Introduction to microbial diversity in soil

Soils are among the most complex habitats on Earth. Biodiversity in soil environments is immense, with the vast majority of this diversity with microorganisms (Daniel, 2004). One gram of soil may contain up to 10 billion microorganisms of possibly thousands of different species (Torsvik & Ovreas, 2002). With the introduction of molecular methods in microbial ecology, the genetic complexity of soil microbial communities is now usually estimated on the basis of rRNA gene biomarkers.

Unfortunately, less is known about the functioning of microbial communities that dominate soil. To a large extent this knowledge gap stems from the difficulties associated with culturing the majority of soil microorganisms. It has been estimated that the fraction of bacteria in soil that can be accessed through standard cultivation techniques is generally low and might be less than 1% (Torsvik & Ovreas, 2002; Curtis & Sloan, 2004). We can only speculate about the importance of the uncultivated majority of microbial communities for ecosystem functioning. Thus, a large part of our understanding of terrestrial ecosystems is still missing. Also, over half of the natural products of economical value (including antibiotics and pharmaceutics) is derived from soil-borne microbes. Thus, assuming that only 1% is typically considered as culturable, an enormous potential of useful products and activities remains unexplored and locked within yet-to-be-cultured soil microorganisms. This expectation has led to renewed interest in examining soil microorganisms for agricultural, pharmaceutical and industrial purposes. To assess and exploit the diversity of
soil microbial communities novel cultivation methods, as well as molecular cultivation-independent approaches have been designed.

Regarding novel methodologies for culturing soil microbes, relatively simple technologies that mimick the natural niche conditions have been suggested (Janssen et al., 2002). This usually involves diluted growth media that supports very low levels of cell growth and consequently involves prolonged incubation times (Keller & Zengler, 2004), or the use of solutions extracted from the environment in which the inoculum was previously present (Kaeberlein et al., 2002). Even with such improvements, there are, however, still serious problems associated with the use of these culturing methods, including biases and failure to detect essential microbial interactions and symbioses (Stevenson et al., 2004).

These limitations have stimulated the development of several approaches designed to overcome cultivation-related problems, thereby providing access to information of soil microbial identity, and more recently function, without the requisite need for laboratory cultivation. Based on phylogenetic markers genes, such as 16S rRNA genes, these developments have provided clear evidence of the existence of numerous bacterial lineages that appear to be quite common in the environment, but are clearly underrepresented in bacterial culture collections (Figure 1). Two of the most common bacterial phyla in soils that fall into this category are the Acidobacteria and Verrucomicrobiales.
Figure 1. Phylogenetic tree of the Bacteria showing established phyla (italicized Latinized names) and candidate phyla described previously using the November 2003 ARB database (http://arb-home.de) with 16,964 sequences that are over 1,000 bp. The vertex angle of each wedge indicates the relative abundance of sequences in each phylum, and the length of each side of the wedge indicates the range of branching depth found in that phylum. The density of shading of each wedge corresponds to the proportion of sequences in that phylum obtained from cultured representatives. None of the candidate phyla have cultured representatives (Schloss and Handelsman (2004) Microbiol Mol Biol Rev 68(4): 686–691).
Although the knowledge of these two groups and of their roles in global biogeochemical cycles is very limited, their abundance and diversity suggests their potentially great ecological importance (Hugenholtz et al., 1998; Stevenson et al., 2004).

The main objective of the research described in this thesis is to provide new information on the functioning of *Acidobacteria* and *Verrucomicrobiales* in soil using metagenomic approaches.

**Acidobacteria**

The designation, Acidobacteria, was coined after the first described species of this phylum, *Acidobacterium capsulatum*, which was isolated in the late 1990s from an acidic mine drainage (Ludwig et al., 1997). The phylum name is somewhat misleading since Acidobacteria are not restricted to only acidic environments. Members of this phylum have also been found in numerous other habitats including ocean water, sediments, hot springs, peat bogs, and soil (Barns et al., 1999; Lopez-Garcia et al., 2003; Polymenakou et al., 2005; Chan et al., 2006; Dedysh et al., 2006; Penn et al., 2006; Sogin et al., 2006; Gomez-Alvarez et al., 2007). Members of the Acidobacteria phylum seem to be particularly dominant in the soil, representing up to 52% of 16S rRNA gene sequences from clone libraries (Sait et al., 2002) and typically represent 20 to 30% of the 16S rRNA sequences amplified by PCR from soil DNA (Handelsman, 2004; Janssen, 2006). A total of eight acidobacterial species have been described, and four of these, *A. capsulatum*, *T. roseus*, *Edaphobacter modestus*, and *Edaphobacter aggregans*, were isolated from soil habitats.

By means of phylogenetic analyses, Acidobacteria were originally described as having 4 to 5 subdivisions (Kuske et al., 1997; Ludwig et al., 1997). With expanding sequencing efforts, the number of subdivisions grew to eight (Sait et al., 2002; Handelsman, 2004) then to eleven (Zimmermann et al., 2005), and a new classification scheme, covering the latest expanded diversity, suggests the presence of even 26 subdivisions within the Acidobacteria (Barns et al., 2007).

However, our knowledge to date has been mostly limited to 16S rRNA gene-based studies due to the problems with cultivation, storage and propagation of members of this phylum. With the introduction of new cultivation methods (Schoenborn et al., 2004; Stevenson et al., 2004; Davis et al., 2005; Sangwan et al., 2005; Stott et al., 2008), the number of acidobacterial isolates is growing but cultures still offer a skewed view of acidobacterial diversity. For instance, subdivision 6 is the most dominant one in most soils, yet not a single member of this subdivision has been entered into a culture collection. Recently, through the developments of molecular techniques and automatization of high throughput processes, culture-independent techniques are employed to fill the gap in knowledge with respect to acidobacterial distribution, diversity and activities. More detailed information on the phylogeny and potential ecology of *Acidobacteria* in soils is given in chapter 2 of this thesis.
**Verrucomicrobia**

Similar to *Acidobacteria*, members from the phylum *Verrucomicrobia* are widely distributed and abundant in various environments (Hugenholtz *et al*., 1998). They have been recovered from a range of soil, water, and sediment samples (Janssen *et al*., 2002). The lifestyle of some of *Verrucomicrobia* is strongly related with eukaryotes (Sakai *et al*., 2003; Wang *et al*., 2005), and phylum members have been described that live as obligate endosymbionts of nematodes from the genus *Xiphinema* (Vandekerckhove *et al*., 2000).

It has been estimated that in the soil environment the *Verrucomicrobia* represent 1% to 11% of the total bacteria, accounting for 5% of all characterized 16S rRNA genes (Sangwan *et al*., 2004), which suggests that they may be important for the functioning of plant/soil systems.

This phylum is characterized by seven subdivisions, but to date only four of them contain validly described representatives. Similar to the *Acidobacteria*, the use of new dilution culture techniques has led to the isolation of several new isolates of the *Verrucomicrobiales* group, which have further been identified on the basis of 16S rRNA gene sequences (Janssen *et al*., 1997).

Recently, the sequencing of the genomes of seven *Verrucomicrobia* members (bacterium Ellin514, *Verrucomicrobiaceae bacterium* DG1235, *Akkermansia muciniphila* ATCC BAA-835, *Opitutaceae bacterium* TAV2, *Opitutus terrae* PB90-1, *Methylacidiphilum infernorum* V4, *Chthoniobacter flavus* Ellin428, *Verrucomicrobiium spinosum* DSM 4136) was started and the drafts of the first genome sequences have recently become available through the National Center for Biotechnology Information (NCBI).

**The metagenome approach**

As stated above, the majority of microorganisms cannot be, or are recalcitrant to being cultured. An alternative strategy to gain insight into the functions of the not-yet-cultured majority is offered by metagenomics, where the collective genomes of an entire community, the metagenome, are the target of study. Metagenomic approaches have been developed to study the community structure and functional content on the basis of its genetic material obtained directly from the environment, without need to culture individual organisms (Handelsman *et al*., 1998). A wide range of approaches has been employed to gain access to environmental metagenomes. The choice of strategy depends on a number of factors, including the complexity of the community, the amount of sample material available, the nature of the substrate, the density of microorganisms in a habitat, and of course the goal and scope of the study (Kowalchuk *et al*., 2007).

Metagenomic strategies can be divided into two groups according to the research focus. The first one involves constructions of small- or large-insert libraries, which are further screened for specific activities of interest (Henne *et al*., 1999; Rondon *et al*., 2000;
Daniel, 2004; Yun et al., 2004; Kim et al., 2006; Hardeman & Sjoling, 2007; Riaz et al., 2008). Such screening can be performed using classical growth-selection based expression strategies. Usually hundreds of thousands of clones have to be screened to detect a single active clone, making the development of good screening strategies, preferably automated, most essential. The target genes and their products are usually of low prevalence in the total complex community. One major stumbling block in such strategies is the expression of the activity of interest in the foreign host, typically E. coli, into which the metagenomic library is introduced. Indeed, a combination of gene expression issues, poorly suited transcription and translation machinery and poor secretion of foreign proteins by the host strain, often make it difficult to examine the activities of the transgenes of interest. It also can be that required chaperones for proper protein folding are not present in the host strain, and further cofactors may not be synthesized or incorrectly inserted into the protein. One way to increase chances of successful gene expression is to use more appropriate host species for library construction (Courtois et al., 2003; Martinez et al., 2004; Li et al., 2005). The big advantage of functional screening approaches is that they do not require extensive sequencing. Functional metagenomics approaches can be powerful approach tools, with the potential to identify entirely new classes of genes with novel or known functions.

The second metagenomic strategy focuses on sequence information that can be obtained by sequencing of constructed metagenomic libraries or large-scale pyrosequencing of bulk DNA. Screening of clone libraries may require the development of specific hybridization probes or specific PCR primers (Leveau et al., 2004). Large clones containing target genes can be fully sequenced and subsequently annotated to describe the genes and metabolic pathways they encode. Likewise, full sequences of large clones with phyllogenetic markers of interest may reveal specific genes that may give clues into the physiology and ecology of the original host organism (Liles et al., 2003; Quaiser et al., 2003; Lopez-Garcia et al., 2004; Quaiser et al., 2008). The use of pyrosequencing methods although extremely fast still give only short reads (250 - 450 bp), making the sequences assembly problematic. Unfortunately, in both cases, analyses of recovered genomic information rely mainly on homology to previously identified genes or proteins and are therefore dependent on published genome annotation and databanks. Due to the vast diversity of bacterial genes, novel sequence information often consists mostly of unknown genes.

**Landmark metagenomic examples:**

**Proteorhodopsin**

As described above, one of the strengths of metagenomic analysis is that it can reveal unknown/unexpected microbial physiology. An excellent example is the discovery of rhodopsin-like photoreceptors in marine bacteria (Beja et al., 2000a). The group of Ed DeLong first used the potential of the metagenomic approach to provide a link between phylogeny and key functional genes. In this research, BAC libraries were constructed from
General Introduction

Genome fragments recovered directly from surface waters off the Californian coast (Beja et al., 2000b). Subsequent screening with domain- or group- specific primers led to the identification of an insert containing a 16S rRNA gene of a marine γ-Proteobacteria (Beja et al., 2000b; Beja et al., 2001; Handelsman, 2004). Sequencing and annotation of this 140-kb genomic fragment also revealed an open reading frame encoding a bacteriorhodopsin-like gene (Beja et al., 2000a). This bacteriorhodopsin was subsequently shown to be a light-sensitive integral membrane biomolecule, which works in a similar way as light sensitive cells of the human eye (Beja et al., 2001). Bacteriorhodopsin captures light energy and subsequently converts it into chemical energy. Previously, rhodopsin had been found only in Archaea. Expression of this gene in *E. coli* produced a protein that functioned as a photoreceptor (Beja et al., 2000a). Using biophysical techniques, proteorhodopsins were shown to be present in substantial quantities in bacterioplankton membranes isolated from Pacific Ocean, and appear to have profound consequences for the energy budget of the world’s oceans (Sabehi et al., 2007; Walter et al., 2007).

Acid mine

Advances in genome assembly now allow for genomes to be reconstructed from environmental samples after sequencing of small-insert shot-gun clone libraries. An example is the study by Banfield and coworkers of an acid mine drainage biofilm (Tyson et al., 2004). The biofilm samples were collected from pyrite (FeS2) ore. The extreme conditions (pH < 1) resulted in a very simple microbial biofilm community that balances the acidification caused by pyrite dissolution by the biological oxidation of iron. Analysis of approximately 100 Mb of raw shotgun data obtained from small insert libraries (average insert size of 3.2 kb) allowed for reconstruction of the dominant bacterial and archaeal genomes present. The differences in G + C content between clones helped to distinguish them and provided a good indicator of their source. These analyses allowed for the assembly of nearly complete genomes belonging to *Leptospirillum* group II and *Ferroplasma* type II, as well as the partial reconstruction of three other genomes. In this example, the shot-gun metagenome approach essentially allowed for sequencing of the entire community, providing insight into community-wide physiology and function. It also facilitated further functional genomics and proteomic studies evaluating predicted microbial activities (Wilmes et al., 2009). One should realize, however, that the analysis of other, more complex, communities based on genome reconstruction and assessing functions will be hampered by high species richness and much higher phylogenetic and physiological complexity (Handelsman, 2004; Tyson et al., 2004).

Sargasso Sea

The Sargasso Sea is known as the reproduction area for the European and the American eel populations, which use the Sargassum as cover from predation until they are mature. This environment was the subject of the extremely large environmental shotgun
sequencing project, in which genetic sequences were sampled from communities of microorganisms by Venter et al. (2004). By shotgun metagenome sequencing, over 1 billion bp of sequence information was generated (Venter et al., 2004). Overlapping sequences were assembled and sorted based on oligonucleotide frequencies, read depth, and similarity to previously sequenced genomes. Sequence analysis showed that cloned DNA fragments had been derived from around 1800 genomically different species, of which 148 represented previously unknown bacterial phylotypes. Sequence annotation led to the discovery of 1.2 million new genes, including 782 rhodopsin-like genes. Partial genome scaffolds were also constructed for the most dominant populations. This project gives a new insight into the biogeochemistry of marine ecosystems (Venter et al., 2004). The Sargasso Sea project was recently expanded to The Global Ocean Sampling (GOS), where samples were collected across a several-thousand km transect from the North Atlantic through the Panama Canal and ending in the South Pacific yielding a dataset of over 6 billion bp (Rusch et al., 2007; Yooseph et al. 2007). This survey is currently circumnavigating the globe.

Application of metagenomics to soil environments

Not-yet-cultured soil microorganisms provide a rich source of new natural-product chemistry (Daniel, 2004; Streit et al., 2004; Schmeisser et al., 2007). The problem is that most of the novel structures are from organisms present with low abundance in the soil. Several studies (Handelsman et al., 1998; Rondon et al., 2000; Liles et al., 2003) have tried to employ the metagenome cloning techniques to access the biosynthetic pathways of not-yet-cultured soil bacteria. The strategy generally involved the isolation of high-molecular-weight DNA directly from the soil, cloning of large genomics DNA fragments into a readily cultured host, and screening for biological activities. As an example, Rondon et al. (2000), created a metagenome library using bacterial artificial chromosomes (BACs), which were transformed into E.coli followed by a screening for gene expression on selective solid media. Several biological activities were observed, and analysis of the relevant BAC clones revealed genes encoding proteins with DNase, lipase, or amylase activities as well as clones showing antibacterial activities (Rondon et al., 2000). This approach showed that the DNA extraction directly from the environment is a valuable source providing access to the genetic information of uncultured soil bacteria, and that it is possible to find detectable levels of biochemical activities expressed from heterologous DNA.

In an approach analogous to the Sargasso Sea project, Tringe et al. (2005) attempted to characterize the content of the shotgun metagenome library from farm soil. This study again demonstrated the vast diversity of soil systems, as exemplified by the less than 1% overlap observed in the nearly 150,000 sequencing reads performed. This large diversity prevented the reconstruction of any genomic scaffolds. However, sequence analysis and advanced annotation analysis did provide insight into the enzymatic diversity
contained within this soil, as compared to other samples examined from whale fall samples, and the Acid mine and Sargasso Sea examples described above.

With the recent advances in sequencing technologies, and especially with pyrosequencing, it is now possible to apply more sophisticated statistical approaches to estimate microbial diversity. It also shows, that obtaining 90% of taxonomic diversity would still require thousands of times more sequencing than performed to date (Roesch et al., 2007; Quince et al., 2008). Thus, while sequencing advances have allowed for a quantum leap in the amount of DNA sequence information that can be generated, it is still not a feasible task to characterize full soil communities via shot-gun metagenomic approaches.

Aims

The main objective of this thesis was to gain access to the diversity and functional capabilities of two bacterial phyla, the *Acidobacteria* and *Verrucomicrobia*, which are known to represent a large proportion of soil-borne communities, but have generally eluded classical strategies of cultivation and physiologic characterization.

To achieve this main objective, the following experimental approaches were undertaken:

- Accessing diversity and abundance of *Acidobacteria* and *Verrucomicrobia* at former agricultural experimental field site
- Classification of recovered *Acidobacteria* and *Verrucomicrobia* sequences by means of phylogenetic analysis
- Construction of a large-insert metagenomic library and selection of Acidobacterial and Verrucomicrobial genome fragments
- Sequence analysis, annotation and functional prediction of partial genomic information to gain insight into the ecology of those dominant, not-yet-cultured rhizosphere bacteria

Approach

As a site for soil sampling, an experimental field site, originally established in the EU project Changing Land Usage, Enhancement of biodiversity and ecosystem development (CLUE), was chosen (Van der Putten et al., 2000; Bezemer & van der Putten, 2007). The fields are located in the Veluwe region of The Netherlands. The soil type at this location is characterized as loamy sand.

To study the diversity and structure of microbial communities, a combination of 16S rRNA gene-based methods was selected (PCR-DGGE, Q-PCR 16S rRNA gene library analysis). Group-specific genomic information of *Acidobacteria* and *Verrucomicrobia* was obtained by constructing metagenomic libraries and selecting fosmid clones carrying the appropriate phylogenetic markers (Figure 2).
Figure 2. Summary of strategies used in this study.

Outline of this thesis

Although information about the Acidobacteria, and their potential environmental functions, is still scarce, recent advances in culturing methods, as well as the information gained through culture-independent approaches is starting to provide new insight into the
behavior and functioning of this group of bacteria. These recent advances and insights are summarized in Chapter 2. In Chapter 3, the relationship between plant species composition and diversity and microbial community structure was examined by carrying out a molecular analysis in two targeted field experiments. Experiment 1 focussed on bacterial community structure and diversity in response to different aboveground plant diversity treatments. Experiment 2 was performed to study the influence of 12 different native grassland plant species on bacterial community size and structure with special emphasis on the structure of Acidobacteria and Verrucomicrobia communities.

Chapter 4 focuses on the Acidobacteria phylum and in particular on the comparison of the diversity recovered from different soil compartments and via different cloning strategies. The diversity of 16S rRNA genes affiliated with the Acidobacteria in a former arable soil was examined via three independent approaches: 1) screening of a fosmid metagenome library (28,800 clones) for inserts containing Acidobacteria-like 16S rRNA genes; 2) PCR-cloning using general bacterial primers; and 3) PCR-cloning with primers specific for Acidobacterial 16S rRNA gene sequences. Chapter 5 presents the results of screening of 28,000 metagenomic clones for the presence of Acidobacteria-like 16S rRNA genes. Sequence analysis, gene annotation and genome fragment comparison of clones containing Acidobacterial genomic DNA was performed in order to gain functional insight into the role of diverse Acidobacteria populations in soil and glimpses into genomic evolution in this phylum.

Phylogenetic analysis of verrucomicrobial 16S rRNA genes from clonal libraries (phylum specific and general bacterial ones) obtained from the same former arable field is presented in Chapter 6. Additionally, this chapter describes the metagenomic analyses performed to gain more inside into the genomic information of members of that phylum. DNA fragment composition and comparisons based on the tetranucleotide frequencies allowed for the identification of potential sites of past horizontal gene transfers events.

Finally, in Chapter 7 a summary is provided and the results are discussed in relation to current knowledge on the diversity and functions of Acidobacteria and Verrucomicrobia and recent advances in the field of metagenomics.