCT
	Uncultured GNS bacteria GCA112	AF154100	CGGCGGTGCCGTAAGCT
	<i>Acetobacter acetii</i>	AF127399	ACTCAGTGTGCTAGCT
	<i>Gluconobacter asatii</i>	AB024492	ACTCAGTGTGCTAGCT
	<i>Burkholderia</i> sp. JBI	X92188	CCTTAGTAACTGAGCT
	<i>Denitrobacter permanens</i>	Y12639	ATGCTGTGCTGCTAGCT
	<i>Desulfobacter curvatus</i>	M34413	CTGCTGTGCTGCTAGCT
	<i>Desulfobulbus</i> sp. BG25	U85473	CCTCTGTGCTGCTAGCT
	<i>Legionella anisa</i>	X73394	ATGCTGTGCTGCTAGCT
	Benzene mineralizing clone SB-1	AF029039	ATGCTGTGCTGCTAGCT
	<i>Escherichia coli</i>	E05133	CGTGGCTTCCGTAAGCT
	Uncultured Acidobacterium Sub.Div-1	X68464	CCGCGCTGCCGTAAGCT
	Uncultured Acidobacterium Sub.Div-1	Z73363	CGGCTGTGCTGCTAGCT
	Uncultured Acidobacterium Sub.Div-1	Z73365	CCACTGTGCTGCTAGCT
	Uncultured Acidobacterium Sub.Div-1	Z73368	CTGCTGTGCTGCTAGCT
	Uncultured Acidobacterium Sub.Div-1	Z73364	CTGCGGTGCCGTAAGCT
	Uncultured Acidobacterium Sub.Div-1	U68659	CCAAATGTGCTGCTAGCT
	Uncultured Acidobacterium Sub.Div-1	D26171	CCGTCTGTGCTGCTAGCT
	Uncultured Acidobacterium Sub.Div-1	X97101	CCGTCTGTGCTGCTAGCT
	Uncultured Acidobacterium Sub.Div-1	X97098	CTGCCGTGTGCTAGCT
	Uncultured Acidobacterium Sub.Div-1	AF047646	CTCCCGTGTGCTAGCT
	Uncultured Acidobacterium Sub.Div-1	AF050548	CCGCCGTGCTGCTAGCT
	Uncultured Acidobacterium Sub.Div-2	U68612	CTGAGGACGTAAGCT
	Uncultured Acidobacterium Sub.Div-2	X07646	GTGTCGTGCTGCTAGCT

ey are not unique to this sel.

X97097
 X68466
 X68468
 U68648
 X68467
 AF0135
 AF02701

Uncultured Acidobacterium Sub.Div-3
 Uncultured Acidobacterium Sub.Div-3
 Uncultured Acidobacterium Sub.Div-3
 Uncultured Acidobacterium Sub.Div-3
 Uncultured Acidobacterium Sub.Div-3
 Uncultured Acidobacterium Sub.Div-3
 Uncultured Acidobacterium Sub.Div-3

* Sequences shown in bold with shadow indicates th

Table II

Species	GenBank Acc#	Tag Sequence	Position
Crenarchaeota			
<i>Aeropyrum pernix</i>	D83259	CTAGGGGGCGGGAG	614-627
<i>Desulfurococcus mobiles</i>	M36474	CTAGGTGTTGGGTG	856-869
<i>Staphylothermus marinus</i>	X99560	CTAGGTGTTGGGCG	770-783
<i>Metallosphaera sedula</i>	X90481	CTAGGTGTCGCGTA	756-769
<i>Sulfolobus acidocaldarius</i>	D14053	CTAGGTGTCGAGTA	785-798
<i>Sulfolobus metallicus</i>	D85519	CTAGGTGTCACGTG	744-757
<i>Caldivirga maquilingensis</i>	AB013926	CTAGCTGTTGGGTG	773-786
<i>Pyrobaculum islandicum</i>	L07511	CTAGCTGTCGGCCG	781-794
Euryarchaeota			
<i>Archaeoglobus fulgidus</i>	X05567	CTAGGTGTCACCGA	780-793
<i>Archaeoglobus veneficus</i>	Y10011	CTAGGTGTCACCGG	758-771
<i>Haloarcula japonica</i>	D28872	CTAGGTGTGGCGTA	762-775
<i>Halococcus morrhuae</i>	D11106	CTAGGTGTGGCGTT	765-778
<i>Methanococcus jannaschii</i>	M59126	CTAGGTGTCGCGTC	768-781
<i>Methanobacterium bryantii</i>	AF028688	None	
<i>Methanobacterium subterraneum</i>	X99045	None	
<i>Pyrococcus abyssi</i>	Z70246	CTAGGTGTCGGGCG	767-780
<i>Picrophilus oshimae</i>	X84901	CTAGCTGTAAACTC	742-755

Table III

Species	GenBank Acc#	Peptide Sequence	M.W.	Position
<i>Desulfurobacterium thermolithotrophum</i>	AJ001049	RAQPLSLVASG*	1097.40	1079-1136
Uncultured Aquificales OPS132	AF027104	RAQPLSCVTSG*	1117.40	1074-1131
<i>Bacteroides caccae</i>	X83951	RAQPLSSVTNRSC*	1417.70	1069-1126
<i>Actinomyces bovis</i>	X81061	RAQPLSRVASTLWWGLAGD	2083.60	1088-1145
<i>Actinomyces meyeri</i>	X82451	RAQPLPYVASTLWWGLVGD	2128.60	1082-1139
<i>Denitrobacterium detoxificans</i>	AF079507	RAQPLPHVASIRLGTHGG	1866.50	1039-1094
Uncultured GNS bacteria BPC110	AF154084	RAQPLLYVIRVIPD	1652.10	1074-1116
Uncultured GNS bacteria GCA004	AF154104	RAQPSLYVTRIIRD	1687.10	1080-1122
Uncultured GNS bacteria GCA112	AF154100	RAQPSPYVIRVIRD	1669.00	1082-1124
<i>Acetobacter aceti</i>	AF127399	RAQPLSLVASMFGWAL*	1746.30	1038-1095
<i>Gluconobacter asaii</i>	AB024492	RAQPLSLVASTFRWAL*	1815.30	1034-1092
<i>Burkholderia</i> sp. JB1	X92188	RAQPLSLVATQEHSRET	1922.20	1094-1144
<i>Denitrobacter permanens</i>	Y12639	RAQPLPLVATFSWAL*	1669.10	1077-1131
<i>Desulfobacter curvatus</i>	M34413	RAQPLSLVASTLCGNSNET	1960.40	1116-1172
<i>Desulfobulbus</i> sp. BG25	U85473	RAQPLPLVASSSAGHSGKT	1863.40	1114-1170
<i>Legionella anisa</i>	X73394	RAQPLSLVAST*	1141.40	1078-1135
Benzene mineralizing clone SB-1	AF029039	RAQPLPLVANRSSWGL*	1764.20	1077-1134
<i>Escherichia coli</i>	E05133	RAQPLSFVASGPAGNSKET	1916.30	1103-1159
Uncultured Acidobacterium Sub.Div-1	Z73363	RAQPLSLVASGSSRAL*	1612.10	775-832

Species	GenBank Acc#	Peptide Sequence	M.W.	Position
Uncultured Acidobacterium Sub.Div-1	Z73365	RAQPLSSVAIGSSRATLAK	1912.50	777-835
Uncultured Acidobacterium Sub.Div-1	Z73368	RAQPLFASCHH*	1933.50	779-835
Uncultured Acidobacterium Sub.Div-1	Z73364	RAQPLFAQLPSFSWALCRN	2204.80	778-835
Uncultured Acidobacterium Sub.Div-1	U68659	RAQPLLPXAI*	1218.70	573-630
Uncultured Acidobacterium Sub.Div-1	D26171	RAQPLLPVATI*	1177.50	1035- 1090
Uncultured Acidobacterium Sub.Div-1	X97101	RAQPLSPVAIL*	1163.50	943-998
Uncultured Acidobacterium Sub.Div-1	X97098	RAQPLSSVATI*	1141.40	1054- 1109
Uncultured Acidobacterium Sub.Div-1	AP047646	RAQPLFLVATI*	1227.60	1035- 1090
Uncultured Acidobacterium Sub.Div-1	AF050548	RAQPSSLVANTLW*	1441.70	572-629
Uncultured Acidobacterium Sub.Div-2	U68612	RAQPLUVVATRKRELYVD	2150.60	577-630
Uncultured Acidobacterium Sub.Div-2	Y07646	RAQPLHWATPQGGTLRG	1857.30	1085- 1140
Uncultured Acidobacterium Sub.Div-3	X97097	RAQPSSLVANPQGKHPKGT	1972.40	1060- 1116
Uncultured Acidobacterium Sub.Div-3	U68648	RARPLSCVAIL*	1197.70	574-629
Uncultured Acidobacterium Sub.Div-3	AF013515	RAQPLSCVANPQGCTLRR	1969.50	1057- 1112
Uncultured	AF027004	RAQPSPCVATPPRAGALSGD	1950.40	1036-

Species	GenBank Acc#	Peptide Sequence	M.W.	Position
Acidobacterium Sub.Div-3				1096
* Indicates an in-frame stop codon was encountered within the polymorphic sequence.				

SEQUENCE LISTING**[0130]**

<110> Matthew, Ashby N.

<120> Methods for the Survey and Genetic Analysis of Populations

<130> ASHBY-1

<140> PCT/US01/11609

<141> 2001-04-10

<150> US 60/196063

<151> 2000-04-10

<150> US 60/196258

<151> 2000-04-11

<160> 244

<170> PatentIn version 3.1

<210> 1

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 1

acgatgagca ctagct 16

<210> 2

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 2

acgatgagta ctagct 16

<210> 3

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 3

acgatgatga ctagct 16

<210> 4

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 4

acgatggatg ctagct 16

<210> 5

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 5

atgctagtct ggagct 16

<210> 6

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 6

atggctgtcg tcagct 16

<210> 7

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 7

atggttgtcg tcagct 16

<210> 8

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 8

attccgtgcc gtagct 16

<210> 9

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 9

cactagtggc gcagct 16

<210> 10

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 10

cccccggtgcc gaagct 16

<210> 11

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 11

cccccggtgcc gcagct 16

<210> 12

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 12

cccccttcct ccagct 16

<210> 13

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 13

ccccggtgcc gcagct 16

<210> 14

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 14

ccgggtagtc ccagct 16

<210> 15

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 15

cctccgtgcc gaagct 16

<210> 16

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 16

cctccgtgcc gcagct 16

<210> 17

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 17

cctccgtgct gcagct 16

<210> 18

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 18

cctcggcgcc gcagct 16

<210> 19

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 19

cctcggtgcc gcagct 16

<210> 20

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 20

cctcgggtgc gcagct 16

<210> 21

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 21

cctgggtgcc gcagct 16

<210> 22

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 22

cctgtgtgac gaagct 16

<210> 23

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 23

ccttggtaac gaagct 16

<210> 24

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 24

ccttggtacc gaagct 16

<210> 25

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 25

cgccagtgcc gtagct 16

<210> 26

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 26

cgctgtgcc gtagct 16

<210> 27

<211> 16

<212> DNA

<213> Unknown

<220>

<223> unidentified soil organism

<400> 27

cgtccgtgcc gaagct 16

<210> 28

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 28

cgtccgtgcc gcagct 16

<210> 29

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 29

cgtcggtgcc gcagct 16

<210> 30

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 30

ctcccgtgcc gcagct 16

<210> 31

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 31

ctcccgtgcc ggagct 16

<210> 32

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 32

ctccggtgcc gcagct 16

<210> 33

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 33

ctcctgtgcc gaagct 16

<210> 34

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 34

ctcctgtgcc gcagct 16

<210> 35

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 35

ctgccgtgcc gaagct 16

<210> 36

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 36

ctgctgtgcc gaagct 16

<210> 37

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 37

ctgtcgtgcc gaagct 16

<210> 38

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 38

cttcagtatc gaagct 16

<210> 39

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 39

cttccgcgcc ggagct 16

<210> 40

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 40

cttccgtgcc gcagct 16

<210> 41

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 41

cttccgtgcc ggagct 16

<210> 42

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 42

cttcggtgcc gcagct 16

<210> 43

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 43

cttctgtggc gaagct 16

<210> 44

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 44

gatccgtgcc gtagct 16

<210> 45

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 45

gctctgtgcc gaagct 16

<210> 46

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 46

gctgggtgcc caagct 16

<210> 47

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 47

ggtccgtgcc gcagct 16

<210> 48

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 48

ggtgctcttc ggagct 16

<210> 49

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 49

gtaaacgatg gaagct 16

<210> 50

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 50

gtggctgtcg tcagct 16

<210> 51

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 51

gttccgtgcc gaagct 16

<210> 52

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 52

gttccgtgcc gcagct 16

<210> 53

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 53

tatcagtggc gcagct 16

<210> 54

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 54

tctccgtgcc gcagct 16

<210> 55

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 55

tctctgtgcc gcagct 16

<210> 56

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 56

tctctgtgcc gtagct 16

<210> 57

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 57

tgtccgtgcc gtagct 16

<210> 58

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 58

tttccgtgcc gcagct 16

<210> 59

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 59

acgatgataa ctagct 16

<210> 60

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 60

acgatgatga caagct 16

<210> 61

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 61

acgatgggca ctagct 16

<210> 62

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 62

acggctgtcg tcagct 16

<210> 63

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 63

actacgagcg caagct 16

<210> 64

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 64

acttaatgcg ttagct 16

<210> 65

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 65

atgctagtct ggagct 16

<210> 66

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 66

atggctctcg tcagct 16

<210> 67

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 67

atggctgtcg ccagct 16

<210> 68

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 68

atggctgtcg tcagct 16

<210> 69

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 69

atggtgtcg tcagct 16

<210> 70

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 70

atgtagacct ggagct 16

<210> 71

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 71

attccgtgcc gcagct 16

<210> 72

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 72

attccgtgcc gtagct 16

<210> 73

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 73

cacaagcggg ggagct 16

<210> 74

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 74

cactagtggc gcagct 16

<210> 75

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 75

catccgtgcc gaagct 16

<210> 76

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 76

ccccagggcc caagct 16

<210> 77

<211> 16

<212> DNA

<213> Unknown

<220>

<223> unidentified soil organism

<400> 77

ccccggtgcc gcagct 16

<210> 78

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 78

cccgcgtgcc ggagct 16

<210> 79

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 79

ccgcggtgcc gtagct 16

<210> 80

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 80

ccgggtagtc ccagct 16

<210> 81

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 81

ccgggtagtc ctagct 16

<210> 82

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 82

cctccgtgcc gaagct 16

<210> 83

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 83

cctccgtgcc gcagct 16

<210> 84

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 84

cctccgtgcc ggagct 16

<210> 85

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 85

cctcgtaagg ggagct 16

<210> 86

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 86

cctgggtgcc gcagct 16

<210> 87

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 87

cctggtagtc ccagct 16

<210> 88

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 88

cctggtagtc ctagct 16

<210> 89

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 89

ccttagtaac gcagct 16

<210> 90

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 90

ccttggtaac gaagct 16

<210> 91

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 91

cgccagtgcc gaagct 16

<210> 92

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 92

cgccggtgcc gcagct 16

<210> 93

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 93

cgccctgtgcc gtagct 16

<210> 94

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 94

cgctcgtggc gaagct 16

<210> 95

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 95

cggaggcgtc gtagct 16

<210> 96

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 96

cgtcagtgtc gcagct 16

<210> 97

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 97

cgtccgtgcc gaagct 16

<210> 98

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 98

cgtccgtgcc gcagct 16

<210> 99

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 99

cgtccgtgcc ggagct 16

<210> 100

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 100

cgtcggtgcc gcagct 16

<210> 101

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 101

ctccagtgcc gcagct 16

<210> 102

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 102

ctcccgtgcc acagct 16

<210> 103

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 103

ctcccgtgcc gcagct 16

<210> 104

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 104

ctcccgtgcc ggagct 16

<210> 105

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 105

ctccggtgcc gcagct 16

<210> 106

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 106

ctcctgtgcc gcagct 16

<210> 107

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 107

ctgccgcgcc ggagct 16

<210> 108

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 108

ctgccgtgcc gaagct 16

<210> 109

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 109

ctgccgtgcc taagct 16

<210> 110

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 110

ctgcggtgcc gcagct 16

<210> 111

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 111

ctgctgtgcc gaagct 16

<210> 112

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 112

ctgtcgtgcc gaagct 16

<210> 113

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 113

cttccgcgcc ggagct 16

<210> 114

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 114

ctcccgtgcc gaagct 16

<210> 115

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 115

cttccgtgcc gcagct 16

<210> 116

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 116

cttccgtgcc ggagct 16

<210> 117

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 117

cttcggtgcc gcagct 16

<210> 118

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 118

cttcggtgtc gcagct 16

<210> 119

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 119

cttgggtgcc gcagct 16

<210> 120

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 120

ctttagtaac gcagct 16

<210> 121

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 121

gacccgcaag ggagct 16

<210> 122

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 122

gatccgtgcc gcagct 16

<210> 123

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 123

gctccgtgcc gaagct 16

<210> 124

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 124

gctccgtgcc gtagct 16

<210> 125

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 125

gctctgtgcc gaagct 16

<210> 126

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 126

gctctgtgcc gtagct 16

<210> 127

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 127

gggcttgcg tcagct 16

<210> 128

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 128

gtaaacgatg gaagct 16

<210> 129

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 129

gttccgtgcc gcagct 16

<210> 130

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 130

gttcctgtgcc gtagct 16

<210> 131

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 131

gttctgtgcc gcagct 16

<210> 132

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 132

tctcacgaca cgagct 16

<210> 133

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 133

tctcagtaac gtagct 16

<210> 134

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 134

tctccgtgcc gcagct 16

<210> 135

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 135

tctctgtgcc gcagct 16

<210> 136

<211> 16

<212> DNA

<213> Unknown

<220>

<223> unidentified soil organism

<400> 136

tggacgttgc ggagct 16

<210> 137

<211> 16

<212> DNA

<213> unknown

<220>

<223> unidentified soil organism

<400> 137

tttccgtgcc ggagct 16

<210> 138

<211> 20

<212> DNA

<213> Artificial Sequence

<220>

<223> Oligonucleotide

<220>

<221> modified_base

<222> (1)..(1)

<223> Biotin

<400> 138

gtgtaghrgt gaaatdclya 20

<210> 139

<211> 21

<212> DNA

<213> Artificial Sequence

<220>

<223> Oligonucleotide

<400> 139

ytcacgrcay gagctgacga c 21

<210> 140

<211> 28

<212> DNA

<213> Artificial Sequence

<220>

<223> Oligonucleotide

<220>

<221> modified_base

<222> (1)..(1)

<223> 5' Phosphate

<220>

<221> modified_base

<222> (28)..(28)

<223> Dideoxy

<400> 140

gtccagggtc tacatcctag tcaggacc 28

<210> 141

<211> 24

<212> DNA

<213> Artificial Sequence

<220>

<223> Oligonucleotide

<400> 141

ataggtcctg actaggatgt agac 24

<210> 142

<211> 28

<212> DNA

<213> Artificial Sequence

<220>

<223> Oligonucleotide

<220>

<221> modified_base

<222> (1)..(1)

<223> 5' Phosphate

<220>

<221> modified_base

<222> (28)..(28)

<223> Dideoxy

<400> 142

gctccagact agcatccgct gacttgac 28

<210> 143

<211> 24

<212> DNA

<213> Artificial Sequence

<220>

<223> Oligonucleotide

<400> 143

aatgtcaagt cagcggatgc tagt 24

<210> 144

<211> 21

<212> DNA

<213> Artificial Sequence

<220>

<223> Oligonucleotide

<220>

<221> modified_base

<222> (1)..(1)

<223> Biotin

<400> 144

ggctctgact aggatgtaga c 21

<210> 145
 <211> 21
 <212> DNA
 <213> Artificial Sequence

<220>
 <223> Oligonucleotide

<220>
 <221> modified_base
 <222> (1)..(1)
 <223> Biotin

<400> 145
 gtcaagtcag cggatgctag t 21

<210> 146
 <211> 21
 <212> DNA
 <213> Artificial Sequence

<220>
 <223> Oligonucleotide

<220>
 <221> modified_base
 <222> (1)..(1)
 <223> Biotin

<400> 146
 ggattagawa cccbggtagt c 21

<210> 147
 <211> 30
 <212> DNA
 <213> Artificial Sequence

<220>
 <223> Oligonucleotide

<400> 147
 ataggtcctg actaggatgt agacctggag 30

<210> 148
 <211> 30
 <212> DNA

<213> Artificial Sequence

<220>

<223> Oligonucleotide

<400> 148

aatgtcaagt cagcggatgc tagtctggag 30

<210> 149

<211> 27

<212> DNA

<213> Artificial Sequence

<220>

<223> Oligonucleotide

<220>

<221> modified_base

<222> (1)..(1)

<223> 5' Phosphate

<220>

<221> modified_base

<222> (27)..(27)

<223> Dideoxy

<400> 149

ctccaggtct acatcctagt caggacc 27

<210> 150

<211> 30

<212> DNA

<213> Artificial Sequence

<220>

<223> Oligonucleotide

<400> 150

ataggtcctg actaggatgt agacctggag 30

<210> 151

<211> 27

<212> DNA

<213> Artificial Sequence

<220>

<223> Oligonucleotide

<220>

<221> modified_base

<222> (1)..(1)

<223> 5' Phosphate

<220>

<221> modified_base

<222> (27)..(27)

<223> Dideoxy

<400> 151

ctccagacta gcatccgctg acttgac 27

<210> 152

<211> 30

<212> DNA

<213> Artificial Sequence

<220>

<223> Oligonucleotide

<400> 152

aatgtcaagt cagcggatgc tagtctggag 30

<210> 153

<211> 33

<212> DNA

<213> Artificial Sequence

<220>

<223> Oligonucleotide

<220>

<221> modified_base

<222> (1)..(1)

<223> Biotin

<400> 153

ctcgatagtc acggtcctga claggatgta gac 33

<210> 154

<211> 33

<212> DNA

<213> Artificial Sequence

<220>

<223> Oligonucleotide

<220>

<221> modified_base

<222> (1)..(1)

<223> Biotin

<400> 154

caggttgcaa cggtaagtc agcggatgct agt 33

<210> 155

<211> 22

<212> DNA

<213> Artificial Sequence

<220>

<223> Oligonucleotide

<220>

<221> modified_base

<222> (1)..(1)

<223> biotin

<220>

<221> misc_feature

<222> (22)..(22)

<223> A, G, C or T

<400> 155

taytycccar gcggybsrct tn 22

<210> 156

<211> 25

<212> DNA

<213> Artificial Sequence

<220>

<223> Oligonucleotide

<220>

<221> misc_feature

<222> (25)..(25)

<223> A, G, C or T

<400> 156

ggtgbcascm gccgcggtaa yaccn 25

<210> 157

<211> 22

<212> DNA

<213> Artificial Sequence

<220>

<223> Oligonucleotide

<220>

<221> misc_feature

<222> (14)..(22)

<223> A, G, C or T

<400> 157

aacgagcgca accnnnnnnn nn 22

<210> 158

<211> 16

<212> DNA

<213> Desulfurobacterium thermolithotrophum

<400> 158

gtcagttgcc gaagct 16

<210> 159

<211> 16

<212> DNA

<213> Unknown

<220>

<223> Uncultured Aquificales OPS132

<400> 159

gtccgtgccg taagct 16

<210> 160

<211> 16

<212> DNA

<213> Bacteroides caccae

<400> 160

atggtgtcg tcagct 16

<210> 161

<211> 16

<212> DNA

<213> Actinomyces bovis

<400> 161

tttccgcgcc gtagct 16

<210> 162

<211> 16

<212> DNA

<213> Actinomyces meyeri

<400> 162

tttctgcgcc gtagct 16

<210> 163

<211> 16

<212> DNA

<213> Denitrobacterium detoxificans

<400> 163

cctccgcgcc gcagct 16

<210> 164

<211> 16

<212> DNA

<213> Unknown

<220>

<223> Uncultured GNS bacteria BPC110

<400> 164

cccggtagtc ctagct 16

<210> 165

<211> 16

<212> DNA

<213> Unknown

<220>

<223> Uncultured GNS bacteria GCA004

<400> 165

catcggtgcc gcagct 16

<210> 166

<211> 16

<212> DNA

<213> Unknown

<220>

<223> Uncultured GNS bacteria GCA112

<400> 166

cggcggtgcc gtagct 16

<210> 167
 <211> 16
 <212> DNA
 <213> *Acetobacter aceti*

 <400> 167
 actcagtgtc gtagct 16

 <210> 168
 <211> 16
 <212> DNA
 <213> *Gluconobacter asaii*

 <400> 168
 actcagtgtc gaagct 16

 <210> 169
 <211> 16
 <212> DNA
 <213> Unknown

 <220>
 <223> *Burkholderia* sp. JB1

 <400> 169
 ccttagtaac gaagct 16

 <210> 170
 <211> 16
 <212> DNA
 <213> *Denitrobacter permanens*

 <400> 170
 acgatgtcaa ctagct 16

 <210> 171
 <211> 16
 <212> DNA
 <213> *Desulfobacter curvatus*

 <220>
 <221> misc_feature
 <222> (11)..(11)
 <223> A, G, C or T

 <400> 171
 ctgctgtgcc naagct 16

<210> 172

<211> 16

<212> DNA

<213> Unknown

<220>

<223> Desulfobulbus sp. BG25

<400> 172

cctctgtgtc gcagct 16

<210> 173

<211> 16

<212> DNA

<213> Legionella anisa

<400> 173

acgatgtcaa ctagct 16

<210> 174

<211> 16

<212> DNA

<213> Unknown

<220>

<223> Benzene mineralizing clone SB-1

<400> 174

atgggtgtcg gcagct 16

<210> 175

<211> 16

<212> DNA

<213> Escherichia coli

<400> 175

cgtggcttcc ggagct 16

<210> 176

<211> 16

<212> DNA

<213> Unknown

<220>

<223> Uncultured Acidobacterium Sub.Div-1

<400> 176

ccgccgtgcc gaagct 16

<210> 177

<211> 16

<212> DNA

<213> Unknown

<220>

<223> Uncultured Acidobacterium Sub.Div-1

<400> 177

cggctgtgcc gaagct 16

<210> 178

<211> 16

<212> DNA

<213> Unknown

<220>

<223> Uncultured Acidobacterium Sub.Div-1

<400> 178

ccactgtgcc gtagct 16

<210> 179

<211> 16

<212> DNA

<213> Unknown

<220>

<223> Uncultured Acidobacterium Sub.Div-1

<400> 179

ctgctgtgcc gcagct 16

<210> 180

<211> 16

<212> DNA

<213> Unknown

<220>

<223> Uncultured Acidobacterium Sub.Div-1

<400> 180

ctgccgtgcc ggagct 16

<210> 181

<211> 16

<212> DNA

<213> Unknown

<220>

<223> Uncultured Acidobacterium Sub.Div-1

<400> 181

ccaatgtgcc ggagct 16

<210> 182

<211> 16

<212> DNA

<213> Unknown

<220>

<223> Uncultured Acidobacterium Sub.Div-1

<400> 182

ccgtcgtgcc gtagct 16

<210> 183

<211> 16

<212> DNA

<213> Unknown

<220>

<223> Uncultured Acidobacterium Sub.Div-1

<400> 183

ccgtcgtgtc gtagct 16

<210> 184

<211> 16

<212> DNA

<213> Unknown

<220>

<223> Uncultured Acidobacterium Sub.Div-1

<400> 184

ctgccgtgtc gaagct 16

<210> 185

<211> 16

<212> DNA

<213> Unknown

<220>

<223> Uncultured Acidobacterium Sub.Div-1

<400> 185

ctcccggtgc gaagct 16

<210> 186

<211> 16

<212> DNA

<213> Unknown

<220>

<223> Uncultured Acidobacterium Sub.Div-1

<400> 186

ccgccgtgcc ggagct 16

<210> 187

<211> 16

<212> DNA

<213> Unknown

<220>

<223> Uncultured Acidobacterium Sub.Div-2

<400> 187

ctgaggaacg aaagct 16

<210> 188

<211> 16

<212> DNA

<213> Unknown

<220>

<223> Uncultured Acidobacterium Sub.Div-2

<400> 188

gtgtcgtccc ggagct 16

<210> 189

<211> 16

<212> DNA

<213> Unknown

<220>

<223> Uncultured Acidobacterium Sub.Div-3

<400> 189

gggctgtgcc gaagct 16

<210> 190

<211> 16

<212> DNA

<213> Unknown

<220>

<223> Uncultured Acidobacterium Sub.Div-3

<400> 190

ggtcggtgcc ggagct 16

<210> 191

<211> 16

<212> DNA

<213> Unknown

<220>

<223> Uncultured Acidobacterium Sub.Div-3

<400> 191

ggtcggtgcc agagct 16

<210> 192

<211> 16

<212> DNA

<213> Unknown

<220>

<223> Uncultured Acidobacterium Sub.Div-3

<400> 192

ggttcgtgcc ggagct 16

<210> 193

<211> 16

<212> DNA

<213> Unknown

<220>

<223> Uncultured Acidobacterium Sub.Div-3

<400> 193

tgtctgtgcc ggagct 16

<210> 194

<211> 16

<212> DNA

<213> Unknown

<220>

<223> Uncultured Acidobacterium Sub.Div-3

<400> 194

tatccgtgcc ggagct 16

<210> 195

<211> 16

<212> DNA

<213> Unknown

<220>

<223> Uncultured Acidobacterium Sub.Div-3

<400> 195

ggtccgtgcc ggagct 16

<210> 196

<211> 14

<212> DNA

<213> Aeropyrum pernix

<400> 196

ctagggggcg ggag 14

<210> 197

<211> 14

<212> DNA

<213> Desulfurococcus mobilis

<400> 197

ctaggtgttg ggtg 14

<210> 198

<211> 14

<212> DNA

<213> Staphylothermus marinus

<400> 198

ctaggtgttg ggcg 14

<210> 199

<211> 14

<212> DNA

<213> Metallosphaera sedula

<400> 199

ctaggtgtcg cgta 14

<210> 200
 <211> 14
 <212> DNA
 <213> Sulfolobus acidocaldarius

<400> 200
 ctaggtgtcg agta 14

<210> 201
 <211> 14
 <212> DNA
 <213> Sulfolobus metallicus

<400> 201
 ctaggtgtca cgtg 14

<210> 202
 <211> 14
 <212> DNA
 <213> Caldivirga maquilingensis

<400> 202
 ctagctgttg ggtg 14

<210> 203
 <211> 14
 <212> DNA
 <213> Pyrobaculum islandicum

<400> 203
 ctagctgtcg gccg 14

<210> 204
 <211> 14
 <212> DNA
 <213> Archaeoglobus fulgidus

<400> 204
 ctaggtgtca ccga 14

<210> 205
 <211> 14
 <212> DNA
 <213> Archaeoglobus veneficus

<400> 205
 ctaggtgtca ccgg 14

<210> 206
 <211> 14
 <212> DNA
 <213> Haloarcula japonica

<400> 206
 ctaggtgtgg cgta 14

<210> 207
 <211> 14
 <212> DNA
 <213> Halococcus morrhuae

<400> 207
 ctaggtgtgg cggt 14

<210> 208
 <211> 14
 <212> DNA
 <213> Methanococcus jannaschii

<400> 208
 ctaggtgtcg cgtc 14

<210> 209
 <211> 14
 <212> DNA
 <213> Pyrococcus abyssi

<400> 209
 ctaggtgtcg ggcg 14

<210> 210
 <211> 14
 <212> DNA
 <213> Picrophilus oshimae

<400> 210
 ctagctgtaa actc 14

<210> 211
 <211> 11
 <212> PRT
 <213> Desulfurobacterium thermolithotrophum

<400> 211
 Arg Ala Gln Pro Leu Ser Leu Val Ala Ser Gly
 1 5 10

<210> 212

<211> 11

<212> PRT

<213> Unknown

<220>

<223> Uncultured Aquificales OPS132

<400> 212

Arg Ala Gln Pro Leu Ser Cys Val Thr Ser Gly
1 5 10

<210> 213

<211> 13

<212> PRT

<213> Bacteroides caccae

<400> 213

Arg Ala Gln Pro Leu Ser Ser Val Thr Asn Arg Ser Cys
1 5 10

<210> 214

<211> 19

<212> PRT

<213> Actinomyces bovis

<400> 214

Arg Ala Gln Pro Leu Ser Arg Val Ala Ser Thr Leu Trp Trp Gly Leu
1 5 10 15

Ala Gly Asp

<210> 215

<211> 19

<212> PRT

<213> Actinomyces meyeri

<400> 215

Arg Ala Gln Pro Leu Pro Tyr Val Ala Ser Thr Leu Trp Trp Gly Leu
1 5 10 15

Val Gly Asp

<210> 216

<211> 18

<212> PRT

<213> Denitrobacterium detoxificans

<400> 216

Arg Ala Gln Pro Leu Pro His Val Ala Ser Ile Arg Leu Gly Thr His
1 5 10 15

Gly Gly

<210> 217

<211> 14

<212> PRT

<213> Unknown

<220>

<223> Uncultured GNS bacteria BPC110

<400> 217

Arg Ala Gln Pro Leu Leu Tyr Val Ile Arg Val Ile Pro Asp
1 5 10

<210> 218

<211> 14

<212> PRT

<213> Unknown

<220>

<223> Uncultured GNS bacteria GCA004

<400> 218

Arg Ala Gln Pro Ser Leu Tyr Val Thr Arg Ile Ile Arg Asp
1 5 10

<210> 219

<211> 14

<212> PRT

<213> Unknown

<220>

<223> Uncultured GNS bacteria GCA112

<900> 219

Arg Ala Gln Pro Ser Pro Tyr Val Ile Arg Val Ile Arg Asp
1 5 10

<210> 220

<211> 16

<212> PRT

<213> Acetobacter aceti

<400> 220

Arg Ala Gln Pro Leu Ser Leu Val Ala Ser Met Phe Gly Trp Ala Leu
1 5 10 15

<210> 221

<211> 16

<212> PRT

<213> Gluconobacter asaii

<400> 221

```

Arg Ala Gln Pro Leu Ser Leu Val Ala Ser Thr Phe Arg Trp Ala Leu
1           5           10           15

```

<210> 222

<211> 17

<212> PRT

<213> Unknown

<220>

<223> Burkholderia sp. JB1

<400> 222

```

Arg Ala Gln Pro Leu Ser Leu Val Ala Thr Gln Glu His Ser Arg Glu
1           5           10           15

```

Thr

<210> 223

<211> 15

<212> PRT

<213> Denitrobacter permanens

<400> 223

```

Arg Ala Gln Pro Leu Pro Leu Val Ala Thr Phe Ser Trp Ala Leu
1           5           10           15

```

<210> 224

<211> 19

<212> PRT

<213> Desullobacter curvatus

<400> 224

```

Arg Ala Gln Pro Leu Ser Leu Val Ala Ser Thr Leu Cys Gly Asn Ser
1           5           10           15

```

Asn Glu Thr

<210> 225

<211> 19

<212> PRT

<213> Unknown

<220>

<223> Desulfobulbus sp. BG25

<400> 225

Arg Ala Gln Pro Leu Pro Leu Val Ala Ser Ser Ser Ala Gly His Ser
1 5 10 15

Lys Gly Thr

<210> 226

<211> 11

<212> PRT

<213> Legionella anisa

<400> 226

Arg Ala Gln Pro Leu Ser Leu Val Ala Ser Thr
1 5 10

<210> 227

<211> 16

<212> PRT

<213> Unknown

<220>

<223> Benzene mineralizing clone SB-1

<400> 227

Arg Ala Gln Pro Leu Pro Leu Val Ala Asn Arg Ser Ser Trp Gly Leu
1 5 10 15

<210> 228

<211> 19

<212> PRT

<213> Escherichia coli

<400> 228

Arg Ala Gln Pro Leu Ser Phe Val Ala Ser Gly Pro Ala Gly Asn Ser
1 5 10 15

Lys Glu Thr

<210> 229

<211> 16

<212> PRT

<213> Unknown

<220>

<223> Uncultured Acidobacterium Sub.Div-1

<400> 229

Arg Ala Gln Pro Leu Ser Leu Val Ala Ser Gly Ser Ser Arg Ala Leu
1 5 10 15

<210> 230

<211> 19

<212> PRT

<213> Unknown

<220>

<223> Uncultured Acidobacterium Sub.Div-1

<400> 230

Arg	Ala	Gln	Pro	Leu	Ser	Ser	Val	Ala	Ile	Gly	Ser	Ser	Arg	Ala	Thr
1				5					10					15	

Leu Ala Lys

<210> 231

<211> 11

<212> PRT

<213> Unknown

<220>

<223> Uncultured Acidobacterium Sub.Div-1

<400> 231

Arg	Ala	Gln	Pro	Leu	Phe	Ala	Ser	Cys	His	His
1				5					10	

<210> 232

<211> 19

<212> PRT

<213> Unknown

<220>

<223> Uncultured Acidobacterium Sub.Div-1

<400> 232

Arg	Ala	Gln	Pro	Leu	Phe	Ala	Gln	Leu	Pro	Ser	Phe	Ser	Trp	Ala	Leu
1				5					10					15	

Cys Arg Asn

<210> 233

<211> 11

<212> PRT

<213> Unknown

<220>

<223> Uncultured Acidobacterium Sub.Div-1

<220>

<221> MISC_FEATURE

<222> (8)..(8)

<223> Unable to determine

<400> 233

```
Arg Ala Gln Pro Leu Leu Pro Xaa Ala Ile Ile
1           5           10
```

<210> 234

<211> 11

<212> PRT

<213> Unknown

<220>

<223> Uncultured Acidobacterium Sub.Div-1

<400> 234

```
Arg Ala Gln Pro Leu Leu Pro Val Ala Thr Ile
1           5           10
```

<210> 235

<211> 11

<212> PRT

<213> Unknown

<220>

<223> Uncultured Acidobacterium Sub.Div-1

<400> 235

```
Arg Ala Gln Pro Leu Ser Pro Val Ala Ile Ile
1           5           10
```

<210> 236

<211> 11

<212> PRT

<213> Unknown

<220>

<223> Uncultured Acidobacterium Sub.Div-1

<400> 236

```
Arg Ala Gln Pro Leu Ser Ser Val Ala Thr Ile
1           5           10
```

<210> 237

<211> 11

<212> PRT

<213> Unknown

<220>

<223> Uncultured Acidobacterium Sub.Div-1

<400> 237

Arg Ala Gln Pro Leu Phe Leu Val Ala Thr Ile
1 5 10

<210> 238

<211> 13

<212> PRT

<213> Unknown

<220>

<223> Uncultured Acidobacterium Sub.Div-1

<400> 238

Arg Ala Gln Pro Ser Ser Leu Val Ala Asn Thr Leu Trp
1 5 10

<210> 239

<211> 18

<212> PRT

<213> Unknown

<220>

<223> Uncultured Acidobacterium Sub.Div-2

<400> 239

Arg Ala Gln Pro Leu His Val Val Ala Thr Arg Lys Arg Glu Leu Tyr
1 5 10 15

Val Asp

<210> 240

<211> 18

<212> PRT

<213> Unknown

<220>

<223> Uncultured Acidobacterium Sub.Div-2

<400> 240

Arg Ala Gln Pro Leu His Val Val Ala Thr Pro Gln Gly Gly Thr Leu
1 5 10 15

Arg Gly

<210> 241

<211> 19

<212> PRT

<213> Unknown

<220>

<223> Uncultured Acidobacterium Sub.Div-3

<400> 241

Arg Ala Gln Pro Ser Ser Leu Val Ala Asn Pro Gln Gly Lys His Pro
1 5 10 15

Lys Gly Thr

<210> 242

<211> 11

<212> PRT

<213> Unknown

<220>

<223> Uncultured Acidobacterium Sub.Div-3

<400> 242

Arg Ala Arg Pro Leu Ser Cys Val Ala Ile Ile
1 5 10

<210> 243

<211> 18

<212> PRT

<213> Unknown

<220>

<223> Uncultured Acidobacterium Sub.Div-3

<400> 243

Arg Ala Gln Pro Leu Ser Cys Val Ala Asn Pro Gln Gly Cys Thr Leu
1 5 10 15

Arg Arg

<210> 244

<211> 20

<212> PRT

<213> Unknown

<220>

<223> Uncultured Acidobacterium Sub.Div-3

<400> 244

Arg Ala Gln Pro Ser Pro Cys Val Ala Thr Pro Pro Arg Ala Gly Ala
1 5 10 15

Leu Ser Gly Asp
20

REFERENCES CITED IN THE DESCRIPTION

This list of references cited by the applicant is for the reader's convenience only. It does not form part of the European patent document. Even though great care has been taken in compiling the references, errors or omissions cannot be excluded and the EPO disclaims all liability in this regard.

Patent documents cited in the description

- [US5866330A](#) [\[0010\]](#)
- [US0111609W](#) [\[0130\]](#)
- [US60196063B](#) [\[0130\]](#)
- [US60196258B](#) [\[0130\]](#)

Non-patent literature cited in the description

- **LARSEN, L. H. et al.** A microscale NO₃ biosensor for environmental applications *Anal. Chem.*, 1997, vol. 69, 3527-3531 [\[0002\]](#)
- **VIRTA M. et al.** Bioluminescence-based metal detectors *Methods Mol. Biol.*, 1998, vol. 102, 219-229 [\[0002\]](#)
- **STICHER P. et al.** Development and characterization of a whole-cell bioluminescent sensor for bioavailable middle-chain alkanes in contaminated groundwater samples *Appl. Environ. Microbiol.*, 1997, vol. 63, 104053-4060 [\[0002\]](#)
- **RELMAN, D.A. et al.** *New Engl. J. Med.*, 1990, vol. 323, 1573- [\[0004\]](#)
- **WILSON, K.H. et al.** *Lancet*, 1991, vol. 338, 474- [\[0004\]](#)
- **LAWRENCE OSBORNE** *Crime-Scene Forensics; Dead Men Talking* New York Times 20001203 [\[0007\]](#)
- **ROITT, I. et al.** *Immunology*, 1993, 5.1-5.14 [\[0008\]](#)
- **AMANN, R. I. et al.** Phylogenetic identification and in situ detection of individual microbial cells without cultivation *Microbiol. Rev.*, 1995, vol. 59, 143-169 [\[0009\]](#)
- **VELCULESCU, V. E. et al.** Serial analysis of gene expression *Science*, 1995, vol. 270, 484-487 [\[0010\]](#)
- **TROESCH, A. et al.** Mycobacterium species identification and rifampin resistance testing with high-density DNA probe arrays *J. Clin. Microbiol.*, 1999, vol. 37, 49-55 [\[0011\]](#)
- **WOESE, C. R.** Bacterial evolution *Microbiol Rev.*, 1987, vol. 51, 2221-71 [\[0032\]](#)

- **PACE, N.** RA molecular view of microbial diversity and the biosphere *Science*, 1997, vol. 276, 5313734-40 [0032]
- **JEANTHON, C. et al.** *Thermotoga subterranea* sp. nov., a new thermophilic bacterium isolated from a continental oil reservoir *Arch. Microbiol.*, 1995, vol. 164, 91-97 [0033]
- **YEATES et al.** Methods for Microbiological DNA Extraction from Soil for PCR Amplification *Biological Procedures Online*, 1998, vol. 1, [0044]
- **LIU et al.** *Applied and Environmental Microbiology*, 1997, vol. 63, 4516-4522 [0044] [0073]
- **TSAI et al.** *Applied and Environmental Microbiology*, 1992, vol. 58, 2292-2295 [0044]
- **COSTERTON, JW. et al.** Bacterial biofilms: a common cause of persistent infections *Science*, 1999, vol. 284, 54181318-1322 [0045]
- **GILLAN, D.C. et al.** Genetic diversity of the biofilm covering *Montacuta ferruginosa* (Mollusca, bivalvia) as evaluated by denaturing gradient gel electrophoresis analysis and cloning of PCR-amplified gene fragments coding for 16S rRNA *Appl. Environ. Microbiol.*, 1998, vol. 64, 93464-72 [0045]
- **ROITT et al.** *Immunology* 199300005.2-5.14 [0047]
- **VELCELESCU et al.** *Science*, 1995, vol. 270, 484-487 [0056] [0059]
- **POWELL** *Nucleic Acids Research*, 1998, vol. 14, 3445-3446 [0056]
- **TUCHOLSKI, J. et al.** Mmel, a class-IIS restriction endonuclease: purification and characterization *Gene*, 1995, vol. 157, 87-92 [0057]
- **M. J. SCHMIDT** *Understanding and Using Statistics* D. C. Health and Company 19750000131-147 [0075]
- **STICHER, P. et al.** Development and characterization of a whole-cell bioluminescent sensor for bioavailable middle-chain alkanes in contaminated groundwater samples *Appl Environ Microbiol.*, 1997, vol. 63, 4053-4060 [0082]
- **LEU, J.-Y. et al.** The same species of sulphate-reducing *Desulfomicrobium* occur in different oil field environments in the north sea *Lett. Appl. Microbiol.*, 1999, vol. 29, 4246-252 [0084]
- **WENSELEERS, T. et al.** Widespread occurrence of the micro-organism *Wolbachia* in ants *Proc. R Soc. Lond. B. Biol. Sci.*, 1998, vol. 265, 14041447-52 [0086]
- **SCHRODER, D. et al.** Intracellular endosymbiotic bacteria of *Camponotus* species (carpenter ants): systematics, evolution and ultrastructural characterization *Mol. Microbiol.*, 1996, vol. 21, 479-89 [0086]
- **NOUR M.** 1998, 16S-23S and 23S-5S intergenic spacer regions of lactobacilli: nucleotide sequence, secondary structure and comparative analysis *Res. Microbiol.*, vol. 149, 6433-448 [0095]
- **BERTHIER F. et al.** Rapid species identification within two groups of closely related lactobacilli using PCR primers that target the 16S/23S rRNA spacer region *FEMS Microbiol Lett.*, 1998, vol. 161, 197-106 [0095]
- **TILSALA-TIMISJARVI A. et al.** Development of oligonucleotide primers from the 16S-23S rRNA intergenic sequences for identifying different dairy and probiotic lactic acid bacteria *PCR Int. J. Food Microbial.*, 1997, vol. 35, 149-56 [0095]
- **FARRELLY, V. F. et al.** Effect of genome size and *rrn* gene copy number on PCR amplification of 16S rRNA genes from a mixture of bacterial species *Appl. Environ.*

Microbiol., 1995, vol. 61, 72798-2801 [\[0114\]](#)

- **PACE, N. R.** A molecular view of microbial diversity and the biosphere *Science*, 1997, vol. 276, 734-740 [\[0119\]](#)
- **GARVIN, A M. et al.** 2000 MALDI-TOF based mutation detection using tagged in vitro synthesized peptides *Nature Biotechnol*, vol. 18, 95-97 [\[0124\]](#)
- **GUNDERSON, K. L. et al.** Mutation detection by ligation to complete n-mer DNA arrays *Genome Res.*, 1998, vol. 8, 1142-1153 [\[0127\]](#)

Patentkrav

1. Fremgangsmåde til konstruktion af en markørdiversitetsprofil- (MDP) database, hvilken fremgangsmåde omfatter følgende trin:
 - 5 a) sekvensering af en flerhed af rRNA-markører fra en prøve, hvor prøver omfatter en mikrobiel population, og hvor hver rRNA-markør omfatter et mikrobielt rRNA-gens polymorfe sekvens, og hvor flerheden af rRNA-markører tilvejebringer en præcis repræsentation af prøvens mikrobielle struktur;
 - b) bestemmelse af abundansen, i prøven, af hver af flerheden af rRNA-markører;
 - 10 c) transduktion af abundansen af hver rRNA-markør til et elektrisk udgangssignal;
 - d) lagring af de elektriske udgangssignaler frembragt i trin c) i en matrixdatastruktur og associering i strukturen af hvert udgangssignal med den tilsvarende sekvens af rRNA-markøren, hvorfra abundansudgangssignalet var transduceret;
 - 15 e) tilvejebringelse af abundanser af ét eller flere prøveparametre, der er forbundet med prøven;
 - f) transduktion af abundanser af hvert prøveparameter til et elektrisk udgangssignal;
 - g) lagring af de elektriske udgangssignaler frembragt i trin f) i en matrixdatastruktur og associering i strukturen af hvert udgangssignal med prøveparameteret, hvorfra abundansudgangssignalet var transduceret;
 - 20 h) betegnelse af de elektriske udgangssignaler frembragt i trin c) svarende til flerheden af rRNA-markørabundanser og de elektriske udgangssignaler frembragt i trin f) svarende til den ene eller flere prøveparameterabundanser, som en MDP; og
 - 25 i) gentagelse af trin a-h for en flerhed af andre prøver, og betegnelse af flerheden af MDP'er laget i datastrukturen som en MDP-database.
2. Fremgangsmåde ifølge krav 1, hvor rRNA-markørabundansen er en abundans i forhold til den totale abundans af flerheden af rRNA-markører i prøven.
- 30 3. Fremgangsmåde ifølge krav 1 eller 2, hvor det mikrobielle rRNA-gen er et 16S Rrna-gen.

4. Fremgangsmåde ifølge krav 1 eller 2, hvor det mikrobielle rRNA-gens polymorfe sekvens er fra intergen-området mellem et 16S rRNA-gen og et 23S rRNA-gen.
5. Fremgangsmåde ifølge et hvilket som helst af kravene 1 til 4, hvor prøven er udvalgt fra gruppen bestående af: en jordprøve, en klippeprøve, en vandprøve, en luftprøve, en carbonhydridprøve, en petroleumsprøve og en biofilmprøve.
6. Fremgangsmåde ifølge et hvilket som helst af kravene 1 til 5, hvor prøven er opnået fra gruppen bestående af: petroleum, carbonhydrider, et oliereservoir, et gasreservoir, jord, klippe, vand, luft, en bygning og en vej.
7. Fremgangsmåde ifølge krav 6, hvor prøveparameteret er udvalgt fra gruppen bestående af: petroleum, carbonhydrider, olie, gas, en uorganisk forbindelse, et mineral, pH, temperatur og saltholdighed.
8. Fremgangsmåde ifølge et hvilket som helst af kravene 1 til 4, hvor prøven er opnået fra gruppen bestående af: et menneske, en plante, et dyr, en fødevarer, en cellekultur eller en vævskultur.
9. Fremgangsmåde ifølge krav 8, hvor prøveparameteret er udvalgt fra gruppen bestående af: en akut sygdomstilstand, en kronisk sygdomstilstand, en fysiologisk tilstand og et udviklingstrin.
10. Fremgangsmåde ifølge et hvilket som helst af kravene 1 til 4, hvor prøveparameteret er tid, og hvor flerheden af prøver omfatter et tidsforløb.
11. Fremgangsmåde ifølge krav 10, hvor flerheden af prøver omfatter et tidsforløb, der kan anvendes til overvågning af miljømæssige afhjælpningsforanstaltninger eller oliefeltproduktion.
12. MDP-database, der er opbygget og lagret i hukommelsen på en computer ifølge fremgangsmåden ifølge et hvilket som helst af kravene 1 til 11.

DRAWINGS

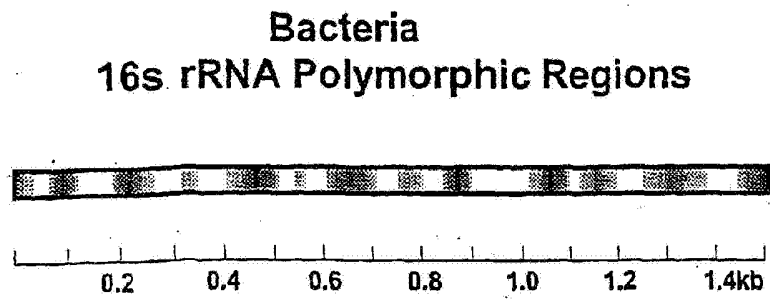


Figure 1.

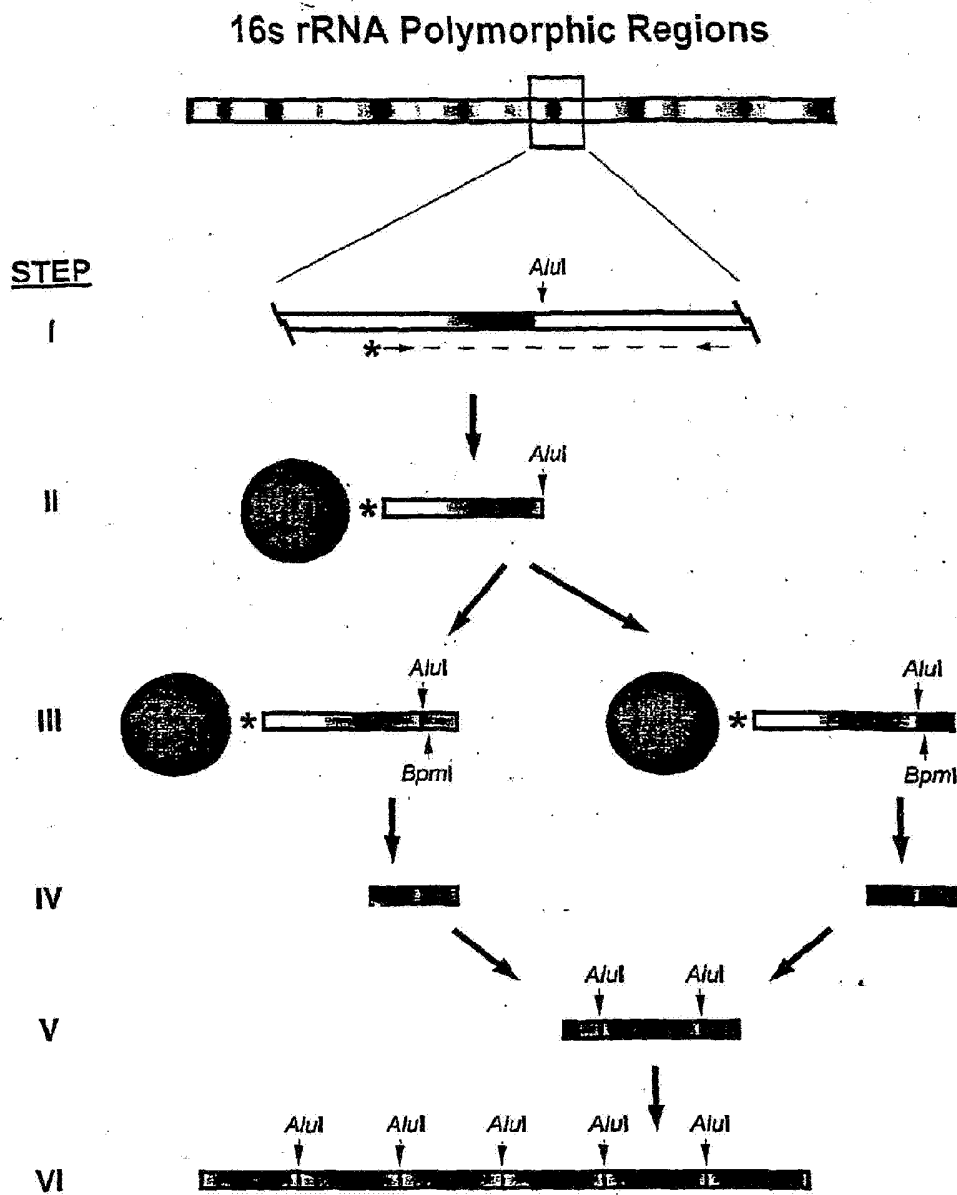


Figure 2.

Bacteria

Acidobacteria Group
 Various uncultured environmental Acidobacteria
 Aquificales
 Desulfurobacterium Group
 Desulfurobacterium
 Desulfurobacterium thermolithotrophum
 environmental samples
 unidentified Aquificales OPS132
 CFB/Green sulfur bacteria group
 Bacteroidaceae
 Bacteroides
 Bacteroides caccae
 Firmicutes(gram-positive bacteria)
 Actinobacteria(high G+C gram-positive bacteria)
 Actinomycetaceae
 Actinomyces
 Actinomyces bovis
 Actinomyces mayeri
 Denitrobacterium
 Denitrobacterium detoxificans
 Green non-sulfur bacteria
 environmental samples
 uncultured HC-seep bacterium BPC110
 uncultured HC-seep bacterium GCA004
 uncultured HC-seep bacterium GCA112
 Proteobacteria(purple non-sulfur bacteria)
 alpha subdivision
 Acetobacteraceae
 Acetobacter
 Acetobacter aceti
 Gluconobacter
 Gluconobacter asahi
 beta subdivision
 Burkholderia group
 Burkholderia
 Burkholderia sp. JB1
 Denitrobacter
 Denitrobacter permanens
 delta subdivision
 Desulfobacter
 Desulfobacter curvatus
 Desulfobulbus
 Desulfobulbus sp. BG25
 Legionellaceae
 Legionella
 Legionella anisa
 unclassified Bacteria
 benzene mineralizing clone SB-1
 uncultured eubacterium env.OPS 1

Fig. 3

- Archaea (archaeobacteria)**
- Crenarchaeota (extremely thermophilic archaeobacteria)**
- Desulfurococcales
 - Desulfurococcaceae
 - Aeropyrum*
 - Aeropyrum pernix*
 - Desulfurococcus
 - Desulfurococcus mobilis*
 - Staphylothermus
 - Staphylothermus marinus*
 - Sulfolobales
 - Metallosphaera
 - Metallosphaera sedula*
 - Sulfolobus
 - Sulfolobus acidocaldarius*
 - Sulfolobus metallicus*
 - Thermoproteales
 - Thermoproteaceae
 - Caldivirga*
 - Caldivirga maquilingensis*
 - Pyrobaculum
 - Pyrobaculum islandicum*
- Euryarchaeota**
- Archaeoglobales
 - Archaeoglobaceae
 - Archaeoglobus
 - Archaeoglobus fulgidus*
 - Archaeoglobus veneficus*
 - Halobacteriales
 - Halobacteriaceae
 - Haloarcula
 - Haloarcula japonica*
 - Halococcus
 - Halococcus morrhuae*
 - Methanococcales
 - Methanococcaceae
 - Methanococcus
 - Methanococcus jannaschii*
 - Methanobacteriales
 - Methanobacteriaceae
 - Methanobacterium
 - Methanobacterium bryanii*
 - Methanobacterium subterraneum*
 - Thermococcales
 - Thermococcaceae
 - Pyrococcus
 - Pyrococcus abyssi*
 - Thermoplasmatales
 - Picrophilaceae
 - Picrophilus
 - Picrophilus oshimae*

Fig. 4

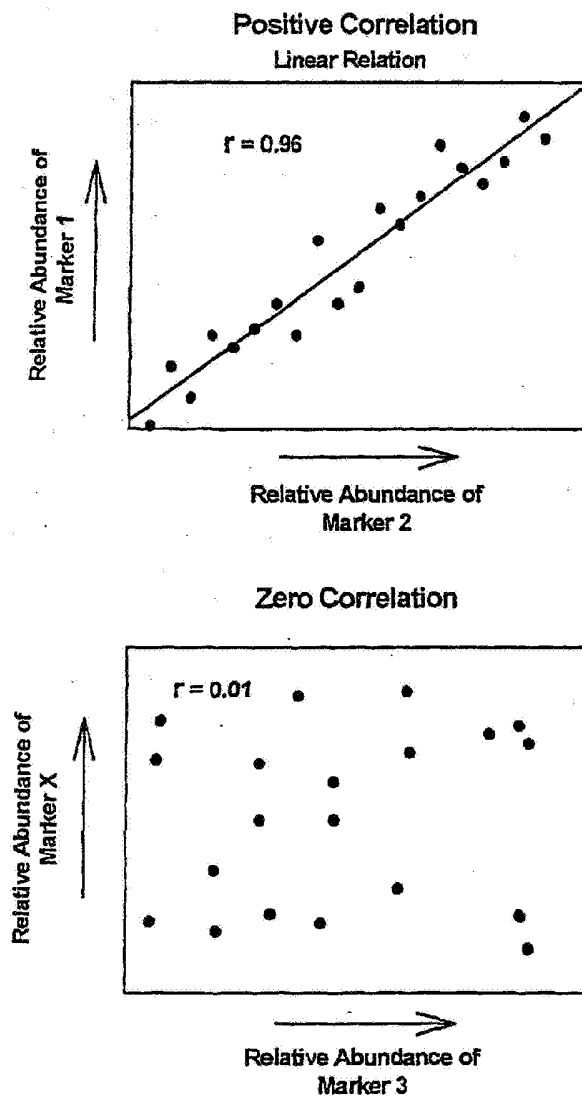


Fig. 5

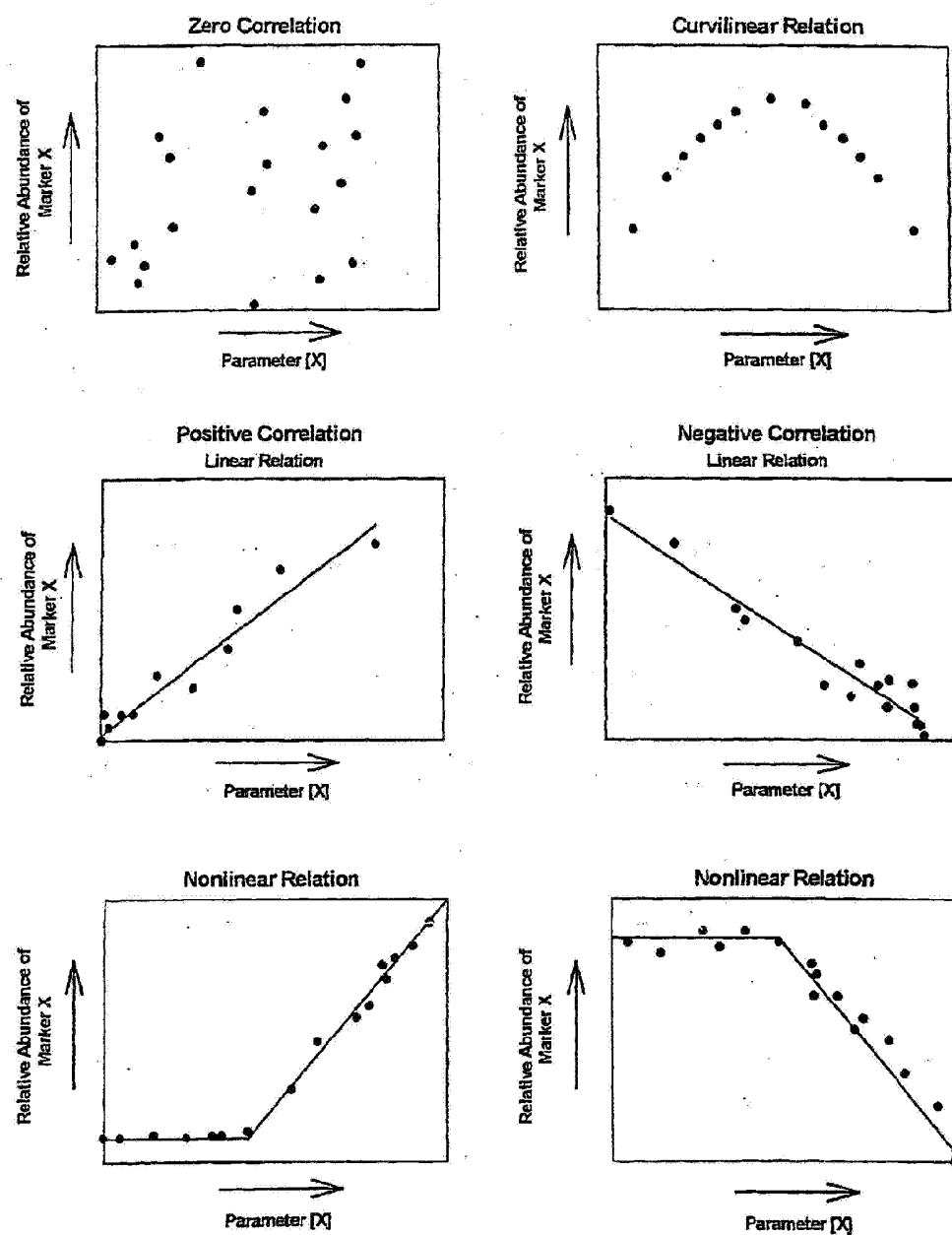


Fig. 6

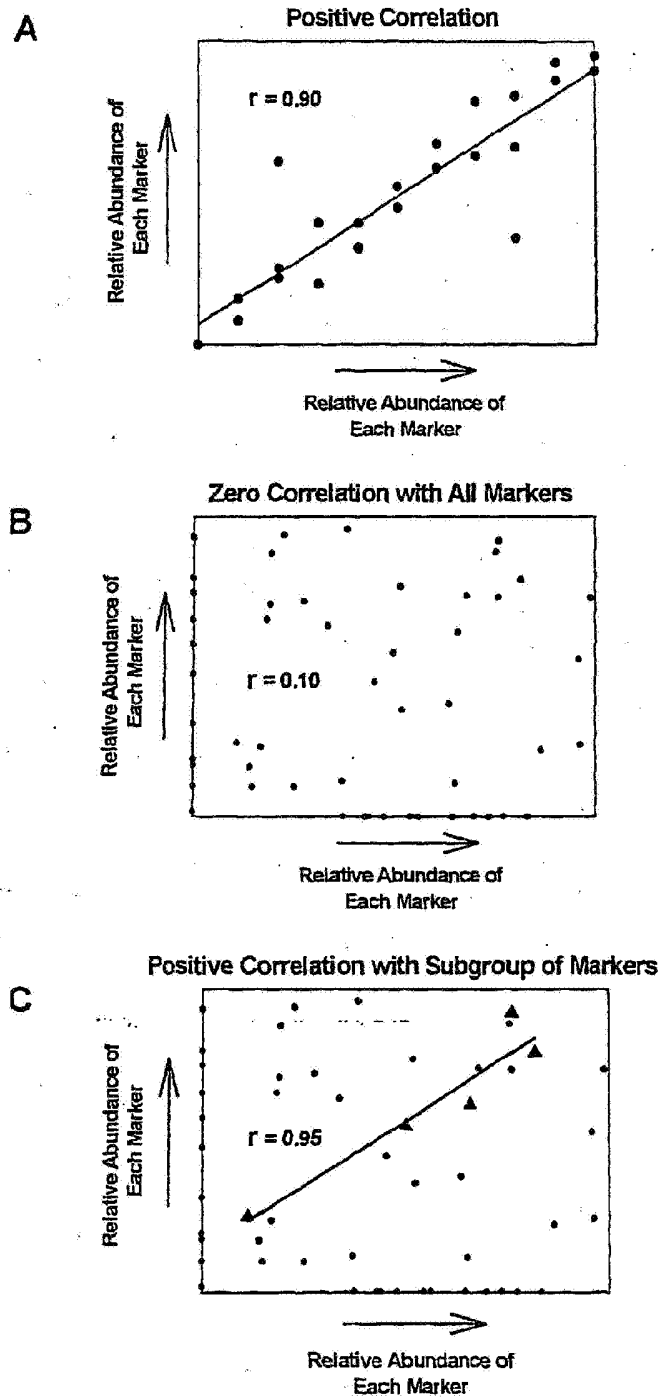


Fig. 7

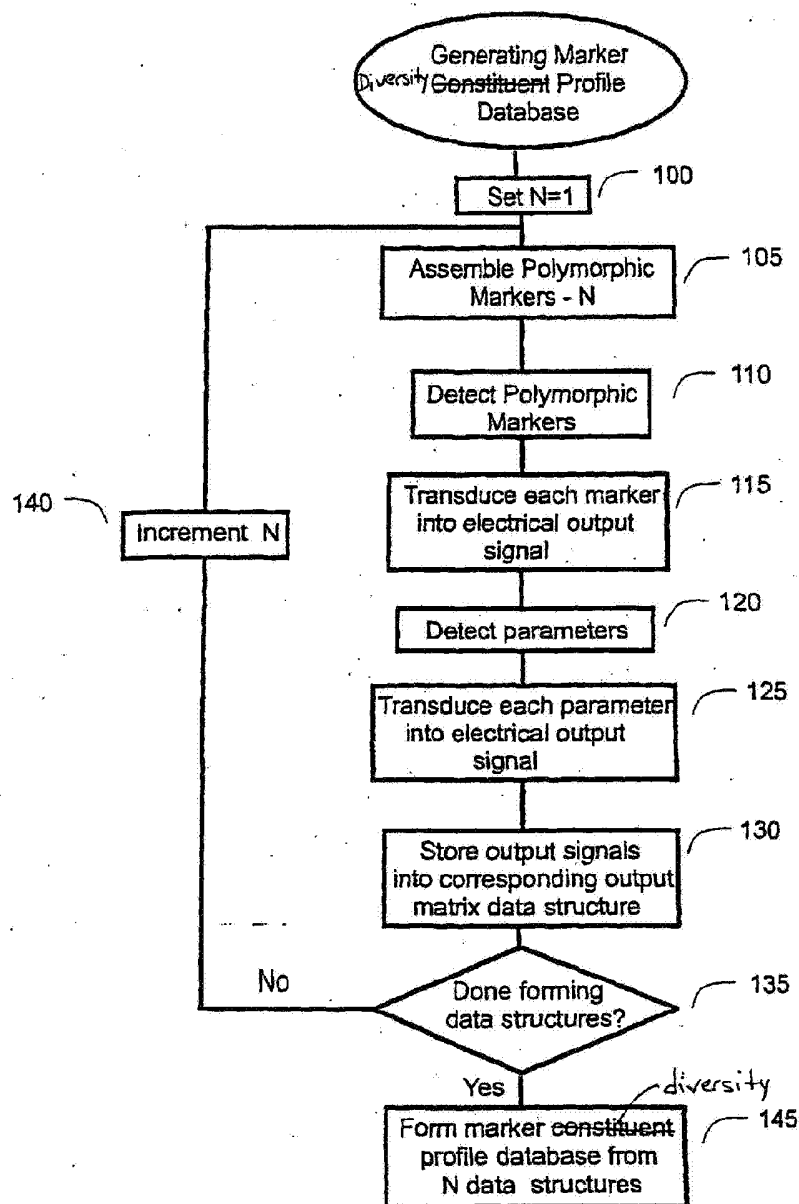


Figure 8.

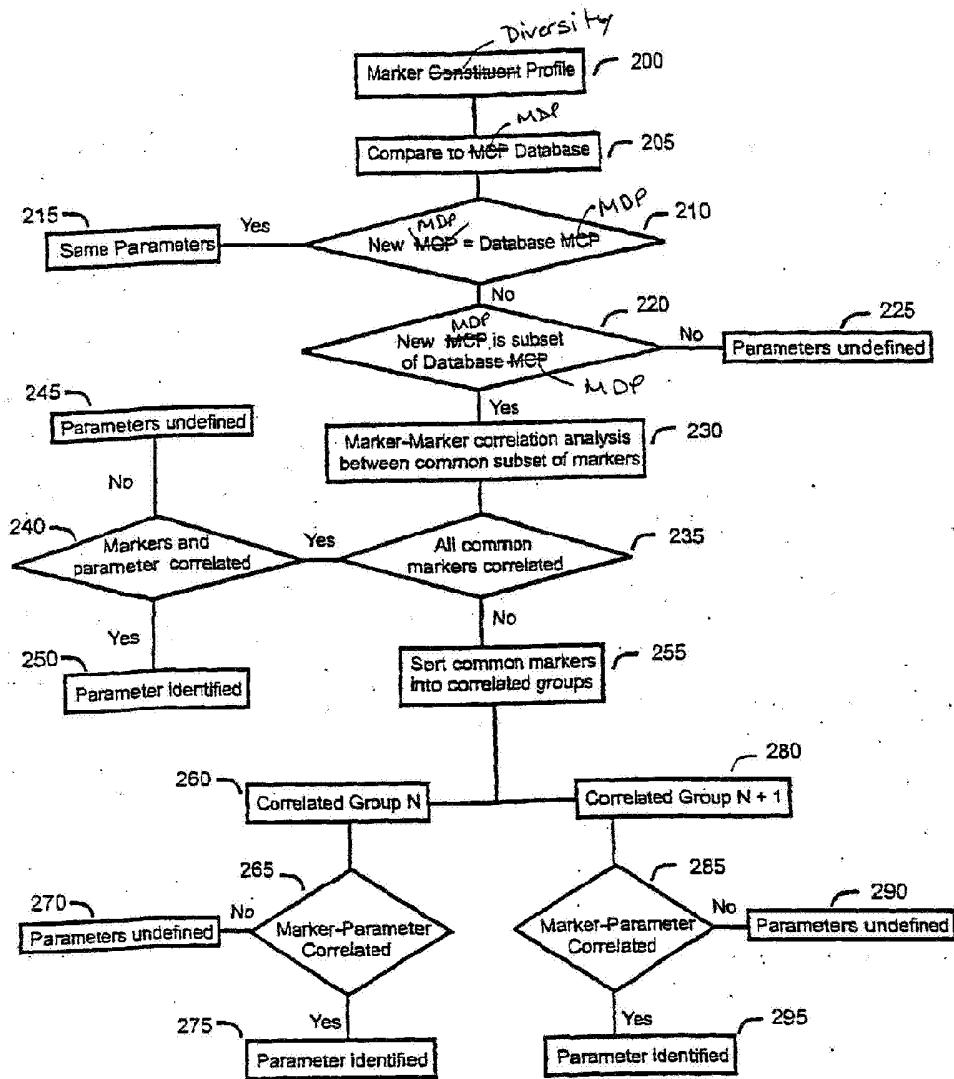


Figure 9.

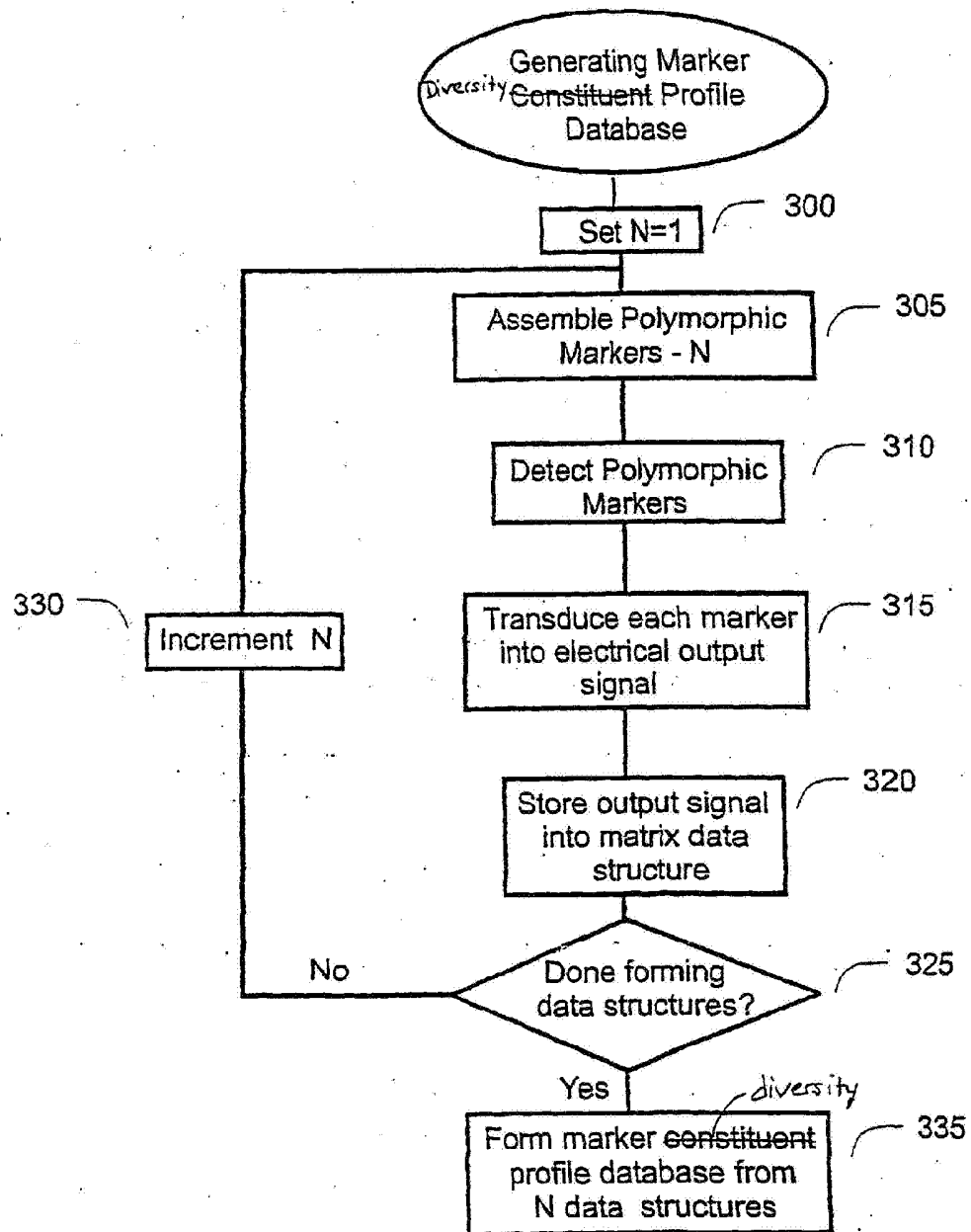


Figure 10.

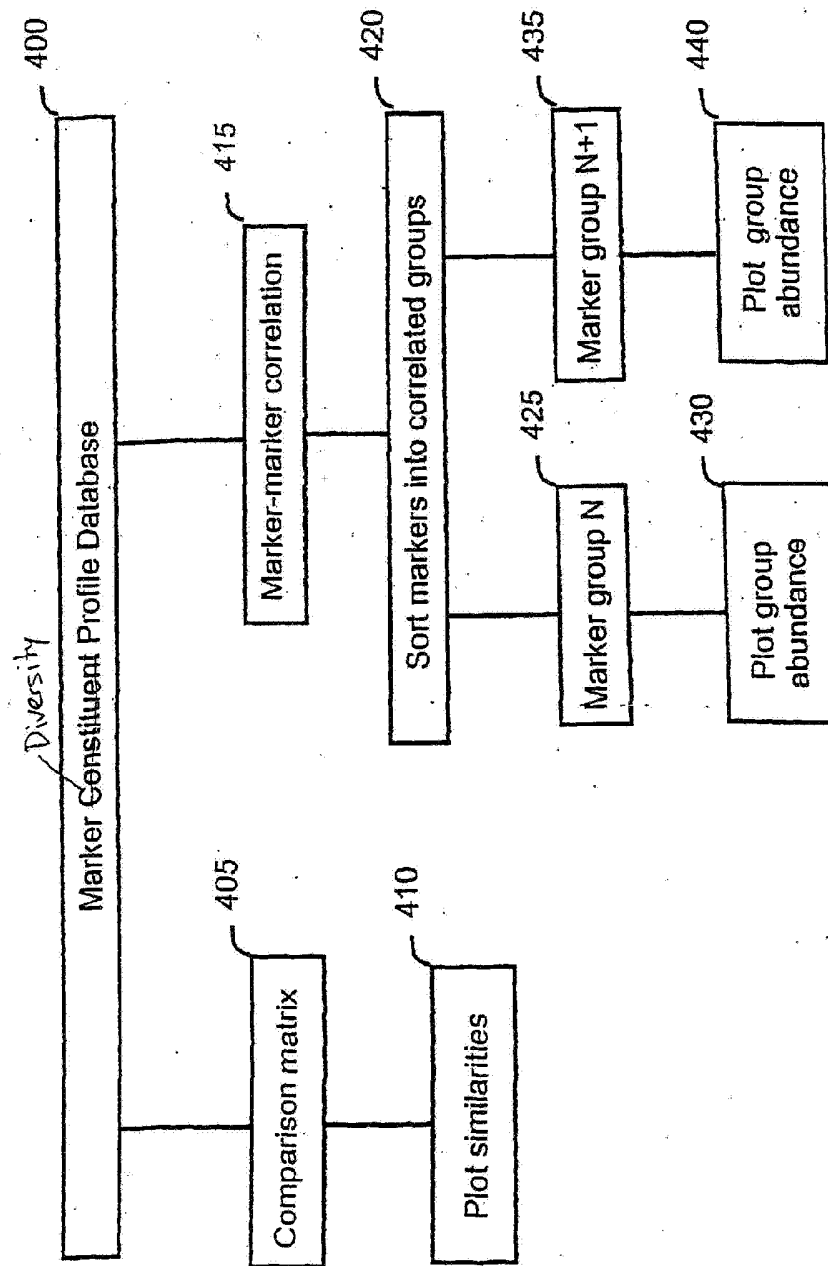


Figure 11.

SARD OLIGONUCLEOTIDES

TX-003	5'-	Phosphate-GCTCCAGGTCTACATCCTAGTCAGGACddc-3'
TX-012	5'-	Phosphate- CTCCAGGTCTACATCCTAGTCAGGACddc
TX-010R	3'	GAGGTCCAGATGTAGGATCACTCCTGGATA-5'
TX-013	3'	GAGGTCCAGATGTAGGATCACTCCTGGATA-5'
TX-004R	3'	CAGATGTAGGATCACTCCTGGATA-5'
TX-007R	3'-	CAGATGTAGGATCACTCCTGG -Biotin-5'
TX-111	3'-	CAGATGTAGGATCACTCCTGGCACTGATAGCTC -Biotin-5'
TX-005	5'-	Phosphate-GCTCCAGACTAGCATCCGCTGACTTGaddc
TX-014	5'-	Phosphate- CTCCAGACTAGCATCCGCTGACTTGaddc
TX-011R	3'	GAGGTCTGATCGTAGGCGACTGAACGTGTA
TX-015	3'	GAGGTCTGATCGTAGGCGACTGAACGTGTA
TX-006R	3'	TGATCGTAGGCGACTGAACGTGTA
TX-008R	3'	TGATCGTAGGCGACTGAACGTG -Biotin-5'
TX-121	3'	TGATCGTAGGCGACTGAACGTGGCAACGTTGGAC -Biotin-5'
TX-001	5'-	Biotin-CTG TAG HRG TGA AAT DCD YA
TX-002	5'-	YTC ACG RCA YGA GCT GAC GAC
TX-003	5'-	Phosphate-GCT CCA GGT CTA CAT CCT AGT CAG GACddc
TX-004	5'-	ATA GGT CCT GAC TAG GAT GTA GAC
TX-005	5'-	Phosphate-GCT CCA GAC TAG CAT CCG CTG ACT TGaddc
TX-006	5'-	AAT GTC AAG TCA GCG GAT GCT AGT
TX-007	5'-	Biotin-GGT CCT GAC TAG GAT GTA GAC
TX-008	5'-	Biotin-GTC AAG TCA GCG GAT GCT AGT
TX-009	5'-	Biotin-GGA TTA GAA ACC CBG GTA GTC
TX-010	5'-	ATA GGT CCT GAC TAG GAT GTA GAC CTG GAG
TX-011	5'	AAT GTC AAG TCA GCG GAT GCT AGT CTG GAG
TX-012	5'-	Phosphate-CTC CAG GTC TAC ATC CTA GTC AGG ACddc
TX-013	5'-	ATA GGT CCT GAC TAG GAT GTA GAC CTG GAG
TX-014	5'-	Phosphate-CTC CAG ACT AGC ATC CGC TGA CTT Gaddc
TX-015	5'	AAT GTC AAG TCA GCG GAT GCT AGT CTG GAG
TX-111	5'	Biotin-CTC GAT AGT CAC GGT CCT GAC TAG GAT GTA GAC
TX-121	5'	Biotin-CAG GTT GCA ACG GTC AAG TCA GCG GAT GCT AGT

Fig. 12

SARD Analysis of Defined Population

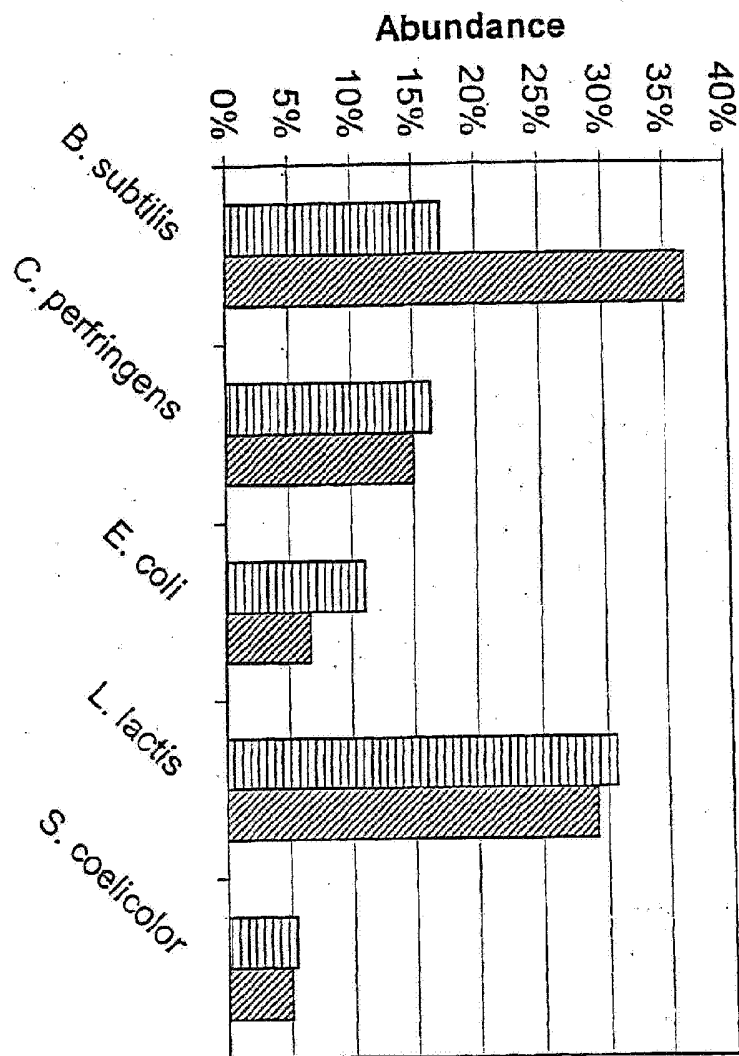


Fig. 14

ACGATGAGCACTAGCT (1)	CCTGTGTGACGAAGCT (1)	CTTCTGTGGCGAAGCT (1)
ACGATGAGTACTAGCT (1)	CCTTGGTAACGAAGCT (1)	GATCCGTGCCGTAGCT (1)
ACGATGATGACTAGCT (2)	CCTTGGTACCGAAGCT (1)	GCTCTGTGCCGAAGCT (1)
ACGATGGATGCTAGCT (1)	CGCCAGTGCCGTAGCT (1)	GCTGGGTGCCCAAGCT (1)
ATGCTAGTCTGGAGCT (6)	CGCCTGTGCCGTAGCT (2)	GGTCCGTGCCGCAGCT (1)
ATGGCTGTGCTCAGCT (50)	CGTCCGTGCCGAAGCT (2)	GGTGTCTTTCGGAGCT (2)
ATGGTTGTGCTCAGCT (1)	CGTCCGTGCCGCAGCT (1)	GTAAACGATGGAAGCT (1)
ATTCCGTGCCGTAGCT (1)	CGTCCGTGCCGCAGCT (3)	GTGGCTGTCTCAGCT (2)
CACTAGTGGCGCAGCT (2)	CTCCCGTGCCGCAGCT (1)	GTTCCGTGCCGAAGCT (2)
CCCCCGTGCCGAAGCT (1)	CTCCCGTGCCGGAGCT (1)	GTTCCGTGCCGCAGCT (3)
CCCCCGTGCCGCAGCT (1)	CTCCCGTGCCGCAGCT (1)	TATCASTGGCGCAGCT (1)
CCCCCTTCCTCCAGCT (1)	CTCCTGTGCCGAAGCT (1)	TCTCCGTGCCGCAGCT (1)
CCCCGTGCCGCAGCT (1)	CTCCTGTGCCGCAGCT (2)	TCTCTGTGCCGCAGCT (2)
CGGGGTAGTCCAGCT (5)	CTGCCGTGCCGAAGCT (9)	TCTCTGTGCCGTAGCT (1)
CCTCCGTGCCGAAGCT (3)	CTGCTGTGCCGAAGCT (2)	TGTCCGTGCCGTAGCT (1)
CCTCCGTGCCGCAGCT (2)	CTGTCGTGCCGAAGCT (1)	TTTCCGTGCCGCAGCT (2)
CCTCCGTGCTGCAGCT (1)	CTTCAGTATCGAAGCT (1)	
CCTCGGCGCCGCAGCT (2)	CTTCCGCGCCGCAGCT (2)	
CCTCGGTGCCGCAGCT (1)	CTTCCGTGCCGCAGCT (3)	
CCTCGGTGTGCGCAGCT (2)	CTTCCGTGCCGGAGCT (1)	
CCTGGGTGCCGCAGCT (2)	CTTCCGTGCCGCAGCT (1)	

Fig. 15

ACGATGATAACTAGCT (2)	CCGGGTAGTCCCAGCT (4)	CTCCAGTGCCCGCAGCT (1)	GATCCGTGCCGCAGCT (2)
ACGATGATGACTAGCT (1)	CCGGGTAGTCCTAGCT (1)	CTCCCGTGCCACAGCT (1)	GCTCCGTGCCGAAGCT (1)
ACGATGGGCACTAGCT (1)	CCTCCGTGCCGAAGCT (2)	CTCCCGTGCCCGAGCT (2)	GCTCCGTGCCGTAGCT (1)
ACGGCTGTCTCAGCT (2)	CCTCCGTGCCCGAGCT (2)	CTCCCGTGCCGGAGCT (3)	GCTCTGTGCCGAAGCT (2)
ACTACGAGCGCAAGCT (1)	CCTCCGTGCCGGAGCT (1)	CTCCCGTGCCCGAGCT (1)	GCTCTGTGCCGTAGCT (1)
ACTTAATGCGTAGCT (1)	CCTCGTAAGGGGAGCT (1)	CTCCTGTGCCCGAGCT (1)	GGGCTTGTCTCAGCT (1)
ATGCTAGTCTGGAGCT (8)	CCTGGGTGCCCGAGCT (1)	CTGCCGCGCCGGAGCT (1)	GTAACGATGGAAGCT (1)
ATGGCTCTCTGTAGCT (1)	CCTGTAGTCCCAGCT (2)	CTGCCGTGCCGAAGCT (5)	GTTCCGTGCCGCAGCT (2)
ATGGCTGTGCCAGCT (3)	CCTGTAGTCCTAGCT (1)	CTGCCGTGCCCTAAGCT (1)	GTTCCGTGCCGTAGCT (2)
ATGGCTGTCTGTAGCT (70)	CCTTAGTAACGCAGCT (1)	CTGCCGTGCCCGAGCT (1)	GTTCTGTGCCGCAGCT (2)
ATGGTTGTCTGTAGCT (2)	CCTTGGTAACGAAGCT (1)	CTGCTGTGCCGAAGCT (7)	TCTCAGCACAGAGCT (2)
ATGTAGACCTGGAGCT (12)	CGCCAGTGCCGAAGCT (1)	CTGTCTGTGCCGAAGCT (1)	TCTCAGTAACGTAGCT (1)
ATTCCGTGCCGCAGCT (6)	CGCCCGTGCCCGAGCT (2)	CTTCCGTGCCCGAGCT (7)	TCTCCGTGCCGCAGCT (2)
ATTCCGTGCCGTAGCT (1)	CGCCTGTGCCGTAGCT (2)	CTTCCGTGCCGAAGCT (2)	TCTCTGTGCCCGAGCT (4)
CACAAGCGGTGGAGCT (1)	CGCTCGTGGCGAAGCT (1)	CTTCCGTGCCCGAGCT (6)	TGGACGTTGCCGAGCT (2)
CACTAGTGGCGCAGCT (4)	CGGAGGCGTCTGTAGCT (1)	CTTCCGTGCCGGAGCT (2)	TTCCTGTGCCGGAGCT (1)
CATCCGTGCCGAAGCT (1)	CGTCAGTGTCTGCAGCT (2)	CTTCCGTGCCCGAGCT (1)	
CCCCAGGGCCCAAGCT (1)	CGTCCGTGCCGAAGCT (7)	CTTCCGTGTCTGCAGCT (1)	
CCCCGTGCCGCAGCT (1)	CGTCCGTGCCCGAGCT (1)	CTTGGGTGCCCGAGCT (2)	
CCCCGTGCCCGAGCT (1)	CGTCCGTGCCGGAGCT (1)	CTTTAGTAACGCAGCT (2)	
CCGCGGTGCCGTAGCT (1)	CGTCCGTGCCCGAGCT (2)	GACCCGCAAGGGAGCT (1)	

Fig. 16

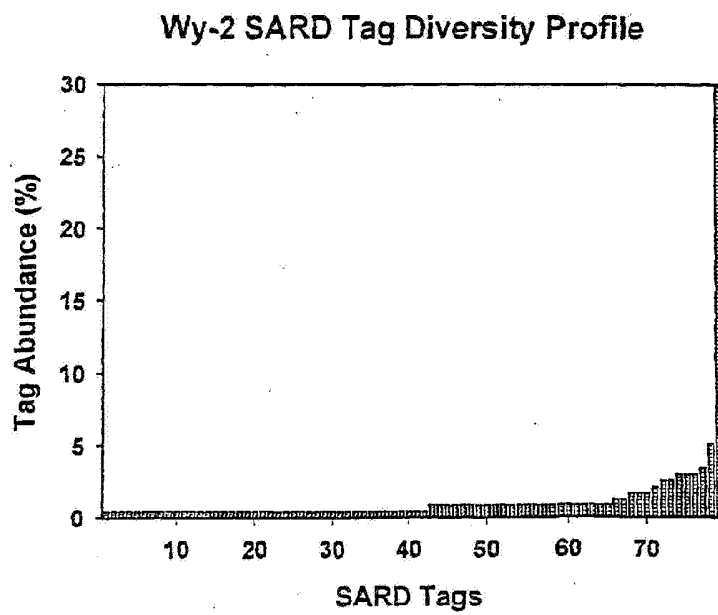
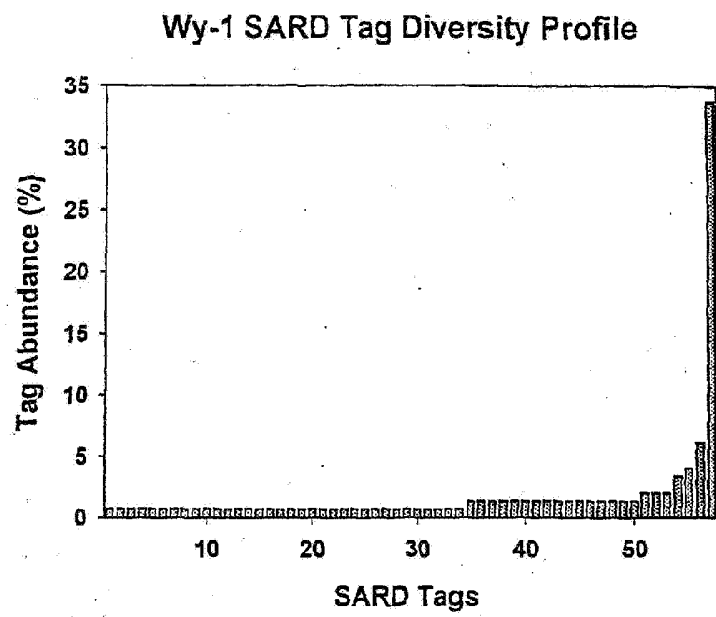


Fig. 17