ASSESSMENT OF THE FUNCTIONAL DIVERSITY OF SOIL MICROBIAL COMMUNITIES IN THE GERMAN BIODIVERSITY EXPLORATORIES BY METAGENOMICS

D i s s e r t a t i o n

zur Erlangung des mathematisch-naturwissenschaftlichen Doktorgrades

"Doctor rerum naturalium"

der Georg-August-Universität Göttingen

vorgelegt von

Christiane Will

aus Essen

Göttingen 2010

D7

Referent:	PD Dr. Rolf Daniel
Korreferent:	PD Dr. Michael Hoppert
Tag der mündlichen Prüfung:	28.01.2011

Table of contents

C. Will, A. Thürmer, A. Wollherr, H. Nacke, N. Herold, M. Schrumpf, J. Gutknecht, T. Wubet, F. Buscot, and R. Daniel

H. Nacke; A. Thürmer; A. Wollherr; C. Will; L. Hodac; N. Herold; I. Schöning; M. Schrumpf; and R. Daniel

H. Nacke*, C. Will*, S. Herzog, B, Nowka, M. Engelhaupt, and R. Daniel

C. Will, H. Nacke, A. Thürmer und R. Daniel

Chapter 1: General introduction

The term "metagenomics" has first been introduced by Handelsman et al. more than ten years ago (Handelsman et al. 1998) and comprises the extraction and the analysis of the entire genetic complement of a microbial habitat. The analysis can either be based on function or on DNA-sequence. The focus of the majority of metagenomic studies is the detection of novel biomolecules, which can be applied to biotechnological processes or medical research.

While Handelsman's group started analyzing the habitat soil, the focus in this rapidly growing field is nowadays more expanded and includes investigations of various prevalent habitats such as diverse soils representing different land use types (Rondon et al. 2000, Daniel 2005, Hong et al. 2007, Bunterngsook et al. 2010), water (Elend et al. 2006) and sediment samples from lakes, rivers and ponds (Kalyuzhnaya et al. 2005, Ranjan et al. 2005, Leroy et al. 2008, Jogler et al. 2009), and marine habitats (Venter et al. 2004, Hu et al. 2010). Moreover, extreme environments such as glacier ice (Simon et al. 2009), permafrost (Yergeau et al. 2010), hypersaline alkaline, or acidic environments (Hamamura et al. 2005, Bodaker et al. 2010, Grant and Heaphy 2010, Xiang et al. 2010) and extreme geothermal samples (Rhee et al. 2005, Tirawongsaroj et al. 2008) have been investigated.

1.1 Demand for enzymes with novel characteristics for biotechnology

It is estimated that the habitat soil harbors 2,000 to 18,000 different prokaryotic genomes per gram but less than 1% of the soil-born microorganisms are cultivable under laboratory conditions (Torsvik et al. 2002, Daniel 2005). Metagenomics is a useful approach to access the genetic information from uncultured microorganisms. Therefore it is possible to benefit from the huge amount of genetic potential that is stored within the enormous variety of habitats. Metagenomics has already been successfully applied in order to discover a high number of novel bioactive molecules of important industrial value in the white (industrial), green (agricultural), and red (medical) biotechnology. Several targets are listed in recent reviews (Ferrer et al. 2009, Simon and Daniel 2009) and comprise e.g. lipolytic enzymes (Henne et al. 2000), cellulases (Voget et al. 2006),

proteases (Waschkowitz et al. 2009), alcohol oxidoreductases (Knietsch et al. 2003), antibiotics (Brady and Clardy 2004), and antibiotic resistance genes (Courtois et al. 2003). Since metagenomic approaches provided access to a variety of biomolecules it is not surprising that many biotechnological processes are catalyzed with enzymes of microbial origin (Ferrer et al. 2009). The company BASF, for example, uses lipases to obtain enantiopure alcohols or R-amides/S-amines from racemic alcohols or amines, respectively. DSM, a Dutch company, produces semisynthetic penicillins applying acylases (Schmid et al. 2001). Furthermore, the Henkel KGaA registered a patent for an esterase that is able to degrade terephtalate esters, an important component of bioplastics (Michels et al. 2007). This company applied for an additional patent for novel glycosyl hydrolases with amylolytic activity for detergent applications and the sequence-based screening method to detect them in metagenomic DNA (Breves et al. 2003). Enzymes from extreme habitats potentially exhibit unusual properties: A metagenomic library constructed from a deep sea hypersaline basin bore esterases which increased or upheld their activity after exposure to a pressure of 40 MPa. They also displayed highest activities at alkaline pH of 8.5 to 9.0 and one retained 80% of its activity at pH 12.0 (Ferrer et al. 2005). A cold-adapted esterase derived from Antarctic desert soil had their temperature optimum at 40°C and retained 100% activity during the assay period (60 min) at 30°C (Heath et al. 2009). However, novel enzymes that enable an environmentally friendly industry, i.e. minimizing the use of hazardous chemicals and waste production and the reduction of energy consumption, are still required. Consequently, the need for novel or improved biocatalysts that meet these requirements still remains (Jaeger and Holliger 2010).

1.2 Applying next-generation sequencing for the analyses of microbial community compositions

The investigation and characterization of microbial community compositions in different environments is of particular interest for biodiversity research. To approach the genetic information of a habitat with low microbial diversity several attempts have already been made by analyzing metagenomic clone libraries. In fact, the reconstruction of a single prokaryotic genome derived from a habitat with low microbial diversity has al-

ready been performed by sequencing a sufficient number of clones from the corresponding metagenomic library. Examples are "*Candidatus* Cloacamonas acidaminovorans" from an anaerobic digester of a municipal wastewater treatment plant in France (Pelletier et al. 2008), *Kuenenia stuttgartiensis* from a Dutch anammox (anaerobic ammonia oxidation) bioreactor (Strous et al. 2006), or the near-complete genomes of *Leptospirillum* group II and *Ferroplasma* type II from the biofilm of acid mines in California (Tyson et al. 2004). The main metabolic pathways as well as special features were subsequently predicted by genome annotation. Reconstructing individual genomes of a microbial community provides information about the composition, functions and interactions within the community.

Nowadays, next-generation sequencing technologies provide a convenient technique to assess metagenomic information without the need of clone libraries. Three major platforms for next-generation sequencing are currently available on the market: (i) The 454-pyrosequencing GenomeSequencer (Roche Applied Science), (ii) the Illumina/Solexa genome analyzer (Illumina), and (iii) the ABI SOLiD system (Applied Biosystems). All provide high throughput in parallel assays and thereby generate a huge amount of data in short time. Since none of these systems requires cloning of DNA fragments they all are less labor- and cost-intensive compared to Sanger-sequencing. In addition, cloning biases are avoided (Ansorge 2009). The sequencing principle differs between the three platforms, illustrative descriptions of the techniques can be found in recent reviews (Mardis 2008, Ansorge 2009). Since the obtained read length varies between approximately 35 bases with Illumina and SOLiD and an average of 400 bases with 454-pyrosequencing (Ansorge 2009), pyrosequencing is employed for phylogenetic analyses of environmental communities based on amplification and analysis of rRNA genes. The principle and the workflow of 454-pyrosequencing are depicted in Figure 1. Adaptors are linked to the DNA fragments, and the fragments are separated into single strands (Fig. 1a). The adaptors cause the binding of single fragments to beads whose surfaces carry complementary oligonucleotides. Emulsion PCR is employed for fragment amplification: Water droplets which are immersed in oil contain PCR reagents and one of the beads each (Fig. 1b). Subsequently, the DNA strands are denaturized and the beads with the single-stranded DNA fragments are deposited into a picotiter plate (PTP), one bead per well. The wells are filled up with smaller beads which contain reactants required for the pyrosequencing reaction (Fig. 1c-e). In the sequencer the PTP is placed opposite of a
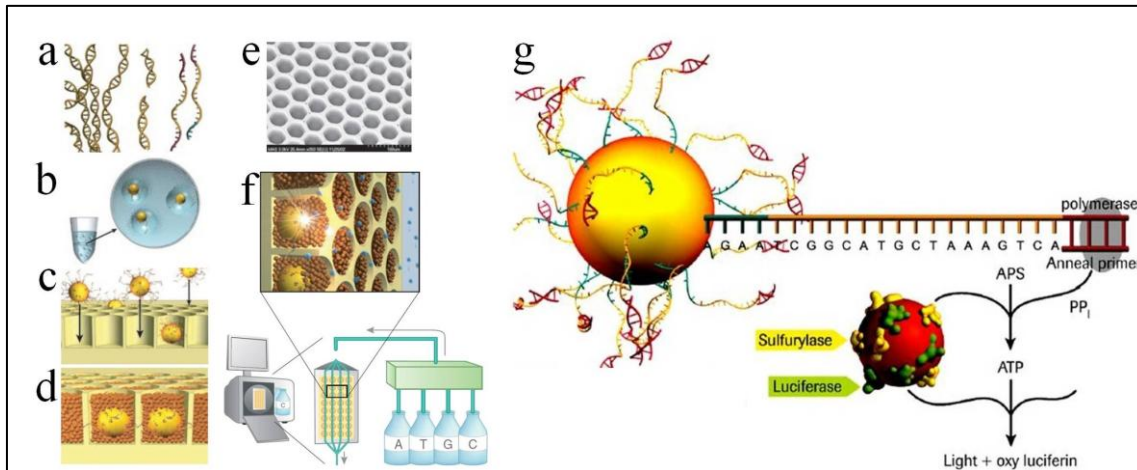
Fig. 1: Overview of the 454 pyrosequencing technology.

(a) single stranded DNA fragments, (b) emulsion-PCR, (c) loading of a PTP, (d) adding enzymes for downstream-applications to PTP, (e) electron micrograph of PTP, (f) the 454 sequencing instruments, (g) scheme of the pyrosequencing reaction where the nucleotide incorporation results in light emission. See text for details. Adapted from www.454.com.

camera and individual nucleotide solutions are provided to the wells by spreading them over the PTP (Fig. 1f). A DNA polymerase synthesizes the complementary strand of the single-stranded template DNA. Each time a nucleotide is incorporated an inorganic pyrophosphate (PPi) is released which in combination with the enzyme luciferase emits light (Fig. 1g). This is detected by the camera in the sequencer and the incorporation of a specific nucleotide is recorded. The amount of light generated by this cascade is proportional to the number of incorporated nucleotides (Margulies et al. 2005, Mardis 2008, Ansorge 2009).

Pyrosequencing has already been employed to obtain insight into the phylogenetic distribution of microorganisms in a wide range of habitats such as ocean water (Sogin et al. 2006, Kirchman et al. 2010), glacier ice (Simon et al. 2009), wastewater treatment plants (Lee et al. 2010), alkaline hot springs (Miller et al. 2009), and several types of soil (Roesch et al. 2007, Acosta-Martínez et al. 2008, Lauber et al. 2009, Rousk et al. 2010, Uroz et al. 2010). In many diversity studies variable regions of the 16S rRNA gene had been amplified prior to pyrosequencing. The sequences often spanned the V1-V2 region (Jones et al. 2009, Lauber et al. 2009, Rousk et al. 2010), the V3 region (Miller et al. 2009, Lee et al. 2010), or the V6 region (Sogin et al. 2006, Kirchman et al. 2010). Sequences of varying read length were generated with the pyrosequencing approaches. Roesch et al. (2007) who published one of the first microbial diversity analyses based on pyrosequencing interpreted sequences with an average read length of

103 bp. Lee et al. (2010) employed the state-of-the-art 454 FLX Titanium pyrosequencing and received sequences with an average read length of 430 bp. However, most recent publications present sequences with read lengths of 220 to 250 bp. Large pyrosequencing data sets have been generated which comprised several tens or hundreds of thousands sequences. Extensive data sets, which contain more than 300,000 partial 16S rRNA sequences, have been produced by Uroz et al. (2010) with 301,887 sequences of rhizosphere and bulk soil and Lee et al. (2010) with 310,901 sequences of anode biofilms in wastewater treatment plants. They analyzed a total of six and four different samples, respectively. Two of the largest data sets from a single sample were derived from a Canadian forest soil sample (53,533 sequences) (Roesch et al. 2007) and a bulk soil sample C3BS (62,022 sequences) (Uroz et al. 2010).

Recently, research started to focus on metatranscriptomic studies to analyze the particular subset of genes in the metagenome of a microbial community that is expressed at the moment of sampling. It has already successfully been employed to several aquatic samples (Frias-Lopez et al. 2008, Gilbert et al. 2008, Gifford et al. 2010), commensal communities (Tartar et al. 2009, Booijink et al. 2010), and soil samples (Bailly et al. 2007, Urich et al. 2008, Shrestha et al. 2009). The variety of habitats whose metatranscriptome has been analyzed is far from reaching the variety of metagenomically analyzed habitats but it is clearly ascending. Different approaches have been applied to gain insight into the active microbial population. Some groups took advantage of polyadenylated mRNA (Bailly et al. 2007, Frias-Lopez et al. 2008, Tartar et al. 2009). Urich et al. (2008) transcribed and sequenced total community RNA and separated ribo-tags from mRNA-tags afterwards. However, many groups successfully employed commercial kits for the subtractive hybridization of rRNA (Gilbert et al. 2008, Shrestha et al. 2009, Booijink et al. 2010, Gifford et al. 2010). The obtained cDNA was either directly applied to pyrosequencing (Frias-Lopez et al. 2008, Gilbert et al. 2008, Urich et al. 2008, Gifford et al. 2010) or ligated in vectors which subsequently were sequenced (Bailly et al. 2007, Shrestha et al. 2009, Tartar et al. 2009). Alternatively, cDNA-ALFP (cDNA amplified fragment length polymorphism) was performed to explore the human gastrointestinal tract (Booijink et al. 2010). The method had limitations obtaining sequence information but was suitable to show differences between the probands and the time points of sampling. The direct pyrosequencing of cDNA from marine bacterioplankton samples revealed that previously unknown microbes play a major role in the

marine ecosystem such as the nitrogen, phosphorous, and sulfur cycles (Gifford et al. 2010). In the picoplanktonic samples analyzed by Frias-Lopez et al. (2008) genes involved in microbial phototrophy were among the most highly expressed genes. When Gilbert et al. (2008) analyzed the metatranscriptome during a phytoplankton bloom they detected primarily housekeeping genes and a large number of sequences that belonged to novel protein clusters. Similarly, 35% of the metatranscriptome sequences from two paddy soil microbial communities could not be assigned and the majority of the assignable 65% were related to cellular housekeeping functions such as ABC-type transporters (Shrestha et al. 2009). Also, 52% of the metatranscriptome sequences from a forest soil eukaryotic community did not correspond to genes with known functions (Bailly et al. 2007). Here, the majority of the assignable sequences belonged to the COG (Clusters of Orthologous Groups of proteins) categories J and O which comprise translation and post-translational modifications. This is in accordance with the findings of Urich et al. (2008): When they compared the metagenome with the metatranscriptome of a German soil sample they detected an overrepresentation of categories involved in RNA and protein metabolism in the metatranscriptome.

Eventually, a combined approach of metagenomics and metatranscriptomics is a promising new way to understand the community structure and the interaction in a complex habitat. Next-generation sequencing as well as bioinformatic databases and software tools are constantly being developed further; advancements are indispensable as the demands on sequence extent and subsequent data analysis are steadily rising.

1.3   Investigations on the microbial habitat soil

When Handelsman et al. published the original definition of the term "metagenome" (Handelsman et al. 1998), cloning was an essential methodology for functional analyses. Due to next-generation sequencing, the necessity for cloning does not exist any longer. A lot of research has already been conducted on the habitat soil and its microbial community. However, the knowledge about the phylogenetic or spatial distributions of microorganisms or about the correlation between soil properties and microbial interactions is still poor. It is a reasonable challenge to investigate the habitat soil in its complexity and the interaction and effect of the biogeochemical factors such as soil struc-

ture, pH, water content, temperature and climatic variations, the availability of oxygen, and the biotic activity. All of these factors affect the character of this habitat with respect to the microbial community size and diversity and hence make soil probably one of the most challenging of all natural environments (Daniel 2005, Rajendhran and Gunasekaran 2008). Bacterial community structure is not only influenced by soil properties but also by land use and modifications of land use (Acosta-Martínez et al. 2008, Wakelin et al. 2008, da C Jesus et al. 2009). It is known that microorganisms are responsible for most nutrient transformation in soils and thus influence plant diversity and productivity. Consequently, microbial communities play an important role in agriculturally managed systems (van der Heijden et al. 2008). Therefore, assessing soil microbial communities is not only of scientific but also of agronomic interest.

Fingerprinting methods such as denaturating gradient gel electrophoresis (DGGE) (Muyzer et al. 1993) or terminal restriction fragment length polymorphism (T-RFLP) (Fierer and Jackson 2006) can efficiently be employed to detect differences in community structure, but insight into diversity is limited (Hartmann and Widmer 2006). 16S rRNA gene analyses give information on the phylogenetic groups present in a sample. Sequencing of cloned (partial) 16S rRNA genes normally reveals information of a few hundreds of clones in the respective metagenomic libraries (Chow et al. 2002, Janssen 2006). In contrast, pyrosequencing allows large-scale analyses of the microbial phylogenetic composition in soil (Roesch et al. 2007, Uroz et al. 2010). Substantial sampling is critical to perform robust estimations of community diversity parameters (Morales et al. 2009) and to assess rare phylogenetic groups that may possess important ecosystem functions (Schloss and Handelsman 2006).

Jones and colleagues studied the diversity of *Acidobacteria* in soils by analyzing 87 different samples from North and South America (Jones et al. 2009). Among a total of 128,477 classified bacterial sequences they detected an average acidobacterial portion of 30.9% (39,707 sequences). The analysis of soil and site characteristics revealed that several of the considered properties such as mean annual precipitation, percent soil organic carbon, and soil C/N ratio correlated with the relative abundances of *Acidobacteria*. The strongest influence, however, was exerted by soil pH which covered a range of 3.56 to 8.86 across all samples and showed a strong negative correlation with the relative abundance of all *Acidobacteria* (Jones et al. 2009). The relationship between soil pH and the relative abundances of the 26 *Acidobacteria* subgroups (Barns et al. 2007)

was also investigated. The relative abundances of most of the subgroups was correlated with the soil pH: While the subgroups 1, 2, 3, 12, 13, and 15 correlated negatively, the subgroups 4, 6, 7, 10, 11, 16, 17, 18, 22, and 25 correlated positively with the pH of the soil samples (Jones et al. 2009). The influence of the soil pH on microbial communities was also investigated by other groups (Lauber et al. 2009, Rousk et al. 2010). According to Lauber et al. (2009), also the overall soil bacterial diversity correlates with soil pH. When testing 88 soil samples from North and South America with a pH range between 3.5 and 9.0 they detected that the phylogenetic diversity is higher at near neutral pH values than in acidic or alkaline samples. Moreover, contrary to the negative correlation of the relative abundance of the *Acidobacteria* with pH, the relative abundances of the *Actinobacteria* and *Bacteroidetes* show a similarly strong positive correlation with soil pH (Lauber et al. 2009). Rousk et al. (2010) surveyed the influence of pH on the abundance and composition of soil bacteria with a liming experiment where the variation of other soil properties had been minimized. They confirmed the correlations of the relative abundances of the *Acidobacteria* subgroups 1, 2, 3, 4, 6, 7, and 16 with the pH as described by Jones et al. (2008). They, like Lauber et al. (2009), detected a positive correlation of the relative abundance of *Bacteroidetes* with pH but it showed no significance in their statistical analysis. The relative abundances of the *Actinobacteria* detected within these 22 soil samples, however, did not correlate with soil pH. The relative abundances of *Nitrospira* and *Alpha-*, *Beta-*, *Gamma-*, and *Deltaproteobacteria* increased with soil pH in their analyses but the trend was significant only for the *Gammaproteobacteria*. Regarding the overall bacterial diversity, Rousk et al. depicted a significant positive correlation of the number of bacterial OTUs (operational taxonomic units) with pH which still increased significantly at pH values above neutral (Rousk et al. 2010).

Apart from the relationship between soil pH and microbial community compositions, the focus of research was recently also on the influence of land use on bacterial communities. Acosta-Martínez et al. (2008) investigated a single soil type under four different land use systems: Two agricultural systems (continuous monoculture with cotton and a rotation of cotton, winter wheat, and corn) and two non-disturbed grass systems (pasture monoculture and a diverse mixture of grasses in a conservation reserve program). The study revealed the *Actinobacteria*, *Bacteroidetes* and *Firmicutes* being predominant (mean values 24.3, 14.2, and 12.2%, respectively) in this high clay soil

across all land use types. Additional taxonomic groups that appeared with a relative abundance of 5 to 10% in all land use systems were the *Acidobacteria* and the *Alpha-*, *Beta-*, and *Gammaproteobacteria*. The main phylogenetic difference between the agricultural and the non-disturbed land use systems was a higher relative abundance of *Actinobacteria* in the undisturbed sampling sites whereas the *Bacteroidetes* were more abundant in the agriculturally managed sampling sites (Acosta-Martínez et al. 2008). Roesch et al. (2007) compared three agricultural and one forest soil sample. The most abundant phylotypes in this study were also the *Bacteroidetes* (approximately 20%) followed by the *Beta-*, *Alpha-*, *Gamma-*, *Deltaproteobacteria*, *Acidobacteria*, *Actinobacteria* and *Firmicutes* (18, 10, 10, 8, 7, 6, and 4%, respectively). The difference between forest and agricultural samples regarding the relative abundances of taxonomic groups can be found for the *Gemmatimonadetes* which represent 3.5% of the sequences in the forest sample but only 1% or less in the agricultural samples. The most noticeable effect of land use in this study is that the forest sample was phylum rich but species poor whereas the agricultural samples were phylum poor and species rich (Roesch et al. 2007). Uroz et al. (2010) focused on the phylum rich forest soil by analyzing oak rhizosphere in comparison with the surrounding bulk soil. The by far most abundant phyla across all samples were the *Proteobacteria* (38%), *Acidobacteria* (24%), and *Actinobacteria* (12%). Other phyla appeared with a relative abundance of less than 2%. No significant difference between the niches "rhizosphere" and "bulk soil" has been detected for any listed phylum. Although the relative portions of the proteobacterial classes have not been listed, a significantly higher abundance of the *Beta-* and *Gammaproteobacteria* in the rhizosphere than in the surrounding soil (Uroz et al. 2010) has been described.

Many studies on soil microbial communities have been conducted so far. Nevertheless, an aspect that had marginally been regarded to date is the sampling depth. Nearly all research was conducted on topsoil samples (A horizon). Though the subsoil (B horizon) also contains an important part of the soil microbial biomass (Fierer et al. 2003). Additionally, agricultural production affects the microbial population in the shallow subsurface similarly to its impact on topsoil (Blume et al. 2002).

An example for different soil horizons is presented in Figure 2, which shows a drill core from a grassland sampling site in the Hainich region (Germany). The different soil horizons of that specific sample (soil type: Stagnosol) are depicted on the right. The A hori-

zon is the top layer of mineral soil which is characterized by its dark brown color. Here, most organic matter accumulation and soil life takes place; in grassland, this layer is completely penetrated by roots. The AB horizon is the transitional horizon between top-soil and subsoil. Its properties are more similar to those of the overlying A horizon but its transition to the B horizon is apparent due to the lighter color. The B horizon is the soil layer between the topsoil and the parent rock material. The original rock structure has mostly been obliterated. The mottled colorization in this soil profile is caused by the accumulation of clay, which is specific for a Stagnosol (FAO 2006, Jahn et al. 2006).

The sampling sites in the Hainich region are embedded in the nationwide conducted functional biodiversity research project entitled "The German Biodiversity Exploratories" (project web page: www.biodiversity-exploratories.de). This project addresses the feedback between land use, biodiversity, and ecosystem processes (Fischer et al. 2010). It is a large-scale research platform, which is designed for long run analysis. Standardized field plots have been set up in three exemplary regions: The Schwäbische Alb in the south west (Baden-Württemberg), the Hainich in the center (Thuringia), and the Schorfheide in the north east (Brandenburg) of Germany. The map (Figure 3) gives more detailed information on the locations of the respective regions. Both, the Schorfheide and the Hainich encompass approximately 1,300 km² whereas the Schwäbische Alb is considerably smaller (422 km²). The altitudes range from 3-140 m above sea level (Schorfheide) to 460-860 m (Schwäbische Alb) (Fischer et al. 2010). The expression "Exploratory" indicates that, in contrast to an observatory, comparative observations and also field experiments are performed. Ideally, the outcomes can be compared and correlated interdisciplinary. Hence, an equal number of field plots have been set up in each of the three Exploratories: Approximately 1000 grid plots per location were chosen, which are mainly used for large-scale biodiversity analyses. A selected subset of 100 experimental plots enables more thorough environmental monitoring and assessment. Eighteen very intensive plots, another subset of the experimental
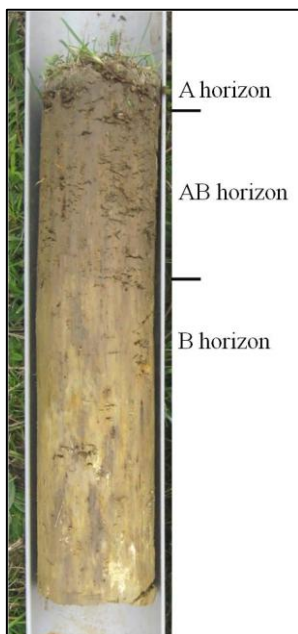

Fig. 2: Drill core of a grassland sampling site in the Hainich region, Germany.

Depicted is a soil sample of a Vertic Stagnosol, which is the most frequent soil type among the analyzed soil samples. The A horizon is the topsoil, the AB horizon a transitional horizon and the B horizon is the subsoil.

Fig. 3: Map of Germany with the location of the three German Biodiversity Exploratories.

The Exploratory Schwäbische Alb is depicted in blue, the Hainich area in red and the Schorfheide in orange. The zoomed maps with the colored frames show details of each Exploratory and the small colored boxes their respective position in Germany. The red and blue spots within every detailed map represent sampling sites of the Exploratory. Sources: http://commons.wikimedia.org (Map of Germany) and project database BExIS (detailed maps of the Exploratories).

plots, provide a possibility to investigate ecological processes and what is connected with that in the minutest detail. On every hierarchical level, half of the sampling sites are situated in grassland, half in forests. They comprise various types of vegetation and management, which, again, are as comparable as possible between the Exploratories.

1.4 Aim of the thesis

The aim of this thesis was the analysis of the microbial diversity in soil samples derived from the Hainich area (Thuringia, Germany). The focus was set on grassland samples of varying sampling depths (topsoil and subsoil) which had been subjected to different management types prior to sampling. Additionally, the difference between soil samples derived from grassland and forest sites was object of research. For this purpose phylogenetic analyses of partial 16S rRNA gene sequences were performed using deep-sequencing approaches.

Furthermore, metagenomic small-insert and large-insert libraries were constructed using DNA isolated from the soil samples. These libraries were employed in functional screening for lipolytic enzymes. The screening was based on the phenotypical detection of active recombinant clones, which were characterized with respect to the substrate spectrum.

1.5   References for chapter 1

Acosta-Martínez, V., S. Dowd, Y. Sun, and V. Allen. 2008. Tag-encoded pyrosequencing analysis of bacterial diversity in a single soil type as affected by management and land use. Soil Biol Biochem **40**:2762–2770.

Ansorge, W. 2009. Next-generation DNA sequencing techniques. N Biotechnol **25**:195-203.

Bailly, J., L. Fraissinet-Tachet, M. Verner, J. Debaud, M. Lemaire, M. Wésolowski-Louvel, and R. Marmeisse. 2007. Soil eukaryotic functional diversity, a metatranscriptomic approach. ISME J **1**:632-642.

Barns, S., E. Cain, L. Sommerville, and C. Kuske. 2007. Acidobacteria phylum sequences in uranium-contaminated subsurface sediments greatly expand the known diversity within the phylum. Appl Environ Microbiol **73**:3113-3116.

Blume, E., M. Bischoff, J. Reichert, T. Moorman, A. Konopka, and R. Turco. 2002. Surface and subsurface microbial biomass, community structure and metabolic activity as a function of soil depth and season. Appl Soil Ecol **20**:171-181.

Bodaker, I., I. Sharon, M. Suzuki, R. Feingersch, M. Shmoish, E. Andreishcheva, M. Sogin, M. Rosenberg, M. Maguire, S. Belkin, A. Oren, and O. Béjà. 2010. Comparative community genomics in the Dead Sea: an increasingly extreme environment. ISME J **4**:399-407.

Booijink, C., J. Boekhorst, E. Zoetendal, H. Smidt, M. Kleerebezem, and W. de Vos. 2010. Metatranscriptome analysis of the human fecal microbiota reveals subject-specific expression profiles, with genes encoding proteins involved in carbohydrate metabolism being dominantly expressed. Appl Environ Microbiol **76**:5533-5540.

Brady, S. and J. Clardy. 2004. Palmitoylputrescine, an antibiotic isolated from the heterologous expression of DNA extracted from bromeliad tank water. J Nat Prod **67**:1283-1286.

Breves, R., K. Maurer, J. Eck, P. Lorenz, and H. Zinke. 2003. New glycosyl hydrolases. Patent WO 03/054177.

Bunterngsook, B., P. Kanokratana, T. Thongaram, S. Tanapongpipat, T. Uengwetwanit, S. Rachdawong, T. Vichitsoonthonkul, and L. Eurwilaichitr. 2010. Identification and characterization of lipolytic enzymes from a peat-swamp forest soil metagenome. Biosci Biotechnol Biochem **74**:1848-1854.

Chow, M., C. Radomski, J. McDermott, J. Davies, and P. Axelrood. 2002. Molecular characterization of bacterial diversity in Lodgepole pine (*Pinus contorta*) rhizosphere soils from British Columbia forest soils differing in disturbance and geographic source. FEMS Microbiol Ecol **42**:347-357.

Courtois, S., C. Cappellano, M. Ball, F. Francou, P. Normand, G. Helynck, A. Martinez, S. Kolvek, J. Hopke, M. Osburne, P. August, R. Nalin, M. Guérineau, P. Jeannin, P. Simonet, and J. Pernodet. 2003. Recombinant environmental libraries provide access to microbial diversity for drug discovery from natural products. Appl Environ Microbiol **69**:49-55.

da C Jesus, E., T. Marsh, J. Tiedje, and F. de S Moreira. 2009. Changes in land use alter the structure of bacterial communities in Western Amazon soils. ISME J **3**:1004-1011.

Daniel, R. 2005. The metagenomics of soil. Nat Rev Microbiol **3**:470-478.

Elend, C., C. Schmeisser, C. Leggewie, P. Babiak, J. Carballeira, H. Steele, J. Reymond, K. Jaeger, and W. Streit. 2006. Isolation and biochemical characterization of two novel metagenome-derived esterases. Appl Environ Microbiol **72**:3637-3645.

FAO. 2006. IUSS Working Group WRB. 2006. World reference base for soil resources 2006. 2nd edition. World Soil Resources Reports No. 103. FAO, Rome.

Ferrer, M., A. Beloqui, K. Timmis, and P. Golyshin. 2009. Metagenomics for mining new genetic resources of microbial communities. J Mol Microbiol Biotechnol **16**:109-123.

Ferrer, M., O. Golyshina, T. Chernikova, A. Khachane, V. Martins Dos Santos, M. Yakimov, K. Timmis, and P. Golyshin. 2005. Microbial enzymes mined from the Urania deep-sea hypersaline anoxic basin. Chem Biol **12**:895-904.

Fierer, N. and R. Jackson. 2006. The diversity and biogeography of soil bacterial communities. Proc Natl Acad Sci U S A **103**:626-631.

Fierer, N., J. Schimel, and P. Holden. 2003. Variations in microbial community composition through two soil depth profiles. Soil Biol Biochem **35**:167–176.

Fischer, M., O. Bossdorf, S. Gockel, F. Hänsel, A. Hemp, D. Hessenmöller, G. Korte, J. Nieschulze, S. Pfeiffer, D. Prati, S. Renner, I. Schöning, U. Schumacher, K. Wells, F. Buscot, E. Kalko, K. Linsenmair, E.-D. Schulze, and W. Weisser. 2010. Implementing large-scale and long-term functional biodiversity research: The Biodiversity Exploratories. Basic Appl Ecol **11**:473–485.

Frias-Lopez, J., Y. Shi, G. Tyson, M. Coleman, S. Schuster, S. Chisholm, and E. Delong. 2008. Microbial community gene expression in ocean surface waters. Proc Natl Acad Sci U S A **105**:3805-3810.

Gifford, S., S. Sharma, J. Rinta-Kanto, and M. Moran. 2010. Quantitative analysis of a deeply sequenced marine microbial metatranscriptome. ISME J.

Gilbert, J., D. Field, Y. Huang, R. Edwards, W. Li, P. Gilna, and I. Joint. 2008. Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. PLoS One **3**:e3042.

Grant, W. and S. Heaphy. 2010. Metagenomics and recovery of enzyme genes from alkaline saline environments. Environ Technol **31**:1135-1143.

Hamamura, N., S. Olson, D. Ward, and W. Inskeep. 2005. Diversity and functional analysis of bacterial communities associated with natural hydrocarbon seeps in acidic soils at Rainbow Springs, Yellowstone National Park. Appl Environ Microbiol **71**:5943-5950.

Handelsman, J., M. Rondon, S. Brady, J. Clardy, and R. Goodman. 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol **5**:R245-249.

Hartmann, M. and F. Widmer. 2006. Community structure analyses are more sensitive to differences in soil bacterial communities than anonymous diversity indices. Appl Environ Microbiol **72**:7804-7812.

Heath, C., X. Hu, S. Cary, and D. Cowan. 2009. Identification of a novel alkaliphilic esterase active at low temperatures by screening a metagenomic library from antarctic desert soil. Appl Environ Microbiol **75**:4657-4659.

Henne, A., R. Schmitz, M. Bömeke, G. Gottschalk, and R. Daniel. 2000. Screening of environmental DNA libraries for the presence of genes conferring lipolytic activity on *Escherichia coli*. Appl Environ Microbiol **66**:3113-3116.

Hong, K., H. Lim, E. Chung, E. Park, M. Lee, J. Kim, G. Choi, K. Cho, and S. Lee. 2007. Selection and characterization of forest soil metagenome genes encoding lipolytic enzymes. J Microbiol Biotechnol **17**:1655-1660.

Hu, Y., C. Fu, Y. Huang, Y. Yin, G. Cheng, F. Lei, N. Lu, J. Li, E. Ashforth, L. Zhang, and B. Zhu. 2010. Novel lipolytic genes from the microbial metagenomic library of the South China Sea marine sediment. FEMS Microbiol Ecol **72**:228-237.

Jaeger, K. and P. Holliger. 2010. Chemical biotechnology-a marriage of convenience and necessity. Curr Opin Biotechnol.

Jahn, R., H. Blume, V. Asio, O. Spaargaren, and P. Schad. 2006. Guidelines for soil description. Fourth edition edition. FAO, Rome (Italy).

Janssen, P. 2006. Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes. Appl Environ Microbiol **72**:1719-1728.

Jogler, C., W. Lin, A. Meyerdierks, M. Kube, E. Katzmann, C. Flies, Y. Pan, R. Amann, R. Reinhardt, and D. Schüler. 2009. Toward cloning of the magnetotactic metagenome: identification of magnetosome island gene clusters in uncultivated magnetotactic bacteria from different aquatic sediments. Appl Environ Microbiol **75**:3972-3979.

Jones, R., M. Robeson, C. Lauber, M. Hamady, R. Knight, and N. Fierer. 2009. A comprehensive survey of soil acidobacterial diversity using pyrosequencing and clone library analyses. ISME J **3**:442-453.

Kalyuzhnaya, M., S. Bowerman, O. Nercessian, M. Lidstrom, and L. Chistoserdova. 2005. Highly divergent genes for methanopterin-linked C1 transfer reactions in Lake Washington,

assessed via metagenomic analysis and mRNA detection. Appl Environ Microbiol **71**:8846-8854.

Kirchman, D., M. Cottrell, and C. Lovejoy. 2010. The structure of bacterial communities in the western Arctic Ocean as revealed by pyrosequencing of 16S rRNA genes. Environ Microbiol **12**:1132-1143.

Knietsch, A., T. Waschkowitz, S. Bowien, A. Henne, and R. Daniel. 2003. Construction and screening of metagenomic libraries derived from enrichment cultures: generation of a gene bank for genes conferring alcohol oxidoreductase activity on *Escherichia coli*. Appl Environ Microbiol **69**:1408-1416.

Lauber, C., M. Hamady, R. Knight, and N. Fierer. 2009. Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. Appl Environ Microbiol **75**:5111-5120.

Lee, T., T. Van Doan, K. Yoo, S. Choi, C. Kim, and J. Park. 2010. Discovery of commonly existing anode biofilm microbes in two different wastewater treatment MFCs using FLX Titanium pyrosequencing. Appl Microbiol Biotechnol **87**:2335-2343.

Leroy, M., M. Prigent, M. Dutertre, F. Confalonieri, and M. Dubow. 2008. Bacteriophage morphotype and genome diversity in Seine River sediment. Freshwater Biol **53**:1176–1185.

Mardis, E. 2008. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet **9**:387-402.

Margulies, M., M. Egholm, W. Altman, S. Attiya, J. Bader, L. Bemben, J. Berka, M. Braverman, Y. Chen, Z. Chen, S. Dewell, L. Du, J. Fierro, X. Gomes, B. Goodwin, W. He, S. Helgesen, C. Ho, G. Irzyk, S. Jando, M. Alenquer, T. Jarvie, K. Jirage, J. Kim, J. Knight, J. Lanza, J. Leamon, S. Lefkowitz, M. Lei, J. Li, K. Lohman, H. Lu, V. Makhijani, K. McDade, M. McKenna, E. Myers, E. Nickerson, J. Nobile, R. Plant, B. Puc, M. Ronan, G. Roth, G. Sarkis, J. Simons, J. Simpson, M. Srinivasan, K. Tartaro, A. Tomasz, K. Vogt, G. Volkmer, S. Wang, Y. Wang, M. Weiner, P. Yu, R. Begley, and J. Rothberg. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature **437**:376-380.

Michels, A., A. Puetz, K. Maurer, T. Eggert, and K. Jaeger. 2007. Use of esterases for separating plastics. Patent WO2007/017181.

Miller, S., A. Strong, K. Jones, and M. Ungerer. 2009. Bar-coded pyrosequencing reveals shared bacterial community properties along the temperature gradients of two alkaline hot springs in Yellowstone National Park. Appl Environ Microbiol **75**:4565-4572.

Morales, S., T. Cosart, J. Johnson, and W. Holben. 2009. Extensive phylogenetic analysis of a soil bacterial community illustrates extreme taxon evenness and the effects of amplicon length, degree of coverage, and DNA fractionation on classification and ecological parameters. Appl Environ Microbiol **75**:668-675.

Muyzer, G., E. de Waal, and A. Uitterlinden. 1993. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. Appl Environ Microbiol **59**:695-700.

Pelletier, E., A. Kreimeyer, S. Bocs, Z. Rouy, G. Gyapay, R. Chouari, D. Rivière, A. Ganesan, P. Daegelen, A. Sghir, G. Cohen, C. Médigue, J. Weissenbach, and D. Le Paslier. 2008. "*Candidatus* Cloacamonas acidaminovorans": genome sequence reconstruction provides a first glimpse of a new bacterial division. J Bacteriol **190**:2572-2579.

Rajendhran, J. and P. Gunasekaran. 2008. Strategies for accessing soil metagenome for desired applications. Biotechnol Adv **26**:576-590.

Ranjan, R., A. Grover, R. Kapardar, and R. Sharma. 2005. Isolation of novel lipolytic genes from uncultured bacteria of pond water. Biochem Biophys Res Commun **335**:57-65.

Rhee, J., D. Ahn, Y. Kim, and J. Oh. 2005. New thermophilic and thermostable esterase with sequence similarity to the hormone-sensitive lipase family, cloned from a metagenomic library. Appl Environ Microbiol **71**:817-825.

Roesch, L., R. Fulthorpe, A. Riva, G. Casella, A. Hadwin, A. Kent, S. Daroub, F. Camargo, W. Farmerie, and E. Triplett. 2007. Pyrosequencing enumerates and contrasts soil microbial diversity. ISME J **1**:283-290.

Rondon, M., P. August, A. Bettermann, S. Brady, T. Grossman, M. Liles, K. Loiacono, B. Lynch, I. MacNeil, C. Minor, C. Tiong, M. Gilman, M. Osburne, J. Clardy, J. Handelsman, and R. Goodman. 2000. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Appl Environ Microbiol **66**:2541-2547.

Rousk, J., E. Bååth, P. Brookes, C. Lauber, C. Lozupone, J. Caporaso, R. Knight, and N. Fierer. 2010. Soil bacterial and fungal communities across a pH gradient in an arable soil. ISME J **4**:1340-1351.

Schloss, P. and J. Handelsman. 2006. Toward a census of bacteria in soil. PLoS Comput Biol **2**:e92.

Schmid, A., J. Dordick, B. Hauer, A. Kiener, M. Wubbolts, and B. Witholt. 2001. Industrial biocatalysis today and tomorrow. Nature **409**:258-268.

Shrestha, P., M. Kube, R. Reinhardt, and W. Liesack. 2009. Transcriptional activity of paddy soil bacterial communities. Environ Microbiol **11**:960-970.

Simon, C. and R. Daniel. 2009. Achievements and new knowledge unraveled by metagenomic approaches. Appl Microbiol Biotechnol **85**:265-276.

Simon, C., A. Wiezer, A. Strittmatter, and R. Daniel. 2009. Phylogenetic diversity and metabolic potential revealed in a glacier ice metagenome. Appl Environ Microbiol **75**:7519-7526.

Sogin, M., H. Morrison, J. Huber, D. Mark Welch, S. Huse, P. Neal, J. Arrieta, and G. Herndl. 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci U S A **103**:12115-12120.

Strous, M., E. Pelletier, S. Mangenot, T. Rattei, A. Lehner, M. Taylor, M. Horn, H. Daims, D. Bartol-Mavel, P. Wincker, V. Barbe, N. Fonknechten, D. Vallenet, B. Segurens, C. Schenowitz-Truong, C. Médigue, A. Collingro, B. Snel, B. Dutilh, H. Op den Camp, C. van der Drift, I. Cirpus, K. van de Pas-Schoonen, H. Harhangi, L. van Niftrik, M. Schmid, J. Keltjens, J. van de Vossenberg, B. Kartal, H. Meier, D. Frishman, M. Huynen, H. Mewes, J. Weissenbach, M. Jetten, M. Wagner, and D. Le Paslier. 2006. Deciphering the evolution and metabolism of an anammox bacterium from a community genome. Nature **440**:790-794.

Tartar, A., M. Wheeler, X. Zhou, M. Coy, D. Boucias, and M. Scharf. 2009. Parallel metatranscriptome analyses of host and symbiont gene expression in the gut of the termite *Reticulitermes flavipes*. Biotechnol Biofuels **2**:25.

Tirawongsaroj, P., R. Sriprang, P. Harnpicharnchai, T. Thongaram, V. Champreda, S. Tanapongpipat, K. Pootanakit, and L. Eurwilaichitr. 2008. Novel thermophilic and thermostable lipolytic enzymes from a Thailand hot spring metagenomic library. J Biotechnol **133**:42-49.

Torsvik, V., L. Øvreås, and T. Thingstad. 2002. Prokaryotic Diversity–Magnitude, Dynamics, and Controlling Factors. Science **296**:1064-1066.

Tyson, G., J. Chapman, P. Hugenholtz, E. Allen, R. Ram, P. Richardson, V. Solovyev, E. Rubin, D. Rokhsar, and J. Banfield. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature **428**:37-43.

Urich, T., A. Lanzén, J. Qi, D. Huson, C. Schleper, and S. Schuster. 2008. Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. PLoS One **3**:e2527.

Uroz, S., M. Buée, C. Murat, P. Frey-Klett, and F. Martin. 2010. Pyrosequencing reveals a contrasted bacterial diversity between oak rhizosphere and surrounding soil. Environ Microbiol Rep **2**:281-288.

van der Heijden, M., R. Bardgett, and N. van Straalen. 2008. The unseen majority: soil microbes as drivers of plant diversity and productivity in terrestrial ecosystems. Ecol Lett **11**:296-310.

Venter, J., K. Remington, J. Heidelberg, A. Halpern, D. Rusch, J. Eisen, D. Wu, I. Paulsen, K. Nelson, W. Nelson, D. Fouts, S. Levy, A. Knap, M. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y. Rogers, and H. Smith. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science **304**:66-74.

Voget, S., H. Steele, and W. Streit. 2006. Characterization of a metagenome-derived halotolerant cellulase. J Biotechnol **126**:26-36.

Wakelin, S., L. Macdonald, S. Rogers, A. Gregg, T. Bolger, and J. Baldock. 2008. Habitat selective factors influencing the structural composition and functional capacity of microbial communities in agricultural soils. Soil Biol Biochem **40**:803–813.

Waschkowitz, T., S. Rockstroh, and R. Daniel. 2009. Isolation and characterization of metalloproteases with a novel domain structure by construction and screening of metagenomic libraries. Appl Environ Microbiol **75**:2506-2516.

Xiang, W., J. Zhang, L. Li, H. Liang, H. Luo, J. Zhao, Z. Yang, and Q. Sun. 2010. Screening a novel Na$^+$/H$^+$ antiporter gene from a metagenomic library of halophiles colonizing in the Dagong Ancient Brine Well in China. FEMS Microbiol Lett **306**:22-29.

Yergeau, E., H. Hogues, L. Whyte, and C. Greer. 2010. The functional potential of high Arctic permafrost revealed by metagenomic sequencing, qPCR and microarray analyses. ISME J **4**:1206-1214.

Chapter 2: Horizon-specific bacterial community composition of German grassland soils as revealed by pyrosequencing-based analysis of 16S rRNA genes

Christiane Will, Andrea Thürmer, Antje Wollherr, Heiko Nacke, Nadine Herold, Marion Schrumpf, Jessica Gutknecht, Tesfaye Wubet, François Buscot, and Rolf Daniel

In the first part of the thesis the microbial phylogenetic community composition of different soil samples from the Hainich region (Thuringia, Germany) was investigated. The 9 grassland sampling sites had been subjected to different management types. Topsoil as well as the corresponding subsoil of each site was employed in the analysis. A substantial pyrosequencing approach yielded a large amount of 16S rRNA gene sequence data which exceeded that of previous soil studies of varying sampling depths. The taxonomic composition was determined and compared to biogeochemical properties of the site. Significant differences in community composition between topsoil and subsoil samples were detected. The organic carbon content was the major driver affecting soil microbial composition.

Author contributions to the work:

Performed the experiments: CW, AT, HN.

Analyzed data: CW.

Bioinformatic support: AW.

Contributed data on soil properties and analysis of these data: NH, MS.

Contributed data on microbial biomass and analysis of these data: JG, TW, FB.

Wrote the paper: CW, RD.

Conceived and designed the experiments: RD.

Horizon-Specific Bacterial Community Composition of German Grassland Soils, as Revealed by Pyrosequencing-Based Analysis of 16S rRNA Genes[∇][†]

Christiane Will,[1] Andrea Thürmer,[2] Antje Wollherr,[2] Heiko Nacke,[1] Nadine Herold,[3] Marion Schrumpf,[3] Jessica Gutknecht,[4] Tesfaye Wubet,[4] François Buscot,[4] and Rolf Daniel[1,2]*

Department of Genomic and Applied Microbiology[1] and Göttingen Genomics Laboratory,[2] Institute of Microbiology and Genetics, Georg-August-Universität Göttingen, Grisebachstr. 8, D-37077 Göttingen, Germany; Max Planck Institute for Biogeochemistry, Hans-Knöll-Str. 10, D-07745 Jena, Germany[3]; and UFZ-Helmholtz Centre for Environmental Research, Department of Soil Ecology, Theodor-Lieser-Str. 4, D-06120 Halle/Saale, Germany[4]

Received 3 May 2010/Accepted 9 August 2010

The diversity of bacteria in soil is enormous, and soil bacterial communities can vary greatly in structure. Here, we employed a pyrosequencing-based analysis of the V2-V3 16S rRNA gene region to characterize the overall and horizon-specific (A and B horizons) bacterial community compositions in nine grassland soils, which covered three different land use types. The entire data set comprised 752,838 sequences, 600,544 of which could be classified below the domain level. The average number of sequences per horizon was 41,824. The dominant taxonomic groups present in all samples and horizons were the *Acidobacteria*, *Betaproteobacteria*, *Actinobacteria*, *Gammaproteobacteria*, *Alphaproteobacteria*, *Deltaproteobacteria*, *Chloroflexi*, *Firmicutes*, and *Bacteroidetes*. Despite these overarching dominant taxa, the abundance, diversity, and composition of bacterial communities were horizon specific. In almost all cases, the estimated bacterial diversity (H') was higher in the A horizons than in the corresponding B horizons. In addition, the H' was positively correlated with the organic carbon content, the total nitrogen content, and the C-to-N ratio, which decreased with soil depth. It appeared that lower land use intensity results in higher bacterial diversity. The majority of sequences affiliated with the *Actinobacteria*, *Bacteroidetes*, *Cyanobacteria*, *Fibrobacteres*, *Firmicutes*, *Spirochaetes*, *Verrucomicrobia*, *Alphaproteobacteria*, *Betaproteobacteria*, and *Gammaproteobacteria* were derived from A horizons, whereas the majority of the sequences related to *Acidobacteria*, *Chloroflexi*, *Gemmatimonadetes*, *Nitrospira*, TM7, and WS3 originated from B horizons. The distribution of some bacterial phylogenetic groups and subgroups in the different horizons correlated with soil properties such as organic carbon content, total nitrogen content, or microbial biomass.

Soil is probably the most complex microbial environment on Earth with respect to species richness and community size. The microbial richness in soils exceeds that of other environments (44) and is higher by orders of magnitude than the biodiversity of plants and animals. Cultivated soil or grassland soil contains an estimated 2×10^9 prokaryotic cells per gram (12). Soil microbial communities are an important factor of agriculturally managed systems, as they are responsible for most nutrient transformations in soil and influence the above-ground plant diversity and productivity (53).

To analyze the bacterial community in soils, most approaches target the 16S rRNA gene by PCR amplification and subsequent analysis employing sequencing of clone libraries (10, 24), denaturing gradient gel electrophoresis (DGGE) (38), or terminal restriction fragment length polymorphism (T-RFLP)

(17, 52). Most of these approaches provided limited insights into the structure of soil bacterial communities, as the survey sizes and the number of compared sampling sites were small with respect to the enormous bacterial diversity present in different soil samples. For example, the reported clone libraries vary considerably in size, but small sample sizes (500 or fewer 16S rRNA gene sequences) are usually analyzed and employed for the theoretical estimation of species richness (39). This provides snapshots of the predominant bacterial community members, but phylogenetic groups that are present in a low abundance and which may possess important ecosystem functions are not assessed (47). In addition, it has been shown that rich sampling (several thousands of clones) of complex bacterial communities is required to perform robust measurements and estimations of community diversity parameters (37). Thus, the detection bias accompanying analyses of small sample sizes can lead to invalidated assumptions. Genetic profiling techniques such as DGGE and T-RFLP have high-throughput capability. These approaches allow researchers to unravel differences in community structure but are limited for assessing diversity (23, 40). To deeply survey the diversity and the composition of the bacterial communities within different soil samples, large-scale pyrosequencing of partial 16S rRNA

TABLE 1. Physical and geochemical characteristics of the soil samples from two different soil horizons

Horizon	Plot	Land use type	pH[a]	Water content (%)[a]	Microbial biomass (nmol PLFA/g dry soil)	OC content (g/kg)[a]	N content (g/kg)[a]	C-to-N ratio[a]	Soil texture (%)		
									Sand	Silt	Clay
A	1	Fertilized meadow	6.63	38.50	81.85	66.20	6.24	10.61	6.8	45.8	47.4
A	2	Fertilized meadow	7.12	25.97	35.83	32.60	3.34	9.75	8.8	37.1	54.1
A	3	Fertilized meadow	7.20	25.49	53.83	26.02	2.90	8.97	5.6	37.2	57.2
A	4	Fertilized mown pasture, cattle	6.49	45.68	131.00	66.95	6.02	11.11	6.7	51.4	41.9
A	5	Fertilized mown pasture, cattle	6.91	35.96	95.19	53.41	5.09	10.50	7.0	46.5	46.5
A	6	Fertilized mown pasture, cattle	6.03	21.22	24.13	14.24	1.63	8.71	7.0	66.4	26.6
A	7	Unfertilized pasture, cattle	6.91	43.06	117.60	70.08	6.36	11.02	6.2	41.7	52.1
A	8	Unfertilized pasture, cattle	6.97	41.45	139.82	74.84	6.90	10.85	6.7	44.4	48.9
A	9	Unfertilized pasture, cattle	6.62	30.94	119.91	48.27	4.13	11.69	7.9	51.1	41.0
B	1	Fertilized meadow	7.10	23.16	27.99	17.77	2.00	8.88	7.0	46.2	46.8
B	2	Fertilized meadow	7.32	22.47	1.34	4.38	0.58	7.54	22.5	27.2	50.3
B	3	Fertilized meadow	7.40	22.14	5.82	7.19	0.99	7.25	2.9	33.0	64.1
B	4	Fertilized mown pasture, cattle	7.35	23.19	9.43	6.13	0.83	7.34	4.4	36.8	58.8
B	5	Fertilized mown pasture, cattle	7.18	22.32	12.87	10.26	1.19	8.62	8.6	53.3	38.1
B	6	Fertilized mown pasture, cattle	6.30	20.10	5.64	4.23	0.66	6.44	5.6	67.7	26.8
B	7	Unfertilized pasture, cattle	7.26	26.29	60.36	34.39	3.62	9.51	5.3	44.7	50.0
B	8	Unfertilized pasture, cattle	7.28	21.75	17.36	19.12	2.22	8.59	11.1	45.6	43.3
B	9	Unfertilized pasture, cattle	7.35	19.94	7.92	5.02	0.63	7.93	8.5	50.4	41.1

[a] Statistically significant differences between the A and B horizons ($P \leq 0.01$).

genes has been employed recently. Previous pyrosequencing-based studies of soil (1, 30, 34, 43) have generated large data sets, which comprised 39,707 (30) to 152,359 (34) 16S rRNA partial gene sequences. Those studies provided comprehensive insights into the biogeography of bacterial soil communities and taxa that were present in a low abundance. However, all those studies focused on the analysis of microbial communities present in topsoil. The subsoil is also known to harbor an important part of the soil microbial biomass (18). It has been shown that the microbial population in the shallow subsurface is impacted by agricultural production to a similar extent as that in topsoil (5).

In this study, we performed large-scale pyrosequencing-based analyses of 16S rRNA genes to assess the bacterial community composition in topsoil and the corresponding subsoil of nine different grassland sites in the Hainich region (Thuringia, Germany). To provide a high level of coverage at the species level (97% genetic distance) and minimize detection bias, we exceeded the above-described numbers of analyzed 16S rRNA gene sequences (752,838 in this study). To examine the impact of land use on bacterial diversity and community composition, the selected grassland sites covered a range of three different land use types, including samples from unfertilized pastures grazed by cattle, fertilized mown pastures grazed by cattle, and fertilized meadows. In many recent studies, surveys were focused on comprehensive analyses of a single soil or a few soil samples (1, 14, 37, 43). This allowed the determination of overall bacterial species richness and community composition, but the assessment of spatial patterns and environmental factors that drive these patterns is hampered by the limited number of examined soils. To assess spatial distribution and the impact of soil edaphic factors and land use on community structure, we used triplicate samples of each land use type from different locations. In addition, composite samples derived from five soil cores after the separation of soil horizons were employed.

MATERIALS AND METHODS

Study sites and soil sampling. Soil samples were collected from nine different grassland sites of the Hainich region in Germany, which is located in the west of Thuringia near the border to Hessen (latitude, 51.2167/N 51°13′0″; longitude, 10.45/E 10°27′0″). The Hainich region is one of the three locations investigated within the framework of the German Biodiversity Exploratories initiative (www.biodiversity-exploratories.de). The nine sampling sites encompassed the following three different land use types: fertilized meadow (plots 1 to 3), fertilized mown pasture grazed by cattle (plots 4 to 6), and unfertilized pasture grazed by cattle (plots 7 to 9) (for coordinates, see Table S1 in the supplemental material). Sampling was performed in April and May 2008. At each sampling site, five soil cores (8.3 cm in diameter) were sampled with a motor-driven soil column cylinder at each corner and in the center of the plot within a given area of 20 m by 20 m. The soil was classified using the World Reference Base of Soil Resources (27). The predominant soil type in the studied plots is Stagnosol, which is characterized by a perched water table, strong mottling, and reducing conditions (27). For each soil core, we determined soil horizons according to the *Guidelines for Soil Description* (28). The horizons were homogenized and pooled into one composite sample per plot and horizon. Coarse roots and stones (>5 mm) were removed from the samples. In the majority of the samples, the horizons Ah (topsoil) and Btg (subsoil) were detected. In plots 2 and 3 the top horizon was a transitional horizon (ABth) that was employed instead of the missing Ah horizon. In plots 7 and 8 only a transitional horizon (ABth) between topsoil and parent rock material was present. This horizon was employed instead of the missing Btg horizon. Throughout the study the topsoil horizon and the subsoil horizon were designated horizon A and horizon B, respectively.

Edaphic properties of the soil samples. For determinations of organic carbon (OC) content, total nitrogen (N) content, and soil texture, subsamples from the same composite sample were dried at 40°C and sieved to <2 mm. Total carbon and nitrogen were measured after grinding subsamples to a size of <100 μm in a ball mill. The ground samples were analyzed for total carbon and nitrogen by dry combustion with a Vario Max CN analyzer (Elementar Analysensysteme GmbH, Hanau, Germany). Inorganic carbon was quantified by measuring the total amount of carbon after the removal of organic carbon by the ignition of samples at 450°C for 16 h. To determine soil pH, the subsamples were suspended at a soil-to-liquid ratio of 1:2.5 (soil/0.01 M CaCl₂). Subsequently, pH was measured in the supernatant with a glass electrode. Soil texture was determined on 30 g soil according to a method described previously by Schlichting and Blume (45). The edaphic properties are depicted in Table 1.

Determination of microbial biomass. To determine microbial biomass, we performed phospholipid fatty acid analysis (PLFA) on soil samples from the A and B horizons of the sampling sites. The composite samples were kept frozen at −80°C after sampling and freeze-dried prior to PLFA extractions. PLFA extractions were performed by using a modified Bligh and Dyer (4) method. Briefly, 2 g

of freeze-dried sample was extracted twice in a chloroform-methanol-citrate buffer (1:2:0.8), followed by overnight phase separation. Fatty acids in the organic phase were then separated by using a silica-bonded phase column (silica-based solid-phase extraction [SPE-SI] Bond Elut, 3 ml, 500 mg; Varian Inc., Darmstadt, Germany) to remove glycolipids and neutral lipids. The polar lipids were then converted to fatty acid methyl esters by mild alkaline methanolysis. Methyl-esterified fatty acids were analyzed by using a Hewlett-Packard 6890 gas chromatograph equipped with a DB-5MS column (60-m length; Agilent Technologies, Böblingen, Germany) and interfaced with an Agilent 5973 mass selective detector. Peak areas of each lipid were converted to nmol/g soil using internal standards (19:0 nonadecanoic methyl ester). The total nmol lipid/g dry soil (sum of all lipids present, 20 or fewer carbons in length) was used as an index of microbial biomass (19, 25).

DNA extraction, amplification of 16S rRNA genes, and pyrosequencing. Total microbial community DNA was isolated from approximately 10 g of soil per sample. For this purpose, the MoBio Power Max soil DNA extraction kit (MoBio Laboratories, Carlsbad, CA) was used according to the manufacturer's instruction. To analyze the taxonomic composition of the soil bacterial community, the V2-V3 region of the 16S rRNA gene (*Escherichia coli* positions 101 to 536) was chosen for the amplification and subsequent pyrosequencing of the PCR products. The V2-V3 region was amplified with the following primer set, containing the Roche 454 pyrosequencing adaptors (underlined): V2for (5′-GCCTCCCTCGCGCCATCAGAGTGGCGGACGGGTGAGTAA-3′) (modified from that described previously by Schmalenberger et al. [48]) and V3rev (5′-GCCTTGCCAGCCCGCTCAGCGTATTACCGCGGCTGCTG-3′) (7).

For each sample, three independent PCRs were performed. The PCR mixture (final volume, 50 µl) contained 5 µl 10-fold reaction buffer (MBI Fermentas GmbH, St. Leon-Rot, Germany), 30 to 70 ng of soil DNA, 0.4 µM each primer, 0.5 U *Pfu* polymerase (MBI Fermentas), and 800 µM concentration of each of the four deoxynucleoside triphosphates. In some cases, to achieve amplification of 16S rRNA genes, a different DNA polymerase was used as recommended by the manufacturer (PCR Extender system; VWR International, Hannover, Germany). The polymerase was applied to samples derived from the A horizons of plots 2, 3, 4, and 6 and from the B horizons of plots 4, 6, and 8. Negative-control reactions lacked template DNA. The following thermal cycling scheme was used: initial denaturation at 94°C for 2 min and 25 cycles of denaturation at 94°C for 1 min, annealing using a temperature gradient ranging from 60.9°C to 68.2°C, and extension at 72°C for 1.25 min, followed by a final extension period at 72°C for 10 min. Subsequently, the three PCR products per soil sample were pooled in equal amounts and purified by employing the peqGOLD gel extraction kit (Peqlab Biotechnologie GmbH, Erlangen, Germany). Quantification of the PCR products was performed by using the Quant-iT dsDNA BR assay kit and a Qubit fluorometer (Invitrogen GmbH, Karlsruhe, Germany) as recommended by the manufacturer. The Göttingen Genomics Laboratory determined the sequences of the partial 16S rRNA genes by employing the Roche GS-FLX 454 pyrosequencer (Roche, Mannheim, Germany) and using picotiter sequencing plates subdivided into 8 parts (1 part per sample). Amplicons were sequenced as recommended in the instructions of the manufacturer for amplicon sequencing.

Analysis of pyrosequencing-derived data. Sequences that were shorter than 200 bp in length or of low quality were removed from the pyrosequencing-derived data sets. For taxonomy-based analysis, the Naïve Bayesian rRNA classifier of the Ribosomal Database Project (RDP) (55; http://rdp.cme.msu.edu/) was used. In this way, a rapid taxonomic classification of large partial and full-length rRNA gene sequence data sets according to the new Bergey's bacterial taxonomy (20) was feasible. The bootstrap value was set to ≥80%. Rarefaction curves were calculated by using the tools Aligner, Complete Linkage Clustering, and Rarefaction of the RDP pyrosequencing pipeline (11). We calculated Shannon (50) and Chao1 (8) indices based on the Complete Linkage Clustering data.

Statistical analyses of the pyrosequencing-derived data were carried out with STATISTICA 8.0 (StatSoft Inc., Tulsa, OK). P values of ≤0.05 were considered significant. Significant effects of soil horizon on edaphic soil properties were determined by using the Mann-Whitney U test for nonparametric data. Correlations between phylogenetic groups and soil properties were tested for significance by using Pearson's correlation coefficient. The distributions of phyla between the two soil horizons were calculated by employing the chi-square test. Microbial community compositions were compared to the land use types by using a post hoc Tukey honestly significant difference (HSD) analysis of variance. If the normality test revealed a non-Gaussian distribution, data were transformed.

Nucleotide sequence accession number. The 16S rRNA gene sequences derived from pyrosequencing have been deposited in the NCBI Sequence Read Archive under accession number SRA020168.1.

RESULTS AND DISCUSSION

General characteristics of the soil samples and the pyrosequencing-derived data set. In this study, we assessed and compared the compositions of soil bacterial communities present in the A and B horizons of nine different grassland sites of the Hainich region in Germany by a pyrosequencing-based analysis of the 16S rRNA gene sequences. The grassland sites covered a range of the following three different land use types: fertilized meadow (plots 1 to 3), fertilized mown pasture grazed by cattle (plots 4 to 6), and unfertilized pasture grazed by cattle (plots 7 to 9). The soil type of all samples was Stagnosol, except for plot 1, which was a Vertic Cambisol. In addition, further analysis of the Stagnosols revealed that plot 6 was a Luvic Stagnosol, whereas the other plots were Vertic Stagnosols (see Table S1 in the supplemental material).

The two analyzed soil horizons showed significant differences with respect to edaphic properties such as soil pH, OC content, N content, C-to-N ratio, and water content (Table 1). The pH in the A horizons ranged from 6.03 to 7.20, and the pH in the B horizons ranged from 6.30 to 7.40. In general, the pH value of the B horizon was higher than that of the corresponding A horizon, whereas the water content, the amounts of OC and N, and the C-to-N ratio showed 1.1- to 2.0-fold, 2.0- to 10.9-fold, 1.8- to 7.3-fold, and 1.2- to 1.5-fold decreases with depth, respectively.

The microbial biomass in the B horizons of all samples was lower by 48.7 to 96.3% than that in the corresponding A horizons (Table 1). A decrease in the total microbial biomass with soil depth was previously reported (5, 9, 16, 18). The total microbial biomass was positively correlated with the concentration of OC ($r = 0.88$; $P < 0.01$). This supports the assumption reported previously by Blume et al. (5), that carbon availability is closely associated with microbial biomass. In addition, significant correlations of microbial biomass with the concentration of N ($r = 0.84$; $P < 0.01$) and the C-to-N ratio ($r = 0.89$; $P < 0.01$) were detected.

The pyrosequencing-based analysis of the V2-V3 region of the 16S rRNA gene was employed for assessments of bacterial community compositions from the A and B horizons of the nine sampling sites. Short pyrosequencing reads assess the microbial diversity almost as reliably as near-full-length sequences when appropriate primers are chosen. Primers derived from V2-V3 region of the 16S rRNA gene were shown previously to be suitable for this purpose (36). In addition, this region is the most effective region for universal genus identification (7, 42). Across all 18 samples, we recovered 752,838 quality sequences with a read length of ≥200 bp. The average read length was 262 bp. The number of sequences per sample ranged from 25,851 to 61,366, with an average of 41,824 (see Table S2 in the supplemental material). We were able to classify 600,544 (79.77%) of the quality sequences below the domain level. The percentage of classified 16S rRNA gene sequences was in the range of those of other pyrosequencing-based studies (35), but the average number of sequences per sample and the total number of analyzed sequences exceeded those of other previously reported studies of pyrosequencing-based determinations of soil bacterial community composition (34, 43).

Bacterial richness and diversity indices. To determine rarefaction curves, richness, and diversity, we identified oper-
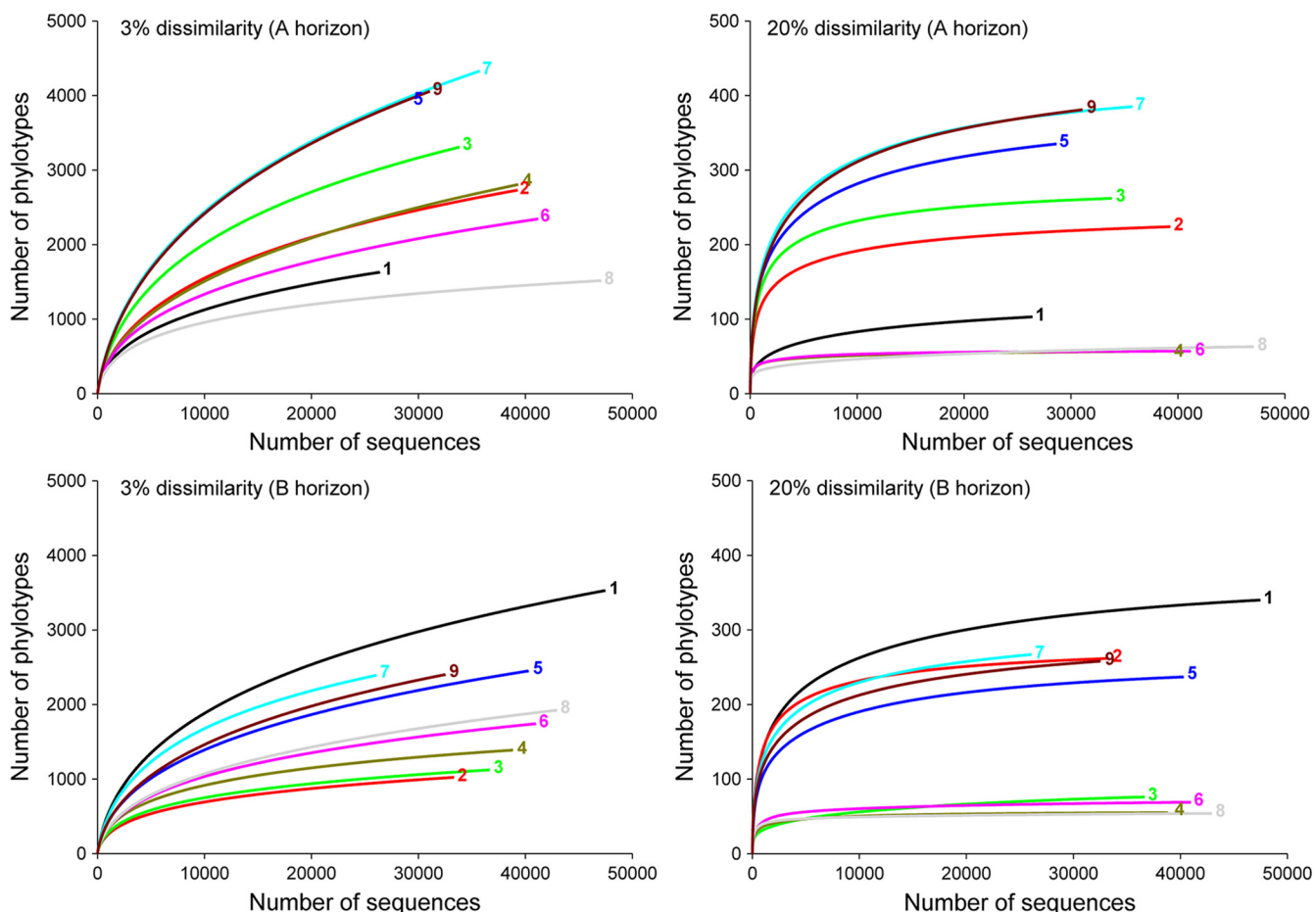
FIG. 1. Rarefaction curves indicating the observed number of OTUs within the 16S rRNA gene sequences of the A and B horizons, derived from nine German grassland sites. OTUs are shown at the 3 and 20% genetic distance levels. The rarefaction curves were calculated by employing the tools Aligner, Complete Linkage Clustering, and Rarefaction of the RDP pyrosequencing pipeline (11). The colored numbers mark the different sampling sites. A description of the sampling sites is given in Table 1.

ational taxonomic units (OTUs) at sequence divergences of 3% (species level) and 20% (phylum level). The rarefaction analysis of bacterial communities derived from the A and B horizons of the nine sampling sites is depicted in Fig. 1. At a 20% genetic distance, almost all curves showed saturation, indicating that the surveying effort covered almost the full extent of taxonomic diversity at this genetic distance. In addition, a comparison of rarefaction analyses with the number of OTUs estimated by the Chao1 richness estimator revealed that 83 to 100% of the estimated taxonomic richness was covered by the sequencing effort. At a 3% genetic distance, the observed richness was 63 to 80% of that predicted by the Chao1 richness estimator (Table 2). Thus, we did not survey the full extent of taxonomic diversity at the species level. Taking into account that at genetic distances below 5%, rarefaction analyses underestimate the bacterial richness whereas Chao1 estimators overestimate it (43), a substantial fraction of the bacterial diversity at the species level was assessed by the surveying effort. It is important that pyrosequencing provides an unprecedented sampling depth compared to that of traditional Sanger sequencing of 16S rRNA genes (51), but the intrinsic error of pyrosequencing could result in the overestimation of rare phy-

lotypes, since each pyrosequencing read is treated as a unique identifier of a community member and correction by assembly and sequencing depth, which is typically applied during genome projects, is not feasible (26, 32). To ensure per-base error rates lower than that of conventional Sanger sequencing, we used quality filtering of the pyrosequencing-derived data set, such as the removal of reads with atypical lengths (26). In addition, to minimize the overestimation of rare phylotypes, clustering and diversity estimates were performed only at genetic divergences of ≥3% (32).

Acosta-Martínez et al. (1) postulated previously that in managed soils, the maximum number of OTUs is less than 3,400 at a genetic distance of 3%. This is in contrast to our results, as up to 4,781 and 6,231 OTUs were predicted for fertilized meadows (plot 3, A horizon) and fertilized mown pasture grazed by cattle (plot 5, A horizon), respectively (Table 2). The differences in the results might be explained by the different surveying efforts. Several studies (13, 37, 43, 46, 56) showed that the number of analyzed sequences per sample has an effect on the predicted number of OTUs. For example, Roesch et al. (43) previously plotted the number of observed OTUs against the sequencing effort using the bacterial community present in a

TABLE 2. Species richness estimates obtained at genetic distances of 3% and 20%[c]

Horizon	Plot	Shannon index (H′)[a]		Rarefaction (no. of OTUs)		Chao1[b] (no. of OTUs)		Coverage (%)	
		3%	20%	3%	20%	3%	20%	3%	20%
A	1	5.92	2.67	1,629	103	2,335	124	69.8	83.1
A	2	6.11	4.03	2,730	224	4,084	236	66.9	94.8
A	3	6.76	4.27	3,307	262	4,781	272	69.2	96.2
A	4	6.30	2.93	2,805	57	4,395	59	63.8	96.6
A	5	7.07	4.49	3,937	335	6,231	366	63.2	91.6
A	6	6.15	2.91	2,344	57	3,551	57	66.0	100
A	7	7.16	4.50	4,329	385	6,487	407	66.7	94.7
A	8	5.65	2.51	1,516	63	1,924	68	78.8	93.3
A	9	7.05	4.41	4,056	381	6,232	438	65.1	87.0
B	1	6.72	4.34	3,528	340	5,168	360	68.3	94.5
B	2	5.01	2.64	1,022	84	1,399	94	73.0	89.6
B	3	5.14	2.50	1,122	76	1,509	82	74.4	92.6
B	4	5.57	2.72	1,388	55	1,745	55	79.5	99.4
B	5	6.15	3.98	2,450	237	3,635	252	67.4	94.0
B	6	5.64	2.83	1,741	69	2,420	73	72.0	94.8
B	7	6.51	4.15	2,392	267	3,293	293	72.6	91.2
B	8	5.57	2.88	1,923	54	2,854	57	67.4	94.7
B	9	6.09	4.10	2,402	258	3,606	280	66.6	92.1

[a] A higher number indicates more diversity.
[b] Nonparametric richness estimator based on the distribution of singletons and doubletons.
[c] The estimates were calculated by employing the tools Aligner, Complete Linkage Clustering, and Rarefaction of the RDP pyrosequencing pipeline (11). The results from the rarefaction analyses are also depicted in Fig. 1.

Canadian forest soil sample. The employment of the whole data set (53,632 sequences) revealed 5,500 OTUs at a genetic distance of 3%, whereas the reduction of the same data set to 30,000 sequences yielded 3,500 OTUs. In general, fewer sequences result in lower curve progression and a lower number of predicted OTUs. In addition, the comparison of richness estimates between different surveys might be hampered by the differences in sequence conservation and sequence length of the analyzed 16S rRNA gene regions. Recently, Engelbrektson et al. (15) showed that amplicon length and differences in the analyzed 16S rRNA gene regions markedly influenced estimates of richness and evenness.

The Shannon index of diversity (H′) was determined for all samples (Table 2). At a genetic distance of 3%, it ranged from 5.65 to 7.16 in the A horizons and from 5.01 to 6.72 in the B horizons. The predicted diversity in the topsoil exceeded that of the corresponding subsoil, except for plot 1 (Table 2). To our knowledge, no other study assessing bacterial diversity along a soil profile was conducted with a comparable surveying effort. However, a significant decrease of bacterial diversity with soil depth was also recorded by a community analysis employing terminal restriction fragment length polymorphism (33) and phospholipid fatty acid analysis (18) of soil profiles derived from Californian grassland and soil samples of the Sedgwick Reserve (California), respectively.

The bacterial diversity at a genetic distance of 3% was strongly related to the content of OC and N as well as to the C-to-N ratio. Positive correlations between the H′ and the OC content ($r = 0.60$; $P < 0.01$), the N content ($r = 0.58$; $P < 0.05$), and the C-to-N ratio ($r = 0.65$; $P < 0.01$) were observed. Similar correlations were detected by analyzing soil samples from South American grasslands, in which the H′ correlated positively with the microbial biomass C and N ($r = 0.53$ to 0.58; $P = 0.02$ to 0.03) (3). Interestingly, the mean H′ was lower in

fertilized meadows (plots 1 to 3), with intermediate values in fertilized mown meadows (plots 4 to 6) and the highest values in unfertilized pastures (plots 7 to 9), over both horizons (Table 2). Thus, a higher bacterial diversity in samples from unfertilized plots, which represent the lowest land use intensity in this study, is indicated. Nevertheless, within the same land use types, strong variations in diversity were observed.

Distribution of taxa and phylotypes across all samples. The 600,544 classifiable sequences were affiliated with 23 phyla across the entire data set. The dominant phyla across all samples were *Acidobacteria*, *Betaproteobacteria*, *Actinobacteria*, *Gammaproteobacteria*, *Alphaproteobacteria*, *Deltaproteobacteria*, *Chloroflexi*, *Firmicutes*, and *Bacteroidetes*, representing 26.98, 15.76, 11.62, 11.10, 9.69, 5.09, 3.85, 3.22, and 1.45%, respectively, of all sequences that were classified below the domain level (Fig. 2 and see Table S3 in the supplemental material). In addition, the dominant phyla were present in all samples. These results are in accordance with results from a previously reported meta-analysis of bacterial community composition in soils (29). In addition, the abundances of the five dominant phyla, *Acidobacteria*, *Actinobacteria*, *Alphaproteobacteria*, *Betaproteobacteria*, and *Gammaproteobacteria*, which represented approximately 75% of all classified sequences, corresponded roughly to those found by other studies (21, 34). Thus, despite the different surveying efforts and sampling sites used in the different studies, it is indicated that a variety of soils contain the same dominant bacterial groups.

The most abundant phylotypes in the A and B horizons were a member of the *Alphaproteobacteria* (*Acetobacteraceae*) and a member of acidobacterial subgroup 4, respectively. The corresponding sequences represented 0.74 and 2.86% of all classified sequences in each respective horizon. The most abundant phylotype within one individual sample (plot 3, B horizon) was
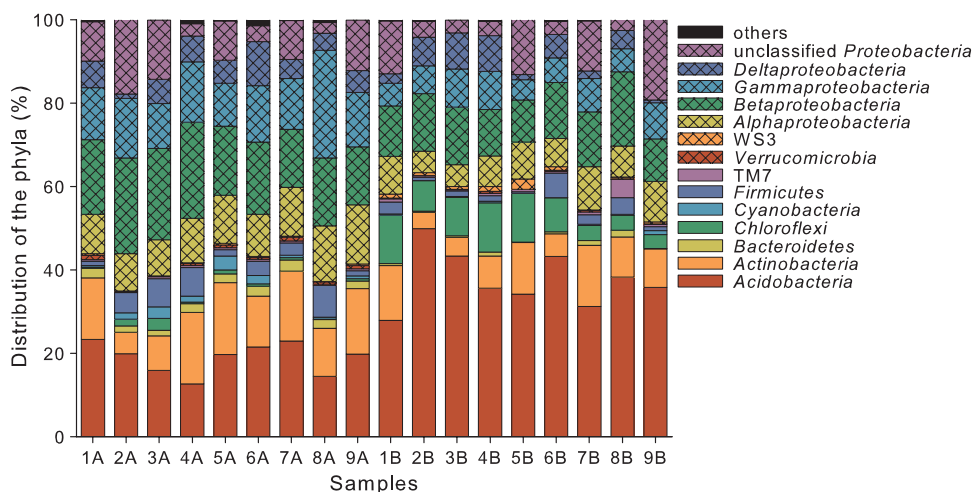
FIG. 2. Distribution of phylogenetic groups in the A and B horizons derived from the different grassland sampling sites. Plot numbers are given below the graph. A description of the plots is given in Table 1. A and B indicate the different horizons. Shown are the percentages of the classified sequences. Phylogenetic groups accounting for ≤0.25% of the classified sequences are summarized in the artificial group "others."

the above-mentioned member of acidobacterial subgroup 4, representing 8.36% of the sequences from that soil.

Distribution and abundance of the predominant phylum *Acidobacteria*. Members of the phylum *Acidobacteria* were predominant across all samples. This finding is in accordance with findings of other studies of the composition of soil-derived bacterial communities from a variety of environments, such as pristine forest, grassland, and agricultural soils (29). Here, members of the *Acidobacteria* form a significant fraction (12.68 to 49.86%) of the 16S rRNA gene sequences in both horizons from all land use types (see Tables S3 and S4 in the supplemental material). Correspondingly, members of this phylum have been reported to constitute an average of 20% and a maximum of approximately 50% of bacterial communities derived from various soils (13). Thus, based on their abundance and their presence in various soil types, the *Acidobacteria* appear to play an important role in the ecosystem function of soils.

The phylum *Acidobacteria* is divided into 26 subgroups, but only little is known with respect to the physiological and metabolic capabilities of the different subgroups (2). We detected 18 and 22 of these subgroups in the A and B horizons, respectively. Most abundant in both horizons were subgroups 6, 4, 16, and 7 (see Tables S4 and S5 in the supplemental material). In the A horizons, these subgroups were represented by 47, 21, 13, and 7% of all acidobacterial sequences, respectively, and in the B horizons, these subgroups were represented by 26, 38, 5, and 14% of all acidobacterial sequences, respectively (see Table S5 in the supplemental material). These results are in contrast to a previous study by Hansel et al. (21) of samples from a continuous watershed soil profile (Oak Ridge), which is the only other report of acidobacterial diversity with respect to soil horizon. In the A horizon, those researchers detected primarily subgroups 3 (21%), 4 (29%), and 6 (29%), whereas in our samples, these subgroups were represented by 5, 21, and 47%, respectively. In the B horizon, Hansel et al. (21) detected primarily subgroups 1 (32%) and 2 (61%), which were represented by less than 1% of all acidobacterial sequences derived

from our soil samples. The predominant subgroups in the B horizons from the Hainich region were subgroups 4, 5, and 7. The major differences in the occurrences of acidobacterial subgroups in the B horizon might be due to the dissimilar pH values of the samples used in both studies. The pH in our subsoil samples ranged from 6.30 to 7.40 (Table 1) whereas the pH of the soil samples studied by Hansel et al. (21) was 4.5. It was reported previously that the abundance of the phylum *Acidobacteria* correlates with the soil pH (22, 30). Lauber et al. (34) showed previously that acidobacterial subgroups 1 and 2 were most abundant in acidic soils and decreased with the increase of the pH. Here, no significant correlations of changes in the abundance of the dominant acidobacterial subgroups and other phylogenetic groups with pH were observed. A reason for this finding might be that the sampling effort in most of the other studies was much less than that of this study. Another possibility is that almost all the pH values of our samples were near neutral. Correspondingly, a relatively small pH range was covered by our soil samples (Table 1), so there is simply a lower pH range from which to determine correlations. Nevertheless, we observed negative correlations between the abundant acidobacterial subgroup 4 and the OC content ($r = -0.84$; $P < 0.01$), N content ($r = -0.83$; $P < 0.01$), or C-to-N ratio ($r = -0.77$; $P < 0.05$) in the A horizons (see Fig. S1 in the supplemental material). In the B horizons, subgroup 4 also correlated negatively with the C-to-N ratio ($r = -0.70$; $P < 0.05$), whereas the relative abundance of subgroup 6 showed a positive correlation with the C-to-N ratio ($r = 0.70$; $P < 0.05$). Thus, the subgroup distribution varied with respect to the soil profile (horizon) and soil properties and provided some insights into the conditions that are required by the different subgroups. For example, low nutrient/OC conditions (B horizons) appear to favor subgroups 4 and 7, whereas higher nutrient/OC conditions (A horizons) favor subgroup 16. Interestingly, for members of subdivision 6, a high tolerance to nutrient/OC availability was indicated, as they constituted a substantial fraction in the A horizon and the B horizon (8.87 and 9.73% of all classified sequences, respectively).
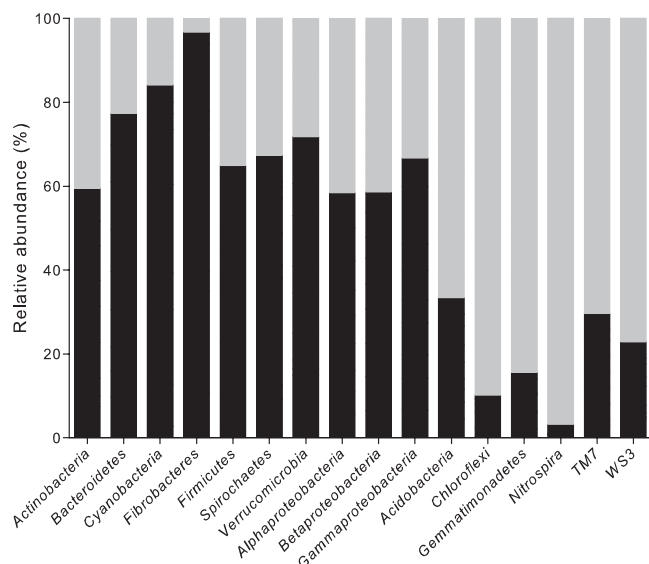
FIG. 3. Comparison of the overall distribution of selected phyla within the A and B horizons. The black bar represents the sum of all members of a phylum in the A horizon, while the gray bar represents the sum of all members in the B horizon. Box-and-whisker plots of the data are depicted in Fig. S2 in the supplemental material.

Taxonomic compositions in A horizons and the corresponding B horizons. The compositions of the bacterial community and the distributions of the phyla varied between A and B soil horizons. The most frequently present phyla in the A horizons were *Acidobacteria* (13 to 23%), *Betaproteobacteria* (14 to 23%), *Gammaproteobacteria* (10 to 26%), *Actinobacteria* (5 to 17%), and *Alphaproteobacteria* (9 to 14%). The most abundant phyla in the B horizons were *Acidobacteria* (28 to 50%), *Betaproteobacteria* (10 to 18%), *Actinobacteria* (4 to 15%), *Chloroflexi* (3 to 12%), and *Alphaproteobacteria* (5 to 10%) (Fig. 2).

We analyzed the respective abundances of the 15 most represented phyla in the A and B horizons. For almost all phyla and land use types, a significant ($P < 0.00001$) difference in distribution between the two horizons was apparent (see Table S6 in the supplemental material). The distribution of selected phyla in the two horizons is shown in Fig. 3 (see also Fig. S2 in the supplemental material). The majority of sequences affiliated with the *Actinobacteria*, *Bacteroidetes*, *Cyanobacteria*, *Fibrobacteres*, *Firmicutes*, *Spirochaetes*, *Verrucomicrobia*, *Alphaproteobacteria*, *Betaproteobacteria*, and *Gammaproteobacteria* were derived from A horizons, whereas the majority of the sequences related to *Acidobacteria*, *Chloroflexi*, *Gemmatimonadetes*, *Nitrospira*, TM7, and WS3 originated from B horizons. In many other reports, a pH gradient was identified as a major factor for changes in soil community structure, but as mentioned above, this trend was not observed in this study. Therefore, other factors appear to control the distribution of the phyla along the soil profile. As stated above, the total biomass decreased with soil depth (Table 1). A significant correlation between the total microbial biomass and the occurrence of several phyla was recorded. The relative abundance of the *Acidobacteria*, *Chloroflexi*, and *Nitrospira*, which increased with depth, correlated negatively with total biomass ($r = -0.53$ to

-0.79; $P < 0.05$). The relative abundances of the *Actinobacteria*, *Bacteroidetes*, *Verrucomicrobia*, *Alphaproteobacteria*, and *Gammaproteobacteria*, which decreased with depth, showed a positive correlation with biomass ($r = 0.69$ to 0.84; $P < 0.01$). A positive correlation of some of the dominant acidobacterial groups (see above) and the alphaproteobacterial order *Rhizobiales* with the concentration of OC ($r = 0.77$; $P < 0.05$), the concentration of N ($r = 0.73$; $P < 0.05$), and the C-to-N ratio ($r = 0.87$; $P < 0.01$) was detectable in the B horizon (see Fig. S3 in the supplemental material). A statistically significant positive correlation of the *Rhizobiales* with the C-to-N ratio ($r = 0.80$; $P < 0.01$) was also detected in the A horizon. Taking into account that the *Rhizobiales* include the genera *Rhizobium* and *Bradyrhizobium*, which comprise members that are able to fix nitrogen and are associated with roots of legumes, a positive correlation was expected (49). In addition, land use and management regimens seem to have an impact on the *Rhizobiales*, as the fertilized plots cluster and the plots with cattle cluster (data not shown). An impact of fertilization on the structure and diversity of rhizobial populations was observed previously in other studies (6, 41). For example, rhizobial populations differed between cultivated and uncultivated Mexican soils. In addition, the affinity of host cultivars for different members of the *Rhizobiales* influenced the composition of rhizobial populations (54).

Striking is the distribution of the phyla *Fibrobacteres* and *Nitrospira*, which occurred almost exclusively in the A or the B horizon, respectively. Members of the *Fibrobacteres* are part of the microbial community in the first stomach of ruminant animals and degrade plant-based cellulose (31). Taking into account that members of the *Fibrobacteres* hardly occurred in grassland samples without the presence of cattle (i.e., plots 1 to 3) and almost exclusively in the topsoil, it can be assumed that members of this phylum were introduced into the samples by cattle. Members of the *Nitrospira* are found in interspace soils and rarely in the rhizosphere (13). In the latter environment, heterotrophic root-associated microorganisms suppress the growth of autotrophic *Nitrospira*. This might explain why in our samples, members of the *Nitrospira* occurred in the subsoil, with less rooting than in the topsoil. In addition, the concentrations of OC and N decreased with soil depth (Table 1). Thus, chemolithoautotrophic organisms adapted to darkness, like *Nitrospira*, have a selective advantage in subsoil samples.

Conclusions. Although we recovered an average of 41,824 sequences per sample, we did not survey the full extent of bacterial richness at the species level within an individual soil or horizon. Thus, an increase in surveying efforts would probably result in the identification of more bacterial taxa, which are present in a low abundance. In most cases, the B horizons showed a lower estimated bacterial diversity than the corresponding A horizons (Table 2). Correspondingly, a greater coverage of the bacterial community in the B horizons can be achieved by using the same surveying effort. In addition, the identification of bacterial taxa at the finest level of taxonomic resolution is currently not feasible by applying large-scale pyrosequencing. However, the advancement of the technology will result in an increase of the read length, and this limitation will become less relevant in the near future.

To provide a robust assessment of the impact of land use, soil factors, or soil depth on bacterial diversity, distribution,

and community composition, we used triplicate samples of each land use type from different locations. Overall, the abundance, composition, and diversity of the bacterial communities were strongly depth dependent. The Shannon index of diversity along with the nutrient content (N and OC), water content, and biomass decreased with depth (Tables 1 and 2). Based on the sharp decrease of the OC content (up to 10.9-fold) in the B horizons compared to the corresponding A horizons (Table 1) and other surveys (18, 33, 57), the concentration of OC appears to be the major driver for the diversity and structure of bacterial communities along the soil profile at near-neutral pH values. Nevertheless, we observed a variability of bacterial communities within an individual land use type, and exceptions to the above-mentioned general results were found; i.e., a slightly higher Shannon index was recorded for the B horizon of plot 1. Thus, it is advisable to survey as many soil samples as possible for the identification of general patterns and comparison of the results with those of other soil surveys. One caveat of the latter, however, is the limited comparability of different surveys, as sampling strategy, survey effort, number and type of soil factors measured, and approaches used to analyze the sequence data vary considerably (37). To take full advantage of the increasing number of data sets on soil bacterial communities, minimal requirements for sampling and the set of analyzed soil factors as well as rules for sequence analysis and phylogenetic assignment should be defined.

ACKNOWLEDGMENTS

REFERENCES

1. **Acosta-Martínez, V., S. Dowd, Y. Sun, and V. Allen.** 2008. Tag-encoded pyrosequencing analysis of bacterial diversity in a single soil type as affected by management and land use. Soil Biol. Biochem. **40:**2762–2770.
2. **Barns, S., E. Cain, L. Sommerville, and C. Kuske.** 2007. *Acidobacteria* phylum sequences in uranium-contaminated subsurface sediments greatly expand the known diversity within the phylum. Appl. Environ. Microbiol. **73:**3113–3116.
3. **Berthrong, S., C. Schadt, G. Piñeiro, and R. Jackson.** 2009. Afforestation alters the composition of functional genes in soil and biogeochemical processes in South American grasslands. Appl. Environ. Microbiol. **75:**6240–6248.
4. **Bligh, E., and W. Dyer.** 1959. A rapid method of total lipid extraction and purification. Can. J. Biochem. Physiol. **37:**911–917.
5. **Blume, E., M. Bischoff, J. M. Reichert, T. Moorman, A. Konopka, and R. F. Turco.** 2002. Surface and subsurface microbial biomass, community structure and metabolic activity as a function of soil depth and season. Appl. Soil Ecol. **20:**171–181.
6. **Caballero-Mellado, J., and E. Martínez-Romero.** 1999. Soil fertilization limits the genetic diversity of *Rhizobium* in bean nodules. Symbiosis **26:**111–121.
7. **Chakravorty, S., D. Helb, M. Burday, N. Connell, and D. Alland.** 2007. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. J. Microbiol. Methods **69:**330–339.
8. **Chao, A., and J. Bunge.** 2002. Estimating the number of species in a stochastic abundance model. Biometrics **58:**531–539.
9. **Cho, S., S. Tsai, A. Ravindran, A. Selvam, and S. Yang.** 2008. Seasonal variation of microbial populations and biomass in Tatachia grassland soils of Taiwan. Environ. Geochem. Health **30:**255–272.
10. **Chow, M., C. Radomski, J. McDermott, J. Davies, and P. Axelrood.** 2002. Molecular characterization of bacterial diversity in Lodgepole pine (*Pinus contorta*) rhizosphere soils from British Columbia forest soils differing in disturbance and geographic source. FEMS Microbiol. Ecol. **42:**347–357.
11. **Cole, J., Q. Wang, E. Cardenas, J. Fish, B. Chai, R. Farris, A. Kulam-Syed-Mohideen, D. McGarrell, T. Marsh, G. Garrity, and J. Tiedje.** 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res. **37:**D141–D145.
12. **Daniel, R.** 2005. The metagenomics of soil. Nat. Rev. Microbiol. **3:**470–478.
13. **Dunbar, J., S. Barns, L. Ticknor, and C. Kuske.** 2002. Empirical and theoretical bacterial diversity in four Arizona soils. Appl. Environ. Microbiol. **68:**3035–3045.
14. **Elshahed, M., N. Youssef, A. Spain, C. Sheik, F. Najar, L. Sukharnikov, B. Roe, J. Davis, P. Schloss, V. Bailey, and L. Krumholz.** 2008. Novelty and uniqueness patterns of rare members of the soil biosphere. Appl. Environ. Microbiol. **74:**5422–5428.
15. **Engelbrektson, A., V. Kunin, K. Wrighton, N. Zvenigorodsky, F. Chen, H. Ochman, and P. Hugenholtz.** 2010. Experimental factors affecting PCR-based estimates of microbial species richness and evenness. ISME J. **4:**642–647.
16. **Federle, T., D. Dobbins, J. Thornton-Manning, and D. Jones.** 1986. Microbial biomass, activity, and community structure in subsurface soils. Ground Water **24:**365–374.
17. **Fierer, N., and R. Jackson.** 2006. The diversity and biogeography of soil bacterial communities. Proc. Natl. Acad. Sci. U. S. A. **103:**626–631.
18. **Fierer, N., J. P. Schimel, and P. A. Holden.** 2003. Variations in microbial community composition through two soil depth profiles. Soil Biol. Biochem. **35:**167–176.
19. **Frostegård, A., and E. Bååth.** 1996. The use of fatty acid analysis to estimate bacterial and fungal biomass in soil. Biol. Fertil. Soils **22:**59–65.
20. **Garrity, G. M., J. A. Bell, and T. G. Lilburn.** 2004. Taxonomic outline of the prokaryotes, p. 4–23. *In* G. M. Garrity (ed.), Bergey's manual of systematic bacteriology, 2nd ed., vol. 2. Springer-Verlag, New York, NY.
21. **Hansel, C., S. Fendorf, P. Jardine, and C. Francis.** 2008. Changes in bacterial and archaeal community structure and functional diversity along a geochemically variable soil profile. Appl. Environ. Microbiol. **74:**1620–1633.
22. **Hartman, W., C. Richardson, R. Vilgalys, and G. Bruland.** 2008. Environmental and anthropogenic controls over bacterial communities in wetland soils. Proc. Natl. Acad. Sci. U. S. A. **105:**17842–17847.
23. **Hartmann, M., and F. Widmer.** 2006. Community structure analyses are more sensitive to differences in soil bacterial communities than anonymous diversity indices. Appl. Environ. Microbiol. **72:**7804–7812.
24. **He, J., Z. Xu, and J. Hughes.** 2006. Molecular bacterial diversity of a forest soil under residue management regimes in subtropical Australia. FEMS Microbiol. Ecol. **55:**38–47.
25. **Hill, T., E. McPherson, J. Harris, and P. Birch.** 1993. Microbial biomass estimated by phospholipid phosphate in soils with diverse microbial communities. Soil Biol. Biochem. **25:**1779–1786.
26. **Huse, S., J. Huber, H. Morrison, M. Sogin, and D. Welch.** 2007. Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol. **8:**R143.
27. **IUSS Working Group.** 2006. World reference base for soil resources 2006, 2nd ed. World soil resources reports no. 103. FAO, Rome, Italy.
28. **Jahn, R., H.-P. Blume, V. B. Asio, O. Spaargaren, and P. Schad.** 2006. Guidelines for soil description, 4th ed. FAO, Rome, Italy.
29. **Janssen, P.** 2006. Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes. Appl. Environ. Microbiol. **72:**1719–1728.
30. **Jones, R., M. Robeson, C. Lauber, M. Hamady, R. Knight, and N. Fierer.** 2009. A comprehensive survey of soil acidobacterial diversity using pyrosequencing and clone library analyses. ISME J. **3:**442–453.
31. **Kobayashi, Y., T. Shinkai, and S. Koike.** 2008. Ecological and physiological characterization shows that *Fibrobacter succinogenes* is important in rumen fiber digestion—review. Folia Microbiol. (Praha) **53:**195–200.
32. **Kunin, V., A. Engelbrektson, H. Ochman, and P. Hugenholtz.** 2010. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. Environ. Microbiol. **12:**118–123.
33. **LaMontagne, M., J. Schimel, and P. Holden.** 2003. Comparison of subsurface and surface soil bacterial communities in California grassland as assessed by terminal restriction fragment length polymorphisms of PCR-amplified 16S rRNA genes. Microb. Ecol. **46:**216–227.
34. **Lauber, C., M. Hamady, R. Knight, and N. Fierer.** 2009. Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. Appl. Environ. Microbiol. **75:**5111–5120.
35. **Lazarevic, V., K. Whiteson, S. Huse, D. Hernandez, L. Farinelli, M. Østerås, J. Schrenzel, and P. François.** 2009. Metagenomic study of the oral microbiota by Illumina high-throughput sequencing. J. Microbiol. Methods **79:**266–271.
36. **Liu, Z., C. Lozupone, M. Hamady, F. Bushman, and R. Knight.** 2007. Short pyrosequencing reads suffice for accurate microbial community analysis. Nucleic Acids Res. **35:**120–130.
37. **Morales, S., T. Cosart, J. Johnson, and W. Holben.** 2009. Extensive phylogenetic analysis of a soil bacterial community illustrates extreme taxon evenness and the effects of amplicon length, degree of coverage, and DNA

fractionation on classification and ecological parameters. Appl. Environ. Microbiol. **75:**668–675.

38. **Muyzer, G., E. de Waal, and A. Uitterlinden.** 1993. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. Appl. Environ. Microbiol. **59:**695–700.

39. **Narang, R., and J. Dunbar.** 2004. Modeling bacterial species abundance from small community surveys. Microb. Ecol. **47:**396–406.

40. **Osborn, A., E. Moore, and K. Timmis.** 2000. An evaluation of terminal-restriction fragment length polymorphism (T-RFLP) analysis for the study of microbial community structure and dynamics. Environ. Microbiol. **2:**39–50.

41. **Palmer, K., and J. Young.** 2000. Higher diversity of *Rhizobium leguminosarum* biovar viciae populations in arable soils than in grass soils. Appl. Environ. Microbiol. **66:**2445–2450.

42. **Petrosino, J., S. Highlander, R. Luna, R. Gibbs, and J. Versalovic.** 2009. Metagenomic pyrosequencing and microbial identification. Clin. Chem. **55:**856–866.

43. **Roesch, L., R. Fulthorpe, A. Riva, G. Casella, A. Hadwin, A. Kent, S. Daroub, F. Camargo, W. Farmerie, and E. Triplett.** 2007. Pyrosequencing enumerates and contrasts soil microbial diversity. ISME J. **1:**283–290.

44. **Rosselló-Mora, R., and R. Amann.** 2001. The species concept for prokaryotes. FEMS Microbiol. Rev. **25:**39–67.

45. **Schlichting, E., and H.-P. Blume.** 1966. Bodenkundliches Praktikum: eine Einführung in pedologisches Arbeiten für Ökologen, insbesondere Land- und Forstwirte und für Geowissenschaftler. Paul Parey, Hamburg, Germany.

46. **Schloss, P., and J. Handelsman.** 2005. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Appl. Environ. Microbiol. **71:**1501–1506.

47. **Schloss, P., and J. Handelsman.** 2006. Toward a census of bacteria in soil. PLoS Comput. Biol. **2:**e92.

48. **Schmalenberger, A., F. Schwieger, and C. Tebbe.** 2001. Effect of primers hybridizing to different evolutionarily conserved regions of the small-subunit rRNA gene in PCR-based microbial community analyses and genetic profiling. Appl. Environ. Microbiol. **67:**3557–3563.

49. **Sessitsch, A., J. G. Howieson, X. Perret, H. Antoun, and E. Martínez-Romero.** 2002. Advances in *Rhizobium* research. Crit. Rev. Plant Sci. **21:**323–378.

50. **Shannon, C.** 1997. The mathematical theory of communication. 1963. MD Comput. **14:**306–317.

51. **Sogin, M., H. Morrison, J. Huber, D. Mark Welch, S. Huse, P. Neal, J. Arrieta, and G. Herndl.** 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere." Proc. Natl. Acad. Sci. U. S. A. **103:**12115–12120.

52. **Tiedje, J. M., S. Asuming-Brempong, K. Nüsslein, T. L. Marsh, and S. J. Flynn.** 1999. Opening the black box of soil microbial diversity. Appl. Soil Ecol. **13:**109–122.

53. **van der Heijden, M., R. Bardgett, and N. van Straalen.** 2008. The unseen majority: soil microbes as drivers of plant diversity and productivity in terrestrial ecosystems. Ecol. Lett. **11:**296–310.

54. **Wang, E., J. Martínez-Romero, and E. Martínez-Romero.** 1999. Genetic diversity of rhizobia from *Leucaena leucocephala* nodules in Mexican soils. Mol. Ecol. **8:**711–724.

55. **Wang, Q., G. Garrity, J. Tiedje, and J. Cole.** 2007. Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol. **73:**5261–5267.

56. **Youssef, N., and M. Elshahed.** 2008. Species richness in soil bacterial communities: a proposed approach to overcome sample size bias. J. Microbiol. Methods **75:**86–91.

57. **Zhou, J., B. Xia, D. Treves, L. Wu, T. Marsh, R. O'Neill, A. Palumbo, and J. Tiedje.** 2002. Spatial and resource factors influencing high microbial diversity in soil. Appl. Environ. Microbiol. **68:**326–334.

Chapter 3: Pyrosequencing-based Assessment of Bacterial Community Structure Along Different Management Types in German Forest and Grassland Soils

Heiko Nacke; Andrea Thürmer; Antje Wollherr; **Christiane Will**; Ladislav Hodac; Nadine Herold; Ingo Schöning; Marion Schrumpf; and Rolf Daniel

In this part of the thesis pyrosequencing was employed to analyze the microbial phylogenetic composition of different soil samples from the Schwäbische Alb. The focus was set on 18 soil samples which were all obtained from topsoil of nine forest and nine grassland sites. Significant differences in the microbial community compositions were detected between the two land use types. More precisely, the soil bacterial community structure varied between two different tree species. Moreover, strong correlations of several taxonomic groups with soil pH were observed.

Author contributions to the work:

Performed the experiments: HN, AT, CW.

Analyzed data: HN, AT, AW, LH.

Contributed data on soil properties and analysis of these data: IS, NH, MS.

Wrote the paper: HN, RD

Conceived and designed the experiments: RD.

PLoS one

Pyrosequencing-Based Assessment of Bacterial Community Structure Along Different Management Types in German Forest and Grassland Soils

Heiko Nacke[1], Andrea Thürmer[2], Antje Wollherr[2], Christiane Will[1], Ladislav Hodac[3], Nadine Herold[4], Ingo Schöning[5], Marion Schrumpf[4], Rolf Daniel[1,2]*

1 Department of Genomic and Applied Microbiology, Institute of Microbiology and Genetics, Georg-August University Göttingen, Göttingen, Germany, 2 Göttingen Genomics Laboratory, Institute of Microbiology and Genetics, Georg-August University Göttingen, Göttingen, Germany, 3 Experimental Phycology and Culture Collection of Algae at the University of Göttingen, Göttingen, Germany, 4 Max Planck Institute for Biogeochemistry, Jena, Germany, 5 Institute of Ecology, Friedrich-Schiller-University, Jena, Germany

Abstract

Background: Soil bacteria are important drivers for nearly all biogeochemical cycles in terrestrial ecosystems and participate in most nutrient transformations in soil. In contrast to the importance of soil bacteria for ecosystem functioning, we understand little how different management types affect the soil bacterial community composition.

Methodology/Principal Findings: We used pyrosequencing-based analysis of the V2-V3 16S rRNA gene region to identify changes in bacterial diversity and community structure in nine forest and nine grassland soils from the Schwäbische Alb that covered six different management types. The dataset comprised 598,962 sequences that were affiliated to the domain Bacteria. The number of classified sequences per sample ranged from 23,515 to 39,259. Bacterial diversity was more phylum rich in grassland soils than in forest soils. The dominant taxonomic groups across all samples (>1% of all sequences) were *Acidobacteria*, *Alphaproteobacteria*, *Actinobacteria*, *Betaproteobacteria*, *Deltaproteobacteria*, *Gammaproteobacteria*, and *Firmicutes*. Significant variations in relative abundances of bacterial phyla and proteobacterial classes, including *Actinobacteria*, *Firmicutes*, *Verrucomicrobia*, *Cyanobacteria*, *Gemmatimonadetes* and *Alphaproteobacteria,* between the land use types forest and grassland were observed. At the genus level, significant differences were also recorded for the dominant genera *Phenylobacter*, *Bacillus*, *Kribbella*, *Streptomyces*, *Agromyces*, and *Defluviicoccus*. In addition, soil bacterial community structure showed significant differences between beech and spruce forest soils. The relative abundances of bacterial groups at different taxonomic levels correlated with soil pH, but little or no relationships to management type and other soil properties were found.

Conclusions/Significance: Soil bacterial community composition and diversity of the six analyzed management types showed significant differences between the land use types grassland and forest. Furthermore, bacterial community structure was largely driven by tree species and soil pH.

Introduction

Soils are considered to be the most diverse microbial habitat on Earth with respect to species diversity and community size. Bacteria are the most abundant group of microorganisms in soil [1]. The calculated number of distinct bacterial genomes ranges from 2,000 to 18,000 per gram of soil [2]. Although the importance of bacteria for ecosystem functions and maintaining soil quality in agriculturally managed systems has long been recognized, the influence of land use type and management type on soil bacterial communities is poorly understood. In a recent pyrosequencing survey, bacterial diversity of forest soil was more phylum rich compared to agricultural soils, which were more species rich [3]. Furthermore, it has been described that

Bacteroidetes were more predominant in Pullman soil in agricultural systems than in the same soil under non-disturbed conditions, whereas the opposite trend was found for *Actinobacteria* [4]. It has been reported that land use indirectly affects the bacterial community structure by modification of soil properties [5]. Other studies also indicated that soil properties are important drivers of soil bacterial community structure [6], but soil pH appears to be a major factor influencing community composition [7]. This influence of soil pH has been recognized at coarse levels of taxonomic resolution [8], but also within individual phyla [9]. In addition, it has been shown that the type of plant species [10], soil type [11], soil texture [12], and nitrogen availability [13] can affect bacterial community structure. Tree species influences on soil bacterial communities are indicated by previous studies [14], but

detailed information on the affected bacterial groups and degree of these influences is still lacking.

In most previous studies the effects of land use and soil properties on soil bacterial communities have been assessed by employing traditional molecular methods such as Sanger sequencing-based analysis of 16S rRNA gene libraries or fingerprinting methods [15]. These approaches are often limited to the analysis of a relatively small number of clones and a few different soil samples. Taking into account the large bacterial community size and the heterogeneity of soils, only a tiny fraction of the bacterial diversity was unraveled by these studies. Recently, high-throughput pyrosequencing of 16S rRNA gene fragments has been applied for in-depth analysis of soil bacterial communities [3,4]. However, most of the available pyrosequencing studies do not allow a statistical assessment of land use and management effects on soil bacterial communities, as analyses of replicates were often not performed.

In this report, we applied pyrosequencing of the V2-V3 16S rRNA gene region to analyze bacterial community structure in A horizons of forest and grassland sites, which varied in management type. A horizons are mineral soil horizons formed at the surface or below an O horizon, which is dominated by organic material consisting of undecomposed or partially decomposed litter. A horizons are often characterized by accumulation of humidified organic matter [16]. It has been shown that analysis of the V2-V3 region provides a taxonomic resolution ranging from the phylum level to the genus level [17]. Thus, it is possible to detect variations in bacterial communities at different taxonomic levels. We analyzed 18 different soil samples derived from the Schwäbische Alb, which is one of the three German Biodiversity Exploratories [18]. Schwäbische Alb is a mosaic of forest and grasslands with a higher proportion of grassland. This is due to traditional sheep herding. We determined soil bacterial community structure in A

horizons of 9 forest and 9 grassland sites. The selected grassland and forest sites covered a range of 6 different management types. Triplicates of the different management types were analyzed, which is an important feature of this study, as it allows statistical analysis of management effects on soil bacterial communities. For each sample, the relative abundance and the distribution of bacterial groups were determined. Subsequently, we correlated variations in the relative abundances with land use type, management type, and soil properties.

Results and Discussion

General characteristics of the soil samples

In this study, we assessed and compared the composition of soil bacterial communities present in the A horizons of 18 soil samples derived from forest and grassland sites of the Schwäbische Alb (Germany) by large-scale pyrosequencing-based analysis of 16S rRNA gene sequences. The soil samples represented triplicates of 6 different management types, which encompassed spruce age class forest (SAF1-3), beech age class forest (BAF1-3), unmanaged beech forest (BF1-3), fertilized intensely managed grassland (FUG1-3), fertilized mown pasture grazed by horse and cattle (FMG1-3), and unfertilized pasture grazed by sheep (UPG1-3) (Tables 1 and S1). The soil groups of the forest soils and the grassland soils were Cambisols and Leptosols, respectively (Table 1). In addition, soil properties such as total nitrogen (N) content, organic carbon (OC) content, pH, and soil texture were determined. The soils had overall low sand (71 ± 64 g kg^{-1}) and highly variable clay contents with values ranging from 188 to 670 g kg^{-1} (average 412 g kg^{-1}). Similarly, OC contents showed a huge variability (68 ± 16 g kg^{-1}). Total N contents were on average lower in forest sites than in grassland sites and C/N ratios were accordingly higher (14 ± 1 forest and 11 ± 1 grassland)

Table 1. Physical and geochemical characteristics of the analyzed grassland and forest soil samples.

Management type	Sample	Soil group	pH	OC (g kg^{-1})	Total N (g kg^{-1})	C:N ratio	Gravimetric water content (%)	Particle size (g kg^{-1})		
								Sand	Silt	Clay
Spruce age class forest	SAF1	Cambisol	3.30	64.57	3.97	16.26	62.8	26	668	306
Spruce age class forest	SAF2	Cambisol	4.55	65.19	4.35	14.99	65.2	43	446	511
Spruce age class forest	SAF3	Cambisol	5.04	74.68	5.14	14.53	76.5	60	445	495
Beech age class forest	BAF1	Cambisol	6.38	78.50	6.01	13.06	75.1	70	534	396
Beech age class forest	BAF2	Cambisol	4.52	57.53	4.45	12.93	70.4	47	587	368
Beech age class forest	BAF3	Cambisol	5.36	39.05	3.15	12.40	50.8	107	575	318
Unmanaged beech forest	BF1	Cambisol	4.87	77.62	5.54	14.01	75.7	109	371	520
Unmanaged beech forest	BF2	Cambisol	5.10	105.00	6.77	15.51	96.6	34	296	670
Unmanaged beech forest	BF3	Cambisol	6.37	60.03	4.49	13.37	54.9	56	495	449
Fertilized intensely managed grassland	FUG1	Leptosol	6.71	77.09	7.58	10.17	66.2	38	543	419
Fertilized intensely managed grassland	FUG2	Leptosol	6.92	72.25	7.18	10.06	59.6	139	646	215
Fertilized intensely managed grassland	FUG3	Leptosol	6.32	53.74	5.18	10.37	57.2	25	449	526
Fertilized mown pasture, horse and cattle	FMG1	Leptosol	5.11	51.61	5.35	9.65	57.5	80	475	445
Fertilized mown pasture, horse and cattle	FMG2	Leptosol	6.36	85.16	7.87	10.82	76.4	56	694	250
Fertilized mown pasture, horse and cattle	FMG3	Leptosol	6.14	68.17	6.67	10.22	64.0	32	492	476
Unfertilized pasture, sheep	UPG1	Leptosol	7.24	40.85	3.65	11.19	46.7	282	530	188
Unfertilized pasture, sheep	UPG2	Leptosol	6.45	81.15	7.41	10.95	74.3	18	384	598
Unfertilized pasture, sheep	UPG3	Leptosol	6.65	68.89	5.82	11.84	67.6	44	684	272

(Table 1). The forest samples showed lower pH values than the grassland soils, which were all, except FMG1, near neutral. The analysis of differences of soil properties and management types by employing one-way analysis of variance and Tukey pair-wise comparisons showed that the analyzed management types did not vary significantly in OC, total N, and soil texture (Table S2). The only significant difference between management types was observed for the pH values, which were higher in unfertilized pastures grazed by sheep (6.9±0.4) than in spruce age class forests (4.7±0.9).

General analyses of the pyrosequencing-derived dataset

Profiling of pylogenetic diversity and community composition by large-scale pyrosequencing of 16S rRNA gene sequences provides more sequence information compared to traditional Sanger sequencing of 16S rRNA gene clone libraries [19]. Although the per-base error rate of pyrosequencing of 16S rRNA genes is not higher than that of Sanger sequencing, the intrinsic error rate of pyrosequencing might lead to overestimation of the number of rare phylotypes. Since each pyrosequencing read is treated as an unique identifier of a community member and correction by assembly and sequencing depth applied during genome projects is not feasible, errors can result in overestimation of diversity [20,21]. To minimize the overestimation of rare phylotypes, we used quality filtering of the pyrosequencing-derived dataset, and clustering and diversity estimates were performed at genetic divergences of ≥3% [21]. Alpha diversity analysis was performed at the same level of surveying effort (22,000 sequences per sample). In addition, denoising of each sequence subset was performed to avoid overestimation of operational taxonomic units (OTUs) and diversity [22,23]. The pyrosequencing-based analysis of the V2-V3 region of the 16S rRNA genes resulted in recovery of 599,284 high quality sequences with a read length of ≥200 bp across all 18 samples. The average read length was 255 bp. The number of sequences per sample ranged from 23,519 to 39,273 with an average of 33,275 (Table S1). We were able to assign 598,962 sequences to the domain Bacteria and to classify 474,868 (79.3%) of these sequences below the domain level. Taking into account the number of sequences per sample and the number of analyzed sequences, the size of this study exceeded other published studies on pyrosequencing-based determination of soil bacterial community composition [3,4,7].

Bacterial diversity and richness

To determine rarefaction curves, richness, and diversity, OTUs were identified at genetic distances of 3, 5, and 20% by using 22,000 randomly selected and denoised sequences per sample. At 20% sequence divergence most rarefaction curves reached saturation, indicating that the surveying effort covered almost the full extent of taxonomic diversity at this genetic distance (Figure S1). Comparison of the rarefaction analyses with the number of OTUs determined by Chao1 and ACE richness estimators revealed that 50.0 to 100% (20% genetic distance) of the estimated taxonomic richness was covered by the surveying effort (Table S3). At 3 and 5% genetic distance, the rarefaction curves were not saturated and the richness estimators indicated that 35.5 to 89.3% and 38.9 to 84.8% of the estimated richness, respectively, were recovered by the sequencing effort (Figures 1, 2 and S1, and Table S3). Thus, we did not survey the full extent of taxonomic diversity at these genetic distances, but a substantial fraction of the bacterial diversity within individual soil samples was assessed at species and genus level by the surveying effort (Figure 1 and Table S3). The comparison of mean Chao1 richness estimates of all forest soils with all grassland soils showed similar values at

genetic distances of 3% (3,219 OTUs and 2,611 OTUs, respectively) and 5% (2,331 OTUs and 2,095 OTUs, respectively) but at a genetic distance of 20% (75 OTUs and 153 OTUs, respectively) the richness was higher in grassland (P<0.05). The analysis of differences of richness estimates at genetic distances of 3% and 20% and the six management types by employing one-way analysis of variance showed that the analyzed management types did not vary significantly in the predicted number of OTUs (P>0.05 in both cases). Comparing this result to previous studies is difficult, as the number of analyzed sequences per sample has an effect on the predicted number of OTUs. In addition, denoising of amplicon sequences was not performed in other studies employing soil-derived pyrosequencing datasets [3,24]. In our study, richness estimates at 3% sequence divergence were approximately 2-fold higher in non-denoised datasets than in the corresponding denoised datasets (data not shown). In addition, in most other studies far fewer 16S rRNA fragments derived from a few soil samples have been analyzed.

The Shannon index of diversity (H') was determined for all samples (Table S3). At a genetic distance of 3%, the Shannon index ranged from 4.96 to 5.92 in the grassland samples and from 4.74 to 5.99 in the forest samples. Comparison of the mean H' of the different management types revealed that the highest bacterial diversity at a genetic distance of 3% was found in unmanaged beech forest, followed by fertilized intensely managed grassland, fertilized mown pastures grazed by horse and cattle, beech age class forest, spruce age class forest, and unfertilized pastures grazed by sheep (Table S3). In forest soils, the sample with the lowest pH (SAF1; pH 3.3) showed the lowest predicted diversity of all forest samples at all analyzed genetic distances (Figures 1, 2 and S1, and Table S3). Similar results were obtained by Fierer and Jackson [25] but a peak of diversity in soils with near-neutral pH values (BAF1 and BF3) that has been found in other studies [7] was not recorded. The spruce forest samples SAF2 and SAF3 showed higher diversity and richness estimates at phylum level but lower richness estimates at species level than the beech forest samples (Figure 2 and Table S3). Thus, an influence of the tree species on bacterial diversity is indicated. In addition, the rarefaction curves and the H' values derived from beech age class forest soils and unmanaged beech forest soils were not separated at all analyzed genetic distances (Figures 1 and S1, and Table S3), indicating that harvesting type (age class forest or unmanaged forest) has a minor or no impact on overall bacterial diversity and richness.

In grassland soils, similar values for estimated bacterial richness were obtained for the three samples derived from fertilized mown pastures grazed by horse and cattle whereas the replicated samples from the other two management types showed strong variations in estimated bacterial richness (Figure 2 and Table S3). At a genetic distance of 3%, the highest average bacterial richness according to Chao1 richness estimator was predicted for fertilized intensely managed grassland (2,887 OTUs), followed by fertilized mown pastures grazed by horse and cattle (2,720 OTUs), and unfertilized pastures grazed by sheep (2,226 OTUs). Nevertheless, the soil sample UPG3 derived from an unfertilized pasture grazed by sheep showed the second highest OTU estimate of all grassland soils (3,413 OTUs). Thus, bacterial diversity showed strong variations within management types in grassland soils.

Distribution of taxa and phylotypes across all samples

The 474,868 sequences classified below domain level were affiliated to 17 bacterial phyla and 4 proteobacterial classes (Tables S4 and S5). The dominant phyla and proteobacterial classes across all samples were *Acidobacteria*, *Alphaproteobacteria*, *Actinobacteria*, *Betaproteobacteria*, *Deltaproteobacteria*, *Gammaproteobacteria*, and *Firmi-*
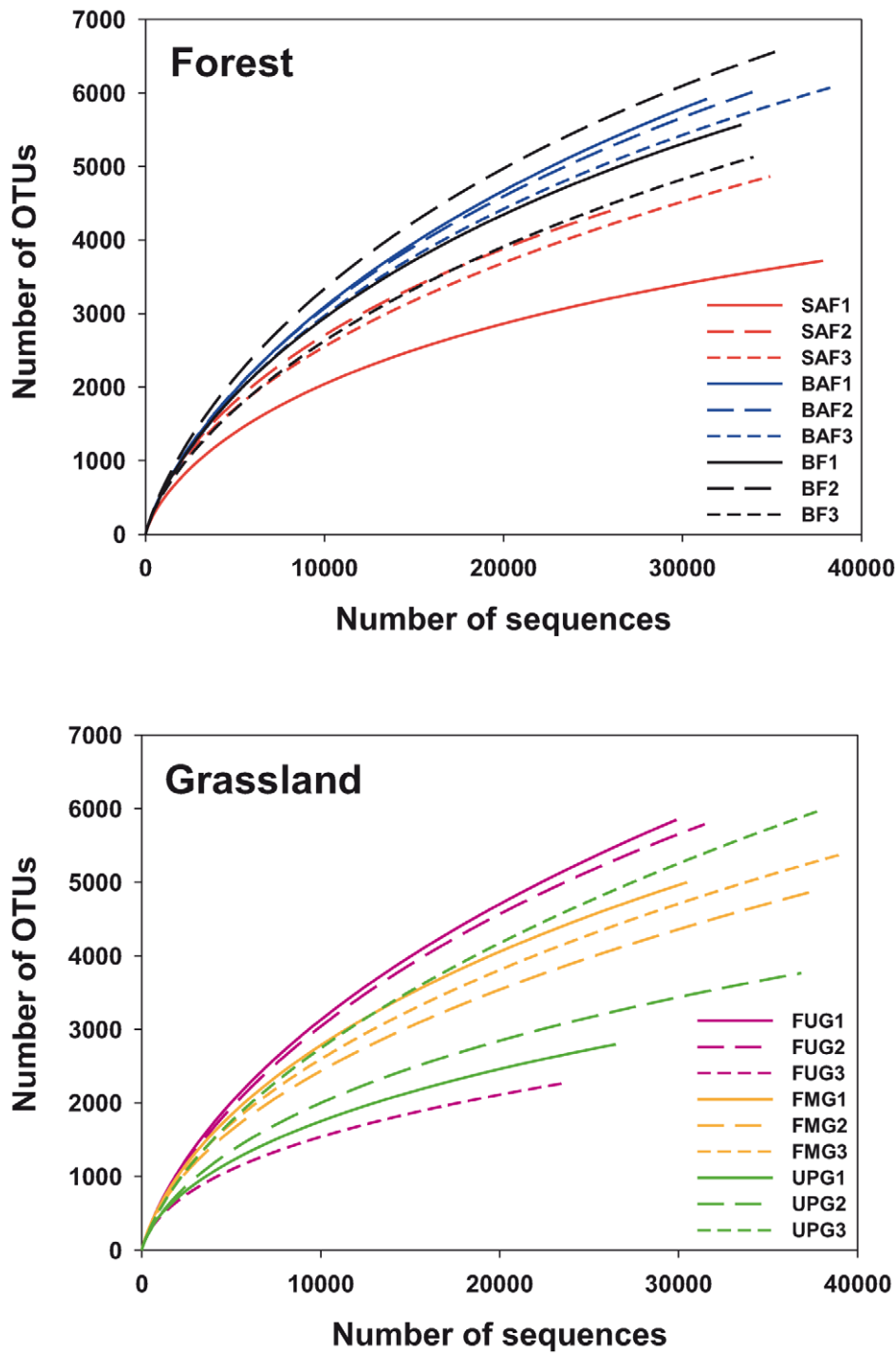
Figure 1. Rarefaction curves indicating the observed number of operational taxonomic units (OTUs) at a genetic distance of 3% in different forest and grassland soils. The spruce age class forest (SAF1-3), beech age class forest (BAF1-3), and unmanaged beech forest (BF1-3) sampling sites are marked by the red, blue, and black color, respectively. The fertilized intensely managed grassland (FUG1-3), fertilized mown pasture grazed by horse and cattle (FMG1-3), and unfertilized pasture grazed by sheep (UPG1-3) sampling sites are shown in purple, orange, and green, respectively.
doi:10.1371/journal.pone.0017000.g001

cutes, representing 19.6, 18.3, 16.1, 5.9, 3.4, 2.9, and 1.2%, respectively, of all sequences that were assigned to the domain Bacteria. The dominant taxa were present in all samples and corresponded roughly with those reported in other studies on soil bacterial community composition [26]. The members of rare phyla (<1% of all classified sequences) included WS3, *Bacteroidetes*,

TM7, *Chloroflexi*, *Verrucomicrobia*, *Cyanobacteria*, *Fibrobacteres*, *Spirochaetes*, *Gemmatimonadetes*, *Planctomyces*, OP11, *Deinococcus-Thermus*, and *Fusobacteria* (Figures 3 and 4, and Tables S4 and S5). The most abundant phylotype at a genetic distance of 3% across all samples was an unclassified member of the *Alphaproteobacteria*, representing 2.9% of all sequences. The most abundant phylotype at a genetic
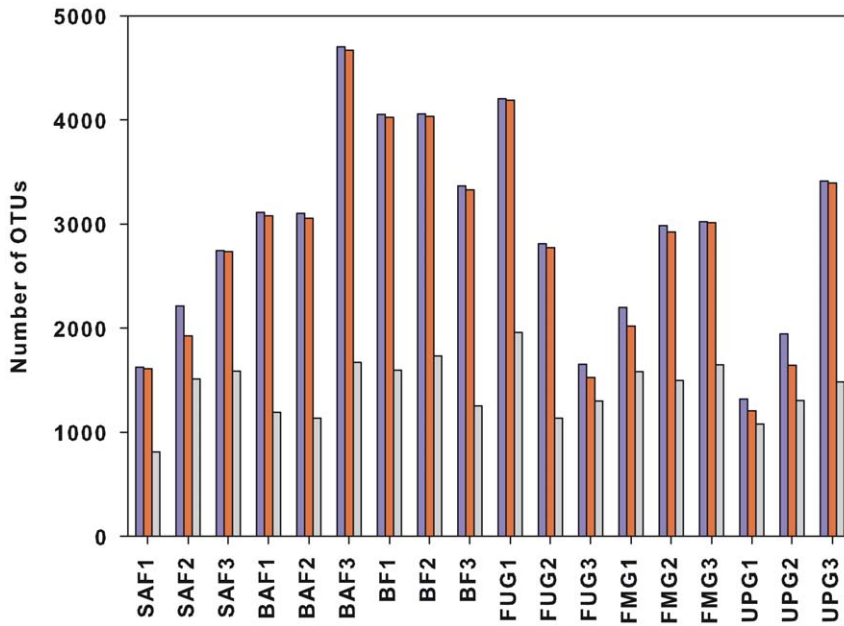
Figure 2. Bacterial richness estimates of German grassland and forest soils representing different management types at a genetic distance of 3%. Richness is expressed as number of observed unique OTUs. In addition, richness has been estimated by the abundance-based coverage estimator (ACE), which is a nonparametric richness estimator based on distribution of abundant (>10) and rare (≤10) OTUs, and the richness estimator Chao1, which is a nonparametric richness estimator based on distribution of singletons and doubletons. Richness prediction from Chao1 is colored in blue, richness prediction from ACE is colored in red, and richness observed is colored in grey. Sample numbers indicating the different management types are given below the graph. A description of the samples is shown in Table 1.
doi:10.1371/journal.pone.0017000.g002

distance of 3% within one individual forest soil sample (SAF1) was a member of the family *Caulobacteraceae*, representing 7.9% of the sequences from that soil. In grassland, an unclassified member of the *Proteobacteria* was the predominant phylotype (22.5% of all sequences) within an individual soil sample (UPG2).

Differences in community structure between forest and grassland soils

The relative abundances of dominant taxa varied between grassland and forest soils. The dominant taxa in forest soils were

Alphaproteobacteria (25.1±8.9%), *Acidobacteria* (20.4±3.0%), *Actinobacteria* (12.7±2.1%), and *Betaproteobacteria* (6.0±2.1%), whereas in grassland soils the predominant phylogenetic group was *Actinobacteria* (19.6±6.5%) followed by *Acidobacteria* (18.7±4.4%), *Alphaproteobacteria* (11.4±4.4%), and *Betaproteobacteria* (5.9%±1.2) (Figure 3, and Tables S4 and S5). The bacterial phyla and proteobacterial classes observed in our forest and grassland soils were also present in similar relative abundances in a meta-analysis of 32 bacterial 16S rRNA gene libraries derived from a variety of different soils, including samples from pristine forest, grassland and
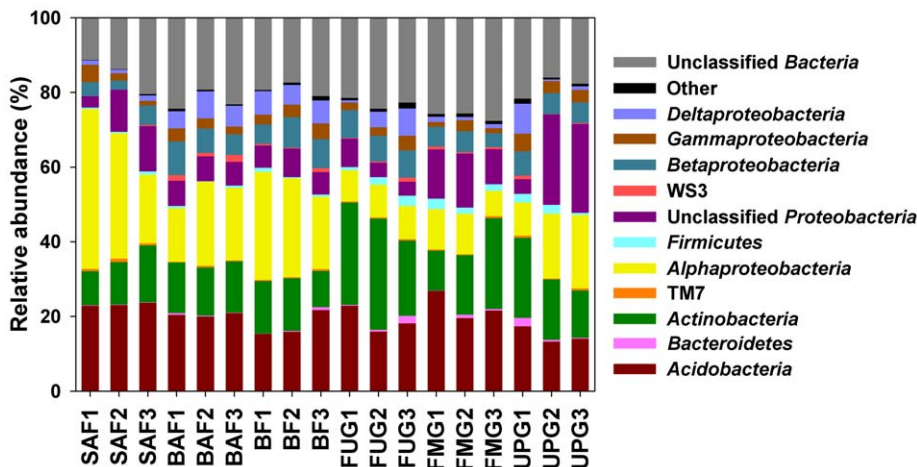
Figure 3. Relative abundances of phylogenetic groups in soils derived from the different grassland and forest sampling sites. Sample numbers indicating the different management types are given below the graph. A description of the samples is shown in Table 1. Phylogenetic groups accounting for ≤0.4% of all classified sequences are summarized in the artificial group 'others'.
doi:10.1371/journal.pone.0017000.g003
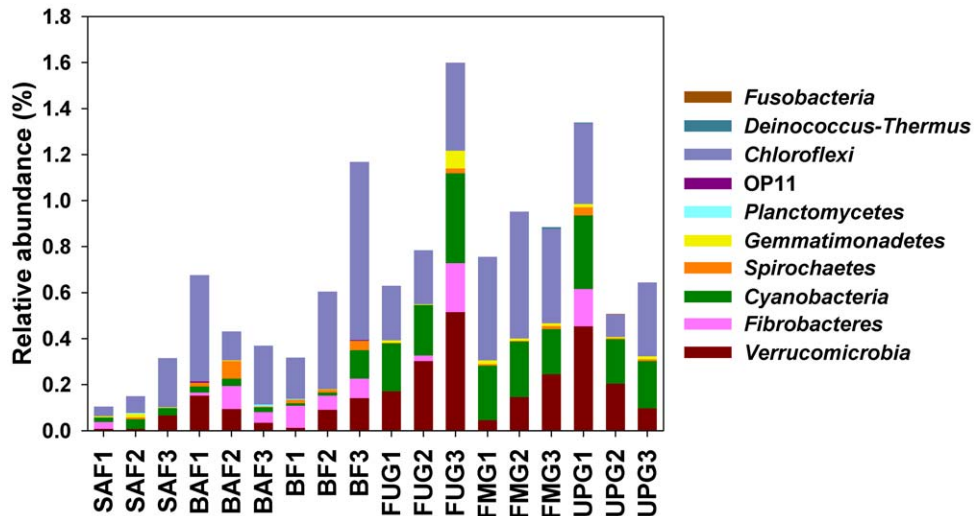
Figure 4. Relative abundances of rare phylogenetic groups of all sequences that were assigned to the domain Bacteria in soils derived from the different grassland and forest sampling sites. A description of the samples is shown in Table 1.

doi:10.1371/journal.pone.0017000.g004

agricultural soils [26]. Principal components analysis (PCA) based on the relative abundances of the different bacterial phyla and proteobacterial classes confirmed that the bacterial communities in grassland soils, except the one in sample UPG3, differed from communities in forest soils (Figure 5). We observed significant higher relative abundances of *Actinobacteria*, *Firmicutes*, *Verrucomicrobia*, *Cyanobacteria*, and *Gemmatimonadetes* in grassland soils than in forest soils whereas *Alphaproteobacteria* showed the opposite pattern ($P<0.05$ in all cases) (Figures 3 and 4). Thus, the shifts in soil bacterial community composition correlated with a change from forest to grassland. A similar trend was also found by comparison of Typic Placandept soils derived from a forest site and a pasture grazed by cattle [27]. In addition, sequences affiliated to *Alphaproteobacteria* dominated in 16S rRNA clone libraries of a spruce-fir-beech forest soil in Austria as well as in a Canadian boreal forest soil [14,28].

Differences of bacterial community structure between grassland and forest soils were also found in the phylogenetic structure within individual lineages. Members of the phylum *Acidobacteria* were predominant across all samples and the second most abundant group in forest and grassland soils, representing approximately 20% of all classified sequences. Correspondingly, members of this phylum have been reported to constitute an average of 20% in bacterial communities derived from various soil types [29]. Based on their abundance and the presence in various soil types, *Acidobacteria* appear to play an important role in ecosystem functions of soils, but little is known about physiology and metabolic functions of acidobacterial species. The phylum *Acidobacteria* is divided into 26 subgroups [30] with subgroups 1, 2, 3, 4, and 6 being most abundant within a variety of diverse soils [26,31]. Here, we detected 18 and 22 of these subgroups in grassland soils and forest soils, respectively. Most abundant in the grasslands soils were subgroups 16, 6, 4, 3, and 7, which represented 6.8, 4.4, 2.8, 1.8, and 1.4%, respectively, of all sequences that were classified in grassland. In forest soils, the dominant subgroups were 3, 16, 6, 1, and 4, representing 7.0, 3.0, 2.9, 2.9, and 2.1%, respectively, of all sequences that were classified (Tables S6 and S7).

Most of the sequences belonging to the second most abundant phylum *Alphaproteobacteria* across all samples were affiliated on the

order level to the *Rhodospirillales* in forest soils and to *Rhizobiales* in grassland soils. *Actinobacteridae* and *Rubrobacteridae* were the most abundant subclasses within the *Actinobacteria* in both land use types, but the actinobacterial subclass *Coriobacteridae* was only detected in grassland (Tables S8 and S9). Taking into account that members of this subclass are frequently found in gut or rumen samples [32,33] it is possible that they were introduced in the grassland sites by cattle or sheep.

At the genus level, comparison of the relative abundances revealed significant differences between grassland and forest soil bacterial communities. *Mycobacterium* was the most abundant genus across all soil samples, representing 3.7% of all classified sequences in forest soils and 5.7% in grassland soils. *Mycobacteria* are free-living saprophytes and well adapted to a variety of different environments including soils [34]. The distribution of the other dominant genera *Phenylobacter*, *Bacillus*, *Kribbella*, *Agromyces*, and *Defluviicoccus* varied significantly between forest and grassland soils ($P<0.05$). *Phenylobacter* showed a higher relative abundance in forest soils than in grassland soils whereas *Bacillus*, *Kribbella*, *Agromyces*, and *Defluviicoccus* showed the opposite pattern (Figure 6). *Rubrobacter* and *Streptomyces* were present in higher proportions in grassland soils compared to forest soils ($P<0.05$) (Figure 6). Consistently, Acosta-Martínez et al. [4] found *Rubrobacter* and *Streptomyces* among the top 20 predominant bacteria in two non-disturbed grass systems derived from Texas High Plains.

In summary, significant differences of the community structure between the two analyzed land use types forest and grassland were visible. Here, the different analyzed management types in grassland and forest were not reflected by significant changes in bacterial community structure. Thus, soils derived from an identical management type, i.e., UPG1 to UPG3 do not necessarily harbor similar bacterial communities. An exception was the significant impact of tree species (beech or spruce) on community structure in our forest soils. The comparison of relative abundances of bacterial phyla and proteobacterial classes with respect to tree species revealed significant differences between soils derived from spruce and beech forests (Figure 5). Based upon two sample t-test analyses, *Deltaproteobacteria* were less abundant in spruce forest than in beech forests ($P<0.05$) (Figure 3). At the genus level, *Methylocapsa* and *Burkholderia* were more abundant in
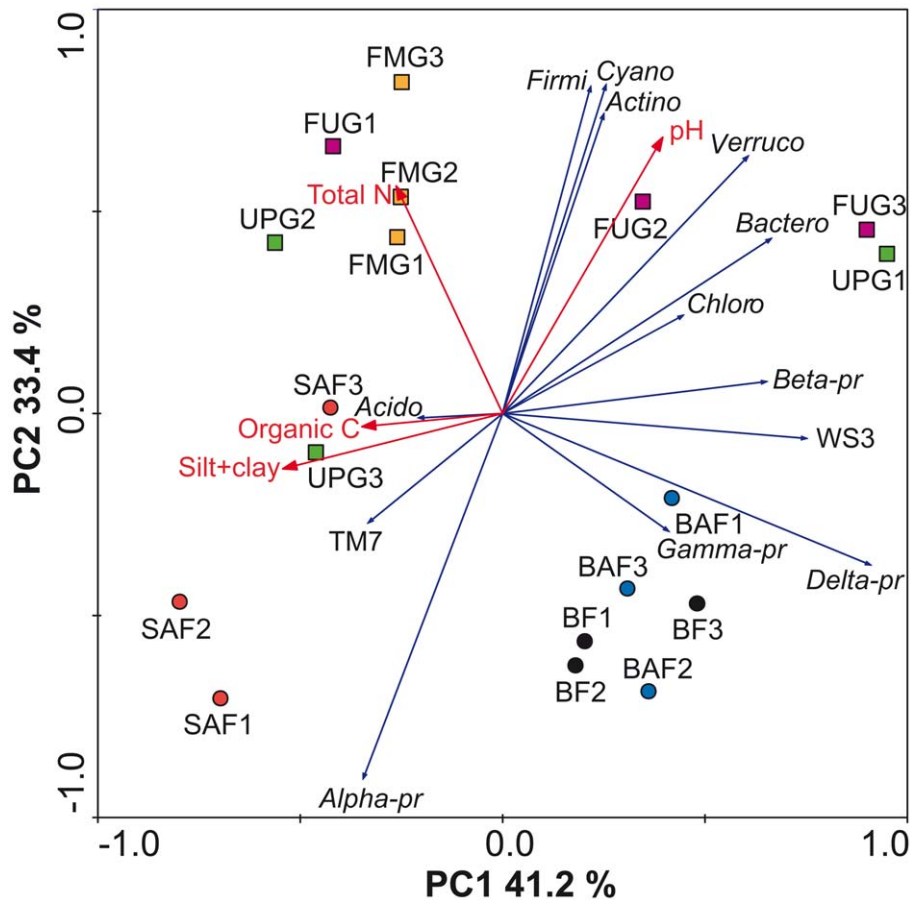
Figure 5. Principal components analysis of bacterial communities as affected by land use, based on the relative abundance of bacterial phyla and proteobacterial classes. Every vector points to the direction of increase for a given variable so that soil samples with similar bacterial communities are localized in similar positions in the diagram. The spruce age class forest (SAF1-3), beech age class forest (BAF1-3), and unmanaged beech forest (BF1-3) sampling sites are marked by the red, green, and black circles, respectively. The fertilized intensely managed grassland (FUG1-3), fertilized mown pasture grazed by horse and cattle (FMG1-3), and unfertilized pasture grazed by sheep (UPG1-3) sampling sites are depicted by red, green, and black squares, respectively. Abbreviations in figure: *Firmi, Firmicutes; Cyano, Cyanobacteria; Actino, Actinobacteria; Verruco, Verrucomicrobia; Bactero, Bacteroidetes; Chloro, Chloroflexi; Beta-pr, Betaproteobacteria; Delta-pr, Deltaproteobacteria; Gamma-pr, Gammaproteobacteria; Alpha-pr, Alphaproteobacteria; Acido, Acidobacteria.*
doi:10.1371/journal.pone.0017000.g005

spruce forest soil than in beech forest soil, whereas *Nocardioides*, *Leptothrix*, and *Amaricoccus* showed the opposite pattern (Figure 6). Thus, tree species appear to be an important driver of soil bacterial community structure, but the type of harvesting (age class forest or unmanaged forest) does not significantly affect bacterial community composition (Figure 5).

Impact of soil properties on the relative abundances of bacterial taxa

Previous studies indicated that soil properties such as pH value or soil texture are important drivers of bacterial community structure [12,35]. We used correlation analysis to identify relationships between the relative abundances of bacterial groups and soil properties. The relative abundances of bacterial groups at different taxonomic levels responded strongly to soil pH. This is in accordance to other surveys on soil bacterial communities derived from different management types in which pH-dependent changes in abundance and distribution of bacterial phyla were observed [36,37]. At the phylum level, relative abundances of *Bacteroidetes* and *Actinobacteria* in the analyzed soils significantly increased with higher pH values ($P<0.05$ in both cases) (Table 2).

As described for a freshwater lake [38] and diverse soils [9], we also found strong correlations of pH and relative abundances of bacterial groups below the phylum level. The relative abundances of the proteobacterial classes *Alphaproteobacteria* and *Betaproteobacteria* were significantly correlated to pH ($P<0.05$). The abundances of *Alphaproteobactia* were negatively correlated with soil pH, whereas the abundances of *Betaproteobacteria* increased with pH (Table 2). Within the *Alphaproteobacteria*, the relative abundances of the order *Caulobacterales* and the family *Acetobacteraceae* showed similar correlations to soil pH as the *Alphaproteobacteria* in general ($P<0.05$ in both cases) (Figure 7). This result corresponded to a cultivation-dependent study of Jimenez-Salgado et al. [39], in which more members of the *Acetobacteraceae* were isolated from low pH soils than from high pH soils. Although relative abundances of *Gammaproteobacteria* showed no significant correlation to soil pH at the class level, the relative abundances of the gammaproteobacterial genus *Dyella* significantly increased with lower pH values ($P<0.05$) (Figure 7). The genus *Dyella* has been recently described by Xie and Yokota [40]. So far, it includes seven species isolated from soil, but no growth of these isolates below pH 4.0 was described [41,42]. In contrast, the highest relative abundances for sequences affiliated to the genus *Dyella* (0.6% of all classified
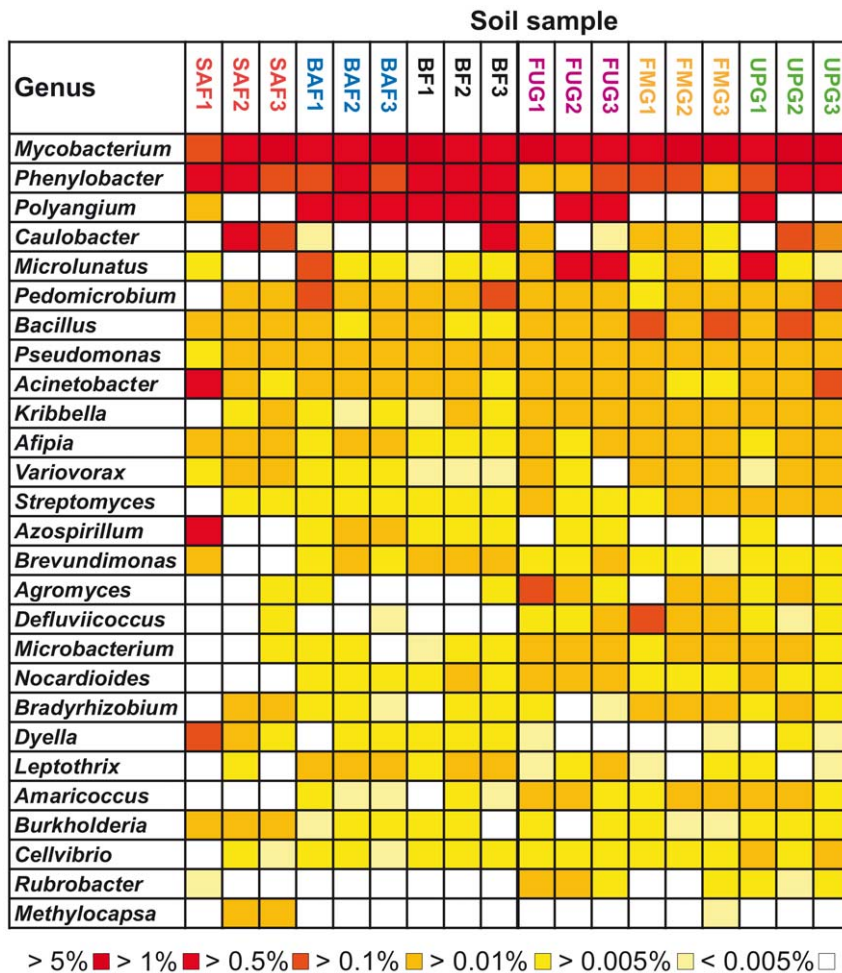
Figure 6. Relative abundances of the most abundant genera as affected by land use. Percentages below the map indicate the abundance of each genus relative to all bacterial sequences that were classified in each of the 18 soils. A description of the samples is shown in Table 1. Grassland and forest samples are separated by a bold line. Samples of different management types are colored in red (SAF1-3), blue (BAF1-3), black (BF1-3), purple (FUG1-3), orange (FMG1-3), and green (UPG1-3).
doi:10.1371/journal.pone.0017000.g006

sequences) were found in sample SAF1, which exhibited the lowest pH value of all samples (pH 3.3). Furthermore, we obtained the highest relative abundances for genera *Azospirillum* and *Acinetobacter* (each representing more than 0.5% of all classified *Bacteria*) in soil sample SAF1 (Figure 6). Thus, our results might help to identify conditions that are best suited for a targeted cultivation of members belonging to these genera.

The occurrence of several subgroups of the *Acidobacteria*, which were predominant across all samples, was also dependent on soil pH. The relative abundances of acidobacterial subgroups 1, 3, 6, 13, 17, and 18 showed strong significant correlations to soil pH ($P<0.001$ in all cases). The relative abundances of subgroups 1, 3, and 13 decreased with pH whereas those of subgroups 6, 17, and 18 were positively correlated with pH (Figure 7 and Table S10). Similar correlations of soil pH and the abundances of acidobacterial subgroups 1, 3, 6, 13, 17, and 18 have been reported by Jones et al. [9]. In addition, the inverse relationship of soil pH on the abundance of members affiliated to subgroup 1 has been reported for soils derived from rotationally grazed perennial ryegrass and white clover pasture [43].

In general, more groups at different taxonomic levels showed significant correlations to soil pH in forest soils than in grassland soils (data not shown). This might be due to the different pH range

covered by the analyzed forest and grassland soils. The pH in our forest samples ranged from pH 3.30 to 6.37 (Table 1) whereas the pH values of the grassland samples were all, except sample FMG1, near neutral. Thus, a relatively small pH range was covered by our grassland samples (Table 1), so there is simply less pH range from which to determine correlations. Significant correlations of relative abundances with other soil properties were found for *Deltaproteobacteria* and *Actinobacteria*. The *Deltaproteobacteria* showed a significant correlation to OC ($P<0.05$) with higher abundances in soils with low OC content, whereas *Actinobacteria* showed a significant correlation to total N ($P<0.05$) with higher abundances in soils with high total N content (Table 2), but a connection to the observed correlations was not evident.

Conclusion

The analysis of one of the largest bacterial 16S rRNA-based datasets from soils revealed statistically significant differences in soil bacterial diversity and community structure between the two land use types forest and grassland. Additionally, the occurrence of different tree species had statistically significant effects on soil bacterial diversity, richness, and community composition in forest. The analysis of influences of soil properties on bacterial community structure revealed that pH had the strongest effect

Table 2. Spearman's rank correlations between the relative abundances of the six most abundant bacterial phyla and proteobacterial classes and the soil properties in grassland and forest soils.

Taxonomic group	Correlation			
	pH	OC	Total N	Sand/Silt/Clay
Actinobacteria	**0.58**	0.26	**0.52**	0.02/−0.08/−0.02
Bacteroidetes	<u>**0.71**</u>	0.14	0.33	−0.08/0.17/−0.19
Alphaproteobacteria	**−0.68**	0.05	−0.44	−0.12/−0.13/0.22
Betaproteobacteria	**0.56**	0.22	0.35	0.04/0.04/0.00
Deltaproteobacteria	−0.10	**−0.48**	−0.55	0.43/−0.15/−0.04
Gammaproteobacteria	0.27	−0.04	−0.17	−0.13/0.19/−0.19

Correlations for *Acidobacteria* are shown at higher taxonomic resolution Table S10.
Bold numbers: $P < 0.05$; Bold and underlined numbers $P < 0.001$.
doi:10.1371/journal.pone.0017000.t002

on bacterial community structure of the analyzed soil properties. Management type and other soil properties appear to have a minor impact on soil bacterial community structure and diversity.

In this survey, the correlations between land use type and community composition were obvious. The relative abundances of a number of taxonomic groups changed significantly between forest and grassland soils (e.g., *Actinobacteria*), but the abundances of other taxa (e.g., *Gammaproteobacteria*) were almost unaffected by land use type, indicating that the abundances of the latter groups are influenced by other factors. Specific bacterial groups such as *Amaricoccus* or *Methylocapsa* showed significantly higher abundances in beech or spruce forest soils. Finally, we cannot determine whether pH has a direct or indirect effect on community composition, as a number of soil properties (e.g., OC) are directly or indirectly related to pH [44]. Thus, the effect of a number of different factors is reflected by soil pH and these factors may also drive community composition.

Availability

The 18 pyrosequencing-derived 16S rRNA gene sequence datasets have been deposited in the GenBank short-read archive under accession number SRA022075.

Materials and Methods

Site description, sampling, DNA extraction, and soil characterization

In the frame of the German Biodiversity Exploratories, initiative soil samples were collected from 9 forest and 9 grassland plots of the German Biodiversity Exploratory Schwäbische Alb. The Schwäbische Alb covers more than 450 km×450 km in the state of Baden-Württemberg (southwestern Germany). Soil samples were collected in April 2008. The forest sampling sites included 3 spruce age class forests (SAF1-3), 3 beech age class forests (BAF1-3), and 3 unmanaged beech forests (BF1-3). Grassland sampling sites comprised 3 fertilized intensely managed grasslands (FUG1-3), 3 fertilized mown pastures grazed by horse and cattle (FMG1-3), and 3 unfertilized pastures grazed by sheep (UPG1-3) (Table S1). The dominant grasses included *Poa trivialis*, *Trisetum flavescens*, and *Arrhenaterum elatius* in sites FUG1-3, *Poa trivialis*, *Alopecurus pratensis*, *Trisetum flavescens*, *Dactylis glomerata*, *Festuca pratensis*, *Lolium perenne*, and *Arrhenaterum elatius* in sites FMG1-3, and *Brachypodium pinnatum*, *Bromus erectus*, and *Festuca guestfalica* in sites UPG1-3. A

detailed description of the dominant grasses of the individual plots is provided in Table S11.

Soil samples were collected and classified at each of the grassland and forest sites as described by Will et al. [45]. Briefly, five soil cores (8.3 cm in diameter) were sampled with a motor driven soil column cylinder at each corner and in the center of each plot within a given area of 20 m×20 m. Composite samples of the five collected A horizons per plot were used for DNA extraction, after the soils were homogenized and coarse roots and stones (>5 mm) were removed. Total microbial community DNA was extracted from approximately 8 g soil derived from the A horizons of each plot by employing the MoBio PowerMax Soil DNA isolation kit (MoBio Laboratories, Carlsbad, CA, USA) as recommended by the manufacturer. DNA concentrations were quantified by using a NanoDrop ND-1000 UV-Vis Spectrophotometer (NanoDrop Technologies, USA) according to the manufacturer's protocol.

OC content, total N content, soil texture, and soil pH were measured as described by Will et al. [45]. To determine the gravimetric water content, 10 g of moist soil were dried to constant weight at 105°C for 24 h. The mass of water was calculated per mass of dry soil.

Amplification of 16S rRNA genes and pyrosequencing

The V2-V3 region of the 16S rRNA gene was amplified by PCR. The PCR reaction mixture (33 µl) contained 3.3 µl 10-fold reaction buffer (Fusion GC buffer, FINNZYMES, Espoo, Finland), 800 µM of each of the four deoxynucleoside triphosphates, 3% DMSO, 1.2 µM of each of the primers, 0.5 U of Phusion hot start high-fidelity DNA Polymerase (FINNZYMES), and 20 ng of isolated DNA as template. The V2-V3 region was amplified with the following set of primers containing the Roche 454 pyrosequencing adaptors (underlined): V2for 5′-<u>GCCTCCCTCGCGCCATC</u>-AGAGTGGCGGACGGGTGAGTAA-3′ and V3rev 5′-<u>GCC</u><u>TTGCCAGCCCGCTCAGC</u>GTATTACCGCGGCTGCTG-3′ (modified from Schmalenberger et al. [46]). The following thermal cycling scheme was used: initial denaturation at 98°C for 5 min, 25 cycles of denaturation at 98°C for 45 s, annealing at 68°C for 45 s, and extension at 72°C for 25 s followed by a final extension period at 72°C for 5 min. All samples were amplified in triplicate, pooled in equal amounts, and purified using the peqGold gel extraction kit as recommended by the manufacturer (Peqlab Biotechnologie GmbH, Erlangen, Germany). Quantification of the PCR products was performed using the Quant-iT dsDNA BR assay kit and a Qubit fluorometer (Invitrogen GmbH, Karlsruhe, Germany) as recommended by the manufacturer. The Göttingen Genomics Laboratory determined the sequences of the partial 16S rRNA genes by using a Roche GS-FLX 454 pyrosequencer (Roche, Mannheim, Germany) and the instructions of the manufacturer for amplicon sequencing.

Analysis of pyrosequencing data

Sequences that were shorter than 200 bp in length and reads containing any unresolved nucleotides were removed from the 18 pyrosequencing-derived datasets. For taxonomy-based analysis, the RDP Classifier of the Ribosomal Database Project (RDP) was used [47] at a confidence threshold of 80%. Pyrosequencing noise was removed for alpha diversity analyses by using the denoiser program [23]. For the determination of OTUs, we defined species, genus, and phylum level at 3, 5, and 20%, respectively, sequence divergence according to Schloss and Handelsman [48]. OTUs were determined for each denoised sequence dataset by using the uclust OTU picker version 1.2.21q of the QIIME software pipeline [49]. We calculated rarefaction curves as well as the
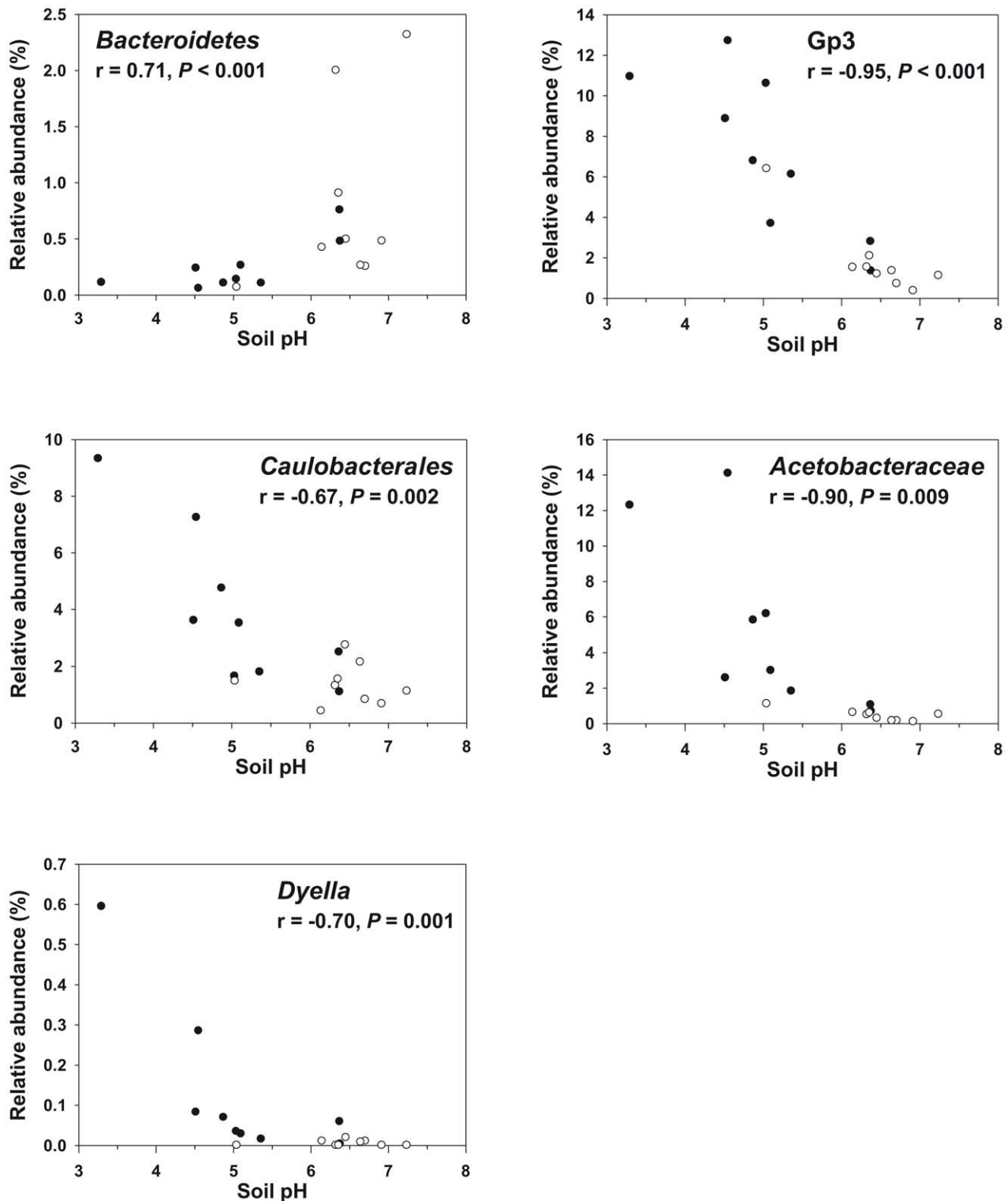
Figure 7. Correlations between relative abundances of different taxonomic groups and soil pH. Black circles represent forest sites and white circles represent grassland sites. Spearman's rank correlation coefficients (r) with the associated *P* values are shown for each taxonomic group. Abbreviation: Gp3, acidobacterial subgroup 3.
doi:10.1371/journal.pone.0017000.g007

Shannon [50] index based on OTU picker data, by employing the RDP pyrosequencing pipeline [51]. ACE and Chao1 [52] indices were calculated using the EstimateS program version 8.2.0 (http://purl.oclc.org/estimates).

Statistical analyses

Normality tests (Shapiro-Wilk) were performed with data that were used for principal component analysis (PCA), and one-way analysis of variance. Data that did not pass normality test were log transformed

and normality test was repeated. Only data that passed normality test were used for further analyses. For each soil attribute and each richness estimate at 3 and 20% genetic distance, one-way analysis of variance and Tukey pair-wise comparisons were used to determine the minimum significant difference ($P<0.05$) between management types by employing STATISTICA 8.0 (StatSoft, Inc., Tulsa, USA). To compare bacterial community structures across all samples based on the relative abundance of bacterial phyla and proteobacterial classes, PCA was performed by using CANOCO for Windows [53]. To correlate bacterial taxonomic groups with soil properties, Spearman's rank correlations were determined by using the SigmaPlot program version 11.0 (Systat Software, Inc., San Jose, CA). We used two sample t-test analyses and M-W-U-Test for non-parametric data to compare relative abundances of bacterial groups and richness estimates between grassland and forest, and on a second level between different management types using the software package PAST [54].

Supporting Information

Figure S1 Rarefaction curves indicating the observed number of OTUs at genetic distances of 5 and 20% in the different forest and grassland soils. The spruce age class forest (SAF1-3), beech age class forest (BAF1-3), and unmanaged beech forest (BF1-3) sampling sites are marked by the red, blue, and black color, respectively. The fertilized intensely managed grassland (FUG1-3), fertilized mown pasture grazed by horse and cattle (FMG1-3), and unfertilized pasture grazed by sheep (UPG1-3) sampling sites are shown in purple, orange, and green, respectively.
(DOC)

Table S1 Localization of the sampling sites and number of 16S rRNA gene sequences derived from the analyzed grassland and forest soil samples.
(DOC)

Table S2 Mean values of soil properties and standard deviation for each management type and ANOVA P values. Differences of soil properties between management types were analyzed by employing one-way analysis of variance and Tukey pair-wise comparisons. Significant ANOVA P values are shown in bold ($P<0.05$). Figures followed by different letters indicate differences among management types ($P<0.05$). Abbreviations: SAF, spruce age class forest; BAF, beech age class forest; BF, unmanaged beech forest; FUG, fertilized intensely managed grassland; FMG, fertilized mown pasture grazed by horse and cattle; UPG, unfertilized pasture grazed by sheep. Complete soil and site information for all 18 sampling sites is provided in Table 1.
(DOC)

Table S3 Bacterial diversity as assessed by Shannon index (H') and species richness estimation in all forest and grassland soils. The results from the rarefaction analyses are also depicted in Figure 1 and Figure S1.
(DOC)

Table S4 Relative abundances of bacterial phyla and proteobacterial classes in the analyzed forest soils. Values represent percentages of all sequences assigned to the domain Bacteria for all forest soils or individual forest soils. Groups labeled with asterisks could not be assigned to a specific phylum or a proteobacterial class.
(DOC)

Table S5 Relative abundances of bacterial phyla and proteobacterial classes in the analyzed grassland soils. Values represent percentages of all sequences assigned to the domain Bacteria for all grassland soils or individual grassland soils. Groups labeled with asterisks could not be assigned to a specific phylum or a proteobacterial class.
(DOC)

Table S6 Relative abundances of acidobacterial subgroups in the analyzed forest soils. Values represent percentages of all sequences assigned to the domain Bacteria for all forest soils or individual forest soils. Groups labeled with asterisks could be assigned to the phylum level only.
(DOC)

Table S7 Relative abundances of acidobacterial subgroups in the analyzed grassland soils. Values represent percentages of all sequences assigned to the domain Bacteria for all grassland soils or individual grassland soils. Groups labeled with asterisks could be assigned to the phylum level only.
(DOC)

Table S8 Relative abundances of taxonomic groups within the phylum *Actinobacteria* and within proteobacterial classes in the analyzed forest soils. Values represent percentages of all sequences assigned to the domain Bacteria for all forest soils or individual forest soils. Groups labeled with asterisks could be assigned to the phylum level only.
(DOC)

Table S9 Relative abundances of taxonomic groups within the phylum *Actinobacteria* and within proteobacterial classes in the analyzed grassland soils. Values represent percentages of all sequences assigned to the domain Bacteria for all grassland soils or individual grassland soils. Groups labeled with asterisks could be assigned to the phylum level only.
(DOC)

Table S10 Spearman's rank correlations between relative abundances of *Acidobacteria* subgroups and soil properties. Only relative abundances of acidobacterial subgroups that represented $\geq 0.029\%$ of all analyzed sequences were considered.
(DOC)

Table S11 Dominant grasses of the analyzed grassland sites.
(DOC)

Author Contributions

Conceived and designed the experiments: RD. Performed the experiments: HN AT CW. Analyzed the data: HN AT AW LH. Wrote the paper: HN RD. Contributed data on soil properties and analysis of these data: IS NH MS.

References

1. Gans J, Wolinsky M, Dunbar J (2005) Computational improvements reveal great bacterial diversity and high metal toxicity in soil. Science 309: 1387–1390.
2. Daniel R (2005) The metagenomics of soil. Nat Rev Microbiol 3: 470–478.
3. Roesch LFW, Fulthorpe RR, Riva A, Casella G, Hadwin AKM, et al. (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. ISME J 1: 283–290.

4. Acosta-Martínez V, Dowd S, Sun Y, Allen V (2008) Tag-encoded pyrosequencing analysis of bacterial diversity in a single soil type as affected by management and land use. Soil Biol Biochem 40: 2762–2770.

5. Jesus ED, Marsh TL, Tiedje JM, Moreira FMD (2009) Changes in land use alter the structure of bacterial communities in Western Amazon soils. ISME J 3: 1004–1011.

6. Wakelin SA, Macdonald LM, Rogers SL, Gregg AL, Bolger TP, et al. (2008) Habitat selective factors influencing the structural composition and functional capacity of microbial communities in agricultural soils. Soil Biol Biochem 40: 803–813.

7. Lauber CL, Hamady M, Knight R, Fierer N (2009) Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. Appl Environ Microbiol 75: 5111–5120.

8. Bååth E, Anderson TH (2003) Comparison of soil fungal/bacterial ratios in a pH gradient using physiological and PLFA-based techniques. Soil Biol Biochem 35: 955–963.

9. Jones RT, Robeson MS, Lauber CL, Hamady M, Knight R, et al. (2009) A comprehensive survey of soil acidobacterial diversity using pyrosequencing and clone library analyses. ISME J 3: 442–453.

10. Wieland G, Neumann R, Backhaus H (2001) Variation of microbial communities in soil, rhizosphere, and rhizoplane in response to crop species, soil type, and crop development. Appl Environ Microbiol 67: 5849–5854.

11. Girvan MS, Bullimore J, Pretty JN, Osborn AM, Ball AS (2003) Soil type is the primary determinant of the composition of the total and active bacterial communities in arable soils. Appl Environ Microbiol 69: 1800–1809.

12. Sessitsch A, Weilharter A, Gerzabek MH, Kirchmann H, Kandeler E (2001) Microbial population structures in soil particle size fractions of a long-term fertilizer field experiment. Appl Environ Microbiol 67: 4215–4224.

13. Frey SD, Knorr M, Parrent JL, Simpson RT (2004) Chronic nitrogen enrichment affects the structure and function of the soil microbial community in temperate hardwood and pine forests. Forest Ecol Manage 196: 159–171.

14. Hackl E, Zechmeister-Boltenstern S, Bodrossy L, Sessitsch A (2004) Comparison of diversities and compositions of bacterial populations inhabiting natural forest soils. Appl Environ Microbiol 70: 5057–5065.

15. McCaig AE, Glover LA, Prosser JI (2001) Numerical analysis of grassland bacterial community structure under different land management regimens by using 16S ribosomal DNA sequence data and denaturing gradient gel electrophoresis banding patterns. Appl Environ Microbiol 67: 4554–4559.

16. FAO (2006) Guidelines for soil description, 4th edn. Rome, Italy: FAO.

17. Chakravorty S, Helb D, Burday M, Connell N, Alland D (2007) A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. J Microbiol Meth 69: 330–339.

18. Fischer M, Bossdorf O, Gockel S, Hänsel F, Hemp A, et al. (2010) Implementing large-scale and long-term functional biodiversity research: The Biodiversity Exploratories. Basic Appl Ecol 11: 473–485.

19. Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, et al. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci USA 103: 12115–12120.

20. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM (2007) Accuracy and quality of massively parallel DNA pyrosequencing. Gen Biol 8: R143.

21. Kunin V, Engelbrektson A, Ochman H, Hugenholtz P (2010) Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. Environ Microbiol 12: 118–123.

22. Quince C, Lanzen A, Curtis TP, Davenport RJ, Hall N, et al. (2009) Noise and the accurate determination of microbial diversity from 454 pyrosequencing data. Nat Meth 6: 639–641.

23. Reeder J, Knight R (2010) Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions. Nat Meth 7: 668–669.

24. Morales S, Cosart T, Johnson J, Holben W (2009) Extensive phylogenetic analysis of a soil bacterial community illustrates extreme taxon evenness and the effects of amplicon length, degree of coverage, and DNA fractionation on classification and ecological parameters. Appl Environ Microbiol 75: 668–675.

25. Fierer N, Jackson RB (2006) The diversity and biogeography of soil bacterial communities. Proc Natl Acad Sci USA 103: 626–631.

26. Janssen PH (2006) Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes. Appl Environ Microbiol 72: 1719–1728.

27. Nüsslein K, Tiedje JM (1999) Soil bacterial community shift correlated with change from forest to pasture vegetation in a tropical soil. Appl Environ Microbiol 65: 3622–3626.

28. Dimitriu PA, Grayston SJ (2010) Relationship between soil properties and patterns of bacterial β-diversity across reclaimed and natural boreal forest soils. Microb Ecol 59: 563–573.

29. Dunbar J, Barns SM, Ticknor LO, Kuske CR (2002) Empirical and theoretical bacterial diversity in four Arizona soils. Appl Environ Microbiol 68: 3035–3045.

30. Barns SM, Cain EC, Sommerville L, Kuske CR (2007) Acidobacteria phylum sequences in uranium-contaminated subsurface sediments greatly expand the known diversity within the phylum. Appl Environ Microbiol 73: 3113–3116.

31. Barns SM, Takala SL, Kuske CR (1999) Wide distribution and diversity of members of the bacterial kingdom Acidobacterium in the environment. Appl Environ Microbiol 65: 1731–1737.

32. Anderson RC, Rasmussen MA, Jensen NS, Allison MJ (2000) Denitrobacterium detoxificans gen. nov., sp. nov., a ruminal bacterium that respires on nitrocompounds Int J Syst Evol Microbiol 50: 633–638.

33. Collado MC, Sanz Y (2007) Quantification of mucosa-adhered microbiota of lambs and calves by the use of culture methods and fluorescent in situ hybridization coupled with flow cytometry techniques. Vet Microbiol 121: 299–306.

34. Ventura M, Canchaya C, Tauch A, Chandra G, Fitzgerald GF, et al. (2007) Genomics of Actinobacteria: tracing the evolutionary history of an ancient phylum. Microbiol Mol Biol Rev 71: 495–548.

35. Rousk J, Bååth E, Brookes P, Lauber C, Lozupone C, et al. (2010) Soil bacterial and fungal communities across a pH gradient in an arable soil. ISME J 4: 1340–1351.

36. Lauber CL, Strickland MS, Bradford MA, Fierer N (2008) The influence of soil properties on the structure of bacterial and fungal communities across land-use types. Soil Biol Biochem 40: 2407–2415.

37. Hartman WH, Richardson CJ, Vilgalys R, Bruland GL (2008) Environmental and anthropogenic controls over bacterial communities in wetland soils. Proc Natl Acad Sci USA 105: 17842–17847.

38. Newton RJ, Jones SE, Helmus MR, McMahon KD (2007) Phylogenetic ecology of the freshwater Actinobacteria acI lineage. Appl Environ Microbiol 73: 7169–7176.

39. Jimenez-Salgado T, Fuentes-Ramirez LE, Tapia-Hernandez A, Mascarua-Esparza MA, Martinez-Romero E, et al. (1997) Coffea arabica L., a new host plant for Acetobacter diazotrophicus, and isolation of other nitrogen-fixing acetobacteria. Appl Environ Microbiol 63: 3676–3683.

40. Xie CH, Yokota A (2005) Dyella japonica gen. nov., sp. nov., a gamma-proteobacterium isolated from soil. Int J Syst Evol Microbiol 55: 753–756.

41. Weon HY, Anandham R, Kim BY, Hong SB, Jeon YA, et al. (2009) Dyella soli sp. nov. and Dyella terrae sp. nov., isolated from soil. Int J Syst Evol Microbiol 59: 1685–1690.

42. Anandham R, Kwon SW, Indira Gandhi P, Kim SJ, Weon HY, et al. (2010) Dyella thiooxydans sp. nov., a facultative chemolithotrophic thiosulfate oxidizing bacterium isolated from rhizosphere soil of sunflower (Helianthus annuus L.). Int J Syst Evol Microbiol: doi:10.1099/ijs.0.022012-0.

43. Sait M, Davis KER, Janssen PH (2006) Effect of pH on isolation and distribution of members of subdivision 1 of the phylum Acidobacteria occurring in soil. Appl Environ Microbiol 72: 1852–1857.

44. Brady NC, Weil RR (2007) The nature and properties of soil, 13th edn. Upper Saddle River, New Jersey, USA: Prentice Hall.

45. Will C, Thürmer A, Wollherr A, Nacke H, Herold N, et al. (2010) Horizon-specific bacterial community composition of German grassland soils, as revealed by pyrosequencing-based analysis of 16S rRNA genes. Appl Environ Microbiol 76: 6751–6759.

46. Schmalenberger A, Schwieger F, Tebbe C (2001) Effect of primers hybridizing to different evolutionarily conserved regions of the small-subunit rRNA gene in PCR-based microbial community analyses and genetic profiling. Appl Environ Microbiol 67: 3557–3563.

47. Wang Q, Garrity G, Tiedje J, Cole J (2007) Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 73: 5261–5267.

48. 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.

49. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, et al. (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Meth 7: 335–336.

50. Shannon CE, Weaver W (1949) The mathematical theory of communication. MD Comput 14: 306–317.

51. Cole JR, Wang Q, Cardenas E, Fish J, Chai B, et al. (2009) The ribosomal database project: improved alignments and new tools for rRNA analysis. Nucl Acids Res 37: D141–145.

52. Chao A, Bunge J (2002) Estimating the number of species in a stochastic abundance model. Biometrics 58: 531–539.

53. ter Braak CJF, Smilauer P (2002) CANOCO reference manual and CanoDraw for windows user's guide: software for canonical community Ordination (version 4.5). Ithaca, NY, USA: Microcomputer Power.

54. Hammer Ø, Harper DAT, Ryan PD (2001) PAST: Paleontological Statistics Software Package for Education and Data Analysis. Palaeontologia Electronica 4: 1–9.

Chapter 4: Identification of novel lipolytic genes and gene families by screening of metagenomic libraries derived from soil samples of the German Biodiversity Exploratories

Heiko Nacke*, **Christiane Will***, Sarah Herzog, Boris Nowka, Martin Engelhaupt, and Rolf Daniel

The aim of this part of the thesis was the detection and partial characterization of novel lipolytic enzymes from soil metagenomes. Soil has proven to be a rich source for novel biocatalysts which can be employed to various industrial applications. The screen of 12 Gb of environmental DNA (21 metagenomic libraries) resulted in the detection of 37 different enzymes, which exhibit esterase activity. The initial characterization revealed a preference for short-chain triacylglycerols of most of the lipolytic enzymes.

Author contributions to the work:

* H.N. and C.W. contributed equally to this work.

Performed the experiments: CW, HN, SH, BN, ME.

Analyzed data: CW, HN.

Wrote the paper: CW, HN, RD.

Conceived and designed the experiments: CW, HN, RD.

Identification of novel lipolytic genes and gene families by screening of metagenomic libraries derived from soil samples of the German Biodiversity Exploratories

Heiko Nacke[1], Christiane Will[1], Sarah Herzog[1], Boris Nowka[1], Martin Engelhaupt[1] & Rolf Daniel[1,2]

[1]Department of Genomic and Applied Microbiology, Institute of Microbiology and Genetics, Georg-August University Göttingen, Göttingen, Germany; and [2]Göttingen Genomics Laboratory, Institute of Microbiology and Genetics, Georg-August University Göttingen, Göttingen, Germany

Correspondence: Rolf Daniel, Department of Genomic and Applied Microbiology, Institute of Microbiology and Genetics, Georg-August University Göttingen, Grisebachstr. 8, D-37077 Göttingen, Germany. Tel.: +49 551 393 827; fax: +49 551 391 2181; e-mail: rdaniel@gwdg.de

Present address: Boris Nowka, Universität Hamburg, Biozentrum Klein Flottbek, Ohnhorststraße 18, D-22609 Hamburg, Germany.

Abstract

Microbial metagenomes derived from soils are rich sources for the discovery of novel genes and biocatalysts. Fourteen environmental plasmid and seven fosmid libraries obtained from 10 German forest soils (A horizons) and six grassland soils (A and B horizons) were screened for genes conferring lipolytic activity. The libraries comprised approximately 29.3 Gb of cloned soil DNA. Partial activity-based screening of the constructed libraries resulted in the identification of 37 unique lipolytic clones. The amino acid sequences of the 37 corresponding lipolytic gene products shared 29–90% identity to other lipolytic enzymes, which were mainly uncharacterized or derived from uncultured microorganisms. Multiple sequence alignments and phylogenetic tree analysis revealed that 35 of the predicted proteins were new members of known families of lipolytic enzymes. The remaining two gene products represent two putatively new families. In addition, sequence analysis indicated that two genes encode true lipases, whereas the other genes encode esterases. The determination of substrate specificity and chain-length selectivity using different triacylglycerides and *p*-nitrophenyl esters of fatty acids as substrates supported the classification of the esterases.

Introduction

Lipolytic enzymes such as lipases (EC 3.1.1.3) and esterases (EC 3.1.1.1.) are ubiquitous enzymes found in animals, plants, and microorganisms. These enzymes exhibit broad substrate specificity and catalyze both the hydrolysis and the synthesis of esters formed from glycerol and fatty acids. Lipases resemble esterases, but differ from them in their ability to act on water-insoluble esters (Arpigny & Jaeger, 1999). Lipolytic enzymes have been recognized as very useful biocatalysts because of their wide-ranging versatility in industrial applications, including food technology, detergent production, biodiesel formation, fine chemistry, and biomedical sciences (Jaeger & Eggert, 2002).

Soils harbor enormously diverse microbial communities and are a major reservoir of microbial genomic and taxo-nomic diversity. The microbial diversity in soils exceeds that of other environments and, by far, that of eukaryotic organisms. One gram of soil may contain up to 10 billion microorganisms of possibly thousands of different species (Rosselló-Mora & Amann, 2001). Soil microorganisms have been the major source for lipolytic enzymes and other biomolecules of industrial importance (Strohl, 2000). However, of late, the discovery rate of novel biomolecules is extremely low by applying traditional cultivation techniques, because most of the soil microorganisms cannot be cultured and only a small fraction of soil microbial diversity is assessed in this way. Culture-dependent methods have been complemented or replaced by culture-independent metagenomic approaches, which theoretically provide access to the collective nucleic acids of all indigenous microorganisms present in the studied environment (Handelsman, 2004;

Daniel, 2005). Functional metagenomics comprising the isolation of DNA from environmental samples without prior enrichment of individual microorganisms, construction of libraries from the recovered DNA, and function-driven screening of the generated libraries has led to the identification and characterization of a variety of novel enzymes (Ferrer *et al.*, 2005; Simon & Daniel, 2009; Steele *et al.*, 2009), including lipolytic enzymes (Roh & Villate, 2008; Rashamuse *et al.*, 2009). Lipolytic enzymes have been derived from different environmental samples such as soils (Henne *et al.*, 2000; Lee *et al.*, 2004; Elend *et al.*, 2006), sea water (Chu *et al.*, 2008), and sediments (Jeon *et al.*, 2008, 2009).

In this study, we used the soil metagenome as a source for the recovery of novel genes encoding lipolytic enzymes. We report on the construction of small-insert and large-insert metagenomic libraries from 16 different forest and grassland soil samples, which were derived from the three German Biodiversity Exploratories Hainich-Dün, Schorfheide-Chorin, and Schwäbische Alb (Fischer *et al.*, 2010). Subsequently, the constructed libraries were subjected to activity-based screening for genes encoding lipolytic enzymes. In this way, 37 novel lipolytic enzymes were identified. Lipolytic activities of the genes were confirmed by analysis of substrate specificity and chain-length selectivity. Our results demonstrate that function-driven soil-based metagenomics is a very powerful approach for the discovery of novel biomolecules and soil microorganisms can continue to play a major role as a resource for natural product discovery.

Materials and methods

Soil sampling

Soil samples used for metagenomic library construction were derived from the A horizons of 10 forest and six grassland sites of the three German Biodiversity Exploratories Hainich-Dün (samples HEG1, HEG9, HEW2, HEW5, HEW9), Schorfheide-Chorin (samples SEG2, SEG6, SEG9, SEW2, SEW5, SEW8), and Schwäbische Alb (samples AEG2, AEW1, AEW4, AEW5, AEW9). In addition, the B horizons from the Hainich-Dün samples HEG1 and HEG9 were used. Samples were collected in April and May 2008. Sampling was performed as described by Will *et al.* (2010) and Nacke *et al.* (2011). Descriptions of the sampling sites and soil characteristics are provided in Supporting Information, Table S1. Names of the metagenomic libraries refer to the designation of the samples from which the libraries were derived.

Isolation of soil DNA and construction of metagenomic DNA libraries

Total microbial community DNA was isolated from 10 g of soil per sample. For this purpose, the MoBio Power Max Soil DNA extraction kit (MoBio Laboratories, Carlsbad, CA) was used according to the instructions of the manufacturer.

Small-insert libraries were constructed using the plasmid pCR-XL-TOPO as a vector (TOPO XL PCR Cloning Kit; Invitrogen GmbH, Karlsruhe, Germany). Approximately 10 µg extracted DNA of each soil sample was separated by agarose gel electrophoresis. Subsequently, DNA fragments > 6 kb were recovered and purified from the gels using the peqGold Gel Extraction Kit (Peqlab Biotechnologie GmbH, Erlangen, Germany). The purified DNA fragments were subjected to blunt-end polishing using T4 DNA polymerase (MBI Fermentas, St Leon-Rot, Germany) as suggested by the manufacturer. Subsequently, the DNA was purified using SureClean solution (Bioline GmbH, Luckenwalde, Germany) and the resulting DNA pellet was suspended in 35 µL H_2O. Subsequently, a deoxyadenosine was added to the 3′ termini of the DNA to facilitate cloning by the TA method. For this purpose, 1 µL dATP solution (100 mM), 6 µL $MgCl_2$ solution (25 mM), 7 µL of 10-fold $(NH_4)_2SO_4$-containing *Taq* DNA polymerase buffer (MBI Fermentas), 1 µL of *Taq* DNA polymerase (5 U), and 20 µL of H_2O were mixed with the DNA solution, incubated at 72 °C for 30 min, and purified using SureClean solution (Bioline GmbH). The resulting DNA pellet was suspended in 15 µL H_2O and dephosphorylated using 5 U Antarctic Phosphatase (NEB, Ipswich, MA) as described by the manufacturer. Finally, the recovered DNA fragments were inserted into pCR-XL-TOPO using the TOPO XL PCR cloning kit (Invitrogen). To screen the small-insert metagenomic libraries for lipolytic activity, *Escherichia coli* DH5α (Ausubel *et al.*, 1987) was used as a host. Large-insert metagenomic libraries were constructed using the fosmid pCC1FOS as a vector and the Copy Control Fosmid Library Production kit (Epicentre Biotechnologies, Madison, WI) as recommended by the manufacturer. The extracted DNA (5 µg) was directly inserted into the fosmid without prior size fractionation (Simon & Daniel, 2010). Subsequently, the resulting recombinant fosmids were packaged into λ phages using MaxPlax Lambda Packaging Extracts (Epicentre Biotechnologies), and used to infect *E. coli* EPI300-T1^R cells according to the protocol of the manufacturer.

Growth condition and activity-based screening

Escherichia coli strains were routinely grown in Luria–Bertani (LB) medium at 30 °C. For activity-based screening of metagenomic libraries, recombinant *E. coli* strains were grown under aerobic conditions in LB medium, which was supplemented with 1% tributyrin and solidified with agar (15 g L^{-1}). For the determination of substrate specificity, the following compounds were added instead of tributyrin: tricaproin, tricaprylin, tricaprin, trilaurin, trimyristin, and tripalmitin. In addition, to maintain the presence of

recombinant plasmids and fosmids, the medium contained $50 \, \text{mg} \, \text{L}^{-1}$ kanamycin or $12.5 \, \text{mg} \, \text{L}^{-1}$ chloramphenicol, respectively. Clones showing lipolytic activity were identified by the formation of clear zones (halos) against the creamy background after incubation for 1–7 days at $37 \, ^{\circ}\text{C}$. To avoid the isolation of false-positive clones and to confirm that the lipolytic activity of the positive clones was plasmid encoded, the recombinant plasmids were isolated and used to transform *E. coli*. The resulting *E. coli* strains were screened again on tributyrin-containing agar.

Subcloning and sequence analysis

To subclone DNA fragments containing the lipolytic genes from large-insert fosmids, the recombinant fosmids from positive clones were sheared by sonication (UP200S Sonicator, Dr Hielscher GmbH, 5 s at 30% amplitude, cycle 0.5). Subsequently, sheared DNA fragments were separated by agarose gel electrophoresis, and 2–6-kb fragment were excised and extracted using the peqGold gel extraction kit (Peqlab Biotechnologie GmbH). The resulting DNA fragments were ligated into pCR-XL-TOPO or pCR4-TOPO (Invitrogen), and used to transform *E. coli* as recommended by the manufacturer (Invitrogen). The resulting recombinant *E. coli* strains were screened on tributyrin-containing indicator agar for the presence of genes conferring lipolytic activity.

The recombinant plasmids derived from all 37 positive clones were sequenced by the Göttingen Genomics Laboratory (Göttingen, Germany). The initial prediction of ORFs located on the inserts of plasmids pLE01–pLE08 and pLE10–pLE38 was accomplished using the ORF FINDER program (http://www.ncbi.nlm.nih.gov/gorf/gorf.html) provided by the National Center for Biotechnology Information and the ARTEMIS program (Rutherford *et al.*, 2000). The results were verified and improved manually using criteria such as the presence of a ribosome-binding site, GC frame plot analysis, and similarity to known lipolytic-protein-encoding sequences. Initial annotation of the deduced proteins was performed by searching the amino acid sequences against the public GenBank database using the BLAST program (Ye *et al.*, 2006). All coding sequences were examined for similarities to protein families and domains using searches against the CDD databases (Marchler-Bauer *et al.*, 2007). Signal peptides of putative lipolytic proteins were predicted using the SIGNALP 3.0 server (Bendtsen *et al.*, 2004). To construct a phylogenetic tree of the lipolytic proteins recovered, multiple alignments of the deduced protein sequences were performed using CLUSTALW version 2.0.12 (Thompson *et al.*, 1994) and examined with the BIOEDIT program (Hall, 1999). The phylogenetic tree was constructed using the program MEGA version 4.0.2 (Tamura *et al.*, 2007) using the neighbor-joining method. Bootstrap-

ping based on 1000 resamplings was used to estimate the robustness of the tree (Felsenstein, 1985).

The nucleotide sequences of the recombinant plasmids harboring the esterase genes *est01–est08* and *est10–est38* (pLE01–pLE08 and pLE10–pLE38) have been submitted to GenBank under accession numbers HQ156900–HQ156907 and HQ156909–HQ156937.

Lipase/esterase activity assay

To analyze the lipolytic activity of *E. coli* cells harboring the individual plasmids pLE01–pLE08 and pLE10–pLE38, *p*-nitrophenyl esters of fatty acids were used as substrates. *Escherichia coli* strains carrying the cloning vector pCR-XL-TOPO were used as controls. The *E. coli* clones were grown in LB medium to an $OD_{600 \, \text{nm}}$ of 3.0–4.0 and the cell cultures were directly used for the lipolytic activity assay described by Lee *et al.* (2006), with modifications. The activity was determined by measuring *p*-nitrophenol formation from enzymatic hydrolysis of fatty acid *p*-nitrophenyl esters (C_4, *p*-nitrophenyl butyrate; C_6, *p*-nitrophenyl caproate; C_8, *p*-nitrophenyl caprylate; C_{10}, *p*-nitrophenyl caprate; C_{12}, *p*-nitrophenyl laurate; and C_{16}, *p*-nitrophenyl palmitate). Measurements were performed at 410 nm using a Cary 100 UV-visible spectrophotometer with a dual cell peltier accessory (Varian Inc., Vic., Australia). Enzyme activity was measured at $25 \, ^{\circ}\text{C}$. The reaction mixture contained $890 \, \mu\text{L}$ 50 mM Tris-HCl (pH 7.5), $100 \, \mu\text{L}$ culture supernatant, and $10 \, \mu\text{L}$ 1 mM *p*-nitrophenyl ester as a substrate. The reaction was initiated by substrate addition. One unit (U) of enzyme activity was defined as the amount of activity required for the release of $1 \, \mu\text{mol} \, p$-nitrophenol min^{-1} from *p*-nitrophenyl ester.

Results and discussion

Construction of environmental DNA libraries

DNA derived from soil samples of six grassland and 10 forest plots was used for the construction of 21 metagenomic DNA libraries (Table 1). DNA was directly isolated from the samples without previous enrichment or extraction of microbial cells. The DNA yield ranged in the A horizons from 17 to $56 \, \mu\text{g} \, \text{g}^{-1}$ soil and in the B horizons from 7 to $9 \, \mu\text{g} \, \text{g}^{-1}$ soil (Table S1). The yield of the soil DNA derived from the A horizons exceeded the $2–15 \, \mu\text{g} \, \text{g}^{-1}$ soil described in other publications (Lee *et al.*, 2004; Hong *et al.*, 2007; Lämmle *et al.*, 2007). In recently published studies, soil-derived metagenomic libraries comprised 11 000–80 000 and 2400–200 000 clones using plasmids and fosmids as vectors, respectively (Hong *et al.*, 2007; Kim *et al.*, 2007; Tirawongsaroj *et al.*, 2008; Cieśliński *et al.*, 2009; Waschkowitz *et al.*, 2009; Couto *et al.*, 2010). In this study, the 14 plasmid libraries and the seven fosmid libraries contained

Table 1. Characterization of constructed metagenomic libraries and screening for genes conferring lipolytic activity on *Escherichia coli*

Library*	Sample site	Vector	Number of clones	Average insert size (kb)†	Insert frequency (%)	Estimated library size (Gb)	Screening extend (Gb)	No. of lipolytic *E. coli* clones (designation)
SEG2	Schorfheide	pCC1FOS	86 944	24.0	100	2.09	0.22	1 (pLE01)
SEG6	Schorfheide	pCR-XL-TOPO	39 825	6.0	91	0.22	0.22	1 (pLE02)
SEG9	Schorfheide	pCR-XL-TOPO	68 770	7.3	94	0.47	0.47	1 (pLE03)
SEG9	Schorfheide	pCC1FOS	147 888	23.3	100	3.45	0.29	1 (pLE04)
SEW2	Schorfheide	pCR-XL-TOPO	135 240	5.7	95	0.73	0.45	1 (pLE05)
SEW5	Schorfheide	pCR-XL-TOPO	166 040	4.0	95	0.63	0.34	1 (pLE06)
SEW8	Schorfheide	pCR-XL-TOPO	69 984	5.5	90	0.35	0.35	1 (pLE07)
HEG1	Hainich	pCR-XL-TOPO	70 313	2.6	98	0.18	0.18	1 (pLE08)
HEG9	Hainich	pCR-XL-TOPO	161 940	6.4	69	0.72	0.72	6 (pLE10-pLE15)
HEG1‡	Hainich	pCR-XL-TOPO	510 808	5.7	97	2.80	2.80	2 (pLE16; pLE17)
HEG9‡	Hainich	pCR-XL-TOPO	150 782	9.4	96	1.36	0.54	4 (pLE18-pLE21)
HEW2	Hainich	pCR-XL-TOPO	340 990	4.9	88	1.48	1.48	2 (pLE22; pLE23)
HEW5	Hainich	pCR-XL-TOPO	181 958	6.7	92	1.13	1.13	4 (pLE24-pLE27)
HEW9	Hainich	pCC1FOS	60 000	27.8	100	1.67	0.13	1 (pLE28)
AEG2	Schwäbische Alb	pCC1FOS	299 880	26.3	100	7.89	0.64	2 (pLE29; pLE30)
AEW1	Schwäbische Alb	pCR-XL-TOPO	129 748	6.7	91	0.79	0.79	2 (pLE31; pLE32)
AEW1	Schwäbische Alb	pCC1FOS	50 952	27.8	100	1.42	0.63	2 (pLE33; pLE34)
AEW4	Schwäbische Alb	pCC1FOS	61 530	19.0	100	1.17	0.18	1 (pLE35)
AEW5	Schwäbische Alb	pCR-XL-TOPO	90 300	5.2	89	0.42	0.42	1 (pLE36)
AEW5	Schwäbische Alb	pCC1FOS	4600	30.0	100	0.14	0.14	1 (pLE37)
AEW9	Schwäbische Alb	pCR-XL-TOPO	100 950	2.6	89	0.23	0.13	1 (pLE38)

*Names of the metagenomic libraries refer to the designation of the sampling sites. Descriptions of the sampling sites are provided in Table S1.
†The average insert size was determined by analysis of 20 insert-containing recombinant plasmids or fosmids.
‡Libraries constructed from soil derived from B horizon.

approximately 40 000–341 000 clones and 4600–300 000 clones, respectively (Table 1). The quality of the 21 different environmental libraries was controlled by determination of the average insert sizes and the percentage of insert-bearing *E. coli* clones. The average insert sizes ranged from 2.6 to 9.4 kb (plasmids) and 19 to 30 kb (fosmids). These values are in accordance with those published for other soil-derived small-insert or large-insert metagenomic libraries (Henne *et al.*, 2000; Lämmle *et al.*, 2007; Waschkowitz *et al.*, 2009; Couto *et al.*, 2010). The percentage of insert-carrying clones was 69–98% (plasmids) and 100% (fosmids). Thus, approximately 11.5 and 17.8 Gb of cloned soil DNA were stored in the constructed plasmid and fosmid libraries, respectively. Assuming an average prokaryotic genome size of 5 Mb (Hårdeman & Sjöling, 2007), both types of metagenomic libraries represented approximately 5500 prokaryotic genomes.

Screening for genes conferring lipolytic activity

To isolate genes encoding lipolytic activity from the constructed libraries, a function-driven approach was chosen. As sequence information is not required before screening, this is the only strategy that bears the potential to discover entirely novel lipolytic genes (Daniel, 2005; Simon & Daniel, 2009). In addition, it is selective for full-length genes and

functional gene products. The screen for genes exhibiting lipolytic activity was based on the ability of library-bearing *E. coli* clones to form halos when grown on agar medium containing tributyrin. Halo formation is caused by the hydrolysis of tributyrin. This function-based screen has been used to identify the lipolytic activity of individual microorganisms (Seo *et al.*, 2005; Hantsis-Zacharov & Halpern, 2007), and recombinant *E. coli* strains that harbor gene libraries from single microorganisms (Hotta *et al.*, 2002) or metagenomic libraries (Heath *et al.*, 2009; Hu *et al.*, 2010). In the case of metagenomic libraries, genes conferring lipolytic activity have been recovered from diverse environments such as mangrove sediment (Couto *et al.*, 2010), marine sediment (Hårdeman & Sjöling, 2007; Hu *et al.*, 2010), water samples (Ranjan *et al.*, 2005; Chu *et al.*, 2008; Wu & Sun, 2009), compost (Lämmle *et al.*, 2007), and soils (Henne *et al.*, 2000; Lee *et al.*, 2004; Kim *et al.*, 2005; Hong *et al.*, 2007).

In this study, the constructed plasmid libraries and fosmid libraries were partially screened using the abovementioned activity-based approach. The screening effort comprised approximately 10.0 Gb (plasmids) and 2.2 Gb (fosmids) of the cloned soil DNA. Positive *E. coli* clones were collected after incubation for 1–7 days at 37 °C on tributyrin-containing indicator agar. In order to confirm that the lipolytic activity of the positive clones was vector

encoded, the recombinant plasmids or fosmids were isolated from the positive clones and used to transform *E. coli*. The resulting *E. coli* strains were screened again on indicator agar. Twenty-eight different recombinant plasmids and nine fosmids conferred a stable lipolytic phenotype (Table 1). Eighteen of these were derived from A horizons of forest soil samples, 13 from A horizons of grassland soil samples, and six from B horizons of grassland soil samples.

In this study, the average hit rate was approximately one lipolytic gene per 240 Mb (fosmids) and 360 Mb (plasmids), respectively, of screened soil DNA. For comparison, one lipolytic gene per 480 Mb of screened soil DNA, which was cloned into a high-copy plasmid vector, was identified during screening on tributyrin agar in another study (Henne *et al.*, 2000). Other soil metagenomic studies using fosmid or bacterial artificial chromosomes as vectors achieved hit rates of one lipolytic gene per 148 Mb (Lee *et al.*, 2004) or 50 Mb of screened soil DNA (Rondon *et al.*, 2000).

Molecular analysis

The inserts of all 28 recombinant plasmids (pLE02, pLE03, pLE05–pLE08, pLE10–pLE27, pLE31, pLE32, pLE36, and pLE38) recovered from the positive clones were sequenced. The insert sizes of the plasmids ranged from 1107 to 11 172 bp (Table S2). The nine fosmid-harboring lipolytic clones carried insert DNA ranging from 19 800 to 36 500 bases in size (Table S2). DNA fragments of fosmids containing the genes conferring lipolytic activity were identified by subcloning and screening for lipolytic subclones before sequencing. In this way, recombinant plasmids carrying the desired DNA fragments for all nine fosmids were recovered. The insert sizes of the nine corresponding plasmids (pLE01, pLE04, pLE28, pLE29, pLE30, pLE33, pLE34, pLE35, and pLE37) ranged from 1511 to 3568 bp (Table S2). The insert sequences of pLE01–pLE08 and pLE10–pLE38 were sequenced and analyzed. In all 37 cases, a putative gene showing similarities to known genes encoding lipases or esterases was found. The amino acid sequences deduced from the 37 identified genes (*est01–est08* and *est10–est38*) comprised 230–556 amino acids with calculated molecular masses from 25.1 to 57.9 kDa. The sequence identities to the closest similar known lipolytic protein ranged from 29% to 90% (Table 2). Fourteen of the 37 putative lipolytic proteins showed the highest similarity to esterases/lipases from uncultured bacteria and the remaining 23 to lipolytic proteins deduced from genome sequences of individual microorganisms. Interestingly, 50% of the lipolytic genes derived from forest soils (nine genes), but only 30% of those recovered from grassland soils (five genes) showed the closest similarity to esterases/lipases from uncultured bacteria. In addition, almost all of the forest soil-derived enzymes showed the closest amino acid identity (48–87%) to six putative

lipolytic enzymes that have been recovered during activity-based screening of a Korean forest soil-derived metagenomic library (Lee *et al.*, 2004). In the remaining cases in which a lipolytic protein from an uncultured bacterium was the best hit, the matching lipase/esterase was detected in other recently published metagenomic surveys (Hong *et al.*, 2007; Hu *et al.*, 2010). Hu *et al.* (2010) used samples from marine sediment and the identities to our lipolytic enzyme sequences were lower than those to our lipolytic enzymes recovered from other soil metagenomic libraries (data not shown). Thus, the similarity of the habitat seems to have an impact on the degree of amino acid identity.

Classification of the lipolytic enzymes

We used the classification system of Arpigny & Jaeger (1999), who subdivided bacterial lipolytic enzymes into eight different families (I–VIII) based on the amino acid sequence similarity and some fundamental biological properties. As shown in Fig. 1, 35 of the enzyme sequences grouped into five of these families. The majority of the classified enzymes were affiliated to family IV, followed by families V, VIII, I, and VI. The remaining two enzyme sequences (Est01 and Est15) could not be assigned to the families described by Arpigny & Jaeger (1999).

Family IV

Most of the enzyme sequences (17 sequences) were affiliated to family IV. Members of this family show significant similarity to the mammalian hormone-sensitive lipase (HSL). Therefore, family IV is also known as the 'HSL family' of lipolytic enzymes (Arpigny & Jaeger, 1999; Hausmann & Jaeger, 2010). All 17 enzyme sequences contained the lipase-conserved catalytic triad residues aspartate, histidine and the nucleophile serine in the consensus pentapeptide motif GXSXG (Fig. 2). The only exceptions were the amino acid sequences of Est05, Est06, Est29, Est34, and Est38, in which the aspartate residue was replaced by a glutamate residue (Fig. 2). This substitution is common for members of family IV (Chu *et al.*, 2008; Hu *et al.*, 2010). The highly conserved motif HGGGF was present in 16 enzyme sequences. In the amino acid sequence of Est35, the motif HGGGF was replaced by PGGGF (Fig. 2). Lipolytic enzymes of the HSL family were predominant and widespread in our samples, as the 17 enzymes belonging to this family originated from forest and grassland samples and were present in all three German Biodiversity Exploratories studied. Lipolytic enzymes of family IV were also predominantly recovered in similar activity-based screens of other metagenomic libraries derived from a variety of environments such as forest soil (Lee *et al.*, 2004; Hong *et al.*, 2007), deep sea sediment (Hu *et al.*, 2010), and arctic seashore sediment

Table 2. Description of the lipolytic gene products and their observed sequence similarities

Gene (accession no.)	No. of encoded amino acids	Closest similar lipolytic protein, accession no. (no. of encoded amino acids), organism	*E* value	Amino acid homology to the closest similar lipolytic protein (% identity)
est01 (HQ156900)	397	Phospholipase/carboxylesterase, ZP_06237474 (349), *Frankia* sp. EuI1c	3e − 23	101/298 (33%)
est02 (HQ156901)	311	Lipase/esterase, ABQ11271 (310), uncultured bacterium	1e − 79	160/311 (51%)
est03 (HQ156902)	333	Putative α/β hydrolase, YP_555239 (276), *Burkholderia xenovorans* LB400	2e − 42	101/267 (37%)
est04 (HQ156903)	294	Lipase/esterase, ABQ11272 (296), uncultured bacterium	1e − 61	133/267 (49%)
est05 (HQ156904)	310	Lipase/esterase, AAS77233 (296), uncultured bacterium	3e − 140	239/284 (84%)
est06 (HQ156905)	296	Lipase/esterase, AAS77236 (296), uncultured bacterium	1e − 136	235/284 (82%)
est07 (HQ156906)	321	Putative lipase, ZP_05767946 (319), *Mycobacterium tuberculosis* T46	8e − 138	243/320 (75%)
est08 (HQ156907)	384	Esterase, ZP_04691113 (371), *Streptomyces ghanaensis* ATCC 14672	9e − 64	144/365 (39%)
est10 (HQ156909)	337	β-Lactamase, YP_001682441 (408), *Caulobacter* sp. K31	0.0	305/337 (90%)
est11 (HQ156910)	314	α/β hydrolase fold-3 domain protein, YP_002946812 (314), *Variovorax paradoxus* S110	2e − 132	239/314 (76%)
est12 (HQ156911)	408	β-Lactamase, YP_001530546 (390), *Desulfococcus oleovorans* Hxd3	1e − 89	179/391 (45%)
est13 (HQ156912)	272	Putative lipase, YP_766845 (293), *Rhizobium leguminosarum* bv. *viciae* 3841	2e − 39	89/232 (38%)
est14 (HQ156913)	377	Putative esterase, ZP_01617169 (381), marine gammaproteobacterium HTCC2143	4e − 98	193/380 (50%)
est15 (HQ156914)	448	Lipase, ZP_06703106 (440), *Xanthomonas fuscans* ssp. *aurantifolii* str. ICPB 11122	1e − 04	52/177 (29%)
est16 (HQ156915)	272	α/β fold family hydrolase, YP_628483 (314), *Myxococcus xanthus* DK 1622	2e − 24	87/262 (33%)
est17 (HQ156916)	556	Lipase class 2, YP_953514 (573), *Mycobacterium vanbaalenii* PYR-1	1e − 113	222/357 (62%)
est18 (HQ156917)	331	Lipolytic enzyme, ACL67843 (311), uncultured bacterium	2e − 73	151/308 (49%)
est19 (HQ156918)	337	Triacylglycerol lipase, ZP_06045720 (317), *Aeromicrobium marinum* DSM 15272	5e − 84	159/298 (53%)
est20 (HQ156919)	443	β-Lactamase, YP_589716 (424), *Candidatus* Koribacter versatilis Ellin345	1e − 108	218/433 (50%)
est21 (HQ156920)	230	Phospholipase/carboxylesterase, YP_413093 (227), *Nitrosospira multiformis* ATCC 25196	1e − 67	128/223 (57%)
est22 (HQ156921)	424	β-Lactamase, YP_577943 (424), *Nitrobacter hamburgensis* X14	0.0	340/424 (80%)
est23 (HQ156922)	310	Lipase/esterase, ABQ11271 (310), uncultured bacterium	1e − 158	268/310 (86%)
est24 (HQ156923)	277	α/β hydrolase fold protein, YP_003321859 (267), *Thermobaculum terrenum* ATCC BAA-798	4e − 37	86/268 (32%)
est25 (HQ156924)	276	α/β hydrolase domain-containing protein, ZP_03632535 (297), bacterium Ellin514	2e − 49	115/273 (42%)
est26 (HQ156925)	257	α/β hydrolase fold, YP_001773617 (254), *Methylobacterium* sp. 4-46	6e − 87	164/252 (65%)
est27 (HQ156926)	402	β-Lactamase, YP_484201 (395), *Rhodopseudomonas palustris* HaA2	6e − 84	155/382 (40%)
est28 (HQ156927)	245	α/β fold family hydrolase, YP_360549 (258), *Carboxydothermus hydrogenoformans* Z-2901	8e − 37	94/232 (40%)
est29 (HQ156928)	300	Lipolytic enzyme, ACL67845 (307), uncultured bacterium	1e − 78	158/299 (52%)
est30 (HQ156929)	363	Esterase, AAY45707 (362), uncultured bacterium	2e − 105	199/351 (56%)
est31 (HQ156930)	312	Lipase/esterase, AAS77247 (311), uncultured bacterium	8e − 127	214/312 (68%)
est32 (HQ156931)	266	Carboxylesterase (est-1), NP_069699 (266), *Archaeoglobus fulgidus* DSM 4304	3e − 30	91/277 (32%)
est33 (HQ156932)	297	Lipase/esterase, ABQ11272 (296), uncultured bacterium	5e − 102	178/296 (60%)
est34 (HQ156933)	296	Lipase/esterase, AAS77236 (296), uncultured bacterium	1e − 144	248/285 (87%)
est35 (HQ156934)	314	α/β hydrolase fold-3 domain protein, ZP_05908953 (314), *Vibrio parahaemolyticus* AQ4037	2e − 51	113/285 (39%)
est36 (HQ156935)	312	Lipolytic enzyme, ACL67843 (311), uncultured bacterium	2e − 99	186/311 (59%)
est37 (HQ156936)	309	Lipase/esterase, AAS77247 (311), uncultured bacterium	1e − 78	151/309 (48%)
est38 (HQ156937)	330	Lipase/esterase, AAX37296 (297), uncultured bacterium	2e − 92	169/272 (62%)

(Jeon *et al.*, 2009). Hong *et al.* (2007) identified four enzymes clustering with family IV lipolytic enzymes by screening of a forest-soil metagenomic library. In addition, two enzymes showed similarity to family V lipolytic enzymes and one to a lysophospholipase from family II. However, 20 of the 37 putative lipolytic enzymes identified in this study were not affiliated to family IV.

Family V

Eight of the lipolytic enzyme sequences grouped into family V (Est03, Est13, Est16, Est24–Est26, Est28, and Est32). The multiple sequence alignment revealed that all of them contained the catalytic triad residues (Fig. 2). The consensus motif GXSXG was present in all protein sequences of the
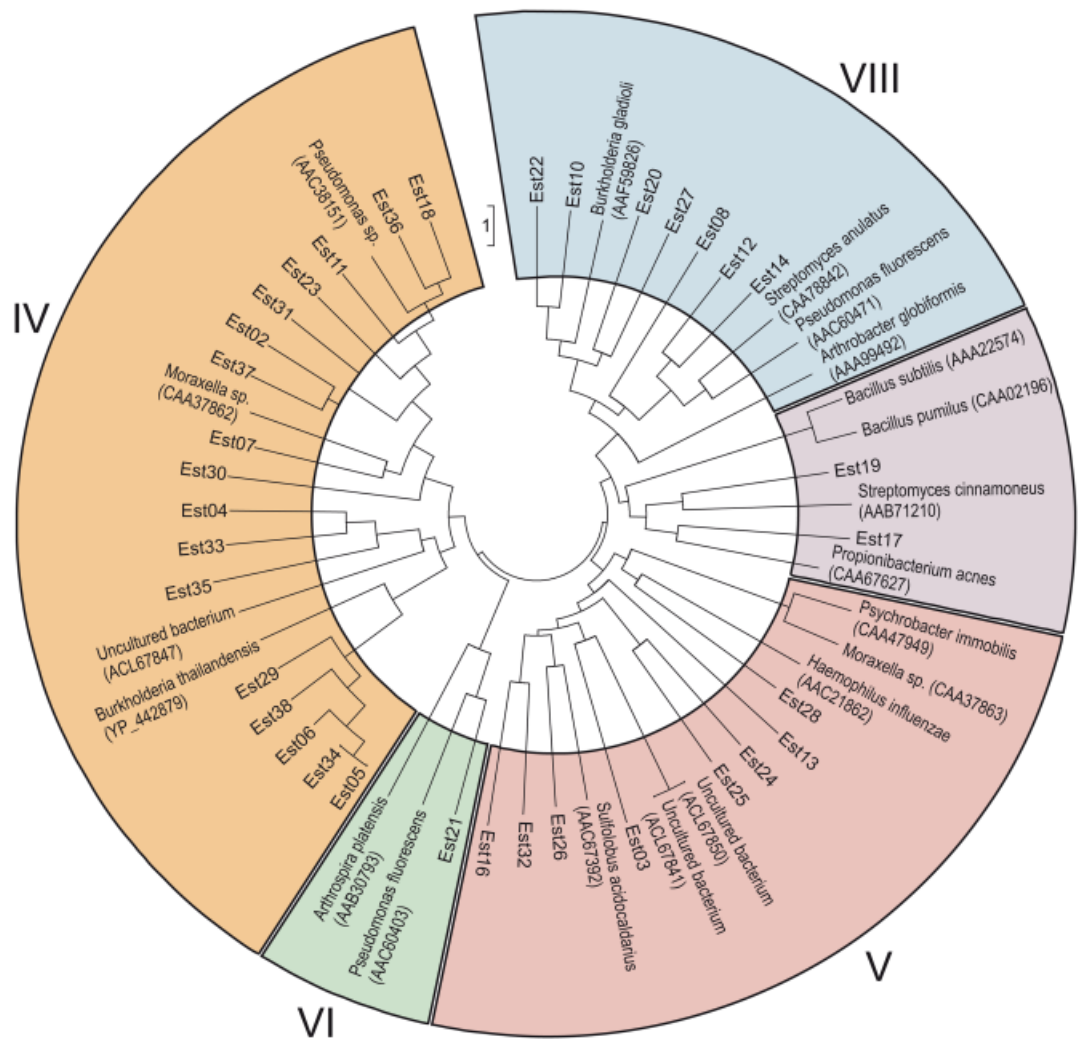
Fig. 1. Unrooted neighbor-joining tree of lipolytic enzymes obtained from forest and grassland soil metagenomes in this study and representative members of families I, IV, V, VI, and VIII. Amino acid sequences of published esterases/lipases were retrieved from GenBank. Lipolytic enzymes belonging to putative new families were analyzed in an alignment with similar proteins retrieved from GenBank (Fig. 3). The length of the branches of the phylogenetic tree indicates the difference of the protein sequences.

family V enzymes. The PTL motif, which is another common motif among family V esterases, was detected in six of the lipolytic enzymes affiliated to family V (Est03, Est13, Est25, Est26, Est28, and Est32) (Fig. 2). In addition, Est24 contained the amino acid sequence PTQ (Fig. 2), which is a known variation of the PTL motif (Arpigny & Jaeger, 1999). Est16 contained the amino acid sequence PAL instead of PTL.

Family VIII

According to the multiple sequence alignment, seven esterases (Est08, Est10, Est12, Est14, Est20, Est22, and Est27) belonged to family VIII (Fig. 2). Esterases belonging to this family show a remarkable similarity to class C β-lactamases

and penicillin-binding proteins (Bornscheuer, 2002). Members of this family comprise approximately 380 amino residues with a molecular mass of approximately 42 kDa (Hausmann & Jaeger, 2010). The length of the amino acid sequences and the molecular mass of all seven enzymes were in the same range (337–443 amino acids and 36.9–48.6 kDa, respectively). The enzymes showed 39–90% identity to the closest similar known lipolytic enzyme, which was, in most cases, a putative β-lactamase (Table 2). The conserved pentapeptide GXSXG, which includes the active-site serine, was only present in the sequences of Est12 and Est14 (Fig. 2). Analysis of the protein sequence of the remaining enzymes revealed that the active site serine is part of the consensus motif SXXK, which is conserved in class C β-lactamases, penicillin-binding proteins, and family VIII

esterases (Arpigny & Jaeger, 1999; Wagner *et al.*, 2002; Hausmann & Jaeger, 2010) (Fig. 2). Moreover, the family VIII motif LLXHXXG described by Ranjan *et al.* (2005) appeared in four of these lipolytic enzymes (Est10, Est12, Est14, and Est20) (Fig. 2). In the remaining three enzymes, the two leucine residues were replaced by alanine and methionine (Est08), leucine and methionine (Est22), or leucine and alanine residues (Est27) (Fig. 2). Thus, these

```
Family I
                                        ▼                                                      ▼                                       ▼
Est19      153 GFVDQVLAATGASEVDIVGHSEGS-   263 YTMVLTKYDEAVVPYTSGLMDAPN--AKNVVLQKICPLDFS-GHG
Est17      371 AEIDRVLEPTGAEKVILIGHSQGGG   479 YTNIITMNDEIVTPYTQQALEGDN--VTNIVLQERYPGYPA-GHL
AAB71210   105 DFVEAVRGATGAAKVDIVGHSQGG-   198 YTVITTRYDEVVIPYASALLTGDKEHLTNVVLQDKCPLDLY-MHD
AAA22574    88 REVQKVLDETGAKKVDIVAHSMGG-   156 YTSIYSSADMIVMNYLSRLDGARN-----------VQIHGV-GHI
CAA02196    57 REVKDVLDKTGAKKVDIVAHSMGG-   125 YTSVYSSADLIVVNSLSRLIGARN-----------ILIHGV-GHI
CAA67627   149 AFVDRVRKATGSEKVDFVGHSQGGG   259 YTVISTRLDMTVTPYTQAFLKGAK---NMTVQDACPLDAY-GHG

Family IV
                              ▼                                  ▼                                               ▼
Est31       81 FFHGGGWV   155 GDSAGGNLAA   244 EAFVMTCEVDPLRD--EGEAYADAURKAGIAVKSKRYNGVCHGELM
Est35       66 YVPGGGFT   140 GESAGGYLVG   236 FFQIFHGNKDQIVSPSQTLVLHNALVAKEASSVRYVVDGANHGDMA
Est30      124 HIHGGGWV   201 GESGGGNLTL   296 GHVISVKLDPLRD--EGLKYYQMPLAAGVRVYSRTVNGTCHAGDV
Est23       81 FFHGGGWV   155 GDSAGGNLSA   242 FALIITAEFDPLRD--EGEAYGEKLRAAGVPVSVTRYEGMIHGEFS
Est36       83 YFHGGGFV   157 GDSAGGNLAA   245 RAYVVTAGFDPLRD--EGKAYIDKLNRAGVAAVYVDYPSMIHGEFG
Est37       80 FFHGGGWV   154 GDSAGGNLSA   241 DTFIATCEFDPLRD--EGEAYGDAPRANCGHVTTKRYDGLIHGVAN
Est02       82 FFHGGGWV   156 GDSAGGNLSA   243 FCFIATCEYDPLRD--EGEAYGAAPRNNGVAAEVKRYDGLIHAAVN
Est11       84 YFHGGGWT   158 GDSAGGNLAA   246 FALVLTAGFDPLRD--EGLAYANALVAAGNRASYLCFERQIHGRIT
Est18      100 YLHGGGWV   174 GDSAGGNLSA   262 FALVITAGFDPLRD--EGRAYADRLSAAGVEVVHREYPGQIHAFVS
Est07       87 FYHGGGFA   161 GDSAGANLAA   253 EAFIGTAEHDPLRD--EGAHYEVLNAAGVPVERSNEPTLVHGYVS
Est33       76 YIHGGGWI   149 GDSAGGNLPA   232 FCFVICGTADDLLP--ESKAMPEALRRANIESELHLMEEMPHAFMQ
Est04       76 YFHGGGWT   149 GDSAGGNLAG   227 FCYVGVGTKDPIYA--ESIKIAAALKAAGIDHDLHVIEGAPHGEFQ
Est05       86 YLHGGGYV   156 GDSAGGGLTV   243 ELLIQVGTAETLLD--DSTRIAERARKAGVKVTLEPWENMIHVEQV
Est06       72 YLHGGGYV   142 GDSAGGGLTV   229 FLLIHVGTAETLLD--DSTRLAERARKAGVKVTLEPWENMIHVEQV
Est34       72 YLHGGGYV   142 GDSAGGGLTI   229 ELLIQVGTAETLLD--DSTRIAERARKAGVKVTLEPWENMIHVEQV
Est38      106 YLHGGGYV   176 GDSAGGGLTV   263 ELLIHVGSDEVLLD--DAIGLAERAKAAGVDATLEQWDRMIHVWHW
Est29       74 YVHGGGYV   144 GDSAGGGLTM   231 ELYIQVGDDETLLD--DSVRLAERAQAAGVDCRLDVFPEMQHVEQM
ACL67847   144 FIHGGGWT   213 GTSAGGGFTL   302 ELFLLSGTAEIVGS--DAIRVMASARSQGVDVTLLVSDGMWHVPIS
CAA37862   163 FFHGGGFC   237 GDSAGGQIAA   352 ISYIVVAELDILRD--EGLAYAELLQKEGVQVQTYTVLGAPHGEIN
AAC38151    79 FFHGGGFV   153 GDSAGGNLAL   241 FTTLITAEFDPLRD--EGEAFALRLQQCAGVSVRVQRCEGMIHGEIS
YP_442879   82 YLHGGGYC   152 GDSSGGGLAL   239 ELFVQASDTEVLLD--DAARVAEKARVADVAVDFKVWHRLPHAWPT

Family V
                              ▼                                      ▼                                                                               ▼
CAA47949   140 GNSMGGAISVA   255 PTLVVWCDKDQIIKPE-TVNLIKKIIPQA----QVIMMEDV-----------------------------------GH
AAC67392   142 GWSMGGFVAQQ   257 PTLVIGGDSDLLLPPQ-NSQYLAENIPNA----QLYIFSPDA-----------------------------------GH
AAC21862   117 GHSMGGKTAMK   229 PTLFIKGGNSSYIKIE-NSEKILEQFPNA----TAFTINGS-----------------------------------GH
ACL67850   116 GMSMGGMIAQT   235 PTLVIHCADDLLVPVE-GGKDTAAAVPEA----ELIIIDGM-----------------------------------GH
ACL67841   116 GMSMGGMIAQT   235 PTLVIHCADDLLVPVE-GGKDTAAAVPEA----ELIIIDGM-----------------------------------GH
CAA37863   140 GNSMGGAISVA   255 PTLVVWCDKDQVIKPE-TTELIKEIIPQA----QVIMMNDV-----------------------------------GH
Est32       92 GNSIGGMIAMQ   206 PTLVLACEEDQATPVA-ANKFLADNIPDA----KLNVLKDI-----------------------------------GH
Est16       93 GHSFGGPISLQ   205 PALILCCDADKPVPLEAASKVLAEKIPDA----RLVVIKDT-----------------------------------GH
Est03      162 GNSLGGFTAAA   274 PTLIIWCREDGLTQLA-MGQRFNKEIAGS----QLFIIEKC-----------------------------------GH
Est24       91 GLSLGGAIGLW   208 PTQITFCRHDMITSTR-FAAAMTDNIPNT----ELIVFENA-----------------------------------SH
Est28       67 GISMGGMIAQE   186 PTLIIHGDDSIVPYP-NAEVLQGAIAGS----KKHTLKSA-----------------------------------GH
Est26      102 CYSMGARIVAF   199 PTLIAVCTKDPVAG---SAQELAALMPHA----IALDIPGR-----------------------------------DH
Est13      131 GFSQGGAVLA   176 PTLVLIGNADDWTPASACEAMAAGRTEFGA---PRTP----GDRSLVELVIYPGVHHSFDVLDLSLAPTRGTTAHGH
Est25      129 GISTGGHLSLM   213 PTLIIHCDSDKVVPLQ-QARAMDHALAKAGVEHKLEVIPGG-----------------------------------GH

Family VI
                              ▼                                  ▼                                     ▼
Est21       24 WMHG----LG--A-D   119 GFSQGGAIAL   166 PIFLGHGKQDNVIPFTAGANSKERITELGYRVQ-WHEYPMAHTVS
AAC60403    19 WLHG----LG--A-D   112 GFSQGGAVVF   159 PALCLHGQYDDVVQNAMCRSAFEHLKSRGVTVT-WQEYPMGHEVL
AAB30793    22 FLHG----WG--A-N   105 GFSQGGAMTF   145 PILMAHGKQDMVVPLGAAHQARDSFQKLGATVE-YHEYNMGHEIC

Family VIII
Est22       69 DTIFRIYSMTKPI   134 VDLMRHTSGLTY   377 GAATTSFW-IDPV
Est08       58 DTLFPSFSTTKGV   108 AQAMSHQS----   337 GAGGSIAF-ADPE
Est10       24 PVHRFIPSWKKLG    58 IDLLRHTAGLTY   290 GMASTAFW-IDPV
Est20       96 DAIFRIASMTKPV   161 RDLLRHTSCISY   400 GIYNTRFW-IDPK
Est12       84 DTIAHVYSVSKGI   134 RWLLSHRAGL-P   362 GAGGSLGF-ADPD
Est14       66 DTIACVFSTTKAA   116 AQLLSHQAGL-P   333 GAGGSLGI-ADMD
Est27       65 DTVFRIFSMTKAV   131 RHLATHTSGLVY   355 GVCNTHFW-FDPA
CAA78842    61 DTLVNVWSTGKGP   111 RHLLSHRSGV-A   339 GAGGSCGL-ADPD
AAC60471    60 DTIVNLFSCTKTF   110 RQLLCHQAGL-P   338 RSAPVRVW-VLPE
AAA99492    52 DSVTGVFSCSKGV   102 AQLLSHQAGLLG   327 GASASLGF-ADPV
AAF59826    68 DTLFRLASVTKPI   122 HHLLTHTSGLGY   350 GVYGHSWF-VDRA
```

variations indicated that the proposed motif is not as conserved as previously assumed. However, Est08, Est22, and Est27 did not form a separate branch within family VIII (Fig. 1). Interestingly, family VIII members were derived only from plots in the Hainich region and were absent in the other two Exploratories.

Families I and VI

Two enzyme sequences (Est17 and Est19) were affiliated to family I containing the true lipases and one to family VI (Est21). All three enzymes were derived from metagenomic libraries constructed from subsoil samples (B horizons). To our knowledge, no other studies on screening of subsoil-derived metagenomic libraries for genes conferring lipolytic enzymes have been published. True lipases (family I members) have rarely been discovered during activity-based screens of soil-derived and other metagenomic libraries for lipolytic enzymes. So far, cold-adapted and thermostable lipases (Wei et al., 2009), as well as a lipase showing similarity to the lipase (Lip) of *Streptomyces albus* (Henne et al., 2000) were derived from soil metagenomic libraries. Furthermore, an enzyme isolated from mangrove sediment formed a unique branch within family I of true lipases (Couto et al., 2010). In this study, we detected two putative lipases. As it is typical for family I lipases, the active-site serine of Est17 and Est19 is embedded in the motif GHSXG (Arpigny & Jaeger, 1999) (Fig. 2).

Esterases belonging to family VI have also been recovered from metagenomic libraries that have been constructed from soil of a hot spring area (Kim et al., 2005) and compost (Lämmle et al., 2007). Arpigny & Jaeger (1999) described a molecular mass in the range of 23–26 kDa for esterases belonging to family VI. This is in accordance with the calculated molecular mass of Est21 (25.1 kDa). The multiple sequence alignment and the phylogenetic tree with two known esterases from family VI indicated that Est21 is a new member of family VI (Figs 1 and 2).

Novel families

Two of the lipolytic enzyme sequences (Est01 and Est15) could not be classified according to Arpigny & Jaeger (1999).

In addition, these enzymes also did not group into other novel families of lipolytic enzymes, which have been described recently (Lee et al., 2006; Chu et al., 2008; Kim et al., 2009; Bayer et al., 2010; Hu et al., 2010). Est01 has a size of 397 amino acids and a predicted molecular mass of 40.7 kDa. A phospholipase/carboxylesterase from *Frankia* sp. was the closest similar known lipolytic protein (33% amino acid identity) (Table 2). An alignment (Fig. 3a) of Est01 with similar known lipolytic proteins revealed a conserved GHSXG (amino acids 195–199) motif containing the active-site serine. The other putative residues of the catalytic triad, histidine and aspartate, are also highly conserved in the aligned sequences (Fig. 3a). The overall low similarities to members of already existing families of lipolytic enzymes indicated that Est01 is a member of a new family of lipolytic enzymes. As Est01 was derived from the German Biodiversity Exploratory Schorfheide, we propose to designate the putative new family EstGS. The protein sequence of Est01 showed the closest similarity (31% and 33%) to two hypothetical proteins from *Congregibacter litoralis* KT71 (ZP_01103967) and gammaproteobacterium NOR5-3 (ZP_05127748). These proteins are probably also members of the new EstGS family.

Est15 is one of the largest lipolytic enzymes detected within this study. It comprises 448 amino acids with a predicted molecular mass of 45.9 kDa (Table 2). The closest similar lipolytic enzyme is a lipase from *Xanthomonas fuscans* (29% identity; ZP_06703106). The deduced protein sequence of Est15 exhibited no significant similarity to conserved motifs described for true lipases or other lipolytic enzyme families. An alignment (Fig. 3b) with the four best-matching protein sequences (55–66% identity), which were all hypothetical proteins from *Mycobacteria* (YP_001073968, YP_642492, YP_890989, and YP_951301), revealed highly conserved sequence regions. The putative active-site serine is embedded in a GHSLG motif (amino acids 273–277). For the other conserved residues of the catalytic triad, histidine and aspartate, analysis of the alignment revealed several candidate histidine and aspartate residues. Thus, the results indicated that Est15 and the hypothetical proteins of *Mycobacteria* are members of a new family of lipolytic enzymes. As Est15 was derived from the German Biodiversity

---

Fig. 2. Multiple sequence alignment of conserved regions of lipolytic enzymes belonging to families I, IV, V, VI, and VIII. Identical amino acid residues are shown as white letters on a dark background. Triangles indicate amino acid residues belonging to the catalytic triad. References: Est02–Est14 and Est16–Est37 (this study); AAB71210, lipase LipA from *Streptomyces cinnamoneus*; AAA22574, lipase from *Bacillus subtilis*; CAA02196, lipase from *Bacillus pumilus*; CAA67627, triacylglycerol lipase from *Propionibacterium acnes*; ACL67847, lipolytic enzyme from uncultured bacterium; CAA37862, triacylglycerol lipase from *Moraxella* sp.; AAC38151, lipase from *Pseudomonas* sp. B11-1; YP_442879, lipase/esterase from *Burkholderia thailandensis* E264; CAA47949, triacylglycerol lipase from *Psychrobacter immobilis*; AAC67392, lipolytic enzyme from *Sulfolobus acidocaldarius*; AAC21862, putative esterase/lipase from *Haemophilus influenzae* Rd KW20; ACL67850, lipolytic enzyme from uncultured bacterium; ACL67841, lipolytic enzyme from uncultured bacterium; CAA37863, triacylglycerol lipase from *Moraxella* sp.; AAC60403, esterase II from *Pseudomonas fluorescens*; AAB30793, serine esterase from *Arthrospira platensis*; CAA78842, esterase A from *Streptomyces anulatus*; AAC60471, esterase III from *Pseudomonas fluorescens*; AAA99492, carboxylic ester hydrolase from *Arthrobacter globiformis*; AAF9826, esterase EstB from *Burkholderia gladioli*.
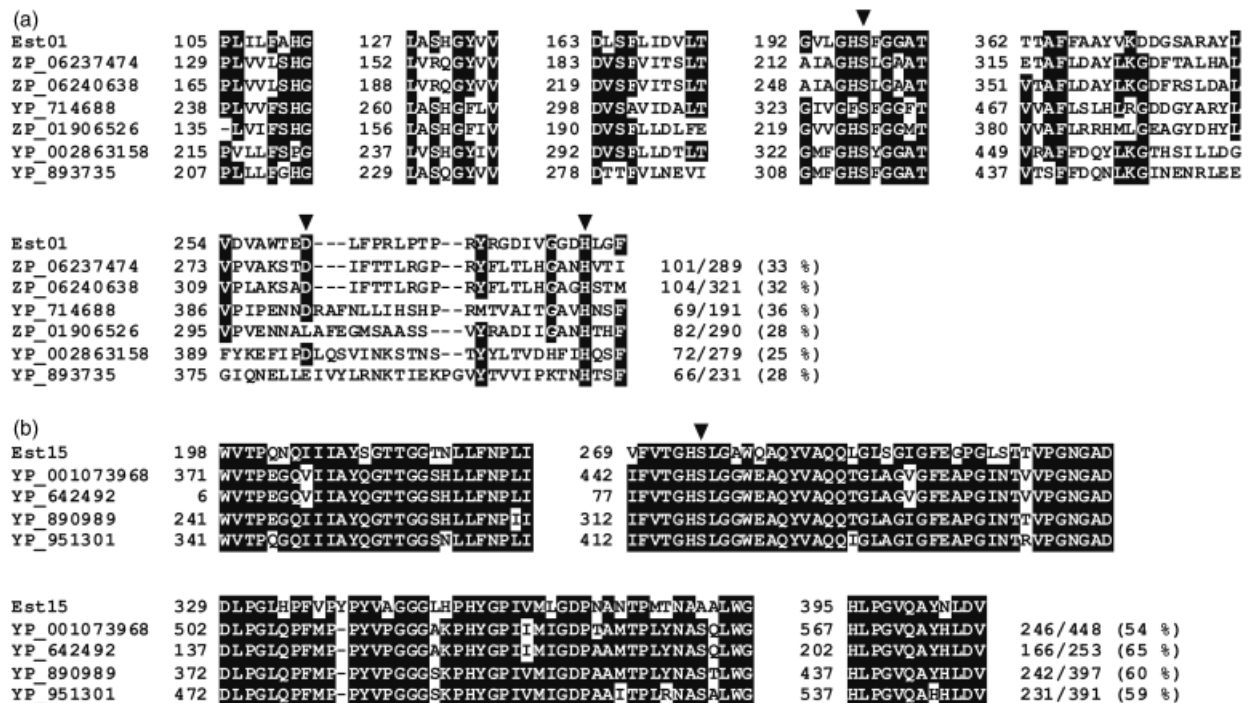
Fig. 3. Multiple sequence alignment of partial amino acid sequences harboring conserved regions of homology. Identical amino acid residues are shown as white letters on a dark background. The identity of closely related proteins to putative lipolytic enzymes identified in this study is indicated behind each sequence. Triangles indicate amino acid residues belonging to the catalytic triad. (a) Sequence alignment of Est01 with similar proteins. (b) Sequence alignment of Est15 with closely related proteins. References: Est01 and Est15 (present study); ZP_06237474, phospholipase/carboxylesterase from *Frankia* sp. EuI1c; ZP_06240638, phospholipase/carboxylesterase from *Frankia* sp. EuI1c; YP_714688, putative secreted lipase from *Frankia* ACN14a; ZP_01906526, putative secreted lipase from *Plesiocystis pacifica* SIR-1; YP_002863158, carboxylic ester hydrolase from *Clostridium botulinum* Ba4 str. 657; YP_893735, carboxylic ester hydrolase from *Bacillus thuringiensis* str. Al Hakam; YP_001073968, hypothetical protein Mjls_5714 from *Mycobacterium* sp. JLS; YP_642492, hypothetical protein Mmcs_5335 from *Mycobacterium* sp. MCS; YP_890989, hypothetical protein MSMEG_6781 from *Mycobacterium smegmatis* str. MC2 155; YP_951301, hypothetical protein Mvan_0447 from *Mycobacterium vanbaalenii* PYR-1.

Exploratory Hainich, we propose to designate the putative new family EstGH.

Putative secreted lipolytic enzymes

A potential signal peptide was predicted at the N terminus for six of the 37 deduced amino acid sequences (Est01, Est03, Est13, Est19, Est20, and Est25) using the SIGNALP 3.0 server (Bendtsen *et al.*, 2004). The number of amino acid residues of the predicted signal peptides ranged from 19 (Est25) to 32 (Est03). In addition, the amino acid sequences of all putative signal peptides showed the typical orientation of signal peptides with three distinct parts (N, H, and C domains) (Pugsley, 1993). This observation suggested that these lipolytic enzymes are secreted and function outside of the cell. Three of the putative signal peptide containing lipolytic enzymes, Est03, Est13, and Est25, belong to family V. To our knowledge, no family V lipases/esterases containing putative signal peptides have been identified in previous metagenomic studies, but signal peptide-containing family V members derived from genome sequences of individual

microorganisms such as *Psychrobacter immobilis* (CAA47949) or *Moraxella* sp. (CAA37863) are known. The three remaining signal peptide-containing lipolytic enzymes grouped into families EstGS (Est01), I (Est19), and VIII (Est20). Signal peptide-harboring enzymes of the latter two families were also observed in other metagenomic studies (Meilleur *et al.*, 2009; Rashamuse *et al.*, 2009). No signal peptides were predicted for the 17 members of family IV recovered in this study. Thus, family IV lipolytic enzymes seem to play a minor role in degrading extracellular lipids in the analyzed grassland and forest soils.

Hydrolysis of triacylglycerides and *p*-nitrophenyl esters varying in chain length

The ability to hydrolyze different triacylglycerides and *p*-nitrophenyl esters was used to determine the chain-length selectivity of the 37 unique lipolytic clones and the corresponding lipolytic gene products. Each unique lipolytic clone was plated on LB agar emulsified with tributyrin (C_4), tricaproin (C_6), tricaprylin (C_8), tricaprin (C_{10}),

trilaurin (C_{12}), trimyristin (C_{14}), or tripalmitin (C_{16}). Clones with hydrolytic activity were identified after 7 days of incubation at 37 °C by halo formation. All clones showed hydrolysis activity toward the screening substrate tributyrin (Table 3), but only *E. coli* strains carrying pLE19 exhibited hydrolysis activity toward long-chain acylglycerides such as tricaprin and trilaurin (Table 3). This result supports the sequence-based classification of Est19 into the family of true lipases (family I), as true lipases show activity toward water-insoluble long-chain triacylglycerides (Arpigny & Jaeger, 1999). The other *E. coli* clone carrying a gene encoding a putative true lipase (Est17) hydrolyzed tributyr-in, but none of the other tested triacylglycerides (Table 3). Besides tributyrin, tricaproin was hydrolyzed by the major-ity of the *E. coli* clones (21 clones) (Table 3). Thus, the substrate preference for substrates containing short-chain fatty acids ($\leq C10$) indicated that almost all enzymes were esterases.

Hydrolysis of different *p*-nitrophenyl esters was used to further analyze the substrate specificity of the 37 unique lipolytic clones using cell culture supernatant directly. The enzyme activities determined were in the range of those previously reported by Lee *et al.* (2004), who used a similar assay to initially characterize soil-derived lipolytic enzymes. All clones, except *E. coli*/pLE12, exhibited higher activity toward *p*-nitrophenyl butyrate (C_4) than the negative control strain harboring the cloning vector (Table 3). Furthermore, the analyzed clones, except *E. coli*/pLE01, showed the highest activity toward *p*-nitrophenyl butyrate (C_4) compared with the other tested *p*-nitrophenyl esters *p*-nitrophenyl caproate (C_6), *p*-nitrophenyl caprylate (C_8), *p*-nitrophenyl caprate (C_{10}), *p*-nitrophenyl laurate (C_{12}), and *p*-nitrophenyl palmitate (C_{16}). *Escherichia coli* strains carrying pLE01–pLE07, pLE10, pLE11, pLE23, pLE24, and pLE30–pLE37 also showed activity toward *p*-nitrophenyl caproate ($\geq 0.15\,U\,mL^{-1}$) and *E. coli* carrying pLE01, pLE03, pLE10, pLE11, and pLE32 toward *p*-nitrophenyl caprylate ($\geq 0.15\,U\,mL^{-1}$) (Table 3). These results sup-ports those derived from the plate assays using triacylgly-cerides in which also a preference for substrates containing short-chain fatty acids was obvious. Taking only the activities with *p*-nitrophenyl butyrate into account, *E. coli* clones harboring genes encoding family IV lipolytic en-zymes showed higher activity than the clones carrying genes encoding members of other families. In summary, as expected from the sequence analysis, the determination of the substrate specificity revealed that almost all genes encode esterases. In the case of both putative lipase-encoding genes (*est17* and *est19*), further biochemical characterization is required to verify that both enzymes are true lipases. In addition, although the activity assay did not provide quantitative data to compare activity between enzymes, it is suitable to determine substrate

Table 3. Determination of substrate specificity and chain-length selec-tivity of *Escherichia coli* cells harboring pLE01 to pLE08 and pLE10 to pLE38

Plasmid	Enzyme activity ($U\,mL^{-1}$)				
	C_4	C_6	C_8	C_{10}	C_{12}
Cloning vector	0.018	0.001	0.004	ND	ND
pLE01	0.034*	0.037*	0.018	0.001	0.004
pLE02	0.169*	0.030*	0.010	0.004	0.001
pLE03	0.119*	0.039*	0.021	0.009	0.001
pLE04	0.141*	0.024*	0.011	0.005	0.002
pLE05	0.120*	0.022*	0.007	0.007	0.004
pLE06	0.151*	0.029*	0.008	0.005	0.005
pLE07	0.116*	0.017*	0.008	ND	0.004
pLE08	0.029*	0.005	0.006	0.001	0.003
pLE10	0.204*	0.083*	0.033	ND	0.004
pLE11	0.736*	0.352	0.036	0.007	0.002
pLE12	0.018*	0.003	0.004	ND	ND
pLE13	0.024*	0.001	0.005	0.003	0.003
pLE14	0.021*	ND	0.003	0.001	0.002
pLE15	0.023*	ND	0.002	0.001	ND
pLE16	0.031*	0.002	0.005	0.005	0.001
pLE17	0.021*	ND	0.003	ND	ND
pLE18	0.028*	ND	0.004	0.001	0.001
pLE19	0.063*	0.011*	0.014*	0.006*	0.005*
pLE20	0.047*	0.004*	0.003	0.001	0.003
pLE21	0.019*	0.003	0.006	ND	ND
pLE22	0.018*	0.006*	0.008	0.006	0.003
pLE23	0.057*	0.021*	0.008	0.003	0.003
pLE24	0.080*	0.028	0.003	0.003	0.003
pLE25	0.022*	0.004*	0.003	0.003	ND
pLE26	0.034*	0.006	0.008	0.003	0.003
pLE27	0.063*	0.011	0.007	0.003	0.002
pLE28	0.037*	0.006*	0.006	0.001	0.001
pLE29	0.098*	ND	0.001	0.004	ND
pLE30	0.835*	0.289*	0.006	0.037	0.007
pLE31	0.217*	0.055*	0.003	0.004	ND
pLE32	0.110*	0.016	0.017	0.003	0.003
pLE33	0.220*	0.056*	0.006	0.008	ND
pLE34	0.238*	0.031*	0.007	0.005	0.003
pLE35	0.371*	0.097*	0.009	ND	0.002
pLE36	0.142*	0.016*	ND	0.001	ND
pLE37	0.150*	0.031*	0.005	0.002	0.002
pLE38	0.115*	ND	0.002	ND	0.001

For this purpose, a plate assay using triacylglycerides and an activity assay using *p*-nitrophenyl (pNP) esters of fatty acids as substrates were used. Hydrolysis of triacylglycerides was identified by the formation of halos on agar plates. The following pNP esters of fatty acids and triacylglycerides were used as substrates: C_4, pNP-butyrate and tributyrin; C_6, pNP-caproate and tricaproin; C_8, pNP-caprylate and tricaprylin; C_{10}, pNP-caprate and tricaprin; C_{12}, pNP-laurate and trilaurin. Trimyristin (C_{14}), pNP-palmitate (C_{16}), and tripalmitin (C_{16}) were also tested, but resulted in no activity or halo formation.
*Activity toward triacylglycerides.
ND, no activity detected.

specificity and chain-length selectivity of a large numbers of clones, which are usually recovered during metage-nomic screens.

Conclusions

A total of 37 clones conferring lipolytic activity were identified by function-driven screening of soil-derived metagenomic libraries. All of the corresponding 37 lipolytic enzymes were new members of known or putatively new lipase/esterase families, and most of the enzymes were assigned to family IV. Amino acid sequence analysis and substrate specificity showed that mainly esterases that hydrolyze esters containing short-chain fatty acids were identified.

Despite the inherent limitations and biases of cloning and activity-based screens, a diverse set of genes conferring the targeted reaction was recovered using the metagenomic approach. The novelty of the lipolytic enzymes encountered arises from the largely untapped enormous genetic diversity of uncultured soil microorganisms. This study demonstrated that soils are an important source of novel lipolytic enzymes. In addition, the results presented here showed that soil-based metagenomics and the use of a simple activity-based screening system is a method for the isolation of a large number of diverse genes conferring the targeted reaction. The number of genes and gene products performing this reaction can be extended by increasing the number of screened clones. In addition, as significant differences exist in expression modes between different taxonomic groups of prokaryotes, different hosts for the constructed libraries can be used to further expand the diversity of lipolytic and other enzymes recovered during functional screening. In this way, gene banks consisting of several hundreds of genes conferring lipolytic activity or other activities can be prepared rapidly. These gene banks or the corresponding clones can serve as the starting material for the development of novel processes and products.

Acknowledgements

Authors'contribution

H.N. and C.W. contributed equally to this work.

References

Arpigny J & Jaeger KE (1999) Bacterial lipolytic enzymes: classification and properties. *Biochem J* **343**: 177–183.

Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA & Struhl K (1987) *Current Protocols in Molecular Biology.* John Wiley & Sons Inc., New York, NY.

Bayer S, Kunert A, Ballschmiter M & Greiner-Stoeffele T (2010) Indication for a new lipolytic enzyme family: isolation and characterization of two esterases from a metagenomic library. *J Mol Microb Biotech* **18**: 181–187.

Bendtsen JD, Nielsen H, von Heijne G & Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* **340**: 783–795.

Bornscheuer UT (2002) Microbial carboxyl esterases: classification, properties and application in biocatalysis. *FEMS Microbiol Rev* **26**: 73–81.

Chu X, He H, Guo C & Sun B (2008) Identification of two novel esterases from a marine metagenomic library derived from South China Sea. *Appl Microbiol Biot* **80**: 615–625.

Cieśliński H, Białkowskaa A, Tkaczuk K, Długołecka A, Kur J & Turkiewicz M (2009) Identification and molecular modeling of a novel lipase from an Antarctic soil metagenomic library. *Pol J Microbiol* **58**: 199–204.

Couto G, Glogauer A, Faoro H, Chubatsu L, Souza E & Pedrosa F (2010) Isolation of a novel lipase from a metagenomic library derived from mangrove sediment from the south Brazilian coast. *Genet Mol Res* **9**: 514–523.

Daniel R (2005) The metagenomics of soil. *Nat Rev Microbiol* **3**: 470–478.

Elend C, Schmeisser C, Leggewie C, Babiak P, Carballeira JD, Steele HL, Reymond J-L, Jaeger KE & Streit WR (2006) Isolation and biochemical characterization of two novel metagenome-derived esterases. *Appl Environ Microb* **72**: 3637–3645.

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**: 783–791.

Ferrer M, Golyshina OV, Chernikova TN, Khachane AN, Martins dos Santos VAP, Yakimov MM, Timmis KN & Golyshin PN (2005) Microbial enzymes mined from the Urania deep-sea hypersaline anoxic basin. *Chem Biol* **12**: 895–904.

Fischer M, Bossdorf O, Gockel S *et al.* (2010) Implementing large-scale and long-term functional biodiversity research: the biodiversity exploratories. *Basic Appl Ecol* **11**: 473–485.

Hall A (1999) Bioedit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl Acids Symp Ser* **41**: 95–98.

Handelsman J (2004) Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol R* **68**: 669–685.

Hantsis-Zacharov E & Halpern M (2007) Culturable psychrotrophic bacterial communities in raw milk and their proteolytic and lipolytic traits. *Appl Environ Microb* **73**: 7162–7168.

Hårdeman F & Sjöling S (2007) Metagenomic approach for the isolation of a novel low-temperature-active lipase from uncultured bacteria of marine sediment. *FEMS Microbiol Ecol* **59**: 524–534.

Hausmann S & Jaeger KE (2010) Lipolytic enzymes from bacteria. *Handbook of Hydrocarbon and Lipid Microbiology*, *Vol. 2* (Timmis KN, ed), pp. 1099–1126. Springer, Berlin.

Heath C, Hu X, Cary S & Cowan D (2009) Identification of a novel alkaliphilic esterase active at low temperatures by screening a metagenomic library from Antarctic desert soil. *Appl Environ Microb* **75**: 4657–4659.

Henne A, Schmitz R, Bömeke M, Gottschalk G & Daniel R (2000) Screening of environmental DNA libraries for the presence of genes conferring lipolytic activity on *Escherichia coli*. *Appl Environ Microb* **66**: 3113–3116.

Hong KS, Lim HK, Chung EJ, Park EJ, Lee MH, Kim JC, Choi GJ, Cho KY & Lee SW (2007) Selection and characterization of forest soil metagenome genes encoding lipolytic enzymes. *J Microbiol Biotechn* **17**: 1655–1660.

Hotta Y, Ezaki S, Atomi H & Imanaka T (2002) Extremely stable and versatile carboxylesterase from a hyperthermophilic archaeon. *Appl Environ Microb* **68**: 3925–3931.

Hu Y, Fu C, Huang Y *et al.* (2010) Novel lipolytic genes from the microbial metagenomic library of the South China Sea marine sediment. *FEMS Microbiol Ecol* **72**: 228–237.

Jaeger KE & Eggert T (2002) Lipases for biotechnology. *Curr Opin Biotech* **13**: 390–397.

Jeon JH, Kim JT, Kang SG, Lee JH & Kim SG (2008) Characterization and its potential application of two esterases derived from the arctic sediment metagenome. *Mar Biotechnol* **11**: 307–316.

Jeon JH, Kim JT, Kim YJ, Kim HK, Lee HS, Kang SG, Kim SJ & Lee JH (2009) Cloning and characterization of a new cold-active lipase from a deep-sea sediment metagenome. *Appl Microbiol Biot* **81**: 865–874.

Kim BS, Kim SY, Park J, Park W, Hwang KY, Yoon YJ, Oh WK, Kim BY & Ahn JS (2007) Sequence-based screening for self-sufficient P450 monooxygenase from a metagenome library. *J Appl Microbiol* **102**: 1392–1400.

Kim E, Oh K, Lee M, Kang C, Oh T & Yoon J (2009) Novel cold-adapted alkaline lipase from an intertidal flat metagenome and proposal for a new family of bacterial lipases. *Appl Environ Microb* **75**: 257–260.

Kim J, Seo M, Cho E, Lee S, Kim S, Cheigh C & Pyun Y (2005) Screening and characterization of an esterase from a metagenomic library. *J Microbiol Biotechn* **15**: 1067–1072.

Lämmle K, Zipper H, Breuer M, Hauer B, Buta C, Brunner H & Rupp S (2007) Identification of novel enzymes with different hydrolytic activities by metagenome expression cloning. *J Biotechnol* **127**: 575–592.

Lee M, Lee C, Oh T, Song J & Yoon J (2006) Isolation and characterization of a novel lipase from a metagenomic library of tidal flat sediments: evidence for a new family of bacterial lipases. *Appl Environ Microb* **72**: 7406–7409.

Lee S, Won K, Lim H, Kim J, Choi G & Cho K (2004) Screening for novel lipolytic enzymes from uncultured soil microorganisms. *Appl Microbiol Biot* **65**: 720–726.

Marchler-Bauer A, Anderson JB, Derbyshire MK *et al.* (2007) CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res* **35**: D237–D240.

Meilleur C, Hupé J, Juteau P & Shareck F (2009) Isolation and characterization of a new alkali-thermostable lipase cloned from a metagenomic library. *J Ind Microbiol Biot* **36**: 853–861.

Nacke H, Thürmer A, Wollherr A, Will C, Hodac L, Herold N, Schöning I, Schrumpf M & Daniel R (2011) Pyrosequencing-based assessment of bacterial community structure along different management types in German forest and grassland soils. *PLoS One* **6**: e17000.

Pugsley AP (1993) The complete general secretory pathway in gram-negative bacteria. *Microbiol Rev* **57**: 50–108.

Ranjan R, Grover A, Kapardar R & Sharma R (2005) Isolation of novel lipolytic genes from uncultured bacteria of pond water. *Biochem Bioph Res Co* **335**: 57–65.

Rashamuse K, Magomani V, Ronneburg T & Brady D (2009) A novel family VIII carboxylesterase derived from a leachate metagenome library exhibits promiscuous beta-lactamase activity on nitrocefin. *Appl Microbiol Biot* **83**: 491–500.

Roh C & Villate F (2008) Isolation of a low-temperature adapted lipolytic enzyme from uncultivated micro-organism. *J Appl Microbiol* **105**: 116–123.

Rondon MR, August PR, Bettermann AD *et al.* (2000) Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl Environ Microb* **66**: 2541–2547.

Rosselló-Mora R & Amann R (2001) The species concept for prokaryotes. *FEMS Microbiol Rev* **25**: 39–67.

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA & Barrel B (2000) Artemis: sequence visualization and annotation. *Bioinformatics* **16**: 944–945.

Seo HJ, Bae SS, Yang SH, Lee J-H & Kim S-J (2005) *Photobacterium aplysiae* sp. nov., a lipolytic marine bacterium isolated from eggs of the sea hare *Aplysia kurodai*. *Int J Syst Evol Micr* **55**: 2293–2296.

Simon C & Daniel R (2009) Achievements and new knowledge unraveled by metagenomic approaches. *Appl Microbiol Biot* **85**: 265–276.

Simon C & Daniel R (2010) Construction of small-insert and large-insert metagenomic libraries. *Methods Mol Biol* **668**: 39–50.

Steele HL, Jaeger KE, Daniel R & Streit WR (2009) Advances in recovery of novel biocatalysts from metagenomes. *J Mol Microb Biotech* **16**: 25–37.

Strohl WR (2000) The role of natural products in modern drug discovery. *Drug Discov Today* **5**: 39–41.

Tamura K, Dudley J, Nei M & Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* **24**: 1596–1599.

Thompson J, Higgins D & Gibson T (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**: 4673–4680.

Tirawongsaroj P, Sriprang R, Harnpicharnchai P, Thongaram T, Champreda V, Tanapongpipat S, Pootanakit K & Eurwilaichitr L (2008) Novel thermophilic and thermostable lipolytic enzymes from a Thailand hot spring metagenomic library. *J Biotechnol* **133**: 42–49.

Wagner U, Petersen E, Schwab H & Kratky C (2002) EstB from *Burkholderia gladioli*: a novel esterase with a beta-lactamase fold reveals steric factors to discriminate between esterolytic and beta-lactam cleaving activity. *Protein Sci* **11**: 467–478.

Waschkowitz T, Rockstroh S & Daniel R (2009) Isolation and characterization of metalloproteases with a novel domain structure by construction and screening of metagenomic libraries. *Appl Environ Microb* **75**: 2506–2516.

Wei P, Bai L, Song W & Hao G (2009) Characterization of two soil metagenome-derived lipases with high specificity for *p*-nitrophenyl palmitate. *Arch Microbiol* **191**: 233–240.

Will C, Thürmer A, Wollherr A, Nacke H, Herold N, Schrumpf M, Gutknecht J, Wubet T, Buscot F & Daniel R (2010) Horizon-specific bacterial community composition of German grassland soils as revealed by pyrosequencing-based analysis of 16S rRNA genes. *Appl Environ Microb* **76**: 6751–6759.

Wu C & Sun B (2009) Identification of novel esterase from metagenomic library of Yangtze river. *J Microbiol Biotechn* **19**: 187–193.

Ye J, McGinnis S & Madden TL (2006) BLAST: improvements for better sequence analysis. *Nucleic Acids Res* **34**: W6–W9.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1. DNA yields, and physical and geochemical characteristics of the analyzed grassland and forest soil samples.
Table S2. Insert sizes of plasmids pLE01–pLE08 and pLE10–pLE38.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Chapter 5: Charakterisierung und Nutzung der bakteriellen Diversität in Bodenmetagenomen

Christiane Will, Heiko Nacke, Andrea Thürmer und Rolf Daniel

The purpose of this article was to give a brief summary of the current state of research. Moreover, it offered a possibility to introduce the Biodiversity Exploratories initiative and our task within this project. This article emphasizes the importance of studying soil microbial communities and illustrates phylogenetic differences between different land use types and varying sampling depth. Furthermore, it points out the metagenomic potential of soil for finding novel biocatalysts such as esterases and lipases.

Status: Published 2010 in *GenomXpress*, Vol. 1.10, p 9-11

Author contributions to the work:

Performed the experiments: CW, HN, AT.

Analyzed data: CW, HN.

Wrote the paper: CW, HN, RD.

Conceived and designed the experiments: CW, HN, RD.

Schlaglicht Biodiversität
Charakterisierung und Nutzung der bakteriellen Diversität in Bodenmetagenomen

Seit 2006 fördert die Deutsche Forschungsgemeinschaft in einem Schwerpunktprogramm das Verbundprojekt „Exploratorien zur funktionellen Biodiversitätsforschung", kurz Biodiversitäts-Exploratorien (www.biodiversity-exploratories.de). Drei Exploratorien dienen als offene Forschungsplattform für Wissenschaftler aus ganz Deutschland: das Biosphärenreservat Schorfheide-Chorin in Brandenburg, der Nationalpark Hainich und seine Umgebung in Thüringen und das Biosphärengebiet Schwäbische Alb in Baden-Württemberg. Untersucht werden die Beziehungen zwischen der Biodiversität verschiedener Taxa und Ebenen, die Rolle von Landnutzung und Management für die Biodiversität und die Rolle der Biodiversität für Ökosystemprozesse. In den Exploratorien waren und sind über 330 Mitarbeiter aus 61 Arbeitsgruppen von ingesamt 33 Forschungseinrichtungen tätig. Im Folgenden werden erste Ergebnisse aus dem Teilprojekt Boden (Biotik) mit Schwerpunkt auf der bodenmikrobiellen Ökologie dargestellt. Geplant sind der Aufbau von metagenomischen Banken, das Screening für Targets (Organismen und Funktionen) zur Idenfizierung von *key players* und parallel die bodenökologische Charakterisierung.

Christiane Will, Heiko Nacke, Andrea Thürmer und Rolf Daniel

Die mikrobielle Diversität in Böden ist sehr viel größer als die in anderen Habitaten und übertrifft um Größenordnungen die Diversität von Pflanzen und Tieren. Bakterien bilden die häufigste Gruppe der Mikroorganismen in Böden. Es wird geschätzt, dass 2.000 bis 18.000 bakterielle Arten und bis zu 109 bakterielle Zellen ein Gramm Boden besiedeln (Daniel, 2005). Bodenbakterien sind unverzichtbar für die Funktionalität von geochemischen Stoffkreisläufen, stabilisieren die Bodenstruktur und verbessern die Speicherung von Wasser im Boden. Durch die wechselnde physische, chemische und biologische Beschaffenheit des Bodens kann die Zusammensetzung von mikrobiellen Gemeinschaften mit zunehmender Bodentiefe und in unterschiedlichen Böden variieren.

Die Metagenomik ermöglicht trotz der enormen Komplexität von Bodenhabitaten die Gewinnung tiefgehender Erkenntnisse über Struktur und Funktion von mikrobiellen Gemeinschaften im Boden. Zur Erforschung der Diversität in Bodenproben werden direkte Kultivierungsansätze und indirekte molekulare Verfahren verwendet. Da aber nur ca. 1% der Bodenbakterien mit Standardmethoden kultivierbar sind, wurden zur Erschließung der Komplexität der mikrobiellen Bodengemeinschaft indirekte molekulare Verfahren entwickelt, die auf der direkten Isolierung von Nukleinsäuren (Metagenomen) aus Bodenproben basieren. Das Metagenom umfasst die Gesamtheit der mikrobiellen genetischen Information eines Standortes. Im Rahmen der hier vorliegenden Studie wurden Bodenproben der deutschen Biodiversitäts-Exploratorien Hainich-Dün, Schorfheide-Chorin und Schwäbische Alb analysiert.

Im Verlauf der Metagenomanalyse wurde die phylogenetische Diversität der Bakterien im A-Horizont (Oberboden) und im B-Horizont (Unterboden) von Grünlandbodenproben aus dem Hainich-Dün Exploratorium untersucht. Dazu wurde die Amplikon-Sequenzierung, eine Variante der Pyrosequenzierung, etabliert und angewendet. Zusätzlich wurde eine analoge phylogenetische Analyse mit A-Horizont Wald- und Grünlandbodenproben aus der Schwäbischen Alb durchgeführt, die unterschiedliche Landnutzungstypen repräsentieren. Neben der phylogeneti-

schen Analyse wurde das genetische Potential der Standorte erschlossen. In diesem Zusammenhang wurden Metagenom-Bibliotheken konstruiert und auf die Existenz von lipolytischen Enzymen durchmustert.

Phylogenetische Analyse von bakteriellen Gemeinschaften im Boden
Wald- und Grünlandflächen mit unterschiedlichen Landnutzungsintensitäten wurden für die Entnahme von Bodenproben ausgewählt. Im Wald umfasste dies vorwiegend Buchenforste aber auch Flächen mit Kiefern und Fichten. Die Nutzungsintensität reichte von Altersklassenwald über Plenterwald (d.h. ein sich stetig verjüngender Dauerwald, in dem Bäume aller Altersklassen kleinstflächig bis einzelstammweise vermischt sind) bis hin zu Naturwald. Im Grünland wurden sowohl gedüngte als auch ungedüngte Flächen herangezogen, welche als Mähweiden und Wie-

Abb. 1: Bohrkerne von Waldproben (oben) und Grünlandproben (unten). Der Durchmesser der Bohrkerne beträgt 8,3 cm und es wurden Ober- und Unterboden beprobt.

Abb. 2: Relative Verteilung von bakteriellen Phyla im Wald und Grünland sowie im A- und B- Horizont.

Schwäbische Alb (Grünland) · **Schwäbische Alb (Wald)** · **Hainich-Dün (Grünland, A-Horizont)**
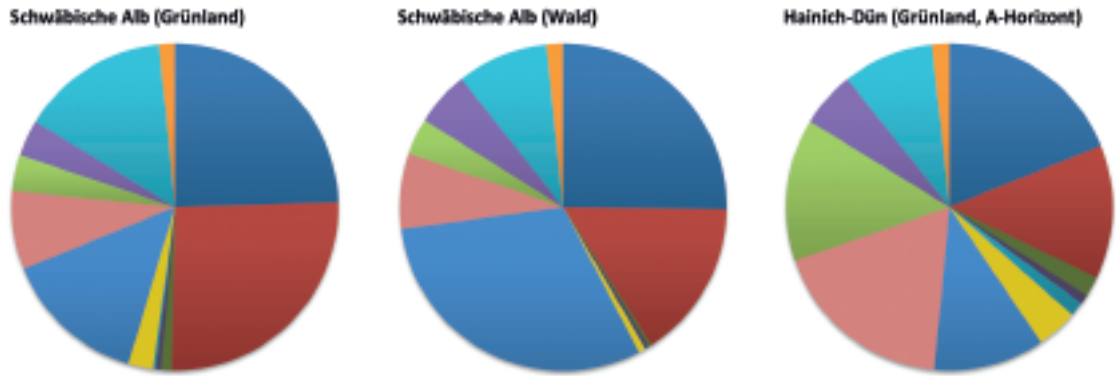
sen oder als Schafs-, Pferde- bzw. Rinderweiden dienten.

Auf jeder Entnahmestelle wurden innerhalb einer 20 x 20 m Fläche fünf Bohrkerne entnommen (Abbildung 1). Die Bohrkerne wurden in Bodenhorizonte getrennt und zu einem Feldlabor transportiert. Steine und Grobwurzeln wurden aus den Bodenproben entfernt und größere Bodenpartikel homogenisiert. Die Proben der einzelnen Entnahmestellen wurden horizontweise zu Mischproben vereint. Aus je 10 g Mischprobe des A-Horizonts wurde mikrobielle DNA isoliert. Zusätzlich wurde mikrobielle DNA aus dem B-Horizont aus den Bodenproben des Hainich-Dün-Exploratoriums isoliert. Für die phylogenetische Analyse der Bakteriengemeinschaften wurde aus der gewonnenen DNA der Bodenproben aus der Schwäbischen Alb sowie dem Grünland des Hainich-Dün eine Teilregion (V2-V3 Region) des 16S rRNA-Gens durch PCR amplifiziert und anschließend sequenziert und analysiert. Durch Vergleiche mit einer 16S rRNA-Sequenzdatenbank wurden in den 36 Proben insgesamt 1.348.962 bakterielle Sequenzen mit einer durchschnittlichen Leselänge von 259 bp identifiziert. Die dominanten bakteriellen Phyla umfassen die *Proteobacteria, Acidobacteria* und *Actinobacteria*. Es konnten signifikante Unterschiede in der Verteilung einzelner Phyla zwischen den Proben aus dem Wald und dem Grünland festgestellt werden sowie zwischen den Proben von A- und B-Horizont (Abbildung 2).

Beim Vergleich der Wald- und Grünlandproben aus der Schwäbischen Alb ist auffällig, dass *Actinobacteria* und *Firmicutes* im Grünland einen höhere Abundanz aufweisen als im Wald, wogegen die *Alphaproteobacteria* im Wald im Vergleich zum Grünland eine höhere Abundanz aufweisen. Außerdem zeigte sich, dass die relative Abundanz einiger bakterieller Phyla, Ordnungen, Familien und Gattungen innerhalb der Wald- und Grünlandproben mit dem pH-Wert des Bodens korrelierte. Im Wald nahm die relative Abundanz der *Alphaproteobacteria*, *Caulobacterales* und *Acetobacteraceae* mit zunehmendem pH-Wert ab, wogegen die der *Bacteroidetes* und *Betaproteobacteria* mit zunehmendem pH-Wert anstieg. Im Grünland erhöhte sich die relative Abundanz der *Acidobacteria* und *Acetobacteraceae* mit abnehmendem pH-Wert.

Beim Vergleich von Proben aus dem Oberboden mit denen aus dem Unterboden des Hainich-Dün Grünlands zeigt sich, dass eine größere Umverteilung der relativen Häufigkeiten stattfindet. Auffällig ist, dass die relative Abundanz der *Acidobacteria* und die der *Chloroflexi* von oben nach unten in der Bodensäule stark zunimmt. Dafür reduziert sich der Anteil der *Actinobacteria*, *Alpha-*, *Beta-* und *Gammaproteobacteria*. Insgesamt gesehen nimmt sowohl die Biomasse als auch die Anzahl der auftretenden Phyla oder Spezies im Bodenprofil von oben nach unten ab.

Lipase/ Esterase-Familie	Anzahl der gefundenen Vertreter	Proteinsequenz-ähnlichkeit zu bekannten lipolytischen Enzymen (%)	Bekanntes Protein mit der höchsten Ähnlichkeit (Zugangsnummer in GenBank-Datenbank)	Bekanntes Protein mit der geringsten Ähnlichkeit (Zugangsnummer in GenBank-Datenbank)
I	2	44 – 62	Lipase, Klasse 2 aus *Mycobacterium vanbaalenii* PYR-1 (YP_953514)	Vorhergesagte Acetyltransferase/ Hydrolase aus *Tsukamurella paurometabola* DSM 20162 (ZP_04028715)
IV	17	48 – 87	Lipase/Esterase aus einem unkultivierten Bakterium (AAS77236)	Lipase/Esterase aus einem unkultivierten Bakterium (AAS77247)
V	10	32 – 65	Alpha/Beta Hydrolase aus *Methylobacterium* sp. 446 (YP_001773617)	Vorhergesagte Hydrolase oder Acyltransferase aus *Lentisphaera araneosa* HTCC2155 (ZP_01875730)
VI	1	57	Phospholipase/Carboxylesterase aus *Nitrosospira multiformis* ATCC 25196 (YP_413093)	Phospholipase/Carboxylesterase aus *Nitrosospira multiformis* ATCC 25196 (YP_413093)
VIII	6	39 – 90	Beta-Lactamase aus *Caulobacter* sp. K31 (YP_001682441)	Beta-Lactamase aus *Burkholderia* sp. H160 (ZP_03264354)

Tab. 1. Einordnung der identifizierten Genprodukte in bekannte Lipase/Esterase-Familien und Sequenzähnlichkeiten.

Hainich-Dün (Grünland, B-Horizont)

- Acidobacteria
- Actinobacteria
- Bacteroidetes
- Chloroflexi
- Cyanobacteria
- Firmicutes
- Alphaproteobacteria
- Betaproteobacteria
- Gammaproteobacteria
- Deltaproteobacteria
- unclassified Proteobacteria
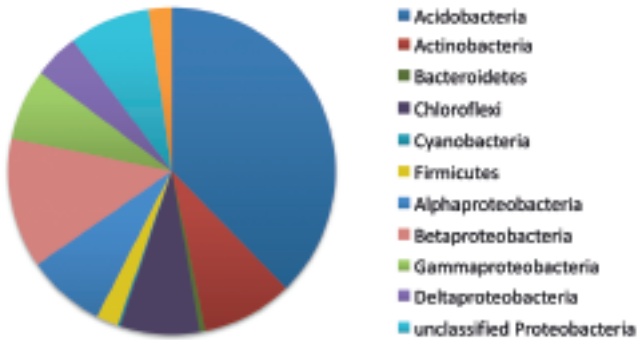
Mit dieser umfangreichen Studie über verschiedene Bodenproben aus Deutschland konnte ein tiefgehender Einblick in die bakterielle Biodiversität und die Veränderung der Verteilung im Habitat Boden gewonnen werden. Dabei zeigte sich, dass der pH-Wert und der Bodenhorizont einen Einfluss auf die Zusammensetzung von bakteriellen Gemeinschaften haben.

Das genetische Potential des Boden-Metagenoms

In Anbetracht der Tatsache, dass ein Gramm Boden bis 2×10^9 prokaryotische Zellen beinhaltet, die zum überwiegenden Teil im Labor nicht kultivierbar sind, bietet dieses Habitat ein nahezu unerschöpfliches Potential an Genen für neuartige Biokatalysatoren und andere Biomoleküle.

Eine industriell bedeutende Rolle spielen dabei Lipasen (EC 3.1.1.3) und Esterasen (EC 3.1.1.1). Diese finden biotechnologische Anwendung als Katalysator bei der Synthese von Polymeren und bei der Herstellung von Biodiesel. Des Weiteren können Esterasen bei der Produktion von chemischen Reinstoffen wie zum Beispiel Arzneimitteln, Herbiziden, Kosmetika, Aroma- und Duftstoffen eingesetzt werden.

Um die Bodenproben auf neuartige Lipasen und Esterasen zu durchmustern, wurden metagenomische Genbibliotheken in Plasmiden und Fosmiden aus der isolierten DNA angelegt. Die insgesamt 32 Plasmidbanken umfassen 29 Gb klonierte DNA, die 19 Fosmidbanken 64 Gb. Davon wurden 11 Gb DNA (Plasmidbanken) bzw. 1,17 Gb DNA (Fosmidbanken) auf das Vorhandensein von Genen, die für lipolytische Aktivität kodieren, untersucht. Das Screening erfolgte mittels Plattentest, indem das Triglycerid Tributyrin als Indikator-

substanz fungiert. Auf diesen Testplatten wurden die metagenomischen Banken in *E. coli* als Wirt ausplattiert. Klone, welche in der Lage sind, die Esterbindungen des Testsubstrats in Glycerin und Butyrat zu spalten, zeigen einen Aufklärungshof um die Kolonie (Abbildung 3). Insert-DNA von positiven Klonen wurde sequenziert und analysiert. Mit dieser Methode konnten 36 unterschiedliche Gene für potentielle Esterasen gefunden werden. Die korrespondierenden Klone wurden auch auf Agarplatten mit längerkettigen Triglyceriden getestet. Dabei zeigte sich, dass Trihexanoat (C6) überwiegend abgebaut werden konnte, während allerdings auf Agarplatten mit Trioctanoat (C8) nur einmal Aktivität zu erkennen waren. Einige der Esterase-Gene wurden in Expressionsvektoren subkloniert und exprimiert. Auch hier zeigte sich, dass Substrate mit kurzkettigen Fettsäuren in der Regel bevorzugt wurden. Es konnte für alle Substrate eine erhöhte Enzymaktivität bei steigenden Temperaturen festgestellt werden. Es wurden Esterasen gefunden die bei extremen pH-Werten (pH 3 bis 4 und pH 10 bis 12) sowie über einen längeren Zeitraum bei 60 °C keinen nennenswerten Aktivitätsverlust zeigten.

Die 36 gefundenen Esterasen wurden in die acht Familien der lipolytischen Enzyme einsortiert (siehe Tabelle 1). Sie entstammen zum größten Teil den Familien IV, V und VIII, weiterhin sind auch die Familien I und VI vertreten. Familie I umfasst echte Lipasen. Diese zeigen eine höhere Aktivität gegenüber wasserunlöslichen längerkettigen Triglyceriden während Esterasen kurzkettige wasserlösliche Triglyceride bevorzugen. Die Ähnlichkeiten zu bereits bekannten Esterasen reichen von 32 bis 90 %. Die Größen der lipolytischen Genprodukte bewegen sich zwischen 111 und 556 Aminosäuren. Da Esterasen weit verbreitet sind und in allen untersuchten Bodenproben vorkamen, eignen sie sich gut, um die Diversität eines Standortes widerzuspiegeln. Einige der hier gefundenen Esterasen haben aufgrund ihrer Temperatur- und pH-Stabilität biotechnologisches Anwendungspotential.
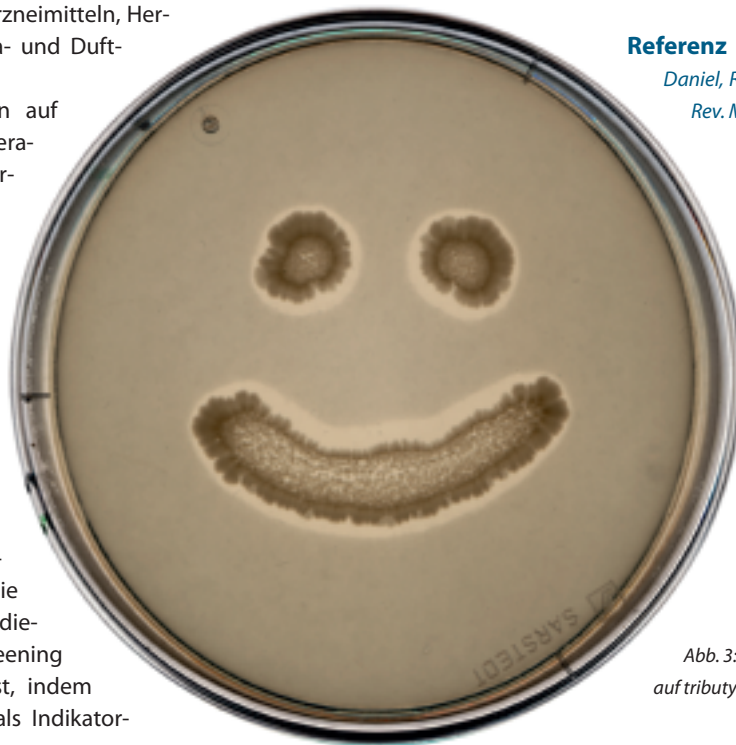
Referenz

Daniel, R. (2005) *The metagenomics of soil. Nat. Rev. Microbiol.* 3:470-478.

Kontakt

PD Dr. Rolf Daniel
Georg-August-Universität Göttingen
E-Mail: rdaniel@gwdg.de

Abb. 3: Klon mit lipolytischer Aktivität auf tributyrinhaltigem Agar.

Chapter 6: Discussion

Metagenomics combined with next-generation sequencing approaches is a powerful method to provide an in-depth analysis of microbial diversity in different habitats. In addition, metagenomics has proven to be an effective tool for the detection of novel bioactive molecules. In this way, the metabolic potential of uncultured microorganisms has been made available for biotechnological applications (Ferrer et al. 2009, Simon and Daniel 2009).

In this thesis, multiple soil samples have been analyzed. These comprised different land use and management types across two soil horizons. In one approach, grassland samples from topsoil and the corresponding subsoil samples were investigated with respect to the prokaryotic diversity and community composition. In another approach, topsoil samples from grasslands were compared with those from forest. The environmental DNA was isolated from each sample and employed in a PCR reaction to amplify partial 16S rRNA gene sequences. These were subsequently subjected to 454-pyrosequencing. The obtained sequence data were analyzed with various statistical methods to assess taxonomic compositions in these habitats. Moreover, correlations of soil properties with the sequence data were conducted to discover their influence on the prokaryotic diversity.

Another objective of this thesis was the discovery of novel enzymes with lipolytic activity in the soil samples. For this purpose, metagenomic libraries were constructed from the environmental DNA. Both, small-insert libraries in plasmids and large-insert libraries in fosmids were screened for genes that confer esterase activity. The inserts of the positive recombinant clones were sequenced and characterized partially. Figure 4 gives a schematic overview of the tasks fulfilled during this thesis.
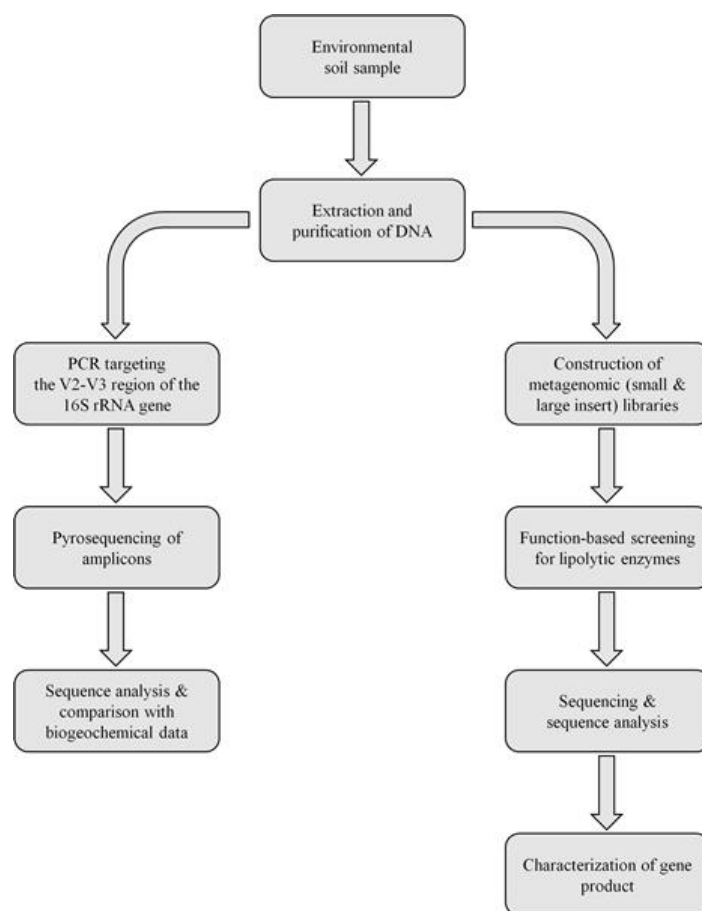
Fig. 4: Workflow of the approaches for the metagenomic analyses of different soil samples.

6.1 Phylogenetic analyses of the habitat soil

The dimension of microbial diversity in soil ranges from 2,000 to 18,000 different genomes per gram (Daniel 2005). This emphasizes that microorganisms form the overall majority in soil. Moreover, they comprise a large portion of the genetic diversity on Earth (van der Heijden et al. 2008). The question about the dimension of microbial diversity in soil remains unanswered because the determination of the amount of species could not at all be addressed with the present approaches (López-García and Moreira 2008). With the availability of the recently established next-generation sequencing techniques the amount of sequence information has increased considerably. For example, 87 data sets with the keyword "soil" have already been submitted to the NCBI Sequence

Read Archive (http://www.ncbi.nlm.nih.gov/sra, visited 11/18/2010). However, these are still not sufficient to cover the prokaryotic diversity in a complex habitat such as soil on a species level. At a genetic distance of 3%, which represents species level, the performed sampling effort is still far from uncovering the full extent of soil species diversity for the majority of the published pyrosequencing-based studies. Acosta-Martínez et al. (2008) plotted 6,000-7,000 sequences from four soil samples derived from different land use and management types against the number of observed OTUs to generate rarefaction curves (Fig. 5A). At 6,000 sequences saturation is not yet detectable. Roesch et al. (2007), however, detected a stagnation of the rarefaction curves for their Canadian forest sample at approximately 40,000 sequences (Fig. 5B). They characterize this sample as phylum rich and species poor, but in the studied grassland samples no saturation of the curves was visible (Fig. 5B). In the oak rhizosphere and bulk soil in the study of Uroz et al. (2010) a steady increase in the number of OTUs up to 37,358-62,022 sequences was found. All these results are in accordance with the findings obtained in this thesis: Neither in the Hainich topsoil samples nor in the forest or grassland samples from the Schwäbische Alb the full species extent was covered (see Chapter 2, Fig. 1 and Chapter 3, Fig. 1). Pyrosequencing-based assessments of soil bacterial diversity yielded varying sequence extents and the number of OTUs in the rarefaction curves differed strongly (Roesch et al. 2007, Acosta-Martínez et al. 2008, Lauber et al. 2009, Uroz et al. 2010). It is proven that the number of OTUs is strongly affected by the number of analyzed sequences (Schloss and Handelsman 2005, Roesch et al. 2007). An
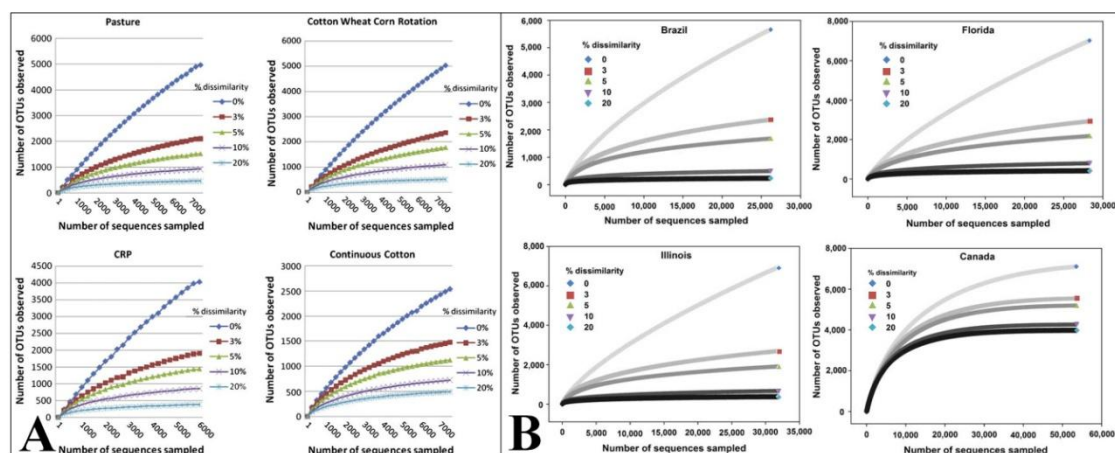
Fig. 5: Rarefaction curves from two other studies depicting the effect of % dissimilarity on the number of OTUs they identified.

A: Study of soil samples derived from different land use and management types (Acosta-Martínez et al. 2008) B: four soils of different land use across America (modified from Roesch et al. 2007).

artificial reduction of a large pyrosequencing data set from 53,533 sequences to 10,000 sequences resulted in two rarefactions curves, each with a saturation either at 5,500 OTUs (original data set) or 1,500 OTUs (reduced data set) (Roesch et al. 2007).

The sequence amount for the grassland samples derived from Hainich subsoil comprised 25,851 to 47,264 sequences, which was in the same range as the sequence amount of the topsoil samples. The bacterial diversity represented by the detected number of OTUs was significantly lower in the subsoil, though (Chapter 2, Table 2). The outcomes of this study are consistent with the findings of a continuous soil profile in Oak Ridge (TN, USA). Ninety-six 16S rDNA clones per soil horizon were analyzed and the detected number of species decreased significantly with depth (Hansel et al. 2008). A decrease of PLFA (phospholipid fatty acid) richness, which the authors considered as an indicator of overall microbial community richness, was also detected for two soil profiles in the Sedgwick Reserve, Santa Barbara (CA, USA) from topsoil to subsoil (Fierer et al. 2003).

At a genetic distance of 20%, which represents phylum level, the surveying effort covered (almost) the full extent of taxonomic diversity (Roesch et al. 2007, Acosta-Martínez et al. 2008). This is in accordance with the findings obtained in this thesis: Both, for the Hainich grassland samples as well as for the grassland and forest samples from the Schwäbische Alb a (near) saturation of the rarefaction curves could be detected at phylum level (Chapter 2, Fig. 1 and Chapter 3, Fig. S1 in the SI [Supplemental Information]). The achieved coverage according to the richness estimator Chao1 was 98% for forest and 93% and 96% for grassland in the Hainich and the Schwäbische Alb, respectively (Chapter 3, SI Table S3 and Chapter 2, Table 2). Interestingly, the average number of detected OTUs at phylum level is twice as high in grassland (207 OTUs in the Hainich and 193 in the Schwäbische Alb) compared to forest (99 OTUs). Nevertheless, the coverage of the grassland sites is below that of the forest sites. Thus, grassland has a higher complexity of the bacterial community at phylum level. This is in contrast to the result by Roesch et al. (2007) who described the bacterial diversity in forest to be phylum rich and species poor and in agricultural soils species rich and phylum poor. However, the results from this thesis confirm the finding that the oak rhizosphere is less complex than the surrounding soil (Uroz et al. 2010). In Chapter 3 an influence of the tree species is assumed to explain the disparity in the bacterial diversity between spruce and beech forest. The spruce forest samples show a higher bacterial diversity at phylum

level than the beech samples (Chapter 3, SI Table S3 and Fig. S1). Unfortunately, the tree species which grows in the Canadian forest is only described as boreal forest (Roesch et al. 2007). Anyhow, soil samples from coniferous forests such as spruce (Chapter 3) or boreal forest (Roesch et al. 2007) contain more phyla than soil samples from deciduous forests such as beech (Chapter 3) or oak forests (Uroz et al. 2010). Since pyrosequencing can currently not provide a satisfying coverage on a species level the initial question about the dimension of soil-borne microorganisms cannot yet be resolved with the available techniques.

Although the analysis of the 16S rRNA gene has been established to access the diversity and community composition in a wide range of habitats, this method has some limitations. First of all the results of any PCR-based method are biased due to the amplification reaction. Sequencing errors can cause an overestimation of the diversity as every sequence read is treated as an unique identifier of a community member. This is crucial especially for large-scale pyrosequencing analyses (Kunin et al. 2010). Additionally, the resolving power is limited close to the species level and even inter-strain differences of up to 16% have been observed (Klenk and Göker 2010). Moreover, the copy number of the 16S rRNA gene varies between the different phylogenetic groups (Rastogi et al. 2009). However, the 16S rRNA gene is highly conserved and easy to access. Sequencing, especially pyrosequencing, can be performed at reasonable cost (Klenk and Göker 2010). Since the phylogenetic classification by 16S rRNA gene analyses is well established, a wide range of software tools and databases for the analyses and comparison are currently available (DeSantis et al. 2006, Pruesse et al. 2007, Cole et al. 2009).

To assess the overall distribution of bacterial taxa in the habitat soil, Janssen conducted a meta-analysis of numerous 16S rRNA clone libraries (Janssen 2006). He analyzed different libraries of 16S rRNA and 16S rRNA genes which had been constructed from a variety of soil samples. The survey comprised 2,920 clones derived from different vegetation types across Europe and America. Janssen observed the appearance of at least 32 phyla in this data set of the soil samples. However, an average of 92% of a soil microbial community comprised nine dominant phyla: *Proteobacteria*, *Acidobacteria*, *Actinobacteria*, *Verrucomicrobia*, *Bacteroidetes*, *Chloroflexi*, *Planctomycetes*, *Gemmatimonadetes*, and *Firmicutes*. These were omnipresent and accounted for an average of 39, 20, 13, 7, 5, 3, 2, 2, and 2%, respectively, of a soil bacterial community (Fig. 6). However, the variance of the relative abundances in a single soil was very high: For example the

relative abundance of *Proteobacteria* ranged from 10 to 77% and that of *Acidobacteria* from 5 to 46% (Fig. 6). Although this analysis of soil bacterial diversity has been conducted on clone libraries with an overall number of 2,920 sequences, it is astonishingly congruent with large-scale pyrosequencing studies on a phylum level. *Proteobacteria* are likewise the most abundant phylum in the surveys of one soil type under four different management regimes (Acosta-Martínez et al. 2008), in the agricultural and forest soils analyzed by Roesch et al. (2007), or in the oak rhizosphere and surrounding soil from Uroz et al.
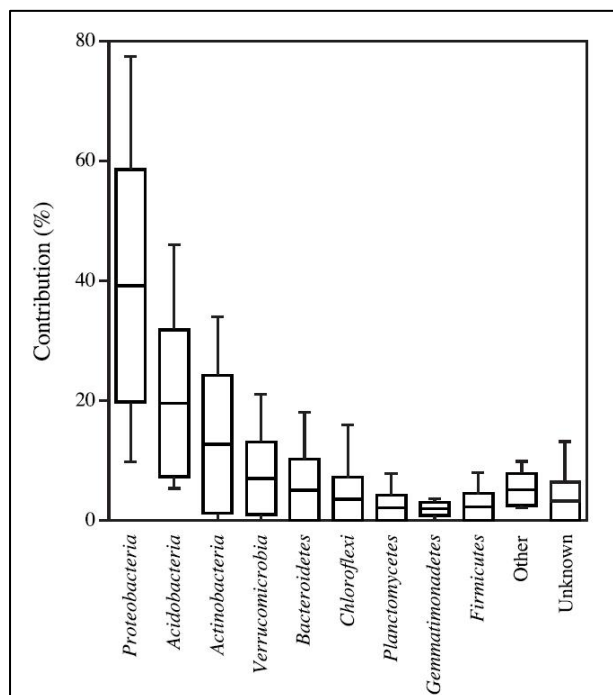
Fig. 6: Contributions of 16S rRNA genes from members of different phyla in a meta-analysis of soil bacterial communities in Europe and America.

The horizontal line in the middle of each block indicates the mean, the block represents 1 standard deviation on either side of the mean, and the vertical lines extending above and below each block indicate the minimum and maximum contributions of each phylum (Janssen, 2006).

(2010). These studies obtained an average of approximately 30 to 45% of proteobacterial sequences (Fig. 7). The relative abundance of *Proteobacteria* in the Hainich topsoil samples (Chapter 2) was 57.8% and thereby 1.5-times higher than that reported by Janssen (2006), but the amount is within the observed range of 10 to 77%. The portions of *Proteobacteria* in the forest and grassland samples from the Schwäbische Alb (Chapter 3) are intermediate with 45.5 and 34.9%, respectively. The only pyrosequencing-based study in which *Proteobacteria* were not the most abundant phylum was the analysis of the impact of pH on soil bacterial communities (Lauber et al. 2009). There, the average portion of *Proteobacteria* was 29.6% and thereby also in range of the results by Janssen (2006), but the phylum *Acidobacteria* was more abundant (30.9%). This is caused by the sharp increase of the relative abundance of *Acidobacteria* in soils with a pH ≤5, which occurred frequently in the Lauber study. Comparisons of the relative abundances of the individual proteobacterial classes again were in good agreement with the results from Janssen (2006) and the results obtained in this thesis. In the meta-analysis the majority within the *Proteobacteria* was formed by *Alphaproteobacteria*,
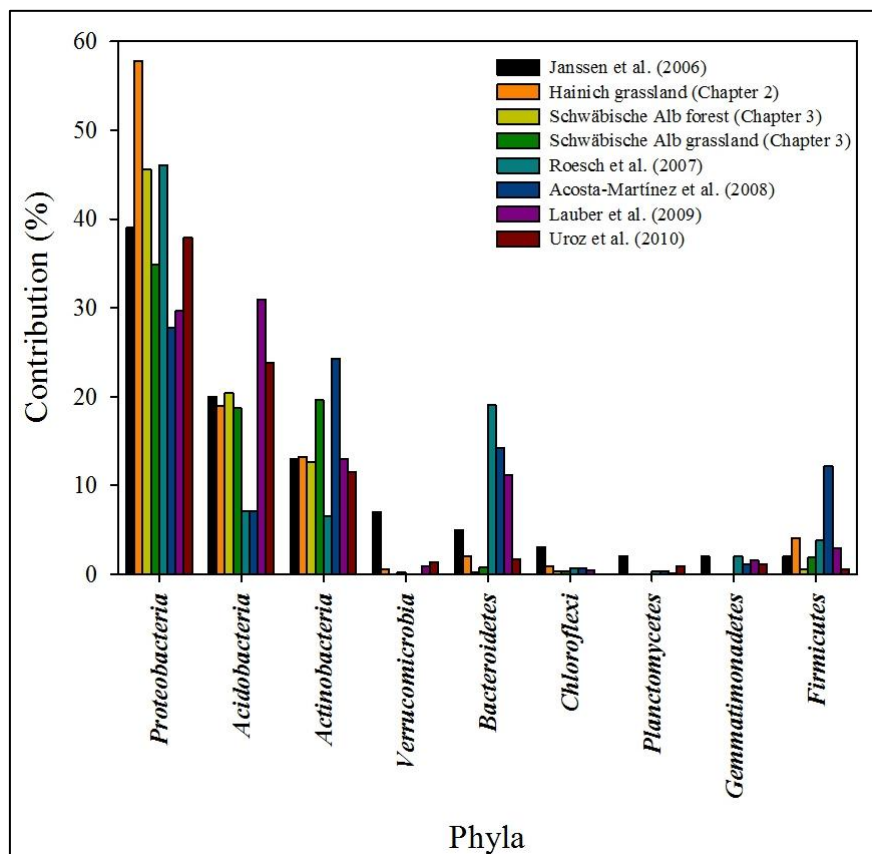
Fig. 7: Contributions of 16S rRNA genes from members of dominant phyla in different studies which assessed the bacterial community composition in soil.

Those nine phyla that make up for 92% of a soil bacterial community according to the meta-analysis conducted by Janssen (2006) were chosen for the comparison. Depicted is the relative abundance of each phylum across all samples of the respective study.

followed by *Betaproteobacteria*, and *Gammaproteobacteria*, representing a relative abundance of 18.8, 10.0, and 8.1%, respectively (Fig. 8). *Delta-* and *Epsilonproteo-bacteria* accounted for 2.3 and 0.04%. The only considerable difference in the present study was that *Betaproteobacteria* were more abundant in the Hainich samples than *Alphaproteobacteria* (18.2 and 11.0%, respectively), but these results and all findings from the Schwäbische Alb were consistent with the ranges described by Janssen (2006). However, the relative abundances of *Alpha-* and *Betaproteobacteria* in the Hainich samples resemble those in the study conducted by Roesch et al. (2007) in which *Alpha-proteobacteria* comprised an average of approximately 11% and *Betaproteobacteria* 18% of all classified sequences (Fig. 8).

The second most abundant phylum in soil according to Janssen are the *Acidobacteria* which accounted for an average of 20% among all sequences (Fig. 6). This portion was also obtained in the samples of this thesis as well as in the survey of Uroz et al. (2010)

(Fig. 7). *Acidobacteria* are strongly pH-dependent, which causes those discrepancies. Within this phylum the most abundant subgroups in Janssen's analysis were subgroup 4, 6, and 1, which represent 39.1, 23.8, and 16.8% of all detected acidobacterial sequences (Janssen 2006) (Fig. 9). A pyrosequencing-based research that focused on the distribution of the acidobacterial subgroups in soil (Jones et al. 2009) detected higher relative abundances of subgroups 1, 2, and 3 than Janssen but lower relative abundances of subgroups 4, 6, and 7. However, they also found that the subgroups 1-3 correlate negatively with pH while the other three correlate positively. Consequently, the average soil pH in the samples of Janssen was probably higher than the average pH in the samples of Jones et al. (2009). A comparison of the distributions of the individual *Acidobacteria* subgroups revealed that subgroup 1 occurred neither in the Hainich grassland nor in the Schwäbische Alb grassland although it formed a major portion in the other studies (Janssen 2006, Jones et al. 2009, Lauber et al. 2009) (Fig. 9). Since members of subgroup 1 comprised 14.4% of

Fig. 8: Average relative abundances of proteobacterial classes within this thesis and two other studies.

The second column represents the results from Chapter 2, the third and fourth those from Chapter 3. Depicted are only the four major classes*: Alphaproteobacteria* (red), *Betaproteobacteria* (orange), *Gammaproteobacteria* (yellow), and *Deltaproteobacteria* (green).

all acidobacterial sequences in the forest samples of the Schwäbische Alb (Fig. 9), a bias in the methods for DNA extraction or sequencing employed in the present study can be excluded. The negative correlation of subgroup 1 with pH (Jones et al. 2009, Lauber et al. 2009) explained that almost no members of that acidobacterial subgroup occur in the grassland samples with (near) neutral pH. Another abundant acidobacterial subgroup is subgroup 6. The abundance of this subgroup is also strongly pH-dependent and correlates positively with increasing pH. The relative abundance among all acidobacterial sequences of this subgroup ranges between 7.8% in rather acidic soil samples (Lauber et al. 2009) and 46.2% in the neutral Hainich grassland (Fig. 9). In 2007, the knowledge on *Acidobacteria* had improved with the discovery of 26 subgroups instead of the previously known 7 subgroups (Barns et al. 2007). With this background, the classification performed by Janssen (2006) is problematic from today's perspective because some of their acidobacterial sequences in subgroups 1-7 might be now classified as members of

subgroups 8-26. However, the only one of the newly discovered *Acidobacteria* subgroups whose abundance is of importance is subgroup 16. The relative abundance of this subgroup increases with pH. It accounts for 13.4 to 36.3% in the (near) neutral grassland samples from this study (Fig. 9).

The third most abundant phylum in soil according to the survey of Janssen are the *Actinobacteria* (Janssen 2006). He obtained a relative abundance of 13% average ranging from 0 to 34% (Fig. 6 and 7). This corresponded with the results from this thesis: The grassland samples from the Hainich and the Schwäbische Alb showed an average relative abundance of 13.2 and 12.7%, respectively. Other studies revealed relative abundances in a similar range of 12 to 13% (Lauber et al. 2009, Uroz et al. 2010) (Fig. 7). Only the study of Roesch et al. (2007) obtained a significantly lower relative abundance of *Actinobacteria* (6.5%), but an explanation was not evident.

Fig. 9: Average relative abundances of acidobacterial subgroups within this thesis and three other studies.

The second column represents the results from Chapter 2, the third and fourth those from Chapter 3. Depicted are the relative abundances of the respective subgroups according to all acidobacterial sequences in the specific survey.

The remaining six phyla which together with *Proteobacteria*, *Acidobacteria*, and *Actinobacteria* account for 92% of a soil microbial community all occur with a relative abundance of 2-7% in the meta-analysis conducted by Janssen (2006) (Fig. 6 and 7). In the samples from the present thesis all of these phyla except for the *Firmicutes* have a relative abundance of less than 1%. The portions of these phyla in other studies are generally very low (Roesch et al. 2007, Lauber et al. 2009, Uroz et al. 2010). However, discrepancies in the relative abundances of these phyla between the studies are visible (Fig. 7). These discrepancies have several explanations: A cloning bias in the libraries analyzed by Janssen and the choice of different primer sets, which had been employed in all surveys, are factors that potentially influence the relative abundances of minor phyla crucially. Another explanation is the enormous differences in the sequence

extents. Janssen investigated a total of 2,920 clones while the pyrosequencing-based studies comprise up to 752,838 sequences like the analysis of Hainich grassland soils (Chapter 2). Nevertheless, also the influence of soil properties must not be disregarded when studying the bacterial composition in soils. It is known that various soil properties are important drivers of bacterial community structures (Fierer and Jackson 2006, Lauber et al. 2008, Mocali and Benedetti 2010). As mentioned above, soil pH has major influence on the relative abundance of *Acidobacteria* as well as on the composition of the subgroups (Jones et al. 2009). The importance of pH for the soil bacterial community structure was also detected for the Schwäbische Alb samples on several phylogenetic resolutions (Chapter 3, Fig. 8 and SI Table 10). An interdependency of soil characteristics such as organic carbon (OC) content or soil carbon-to-nitrogen (C/N) ratio with pH have been described previously (Jones et al. 2009). Some researchers assumed particularly the carbon content to be a major driver for microbial community compositions (Zhou et al. 2002, Fierer et al. 2007, Rousk et al. 2010). Since the Hainich soil samples were all near neutral and no correlation of pH values and the bacterial community composition were found, the OC content was the main driver for community structure in these samples (Chapter 2).

In conclusion pH is a major characteristic regarding the composition of microbial soil communities. The relative abundances of some phylogenetic groups are related to other factors such as the OC content. A lot of research is still necessary to understand soil microbial processes and interactions with soil properties and to bring the results of different studies in accordance. However, the latter is difficult as long as lots of parameters vary between the surveys.

6.2 Relevance of lipolytic enzymes for biotechnology processes

In 2003 global industrial enzyme sales were estimated at \$2.3 billion. The sectors that contributed to the majority of these sales were detergents (\$789 million), food applications (\$634 million), agriculture/feed (\$376 million), textile processing (\$237 million), and pulp/paper, leather and other applications including enzymes for the production of fine and bulk chemicals (\$222 million) (Lorenz and Eck 2005). In 2003, the market share of biotechnology was 5% but it had already increased to 10% this year. Due to the

need for environmentally friendly, resource-conserving, and profitable biotechnology to replace or complement current industrial production processes a further increase in the market share of biotechnology is predicted for the coming years (Jaeger and Holliger 2010). A prerequisite, however, is the continuous discovery of novel bioactive molecules with industrial potential.

Two general types for screening metagenomes have been established to detect novel enzymes: The sequence-driven approach, which is based on the nucleotide sequence and the function-driven approach, which is based on metabolic activities. The sequence-based approach relies on sequence homologies to known genes or corresponding enzymes. PCR or hybridization are the common methods for this strategy. The genes of interest are targeted either by primers or probes from conserved regions. Although the sequence-based screening method has the advantage that it can be performed without the previous construction of metagenomic libraries it is limited to the detection of new members of already known gene families (Daniel 2005, Simon and Daniel 2009). The function-based approach instead requires the construction of metagenomic libraries and is dependent on a simple activity-based screening method as a high number of clones usually needs to be tested. However, since the knowledge of sequence information is not required, it has the advantage that entirely new genes or classes of genes can be detected (Kim et al. 2009, Lee et al. 2010). In addition, positive clones express functional gene products. Nevertheless, the function-driven screening approach is limited by the fact that the metagenomic gene of interest must be heterologously expressed in the host organism. Usually, *Escherichia coli* is employed as host organism, although it has been demonstrated that only approximately 40% of the enzymes from other organisms can be heterologously expressed in *E. coli* (Gabor et al. 2004). However, *E. coli* is easily manageable under laboratory conditions.

The functional screening approach has already been employed to obtain lipolytic enzymes from numerous metagenomic libraries. Lipolytic enzymes are known to be versatile enzymes which can serve as biocatalysts in a wide range of applications. Fields of applications are food, detergent, chemical, and pharmaceutical industries (Gupta et al. 2004, Kim et al. 2009). Particularly the fine chemicals industry benefits from the enantioselectivity of lipases. As a general rule only one out of two possible enantiomers functions, e.g., as therapeutic, cosmetics, flavor, or agrochemical and especially in the pharmacy sector an enantiopure synthesis of drugs is compulsory (Jaeger and Eggert

2002). A broad spectrum of habitats has already successfully been screened for the presence of lipolytic enzymes such as various sediments (Hårdeman and Sjöling 2007, Jeon et al. 2009, Couto et al. 2010, Hu et al. 2010) as well as several aquatic samples from freshwater (Ranjan et al. 2005, Wu and Sun 2009) and marine environments (Ferrer et al. 2005a, Chu et al. 2008). Moreover, hot springs (Kim et al. 2005, Rhee et al. 2005, Tirawongsaroj et al. 2008), the bovine and sheep rumen microflora (Ferrer et al. 2005b, Bayer et al. 2010), and compost samples (Lämmle et al. 2007, Kim et al. 2010) have been subjected to functional screening for esterases. Also a great variety of soil samples (Lee et al. 2004, Hong et al. 2007, Heath et al. 2009, Wei et al. 2009, Bunterngsook et al. 2010, Lee et al. 2010) has already been screened for genes that confer lipolytic activity. This demonstrates that the habitat soil harbors an enormous amount of biocatalysts, lipolytic enzymes in this particular case. By employing meta-genomic approaches, soil-born microorganisms will continue to be the main reservoir of novel bioactive molecules (Daniel 2005, Li et al. 2008).

Some of the lipases and esterases that were derived from soil metagenomes displayed unusual properties which could partially be related to the habitat they originated from. Two lipases that were detected in soil samples from different altitudes (400 and 1,200m) and had been obtained through screening at different temperatures (50°C and 28°C, respectively) showed their maximum activity and stability at 40°C and 20°C, respectively (Wei et al. 2009). A thermostable esterase from mud of a hot spring displayed activity at a temperature range of 30 to 90°C (Byun et al. 2006). Esterases that were obtained from a acidic peat-swamp forest soil (pH 5) were active at pH 5. Interestingly, some of the clones were also able to hydrolyze the screening substrate tributyrin in alkaline medium at pH 10 (Bunterngsook et al. 2010). A fosmid library constructed from Antarctic desert soil bore an esterase with an optimum pH of 9, an optimum temperature of 40°C, and a constant stability at 30°C. These values seem to be moderate, nevertheless this esterase is a potential candidate for industrial application, since it exhibits activity between 7°C and 54°C and over a pH range of 5.5 to 11.0 (Heath et al. 2009). Another esterase derived from vegetable soil was stable over a pH range of 5.5 to 9.0 (Li et al. 2008).

By employing extreme habitats with more unusual properties than common soil for the construction of metagenomic libraries and subsequent screening for lipolytic activity, the chance to detect enzymes with unusual characteristics might be higher. For example

two esterases obtained from hot springs showed maximum activity at 70°C (Tirawongsaroj et al. 2008). Additionally, another esterase detected in deep sea had maximum activity under marine conditions regarding pH and salinity (Hu et al. 2010). However, extreme habitats are disproportionately more difficult to access than the habitat soil. The functional screening method for the detection of lipolytic enzymes is usually a plate screening with tributyrin-containing media. The ability to hydrolyze the substrate results in a clear zone around the colony. Thus, the screening method is easily applicable on a large scale. Consequently, a high number of clones can be screened with a reasonable effort. It has been shown that soil-borne esterases can exhibit unusual and biotechnologically valuable properties. We obtained 37 unique enzymes with lipolytic activity by (partial) screening of several metagenomic libraries from German soil samples (Chapter 4). This is the highest number of soil-borne esterases published in one article. Thereby, we demonstrated how to generate a large reservoir of different lipolytic enzymes with reasonable effort.

6.3   References for chapter 6

Acosta-Martínez, V., S. Dowd, Y. Sun, and V. Allen. 2008. Tag-encoded pyrosequencing analysis of bacterial diversity in a single soil type as affected by management and land use. Soil Biol Biochem **40**:2762–2770.

Barns, S., E. Cain, L. Sommerville, and C. Kuske. 2007. Acidobacteria phylum sequences in uranium-contaminated subsurface sediments greatly expand the known diversity within the phylum. Appl Environ Microbiol **73**:3113-3116.

Bayer, S., A. Kunert, M. Ballschmiter, and T. Greiner-Stoeffele. 2010. Indication for a new lipolytic enzyme family: isolation and characterization of two esterases from a metagenomic library. J Mol Microbiol Biotechnol **18**:181-187.

Bunterngsook, B., P. Kanokratana, T. Thongaram, S. Tanapongpipat, T. Uengwetwanit, S. Rachdawong, T. Vichitsoonthonkul, and L. Eurwilaichitr. 2010. Identification and characterization of lipolytic enzymes from a peat-swamp forest soil metagenome. Biosci Biotechnol Biochem **74**:1848-1854.

Byun, J., J. Rhee, D. Kim, J. Oh, and H. Cho. 2006. Crystallization and preliminary X-ray crystallographic analysis of EstE1, a new and thermostable esterase cloned from a metagenomic library. Acta Crystallogr Sect F Struct Biol Cryst Commun **62**:145-147.

Chu, X., H. He, C. Guo, and B. Sun. 2008. Identification of two novel esterases from a marine metagenomic library derived from South China Sea. Appl Microbiol Biotechnol **80**:615-625.

Cole, J. R., Q. Wang, E. Cardenas, J. Fish, B. Chai, R. J. Farris, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, T. Marsh, G. M. Garrity, and J. M. Tiedje. 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res **37**:D141-145.

Couto, G., A. Glogauer, H. Faoro, L. Chubatsu, E. Souza, and F. Pedrosa. 2010. Isolation of a novel lipase from a metagenomic library derived from mangrove sediment from the south Brazilian coast. Genet Mol Res **9**:514-523.

Daniel, R. 2005. The metagenomics of soil. Nat Rev Microbiol **3**:470-478.

DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol **72**:5069-5072.

Ferrer, M., A. Beloqui, K. Timmis, and P. Golyshin. 2009. Metagenomics for mining new genetic resources of microbial communities. J Mol Microbiol Biotechnol **16**:109-123.

Ferrer, M., O. Golyshina, T. Chernikova, A. Khachane, V. Martins Dos Santos, M. Yakimov, K. Timmis, and P. Golyshin. 2005a. Microbial enzymes mined from the Urania deep-sea hypersaline anoxic basin. Chem Biol **12**:895-904.

Ferrer, M., O. Golyshina, T. Chernikova, A. Khachane, D. Reyes-Duarte, V. Santos, C. Strompl, K. Elborough, G. Jarvis, A. Neef, M. Yakimov, K. Timmis, and P. Golyshin. 2005b. Novel hydrolase diversity retrieved from a metagenome library of bovine rumen microflora. Environ Microbiol **7**:1996-2010.

Fierer, N., M. Bradford, and R. Jackson. 2007. Toward an ecological classification of soil bacteria. Ecology **88**:1354-1364.

Fierer, N. and R. Jackson. 2006. The diversity and biogeography of soil bacterial communities. Proc Natl Acad Sci U S A **103**:626-631.

Fierer, N., J. Schimel, and P. Holden. 2003. Variations in microbial community composition through two soil depth profiles. Soil Biol Biochem **35**:167–176.

Gabor, E., W. Alkema, and D. Janssen. 2004. Quantifying the accessibility of the metagenome by random expression cloning techniques. Environ Microbiol **6**:879-886.

Gupta, R., N. Gupta, and P. Rathi. 2004. Bacterial lipases: an overview of production, purification and biochemical properties. Appl Microbiol Biotechnol **64**:763-781.

Hansel, C., S. Fendorf, P. Jardine, and C. Francis. 2008. Changes in bacterial and archaeal community structure and functional diversity along a geochemically variable soil profile. Appl Environ Microbiol **74**:1620-1633.

Heath, C., X. Hu, S. Cary, and D. Cowan. 2009. Identification of a novel alkaliphilic esterase active at low temperatures by screening a metagenomic library from antarctic desert soil. Appl Environ Microbiol **75**:4657-4659.

Hong, K., H. Lim, E. Chung, E. Park, M. Lee, J. Kim, G. Choi, K. Cho, and S. Lee. 2007. Selection and characterization of forest soil metagenome genes encoding lipolytic enzymes. J Microbiol Biotechnol **17**:1655-1660.

Hu, Y., C. Fu, Y. Huang, Y. Yin, G. Cheng, F. Lei, N. Lu, J. Li, E. Ashforth, L. Zhang, and B. Zhu. 2010. Novel lipolytic genes from the microbial metagenomic library of the South China Sea marine sediment. FEMS Microbiol Ecol **72**:228-237.

Hårdeman, F. and S. Sjöling. 2007. Metagenomic approach for the isolation ofa novel low-temperature-active lipase fromuncultured bacteria of marine sediment. FEMS Microbiol Ecol **59**:524–534.

Jaeger, K. and T. Eggert. 2002. Lipases for biotechnology. Curr Opin Biotechnol **13**:390-397.

Jaeger, K. and P. Holliger. 2010. Chemical biotechnology-a marriage of convenience and necessity. Curr Opin Biotechnol **21**:711-712.

Janssen, P. 2006. Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes. Appl Environ Microbiol **72**:1719-1728.

Jeon, J., J. Kim, S. Kang, J. Lee, and S. Kim. 2009. Characterization and its potential application of two esterases derived from the arctic sediment metagenome. Mar Biotechnol (NY) **11**:307-316.

Jones, R., M. Robeson, C. Lauber, M. Hamady, R. Knight, and N. Fierer. 2009. A comprehensive survey of soil acidobacterial diversity using pyrosequencing and clone library analyses. ISME J **3**:442-453.

Kim, E., K. Oh, M. Lee, C. Kang, T. Oh, and J. Yoon. 2009. Novel cold-adapted alkaline lipase from an intertidal flat metagenome and proposal for a new family of bacterial lipases. Appl Environ Microbiol **75**:257-260.

Kim, J., M. Seo, E. Cho, S. Lee, S. Kim, C. Cheigh, and Y. Pyun. 2005. Screening and Characterization of an Esterase from a Metagenomic Library. J Microbiol Biotechnol **15**:1067–1072.

Kim, Y., E. Kwon, S. Kim, Y. Jeong, J. Kim, H. Yun, and H. Kim. 2010. Molecular cloning and characterization of a novel family VIII alkaline esterase from a compost metagenomic library. Biochem Biophys Res Commun **393**:45-49.

Klenk, H. P. and M. Göker. 2010. En route to a genome-based classification of Archaea and Bacteria? Syst Appl Microbiol **33**:175-182.

Kunin, V., A. Engelbrektson, H. Ochman, and P. Hugenholtz. 2010. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. Environ Microbiol **12**:118-123.

Lauber, C., M. Hamady, R. Knight, and N. Fierer. 2009. Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. Appl Environ Microbiol **75**:5111-5120.

Lauber, C., M. Strickland, M. Bradford, and N. Fierer. 2008. The influence of soil properties on the structure of bacterial and fungal communities across land-use types. Soil Biol Biochem **40**:2407–2415.

Lee, M., K. Hong, S. Malhotra, J. Park, E. Hwang, H. Choi, Y. Kim, W. Tao, and S. Lee. 2010. A new esterase EstD2 isolated from plant rhizosphere soil metagenome. Appl Microbiol Biotechnol **88**:1125-1134.

Lee, S., K. Won, H. Lim, J. Kim, G. Choi, and K. Cho. 2004. Screening for novel lipolytic enzymes from uncultured soil microorganisms. Appl Microbiol Biotechnol **65**:720-726.

Li, G., K. Wang, and Y. Liu. 2008. Molecular cloning and characterization of a novel pyrethroid-hydrolyzing esterase originating from the Metagenome. Microb Cell Fact **7**:38.

Lorenz, P. and J. Eck. 2005. Metagenomics and industrial applications. Nat Rev Microbiol **3**:510-516.

Lämmle, K., H. Zipper, M. Breuer, B. Hauer, C. Buta, H. Brunner, and S. Rupp. 2007. Identification of novel enzymes with different hydrolytic activities by metagenome expression cloning. J Biotechnol **127**:575-592.

López-García, P. and D. Moreira. 2008. Tracking microbial biodiversity through molecular and genomic ecology. Res Microbiol **159**:67-73.

Mocali, S. and A. Benedetti. 2010. Exploring research frontiers in microbiology: the challenge of metagenomics in soil microbiology. Res Microbiol **161**:497-505.

Pruesse, E., C. Quast, K. Knittel, B. M. Fuchs, W. Ludwig, J. Peplies, and F. O. Glöckner. 2007. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res **35**:7188-7196.

Ranjan, R., A. Grover, R. Kapardar, and R. Sharma. 2005. Isolation of novel lipolytic genes from uncultured bacteria of pond water. Biochem Biophys Res Commun **335**:57-65.

Rastogi, R., M. Wu, I. Dasgupta, and G. E. Fox. 2009. Visualization of ribosomal RNA operon copy number distribution. BMC Microbiol **9**:208.

Rhee, J., D. Ahn, Y. Kim, and J. Oh. 2005. New thermophilic and thermostable esterase with sequence similarity to the hormone-sensitive lipase family, cloned from a metagenomic library. Appl Environ Microbiol **71**:817-825.

Roesch, L., R. Fulthorpe, A. Riva, G. Casella, A. Hadwin, A. Kent, S. Daroub, F. Camargo, W. Farmerie, and E. Triplett. 2007. Pyrosequencing enumerates and contrasts soil microbial diversity. ISME J **1**:283-290.

Rousk, J., E. Bååth, P. Brookes, C. Lauber, C. Lozupone, J. Caporaso, R. Knight, and N. Fierer. 2010. Soil bacterial and fungal communities across a pH gradient in an arable soil. ISME J **4**:1340-1351.

Schloss, P. and J. Handelsman. 2005. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Appl Environ Microbiol **71**:1501-1506.

Simon, C. and R. Daniel. 2009. Achievements and new knowledge unraveled by metagenomic approaches. Appl Microbiol Biotechnol **85**:265-276.

Tirawongsaroj, P., R. Sriprang, P. Harnpicharnchai, T. Thongaram, V. Champreda, S. Tanapongpipat, K. Pootanakit, and L. Eurwilaichitr. 2008. Novel thermophilic and thermostable lipolytic enzymes from a Thailand hot spring metagenomic library. J Biotechnol **133**:42-49.

Uroz, S., M. Buée, C. Murat, P. Frey-Klett, and F. Martin. 2010. Pyrosequencing reveals a contrasted bacterial diversity between oak rhizosphere and surrounding soil. Environ Microbiol Rep **2**:281-288.

van der Heijden, M., R. Bardgett, and N. van Straalen. 2008. The unseen majority: soil microbes as drivers of plant diversity and productivity in terrestrial ecosystems. Ecol Lett **11**:296-310.

Wei, P., L. Bai, W. Song, and G. Hao. 2009. Characterization of two soil metagenome-derived lipases with high specificity for p-nitrophenyl palmitate. Arch Microbiol **191**:233-240.

Wu, C. and B. Sun. 2009. Identification of novel esterase from metagenomic library of Yangtze river. J Microbiol Biotechnol **19**:187-193.

Zhou, J., B. Xia, D. Treves, L. Wu, T. Marsh, R. O'Neill, A. Palumbo, and J. Tiedje. 2002. Spatial and resource factors influencing high microbial diversity in soil. Appl Environ Microbiol **68**:326-334.

Chapter 7: Summary

Soil is known to be the most complex microbial habitat on Earth with respect to species diversity and community size. Bacteria are the most abundant group of microorganisms in this habitat. Therefore, assessing the bacterial community composition in soil is an important task but also a considerable challenge. Furthermore, soil-born microorganisms harbor a largely untapped biotechnological potential. The present thesis focused on the metagenomic analysis of soil samples from the three German Biodiversity Exploratories, the Hainich region, the Schwäbische Alb, and the Schorfheide. Whole metagenomic DNA was isolated for the subsequent analyses.

Phylogenetic analysis of microbial soil community compositions was performed using 18 grassland samples from the Hainich (9 from topsoil and 9 from subsoil) and 18 samples from the Schwäbische Alb (9 from grassland topsoil and 9 from forest topsoil). The V2-V3-region of the 16S rRNA gene was amplified prior to pyrosequencing which yielded 752,838 high quality sequences from the Hainich and 599,284 from the Schwäbische Alb. The bacterial community structure was assessed by sequence analysis using the RDP classifier. The relative abundances of several taxonomic groups, such as *Actinobacteria*, *Alphaproteobacteria*, *Cyanobacteria*, *Firmicutes*, and *Gemmatimonadetes* changed significantly between topsoil and subsoil grassland samples and between topsoil samples from forest and grassland. The RDP pyrosequencing pipeline was employed for calculating diversity and richness estimates such as the Shannon index, Chao1 estimator, or rarefaction curves. A higher bacterial diversity was determined in topsoil than in subsoil on phylum and species level. Furthermore, a higher bacterial diversity was detected in grassland soil than in forest soil on the phylum level. Within the forest samples the tree species affected the soil bacterial diversity whereas management intensities did not exert any influence on community compositions in the investigated samples. Correlation analyses between abundances of individual phylogenetic groups and geochemical soil characteristics revealed a strong influence of pH on bacterial community compositions for soil samples that cover a broad pH range. At near neutral soil samples the organic carbon content was the major driver of microbial community structure.

In addition, the metagenomic DNA from all three Exploratories was used for the construction of 14 small-insert and seven large-insert libraries in plasmids and fosmids,

respectively. The small-insert libraries comprised approximately 40,000 to 511,000 clones with insert sizes of 3 to 9 kb and the large-insert libraries 4,600 to 300,000 clones with inserts of 19 to 30 kb. Thus, 11.5 Gb of soil metagenomic information were stored in plasmids and 17.8 Gb in fosmids. Subsequently, the constructed libraries were partially subjected to functional screening for lipolytic enzymes. Thirty-seven unique clones with esterase activity have been recovered by a function-driven screen that was based on the hydrolysis of the triglyceride tributyrin. The inserts have been sequenced and analyzed. From the 37 identified lipolytic genes deduced enzymes represent new members of known or putatively new lipase/esterase families. Most of the enzymes were assigned to family IV but also members of the families I (true lipases), V, VI, VIII, and two presumably new families have been detected. Thus, soil is a rich source for the discovery of novel bioactive molecules with biotechnological potential that originate from uncharacterized microorganisms. Furthermore, the phylogenetic analysis of the habitat soil provides insight into soil microbial processes and ecosystem functions.

Chapter 8: Appendix

8.1 Appendix A: Supplemental Information for chapter 2

Table S1. Soil type, land use type, and coordinates of the sampling sites.

Plot	Soil type	Land use type	Latitude	Longitude
1	Vertic Cambisol (eutric)	Fertilized meadow	50°58'17.99"	10°24'19.24"
2	Vertic Stagnosol (eutric)	Fertilized meadow	51°0'2.75"	10°25'47.96"
3	Vertic Stagnosol (eutric)	Fertilized meadow	50°59'53.19"	10°25'58.54"
4	Vertic Stagnosol (eutric)	Fertilized mown pasture, cattle	51°6'48.17"	10°26'10.17"
5	Vertic Stagnosol (eutric)	Fertilized mown pasture, cattle	51°12'57.28"	10°19'21.1"
6	Luvic Stagnosol (siltic)	Fertilized mown pasture, cattle	51°12'53.83"	10°23'28.31"
7	Vertic Stagnosol (eutric)	Unfertilized pasture, cattle	51°16'24.96"	10°24'37.40"
8	Vertic Stagnosol (eutric)	Unfertilized pasture, cattle	51°16'16.59"	10°25'4.52"
9	Vertic Stagnosol (eutric)	Unfertilized pasture, cattle	51°13'26.9"	10°22'50.75"

Table S2. Number of 16S rRNA sequences derived from the A and B horizons of nine different grassland sampling sites. For a description of the sampling sites, see Table 1 and supplemental Table S1.

Horizon	Plot	No. of obtained sequences	No. of classified sequences
A	1	47,063	38,693
A	2	39,270	32,702
A	3	33,804	27,960
A	4	39,328	34,651
A	5	53,422	43,443
A	6	41,239	35,062
A	7	61,366	49,383
A	8	47,068	41,810
A	9	51,870	42,243
B	1	47,264	33,558
B	2	33,209	23,875
B	3	36,598	28,428
B	4	38,856	30,565
B	5	40,189	25,511
B	6	41,027	31,572
B	7	25,851	19,662
B	8	42,994	37,439
B	9	32,420	23,987
Sum		752,838	600,544

Table S3. Relative abundance of the phylogenetic groups. Values are given as percentage. Phylogenetic groups accounting for ≤0.25% of the classified sequences are summarized in the artificial group 'others', containing the phyla *Deinococcus-Thermus, Fibrobacteres, Fusobacteria, Gemmatimonadetes, Nitrospira,* OP11, *Planctomycetes, Spirochaetes, Tenericutes,* and *Epsilonproteobacteria.*

Horizon	Plot	Acido-bacteria	Actino-bacteria	Bactero-idetes	Chloro-flexi	Cyano-bacteria	Firmicutes	TM7	Verruco-microbia	WS3	Alpha-proteo-bacteria	Beta-proteo-bacteria	Gamma-proteo-bacteria	Delta-proteo-bacteria	unclassified Proteo-bacteria	others
A	1	23.27	14.80	2.26	0.30	0.43	1.04	0.34	1.03	0.43	9.40	17.89	12.51	6.33	9.44	0.50
A	2	19.88	5.21	1.46	1.59	1.57	4.84	0.20	0.11	0.05	9.00	22.95	14.13	1.00	17.79	0.05
A	3	15.90	8.26	1.33	2.86	2.78	6.75	0.44	0.14	0.09	8.68	21.91	10.77	5.80	14.13	0.17
A	4	12.68	17.14	2.09	0.41	1.37	6.91	0.45	0.44	0.17	10.70	23.01	14.42	6.20	2.92	1.02
A	5	19.73	17.17	2.11	0.96	3.31	1.54	0.47	0.66	0.43	11.54	16.57	10.23	5.55	9.46	0.26
A	6	21.51	12.22	2.36	0.53	2.04	3.39	0.50	0.44	0.22	10.10	17.37	13.48	10.60	3.82	1.43
A	7	22.97	16.73	2.65	0.54	0.60	3.00	0.46	0.83	0.30	11.72	13.89	12.24	4.52	9.33	0.23
A	8	14.40	11.59	2.05	0.17	0.42	7.64	0.30	0.54	0.12	13.31	16.32	25.80	4.09	2.59	0.66
A	9	19.79	15.73	1.75	0.61	0.66	1.28	0.46	0.76	0.38	14.22	13.91	13.04	5.19	12.08	0.16
B	1	27.85	13.19	0.43	11.64	0.32	2.86	0.71	0.26	0.87	9.08	12.13	5.42	2.26	12.70	0.30
B	2	49.86	3.95	0.26	7.30	0.05	0.77	0.34	0.09	0.65	5.13	13.91	6.64	6.85	3.78	0.41
B	3	43.26	4.53	0.32	9.13	0.20	1.26	0.34	0.70	0.69	5.24	13.83	9.14	8.61	3.01	0.17
B	4	35.61	7.73	0.92	11.81	0.41	1.26	0.61	0.52	1.07	7.35	11.13	9.14	8.63	3.43	3.67
B	5	34.19	12.40	0.09	11.65	0.04	0.34	0.58	0.04	2.44	8.89	10.13	4.81	1.26	13.05	0.11
B	6	43.20	5.44	0.48	8.17	0.07	5.84	0.36	0.30	0.83	6.85	13.37	5.89	5.60	3.16	0.43
B	7	31.23	14.70	1.13	3.57	0.26	2.35	0.69	0.34	0.08	10.33	1.319	7.96	1.84	12.02	0.29
B	8	38.28	9.66	1.55	3.65	0.19	4.01	4.39	0.26	0.26	7.44	17.76	5.61	4.38	2.43	0.12
B	9	35.81	9.22	0.10	3.34	0.93	0.95	0.57	0.06	0.51	9.71	10.25	8.61	0.66	19.14	0.14

Table S4. Percentages of the different acidobacterial subgroups in the A and B horizons of the German grassland soils derived from each plot. Percentages are given relative to the total number of classified sequences in that soil. Sequences of acidobacterial subgroups 14, 23, 24, and 26 were not detected.

Plot[a]	Total No.[b]	Gp1	Gp2	Gp3	Gp4	Gp5	Gp6	Gp7	Gp8	Gp9	Gp10	Gp11	Gp12	Gp13	Gp15	Gp16	Gp17	Gp18	Gp19	Gp20	Gp21	Gp22	Gp25
1A	38,693	0.05	0.00	1.22	2.55	0.28	13.30	1.67	0.00	0.00	0.00	0.29	0.00	0.01	0.00	2.85	0.46	0.02	0.00	0.00	0.00	0.42	0.02
2A	32,702	0.02	0.00	0.59	7.70	0.10	6.64	1.86	0.00	0.00	0.00	0.81	0.00	0.00	0.00	1.10	0.13	0.01	0.00	0.50	0.00	0.33	0.05
3A	27,960	0.01	0.00	1.14	4.46	0.09	7.01	0.64	0.00	0.00	0.02	0.34	0.00	0.00	0.00	1.53	0.25	0.03	0.00	0.08	0.00	0.25	0.03
4A	34,651	0.01	0.00	0.37	2.12	0.11	6.11	0.44	0.01	0.00	0.02	0.23	0.00	0.00	0.00	2.59	0.34	0.01	0.00	0.02	0.00	0.16	0.03
5A	43,443	0.09	0.00	0.84	3.51	0.26	8.63	1.62	0.00	0.00	0.03	0.24	0.00	0.00	0.00	3.34	0.64	0.00	0.00	0.03	0.00	0.34	0.02
6A	35,062	0.08	0.00	1.17	7.69	0.13	7.70	1.22	0.00	0.00	0.10	0.31	0.00	0.01	0.01	2.14	0.26	0.02	0.00	0.07	0.00	0.37	0.12
7A	49,383	0.02	0.00	1.29	3.73	0.38	11.25	1.56	0.00	0.00	0.03	0.27	0.00	0.00	0.00	3.21	0.71	0.02	0.00	0.01	0.00	0.31	0.07
8A	41,810	0.01	0.00	0.29	2.36	0.11	7.74	0.81	0.00	0.00	0.03	0.18	0.00	0.00	0.00	2.39	0.12	0.01	0.00	0.02	0.00	0.16	0.01
9A	42,243	0.03	0.00	1.10	2.91	0.41	9.56	1.45	0.00	0.00	0.05	0.23	0.00	0.00	0.00	2.81	0.71	0.02	0.00	0.00	0.00	0.36	0.03
1B	33,558	0.01	0.00	1.59	6.65	0.22	10.79	2.36	0.00	0.02	0.01	0.88	0.00	0.00	0.01	2.43	1.14	0.59	0.00	0.09	0.00	0.72	0.22
2B	23,875	1.91	0.00	1.10	16.16	0.24	14.00	10.06	0.00	0.05	0.04	1.53	0.00	0.35	0.09	1.78	0.91	1.08	0.00	0.04	0.00	0.06	0.04
3B	28,428	0.10	0.00	1.38	19.45	0.14	9.98	7.29	0.00	0.02	0.06	2.05	0.00	0.04	0.04	0.82	0.70	0.66	0.00	0.04	0.00	0.11	0.21
4B	30,565	0.12	0.00	1.38	15.95	0.95	7.66	3.09	0.02	0.03	0.04	3.15	0.00	0.02	0.04	0.86	0.77	0.60	0.01	0.27	0.00	0.38	0.23
5B	25,511	0.05	0.00	1.27	8.48	0.20	7.58	7.98	0.00	0.06	0.02	1.52	0.00	0.02	0.08	3.00	2.59	0.54	0.00	0.05	0.00	0.58	0.11
6B	31,572	3.22	0.01	1.69	19.39	0.77	7.82	2.88	0.00	0.01	0.09	1.73	0.08	0.02	0.10	1.38	1.06	0.91	0.02	0.93	0.00	0.38	0.61
7B	19,662	0.01	0.00	1.29	7.36	0.35	14.67	2.31	0.00	0.00	0.08	0.43	0.00	0.00	0.03	2.85	0.74	0.29	0.00	0.19	0.00	0.38	0.14
8B	37,439	0.04	0.00	0.87	20.45	0.61	9.39	2.53	0.00	0.00	0.08	1.67	0.00	0.00	0.03	1.23	0.24	0.20	0.00	0.15	0.01	0.32	0.27
9B	23,987	0.04	0.00	1.73	9.70	0.15	7.65	11.96	0.00	0.03	0.01	0.54	0.00	0.00	0.03	2.33	0.61	0.37	0.00	0.14	0.00	0.34	0.10

Relative abundance of acidobacterial subgroups[c] (%)

[a], A and B represent samples from A and B horizon, respectively
[b], No. of all classified sequences in that soil
[c], Only subgroups that were detected are depicted

Table S5. Relative abundance of the different acidobacterial subgroups in the A and B horizons of the German grassland soils within the *Acidobacteria*. Percentages are given relative to all acidobacterial sequences per plot. Sequences of acidobacterial subgroups 14, 23, 24, and 26 were not detected.

Plot[a]	Total No.[b]	Gp1	Gp2	Gp3	Gp4	Gp5	Gp6	Gp7	Gp8	Gp9	Gp10	Gp11	Gp12	Gp13	Gp15	Gp16	Gp17	Gp18	Gp19	Gp20	Gp21	Gp22	Gp25
1A	8954	0.23	0.00	5.27	11.01	1.19	57.47	7.23	0.00	0.00	0.01	1.26	0.00	0.02	0.01	12.32	1.98	0.10	0.00	0.00	0.00	1.80	0.09
2A	6488	0.08	0.00	2.97	38.81	0.51	33.49	9.36	0.00	0.00	0.02	4.08	0.00	0.00	0.00	5.53	0.63	0.06	0.00	2.54	0.00	1.68	0.23
3A	4438	0.07	0.00	7.19	28.12	0.56	44.19	4.03	0.00	0.00	0.14	2.12	0.00	0.00	0.00	9.62	1.55	0.18	0.00	0.50	0.00	1.55	0.18
4A	4355	0.05	0.00	2.96	16.83	0.90	48.59	3.47	0.05	0.00	0.16	1.86	0.00	0.00	0.02	20.57	2.71	0.07	0.00	0.18	0.00	1.31	0.28
5A	8520	0.46	0.00	4.31	17.89	1.34	44.01	8.26	0.00	0.00	0.15	1.24	0.00	0.00	0.00	17.05	3.29	0.01	0.00	0.13	0.00	1.75	0.11
6A	7501	0.36	0.00	5.45	35.96	0.59	36.01	5.71	0.00	0.01	0.45	1.43	0.00	0.04	0.05	10.00	1.20	0.09	0.00	0.35	0.00	1.75	0.56
7A	11287	0.08	0.00	5.66	16.30	1.67	49.21	6.80	0.00	0.01	0.12	1.18	0.00	0.00	0.02	14.03	3.13	0.09	0.00	0.04	0.00	1.36	0.31
8A	5952	0.10	0.00	2.03	16.55	0.74	54.35	5.70	0.00	0.00	0.20	1.26	0.00	0.00	0.03	16.77	0.84	0.10	0.00	0.13	0.00	1.09	0.10
9A	8316	0.17	0.00	5.60	14.79	2.10	48.58	7.35	0.00	0.00	0.24	1.15	0.00	0.01	0.01	14.30	3.60	0.11	0.00	0.02	0.00	1.80	0.16
1B	9303	0.05	0.00	5.72	23.97	0.78	38.92	8.50	0.01	0.06	0.05	3.18	0.00	0.00	0.04	8.75	4.10	2.13	0.01	0.33	0.00	2.58	0.80
2B	11807	3.85	0.00	2.23	32.68	0.49	28.31	20.34	0.00	0.09	0.08	3.10	0.01	0.70	0.19	3.59	1.85	2.19	0.00	0.08	0.01	0.13	0.08
3B	12245	0.24	0.00	3.19	45.14	0.32	23.18	16.91	0.00	0.06	0.13	4.77	0.01	0.08	0.10	1.89	1.62	1.53	0.00	0.09	0.01	0.25	0.48
4B	10873	0.34	0.00	3.88	44.84	2.68	21.54	8.68	0.06	0.07	0.12	8.86	0.01	0.06	0.12	2.42	2.16	1.69	0.02	0.74	0.00	1.08	0.64
5B	8713	0.16	0.00	3.72	24.82	0.59	22.21	23.38	0.00	0.17	0.05	4.44	0.00	0.05	0.24	8.79	7.60	1.58	0.00	0.16	0.00	1.71	0.33
6B	13607	7.47	0.01	3.92	44.98	1.78	18.14	6.69	0.00	0.01	0.21	4.01	0.18	0.05	0.24	3.21	2.47	2.11	0.05	2.15	0.00	0.87	1.42
7B	6121	0.03	0.00	4.13	23.66	1.13	47.12	7.43	0.00	0.00	0.25	1.39	0.00	0.00	0.10	9.17	2.39	0.95	0.00	0.62	0.00	1.21	0.44
8B	14264	0.10	0.00	2.29	53.68	1.61	24.64	6.65	0.00	0.01	0.21	4.38	0.00	0.01	0.08	3.22	0.64	0.52	0.00	0.40	0.01	0.84	0.70
9B	8567	0.11	0.00	4.83	27.16	0.42	21.43	33.48	0.00	0.07	0.04	1.51	0.00	0.00	0.08	6.53	1.70	1.03	0.00	0.40	0.00	0.96	0.27

Relative abundance of acidobacterial subgroups[c] relating to the total number of classified acidobacterial sequences in that soil (%)

[a], A and B represent samples from A and B horizon, respectively
[b], No. of classified acidobacterial sequences in that soil
[c], Only subgroups that were detected are depicted

Table S6. *P* values estimating the difference of the abundance of selected phyla between A and B horizon with respect to different land use types.

Phylum	Land use type		
	Fertilized meadow (plot 1 to 3)	Fertilized mown pasture, cattle (plot 4 to 6)	Unfertilized pasture, cattle (plot 7 to 9)
Acidobacteria	*P* <0.00001	*P* <0.00001	*P* <0.00001
Actinobacteria	*P* <0.00001	*P* <0.00001	*P* <0.00001
Bacteroidetes	*P* <0.00001	*P* <0.00001	*P* <0.00001
Chloroflexi	*P* <0.00001	*P* <0.00001	*P* <0.00001
Cyanobacteria	*P* <0.00001	*P* <0.00001	*P* 0.00003
Fibrobacteres	*P* <0.00001	*P* <0.00001	*P* < 0.00001
Firmicutes	*P* <0.00001	*P* <0.00001	*P* <0.00001
Nitrospira	*P* <0.00001	*P* <0.00001	*P* 0.47635
TM7	*P* <0.00001	*P* 0.21149	*P* <0.00001
Verrucomicrobia	*P* <0.00001	*P* <0.00001	*P* <0.00001
WS3	*P* <0.00001	*P* <0.00001	*P* 0.29925
Alphaproteobacteria	*P* <0.00001	*P* <0.00001	*P* <0.00001
Betaproteobacteria	*P* <0.00001	*P* <0.00001	*P* 0.15213
Gammaproteobacteria	*P* <0.00001	*P* <0.00001	*P* <0.00001
Deltaproteobacteria	*P* <0.00001	*P* <0.00001	*P* <0.00001
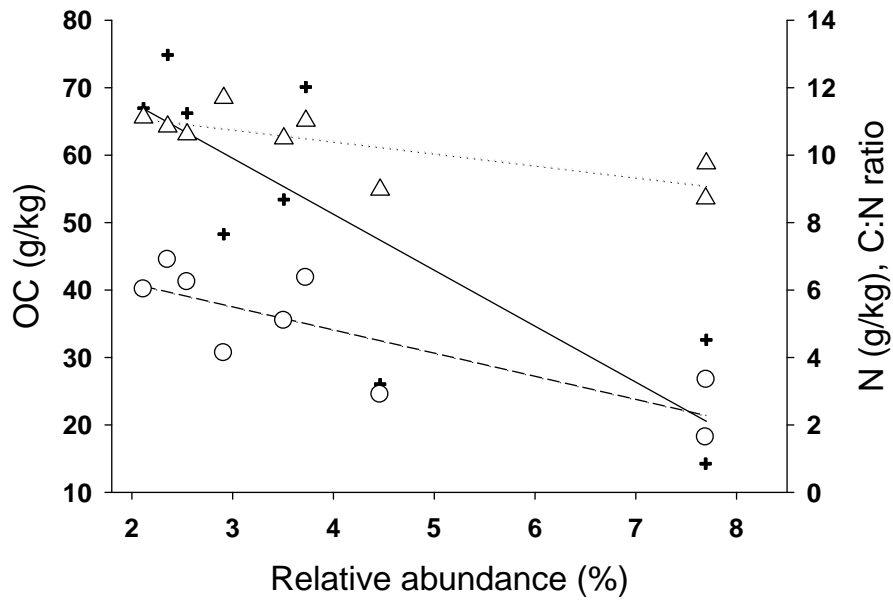
Fig. S1. Correlation of the relative abundance of the acidobacterial subgroup 4 and the edaphic soil properties OC content, N content, and C:N ratio in the A horizon.

Crosses and open circles indicate the OC and N content, respectively. The C:N ratio is marked by triangles.
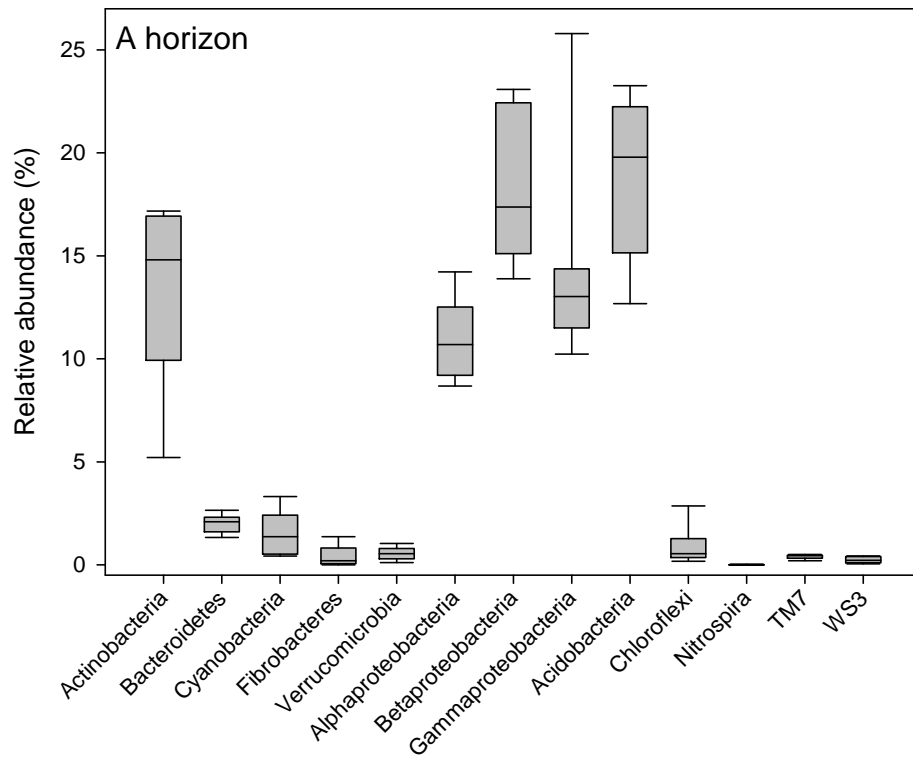
Fig. S2. Box-and-whiskers plot of relative distribution between A and B horizon for selected phyla.
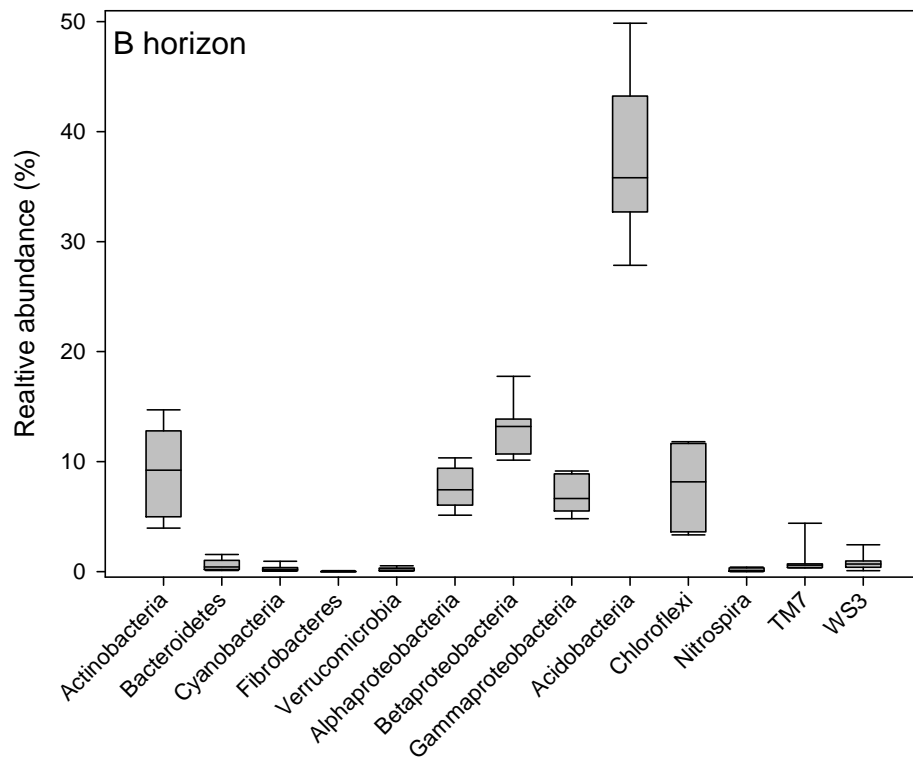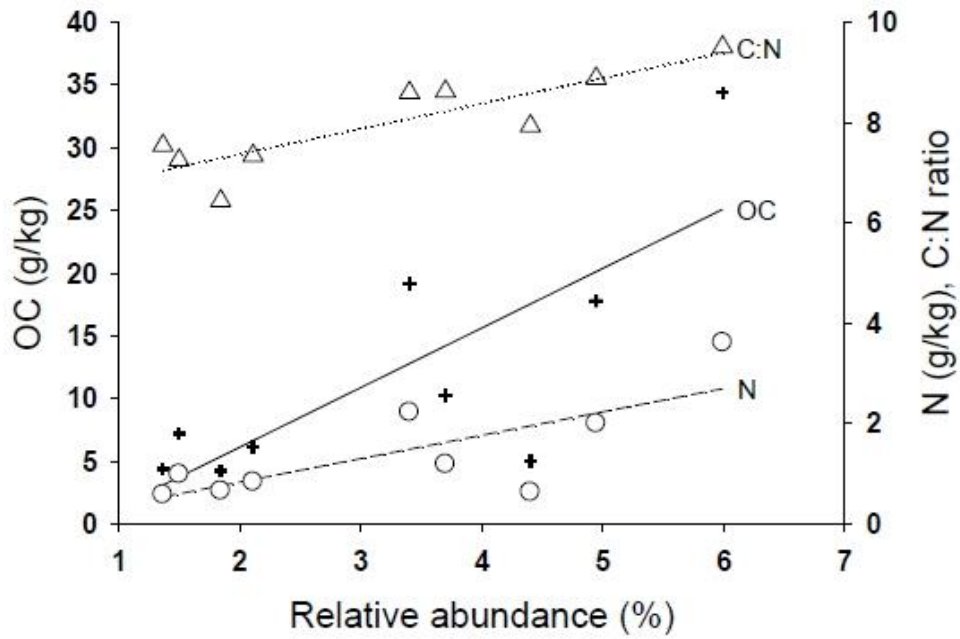
Fig. S3. Correlation of the relative abundance of *Rhizobiales* and the edaphic soil properties OC content, N content, and C:N ratio in the B horizon.

Crosses and open circles indicate the OC and N content, respectively. The C:N ratio is marked by triangles.

8.2 Appendix B: Supplemental Information for chapter 3
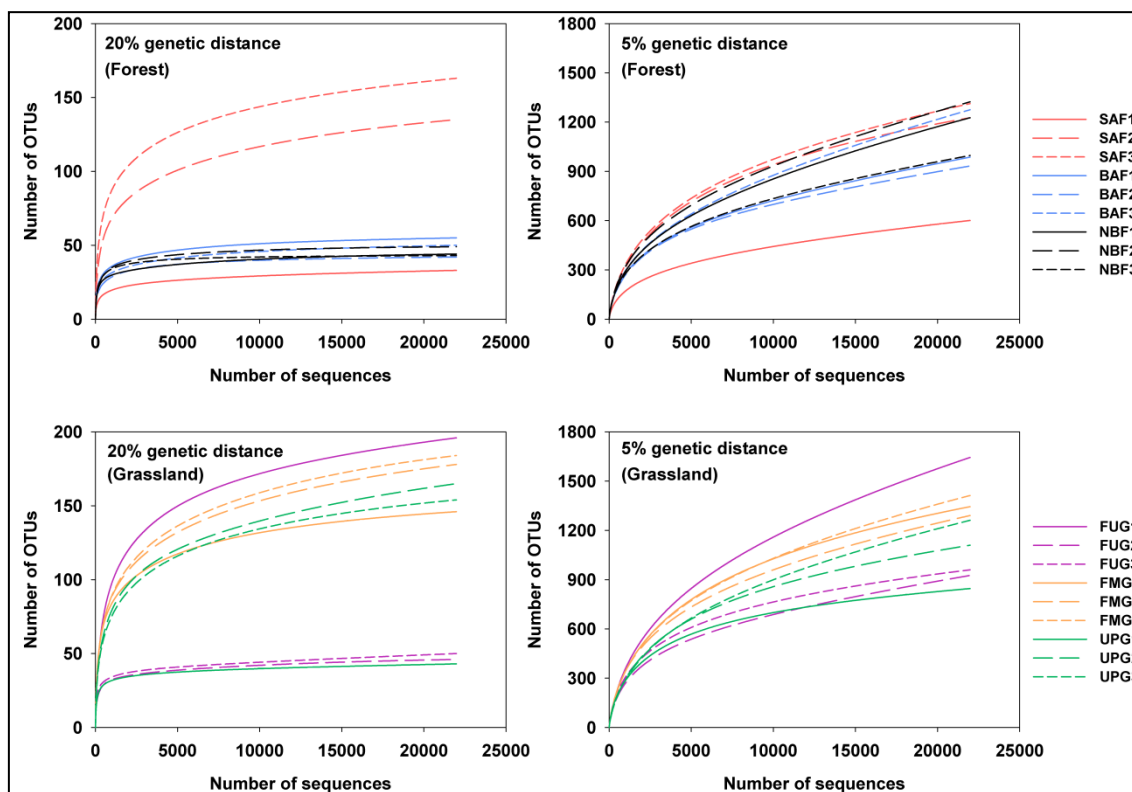
Fig. S1: Rarefaction curves indicating the observed number of operational taxonomic units (OTUs) at a genetic distance of 5% and 20% in different forest and grassland soils.

The spruce age class forest (SAF1-3), beech age class forest (BAF1-3), and unmanaged beech forest (BF1-3) sampling sites are marked by the red, blue, and black color, respectively. The fertilized intensely managed grassland (FUG1-3), fertilized meadow grazed by horse and cattle (FMG1-3), and unfertilized pasture grazed by sheep (UPG1-3) sampling sites are shown in purple, orange, and green, respectively.

Table S1. Localization of the sampling sites and number of 16S rRNA gene sequences derived from the analyzed grassland and forest soil samples.

Management type	Sample	Localization of the plots (deg min sec)		No. of sequences ≥ 200 bp	No. of sequences assigned to domain Bacteria	No. of sequences classified below domain level
		Latitude (N)	Longitude (E)			
Spruce age class forest	SAF1	48 28 41.0628	9 20 3.876	37,861	37,861	33,562
Spruce age class forest	SAF2	48 22 47.9634	9 21 5.2236	25,987	25,974	22,399
Spruce age class forest	SAF3	48 24 44.1468	9 21 20.1276	34,903	34,812	27,681
Beech age class forest	BAF1	48 23 56.7558	9 14 41.3772	31,373	31,372	23,723
Beech age class forest	BAF2	48 25 10.6242	9 24 52.8552	34,095	34,092	27,518
Beech age class forest	BAF3	48 23 38.583	9 26 45.3732	38,580	38,580	29,621
Unmanaged beech forest	BF1	48 23 46.4562	9 15 40.881	33,298	33,298	26,859
Unmanaged beech forest	BF2	48 22 57.324	9 22 56.5824	35,373	35,372	29,228
Unmanaged beech forest	BF3	48 22 9.6456	9 24 54.7842	33,974	33,974	26,839
Fertilized intensely managed grassland	FUG1	48 23 52.8174	9 20 31.1526	29,890	29,865	23,444
Fertilized intensely managed grassland	FUG2	48 22 36.6852	9 28 22.0224	31,483	31,481	23,806
Fertilized intensely managed grassland	FUG3	48 24 31.9716	9 31 56.5644	23,519	23,515	18,173
Fertilized meadow grazed by horse and cattle	FMG1	48 22 51.1962	9 25 8.0004	30,498	30,471	22,617
Fertilized meadow grazed by horse and cattle	FMG2	48 23 45.1536	9 26 21.1236	37,798	37,745	28,060
Fertilized meadow grazed by horse and cattle	FMG3	48 24 4.5432	9 26 30.0402	39,273	39,259	28,413
Unfertilized pasture grazed by sheep	UPG1	48 23 29.1156	9 22 36.6486	26,510	26,500	20,761
Unfertilized pasture grazed by sheep	UPG2	48 25 21.504	9 29 31.6494	36,854	36,79	30,888
Unfertilized pasture grazed by sheep	UPG3	48 23 40.8156	9 30 10.0506	38,015	38,001	31,276

Table S2. Mean values of soil properties (± Standard deviation) for each management type and ANOVA *P* values.

Soil property	Management type						ANOVA
	SAF	BAF	BF	FUG	FMG	UPG	*P*
pH	4.7±0.9 b	5.9±0.9 ab	5.9±0.8 ab	6.7±0.3 a	6.1±0.7 ab	6.9±0.4 a	**0.011**
Organic C (g kg^{-1})	68.2±5.7 a	58.4±19.7 a	80.9±22.7 a	67.7±12.3 a	68.3±16.8 a	63.6±20.7 a	0.731
Total N (g kg^{-1})	4.5±0.6 a	4.5±1.4 a	5.6±1.1 a	6.7±1.3 a	6.6±1.3 a	5.6±1.9 a	0.247
Sand (g kg^{-1})	43±17.0 a	75±30.3 a	66±38.6 a	67±62.4 a	56±24 a	115±145.5 a	0.967
Silt (g kg^{-1})	520±128.5 a	565±27.8 a	387±100.5 a	546±98.5 a	554±121.8 a	533±150.0 a	0.436
Clay (g kg^{-1})	437±114.0 a	361±39.5 a	546±112.8 a	387±158.0 a	390±122.5 a	353±216.6 a	0.559

Differences of soil properties between management types were analyzed by employing one-way analysis of variance and Tukey pair-wise comparisons. Significant ANOVA *P* values are shown in bold (*P* < 0.05). Figures followed by different letters indicate differences among management types (*P* < 0.05). Abbreviations: SAF, spruce age class forest; BAF, beech age class forest; BF, unmanaged beech forest; FUG, fertilized intensely managed grassland; FMG, fertilized meadow grazed by horse and cattle; UPG: unfertilized pasture grazed by sheep. Complete soil and site information for all 18 sample sites is provided in Table 1.

Table S3. Bacterial diversity as assessed by Shannon index (H') and species richness estimation in all forest and grassland soils. The results from the rarefaction analyses are also depicted in Figure 1 and Figure S1.

Management type	Sample	Genetic distance (%)	H´	Rarefaction	Chao1	ACE
				No. of operational taxonomic units		
Spruce age class forest	SAF1	3	4.74	810	1625	1608
		5	4.21	601	1119	1108
		20	1.62	33	37	36
Spruce age class forest	SAF2	3	5.75	1509	2214	1924
		5	5.40	1227	1786	1781
		20	2.98	135	149	154
Spruce age class forest	SAF3	3	5.81	1584	2745	2735
		5	5.55	1313	2137	2144
		20	3.66	163	193	183
Beech age class forest	BAF1	3	5.55	1192	3112	3076
		5	5.21	987	2328	2305
		20	2.50	55	60	58
Beech age class forest	BAF2	3	5.46	1134	3103	3056
		5	5.13	933	2399	2366
		20	2.06	42	42	43
Beech age class forest	BAF3	3	5.87	1669	4703	4669
		5	5.37	1276	3127	3107
		20	2.26	50	53	53
Unmanaged beech forest	BF1	3	5.68	1594	4050	4024
		5	5.23	1227	2751	2735
		20	2.00	44	45	46
Unmanaged beech forest	BF2	3	5.99	1734	4056	4033
		5	5.52	1324	3072	3054
		20	2.22	49	49	50
Unmanaged beech forest	BF3	3	5.66	1254	3366	3326
		5	5.24	997	2263	2242
		20	2.33	43	43	43

Table S3 (continued)

Management type	Sample	Genetic distance (%)	H´	Rarefaction	Chao1	ACE
				No. of operational taxonomic units		
Fertilized intensely managed grassland	FUG1	3	5.86	1960	4203	4188
		5	5.55	1645	3236	3224
		20	3.47	196	234	221
Fertilized intensely managed grassland	FUG2	3	5.46	1134	2807	2773
		5	5.10	926	2163	2139
		20	2.35	46	48	47
Fertilized intensely managed grassland	FUG3	3	5.92	1301	1652	1524
		5	5.52	960	1359	1353
		20	2.60	50	100	73
Fertilized mown pasture, horse and cattle	FMG1	3	5.79	1580	2199	2017
		5	5.54	1345	1912	1907
		20	3.67	146	156	157
Fertilized mown pasture, horse and cattle	FMG2	3	5.71	1498	2938	2923
		5	5.53	1291	2320	2309
		20	3.60	178	207	205
Fertilized mown pasture, horse and cattle	FMG3	3	5.60	1648	3022	3011
		5	5.37	1413	2561	2551
		20	3.50	184	202	209
Unfertilized pasture, sheep	UPG1	3	5.64	1078	1320	1207
		5	5.35	846	1119	998
		20	2.36	43	56	48
Unfertilized pasture, sheep	UPG2	3	4.96	1302	1945	1644
		5	4.79	1110	1681	1674
		20	3.23	165	201	198
Unfertilized pasture, sheep	UPG3	3	4.99	1482	3413	3394
		5	4.81	1262	2507	2495
		20	3.10	154	171	171

Table S4. Relative abundances of bacterial phyla and proteobacterial classes in the analyzed forest soils. Values represent percentages of all sequences assigned to the domain Bacteria for all forest soils or individual forest soils. Groups labeled with asterisks could not be assigned to a specific phylum or a proteobacterial class.

Phylogenetic group	Relative abundance (%)									
	Average	Spruce age class forests			Beech age class forests			Unmanaged beech forests		
		SAF1	SAF2	SAF3	BAF4	BAF5	BAF6	BF7	BF8	BF9
Proteobacteria	45.539	55.355	50.204	38.366	38.152	45.609	38.955	49.018	51.040	43.477
Alphaproteobacteria	25.072	42.759	33.603	18.272	14.338	22.480	19.453	28.957	26.515	19.388
Acidobacteria	20.391	22.852	23.081	23.661	20.404	19.981	20.910	15.325	15.868	21.725
*Bacteria**	18.964	11.355	13.764	20.484	24.382	19.283	23.222	19.337	17.370	21.001
Actinobacteria	12.655	9.165	11.408	15.253	13.550	12.871	13.735	14.052	14.104	9.690
*Proteobacteria**	7.155	3.056	11.246	12.171	6.732	6.650	6.353	5.925	7.633	5.972
Betaproteobacteria	5.991	3.735	2.476	5.191	8.992	6.482	5.534	5.220	8.187	7.738
Deltaproteobacteria	4.356	1.138	0.970	1.454	4.568	7.257	5.568	6.271	5.335	6.125
Gammaproteobacteria	2.966	4.667	1.910	1.278	3.522	2.740	2.048	2.646	3.370	4.253
WS3	0.728	0.005	0.046	0.273	1.540	0.930	1.861	0.478	0.209	1.071
Firmicutes	0.562	0.409	0.354	0.942	0.583	0.170	0.591	1.078	0.271	0.636
TM7	0.450	0.642	0.936	0.566	0.236	0.487	0.254	0.288	0.269	0.474
Chloroflexi	0.285	0.042	0.073	0.215	0.462	0.126	0.257	0.180	0.424	0.774
Bacteroidetes	0.249	0.111	0.058	0.138	0.478	0.238	0.104	0.105	0.263	0.756
Verrucomicrobia	0.067	0.008	0.008	0.066	0.153	0.094	0.034	0.012	0.090	0.141
Fibrobacteres	0.049	0.029	0.000	0.000	0.013	0.100	0.047	0.096	0.062	0.085
Cyanobacteria	0.035	0.021	0.042	0.032	0.026	0.032	0.021	0.012	0.014	0.124
Spirochaetes	0.019	0.000	0.008	0.000	0.016	0.076	0.003	0.015	0.011	0.041
Gemmatimonadetes	0.003	0.005	0.015	0.003	0.000	0.003	0.000	0.000	0.003	0.000
Planctomycetes	0.002	0.000	0.004	0.000	0.000	0.000	0.008	0.003	0.000	0.000
OP11	0.001	0.000	0.000	0.000	0.006	0.000	0.000	0.000	0.000	0.003

Table S5. Relative abundances of bacterial phyla and proteobacterial classes in the analyzed grassland soils. Values represent percentages of all sequences assigned to the domain Bacteria for all grassland soils or individual grassland soils. Groups labeled with asterisks could not be assigned to a specific phylum or a proteobacterial class.

Phylogenetic group	Relative abundance (%)									
	Average	Fertilized intensely managed grassland			Fertilized meadows grazed by horse and cattle			Unfertilized pastures grazed by sheep		
		FUG1	FUG2	FUG3	FMG4	FMG5	FMG6	UPG7	UPG8	UPG9
Proteobacteria	34.863	26.044	26.028	31.218	31.906	34.699	22.318	31.989	50.886	53.356
*Bacteria**	22.542	21.500	24.380	22.717	25.775	25.659	27.627	21.657	16.042	17.697
Actinobacteria	19.625	27.420	29.796	20.174	10.722	15.949	24.374	21.460	16.167	12.684
Acidobacteria	18.710	22.873	15.946	18.150	26.835	19.571	21.572	17.306	13.226	14.034
*Proteobacteria**	12.437	7.604	3.888	3.776	13.190	14.428	9.361	3.921	24.224	23.799
Alphaproteobacteria	11.434	8.321	8.758	8.930	10.768	10.862	6.765	8.830	17.374	19.636
Betaproteobacteria	5.863	7.504	6.953	7.272	5.274	5.595	3.739	6.408	5.624	5.584
Gammaproteobacteria	2.743	2.032	2.271	3.934	1.303	2.967	1.409	4.751	3.284	3.339
Deltaproteobacteria	2.387	0.583	4.158	7.306	1.372	0.845	1.044	8.079	0.381	0.997
Firmicutes	1.845	0.837	1.992	2.760	2.822	1.674	1.752	2.366	2.324	0.603
Bacteroidetes	0.723	0.254	0.480	1.999	0.069	0.906	0.423	2.317	0.495	0.263
WS3	0.459	0.161	0.305	1.050	0.791	0.376	0.540	0.992	0.068	0.192
TM7	0.371	0.281	0.289	0.332	0.325	0.215	0.515	0.577	0.277	0.526
Chloroflexi	0.338	0.238	0.235	0.383	0.450	0.551	0.410	0.351	0.098	0.321
Cyanobacteria	0.237	0.208	0.219	0.391	0.236	0.241	0.196	0.321	0.193	0.205
Verrucomicrobia	0.226	0.171	0.302	0.515	0.046	0.146	0.245	0.453	0.204	0.097
Fibrobacteres	0.034	0.000	0.025	0.213	0.000	0.000	0.000	0.162	0.000	0.000
Gemmatimonadetes	0.016	0.010	0.003	0.077	0.016	0.011	0.013	0.015	0.005	0.013
Spirochaetes	0.010	0.003	0.000	0.021	0.007	0.003	0.013	0.034	0.005	0.008
Deinococcus-Thermus	0.001	0.000	0.000	0.000	0.000	0.000	0.008	0.003	0.000	0.000
Fusobacteria	0.0003	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.000

Table S6. Relative abundances of acidobacterial subgroups in the analyzed forest soils. Values represent percentages of all sequences assigned to the domain Bacteria for all forest soils or individual forest soils. Groups labeled with asterisks could be assigned to the phylum level only.

Phylogenetic group	Relative abundance (%)									
	Average	Spruce age class forests			Beech age class forests			Unmanaged beech forests		
		SAF1	SAF2	SAF3	BAF4	BAF5	BAF6	BF7	BF8	BF9
Acidobacteria	**20.391**	**22.852**	**23.081**	**23.661**	**20.404**	**19.981**	**20.910**	**15.325**	**15.868**	**21.725**
Gp3	7.024	10.937	12.705	10.606	1.348	8.861	6.115	6.784	3.684	2.796
Gp16	2.951	0.103	2.495	4.010	5.049	2.256	3.595	2.204	3.017	4.080
Gp6	2.942	0.127	0.574	1.703	8.068	1.596	1.903	1.697	4.388	6.676
Gp1	2.931	10.626	5.532	3.143	0.064	2.443	1.063	1.535	1.179	0.439
Gp4	2.121	0.164	0.543	2.172	2.630	1.296	4.217	1.009	1.563	5.101
Gp7	1.145	0.021	0.562	1.129	1.001	2.074	2.447	1.030	0.958	0.889
*Acidobacteria**	0.330	0.449	0.377	0.149	0.354	0.422	0.384	0.291	0.263	0.280
Gp17	0.289	0.008	0.096	0.293	0.771	0.170	0.280	0.126	0.300	0.580
Gp5	0.201	0.032	0.108	0.178	0.233	0.340	0.241	0.258	0.136	0.286
Gp11	0.168	0.016	0.015	0.118	0.434	0.150	0.308	0.120	0.102	0.238
Gp22	0.160	0.005	0.012	0.092	0.360	0.138	0.228	0.150	0.195	0.253
Gp10	0.037	0.011	0.000	0.006	0.032	0.070	0.044	0.060	0.051	0.056
Gp2	0.030	0.158	0.015	0.020	0.006	0.041	0.005	0.009	0.000	0.000
Gp13	0.026	0.114	0.012	0.011	0.000	0.041	0.013	0.024	0.008	0.000
Gp15	0.022	0.026	0.035	0.020	0.010	0.038	0.041	0.006	0.011	0.006
Gp25	0.007	0.000	0.000	0.003	0.010	0.012	0.008	0.009	0.006	0.018
Gp18	0.006	0.000	0.000	0.009	0.019	0.000	0.000	0.006	0.006	0.015
Gp14	0.006	0.045	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Gp12	0.004	0.008	0.000	0.000	0.000	0.015	0.013	0.000	0.000	0.000
Gp20	0.003	0.003	0.000	0.000	0.010	0.000	0.003	0.003	0.003	0.009
Gp8	0.002	0.000	0.000	0.000	0.000	0.018	0.000	0.003	0.000	0.000
Gp19	0.001	0.000	0.000	0.000	0.000	0.000	0.003	0.000	0.000	0.006
Gp9	0.001	0.000	0.000	0.000	0.006	0.000	0.000	0.000	0.000	0.000

Table S7. Relative abundances of acidobacterial subgroups in the analyzed grassland soils. Values represent percentages of all sequences assigned to the domain Bacteria for all grassland soils or individual grassland soils. Groups labeled with asterisks could be assigned to the phylum level only.

Phylogenetic group	Relative abundance (%)									
	Average	Fertilized intensely managed grassland			Fertilized meadows grazed by horse and cattle			Unfertilized pastures grazed by sheep		
		FUG1	FUG2	FUG3	FMG4	FMG5	FMG6	UPG7	UPG8	UPG9
Acidobacteria	**18.710**	**22.873**	**15.946**	**18.150**	**26.835**	**19.571**	**21.572**	**17.306**	**13.226**	**14.034**
Gp16	6.798	9.925	4.380	3.321	9.137	8.197	11.816	3.925	4.629	4.147
Gp6	4.408	8.331	7.967	5.192	1.549	3.624	3.724	5.113	2.422	3.100
Gp4	2.798	1.771	1.372	3.593	3.669	2.567	2.331	3.638	3.025	3.505
Gp3	1.795	0.717	0.365	1.535	6.390	2.090	1.516	1.117	1.188	1.358
Gp7	1.386	1.005	0.588	2.135	3.771	1.258	0.963	1.721	0.889	0.784
Gp17	0.460	0.321	0.394	0.863	0.289	0.440	0.469	0.668	0.321	0.516
Gp11	0.304	0.321	0.337	0.315	0.269	0.400	0.252	0.238	0.307	0.287
Gp5	0.221	0.080	0.098	0.621	0.167	0.400	0.143	0.275	0.247	0.071
Gp1	0.183	0.003	0.000	0.166	1.267	0.064	0.059	0.102	0.024	0.076
Gp22	0.160	0.167	0.108	0.162	0.062	0.368	0.176	0.275	0.060	0.066
*Acidobacteria**	0.122	0.177	0.254	0.140	0.167	0.053	0.076	0.136	0.043	0.100
Gp18	0.024	0.010	0.029	0.034	0.010	0.042	0.015	0.038	0.030	0.011
Gp25	0.017	0.020	0.016	0.021	0.030	0.008	0.005	0.030	0.016	0.013
Gp20	0.016	0.010	0.019	0.000	0.026	0.048	0.010	0.008	0.019	0.000
Gp10	0.006	0.003	0.010	0.043	0.000	0.000	0.000	0.015	0.000	0.000
Gp15	0.005	0.000	0.000	0.009	0.016	0.005	0.015	0.000	0.003	0.000
Gp13	0.003	0.003	0.006	0.000	0.010	0.005	0.003	0.004	0.000	0.000
Gp9	0.002	0.007	0.003	0.000	0.000	0.000	0.000	0.004	0.003	0.000
Gp8	0.001	0.000	0.000	0.000	0.007	0.000	0.000	0.000	0.000	0.000

Table S8. Relative abundances of taxonomic groups within the phylum *Actinobacteria* and within proteobacterial classes in the analyzed forest soils. Values represent percentages of all sequences assigned to the domain Bacteria for all forest soils or individual forest soils. Groups labeled with asterisks could be assigned to the phylum level only.

Phylogenetic group	Relative abundance (%)									
	Average	Spruce age class forests			Beech age class forests			Unmanaged beech forests		
		SAF1	SAF2	SAF3	BAF4	BAF5	BAF6	BF7	BF8	BF9
Actinobacteria	**12.655**	**9.165**	**11.408**	**15.253**	**13.550**	**12.871**	**13.735**	**14.052**	**14.104**	**9.690**
Actinobacteridae	10.858	6.360	11.173	14.860	11.211	11.114	11.659	12.367	11.874	7.497
*Actinobacteria**	1.658	2.697	0.100	0.221	2.180	1.590	1.947	1.589	2.069	2.060
Rubrobacteridae	0.139	0.103	0.135	0.172	0.159	0.167	0.130	0.096	0.161	0.132
Acidimicrobidae	0.001	0.005	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Alphaproteobacteria	**25.072**	**42.759**	**33.603**	**18.272**	**14.338**	**22.480**	**19.453**	**28.957**	**26.515**	**19.388**
*Alphaproteobacteria**	10.729	10.924	8.628	6.788	7.988	11.492	11.801	12.451	14.271	11.323
Rhodospirillales	7.106	20.359	14.191	6.256	1.093	4.376	2.934	8.475	4.577	2.107
Caulobacterales	3.904	9.318	7.238	1.646	1.097	3.605	1.796	4.742	3.514	2.502
Rhizobiales	3.234	2.155	3.469	3.433	4.004	2.919	2.869	3.222	3.983	3.303
Rhodobacterales	0.081	0.003	0.073	0.147	0.124	0.053	0.034	0.048	0.141	0.115
Other	0.016	0.000	0.004	0.003	0.000	0.018	0.013	0.015	0.011	0.021
Betaproteobacteria	**5.991**	**3.735**	**2.476**	**5.191**	**8.992**	**6.482**	**5.534**	**5.220**	**8.187**	**7.738**
*Betaproteobacteria**	2.981	1.706	0.805	2.407	4.906	2.957	3.258	2.505	3.729	4.271
Burkholderiales	2.886	2.013	1.602	2.775	3.777	3.394	2.182	2.568	4.266	3.300
Other	0.124	0.016	0.069	0.009	0.309	0.132	0.093	0.147	0.192	0.168
Gammaproteobacteria	**2.966**	**4.667**	**1.910**	**1.278**	**3.522**	**2.740**	**2.048**	**2.646**	**3.370**	**4.253**
*Gammaproteobacteria**	1.940	2.242	0.616	0.755	2.783	1.980	1.369	2.039	2.242	3.247
Pseudomonadales	0.569	1.144	0.501	0.169	0.555	0.525	0.568	0.294	0.670	0.612
Xanthomonadales	0.267	0.977	0.628	0.190	0.019	0.103	0.039	0.120	0.221	0.127
Other	0.190	0.304	0.166	0.164	0.166	0.132	0.073	0.192	0.237	0.268
Deltaproteobacteria	**4.356**	**1.138**	**0.970**	**1.454**	**4.568**	**7.257**	**5.568**	**6.271**	**5.335**	**6.125**
Myxococcales	3.354	0.961	0.516	1.048	3.232	5.960	4.150	4.922	4.252	4.671
*Deltaproteobacteria**	0.969	0.169	0.262	0.399	1.316	1.258	1.407	1.333	1.069	1.416
Other	0.033	0.008	0.193	0.006	0.019	0.038	0.010	0.015	0.014	0.038

Table S9. Relative abundances of taxonomic groups within the phylum *Actinobacteria* and within proteobacterial classes in the analyzed grassland soils. Values represent percentages of all sequences assigned to the domain Bacteria for all grassland soils or individual grassland soils. Groups labeled with asterisks could be assigned to the phylum level only.

Phylogeneticgroup	Relative abundance (%)									
	Average	Fertilized intensely managed grassland			Fertilized meadows grazed by horse and cattle			Unfertilized pastures grazed by sheep		
		FUG1	FUG2	FUG3	FMG4	FMG5	FMG6	UPG7	UPG8	UPG9
Actinobacteria	**19.625**	**27.420**	**29.796**	**20.174**	**10.722**	**15.949**	**24.374**	**21.460**	**16.167**	**12.684**
Actinobacteridae	17.595	26.322	23.719	14.497	10.374	15.451	23.668	16.106	15.515	12.281
*Actinobacteria**	1.728	0.573	5.387	5.252	0.282	0.363	0.555	4.864	0.367	0.279
Rubrobacteridae	0.281	0.465	0.686	0.425	0.062	0.125	0.135	0.491	0.234	0.095
Acidimicrobidae	0.015	0.060	0.000	0.000	0.003	0.008	0.015	0.000	0.016	0.029
Coriobacteridae	0.005	0.000	0.003	0.000	0.000	0.003	0.000	0.000	0.035	0.000
Alphaproteobacteria	**11.434**	**8.321**	**8.758**	**8.930**	**10.768**	**10.862**	**6.765**	**8.830**	**17.374**	**19.636**
*Alphaproteobacteria**	5.737	2.726	4.056	3.849	5.192	4.581	2.838	3.936	9.508	12.850
Rhizobiales	3.115	3.817	2.935	2.173	2.251	3.245	2.308	2.457	4.142	4.158
Caulobacterales	1.390	0.824	0.670	1.314	1.474	1.539	0.415	1.117	2.748	2.145
Rhodospirillales	0.816	0.378	0.756	1.301	1.684	1.062	0.909	0.940	0.359	0.232
Rhodobacterales	0.354	0.559	0.327	0.259	0.154	0.419	0.283	0.351	0.576	0.226
Other	0.022	0.017	0.012	0.034	0.013	0.016	0.013	0.031	0.041	0.026
Betaproteobacteria	**5.863**	**7.504**	**6.953**	**7.272**	**5.274**	**5.595**	**3.739**	**6.408**	**5.624**	**5.584**
Burkholderiales	4.282	6.553	4.622	4.257	3.029	4.125	2.838	3.894	4.795	4.658
*Betaproteobacteria**	1.465	0.847	2.011	2.909	1.979	1.367	0.825	2.445	0.818	0.892
Other	0.116	0.104	0.321	0.106	0.266	0.103	0.076	0.068	0.011	0.034
Gammaproteobacteria	**2.743**	**2.032**	**2.271**	**3.934**	**1.303**	**2.967**	**1.409**	**4.751**	**3.284**	**3.339**
*Gammaproteobacteria**	1.791	1.122	1.544	2.717	0.748	2.143	1.032	3.479	2.088	1.758
Pseudomonadales	0.631	0.720	0.524	0.642	0.371	0.487	0.224	0.657	0.628	1.400
Other	0.177	0.137	0.165	0.374	0.089	0.143	0.107	0.434	0.139	0.129
Xanthomonadales	0.143	0.054	0.038	0.200	0.095	0.193	0.046	0.181	0.429	0.053
Deltaproteobacteria	**2.387**	**0.583**	**4.158**	**7.306**	**1.372**	**0.845**	**1.044**	**8.079**	**0.381**	**0.997**
Myxococcales	1.321	0.285	2.935	1.761	0.545	0.419	0.644	6.268	0.188	0.389
*Deltaproteobacteria**	0.978	0.275	1.191	5.426	0.541	0.363	0.242	1.713	0.182	0.579
Other	0.088	0.023	0.032	0.119	0.286	0.064	0.158	0.098	0.011	0.029

Table S10. Spearman's rank correlations between relative abundances of *Acidobacteria* subgroups and soil properties. Only relative abundances of acidobacterial subgroups that represented ≥ 0.029% of all analyzed sequences were considered.

Acidobacteria subgroup	Correlation			
	pH	Organic C	Total N	Sand/Silt/Clay
1	**-0.87**	-0.32	**-0.71**	0.02/-0.26/0.23
2	None	None	None	None
3	**-0.95**	-0.21	**-0.56**	-0.07/-0.22/0.29
4	**0.49**	-0.40	-0.11	0.20/0.04/-0.11
5	0.01	-0.16	-0.12	0.16/-0.15/0.12
6	**0.80**	0.24	**0.47**	0.08/0.11/-0.12
7	-0.07	-0.40	-0.19	0.35/-0.06/-0.04
10	-0.23	-0.18	-0.39	0.25/-0.13/0.10
11	**0.67**	0.15	**0.54**	0.14/0.43/-0.34
13	**-0.75**	-0.20	-0.45	0.25/0.02/-0.09
16	**0.60**	0.22	**0.64**	0.01/0.18/-0.15
17	**0.77**	0.03	0.30	0.02/0.14/-0.16
18	**0.76**	0.20	**0.49**	0.05/0.12/-0.19
22	0.41	0.09	0.15	0.34/0.15/-0.22

Bold numbers: $P < 0.05$; Bold and underlined numbers: $P < 0.001$. None: subgroup 2 was not detected in grassland.

Table S11. Dominant grasses of the analyzed grassland sites.

Management type	Sample	Dominant grasses
Fertilized intensely managed grassland	FUG1	*Arrhenaterum elatius*, *Trisetum flavescens*, *Poa trivialis*
Fertilized intensely managed grassland	FUG2	*Poa trivialis*, *Trisetum flavescens*
Fertilized intensely managed grassland	FUG3	*Poa trivialis*, *Trisetum flavescens*
Fertilized meadow grazed by horse and cattle	FMG1	*Poa trivialis*, *Alopecurus pratensis*, *Festuca pratensis*, *Lolium perenne*
Fertilized meadow grazed by horse and cattle	FMG2	*Poa trivialis*, *Trisetum flavescens*, *Alopecurus pratensis*, *Dactylis glomerata*
Fertilized meadow grazed by horse and cattle	FMG3	*Dactylis glomerata*, *Arrhenaterum elatius*, *Trisetum flavescens*, *Poa trivialis*, *Alopecurus pratensis*
Unfertilized pasture grazed by sheep	UPG1	*Brachypodium pinnatum*, *Festuca guestfalica*
Unfertilized pasture grazed by sheep	UPG2	*Bromus erectus*
Unfertilized pasture grazed by sheep	UPG3	*Bromus erectus*, *Brachypodium pinnatum*

8.3   Appendix C: Supplemental Information for chapter 4

Table S1. DNA yields, and physical and geochemical characteristics of the analyzed grassland and forest soil samples.

Sample	DNA yield (μg/g soil)	Soil group	pH	OC[a] (g kg^{-1})	Total N (g kg^{-1})	C:N ratio	Gravimetric water content (%)
SEG2	54.3	Histosol	7.42	120.8	12.3	9.8	84.8
SEG6	47.8	Histosol	5.22	284.1	23.9	11.9	162.9
SEG9	53.6	Histosol	6.23	229.7	18.7	12.3	114.8
SEW2	35.6	Arenosol	3.46	17.0	1.0	16.7	15.7
SEW5	36.9	Arenosol	3.05	29.6	1.6	18.3	25.9
SEW8	22.4	Albeluvisol	3.09	29.2	1.8	16.1	29.3
HEG1	26.4	Cambisol	6.63	66.2	6.2	10.6	38.5
HEG9	29.6	Stagnosol	6.62	48.3	4.1	11.7	30.9
HEG1[b]	8.6	Cambisol	7.10	17.8	2.0	8.9	23.2
HEG9[b]	6.9	Stagnosol	7.35	5.0	0.6	7.9	19.9
HEW2	17.2	Luvisol	4.45	50.6	3.1	16.2	49.1
HEW5	36.5	Leptosol	4.78	61.8	4.9	12.5	77.9
HEW9	31.6	Luvisol	4.09	48.1	3.4	14.0	64.8
AEG2	37.6	Leptosol	6.92	72.3	7.2	10.1	59.6
AEW1	56.3	Cambisol	3.30	64.6	4.0	16.3	62.8
AEW4	55.9	Cambisol	6.38	78.5	6.0	13.1	75.1
AEW5	49.9	Cambisol	4.52	57.5	4.5	12.9	70.4
AEW9	37.9	Cambisol	6.37	60.0	4.5	13.4	54.9

[a], Organic carbon content

[b], Libraries constructed from soil derived from B horizon

Table S2. Insert sizes of plasmids pLE01 to pLE08 and pLE10 to pLE38. For plasmids resulting from subcloning (pLE01, pLE04, pLE28, pLE29, pLE30, pLE33, pLE34, pLE35, and pLE37), the insert sizes of the corresponding fosmids are also depicted.

Plasmid	Insert size (bp)	Insert size corresponding fosmid (bp)
pLE01	2,300	19,800
pLE02	3,970	Not applicable
pLE03	11,172	Not applicable
pLE04	2,290	36,000
pLE05	2,658	Not applicable
pLE06	3,608	Not applicable
pLE07	2,752	Not applicable
pLE08	6,357	Not applicable
pLE10	1,629	Not applicable
pLE11	1,245	Not applicable
pLE12	4,202	Not applicable
pLE13	3,332	Not applicable
pLE14	4,586	Not applicable
pLE15	2,439	Not applicable
pLE16	8,591	Not applicable
pLE17	6,310	Not applicable
pLE18	1,107	Not applicable
pLE19	4,995	Not applicable
pLE20	8,411	Not applicable
pLE21	7,637	Not applicable
pLE22	3,087	Not applicable
pLE23	2,568	Not applicable
pLE24	3,264	Not applicable
pLE25	4,951	Not applicable
pLE26	2,804	Not applicable
pLE27	2,176	Not applicable
pLE28	2,764	25,000
pLE29	3,568	31,000
pLE30	1,708	30,400
pLE31	7,606	Not applicable
pLE32	3,381	Not applicable
pLE33	1,583	21,500
pLE34	2,448	36,500
pLE35	1,945	34,800
pLE36	2,663	Not applicable
pLE37	1,511	28,000
pLE38	2,716	Not applicable

Acknowledgements

Many thanks to my parents, my sister, and my future family-in-law for their belief in me and their constant support in a moral, financial, and statistical way.

Thanks to my friends for their confidence in me and for making me think outside the (microbiological) box. Thank you, Maike, for your constant friendship over the last 23 years, and, speaking of boxes, for your interest in rRNA ☺. I regret that we live so far apart.

In the end, very special thanks to the most important person in my life, my fiancé Harald. I don't know how I could have made it through the one or other crisis without your encouragement, especially during the last weeks. Thank you for your love, patience and understanding.

Curriculum vitae

Christiane Will, Dipl. biol.

Born on 13. March 1981 in Essen, Germany

Since 10/2007	PhD thesis at the Institute of Microbiology and Genetics, Georg-August University Göttingen
01/2007-10/2007	Scientific assistant at the Institute of Microbiology and Genetics, Georg-August University Göttingen
10/2000-11/2006	Study of Biology (Diploma) at the Georg-August University Göttingen
12/2005-11/2006	Diploma thesis: „Charakterisierung von neuartigen Genen und Genprodukten aus unkultivierten Mikroorganismen, die 4-Hydroxybutyrat-Verwertung vermitteln"
06/2000	Abitur at the Maria-Wächtler-Schule, Essen