# Approaches to analyse dynamic microbial communities such as those seen in cystic fibrosis lung

*Melissa Doud,[1] Erliang Zeng,[2] Lisa Schneper,[3] Giri Narasimhan[4]* and Kalai Mathee[3]*

[1]Department of Biological Sciences, Florida International University, Miami, FL 33199, USA
[2]Department of Computer Science, University of Miami, 1365 Memorial Drive, Coral Gables, FL 33146, USA
[3]Department of Molecular Microbiology and Infectious Diseases, College of Medicine, Florida International University, Miami, FL 33199, USA
[4]Bioinformatics Research Group, School of Computing and Information Sciences, Florida International University, Miami, FL 33199, USA
*Correspondence to: Tel: +1 305 348 3748; Fax: +1 305 348 3549; E-mail: giri@cs.fiu.edu

*Date received (in revised form): 2nd December, 2008*

## Abstract

Microbial communities play vital roles in many aspects of our lives, although our understanding of microbial biogeography and community profiles remains unclear. The number of microbes or the diversity of the microbes, even in small environmental niches, is staggering. Current microbiological methods used to analyse these communities are limited, in that many microorganisms cannot be cultured. Even for the isolates that can be cultured, the expense of identifying them definitively is much too high to be practical. Many recent molecular technologies, combined with bioinformatic tools, are raising the bar by improving the sensitivity and reliability of microbial community analysis. These tools and techniques range from those that attempt to understand a microbial community from their length heterogeneity profiles to those that help to identify the strains and species of a random sampling of the microbes in a given sample. These technologies are reviewed here, using the microbial communities present in the lungs of cystic fibrosis patients as a paradigm.

***Keywords:*** *amplicon length heterogeneity (LH), LH-PCR, bioinformatics, supervised machine learning, support vector machines, T-RFLP, sequencing, pyrosequencing, 454 sequencing*

## Introduction

Microbial communities play important roles in agriculture, bioremediation, and animal and human health, although our understanding of microbial biogeography and community profiles remains unclear. Current microbiological methods used to analyse these communities are limited, in that many microorganisms cannot be cultured or definitively identified. The application of recent molecular and bioinformatics tools is improving the sensitivity and reliability of microbial community analysis. These tools range from those using a 'broad brush strokes' approach to shed light on a microbial community profile to those involving identification of the strains and species of a random sampling of the microbes in a sample. The environmental genome shotgun survey of the Sargasso Sea[1] highlights the tremendous microbial diversity present in nature and the enormity of the effort needed to assess diversity and to understand a meta-community. This review discusses these technologies in the context of analysing the microbial communities present in the lungs of cystic fibrosis (CF) patients.

### CF

CF is a fatal inherited disease primarily affecting Caucasians. In the USA, 3,500 children are born with the disease each year.[2] The gene responsible for CF encodes a protein called the CF transmembrane conductance regulator (CFTR).[3] The CFTR is a secretory epithelial cyclic-AMP-activated chloride channel; mutations in the CFTR lead to decreased fluid secretion and dehydration of epithelial surfaces.[4] Oversecretion of thick mucus in the airway leads to congestion of the respiratory tract and increased susceptibility to chronic broncho-pulmonary infection, which is the major cause of morbidity and mortality among patients with CF.[4] To retard the rate of decreasing lung function, bacterial infections are treated with antibiotics; however, these must be tailored to the particular infection, which is often polymicrobial. For example, anti-pseudomonal drugs are often ineffective for patients treated for *Burkholderia cenocepacia* infection owing to resistance.[5] Thus, it is important to identify the infecting pathogens correctly in order to prescribe an appropriate antibiotic regimen.

### CF sputum bacterial flora

*Staphylococcus aureus*, *Haemophilus influenzae* and *Pseudomonas aeruginosa* are the primary pathogens found in the polymicrobial infection of CF patients.[6] Other opportunistic pathogens have also emerged, such as *B. cenocepacia, Alcaligenes xylosoxidans, Ralstonia pickettii, Burkholderia gladioli, Stenotrophomonas maltophilia* and *Mycobacterium* species.[6,7] *S. aureus*, the predominant pathogen in children, is succeeded by *H. influenzae* during early childhood, and *P. aeruginosa* becomes the predominant pathogen during adolescence, reaching a prevalence rate of 80 per cent in adults.[8] The occurrence of the more recently emerging organisms increases with advancing age and severity of lung disease.[8,9]

### Common assays used for clinical identification of bacteria and their limitations

Currently, the pathogens present in a CF sputum sample, throat swab or bronchoalveolar lavage (BAL) fluid are determined based on commercially available culture-based biochemical and phenotypic identification systems. These systems can either be manual, such as the API 20 NE (BioMérieux, Marcy l'Etoile, France) or fully or partly automated, such as MicroScan (Dade Behring, West Sacramento, CA, USA), BD Phoenix (Becton Dickinson, Sparks, MD, USA), and VITEK (BioMérieux).[10] These systems allow clinical microbiologists to identify bacteria accurately and rapidly, ultimately leading to better and more cost-effective patient management.[11] Misdiagnosis results from the limitation of the system's reference database[10] or from strain variation.[12] Since only about 1 per cent of eubacteria in the environment can be cultured,[13–15] a number of pathogenic species that are potentially present in the CF lung can be missed.[16] With other bacterial species (eg *Mycobacterium*), even though they can be cultured, due to their slow growth and similar phenotypes they can still be easily misdiagnosed.[17] Misidentification problems can be reduced or completely eliminated by using genotype-based molecular identification methods.[18]

## Molecular analysis of isolates

In the CF lung, some bacteria can be identified through culture whereas others would require molecular analysis. Molecular-based assays using polymerase chain reaction (PCR) and molecular markers such as 16S rRNA have been designed to identify pure isolates of many types of bacteria, including *Mycobacterium*, and will be discussed in detail.

### PCR

PCR amplifies template material from minimal amounts of extracted DNA.[19,20] This technique heralded a new era for the detection and identification of various microorganisms in any samples. Thus, the most recent techniques that study microorganisms are molecular based, using both universal and species-specific primers to select molecular markers.[19]

## Molecular marker 16S ribosomal RNA (rRNA)

rRNA plays a catalytic role in protein synthesis. The basic ribosome structure is evolutionarily conserved, although variations in overall proportions and sizes of RNA and protein exist.[21,22] A component of the small ribosomal subunit, 16S rRNA, is composed of alternating evolutionarily conserved and variable regions,[23] and is the most commonly used molecular marker.[24] The conserved regions in 16S rRNA (Figure 1) can be used to link organisms to their distant ancestors, while the highly variable regions can be used to identify evolutionary relationships between closely related organisms, at the genus and species level.[23] Studying these evolutionary relationships, however, requires the sequencing of the 16S rRNA gene.

## *Mycobacterium* spp. identification

DNA-based commercial assays have been developed to identify slow-growing Mycobacteria. *Mycobacterium tuberculosis* can be identified using the Cobas Amplicor assay, which is based on DNA hybridisation of a fragment of the 16S rRNA gene.[25] Hain Lifescience (Baden-Württemberg, Germany) developed a genotype Mycobacteria direct assay for the detection of *M. tuberculosis* complex and four atypical Mycobacteria.[25] This technique uses nucleic acid sequence-based amplification of the 23S rRNA gene. The MicroSeq system (Applied Biosystems, Foster City, CA, USA) is able to identify many *Mycobacterium* species based on the first 500 base pairs of the 16S rRNA gene.[25,26] The most used identification method is AccuProbe (Gen-Probe, San Diego, CA, USA). Isolates are grown either in solid or liquid cultures.

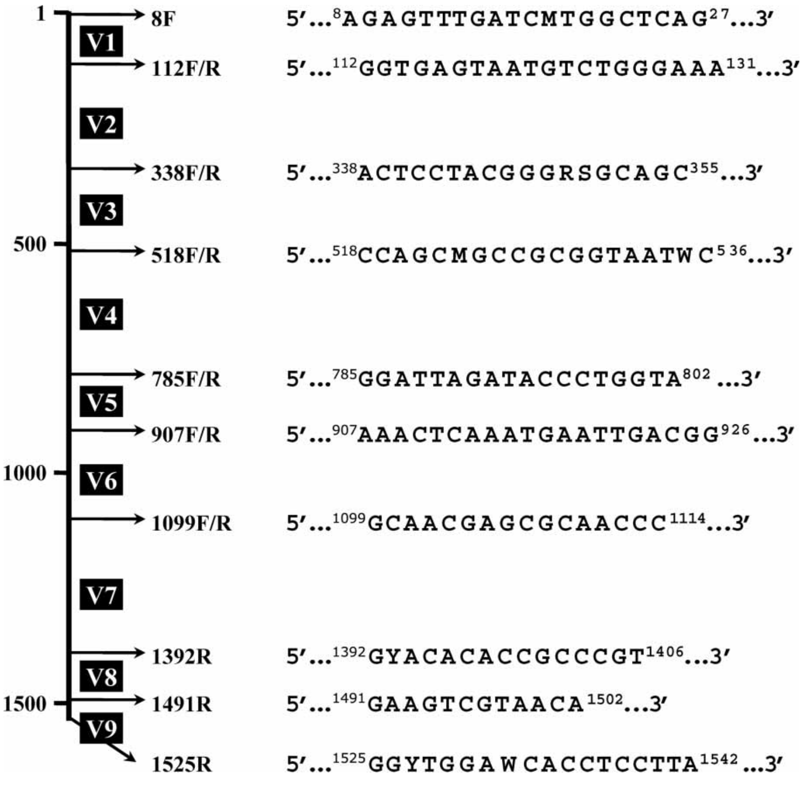
1
500
1000
1500
V1
V2
V3
V4
V5
V6
V7
V8
V9
8F 5′...⁸AGAGTTTGATCMTGGCTCAG²⁷...3′
112F/R 5′...¹¹²GGTGAGTAATGTCTGGGAAA¹³¹...3′
338F/R 5′...³³⁸ACTCCTACGGGRSGCAGC³⁵⁵...3′
518F/R 5′...⁵¹⁸CCAGCMGCCGCGGTAATWC⁵³⁶...3′
785F/R 5′...⁷⁸⁵GGATTAGATACCCTGGTA⁸⁰²...3′
907F/R 5′...⁹⁰⁷AAACTCAAATGAATTGACGG⁹²⁶...3′
1099F/R 5′...¹⁰⁹⁹GCAACGAGCGCAACCC¹¹¹⁴...3′
1392R 5′...¹³⁹²GYACACACCGCCCGT¹⁴⁰⁶...3′
1491R 5′...¹⁴⁹¹GAAGTCGTAACA¹⁵⁰²...3′
1525R 5′...¹⁵²⁵GGYTGGAWCACCTCCTTA¹⁵⁴²...3′

**Figure 1.** Schematic representation of the variable and conserved regions of the 16S rRNA genes, using *Escherichia coli rrsA* (ECDH10B_4040) as a reference. The diagram illustrates the approximate positions of nine variable (V) regions that are interspersed with conserved regions. LH-PCR primer sequences for the conserved regions are included. The 8 F, 112 F/R, 338 F/R, 518 F/R and 785 F/R primers have also been referred to as 27 F, P2, 355 F/R, 536 F/R and 802 F/R, respectively. F and R refer to forward and reverse, respectively. The degenerate nucleotides M, R, S, W and Y stand for A/C, A/G, G/C, A/T and C/T, respectively.

The cells are lysed using sonication and labelled DNA probes bound to the targeted rRNA. The resulting light emission is measured, thus identifying the isolate based on the DNA probe used in the experiment.[25] The emergence of non-tuberculous mycobacteria in CF and immunocompromised patients has created a need to assure accurate identification of these organisms. The sensitivity and accuracy of each of these assays and others vary, based on the species of Mycobacteria being analysed. These assays all rely on the isolation of bacteria and are not used to identify complex samples.

A sample containing two types of bacteria can be analysed using matrix-assisted laser desorption ionisation–time of flight mass spectrometry (MALDI-TOF-MS).[27,28] This method identifies cultivated organisms based upon the profile of proteins and peptides detected from the bacteria. In one study, CF-associated bacteria were analysed using MALDI-TOF-MS.[27] Each organism gave a specific spectrum, irrespective of how the organism had been grown (ie incubation time or media) or the presence of a mucoidy phenotype. The authors concluded that this identification technique is cost-effective, rapid and easy to use. This technique, as mentioned earlier, cannot be used to analyse complex communities.

## Molecular tools for community studies

Microbial diversity in complex microbial communities can be assessed based on the lengths of one or more of the nine variable regions of 16S rRNA (Figure 1). The PCR amplicons can be analysed using other techniques, including: terminal restriction fragment length polymorphism (T-RFLP) analysis and amplicon length heterogeneity (LH).[24,29] The fragments are separated and analysed using a capillary electrophoresis-based genetic analyser. The data generated can be subjected to bioinformatics and statistical analysis to increase their reliability. The resulting output can provide a community profile and can putatively identify individual organisms at the strain, species or genus level. The recent developments in high-throughput sequencing enable rapid sequencing of the amplicons (bacterial and fungal, with the use of appropriate primers), which is likely to lead to a rapid understanding of the community structure of any complex niche.

### T-RFLP analysis

This technique relies on the inherent variation of the sequence of a molecular marker[30] and is the most widely used method in identifying phylogenetic specificity in bacterial communities.[31] T-RFLP analysis includes PCR amplification, using one primer that is fluorescently end-labelled, restriction enzyme digestion of the amplicon and detection of the terminal restriction fragment by an automated DNA sequencer or capillary electrophoresis.[31] The resulting output consists of a microbial profile where each detected length is that of specific fragments from the digested PCR product. Each length represents one or more bacteria that have the same terminal restriction fragment length. T-RFLP profiles can be used for community differentiation, identification of specific organisms in populations and comparison of the relative phylotype richness and community structure.[30]

This method has been successful in the differentiation of bacterial communities present in many environments, including marine samples, soil samples and sputum samples from CF patients.[30–33] Rogers *et al.*[32] analysed T-RFLP amplicons of CF patient sputa and bronchoscopy samples using a computer program called MapSort (Wisconsin Package version 10.3; Accelrys, Inc., San Diego, CA, USA), which contains a database containing restriction patterns and lengths of fragments generated for known 16S rRNA bacterial sequences. The analysis suggested the presence of *P. aeruginosa, B. cenocepacia, S. aureus*, and *H. influenzae* in the CF samples.[32]

The T-RFLP method is fast and data can be easily replicated for statistical analysis. The major disadvantage of T-RFLP is that many bacteria produce similar fragment sizes, and thus not all peaks in the profiles are species specific. Some peaks may even represent more than one genus.[30,32] There are also inherent problems in using restriction enzymes, such as incomplete digestion, which can produce DNA fragments that do not correlate with the correct bacterium.[33] Therefore, to achieve better identification of the organism, further analysis — such as sequencing of the 16S rRNA gene — must be performed.

## LH

LH techniques analyse microbial populations based on the lengths of generated PCR products produced from the hypervariable regions of the 16S rRNA.[33–38] Profiles from one region are produced for the microbial community. These profiles represent the minimum diversity of bacteria present within the eubacterial community. The profiles contain peaks at specific amplicon lengths (Figure 2) representative of the number of nucleotides in the hypervariable region between the conserved regions. The peak heights are representative of the relative abundance of amplicons of that length present in the community. To identify individual bacterial organisms in the community, a database is needed. This can be generated by *in silico* analysis of known 16S rRNA sequences and the expected amplicon fragment length with a particular primer set that would be produced during an LH-PCR. The fragment lengths in the sample profile are compared against the database to identify the putative organisms. A profile resulting from this analysis suggests the presence of certain organisms and the definitive absence of others. In cases where the amplicon length is not species specific, it is often genus specific.[29] LH profiles can also be used to compare community profiles from multiple samples. Previous research has shown that the compositions of bacterial communities are highly specific to the environment in which they are found, and these differences are represented in LH profiles.[33,35] Changes in the community's niche can drastically influence bacteria and thus add specificity to the profile of a bacterial community, showing that the overall bacterial community has many unique features from sample to sample.[33,35]

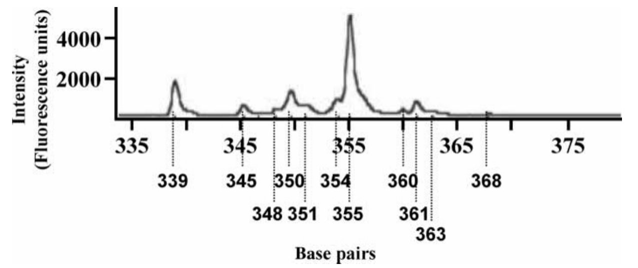

**Figure 2.** A sample amplicon length heterogeneity electropherogram using primers 8F and 338R. The *x*-axis represents amplicon length in base pairs, and relative abundance (proportional to intensity) is represented on the *y*-axis.

The main advantages of LH-PCR are that it rapidly surveys relative gene frequencies within complex mixtures of DNA, is reproducible, requires small sample sizes and can be performed simultaneously with many samples.[29] The LH profiles provide information about the members of the entire bacterial community (not just specific isolates) and their relative abundance. These data allow one to make taxonomic inferences and sample comparisons.[29] A major disadvantage of this technique is that one amplicon in the profile can represent more than one bacterium, therefore, identification at the species level cannot be guaranteed. This is also true with many length-based molecular techniques, such as T-RFLP; however, the fragments are discrete 'units' of information that can be used for comparative analyses.[30] Analysis of different combinations of the 16S rRNA variable regions will increase the power of microbial detection and sample discrimination and lead to more definitive identification.

LH was the first technique used in several ecological research projects to compare microbial communities between samples and to identify members within one community.[33,35,38] Fourteen CF sputum samples were analysed using LH-PCR for the presence of eubacteria.[32] The raw data generated from the genetic analyser were first processed with corresponding software, such as GeneimageIR v.3.56 (Scanalytics, Fairfax, VA, USA.)[32] or GeneMapper (Applied Biosystems),[35] to produce amplicon fragment lengths in base pairs. To identify presumptively the bacteria present in the CF samples, the fragment lengths were compared with a database of theoretical fragment lengths constructed using GAP (Wisconsin Package version 10.3).[32] For example, *P. aeruginosa* was identified presumptively in all 14 CF samples, five of which were confirmed by cloning and sequencing.[32] In another study, LH analysis presumptively identified *P. aeruginosa* in 19 south Florida CF patients, all of which were clinically diagnosed with this pathogen.[39] The LH fragment representing *B. cenocepacia* was not found in any of the patients, and clinical diagnosis and sequencing results confirmed the absence of this organism.[39]

To assist in the identification of individual microbial organisms in a community, we developed a software package called AmpliQué, to be used in conjunction with LH-PCR.[39] For all the bacterial and archaeal 16S rRNA sequences available from the Ribosomal Database Project (RDP) (http://rdp.cme.msu.edu/), AmpliQué computes the length of the amplicon for any specified (degenerate) primer sequence pair. For a given sample on which PCR has been performed with a fixed pair of primers, and given the lengths of the PCR products, AmpliQué infers the bacterial and archaeal organisms present in the sample. AmpliQué has recently been generalised also to handle lengths of PCR products from more than one pair of primers, enhancing the power of this *in silico* identification method. AmpliQué was used to determine the presumptive identity of organisms present in 19 south Florida CF patients based on the fragment lengths produced by LH-PCR. Oral-associated bacteria, such as *Lactobacillus mali*, *Capnocytophaga gingivalis, Porphyromonas* spp. and *Prevotella* spp. and the known CF-associated lung pathogens *P. aeruginosa, H. influenzae*, *B. cenocepacia*, *Achromobacter xylosoxidans*, *Serratia marcesens*, *S. maltophilia* and *Sarcina ventriculi*, were presumptively identified.[39]

To expand the use of LH-PCR in clinical settings, Bjerketorp *et al.*[40] combined it with a lab-on-a-chip (LOC) system, which is used for sizing and quantifying DNA, to analyse samples containing mixtures of known human gut microbes. An Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA), a bench-top instrument that uses microfluidics-based separation, was used to detect the LH fragments. This modified method allows LH-PCR to be more affordable and faster, and thus more convenient and suitable for clinical and diagnostic situations.[40] To test this system, samples containing mixtures of human gut microbes and known human gut bacteria isolates were analysed using both LOC and a capillary electrophoresis-based genetic analyser. The latter method had a higher resolution and was thus able to resolve more peaks or fragments from one another. It is important to separate PCR fragments clearly, as LH identification is based on the lengths of PCR products. Single base pair length differences are known to occur between species and even at the genus level. The level of resolution for the LOC LH-PCR technique is a weakness but the technique is rapid, economical and easier to analyse than the traditional system. Future modifications may improve the resolution, making it more useful for clinical diagnosis.[40]

## LH-related bioinformatics

Regardless of whether LH is being used to compare communities or to identify members of a community, statistics and bioinformatics must be used to derive any information produced by the technique. The first aspect of the LH-PCR system is that it profiles a community based on the patterns of lengths of amplified products (amplicons) and allows one community to be distinguished among other communities, without necessarily identifying individual species or genera.

Microbial diversity and community dynamics were first studied using computing measures, such as species richness and dominance or evenness indices.[41] Theoretical models of microbial diversity based on the log-normal distributions have also been used.[42] LH and T-RFLP data derived from soil communities have been clustered using the unweighted pair-group method using arithmetic averages (UPGMA) algorithm based on the use of distance metrics (such as the Jaccards, Hellinger or Pearson distances).[43–45] Such unsupervised methods have been used to support claims that certain relationships between communities can be discerned, that the groupings are natural and that outliers can be identified.

The statistical analysis of LH profiles is used to differentiate between two or more microbial communities. Without rigorous statistical analysis, it is impossible to differentiate between significant differences and random events. The identification of individual organisms in the community will be discussed later.

### Statistical analysis based on ecological indices

Many statistical techniques have been applied to ecological indices that measure the diversity of microbial communities. A number of diversity indices have been used with microbial communities.[41] Traditional indices include the richness (S), the Shannon information index (H) and the evenness (E) derived from it, and are defined as follows in Equations (1), (2) and (3), respectively:

$$S = \text{number of peaks of in each sample} \tag{1}$$

$$H = -\sum_i p_i(\ln p_i) \tag{2}$$

where $p_i$ is the ratio of individual peak height to the sum total of the heights of all the peaks in the LH profile.

$$E = \frac{H}{H_{max}}, \tag{3}$$

where $H_{\max} = \ln(S)$. Note that the traditional diversity indices are based on the clear definition of an ecological description of an individual species. Here, the definitions have been modified for presumptive identification of LH profiles by replacing the definition of an individual species with that of individual peaks in LH profiles.

Once appropriate diversity indices are chosen, multivariate statistical techniques, such as analysis of variance (ANOVA), can be applied to compare microbial communities.

### Statistical analysis based on abundance models

Even with the availability of the numerous diversity indices, analysing microbial diversity and communities merely using ecological indices has its shortcomings.[46] Although each index represents an attempt to distil diversity information into a single quantity, each one ends up measuring specific aspects of diversity. Diversity indices vary in their sensitivity to different abundance classes. Species abundance models are considered to be more sophisticated tools to investigate diversity because they examine the distribution of abundances in a population.

Statistical models used for species abundance of microbial communities include log series distribution, log-normal distribution,[47] the broken stick model and the overlapping niche model.[41] The most frequently used statistical model for species abundance distributions is the log-normal distribution. In log-normal communities, the null model for bacterial species abundance is a log-normal distribution defined as follows:

$$S(R) = (S_T/[\sigma(2\pi)^{0.5}])e^{[-R^2/2\sigma^2]},$$

where $S(R)$ is the number of species that contain $R$ individuals, $S_T$ is the total number of species in the community, and $\sigma^2$ is the variance of the distribution. The parameters $S_T$ and $\sigma^2$ can be estimated from a sample of measured species abundance data by using statistical techniques such as the method of moments or least squares analysis.[47]

### Supervised analysis of LH profiles

In addition to the unsupervised methods introduced above, computational tools based on supervised classification methods from machine learning have also been used for analyses based on microbial diversity.[38] These methods are used to 'learn' the differences between the diversities in the microbial communities of two sets of samples. Two well-known supervised classification tools include support vector machines (SVM) and the k-nearest neighbour method (KNN). These tools have the ability to 'learn' to classify samples after being 'trained' with 'features' from a collection of known, labelled samples. Both are computational machine-learning tools that treat the data as points or vectors in Euclidean space. These vectors are usually referred to as 'feature vectors' because their coordinates correspond to quantified 'features' of the data. These features are usually obtained after a feature extraction process. Given a new sample, it too is represented by a feature vector. In both methods, classification of the new sample is based on the location of its feature vector in relation to the location of the labelled feature vectors in the feature space.[48–51] SVMs have been shown to perform well in a variety of research areas, including pattern recognition,[52] face recognition,[53] classifications based on microarray gene expression data,[54–58] detecting remote protein homologies[59] and classifying G-protein-coupled receptors.[60] In particular, SVMs are well suited for dealing with high-dimensional data.[48,51] KNN classifiers have been successfully used in applications such as classification of handwritten digits and satellite image scenes.[50]

Computational machine learning classifiers based on SVMs and KNNs have been used to identify and compare microbial communities from different types of soil samples.[38] After a learning phase, the resulting classifiers were able to classify with high accuracy. Detailed studies using these tools revealed the limitations of the data and the minimum amount of information from LH assays that was necessary to perform reliable classification for microbial communities.[38]

## Sequencing

Even with the combined use of bioinformatics tools and LH, certain members of a community may not be identified. Sequencing of the 16S rRNA gene is imperative to identify an organism with near certainty. The most common method of sequencing is the Sanger method, developed in 1977.[61] Once the sequences are generated they are compared with known 16S rRNA sequences (stored in the Ribosomal Database Project II,[62] Greengenes[63] and GenBank[64]) to identify organisms in any samples, including the CF lung.[10,65] Sequencing of the RFLP-PCR products from the total metagenomic DNA from BAL samples of CF children identified known CF pathogens, such as *P. aeruginosa*, *S. aureus*, *S. maltophilia* and *H. influenzae*.[65] Potentially novel pathogens from the genera *Lysobacter*, *Coxiellaceae* and *Rickettsiales* were also found.[65]

Another study which involved the sequencing of the 16S rRNA gene has shown that CF sputum contains *Streptococcus mitis, S. pneumoniae, Prevotella melaninogenica, Veionella* spp., *Granulicatella para-adiacens* and *Exiguobacterium* spp., besides the normal CF pathogens, such as *P. aeruginosa*. In this study, clones were screened using LH-PCR to ensure that plasmids containing a wide array of 16S rRNA genes were sequenced.

Although sequencing technologies are able to identify bacteria in a sample more accurately, the high cost of reagents and labour may be too expensive for widespread clinical use.[66] For some bacteria, partial sequencing of the gene would lead to identification; for others, the entire gene would need to be analysed. Sequencing isolates can be performed in a timely manner and the data produced are fairly easy to analyse, especially with the use of commercial sequencing kits;[67] however, sequencing cannot differentiate between some species (eg *Mycobacterium chelonae* and *M. abscessus* are 99 per cent similar).[66] Bacterial identification would still have to be achieved using a polyphasic approach.

As with most molecular methods, non-culturable bacteria can be sequenced but this requires additional protocols, reagents and time. With traditional sequencing methods, cloning must be performed to isolate individual 16S rRNA genes amplified by PCR. Even then, further screening must be performed to ensure that multiple copies of the same 16S rRNA gene are not repetitively sequenced, thereby wasting time, reagents and money. LH can be used as a screening method to ensure that only clones of interest are sequenced. Thus, efficient identification of non-isolates poses many challenges.

## Pyrosequencing

New developments in sequencing technologies are revolutionising the way that microbial communities are being studied.[68,69] Recently developed pyrosequencing techniques that allow faster sequencing at a lower cost are opening doors for many laboratories to use sequence data for microbial identification. Pyrosequencing relies on a process referred to as sequencing-by-synthesis,[70] a technique that allows for real-time monitoring of DNA synthesis.[71] Pyrosequencing is based on the principle that pyrophosphate (PPi) is released when the

DNA polymerase adds a nucleotide to the growing complementary strand. The PPi is converted to adenosine triphosphate (ATP), which is used as a substrate in a chemical reaction that results in visible light emission. The detectable amount of light produced is relative to the amount of synthesis.[71] As with the Sanger method, pyrosequencing can only sequence individual PCR products, and thus must be used in conjunction with cloning to study microbial communities.

Pyrosequencing has been used to identify bacterial isolates by using the first and the third variable regions of the 16S rRNA.[72,73] Importantly, pyrosequencing surpassed traditional methods of detection in a clinical setting by identifying 90 per cent of the isolates at least at the genus level.[74] The remaining 10 per cent of the isolates could not be identified owing to the short sequencing reads, a clear drawback of pyrosequencing.[74] Pyrosequencing may help bacterial identification in samples that do not lend themselves to polyphasic approaches.[75,76] This technique has also been shown to distinguish clearly between multiple species of *Mycobacterium*. Three species, *Mycobacterium kansasii*, *M. scrofulaceum* and *M. gordonae*, require further sequencing analyses to obtain accurate identifications.[75] To implement pyrosequencing successfully as a diagnostic tool, the technique needs to be improved to address its limitations. Bioinformatics tools need to be refined or newly designed to handle the large amounts of data. Also, further research needs to be performed to validate the technique. In addition, issues regarding management and use of pyrosequencing in a clinical laboratory need to be addressed.[74]

### 454 sequencing

This is a new technique which allows whole-genome sequencing in a matter of days. To circumvent the need for cloning, 454 sequencing, which performs many PPi-sequencing reactions in parallel, was developed.[77] The 454 sequencing combines an emulsion-based method that isolates and amplifies DNA fragments *in vitro* with an instrument that performs pyrosequencing in picolitre-sized wells.[77] The reactions are resolved on a Genome Sequencer FLX (454 Life Sciences, Inc., Bradford, CT, USA), which reads 200–300 bases and in one run can read up to 400,000 bases.[78] This method has been used to study the microbial diversity of the deep sea[79] and the metagenome found in honey bees, which led to the discovery of a possible causative agent of colony collapse disorder.[80] A large number of microbial communities can be studied quickly and efficiently with 454 sequencing.

## Conclusion

The members of a microbial community and the associated dynamics of the niche can be studied using various methods. LH, T-RFLP and sequencing have all been used to study microbial community profiles, as well as to identify bacteria found in the CF lung. Each of these techniques has its drawbacks but can produce data that can be used (with the help of bioinformatics) to understand the composition of the community and the factors that drive it. Recent advances in technology are now the driving force behind community profiling. With the advances in high-throughput sequencing-based technologies, entire niches of organisms can be studied in a relatively short period of time. As a result, a vast amount of complex data is produced from these experiments. With the use of newly designed bioinformatics tools, data can be interpreted correctly and provide researchers with information that can ultimately be used to address community interactions that dictate the outcomes of human health studies.

## References

1. Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L. *et al.* (2004), 'Environmental genome shotgun sequencing of the Sargasso Sea', *Science* Vol. 304, pp. 66–74.
2. Cystic Fibrosis Foundation. (2006), '*Patient Registry Annual Data Report*', Cystic Fibrosis Foundation, Bethesda, MD, USA.
3. Kunzelmann, K. (1999), 'The cystic fibrosis transmembrane conductance regulator and its function in epithelial transport', *Rev. Physiol. Biochem. Pharmacol.* Vol. 137, pp. 1–70.
4. Frizzell, R.A. (1999), '*Physiology of Cystic Fibrosis*', American Physiology Society, New Haven, CT, USA.
5. McGowan, J.E., Jr. (2006), 'Resistance in nonfermenting gram-negative bacteria: Multidrug resistance to the maximum', *Am. J. Infect. Control* Vol. 34, pp. S29–S37.
6. Whittier, S. (2001), 'Update on the microbiology of cystic fibrosis: Traditional and emerging pathogens', *Clin. Microbiol. Newslett.* Vol. 23, pp. 67–71.

7. Pierre-Audigier, C., Ferroni, A., Sermet-Gaudelus, I., Le Bourgeois, M. *et al*. (2005), 'Age-related prevalence and distribution of nontuberculous mycobacterial species among patients with cystic fibrosis', *J. Clin. Microbiol*. Vol. 43, pp. 3467–3470.
8. Beringer, P.M. and Appleman, M.D. (2000), 'Unusual respiratory bacterial flora in cystic fibrosis: Microbiologic and clinical features', *Curr. Opin. Pulm. Med*. Vol. 6, pp. 545–550.
9. Lyczak, J.B., Cannon, C.L. and Pier, G.B. (2002), 'Lung infections associated with cystic fibrosis', *Clin. Microbiol. Rev.* Vol. 15, pp. 194–222.
10. Bosshard, P.P., Zbinden, R., Abels, S., Boddinghaus, B. *et al*. (2006), '16S rRNA gene sequencing versus the API 20 NE system and the VITEK 2 ID-GNB card for identification of nonfermenting gram-negative bacteria in the clinical laboratory', *J. Clin. Microbiol*. Vol. 44, pp. 1359–1366.
11. O'Hara, C.M. and Miller, J.M. (2003), 'Evaluation of the Vitek 2 ID-GNB assay for identification of members of the family Enterobacteriaceae and other nonenteric gram-negative bacilli and comparison with the Vitek GNI+ card', *J. Clin. Microbiol*. Vol. 41, pp. 2096–2101.
12. Wigley, P. and Burton, N.F. (1999), 'Genotypic and phenotypic relationships in *Burkholderia cepacia* isolated from cystic fibrosis patients and the environment', *J. Appl. Microbiol*. Vol. 86, pp. 460–468.
13. Roszak, D.B. and Colwell, R.R. (1987), 'Survival strategies of bacteria in the natural environment', *Microbiol. Rev.* Vol. 51, pp. 365–379.
14. Ward, D.M., Weller, R. and Bateson, M.M. (1990), '16S rRNA sequences reveal numerous uncultured microorganisms in a natural community', *Nature* Vol. 345, pp. 63–65.
15. Ward, D.M., Weller, R. and Bateson, M.M. (1990), '16S rRNA sequences reveal uncultured inhabitants of a well-studied thermal community', *FEMS. Microbiol. Rev.* Vol. 6, pp. 105–115.
16. van Belkum, A., Renders, N.H., Smith, S., Overbeek, S.E. *et al*. (2000), 'Comparison of conventional and molecular methods for the detection of bacterial pathogens in sputum samples from cystic fibrosis patients', *FEMS Immunol. Med. Microbiol*. Vol. 27, pp. 51–57.
17. Richter, E., Rusch-Gerdes, S. and Hillemann, D. (2006), 'Evaluation of the GenoType *Mycobacterium* assay for identification of mycobacterial species from cultures', *J. Clin. Microbiol*. Vol. 44, pp. 1769–1775.
18. Spilker, T., Coenye, T., Vandamme, P. and LiPuma, J.J. (2004), 'PCR-based assay for differentiation of *Pseudomonas aeruginosa* from other *Pseudomonas* species recovered from cystic fibrosis patients', *J. Clin. Microbiol*. Vol. 42, pp. 2074–2079.
19. Mullis, K., Faloona, F., Scharf, S., Saiki, R. *et al*. (1986), 'Specific enzymatic amplification of DNA in vitro: The polymerase chain reaction', *Cold Spring Harb. Symp. Quant. Biol*. Vol. 51, pp. 263–273.
20. Erlich, H.A., Gelfand, D. and Sninsky, J. (1991), 'Recent advances in the polymerase chain reaction', *Science* Vol. 252, pp. 1643–1651.
21. Nomura, M. and Erdmann, V.A. (1970), 'Reconstitution of 50S ribosomal subunits from dissociated molecular components', *Nature* Vol. 228, pp. 744–748.
22. Nomura, M., Morgan, E.A. and Jaskunas, S.R. (1977), 'Genetics of bacterial ribosomes', *Annu. Rev. Genet*. Vol. 11, pp. 297–347.
23. Van de Peer, Y., Chapelle, S. and De Wachter, P. (1996), 'A quantitative map of nucleotide substitution rates in bacterial rRNA', *Nucl. Acids Res.* Vol. 24, pp. 3381–3391.
24. Klappenbach, J.A., Saxman, P.R., Cole, J.R. and Schmidt, T.M. (2001), 'rrndb: The ribosomal RNA operon copy number database', *Nucl. Acids Res.* Vol. 29, pp. 181–184.
25. Neonakis, I.K., Gitti, Z., Krambovitis, E. and Spandidos, D.A. (2008), 'Molecular diagnostic tools in mycobacteriology', *J. Microbiol. Meth*. Vol. 75, pp. 1–11.
26. Cloud, J.L., Neal, H., Rosenberry, R., Turenne, C.Y. *et al*. (2002), 'Identification of *Mycobacterium* spp. by using a commercial 16S ribosomal DNA sequencing kit and additional sequencing libraries', *J. Clin. Microbiol*. Vol. 40, pp. 400–406.
27. Degand, N., Carbonnelle, E., Dauphin, B., Beretti, J.L. *et al*. (2008), 'Matrix-assisted laser desorption ionization-time of flight mass spectrometry for identification of nonfermenting gram-negative bacilli isolated from cystic fibrosis patients', *J. Clin. Microbiol*. Vol. 46, pp. 3361–3367.
28. Demirev, P.A. and Fenselau, C. (2008), 'Mass spectrometry for rapid characterization of microorganisms', *Ann. Rev. Anal. Chem*. Vol. 1, pp. 71–93.
29. Suzuki, M., Rappe, M.S. and Giovannoni, S.J. (1998), 'Kinetic bias in estimates of coastal picoplankton community structure obtained by measurements of small-subunit rRNA gene PCR amplicon length heterogeneity', *Appl. Environ. Microbiol*. Vol. 64, pp. 4522–4529.
30. Dunbar, J., Ticknor, L.O. and Kuske, C.R. (2001), 'Phylogenetic specificity and reproducibility and new method for analysis of terminal restriction fragment profiles of 16S rRNA genes from bacterial communities', *Appl. Environ. Microbiol*. Vol. 67, pp. 190–197.
31. Hiraishi, A., Iwasaki, M. and Shinjo, H. (2000), 'Terminal restriction pattern analysis of 16S rRNA genes for the characterization of bacterial communities of activated sludge', *J. Biosci. Bioeng*. Vol. 90, pp. 148–156.
32. Rogers, G.B., Hart, C.A., Mason, J.R., Hughes, M. *et al*. (2003), 'Bacterial diversity in cases of lung infection in cystic fibrosis patients: 16S ribosomal DNA (rDNA) length heterogeneity PCR and 16S rDNA terminal restriction fragment length polymorphism profiling', *J. Clin. Microbiol*. Vol. 41, pp. 3548–3558.
33. Mills, D.K., Fitzgerald, K., Litchfield, C.D. and Gillevet, P.M. (2003), 'A comparison of DNA profiling techniques for monitoring nutrient impact on microbial community composition during bioremediation of petroleum-contaminated soils', *J. Microbiol. Meth*. Vol. 54, pp. 57–74.
34. Bernhard, A.E., Colbert, D., McManus, J. and Field, K.G. (2005), 'Microbial community dynamics based on 16S rRNA gene profiles in a Pacific Northwest estuary and its tributaries', *FEMS Microbiol. Ecol*. Vol. 52, pp. 115–128.
35. Moreno, L.I., Mills, D.K., Entry, J. *et al*. (2006), 'Microbial metagenome profiling using amplicon length heterogeneity-polymerase chain reaction proves more effective than elemental analysis in discriminating soil specimens', *J. Forensic Sci*. Vol. 51, pp. 1315–1322.
36. Ritchie, N.J., Schutter, M.E., Dick, R.P. and Myrold, D.D. (2000), 'Use of length heterogeneity PCR and fatty acid methyl ester profiles to characterize microbial communities in soil', *Appl. Environ. Microbiol*. Vol. 66, pp. 1668–1675.
37. Suzuki, M.T., Rappe, M.S., Haimberger, Z.W., Winfield, H. *et al*. (1997), 'Bacterial diversity among small-subunit rRNA gene clones and cellular isolates from the same seawater sample', *Appl. Environ. Microbiol*. Vol. 63, pp. 983–989.
38. Yang, C., Mills, D., Mathee, K., Wang, Y. *et al*. (2006), 'An eco-informatics tool for microbial community studies: Supervised classification of amplicon length heterogeneity (ALH) profiles of 16S rRNA', *J. Microbiol. Meth*. Vol. 65, pp. 49–62.
39. Doud, M. (2006), 'The role of amplicon length heterogeneity–polymerase chain reaction in microbial community profiling and presumptive testing of bioagents', Masters Thesis in Forensic Science, Florida International University, Miami, FL, USA.
40. Bjerketorp, J., Ng Tze Chiang, A., Hjort, K. *et al*. (2008), 'Rapid lab-on-a-chip profiling of human gut bacteria', *J. Microbiol. Meth*. Vol. 72, pp. 82–90.
41. Hill, T.C.J., Walsh, K.A., Harris, J.A. and Moffett, B.F. (2003), 'Using ecological diversity measures with bacterial communities', *FEMS Microbiol. Ecol*. Vol. 43, pp. 1–11.
42. Dunbar, J., Barns, S.M., Ticknor, L.O. and Kuske, C.R. (2002), 'Empirical and theoretical bacterial diversity in four Arizona soils', *Appl. Environ. Microbiol*. Vol. 68, pp. 3035–3045.
43. Blackwood, C.B., Marsh, T., Kim, S.-H. and Paul, E.A. (2003), 'Terminal restriction fragment length polymorphism data analysis for quantitative comparison of microbial communities', *Appl. Environ. Microbiol*. Vol. 69, pp. 926–932.
44. Dunbar, J., Ticknor, L.O. and Kuske, C.R. (2000), 'Assessment of microbial diversity in four southwestern United States soils by 16S rRNA gene terminal restriction fragment analysis', *Appl. Environ. Microbiol*. Vol. 66, pp. 2943–2950.
45. Griffiths, R.I., Whiteley, A.S., O'Donnell, A.G. and Bailey, M.J. (2000), 'Rapid method for coextraction of DNA and RNA from natural

environments for analysis of ribosomal DNA- and rRNA-based microbial community composition', *Appl. Environ. Microbiol.* Vol. 66, pp. 5488–5491.

46. Mills, D.K., Entry, J.A., Voss, J.D., Gillever, P.M. *et al.* (2006), 'An assessment of the hypervariable domains of the 16S rRNA genes for their value in determining microbial community diversity: The paradox of traditional ecological indices', *FEMS Microbiol. Ecol.* Vol. 57, pp. 496–503.
47. Curtis, T.P., Sloan, W.T. and Scannell, J.W. (2002), 'Estimating prokaryotic diversity and its limits', *Proc. Natl. Acad. Sci. USA* Vol. 99, pp. 10494–10499.
48. Cristianini, N. and Shawe-Taylor, J. (2000), '*Support Vector Machines*', Cambridge University Press, Cambridge, UK.
49. Hastie, R. and Friedman, J. (2001), '*The Elements of Statistical Learning*', Springer, New York, NY, USA.
50. Michie, D., Spiegelhalter, D. and Taylor, C. (1994), '*Machine Learning, Neural and Statistical Classification*', Ellis Horwood, West Sussex, UK.
51. Noble, W. (2004), '*Kernal Methods in Computational Biology*', MIT Press, Cambridge, MA, USA.
52. Burges, C.J.C. (1998), 'A tutorial on support vector machines for pattern recognition', *Data Mining Knowledge Discov.* Vol. 2, pp. 121–167.
53. Waring, C.A. and Liu, X. (2005), 'Face detection using spectral histograms and SVMs', *IEEE Trans. Syst. Man. Cybern. B. Cybern.* Vol. 35, pp. 467–476.
54. Brown, M.P., Grundy, W.N., Lin, D., Cristianini, N. *et al.* (2000), 'Knowledge-based analysis of microarray gene expression data by using support vector machines', *Proc. Natl. Acad. Sci. USA* Vol. 97, pp. 262–267.
55. Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W. *et al.* (2000), 'Support vector machine classification and validation of cancer tissue samples using microarray expression data', *Bioinformatics* Vol. 16, pp. 906–914.
56. Lee, Y. and Lee, C.K. (2003), 'Classification of multiple cancer types by multicategory support vector machines using gene expression data', *Bioinformatics* Vol. 19, pp. 1132–1139.
57. Sturn, A., Quackenbush, J. and Trajanoski, Z. (2002), 'Genesis: Cluster analysis of microarray data', *Bioinformatics* Vol. 18, pp. 207–208.
58. Zheng, G., George, E.O. and Narasimhan, G. (2003), 'Neural network classifiers and gene selection methods for microarray data on human lung adenocarcinoma', in: *Proceedings of Critical Assessment of Techniques for Microarray Data Analysis (CAMDA)*, Raleigh, NC, USA.
59. Vert, J.P. (2002), 'Support vector machine prediction of signal peptide cleavage site using a new class of kernels for strings', *Pac. Symp. Biocomput.* pp. 649–660.
60. Karchin, R., Karplus, K. and Haussler, D. (2002), 'Classifying G-protein coupled receptors with support vector machines', *Bioinformatics* Vol. 18, pp. 147–159.
61. Sanger, F., Nicklen, S. and Coulson, A.R. (1977), 'DNA sequencing with chain-terminating inhibitors', *Proc. Natl. Acad. Sci. USA* Vol. 74, pp. 5463–5467.
62. Maidak, B.L., Cole, J.R., Parker, C.T., Jr., Garrity, G.M. *et al.* (1999), 'A new version of the RDP (Ribosomal Database Project)', *Nucl. Acids Res.* Vol. 27, pp. 171–173.
63. DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M. *et al.* (2006), 'Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB', *Appl. Environ. Microbiol.* Vol. 72, pp. 5069–5072.
64. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. *et al.* (2008), 'GenBank', *Nucl. Acids Res.* Vol. 36, pp. 25–30.
65. Harris, J.K., De Groote, M.A., Sagel, S.D., Zemanick, E.T. *et al.* (2007), 'Molecular identification of bacteria in bronchoalveolar lavage fluid from children with cystic fibrosis', *Proc. Natl. Acad. Sci. USA* Vol. 104, pp. 20529–20533.
66. Patel, J.B. (2001), '16S rRNA gene sequencing for bacterial pathogen identification in the clinical laboratory', *Mol. Diagnosis* Vol. 6, pp. 313–321.
67. Vliegen, I., Jacobs, J.A., Beuken, E., Bruggeman, C.A. *et al.* (2006), 'Rapid identification of bacteria by real-time amplification and sequencing of the 16S rRNA gene', *J. Microbiol. Meth.* Vol. 66, pp. 156–164.
68. Mardis, E.R. (2008), 'Next-generation DNA sequencing methods', *Annu. Rev. Genomics Hum. Genet.* Vol. 9, pp. 387–402.
69. Schuster, S.C. (2008), 'Next-generation sequencing transforms today's biology', *Nature Meth.* Vol. 5, pp. 16–18.
70. Ahmadian, A., Ehn, M. and Hober, S. (2006), 'Pyrosequencing: History, biochemistry and future', *Clin. Chim. Acta* Vol. 363, pp. 83–94.
71. Ronaghi, M., Uhlen, M. and Nyren, P. (1998), 'A sequencing method based on real-time pyrophosphate', *Science* Vol. 281, pp. 363–365.
72. Monstein, H.-J., Nikpour-Badr, S. and Jonasson, J. (2001), 'Rapid molecular identification and subtyping of *Helicobacter pylori* by pyrosequencing of the 16S rDNA variable V1 and V3 regions', *FEMS Microbiol. Lett.* Vol. 199, pp. 103–107.
73. Jonasson, J., Olofsson, M. and Monstein, H.-J. (2002), 'Classification, identification and subtyping of bacteria based on pyrosequencing and signature matching of 16S rDNA fragments', *APMIS* Vol. 110, pp. 263–272.
74. Luna, R.A., Fasciano, L.R., Jones, S.C., Boyanton, B.L., Jr. *et al.* (2007), 'DNA pyrosequencing-based bacterial pathogen identification in a pediatric hospital setting', *J. Clin. Microbiol.* Vol. 45, pp. 2985–2992.
75. Tuohy, M.J., Hall, G.S., Sholtis, M. and Procop, G.W. (2005), 'Pyrosequencing$^{(TM)}$ as a tool for the identification of common isolates of *Mycobacterium* sp', *Diagn. Microbiol. Infect. Dis.* Vol. 51, pp. 245–250.
76. Heller, L.C., Jones, M. and Widen, R.H. (2008), 'Comparison of DNA pyrosequencing with alternative methods for identification of *Mycobacteria*', *J. Clin. Microbiol.* Vol. 46, pp. 2092–2094.
77. Margulies, M., Egholm, M., Altman, W.E., Attiya, S. *et al.* (2005), 'Genome sequencing in microfabricated high-density picolitre reactors', *Nature* Vol. 437, pp. 376–380.
78. Patrick, K. (2007), '454 life sciences: Illuminating the future of genome sequencing and personalized medicine', *Yale J. Biol. Med.* Vol. 80, pp. 191–194.
79. Sogin, M.L., Morrison, H.G., Huber, J.A., Mark Welch, D. *et al.* (2006), 'Microbial diversity in the deep sea and the underexplored "rare biosphere"', *Proc. Natl. Acad. Sci. USA* Vol. 103, pp. 12115–12120.
80. Cox-Foster, D.L., Conlan, S., Holmes, E.C., Palacios, G. *et al.* (2007), 'A metagenomic survey of microbes in honey bee colony collapse disorder', *Science* Vol. 318, pp. 283–287.