

Exploring Microbial Diversity Using 16S rRNA High-Throughput Methods

Fabrice Armougom and Didier Raoult*

URMITE - UMR CNRS 6236, IRD 3R198, Université de la Méditerranée, Faculté de Médecine, 27 Boulevard Jean Moulin, 13005 Marseille, France

*Corresponding author: URMITE - UMR CNRS 6236, IRD 3R198, Université de la Méditerranée, Faculté de Médecine, 27 Boulevard Jean Moulin, 13005 Marseille, France, Tel: (33).04.91.38.55.17; Fax: (33).04.91.83.03.90; E-mail: didier.raoult@gmail.com

Received January 21, 2009; Accepted February 24, 2009; Published February 27, 2009

Citation: Fabrice A, Didier R (2009) Exploring Microbial Diversity Using 16S rRNA High-Throughput Methods. J Comput Sci Syst Biol 2: 074-092.

Copyright: © 2009 Fabrice A, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

As a result of advancements in high-throughput technology, the sequencing of the pioneering 16S rRNA gene marker is gradually shedding light on the taxonomic characterization of the spectacular microbial diversity that inhabits the earth.

16S rRNA-based investigations of microbial environmental niches are currently conducted using several technologies, including large-scale clonal Sanger sequencing, oligonucleotide microarrays, and, particularly, 454 pyrosequencing that targets specific regions or is linked to barcoding strategies. Interestingly, the short read length produced by next-generation sequencing technology has led to new computational efforts in the taxonomic sequence assignment process.

From a medical perspective, the characterization of the microbial composition of the skin surface, oral cavity, and gut in both healthy and diseased people enables a comparison of microbial community profiles and also contributes to the understanding of the potential impact of a particular microbial community.

Introduction

Until recently, the vast majority of global microbial diversity was inaccessible or largely underestimated by culture-dependent methods, since the cultivated fraction of the 4×10^{31} prokaryotic genomes moving around the biosphere (Whitman et al, 1998) is currently estimated to be 1% (Giovannoni and Stingl, 2005). However, the development of culture-independent methods and the commercialization of next-generation sequencing technology (Mardis, 2008, Rothberg and Leamon, 2008) have yielded powerful new tools in terms of time savings, cost effectiveness, and data production capability. These new tools allow for the gradual characterization of the unseen majority of environmental microbial communities. Microbial diversity has recently been

explored in a great variety of environments, including soil (Roesch et al, 2007, Yergeau et al, 2008), sea (Huber et al, 2007, Sogin et al, 2006), air (Wilson et al, 2002), and the human body, including from a medical perspective, the gastrointestinal tract (Andersson et al, 2008, Ley et al, 2006), oral cavity (Jenkinson and Lamont, 2005), vaginal tract (Zhou et al, 2004), and skin surface (Fierer et al, 2008). These microbial communities have been characterized in terms of community structure, composition, metabolic function, and ecological roles. Investigations of environmental microbial diversity have employed the 16S rRNA (16S) gene marker, which offers phylogenetic taxonomic classification without requiring isolation and cultivation. Although the use of the

16S phylogenetic marker is often criticized, due to its heterogeneity among operons of the same genome (Acinas et al, 2004) or its lack of resolution at the species level (Pontes et al, 2007), it is still considered as a 'gold standard' for bacterial identification. The use of next-generation sequencing technology has increased the size of 16S sequence databases at an impressive speed (Tringe and Hugenholtz, 2008).

Supported by new high-throughput methods (454 pyrosequencing, PhyloChip microarrays) and strategies (barcoding); the surveys of 16S gene in the human microbiota attempt to provide a comprehensive picture of the community differences between healthy and diseased states. In this review, we focus on the 16S-gene-based characterization of microbial communities using clonal Sanger sequencing, phylogenetic oligonucleotide microarrays, and 454 pyrosequencing strategies as applied to medical research. Propelled by the launch of the Human Microbiome Project, 16S high-throughput methods show tremendous potential for identifying uncultivated or rare pathogenic agents, finding shifts in the bacterial community associated with disease states (Ley et al, 2006), understanding how microbiota are affected by environmental factors of the human host (diet, lifestyle, sex, age) (Fierer et al, 2008), and differentiating between a core human microbial community and inter-individual variability (Gao et al, 2007, Turnbaugh et al, 2008). These advances will contribute to a more comprehensive picture of both healthy and diseased states and will lead to the use of more appropriate medical treatments, such as targeted antibiotic therapy rather than the use of broad-range antibiotics.

Bacterial Taxonomic Classification

The 16S rRNA Gene: A Phylogenetic Marker

In the mid-1980s, major enhancements in bacterial typing and characterization of phylogenetic relationships were achieved, using new molecular approaches based on full-length sequencing of ribosomal genes. Pioneering work by Woese and colleagues (Woese, 1987) described bacterial rRNA genes as 'molecular clocks', due to their uncommon features such as universality, activity in cellular functions, and extremely conserved structure and nucleotide sequence. The three types of rRNA in prokaryotic ribosomes are classified as 23S, 16S, and 5S, according to their sedimentation rates, and have sequence lengths of about 3300, 1550, and 120 nucleotides, respectively (Rossello-Mora and Amann,

2001). Initially, microbial diversity studies involved sequencing the 5S rRNA gene obtained from environmental samples (Lane et al, 1985, Stahl et al, 1985). However, the relatively short sequence length of the 5S gene contains few phylogenetically informative sites, which limits its usefulness for taxonomic classification purposes. In addition, although the information content of the 23S rRNA gene is larger than that of the 16S gene, it is the 16S gene that has become a standard in bacterial taxonomic classification because it is more easily and rapidly sequenced (Spiegelman et al, 2005). It is widely accepted that a compelling classification of prokaryotes should be based on a 'polyphasic approach' that includes genomic, phenotypic, and phylogenetic information (Vandamme et al, 1996). However, most bacterial diversity surveys exclusively target the 16S gene in a single-step phylogenetic approach (Pace, 1997).

The 1550 base pairs of the 16S gene are a structural part of the 30S ribosomal small subunit (SSU) and consist of eight highly conserved regions (U-U8) and nine variable regions across the bacterial domain (Jonasson et al, 2002). As no lateral gene transfer seems to occur between 16S genes (Olsen et al, 1986) and as their structure contains both highly conserved and variable regions with different evolution rates, the relationships between 16S genes reflect evolutionary relationships between organisms. A comparison of 16S gene sequence similarities is usually used as the 'gold standard' for taxonomic identification at the species level. Although thresholds are arbitrary and controversial, a range of 0.5% to 1% sequence divergence is often used to delineate the species taxonomic rank (Clarridge, III, 2004). Sequencing the 16S gene is currently the most common approach used in microbial classification as a result of its phylogenetic properties and the large amount of 16S gene sequences available for comparison analyses.

16S Gene Sequence Databases

Accurate identification of organisms by comparative analysis of 16S gene sequences is strongly dependent on the quality of the database used. The curated Ribosomal Database Project (RDP-II, <http://rdp.cme.msu.edu/>) provides 623,174 bacterial and archaeal small subunit rRNA gene sequences in an aligned and annotated format and has achieved major improvements in the detection of sequence anomalies (Cole et al, 2007). Notably, among all of the online tools provided by the RDP-II web site, the RDP classifier tool has demonstrated effective taxonomic classification of short sequences produced by the new pyrosequencing technology. The

Greengenes project (<http://greengenes.lbl.gov/>) offers annotated, chimera-checked, full-length 16S gene sequences in standard alignment formats (DeSantis et al, 2006) and has a particularly useful tool for 16S microarray design. The Silva project (<http://www.arb-silva.de>) (Pruesse et al, 2007) provides SSU as well as large subunit (LSU) rRNA sequences from Bacteria, Archaea, and Eukarya in a format that is fully compatible with the ARB package (Ludwig et al, 2004). The ARB package (www.arb-home.de) has been used in major 16S surveys (Eckburg et al, 2005, Ley et al, 2005, Ley et al, 2006, McKenna et al, 2008, Turnbaugh et al, 2006) and notably allows phylogenetic tree constructions by insertion of partial or near-full sequences into a pre-established phylogenetic tree using a parsimony insertion tool.

However, the lack of quality control of sequence entries (ragged sequence ends and outdated or faulty entries) in these major public sequence databases has led to the development of high quality commercial databases, including MicroSeq 500 and the RIDOM *Mycobacteria* project (<http://www.ridom-rdna.de>) (Harmsen et al, 2002). The MicroSeq 500 database targets the first 527-bp fragment of the 16S gene and is able to identify most of the clinically important bacterial strains with ambiguous biochemical profiles (Woo et al, 2003). The ribosomal differentiation of the medical microorganism (RIDOM) database targets the 5' end of the 16S sequence and is dedicated to *Mycobacteria* family analyses. However, although these commercial databases are continually expanding, the current total number of 16S entries remains uncertain, and the representation of taxonomic divisions is limited.

Measures of Microbial Diversity

The assessment of microbial diversity in a natural environment involves two aspects, species richness (number of species present in a sample) and species evenness (distribution of relative abundance of species) (Magurran, 2005). In order to estimate species richness, researchers widely rely on the assignment of 16S sequences into Operational Taxonomic Unit (OTU or phylotype) clusters, for instance, as performed by DOTUR (Schloss and Handelsman, 2005). The criterion used to define an OTU at the species level is the percentage of nucleotide sequence divergence; the cut-off values vary between 1%, 3%, or 5%, depending on the study. As a result of these inconsistencies, reliable statistical comparisons or descriptions of species richness across studies are restricted (Martin, 2002). The total community diversity of a single environment, or the α -diversity, is often

represented by rarefaction curves. These curves plot the cumulative number of OTUs or phylotypes captured as a function of sampling effort and, therefore, indicate only the OTU richness observed in a given set of samples (Eckburg et al, 2005). In contrast, nonparametric methods, including Chao1 or ACE, are richness estimators of overall α -diversity (Roesch et al, 2007). In addition, quantitative methods such as the Shannon or Simpson indices measure the evenness of the α -diversity. However, although these estimators can describe the diversity of the microbiota associated with a healthy or diseased state, they are not informative of the (phylo)genetic diversity of an environmental sample (Martin, 2002).

Contrary to the α -diversity, the β -diversity measure offers a community structure comparison (taxon composition and relative abundance) between two or more environmental samples. For instance, β -diversity indices can compare similarities and differences in microbial communities in healthy and diseased states. A broad range of qualitative (presence/absence of taxa) and quantitative (taxon abundance) measures of community distance are available using several tools, including LIBHUFF, P-test, TreeClimber (Schloss and Handelsman, 2006b), SONS (Schloss and Handelsman, 2006a), DPCoA, or UniFrac (Lozupone et al, 2006, Lozupone and Knight, 2005); these methods have been thoroughly detailed in a previous review (Lozupone and Knight, 2008). For example, the robust unweighted UniFrac tool (Liu et al, 2007) measures the phylogenetic distance between two communities as the fraction of phylogenetic tree branch lengths leading to a descendant from either one community or the other. While UniFrac efficiently detects differences in the presence or absence of bacterial lineages, the recently developed weighted UniFrac is the qualitative version of original UniFrac and provides an efficient detection of differences in the relative abundance of bacterial lineages (Lozupone et al, 2007).

16S High-throughput Methods

16S Clonal Sanger Sequencing

Until the appearance in the two last decades of sequencing-by-synthesis methods (Ronaghi et al, 1996) such as those used in pyrosequencing, the Sanger sequencing method (Sanger et al, 1977) was the cornerstone of DNA sequence production. The Sanger method is based on DNA synthesis on a single-stranded template and di-deoxy chain-termination (Hall, 2007, Pettersson et al, 2008). Improvements in

cost-effectiveness and the development of high-throughput techniques (e.g, fluorescent-labeled terminators, capillary separation, template preparation) (Hunkapiller et al, 1991) have enabled direct sequencing of clones, without laborious prior screening by restriction analysis. As a result, Sanger sequencing has produced the earliest in-depth analyses of microbial communities (Eckburg et al, 2005, Ley et al, 2006, Turnbaugh et al, 2006, Zhou et al, 2004). All 16S genes in a sample are amplified using a universally conserved primer pair that targets most of the species that have been sequenced and deposited in the ribosomal databases. After the cloning of amplified PCR products into specific vectors, the inserts are sequenced. Sanger sequencing offers high phylogenetic resolution power, as the method yields the longest sequencing read available, up to 1000 base pairs (bp) (Shendure and Ji, 2008).

16S Microarray-based Approach

The DNA microarray is a powerful technology that can simultaneously detect thousands of genes on a single glass slide or silicon surface (Gentry et al, 2006). Mainly used in gene expression profiling (DeRisi et al, 1997, Schena et al, 1996), DNA microarray technology has also been employed in bacterial identification and, more recently, has been adapted for exploring microbial community diversity in environmental niches. Microarrays used for microbial identification rely on the 16S gene and use short 20- to 70-mer oligonucleotide probes (Bae and Park, 2006) for multi-species detection. These are referred to as phylogenetic oligonucleotide microarrays (POAs).

Due to the phenomenal microbial diversity that might be

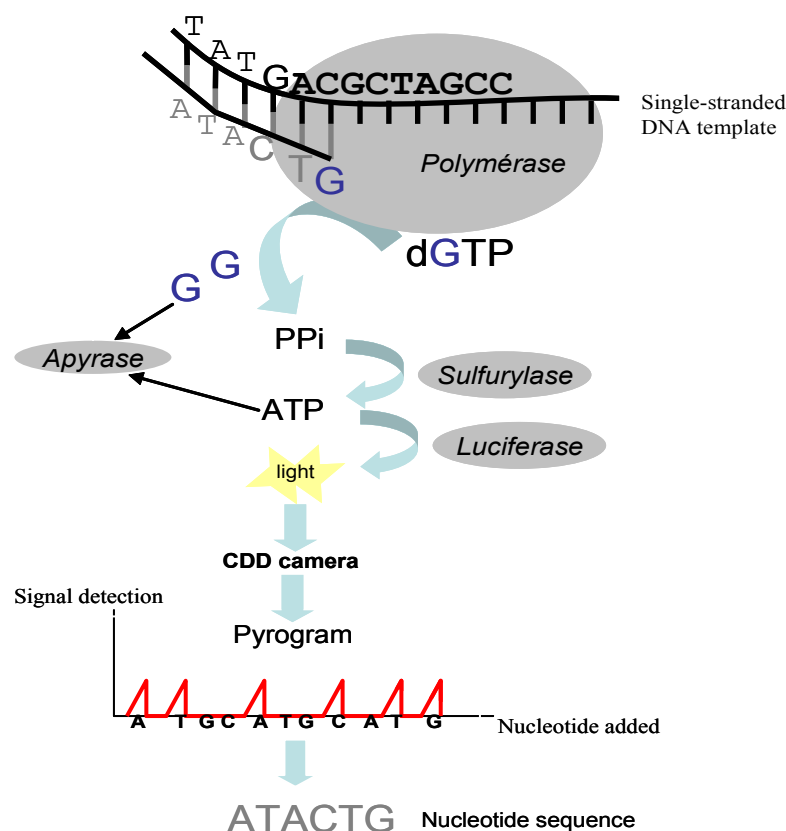


Figure 1: Principle of pyrosequencing technology

The primer for the sequencing step is hybridized to a single-stranded DNA template, and incubated with the enzymes, DNA polymerase, ATP sulfurylase, luciferase and apyrase, and the substrates. Deoxyribonucleotide triphosphate (dNTP) is added, one at a time, to the pyrosequencing reaction. The incorporation of a nucleotide is accompanied by release of pyrophosphate (PPi). The ATP sulfurylase quantitatively converts PPi to ATP. The signal light produced by the luciferase-catalyzed reaction in presence of ATP is detected by a charge coupled device (CCD) camera and integrated as a peak in a Pyrogram. The nucleotide degrading Apyrase enzyme continuously degrades ATP excess and unincorporated dNTPs. The process continues with addition of the next dNTP and the nucleotide sequence of the complementary DNA strand is inferred from the signal peaks of the pyrogram.

present in an environmental sample and the lack of prior knowledge of its population composition, the oligonucleotide probe design strategy has been modified for use in community diversity analysis. Instead of using one unique probe that targets a specific region of one taxon, multiple probe sets are employed to target microorganisms at different taxonomic levels. An efficient probe design based on a hierarchical phylogenetic framework was established in 2003, using a database of curated and aligned 16S sequences known as ProkMSA (DeSantis et al, 2003). Desantis *et al.* defined 9,020 OTUs in the ProkMSA alignment using sequence identity clustering and demonstrated the capability of microarrays to correctly assign two species to their OTU using multiple probe sets (DeSantis et al, 2003).

High-throughput microorganism detection by microarray technology was first achieved with high-density photolithography microarrays using 31,179 20-mer oligonucleotide probes in a deep air investigation (Wilson et al, 2002). These high density microarrays present an alternative to clone library sequencing, since they can more deeply assess microbial diversity (DeSantis et al, 2007) without a cloning bias. With advancements in technology, the PhyloChip platform, an Affymetrix microarray product, was developed by the Lawrence Berkeley National Laboratory. The PhyloChip has rapidly identified up to 8,900 distinct environmental microorganisms at different taxonomy levels from soil (Yergeau et al, 2008), air (Brodie et al, 2007), or uranium-contaminated site samples (Brodie et al, 2006) in a single experimental run. The PhyloChip is a glass slide with a small surface area, containing a high density microarray of hundreds of immobilized oligonucleotides (15- to 25-mer length). The method employs parallel hybridization reactions, using a flow of fluorescently labeled DNA targets. The microarray slide is analyzed with a fluorescence microscope equipped with a cooled CCD camera.

Pyrosequencing Technology

Sequencing Chemistry

Pyrosequencing is a DNA sequencing method (Clarke, 2005, Ronaghi, 2001, Ronaghi and Elahi, 2002) based on the sequencing-by-synthesis principle, which was first described in 1985 (Melamed and Wallace, 1985). This method relies on efficient detection of the sequential incorporation of natural nucleotides during DNA synthesis (Ronaghi et al, 1996, Ronaghi et al, 1998) (Figure 1). The pyrosequencing technique includes four enzymes that are involved in a cascade

reaction system (Figure 1).

During the reaction, the Klenow fragment of DNA polymerase I releases inorganic pyrophosphate molecules (PPi) upon the addition of one nucleotide to a primer hybridized to a single-stranded DNA template. The second reaction, catalyzed by ATP sulfurylase, produces ATP, using the released PPi as a substrate. The ATP molecules are then converted to a luminometric signal by the luciferase enzyme. Therefore, the signal light is detected only if a base pair is formed with the DNA template, and the signal strength is proportional to the number of nucleotides incorporated in a single nucleotide flow. Finally, the unincorporated nucleotides and excess ATP are degraded between base additions by a nucleotide-degrading enzyme such as apyrase (Ronaghi et al, 1998); at this point, another dNTP is added and a new cycle begins. The earliest attempts at pyrosequencing were performed using the PSQ96 system (Biotage AB, Uppsala, Sweden) and targeted the short 16S variable regions V1 and V3 (Jonasson et al, 2002, Tarnberg et al, 2002) or the human pathogen *H. pylori* (Hjalmarsson et al, 2004). This system produced reads of an average length of around 20-40 bases.

The 454 Life Sciences Pyrosequencing Platform

In 2005, Margulies *et al.* first described a highly parallel sequencing platform (GS 20 454 Life Sciences) using a pyrosequencing protocol optimized for solid support. The authors demonstrated the ability of the system to assemble complete genomes (*Mycoplasma genitalium* and *Streptococcus pneumoniae*) from short sequencing reads (Margulies et al, 2005). In 454 pyrosequencing, the DNA template is fragmented, and the resulting fragments are individually immobilized onto a bead by limiting dilution. Emulsion PCR is performed for the DNA amplification step, in which each DNA fragment is independently confined into a droplet of oil and water containing PCR reagents. The clonally amplified fragments are distributed into a picotiter plate, which contains ~1.6 million picoliter wells, with a well diameter allowing one bead per well.

Using the pyrosequencing protocol previously described, the chemiluminescent signal obtained from an incorporated nucleotide is recorded by a charge-coupled device (CCD) camera, and data analysis, such as image processing or *de novo* sequence/genome assembly, is performed with provided bioinformatics tools. Other massively parallel platforms (Solexa and SOLID) using flow cell and other sequencing

chemistries were reviewed previously (Holt and Jones, 2008).

While the first generation 454 Life Sciences apparatus (GS20) provided up to 25 megabases of data with an average read length of 100 base pairs (bp), the new GS FLX Titanium provides up to 400 megabases of data with an average read length of 400 bp. The 454 Life Sciences high-throughput sequencing platform outperforms the clonal Sanger sequencing method in terms of time and cost per sequenced nucleotide (Hugenholtz and Tyson, 2008), capacity of sequencing per run, as well as time savings in library preparation and data analysis. The contribution of 454 pyrosequencing to the literature is considerable, as 283 publications since 2005 have reported the use of this technology (Figure 2), and more than 20% of those publications belong to Science or Nature group journals. When applied to the field of microbial diversity, 454 pyrosequencing offers major insights, particularly in the investigation of human gut flora (Dethlefsen et al, 2008, Turnbaugh et al, 2006, Turnbaugh et al, 2008), vaginal microbiota (Spear et al, 2008), hand surface bacteria (Fierer et al, 2008), soil diversity (Roesch et al, 2007), deep sea ecosystems (Huber et al, 2007, Sogin et al, 2006), and viral and phage populations (Desnues et al, 2008, Eriksson et al, 2008, La et al, 2008). Due to the short read length generated by the 454 platform and in order to increase sequencing capacity, new strategies for exploring microbial diversity by 16S pyrosequencing were developed.

16S Pyrosequencing Strategies

Targeting Specific Regions

Due to the short sequencing read length generated by pyrosequencing technology (e.g, 100 bp for GS 20 and 200 bp for GS FLX) and due to the small amount of nucleotide variability in the 16S gene throughout the bacterial domain, full 16S gene sequence assembly and the taxonomic assignment of species present in a mixed microbial sample remain a computational challenge (Armougom and Raoult, 2008). One strategy of addressing this problem consists of targeting a specific variable 16S region that exhibits a sufficient phylogenetic signal to be accurately assigned at the genus level or below. Surprisingly, short sequence fragments (including 100 bases) can provide substantial phylogenetic resolution (Liu et al, 2007). By choosing appropriate 16S regions and simulating the read length obtained with GS FLX (250 bases), Liu *et al.* reproduced the same results obtained from full length 16S sequences using the UniFrac clustering tool (Liu et al, 2007). The authors suggested that the F8-R357 primer set, which amplifies a sequence spanning the V2 and V3 hypervariable regions and generates a 250-bp amplicon, could be the optimal primer set for exploring microbial diversity using 16S 454 pyrosequencing with GS FLX. These primers were also used in a study of the macaque gut microbiome (McKenna et al, 2008). A more recent study reported by Liu *et al.*, focused on the capabilities of different taxonomic assignment methods (as a function of se-

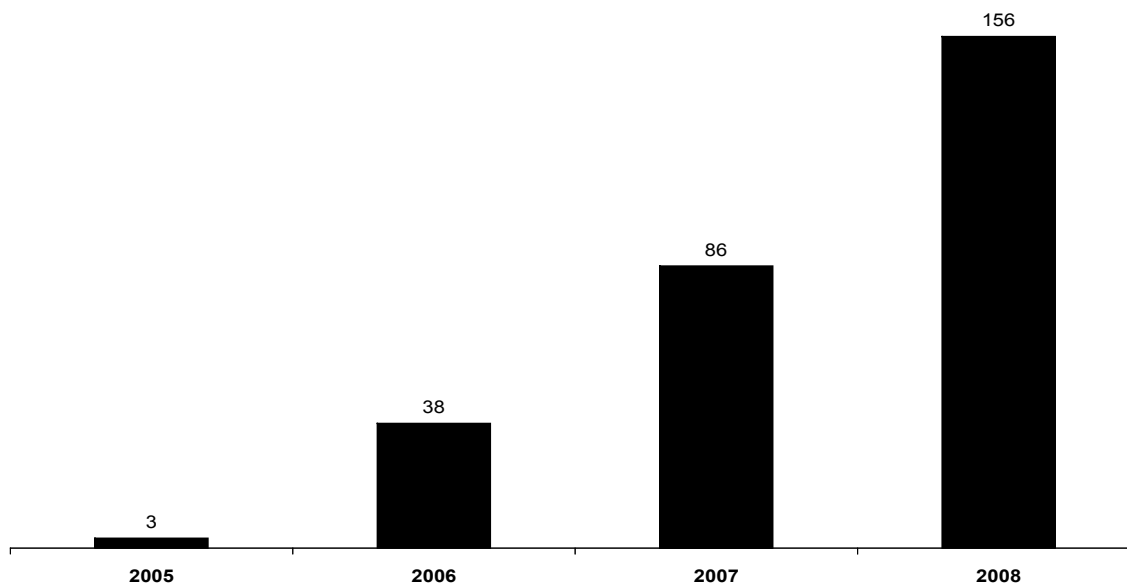


Figure 2: Number of publications enabled by the 454 pyrosequencing technology since 2005.

Reference	Samples	Forward Primer	Reverse Primer	16S regions
(Roesch et al, 2007)	Soil	787F GNTACCTTGTTACGACTT	1492R ATTAGATACCCNGGTAG	V5 to V9
(Andersson et al, 2008)	Human gut	784F: GCCTTGCCAGCCCGCTCAG	1061R: AGGATTAGATACCCTGGTA	*V5-V6
(Dowd et al, 2008a)	Cattle feces	530F: GTGCCAGCMGCNGCGG	1100R: GGGTTNCGNTCGTTG	*V4 to V6
(Huse et al, 2008)	Human gut & Deep sea vent	338F: ACTCCTACGGGAGGCAGCAG 967F: CAACGCGAAGAACCTTACC & ATACGCGA[AG]GAACCTTACC	533R: TTACCGCGGCTGCTGGCAC 1046R: AGGTGNTGCATGGCTGTTCG & 1046R1: AGGTGNTGCATGGTTGTTCG	V3 — V6
(McKenna et al, 2008)	Macaque gut	F8: AGAGTTTGATCCTGGCTCAG	R357: CTGCTGCCTYCCGTA	*V2-V3
(Sogin et al, 2006)	Deep sea	967F: CAACGCGAAGAACCTTACC	1046R: CGACAGCCATGCANCACCT	V6
(Kim et al, 2008)	Tidal flat sediments	bact363F:CAATGGRSGVRASYCTGAH S Arch339F: GGYGCASCAGGCGCGVAW	Bact531R:CTNYGTMTTACCGCGGCTG C Arch523R: TMCCGCGGCKGCTGVCASC	V3
(Turnbaugh et al, 2008)	human Gut	8F: AGAGTTTGATCCTGGCTCAG F : See (Huber et al, 2007)	338R: RTGCTGCCTCCCGTAGGAGT R: See (Huber et al, 2007)	*V2 *V6
(Keijser et al, 2008)	Oral cavity	909F: AAACYAAARRAATTGACGG & 917F: GAATTGACGGGGRCCCGCA	1061R: TCACGRCACGAGCTGACGAC	V6
(Fierer et al, 2008)	Hand surface	27F: AGAGTTTGATCCTGGCTCAG	338R: CATGCTGCCTCCCGTAGGAGT	*V2-V3
(Dethlefsen et al, 2008)	Human Gut	F1: CAACGCGAAGAACCTTACC & F2: ATACGCGAGGAACCTTACC F: ACTCCTACGGGAGGCAGCAG	R: CGACARCCATGCASCACCT R: TTACCGCGGCTGCTGGCAC	*V6 *V3
(Spear et al, 2008)	vagina	27F: AGAGTTTGATCCTGGCTCAG	355R: GCTGCCTCCCGTAGGAGT	*V2-V3
(Huber et al, 2007)	Deep sea	967F-PP: CNACGCGAAGAACCTTANC & 967F-UC1: CAACGCGAAAAACCTTACC & 967F-UC2: CAACGCGCAGAACCTTACC & 967F-UC3: ATACGCGARGAACCTTACC & 967F-AQ: CTAACCGANGAACCTYACC 958arcF: AATTGGANTCAACGCCGG	1046R CGACAGCCATGCANCACCT & 1046RPP CGACAACCATGCANCACCT & 1046RAQ1 CGACGGCCATGCANCACCT & 1046RAQ2 CGACGACCATGCANCACCT 1048arcR: CGRCGGCCATGCACCWC & 1048arcR: CGRCRGCCATGYACCWC	*V6

Table 1: Choice of primer set in microbial diversity investigations using 16S 454 pyrosequencing

A Mixture of Forward (F) or Reverse (R) primers is indicated by “&”. In italic, primer set targeting the Archaeal domain. Studies that used the barcoding strategy are indicated with an asterisk (*). The 16S rRNA variable region (Vx) targeted by the study is indicated in the last column.

lected 16S regions), confirmed their previous conclusions recommending the F8-R357 primer set. The authors also suggested that regions surrounding the V6 hypervariable region were not appropriate for taxonomic assignment using a 16S 454 pyrosequencing strategy (Liu et al, 2008).

Whereas Liu *et al.* laud the use of the F8-R357 primer pair the study that introduced the concept of tag pyrosequencing was performed using the V6 hypervariable region amplified from deep water samples of the North Atlantic (Sogin et al, 2006). Based on the Shannon entropy measure, the 16S gene shows high variability in the V6 region. This region was selected in an analysis of the human gut microbiome combined with a barcoding strategy (Andersson et al, 2008). In addition, in their recent study of a gut microbial community, Huse *et al.* showed that the use of tags of the V3 and V6 regions in 454 pyrosequencing notably provides taxonomic assignments equivalent to those obtained by full length sequences generated using clonal Sanger sequencing (Huse et al, 2008).

The taxonomic assignment of pyrosequencing reads is sensitive to the classification method employed (Liu et al, 2008). Wang *et al.* showed a classification accuracy map of their RDP classifier tool along the 16S sequence position (Wang et al, 2007), which revealed that variable V2 and V4 regions of 16S provided the best taxonomic assignment results at the genus level. No true consensus seems to emerge among the 16S 454 pyrosequencing studies that target variable regions (Table1). However, this is not surprising since the power of phylogenetic resolution of variable 16S regions might differ depending on the taxa present in the mi-

crobial community studied; these results may also be due to the under-representation of 16S sequences of certain environments in the reference databases (Huse et al, 2008).

16S Barcoding Strategy

The currently available 454 GS FLX pyrosequencer can accommodate a maximum of sixteen independent samples, since a picotiter plate contains sixteen physically separated regions.

To overcome this limitation and expand the capacity by pooling DNA from independent samples in a single sequencing run, a barcoding approach was developed, which associates a short unique DNA sequence tag (barcode) with each DNA template origin. In contrast with Sanger sequencing, pyrosequencing technology such as GS FLX generates sequencing reads from the first position of the DNA template fragment. Therefore, the sequencing reaction driven by an oligonucleotide that is complementary to adaptor A and B first reads the barcode sequence, allowing the identification of the original DNA template source (Andersson et al, 2008, McKenna et al, 2008). Using the DNA barcode strategy, each primer is composed of the 5'-adaptor A or B (required for the PCR emulsion), followed by the DNA barcode and the primer targeting the DNA region (McKenna et al, 2008) (Figure 3). By selecting a DNA target with a nucleotide length inferior to the average read length of pyrosequencing, an unidirectional barcoding strategy can be achieved, adding the barcode sequence only to the forward or reverse primer (Andersson et al, 2008). The length of the barcode sequence varies from 2-4 nucleotides

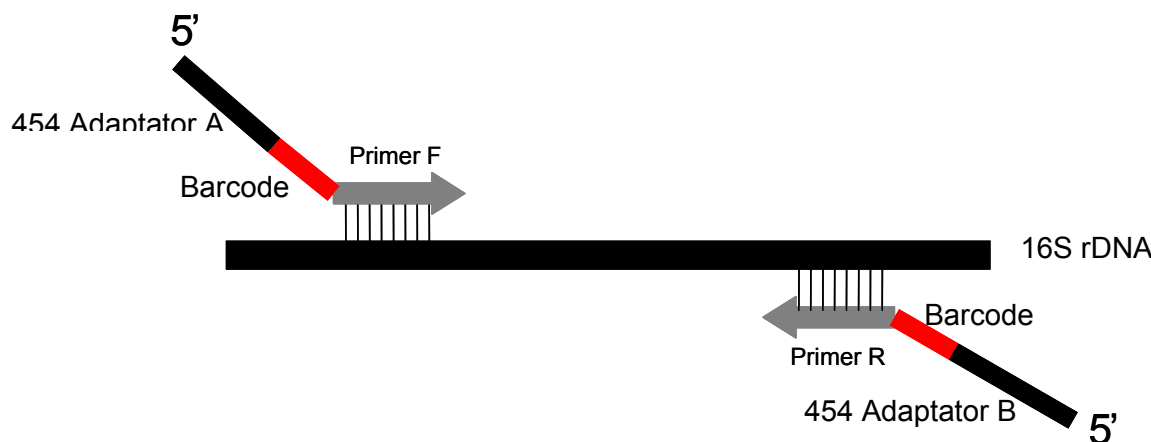


Figure 3: The barcoding strategy used in 454 pyrosequencing

Each primer contains a region complementary to the 454 A or B adaptor sequencing primers, a variable barcode sequence length (in red) and a region complementary to ends of the 16S rRNA region targeted (Primer R or F)

(Andersson et al, 2008, Binladen et al, 2007, McKenna et al, 2008) to 10 nucleotides (Dowd et al, 2008b, Dowd et al, 2008a) or 12 nucleotides (Fierer et al, 2008, Turnbaugh et al, 2008); the number of DNA samples that can be combined in a single sequencing run increases with barcode length.

Finally, the efficiency of barcoding strategy was investigated using a set of eight nucleotide barcodes based on error-correcting codes (called Hamming codes). The ability to detect sequencing errors that change sample assignments and to correct errors in the barcodes was evaluated to 92% (Hammady et al, 2008).

Taxonomy Assignment of Short Pyrosequencing Reads

Taxonomy assignment using standard phylogenetic methods such as likelihood- or parsimony- based tree construction is inconceivable, given the large amount of sequence data (400,000 reads for GS FLX Titanium) generated by high-throughput pyrosequencing (Liu et al, 2008). In addition, due to the short read length resulting from pyrosequencing (~ 100 bp for GS 20 and 250 bp for GS FLX), full 16S sequence assembly is a computational challenge, especially for closely related species in a mixed bacterial sample. Therefore, current taxonomy classification tools such as the naïve Bayesian RDP-II classifier (Wang et al, 2007) or the Greengenes classifier (DeSantis et al, 2006) employ rapid but approximate methods. In contrast to the RDP-II classifier, the Greengenes classifier requires a pre-computed alignment for its taxonomic classification. While, tree-based methods are subject to large variations in assignment accuracy according to the DNA region examined (Liu et al, 2008), the tree-independent methods used in the RDP-II and Greengene classification tools yield stable and accurate taxonomy assignment results. However, although these methods provide satisfactory assignment results at the genus level, they have a limited resolution power at the species level. This limitation could be reduced, but not yet solved, by the 400- to 500-base read length generated by the new GS Titanium pyrosequencing platform.

Pyrosequencing read classification based on a sequence similarity search using the BLAST algorithm can yield reliable results; however, the closest match is not inevitably the nearest phylogenetic neighbor (Koski and Golding, 2001). Sundquist *et al.* (Sundquist et al, 2007) recently proposed a method based on a similarity search with BLAT, a BLAST-like alignment tool (Kent, 2002), and inferred specific phy-

logeny placement using a preliminary set of best BLAT match scores. Finally, a tag-mapping methodology has been recently introduced with GAST (Global Alignment for Sequence Taxonomy), which is a taxonomic assignment tool combining BLAST, multiple sequence alignment, and global distance measures. GAST has been used to identify taxa present in deep sea vent and human gut samples (Huse et al, 2008, Sogin et al, 2006).

The accuracy rate of classification methods of short 16S rRNA sequences is measured as the percentage of sequences correctly classified from a representative set of bacteria sequences of known classification. This accuracy rate can be associated with the misclassified sequence rate. Overall, the simulation tests showed efficient classification down to the genus level. Simulating 200-base segments (such as generated by the GS FLX), the RPD-II classifier tool indicated an overall taxonomic assignment accuracy of 83.2% at the genus level (Wang et al, 2007). Likewise, using the methodology of Sundquist and co-workers (Sundquist et al, 2007), the simulation of read resolution for 200 base segments in diverse and representative samples yield an accuracy rate around 80% as obtained for the RDP-II classifier. However, while the benchmark of the Sundquist method identified the V1 and the V2 regions of the 16S as the best targets for the pyrosequencing of 100 bases (such as generated by the GS 20), the benchmark of the RDP-II classifier method identified the V2 and the V4 regions. Finally, the comparison of the classification methods is rather difficult since the representative bacteria sequences of known classification used for the benchmarking is different for each method. A collection of reference sequences of known classification (defined as gold standard) is required for classification method comparisons. For instance, collections of reference sequence alignments are generally used for the benchmarking of multiple sequence alignment methods (Armougom et al. 2006).

Metagenomic Approach

Metagenomic is a culture-independent genomic analysis of entire microbial communities inhabiting a particular niche, such as the human gut (Riesenfeld et al, 2004, Schmeisser et al, 2007). Metagenomic investigations aimed at finding “who’s there and what are they doing” (Board On Life Sciences, 2007) are providing new insight into the genetic variability and metabolic capabilities of unknown or uncultured microorganisms (Turnbaugh et al, 2006). The inclusion of the metagenomic approach in this review was re-

quired, since metagenomic studies include the analysis of the 16S sequences contained in metagenome data, in order to identify community composition and determine bacterial relative abundance (Biddle et al, 2008, Edwards et al, 2006, Krause et al, 2008, Wegley et al, 2007). However, we distinguished the metagenomic approach from 16S high-throughput methods, as the principal purpose of metagenomic is to explore the entire gene content of metagenomes for metabolic pathways and to understand microbial community interactions (Tringe et al, 2005), rather than targeting a single gene such as 16S. In contrast with some scientific literature examples and because it can only answer “who is there”, 454 pyrosequencing investigations based on the 16S gene should not be assimilated to a metagenomic case.

Before the commercialization of the 454 pyrosequencing platform, microbial community sequencing in early metagenomic studies, such as in the Sargasso sea (Venter et al, 2004) or in acid mine drainage (Edwards et al, 2000), involved preliminary clone library construction and capillary sequencing. This approach limited the expansion of microbial diversity knowledge due to cloning bias and the cost of capillary sequencing. In contrast, the inexpensive next-generation 454 pyrosequencing technology can perform direct sequencing without requiring preliminary PCR amplification or library construction. By excluding cloning and PCR bias, 454 sequencing revolutionized the metagenomic field by capturing up to 100% (depending on the quality of DNA extraction and environmental sampling) of the microbial diversity present in a sample.

Case Studies

The Human Microbiome Project

Launched by the NIH Roadmap for Medical Research, the Human Microbiome Project (HMP) seeks to comprehensively characterize the human microbiota in order to better understand its role in human health and disease states (<http://nihroadmap.nih.gov/>) The HMP project mainly focuses on the gastrointestinal, oral, vaginal, and skin microbiota.

The Gastrointestinal Microbiota

Until recently, characterization of the gut microbiota diversity was restricted to culture-based methods (Simon and Gorbach, 1986). While the cultivable fraction is currently estimated to be 442 bacterial, 3 archaeal, and 17 eukaryotic

species (Zoetendal et al, 2008), the species richness of the gut microbiota is estimated to be 15,000 or 36,000 bacterial species, depending on the similarity cut-off applied in OTU clustering (Frank et al, 2007). With the development of culture-independent methods, 16S gene surveys have deeply enhanced the microbial diversity map of the human gut microbiota (Eckburg et al, 2005, Gill et al, 2006, Ley et al, 2006, Palmer et al, 2007). A major large-scale 16S investigation by Eckburg *et al.* indicated that the gut of healthy individuals was mainly composed of the *Bacteroidetes* and *Firmicutes* bacterial phyla (83% of sequences). The sequences also included the archaeal *Metanobrevibacter* species, as well as a majority of uncultivated species and novel microorganisms (Eckburg et al, 2005). Surprisingly, at the phylum level, the bacterial diversity of the gut microbiota is low; only 8 of 70 known bacterial phyla are represented. Despite the predominance of *Firmicutes* and *Bacteroidetes* phylotypes, the gut microbiota displays a great inter-individual specificity in its composition (Ley et al, 2006), especially in newborn babies (Palmer et al, 2007). Within elderly populations, the *Bacteroidetes* proportion can decline (Woodmansey, 2007). Recently, a barcoded pyrosequencing study of the gut microbiota of six elderly individuals showed that *Actinobacteria* was the second most abundant phylum, not *Bacteroidetes* as expected (Andersson et al, 2008). Age, caloric intake, antibiotic treatment (Dethlefsen et al, 2008) and diet are a few environmental factors that can influence the gut microbiota composition and thus affect human health.

Through the comparison of lean and obese individuals, a possible relationship between obesity and the composition of (and changes in) gut microbiota has been investigated (Ley et al, 2006, Turnbaugh et al, 2008) and reviewed (Dibaise et al, 2008). Ley *et al.* showed that obese subjects have a higher *Firmicutes/Bacteroidetes* ratio than lean controls (Ley et al, 2006). By testing dietary factors, the authors demonstrated a shift in *Bacteroidetes* and *Firmicutes* relative abundance that correlated with weight loss. Reduced bacterial diversity and familial similarity of gut microbiota composition within obese individuals have also been recently reported (Turnbaugh et al, 2008).

The Oral Cavity Microbiota

The understanding of healthy oral cavity microflora is essential for the prevention of oral diseases and requires unambiguous identification of microorganism(s) associated with pathology. For instance, it is accepted that *Streptococcus*

mutans is the etiological agent in dental caries (Jenkinson and Lamont, 2005). Limited by the cloning and sequencing approach, the characterization of the diversity in human oral microflora was radically enhanced by the first 16S 454 pyrosequencing of saliva and supragingival plaque (Keijser et al, 2008). Keijser *et al.* revealed 3,621 and 6,888 phylotypes in saliva and plaque samples, respectively, and estimated the total microbial species richness to be 19,000 (3% similarity cut-off). Within the 22 phyla identified, the main taxa are *Firmicutes* (genus *Streptococcus* and *Veillonella*) and *Bacteroidetes* (genus *Prevotella*) in saliva, while *Firmicutes* and *Actinobacteria* (genus *Corynebacterium* and *Actinomyces*) are the most common in plaque.

The Vaginal Microbiota

By its exposure to the external environment, the female genital tract can be easily affected in its reproductive functions. Previous surveys of the human vaginal microbiota proposed that the normal vaginal microbiota can act as a defense mechanism, playing an essential role in preventing infections such as bacterial vaginosis or sexually transmitted diseases in women. Zhou *et al.* first characterized the vaginal microbiota by 16S clone library sequencing (Zhou et al, 2004). As found in culture-dependent studies, the authors showed that *Lactobacilli* and *Atopobium* were generally the predominant organisms; they also reported the first identification of a *Megasphaera* species in the vagina. By employing a full 16S pyrosequencing strategy and developing a classification method of short 16S pyrosequencing reads, Sundquist *et al.* studied the human vagina during pregnancy and corroborated previous results. The authors identified *Lactobacillus* as the dominant genus and detected a significant presence of other genera including *Psychrobacter*, *Magnetobacterium*, *Prevotella*, *Bifidobacterium*, and *Veillonella* (Sundquist et al, 2007). However, *Lactobacillus* can be missing in healthy vaginal microbiota and replaced by other predominant genera such as *Gardnerella*, *Pseudomonas*, or *Streptococcus* (Hyman et al, 2005). Although inter-individual variability of the vaginal microbiota was demonstrated, the function of the communities was conserved and was shown to be involved in the production of lactic acid (Zhou et al, 2004). The maintenance of the vaginal acidity preserves an unfriendly environment for the growth of many pathogenic organisms (Zhou et al, 2004). Important shifts in relative abundances and types of bacteria, especially the decrease in lactic acid bacteria, in the healthy vagina are associated with bacterial vaginosis infections (Spiegel, 1991). Compared to the healthy vaginal

microbiota, bacterial vaginosis-associated microbiota showed greater specie richness, different bacterial community structures, and a strong association with members of the *Bacteroidetes* and *Actinobacteria* phyla (Oakley et al, 2008). To enhance the understanding of the important variations in the incidence of bacterial vaginosis among racial or ethnic groups, the vaginal microbiota of Caucasian and black women were explored. Striking differences were demonstrated in community abundance and also in composition; for instance, the predominance of *Lactobacillus* in black women is lower than in Caucasian women (Zhou et al, 2007).

The Skin Microbiota

The skin probably offers one of the largest human-associated habitats and has a bacterial density of around 10⁷ cells per square centimeter (Fredricks, 2001). The commensal bacterial communities and pathogenic microorganisms harbored by the skin's surface suggest an association with healthy, infectious or noninfectious (psoriasis, eczema) (Grice et al, 2008) skin states. Until recently, because skin infections generally involve rare pathogenic isolates, the resident skin bacteria remained poorly described and limited to culture-dependent studies that under-represented the extent of bacterial diversity.

A recent 16S survey of the resident skin microbiota of the inner elbow region, an area subjected to atopic dermatitis, from five healthy humans generated 5,373 nearly complete 16S rRNA sequences. These sequences were assigned into 113 phylotypes belonging to the *Proteobacteria* (49%), *Actinobacteria* (28%), *Firmicutes* (12%), *Bacteroidetes* (9.7%), *Cyanobacteria* (<1%), and *Acidobacteria* (<1%) phyla (Grice et al, 2008). Most of the 16S sequences (90%) belong to the *Proteobacteria* phylum and, more accurately, to the *Pseudomonas* and *Janthinobacterium* genera. Finally, the authors' results indicated a low level of deep evolutionary lineage diversity and a similar diversity profile for all of the subjects, suggesting a common core skin microbiota among healthy subjects (Grice et al, 2008). Interestingly, the few cultivated commensal skin bacteria, including *S. epidermidis* and *P. acnes*, accounted for only 5% of the microbiota captured in the Grice *et al.* study (Grice et al, 2008).

In contrast, a study examining the diversity of skin microbiota from the superficial forearms in six healthy subjects indicated a small core set of phylotypes (2.2%) and a high degree of inter-individual variability in the microbiota

(Gao et al, 2007). The distribution of the 182 identified phylotypes at the phylum level was 29% *Proteobacteria*, 35% *Actinobacteria*, 24% *Firmicutes*, 8% *Bacteroidetes*, 1.6% *Deinococcus-Thermus*, 0.5% *Termomicrobia*, and 0.5% *Cyanobacteria* (Gao et al, 2007). However, only the three most abundant phyla were observed in all subjects. Although many phylotypes overlap between the inner elbow and forearm skin microbiota, the predominant phylotype belonging to the *Proteobacteria* phyla of the inner elbow microbiota is missing in the forearm microbiota. In addition, the forearm skin microbiota possesses more members of the *Actinobacteria* and *Firmicutes* phyla (Grice et al, 2008). It is difficult, however, to compare studies that employed different methods and skin sample locations.

The characterization of skin microbial communities and their interactions is still in its infancy, since the impacts of sex, age, clothing, and others factors have not been clearly determined. However, a recent study on the reduction of disease transmission by hand washing used a barcoded pyrosequencing approach to characterize the hand surface microbiota of 51 healthy men and women to determine how specific factors could affect the community structure (Fierer et al, 2008). Although the authors detected a core set of bacterial taxa on the hand surface, the results mainly revealed a high intra- and inter-individual variation in community structure when sex, hand washing, or handedness factors were considered. In addition, though the diversity observed in hand surfaces is high (sequences from >25 phyla were identified), 94% of the sequences belong only to three of these phyla (*Actinobacteria*, *Firmicutes* and *Proteobacteria*) (Fierer et al, 2008). Finally, independently of the skin site sampled (Fierer et al, 2008, Gao et al, 2007, Grice et al, 2008), all of the studies shared the same predominant phyla: *Proteobacteria*, *Actinobacteria*, and *Firmicutes*.

Limits of 16S Analyses

The 16S is an efficient phylogenetic marker for bacteria identification and microbial community analyses. However, the multiple pitfalls of PCR-based analyses, including sample collection, cell lysis, PCR amplification, and cloning, can affect the estimation of the community composition in mixed microbial samples (Farrelly et al, 1995, von et al, 1997).

Although one or two 16S gene copies are commonly exhibited by a single genome, multiple and heterogeneous 16S genes in a single microbial genome are not rare and can

lead to an overestimation of the abundance and bacterial diversity using culture-independent approaches (Acinas et al, 2004). Multiple copies of rRNA operons (*rrn*) per genome are generally found in rapidly growing microorganisms, especially in soil bacteria (Klappenbach et al, 2000). As a result, the number of copies of the 16S gene in a microbial genome can reach 10 or 12 copies in *Bacillus subtilis* (Stewart et al, 1982) or *Bacillus cereus* (Johansen et al, 1996), respectively, and up to 15 copies in *Clostridium paradoxum* (Rainey et al, 1996). In addition to the heterogeneity of the 16S gene copy number per genome, a bacterial species can display important nucleotide sequence variability among its 16S genes (Acinas et al, 2004, Rainey et al, 1996, Turova et al, 2001). Furthermore, it is well known that 16S gene sequencing lacks taxonomic resolution at the species level for some closely related species (Janda and Abbott, 2007), subspecies, or recently diverged species (Fox et al, 1992). In this way, *Escherichia coli* and *E. fergusonii* species, as well as the subspecies *Bartonella vinsonii subsp. arupensis* and *B. vinsonii subsp. vinsonii*, are indistinguishable, when using 16S sequence similarity comparisons (Adekambi et al, 2003).

Likewise, PCR amplification bias in a mixed microbial sample causes the taxon amplicon abundance to differ from the real proportions present in the community. Notably, PCR amplification bias can be induced by the choice of primer set, the number of replication cycles, or the enzyme system used (Qiu et al, 2001, Suzuki and Giovannoni, 1996). An obvious example is that the sensitivity of universal primers is limited to the currently known 16S sequences and does not reach 100% coverage. What is the primer sensitivity for unknown microbial sequences, and how can the same hybridization efficiency be guaranteed for all targets in the sample? However, metagenomic studies, which can theoretically access up to 100% of the microbial diversity in a sample, have yielded powerful alternatives to bypass primer and cloning bias using 454 pyrosequencing.

Another aspect limiting the capture of the true microbial diversity in a mixed sample by 16S surveys is inherent to the cloning step. The efficiency of ligation to the plasmid, transformation, and amplification in the host can all have an effect. For instance, it has been suggested that many unclonable genes in the *E. coli* host are present in a single copy per genome and hence are under-represented in clone libraries due to inactive promoters or toxic effects induced by gene transcription/translation into the host (Kunin et al,

2008, Sorek et al, 2007). Finally, horizontal gene transfer and recombination events have also been reported in ribosomal genes (Miller et al, 2005), which distort phylogenetic signals and thereby affect the phylogenetic classification.

Conclusions and Perspectives

Culture-independent methods based on the 16S rRNA gene yield a useful framework for exploring microbial diversity, by establishing the taxonomic composition and/or structure present in environmental samples using both α and β -diversity measures, phylogenetic tree construction, and sequence similarity comparison.

Unlikely to culture methods, these recent high-throughput methods allow accessing to the true microbial diversity. In a point of view of clinical research, new or uncultured etiologic agents from poly-microbial samples (pulmonary infections, brain abscess) of disease states can be identified and will lead to elaborate more appropriate antibiotherapies rather than the use of broad range antibiotics. In addition, compared to lean controls, the reduction of the *Bacteroidetes* members and the increase of the methanogen *Methanobrevibacter smithii* in obese patients were revealed by high-throughput sequencing methods. These results suggested that modulate the relative abundance of some microbial groups of the gut microbiota could be beneficial for obese treatment. The huge amount of sequences provided by these new sequencing methods hugely increase the number of 16S sequences in databases, and thus improve the ability of 16S sequence identification using sequence similarity search tools. In a near future, the accuracy of classification methods of short 16S sequences will be improved by the increase of read length (450pb) produced by the new 454 FLX Titanium apparatus. In addition, the increase in sequence production capabilities of the 454 FLX Titanium associated with the barcoding strategy will allow examining much more different samples in a single pyrosequencing run.

However, 16S high-throughput methods can not characterize the functional component (defined as the microbiome) of an environmental sample. Such limitations arise by targeting a sole gene marker. This limitation can be overcome by a metagenomic approach, which focuses on the full gene content (gene-centric analysis) of a sample. Therefore, in addition to providing species richness and evenness information, the relatively unbiased metagenomic approach can also identify the metabolic capabilities of a microbial com-

munity and disclose specific adaptive gene sets that are potentially beneficial for survival in a given habitat.

References

1. Acinas SG, Marcelino LA, Klepac CV, Polz MF (2004) Divergence and redundancy of 16S rRNA sequences in genomes with multiple *rrn* operons. *J Bacteriol* 186: 2629-2635.
2. Adekambi T, Colson P, Drancourt M (2003) RpoB-based identification of nonpigmented and late-pigmenting rapidly growing mycobacteria. *J Clin Microbiol* 41: 5699-5708.
3. Andersson AF, Lindberg M, Jakobsson H, Backhed F, Nyren P, et al. (2008) Comparative analysis of human gut microbiota by barcoded pyrosequencing. *PLoS ONE* 3: e2836.
4. Armougom F, Raoult D (2008) Use of pyrosequencing and DNA barcodes to monitor variations in Firmicutes and Bacteroidetes communities in the gut microbiota of obese humans. *BMC Genomics* 9: 576.
5. Armougom F, Moretti S, Keduas V, Notredame C (2006) The iRMSD: a local measure of sequence alignment accuracy using structural information. *Bioinformatics* 22: e35-e39.
6. Bae JW, Park YH (2006) Homogeneous versus heterogeneous probes for microbial ecological microarrays. *Trends Biotechnol* 24: 318-323.
7. Biddle JF, Fitz GS, Schuster SC, Brenchley JE, House CH (2008) Metagenomic signatures of the Peru Margin subseafloor biosphere show a genetically distinct environment. *Proc Natl Acad Sci USA* 105: 10583-10588.
8. Binladen J, Gilbert MT, Bollback JP, Panitz F, Bendixen C, et al. (2007) The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS ONE* 2: e197.
9. Board On Life Sciences (2007) *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet*. NATIONAL RESEARCH COUNCIL OF THE NATIONAL ACADEMIES ed.
10. Brodie EL, DeSantis TZ, Joyner DC, Baek SM, Larsen

- JT, et al. (2006) Application of a high-density oligonucleotide microarray approach to study bacterial population dynamics during uranium reduction and reoxidation. *Appl. Environ. Microbiol* 72: 6288-6298.
11. Brodie EL, DeSantis TZ, Parker JP, Zubieta IX, Piceno YM, et al. (2007) Urban aerosols harbor diverse and dynamic bacterial populations. *Proc Natl Acad Sci USA* 104: 299-304.
 12. Clarke SC (2005) Pyrosequencing: nucleotide sequencing technology with bacterial genotyping applications. *Expert Rev Mol Diagn* 5: 947-953.
 13. Clarridge JE III (2004) Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clin Microbiol Rev* 17: 840-62.
 14. Cole JR, Chai B, Farris RJ, Wang Q, Kulam SMAS, et al. (2007) The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. *Nucleic Acids Res* 35: D169-D172.
 15. DeRisi JL, Iyer VR, Brown PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278: 680-686.
 16. DeSantis TZ, Brodie EL, Moberg JP, Zubieta IX, Piceno YM, et al. (2007) High-density universal 16S rRNA microarray analysis reveals broader diversity than typical clone library when sampling the environment. *Microb Ecol* 53: 371-383.
 17. DeSantis TZ, Dubosarskiy I, Murray SR, Andersen GL (2003) Comprehensive aligned sequence construction for automated design of effective probes (CASCADE-P) using 16S rDNA. *Bioinformatics* 19: 1461-1468.
 18. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, et al. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* 72: 5069-5072.
 19. Desnues C, Rodriguez BB, Rayhawk S, Kelley S, Tran T, et al. (2008) Biodiversity and biogeography of phages in modern stromatolites and thrombolites. *Nature* 452: 340-343.
 20. Dethlefsen L, Huse S, Sogin ML, Relman DA (2008) The Pervasive Effects of an Antibiotic on the Human Gut Microbiota, as Revealed by Deep 16S rRNA Sequencing. *PLoS Biol* 6: e280.
 21. Dibaise JK, Zhang H, Crowell MD, Krajmalnik BR, Decker GA, et al. (2008) Gut microbiota and its possible relationship with obesity. *Mayo Clin Proc* 83: 460-469.
 22. Dowd SE, Callaway TR, Wolcott RD, Sun Y, McKeehan T, et al. (2008a) Evaluation of the bacterial diversity in the feces of cattle using 16S rDNA bacterial tag-encoded FLX amplicon pyrosequencing (bTEFAP). *BMC Microbiol* 8: 125.
 23. Dowd SE, Sun Y, Wolcott RD, Domingo A, Carroll JA (2008b) Bacterial tag-encoded FLX amplicon pyrosequencing (bTEFAP) for microbiome studies: bacterial diversity in the ileum of newly weaned Salmonella-infected pigs. *Foodborne Pathog Dis* 5: 459-472.
 24. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, et al. (2005) Diversity of the human intestinal microbial flora. *Science* 308: 1635-1638.
 25. Edwards KJ, Bond PL, Gihring TM, Banfield JF (2000) An archaeal iron-oxidizing extreme acidophile important in acid mine drainage. *Science* 287: 1796-1799.
 26. Edwards RA, Rodriguez BB, Wegley L, Haynes M, Breitbart M, et al. (2006) Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* 7: 57.
 27. Eriksson N, Pachter L, Mitsuya Y, Rhee SY, Wang C, et al. (2008) Viral population estimation using pyrosequencing. *PLoS Comput Biol* 4: e1000074.
 28. Farrelly V, Rainey FA, Stackebrandt E (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: 2798-2801.
 29. Fierer N, Hamady M, Lauber CL, Knight R (2008) The influence of sex, handedness, and washing on the diversity of hand surface bacteria. *Proc Natl Acad Sci USA* 105: 17994-17999.
 30. Fox GE, Wisotzkey JD, Jurtshuk P Jr (1992) How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. *Int J Syst Bacteriol* 42: 166-170.

31. Frank DN, St Amand AL, Feldman RA, Boedeker EC, Harpaz N, et al. (2007) Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc Natl Acad Sci USA* 104: 13780-13785.
32. Fredricks DN (2001) Microbial ecology of human skin in health and disease. *J Investig Dermatol Symp Proc* 6: 167-169.
33. Gao Z, Tseng CH, Pei Z, Blaser MJ (2007) Molecular analysis of human forearm superficial skin bacterial biota. *Proc Natl Acad Sci USA* 104: 2927-2932.
34. Gentry TJ, Wickham GS, Schadt CW, He Z, Zhou J (2006) Microarray applications in microbial ecology research. *Microb Ecol* 52: 159-175.
35. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, et al. (2006) Metagenomic analysis of the human distal gut microbiome. *Science* 312: 1355-1359.
36. Giovannoni SJ, Stingl U (2005) Molecular diversity and ecology of microbial plankton. *Nature* 437: 343-348.
37. Grice EA, Kong HH, Renaud G, Young AC, Bouffard GG, et al. (2008) A diversity profile of the human skin microbiota. *Genome Res* 18: 1043-1050.
38. Hall N (2007) Advanced sequencing technologies and their wider impact in microbiology. *J Exp Biol* 210: 1518-1525.
39. Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat Methods* 5: 235-237.
40. Harmsen D, Rothganger J, Frosch M, Albert J (2002) RIDOM: Ribosomal Differentiation of Medical Microorganisms Database. *Nucleic Acids Res* 30: 416-417.
41. Hjalmarsson S, Alderborn A, Fock C, Muldin I, Kling H, et al. (2004) Rapid combined characterization of microorganism and host genotypes using a single technology. *Helicobacter* 9: 138-145.
42. Holt RA, Jones SJ (2008) The new paradigm of flow cell sequencing. *Genome Res* 18: 839-846.
43. Huber JA, Mark WDB, Morrison HG, Huse SM, Neal PR, et al. (2007) Microbial population structures in the deep marine biosphere. *Science* 318: 97-100.
44. Hugenholtz P, Tyson GW (2008) Microbiology: metagenomics. *Nature* 455: 481-483.
45. Hunkapiller T, Kaiser RJ, Koop BF, Hood L (1991) Large-scale and automated DNA sequence determination. *Science* 254: 59-67.
46. Huse SM, Dethlefsen L, Huber JA, Welch DM, Relman DA, et al. (2008) Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genet* 4: e1000255.
47. Hyman RW, Fukushima M, Diamond L, Kumm J, Giudice LC, et al. (2005) Microbes on the human vaginal epithelium. *Proc Natl Acad Sci USA* 102: 7952-7957.
48. Janda JM, Abbott SL (2007) 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *J Clin Microbiol* 45: 2761-2764.
49. Jenkinson HF, Lamont RJ (2005) Oral microbial communities in sickness and in health. *Trends Microbiol* 13: 589-595.
50. Johansen T, Carlson CR, Kolsto AB (1996) Variable numbers of rRNA gene operons in *Bacillus cereus* strains. *FEMS Microbiol Lett* 136: 325-328.
51. Jonasson J, Olofsson M, Monstein HJ (2002) Classification, identification and subtyping of bacteria based on pyrosequencing and signature matching of 16S rDNA fragments. *APMIS* 110: 263-272.
52. Keijser BJ, Zaura E, Huse SM, van d V, Schuren FH, et al. (2008) Pyrosequencing analysis of the oral microflora of healthy adults. *J Dent Res* 87: 1016-1020.
53. Kent WJ (2002) BLAT-the BLAST-like alignment tool. *Genome Res* 12: 656-664.
54. Kim BS, Kim BK, Lee JH, Kim M, Lim YW, et al. (2008) Rapid phylogenetic dissection of prokaryotic community structure in tidal flat using pyrosequencing. *J Microbiol* 46: 357-363.
55. Klappenbach JA, Dunbar JM, Schmidt TM (2000) rRNA operon copy number reflects ecological strategies of

- bacteria. *Appl Environ Microbiol* 66: 1328-1333.
56. Koski LB, Golding GB (2001) The closest BLAST hit is often not the nearest neighbor. *J Mol Evol* 52: 540-542.
57. Krause L, Diaz NN, Edwards RA, Gartemann KH, Kromeke H, et al. (2008) Taxonomic composition and gene content of a methane-producing microbial community isolated from a biogas reactor. *J Biotechnol* 136: 91-101.
58. Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P (2008) A bioinformatician's guide to metagenomics. *Microbiol. Mol Biol Rev* 72: 557-78.
59. La SB, Desnues C, Pagnier I, Robert C, Barrassi L, et al. (2008) The virophage as a unique parasite of the giant mimivirus. *Nature* 455: 100-104.
60. Lane DJ, Stahl DA, Olsen GJ, Heller DJ, Pace NR (1985) Phylogenetic analysis of the genera *Thiobacillus* and *Thiomicrospira* by 5S rRNA sequences. *J Bacteriol* 163: 75-81.
61. Ley RE, Backhed F, Turnbaugh P, Lozupone CA, Knight RD, et al. (2005) Obesity alters gut microbial ecology. *Proc Natl Acad Sci USA* 102: 11070-11075.
62. Ley RE, Turnbaugh PJ, Klein S, Gordon JI (2006) Microbial ecology: human gut microbes associated with obesity. *Nature* 444: 1022-1023.
63. Liu Z, DeSantis TZ, Andersen GL, Knight R (2008) Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res* 36: e120.
64. Liu Z, Lozupone C, Hamady M, Bushman FD, Knight R (2007) Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Res* 35: e120.
65. Lozupone C, Hamady M, Knight R (2006) UniFrac—an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics* 7: 371.
66. Lozupone C, Knight R (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* 71: 8228-8235.
67. Lozupone CA, Hamady M, Kelley ST, Knight R (2007) Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Appl Environ Microbiol* 73: 1576-1585.
68. Lozupone CA, Knight R (2008) Species divergence and the measurement of microbial diversity. *FEMS Microbiol Rev* 32: 557-578.
69. Ludwig W, Strunk O, Westram R, Richter L, Meier H, et al. (2004) ARB: a software environment for sequence data. *Nucleic Acids Res* 32: 1363-1371.
70. Magurran AE (2005) Biological diversity. *Curr Biol* 15: R116-R118.
71. Mardis ER (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet* 24: 133-141.
72. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437: 376-380.
73. Martin AP (2002) Phylogenetic approaches for describing and comparing the diversity of microbial communities. *Appl Environ Microbiol* 68: 3673-3682.
74. McKenna P, Hoffmann C, Minkah N, Aye PP, Lackner A, et al. (2008) The macaque gut microbiome in health, lentiviral infection, and chronic enterocolitis. *PLoS Pathog* 4: e20.
75. Melamede RJ, Wallace SS (1985) A possible secondary role for thymine-containing DNA precursors. *Basic Life Sci* 31: 67-102.
76. Miller SR, Augustine S, Olson TL, Blankenship RE, Selker J, et al. (2005) Discovery of a free-living chlorophyll d-producing cyanobacterium with a hybrid proteobacterial/cyanobacterial small-subunit rRNA gene. *Proc Natl Acad Sci USA* 102: 850-855.
77. Oakley BB, Fiedler TL, Marrazzo JM, Fredricks DN (2008) Diversity of human vaginal bacterial communities and associations with clinically defined bacterial vaginosis. *Appl Environ Microbiol* 74: 4898-4909.
78. Olsen GJ, Lane DJ, Giovannoni SJ, Pace NR, Stahl DA (1986) Microbial ecology and evolution: a ribosomal RNA approach. *Annu Rev Microbiol* 40: 337-365.

79. Pace NR (1997) A molecular view of microbial diversity and the biosphere. *Science* 276: 734-740.
80. Palmer C, Bik EM, Digiulio DB, Relman DA, Brown PO (2007) Development of the Human Infant Intestinal Microbiota. *PLoS Biol* 5: e177.
81. Pettersson E, Lundeberg J, Ahmadian A (2008) Generations of sequencing technologies. *Genomics*.
82. Pontes DS, Lima BCI, Chartone SE, maral Nascimento AM (2007) Molecular approaches: advantages and artifacts in assessing bacterial diversity. *J Ind Microbiol Biotechnol* 34: 463-473.
83. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, et al. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* 35: 7188-7196.
84. Qiu X, Wu L, Huang H, McDonel PE, Palumbo AV, Tiedje JM, et al. (2001) Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16S rRNA gene-based cloning. *Appl. Environ. Microbiol* 67: 880-887.
85. Rainey FA, Ward RNL, Janssen PH, Hippe H, Stackebrandt E (1996) *Clostridium paradoxum* DSM 7308T contains multiple 16S rRNA genes with heterogeneous intervening sequences. *Microbiology* 142: 2087-2095.
86. Riesenfeld CS, Schloss PD, Handelsman J (2004) Metagenomics: genomic analysis of microbial communities. *Annu Rev Genet* 38: 525-552.
87. Roesch LF, Fulthorpe RR, Riva A, Casella G, Hadwin AK, et al. (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J* 1: 283-290.
88. Ronaghi M (2001) Pyrosequencing sheds light on DNA sequencing. *Genome Res* 11: 3-11.
89. Ronaghi M, Elahi E (2002) Pyrosequencing for microbial typing. *J Chromatogr B Analyt Technol Biomed Life Sci* 782: 67-72.
90. Ronaghi M, Karamohamed S, Pettersson B, Uhlen M, Nyren P (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem* 242: 84-89.
91. Ronaghi M, Uhlen M, Nyren P (1998) A sequencing method based on real-time pyrophosphate. *Science* 281: 363-365.
92. Rossello MR, Amann R (2001) The species concept for prokaryotes. *FEMS Microbiol Rev* 25: 39-67.
93. Rothberg JM, Leamon JH (2008) The development and impact of 454 sequencing. *Nat Biotechnol* 26: 1117-1124.
94. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, et al. (1977) Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* 265: 687-695.
95. Schena M, Shalon D, Heller R, Chai A, Brown PO, et al. (1996) Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc Natl Acad Sci USA* 93: 10614-10619.
96. Schloss PD, Handelsman J (2005) Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl Environ Microbiol* 71: 1501-1506.
97. Schloss PD, Handelsman J (2006a) Introducing SONS, a tool for operational taxonomic unit-based comparisons of microbial community memberships and structures. *Appl Environ Microbiol* 72: 6773-6779.
98. Schloss PD, Handelsman J (2006b) Introducing TreeClimber, a test to compare microbial community structures. *Appl Environ Microbiol* 72: 2379-2384.
99. Schmeisser C, Steele H, Streit WR (2007) Metagenomics, biotechnology with non-culturable microbes. *Appl Microbiol Biotechnol* 75: 955-962.
100. Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nat Biotechnol* 26: 1135-1145.
101. Simon GL, Gorbach SL (1986) The human intestinal microflora. *Dig Dis Sci* 31: 147S-162S.
102. Sogin ML, Morrison HG, Huber JA, Mark WD, Huse SM, et al. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc Natl Acad Sci USA* 103: 12115-12120.
103. Sorek R, Zhu Y, Creevey CJ, Francino MP, Bork P, et al. (2007) Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* .318: 1449-1452.

104. Spear GT, Sikaroodi M, Zariffard MR, Landay AL, French AL, et al. (2008) Comparison of the diversity of the vaginal microbiota in HIV-infected and HIV-uninfected women with or without bacterial vaginosis. *J Infect Dis* 198: 1131-1140.
105. Spiegel CA (1991) Bacterial vaginosis. *Clin Microbiol Rev* 4: 485-502.
106. Spiegelman D, Whissell G, Greer CW (2005) A survey of the methods for the characterization of microbial consortia and communities. *Can J Microbiol* 51: 355-386.
107. Stahl DA, Lane DJ, Olsen GJ, Pace NR (1985) Characterization of a Yellowstone hot spring microbial community by 5S rRNA sequences. *Appl Environ Microbiol* 49: 1379-1384.
108. Stewart GC, Wilson FE, Bott KF (1982) Detailed physical mapping of the ribosomal RNA genes of *Bacillus subtilis*. *Gene* 19: 153-162.
109. Sundquist A, Bigdeli S, Jalili R, Druzin ML, Waller S, et al. (2007) Bacterial flora-typing with targeted, chip-based Pyrosequencing. *BMC Microbiol* 7: 108.
110. Suzuki MT, Giovannoni SJ (1996) Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl Environ Microbiol* 62: 625-630.
111. Tarnberg M, Jakobsson T, Jonasson J, Forsum U (2002) Identification of randomly selected colonies of lactobacilli from normal vaginal fluid by pyrosequencing of the 16S rDNA variable V1 and V3 regions. *APMIS* 110: 802-810.
112. Tringe SG, Hugenholtz P (2008) A renaissance for the pioneering 16S rRNA gene. *Curr Opin Microbiol* 11: 442-446.
113. Tringe SG, von MC, Kobayashi A, Salamov AA, Chen K, et al. (2005) Comparative metagenomics of microbial communities. *Science* 308: 554-557.
114. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A et al. (2008) A core gut microbiome in obese and lean twins. *Nature*.
115. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis E.R, Gordon JI (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444: 1027-1031.
116. Turova TP, Kuznetsov BB, Novikova EV, Poltarau AB, Nazina TN (2001) [Heterogeneity of nucleotide sequences of 16S ribosomal RNA genes from the *Desulfotomaculum kuznetsovii* type strain]. *Mikrobiologiya* 70: 788-795.
117. Vandamme P, Pot B, Gillis M, de VP, Kersters K, Swings J (1996) Polyphasic taxonomy, a consensus approach to bacterial systematics. *Microbiol Rev* 60: 407-438.
118. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304: 66-74.
119. von WF, Gobel UB, Stackebrandt E (1997) Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS Microbiol Rev* 21: 213-229.
120. Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 73: 5261-5267.
121. Wegley L, Edwards R, Rodriguez BB, Liu H, Rohwer F (2007) Metagenomic analysis of the microbial community associated with the coral *Porites astreoides*. *Environ Microbiol* 9: 2707-2719.
122. Whitman WB, Coleman DC, Wiebe WJ (1998) Prokaryotes: the unseen majority. *Proc Natl Acad Sci USA* 95: 6578-6583.
123. Wilson KH, Wilson WJ, Radosevich JL, DeSantis TZ, Viswanathan VS, et al. (2002) High-density microarray of small-subunit ribosomal DNA probes. *Appl Environ Microbiol* 68: 2535-2541.
124. Woese CR (1987) Bacterial evolution. *Microbiol Rev* 51: 221-271.
125. Woo PC, Ng KH, Lau SK, Yip KT, Fung AM, et al. (2003) Usefulness of the MicroSeq 500 16S ribosomal DNA-based bacterial identification system for identification of clinically significant bacterial isolates with

- ambiguous biochemical profiles. *J Clin Microbiol* 41: 1996-2001.
126. Woodmansey EJ (2007) Intestinal bacteria and ageing. *J Appl Microbiol* 102: 1178-1186.
127. Yergeau E, Schoondermark SSA, Brodie EL, Dejean S, DeSantis TZ, et al. (2008) Environmental microarray analyses of Antarctic soil microbial communities. *ISME J*.
128. Zhou X, Bent SJ, Schneider MG, Davis CC, Islam MR, et al. (2004) Characterization of vaginal microbial communities in adult healthy women using cultivation-independent methods. *Microbiology* 150: 2565-2573.
129. Zhou X, Brown CJ, Abdo Z, Davis CC, Hansmann MA, et al. (2007) Differences in the composition of vaginal microbial communities found in healthy Caucasian and black women. *ISME J* 1: 121-133.
130. Zoetendal EG, Rajilic SM, de Vos WM (2008) High-throughput diversity and functionality analysis of the gastrointestinal tract microbiota. *Gut* 57: 1605-1615.