Technische Universität München
Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt
Lehrstuhl für Bodenkunde

Functional and phylogenetic diversity of cellulase genes in agricultural soil under two different tillage treatments

Maria de Vries

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender:
Prof. Dr. Siegfried Scherer

Prüfer der Dissertation:
1. Hon. Prof. Dr. Michael Schloter
2. Prof. Dr. Wolfgang Liebl

Die Dissertation wurde am 17.04.2018 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 27.07.2018 angenommen.

Problems cannot be solved with the same mindset that created them.
— Albert Einstein

Table of Contents

Summary

The loss of stable organic matter and structure in agricultural soil is termed soil degradation and leads to the loss of soil fertility. Conventional tillage with concomitant soil inversion and soil aggregate disruption is regarded as one of the key factors facilitating this process. The application of conservation or non-inversion tillage with a reduced depth has been shown to mitigate the loss of native soil organic matter to the atmosphere. Soil microorganisms are key players in the global carbon cycle because of their involvement in the degradation of bioavailable organic matter in soil. The majority of organic matter input to agricultural soil is in the form of dead plant material, of which the organic polymer cellulose is one of the major components. Indeed, soil microorganisms are well equipped with a diverse array of genes coding for cellulose-degrading enzymes. The aim of this research was to identify differences in the genetic potential to degrade cellulose of soils subjected to conventional or reduced tillage treatment. It was hypothesized that soil under reduced tillage would be enriched in and comprise a higher diversity of cellulase genes compared to soil under conventional tillage, thereby functioning as a biomarker for soil fertility. To this end, soil was sampled from two agricultural experiments in Western Europe applying both tillage treatments. These agricultural fields differed in farming management regimes and experiment duration; the twenty-year experiment was managed using conventional farming practices while the four-year experiment was based on an organic farming management-regime. From both fields the tilled surface horizon of soil under both tillage treatments was sampled. From the organic farming experimental field, samples were additionally taken from the soil layer beneath the tillage layer. DNA was extracted from these samples and the metagenomes were subsequently shotgun-sequenced using high-throughput next-generation sequencing techniques.

Metagenome analysis results showed that tillage intensity did not affect the genetic potential for endoglucanases or exoglucanases in the metagenome of either experiment. However, in the organic farming experiment, the surface soil horizon of plots with reduced tillage treatment was enriched mainly in β-glucosidase genes (including glycoside hydrolase (GH)1-genes), GH26-genes and genes coding for carbohydrate binding modules (CBM2, CBM3 and CBM6) compared to that of plots with conventional tillage treatment. In contrast to results on DNA-level, measurements of enzyme activity in soil from the conventional farming experiment demonstrated a higher potential β-glucosidase- and exoglucanase-enzyme activity under reduced tillage compared to conventional tillage in the surface soil horizon. The effects of tillage intensity observed on the whole microbial community in both agricultural soil metagenomes indicated a secondary role for tillage in shaping the microbial community structure and a possibly higher importance of factors like crop residue incorporation, plant root development and time point of sampling after tillage treatment. In the soil metagenomes of both agricultural experiments, the most abundant cellulases were β-glucosidases (GH1 and GH3), cellulose phosphorylases (GH94) and genes from CBM-family 2. Endoglucanase genes belonging to GH5 and GH74 were most abundant, while exoglucanase genes were infrequent. Most cellulase genes were derived from members of the Actinobacteria (Actinomycetales) and Proteobacteria (e.g. Xanthomonodales, Rhizobiales and Myxococcales), but also of Bacteroidetes. Furthermore, based on cellulase gene sequences identified in the conventional farming metagenome, a primer set was developed for amplification of GH5-subfamily 2-cellulase genes from agricultural soil DNA. In-depth analysis of amplified GH5-subfamily 2-genes by high-throughput

sequencing revealed a high diversity of unexplored putative GH5-cellulase sequences and indications of cross-kingdom horizontal gene transfer.

In light of these results it can be concluded that tillage intensity mainly affects genes coding for proteins involved in cellulose binding and the degradation of cellulose oligomers. Although cellulase gene diversity and the genetic potential for cellulose polymer degradation were not strongly affected by tillage intensity, enzyme activity measurements do indicate a tillage effect on transcriptional or translational level. Therefore, to elucidate effects of agricultural practices like tillage intensity on the role of soil microorganisms in the global carbon cycle, investigation of transcriptional and enzymatic responses by the cellulolytic soil microorganisms to different sources of organic compounds over the course of degradation is recommended as future focus of research.

Zusammenfassung

Das Nachlassen der Bodenstruktur und die Verringerung von stabiler organischer Substanz in landwirtschaftlichen Böden wird als Bodendegradation bezeichnet und führt zum Verlust der Bodenfruchtbarkeit. Konventionelle oder wendende Bodenbearbeitung und damit einhergehender Zerstörung der Bodenaggregate wird als einer der Hauptfaktoren für diesen Prozess angesehen. Es hat sich gezeigt, dass die Anwendung konservierender oder nichtwendender Bodenbearbeitung mit reduzierter Tiefe den atmosphärischen Verlust von nativer organischer Substanz reduziert. Bodenmikroorganismen spielen eine Schlüsselrolle im globalen Kohlenstoffkreislauf, da sie am Abbau biologisch verfügbarer organischer Substanz im Boden beteiligt sind. Der Großteil der organischen Substanz, der dem landwirtschaftlichen Boden hinzugefügt wird, besteht aus nicht-lebendem Pflanzenmaterial, von dem das organische Polymer Cellulose eine der Hauptkomponenten ist. In der Tat sind Bodenmikroorganismen bestens ausgestattet mit einer Vielfalt von Genen, die Cellulose-abbauende Enzyme kodieren.

Ziel dieser Forschung war es, Unterschiede im genetischen Potential für den Abbau von Cellulose im Vergleich bestimmter Böden zu identifizieren, die einer konventionellen oder konservierenden Bodenbearbeitung unterzogen wurden. Es wurde die Hypothese aufgestellt, dass Boden unter konservierender Bodenbearbeitung eine Anreicherung und eine höhere Diversität an Cellulase-Genen aufweisen würde als Boden unter konventioneller Bodenbearbeitung und somit als Biomarker für Bodenfruchtbarkeit fungierte. Zu diesem Zweck wurden Bodenproben aus zwei landwirtschaftlichen Experimenten in Westeuropa entnommen, die beide Bodenbearbeitungstechniken anwendeten. Diese landwirtschaftlichen Experimenten unterschieden sich in den Bewirtschaftungssystem und der Versuchsdauer. Das zwanzigjährige Experiment wurde mit konventionellen Anbaumethoden durchgeführt, während das Vierjahresexperiment auf einem Managementsystem für ökologische Landwirtschaft basierte. Von beiden Feldern wurde der bearbeitete Oberflächenhorizont des Bodens unter beiden Bodenbearbeitungstechniken beprobt. Aus dem Versuchsfeld des ökologischen Landbaus wurden zusätzlich Proben aus der Bodenschicht unterhalb der Bodenbearbeitungsschicht entnommen. Die DNA wurde aus diesen Bodenproben extrahiert, und die Metagenome wurden anschließend Schrotschuss-sequenziert – unter Verwendung von Hochdurchsatz-Sequenzierungstechniken.

Metagenomanalyseergebnisse zeigten, dass die Bodenbearbeitungsintensität das genetische Potenzial für Endoglucanasen oder Exoglucanasen im Metagenom beider Experimente nicht beeinflusste. Dennoch, im ökologischen Landbauexperiment war der Oberflächenbodenhorizont – beim Vergleich von konservierender zu konventioneller Bodenbearbeitung – mit β-Glucosidase-Genen (einschließlich Glykosidhydrolase (GH) 1-Genen), GH26-Genen und Genen für Kohlenhydrat-Bindungsmodule (CBM2, CBM3 und CBM6) angereichert. Im Gegensatz zu den Ergebnissen auf DNA-Ebene zeigten Messungen der Enzymaktivität im Boden des konventionellen Landwirtschaftsexperiments eine höhere potentielle β-Glucosidase- und Exoglucanase-Enzymaktivität bei reduzierter Bodenbearbeitung im Vergleich zur konventionellen Bodenbearbeitung an der Bodenoberfläche. Die beobachteten Auswirkungen der Bodenbearbeitungsintensität auf die gesamte mikrobielle Gemeinschaft in beiden landwirtschaftlichen Bodenmetagenomen zeigten eine sekundäre Rolle für Bodenbearbeitung bei der Gestaltung der mikrobiellen Gemeinschaftsstruktur und eine möglicherweise höhere Bedeutung von Faktoren wie Pflanzenrückstandsbildung, Pflanzenwurzelentwicklung und Zeitpunkt der

Probenahme nach Bodenbearbeitung. In den Bodenmetagenomen beider landwirtschaftlicher Experimente waren die am häufigsten vorkommenden Cellulasen β-Glucosidasen (GH1 und GH3), Cellulose-Phosphorylasen (GH94) und Gene aus der CBM-Familie 2. Endoglucanase-Gene, die zu GH5 und GH74 gehören, waren am häufigsten vorhanden, während Exoglucanase-Gene selten waren. Die meisten Cellulase-Gene stammten von den Actinobacteria (Actinomycetales) und Proteobacteria (z.B. Xanthomonodales, Rhizobiales und Myxococcales), aber auch von Bacteroidetes. Basierend auf Cellulase-Gensequenzen, die im konventionellen Landwirtschafts-Metagenom identifiziert wurden, wurde außerdem ein Primer-Set zur Amplifikation von GH5-Subfamilie-2-Cellulase-Genen aus der landwirtschaftlichen Boden-DNA entwickelt. Eine detaillierte Analyse von amplifizierten GH5-Subfamilie-2-Genen durch Hochdurchsatz-Sequenzierung ergab eine große Diversität von unerforschten vermutlichen GH5-Cellulase-Sequenzen und Hinweise auf Domän-überschreitenden horizontalen Gentransfer.

Angesichts dieser Ergebnisse kann geschlussfolgert werden, dass die Bearbeitungsintensität hauptsächlich Gene beeinflusst, die für Proteine kodieren, die an der Cellulose-Bindung und dem Abbau von Cellulose-Oligomeren beteiligt sind. Obwohl die Cellulase-Gen-Diversität und das genetische Potenzial für den Abbau von Cellulosepolymeren nicht stark von der Intensität der Bodenbearbeitung beeinflusst wurden, deuten Messungen der Enzymaktivität auf einen Bodenbearbeitungs-Effekt auf der Transkriptions- oder Translations-Ebene hin. Um die Auswirkungen von landwirtschaftlichen Methoden wie der Bodenbearbeitungsintensität auf die Rolle von Bodenmikroorganismen im globalen Kohlenstoffkreislauf aufzuklären, wird daher die Untersuchung der transkriptionellen und enzymatischen Reaktionen der cellulolytischen Bodenmikroorganismen auf verschiedene Quellen organischer Verbindungen im Verlauf des Abbaus als zukünftiger Forschungsschwerpunkt empfohlen.

1 - Introduction

Soil degradation

Soil carbon stocks are more than twice as high as carbon amounts in the atmosphere (1, 2), amounting to around 2500 Gt of carbon in the top 2 m of soil (3). Of this carbon stock, around 0.8 Gt is estimated to be lost yearly (4), leading to the degradation of soil quality. Soil degradation is the loss of soil native fertility, characterized by loss of soil structure and stable organic matter and the concomitant loss of nutrients and nutrient-holding capacity. The start and intensification of agriculture has often been considered as one of the greatest causes of the increase in soil degradation (4, 5). Conversion to agricultural systems has been estimated to cause the depletion of at least 60% of the soil organic carbon pool (5, 6). Agricultural intensification in the western part of the world has led to increasing yield and energy efficiency but also to non-holistic and short-term nutrient and soil management strategies. Agricultural intensification is generally characterised by a low carbon or organic matter return to soil, because fertilization strategies are currently focussed on chemical nitrogen and phosphorous addition. Low soil organic matter content together with long periods of fallow induces high soil loss by erosion. In non-western areas, where farmers often cannot afford to use chemical or organic fertilizers, soil degradation is prevalent because of a general lack of nutrient return to soil (7).

Several strategies exist to overcome soil degradation like growing cover crops and catch crops. In addition, soil degradation can be mitigated by directly increasing the return of organic matter to the soil. A well-known practice is crop residue retention on the soil surface or incorporation into the top soil layer. Carbon derived from crop residues functions as a substrate to the soil microbiome and is partly immobilized in the microbial biomass (8), partly lost as carbon dioxide, and partly ends up as stable soil organic matter. Substrates which are more difficult to degrade, e.g. cellulose, are incorporated into microbial biomass in higher amounts than easily-utilizable carbon sources, e.g. glucose (9). The part of crop residues which ends up as stable organic matter originates from degradation-recalcitrant molecules in the fresh substrate or from microbial biomass (10). Degradation-recalcitrant molecules of fresh organic matter are less bioavailable due to their structure or because of limiting soil conditions like low soil moisture (11, 12) or presence of micro-aggregates (13)). Furthermore, reduction of tillage intensity has been proposed in order to mitigate the loss of native soil organic matter to the atmosphere. Tillage of soil is primarily done as seedbed preparation, as it loosens and dries the soil and therefore provides a suitable environment for crop roots to develop. Indeed, tillage has been shown to increase soil aeration and reduce soil compaction or soil bulk density in the tilled layer (14, 15) which leads to a lower soil volumetric water content (13). Furthermore, tillage is applied for weed control by disrupting the roots of growing vegetation and turning the soil to cut off its access to light. Finally, tillage leads to homogeneous nutrient distribution in the soil by mixing. However, a negative effect of intensive tillage which impacts the quality of the soil is the lower organic carbon content in the surface or tilled soil horizon and the lack of organic matter stratification compared to non-tilled soil, (16–20). Therefore, combining crop residue retention with no-tillage techniques has become the main recommendation to restore soil organic carbon stocks (17, 21). Although this method includes lower

costs by saving on fuel for tilling, it is not always popular because of lower yields (22, 23). Other alternatives to conventional tillage are tillage methods with reduced intensity and depth, for example chisel plough, rotary harrow and cultivator. Whereas conventional tillage is mostly done with a mouldboard plough to overturn the soil until a depth of 20-35 cm, reduced tillage methods are non-overturning tillage techniques with a shallower working depth (14). Depending on the climate conditions and soil type, reduced tillage methods lead to higher carbon levels in the tilled layer than conventional tillage practices and they do not show the same strong negative impact on yield as no-tillage methods (18, 22, 24, 25).

Degradation of cellulose by soil microorganisms

Microorganisms play an important role in the degradation of native and fresh organic matter in soil. While the availability of native organic matter is governed mainly by edaphic factors, actual degradation of available organic matter is performed by the soil microbiota (26). The soil microbiota consists for a large part of micro-fauna, like insects, earthworms, termites, nematodes and other small animals, but the majority of the soil biomass consists of microorganisms, including protists, fungi, Archaea and bacteria. Hence, their usage of the organic material as carbon and energy source is a significant contribution to the soil carbon cycle.

Organic matter enters the soil partly through living plants which provide the soil microbial community with root exudates (27). However, the bulk of organic matter entering the soil is in the form of dead plant biomass (28, 29). Plant biomass generally consists of a mixture of organic polymers like hemicellulose (xylan, mannan and xyloglucan), lignin, pectin and, most importantly, cellulose. It is one of the most abundant glycans on earth, representing a major element of plant biomass (30) and thus a considerable portion of the organic matter input to soil. Cellulose is a relatively simple compound, but quite inaccessible for breakdown as it consists of closely stacked fibres of β-1,4-linked glucose disaccharides called cellobiose. Cellulose fibres are stacked in highly ordered crystalline structures which are now and then interspersed by less well-ordered amorphous regions, where the cellulose fibres are not so closely packed together (reviewed by (31)). Because of this organization and the close proximity of many other plant carbohydrates, the bio-availability of cellulose is relatively low. Different enzymes exist which catalyse the cleavage of the cellulose-polymer on different sites and using different mechanisms (summarized in (32)). Endoglucanases, exoglucanases and β-glucosidases act on cellulose through hydrolysis (Figure 1). Of these, endoglucanases (Enzyme Commission (EC) number: 3.2.1.4) cut internal sites on amorphous cellulose regions or on soluble forms of cellulose like carboxymethyl cellulose (CMC), by which new chain ends are produced for exoglucanases to bind to. Exoglucanases cleave the ends of crystalline cellulose chains to release cellobiose and are also called cellobiohydrolases or cellobiosidases. Of these, EC: 3.2.1.176 acts on the reducing end of the cellulose chain and EC: 3.2.1.91 on the non-reducing end. Exoglucanases which cut the ends of cellooligomers (cellodextrins), releasing cellobiose or glucose, are called cellodextrinases (EC: 3.2.1.74). Finally, cellobiose is degraded to glucose monomers by β-glucosidases (EC: 3.2.1.21). Most endoglucanases and exoglucanases are excreted extracellular, whereas most β-glucosidases are located intracellular, after the uptake of cellobiose or chemical modifications thereof. Although most cellulases cleave the β-1,4-bonds by hydrolysis, cellulases using phosphorylation-mediated cellulose cleavage also exist (33). Cellodextrin phosphorylases (EC: 2.1.4.49) release glucose monomers from cellodextrin molecules. Cellobiose

phosphorylases (EC: 2.1.4.20) and cellobionic acid phosphorylases (EC: 2.4.1.321) degrade cellobiose (or a chemical modification of cellobiose) to different forms of glucose monomers (34, 35). Oxidoreductases, like cellobiose dehydrogenases (EC: 1.1.99.18) and lytic polysaccharide monooxygenases (LPMOs, recently given the EC: number 1.14.99.54) use an oxidative catalytic mode of action to modify cellobiose respectively cellulose chains (36).

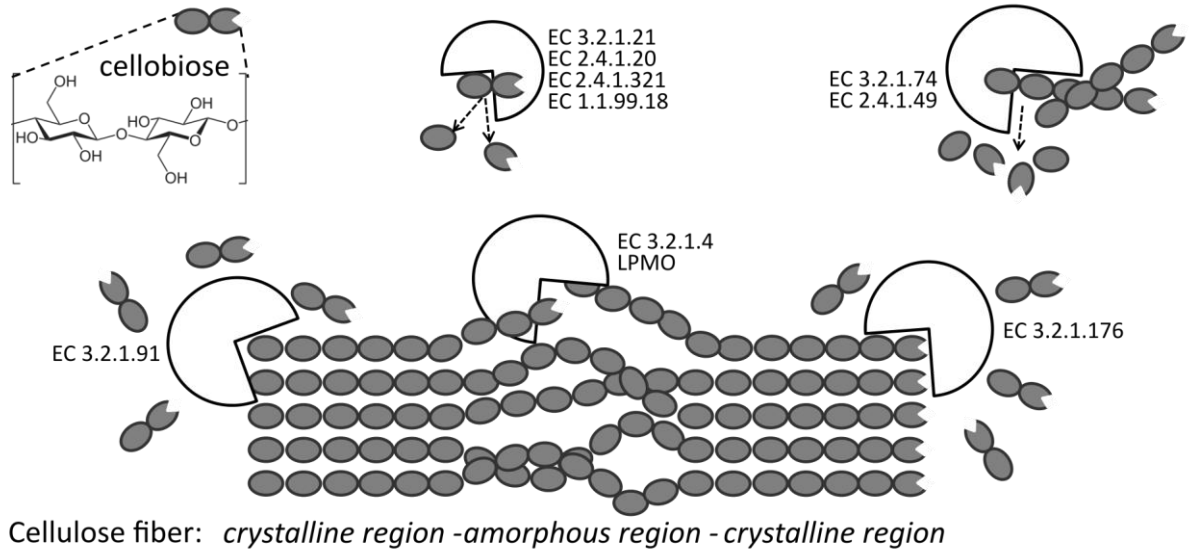


Figure 1: Overview of binding- and cleaving-locations of different cellulase enzymes on the cellulose substrate. β-glucosidase (EC: 3.2.1.21), cellobiose phosphorylase (EC: 2.1.4.20), cellobionic acid phosphorylase (EC: 2.4.1.321) and cellobiose dehydrogenase (EC: 1.1.99.18) act on cellobiose, while cellodextrinase (EC: 3.2.1.74) and cellobiose phosphorylase (EC: 2.1.4.49) act on cellulosic oligomers. The cellulose fiber is internally cleaved by endoglucanase (EC: 3.2.1.4, amorphous region) and LPMO, which can cleave the cellulose polymer both in the amorphous and the crystalline region. Cellulose chain ends can be cleaved by reducing-end exoglucanases (EC: 3.2.1.176) or by non-reducing-end exoglucanases (EC: 3.2.1.91).

Many microorganisms have been physiologically characterized as cellulose degraders. These include various bacterial, fungal and metazoan taxa which have been regularly described in reviews (e.g. (32, 37)). Bacterial orders with most cellulose degraders are the Clostridiales (phylum Firmicutes) and the actinomycetes (phylum Actinobacteria) (38). Other well-known cellulose-degrading bacterial genera from the Firmicutes are *Bacillus* and *Caldicellulosiruptor*, from the Bacteroidetes *Cytophaga*, *Sporocytophaga* and *Bacteroides*, *Fibrobacter* from Fibrobacteres and *Pseudomonas, Myxococcus* and *Cellvibrio* from the Proteobacteria (39–41). Among fungi, well-known degraders are *Trichoderma*, *Penicillium*, *Aspergillus*, *Fusarium* and *Chaetomium* from Ascomycota, and *Phanerochaete*, *Sclerotium, Volvariella* and *Gloeophyllum* from the Basidiomycota and the anaerobic *Neocallimastix* (32, 42, 43). Finally, animals able to degrade cellulose are for example (parasitic) nematodes (44, 45), termites (46, 47), insects (48) and some (mostly symbiotic) protists (49, 50). New genera of cellulolytic bacteria are discovered every day by analysis of sequenced genomes, for example *Actinospica*, *Microbispora* and *Hamadaea* (Actinobacteria) and *Cystobacter* (Deltaproteobacteria) (51). Analysis of metagenomic datasets has additionally resulted in the identification of many potential degraders (52).

Cellulase activity and gene expression by cellulolytic microorganisms is regulated by the amount and type of available substrate. In bacteria, pure cellulose has been demonstrated to induce

cellulase enzyme activity whereas glucose and cellobiose inhibit it (53), while cellulase gene expression was shown to be enhanced during growth on cellulose but repressed by glucose in *Thermobifida curvata* (54). In fungi, cellulase gene expression is induced by cellulose while glucose inhibits it (55–58). Disaccharides like cellobiose, lactose and sophorose can also induce cellulase gene expression in some fungi (55). Conversely, monosaccharides like fructose and 2-deoxyglucose seem to generally repress cellulase gene expression, whereas glycerol, mannitol, sorbitol, and maltose did not affect expression (58). All in all, available data suggest that cellulose production occurs on basal levels without direct need for induction but is repressed by catabolites (37).

Cellulase domain families and substrate specificity

Much work has been put in attempts to predict cellulolytic ability of microbial groups by analysing the respective gene clusters. However, the variability in cellulase-coding gene sequences and even protein sequences leads to major difficulties in the prediction of function. A useful classification system has been developed by Henrissat *et al.*, based on amino acid sequence similarities on a secondary structure-level (59). Cellulases display a modular structure, where one or more catalytic domains and often a cellulose binding module are present in the enzyme. These catalytic and binding domains of cellulases and other carbohydrate-active enzymes have been classified into different families. Currently, the CAZy database (Carbohydrate-active enzymes database, www.cazy.org, (60)) contains and updates a vast quantity of carbohydrate-active enzyme sequences classified into a module family. Different classes of catalytic and binding modules are recognized: Carbohydrate Binding Modules (CBM), Glycoside Hydrolases (GH, generally comprising enzymes with a hydrolytic activity (EC: 3.2.1.-) but also several with phosphotransferase activity (EC: 2.4.1.-)) and the Auxiliary Activities (AA)-class. The AA-class comprises enzymes related to lignocellulose conversion; mainly Lytic Polysaccharide Mono-oxygenases (LPMO's) and dehydrogenases (61).

Cellulase mode of action and substrate specificity

Enzymes within one catalytic or binding domain family display similarities in amino acid sequence, secondary and tertiary structure (overall fold) and catalytic mechanism. Cellulolytic catalysis is mostly performed by two highly conserved amino acid residues: a general acid (proton donor) and a nucleophile/base (62), defining the catalytic mechanism. In addition, there are between-family similarities in 3D-structure and catalytic machinery, indicating that some families are more related than others (i.e. superfamilies (63)). Enzymes within one catalytic or binding domain family do not, however, necessarily perform the same enzymatic function or act on the same substrate. Many of these families contain enzymes which act on different carbohydrates, and only some contain enzymes acting on cellulose. Family GH5, for example, is harboured by enzymes catalysing twenty different enzymatic reactions, many of which are even bifunctional (www.cazy.org, (60)). For a summary of the information found on the CAZy database (October 2017), see Figure 2. In fact, there appear to be only two domain families containing enzymes acting on cellulose, with a single cellulolytic activity: GH45 and GH124. The substrate specificity and, therefore, enzymatic function is thought to depend on both the catalytic domain and the CBM. There are many indications that the substrate specificity of many cellulases is governed by the associated CBM in the protein (64). Generally, the function of the carbohydrate binding module is threefold (65); First, CBMs ensure close contact between the catalytic domain and the cellulose molecules. Second, they ensure the binding of the right cellulose substrate type, i.e. crystalline or amorphous cellulose, or cellodextrins.

In fact, at least some CBMs not only select for substrate type, but also for specific regions on the same substrate type, as was shown by Carrard *et al.* for CBMs 1 and 3 (66). Third, CBMs function by aiding in the disruption of the cellulose fibres. This CBM-mediated disruption (often CBM3a) is probably the feature defining whether or not some exoglucanases and even endoglucanases show a processive mode of action (67).

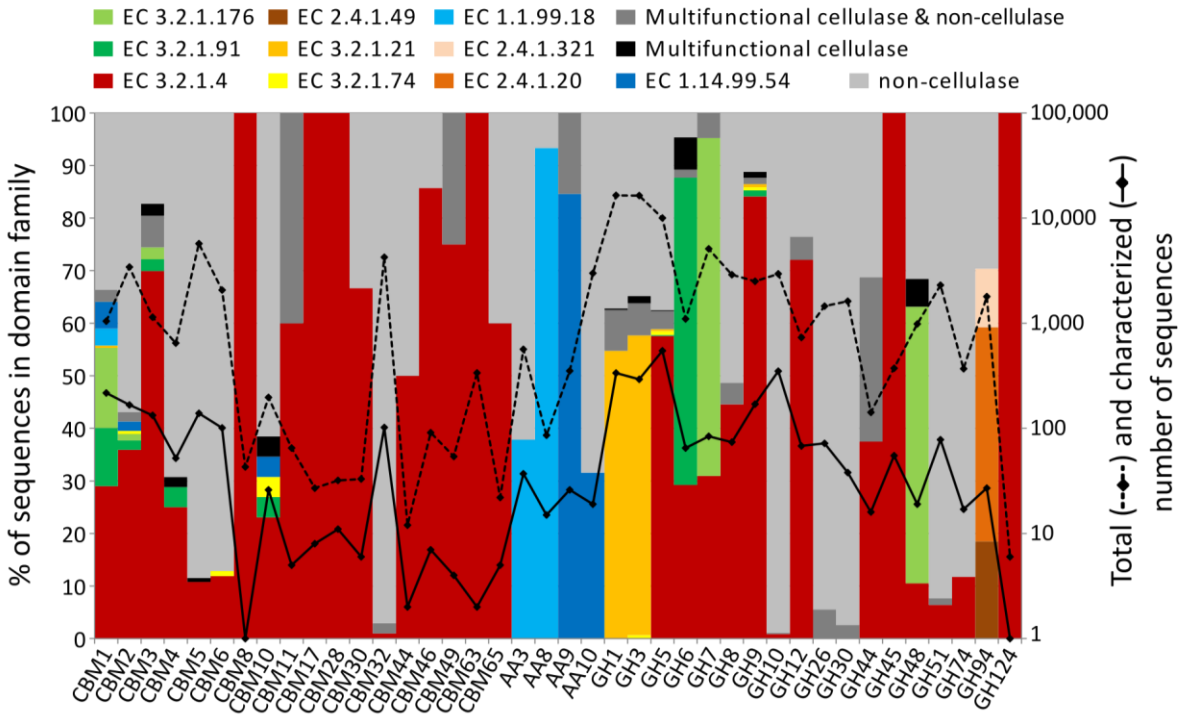


Figure 2: Distribution of enzymatic functions (EC-number) involved in cellulose degradation among sequences classified in domain families found in the CAZy database. Shown is the percentage of sequences per domain family of which the cellulase function has been functionally characterized, as well as the percentage of multifunctional cellulase- and of non-cellulase-sequences (primary y-axis). In addition, the total number of (both functionally characterized and non-characterized) sequences and the number of functionally characterized sequences classified into the respective domain family is shown (secondary y-axis). Data was extracted from the CAZy database (www.cazy.org, (60) - October 2017).

Considering this, the importance of a CBM for cellulase efficiency appears significant and raises the question why some cellulases do not contain a CBM. It was postulated that cellulase enzymatic groups without a binding module can have several substrate specificities (42), or that they may play an important role in amorphous cellulose degradation where larger enzymes (i.e. with CBM) have no access (68). In any case, identification of a gene's binding module should give insight into its substrate recognition characteristics, although the overall fold-classification into a CBM-family does not predict its function (65), much like the catalytic modules. Interestingly, the catalytic domain of a cellulase can also select for a specific substrate. As was discussed by Davies and Henrissat (62), details in the structural fold of the catalytic module define the substrate specificity and mechanistic action of the enzyme. They explain that the substrate specificity of the catalytic module seems to depend on the 3D-topology of the active site, of which there are three types: a pocket, a groove and a tunnel. The pocket is designed to fit a saccharide non-reducing chain end, whereas the groove allows the binding of a random part of the polymer. Finally, the tunnel-topology is formed when a long loop covers the groove otherwise left open. It allows the protein to processively continue its

cleavage along the polysaccharide chain, and it is found mostly in cellobiohydrolases but also in some endoglucanases (62).

These details do seem to have quite an effect on substrate specificity, as was shown for GH3-β-glucosidases (69). While substrate specificity of the catalytic module seems indispensible for β-glucosidases as they generally do not possess a CBM (www.cazy.org, (60)), it has also been shown to be present in catalytic modules of endo- or exoglucanases. For GH48-family enzymes, a single residue has been identified defining its specificity for cellulose versus non-cellulosic substrates (70). In addition, Chen *et al*. (71) identified six active site-motifs which strongly influenced the substrate-recognition promiscuity of GH5-cellulases. Furthermore, the importance of the size and amino acid density and composition of the active site for substrate-binding affinity was elucidated by Tian *et al*. (72). Exoglucanases tended to show longer active sites and higher binding energies, while β-glucosidase-binding affinity was concentrated in a small active site. Moreover, they described the ligand-binding properties of the conserved residues in the active sites of several GH-families, elucidating causes of substrate affinities between cellulases, xylanases and β-glucosidases (72).

Functional redundancy of cellulases

Despite the difficulties in identification of cellulolytic function purely from gene sequences, many studies put effort in characterizing the cellulolytic trait-dispersion among not-physiologically characterized microorganisms. By analysing publically available complete bacterial genomes, Berlemont and Martiny showed that cellulase genes are present in the majority of bacterial genomes and in most microbial groups (73). 56% of the analysed bacterial genomes contained β-glucosidases and no endo- or exoglucanases, indicating that the opportunistic use of cellulose degradation products is widespread among bacteria (73). However, cellulose degraders, which possess both endo- or exoglucanases and β-glucosidases, were only present in 24% of the analysed bacterial genomes (73). In addition, they could show that potential cellulose degraders formed phylogenetic clusters at the maximum taxonomic level of genus (73, 74). All in all, this indicates that cellulose utilisation might be functionally redundant among heterotrophic and non-parasitic bacteria. Cellulose degradation is, however, more restricted to certain taxonomic assemblages, although these degrading communities also show a certain functional redundancy (52).

Indeed, substrates containing different carbohydrate amounts or types select for a different degrading community composition, as was shown for fungi (75). In addition, substrate type was shown to influence the synergism among microorganisms (76). Considering that the plant cell wall consists of many different glycan polymers, microorganisms are therefore at an advantage when they can degrade many different types of carbohydrates. Indeed, adaptation of microorganisms to different carbon sources is characterized by the use of different carbohydrate-active enzymes. For example, differences in tree species affected the transcriptional diversity of the fungal lignocellulolytic genes GH7, GH11 & AA2 (77) and the leaf litter horizon in forest soil exhibits a higher diversity and proportion of *cbhI*-genes than the organic horizon, as shown by metatranscriptomics (78). Furthermore, wheat straw and tree leaf litter induce different patterns of degrading enzyme production, with wheat straw inducing higher cellulase production than leaves (79). Moreover, the composition of cellulosomes was shown to vary according to the substrate type (80, 81), also illustrating the adaptive abilities of microorganisms to the nutritional status of their environment.

The advantages of cellulase functional diversity, i.e. endoglucanases, exoglucanases and β-glucosidases, for cellulose degradation are evident considering their synergistic activity. For example, LPMO's can enhance the activity of endoglucanases and exoglucanases (82, 83). In addition, optimal cellulose degradation is often achieved using both endoglucanases and exoglucanases (84, 85); microorganisms containing both enzymes or dual enzymes are very efficient degraders (86). Nevertheless, bacterial genomes containing genes for both types of cellulases are not altogether common; Whereas predicted cellulolytic bacteria were shown to all harbour endoglucanase- and multiple β-glucosidase and/or cellobiose phosphorylase genes, not all possessed exoglucanases (39, 73). Moreover, exoglucanases show additional substrate specificity by exhibiting either reducing-end (EC: 3.2.1.176, GH7, GH9 and GH48) or non-reducing end activity (EC: 3.2.1.91, GH5, GH6 and GH9) and thus further induce synergy in degradation (87). In summary, this suggests that cellulose-degrading microbial consortia greatly enhance degradation efficiency when containing an exoglucanase-producing member.

Understanding the role of cellulases which have the same substrate-specificity but belong to different domain families is, however, less straightforward. To date, much knowledge is lacking about the function of the diversity of enzymes with the same substrate specificity. Nevertheless, the importance of enzyme phylogenetic diversity for cellulose degradation is becoming increasingly clear. A high phylogenetic variability of cellulase genes may increase the chance of substrate recognition and degradation (43, 88). Indeed, combining two exoglucanases from different organisms but from the same cellulase domain family resulted in synergistic improvement of degradation rate compared to the performance of either single exoglucanase (89). The synergy of different cellulase domain families is especially well illustrated by the cellulase-gene repertoire of cellulosome-harbouring bacteria. For example, Raman *et al.* (80) showed that the abundance of GH5- or GH9-enzymes in the cellulosomes of *Clostricium thermocellum* increased when cellulose and crystalline cellulose, respectively, were present as carbon source. While these results indicate a specificity for a certain region on the cellulose chain, these two GH-families were in fact generally found in relatively high abundance among cellulosome-enzymes, regardless of substrate type, also suggesting a degree of functional redundancy within these domain families (80). In addition, whereas cellulosomes all harbour a single copy of the GH48-exoglucanase, they generally harbour multiple copies of the GH9-gene, of which the expressed proteins all show different functional properties (90). A reason for this synergy between cellulases from the same domain family may be differences in substrate-binding affinity or specificity for different regions on the same substrate. For instance, cellulose-binding CBMs of the same domain family were found to be specific for a different region on the cellulose chain (66). In addition, several GH9-endoglucanases from *Clostridium cellulolyticum* showed different catalytic activities and binding affinities to amorphous cellulose, despite them belonging to same GH-family and harbouring the same carbohydrate binding module (91). Besides putatively different substrate-binding affinities, steric characteristics of cellulase domains appear to influence their contribution to cellulose degradation efficiency, as was elucidated by experiments with designer-cellulosomes (92, 93).

The adoption of cellulase-coding genes that harbour multiple different domain families might in fact be an adaptive strategy to adjust to the high diversity of carbohydrate molecules present in the environment and the high diversity of their steric conformations (43). Cellulase genes are optimally suited to readily evolve into performing a similar but different degradation reaction, as only a small number of residues in the active site seem to be important for substrate specificity and catalytic

action (69–72). This is reflected in the high protein sequence-variation found within cellulase domain families, indicating that single changes in amino acid residues should not have such strong effects on total protein fold, thereby maintaining the evolutionary recognizable family fold (63). At the same time, there is a chance that single residue changes elicit a change in the binding site topography (65) (71), leading to alteration of the substrate recognition, whereas the chance of altering the two conserved catalytic residues is very small, safeguarding the catalytic activity. Moreover, it has been postulated that certain protein families (e.g. GH5) exhibit inherent properties for multispecificity, i.e. their active/binding site topography is such that single residue changes can readily elicit additional substrate recognitions (71).

Another possible strategy for adaptation is to share highly transferrable genes with other (related) microorganisms by horizontal gene transfer (HGT). Indeed, the concept of a common gene pool which is shared between microorganisms has been proposed before as an ecologically relevant mechanism (for a review see Boon *et al.* (94)). As cellulase domain families are defined based on amino acid sequence similarity (59), it follows that proteins within one family are phylogenetically more related than proteins from different cellulase domain families. The fact that many cellulase domain families are found over a large range of microorganisms (e.g. GH5) and that different, non-related, GH-families are observed among closely related microorganisms does indicate a high occurrence of HGT. This transfer does not have to be restricted to the complete cellulase gene (which harbours multiple domains), but might be domain-specific (partial HGT). Partial HGT is also suggested by the high variation in catalytic domain- and CBM-combinations (see Figure 2). Furthermore, Danchin *et al.* observed that the similarity between the CBMs of different proteins was higher than the similarity between the catalytic domains of the same proteins (95), also indicating a distinct evolutionary history for both domains.

Aims and hypotheses of this thesis

Thus, the exact contribution of soil microorganisms and their diversity to the degradation of cellulose and the soil carbon cycle is not clear yet. Therefore, the primary aim of this research was to explore the genetic potential for cellulose degradation in agricultural soil and how this is affected by different tillage treatments. More specifically, it was investigated if the effects of reduced tillage on soil carbon stock are mediated by a change in the microbial cellulose-degrading potential. Results were expected to elucidate possible biomarkers for soil fertility and suitability for crop production. The second aim was to assess the phylogenetic diversity of the GH5-coding gene in agricultural soil, to get insight into its evolutionary history. This would increase the understanding of the redundancy of the cellulose-degrading trait and help in the identification of strategies to manage soil fertility and counteract further soil degradation.

To investigate the effect of different tillage intensities on the soil microbial potential to degrade cellulose, a metagenomic shotgun-sequencing approach was implemented. This method allows the capture of all the genetic information in the soil, avoiding cultivation- or marker gene-biases invoked by growing soil microorganisms on different media (96) or by amplifications. With shotgun-metagenomics, genetic information from organisms of all domains of life can be obtained, allowing for a holistic assessment of the soil ecology. Soils from two field experiments located in the temperate climate region in Western Europe were used. To understand the lasting effects of tillage on the soil microbiome, first a long-term field experiment, on which reduced (RT) and conventional

tillage (CT) was applied for twenty subsequent years, was chosen for investigation. Samples were taken from the surface soil layer, up to the depth of the reduced tillage-working, to target the soil layer which is under influence of tillage. As new sequencing technologies with higher capacity became available, a second field experiment was chosen for analysis. Analysis of another field experiment would provide a more robust assessment of tillage effects on the soil microbiome, especially when sampling was performed in another season and at a different time point after tillage treatment. Stronger statistics and co-occurrence analyses would be able to be applied using the higher amount of data generated with the new sequencing technique. As the long-term field experiment was managed using conventional farming practices (hereafter designated as the conventional farming experiment), the second field experiment was chosen based on an organic farming management-regime (hereafter designated as the organic farming experiment). A higher contribution of especially fungi to the soil microbial community composition was expected to be found in the organic field experiment (97–99). To investigate the differences in the soil layer beneath the reduced-tillage depth, the soil of the organic field experiment was sampled both from the surface soil layer and the soil layer which was below the working depth of reduced tillage but within the working depth of conventional tillage (hereafter designated as the deeper soil layer or horizon). Because the soil layer below the reduced tillage-working depth can be considered a no-tillage zone, it was expected that differences between tillage treatments are more pronounced in the deeper soil layer than in the surface soil layer, which is the tilled horizon. Ultimately, the analysis of the soil of these two field experiments and the outstanding sequencing effort has yielded a great amount of sequence data which can be related to the recorded soil- and climatic metadata. The obtained metagenomes allowed for a robust assessment of the effects of tillage intensity on the soil genetic potential. Results from both metagenome studies are shown and discussed from both a general perspective and focussing on cellulase-encoding genes (from here on designated as cellulase genes). The main hypothesis was that soil under reduced tillage contains a higher diversity and relative abundance of cellulose-degrading microorganisms and cellulase genes than soil under conventional tillage, as a response to a higher organic carbon input and content.

To be able to quantitatively assess the abundance of cellulase genes in soil under different tillage treatments, a cellulase gene-amplification assay was developed. The focus was placed on capturing the *in situ* cellulase diversity, independent from database sequence biases, which led to the choice of using a metagenome-read as a basis for primer design. A primer set was developed for GH5-family cellulase genes with the final goal of application in real-time quantitative PCRs. The GH5-family includes enzymes with twenty different catalytic functions, of which the endoglucanases, cellobiosidases and cellodextrinases comprise the ones with cellulolytic functions (CAZy database, www.cazy.org, (100)). It is one of the most abundant endoglucanase-families found in soil (101–103). Furthermore, the GH5-endoglucanases are highly phylogenetically divergent, reflected by its broad distribution among the kingdoms of life, notably Archaea, bacteria, fungi and plants (CAZy database, www.cazy.org, (100)) and by its low sequence conservation, with only seven conserved residues (104). Therefore, a metagenome read from the conventional farming experiment, annotated with high reliability to this catalytic domain family, was chosen to capture cellulase sequences with a high phylogenetic variability. The results of the primer development process are reported here and recommendations for further development are given. The applied amplification and high-throughput-sequencing method has yielded a large amount of highly diverse partial GH5-genes, which were used for phylogenetic analysis. Here, the diversity of a specific group of GH5-

genes in agricultural soil and its phylogenetic range is presented. It was expected that the phylogenetic diversity of cellulase genes in soil exceeds the known diversity of cellulase-database sequences and that several indications of HGT would be identified, suggesting a high (possibly shared) genetic potential to adjust to the substrate-type presented as carbon source. The evolutionary implications of these results in an ecological context are discussed.

2 - Materials and Methods

This chapter describes the materials and methods used to obtain the presented data. Contributions of others to the data generation is either explicitly stated in the corresponding section or mentioned in the section 'Contributors to the data presented'. The majority of the conventional farming experiment-data has been published in 2015 in FEMS Microbiology Ecology (101). Here, an adjusted description of the materials and methods is given.

Site description and soil sampling

The conventional farming experiment

The conventional farming experiment was part of an agricultural field experiment in Scheyern (ca. 460m above sea level), 40km north of Munich (Germany), established in 1992 (24, 105). Soil sampling was done with help of Anne Schöler and Julien Ollivier, both working at the department of Environmental Genomics at the Helmholtz Zentrum München. Local precipitation and temperature were respectively 792.3 mm and 8.3 °C (average over 2000-2010). The soil is classified as a Luvisol, has a pH ($CaCl_2$) of 6.3 and consists of 2.2% coarse sand, 17.0% fine sand, 55.4% silt and 25.4% clay (silty clay loam). A crop rotation including winter wheat-potato-winter wheat-maize was applied on this parcel. The experimental design is given in Figure 3, showing the layout of plots treated with two experimental factors: tillage intensity and nitrogen fertilization amount.

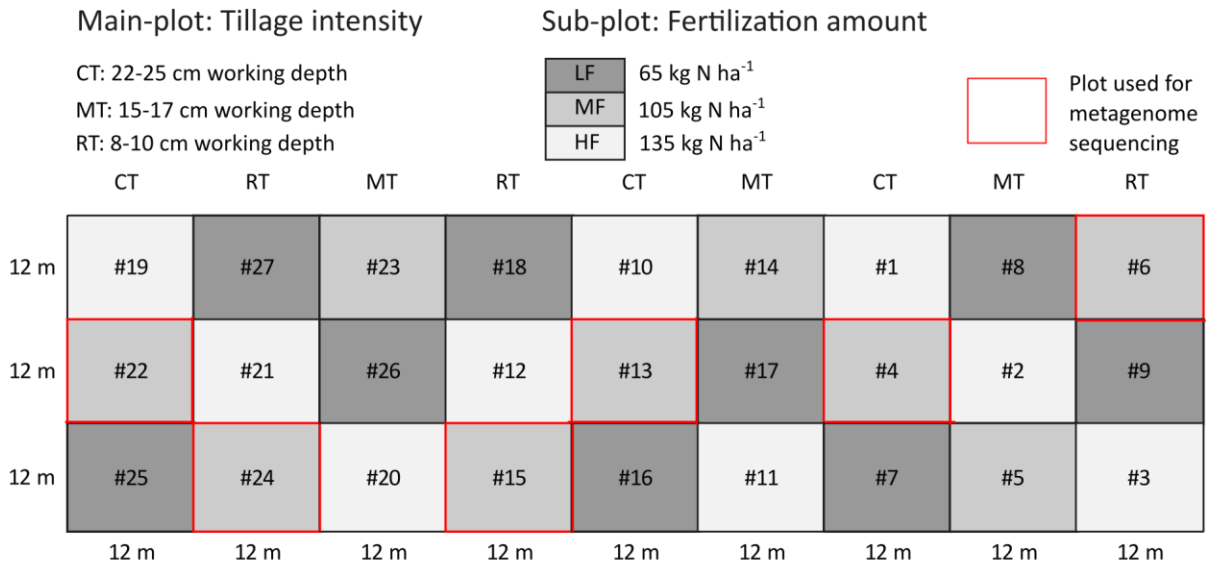


Figure 3: Experimental design of the conventional farming experiment. The sampling site consisted of 27 square plots, each 12 m by 12 m, organized in a split-plot design. As main treatment, three different tillage intensities were applied (RT: 8-10 cm, MT: 15-17 cm and CT: 22-25 cm working depth). As sub-plot treatment, three concentrations of fertilizer were applied (LF: 65 kg N ha^{-1}, MF: 105 kg N ha^{-1} and HF: 135 kg N ha^{-1}) which respectively corresponded to 33% less than, same as, or 33% more than conventional fertilization practice. Soil was sampled from all plots and the soil metagenome was sequenced from the plots with a red border.

Three different levels of tillage intensity (reduced tillage (RT, tilled with rotary harrow with 8 cm working depth), medium tillage (MT, tilled with cultivator with 15 cm working depth) and conventional tillage (CT, tilled with mouldboard plough with 25 cm working depth) and three levels

of fertilization amount (low fertilization (LF), medium fertilization (MF) and high fertilization (HF)) were applied to plots arranged according to a split-plot design, with tillage intensity as main plot and fertilization amount as sub-plot. Sampling was done in November 2012, one month after corn harvest, when the maize plant residues were incorporated into the soil by tillage and winter wheat was sown. At the time, soil temperature was 6.7 °C. A composite sample of five soil cores (0-10 cm, the surface soil layer) was taken of every plot with a soil auger of 5 cm diameter. The composite sample was sieved through a 3-mm sieve and subsequently mixed. A subset was immediately stored on dry ice and afterwards at -80 °C for subsequent DNA extraction. The leftover soil was kept at 4 °C for chemical and biological analysis.

The organic farming experiment

Soil was sampled from an agricultural field experiment, established at the Swiss federal agricultural research station Agroscope (ca. 410m above sea level), Reckenholz station near Zurich (Switzerland), called "Farming System and Tillage experiment (FAST)-II". Sampling was done with help of Raphael Wittwer from the Agroscope Reckenholz research station. Ten-year-average precipitation and temperature are 976 mm respectively 10°C (2009-2015, (106)). The soil is a calcareous Cambisol with a pH (H_2O) of 7.9 (0-16 cm), consisting of 43% sand, 33% silt and 24% clay (clay loam soil). The experimental design is given in Figure 4, showing the layout of plots treated with three experimental factors: farming system (organic or conventional), tillage intensity (reduced tillage under organic farming (RT, rotary harrow with 5 cm working depth), no-tillage under conventional farming (NT), intensive tillage under organic farming (CT, mouldboard plough with 20 cm working depth) and intensive tillage under conventional farming (PT, mouldboard plough with 20 cm working depth)) and cover crop (legume (L), non-legume (NL), a mixture (M) or none (NO)). These treatments were applied to plots in quadruplicates, arranged according to a split-plot design, with production system (farming system*tillage intensity) as main plot and cover crop as sub-plot. This experiment has been established in 2010 and will continue until at least 2020. Additional details of previous crops, yields and soil characteristics can be found in a recent publication by Wittwer *et al*. (106). From this experiment, soil samples were taken of the plots with legume cover crop and no cover crop (blue-squared plots in Figure 4) in August 2014, one week after harvest of winter wheat (wheat straw was removed from the field) and short stubble mulching. Five months before sampling, weed-control had taken place, consisting of shallow harrowing on the RT- and CT-plots and herbicide application on the NT- and PT-plots. At the time, soil temperature was 22 °C. Of every plot a composite sample of six cores was taken of soil on a depth of 0-6 cm (the surface soil layer) and 10-16 cm (the deeper soil layer) under field level, yielding a total of 64 samples. Soil was taken using a soil auger of 2 cm diameter and the composite sample was mixed on field and stored immediately at 4°C. Afterwards, the soil was sieved with a 2-mm sieve and a subset was shot-frozen in liquid nitrogen and subsequently transported on dry ice and stored at -80°C before DNA extraction. The remaining soil was stored at 4°C, transported in cooling boxes and afterwards stored at -20°C for subsequent chemical and biological analysis.

Soil chemical and biological analysis

The pH of the soil samples was measured using a glass electrode (WTW Inolab® pH 720) submerged in a prepared soil solution. The solution was prepared by adding 25 mL of 0.01M $CaCl_2$ or demi-H_2O to 5.0 g of fresh soil, shaking the mixture overhead for 5 min and incubating it at room temperature for 20 h.

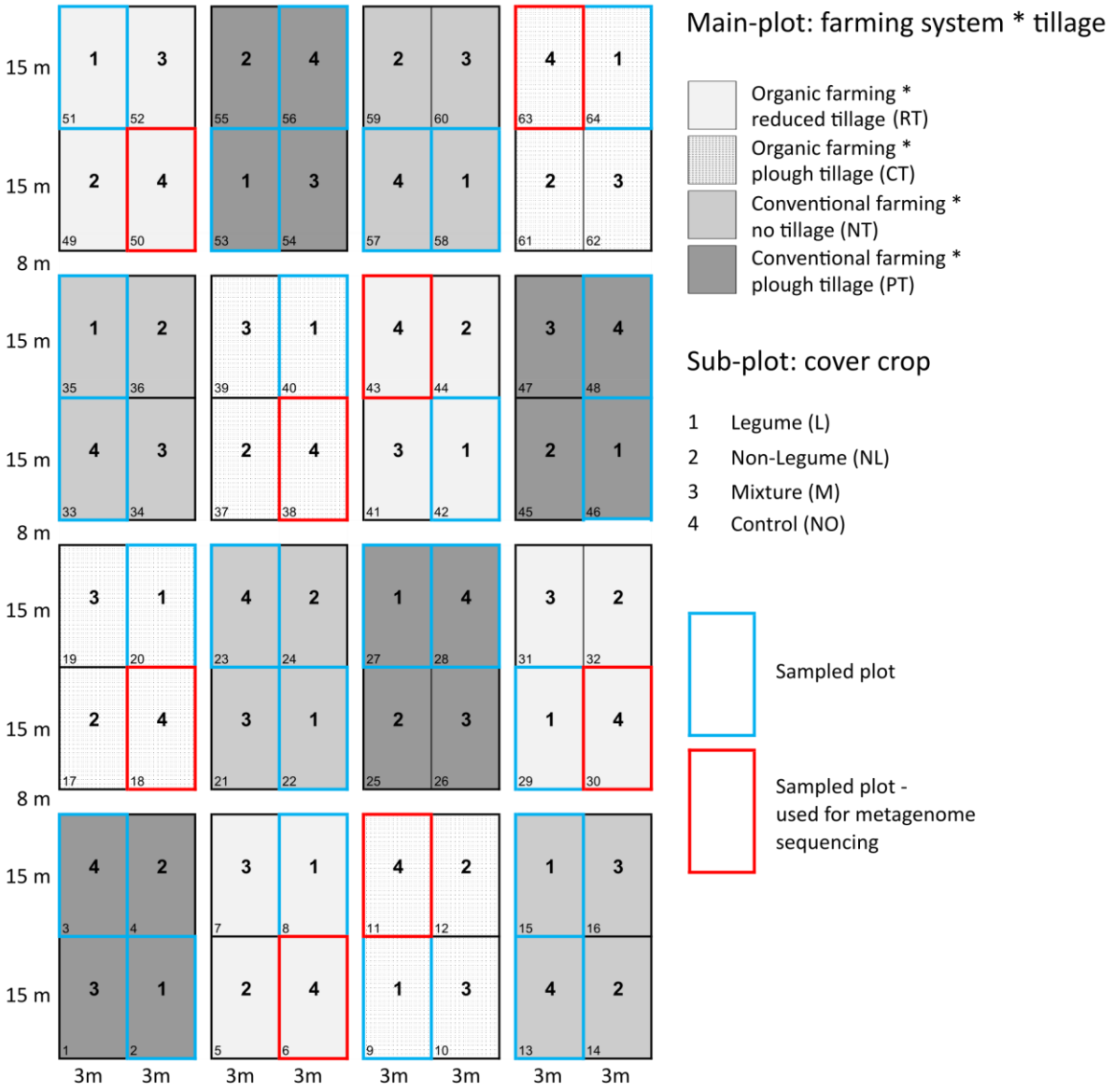


Figure 4: Experimental design of the organic farming experiment. The sampling site consisted of 64 plots, each 15 m by 3 m, organized in a split-plot design. In the main plot, four different management systems were randomly applied; RT: organic farming with reduced tillage (5 cm working depth), CT: organic farming with plough tillage (20 cm working depth), NT: conventional farming without tillage and PT: conventional farming with plough tillage (20 cm working depth). As sub-plot treatment, four different cover crop treatments were applied; L: leguminous crops, NL: non-leguminous crops, M: mixture of legume and non-leguminous crops and NO: no cover crop (control). Soil was sampled from all plots with a blue and red border and the soil metagenome was sequenced from the plots with a red border.

For microbial biomass measurement, the water extractable organic carbon (WEOC) was determined before and after chloroform fumigation according to Joergensen, 1996 (107). In short, triplicates of 5.0 g of fresh soil in glass vials were incubated in chloroform gas for 24 h. To extract the WEOC, 5.0 g of fresh soil was mixed with 20 mL of 0.01M $CaCl_2$, incubated in an overhead shaker for 40 min at room temperature, filtered using a pre-rinsed Whatman-filter, and stored at -20 °C until measurement with the Dimatoc 100 (DimatecAnalysentechnik GmbH, Germany). Soil humidity was

measured by drying soil at 40°C for ca. 5 days or at 106°C for 36 h. Total carbon (TC) and nitrogen (TN) was determined grinding dry soil to a homogeneous powder by shaking it in a Tissue lyzer at 30 Hz for 3-9 min together with steel balls. Duplicates of 20-25 mg of ground soil powder were subjected to elemental analysis using an Elementar Vario EL III instrument in combustion mode. Additionally, for calcareous soil (the organic farming experiment), 2.0-2.5 mg of ground soil powder treated with 10-20 µL of 2M HCl was analysed to correct for possible carbonate content.

Enzymatic activity assays

For the conventional farming experiment the potential activity of cellulase- and xylosidase-enzymes was measured. The soil stored at 4 °C was assessed for potential enzymatic activity one day after sampling. This was done by adding methylumbeliferone (MU)-complexed substrates to soil solutions (108). The three substrates used were MU-β-1,4-glucopyranoside (targeting β-glucosidase (EC:3.2.1.21)), MU-β-cellobiohydrofurane (targeting cellobiosidase (cellobiohydrolase, EC:3.2.1.91)) and MU-β-D-xylopyranoside (targeting β-D-xylosidase) (Sigma-Aldrich). First, pilot experiments were performed to determine the optimum substrate concentration (C_{opt}) and incubation time (T_{opt}) for degradation of each substrate: C_{opt} was 800 µM of MU-β-cellobiohydrofurane and 500 µM of MU-β-D-glucopyranoside and MU-β-D-xylopyranoside, T_{opt} was 120 min for each substrate. Subsequently, a soil solution of each sample was incubated with each substrate for 120 min in triplicate in a black 96-well plate with transparent underside. Subsequently, they measured for MU-fluorescence, using a spectrophotometer with excitation wavelength λ=365 nm and emission wavelength of λ=450 nm. The obtained fluorescence values were corrected for inhibition by soil particles (quenching) and background noise, as well as for soil-autofluorescence. Finally, they were converted to MU-concentrations using a calibration line calculated from reference measurements on the same plate. The final fluorescence values in pmol MU were calculated per hour and per gram dry weight of soil. To compensate for microbial biomass, additional fluorescence values were calculated per µg microbial carbon, by dividing the fluorescence values per hour and gram soil with the microbial carbon-values per gram dry weight of soil.

DNA-extraction, library preparation and sequencing

For the conventional farming experiment (section 3) and the GH5-amplicon study (section 5), DNA was extracted of the soil samples taken from the RT/MF and CT/MF-treatment (6 samples). For the organic farming experiment (section 4), DNA was extracted of the soil samples taken from the RT/NO- and CT/NO-treatments and two soil depths, the surface soil layer and the deeper soil layer (16 samples, red-squared plots in Figure 4). DNA for the surface soil layer-samples of the organic farming experiment was extracted by Irina Tanuwidjaja, working at the department of Environmental Genomics at the Helmholtz Zentrum München. Around 300 mg of frozen soil was taken and DNA was extracted using the NucleoSpin Soil DNA-isolation kit 'Genomic DNA from soil' (Macherey– Nagel, Germany), buffer 1, according to the protocol of the manufacturer. For the organic farming experiment, a negative control was included for each extraction run (one for each soil depth). Extracted DNA was stored at -20°C until further processing.

The 6 sample-library of the conventional farming experiment was prepared for shotgun sequencing according to the Roche protocol "Rapid Library Preparation Method Manual", employing Roche Molecular Identifier (MID)-adapters as barcodes. The Roche protocol "emPCR Method Manual - Lib-L Large Volume (LV)" was followed for DNA-amplification in an emulsion PCR. Finally, the amplified libraries were sequenced on the Genome Sequencing (GS) FLX+ instrument, using a GS

FLX Titanium sequencing kit XL+. Pyrosequencing signals and images were processed by the supplied Roche software and the raw sequences have been deposited in the NCBI-Sequence Read Archive (SRA, BioProject ID: PRJNA235154).

The 18 samples of the organic farming experiment (including negative controls) were all prepared for shotgun sequencing on the MiSeq® sequencing machine (Illumina) using the exact same procedure, but the samples of the surface soil layer were prepared by Anne Schöler. To obtain an average sequence length of 300 base pairs (bp), 60 µL sample-DNA was sheared with an E220 Focused-ultrasonicator (Covaris® Inc., Woburn, Massachusetts, USA) for 80 s using the following settings: Peak Incident Power= 175W, Duty Factor = 10%, 200 cycles per burst, 7°C, water level 6 and with Intensifier. Afterwards, library preparation was performed according to the "NEBNext® Ultra DNA Library Prep Kit for Illumina (for Metagenome)"-manual provided by Illumina, with the following adaptations: a library size selection of 400-500 bp was done, for PCR amplification 10 cycles were used, the NEBNext High Fidelity 2x PCR Master Mix was used instead of the Q5-mix, 2.5 µL (instead of 1 µL) of primers was used, and 20 µL (instead of 25 µL) of DNA input was used. The samples were indexed using 1:5 diluted NEBNext® Multiplex Oligos for Illumina (Dual Index) (for Metagenome), provided by Illumina. Finally, 4 nM DNA-aliquots of each sample were pooled and sequenced in two runs (one run for all samples of each depth layer). Sequencing was performed with the MiSeq® Reagent Kit v3 (600 cycles) for paired end- sequencing on a MiSeq® sequencing machine.

GH5-primer design, amplification and sequencing

For the development of the GH5-cellulase primer set, metagenome read H3PB9EL02JLGQQ (top BLASTx-hit: endoglucanase [Gynuella sunshinyii YC6258], max score: 70.5, total score: 70.5, query coverage: 97%, expect (e)-value: $4e^{-13}$, identity: 79%, accession: AJQ95033.1 (October 2013)) from the conventional farming experiment dataset was aligned to as many nucleotide sequences as possible of GH5-endoglucanases found in the CAZy database (www.cazy.org, (60)), in order to determine suitable primer-binding regions shared by all sequences. However, the amount of database sequences was reduced to only two which were suitable for primer-binding region identification: an endoglucanase (cel5B, ACE84076) from *Cellvibrio japonicus* strain Ueda107 and a bifunctional endoglucanase/cellobiohydrolase (celAB, ABS72374) from *Teredinibacter turnerae* strain T7901. For alignment of these templates and determination of suitable primer sequences, the Clustal Omega service on the EMBL-EBI-website (https://www.ebi.ac.uk/Tools/msa/clustalo/) with default parameters was used (109). Ultimately, a forward and reverse primer sequence with degeneracies of 8 respectively 96 were obtained. The primer sequences and the binding regions on the catalytic GH5-domain of ACE84076, as identified by a conserved domain database-search (110), are shown in Figure 31. To test primer specificity, an *in-silico* PCR was done with the De-MetaST-software (111), allowing no mismatches and using the metagenome from the conventional farming experiment and the National Center for Biotechnology Information (NCBI, 2013) nucleotide database as template sequences. A second *in-silico* PCR was done with the Genomatix software suite using the FastM and ModelInspector tool (112), allowing for 1 mismatch and using the Genbank bacterial and plant database as template sequences. Furthermore, PCR-conditions were optimized using temperature gradients and different PCR-mix components. In addition, primer specificity was confirmed by amplification of the GH5-cellulase gene of the obtained reference strain of *C. japonicus* (DSMZ, Braunschweig, Germany). Genomic DNA from *C. japonicus* was obtained by DNA-extraction from liquid culture (DSMZ carboxymethyl cellulose medium 1111, pH-adjusted to 7.0, Temperature: 37 °C, 150 rpm) using a modified version of the DNA and RNA co-extraction protocol by Griffiths *et*

al. (113). Amplification was performed using a hot start (94 °C for 7 min), followed by forty cycles with a denaturation phase at 94 °C for 45 s, an annealing phase at 57°C for 45 s and an elongation phase at 68 °C for 30 s, while the final elongation was done at 68 °C for 7 min. The following reagents were used for one reaction volume of 50 µL: 0.5 µL Taq polymerase (5 units μL^{-1}) (Qiagen, Germany), 5 µL PCR buffer (10x, without $MgCl_2$ or $MgSO_4$) (Qiagen, Germany), 5 µL dNTP-mix (2 mM) (Qiagen, Germany), 2 µL $MgCl_2$ (50 mM), 1 µL of forward primer (10 µM) (Metabion, Martinsried, Germany), 1 µL of reverse primer (10 µM) (Metabion, Martinsried, Germany), 0.5 µL of template DNA (50 ng μL^{-1}) and 35 µL of demineralised H2O. Amplification was confirmed by loading of products on a 2%-agarose gel.

The 6 DNA samples from soil of the conventional farming experiment were pooled to sample the complete diversity of the soil and not have differences because of treatments. This pooled sample was used as template for the amplification of GH5-cellulase genes. To increase amplification specificity, the amplification of GH5-genes was performed in two consecutive PCR runs; a first amplification round using the target-specific primers (see Figure 31-B) and a second amplification using the same primers with an Illumina Nextera V2 adapter overhang attached (see Figure 31-C). The following reagents were used in the first PCR run (for one reaction volume of 25 µL): 12.5 µL NebNext Polymerase (2x) (New England Biolabs GmbH, Frankfurt, Germany), 0.75 µL of forward primer (10 µM) (Metabion, Martinsried, Germany), 0.75 µL of reverse primer (10 µM) (Metabion, Martinsried, Germany), 0.5 µL of 3% bovine serum albumin, 2 µL of template DNA (100 ng μL^{-1}) and 8.5 µL of demineralised H2O. The PCR run had a first denaturation phase at 98 °C for 5 min, then thirty cycles with a denaturation phase at 98 °C for 10 s, an annealing phase at 57°C for 30 s and an elongation phase at 68 °C for 30 s. The final elongation was done at 68 °C for 5 min. The products were excised from a 1%-agarose gel and purified using the NucleoSpin gel- and PCR-clean-up kit (Macherey-Nagel, Germany). 2 µL of the purified PCR products from the first PCR were used as template for the second PCR. The reagent mix and PCR conditions were identical to the first PCR, but instead of thirty cycles only ten cycles were run. The products were excised from a 1%-agarose gel and purified using the NucleoSpin gel- and PCR-clean-up kit (Macherey-Nagel, Germany). Afterwards, library preparation was performed according to the Illumina "16S Metagenomic Sequencing Library preparation" manual, starting from the Indexing PCR step onwards and using the NebNext 2x High Fidelity Master Mix instead of the Kappa HiFi Hotstart Ready Mix, which is the default mix mentioned in the manual. For sample indexing the Nextera® XT Index Kit Set C for Amplicons (Illumina) was used. Sequencing was performed with the MiSeq® Reagent Kit v3 (600 cycles) for paired end- sequencing on a MiSeq® sequencing machine.

Quantitative real-time PCR assay

For quantification of bacterial and fungal abundances in the soil of both metagenome studies, the 16S ribosomal RNA-genes (16S rRNA genes) and the fungal Internal Transcribed Spacer (ITS) regions were amplified during quantitative real-time PCR (qPCR). For the conventional farming experiment, the qPCR was carried out together with Anne Schöler. For 16S rRNA gene quantification, the 25 µL reaction mixtures comprised: 12.5 µL Power SYBR Green master mix (Life Technologies), 5 pmol of each primer (114), 0.5 µL 3% bovine serum albumin and 2 µL soil DNA template. For ITS rRNA gene quantification, the 25 µL reaction mixtures comprised: 12.5 µL Power SYBR Green master mix (Life Technologies), 10 pmol of each primer (115) and 2 µL soil DNA template. Standard curves were composed from the amplification of serial dilutions (10^1 to 10^7 gene copies mL^{-1}) of a plasmid containing *Fusarium oxysporum*-DNA (the conventional farming experiment)/ *Trichoderma reesei*-

DNA (the organic farming experiment) (for fungal qPCR) or *Pseudomonas putida*-DNA (for bacterial qPCR). The qPCR detection limit was defined at 10 gene copies mL^{-1}. For each amplification assay, several dilutions of the soil DNA were employed to detect PCR inhibition by high DNA-concentrations. All qPCR assays were carried out in 96-well plates. PCR conditions were: 95˚C for 10 minutes, then 40 cycles of 30 s at 94˚C, 30 s at 55˚C and 30 s at 72˚C for ITS rRNA gene quantification, or 40 cycles of 45 s at 95°C, 45 s at 58°C and 45 s at 72°C for 16S rRNA genes quantification. In the end, final elongation and melting curve analysis was performed for all samples by a final cycle of 15 s at 95˚C, 30 s at 60˚C and 15 s at 95˚C. The amplification efficiency was calculated from the formula Eff = $[10^{(-1/slope)} -1]*100$. The efficiency of the qPCR on samples from the conventional farming experiment was for 16S rRNA genes: 102% and for the ITS rRNA gene: 87%, and on samples from the organic farming experiment for 16S rRNA genes: 94% and for the ITS rRNA gene: 92%.

Analysis of metagenome data and prediction of cellulase genes

Most of the bioinformatics analyses described were performed using the program GNU Parallel (116). For the analysis of the conventional farming experiment metagenome, the provided Roche SFF-Tools were used to separate the obtained reads according to the MID adapters. Read trimming was performed with DynamicTrim (117), as supplied by MG-RAST (118), with the following settings: minimum Phred score= 15, maximum number of bases below minimum Phred score=5, minimum read length=50. Any remaining adapter sequences and duplicate reads were removed using Biopieces (www.biopieces.org) and cd-hit (119), respectively. To assign taxonomy and function to the clean reads, they were compared to the NCBI (2013) non-redundant protein database and the Kyoto Encyclopedia of Genes and Genomes (KEGG) Orthology-database (June 2011) using Diamond (120) using a minimum score cut-off of 50. In addition, they were compared to the SILVA SSU-database (September 2013) using BLASTn (e-value threshold= 10^{-4}, (121, 122)). Evaluation of the top BLAST hits (i.e. up to 25 hits with lowest e-value) of each read for taxonomic and functional mapping was done using the MetaGenome ANalyzer (MEGAN, Version 5.2.3) software (http://ab.inf.uni-tuebingen.de/software/megan5/ (123)), in the context of taxonomic hierarchy respectively KEGG-pathways. The following settings applied in MEGAN: Min support=1, Min score=50, Top%=10, Min-Complexity Filter= 0, LCA-percent=100. Further data analysis was performed in R (124) or MS Excel®. General taxonomic and functional annotation of this metagenome was performed by Anne Schöler.

To find genes encoding cellulase domain families (GH-, AA- and CBM-genes), protein open-reading frames (ORFs) were first predicted from the metagenome reads using FragGeneScan (125) and then translated to protein sequences. A set of CBMs and catalytic domain families which are found in endoglucanases, exoglucanases, β-glucosidases, cellodextrin phosphorylases and cellobiose dehydrogenases, were selected from the Carbohydrate Active Enzymes (CAZy)-database (www.cazy.org, (60)). To identify these domain families in the metagenome sequences, they were scanned against a database of protein profile-Hidden Markov Models (HMMs) using hmmscan (126). This database had been assembled in the following manner; First, self-built HMMs were created for each of the selected domain families from alignments of sequences containing the target protein domain which were used as input for hmmbuild (which is contained within the HMMER version 3.0 (March 2010), packaged together with hmmscan; http://hmmer.org/). The alignment files were generated using sequences from the CAZy database for each corresponding family. Then, these self-built HMMs were tested for their sensitivity and specificity to detect cellulases by comparison with the available HMMs in the Protein family (Pfam)-A 26.0 database (127) and with those from the

Database for automated Carbohydrate-active enzyme Annotation (dbCAN, (128)). This was done by scanning the HMMs against a set of "positive" or "negative" cellulase sequences, which were compiled by downloading the protein sequences of all CAZy entries (January 2015) for GH-, CBM- and AA-families that contain cellulolytic proteins, from the NCBI-database. Of these, 1283 entries were labelled "positive", having a matching EC number (EC: 3.2.1.4, EC: 3.2.1.74, EC: 3.2.1.176, EC: 3.2.1.91, EC: 3.2.1.21, EC: 1.1.99.18, EC: 2.1.4.49, EC: 2.1.4.20 and EC: 2.4.1.321), whereas 1390 entries were labelled "negative" as the EC number did not match. As no EC number was available for the functionally characterized proteins of domain families AA8 and AA9, descriptions which mentioned activity on cellulose were regarded as "positive". Thus, after scanning these sets of cellulase and non-cellulase sequences with the self-built HMMs, those from the Pfam A-database and those from the dbCAN, the best-performing HMM for each domain family was selected for further analysis. The selection criteria included highest sensitivity (percentage of positive sequences annotated, see shaded cells in Table A1 in the appendix) and, if possible, lowest false-positive prediction. The selected HMMs were afterwards applied to scan the translated protein sequences from the metagenomes (hmmscan e-value threshold: 10^{-4}) for cellulase domain families. To further limit the number of false-positive predictions, an additional quality control step was performed. All hmmscan-annotated sequences were BLASTed against the above-mentioned positive cellulase sequences-database using BLASTp (BLASTing e-value threshold= 10^{-5}). Finally, these identified cellulases were taxonomically assigned by BLASTing their corresponding nucleotide sequences against the NCBI non-redundant protein database using Diamond (minscore=50) and mapping their taxonomy using MEGAN5 (parameters as described above).

For the analysis of the organic farming experiment metagenome, the obtained paired reads remaining adaptor sequences were removed, reads were trimmed and merged using AdapterRemoval v2 (129), using the following settings: -q 15 -l 50. Afterwards, contaminant sequences (PhiX Illumina control sequence) were removed using DeconSeq version 0.4.3 (130). The high-quality merged reads (more than 95% of the read-pairs were merged) were used for downstream analysis. Further general taxonomic and functional annotation and mapping using MEGAN was done as described for the conventional farming experiment, but with updated NCBI (January 2015) and SILVA database (August 2015). Likewise, the cellulase domain family prediction using hmmscan and further filtering using BLASTp was done as described for the conventional farming experiment. Furthermore, an overview of the tillage effects observed on the microbial community members and the cellulase enzymatic groups or domain families in the metagenomic dataset was created by plotting them as connected nodes in Cytoscape 3.6.0 (131) using the Prefuce Force Directed Layout and subsequent manual adjustment of the layout. Most edge and node settings (e.g. colour and size) were customized according to tillage effects and designations as cellulose "degraders" or "utilisers".

To discover positive relationships between the relative abundances of annotated microbial families, a co-occurrence analysis was performed by David Endesfelder, working at the Scientific Computing department of the Helmholtz Zentrum München at the time of cooperation. In addition, the co-occurrence between microbial families and protein-coding genes was assessed. Different abundance cut-offs were applied; For the co-occurrence between microbial families, only those with a minimum abundance of 50 reads in at least 3 of the 16 samples were considered, whereas for the co-occurrence between microbial families and protein-coding genes an abundance cut-off for an observed connection was applied: only connections with a minimum occurrence of 5 times within 1

sample and in at least 5 samples were considered. Furthermore, only significant ($P<0.05$) correlations of abundances between any pair of microbial families or between a microbial family and a protein-coding gene, with a correlation coefficient of >0.6, was considered a co-occurrence. Correlation significance from Spearman's rank correlation coefficients was estimated using the CCREPE (132) method, to eliminate possible spurious correlations caused by the compositional structure of the relative abundances. Setting an edge between pairs of co-occurring entities led to the formation of the corresponding co-occurrence network. Subsequently, clusters of co-occurring microbial families were defined from the network by grouping families with high intra-cluster and low inter-cluster connectivity. This was done using the Markov Dynamics clustering algorithm (133), executed in MATLAB®. In short, this algorithm identifies clusters of communities within a continuous range of the Markovtime-parameter. Then, the number of community clusters with the longest stable assignment over a range of Markov time points is chosen as the number of communities separated in the network.

Statistical analysis of metagenome sequencing data

To be able to compare the annotation counts of the metagenomic data samples, relative abundances per sample were calculated. To obtain the relative abundance of annotations, the number of annotated reads within one sample was divided by the initial number of clean reads of that sample and multiplied by 100,000. This resulted in relative abundances given in % of clean reads $x10^{-3}$. Only taxa and functions with a minimum total abundance (summed up over all samples) of at least 0.01% of all metagenome reads ($10x10^{-3}$ %) were considered relevant for discussion and are shown in the figures. An exception was made, however, for the annotation of cellulase genes and their taxonomic assignments. Statistical analyses were performed in R (124) or MS Excel®. To detect overall differences between the treatments, a principal component analysis was performed (R-function prcomp) on the scaled and centred relative abundance-data of metagenomic reads annotated on order (the conventional farming experiment) or family (the organic farming experiment) level, after rarefying (using the Rarefy-function in the R-package vegan (134)) and log-transformation. Furthermore, Shannon indices were calculated (using the diversity-function in the R-package vegan (134)) from the relative abundances of metagenomic reads annotated on order (the conventional farming experiment) or family (the organic farming experiment) level, also after rarefying. Finally, the coverage of the total known taxonomical (NCBI-referenced) diversity was calculated by subsampling from the relative abundances of metagenomic reads annotated on order (the conventional farming experiment) or family (the organic farming experiment) level using the rarec-function in the R-package vegan (134). In addition, a database-independent coverage of the total sequence diversity was estimated by uploading clean reads in FastA-(the conventional farming experiment) or FastQ-(the organic farming experiment) files to the Nonpareil online tool (http://enve-omics.ce.gatech.edu/nonpareil/submit (135)) with the following parameters: Overlap: 25%, Identity: 95%, Single-stranded - not checked, N as mismatch - checked, Random seed: 12, Query set size: 1000.

Parametric statistics on soil characteristics between treatments were performed after testing for normal distribution of the (modelled) data using QQ-plots and Shapiro-Wilk tests (R-functions qqnorm respectively shapiro.test). Differences in soil characteristics between treatments were determined by analysis of variance (ANOVA), taking into account the split-plot design of the conventional farming experiment (main-plot: tillage intensity, sub-plot: fertilization amount) and the split-split-plot relationship of the samples taken from the field of the organic farming experiment

(main-plot: production system, sub-plot: cover crop, sub-sub-plot: soil depth layer).The following R-functions were used for the ANOVA-model calculations: aov-function, sp.plot-function of the R-package agricolae and aovp-function of the R-package lmPerm (when model residuals were not normally distributed). To determine which levels of treatments were different, the post-hoc least-significant-difference (LSD)-test was performed (LSD.test-function of the R-package agricolae). For metagenomic data of the conventional farming experiment, differences in relative abundances of taxonomic or functional assignments between CT and RT samples were determined by two-tailed paired t-test statistics. For metagenomic data of the organic farming experiment, differences between CT-, RT- and the depth-samples were determined by ANOVA, handling the data according to a split-plot design (main-plot: tillage treatment, sub-plot: depth layer). In addition, the LSD-test was performed after the ANOVA to determine which levels of treatments differed. Differences were regarded significant when the *P*-value was smaller than 0.05. In addition, a multiple-testing correction was performed in the following manner: first, the number of variables was calculated which would show a significant difference by chance (false-positives). Then, all variables showing significant differences were ordered based on *P*-value. All variables, starting from the variable with the highest *P*-value and continuing until the predicted number of false positives was reached, were considered falsely significant. The data and results of the statistical analysis (relative abundances per treatment, total relative abundance, *P*-values of the used statistical test) are summarized in Tables A3 and A6 in the appendix. Due to the low amount of predicted cellulases and their taxonomic assignments, no abundance cut-off was used to detect differences between treatments. However, only the taxonomic assignments of the most abundant cellulase genes were considered for calculation of differences between tillage treatments.

Amplicon data analysis

From the obtained reads the remaining adaptor sequences were removed using Biopieces (www.biopieces.org) and afterwards, the read pairs were merged using FLASH v1.2.11 (136). Merging was successful for more than 90% of the read pairs. For downstream analysis only the merged reads were used. The majority of the obtained merged reads had a relatively high sequencing quality (above phred-score 25) and was between 40 and 120 bp long (see Figure 5-A). They were further filtered for quality (minimum average quality of 35 in a sliding window of 45 bases) and for a minimum length of 80 and a maximum length of 185 bp using Biopieces, to obtain clean reads of the expected amplicon length and with a consistent high quality (see Figure 5-B). Subsequently, the clean reads were translated to amino acid sequences using FragGeneScan (125) and were subjected to pairwise alignment with ssearch36 (137) and Markov Clustering (138), together with protein sequences from the CAZy database (see Table A2 in the appendix). These database sequences consisted of only functionally characterized cellulase (EC: 3.2.1.4, EC: 3.2.1.74, EC: 3.2.1.176, EC: 3.2.1.91 or EC: 3.2.1.21) and non-cellulase sequences (other functional classifications) of GH5. Furthermore, to improve the exclusion of false positives, cellulases and non-cellulase sequences from four other GH families (GH48, GH7, GH74 and GH9) were additionally included. Markov Clustering was done using cut-offs which were based on benchmarking for successful separation of database sequences according to enzymatic function (cellulase vs. non-cellulase functions). The applied cut-offs were further based on calculations of length difference between sequencing reads and database reads: e-value cut-off of 0.01, an ORF length difference-value between 0.06 and 16.3, alignment coverage of 0.1 and an inflation value of 2. Reads that

clustered with GH5-cellulases from the database sequences were considered cellulase amplicon sequences and used for further analysis. Duplicate sequences were removed using cd-hit (119).

Taxonomical assignment of the cellulase amplicon sequences was done by BLASTing the translated cellulase-sequences to the NCBI non-redundant protein database using BLASTp, contained within the BLASTall 2.2.26 package, using an e-value cut-off of 10^{-4}. Mapping the BLAST hits to taxonomical groups was done with the MEtaGenome ANalyzer (MEGAN, Version 5.7.1) software (http://ab.inf.unituebingen.de/software/megan5/), using the following settings: Min support = 1, Min score = 50, Top% = 10, Min-Complexity Filter = 0, LCA-percent=100. A second mapping was performed with the same parameters, but with LCA-percent (lowest common ancestor-assignment cut-off (123)) set to 50.

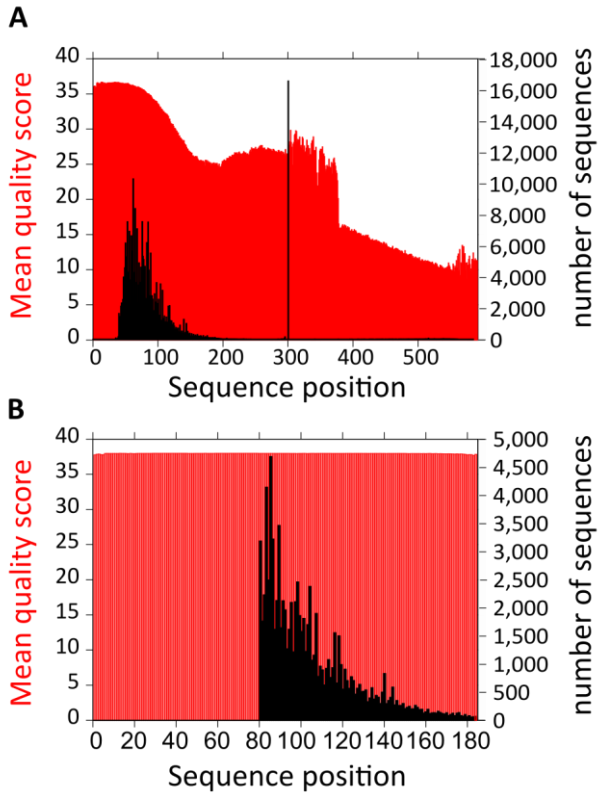


Figure 5: Mean phred-score (red) on each sequence position and amount of sequences (black) with the corresponding sequence length. (A) Raw reads, before sequence quality- and length-trimming, and **(B)** clean reads, after quality- and length-trimming.

To group the cellulase amplicon sequences based on similarity, a second Markov Clustering was performed based on the pairwise similarities of the cellulase amplicon sequences and the GH5-cellulases from the database. Different cut-offs were tested to discover the most meaningful sub-clustering. Finally, the subclustering with the following cut-offs was deemed most meaningful: e-value cut-off of 10^{-11}, ORF length difference-value between 0.06 and 16.3, alignment coverage of 0.1 and an inflation value of 2. The cellulase amplicon sequences clustered according to taxonomical assignment, together with GH5-cellulases from the database and on a high similarity level. Nevertheless, many cellulase amplicon sequences did not cluster with other database- or amplicon sequences, and these were left out of further analysis. For a detailed account of the subclustering result, see Table A10 in the appendix.

Phylogenetic analysis of GH5 protein sequences

GH5-database sequences

Phylogenetic analysis of GH5-database sequences was done using only functionally characterized cellulase sequences (EC: 3.2.1.4, EC: 3.2.1.21, EC: 3.2.1.91 and EC: 3.2.1.74) from the manually curated database, as indicated in Table A2 in the appendix, extracted from the CAZy database (www.cazy.org,, (60)). In addition, two actinobacterial endoglucanase sequences (*Micromonospora*) from the NCBI-database were added because of their similarity to the amplified sequences. These protein sequences were searched for their conserved domains using the batch CD-search tool of the

NCBI (110) and subsequently trimmed using Biopieces to retrieve only the GH5-domain. Finally, all the sequences were aligned with MUSCLE, contained within the MEGA6 software (139), using a Neighbour-Joining clustering method and further default settings. The resulting multiple sequence alignment (MSA) was used for the calculation of two reference-phylogenetic trees of the GH5-database sequences. The first reference tree was calculated using the complete GH5-database sequences and the second reference tree using only the amplicon-region of the GH5-database sequences. To obtain the latter alignment of partial GH5-database sequences, the aligned complete GH5-database sequences were manually cut to leave only the amplification region, including the primer-binding sites. These two alignments were submitted to phylogenetic analysis. The MEGA6-built-in model-selection tool was used to find the most suitable protein substitution model using the maximum likelihood (ML) method and further default parameters. The subsequent phylogenetic ML-tree was estimated using the Whelan And Goldman (WAG) + Frequency (F)-substitution model (140). The initial tree for the heuristic search was obtained by applying the Neighbour-Joining method to a matrix of pairwise distances estimated using a JTT model. In addition, a discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories), and the degree of statistical branching support was obtained by a bootstrap procedure (100 replicates). Only those alignment positions were used for phylogenetic calculation, which were covered by 95% of the sequences. Thus, if fewer than 95% of the sequences contained alignment gaps, missing data or ambiguous residues at a certain position, this position was included in the calculation. After tree-calculation, a GH5-cellulase sequence from Crenarchaeota was selected as outgroup to root the tree.

GH5-amplicon sequences

The “Add unaligned sequence(s) to an existing alignment”-page of the online MAFFT server (MAFFT version 7, http://mafft.cbrc.jp/alignment/server/add_sequences.html, (141)) was used to align the cellulase amplicon sequences to the database sequences. This was done in order to make sure that the alignment of the database sequences would not change. The alignment of partial GH5-database sequences was uploaded as the existing reference alignment and the unaligned cellulase amplicon sequences as the new sequences to align to the existing alignment. The alignment job was submitted with the following parameter settings: --inputorder, --anysymbol, --add new_sequences, --localpair, which uses the L-INS-1 alignment strategy. The output alignment was imported into MEGA6 (139) and submitted for phylogenetic analysis using the same model and settings as described for the GH5-reference trees.

Top BLAST hits of GH5-amplicon sequences

Several cellulase amplicon sequences which are phylogenetically closely related to a partial GH5-database sequence (see section 5: results of phylogenetic analysis of amplified GH5-genes) were blasted against the NCBI non-redundant protein database using BLASTp. The top BLAST hits of these cellulase amplicon sequences were extracted to obtain additional complete GH5-sequences. These were aligned to the existing alignment of complete GH5-database sequences, in the same manner as described above by uploading them to the MAFFT server. The output alignment was imported into MEGA6 (139) and submitted for phylogenetic analysis using the same model and settings as described for the other GH5-reference trees.

3 - Results of the conventional farming experiment

Here, the analyses of the conventional farming experiment samples are presented. Most of this data has been published in 2015 in FEMS Microbiology Ecology (101). For the Ph.D. thesis, the figures and tables have been changed or adapted from the publication and additional data is presented. In short, the surface layer of the soil (0-10 cm) of a field experiment in Scheyern was analysed. The experiment included two treatment factors: three different levels of tillage intensity (reduced tillage (RT), medium tillage (MT) and conventional tillage (CT)) and three levels of fertilization amount (low fertilization (LF), medium fertilization (MF) and high fertilization (HF)), organized in a split-plot-design in three blocks as replicates. For an overview of the experimental setup and more details on the treatments and methodology, see Figure 3 and section 2: Materials and Methods. Soil chemical and biological analyses were done on all treatments.

Soil description

The soil type and climatic conditions at the time of sampling are given in section 2: Materials and Methods. Results of ANOVA of tested soil characteristics are given in Table A3 in the appendix. As was shown before by Küstermann *et al*. (24), the soil under RT had a higher total soil carbon and nitrogen content (TC = 1.64 ± 0.12%, TN = 0.17 ± 0.01%) than the soil under CT (TC = 1.10 ± 0.13%, TN = 0.12 ± 0.01%, $P = 0.004$ and $P = 0.004$, respectively). MT was not different from either RT or CT treatment (TC = 1.37 ± 0.11%, TN = 0.15 ± 0.01%). Fertilization treatment did not affect total soil carbon or nitrogen content. Contrary to total nitrogen content, mineral dissolved nitrogen was not different between treatments. Nitrate was on average 3.2 µg g^{-1} dry weight soil and ammonium 0.6 µg g^{-1} dry weight soil. However, dissolved organic carbon was affected by tillage treatment, with higher levels in soil under RT (7.5 ± 3.7 µg g^{-1} dry weight soil) than in soil under MT (2.0 ± 3.3 µg g^{-1} dry weight soil) or CT (0.8 ± 5.0 µg g^{-1} dry weight soil, $P = 0.013$). In addition, gravimetric soil water content was significantly higher in the surface layer of soil under RT than under MT and under CT (respectively 22.8 ± 0.7, 21.9 ± 0.3 and 19.2 ± 0.7 %, $P = 0.000$).

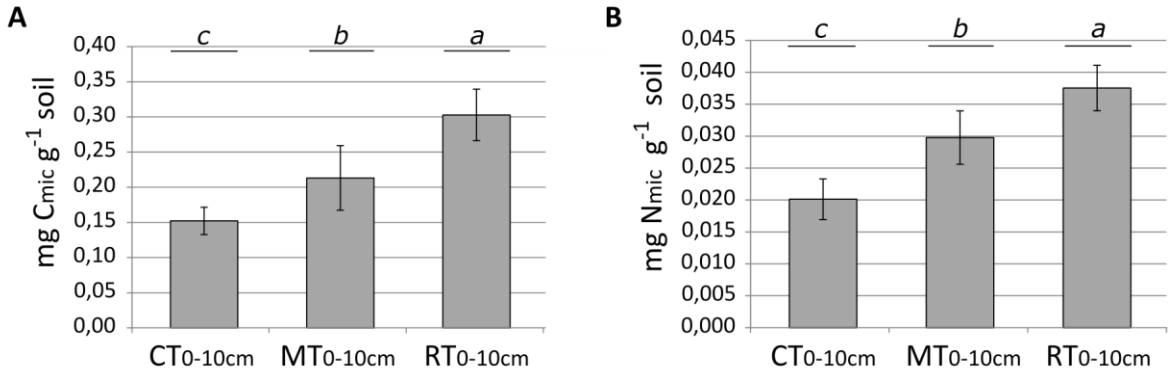


Figure 6: Microbial biomass carbon (C_{mic}, A) and nitrogen (N_{mic}, B) in soil under different tillage treatments. Shown is the amount of microbial biomass carbon or nitrogen detected in in the top 10 cm of soil under conventional tillage (CT), medium tillage (MT) or reduced tillage (RT) in mg per g dry weight soil. Significant differences between tillage treatments are indicated by different letters above the bars ($P<0.05$).

Similarly, soil under RT contained a significantly higher microbial biomass compared to soil under MT, which was again higher than the microbial biomass in soil under CT. Microbial biomass carbon was 0.30 ± 0.04 mg g^{-1} dry weight soil under RT, 0.21 ± 0.05 mg g^{-1} dry weight soil under MT and

0.15 ± 0.02 mg g^{-1} dry weight soil under CT (P = 0.003, Figure 6-A). The same pattern in differences between RT, MT and CT was seen for microbial biomass nitrogen (0.038 ± 0.004, 0.030 ± 0.004 and 0.020 ± 0.003 mg g^{-1} dry weight soil, respectively, P = 0.004, Figure 6-B). qPCR of ribosomal RNA-genes also showed that under RT a higher amount of microbial rRNA-genes is present per gram dry weight soil than under CT (8.37 ± 0.49 × 10^{10} respectively 6.82 ± 0.19 × 10^{10} 16S rRNA gene copies (P = 0.043) and 5.58 ± 1.03 × 10^{8} respectively 3.40 ± 0.40 × 10^{8} ITS rRNA gene copies (P = 0.028)). The ratio of bacterial 16S- to fungal ITS rRNA genes was, however, higher under CT than RT (202 ± 21 respectively 155 ± 39, P=0.047).

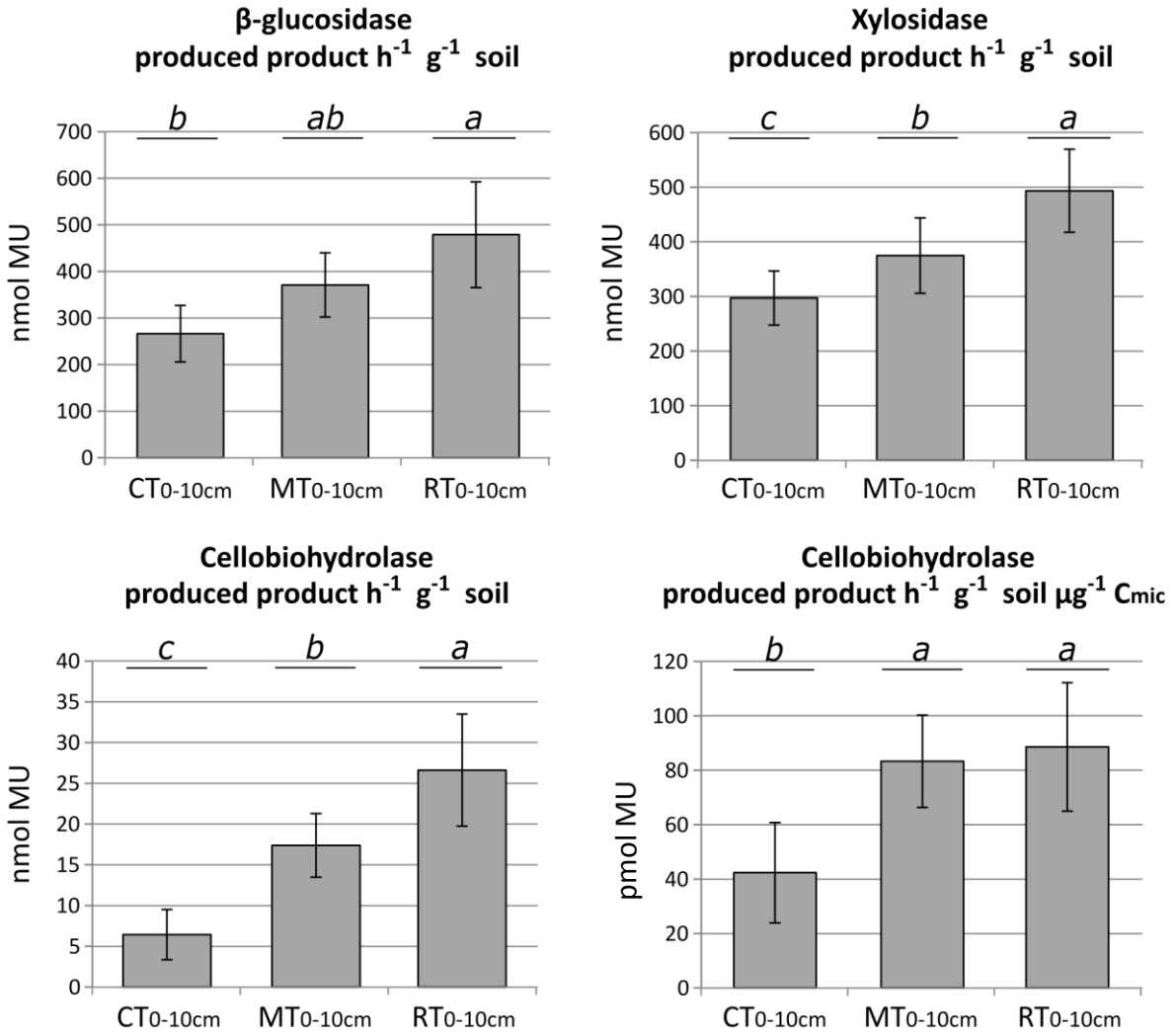


Figure 7: Potential enzymatic activities in soil under different tillage treatments. Shown is the amount of methylumbelliferone (MU) produced per h and per g soil, that has been released by the enzymatic degradation of MU-β-1,4-glucopyranoside (by β-glucosidase), MU-β-D-xylopyranoside (by xylosidase) and MU-β-cellobiohydrofurane (by cellobiohydrolase). For cellobiohydrolase, also the amount of MU produced per h per g soil and per µg of microbial carbon (Cmic) is shown. Soil from the top 10 cm under conventional tillage (CT), medium tillage (MT) or reduced tillage (RT) treatment was measured. Significant differences between treatment means are indicated by different letters above the bars (P<0.05).

The potential activity of extracellular enzymes β-glucosidase, cellobiohydrolase and xylosidase per gram dry weight soil was also affected by tillage (see Figure 7). Potential cellobiohydrolase activity was lower in soil under more intense tillage with lowest activity under CT, intermediate activity

under MT and highest activity under RT treatment (6.44 ± 3.08, 17.39 ± 3.90 and 26.62 ± 6.88 nmol MU h^{-1} g^{-1} dry weight soil, respectively, *P* = 0.003). The same pattern was visible for potential xylosidase activity (297.15 ± 49.43, 375.00 ± 69.07 and 493.60 ± 76.04 nmol MU h^{-1} g^{-1} dry weight soil, under CT, MT and RT respectively, *P* = 0.001). The potential β-glucosidase activity, however, differed significantly (*P* = 0.015) only between soil under CT (266.17 ± 60.69 nmol MU h^{-1} g^{-1} dry weight soil) and RT (479.01 ± 113.55 nmol MU h^{-1} g^{-1} dry weight soil) with intermediate values for soil under MT treatment (370.93 ± 68.85 nmol MU h^{-1} g^{-1} dry weight soil). Potential xylosidase activity was additionally affected by fertilization treatment (*P* = 0.000), with highest values under the HF-treatment than under either MF- and LF-treatment (433.3 ± 116.8, 374.6 ± 97.8 and 357.9 ± 91.1 nmol MU h^{-1} g^{-1} dry weight soil, respectively). After correcting the potential enzyme activities for the differences in soil microbial biomass between samples, no effect of tillage intensity was visible anymore for potential β-glucosidase or xylosidase activity. However, potential cellobiohydrolase activity per microgram of microbial carbon was lower under CT than under either MT or RT (42.3 ± 18.5, 83.3 ± 16.9 and 88.6 ± 23.6 pmol MU h^{-1} g^{-1} dry weight soil μg^{-1} C_{mic} respectively, *P* = 0.014, Figure 7). Also, potential xylosidase activity per microgram of microbial biomass carbon was higher in soil under HF-treatment than under either MF- or LF-treatment (2,098.6 ± 410.7, 1,702.7 ± 216.1 and 1,643.0 ± 443.4 pmol MU h^{-1} g^{-1} dry weight soil μg^{-1} C_{mic} respectively, *P* = 0.036). The overall average potential β-glucosidase activity per microgram of microbial carbon was 1,747.5 ± 550.9 pmol MU h^{-1} g^{-1} dry weight soil.

Microbial community structure

Sequencing statistics

Metagenome sequencing was performed on soil samples of RT and CT treatment with medium fertilization, from now on indicated with RT- and CT-treatment. Sequencing resulted in 0.7 Gbp of raw sequencing data, leading to an average of 157,106 clean reads per sample with an average length of 410 bp (see Table A4 in appendix). To see how much of the soil genetic diversity was sequenced, the level of coverage was analysed using the Nonpareil-method and the taxonomic assignment method. The Nonpareil-method estimated that the metagenome covered around 2.8 % (sample average) of the soil genetic diversity, which was too low to fit a coverage prediction as a function of sequencing effort. The per-sample average coverage estimates by Nonpareil do not significantly differ between tillage treatments (*P*=0.71). However, when comparing to the known diversity of microorganisms in the public database NCBI, rarefaction plots of taxonomic assignments showed that enough coverage has been reached on order level to compare samples and describe most of the diversity (Figure 8-A). An overall analysis of the relative abundance of microorganisms identified on order level using a principal component analysis (Figure 8-B) shows that CT and RT do not show different communities characterizing tillage intensity. In accordance, no significant change in alpha-diversity (Shannon indices: CT, 1.70 ± 0.08; RT, 1.67 ± 0.03, Figure 8-C)) between the metagenomes was detected.

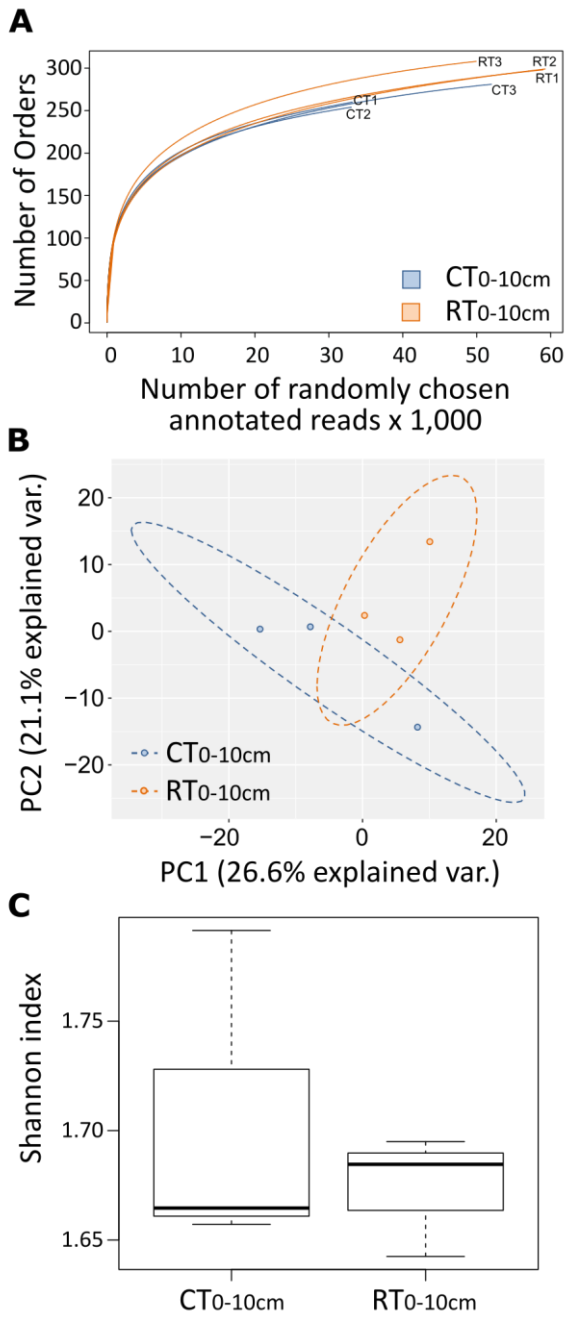


Figure 8: Taxonomic diversity of the conventional farming experiment-samples under conventional tillage (CT) or reduced tillage (RT). (A) Rarefaction curve of the number of orders detected as a function of the number of annotated reads (randomly subsampled), determined by sequence comparison with the NCBI-non-redundant protein database. **(B)** Principal component analysis with 95% confidence range of the rarefied relative order abundances of all 6 metagenome samples. **(C)** Box-plot of Shannon diversity index of rarefied relative order abundances.

Taxonomic composition

Taxonomic assignment of reads was done by comparing them to sequences in the SILVA database and NCBI protein database. Comparison to the SILVA database, using the 0.12% of ribosomal sequences in the metagenome, revealed a dominance of Bacteria (86.15% of all annotated reads) followed by Eukaryota (13.39%, of which 2.8% fungi) and Archaea (0.45%). When comparing to the NCBI database, to which 59% of the reads could be assigned on kingdom level, the percentage of annotated reads that mapped to Bacteria was 98.03%, and those assigned as Eukaryota was 1.08%, of which 0.45% to fungi. To Archaea 0.73% was assigned and to Viruses 0.05%. Figure 9 shows the microbial phyla and orders to which most of the reads were assigned, given in percentage of all metagenome reads times 10^{-3}. Proteobacteria and Actinobacteria were most abundant and together make up 26.4% of all reads. Further prevalent bacterial phyla were Bacteroidetes (3.4%), Acidobacteria (3.3%), Verrucomicrobia (1.7%), Gemmatimonadetes (1.4%), Planctomycetes (1.3%) and Chloroflexi (1.0%). At the order level, Actinomycetales, Rhizobiales, Myxococcales, Burkholderiales and Planctomycetales accounted for 13.6% of all annotated reads.

Several microbial groups showed a different abundance as reaction to tillage treatment. An overview of the relative abundances of these microbial groups and the *P*-values of their responses to tillage treatment is given in Table A5 in the appendix. Here, the most abundant microbial groups reacting to tillage are mentioned. Members of the phylum Chloroflexi showed a higher abundance under CT than under RT (Figure 9-A) and this response was reflected on class level (Chloroflexia and Ktedonobacteria) and order level (Chloroflexales and Ktedonobacterales), although after *P*-value correction only the tillage effect on Chloroflexia remained significant. However, as the effect on members of this phylum is so consistent over different taxonomic levels, one might consider the phylum as a whole being potentially affected by

tillage. In addition, members of the Armatimonadetes showed a higher abundance under CT than RT (Figure 9-A), reflected on order level as well (Fimbriimonas), although not significant after *P*-value correction. Likewise, the tillage effect observed for members of the Solibacterales (Acidobacteria), Stigonematales (Cyanobacteria) (higher abundance under CT than RT) and Betaproteobacteria (higher abundance under RT than CT), was not significant anymore after *P*-value correction.

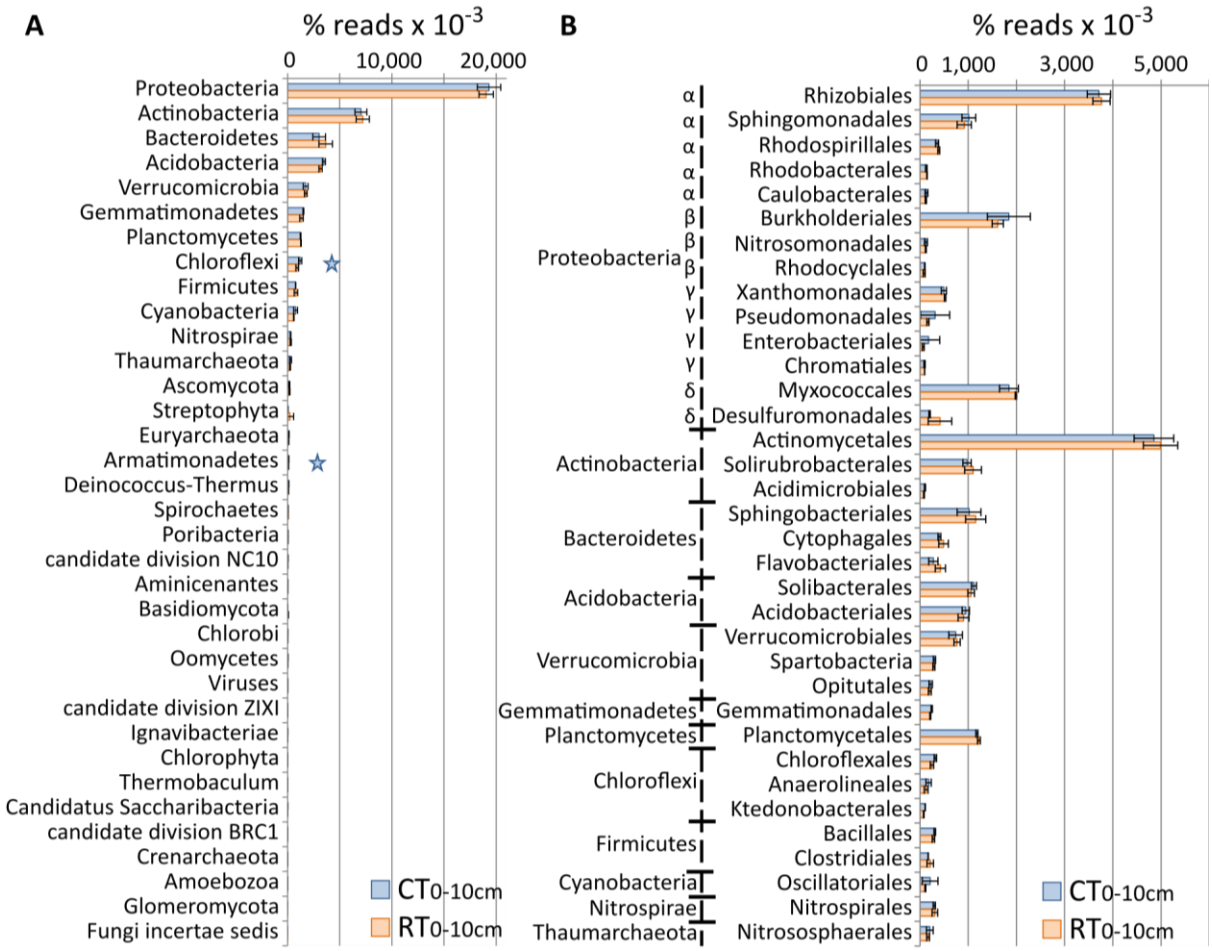


Figure 9: Taxonomic analysis of metagenomes of soil under conventional (CT) and reduced tillage (RT). Shown are the percentage of reads x10-3 of the thirty-five most abundant phyla **(A)** and orders **(B)**, according to the NCBI non-redundant protein database. Stars indicate a significant difference in abundance between treatments (*P*<0.05); light blue: higher abundance under CT, light orange: higher abundance under RT.

Functional analysis

Functional annotation of the metagenome reads was done by comparing them to sequences in the KEGG-database. 35.2% of the reads had a significant similarity to a sequence in the KEGG-database. The thirty-five most abundant KEGG Level4-pathways are shown in Figure 10. Two-component system-enzymes, ABC transporters and enzymes involved in breaking down and synthesizing nucleotides were especially abundant. Together, they accounted for 5.12% of all reads. As was seen for microbial taxonomic assignments, tillage led to a higher abundance of several groups of protein-coding genes under CT than under RT. An overview of the relative abundances of these protein-coding genes and the *P*-values of their responses to tillage treatment is given in Table A5 in the appendix. Soils under CT contained significantly more genes involved in Xenobiotics biodegradation and metabolism (KEGG-Level3, $P = 0.0145$), Methane metabolism ($P = 0.0124$, Figure 10), Fatty acid metabolism ($P = 0.0018$, Figure 10) and Drug metabolism ($P = 0.004$). On the other hand,

Nonribosomal peptide structure-related genes were more abundant under RT than under CT (*P* = 0.017).

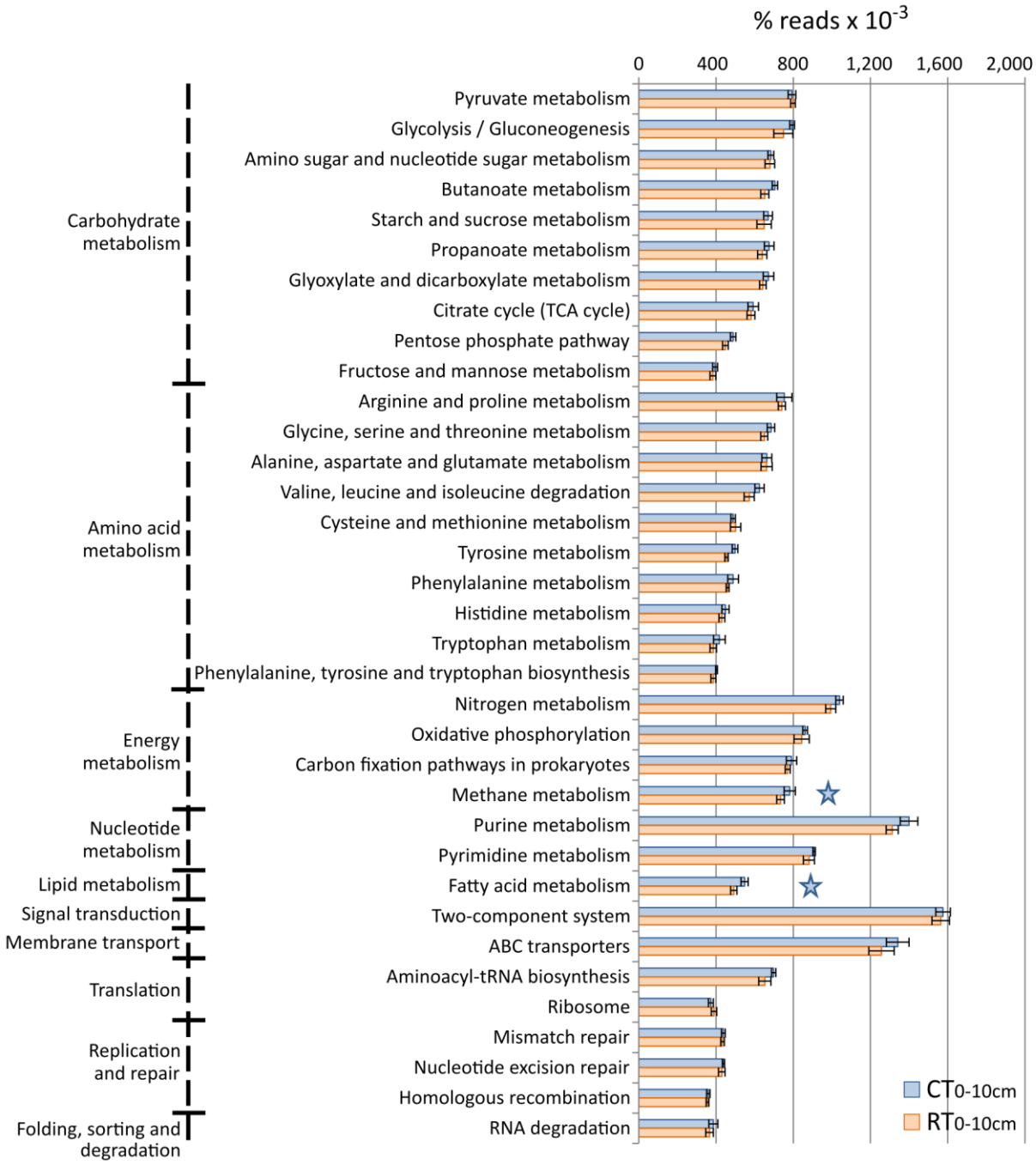


Figure 10: Functional analysis of metagenomes of soil under conventional (CT) and reduced tillage (RT). Shown are the percentages of reads x10^{-3} of the thirty-five most abundant KEGG Level4-pathways. Stars indicate a significant difference in abundance between treatments (P<0.05); light blue: higher abundance under CT, light orange: higher abundance under RT.

Cellulase enzymatic groups and taxonomic assignment

Focusing on cellulose degradation, the annotations to cellulase enzymatic groups using KEGG-orthologous groups were analysed (Figure 11-A). The number of reads annotated to β-glucosidases in the KEGG-database (EC: 3.2.1.21) was 270 (0.071 ± 0.008% of metagenome reads) under CT and

410 reads (0.073 ± 0.003% of metagenome reads) under RT. A total of 123 reads (0.032 ± 0.004% of metagenome reads) for CT and 186 reads (0.033 ± 0.001% of metagenome reads) for RT were annotated to endoglucanase (EC: 3.2.1.4). Far fewer reads were annotated to cellobiohydrolases (EC: 3.2.1.91: CT, 0.001 ± 0.0006%; and RT 0.002 ± 0.0018%) and cellobiose phosphorylase (EC: 2.4.1.20: CT 0.0019 ± 0.0014%; and RT 0.001 ± 0.0005%) and none to reducing-end-acting cellobiohydrolase (EC: 3.2.1.176), cellobiose dehydrogenase (EC: 1.1.99.18) or lytic polysaccharide monooxygenases (EC: 1.14.99.54). In the KEGG-database there were no orthologous groups representing cellodextrinase (EC: 3.2.1.74) or cellodextrin phosphorylase (EC: 2.4.1.49). No significant differences between tillage treatments were detected for these cellulases.

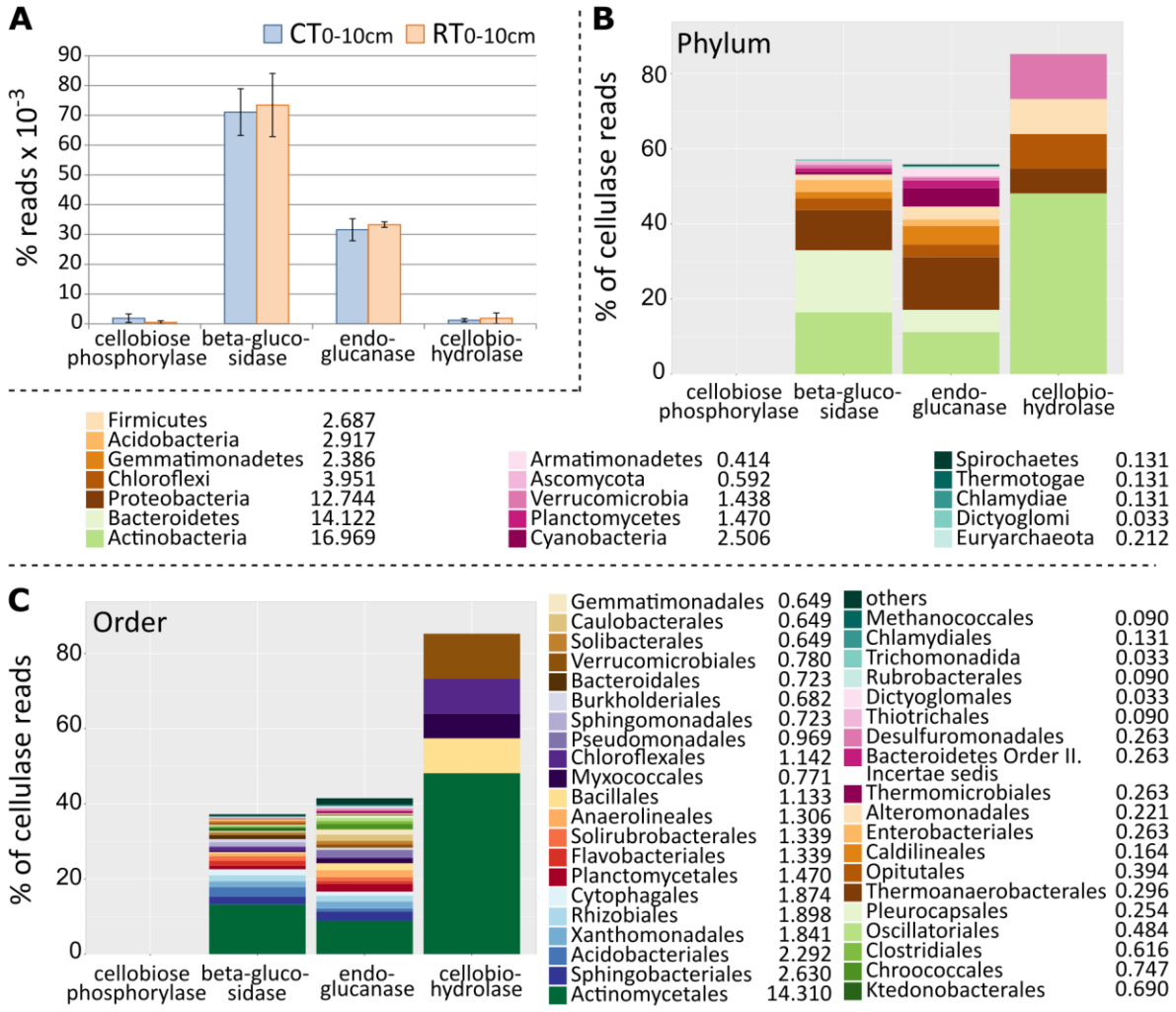


Figure 11: (A) Abundance of cellulase enzymatic groups in metagenome of soil under conventional (CT) or reduced tillage (RT) in the surface soil layer. Shown is the percentage of reads x10^{-3} annotated to cellulase enzymatic functions as mean values for each tillage treatment. Furthermore, the taxonomic composition of reads annotated to cellulase enzymatic groups on phylum **(B)** and order **(C)** level is shown**.** The reads annotated to each cellulase enzymatic group were pooled over all samples and taxonomically assigned. Shown are the forty most abundant taxa, as percentage of reads annotated to the cellulase enzymatic group. The taxonomically unassigned reads are not shown but represented by the rest of the bar until 100%. In addition, the contribution of each microbial group to the cellulase degradation-potential is given behind the name of the microbial group as percentage of all reads annotated as cellulase enzymatic groups.

The taxonomic assignment of sequences annotated as cellulase enzymatic groups on phylum and order level (Figure 11-B and -C) shows that the potential to degrade cellulose using different enzymatic mechanisms is spread widely across different phyla and orders, especially for β-glucosidases and endoglucanases. Fungi are represented in the β-glucosidase and endoglucanase genes belonging to Ascomycota, but make up only a small part of the cellulolytic microorganisms here (0.7% of cellulase reads). Cellobiohydrolase genes (EC: 3.2.1.91) were assigned to a less diverse group of microorganisms and especially to Actinomycetales (Actinobacteria). Unfortunately, none of the reads annotated as cellobiose phosphorylase could be taxonomically assigned on phylum level, although all were found to have bacterial origin. The majority of endoglucanases was assigned to Actinobacteria and Proteobacteria, which were the phyla to which also most β-glucosidase genes were assigned. However, β-glucosidase genes derived from Bacteoidetes were more abundant than those derived from Proteobacteria. No effect of tillage treatment was observed on taxonomically assigned genes encoding cellulase enzymatic groups.

Annotation method of cellulase domain families

As the KEGG-database unfortunately contains a limited number of orthologous groups designated as cellulase enzymatic groups, the translated metagenome sequences were scanned using HMMs to detect a higher variety of cellulase genes. This is a common method to annotate protein sequences from sequence libraries. Several catalytic domain- and CBM-families from the CAZy database (www.cazy.org, (60)) were selected as queries to search the translated metagenome sequences. Because most of the selected domain families contain a varying amount of enzymatic functions other than cellulases (i.e. endoglucanases, exoglucanases, β-glucosidases, cellodextrin phosphorylases and cellobiose dehydrogenases), the available corresponding HMMs are not always specific for cellulases. Therefore, for each domain family, HMMs from different sources were assessed for cellulase-specificity and the most specific and sensitive HMM was selected to scan the metagenome data (see Table A1 in the appendix for the results of the assessment).

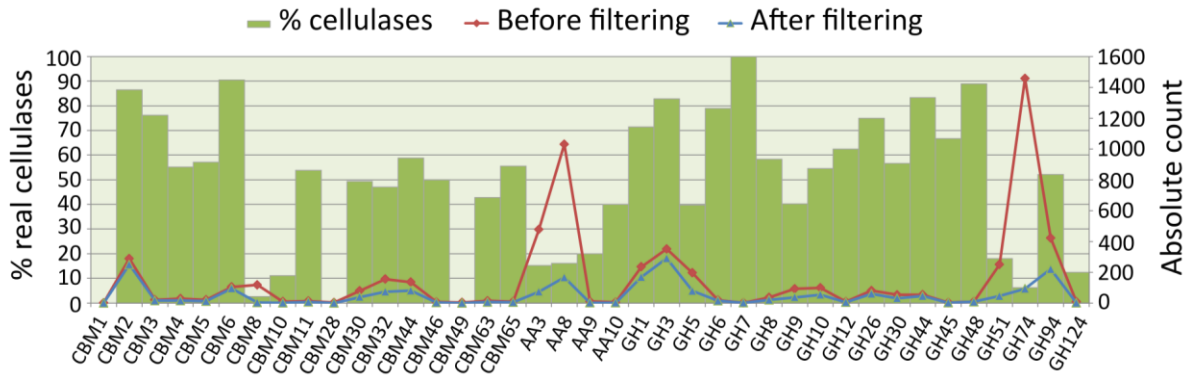


Figure 12: Results of the filtering step of HMM-annotated reads in the metagenome of the conventional farming experiment to cellulase catalytic and binding domain families. Shown is the percentage of HMM-annotated reads with a significant BLAST hit to a sequence in the positive cellulase sequences-database (bars). In addition, the absolute amount of reads annotated to cellulase catalytic and binding domain families before and after the filtering step is shown (lines).

Using this method, 5,906 reads (0.63 % of metagenome reads) could be annotated to the selected domain families. However, as many of the selected HMMs remained relatively unspecific for cellulases, a second filtering step of the annotated reads was performed by BLASTing them against a

database containing only confirmed cellulase sequences (see section 2: Materials and Methods). The result of this filtering step is shown in Figure 12, and highlights the low specificity achieved by identification of cellulases using HMMs.

Cellulase domain families

In total, 2,021 reads were annotated as cellulase catalytic (GH- and AA-) and binding (CBM-) domain families, which corresponds to 0.21% of the total amount of metagenome reads. These reads were annotated to 18 GHs, four AAs and 17 CBMs (Figure 13), which are all domain families known to harbour enzymes involved in cellulose degradation (see Figure 2). Most hits were found for GH1, GH3, GH94, AA8, CBM2 and CBM6. There were no differences between tillage treatments in total amount of annotated cellulases. In addition, no differences between CT and RT were encountered for the separate cellulase domain families, except for CBM11, which was more abundant in conventional tillage soil than in reduced tillage soil (P = 0.044). However, this cellulase domain family was not very abundant (<0.01% of metagenome reads) and after correcting for multiple testing this difference was not significant anymore.

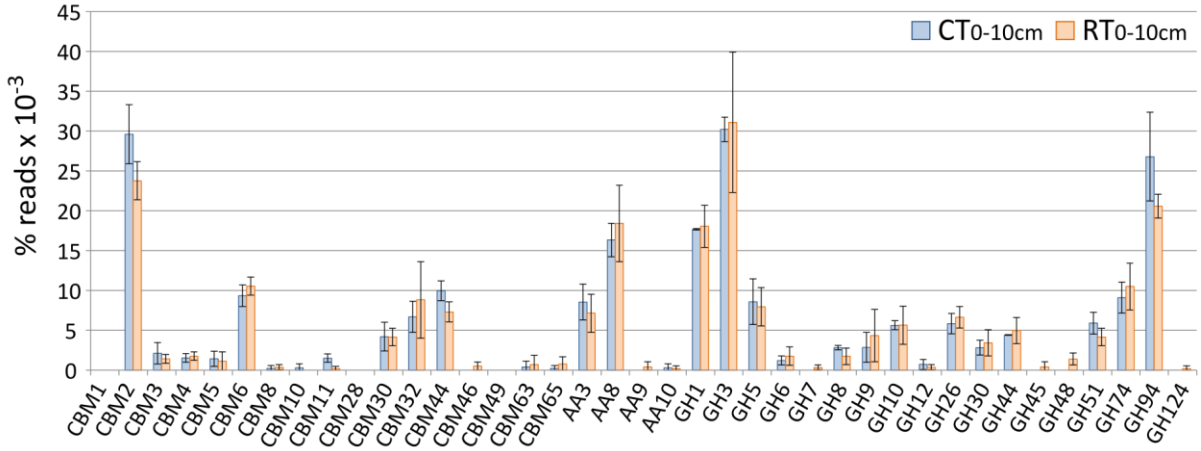


Figure 13: Abundance of cellulase domain families in metagenomes of soil under conventional (CT) or reduced tillage (RT). Shown is the percentage of reads x10-3 annotated to cellulase catalytic modules (AA and GH) and CBMs in both metagenomes. Annotation was done by scanning the reads with HMMs and subsequent BLASTing against a positive cellulase sequence database.

Taxonomic assignment of cellulase domain families

The taxonomic affiliation of the sequences annotated to cellulase domain families is shown in Figure 14 on phylum and order level, and reflects the overall abundance of phyla in the metagenome. A total of 18.6% of all cellulase reads mapped to the Proteobacteria, 11.2% to Actinobacteria and fewer to Bacteroidetes (8.6%), Cyanobacteria (2.2%), Gemmatimonadetes (2.2%), Verrucomicrobia (2.1%), Acidobacteria (1.7%), Firmicutes (1.5%) and Chloroflexi (1.53%). Compared to the relative abundance of Actinobacteria, Cyanobacteria, Firmicutes and Bacteroidetes among all metagenome reads (7.1, 0.7, 0.8 and 3.3% of the total reads, respectively), the cellulase genes derived from these phyla were relatively more abundant. 0.4% of the cellulase reads mapped to fungi (Ascomycota and Basidiomycota). Results on order level support the importance of the role which Actinobacteria (Actinomycetales), Proteobacteria (Rhizobiales, Burkholderiales, Xanthomonadales, Myxococcales),

Bacteroidetes (Cytophagales, Sphingobacteriales) and Verrucomicrobia (Verrucomicrobiales) play in cellulose degradation.

Whereas sequences in many different cellulase domain families, including the most abundant ones, were derived from Actinobacteria (Actinomycetales), Proteobacteria (Rhizobiales) and Bacteroidetes (Sphingobacteriales), some microbial groups appeared to harbour a higher abundance of cellulase genes from a specific cellulase domain family than from other domain families. These observations offer possibilities to recognize microbial specialization towards certain cellulase domain families. The microbial groups to which a higher abundance of assigned reads can be found in specific cellulase domain families compared to other cellulase domain families are listed in Table 1. A low abundance of cellulase sequences were found assigned to fungi; in domain families AA8, AA9, CBM6, GH1, GH3 and GH5 sequences were found which were derived from Ascomycota. In addition, GH5 contained a sequence derived from Basidiomycota and GH12 contained a sequence derived from Streptophyta.

Table 1: Microbial taxonomic groups to which a higher abundance of sequences was assigned in the respective cellulase domain family than in other domain families.

Microbial taxonomic group		
Phylum level	**Order level**	**Cellulase domain family**
Actinobacteria	Solirubrobacterales	AA8 and GH1
Acidobacteria	Acidobacteriales	GH3
	not identified	GH74
Bacteroidetes	Bacteroidales	GH3
	Cytophagales	CBM44
	Flavobacteriales	GH74
	Sphingobacteriales	CBM44
Chloroflexi	Anaerolineales	GH1
	Chloroflexales	GH3
	Ktedonobacterales	GH3
	not identified	GH74
Cyanobacteria	Chroococcales	CBM2
	Oscillatoriales	CBM2
Gemmatimonadetes	Gemmatimonadales	GH3
	not identified	GH94
Planctomycetes	Planctomycetales	CBM2
Proteobacteria	Burkholderiales	AA3 and AA8
	Enterobacteriales	GH1
	Methylococcales	GH94 and GH26
	Pseudomonadales	GH5
	Rhizobiales	GH94 and AA8
	Sphingomonadales	GH3
	Xanthomonadales	CBM2
Verrucomicrobia	Opitutales	GH51
	Verrucomicrobiales	GH9

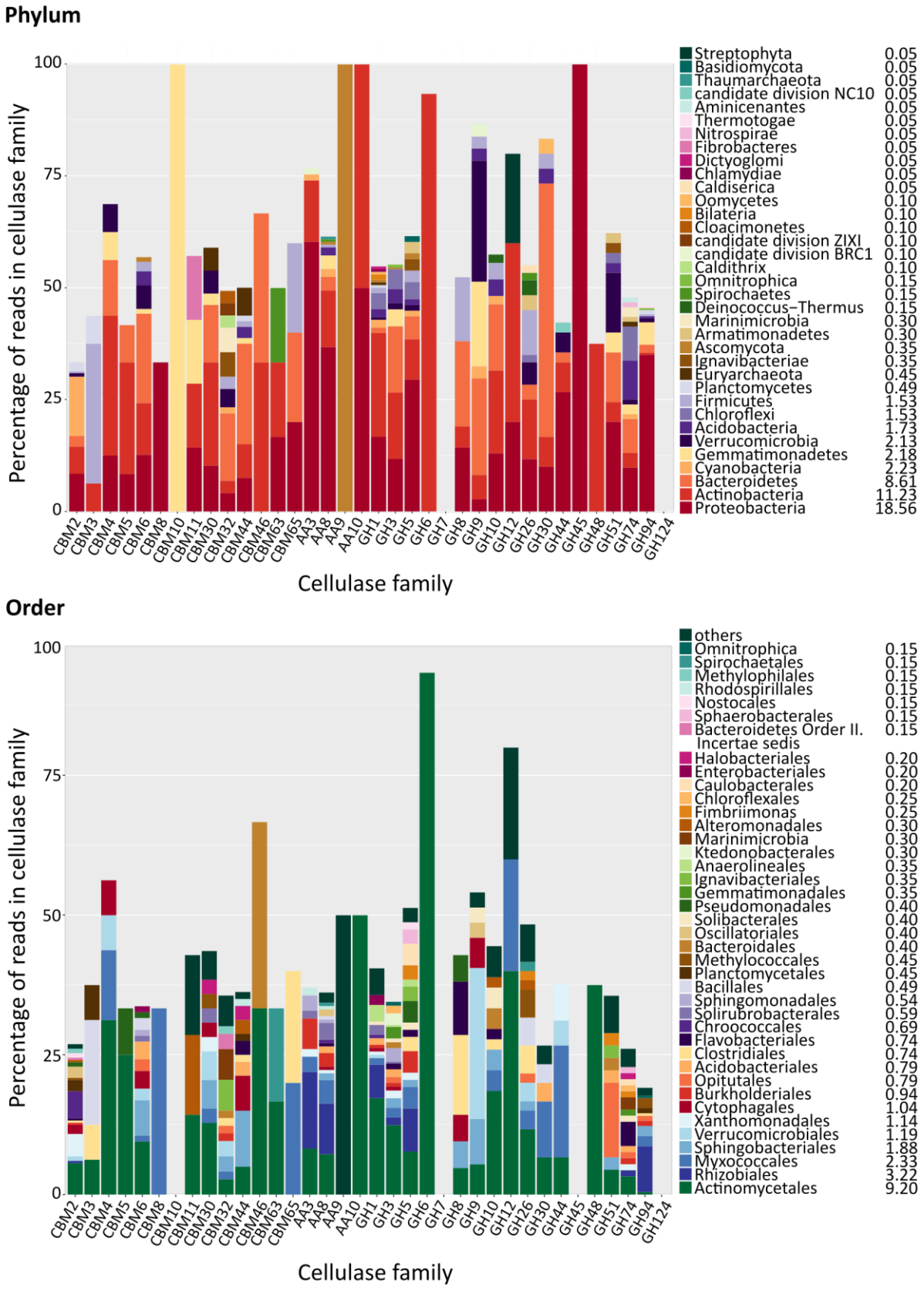


Figure 14: Taxonomic assignment of cellulase domain family-reads on phylum and order level. The reads annotated to cellulase domain families were pooled over all samples and taxonomically assigned. Shown is the percentage of reads in the cellulase domain family annotated to the most abundant phyla or orders. The taxonomically unassigned reads are not shown but represented by the rest of the bar until 100%. In addition, the contribution of each taxonomic group to the cellulose degradation potential is given as percentage of all cellulase domain family-reads.

On the other hand, cellulase genes in some cellulase domain families were assigned in higher abundance to certain microbial taxa than to other taxa (Figure 14); the largest part of AA3-, AA8- and GH94-sequences was harboured by Rhizobiales (Alphaproteobacteria), while the majority of GH44-genes were assigned to Myxococcales (Deltaproteobacteria). Furthermore, all GH6-genes were assigned to Actinobacteria (Actinomycetales), whereas the largest part of GH30- and CBM44-sequences was derived from Bacteroidetes (CBM44: Sphingobacteriales, GH30: not resolved on order level). In addition, CBM3-genes were mostly assigned to Bacillales (Firmicutes), while the largest part of GH8-genes was assigned to Clostridiales (Firmicutes) and Flavobacteriales (Bacteroidetes). Furthermore, a considerably high abundance of GH51-genes was harboured by Opitutatales (Verrucomicrobia) and most GH9-genes to Verrucomicrobiales (Verrucomicrobia). For low-abundant cellulase domain families, it could be observed that sequences in AA9 were solely derived from Ascomycota (Sordariales), those in AA10 only to Actinobacteria and Proteobacteria, those in CBM10 exclusively to Gemmatimonadetes and those in CBM8 and GH45 solely to Proteobacteria. Furthermore, all sequences in CBM46 were harboured by Actinobacteria and Bacteroidetes and the majority of sequences in CBM5, GH12 and GH48 to Actinobacteria (Actinomycetales).

Finally, the taxonomic affiliations of the sequences in the most abundant cellulase domain families (GH1, GH3, GH94, AA8, CBM2 and CBM6) are shown separately in Figure 15 on phylum level and in Figure 16 on order level. Due to high variation in the amount of assigned reads among replicates and treatments, not many treatment effects could be detected. However, some differences in abundances of cellulase genes assigned to microbial taxa between tillage treatments can be observed; GH3-genes derived from Chloroflexi were more abundant under RT than under CT, while AA8-genes harboured by Solirubrobacterales (Actinobacteria) were more abundant under CT than under RT.

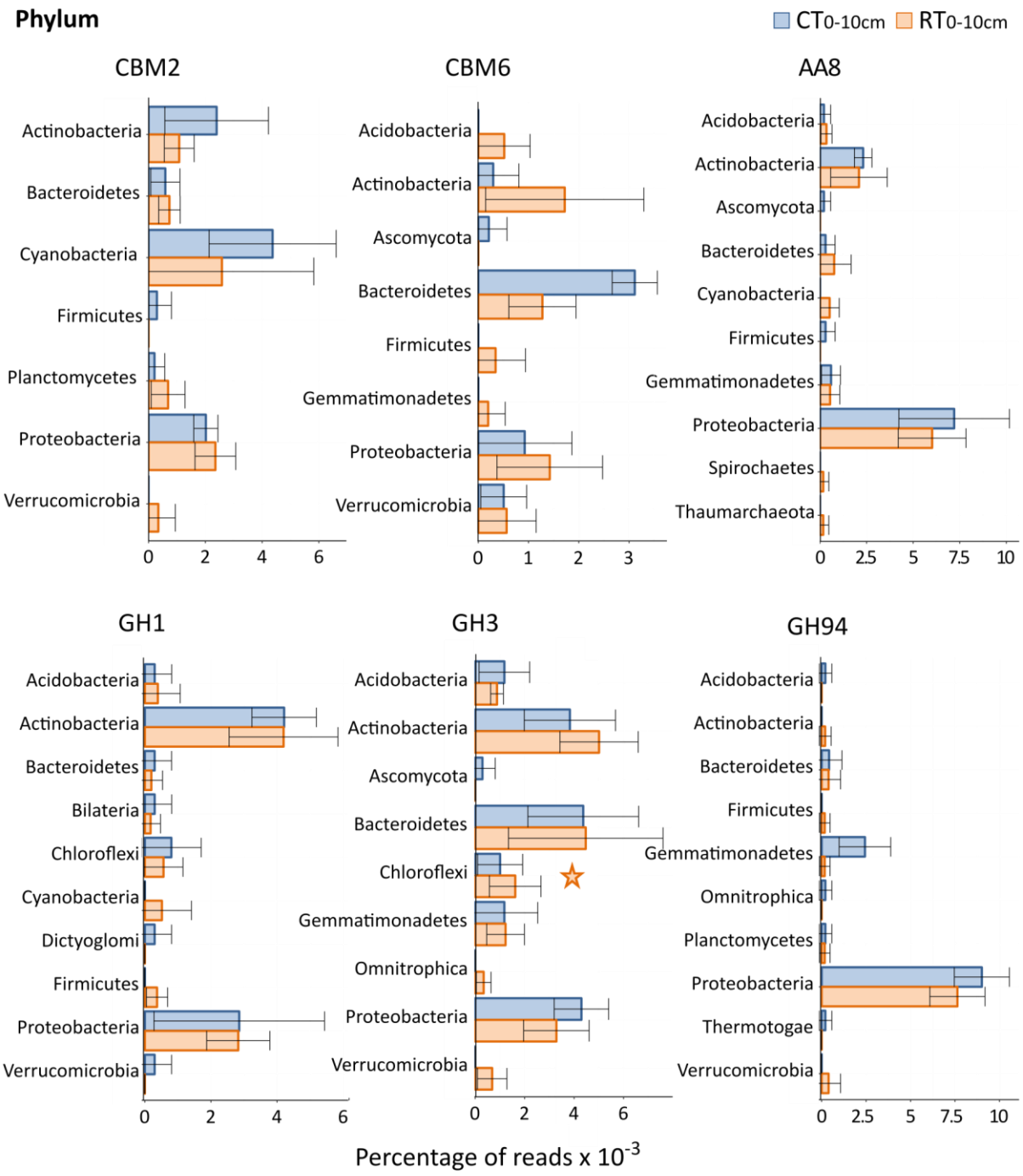


Figure 15: Taxonomic analysis of the six most abundant cellulase domain families on phylum level under conventional (CT) or reduced (RT) tillage. The reads annotated to the cellulase domain families in each sample were taxonomically assigned and treatment averages were calculated. For each cellulase domain family, the percentage of metagenome reads x10^{-3} annotated to the ten most abundant phyla is shown. Stars indicate a significant difference in abundance between treatments (P<0.05); light blue: higher abundance under CT, light orange: higher abundance under RT

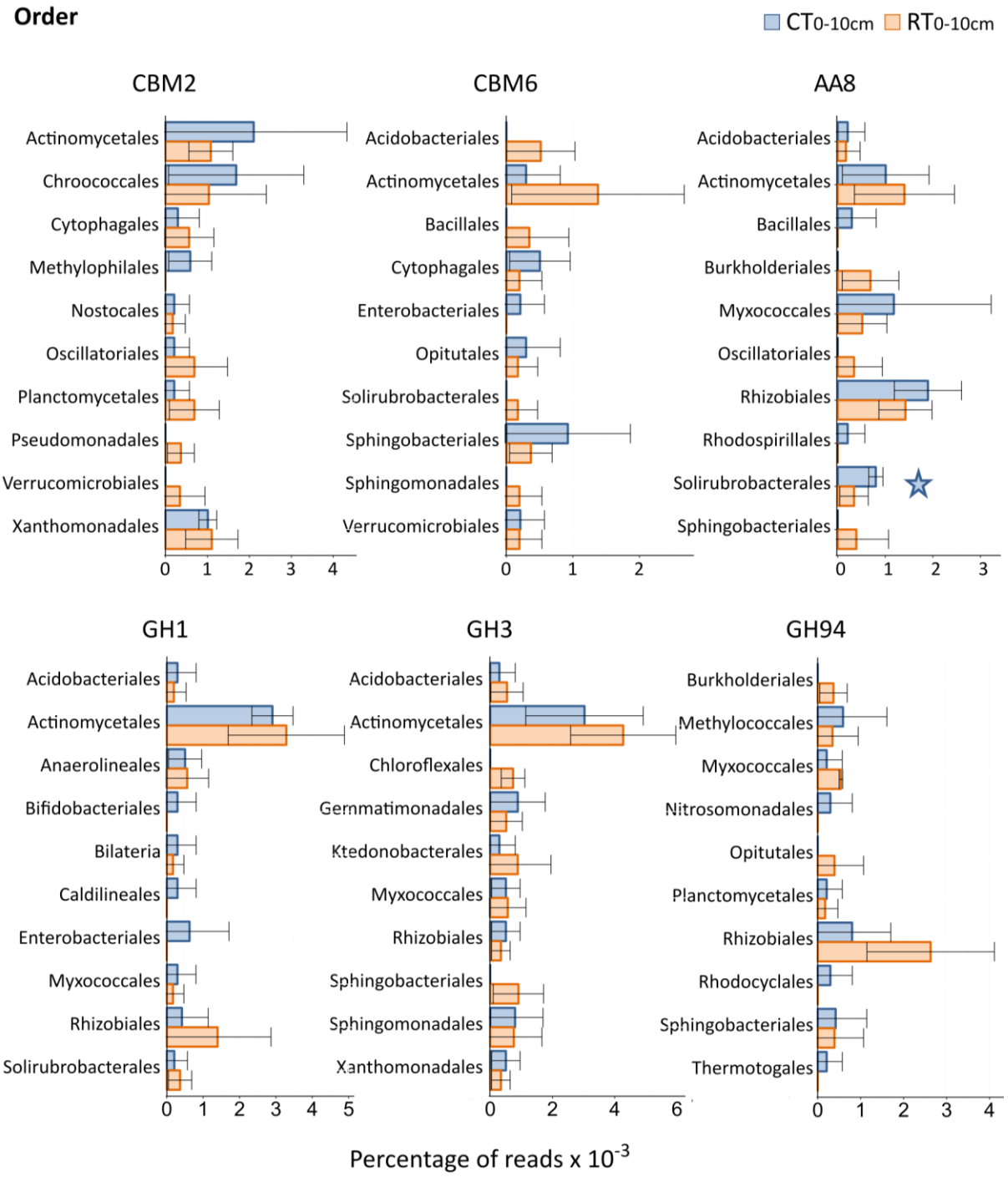


Figure 16: Taxonomic analysis of the six most abundant cellulase domain families on order level under conventional (CT) or reduced (RT) tillage. The reads annotated to the cellulase domain families in each sample were taxonomically assigned and treatment averages were calculated. For each cellulase domain family, the percentage of metagenome reads x10^{-3} annotated to the ten most abundant orders is shown. Stars indicate a significant difference in abundance between treatments (P<0.05); light blue: higher abundance under CT, light orange: higher abundance under RT.

4 - Results of the organic farming experiment

Here, results are presented from the analysis of soil characteristics and metagenomic data of the surface soil layer (0-6 cm) and the deeper soil layer (10-16 cm) of a field experiment located at the Agroscope Reckenholz research institution. In this experiment, the effect of different farming systems and different cover crop treatments on yield and several plant parameters is investigated (106). The different farming systems consisted of "organic management x plough tillage" (designated here as conventional tillage, CT), "organic management x reduced tillage" (designated here as RT), "conventional management x plough tillage" (designated here as PT) and "conventional management x no-tillage" (designated here as NT). The different cover crop treatments are legumes (L), non-legumes (NL), mixture (M) and no cover crop (NO). Treatments were organized in a split-plot-design with farming systems as main plot and cover crop treatment as subplot, in four replicate blocks. For investigation of the effect of tillage on the soil microbial community, soil was sampled from all farming system treatments on plots treated with legumes as cover crop (L) or without cover crop (NO) and from two soil depths (the surface soil layer and the deeper soil layer, as mentioned above). For an overview of the experimental setup and more details on the treatments and methodology, see Figure 4 and section 2: Materials and Methods. Soil chemical analysis was done on all soil samples.

Soil description

The results of the analysis of variance (split-split-plot) on soil parameters measured on the 64 samples are given in Table A6 in the appendix (*P*-values and means). No effect of cover crop was detected for any of the soil parameters measured, except for soil dissolved organic carbon (see below). A general depth effect was seen for pH, ammonium, TC and TN. pH (H_2O) was higher in the deeper soil layer than in the surface soil layer (7.91 ± 0.35 respectively 7.76 ± 0.37, $P = 0.000$), while a higher concentration of available ammonium (0.3 ± 0.2 respectively 0.1 ± 0.1 µg g^{-1} dry weight soil, $P = 0.000$), a higher amount of TC (1.77 ± 0.61 respectively 1.45 ± 0.31 %, $P = 0.008$) and a higher amount of TN (0.17 ± 0.04 respectively 0.15 ± 0.03 %, $P = 0.006$) was observed in the surface soil layer than in the deeper soil layer. Contrary to the results seen in the conventional farming experiment, no effect of farming practices (organic vs. conventional management or reduced vs. conventional tillage) was seen on soil TC, TN and ammonium content. However, farming system did have an effect on soil nitrate content, which was higher in the NT-system than in both CT- and PT-systems, independent of soil depth (NT= 5.4 ± 4.0, CT= 2.4 ± 3.6 and PT= 3.6 ± 4.2 µg g^{-1} dry weight soil, $P = 0.020$). Nitrate content in the RT-system was not different from the other treatments (3.8 ± 3.7 µg g^{-1} dry weight soil). The effect of farming system on soil gravimetric water content was significant but dependent on soil depth ($P = 0.001$); soil water content was higher in RT- and NT-systems (19.7 ± 0.8 respectively 19.7 ± 1.3 %) than in CT- and PT-systems (18.4 ± 1.1 respectively 18.3 ± 1.6 %), but only in the surface soil layer. In the deeper soil layer, soil water content was higher in the PT-system (19.7 ± 0.5 %) than in the RT-system (19.0 ± 0.6 %). Differences in water content between the surface soil layer and the deeper soil layer were seen for all farming systems, except for the NT-system, which had an overall high humidity. Dissolved organic carbon levels were different between farming systems, cover crop treatments and soil depth layers ($P = 0.007$). In the surface soil layer of treatments without cover crop, dissolved organic carbon levels were higher under CT than under RT and NT (respectively 205.0 ± 150.1, 29.7 ± 80.0 and 48.7 ± 115.8 µg g^{-1} dry weight soil), with intermediate and not significantly different values under PT (84.9 ± 146.6 µg g^{-1} dry weight soil).

However, these differences in dissolved organic carbon levels between farming systems were not observed in the surface soil layer of treatments with legume as cover crop. In the deeper soil layer of treatments without cover crop, dissolved organic carbon levels were higher under NT than under PT and CT (respectively 154.7 ± 177.7, -0.7 ± 32.1 and -13.7 ± 13.8 µg g^{-1} dry weight soil), with intermediate and not significantly different values under RT (61.6 ± 113.5 µg g^{-1} dry weight soil). Similar to observations in the surface soil layer, however, the differences in dissolved organic carbon levels between farming systems were not observed in the deeper soil layer of treatments with legume as cover crop.

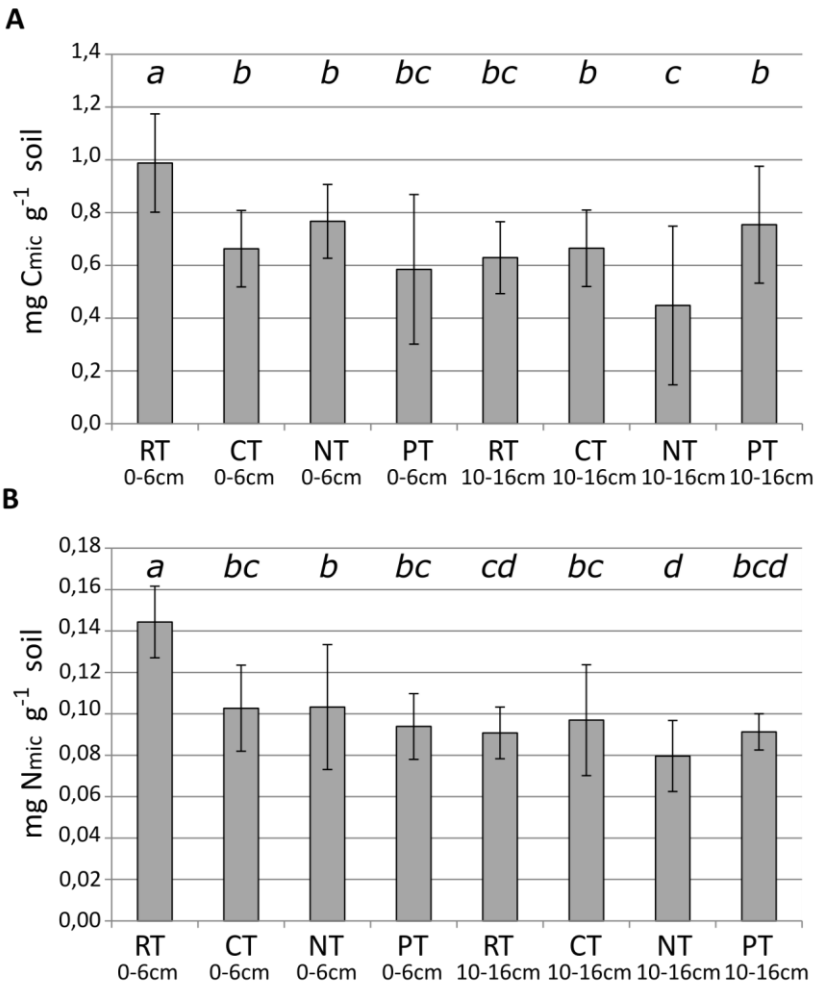


Figure 17: Microbial organic carbon (A) and nitrogen (B) in soil under different tillage treatments and in different soil depths. Shown is the amount of microbial carbon or nitrogen in mg per g dry weight of soil, detected in the top 6 cm of soil under organic management and reduced tillage (RT0-6cm), organic management and conventional tillage (CT0-6cm), conventional management and no tillage (NT0-6cm) or conventional management and conventional tillage (PT0-6cm) and in the lower 10-16 cm of the soil (RT10-16cm, CT10-16cm, NT10-16cm respectively PT10-16cm). Significant differences between tillage treatments are indicated by different letters above the bars (P<0.05).

Soil microbial biomass, measured as microbial carbon (C_{mic}) and nitrogen (N_{mic}) was also affected by farming system and depth (P = 0.001 (C_{mic}) and P = 0.000 (N_{mic}), see Figure 17). C_{mic} and N_{mic} were higher in the RT-system (C: 0.99 ± 0.19, N: 0.14 ± 0.02 mg g^{-1} dry weight soil) than in all other three systems (C: 0.67 ± 0.19, N: 0.10 ± 0.02 mg g^{-1} dry weight soil), but only in the surface soil layer. In the deeper soil layer, C_{mic} and N_{mic} were lower in the NT-system (C: 0.45 ± 0.30, N: 0.08 ± 0.02 mg g^{-1} dry weight soil) than in the CT-system (C: 0.66 ± 0.14, N: 0.10 ± 0.03 mg g^{-1} dry weight soil), but values in the RT- and PT-system did not significantly differ between farming systems (C: 0.69 ± 0.18, N: 0.10 ± 0.01 mg g^{-1} dry weight soil). Higher C_{mic} and N_{mic} levels in the surface soil layer than in the deeper soil layer were seen only for the RT- and NT-systems.

Of all samples taken, only 16 were used for metagenome library generation (CT- and RT-systems from both depths and with four replicates). When analyzing the soil parameters of exclusively these

samples (split-plot), the same effects were observed for soil pH and soil ammonium content. However, no differences in soil TC and TN content were now observed between soil depth layers. While the soil nitrate content did not differ between soil depth layers of all samples, analysis of the samples of the CT- and RT-systems alone revealed higher nitrate content in the deeper soil layer than in the surface soil layer of soil under CT. Nevertheless, no difference in nitrate content was detected between soils under RT or CT. The difference in soil humidity between RT and CT was in accordance with the results observed for all 64 samples. Soil under RT was characterized by a higher soil water content than soil under CT, but only in the surface soil layer. Furthermore, soil water content was higher in the surface layer than in the deeper layer for soil under RT, whereas the opposite was true for soil under CT. In addition, for microbial carbon (C_{mic}) and nitrogen (N_{mic}), generally the same results were found. C_{mic} and N_{mic} were higher under RT than under CT, but only in the surface soil layer. However, in de deeper soil layer N_{mic} was found to be significantly lower under RT than under CT. Furthermore, soil under RT was found to contain higher amounts of C_{mic} and N_{mic} in the surface layer than in the deeper layer. For soil under CT, only N_{mic}, was found to be higher in the surface layer than in the deeper layer.

Bacterial and fungal ribosomal RNA (rRNA)-gene abundances were measured for the soil sampled used for metagenome library generation. Ribosomal RNA-gene abundances were considerably higher in the surface soil layer ($4.04 \pm 0.52 \times 10^{10}$ 16S rRNA gene copies and $6.03 \pm 1.95 \times 10^{9}$ ITS copies) than in the deeper soil layer ($1.59 \pm 0.24 \times 10^{10}$ 16S rRNA gene (P=0.000001) and $2.56 \pm 1.27 \times 10^{9}$ ITS (P=0.00017) copies). No effect of tillage treatment was observed on either bacterial or fungal rRNA-gene abundances. Neither tillage nor depth had an effect on the average bacterial-fungal rRNA-gene ratio (7.30 ± 2.65 16S/ITS).

Microbial community structure

Sequencing statistics

Metagenome sequencing was performed using DNA extracted from soil samples under RT and CT (organic management), from plots without cover crop treatment (NO) and from both depth layers. Sequencing resulted in 16.2 Gbp of raw sequencing data, leading to an average of 3,252,324 clean reads per sample (replicate) with an average length of 258 bp (see Table A7 in the appendix). To analyze the extent to which the soil genetic diversity had been sequenced, the level of coverage was calculated using both the Nonpareil-method and the taxonomic assignment method. The Nonpareil-method estimated that the metagenome covers around 3.6 % (sample average) of the soil microbial diversity, which was too low to fit a coverage prediction as a function of sequencing effort. The per-sample average coverage estimates by Nonpareil did not differ between tillage treatments (surface soil layer: P=0.99, deeper soil layer: P=0.81), but significantly differed between the soil samples of the surface layer (3.1%) and deeper layer (4.2%, two-tailed paired t-test P-value = 0.006). Rarefaction plots of taxonomically assigned reads, obtained by comparison to sequences of known micro-organisms in the public databases (NCBI), showed that enough coverage has been reached on family level to compare samples (Figure 18-A and B). Furthermore, overall analysis of the relative abundance of microorganisms identified on family level using a principal component analysis Figure 18-C) showed that communities were rather different between soil depth layers than between tillage treatments. However, alpha-diversity analysis of family-level taxonomic assignments indicated that the metagenome of soil under RT was more diverse than soil under CT in the surface soil layer,

while the reverse was observed in the deeper soil layer (Shannon indices: RT0-6cm, 1.24 ± 0.03 (*b*); CT0-6cm, 1.17 ± 0.03 (*c*); RT10-16cm, 1.24 ± 0.03 (*b*); CT10-16cm, 1.29 ± 0.02 (*a*)), see Figure 18-D, Simpson indices: RT0-6cm, 0.33 ± 0.01 (*ab*); CT0-6cm, 0.31 ± 0.01 (*c*); RT10-16cm, 0.33 ± 0.01 (*b*); CT10-16cm, 0.34 ± 0.01 (*a*)).

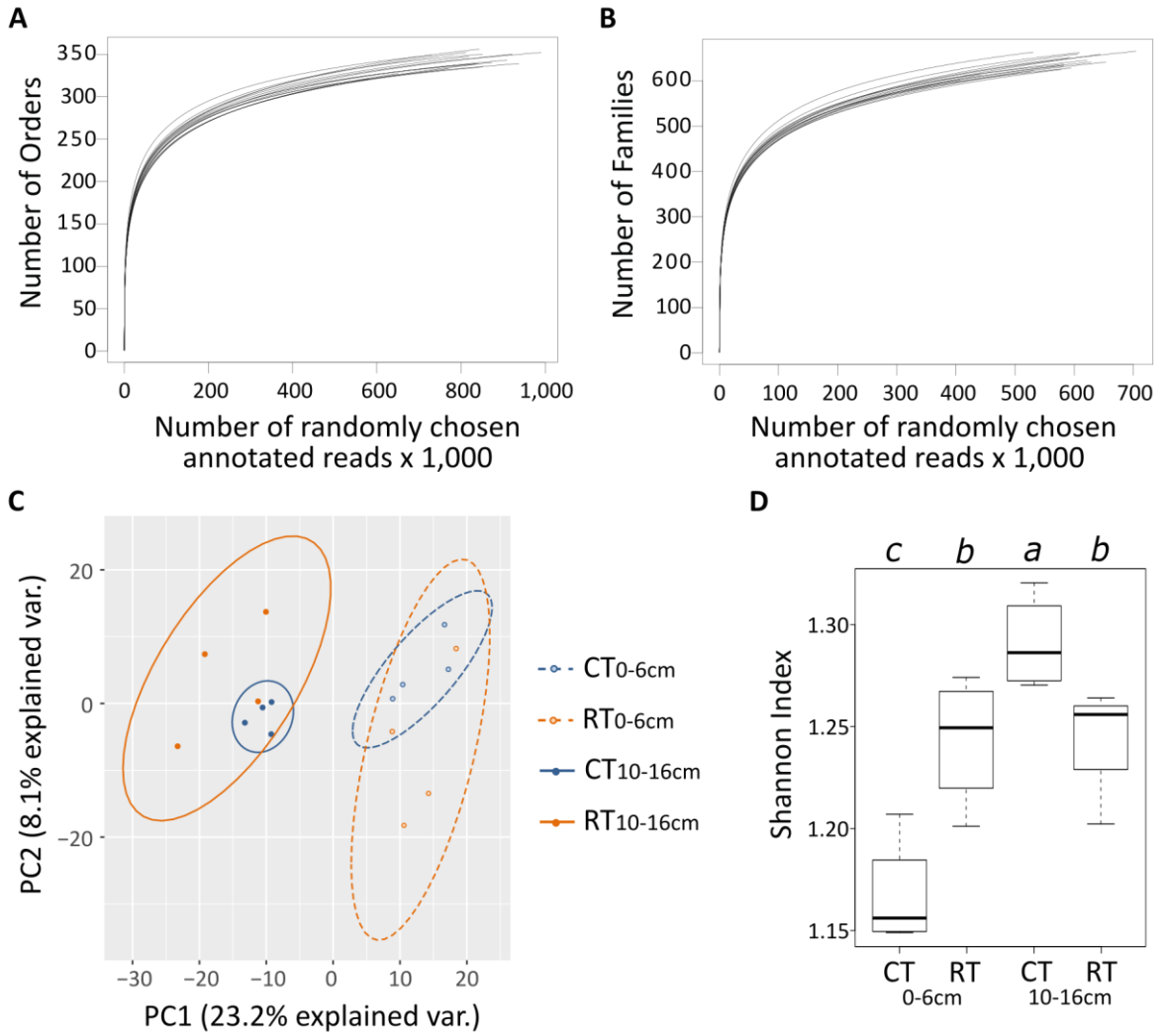


Figure 18: Taxonomic diversity of the organic farming experiment-samples under conventional tillage (CT) or reduced tillage (RT) in the top 6 cm of soil (the surface soil layer) or the top 10-16 cm of soil (the deeper soil layer). (A) Rarefaction curve of the number of orders detected as a function of the number of annotated reads (randomly subsampled), determined by sequence comparison with the NCBI-non-redundant protein database. (B) Rarefaction curve of the number of families detected as a function of the number of annotated reads (randomly subsampled). (C) Principal component analysis with 95% confidence range of the rarefied relative family abundances of all 16 metagenome samples. (D) Box-plot of Shannon diversity index of rarefied relative family abundances. Different letters above the box-plots indicate significantly different means ($P<0.05$).

Taxonomic composition

Taxonomic assignment of reads was done by comparing them to sequences in the SILVA database and NCBI protein database. Comparison to the SILVA database, to which 0.05% of the metagenome reads could be mapped, revealed a dominance of Bacteria (88.4% of all reads assigned) followed by Eukaryota (8.4%, of which 1.5% fungi) and Archaea (3.2%). When comparing to the NCBI database, to which 58.5% of the clean reads could be annotated on kingdom level, the percentage of annotated reads mapped to Bacteria was 98.4% and to Eukaryota 0.5%, of which 0.2% was mapped

to fungi. To Archaea 1.1% of all annotated reads was annotated and to Viruses 0.03%. Figure 19-A shows the microbial phyla and families to which most of the reads were assigned. The Proteobacteria (19.8 % of all metagenome reads) and Actinobacteria (8.5 % of all metagenome reads) were most abundant. Further prevalent bacterial phyla were Bacteroidetes (1.5%), Acidobacteria (2.8%), Verrucomicrobia (1.5%), Gemmatimonadetes (0.8%), Planctomycetes (1.6%) and Chloroflexi (0.7%). The 35 most abundant microbial families (Figure 19-B) were the Planctomycetaceae, the Acidobacteriaceae and Solibacteraceae from the Acidobacteria, the Gemmatimonadaceae, the Mycobacteriaceae and other families from the Actinobacteria, the Bradyrhizobiaceae and other families from the Proteobacteria, two families from the Verrucomicrobia, Nitrospiraceae and the Nitrososphaeraceae from the Thaumarchaeota. Altogether, they accounted for 13.0% of all annotated reads.

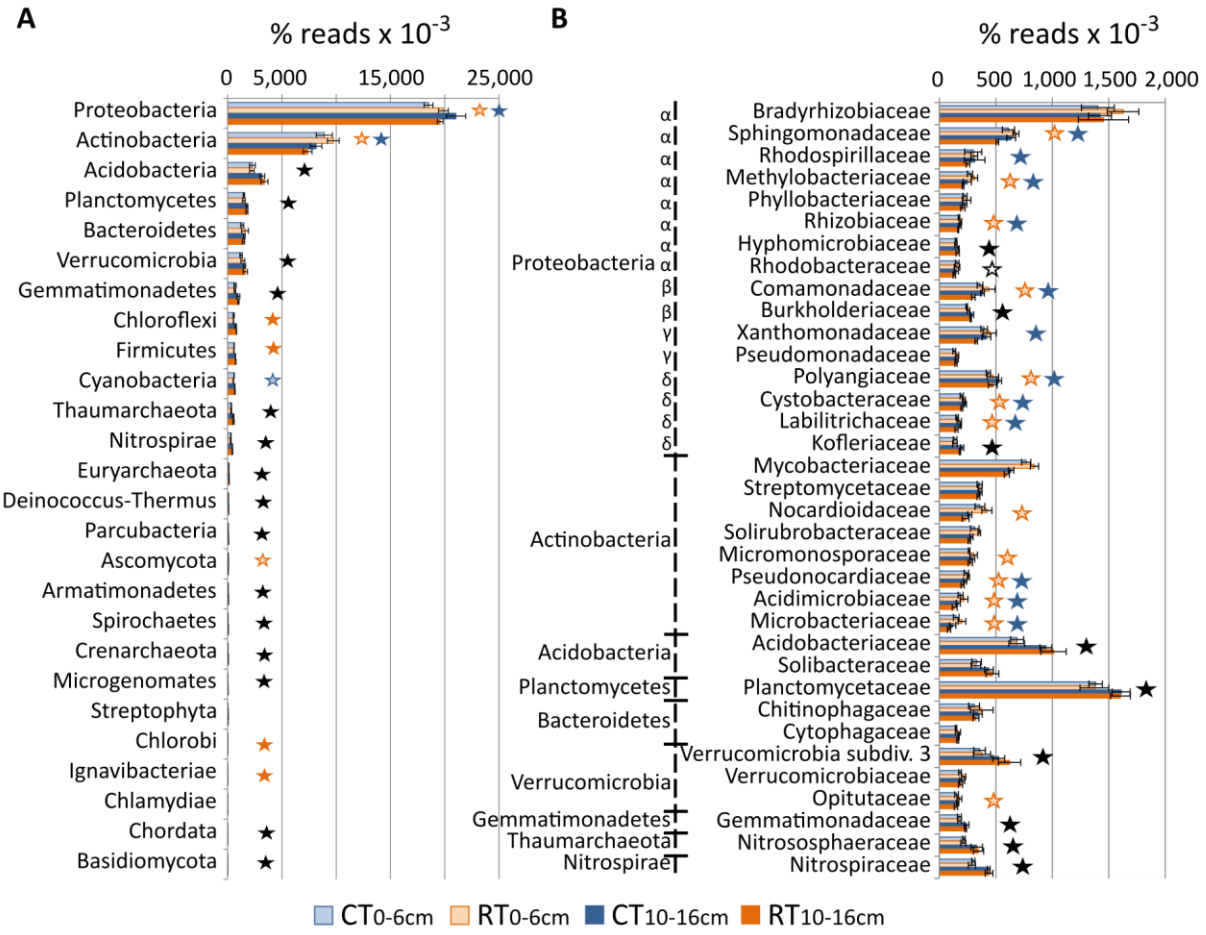


Figure 19: Taxonomic analysis of metagenomes of soil under conventional (CT) and reduced tillage (RT) in the top 6 cm of soil (the surface soil layer) or the top 10-16 cm of soil (the deeper soil layer). Shown are the percentage of reads x10^{-3} of the thirty-five most abundant phyla (A) and families (B), according to the NCBI non-redundant protein database. Stars indicate a significant difference in abundance between treatments (P<0.05); light blue: higher abundance under CT in the surface soil layer, light orange: higher abundance under RT in the surface soil layer, dark blue: higher abundance under CT in the deeper soil layer, dark orange: higher abundance under RT in the deeper soil layer, black: higher abundance in the deeper soil layer than the surface soil layer, white: higher abundance in the surface soil layer than the deeper soil layer.

Statistical analysis of the assessed microbial groups and protein-coding genes resulted in the detection of several significant differences between treatments. These results (relative abundances per treatment, total relative abundance and *P*-values of the used statistical test) are given in Table A8 in the appendix. All of the most abundant microbial phyla were affected by the treatments, except Bacteroidetes, Streptophyta and Chlamydiae. Most phyla were affected by depth, showing a higher abundance in the deeper soil layer than in the surface soil layer. An effect of tillage however was observed on the abundance of Proteobacteria, Actinobacteria, Chloroflexi, Firmicutes, Cyanobacteria, Ascomycota, Chlorobi and Ignavibacteria, depending on the soil layer (for effects, see Figure 19). Tillage effects were more pronounced on the lower taxonomic level. As was visible on phylum level, most of the highly abundant actinobacterial and proteobacterial families were affected by tillage with depth-dependence. Likewise, families from Acidobacteria, Planctomycetes, Verrucomicrobia, Gemmatimonadetes, Thaumarchaeota and Nitrospira were affected by depth, with the exception of the family Opitutaceae which was affected by tillage with depth-dependence. The microbial families affected by depth all showed a higher abundance in the deeper soil layer than in the surface soil layer, except for the Rhodobacteraceae (Proteobacteria). Families belonging to the Actinobacteria (Nocardioidaceae, Micromonosporaceae, Pseudonocardiaceae, Acidimicrobiaceae and Microbacteriaceae) and the Proteobacteria generally showed the same abundance pattern as on phylum level, with the highest abundance under RT in the surface soil layer and the lowest abundance in the deeper soil layer under RT.

Functional analysis

Functional annotation of the metagenome reads was done by comparing them to sequences in the KEGG-database. 22.37% of all reads was significantly similar to a sequence in the KEGG-database. Of these reads, more than half were mapped to general metabolism pathways of the KEGG database. Within metabolism, most reads were annotated to carbohydrate metabolism (pyruvate metabolism and glycolysis/gluconeogenesis), amino acid metabolism (arginine and proline metabolism, more to amino acid degradation than biosynthesis), and energy metabolism (carbon fixation pathways in prokaryotes, nitrogen metabolism and oxidative phosphorylation), see Figure 20. Other large groups of reads were annotated to genetic- and environmental information processing-pathways. Within genetic information processing, most reads were annotated to translation (aminoacyl-tRNA- and ribosome biosynthesis), folding, sorting and degradation (RNA degradation and protein export) and replication and repair (nucleotide excision repair and mismatch repair). Within environmental information processing, most reads were annotated to membrane transport (ABC transporters and bacterial secretion system) and signal transduction (two-component system). Figure 20 shows that a higher number of KEGG-pathways two which most of reads mapped were influenced by tillage with a depth-dependence than the number of microbial taxa, which showed stronger responses to soil depth. These KEGG-pathways generally showed a higher abundance under RT than under CT in the surface soil layer, whereas they were more abundant under CT than under RT in the deeper soil layer, with a few exceptions. While most pathways were more abundant in the deeper soil layer than in the surface soil layer, this effect was often not significant or was dependent on tillage treatment. However, a higher amount of reads in the deeper soil layer were annotated to methane metabolism and the pentose phosphate pathway than in the surface soil layer.

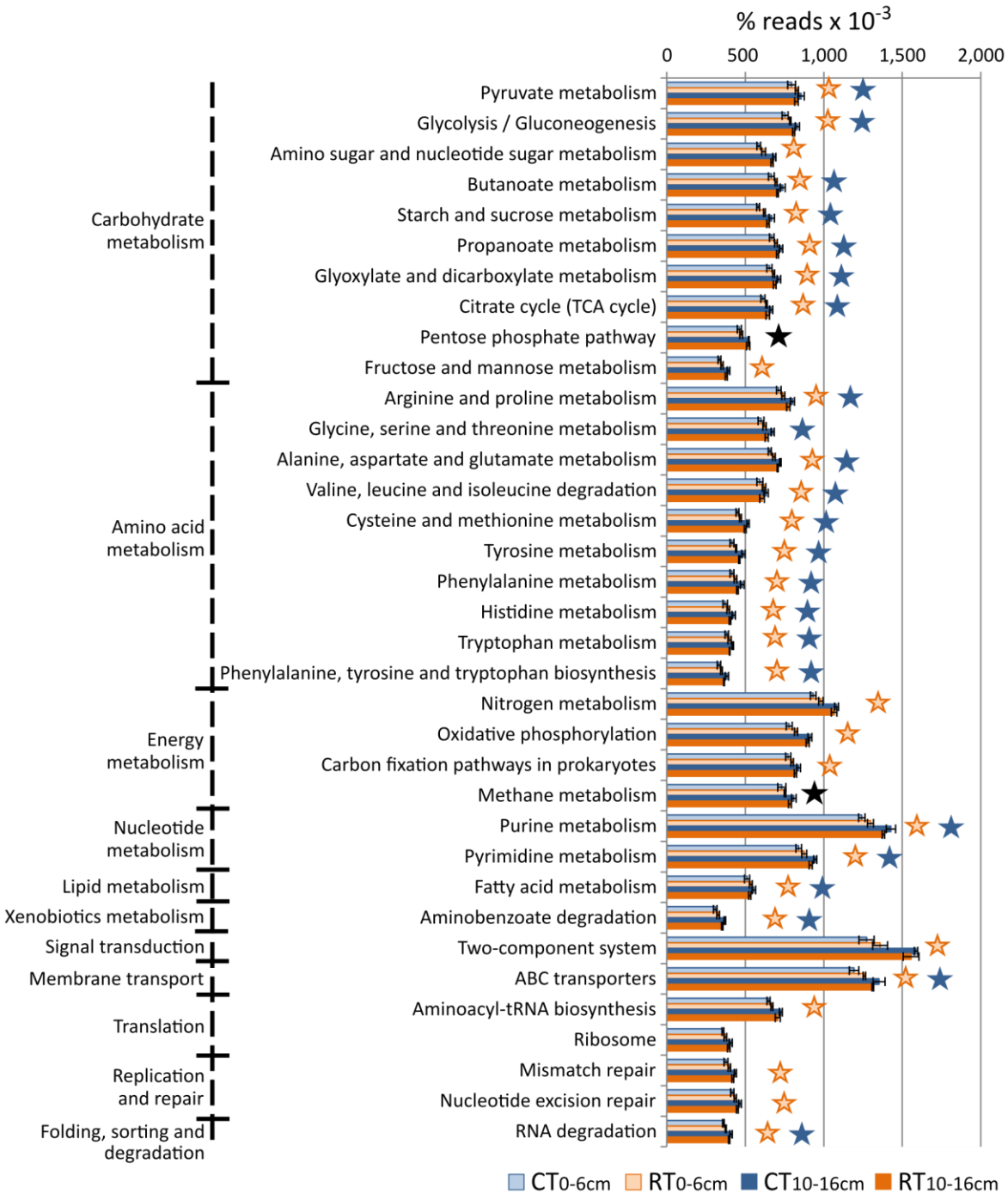


Figure 20: Functional analysis of metagenomes of soil under conventional (CT) and reduced tillage (RT) in the top 6 cm of soil (the surface soil layer) or the top 10-16 cm of soil (the deeper soil layer). Shown are the percentages of reads x10^{-3} of the thirty-five most abundant KEGG Level4-pathways. Stars indicate a significant difference in abundance between treatments; light blue: higher abundance under CT in the surface soil layer, light orange: higher abundance under RT in the surface soil layer, dark blue: higher abundance under CT in the deeper soil layer, dark orange: higher abundance under RT in the deeper soil layer, black: higher abundance in the deeper soil layer than the surface soil layer, white: higher abundance in the surface soil layer than the deeper soil layer.

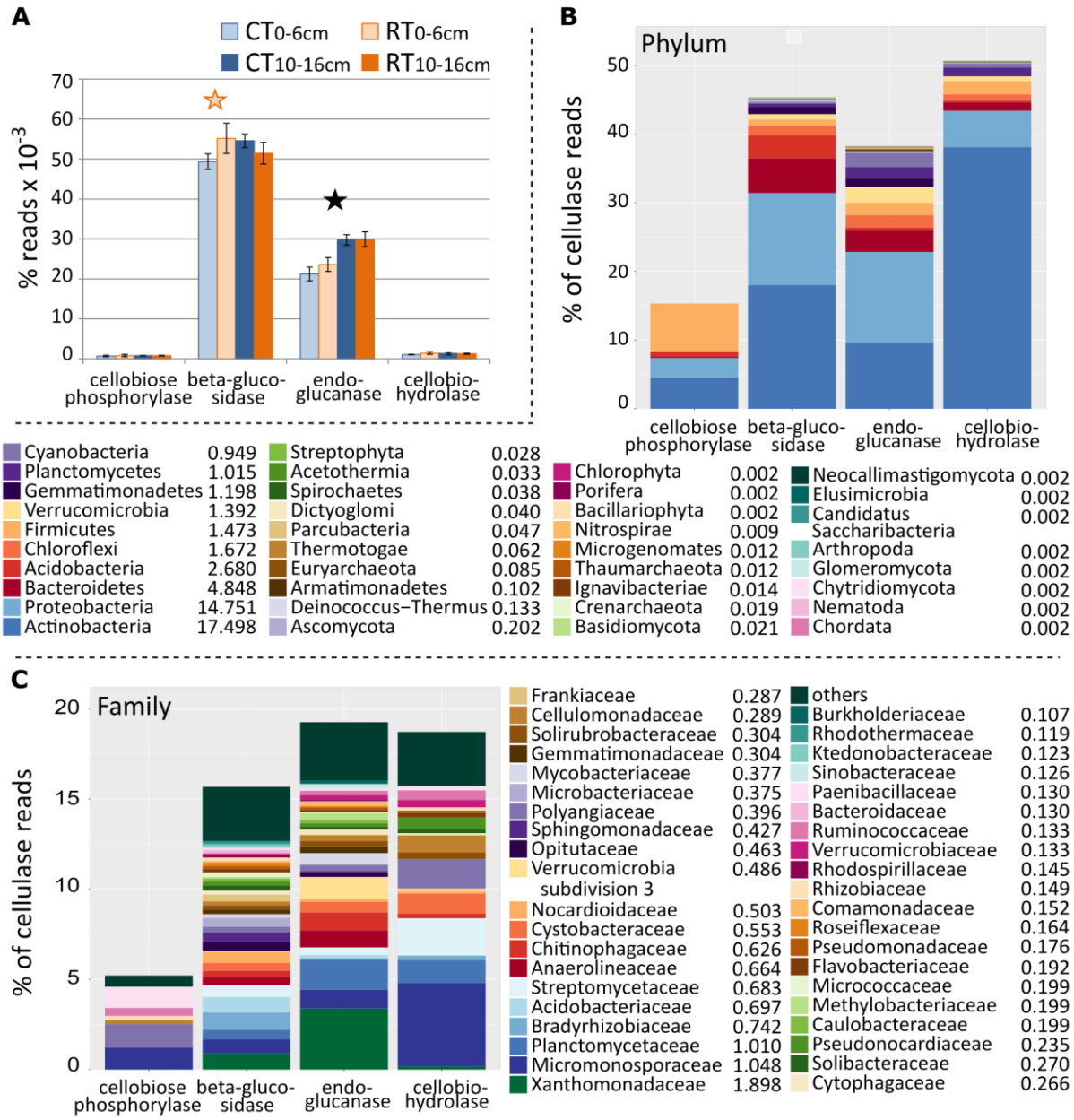


Figure 21: (A) Abundance of cellulase enzymatic groups in metagenomes of soil under conventional (CT) or reduced tillage (RT) in the top 6 cm of soil (the surface soil layer) or the top 10-16 cm of soil (the deeper soil layer). Shown is the percentage of reads x10-3 annotated to cellulase enzymatic functions for each treatment. Stars indicate a significant difference in abundance between treatments; light blue: higher abundance under CT in the surface soil layer, light orange: higher abundance under RT in the surface soil layer, dark blue: higher abundance under CT in the deeper soil layer, dark orange: higher abundance under RT in the deeper soil layer, black: higher abundance in the deeper soil layer than the surface soil layer, white: higher abundance in the surface soil layer than the deeper soil layer. Furthermore, the taxonomic composition of reads annotated to cellulase enzymatic groups on phylum **(B)** and family **(C)** level is shown**.** The reads annotated to each cellulase enzymatic group were pooled over all samples and taxonomically assigned. Shown are the forty most abundant taxa, as percentage of reads annotated to the cellulase enzymatic group. The taxonomically unassigned reads are not shown but represented by the rest of the bar until 100%. In addition, the contribution of each microbial group to the cellulase degradation potential is given behind the name of the microbial group as percentage of all reads annotated as cellulase enzymatic groups.

Cellulase enzymatic groups and taxonomic assignment

The annotations to cellulase enzymatic groups using KEGG-Orthology groups were analysed by comparison of metagenomic reads to cellulase enzymatic groups (EC-numbers: 2.4.1.20 (cellobiose phosphorylase), 3.2.1.21 (β-glucosidase), 3.2.1.4 (endoglucanase), 3.2.1.74 (cellodextrinase), 3.2.1.91 (cellobiohydrolase (non-reducing end)) and 1.1.99.18 (cellobiose dehydrogenase))) in the KEGG-database. This resulted in annotation of in total 42,160 reads (0.081% of all metagenome reads). Unfortunately, no reads were annotated to reducing-end-acting cellobiohydrolase (EC: 3.2.1.176), cellobiose dehydrogenase (EC: 1.1.99.18) or cellobionic acid phosphorylase (EC: 2.4.1.321). Furthermore, no orthologous groups representing cellodextrinase (EC: 3.2.1.74) or cellodextrin phosphorylase (EC: 2.4.1.49) existed in the KEGG-database at the time of analysis. Similar to results in the conventional farming metagenome, annotation to available cellulase enzymatic groups showed that the majority of cellulases in this soil were β-glucosidases (see Figure 21). These showed a higher abundance under RT than under CT but only in the surface soil layer. The relative abundance of endoglucanases, however, was not affected by tillage but was generally higher in the deeper soil layer than the surface soil layer.

Furthermore, the taxonomic affiliations of sequences annotated to cellulase enzymatic groups are shown in Figure 21. Fungi make up only a small part of the cellulolytic microorganisms here (0.23% of cellulase reads). They are represented by Ascomycota, Basidiomycota and Chitridiomycota harbouring β-glucosidases, Ascomycota and Glomeromycota harbouring endoglucanases and Neocallimastigomycota harbouring exoglucanases. Only a relatively small portion of the cellobiose phosphorylases could be taxonomically assigned on phylum level. As was seen in the metagenome of the conventional farming experiment, the potential to degrade cellulose using different mechanisms is widespread across different phyla and families. β-glucosidase genes were assigned to most microbial phyla and to many of these (Acidobacteria, Actinobacteria, Ascomycota, Chloroflexi, Deinococcus-Thermus, Dictyoglomi, Euryarchaeota, Firmicutes, Gemmatimonadetes, Ignavibacteria, Bacteroidetes, Proteobacteria, Streptophyta and Thermotogae) also endoglucanase genes were assigned. Sometimes, the number of endoglucanase sequences derived from a certain phylum was higher than the number of β-glucosidase genes derived from the same phylum (e.g. Cyanobacteria, Planctomycetes and Verrucomicrobia). In addition, certain microbial phyla or families dominated among the taxonomically assigned sequences in each enzymatic group, illustrating a connection between taxonomy and type of enzymatic activity; cellobiose phosphorylase (EC: 2.4.1.20) genes were mostly assigned to Paenibacillaceae and Ruminococcaceae (Firmicutes), Polyangiaceae (Deltaproteobacteria) and Micromonosporaceae (Actinobacteria). No tillage effect could be observed on the relative abundance of the cellobiose phosphorylase genes harboured by these microbial families. β-glucosidase genes were harboured by many microorganisms, but most were assigned to families from the Proteobacteria (e.g. Xanthomonodaceae and Bradyrhizobiaceae), Actinobacteria (Micromonosporaceae and Streptomycetaceae) and Acidobacteria (Acidobacteriaceae) (see Figures 21 and 22). β-glucosidase genes derived from Proteobacteria were more abundant under CT than under RT in the deeper soil layer, whereas those derived from Firmicutes were more abundant under RT than under CT in the surface soil layer (see Figure 22).

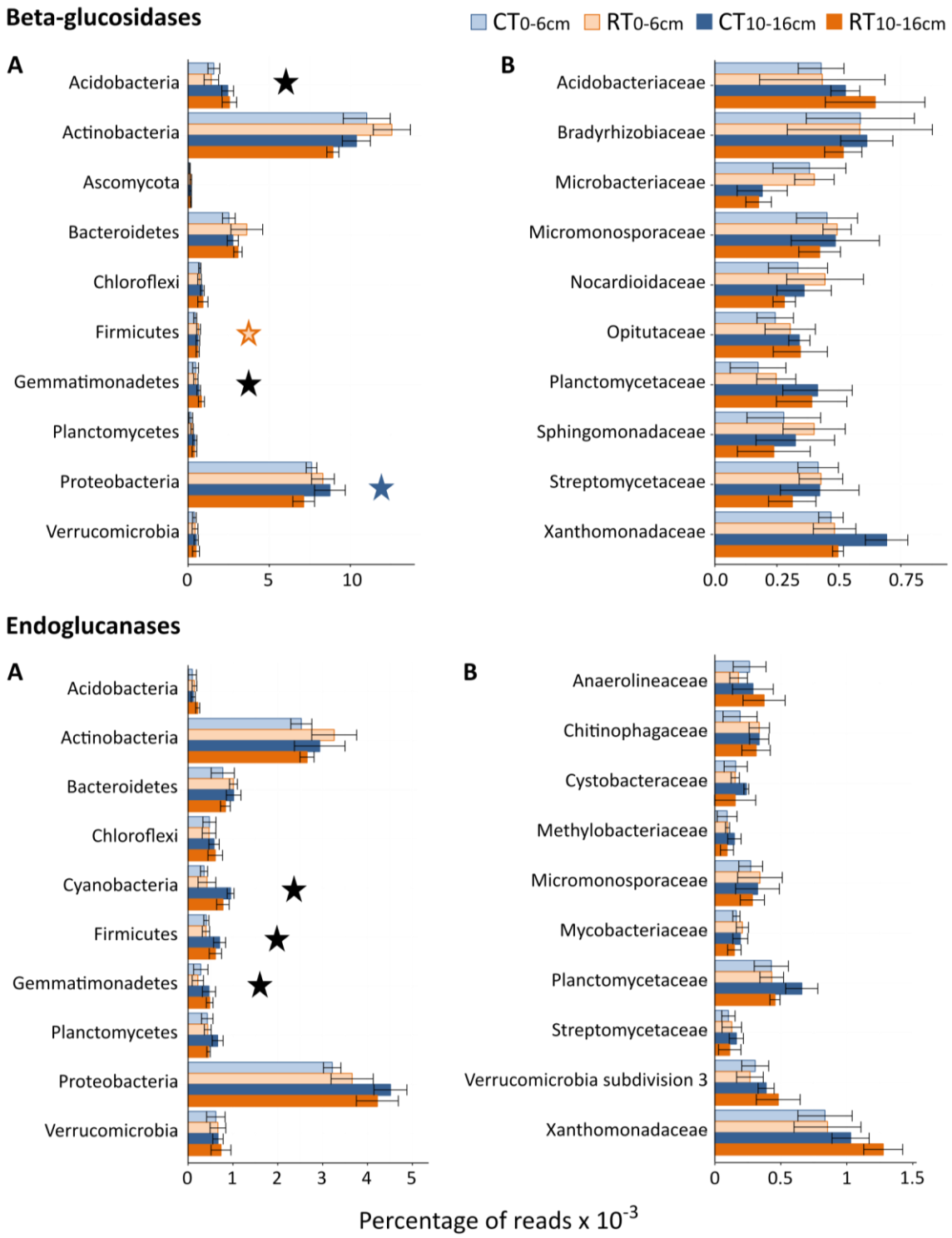


Figure 22: Taxonomic assignment of the reads annotated to the two most abundant cellulase enzymatic groups on phylum (A) and family (B) level under conventional (CT) or reduced (RT) tillage in the surface soil layer (0-6 cm) or the deeper soil layer (10-16 cm). The reads annotated to the cellulase functions in each sample were taxonomically assigned and treatment averages were calculated. For each function, the percentage of metagenome reads x10^{-3} annotated to the ten most abundant taxa is shown. Stars indicate a significant difference in abundance between treatments (P<0.05); light blue: higher abundance under CT in the surface soil layer, light orange: higher abundance under RT in the surface soil layer, dark blue: higher abundance under CT in the deeper soil layer, dark orange: higher abundance under RT in the deeper soil layer, black: higher abundance in the deeper soil layer than the surface soil layer, white: higher abundance in the surface soil layer than the deeper soil layer.

In addition, Figure 22 shows that, on the lower taxonomic level, the β-glucosidase genes harboured by Xanthomonadaceae tended towards a higher abundance under CT than under RT in the deeper soil layer (not significant after *P*-value correction). Furthermore, β-glucosidase genes harboured by the Sphingomonadaceae (Alphaproteobacteria) and Acidimicrobiaceae (Actinobacteria) tended towards a higher abundance under RT than under CT in the surface soil layer (not significant after *P*-value correction). The majority of endoglucanase genes were assigned to Xanthomonadaceae (Gammaproteobacteria) and families from the Actinobacteria (Figures 21 and 22). However, the observed depth effect on the abundance of the endoglucanase genes was likely due to those genes derived from Cyanobacteria, Firmicutes and Gemmatimonadetes, which also showed a higher abundance in the deeper soil layer than in the surface soil layer (see Figure 22). Exoglucanase genes were assigned mostly to Micromonosporaceae and Streptomycetaceae (Actinobacteria). Whereas exoglucanase genes harboured by Planctomycetaceae were more abundant in the deeper soil layer than the surface soil layer, no tillage effect was observed on the abundance of the exoglucanase genes assigned to microbial taxa.

Annotation method of cellulase domain families

As discussed in the results of the conventional farming experiment, metagenome reads were additionally annotated to cellulase catalytic and binding domain families. This was done using benchmarked HMMs (see Table A1 in the appendix for the results of the benchmarking). Using this method, 225,849 reads (0.43 % of metagenome reads) could be annotated to the selected domain families. Afterwards, the HMM-annotated reads were additionally filtered by screening them against a database containing only cellulase sequences (see section 2: Materials and Methods). The result of this filtering step for the metagenome of the organic farming experiment is almost identical to that for the metagenome of the conventional farming experiment and is shown in Figure 23.

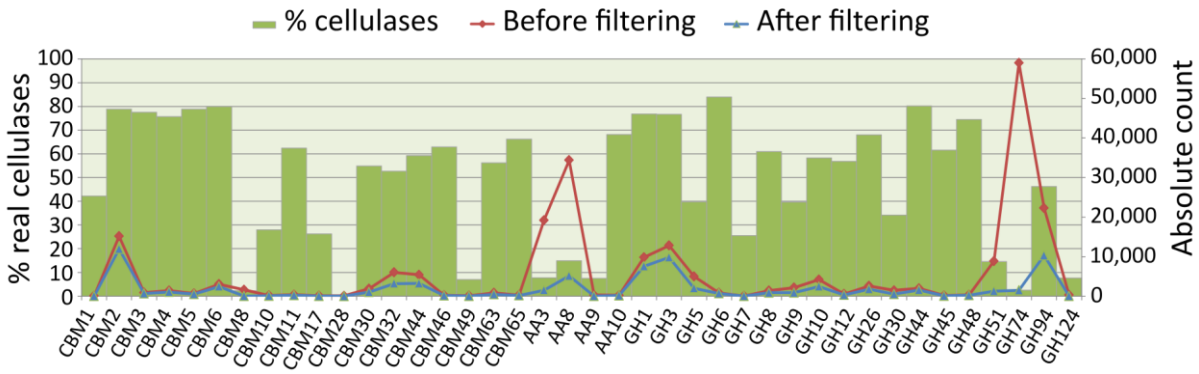


Figure 23: Results of the filtering step of HMM-annotated reads in the metagenome of the organic farming experiment to cellulase catalytic and binding domain families. Shown is the percentage of HMM-annotated reads with a significant BLAST hit to a sequence in the positive cellulase sequences-database (bars). In addition, the absolute amount of reads annotated to cellulase catalytic and binding domain families before and after the filtering step is shown.

Cellulase domain families

A final amount of 73,694 reads (0.14 % of all metagenome reads) was annotated to cellulase catalytic and binding domain-families. Analysis of variance in the treatment means of read abundances annotated to cellulase domain families showed that (Figure 24) five cellulase domain families were more abundant under RT than under CT in the surface soil layer (CBM3, CBM4, CBM6,

GH1 and GH26). GH1 and GH26 contain essentially β-glucosidases respectively bifunctional endoglucanases (see Figure 2). The cellulase domain family AA8 was likewise more abundant under RT than under CT in the surface soil layer, but this effect was not significant after *P*-value correction. At the same time, CBM4 and CBM65 were more abundant under CT than under RT in the deeper soil layer. Furthermore, GH1-genes were also more abundant under CT than under RT in the deeper soil layer, but this effect was not significant after *P*-value correction. The total amount of reads annotated to all cellulase domain families was higher under RT than CT, but only in the surface soil layer, as was the total amount of reads annotated to all catalytic families (Auxiliary Activity- and Glycoside Hydrolase families). However, the total amount of reads annotated to the CBMs was only affected by soil depth (more abundant in the deeper soil layer than the surface soil layer) and not by tillage. In fact, 40% of the modules had a higher abundance in the deeper soil layer than the surface soil layer (see Figure 24). Among these were the most abundant CBMs (CBM2, CBM32 and CBM44) and catalytic modules (GH3 and GH94).

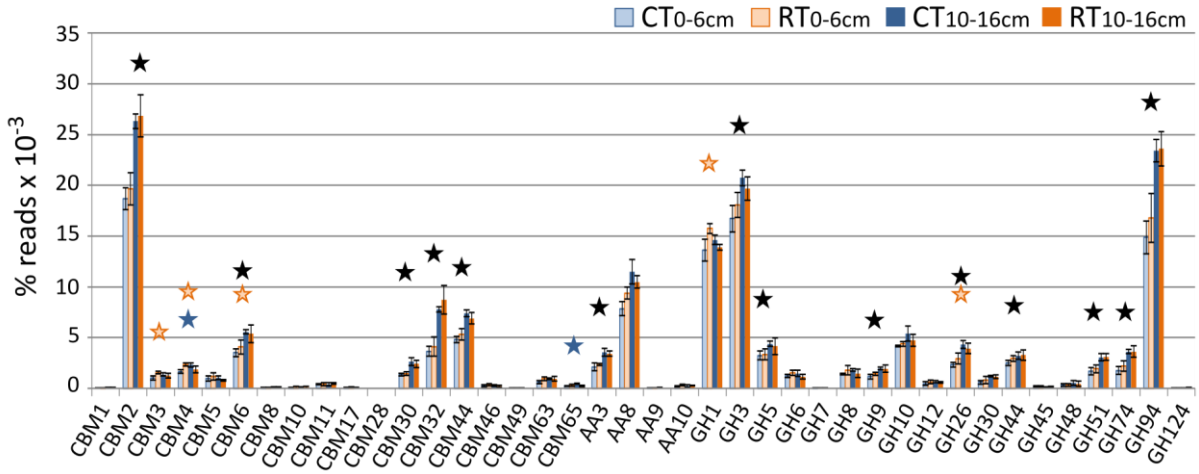


Figure 24: Abundance of cellulase domain families in metagenomes of soil under conventional (CT) or reduced tillage (RT) in the top 6 cm of soil (the surface soil layer) or the top 10-16 cm of soil (the deeper soil layer). Shown is the percentage of reads x10^{-3} annotated to cellulase-catalytic domains (AA & GH) and CBMs for each treatment. Annotation was done by scanning the reads with HMMs and subsequent BLASTing against a positive cellulase sequence database. Stars indicate a significant difference in abundance between treatments; light blue: higher abundance under CT in the surface soil layer, light orange: higher abundance under RT in the surface soil layer, dark blue: higher abundance under CT in the deeper soil layer, dark orange: higher abundance under RT in the deeper soil layer, black: higher abundance in the deeper soil layer than the surface soil layer, white: higher abundance in the surface soil layer than the deeper soil layer.

Taxonomic assignment of cellulase domain families

The taxonomic affiliation of the sequences annotated to cellulase domain families, shown in Figure 25 on phylum and family level, reflects the overall abundance of phyla in the metagenome. However, the abundance of sequences harboured by Acidobacteria and Planctomycetes is slightly lower (1.7% vs. 2.8% respectively 0.6% vs. 1.6%) and the abundance of the Actinobacteria, Bacteroidetes and especially Cyanobacteria is higher (12.8% vs. 8.5%, 3.5% vs. 1.5% respectively 1.7% vs. 0.6%) among the cellulase domain families than in the whole metagenome dataset. 0.18 % of the sequences annotated to cellulase domain families mapped to fungi (Ascomycota, Basidiomycota, Neocallimastigomycota and Glomeromycota).

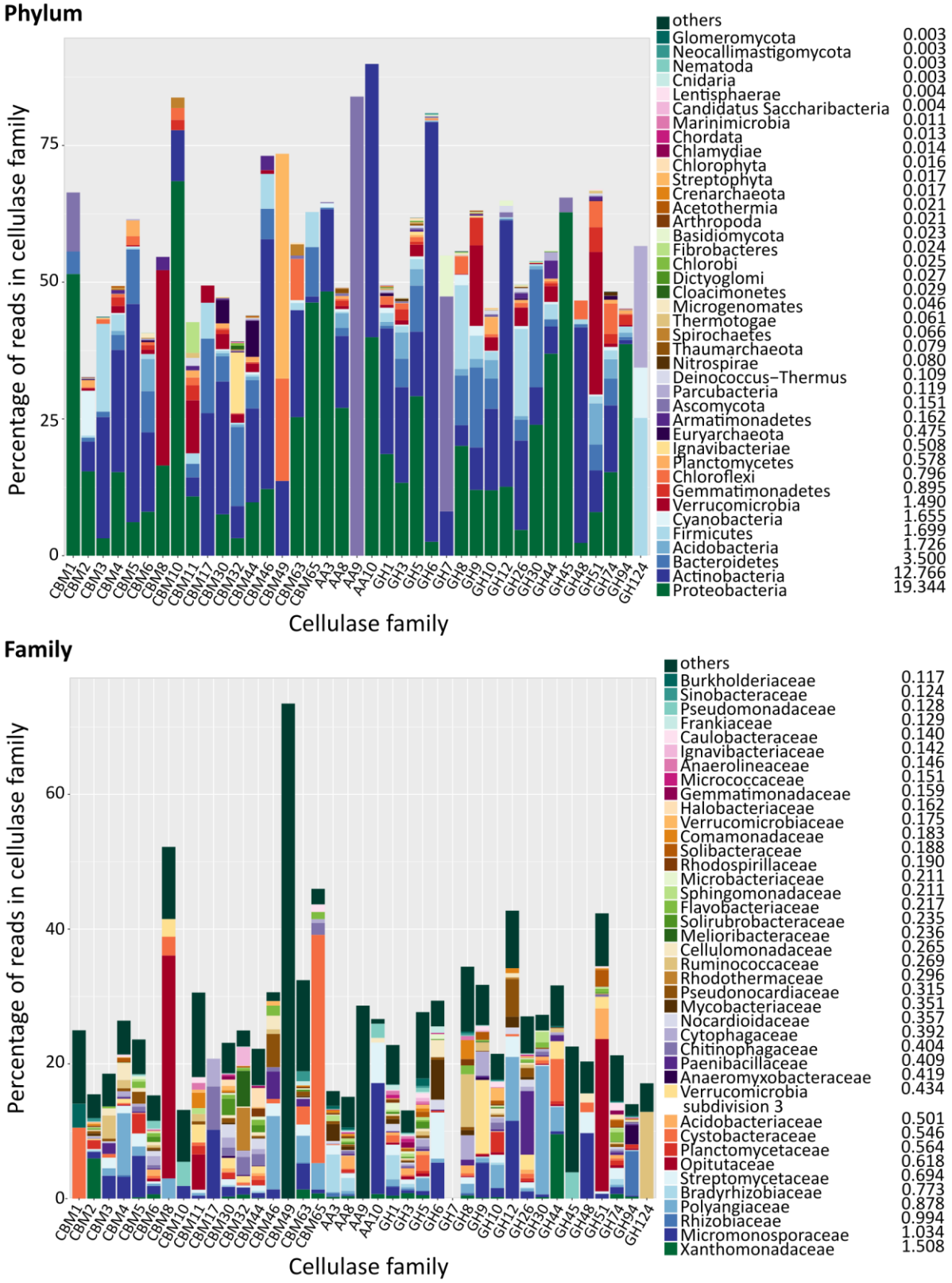


Figure 25: Taxonomic assignment of cellulase domain family-reads on phylum and family level. The reads annotated to cellulase domain families were pooled over all samples and taxonomically assigned. Shown is the percentage of reads in the cellulase domain family annotated to the most abundant microbial phyla or families. The taxonomically unassigned reads are not shown but represented by the rest of the bar until 100%. In addition, the contribution of each taxonomic group to the cellulase degradation potential is given as percentage of all cellulase domain family-reads.

The predicted cellulase genes were mostly derived from the Xanthomonadaceae (Gammaproteobacteria) followed by the Micromonosporaceae (Actinobacteria), the Rhizobiaceae (Alphaproteobacteria), the Polyangiaceae (Deltaproteobacteria) and the Bradyrhizobiaceae (Alphaproteobacteria) (see Figure 25). Moreover, cellulase genes from each of the most abundant domain families were derived from these microbial families. In addition, although the potential to degrade cellulose is spread widely across the different phyla and families, some patterns can be observed connecting microbial groups with certain cellulase domain families. In Table 2 the microbial groups are listed to which a higher abundance of assigned reads can be found in specific cellulase domain families compared to other cellulase domain families. These observations might indicate possible specializations of these microbial groups regarding type of cellulases used in degradation. Affiliations to fungi were found in low abundance but in many domain families for Ascomycota (GH1, 3, 5, 6, 7, 9, 10, 12, 45, 46, 51 and 74, CBM1, 6 and 63 and AA3, 8 and 9) and Basidiomycota (GH1, 3, 5, 6, 7, 9, 10, 12, 44 and 51). However, genes harboured by Glomeromycota were only encountered in cellulase domain families CBM3 and GH1, while sequences assigned to Neocallimastigomycota were only found in GH6.

Table 2: Microbial taxonomic groups to which a higher abundance of sequences was assigned in the respective cellulase domain family than in other domain families.

Microbial taxonomic group		
Phylum level	**Family level**	**Cellulase domain family**
Acidobacteria	Acidobacteriaceae	AA8 and GH3
Actinobacteria (Actinomycetales)	Frankiaceae	GH1
	Microbacteriaceae	GH1
	Mycobacteriaceae	GH1
	Nocardioidaceae	GH1
	Streptomycetaceae	GH1
Bacteroidetes	Rhodothermaceae	CBM32
	Chitinophagaceae	CBM32
Cyanobacteria	Nostocaceae	CBM2
Euryarchaeota	Halobacteriaceae	CBM44
Firmicutes	Paenibacillaceae	GH26
	Ruminococcaceae	GH8
Gemmatimonadetes	Gemmatimonadaceae	GH3
	not identified	GH94
Ignavibacteriae	Ignavibacteriaceae	CBM32
	Melioribacteraceae	CBM32
Planctomycetes	Planctomycetaceae	CBM2
Proteobacteria	Bradyrhizobiaceae	GH1
	Xanthomonadaceae	CBM2
	Sinobacteraceae	CBM2
	Cystobacteraceae	GH44
	Rhizobiaceae	GH94
	Anaeromyxobacteraceae	GH94
Verrucomicrobia	Opitutaceae	GH51
	Verrucomicrobia subdivision 3	GH9

Conversely, Figure 25 also shows that sequences in some cellulase domain families were more harboured by certain microbial groups than to others, also indicating a predictive relationship between type of cellulase domain family found and the respective degrading organisms; the majority of GH6-genes is assigned to Actinobacteria (Streptomycetaceae), while a considerably high abundance of GH94- (Anaeromyxobacteraceae & Rhizobiaceae), GH44- (Xanthomonadaceae), AA8- (Bradyrhizobiaceae), AA3- (Bradyrhizobiaceae), CBM4- (Polyangiaceae) and CBM2- (Xanthomonadaceae) genes are derived from Proteobacteria (see also Figures 26 and 27). In addition, the majority of GH9- and GH51-genes are derived from Verrucomicrobia subdivision 3 respectively Opitutaceae (Verrucomicrobia), while the majority of GH8- and GH26-genes is assigned to Ruminococcaceae respectively Paenibacillaceae (Firmicutes). Moreover, most CBM32-genes are taxonomically assigned to Melioribacteraceae (Ignavibacteria) and Rhodothermaceae (Bacteroidetes). For low-abundant cellulase domain families, it could be observed that sequences in CBM8 were mostly assigned to members of the Verrucomicrobia, while sequences in domain families AA10, CBM17, CBM3, CBM30, CBM46, CBM5, GH12 and GH48 are mostly harboured by Actinobacteria. Also, the majority of ORFs in CBM1, CBM10 and GH45 are assigned to Proteobacteria, while those in GH124 are, in contrast to many other domain families, not harboured by Proteobacteria, but instead to Firmicutes, Cyanobacteria and Parcubacteria. Genes from GH7 were generally derived from Ascomycota, besides to Actinobacteria (Actinomycetales) and Basidiomycota, while genes in AA9 were exclusively assigned to members of the Ascomycota (Nectriaceae, Lasiosphaeriaceae, Sordariaceae, Chaetomiaceae and Pleosporaceae). Finally, sequences with a CBM49 are found in majority derived from Streptophyta.

The taxonomic affiliation of sequences from the most abundant cellulase domain families (GH1, GH3, GH94, AA8, CBM2 and CBM6) is shown separately in Figure 26 on phylum level and in Figure 27 on family level. Several treatment effects on the abundances of these taxonomic groups can be observed; microbial phyla to which genes from CBM2-, CBM6-, AA8-, GH3- and GH94 (Figure 26) are assigned are more abundant in the deeper soil layer than in the surface soil layer, reflecting the abundance-pattern of the corresponding cellulase domain families. In addition, tillage effects were observed for some microbial phyla; Actinobacteria to which sequences from GH1 were assigned showed a higher abundance under RT than under CT in the surface soil layer. The same trend was visible in the abundances of AA8-genes assigned to Proteobacteria, although this effect was not significant anymore after *P*-value correction. Furthermore, AA8- and GH1-genes assigned to Planctomycetes respectively Verrucomicrobia were more abundant under CT than under RT in the deeper soil layer. Interestingly, GH3- and GH94-cellulase genes harboured by Bacteroidetes showed a higher abundance under RT than under CT in both soil layers. No tillage effects were observed on family level however (Figure 27). CBM2-cellulase genes harboured by Nostocaceae, Plantomycetaceae, Sinobacteraceae and Xanthomonadaceae and GH94-genes assigned to Anaeromyxobacteraceae were more abundant in the deeper soil layer than the surface soil layer. The same trend was visible for the GH94-cellulase genes assigned to Rhizobiaceae. In addition, GH1-cellulase genes derived from several microbial families (e.g. Microbacteriaceae, Nocardioidaceae and Polyangiaceae) showed a trend towards higher abundance under RT than under CT in the surface soil layer, while GH1-genes derived from Nocardioidaceae showed a trend towards higher abundance under CT than under RT in the deeper soil layer. Nevertheless, these effects were not significant anymore after *P*-value correction.

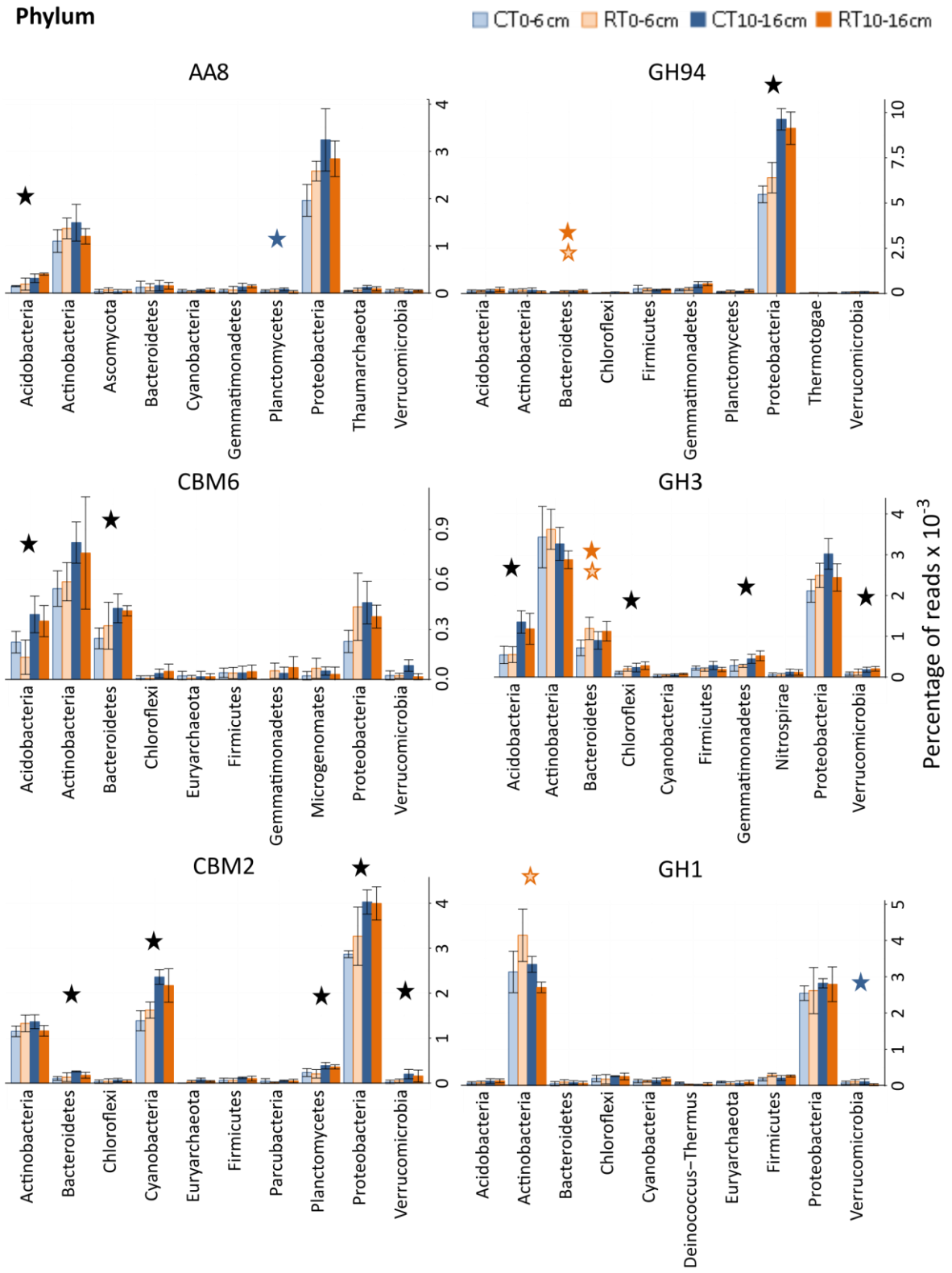


Figure 26: Taxonomic assignment of the six most abundant cellulase domain families on phylum level under conventional (CT) or reduced (RT) tillage in the top 6 cm of soil (the surface soil layer) or the top 10-16 cm of soil (the deeper soil layer). Stars indicate a significant difference in abundance between treatments ($P<0.05$); light blue: higher abundance under CT in the surface soil layer, light orange: higher abundance under RT in the surface soil layer, dark blue: higher abundance under CT in the deeper soil layer, dark orange: higher abundance under RT in the deeper soil layer, black: higher abundance in the deeper soil layer than the surface soil layer, white: higher abundance in the surface soil layer than the deeper soil layer.

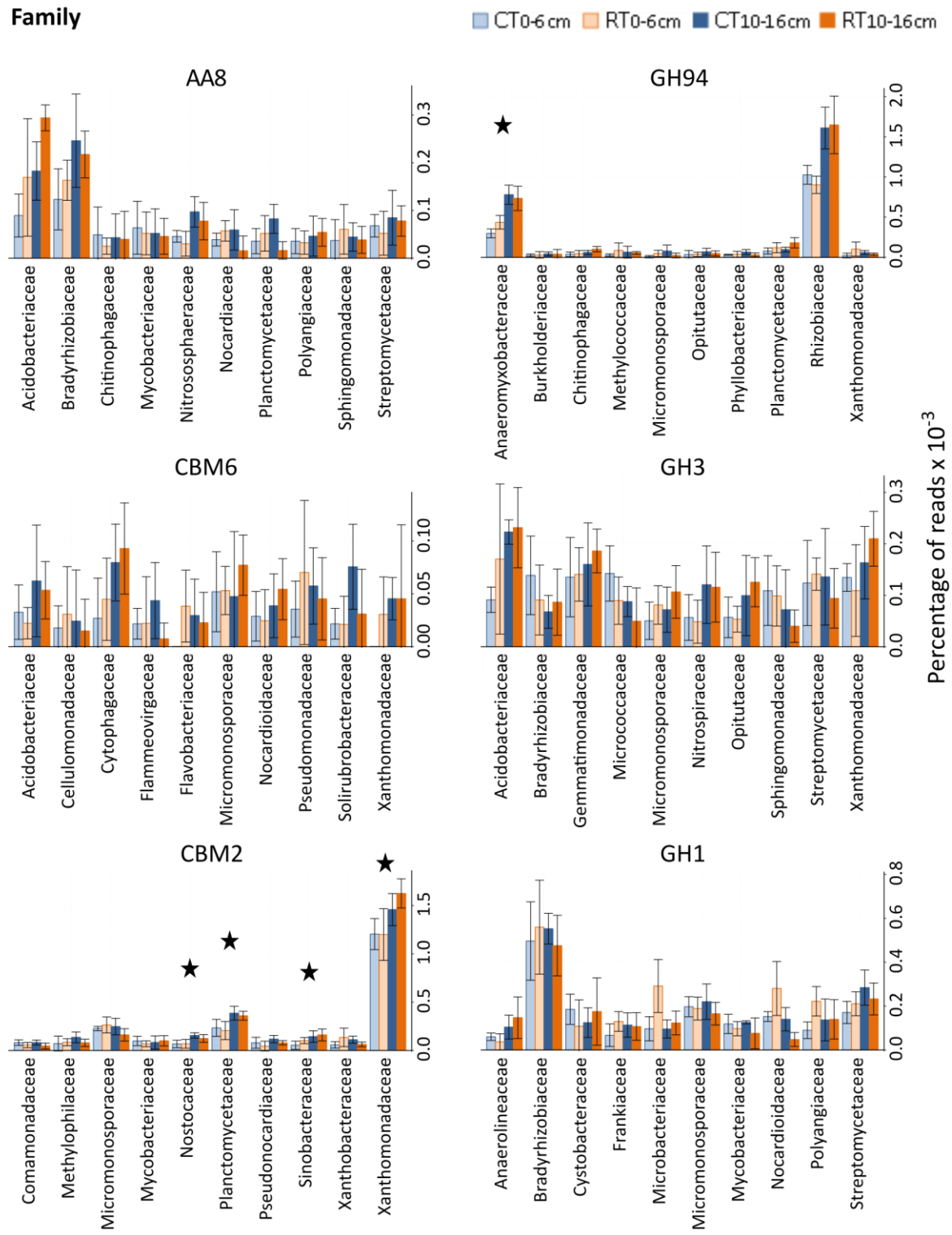


Figure 27: Taxonomic assignment of the six most abundant cellulase domain families on family level under conventional (CT) or reduced (RT) tillage in the top 6 cm of soil (the surface soil layer) or the top 10-16 cm of soil (the deeper soil layer). Stars indicate a significant difference in abundance between treatments (P<0.05); light blue: higher abundance under CT in the surface soil layer, light orange: higher abundance under RT in the surface soil layer, dark blue: higher abundance under CT in the deeper soil layer, dark orange: higher abundance under RT in the deeper soil layer, black: higher abundance in the deeper soil layer than the surface soil layer, white: higher abundance in the surface soil layer than the deeper soil layer.

Co-occurring microbial communities

Co-occurrence analysis of the microbial family abundances in each metagenome sample resulted in the identification of three communities of positively co-occurring families. Of the 917 families identified in the whole metagenome (consisting of 17.9% of all metagenome reads), 265 families (27%) could be assigned to a networking community. These 265 families in total made up 17.5% of all metagenome reads. 19 families were assigned to the smallest community (0.96% of all metagenome reads, red nodes in Figure 28-A), 80 families to the largest community (9.38% of all metagenome reads, green nodes in Figure 28-A) and 156 families to the intermediate community (7.20% of all metagenome reads, blue nodes in Figure 28-A).

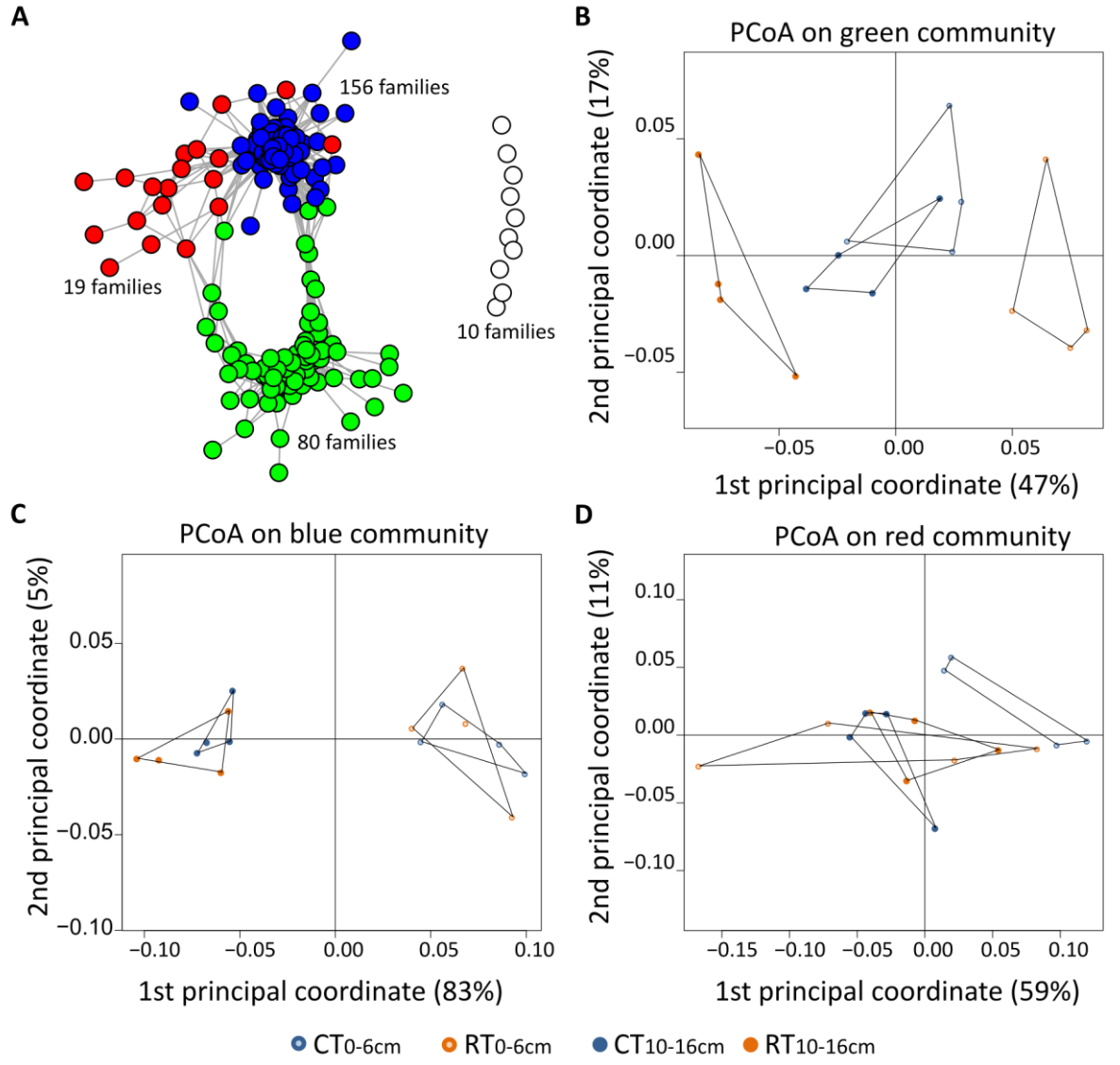


Figure 28: Positive co-occurrence network of three communities in soil metagenome and their response patterns to treatments. (A) Network depiction of the blue, green and red community of co-occurring families, where each sphere stands for one family. 10 families did not show any abundance co-occurrence with the other microbial families (white spheres). Multidimensional scaling of the abundances of the families within the green **(B)**, blue **(C)** and red **(D)** community over all treatments: conventional (CT) or reduced (RT) tillage, in the top 6 cm of soil (the surface soil layer) or the top 10-16 cm of soil (the deeper soil layer).

The blue and red communities consist of relatively lower abundant families, whereas the green community consists of families with high abundance. Some families of a community also co-occur with families of another community. For an overview of all families in each community, see Table A9 in the appendix.

Taxonomic characterisation of communities

An overview of the taxonomic assignments of the communities on phylum level is shown in Figure 29, as percentage of all community reads. The blue community is the most diverse, also on phylum level. The green community contains relatively less Viruses than the red and blue community, whereas the red community contains no microbial families of the Viridiplantae, as opposed to the blue and green community. The blue community, on the other hand, has more Archaea than the green and red communities. The blue community is characterized by, among other phyla, the Planctomycetes, Acidobacteria, Chloroflexi, Nitrospirae, Firmicutes and Thaumarchaeota. The green community can be characterized mainly by the Actinobacteria and Alphaproteobacteria, but also by a few low-abundant phyla like the Fibrobacteres and Chlorophyta. The red community is dominated by the Bacteroidetes (Sphingobacteriia, Cytophagia and Flavobacteriia) and is further characterized by the Cyanobacteria. The Gamma- and Deltaproteobacteria are present in all three communities.

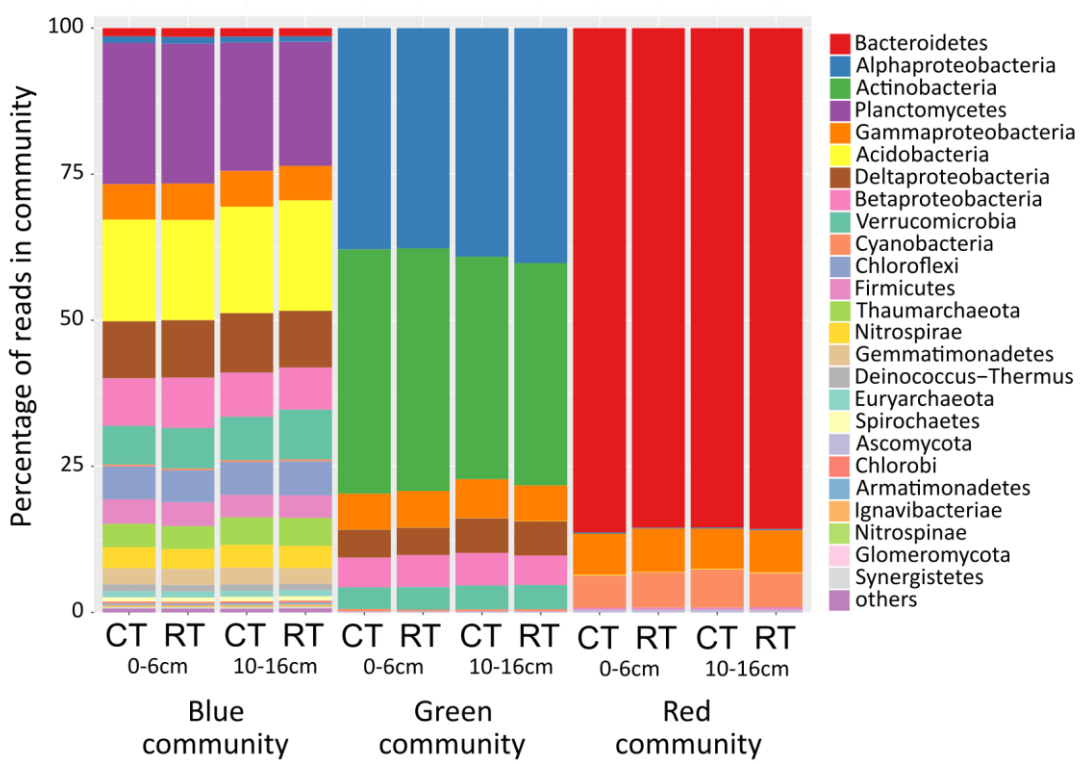


Figure 29: Taxonomic composition of network communities on phylum level. The percentage of reads annotated to the most abundant phyla in each community is shown per treatment: conventional (CT) or reduced (RT) tillage, in the top 6 cm of soil (the surface soil layer) or the top 10-16 cm of soil (the deeper soil layer).

When analysing the abundances of the community-families using multidimensional scaling, specific treatment-response patterns per community can be observed (Figure 28-B, -C and -D). The microbial families from the green community are separated according to different abundance patterns among

samples of different tillage treatments and depth layers (Figure 28-B); they generally had a higher abundance under RT than under CT in the surface soil layer and a lower abundance under RT than CT in the deeper soil layer (see abundances of e.g. Actinobacteria and Alphaproteobacteria in Figure 19). They mostly did not differ in abundance between depths under CT, but often show a depth-effect under RT. Therefore, the abundances of these microbial families appeared to be affected by the action of soil tillage. Although the effect is less evident for the red community-families, tillage treatment also appeared to influence their abundances between soil samples. These microbial families often showed a lower abundance under the CT treatment in the surface soil layer (Figure 28-D) than in all other treatment-depth combinations. Whereas this effect was not visible on the abundances of Bacteroidetes, a tillage effect in the surface soil layer was observed on the abundances of highly abundant families of the Ascomycota and Cyanobacteria belonging to this community (e.g. Trichocomaceae and Microchaetaceae). Conversely, the abundances of microbial families in the blue community were clearly dependent on soil depth (Figure 28-C), where most families showed a higher abundance in the deeper soil layer than in the surface soil layer. Indeed, microbial families from to the Planctomycetes, Acidobacteria and Thaumarchaeota show the same abundance response (see Figure 19).

Cellulose degradation potential within communities

The reads belonging to each community were further analysed for their cellulolytic properties. 8.5%, 15.6% and 2.0% $x10^{-3}$ of the metagenome reads were annotated to cellulase domain families in the blue, green respectively red community, showing that the majority of cellulase genes in the metagenome are harboured by microbial families of the green community. However, relative to the amount of reads belonging to each community, these abundances were 0.12%, 0.17% and 0.21% of all reads in the blue, green and red community, respectively. These results show that the microbial families in the red community are relatively more associated with cellulolytic functions than families in the other two communities. This can also be observed in Figure 30, where a summary of the relative abundance of cellulase domain families and cellulase enzymatic groups annotated to reads in each community is given. In this figure, the size of the groups of cellulase genes reflects the sum of their relative abundance within each community while their position reflects the contribution by each community to the sum. Most cellulase genes are found in all three communities, with some exceptions; the blue community is specifically characterized by the low-abundant GH124-family and by GH26, GH45 and CBM49. The green community, on the other hand, is characterized by the redox cellulase families AA3, AA9 and AA10, CBM8 and the hydrolase families GH1, GH6, GH12 and GH48. Of these, the GH1 domain family and GH1-genes with taxonomic affiliation to microbial families from the green community indeed often showed a higher abundance under RT than under CT in the surface soil layer (see Figure 24 and Figures 26 and 27). Interestingly, the abundant CBM32-family is not well represented in the green community. Conversely, the red community is more or less characterized by higher abundances of CBMs, especially CBM17 and CBM65, but also by CBM44, CBM32, CBM6 and CBM30. GH8 and GH9 are also relatively more abundant in the red community. Strikingly, several cellulase domain families (e.g. CBM8, GH48) or cellobiose phosphorylase (EC:2.4.1.20) are absent in this community. The cellulase domain families which are characteristic for the red community showed little response to tillage, but some were more abundant in the deeper soil layer than the surface soil layer (CBM6, CBM30, CBM32 and CBM44), and CBM65 shows higher

abundance under CT than under RT in the deeper soil layer, while CBM6 shows higher abundance under RT than under CT in the surface soil layer (Figure 24).

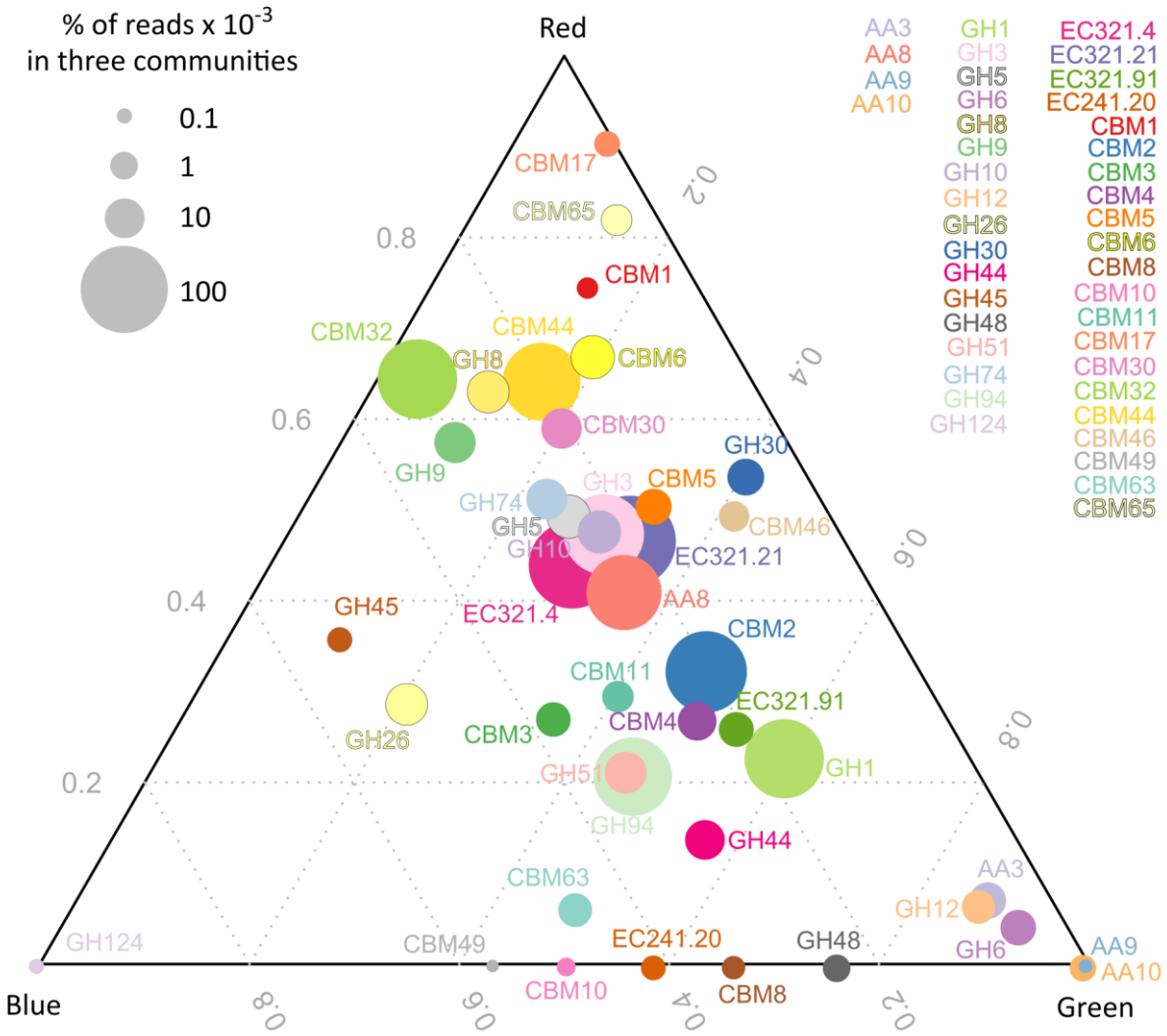


Figure 30: Contribution of each community to the abundance of cellulase domain families or cellulase enzymatic groups in the soil metagenome. Shown is the relative abundance of the cellulase domain families or enzymatic groups as percentage of community reads x10^{-3}. The size of the sphere represents the sum of the relative abundances over all three communities and the location of the sphere in the plot marks the relative contribution of each community to this abundance.

Specialized cellulolytic microorganisms reacting to tillage

The co-occurrence analysis further revealed that the Micromonosporaceae (Actinobacteria), which are more abundant under RT than CT in the surface soil layer (see Figure 19) and belong to the green community, co-occurred positively with two β-glucosidase genes (K05349 & K05350). Total abundances in the whole metagenome of both β-glucosidase genes showed the same response to treatment as the Micromonosporaceae. When analyzing the abundances of β-glucosidase genes harboured by microbial families of the green community, one of the genes (K05350) showed a higher abundance under RT than CT in the surface soil layer. Co-occurrence was also observed between the Opitutaceae (Verrucomicrobia) and a β-glucosidase-gene (K05349). Sequences assigned

to the Opitutaceae as well as to this β-glucosidase gene showed a higher abundance under RT than CT in the surface soil layer. The abundance of the β-glucosidase genes derived from microbial families of the green community, however, did not show this effect anymore. In addition, three microbial families from the blue community co-occurred with an endoglucanase-gene (K01179); Anaerolineaceae (Chloroflexi) showed a higher abundance under RT than CT in the deeper soil layer, whereas the Verrucomicrobia subdivision 3 (Verrucomicrobia) and Planctomycetaceae (Planctomycetes) showed a higher abundance in the deeper soil layer than in the surface soil layer. Likewise, the abundance of the endoglucanase gene was higher in the deeper soil layer than in the surface soil layer. However, the abundance of the endoglucanase gene sequences derived from microbial families of the blue community, however, did not show a response to either depth or tillage treatment.

5 - Results of phylogenetic analysis of amplified GH5-genes

Here, the results of the amplification of putative GH5-cellulase genes from agricultural soil are given. Led by the goal of capturing the *in-situ* GH5-cellulase diversity, the primer design included a metagenome-read obtained from the conventional farming metagenome dataset. The primer set was developed for cellulase genes from GH5 subfamily 2 (GH5_2 cellulase genes). Results of the primer development process, the amplicon sequence analysis and GH5_2 cellulase gene phylogenetic analysis are shown.

GH5-primer design and quality analysis

In order to capture a high diversity of cellulase genes from soil, unbiased by database sequences, a metagenome read from the conventional farming experiment was used as basis for primer development. A metagenome read was selected which had been annotated with high reliability as an endoglucanase from *Gynuella sunshinyii* YC6258 (Genbank accession: AJQ95033.1). This endoglucanase was classified in the CAZy database as a GH5-cellulase in GH5-subfamily 2. The alignment of this metagenome read with two GH5-database sequences (likewise belonging to GH5-subfamily 2) resulted in the selection of suitable regions for primer sequences. The two database sequences used for primer design were endoglucanases from *Cellvibrio japonicus* strain Ueda107 and *Teredinibacter turnerae* strain T7901 (both Gammaproteobacteria, protein IDs ACE84076 respectively ABS72374). The sequences of the degenerate primer pair and its binding regions on the catalytic GH5-domain of ACE84076, as well as the eight conserved residues of a GH5-catalytic domain (104, 142) are shown in Figure 31.

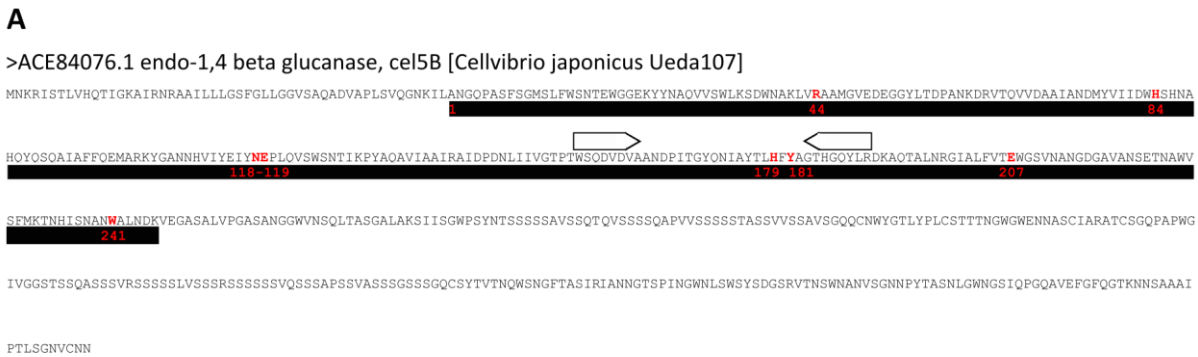


Figure 31: Binding sites and sequences of the GH5_2-primers. (A) Binding sites of the forward and reverse GH5_2 primers (white bars above protein sequence) on the protein sequence of the *Cellvibrio japonicus* Ueda107 endoglucanase, which was used for primer design. The black bar beneath the protein sequence indicates the conserved domain as identified by conserved domain database-search (CDD) (110). Conserved GH5-catalytic domain residues as identified by Wang *et al.* (104) and their position relative to the first residue in the conserved domain are shown in red and bold font. **(B)** Forward and reverse primer pair sequences without and **(C)** with Nextera V2 Adapter sequences (in bold) attached.

As expected*, in-silico* PCR using the metagenome dataset of the conventional farming experiment as template and allowing no mismatches resulted in one 104 bp-long amplicon which was amplified from the metagenome read used as template for design. Similarly, *in-silico* PCR using the NCBI nucleotide database as template and allowing for no mismatches resulted in two 107 bp-long amplicons, amplified from the genes encoding the two endoglucanases used as template for primer design. *In-silico* PCR using the Genbank bacterial and plant database as template and allowing for one mismatch resulted in 26 amplifications; 1 amplicon with a length of 103 bp was amplified from a cellulase-coding region on the complete genome of *Roseateles depolymerans*, a Betaproteobacterium. The encoded cellulase with protein ID: ALV07998.1 is likewise annotated (based on sequence-similarity) as a GH5-cellulase from subfamily 2 (CAZy database). The other 25 amplicons had a length of 106 bp and were amplified from endoglucanase-coding regions on gene sequences or complete genomes of different *Teredinibacter turnerae* or *Cellvibrio japonicus* strains. In addition to *in-silico* specificity analysis, the *in-vitro* specificity of the primer pair was assessed by amplification of the GH5_2 cellulase gene using genomic DNA from *C. japonicus* as template. Analysis of the amplification products on a 2%-agarose gel showed specific bands amplified from *C. japonicus* DNA (Figure A1 in the appendix).

GH5-cellulase identification by amplicon sequencing

To amplify GH5_2 cellulase genes from agricultural soil, the metagenomic DNA extracted from the six soil samples of the conventional farming experiment was pooled and used as template for amplification with the developed GH5_2 primer pair. Sequencing of the amplification products resulted in 305,217 forward and reverse reads. The steps in sequence analysis are described here and the amount of sequences left after each analysis step is summarized in Table 3. First, the obtained forward and reverse reads were merged, resulting in 303,480 (99.4%) merged reads with an average read length of 94 bp. After stringent read quality and length filtering (step 1, Table 3) 91,887 clean reads remained. Thereafter, these nucleotide sequences were translated to protein sequences (step 2, Table 3) and then screened for cellulase function by pairwise alignment and subsequent clustering of related sequences. This resulted in the identification of 2,033 amplicon sequences (2.21% of ORFs) which clustered with functionally characterized database sequences classified as EC:3.2.1.4-cellulases (endoglucanases) from GH5-subfamily 2 (www.cazy.org, (60))(step 3, Table 3). For an overview of the GH5-endoglucanase database sequences used in this analysis step, see Table A2 in the appendix. After identification of the amplicon sequences with predicted GH5_2 cellulase function, duplicate sequences were removed, which resulted in a final amount of 743 unique protein sequences (step 4, Table 3). These unique sequences showed an average pairwise identity among each other of 71% on nucleotide level and 59% on amino acid level. They were subsequently analyzed for taxonomic affiliation and further subclustered using Markov Clustering in order to group sequences on a higher sequence identity level; subclustering of the amplified GH5_2 cellulase genes together with database sequences from GH5-subfamily 2 resulted in 101 clusters, of which amplicon sequences in subcluster 1 showed 74% average sequence identity, sequences in subcluster 2 showed 80% average sequence identity and sequences in the remaining subclusters >90%. Of all subclusters, 74 contained only 1 amplicon sequence which also did not represent a duplicate sequence (step 4, Table 3). Therefore, these sequences were omitted from further analysis (step 5, Table 3) to secure the use of valid amplicon sequences for phylogenetic analysis. For a detailed overview of the subclustering results, see Table A10 in the appendix.

Table 3: Number of GH5-amplicon sequences remaining after each analysis step and their mean sequence length (in base pairs for nucleotide sequences and in amino acids for protein sequences).

		Sequences	**Length**
Step 1	Merged, high quality-nucleotide reads	91,887	105
Step 2	Protein sequences after ORF-prediction	89,741	34
Step 3	Cellulase-annotated protein sequences (by Markov Clustering)	2,033	33
Step 4	Unique cellulase protein sequences after removal of duplicates	743	34
Step 5	Protein sequences with 2 or more representatives at high similarity	669	34

Taxonomic assignment

All amplicon sequences identified as GH5_2 cellulase genes (after analysis step 4, Table 3) were compared to sequences in the NCBI database for taxonomic assignment. First, the usual but stringent method for taxonomic assignment was applied, requiring 100% of the top BLAST hits to agree on assignment to a specific taxonomic level. Then, a secondary taxonomic assignment method was applied, requiring only 50% of the top BLAST hits to agree on an assignment. Annotation percentages mentioned in this chapter are derived from the 50%-annotation method. Furthermore, the method used for GH5 cellulase gene identification in the amplicon dataset (pairwise alignment and Markov Clustering) was applied to identify GH5 cellulase genes in the metagenome of the conventional farming experiment, resulting in the prediction of in total 682 (68.7% $x10^{-3}$ of all metagenome reads) GH5 cellulase genes in the agricultural soil metagenome. The taxonomic assignments of the GH5 cellulase genes identified in the amplicon dataset, metagenome dataset and of functionally characterized GH5 cellulase sequences from the CAZy database are shown here on phylum level in Figure 32.

To a considerable portion of the GH5 cellulase genes identified in the amplicon and metagenome dataset the taxonomy could not be unambiguously assigned, especially if 100% of the BLAST hits were used for annotation. This indicates that a large extent of the GH5 cellulase genes in this environmental sample is as yet unexplored and unknown. The difference in results between a more (Figure 32, LCA-100) and less (Figure 32, LCA-50) stringent annotation also indicates that these partial gene sequences show little monophyletic sequence conservation, as high similarity to database sequences originating from different phyla can be obtained from the same amplicon sequence. Furthermore, the taxonomic diversity captured with the GH5_2 primers (i.e. cellulase genes with affiliation to six different phyla) is much lower than what is known from the database (Figure 32, CAZy: functionally characterized GH5 cellulases with a taxonomic origin in 17 different phyla) and what can be captured by metagenome sequencing (Figure 32, metagenome: GH5 cellulase genes taxonomically assigned to 21 different phyla), probably mainly attributable to the fact that the primer pair was designed to target only GH5 cellulases from subfamily 2. Moreover, some metagenomic GH5 cellulase genes were harboured by microbial phyla (e.g. Cyanobacteria) which were not represented in the functionally characterized GH5 cellulase sequences of the CAZy database. It is worth noticing, however, that GH5 cellulase genes from these phyla may be present among the non-functionally characterized sequences in the public databases.

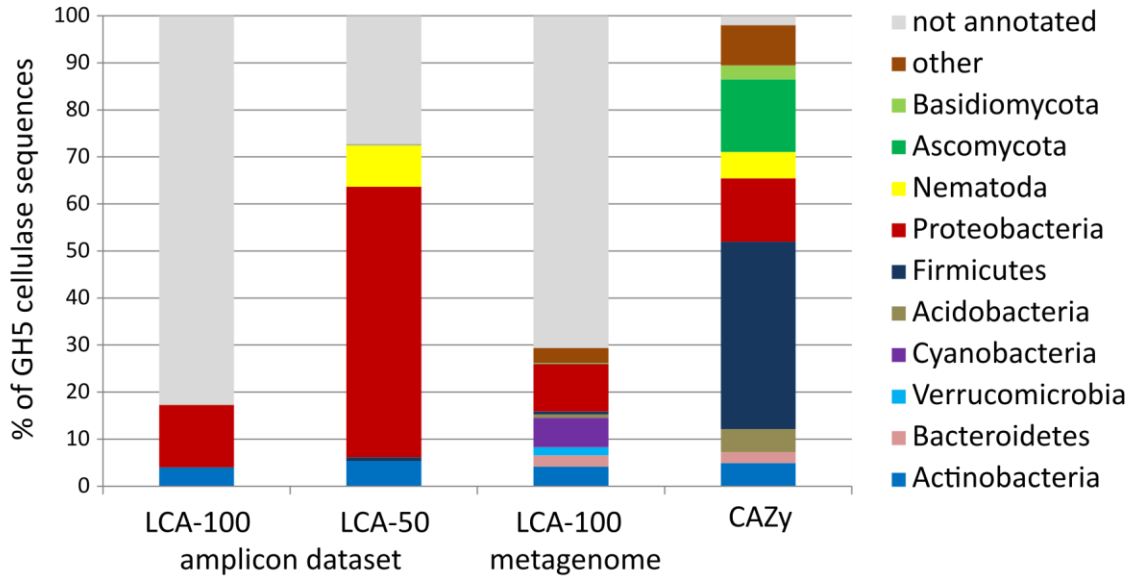


Figure 32: Taxonomic assignment of GH5-cellulases in the amplicon dataset, the conventional farming metagenome dataset and the CAZy database on phylum level. The GH5-cellulase genes identified in the amplicon and metagenome dataset were taxonomically assigned using 100% (LCA-100) or 50% (LCA-50) of the top BLAST hits. Shown is the percentage of GH5 cellulase sequences annotated to the most abundant microbial phyla.

As the GH5_2 primers were designed using proteobacterial sequences (see Figure 31 and section 2: Materials and Methods), it is not surprising that the majority of the taxonomically-assigned amplicon sequences was annotated as Proteobacteria (57.5%). Of these, the majority (91%) was annotated to *Saccharophagus*, *Teredinibacter, Dickeya* and *Vibrio*, but also to eight other proteobacterial genera (*Cystobacter*, *Sorangium*, *Pseudoalteromonas*, *Pantoea*, *Acinetobacter*, *Cellvibrio*, *Arenimonas* and *Rhodanobacter*). GH5 cellulases from Firmicutes were highly abundant among the CAZy database sequences (39.8%) but strikingly few amplified GH5_2 cellulase genes (0.3%) and metagenomic GH5 cellulase genes (0.6%) were harboured by Firmicutes. Most of the amplified GH5_2 cellulase genes which were derived from Firmicutes belonged to the Bacilli. In addition, some were assigned to Actinobacteria (5.4%) and more specifically to the *Micromonospora* (order Actinomycetales) and to a lesser extent to the genus *Janibacter* (order Actinomycetales). Furthermore, using less stringent annotation, several amplicon sequences were derived from Nematoda (8.8%), of which many could be assigned to the plant-parasitic genus *Pratylenchus*. Finally, one amplified GH5_2 cellulase gene showed significant similarity to database sequences from an uncultured protist.

Subclustering of the unique amplicon sequences annotated as GH5_2 cellulase genes resulted in the identification of singletons; sequences which were not similar to other amplicon sequences (single sequence in a subcluster) and did not represent a duplicate were removed to ensure phylogenetic analysis of biologically relevant sequences (see analysis step 5, Table 3). Subclusters containing amplicon sequences with taxonomic assignment are shown in Table 4. Subcluster 1 contained 544 amplicon sequences and all of the database sequences (111) included in the subclustering analysis. Most amplicon sequences in subcluster 1 were assigned to Proteobacteria and, using a less stringent annotation procedure (LCA-50), to Nematoda. The amplicon sequences in subcluster 2, on the other hand, were nearly exclusively assigned to Actinobacteria. These two subclusters were the largest

subclusters and contained most of the taxonomically assigned amplicon sequences. Interestingly, the two conserved amino acid residues located within the amplified region (residues H179 & Y181 on the protein sequence with ID: ACE84076.1, see Figure 31), were only present in the amplicon sequences of subclusters 1, 2, 5 and 9.

Table 4: Taxonomic assignment of unique amplicon sequences annotated as cellulase genes. Given are the number of sequences assigned and the total number of reads within subclusters 1, 2, 12 and 16, using 100% or 50% of the top BLAST hits for the last common ancestor (LCA)-annotation method in MEGAN5.

Subcluster	1		2		12		16	
Taxonomic assignment	LCA-100	LCA-50	LCA-100	LCA-50	LCA-100	LCA-50	LCA-100	LCA-50
Actinobacteria			26	37	1	1		
Proteobacteria	80	422						3
Firmicutes		4		1				
Nematoda		64		1				
uncultured Protist	1	1						
unassigned	463	53	16	3	3	3	3	
Total	**544**	**544**	**42**	**42**	**4**	**4**	**3**	**3**

Phylogenetic tree of GH5-sequences from CAZy database

Two reference trees were calculated from the phylogenetic analysis of the functionally characterized GH5-cellulases from the CAZy database; the tree obtained from complete GH5-sequences is shown in Figure 33 and the tree obtained from partial (amplicon-region) GH5-sequences is shown in Figure 34. As expected, the topology of the complete sequence-reference tree shows high consistency with previous subclassification of GH5-sequences by Aspeborg *et al*. (100), who analysed the phylogenetic distribution of both functionally and non-functionally characterized GH5-sequences in the CAZy database. Both the reference tree presented here (Figure 33) and the tree by Aspeborg *et al.* shows high phylogenetic diversity of GH5 sequences and very little conservation within phylogenetic groups. This polyphyletic distribution offers an explanation for the difficulty of accurate taxonomic assignment of the amplicon sequences, as was mentioned earlier. In the tree of Aspeborg *et al*., sequences in subfamilies 1, 2, 26, 8 and 53 are more related to each other, as well as sequences in subfamilies 4, 25, 37, 38 and 52. Additionally, subfamily 5-sequences are located in between these two groups and subfamily 22-sequences are grouped together with many other subfamilies without an endoglucanase function. In the tree shown here (Figure 33), these characteristics are also clearly present. The large clade of sequences consisting of only GH5-cellulases from subfamily 2 has high bootstrap support (89%) and will further be designated “clade 1”. This clade contains mainly bacterial and metazoan (nematode- and insect-) sequences, and does not contain any fungal sequences, with the exception of a sequence from the Neocallimastigomycota. Within clade 1, a separate clade of sequences (designated “clade 1.1”) is marked as the phylogenetic range to which the amplified GH5_2 cellulase genes belong. This clade also includes the database sequences used in primer design (see arrow in Figure 33) and is fairly well supported statistically by a bootstrap value of 39%.

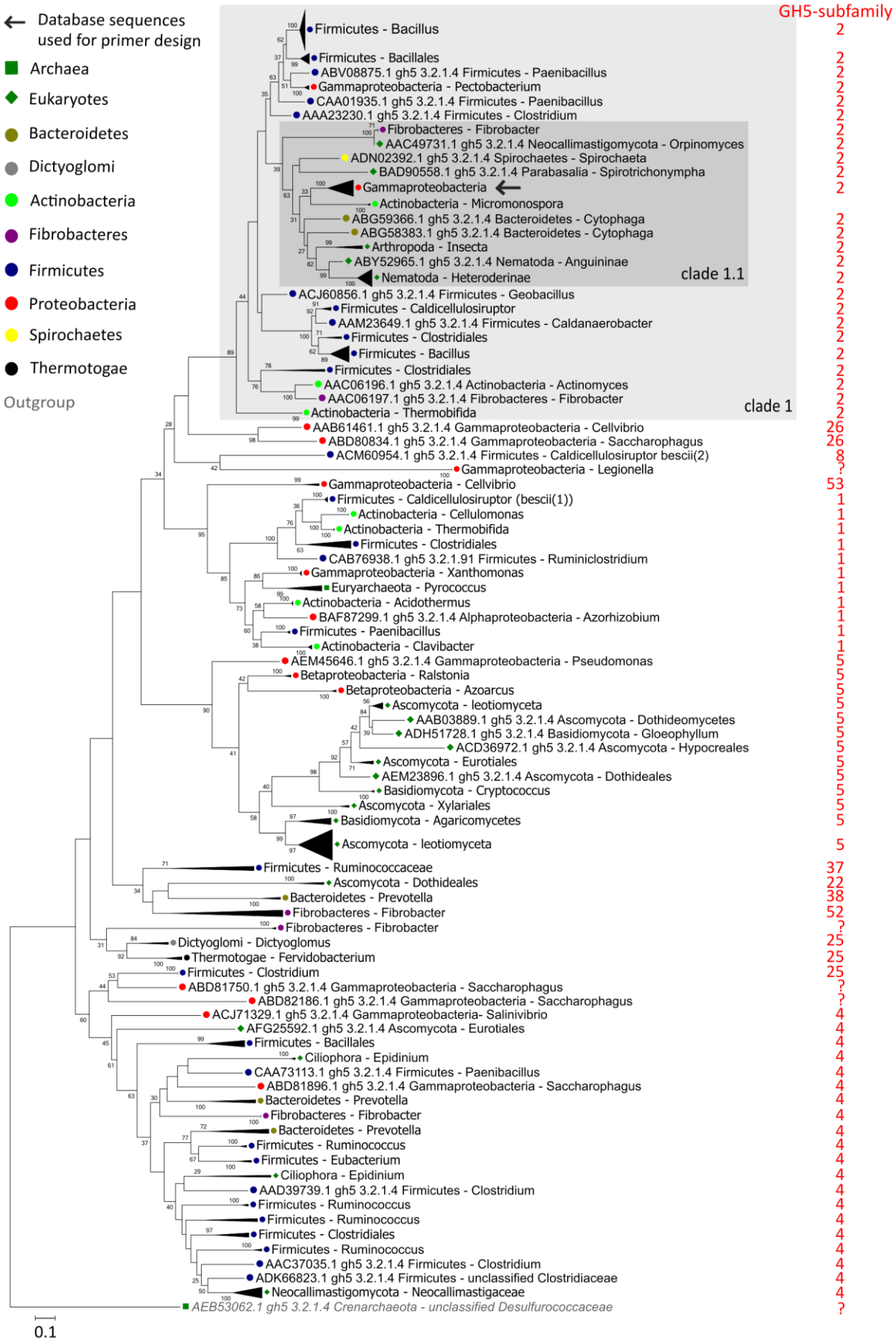

Database sequences used for primer design
Archaea
Eukaryotes
Bacteroidetes
Dictyoglomi
Actinobacteria
Fibrobacteres
Firmicutes
Proteobacteria
Spirochaetes
Thermotogae
Outgroup
GH5-subfamily
clade 1.1
clade 1
Firmicutes - Bacillus
Firmicutes - Bacillales
ABV08875.1 gh5 3.2.1.4 Firmicutes - Paenibacillus
Gammaproteobacteria - Pectobacterium
CAA01935.1 gh5 3.2.1.4 Firmicutes - Paenibacillus
AAA23230.1 gh5 3.2.1.4 Firmicutes - Clostridium
Fibrobacteres - Fibrobacter
AAC49731.1 gh5 3.2.1.4 Neocallimastigomycota - Orpinomyces
ADN02392.1 gh5 3.2.1.4 Spirochaetes - Spirochaeta
BAD90558.1 gh5 3.2.1.4 Parabasalia - Spirotrichonympha
Gammaproteobacteria
Actinobacteria - Micromonospora
ABG59366.1 gh5 3.2.1.4 Bacteroidetes - Cytophaga
ABG58383.1 gh5 3.2.1.4 Bacteroidetes - Cytophaga
Arthropoda - Insecta
ABY52965.1 gh5 3.2.1.4 Nematoda - Anguininae
Nematoda - Heteroderinae
ACJ60856.1 gh5 3.2.1.4 Firmicutes - Geobacillus
Firmicutes - Caldicellulosiruptor
AAM23649.1 gh5 3.2.1.4 Firmicutes - Caldanaerobacter
Firmicutes - Clostridiales
Firmicutes - Bacillus
Firmicutes - Clostridiales
AAC06196.1 gh5 3.2.1.4 Actinobacteria - Actinomyces
AAC06197.1 gh5 3.2.1.4 Fibrobacteres - Fibrobacter
Actinobacteria - Thermobifida
AAB61461.1 gh5 3.2.1.4 Gammaproteobacteria - Cellvibrio
ABD80834.1 gh5 3.2.1.4 Gammaproteobacteria - Saccharophagus
ACM60954.1 gh5 3.2.1.4 Firmicutes - Caldicellulosiruptor bescii(2)
Gammaproteobacteria - Legionella
Gammaproteobacteria - Cellvibrio
Firmicutes - Caldicellulosiruptor (bescii(1))
Actinobacteria - Cellulomonas
Actinobacteria - Thermobifida
Firmicutes - Clostridiales
CAB76938.1 gh5 3.2.1.91 Firmicutes - Ruminiclostridium
Gammaproteobacteria - Xanthomonas
Euryarchaeota - Pyrococcus
Actinobacteria - Acidothermus
BAF87299.1 gh5 3.2.1.4 Alphaproteobacteria - Azorhizobium
Firmicutes - Paenibacillus
Actinobacteria - Clavibacter
AEM45646.1 gh5 3.2.1.4 Gammaproteobacteria - Pseudomonas
Betaproteobacteria - Ralstonia
Betaproteobacteria - Azoarcus
Ascomycota - leotiomyceta
AAB03889.1 gh5 3.2.1.4 Ascomycota - Dothideomycetes
ADH51728.1 gh5 3.2.1.4 Basidiomycota - Gloeophyllum
ACD36972.1 gh5 3.2.1.4 Ascomycota - Hypocreales
Ascomycota - Eurotiales
AEM23896.1 gh5 3.2.1.4 Ascomycota - Dothideales
Basidiomycota - Cryptococcus
Ascomycota - Xylariales
Basidiomycota - Agaricomycetes
Ascomycota - leotiomyceta
Firmicutes - Ruminococcaceae
Ascomycota - Dothideales
Bacteroidetes - Prevotella
Fibrobacteres - Fibrobacter
Fibrobacteres - Fibrobacter
Dictyoglomi - Dictyoglomus
Thermotogae - Fervidobacterium
Firmicutes - Clostridium
ABD81750.1 gh5 3.2.1.4 Gammaproteobacteria - Saccharophagus
ABD82186.1 gh5 3.2.1.4 Gammaproteobacteria - Saccharophagus
ACJ71329.1 gh5 3.2.1.4 Gammaproteobacteria- Salinivibrio
AFG25592.1 gh5 3.2.1.4 Ascomycota - Eurotiales
Firmicutes - Bacillales
Ciliophora - Epidinium
CAA73113.1 gh5 3.2.1.4 Firmicutes - Paenibacillus
ABD81896.1 gh5 3.2.1.4 Gammaproteobacteria - Saccharophagus
Bacteroidetes - Prevotella
Fibrobacteres - Fibrobacter
Bacteroidetes - Prevotella
Firmicutes - Ruminococcus
Firmicutes - Eubacterium
Ciliophora - Epidinium
AAD39739.1 gh5 3.2.1.4 Firmicutes - Clostridium
Firmicutes - Ruminococcus
Firmicutes - Ruminococcus
Firmicutes - Clostridiales
Firmicutes - Ruminococcus
AAC37035.1 gh5 3.2.1.4 Firmicutes - Clostridium
ADK66823.1 gh5 3.2.1.4 Firmicutes - unclassified Clostridiaceae
Neocallimastigomycota - Neocallimastigaceae
AEB53062.1 gh5 3.2.1.4 Crenarchaeota - unclassified Desulfurococcaceae
0.1

Figure 33

Figure 33 (page 68): Maximum Likelihood-phylogenetic tree of complete GH5-database amino acid sequences obtained from the CAZy database (60)**.** The analysis was performed using 308 complete domain sequences of functionally characterized GH5-cellulases and included 428 alignment positions. The percentage of trees in which the associated sequences clustered together is shown next to the branches and was determined by bootstrapping with 100 replicates. Only bootstrap values higher than 25% are shown. The tree is drawn to scale, with branch lengths measured in the number of modelled substitutions per site. The taxonomic affiliation on phylum level of each sequence or collapsed cluster of sequences is indicated in the sequence name and by a coloured marker. Collapsed branches consist of sequences from the same taxonomic group as indicated by the cluster name. GH5-subfamily classification is shown in red besides the tree (57, 97). Sequences belonging to clade 1 (GH5-subfamily 2) and clade 1.1 are outlined by a light- respectively dark-grey box. The location in the tree of the database sequences which were used for primer design is indicated by an arrow.

In the tree calculated from partial GH5-database sequences (Figure 34), clade 1 is partly broken up, with several cellulase sequences from Firmicutes missing and including a few new sequences. Clade 1.1 is almost completely intact, except for three sequences which are located separately within clade 1. However, clade 1.1 has a slightly different topology here from that observed in the complete sequence-reference tree (Figure 33) and no bootstrap support. Thus, the analysis of phylogenetic tree-topology based on partial or complete cellulase sequences suggests that, although a rough indication of the phylogenetic relationships between sequences can be obtained using partial GH5-sequences, these phylogenetic relationships should be confirmed using complete protein sequences.

Phylogenetic tree of GH5-amplicon sequences

The phylogenetic relationships between the amplified GH5_2 cellulase genes and the partial GH5-database sequences are shown in Figure 35 and as a larger image in Figure A2 in the appendix. Database sequences are indicated with filled markers, whereas the amplicon sequences are marked with open markers. The tree shows that the amplified GH5_2 cellulase genes within a subcluster are more closely related to each other than to amplicon sequences from another subcluster (coloured bar beside the tree), corroborating the results of the subclustering analysis. Furthermore, the short branch lengths show that the amplicon sequences are highly related. Analysis of overall tree topology reveals that clade 1 is broken up by other database sequences which did not originally belong to clade 1, as was also observed in the partial-sequence reference tree without amplicon sequences (Figure 34). All amplicon sequences are related to database sequences of the original clade 1.1. However, clade 1.1 is not intact anymore.

Amplicon sequences of subcluster 1 (red open circles and red bar in Figure 35) are generally distributed around gammaproteobacterial database sequences of clade 1.1. As most taxonomically-assigned amplicon sequences of subcluster 1 are assigned to Gammaproteobacteria, this is not surprising. However, database sequences with other taxonomic origin seem to be related to these amplicon sequences as well. Several gammaproteobacterial amplicon sequences are found in close proximity to a GH5- database sequence from *Cytophaga hutchinsonii* (Figure 35-B). Furthermore, amplicon sequences assigned to Gammproteobacteria, Firmicutes and a protist, cluster around a GH5-database sequence from *Spirotrichonympha*, a protist genus (Figure 35-C). Finally, several amplicon sequences harboured by Gammaproteobacteria, are closely related to two database sequences from *Vibrio* and to one from *Spirochaeta* (Figure 35-D). This polyphyletic clustering offers a basis for hypothesis formulations regarding HGT-events.

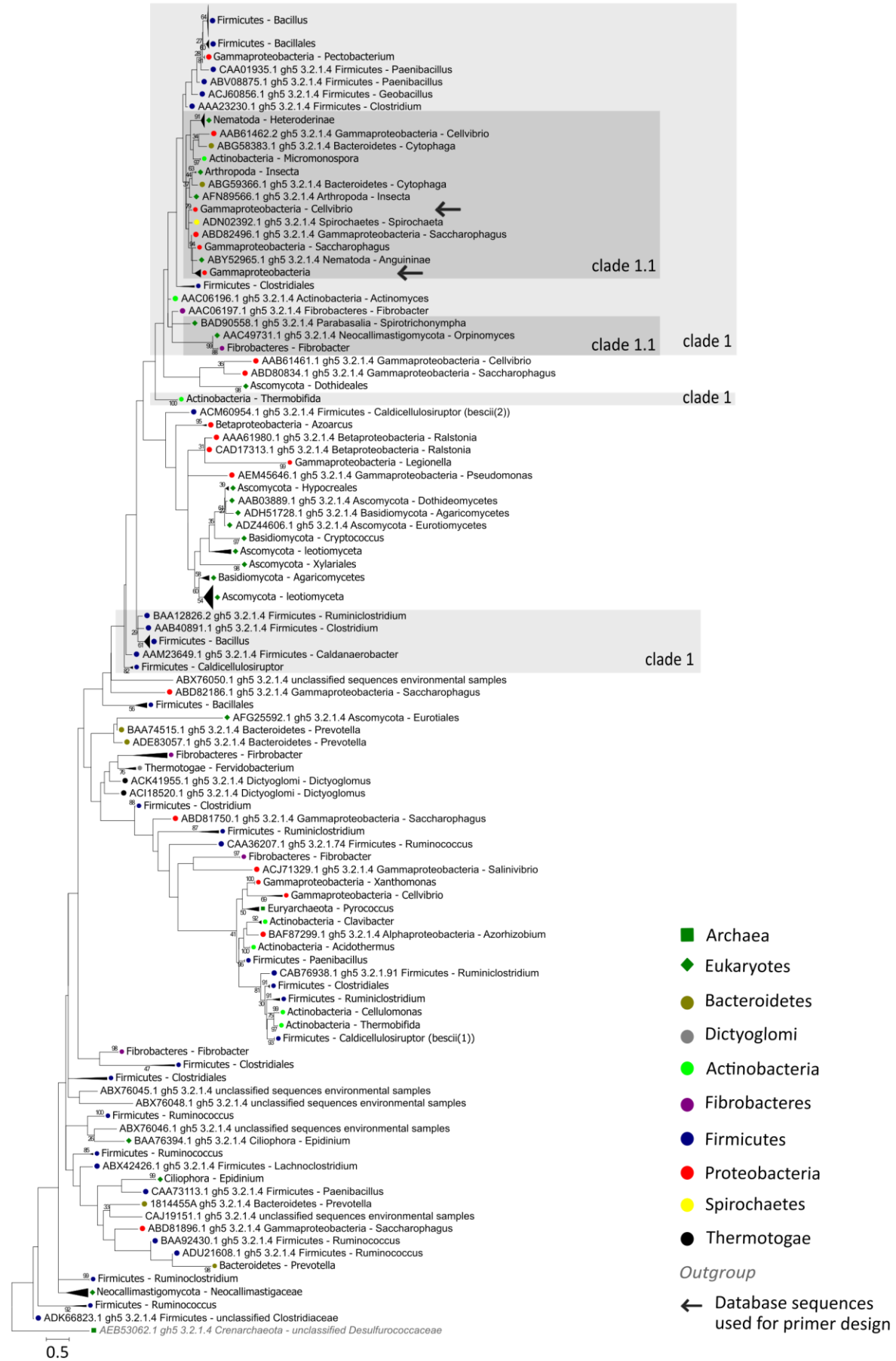

Firmicutes - Bacillus
Firmicutes - Bacillales
Gammaproteobacteria - Pectobacterium
CAA01935.1 gh5 3.2.1.4 Firmicutes - Paenibacillus
ABV08875.1 gh5 3.2.1.4 Firmicutes - Paenibacillus
ACJ60856.1 gh5 3.2.1.4 Firmicutes - Geobacillus
AAA23230.1 gh5 3.2.1.4 Firmicutes - Clostridium
Nematoda - Heteroderinae
AAB61462.2 gh5 3.2.1.4 Gammaproteobacteria - Cellvibrio
ABG58383.1 gh5 3.2.1.4 Bacteroidetes - Cytophaga
Actinobacteria - Micromonospora
Arthropoda - Insecta
ABG59366.1 gh5 3.2.1.4 Bacteroidetes - Cytophaga
AFN89566.1 gh5 3.2.1.4 Arthropoda - Insecta
Gammaproteobacteria - Cellvibrio
ADN02392.1 gh5 3.2.1.4 Spirochaetes - Spirochaeta
ABD82496.1 gh5 3.2.1.4 Gammaproteobacteria - Saccharophagus
Gammaproteobacteria - Saccharophagus
ABY52965.1 gh5 3.2.1.4 Nematoda - Anguininae
Gammaproteobacteria
clade 1.1
Firmicutes - Clostridiales
AAC06196.1 gh5 3.2.1.4 Actinobacteria - Actinomyces
AAC06197.1 gh5 3.2.1.4 Fibrobacteres - Fibrobacter
BAD90558.1 gh5 3.2.1.4 Parabasalia - Spirotrichonympha
AAC49731.1 gh5 3.2.1.4 Neocallimastigomycota - Orpinomyces
Fibrobacteres - Fibrobacter
clade 1.1
clade 1
AAB61461.1 gh5 3.2.1.4 Gammaproteobacteria - Cellvibrio
ABD80834.1 gh5 3.2.1.4 Gammaproteobacteria - Saccharophagus
Ascomycota - Dothideales
Actinobacteria - Thermobifida
clade 1
ACM60954.1 gh5 3.2.1.4 Firmicutes - Caldicellulosiruptor (bescii(2))
Betaproteobacteria - Azoarcus
AAA61980.1 gh5 3.2.1.4 Betaproteobacteria - Ralstonia
CAD17313.1 gh5 3.2.1.4 Betaproteobacteria - Ralstonia
Gammaproteobacteria - Legionella
AEM45646.1 gh5 3.2.1.4 Gammaproteobacteria - Pseudomonas
Ascomycota - Hypocreales
AAB03889.1 gh5 3.2.1.4 Ascomycota - Dothideomycetes
ADH51728.1 gh5 3.2.1.4 Basidiomycota - Agaricomycetes
ADZ44606.1 gh5 3.2.1.4 Ascomycota - Eurotiomycetes
Basidiomycota - Cryptococcus
Ascomycota - leotiomyceta
Ascomycota - Xylariales
Basidiomycota - Agaricomycetes
Ascomycota - leotiomyceta
BAA12826.2 gh5 3.2.1.4 Firmicutes - Ruminiclostridium
AAB40891.1 gh5 3.2.1.4 Firmicutes - Clostridium
Firmicutes - Bacillus
AAM23649.1 gh5 3.2.1.4 Firmicutes - Caldanaerobacter
Firmicutes - Caldicellulosiruptor
clade 1
ABX76050.1 gh5 3.2.1.4 unclassified sequences environmental samples
ABD82186.1 gh5 3.2.1.4 Gammaproteobacteria - Saccharophagus
Firmicutes - Bacillales
AFG25592.1 gh5 3.2.1.4 Ascomycota - Eurotiales
BAA74515.1 gh5 3.2.1.4 Bacteroidetes - Prevotella
ADE83057.1 gh5 3.2.1.4 Bacteroidetes - Prevotella
Fibrobacteres - Firbrobacter
Thermotogae - Fervidobacterium
ACK41955.1 gh5 3.2.1.4 Dictyoglomi - Dictyoglomus
ACI18520.1 gh5 3.2.1.4 Dictyoglomi - Dictyoglomus
Firmicutes - Clostridium
ABD81750.1 gh5 3.2.1.4 Gammaproteobacteria - Saccharophagus
Firmicutes - Ruminiclostridium
CAA36207.1 gh5 3.2.1.74 Firmicutes - Ruminococcus
Fibrobacteres - Fibrobacter
ACJ71329.1 gh5 3.2.1.4 Gammaproteobacteria - Salinivibrio
Gammaproteobacteria - Xanthomonas
Gammaproteobacteria - Cellvibrio
Euryarchaeota - Pyrococcus
Actinobacteria - Clavibacter
BAF87299.1 gh5 3.2.1.4 Alphaproteobacteria - Azorhizobium
Actinobacteria - Acidothermus
Firmicutes - Paenibacillus
CAB76938.1 gh5 3.2.1.91 Firmicutes - Ruminiclostridium
Firmicutes - Clostridiales
Firmicutes - Ruminiclostridium
Actinobacteria - Cellulomonas
Actinobacteria - Thermobifida
Firmicutes - Caldicellulosiruptor (bescii(1))
Fibrobacteres - Fibrobacter
Firmicutes - Clostridiales
Firmicutes - Clostridiales
ABX76045.1 gh5 3.2.1.4 unclassified sequences environmental samples
ABX76048.1 gh5 3.2.1.4 unclassified sequences environmental samples
Firmicutes - Ruminococcus
ABX76046.1 gh5 3.2.1.4 unclassified sequences environmental samples
BAA76394.1 gh5 3.2.1.4 Ciliophora - Epidinium
Firmicutes - Ruminococcus
ABX42426.1 gh5 3.2.1.4 Firmicutes - Lachnoclostridium
Ciliophora - Epidinium
CAA73113.1 gh5 3.2.1.4 Firmicutes - Paenibacillus
1814455A gh5 3.2.1.4 Bacteroidetes - Prevotella
CAJ19151.1 gh5 3.2.1.4 unclassified sequences environmental samples
ABD81896.1 gh5 3.2.1.4 Gammaproteobacteria - Saccharophagus
BAA92430.1 gh5 3.2.1.4 Firmicutes - Ruminococcus
ADU21608.1 gh5 3.2.1.4 Firmicutes - Ruminococcus
Bacteroidetes - Prevotella
Firmicutes - Ruminoclostridium
Neocallimastigomycota - Neocallimastigaceae
Firmicutes - Ruminococcus
ADK66823.1 gh5 3.2.1.4 Firmicutes - unclassified Clostridiaceae
AEB53062.1 gh5 3.2.1.4 Crenarchaeota - unclassified Desulfurococcaceae
0.5
Archaea
Eukaryotes
Bacteroidetes
Dictyoglomi
Actinobacteria
Fibrobacteres
Firmicutes
Proteobacteria
Spirochaetes
Thermotogae
Outgroup
Database sequences used for primer design

Figure 34.

Figure 34 (page 70): Maximum Likelihood-phylogenetic tree of partial GH5-database amino acid sequences obtained from the CAZy database (60)**.** The analysis was done using 308 partial (amplicon region) sequences of functionally characterized GH5-cellulases and included 85 alignment positions. The percentage of trees in which the associated sequences clustered together is shown next to the branches and was determined by bootstrapping with 100 replicates. Only bootstrap values higher than 25% are shown. The tree is drawn to scale, with branch lengths measured in the number of modelled substitutions per site. The taxonomic affiliation on phylum level of each sequence or collapsed cluster of sequences is indicated in the sequence-name and by a coloured marker. Collapsed branches consist of sequences from the same taxonomic group as indicated by the cluster name. Sequences belonging to clade 1 (GH5-subfamily 2) and clade 1.1 are outlined by a light- respectively dark-grey box. The locations in the tree of the database sequences which were used for primer design are indicated by an arrow.

However, very few branching events in this tree are well supported by the bootstrap analysis and the observed topologies must be verified using complete GH5-sequences (results shown below).

Amplicon sequences of subcluster 2 (light green open circles and green bar in Figure 35) cluster with two partial database sequences from *Micromonospora* spp. These sequences were the top BLAST hits of an amplified GH5_2 cellulase gene harboured by Actinobacteria (subcluster 2) but are not functionally characterized. It is therefore not surprising that they show high similarity to the amplicon sequences of subcluster 2. These amplicon sequences are furthermore closely related to partial nematode cellulases. Strikingly, one amplified GH5_2 cellulase gene in subcluster 2 is assigned to nematodes (see Table 4). Moreover, nematode cellulase sequences are present several times among the top BLAST hits of the amplicon sequences of subcluster 2. In clade 1.1 of the partial GH5-database sequences (Figure 34) it can be observed that the *Micromonospora*-sequences are closely related to nematode sequences. However, these relationships are not well statistically supported by bootstrap analysis.

Amplicon sequences of subclusters other than 1 or 2 (black open circles and black bar in Figure 35) do not cluster together with database sequences. They appear to be phylogenetically more distinct from the database sequences than the amplicon sequences from subcluster 1 or 2. This was also evident from the subclustering analysis, as they only clustered with the database sequence when the clustering e-value cut-off was higher than 10^{-11}.

Phylogenetic tree of top BLAST hits and complete database sequences

To obtain additional insight in the phylogenetic relationships between the functionally characterized GH5-database sequences and the amplified GH5_2 cellulase genes which cluster together with database sequences (see Figure 35-B, C, D and E), several of these amplified GH5_2 cellulase genes were compared to the NCBI non-redundant protein database to obtain the full length-sequences of their top BLAST hits. These top BLAST hits were subsequently re-aligned to the existing alignment of functionally characterized GH5-database sequences. The resulting phylogenetic tree is shown in Figure 36 (only clade 1) and in Figure A3 in the appendix (full tree), where the top BLAST hit-sequences are indicated with bold font.

The top BLAST hits of the amplified GH5_2 cellulase genes surrounding *Cytophaga* (bold sequences with “surr. Cytophaga” in Figure 36) and *Spirochaeta* (bold sequences with “surr. Spirochaeta” in Figure 36) belong to several species of Gammaproteobacteria while one is from an uncultured

bacterium, and they cluster together with the functionally characterized database sequences of the Gammaproteobacteria in clade 1.1 with high bootstrap support.

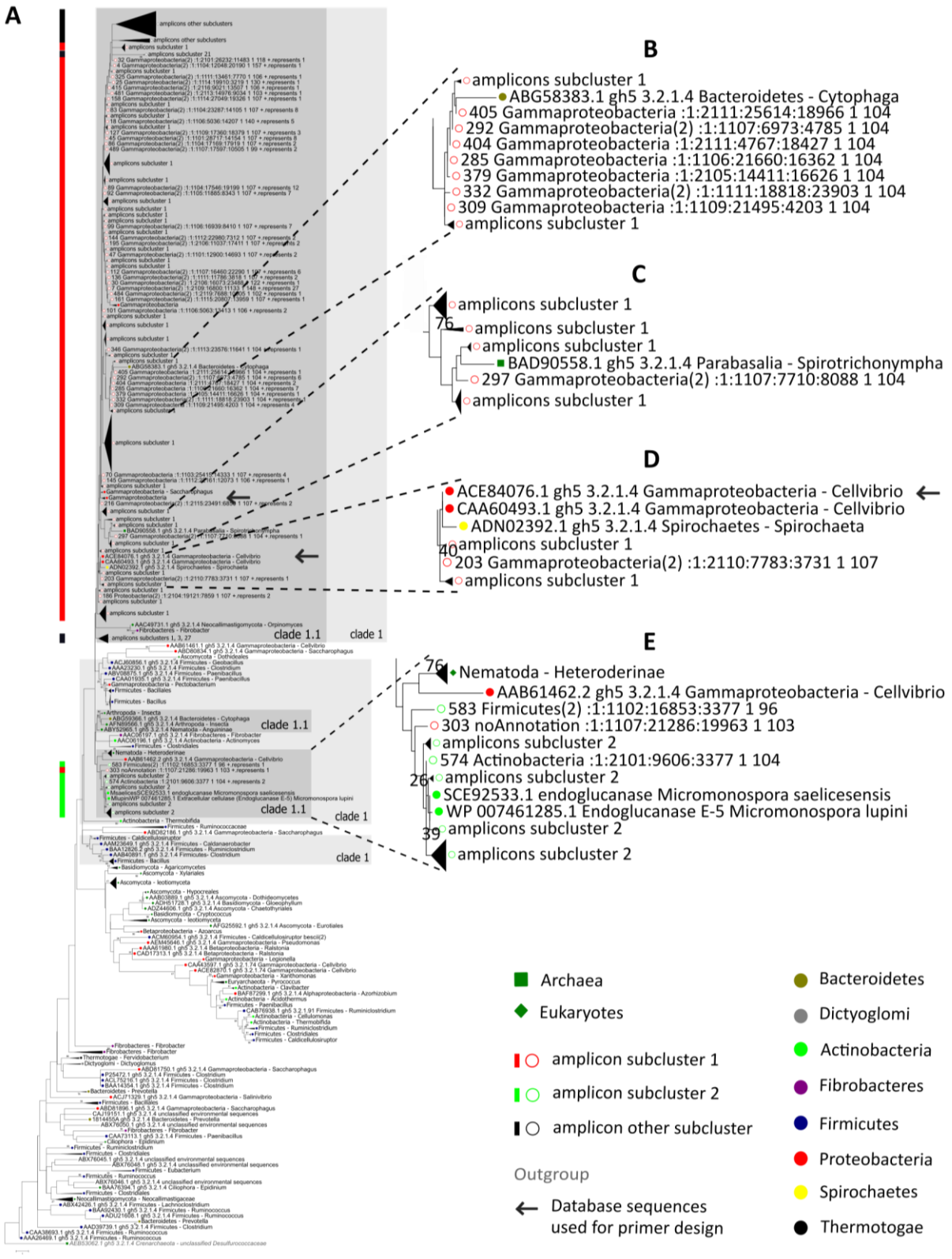


Figure 35.

Figure 35 (page 72): Maximum Likelihood-phylogenetic tree of cellulase-amplicon sequences and partial GH5-database amino acid sequences obtained from the CAZy database (60)**.** The analysis was done using 669 amplicon sequences and 308 partial (amplicon region) sequences of functionally characterized GH5-cellulases (total 977 sequences) and included 64 alignment positions. The percentage of trees in which the associated sequences clustered together is shown next to the branches and was determined by bootstrapping with 100 replicates. Only bootstrap values higher than 25% are shown. The tree is drawn to scale, with branch lengths measured in the number of modelled substitutions per site. The taxonomic affiliation on phylum level of each sequence or collapsed cluster of sequences is indicated in the sequence name and by a coloured marker; a filled marker for database sequences and an open marker for amplicon sequences. Taxonomic affiliations indicated with "(2)" in the sequence name have been assigned using less stringent taxonomic assignment cut-offs (LCA=50%). Collapsed branches consist of sequences from the same taxonomic group as indicated by the cluster name. Sequences belonging to clade 1 (GH5-subfamily 2) and clade 1.1 are outlined by a light- respectively dark-grey box. The locations in the tree of the database sequences which were used for primer design are indicated by an arrow. (A) Complete tree of database and amplicon sequences. Database sequences belonging to clade 1 (GH5-subfamily 2) and clade 1.1 are outlined by a light- respectively dark-grey box. The location of amplicon sequences in the tree is indicated by the bar on the left side of the tree, with the colour corresponding to the subcluster number. (B) Magnification of clade with amplicon sequences from subcluster 1 and a database sequence from *Cytophaga*. (C) Magnification of clade with amplicon sequences from subcluster 1 and a database sequence from *Spirotrichonympha*. (D) Magnification of sequence clade with amplicons from subcluster 1 and database sequences from *Spirochaeta* and *Cellvibrio*. (E) Magnification of clade with amplicon sequences from subcluster 2 and database sequences from Nematoda, *Cellvibrio* and *Micromonospora*.

When analysing the top BLAST hits of the amplicon sequences related to *Spirotrichonympha* (bold sequences with "surr. Spirotrichonympha" in Figure 36), a variable phylogenetic relationship with the functionally characterized database sequences is observed. First, the top BLAST hits of an amplified GH5_2 cellulase gene taxonomically derived from Gammaproteobacteria (amplicon sequence with number 297) were analyzed; two of the top BLAST hits belong to a different phylum (Firmicutes) than two other top BLAST hits, which belong to the Gammaproteobacteria. The latter top BLAST hits cluster accordingly with the functionally characterized complete sequences of the Gammaproteobacteria in clade 1.1, while the two top BLAST hits belonging to Firmicutes are phylogenetically more related to the functionally characterized Firmicutes sequences which are located in clade 1, but not in clade 1.1 (Figure 36). Then, the top BLAST hits of another amplified GH5_2 cellulase gene related to *Spirotrichonympha* (with number 281) but taxonomically harboured by Firmicutes, were analyzed; these two top BLAST hits were the same sequences as the two top BLAST hits of amplicon sequence number 297 (see above). Subsequently, the top BLAST hits were included of an amplified GH5_2 cellulase gene related to *Spirotrichonympha* (with number 581) but taxonomically harboured by an uncultured symbiotic protist. One of these top BLAST hits is, indeed, a putative GH5-sequence from an uncultured protist and is highly related to the functionally characterized *Spirotrichonympha*-sequence in clade 1.1 (Figure 36). The other top BLAST hit, however, belongs to a Gammaproteobacterium and is more related to the functionally characterized database sequences of Firmicutes in clade 1 than to those of Gammaproteobacteria in clade 1.1 (Figure 36). Finally, the top BLAST hits of an amplified GH5_2 cellulase gene related to *Spirotrichonympha* but taxonomically derived from Gammaproteobacteria (with number 497) also showed divergent phylogenetic relationships; one top BLAST hit had an unknown taxonomic origin and was closely related to a functionally characterized database sequence of Firmicutes (not in clade 1.1, see Figure 36). The other top BLAST hit belongs to *Sphingomonas* spp. and was closely related to

the functionally characterized database sequence of *Cytophaga* (Bacteroidetes), but with low bootstrap-support (see Figure 36).

Figure 36: Clade 1 of the Maximum Likelihood-phylogenetic tree of 322 complete domain sequences of functionally characterized GH5-cellulases obtained from the CAZy database (60) **and the protein sequences of top BLAST hits of selected amplicon sequences (indicated in bold)**. The names of the top BLAST hits include: protein ID, functional annotation, taxonomic origin and information on which amplicon sequence they were derived from (BLAST hit, amplicon sequence number, phylogenetic relationship in Figure 35). The analysis included 425 alignment positions. Here, only the topology of sequences of clade 1 is shown (see Figure A3 in the appendix for the complete tree). The percentage of trees in which the associated sequences clustered together is shown next to the branches and was determined by bootstrapping with 100 replicates. Only bootstrap values higher than 25% are shown. The tree is drawn to scale, with branch lengths measured in the number of modelled substitutions per site. The taxonomic affiliation on phylum level of each sequence or collapsed cluster of sequences is indicated by a colour-filled marker. Collapsed branches consist of sequences from the same taxonomic group as indicated by the cluster name. Sequences belonging to clade 1.1 are outlined by a dark-grey box. The locations in the tree of the database sequences which were used for primer design are indicated by an arrow.

6 - Discussion

I - Assessment of the soil genetic diversity

In this research project, different approaches were applied to study the genetic diversity in agricultural soil; shotgun metagenome sequencing was employed to obtain a holistic view of the entire microbial community and its potential for the degradation of cellulose, while amplification and sequencing of the GH5 gene from soil-derived DNA was implemented for in-depth assessment of diversity. Here, the advantages and drawbacks of shotgun metagenome sequencing are discussed first, after which different methods for cellulase gene identification are considered. Finally, the results of the GH5 gene amplification study are discussed.

Potentials and drawbacks of shotgun metagenomic datasets

The aim of this research was to explore the genetic potential for cellulose degradation in agricultural soil under different tillage treatments, for which next generation-shotgun sequencing of the soil metagenome was employed. Cultivation techniques introduce bias by selection of only the cultivable part (143) of the microbial community (for example the active degraders) while phylogenetic marker gene-amplification techniques do not provide functional information and likely introduce bias in the microbial community composition (e.g. (144)). Furthermore, metagenomic-DNA insert-libraries and screening methods are labour-intensive and show low gene-detection frequencies (145–147). Shotgun sequencing of the microbial community DNA circumvents the mentioned drawbacks and provides a wealth of sequencing- and metadata which generates comprehensive information about the genetic potential of the soil microbial community; the putative ecological roles of the most abundant members of the microbial community can thus be described on a taxonomic and functional level. Also, the presence of specific genes can be verified and metabolic pathways can be reconstructed (148). More importantly, links between taxonomy and function can be made using reads which can be mapped to genes of which the putative taxonomic origin can also be predicted. This can reveal new ecological patterns and provide a basis for hypothetical microbial relationships and functionalities (149). Finally, comparing metagenomes from different environments can provide powerful statistical indications of a biome-specific genetic potential in microbial communities. For example, Berlemont and Martiny (52) showed that, across ~2,000 metagenomes from different biomes, the potential for utilization of carbohydrates was discernible according to biome-type. However, there are several pitfalls which, when not considered sufficiently, can undermine the power of the answers given from metagenomic data analysis. Most of these pitfalls are related to conclusions based on predicted gene abundances annotated using public database sequences. It is important to note that these annotations cannot be used as an indication of absolute abundances, as the qualitative and quantitative success of annotation is compromised by several factors (148).

First of all, the fraction of sequencing reads that can be taxonomically and functionally annotated is greatly limited by the available reference sequences in public databases (150). Databases such as the NCBI Genbank, -nucleotide and -protein databases contain many sequences originating from cultivation studies, from gene-isolation, gene-expression and biochemical characterization studies and increasingly from high-throughput sequencing studies of DNA from specific environments (151). The most reliable taxonomic and functional annotations originate from cultivation and biochemical

characterization studies, respectively. High amounts of reliable reference sequences are indeed available; the curated NCBI Reference Sequence (Refseq) database comprised ca. 18,000 species in 2012, of which around 64% were microbial (152) and the SILVA database currently contains ca. 6 million small-subunit ribosomal sequences (corresponding to ~700,000 estimated non-redundant sequences) (153). However, taken that only around 1% of bacteria in environmental samples are predicted to be cultivable (154) and that soils were estimated to contain millions of species (155), it is evident that significant knowledge on microbial diversity in natural samples is missing. Cultivation of fungi is associated with additional difficulties (156, 157) and they are therefore even less represented in the databases. The curated NCBI Refseq-database consists predominantly of bacterial or archaeal sequences (~59% of all accessions), whereas only ~7% of all accessions consist of fungal sequences (152). It is therefore not surprising that a significant portion of metagenome sequences cannot be taxonomically assigned, especially those with fungal origin. In the metagenomes analysed here, the percentage of reads which could be taxonomically assigned on kingdom level using the NCBI database was ca. 59%. Of these annotated sequences, 0.4-1.1% was assigned to eukaryotes and only 0.3-0.4% to fungi. Indeed, despite the estimated high abundance and diversity of fungi in soils (158), annotations of reads with fungal or eukaryotic origin are often very low in soil metagenomes (0.2-1.6% fungi (102, 159) and 0.9% eukaryotes of all sequences (160)). The NCBI database-bias is also illustrated by the higher numbers obtained when using ribosomal databases as reference (ca. 8-14% of all ribosomal sequences annotated as eukaryotes in the metagenomes studied here and in a forest metagenome (161)). The annotation success is generally higher for taxonomic than for functional annotations, as was observed here: 35% and 22% of all sequences in the conventional respectively organic farming metagenome could be functionally annotated. These numbers are not uncommon for soil metagenomes, as annotation percentages ranging from 13-69% of reads are often observed (160–164). The lower amount of functional annotations is probably due to the high amount and diversity of protein coding genes in microbial genomes; the knowledge that the average known bacterial genome consists of 5,000 protein coding regions and that the *Escherichia coli* pan-genome comprises over 60,000 unique and only ~3,000 core gene families (165), offers an indication of the predicted high gene diversity in nature. Although the gene-coding regions in microbial genomes are well predicted, the functional annotation of this variety of genes is not yet exhaustive. While the prokaryotic clusters of orthologous genes (COG)-database, which is based on comparisons of complete sequenced genomes, contained over 4,600 COGs in 2014 (166), today the KEGG database consists of ca. 25 million gene sequences retrieved mainly from the Genbank and Refseq databases and catalogued according to orthologous groups (~21,000) (167). However, only half to three-quarters of the prokaryotic genomes can be specifically functionally annotated using COGs (166). Indeed, even though these databases provide for a reliable functional annotation of a broader array of genomic or protein sequences (with KEGG also including eukaryotic sequences), they still do not seem to sufficiently cover the estimated existing diversity of protein coding genes.

In addition, the reliability of sequences obtained from the DNA-extraction, library preparation and sequencing process can be influenced by several biases. DNA-extraction biases influence the amount and proportion of DNA extracted from soil (160), in a way that a certain selection of a part of the microbial community takes place, e.g. active community (168). Furthermore, sequencing library preparation is prone to selection of DNA sequences with higher stability (due to mechanical shearing of sequences and size selection) and leads to artificial amplification of sequences in an exponential manner during the applied PCR-runs. The Nextera-library preparation from the Illumina platform has been noted to be biased towards GC-sequences

(169), especially with low template abundances and probably linked to PCR-amplification biases (170). Sequencing errors associated with the Illumina technology, e.g. substitution-type miscalls (171), inverted repeats or GGC-sequences (172), additionally influence the quality of obtained reads, depending on their taxonomic source. The 454-pyrosequencing platform, on the other hand, is associated with the introduction of insertions and deletions during sequencing of especially homopolymeric regions (173). DNA-extraction and library preparation biases do influence the reliability of obtained sequencing reads, but do not significantly influence the comparison of microbial communities from different environments, as was shown by Delmont *et al.* (160). Moreover, several of these biases can be accounted for during the process of read quality and filtering and the subsequent annotation cut-offs. Sequencing errors like erroneous base-calling are often reflected by the provided read quality information and low quality-regions or -reads can thus be removed during quality analysis and read filtering. Nevertheless, undetected sequencing errors can affect the annotation precision of metagenome reads. Codon degeneracy leads to a higher variability of nucleotide sequences than of the corresponding protein sequence. Therefore, comparison of nucleotide sequences to databases can lead to decreased homology scores, while protein sequence-comparisons allow for more robust comparisons to database sequences (174). Here, annotation was performed using predicted amino acid sequences derived from the metagenome reads. Nevertheless, sequencing errors like substitutions can lead to nonsynonymous changes to the DNA-sequences and insertion- or deletion-errors can introduce frameshifts in the protein-coding regions, thereby affecting gene-calling precision (175). Therefore, most gene-predicting or translation algorithms account either for sequencing platform-specific errors by entrainment (e.g. FragGeneScan (125)) or for common DNA-sequencing errors by comparison of all six reading frames (e.g. BLASTx (121) and DIAMOND (120)). While the 454-pyrosequencing platform is more susceptible to insertion-deletion errors, the greater average read length obtained with this platform is advantageous for annotation success. A dataset with longer sequences will have higher annotation chances caused by higher probability of including a (partial) conserved protein domain and by the fact that alignment-algorithms take read length into account when calculating expectation values (e.g. the BLAST-algorithm (121)). Read length-effects on annotation success might also be especially influential for annotation of eukaryotic sequences, as intronic sequences may occupy a large part of shorter reads.

Furthermore, the choice of sequencing data analysis methods and conclusions thereon should be based on preliminary diversity- and coverage-exploration. The predictive value of results from annotated gene abundances is greatly influenced by the sequencing depth. The amount of DNA sequenced must be great enough to ensure coverage of most microbial groups and groups of protein-coding genes known in the databases in order to compare annotation abundances between samples. Coverage is generally assessed by sub-sampling from the available annotated reads and counting the annotations reached at the corresponding sequence amount (176). The reported taxonomic assignment-based rarefaction analyses here (see Figures 8 and 18) show that sufficient coverage was reached to describe the diversity at the corresponding taxonomic level and compare the annotations between samples (177). One of the first efforts to sequence an agricultural soil metagenome, showed that coverage of the majority of COG-categories was reached after generation of 100 Mbp of sequencing reads (178). In this study of comparable environmental samples, approximately 60 Mbp for the conventional farming experiment and 1,000 Mbp for the organic farming experiment, per sample, were sequenced, suggesting that a similar coverage of protein-

coding genes was achieved. Nevertheless, Tringe *et al.* (178) estimated that a minimum of 2,000-fold amount of the already sequenced base pairs would be required to cover as much of the genetic diversity within their soil sample to obtain a draft genome of the most abundant genome. Indeed, when considering the average size of bacterial genomes (3-5 Mbp) and the genetic diversity observed even between closely related bacteria (165)), the above-mentioned sequencing depths are clearly not sufficient to cover the predicted genetic diversity in soil. Furthermore, these observations emphasize that, as discussed above, the descriptive power of taxonomically assigned sequences and especially available gene functions in the public databases is likewise rather limited. Therefore, coverage estimations based on comparisons to database sequences are increasingly controversial, as the annotation success depends on the genetic diversity in the sample and the extent to which it constitutes of (relatives of) cultured representatives in the databases; A high genetic diversity will decrease the chance of reliable taxonomic or gene-annotation, as a smaller part of the total diversity is likely to be represented in the public sequence databases. This drawback especially questions the validity of comparison of samples with a different genetic diversity. Therefore, estimation of the genetic diversity within a sample without comparison to database sequences is crucial for comparison of samples (176). The Nonpareil method (135) is one of the developed methods that estimate the genetic diversity within a metagenomic dataset without being biased by databases. This method estimates an average coverage based on the repeated calculation of the overlap between individual reads (redundancy) for different dataset sizes (135). In this study, the Nonpareil method was used as a proxy for genetic diversity and showed a low coverage for both metagenomes, indicating that much of the genetic diversity estimated to be present is not captured by the metagenomic datasets. Furthermore, this method showed for both metagenomes no difference in estimated overall genetic diversity between the samples of both tillage treatments. This means that, for this study, the relative abundances of annotated genes or microbial groups can be compared between tillage treatments. However, the Nonpareil-results also indicated that the soil samples from the organic farming experiment taken from two depths differed in coverage (surface soil horizon: 3.1% of diversity covered, deeper soil horizon: 4.2% of diversity covered). As no differences in sequencing depth existed between samples of both soil horizons, these coverage estimates suggest that the soil samples taken from greater depth contain a lower genetic diversity. When assuming that, in the samples from both depth layers, a similar portion of the genetic diversity has a significant similarity to database sequences, the observed differences in genetic diversity are in accordance with the observation that the relative counts for many microbial groups and protein-coding genes were higher in the samples from the greater depth (see e.g. Figures 19 and 20). Although these relative counts may reflect the true relative abundance for some of the microbial groups, they are unlikely to do so for all. Therefore, comparisons of relative abundances of annotated genes or microbial groups between both depth layers should be regarded with great care and should not be used as sole basis for conclusions.

Altogether, the factors discussed above limit not only the overall taxonomic and functional annotation of the metagenomes, but also the linkage between function and taxonomy which is made several times in this study. Consequently, a large part of the genetic information obtained here is regarded as unknown, but could actually be part of a known microbial group or group of protein-coding genes. Moreover, although with low probability considering the here achieved coverage statistics, chances exist that multiple assignments to protein-coding genes may in fact be multiple partial genes sequenced from a single original gene sequence (179). Therefore, it should be

stressed that the reported abundances here do not exclude an actual higher or lower abundance present in the metagenome. Furthermore, the soil metagenomes may exhibit additional groups of protein-coding genes or microbial groups to those reported here.

Cellulase gene annotation methods

By definition, cellulases are enzymes that catalyze the cleavage of cellulosic substrates, albeit with specificity for different regions on the cellulose polymer (e.g. crystalline or amorphous region). Nevertheless, cellulases show a considerable variability in protein fold and overall amino acid similarity (180), having led to their classification into different GH-families (59). However, most GH-families comprise not only cellulases but also cellulase-homologs with biochemical activities other than cellulolytic (www.cazy.org, (60)). Thus, cellulase identification in genomic datasets based on amino acid sequence similarity (e.g. BLAST) is prone to result in the prediction of both cellulases and non-cellulases. Although the significance of subtle differences in catalytic site architecture for true cellulolytic activity are increasingly appreciated for several GH-families (e.g. GH48 (70), GH5 (71), GH12 (181) and GH61/AA9 (182)), an adequate understanding of the enzyme characteristics leading to functional specificity are absent for most GH-families. Moreover, computational identification of catalytic site-architectural differences requires knowledge of complete protein sequence and fold (70, 183), whereas many genomic datasets (e.g. shotgun-metagenome or amplicon) comprise partial gene sequences.

In this study, profile-Hidden Markov Models (HMMs) were used instead to identify metagenome sequences putatively encoding cellulases. Domain-specific profile HMMs are commonly used to annotate cellulases from partial protein sequences by recognition of the protein domain architecture. HMMs are position-specific residue-scoring profiles built from alignments of multiple amino acid sequences with the same protein domain (126). These HMMs are more sensitive in detecting remote homologous sequences than local alignment tools, as the probability for multiple amino acid residues to be present at a certain position in the domain is taken into account (126). However, currently available HMMs from the Pfam database (184) or dbCAN (185) for detection of GH-family proteins are built from cellulase and non-cellulase protein sequences from the same GH-family and therefore often not specific for cellulases (183). Therefore, here a benchmarking effort was performed (see Table A1 in the appendix) for each GH-family containing cellulases, to identify the HMMs with highest cellulase-specificity from the Pfam database, the dbCAN or the self-built HMMs, generated from alignments of solely cellulase amino acid sequences. The benchmarking showed that, indeed, many self-built HMMs outperformed the HMMs obtained from public databases in recognizing specifically cellulase sequences, although still many non-cellulase sequences were recognized by the best-performing HMMs (Table A1 in the appendix). Therefore, additional filtering to identify true cellulases was performed by amino acid sequence similarity search to CAZy-database sequences (using BLAST (121, 122), see Figures 12 and 23). Using this method 5,906 (0.63% of) reads from the conventional farming metagenome and 225,849 (0.43% of) reads from the organic farming metagenome could be annotated to the selected cellulase domain families before the extra filtering step and 2,021 (0.21% of) respectively 73,694 (0.14% of) reads after the filtering step. These results are comparable to other studies mining for carbohydrate-active enzyme genes in metagenomes using HMMs for annotation; Hess *et al.* predicted 1.1% of a rumen shotgun metagenome to be carbohydrate-active enzyme genes (186). Metagenomic analysis of an enriched thermophilic cellulose-degrading sludge sample revealed approximately 0.9% of all

metagenome reads to consist of carbohydrate-degrading genes (187). A desert soil metagenome yielded ca. 0.5% of total sequence reads annotated as GHs (162) and a study of multiple soil metagenomes reported around 0.7% of combined reads to match carbohydrate active enzymes (159). However, these studies generally targeted all carbohydrate-active enzyme genes in the metagenomes, not only cellulase-containing domain families. Focussing on cellulolytic GH-families, Güllert *et al.* (188) could show an abundance of ca. 0.2-0.3% of sequenced reads from biogas fermenter, elephant faeces and cow rumen metagenomes to be cellulolytic genes. These results indicate that, using the applied method, a realistic amount of cellulase genes was captured in these metagenomes. Moreover, the applied methods (HMM and BLAST-search) are known to reliably annotate homologous protein sequences and to show low false-positive rates (174). In addition, Cardenas *et al.* show that sequence similarity search using BLAST with e-values of 10^{-4} and lower results in accurate prediction of CAZy-sequences (189).

However, while the HMM-BLAST method may show low false-positive rates, it does not ensure the minimization of false-negatives (174). Therefore, another method was explored for annotation of GH5-cellulases, namely a combination of sequence similarity search (ssearch) and sequence clustering (using Markov Clustering, MCL). Ssearch is a more sensitive similarity search method than BLAST and FASTA, but requires more computational power and computation time when including many sequences (137). Using this method, protein sequences predicted from the sequencing reads were subjected to pairwise alignment together with database protein sequences, resulting in similarity scores between any two sequences. Subsequently, the protein sequences were grouped according to customized alignment score cut-offs, e.g. expect-value, overlap, percentage identity (MCL), (138)). Ultimately, when using optimized clustering cut-offs, clusters of related database sequences and related environmental sequences were obtained. These two methods were roughly compared for their sensitivity in annotating conventional farming metagenome-reads to GH5-domain family cellulases. This showed that the ssearch-MCL method yielded a higher amount (682) of annotated reads to GH5-family cellulases than the HMM-method (196 before and 78 after the filtering step). However, of the sequence reads annotated by the ssearch-MCL method ca. 30% could be taxonomically assigned on phylum level by comparison to the NCBI database (see Figure 32), as compared to ca. 60% of reads annotated as cellulases by the HMM-BLAST method (see Figure 14). Still, most GH5-cellulase reads annotated using both methods were assigned to the same microbial phyla: Proteobacteria, Actinobacteria and Bacteroidetes, although the ssearch-MCL method captured a higher abundance of GH5-cellulases from Cyanobacteria and a lower abundance of GH5-cellulases from Chloroflexi than the HMM-BLAST method. As both annotation methods are based on similarity to CAZy-database sequences, it is not completely clear what caused these differences in taxonomic assignment. Nevertheless, there are chances that the somewhat different characteristics of both methods (higher sensitivity of the ssearch-MCL method and lower specificity of the HMM-BLAST method for GH5-family cellulases, see above) resulted in the capture of a slightly different array of GH5-sequences from the environmental sample. Indeed, even though the possibly different sequences captured by both methods are related to the same reference sequences, they do not necessarily have to belong to the same microbial groups; phylogenetic analysis of GH5-family sequences from the CAZy-database shows low phylogenetic conservation among related sequences ((100), Figure 33). Furthermore, the higher sensitivity of the ssearch-MCL method, which led to a higher amount of cellulase-annotated reads than by the HMM-BLAST method, may have resulted in the capture of many novel GH5-cellulase sequences which are not known in the database, accounting for the high proportion of taxonomically unassigned sequences. Indeed, the failure of

taxonomic assignment may either be due to non-significant similarity to sequences in the database or to significant similarity to sequences from different microbial phyla, leading to a failed agreement on phylum level-annotation (defined by the last-common-ancestor (LCA) parameter in MEGAN, see section 2: Materials and Methods).

Whether the increased amount of reads annotated as GH5-cellulases by the ssearch-MCL method are indeed derived from highly diverse, novel GH5-cellulase sequences or in fact false positives remains to be determined. Nevertheless, application of the same method to a dataset of amplified gene sequences obtained from the same soil samples (from the conventional farming experiment) for the identification of GH5-subfamily 2 cellulase genes, suggests a relatively high accuracy of annotation; using the ssearch-MCL method, 743 unique amino acid sequences predicted from the amplified gene sequences were annotated as GH5-subfamily 2 (GH5_2) cellulases. In-depth analysis of these sequences showed high likelihood of successful identification of putative GH5-cellulase sequences, as the majority of these (80%) contained the two conserved residues (see Figure 31 for conserved residues) which are involved in ligand binding in the active site of the cellulase (72). Furthermore, similar to the metagenome reads annotated as GH5-cellulases by the ssearch-MCL method, only a small proportion of the amplified GH5_2 cellulase genes could be taxonomically assigned (20%). To achieve a higher taxonomic assignment success, a less stringent taxonomic assignment cut-off was applied to the amplified GH5_2 cellulase genes. This led to an increase in the taxonomic assignment from 20% to 70% (Figure 32), illustrating the strong effect which ambiguous taxonomic origins of aligned database sequences can have on the taxonomic assignment. This increase in taxonomic assignments also led to the detection of increased phylogenetic diversity among the amplified GH5_2 cellulase genes (Figure 32). Indeed, taxonomic ambiguity of the amplified gene region ranged from differences in taxonomic origin at phylum level (e.g. similar to sequences from Proteobacteria and Firmicutes) to kingdom level, as some amplified gene sequences were similar to sequences from Actinobacteria but also to sequences from nematodes.

Nevertheless, although the ssearch-MCL method seems promising for accurate prediction of cellulase sequences, it may not be preferable for large dataset-analysis due to the relatively high computation time. In addition, other promising methods exist which could be further developed (e.g. peptide pattern recognition, (182)) or are currently being explored (e.g. random forest-mediated prediction, (190)). Nevertheless, great care must be taken with cellulolytic function-predictions even when based on complete-gene sequences, as product-isolation studies have shown that predicted cellulase genes (based on sequence homology) do not appear to code for the designated function (e.g. (191)). This stresses the fact that, in order to conclusively annotate function to gene sequences, the putative functions have to be experimentally verified by for example isolation of organisms with the predicted gene or expression and production of the gene in a host organism (146).

Phylogenetic diversity of GH5 cellulase genes in soil samples

One of the aims of this research project was to assess the abundance of GH5-cellulase genes quantitatively by performing quantitative real-time PCR (qPCR). The GH5-encoding gene was chosen because it is one of the most abundant endoglucanase-related GH-families found in soil (101–103). In addition, it is harboured by a vast array of microorganisms from many different phyla (CAZy

database, (100)) in relatively high copy numbers (73), offering the potential of a diverse and holistic view of the microbial potential to degrade cellulose in soil. However, GH5-genes show a high sequence variability with only seven conserved residues dispersed over the length of the catalytic domain (104). Nevertheless, the identification of subfamilies within the GH5-family (100, 192) has facilitated the comparison of different members of the GH5-family within subfamilies. Here, a degenerate primer pair targeting an internal region of a gene coding for the conserved domain of a GH5-cellulase was developed. The primer-design was based on the alignment of a metagenome sequence from the soil of the conventional farming experiment, it being one of the metagenome sequences with highest alignment scores to database sequences from GH5. As the putative endoglucanase-encoding metagenome read and the two database sequences used for primer design were annotated as proteobacterial GH5-cellulases from subfamily 2, the primer pair was expected to target GH5-genes from subfamily 2. Indeed, *in-silico* primer-specificity analysis showed high specificity of the primer pair for proteobacterial GH5-cellulases from subfamily 2. However, the high phylogenetic diversity within the GH5-family, even within subfamilies (100), led to the expectation that the phylogenetic diversity of the sequences captured from the soil sample would exceed that of the sequences used for primer design. Moreover, the inclusion of a metagenome-derived read in the primer development process, thereby introducing additional diversity of GH5-coding sequences from the environment and, therefore, increased primer degeneracy, led to the expectation that the sequence diversity captured would exceed that of the functionally characterized database sequences from GH5-subfamily 2.

Indeed, a high diversity of amplified sequences was obtained from the amplification of soil-derived DNA using this primer pair. When comparing the *in silico*- and *in vitro*-amplification results, the results strongly suggest that the diversity of genes amplified by the employed primer pair is considerably higher in the studied agricultural soil than that known from the databases. However, sequencing data analysis also showed that the amplification yielded mostly unspecific sequences, which could neither be classified as GH5-cellulases from subfamily 2, nor as GH5-cellulases from another subfamily. Thus, although the target region of the primer pair was located within the conserved domain of GH5-cellulases from subfamily 2 (see Figure 31), the developed primer pair is not specific enough to ensure reliable quantification of GH5-cellulases in environmental samples. Several factors might have contributed to the occurrence of mismatches during PCR: First, the high primer degeneracy (especially of the reverse primer) may have led to unspecific hybridization because of the broad melting temperature range inherent to degenerate primer sets (see review by Kalle *et al.* (193)). Second, mismatches in the 3' region of the reverse primer may have occurred due to the high amount of guanine (G)-residues in this region, considering that G-residue (purine) mismatches are more stable than cytosine (C)-residue (pyrimidine) mismatches (194). Third, the chances for unspecific binding might have been increased additionally by low abundance of the target sequence relative to the amount of DNA in the sample (193). This notion is supported by the observation of a single product in the single-template PCR using the GH5-gene from pure genomic DNA from *C. japonicus* as template (see Figure A1 in the appendix). Indeed, previous studies have reported successful development of primers for the amplification of GH5-genes from environmental samples (77, 195, 196), which likely contained a higher abundance of template sequences. For example, Pereyra *et al.* (195) analyzed the DNA from an anaerobic lignocellulose reactor, which in all likelihood was enriched with anaerobic cellulolytic microorganisms. Their degenerate primers were in fact designed based on cellulase reference sequences from anaerobic cellulose degraders belonging to three different microbial phyla. Moreover, these reference sequences all belonged to

GH5-subfamily 4, which contains mostly cellulases from anaerobic and rumen-dwelling organisms (100), indicating a more specialized environment enriched in target sequences. In addition, Nautiyal *et al.* (197) and Barbi *et al.* (77) reported successful development of primers for the amplification of transcripts of fungal GH5-cellulase genes obtained from chickpea roots respectively forest soil. When transcribing cellulase genes in response to available substrate presence, fungi are known to produce a multifold of transcripts with respect to the amount of cellulase encoding genes harboured (198, 199). Thus, when assuming induction of expression in the environmental samples of Nautiyal *et al.* and Barbi *et al.*, their transcriptome samples might also be considered enriched in template. Moreover, the transcript libraries employed by Barbi *et al.* contained predominantly mRNA due to their use of a poly(A)+ eukaryotic RNA-specific reverse transcriptase (77). Besides, their primers were developed based on fungal protein sequences from GH5-subfamily 5, which, as is shown by their own analysis as well, is a relatively conserved subfamily containing mostly fungal sequences (77, 100). Another study focussed on GH5-cellulase gene amplification from marine DNA-samples (196), which are known to be less diverse than soil samples (176, 178). In addition, they used primers which were developed based on multiple purely metagenomic sequences annotated as GH5-cellulases derived from a similar environment (marine samples) (196), indicating environment-specific targeting. Thus, these studies suggest that the template-primer ratio is a significant factor influencing the success of GH5-cellulase gene amplification, besides primer specificity. Nevertheless, in all of these studies, verification of primer-specificity was performed by sequencing of a limited number of cloned PCR-products. While in the study of Barbi *et al.* (77) the amplification products of other primers (for amplification of e.g. GH11-encoding transcripts) were analysed using high-throughput sequencing, the GH5-subfamily 5-amplification products were only analysed by cloning and sequencing of a subset of clones using Sanger technology. Moreover, they could show that primers (for amplification of GH11 and AA2-transcripts) proven specific for pure culture-DNA resulted in a considerable fraction of sequences harbouring a stop codon among the sequences analysed using high-throughput sequencing (77). Thus, as no high-throughput sequencing of the GH5-cellulase gene- or transcript-amplification products and, therefore, no exhaustive specificity-check of all PCR-products had been applied in the studies discussed above, it cannot be excluded that, in these studies, also non-specific products have been amplified from the environmental (c)DNA.

Nonetheless, a significant amount of the amplified sequences in this study could be predicted to encode GH5-cellulases from subfamily 2 (GH5_2 cellulases). The range of taxonomic affiliations of these sequences was, as expected, broader than that predicted by the *in-silico* primer-specificity analysis. Having targeted GH5-cellulases from subfamily 2, the amplified GH5 cellulase genes in this study were accordingly predicted to belong to mostly Proteobacteria but also to Actinobacteria, Firmicutes and Nematoda (Figure 32). This was not unexpected, as phylogenetic analysis of (functionally characterized) GH5-cellulases from the CAZy database (Figure 33 and (100)) showed that there is little conservation of GH5-cellulase sequences found within the same microbial phyla, or, vice versa, that similar GH5-cellulase sequences can belong to many different microbial phyla, even within the same GH5-subfamily. However, the low amount of amplified GH5_2 cellulase genes derived from Firmicutes is surprising, as the database sequences from GH5-subfamily 2 contain many sequences from Firmicutes (Figure 33). This observation may indicate that the employed primer pair exhibits amplification bias in favour of proteobacterial GH5-cellulases and against GH5-cellulases from Firmicutes. Nevertheless, from the metagenomic GH5-cellulase genes identified from

the same soil and using the same identification method likewise few were harboured by Firmicutes (Figure 32). Therefore, it seems more likely that relatively few GH5-cellulase-harbouring Firmicutes are present in this soil. Conversely, the GH5-cellulase genes harboured by microbial phyla which are detected in the shotgun metagenome but are not by amplification (e.g. GH5-cellulases from Cyanobacteria, see Figure 32), probably belong to other GH5-subfamilies. This is supported by the observation that GH5-cellulases from for example Cyanobacteria have to date been reported to belong to GH5-subfamilies 1 or 55 but none to subfamily 2 in the CAZy-database (www.cazy.org, (60)).

In addition, the results of this study show that the phylogenetic diversity of genes coding for GH5_2 cellulases in the studied soil is higher than that of functionally validated GH5_2 cellulase sequences in the CAZy-database (see Figure 35). Moreover, the amplification approach yielded a higher quantity of detected GH5_2 cellulase genes than the metagenomics approach, emphasizing the benefit of the amplification technique for diversity analysis. Furthermore, the results here show that the amplified GH5_2 cellulase sequences from this soil are generally closer related to each other than to database sequences. The same observation was made by Barbi *et al.* (77), who studied the phylogenetic diversity of fungal transcripts encoding AA2- and GH11-enzymes in forest soil by high-throughput sequencing, and by de Menezes *et al.* (200), who analyzed the actinobacterial GH48-gene diversity and quantity in pasture and woodland soil samples. These results all emphasize the lack of (functionally characterized) protein sequences in the databases. In addition, the phylogenetic analyses here show that the amplified GH5-region, which includes two conserved amino acid residues, is similar between sequences harboured by many different phyla and does not show a monophyletic topology (Figure 35). This non-monophyletic clustering of GH5-cellulases was already apparent from the reference trees with partial (Figure 34) and complete (Figure 33) database sequences, as was also shown by e.g. Aspeborg *et al*. (100). However, the predicted taxonomic affiliations of the amplified GH5_2 cellulase genes from agricultural soil indicate an even more pronounced polyphyletic nature of the GH5_2 cellulase sequences. Moreover, these phylogenetic patterns and the additional observed ambiguity in taxonomic assignments provide indications of possible HGT events. It should be noted, however, that the predicted taxonomic affiliations of the amplified cellulases here rely on database-homology analyses only, which is error-prone especially with sequences of short length. This method can overestimate HGT events and any indications of HGT should be verified by whole genome-analyses (201). In addition, phylogenetic tree analysis is a valid method to identify possible HGT events, but they are most reliable when using complete gene sequences containing the complete open reading frame (202), especially when assuming that HGT mostly occurs with complete genes or domains. Therefore, the observed indications of HGT in this study were further analyzed using complete protein sequences from the NCBI non-redundant protein database which showed highest similarity to the amplified GH5_2 cellulase sequences in question. This indeed resulted in some cases in the recognition of false HGT-indications, but in other cases led to additional interesting phylogenetic insights. For example, reference sequences of Nematoda (Heteroderinae) and amplified GH5_2 cellulase sequences derived from Actinobacteria show homology both in the phylogenetic tree based on partial GH5-sequences and in the tree based on complete GH5-sequences; while the amplified GH5_2 cellulase sequences in question are predicted to be derived from Actinobacteria (using an annotation pipeline based on all top 25 BLAST hits), their closest relative among functionally characterized partial database sequences is a GH5-cellulase from Nematoda (Heteroderinae) (Figure 35-E). Strikingly, the two top BLAST hits of these amplified GH5_2 cellulase sequences, which originate from cultivated

members of the Actinobacteria (*Micromonospora* spp., (203)), are also more similar to the reference sequences isolated from Nematoda (e.g. (45, 204)) and from cultured isolates of *Cytophaga hutchinsonii* (205) than to reference sequences from Actinobacteria (Figure 36). Interestingly, the amplified GH5_2 cellulase sequences exhibit ambiguous taxonomic assignments, as the top 25 BLAST hits of several of these amplicon sequences contained sequences originating from nematodes. Since *Micromonospora* have been identified as plant-endophytes (206) and Heteroderinae are known plant-parasites (207), they share the same plant-associated environment and show high potential to be in close proximity of each other. Therefore, it is not unimaginable that cellulase genes from these Actinobacteria may have been transferred to nematodes. Indeed, transfer of genes encoding plant cell wall-degrading enzymes from bacteria to nematodes has been suggested before (208). Interestingly, many GH5-cellulases in nematodes are found in combination with a CBM2-domain (204) and the closest homologues of these nematode-CBM2-domains were actinobacterial CBM2-domains (95, 204). However, transfer of the GH5-catalytic domain from bacteria to nematodes was also proposed by Danchin *et al.* (95), who identified a GH5-domain from *C. hutchinsonii* as the closest bacterial homologue. These results, together with the topology of the complete domain-phylogenetic trees shown here (Figures 33 and 36), suggest that both the catalytic and cellulose binding domain in the GH5-cellulase gene of nematodes may have been acquired from Actinobacteria or Bacteroidetes. Although separate acquisition of both domains by horizontal transfer and domain shuffling is theoretically possible and has been shown for bacterial genes with GH5- and CBM2-domains (obtained from actinomycetes (209)), research by Kyndt *et al.* (204) has not found support for separate acquisition or shuffling events of domains in the GH5-cellulase gene of nematodes. Nevertheless, the cellulase sequences from *Micromonospora* spp. (top BLAST hits of the amplified GH5_2 cellulase sequences harboured by Actinobacteria) also show high similarity to a database sequence without functional annotation from cultured *Sphingomonas* spp. (210) (Figure 36). Despite the lack of strong statistical support of the tree topology and the fact that this proteobacterial sequence is not functionally characterized, this observation suggests that GH5-cellulases from Proteobacteria (*Sphingomonas*) may also be possible sources of the GH5-cellulases transferred to nematodes. Accordingly, most of the amplified GH5_2 cellulase genes which are assigned to nematodes in this study are closely related to those assigned to Proteobacteria (see Table 4). Moreover, phylogenetic tree analysis by Kyndt *et al.* identified a GH5 catalytic domain from *Cellvibrio japonicus* as the closest homologue to the GH5 cellulases from nematodes (204). Finally, intracellular proteobacterial parasites (*Wolbachia* spp.) have been shown to be a source for HGT to nematodes (211), further supporting the possibility of proteobacterial GH5 gene transfer to nematodes.

Further indicative of HGT events are the amplified GH5_2 cellulase genes showing high similarity to the partial GH5-database sequence of the protist *Spirotrichonympha* (see Figure 35-C). These amplified sequences are predicted to be harboured by Gammproteobacteria, but also to Firmicutes and to a protist. The top BLAST hits of those assigned to Gammaproteobacteria and Firmicutes do not, however, show similarity to the complete GH5-cellulase sequence of the protist *Spirotrichonympha*, (Figure 36). Instead, they show high similarity to GH5-database sequences from cultured members of the Gammaproteobacteria (e.g. *Dickeya dadantii* (212)), Firmicutes (e.g. *Bacillus* spp. (213)) and Bacteroidetes (*C. hutchinsonii*, as discussed above (205)). This indicates that the close relationship between the partial reference sequence of *Spirotrichonympha* and the amplified GH5_2 cellulases derived from Gammaproteobacteria and Firmicutes was attributable to similarities in the amplified region only (Figure 35-B respectively -D) and not to complete-domain

similarities (Figure 36). Nevertheless, both the topology of the reference sequence tree (Figure 33) and the taxonomic ambiguity observed in the top BLAST hits of the amplicon sequences suggest possible evolutionary relationships between GH5 cellulases from Gammaproteobacteria, Firmicutes, Bacteroidetes and protists. Corroborating this, Danchin *et al.* (95) identified several GH5 cellulases from Bacteroidetes (*Bacteroides* and *Flavobacterium* spp.) to be the closest relatives to the same protist-GH5 sequence as was investigated here. Furthermore, based on phylogenetic analysis of GH5 genes obtained from symbiotic protist cDNA-libraries, Todaka *et al.* postulate HGT of GH5 genes from bacteria to protists, with the closest homologue being a GH5 gene from a Gammaproteobacterium (214). These genes may have originated from ectosymbiotic Bacteroidetes (215) or from intranuclear proteobacterial (216) residents of these protists.

Thus, phylogenetic analysis of amplified GH5_2 cellulase sequences from agricultural soil has provided additional incentive to further investigate the evolutionary origins of GH5-cellulases or the GH5-catalytic domain among eukaryotes (protists and nematodes) and bacteria (especially Actinobacteria, Proteobacteria and Bacteroidetes). Whether or not the indications of GH5-HGT events are specific for this soil remains to be explored. However, as the proposed transfers of GH5-cellulases from bacteria to nematodes or protists are thought to be ancient (95, 214), there is no indication that the observed putative phylogenetic patterns of GH5 cellulase genes in this soil are of a recent or local nature. Nevertheless, based on the results presented here it can be postulated that the investigated soil contains many cellulase sequences which may provide insight into the cellulase evolutionary history.

II - Influence of tillage on microbial cellulose degraders

This project focussed on the influence of tillage intensity on the microorganisms involved in the degradation of cellulose in soil. It was expected that soil under reduced tillage comprises a higher diversity and relative abundance of cellulose-degrading microorganisms than soil under conventional tillage, as a response to higher organic carbon input and content in the surface soil horizon under reduced tillage. To investigate this, soil was sampled of two field experiments which applied tillage treatments of different intensity: reduced or shallow non-inversion tillage and conventional or deep inversion tillage in either conventional or organic farming practice. The field experiment under conventional farming practices had been performed for an extended period of time (20 years) and the experiment under organic farming practices for a shorter period (4 years) of time. Here, the results of the analysis of the obtained shotgun-metagenomes of each field experiment are discussed.

Responses of the soil microbiome to tillage in the conventional farming experiment

Long-term tillage effects

In the long-term field experiment on a silty clay loam soil, different tillage intensities had been applied under conventional farming management for twenty years, allowing the investigation of the long-term effects of tillage on the soil carbon status and the soil microbial community. The surface soil horizon under reduced tillage was found to contain a higher amount of organic carbon and nitrogen than under conventional tillage and observations from other studies of tillage effects in silt loam soils in North-western Europe could thus be confirmed (217, 218), in a meta-analysis (219) as well from previous measurements on the same experimental site (24). Presumably, the long-term addition of fresh organic matter from crop residues in this soil layer resulted in higher amounts added in the surface soil horizon under reduced tillage than under conventional tillage (14). Depending on the soil type, organic matter at several stages of decomposition stabilizes in soil aggregates and binds to clay particles (29, 220, 221). The higher content of stable organic matter in soil induces favourable conditions for microbial growth, e.g. higher moisture content (222), as was also observed here, and higher availability of nutrients. In agreement with these processes, dissolved organic carbon levels were higher in the surface soil layer under reduced tillage than under conventional tillage in this study, although no differences in soil mineral nitrogen levels were observed here between tillage treatments. This favourable environment for microbial growth with high soil organic carbon levels therefore leads to higher amounts of microbial biomass (223) under reduced tillage than under conventional tillage, as was shown in numerous studies (217, 218, 224–226). Accordingly, higher amounts of microbial biomass carbon and nitrogen (Figure 6) and higher abundances of 16S- and ITS-rRNA genes were measured in this experiment in soil under reduced tillage than in soil under conventional tillage. As the number of ITS- or 16S-rRNA genes per ng DNA did not differ between the tillage treatments in this study (data not shown), these results indeed show that a higher bacterial and fungal abundance was present in the soil under reduced tillage than under conventional tillage. Moreover, the ratio of 16S- to ITS rRNA genes was higher in soil under conventional tillage, indicating enrichment for bacteria or suppression of fungi under conventional tillage compared to under reduced tillage. These results support the theory that the increased soil disturbance caused by more intense tillage (i.e. inversion in the conventional tillage treatment versus non-inversion in the reduced tillage treatment) leads to a reduction in fungal abundance,

possibly by disrupting their hyphal networks. Indeed, conventional tillage systems have been shown to have a negative effect on fungal abundances (227) and hyphal length (228).

A higher carbon availability also leads to a higher metabolic activity of the soil microbial community (229–231). This relationship is illustrated by the measurement of higher xylose- and cellobiose-metabolism in the organic soil horizon than in the mineral soil horizon in a forest soil (102). Furthermore, the observed linear relationship between polysaccharide hydrolysing activities and plant organic matter availability measured in lake sediments (232) also supports the idea of increased cellulolytic activity in soil with higher organic carbon contents. Indeed, higher potential activities of the extracellular enzymes xylosidase, β-glucosidase and cellobiohydrolase were found here in the surface soil under reduced tillage than under conventional tillage (Figure 7). Moreover, a higher potential cellobiohydrolase activity per unit microbial biomass was observed here under reduced tillage compared to conventional tillage, suggesting that the microbial community under reduced tillage was more specialized in degradation of large carbohydrates like cellulose than the microbial community under conventional tillage in this soil. Therefore, it was expected that those microorganisms involved in degradation of cellulosic (and cellulose-related) compounds were enriched in soil under reduced tillage relative to soil under conventional tillage and that this enrichment would be visible in a relatively higher genetic potential to degrade cellulose. However, no clear differences in relative abundances of genes involved in cellulose degradation between tillage treatments were observed (Figures 11 and 13). In fact, a high overall phylogenetic and functional diversity of cellulase-genes was observed in the metagenome of this soil, irrespective of tillage treatment. This indicates that the microbial potential for cellulose degradation is not strongly influenced by the investigated tillage treatments, even after twenty years of application. Although no other studies have specifically investigated the influence of different tillage intensities on the soil carbohydrate-active genetic potential, differences in carbohydrate degradation potentials between soils with different organic matter or nutrient contents have been observed. For example, a shotgun-metagenome study showed that the organic soil horizon in a forest soil contained relatively more glycoside hydrolase-genes than the mineral soil horizon (102). In another deep-sequencing shotgun-metagenomic study, Cardenas *et al*. (189) showed that cellulose-active oxidoreductase and glycoside hydrolase genes were more abundant in forest soil layer with a higher amount of organic carbon and other nutrient content. The fact that these studies achieved a higher sequencing depth than this study may indicate that not enough genetic diversity was covered here to detect differences between tillage treatments. Nevertheless, both studies mentioned were performed on relatively undisturbed forest soils, whereas the soil investigated here has been under recent influence of agricultural management practices.

Short-term tillage effects overruling long-term effects

In this long-term field experiment, after about twenty years, the short-term effects of soil disturbance by tillage and fresh organic matter incorporation may have been more important for the microbial community structure than the long-term tillage effects. Tillage itself leads to a priming effect of the microbial community by increasing oxygen availability and by the release of stable soil organic matter through destruction of soil aggregates (233). In combination with fresh organic matter incorporation (234), tillage leads to a sudden increase of bio-available old and fresh organic matter. This induces an increase in microbial activity which can last weeks or months in bulk soil (235). As soil sampling for this study was done around five weeks after tillage and maize straw incorporation, short-term effects of tillage and residue incorporation can be expected to a similar

extent under both tillage treatments. Therefore, the observed low degree of differences in microbial community composition between soils under the investigated tillage intensities is not altogether surprising.

Nevertheless, some differences in relative abundances of microbial groups and protein-coding genes between tillage treatments were observed in this experiment (Figures 19 and 20). Most of these microbial groups and protein-coding genes cannot be clearly related to cellulose degradation and are therefore apparently not as influenced by the recent incorporation of fresh cellulose but by more subtle differences in substrate availability affected by tillage treatment. These subtle differences are probably caused by the stage of fresh organic matter decomposition in the soil surface under each tillage treatment, influencing the succession of microorganisms with different life-history strategies. For example, it has been postulated that, during such pulse events of increased fresh substrate availability and microbial activity, *r*-strategists increase in abundance rapidly, whereas *K*-strategists remain in low abundance (236). Chen *et al.* (237) showed that, indeed, the initial degradation of fresh organic matter likely occurs by *r*-strategists. However, after depletion of the fresh organic substrate or nitrogen, the *K*-strategists obtain a competitive advantage, being better able to degrade more recalcitrant native soil organic matter (237). Furthermore, conventional tillage has been shown to favour *K*-strategists, also designated as slow-growers, whereas reduced or non-inversion tillage favours *r*-strategists, or fast-growers (238), indicating that tillage intensity can influence the successional stage of the soil microbial community. This tillage effect might be explained by an earlier exhaustion of plant residue-derived carbon in the surface soil under conventional tillage, because of lower initial amount of incorporated carbon than under reduced tillage. In surface soil under conventional tillage, therefore, the soil circumstances provide a competitive advantage for the *K*-strategists (239), degrading more recalcitrant organic compounds, at an earlier time point after residue incorporation than under reduced tillage. Thus, while at the time of sampling the *r*-strategists may have grown in abundance to a similar extent in soil under both tillage treatments, the *K*-strategists would have started to regain growth advantages under conventional tillage. Indeed, the microbial groups observed to be relatively more abundant under conventional tillage than reduced tillage in this metagenome (members of the Chloroflexi, Armatimonadetes, Crenarchaeota and Solibacteres (Acidobacteria)) have been associated with *K*-strategist characteristics, like slow growth rates or adaptations to relatively nutrient-poor environments; For example, microorganisms of the phylum Armatimonadetes were shown to be enriched in the mineral layer of a forest soil as compared to the organic layer (103). Moreover, the tillage effect observed on Armatimonadetes in this study could be traced back to genus level, where the *Fimbriimonas* showed the observed higher relative abundance under conventional tillage than reduced tillage. A member of this genus, *Fimbriimonas ginsengisoli*, was indeed shown to grow only in low-nutrient media (240). Furthermore, members of the phylum Acidobacteria and Chloroflexi were found to be enriched among very slow-growing soil bacteria (241) and to have relatively small cell sizes (96). Chloroflexi and Crenarchaeota, among other microbial groups, were found to be dominant in savannah soils with 38-times lower carbon and nitrogen content than grassland soils (242). Moreover, negative correlations between relative abundance of members of Chloroflexi and Acidobacteria and increasing soil phosphate or nitrogen contents (243) or increased nitrogen inputs (244) were found. In addition, the relative abundance of Acidobacteria and Armatimonadetes decreased along a pedogenic gradient from less- to well-developed soils, correlating with soil quality parameters like organic carbon and water-stable aggregates (245).

Also the observed tillage effects on protein-coding genes may be explained by the differences in stage of fresh organic matter decomposition under both tillage treatments. For example, genes coding for enzymes involved in fatty acid metabolism (especially long-chain-acyl-CoA dehydrogenase, ferredoxin-NAD+ reductase and cytochrome P450/NADPH-cytochrome P450 reductase (data not shown)) and methane metabolism (especially carbon mono-oxide dehydrogenases, catalase-peroxidases and serine- and coenzyme-biosynthesis (data not shown)) showed a higher relative abundance under conventional than reduced tillage. These genes mainly code for proteins involved in redox-enzyme reactions and reactive oxygen-stress. Dehydrogenases, cytochromes and ferredoxin are involved in energy production or modification of steroids, fatty acids, polyketides and some aromatic compounds (246, 247). As native soil organic matter has higher levels of amides, lipids and organic acids than fresh organic matter (248, 249), the observed increase in redox enzymes under conventional tillage may be related to a higher degradation of native soil organic matter relative to fresh plant residues. In addition, a higher relative abundance of genes coding for enzymes involved in xenobiotics degradation and metabolism (e.g. styrene-, toluene-, (amino)benzoate- and drug-degradation pathways) was observed here in soil under conventional tillage than under reduced tillage. Similar results were found by Degrune *et al.*, who observed a higher abundance of putative herbicide-degrading *Sphingomonas* bacteria in soil under conventional tillage as opposed to reduced tillage (250). Indeed, herbicides had been applied several times (Glyphosate, Bromoterb, Gardogold and Motivell) during growth of the maize plants on all plots of the experimental field studied here (personal communication (Georg Gerl, field records)). It is therefore not surprising that the microbial community in this soil exhibits many genes coding for xenobiotics- or drug-degrading enzymes. Following the same reasoning as above, the higher relative abundance of these genes in soil under conventional tillage is also likely due to a higher degradation activity of recalcitrant native soil organic matter or remnants of herbicides, due to an earlier limitation of easily accessible carbon.

Enzyme activities explained by transcriptional differences

The higher potential extracellular enzyme activity observed in this study under reduced tillage (Figure 7) might also be a result of the short-term effects of the applied tillage treatments. At the time of sampling, only a few weeks after the harvest of corn, the maize plant residues had been recently incorporated into the top 25 cm of soil in the conventionally tilled plots and into the top 8 cm of soil in the reduced tillage plots. As the soil under reduced tillage had been supplied with higher levels of fresh organic matter than the soil under conventional tillage (14), a higher amount of substrate might have led to a differential (post-) transcriptional regulation of enzyme production in the soil under reduced tillage compared to conventional tillage. This higher transcriptional or translational response could explain a higher availability of cellobiohydrolases. In forest soils it was shown that very low-abundant fungi can transcribe a high amount of cellobiohydrolase-sequences (78), suggesting a highly regulated transcriptional response of soil microorganisms to available substrate. Indeed, microbial cellulase production was shown to be strongly induced by the presence of cellulose or its degradation products (53, 55, 251).

Still, it would be expected that the observed higher levels of potential cellobiohydrolase activity leads to a higher relative abundance of the cellulose degraders because of competitive advantage in nutrient acquisition. One reason this is not observed in the metagenome-results might be the broad phylogenetic dispersion of the cellulose-degradation trait. Cellulose degradation has been differentiated into two discernible traits by Berlemont and Martiny (73); the “real” cellulose

degraders, which are bacteria able to produce “real” cellulases (endoglucanases and exoglucanases), and the cellulose utilisers, which are able to only utilise the degradation products (producing for example β-glucosidases). Although not a high proportion of the bacteria they investigated were cellulose degraders (ca. 24%, (73)), the cellulose-degradation trait is a broad trait which is shared among many different microorganisms (73, 74, 100), as is also evident from the CAZy database (www.cazy.org, (60)). This is also highlighted here, where the exoglucanases and especially endoglucanases are harboured by a large diversity of microorganisms (Figures 11 and 14). Whereas these microorganisms share the cellulose-degradation trait, they are likely very different regarding other functional traits influencing their competitiveness. For instance, based on their results, Chen *et al.* hypothesized that cellulose degradation-capability is not specific for *r*- or K-strategists (237). Besides the presumed functional redundancy of the cellulolytic trait, the *in situ* activity of cellobiohydrolase may be diminished by the presence of other recalcitrant plant cell wall-polymers present in the maize straw (252), which can decrease the access of cellulases to their substrate. This may reduce the growth advantage of cellulolytic microorganisms. Adding to the fact that extracellular enzyme action leaves room for non-cellulose degraders to utilise the released cellulose-oligomers offers a possible explanation why organisms harbouring genes encoding real cellulases do not show a competitive growth-advantage compared to the cellulose utilisers. Instead, all microorganisms with the ability to utilise cellulose- and other plant polymer-degradation products seem to be profiting from the increased availability of degradable organic compounds. Therefore, no strong changes in genetic potential to degrade cellulose are observed in the microbial communities.

Whereas the total amount of metagenome reads derived from the phylum Chloroflexi was higher in soil under conventional tillage relative to reduced tillage, the number of GH3-genes harboured by Chloroflexi was relatively higher in soil under reduced tillage compared to conventional tillage. This suggests that the members of Chloroflexi which harbour genes for cellulose utilization exhibit a different response to tillage than the majority of the Chloroflexi-members. In this study, GH3-genes were assigned to four Chloroflexi-genera; *Anaerolinea, Chloroflexus, Roseiflexus* and *Ktedonobacter* (data not shown). Only those harboured by *Chloroflexus* showed a higher relative abundance under reduced tillage than under conventional tillage. Conversely, the total amount of reads derived from *Chloroflexus* genus was not affected by tillage, whereas the total amount of reads derived from *Roseiflexus*, *Oscillochloris* and *Ktedonobacter* was found to be relatively higher under conventional tillage compared to reduced tillage. Thus, these observations suggest that *Chloroflexus* spp. harbouring GH3-genes are in advantage in soil under reduced tillage compared to soil under conventional tillage and that they show a different ecological response than most other genera of the Chloroflexi. This finding is not completely surprising as the phylum Chloroflexi contains many bacteria with different metabolic attributes (253). However, it is not clear if the observed ecological differences between genera is attributable their ability to utilise cellulose. After all, other genera of the Chloroflexi were also found to harbour GH3-genes in this metagenome. Rather, the response to tillage may be dependent on the cellulose-degrading capabilities of microorganisms on a species- or individual level. This is supported by the results of Berlemont and Martiny (73, 74), who found that potential cellulose degraders and utilisers formed phylogenetic clusters at the species respectively genus level.

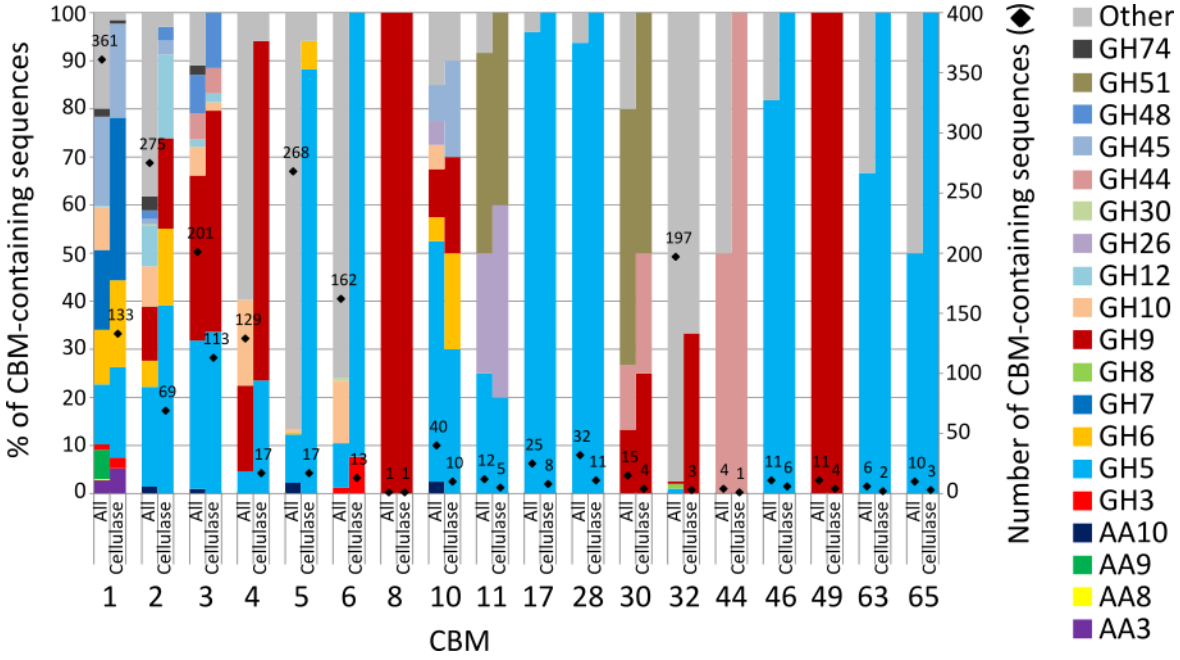


Figure 37: Percentage of CBM-containing functionally characterized sequences in the CAZy database with a nother domain on the same protein sequence, either a cellulase-relevant GH- or AA-domain or another/unknown domain (primary y-axis). This is shown per CBM-family for all functionally characterized sequences with that CBM-domain (All) or only for those of which a cellulase function has been characterized (Cellulase). In addition, the number of available functionally characterized sequences is given within each bar (secondary y-axis). Data was extracted from the CAZy database (www.cazy.org, (60), October 2017).

Responses of the soil microbiome to tillage in the organic farming experiment

In this field experiment conducted on clay loam soil, different tillage intensities had been applied for four years under organic farming management. In these four years before sampling, residues of two cover crops and two main crops (maize and beans) had been incorporated into the soil, but the residues of two winter wheat-crops had been removed (106). Due to the short time period in which the different tillage treatments had been applied and the absence of wheat straw incorporation, no significant differences in soil organic carbon levels were observed between tillage treatments or between the two investigated soil depths. Indeed, recently it was shown that in the same field experiment differences in total soil carbon content between tillage intensities were visible after two more years of experiment duration (254). In addition, D'Haene *et al*. (218) show that an effect of reduced tillage on the soil organic carbon levels in the surface soil horizon only becomes visible after a longer period of different tillage application and is related to the amount of crop residues used as carbon-input. For example, after six years of different tillage treatments in a silt loam soil in Belgium (255) no differences is soil carbon stock were observed. However, nine years of differential tillage treatment in a sandy loam soil in Ireland led to significant differences in soil carbon stocks in the top 30cm of soil (18). Nevertheless, in a loamy sand soil in Denmark, no differences in organic carbon stock were observed in the surface soil layer between conventional and reduced tillage even after ten years of treatment (15), indicating the importance of soil texture to tillage impact on soil carbon stocks. Furthermore, the fact that this field had been farmed under organic management since 2002 (106) may also explain the lack of tillage-effect on soil organic carbon contents, as organic farming is known to improve soil organic carbon levels compared to conventional farming; to illustrate, the

farming system-combinations “organic farming+conventional tillage” and “conventional farming+reduced tillage” have been shown to lead to the same increase in soil organic carbon levels as compared to “conventional farming+conventional tillage” (256). This effect of organic management on soil carbon content is possibly masking the effect of reducing tillage intensity in this experiment, at least after four years of treatment.

Whereas no significant tillage effect on soil organic matter content was observed here, the soil microbial biomass carbon and nitrogen were higher under reduced tillage than under conventional tillage in the surface soil horizon (Figure 17). This is surprising, as soil organic carbon and microbial biomass levels have been found to correlate (257). Nevertheless, as total organic carbon levels did show a trend towards higher levels under reduced tillage than under conventional tillage in the surface soil horizon, this correlation can be endorsed with the data presented here. Moreover, also in the soil horizon below the reduced tillage-working depth (10-16 cm), the soil organic carbon levels paralleled the microbial biomass carbon content, which showed no differences between tillage treatments in this soil horizon. The latter observation is consistent with results obtained in other studies (218, 224, 258), where likewise no differences in soil microbial biomass between tillage treatments were found in the soil horizon beneath the reduced tillage-working depth, in soils with varying texture (respectively silt loam, fine-sandy loam and sandy loam/loamy sand). Here, in the surface soil horizon, it appears that microbial biomass levels show an earlier response to differences in tillage intensity than soil organic carbon levels, as is also suggested by D’Haene *et al.* (218). The observed increased microbial biomass under reduced tillage in the surface soil horizon can, however, not be directly explained by differences in fresh plant residue additions between tillage treatments, as sampling at this site occurred five months after tillage, performed during spring weed management (personal communication), and ten months after plant residue incorporation from the former crop.

Influence of the rhizosphere on soil microbial community composition

Alternatively, the observed microbial biomass response to tillage intensity may be explained by differences in rhizosphere organic carbon content between tillage treatments. At the time of sampling, winter wheat had been harvested 1 week before with the removal of all above-ground plant biomass and without tillage, but leaving the wheat plant roots intact in the soil. As the total organic carbon content presented here was measured from soil samples which cannot be clearly designated as either bulk or rhizosphere soil, only a non-significant tillage effect could be observed. The combined sampling of bulk and rhizosphere soil has probably led to the mixing of heterogeneous soil compartments with regard to organic matter content. The differentiation between bulk and rhizosphere soil is of importance as tillage appears to impact the rhizosphere environment more than the bulk soil environment in terms of soil organic carbon and microbial biomass; in a study of the top 10 cm of a nine year-long field experiment, Yang *et al*. (259) did not find differences in bulk soil organic carbon content between different tillage treatments, but they did show higher organic carbon contents in wheat rhizosphere soil under no-tillage than under conventional tillage. In accordance, rhizosphere soil under reduced tillage has been shown to contain higher bacterial numbers compared to that under conventional tillage (225). Moreover, rhizosphere soil under no-tillage was shown to lead to higher root colonization with rhizosphere bacteria as compared to conventional tillage (260). In addition to this, increasing evidence suggests that the surface soil horizon under reduced tillage contains a higher density of roots than under conventional tillage, as winter wheat adapts its root distribution according to soil compaction under

the tillage horizon; For example, in a calcareous clay loam soil in Slovenia, reduced tillage with 6 cm-working depth led to a higher soil penetration resistance in the top 6-30 cm of soil compared to conventional tillage with 25 cm-working depth (261). In a silt loam soil in Belgium, winter wheat-root density has been shown to be higher under reduced tillage in the surface soil horizon (0-10 cm) but lower in the deeper soil horizon (20-30 cm) compared to conventional tillage (262). Other studies in a sandy loam (263) and silty clay loam (264) soil, investigating winter wheat root distribution at the ripening respectively anthesis stage of development, have shown a similar root length density in the surface soil horizon (0-10 cm) between reduced and conventional tillage treatments, but a lower root length density under reduced tillage in the deeper soil horizon (10-25 cm). Taken together, it may be assumed that, here, in a similar soil (clay loam) and at a similar stage of winter wheat development (shortly after grain filling and harvest), the soil contains a higher root density and, therefore, higher amount of rhizosphere-organic carbon content under reduced tillage than under conventional tillage in the surface soil horizon. This pattern is reversed in the deeper soil horizon, where the winter wheat roots would have grown denser under conventional tillage than under reduced tillage. Thus, the wheat rhizosphere has presumably selected a diverse soil microbial community capable of degrading various organic molecules derived from rhizodeposition, which includes extracellular polysaccharides, dead plant root cells and various root exudates, e.g. phytosiderophores, phenolic compounds, hydroxamic acids and short-chain fatty acids (see reviews by Dennis *et al*. (265) and Bertin *et al.* (266)).

Here, the majority of the most abundant microbial community-members were indeed found to be enriched under reduced tillage in the surface soil horizon or under conventional tillage in the deeper soil horizon, or both (see Figure 19). This tillage-responsive community consisted mainly of highly abundant microbial families including many rhizosphere-associated and plant-symbiotic bacteria from the Actinobacteria and Alphaproteobacteria. Microbial families from these classes of bacteria had indeed been found enriched in the wheat rhizosphere compared to bulk soil in a 16S rRNA gene amplicon study (267). In addition, *Micromonospora* are known plant-endophytes with plant growth-promoting abilities (206). Microbacteriaceae and Streptomycetaceae were found more abundantly in the organic soil layer than the mineral soil layer of a forest soil (103). Actinobacteria are known to produce varying sorts of antibiotics (268), providing pathogen control mechanisms for the plant. In addition, Actinomycetes are known to be able to degrade various recalcitrant or polymeric organic compounds, like pesticides (269) but also cellulose (270), illustrating their probable role in the degradation of dead plant root cells or root exudates. Members of the Alphaproteobacteria are also well-known nitrogen-fixing plant root endophytes (271) and are considered efficient utilisers of a broad range of organic carbon (272). Further microbial groups in this metagenome showing the aforementioned responses to tillage were families from the Beta-, Gamma- and Deltaproteobacteria; the Oxalobacteraceae and Comamonadaceae (Betaproteobacteria), the Xanthomonadaceae (Gammaproteobacteria), the Polyangiaceae, Cystobacteraceae, Labilitrichaceae and Myxococcaceae (Deltaproteobacteria). Members of the Burkholderiales (Betaproteobacteria) were found strongly enriched among wheat-rhizosphere bacteria and capable of phosphate solubilisation and siderophore production (273). They are furthermore well known from their role in nitrification, carrying out the first step as ammonia-oxidizers (274) and to be the major users of plant exudates (275). Pseudomonadaceae and Xanthomonadaceae are known as dominant members of the rhizosphere (273, 276, 277) and as plant-growth- promoters (271), biocontrol agents (*Lysobacter*) (278) but also phytopathogens (279). Members from the Myxobacteria are especially known for their competitive lifestyle producing

bioactive compounds and as bacterial predators (280) and may therefore function in the protection of the plant against phytopathogens. Finally, some other microbial families with a higher relative abundance under reduced tillage in the surface soil horizon or a higher relative abundance under conventional tillage in the deeper soil horizon were found which can be associated with the plant rhizosphere. Among them were a family of the phylum Verrucomicrobia (Opitutaceae), two families of the Bacteroidetes (Bacteroidaceae and Prevotellaceae), the Microchaetaceae (Cyanobacteria) and several important fungal groups like the Nectriaceae, Herpotrichiellaceae and Aspergillaceae (Ascomycota). Ascomycota (for example, members of the order Hypocreales) are known to be rapid rhizodeposition degraders (275). The role of Opitutaceae in the rhizosphere is not yet clear, but they have been found to be enriched in the rhizosphere of several plants, for example wheat (281), rice (282) and maize (283), and are shown to be highly cellulolytic (103) or to contain a high genetic potential for cellulose degradation (51). Moreover, they have been found to be more abundant in soil under reduced tillage than conventional tillage (281), corroborating the abundance patterns observed here in this experiment. Bacteroidetes are predominantly found in soils rich in organic matter and have been shown to be potent cellulose degraders (103). Moreover, Bacteroidetes and Cyanobacteria have been identified as rhizosphere bacteria of maize plants (276) and sugar beet (277). A summary of the tillage effects on microbial groups found in this metagenome can be observed in the overview Figure 38.

Rhizosphere-associated catabolic activities

Thus, many differences in the microbial community composition between soils under reduced or conventional tillage can be related to their association with the wheat rhizosphere. Protein-coding genes predominantly assigned to the rhizosphere-associated microbial community described above are related to nutrient acquisition and uptake, (recalcitrant) organic compound degradation and bacterial warfare; for example, genes involved in membrane transport (ABC-transporters) are more abundant among members of the rhizosphere-associated microbial community than among the rest of the microbial community members (data not shown). The same is true for genes involved in carbohydrate metabolism (butanoate, propanoate and glyoxylate metabolism), amino acid metabolism, xenobiotics metabolism ((amino)benzoate, bisphenol and polycyclic aromatic hydrocarbon degradation), metabolism of terpenoids and polyketides (limonene, pinene and geraniol degradation) and biosynthesis of secondary metabolites (novobiocin and tropane, piperidine and pyridine alkaloid biosynthesis) (data not shown). Most of these groups of protein-coding genes are also relatively more abundant under reduced tillage than under conventional tillage in the surface soil horizon and the other way around in the deeper soil horizon (see Figure 20). These observations clearly indicate that the microbial community in the soil compartments with a higher influence of the wheat rhizosphere is putatively more involved in degradation of various organic compounds and competition among each other compared to the microbial community in soil compartments with lower wheat root density. Indeed, in the rhizosphere soil of *Brachypodium distachyon*, a plant used as a model for wheat, higher relative abundances of genes involved in amino acid metabolism and xenobiotics biodegradation and metabolism were observed compared to bulk soil (284). Similar observations were made by Yang *et al.*, where the higher organic carbon content in mature wheat-rhizosphere soil compared to bulk soil coincided with higher catabolic activities in the rhizosphere soil than in the bulk soil (259). Moreover, they could show that, in the surface soil horizon, no-tillage led to a higher polymer degradation-activity of the soil microbial community compared to conventional tillage in both rhizosphere and bulk soil, and to a higher

amino acid- and carboxylic acid-degradation activity in rhizosphere soil (259). Furthermore, Turner *et al.* identified a higher abundance of cellulolytic and methylotrophic bacteria in the active community of the wheat rhizosphere compared to bulk soil (285).

Accordingly, the total amount of annotated cellulase-specific catalytic or binding domain families was relatively higher under reduced tillage than under conventional tillage in the surface soil horizon. In addition, several groups of cellulase-related genes specifically showed a tillage response; the highly abundant β-glucosidase genes (e.g. GH1) and several less abundant endo- or exoglucanase-related genes (e.g. CBM3, CBM4, CBM6 and GH26) were relatively more abundant under reduced tillage in the surface soil horizon. In addition, genes coding for endo- or exoglucanase-related carbohydrate binding modules (CBM4 and CBM65) were relatively more abundant under conventional tillage in the deeper soil horizon (see Figure 24). Similar trends were observed for the AA8-genes in the surface soil horizon and for the GH1-genes in the deeper soil horizon. These findings indicate a higher use of cellulosic compounds, especially cellulose-degradation products, under reduced tillage in the surface soil horizon and under conventional tillage in the soil horizon below reduced tillage-working depth. A summary of the tillage effects on cellulase genes found in this metagenome can be observed in the overview Figure 38.

Figure 38 (page 97): Overview of connections between and tillage effects on taxonomic groups (outer circles of nodes) and cellulase genes (circular groups of nodes located in the inner part of the circles) in the top 0-6 cm (A) or top 10-16 cm (B) soil of the organic farming experiment. Each taxonomic group (on different taxonomic resolution: from phylum to family level) and cellulase domain family or cellulase enzymatic group is represented by a node. Nodes of taxonomic groups belonging to the same phylum are connected. Furthermore, nodes of cellulase genes are connected to nodes representing the ten most abundant or tillage-affected microbial families harbouring that gene. These nodes are themselves connected to the corresponding nodes of microbial families within the outer circle of nodes. Gray nodes are not affected by tillage; yellow nodes are relatively more abundant under reduced tillage; blue nodes are relatively more abundant under conventional tillage in the corresponding soil horizon. Node border colour: red border indicates a significant tillage effect ($P<0.05$), whereas no red border indicates a tillage effect with lower significance ($P<0.10$) or not significant (gray nodes). The edges between nodes are green if the lower taxonomic group was found to harbour one or more cellulase genes for cellulose degradation ("degraders", harbouring e.g. endoglucanases) and red if the lower taxonomic group was found to harbour only cellulase genes for utilization of cellulose-degradation products ("utilisers", harbouring e.g. β-glucosidases). The node size corresponds to the relative abundances in the metagenome (% of metagenome reads x 10^{-3}) but in case of relative abundances <0.05 % x 10^{-3}, the square root of the relative abundances was used for visualization purposes. The phylum names of the connected taxonomic group-nodes are given or indicated with a number: 1) Nitrospira, 2) Chlorobi, 3) Spirochaetes, 4) Cyanobacteria, 5) Gemmatimonadetes, 6) Ignavibacteriae, 7) Thaumarchaeota, 8) Thermotogae, 9) Verrucomicrobia, 10) Acidobacteria, 11) Armatimonadetes, 12) Fibrobacteres, 13) Chlamydiae, 14) Deinococcus-Thermus, 15) Planctomycetes. The cellulase gene nodes are grouped according to cellulase enzymatic function: 16) β-glucosidase-, AA3- and AA8-genes, 17) cellobiose phosphorylase- and GH94-genes, 18) exoglucanase genes, 19) CBM genes, 20) endoglucanase-, AA9- and AA10 genes.

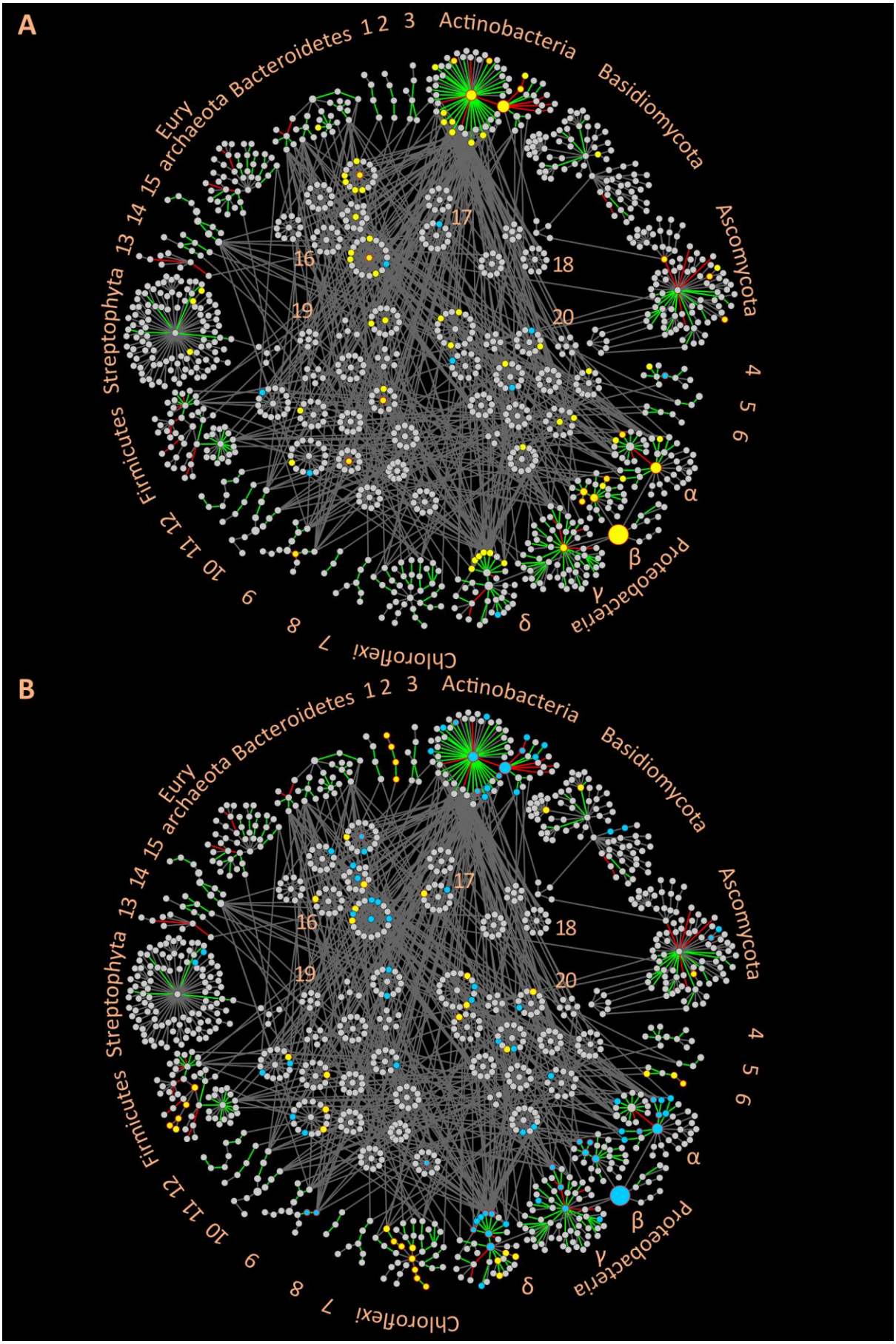

A
Actinobacteria
Basidiomycota
Ascomycota
Proteobacteria
Chloroflexi
Firmicutes
Streptophyta
Eury archaeota
Bacteroidetes
B
Actinobacteria
Basidiomycota
Ascomycota
Proteobacteria
Chloroflexi
Firmicutes
Streptophyta
Eury archaeota
Bacteroidetes

Figure 38.

Diversity of cellulolytic microorganisms affected by tillage treatment

The majority of the cellulase-encoding genes are also taxonomically harboured by the rhizosphere-associated microbial groups described above (see Figures 21 and 25). Sometimes, a large fraction of these genes are harboured by microorganisms showing a response to tillage (see Figures 22, 26 and 27), possibly identifying microbial members responsible for the tillage effects observed on the cellulase-coding genes. For example, several microbial groups to which β-glucosidase- or GH1-genes are assigned show a trend towards a higher abundance under reduced tillage in the surface soil horizon. These include mostly members of the Actinobacteria (the Microbacteriaceae (*Microbacterium*), the Nocardioidaceae (*Nocardioides*), the Thermomonosporaceae (*Actinomadura atramentaria*), the Intrasporangiaceae and the Micromonosporaceae) but also of the Alphaproteobacteria (Sphingomonadaceae (*Novosphingobium*) and Hyphomicrobiaceae (*Devosia*)), Deltaproteobacteria (Polyangiaceae (*Sorangium cellulosum*)), Verrucomicrobia (Opitutaceae) and Firmicutes (Bacillaceae). In addition, the Xanthomonadaceae (*Xanthomonas*), the Hyphomicrobiaceae and the Opitutaceae harbouring β-glucosidase-genes showed a trend towards higher relative abundance under conventional tillage in the deeper soil horizon.

Interestingly, the majority of the actinobacterial and alphaproteobacterial families to which β-glucosidase-genes are assigned and which show a response to tillage, are not known as cellulose degraders (i.e. harbouring endoglucanase genes (73)) and are also not found abundantly among the microbial families to which endoglucanase genes are assigned. Instead, actinobacterial families to which endoglucanase-related genes were assigned included low-abundant Cellulomonadaceae (*Cellulomonas*) to which endoglucanase-, CBM2- and CBM3-genes were assigned, Micromonosporaceae to which CBM3- and CBM44-genes were assigned, Solirubrobacteraceae to which CBM44-genes were assigned and Streptomycetaceae to which GH26-genes were assigned. The CBM- and endoglucanase-genes derived from these actinobacterial families showed a higher relative abundance under reduced tillage in the surface soil horizon or a higher relative abundance under conventional tillage in the deeper soil horizon. Notably, most of these actinobacterial families are not the same as the tillage-responsive Actinobacteria to which β-glucosidase genes were assigned. In addition, some (the Solirubrobacteraceae and the Streptomycetaceae) do not show a response to tillage when considering their annotation in the whole metagenome (see Figure 19). The greatest portion of CBM3-genes was derived from Micromonosporaceae, indicating that this microbial family may be responsible for the observed tillage effect on the abundance of CBM3-genes. However, of the other endoglucanase-related genes mentioned (CBM2, CBM3, CBM44 and GH26), the part assigned to Actinobacteria and showing a tillage response was rather low. These results illustrate that the members of the Actinobacteria did not contribute much to observed tillage effects on endoglucanase-related genes. Furthermore, most of the Actinobacteria which contributed to the observed tillage effect on β-glucosidase-genes probably do not contain endoglucanases, indicating that the majority of the rhizosphere-selected (and, thus, tillage-responsive) Actinobacteria are not particularly involved in long-chain cellulose-degradation but rather in the degradation of cellulose oligomers or chemical modifications thereof. Nevertheless, exoglucanase genes (including GH6- and GH48-genes) harboured by Actinobacteria did show a trend towards relatively higher abundance under reduced tillage in the surface soil horizon, but this observation could not be determined on lower taxonomic levels of the Actinobacteria. Moreover, Actinobacteria to which the exoglucanase genes were assigned constituted the majority of the exoglucanase genes, and included Micromonosporaceae, Streptomycetaceae, Mycobacteriaceae and Cellulomonadaceae (see Figures 21 and 25). Furthermore, the total amount of reads assigned to both the Micromonosporaceae and

Cellulomonadaceae in the whole metagenome (see Figure 19) was higher under reduced tillage in the surface soil horizon. Taken together, it appears that the actinobacterial families Micromonosporaceae and Cellulomonadaceae might in fact play a role in cellulose degradation in the rhizosphere environment.

Several β-glucosidase genes were harboured by families of the Delta- and Gammaproteobacteria, also showing a response to tillage intensity. Of these, the Polyangiaceae (Deltaproteobacteria) and Xanthomonadaceae (Gammaproteobacteria) were additionally found abundantly among the microbial families to which endoglucanase genes and genes with CBM-motifs were assigned. Moreover, CBM-genes derived from Deltaproteobacteria reacted to tillage; CBM4- and CBM6-genes assigned to Polyangiaceae (*Sorangium cellulosum*) showed a trend towards higher relative abundances under reduced tillage in the surface soil horizon, while CBM65-genes assigned to Cystobacteraceae (*Cystobacter fuscus*) and endoglucanase genes assigned to Labilitrichaceae (*Labilithrix luteola*) tended towards higher relative abundances under conventional tillage in the deeper soil horizon. In addition, these microbial families were among the most abundant ones to which CBM4- and CBM65-genes were assigned, indicating that they were primarily responsible for the observed tillage effect on the relative abundances of these genes (see Figure 24). Polyangiaceae (especially from the genus *Sorangium*) were found enriched in the rhizosphere soil of *B. distachyon* (284). In addition, *Sorangium* spp. are capable cellulose degraders (286) and were found to be more abundant in wheat rhizosphere and bulk soil (0-7.5 cm) under conservation tillage compared to conventional tillage (281). In addition to members of the Polyangiaceae (287), also the myxobacterium *Sandaracinus amylolyticus* has been shown to harbour cellulase genes (288) and others may follow, as the myxobacteria have yet to be thoroughly investigated (287). Thus, our results suggest that the three families of the Deltaproteobacteria are involved in cellulose degradation in the rhizosphere environment of this soil. Nevertheless, in this study, no tillage effect could be detected on endoglucanase genes assigned to Polyangiaceae (e.g. GH5-genes). Likewise, most of the endo- or exoglucanase genes derived from other members of the Proteobacteria did not show an effect of tillage, despite them being among the most abundant microbial groups harbouring endoglucanase genes, e.g. Cystobacteraceae harbouring GH5-modules (see Figure 27).

Finally, some other microbial groups (Opitutaceae, Firmicutes) to which β-glucosidase genes were assigned in this study and which showed a tillage response, were found among the microbial families to which endoglucanase genes and genes with CBM-motifs were assigned. In fact, CBM6-genes assigned to Opitutaceae (*Opitutus terrae*) were relatively more abundant under conventional tillage in the deeper soil layer, while GH9-genes assigned to Firmicutes showed a higher abundance under reduced tillage in the surface soil horizon, although not significant after *P*-value correction. Similar to *Sorangium* spp., *Opitutus* spp. were found to be more abundant in wheat rhizosphere under conservation tillage compared to conventional tillage (281) and to be highly cellulolytic (103). Their potential role as important polysaccharide degraders was also highlighted by a study showing a significant enrichment of genes coding for glycoside hydrolases in their genomes (289). Also members of the Firmicutes (*Bacillus* spp. and Clostridia) are known to be potent cellulose degraders (290, 291). The Opitutaceae and members of the Firmicutes (most likely the Bacillaceae) may therefore also be involved in the cellulose degradation activity in the rhizosphere at this site. A summary of the connections found between microbial groups and cellulase genes in this metagenome can be observed in the overview Figure 38.

Role of endoglucanase- and CBM-harbouring microorganisms in rhizosphere soil

Thus, while several microbial families to which endo- or exoglucanase genes are assigned show a weak tillage effect, the majority does not appear to be affected by tillage. This is also evident from the total annotations to the endo- or exoglucanase domain families (see most GH-families in Figure 24) and cellulase enzymatic groups (Figure 21); no clear indications are found of increased endo- or exoglucanase gene abundances under reduced tillage in the surface soil horizon or under conventional tillage in the deeper soil horizon. Moreover, while the total amount of metagenome reads assigned to certain microbial groups are affected by tillage (e.g. annotations to Xanthomonadaceae or Cystobacteraceae, see Figure 19), the amount of cellulase reads assigned to the same microbial groups do not show a tillage effect (e.g. endoglucanase genes assigned to Xanthomonadaceae, see Figure 22, or GH5-genes harboured by Cystobacteraceae). These results indicate that the cellulolytic capability of these endo- or exoglucanase-harbouring microorganisms may not be the selective trait explaining their abundance patterns as rhizosphere-dwellers**.** The exceptions may be the Micromonosporaceae, Cellulomonadaceae, members of the Deltaproteobacteria, Opitutaceae and members of the Bacillaceae. Similar results have been reported by Berlemont *et al.* (12), who found that bacterial sensitivity to drought was the actual trait driving the observed cellulolytic potential in soil. Altogether, these results on DNA-level appear to suggest that no extensive degradation of cellulosic compounds leading to competitive advantage of cellulose degraders had taken place in the time before sampling. This despite the recent wheat plant-harvest, which is expected to remove plant-derived protection against root degradation by the plant immune system (292). In a study on degradation rate of different winter wheat plant components by soil microorganisms, only 28% of wheat root dry matter was decomposed in the 2.5 months after the start of incubation (compared to 65% of wheat stubble, (293)). This low initial root-degradation rate coincided with a low amount (0.1 million cfu) of cellulose degrading microorganisms measured on the root surface 7.5 months after the start of incubation and a much higher amount (6.8 million cfu) after 19.5 months of incubation, around the time where the majority of the roots (>50%) was decomposed (293). These measurements support the idea that the degradation of winter wheat roots had not started in a significant intensity in this soil at the time of sampling. Also the released amounts of cellulose available from plant root-cell walls in rhizodeposition (265) were apparently not high enough to provide a selective advantage for those microorganisms able to not only utilise but also degrade cellulosic compounds. However, as discussed in the previous chapter, cellulose degradation might have taken place through higher abundance of cellulase gene transcripts and enzymes which were not measured in this study. Nevertheless, if this was the case, it did not lead to a detectable enrichment of cellulose degrading microorganisms. Alternatively, the absence of a visible response in the endo- or exoglucanase gene abundances may arise from a lack of coverage of the soil genetic diversity combined with the low phylogenetic depth of this trait (73). Although a considerable amount of soil-derived DNA has been sequenced in this study (16.2 Gbp), coverage analysis indicated that much of the genetic information in the soil had not been sequenced. As endoglucanase- and especially exoglucanase-genes occur in much fewer copies than β-glucosidase genes in bacterial genomes (73), it is possible that many endo- and exoglucanase genes were not detected in this metagenome, despite the great sequencing effort performed. Moreover, the achieved coverage did not allow for a reliable analysis of the microbial community at a taxonomic level lower than the family-level. Therefore, cellulolytic microorganisms on low taxonomic level which are not affected by tillage (possibly because they are not associated with the rhizosphere) can confound the effects observed on the cellulase genes

harboured by the tillage-responsive microbial community. Finally, tillage effects on endo- and exoglucanase gene abundances may have been confounded by the heterogeneity in the soil samples, which were not separated according to rhizosphere or bulk soil.

Interestingly, in the soil metagenome studied here, several CBM-families show a response to tillage intensity (Figure 24). As CBMs are generally part of endo- or exoglucanase proteins and not of β-glucosidases (www.cazy.org, (65) and see Figure 37), it is surprising that they show a tillage response whereas most GH-families do not. The GH-domain family which shows a tillage response and includes endoglucanase sequences (GH26) is found in combination with CBM11 only (see Figure 37), which is a CBM-family not affected by tillage in this study. As the KEGG-annotated endoglucanases or exoglucanases also do not show a reaction to tillage (Figure 21), the annotated tillage-responsive CBMs might be part of yet unknown cellulases or of cellulose-binding proteins with functions other than cellulose degradation. For example, proteins containing only a CBM-module (CBM2) in the cyst nematode *Heterodera schachtii* have been found to play a role in plant cell wall degradation by interacting with the plant's own pectin-methylesterase (294). In this soil, however, no CBM2-genes harboured by nematodes have been found. Nevertheless, cellulose binding proteins (CBPs) or independent CBMs can also occur in bacteria (295), possibly explaining the lack of coherence between observed tillage-effects on CBM- and GH-domains. Another protein with cellulose binding- but not cellulolytic capabilities is the cellulosome-scaffoldin. When considering that CBM3 is one of only two known cellulosome scaffoldin-CBMs (90) and that the composition of the cellulosome can change according to the encountered substrate (80, 81), a possible explanation might be that, here, the tillage-affected cellulosome-scaffoldins contain different compositions of GH-enzymes which then do not appear affected by tillage. However, the relative abundance of CBM3-genes assigned to anaerobes (mostly Ruminococcaceae) is low and not affected by tillage, indicating that the tillage effect observed on the relative abundance of CBM3-genes is not caused by different abundances of cellulosome-harbouring bacteria. Another hypothetical explanation might be the observed role of CBM-like components in sensing the extracellular plant residue degradation status (discussed in (90)). Some cellulosomal genes have been shown to be transcriptionally regulated by sigma factors which contain such CBM-like components (296, 297) and similar sigma factor-gene pairs are also found in non-cellulosomal bacteria (*Bacteroides thetaiotaomicron*, (298)). Thus, in this study, microorganisms to which the CBM-genes affected by tillage (CBM3, CBM4, CBM6 and CBM65) were assigned might be indeed involved in sensing the nutritional status of their environment, providing a competitive advantage in the rhizosphere environment.

Comparison of metagenome-analysis results from both field experiments

Description of each field experiment

This project aimed to investigate the influence of different tillage intensities on the cellulolytic potential of the soil microbial community, by means of shotgun-metagenome-analysis of soils from two different agricultural experiments located in Western Europe. By including datasets from two agricultural experiments, generated using different sequencing platforms and differing in soil type, farming practice and experiment duration, a robust assessment of tillage intensity effects on the soil microbiome was possible. These two agricultural experiments both applied either conventional soil tillage with a working depth of 20-25 cm using a mouldboard plough or reduced non-inversion tillage

with a working depth of 5-8 cm using a rotary harrow. From both experiments the surface soil layer (up to the reduced tillage-working depth) was sampled. From the organic farming experiment, additionally the soil layer beneath the reduced tillage-working depth was sampled to obtain a complete view of the soil conditions under both tillage treatments. However, for comparison of both experiments, only the results of the surface soil layer will be compared. The experiments mainly differed in experiment duration and farming practice, where one experiment was managed using conventional farming practices including chemical fertilizer, fungicide- and herbicide-use since twenty years (as described by (24)) and the other experiment was managed using organic farming practices with cattle slurry as fertilizer and weed management by tillage since four years (as described by (106)). These differences in farming practices (especially of fungicide use) are clearly reflected in the abundances of soil fungi in both experiments. Abundances of fungal ribosomal sequences were considerably lower in the soil of the conventional farming experiment (ca. 4.5×10^8 ITS rRNA gene copies per gram soil) than in the soil of the organic farming experiment (ca. 6.0×10^9 ITS rRNA gene copies per gram soil). This was expected from the known positive effects of organic farming practices on fungal abundances (97, 99). Mean annual precipitation and temperature values are similar for the two sites, although both were slightly higher for the organic farming experiment. The conventional farming experiment was performed on a silty clay loam (Luvisol, pH=6.3) and the organic farming experiment on a clay loam (calcareous Cambisol, pH=7.9) with a similar clay-percentage (25%) but with lower content of silt and higher of sand than that of the conventional farming experiment. Total soil organic carbon and nitrogen levels were slightly higher for the organic farming experiment (1.8% respectively 0.17%) than for the conventional farming experiment (1.4% respectively 0.15%). While dissolved organic nitrogen levels did not differ between sites, dissolved organic carbon levels were much higher in the soil of the organic farming experiment (on average 118 µg g^{-1} dry weight soil) than in the soil of the conventional farming experiment (on average 5 µg g^{-1} dry weight soil). This is presumably caused by the use of cattle slurry for fertilization in the organic farming experiment versus chemical fertilizer-use in the conventional farming experiment, as cattle slurry contains high amounts of bio-available carbon and was shown to significantly increase water-extractable organic carbon in grassland soil (299). Fertilization had taken place around five months before sampling in the conventional farming experiment and four months before sampling in the organic farming experiment. Sampling of the conventional farming experimental plots was done in mid November (soil temperature of 7°C), ca. five weeks after corn harvest and soil tillage with maize plant residue incorporation. In contrast, sampling of the organic farming experimental plots was performed in the beginning of August (soil temperature of 22°C), ca. five months after tillage for weed control and one week after winter wheat harvest without plant residue incorporation. The differences in soil temperature are probably a main cause of the higher microbial biomass measured in the soil of the organic farming experiment (0.8 mg C_{mic} g^{-1} soil on average) than in the soil of the conventional farming experiment (0.22 mg C_{mic} g^{-1} soil on average). In addition, time after harvesting may also have influenced the microbial biomass-differences, as it has been shown that soil microbial biomass is highest in soil during the grain filling stage of wheat growth, with a sharp decrease shortly after harvest and removal of above-ground plant biomass (300).

Similarities between the agricultural experiments

Despite the differences in soil properties between the two agricultural experiments, increasing tillage intensity had a negative effect on the microbial biomass in the surface soil horizon of both

experiments. This coincided with lower soil water content under conventional tillage in both experiments, corroborating results showing the negative effects of tillage on soil humidity and microbial biomass levels (217, 218, 224–226). Furthermore, a striking similarity in microbial community structure was observed between the metagenomes of both agricultural experiments. In both metagenomes, Proteobacteria were most abundant, followed by Actinobacteria. Although preceded by Bacteroidetes in the conventional farming experiment, Acidobacteria were most abundant after Actinobacteria. Planctomycetes, Verrucomicrobia, Gemmatimonadetes and Chloroflexi were among the top eight most abundant bacterial phyla in both metagenomes. Furthermore, the Actinomycetales and Rhizobiales were highly abundant in both soils. In agricultural soil, varying community compositions are found in 16S-pyrosequencing studies; Portillo *et al.* reported a dominance of Firmicutes and Actinobacteria followed by Proteobacteria in an agricultural soil in Michigan, USA (V4-region, annotated to the Greengenes-database) (96), while Acidobacteria and Verrucomicrobia were most abundant in agricultural soils of Argentina (V4-region, annotated to the Greengenes-database) (301) and Proteobacteria, Acidobacteria and Bacteroidetes in an agricultural soil in Belgium (V1-V3-region, annotated to the SILVA-database) (255). Shotgun metagenome studies (annotated using all reads) tend to show a more constant microbial community structure in agricultural soils; in an agricultural soil in Brazil, the most abundant microbial groups were Proteobacteria, Actinobacteria and Acidobacteria (shotgun-pyrosequencing, annotated to the NCBI-NR-database) (302). In addition, Delmont *et al*. identified Proteobacteria as the most abundant phylum followed by Actinobacteria and Acidobacteria in a grassland soil in Europe (shotgun-pyrosequencing, annotated to the SEED-database) (160). In forest soil, however, the most abundant microbial groups were Proteobacteria and Acidobacteria, followed by Actinobacteria (shotgun-pyrosequencing, annotated to the SEED-database) (102).

Furthermore, relative abundances of enzyme-coding genes are very similar between the two investigated metagenomes. In both metagenomes, genes coding for enzymes involved in carbohydrate-, amino acid- and energy-metabolism were most abundant, followed by signal transduction and membrane transport. On pathway-level, genes coding for two-component systems and ABC transporters or for enzymes involved in breaking down and synthesizing nucleotides were most abundant. Similar results were obtained in other studies of agricultural soils; in a farm soil in the USA, genes coding for two-component system, proteasome and inositol-phosphate metabolism were among the most abundant (short-insert metagenome library, annotated to the KEGG-database) (178). In an agricultural soil in Argentina, genes involved in energy production and conversion, amino acid transport and metabolism and protein posttranslational processes were most abundant (shotgun-pyrosequencing, annotated to the COG-database) (301). Furthermore, in an agricultural soil in Brazil (shotgun-pyrosequencing, annotated to the SEED-database), genes annotated to the functional categories of clustering-based subsystems, carbohydrates and amino acids and derivatives were most abundant (303).

In addition, the functional diversity of cellulase genes was similar between the two metagenomes studied here (see Figures 11 and 21 and Figures 13 and 24); the abundance of β-glucosidase- and endoglucanase-coding genes was higher than exoglucanase-coding genes. Furthermore, GH1, GH3 and CBM2 were the most abundant cellulase domain families in both metagenomes. In addition, in both soils a high abundance of cellulose phophorylase- (GH94) and cellulose oxidoreductase- (AA8) coding genes were detected. Similar results have been found previously in a forest soil metagenome (304), a desert soil metagenome (162) and in an agricultural soil metagenome focussing on all carbohydrate active enzymes (178). However, no study is known at

present where cellulase genes have specifically been investigated in agricultural soil metagenomes. The phylogenetic diversity of cellulase-gene harbouring microorganisms found here in both soil metagenomes emphasizes the important role that Actinomycetes, Gamma- and Deltaproteobacteria and Bacteroidetes play in cellulose degradation, whereas Firmicutes show a much lower contribution to the observed cellulase gene abundances. Also Rhizobiales (Alphaproteobacteria) and Burkholderiales (Betaproteobacteria) show a high contribution to the cellulase gene abundances. In addition to the well known cellulase gene harbouring microorganisms, other less-well characterized phyla have come forward in this analysis as cellulolytic, for example Chloroflexi, Cyanobacteria (Chroococcales), Planctomycetes (Planctomycetales), Chlorobi and Verrucomicrobia (subdivision 3 and Opitutales).

Thus, the results obtained here and in other studies suggest that, within similar environments (e.g. agricultural soil), dominant microbial groups and functions are also similarly abundant. Indeed, comparison of metagenomic libraries from different environmental samples (i.e. soil-, whale fall-, sea- and acid mine drainage-samples) by Tringe *et al.* revealed that each environment could be functionally profiled, showing an environment-specific functional fingerprint (178). A similar study, comparing 30 metagenomes of different types of soil (i.e. forest, grassland, tundra, semiarid and desert), showed that within-biome metagenomes show high correlation of functional profiles and that a biome-specific microbial community structure was clearly differentiable (159). This study by Noronha *et al.* further showed that the biome-specific functional profiles were indicative for the respective environmental conditions (159). As soils with a similar land-use type (e.g. agricultural soils) may also be subjected to different environmental conditions, caused by differences in e.g. soil type and structure or climate, it is not self-evident that the agricultural soils investigated here show high similarity in terms of taxonomic and functional microbial community composition. For example, the difference in sampling time points over the yearly season between both agricultural experiments has clearly affected the soil temperature (i.e. 7°C in the soil of the conventional farming experiment and 22°C in the soil of the organic farming experiment). Nevertheless, the metagenome results suggest that the described differences between experiments in soil properties and sampling time point (season) do not greatly influence the microbial community composition. Accordingly, Delmont *et al.* showed that the impact of sampling of a grassland soil in different seasons led to metagenomic-functional dissimilarity comparable to or only slightly higher than that of a replicated metagenome sequencing run (160), indicating that season does not strongly influence microbial community composition. Similar results were obtained by Orellana *et al.*, who showed that metagenomes showed greater differences between depths or locations than between sampling-seasons (305).

Explanation of differences in tillage effects

While many similarities exist between the two agricultural experiments, the effects of tillage on soil carbon levels and relative abundances of microbial groups or genes in the metagenomes differ between the experiments. These differences in tillage effects are mainly caused by differences in soil properties, experiment duration and time between tillage and sampling between the two sites. As was expected from previous observations (24) and from other tillage experiments (217–219), the surface soil horizon of the conventional farming experiment showed a higher organic carbon content under reduced tillage than under conventional tillage. However, tillage intensity had not affected the total soil organic carbon content of the organic farming experiment. As discussed in the previous

chapter, this is presumably due to the short time of application of the tillage treatments and the organic management of the soil under both tillage treatments.

Whereas in the conventional farming experiment the long-term effect of tillage on the soil carbon content was visible, few effects of tillage were visible in the soil microbial community composition. Conversely, many subtle tillage effects on relative abundances of soil microorganisms were detected in the metagenome of the organic farming experiment. Although the higher number of replicates and greater sequencing depth may also have facilitated the detection of subtle differences in the metagenome of the organic farming experiment, the apparent paradox in tillage-responses of soil organic carbon content and microbial community composition in both experiments can probably be related to different indirect factors influencing both measurements. While experiment duration is a likely factor influencing tillage effects on soil carbon content, the tillage effect on the soil microbial community is most likely influenced by different duration of time which had past in each experiment between tillage and sampling and to whether or not crop residues had been incorporated with tillage; whereas the soil of the conventional farming experiment was sampled one month after tillage and incorporation of crop residues, the soil of the organic farming experiment was sampled one week after harvest without crop residue-incorporation and five months after tillage. As discussed above (see section discussing responses of the soil microbiome to tillage in the conventional farming experiment), the absence of the expected tillage effects on the soil microbial community of the conventional tillage experiment are probably related to the overall activating effect of soil tillage and plant residue incorporation on the soil microbial community. This has presumably introduced an overall *r*-strategist-dominated microbial population under both tillage treatments, leading to few differences in relative abundances of microbial groups and no significant differences in alpha-diversity between both tillage treatments. The higher relative abundance of Bacteroidetes in the metagenome of the conventional farming experiment (3.4%) compared to the organic farming experiment (1.5%) further supports this. Bacteroidetes are viewed as clear r-strategists (306) and are frequently found in organic soil horizons with high content of easily accessible carbon substrates (102, 103). Similar observations have been made in a sandy soil, where the abundance of *r*-strategist-microannelids was much higher a few weeks after tillage than five months after tillage (307). The short period of time which had passed since tillage application in the conventional farming experiment may also explain the visible effects of tillage on the absolute abundance of fungi in the surface soil horizon. The recent soil disturbance in this experiment led to a higher abundance of fungi under reduced tillage than under conventional tillage. As in the organic farming experiment several months had passed since tillage, no effect of different tillage intensity on fungal ribosomal gene abundances was detected at the time of sampling. Furthermore, the fact that a higher amount of fresh plant residues was present in the soil of the conventional farming experiment compared to the soil of the organic farming experiment may also explain the relatively higher abundance of cellulase-genes in the metagenome of the conventional (0.11% as cellulase enzymatic groups and 0.63% as cellulase domain families) versus organic farming experiment (0.08% as cellulase enzymatic groups and 0.43% as cellulase domain families). Higher abundances of carbohydrate-active genes have been observed in environments with more (fresh) organic matter e.g. in soil versus freshwater (162) and in the organic horizon versus mineral horizon of forest soil (102, 189). Alternatively, the use of cattle slurry as fertilizer in the organic farming experiment may have decreased the relative amount of cellulase-genes in the soil, as organic fertilization has been shown to decrease endoglucanase activities compared to soil without organic fertilization (308).

Likewise, the encountered tillage effects in the metagenome of the conventional farming experiment were most likely related to an earlier succession under conventional tillage of *r*-strategists by *K*-strategists than under reduced tillage. The *K*-strategists are thought to have regained a growth-advantage after earlier depletion of fresh organic matter under conventional tillage in the surface soil horizon than under reduced tillage. This was reflected in the relatively higher abundance of microbial groups with *K*-strategist-characteristics in soil under conventional versus reduced tillage, like members of the Chloroflexi and Armatimonadetes. Similarly, the higher relative abundance of genes in functional categories related to the degradation of recalcitrant organic substances in soil under conventional versus reduced tillage, also indicated a higher abundance of *K*-strategists more adapted to degrade compounds which are rather present in the native soil organic matter pool (aromatic and phenolic compounds and organic acids) than in fresh organic plant material. Accordingly, no differences were found in cellulase gene abundances or cellulase-harbouring microbial groups between tillage intensities. Thus, whereas the short-term effects of tillage were visible in the metagenome of the conventional farming experiment, no recent soil disturbances had taken place under either tillage treatment in the organic farming experiment. Therefore, no short-term tillage-induced priming effects on the soil microbial community were expected to be seen in this field experiment. Consequently, in the organic farming experiment the effects of tillage intensity on the microbial community at a more developed stage could be analysed. This stage of development could be related to the tillage effects on plant root development, favouring especially rhizosphere-associated microorganisms in the surface soil under reduced tillage and in the deeper soil horizon under conventional tillage. This rhizosphere-associated microbial community consisted mainly of Actinobacteria and Alphaproteobacteria, but also of several families from the Beta-, Gamma- and Deltaproteobacteria. Indeed, recently it was shown that the soil and root microbiome of the wheat plants in this field experiment differed strongly and that the root microbiome tended to comprise more Actinobacteria and Alphaproteobacteria than the soil microbiome (309). Furthermore, in this study, groups of protein-coding genes related to nutrient acquisition and uptake and organic compound degradation were found to be more abundant under reduced tillage compared to conventional tillage in the surface soil horizon, whereas this effect was reversed in the deeper soil horizon. These include β-glucosidase genes and CBM-genes putatively involved in sensing the carbohydrate-content of the environment. These results indicate a higher competition among microorganisms, especially for the acquisition of various organic compounds which presumably are enriched in the rhizosphere environment. However, no differential relative abundances of endo- or exoglucanase genes between soils under different tillage treatments were found, suggesting that decomposition of root tissue had not yet led to enrichment of microbial degraders at the time of sampling. In addition, a higher relative abundance of Ascomycota was found under reduced tillage compared to conventional tillage in the surface soil horizon of the organic farming experiment. As Ascomycota are known to be rhizosphere-dwelling organisms (284) and several members of this phylum are mycorrhizal fungi (310), this observation further supports the theory that, in the surface soil under reduced tillage, a rhizosphere-associated community is predominantly selected. While the same abundance pattern for Ascomycota could be observed in the metagenome of the conventional farming experiment, the tillage effect was not significant at this site (P=0.13). Besides the proposed rhizosphere-effect, however, the low absolute abundance of fungi in the soil of the conventional farming experiment possibly reduced the ability to detect tillage effects on their relative abundances.

7 - Conclusions and final remarks

The role of soil microorganisms in the degradation of cellulose and the soil carbon cycle can be investigated by analysing the genetic potential of microbial communities in the soil. As more intense (conventional) tillage practices have been shown to cause a decrease in soil carbon content in the surface soil horizon compared to less intense (reduced) tillage (18, 22, 24, 25), it was expected that cellulose degrading microorganisms would be enriched in the surface horizon of soil under reduced tillage as compared to soil under conventional tillage. Shotgun metagenomes were generated from soil of two agricultural experiments, subject to similar climatic conditions, which applied both tillage treatments. The results demonstrated that tillage effects on microbial taxa and functions, including cellulases, are small and that the overall microbial community composition of both metagenomes strongly resembled each other. Thus, despite differences in soil characteristics and farming practices, the investigated agricultural soils show a stable microbial community whose composition is not strongly influenced by differences in tillage intensity.

Specifically, metagenome analysis results showed that tillage intensity did not affect the genetic potential for cellulases or the proportion of cellulolytic microorganisms in the soil of the conventional farming experiment. In the soil of the organic farming experiment, however, tillage intensity affected the genetic potential for several groups of cellulase-related enzymes. In particular, β-glucosidase and CBM-genes which were harboured by rhizosphere-associated microorganisms were enriched in the surface soil horizon under reduced tillage compared to soil under conventional tillage. However, tillage intensity did not show pronounced effect on the genetic potential for endo- or exoglucanases or on the proportion of the most important microorganisms harbouring genes coding for these enzymes. Thus, considering the results of both metagenomes, it can be concluded that cellulose degradation potential in the microbial community is not greatly influenced by tillage treatment. The metagenome results therefore do not clearly elucidate the role of cellulose degraders in the observed (24, 217–219) organic matter sequestration-capacity of reduced tillage compared to conventional tillage in the surface soil horizon. Nevertheless, enzyme activity measurements in the soil of the conventional farming experiment showed a higher potential endo- or exoglucanase enzymatic activity in soil under reduced tillage compared to conventional tillage, implying that tillage intensity might have a greater effect on cellulose degradation on a transcriptional or translation level. Moreover, the observed higher microbial biomass in soil under reduced tillage points out that a higher absolute abundance of cellulose degrading microorganisms is present in soil under reduced tillage treatment. Finally, the metagenome analysis data presented here propose some putative mechanisms for carbon sequestration through the reduction of tillage intensity; in the conventional farming experiment, the assumed lower concentration of incorporated plant residues in the surface soil horizon under conventional tillage presumably led to an earlier and possibly increased degradation of native soil organic matter (priming effect (311)) compared to soil under reduced tillage. Besides, deeper incorporation of fresh plant residues under conventional tillage may lead to a priming effect in more soil horizons than the shallow incorporation under reduced tillage, possibly leading to a higher total depletion of native soil organic matter in the complete soil profile. In the organic farming experiment, the observed higher genetic potential for β-glucosidases in surface soil horizon under reduced tillage indicates a higher carbon cycling activity and perhaps sequestration under reduced tillage compared to conventional tillage. Higher potential activity of β-glucosidases in soil has been identified as an indicator of carbon sequestration potential in soil (312, 313), although the exact mechanisms remain to be identified.

Both the analysis of soil shotgun metagenomes and the amplified GH5-subfamily 2 genes from agricultural soil showed that the genetic potential for cellulose degradation and utilization was associated with a broad array of different microorganisms, illustrating the great diversity of potentially cellulolytic microorganisms. In addition, the more in-depth analysis of amplified endoglucanase genes from the GH5-subfamily 2 demonstrated an extraordinary phylogenetic diversity of these genes in agricultural soil, exceeding the diversity of functionally characterized sequences in the public databases. Furthermore, the observed ambiguity in taxonomic assignments of the amplified GH5-subfamily 2 cellulase genes pointed out that many novel cellulolytic microorganisms remain uncovered, which represent suitable candidates for investigation of the redundancy of GH5 genes and the evolutionary principles that govern the cellulose-degradation trait.

When considering the genetic potential of the soil microbial community as a whole, it was observed that effects of tillage intensity were more distinct in the metagenome of the organic farming experiment than in that of the conventional farming experiment. In the conventional farming experiment, few effects of different tillage intensities were found on microbial taxa or general protein-coding genes. Instead, the microbial community of the conventional farming experiment rather reflects the recent crop residue incorporation which occurred a few weeks before sampling. The results suggest that, in this experimental field, an earlier depletion of readily available organic matter in the soil under conventional tillage led to an increase in microorganisms exhibiting a *K*-strategic lifestyle than in the soil under reduced tillage. This earlier depletion of organic matter was probably caused by a lower concentration of crop residues incorporated into the surface horizon of soil under conventional tillage than in that under reduced tillage. In the organic farming experiment, one week after wheat harvest and five months after the last tillage-event, the soil microbial community showed a multitude of differences in genetic potential between tillage treatments. Also for this experiment, the metagenome analysis results suggest that tillage intensity plays a secondary role in shaping the microbial community composition. In other words, rather than being direct results of the action of tillage, the differences in microbial community composition observed are presumably caused by other factors which are influenced by tillage treatment and are dependent on the time past since the last tillage activity. While for the conventional farming experiment the primary factor appeared to be the concentration of fresh organic matter, for the organic farming experiment a high influence was attributed to the wheat plant rhizosphere conditions. However, tillage may have induced changes in characteristics of these factors on the long or short term, thereby indirectly influencing the microbial community composition. Total soil organic matter content was expected to influence the soil genetic potential for cellulose degradation, but the results presented here suggest that availability of fresh organic matter is a more important driver of soil microbes. In the conventional farming experiment this organic matter availability was likely altered by the addition of fresh plant residues, while in the organic farming experiment fresh organic matter was probably available through rhizodeposition.

8 - Outlook

In this study, the genetic potential in agricultural soil and the influence of tillage intensity on the relative abundance of cellulolytic microorganisms was explored. However, the role of the impressive diversity of cellulase genes in soil in organic matter degradation and the carbon cycle remains to be defined. Agricultural management practices shaping the cellulolytic microbial community in soil and thereby improving soil fertility and agricultural sustainability are therefore key identification targets for future research. Experimental field setups testing the effects of type of fertilization (e.g. organic versus chemical), weed management (e.g. herbicide use versus mechanical techniques), cropping systems (e.g. using crops with different rooting strategies), crop residue management (e.g. incorporation, surface-application or removal) and their interactions will be instrumental for investigating the most effective management practices in field situations. However, the influence of climate conditions and soil structure should be taken into account additionally to achieve realistic estimations of global effects. Of particular research interest are factors improving soil carbon sequestration capacity while maintaining crop productivity. In this context, a balance needs to be found between stimulation of fresh organic matter degradation for nutrient cycling and minimization of native organic matter degradation by priming. Different crop residue management techniques may be applicable to different types of residues; for example, residues of different crops were demonstrated to lead to different carbon dioxide emissions from soil (314). Different types of organic matter will likely induce particular expression patterns at different degradation stages, which can elucidate the function of different sets of cellulases. For instance, Güllert *et al.* found that higher expression of certain types of endoglucanase genes in elephant faeces- than in biogas fermentor-community was related to a higher carbohydrate degradation efficiency (188). Measuring treatment effects on the cellulolytic microorganisms in environmental samples has proven to be challenging on the DNA-level. Therefore, more focus should be set on transcriptional or enzymatic responses by the soil microbial community. As these responses can easily change over time in a non-controlled natural environment, measurements over different time sequences are essential to understand their dynamics. In parallel to the microbiological assessment, relevant environmental metadata should be collected on the same spatial and temporal scale, as this information is pivotal in understanding the microbial ecology. Finally, carbon fluxes between different pools (e.g. atmosphere, native soil organic matter and added organic matter) should be meticulously recorded and quantified. Of particular interest will be the contribution of microbial-derived organic matter to the total carbon pool over time. Stable-isotope probing can greatly facilitate this quantification process and additionally reliably identify the microorganisms involved in the organic matter degradation process (see for example (103)). The thus obtained quantitative carbon fluxes will aid in development of carbon cycle-models which include responses the soil microbial community. Combined, results of cellulolytic microbial community responses and carbon fluxes to agricultural management practices and environmental changes can provide a conceptual framework for prediction of soil ecosystem functioning in response to designed agricultural management strategies.

Appendix

Table A1

Table A1: Benchmarking results of HMMs obtained either from the Pfam-A database (Pfam), the DataBase for automated Carbohydrate-active enzyme Annotation (dbCAN (110)) or self-built HMMs based on cellulase-database sequences. Benchmarking was done using "positive" (pos)- and "negative" (neg)-sequence databases, which consisted of functionally characterized sequences of cellulases respectively non-cellulases extracted from the CAZy-database (http://www.cazy.org/, (60)). Given here are the percentages of neg- or pos-database sequences annotated by the HMMs. The HMMs chosen for analysis of the metagenomes are shaded.

Source of HMM	Pfam		dbCAN		Personally built	
CAZy-family	neg	pos	neg	pos	neg	pos
CBM1	2.37	0.08	2.09	0.00	**4.60**	**1.56**
CBM2	1.87	0.08	4.32	0.08	**3.02**	**4.75**
CBM3	0.14	0.00	0.07	0.08	**0.72**	**11.38**
CBM4	3.96	0.00	3.09	0.08	**2.81**	**1.17**
CBM5	1.08	0.00	6.62	0.00	**1.22**	**3.74**
CBM6	3.45	0.31	3.31	0.31	**2.01**	**0.70**
CBM8			0.00	0.00	**0.94**	**0.47**
CBM10	0.22	0.00	0.14	0.00	**0.50**	**0.39**
CBM11	0.07	0.31	0.00	0.31	**0.00**	**0.39**
CBM17	0.00	0.86	0.00	0.55	**0.00**	**0.70**
CBM28			**0.00**	**0.39**	0.14	0.31
CBM30					**0.14**	**0.16**
CBM32	4.39	0.00	3.96	0.00	**0.29**	**0.62**
CBM44			0.07	0.00	**0.14**	**0.08**
CBM46			0.00	0.16	**0.50**	**0.55**
CBM49					**0.00**	**0.31**
CBM63					**0.07**	**0.23**
CBM65			0.00	0.08	**0.79**	**3.27**
AA3	**0.00**	**1.09**	0.00	0.00	0.00	0.94
AA8			**0.07**	**1.09**	0.07	0.16
AA9	0.43	0.94	**0.43**	**0.94**	0.43	0.94
AA10	0.58	0.47	**0.58**	**0.47**	0.58	0.47
GH1	11.51	17.85	**11.51**	**17.85**	12.01	17.85
GH3	6.62	13.48	**6.83**	**13.72**	7.12	13.80
GH5	**13.81**	**24.47**	13.96	23.85	9.42	18.78
GH6	0.07	5.46	**0.07**	**5.46**	0.29	5.38
GH7	0.00	6.39	**0.00**	**6.39**	0.00	6.39
GH8	2.01	2.34	**2.01**	**2.34**	2.01	2.34
GH9	1.08	12.94	**1.08**	**12.94**	1.08	12.24
GH10	23.09	0.31	23.17	0.31	**23.31**	**0.70**
GH12	0.79	3.90	**0.79**	**3.90**	1.15	3.51

GH26	3.53	0.23	3.88	0.23	**3.67**	**0.39**
GH30	**2.59**	**0.08**	3.24	0.08	2.81	0.08
GH44	0.22	0.86	**0.29**	**1.01**	0.22	0.78
GH45	0.00	3.66	**0.00**	**3.74**	0.00	3.90
GH48	0.07	1.09	**0.07**	**1.09**	1.01	0.62
GH51			**5.11**	**0.31**	3.24	0.23
GH74			**1.01**	**0.16**	1.01	0.16
GH94			**0.65**	**1.40**	0.65	1.40
GH124			**0.00**	**0.08**	0.29	0.08

Table A2

Table A2: Protein IDs of sequences used as reference sequences during pair-wise alignments and Markov Clustering with GH5-amplicon sequences. These protein sequences have enzymatic functions which have been functionally characterized (CAZy database, http://www.cazy.org/ (60)) and include cellulolytic enzymes (with EC: 3.2.1.4, EC: 3.2.1.91, EC: 3.2.1.176, EC: 3.2.1.74, EC: 3.2.1.21) and non-cellulolytic enzymes. GH5-protein sequences which were used as input for multiple sequence alignment and phylogenetic tree-calculation are shaded.

GH-family	Protein ID	EC-number	Protein ID	EC-number	Protein ID	EC-number	Protein ID	EC-number
48	AAA73866	3.2.1.175	BAC22065	3.2.1.176	ABD64772	3.2.1.176	ACM60955	3.2.1.176
	AAA23226	3.2.1.176	CAB06786	3.2.1.176	ABN53296	3.2.1.176	AEE47513	3.2.1.176
	AAA50257	3.2.1.176	AAD39947	3.2.1.176	ABX43721	3.2.1.176	AAA72860	3.2.1.4
	AAB00822	3.2.1.176	AAC38571	3.2.1.176	ACH05303	3.2.1.176	AAA91086	3.2.1.4
	AAB41452	3.2.1.176	CAD32945	3.2.1.176	ACH05304	3.2.1.176	ABN51312	3.2.1.4
	CAA93280	3.2.1.176	AAZ55992	3.2.1.176	ACL75108	3.2.1.176	BAE94320	3.2.1.14
74	CAF02212	3.2.1.150	CAD58415	3.2.1.151	BAE44527	3.2.1.151	ACE14921	3.2.1.4
	CAF02249	3.2.1.150	BAC69567	3.2.1.151	ABH71452	3.2.1.151	ACS09151	3.2.1.4
	EAA64249	3.2.1.150	BAC70285	3.2.1.151	ABJ18610	3.2.1.151	AGL49229	3.2.1.4
	CAA20642	3.2.1.150	AAP57752	3.2.1.151	BAF95189	3.2.1.151	AHD17931	3.2.1.4
	AAK77227	3.2.1.151	CAE51306	3.2.1.151	BAA29031	3.2.1.4		
	NP_630626	3.2.1.151	BAD11543	3.2.1.151	AAD35393	3.2.1.4		
	CAA35159	3.2.1.151	AAZ55647	3.2.1.151	NP_228117	3.2.1.4		
7	CAA38274	3.2.1.176	CAB06786	3.2.1.176	EAW16381	3.2.1.176	AAG09047	3.2.1.4
	CAA41780	3.2.1.176	CAC85737	3.2.1.176	XP_001258278	3.2.1.176	BAB64553	3.2.1.4
	AAB46373	3.2.1.176	AAL83303	3.2.1.176	ABN13116	3.2.1.176	BAB64554	3.2.1.4
	CAA37878	3.2.1.176	AAL89553	3.2.1.176	CAK44068	3.2.1.176	BAB64555	3.2.1.4
	CAA80253	3.2.1.176	AAL33603	3.2.1.176	CAK39699	3.2.1.176	BAB64556	3.2.1.4
	CAA77789	3.2.1.176	AAM54070	3.2.1.176	ABS82449	3.2.1.176	BAB64557	3.2.1.4
	CAA77795	3.2.1.176	AAN19007	3.2.1.176	CAM98445	3.2.1.176	BAB64558	3.2.1.4
	CAA82761	3.2.1.176	CAD56667	3.2.1.176	CAM98446	3.2.1.176	BAB64559	3.2.1.4
	CAA82762	3.2.1.176	EAA33262	3.2.1.176	CAM98448	3.2.1.176	BAB64560	3.2.1.4
	CAA49596	3.2.1.176	CAD79780	3.2.1.176	CAM98447	3.2.1.176	BAB64561	3.2.1.4
	AAA19802	3.2.1.176	CAD79781	3.2.1.176	ACF93800	3.2.1.176	BAB64562	3.2.1.4

	AAC49089	3.2.1.176	CAD79782	3.2.1.176	ACH15004	3.2.1.176	AAM54071	3.2.1.4
	BAA09785	3.2.1.176	CAD79785	3.2.1.176	ACH15013	3.2.1.176	BAC07551	3.2.1.4
	CAA68840	3.2.1.176	AAP60302	3.2.1.176	CAR96030	3.2.1.176	BAC07552	3.2.1.4
	BAA25183	3.2.1.176	AAP66263	3.2.1.176	ACV95805	3.2.1.176	AAQ21382	3.2.1.4
	CAA80252	3.2.1.176	AAP66264	3.2.1.176	ADB85438	3.2.1.176	AAQ24882	3.2.1.4
	BAA36215	3.2.1.176	AAQ38146	3.2.1.176	ADX60067	3.2.1.176	AAR60514	3.2.1.4
	AAD11942	3.2.1.176	AAQ76092	3.2.1.176	AEO67172	3.2.1.176	EAA63386	3.2.1.4
	BAA74517	3.2.1.176	AAR60512	3.2.1.176	CDF76454	3.2.1.176	AAX28897	3.2.1.4
	BAA76363	3.2.1.176	AAR79028	3.2.1.176	AGY80096	3.2.1.176	BAE66197	3.2.1.4
	BAA76364	3.2.1.176	EAA66593	3.2.1.176	CAA43059	3.2.1.4	ABE22093	3.2.1.4
	AAD31545	3.2.1.176	CAH10320	3.2.1.176	AAA34212	3.2.1.4	ABM90986	3.2.1.4
	AAD41096	3.2.1.176	AAT99321	3.2.1.176	AAA65586	3.2.1.4	BAF57296	3.2.1.4
	AAF04491	3.2.1.176	AAU96164	3.2.1.176	BAA09786	3.2.1.4	ABY56790	3.2.1.4
	AAF04492	3.2.1.176	AAV65115	3.2.1.176	BAA22589	3.2.1.4	CAR96034	3.2.1.4
	AAF36391	3.2.1.176	AAW64926	3.2.1.176	AAE25068	3.2.1.4	CAR96035	3.2.1.4
	AAE25072	3.2.1.176	EAL73676	3.2.1.176	AAE25069	3.2.1.4	ACT53749	3.2.1.4
	CAC07539	3.2.1.176	AAX84833	3.2.1.176	AAE25070	3.2.1.4	AEB00821	3.2.1.4
	AAL16941	3.2.1.176	AAY89412	3.2.1.176	AAE25071	3.2.1.4	AEO58196	3.2.1.4
9	ACM60955	3.2.1.176_4	AAF06107	3.2.1.4	CAI94607	3.2.1.4	ACL75113	3.2.1.4
	AAF93781	3.2.1.176_4	AAF06109	3.2.1.4	AAY48792	3.2.1.4	ACL75114	3.2.1.4
	NP_230264	3.2.1.21	AAF15367	3.2.1.4	AAN04496	3.2.1.4	ACL75116	3.2.1.4
	CAA28255	3.2.1.21	CAB63115	3.2.1.4	AAU20853	3.2.1.4	ACL75131	3.2.1.4
	CAA31082	3.2.1.4	AAF19168	3.2.1.4	AAZ55662	3.2.1.4	ACL76568	3.2.1.4
	AAA23086	3.2.1.4	CAB76932	3.2.1.4	AAZ56209	3.2.1.4	BAH57006	3.2.1.4
	AAA23088	3.2.1.4	CAB76935	3.2.1.4	BAE20171	3.2.1.4	ACR23658	3.2.1.4
	AAA20892	3.2.1.4	AAF06111	3.2.1.4	AAZ93631	3.2.1.4	ACS45173	3.2.1.4
	AAA24894	3.2.1.4	AAF80584	3.2.1.4	ABA07706	3.2.1.4	ACV59481	3.2.1.4
	AAA24895	3.2.1.4	AAF80585	3.2.1.4	ABA07707	3.2.1.4	CBC93706	3.2.1.4
	AAA52077	3.2.1.4	BAA98160	3.2.1.4	ABA07708	3.2.1.4	CBC93714	3.2.1.4
	CAA39010	3.2.1.4	AAG45157	3.2.1.4	ABB51609	3.2.1.4	CBC93722	3.2.1.4
	AAA68129	3.2.1.4	AAG45158	3.2.1.4	ABB51610	3.2.1.4	CBC93725	3.2.1.4
	AAC06387	3.2.1.4	AAG45160	3.2.1.4	ABB51611	3.2.1.4	ACX73672	3.2.1.4
	AAA02563	3.2.1.4	AAG49558	3.2.1.4	ABD24274	3.2.1.4	ACX75451	3.2.1.4
	AAA69908	3.2.1.4	AAG52329	3.2.1.4	ABD24275	3.2.1.4	ACX75452	3.2.1.4
	AAA69909	3.2.1.4	AAG59608	3.2.1.4	ABD24278	3.2.1.4	ACY24809	3.2.1.4
	AAA91086	3.2.1.4	AAK12339	3.2.1.4	CAK32152	3.2.1.4	ACY24880	3.2.1.4
	AAA73868	3.2.1.4	CAC34051	3.2.1.4	ABG76967	3.2.1.4	EFA05721	3.2.1.4
	CAA39264	3.2.1.4	BAB33148	3.2.1.4	ABG76972	3.2.1.4	ADB12483	3.2.1.4
	CAA40993	3.2.1.4	BAB39482	3.2.1.4	BAF00867	3.2.1.4	ADB82903	3.2.1.4
	CAA43035	3.2.1.4	BAB39483	3.2.1.4	ABH93356	3.2.1.4	ADD61951	3.2.1.4
	CAA46570	3.2.1.4	BAB40693	3.2.1.4	BAF38757	3.2.1.4	ADL24960	3.2.1.4
	AAC44386	3.2.1.4	BAB40694	3.2.1.4	ABM68635	3.2.1.4	ADL25229	3.2.1.4
	AAA96135	3.2.1.4	BAB40695	3.2.1.4	ABN51281	3.2.1.4	ADL26362	3.2.1.4
	AAA80495	3.2.1.4	BAB40696	3.2.1.4	ABN51650	3.2.1.4	ADL52438	3.2.1.4
	AAB60304	3.2.1.4	BAB40697	3.2.1.4	ABN51779	3.2.1.4	ADL52515	3.2.1.4
	BAA06877	3.2.1.4	AAK59818	3.2.1.4	ABN51814	3.2.1.4	ADV16268	3.2.1.4

AAC41523	3.2.1.4	AAK78892	3.2.1.4	ABN51860	3.2.1.4	ADV16269	3.2.1.4
AAC44385	3.2.1.4	AAK82545	3.2.1.4	ABN52060	3.2.1.4	ADV16270	3.2.1.4
CAA65597	3.2.1.4	NP_171779	3.2.1.4	ABN54011	3.2.1.4	ADX05755	3.2.1.4
CAA65600	3.2.1.4	NP_177228	3.2.1.4	BAF62178	3.2.1.4	ADY68794	3.2.1.4
BAA12070	3.2.1.4	NP_199783	3.2.1.4	AAW62376	3.2.1.4	ADY68795	3.2.1.4
AAB42155	3.2.1.4	NP_347552	3.2.1.4	ABU45498	3.2.1.4	AED95850	3.2.1.4
AAB46824	3.2.1.4	AAL30452	3.2.1.4	ABV32557	3.2.1.4	AEE27471	3.2.1.4
AAB46826	3.2.1.4	AAL30453	3.2.1.4	ABX43720	3.2.1.4	AEE35103	3.2.1.4
AAB46828	3.2.1.4	AAL30454	3.2.1.4	ABX76047	3.2.1.4	AEE44171	3.2.1.4
AAC49704	3.2.1.4	BAB79196	3.2.1.4	ACA04897	3.2.1.4	AEE45671	3.2.1.4
CAA72133	3.2.1.4	AAL67092	3.2.1.4	ACE85757	3.2.1.4	AEH04391	3.2.1.4
CAA67156	3.2.1.4	BAB86305	3.2.1.4	ACI45756	3.2.1.4	AEL88496	3.2.1.4
CAA67157	3.2.1.4	AAM41665	3.2.1.4	ACJ68032	3.2.1.4	AEW10553	3.2.1.4
BAA24918	3.2.1.4	NP_637741	3.2.1.4	CAW94451	3.2.1.4	BAM14716	3.2.1.4
AAC16418	3.2.1.4	AAM63370	3.2.1.4	CAW94460	3.2.1.4	AGI61069	3.2.1.4
BAA28815	3.2.1.4	CAD54726	3.2.1.4	CAW94470	3.2.1.4	AGS32241	3.2.1.4
BAA31326	3.2.1.4	CAD54727	3.2.1.4	CAW94474	3.2.1.4	AGT17861	3.2.1.4
AAC33467	3.2.1.4	CAD54728	3.2.1.4	CAX07297	3.2.1.4	AHL27899	3.2.1.4
AAC35344	3.2.1.4	CAD54729	3.2.1.4	CAX07301	3.2.1.4	AHL27900	3.2.1.4
AAC62241	3.2.1.4	CAD54730	3.2.1.4	CAX07305	3.2.1.4	AID55374	3.2.1.4
BAA33708	3.2.1.4	AAN72232	3.2.1.4	CAX07307	3.2.1.4	ABY60376	3.2.1.74
BAA33709	3.2.1.4	AAO30718	3.2.1.4	CAW91525	3.2.1.4	ACX75620	3.2.1.74
AAC64045	3.2.1.4	BAC67186	3.2.1.4	CAW91534	3.2.1.4	ADL26040	3.2.1.74
BAA34050	3.2.1.4	BAC67187	3.2.1.4	CAW91543	3.2.1.4	CAA56918	3.2.1.91
CAA11301	3.2.1.4	AAO61672	3.2.1.4	CAW91547	3.2.1.4	CAA06693	3.2.1.91
BAA34120	3.2.1.4	AAP83128	3.2.1.4	CAX00889	3.2.1.4	AAR87745	3.2.1.91
AAC78293	3.2.1.4	AAQ08018	3.2.1.4	CAX00898	3.2.1.4	ABN51651	3.2.1.91
AAC83240	3.2.1.4	AAQ68347	3.2.1.4	CAX00907	3.2.1.4	ABN51859	3.2.1.151
AAD01959	3.2.1.4	CAD44274	3.2.1.4	CAX00911	3.2.1.4	ACL76949	3.2.1.151
AAD08699	3.2.1.4	AAQ91573	3.2.1.4	CAX06019	3.2.1.4	CAG18943	3.2.1.165
BAA74961	3.2.1.4	CAE51308	3.2.1.4	CAX06023	3.2.1.4	ADH59533	3.2.1.6
BAA74962	3.2.1.4	BAD01504	3.2.1.4	CAX06027	3.2.1.4	BAA10447	3.2.1.73
AAC38572	3.2.1.4	AAR29083	3.2.1.4	CAX06029	3.2.1.4	BAF51695	3.2.1.73
CAB38941	3.2.1.4	AAS87601	3.2.1.4	CAW92415	3.2.1.4	ADQ41951	3.2.1.73
BAA76619	3.2.1.4	AAT66046	3.2.1.4	CAW92424	3.2.1.4	BAK51232	3.2.1.73
BAA77239	3.2.1.4	CAD61242	3.2.1.4	CAW92432	3.2.1.4	AGF52739	3.2.1.73
AAD38027	3.2.1.4	BAD66681	3.2.1.4	CAW92437	3.2.1.4		
AAF02887	3.2.1.4	CAF22221	3.2.1.4	ACL74618	3.2.1.4		
CAA86077	3.2.1.4	BAD95336	3.2.1.4	ACL75110	3.2.1.4		

5

CAA86209	3.2.1.21	NP_638867	3.2.1.4	ACU30064	3.2.1.4	ACP74152	3.2.1.58
CAA47429	3.2.1.21	AAL33639	3.2.1.4	ACW22975	3.2.1.4	AHH92832	3.2.1.58
CAA31936	3.2.1.4	CAD42489	3.2.1.4	ACW22976	3.2.1.4	EGC02962	3.2.1.73
CAA27266	3.2.1.4	AAN03645	3.2.1.4	ACW33392	3.2.1.4	EGC04285	3.2.1.73
CAA49187	3.2.1.4	AAN03646	3.2.1.4	ACX74504	3.2.1.4	CAA55788	3.2.1.75
CAA68604	3.2.1.4	AAN03647	3.2.1.4	ACX75120	3.2.1.4	CAA55789	3.2.1.75
CAA38692	3.2.1.4	AAN03648	3.2.1.4	ACX75816	3.2.1.4	CAA21163	3.2.1.75

CAA38693	3.2.1.4	AAL88714	3.2.1.4	ACX75950	3.2.1.4	CAB50968	3.2.1.75
P25472	3.2.1.4	CAD82873	3.2.1.4	ACY24829	3.2.1.4	AAL84696	3.2.1.75
AAA22299	3.2.1.4	AAP04424	3.2.1.4	ACY24859	3.2.1.4	NP_596224	3.2.1.75
AAA22301	3.2.1.4	AAP51020	3.2.1.4	BAI66446	3.2.1.4	NP_596461	3.2.1.75
AAA22304	3.2.1.4	AAP56348	3.2.1.4	ADB80108	3.2.1.4	AAN04103	3.2.1.75
AAA22305	3.2.1.4	CAB13696	3.2.1.4	ADC54852	3.2.1.4	EAA59985	3.2.1.75
AAA22306	3.2.1.4	AAP88024	3.2.1.4	ADD61853	3.2.1.4	AAT97707	3.2.1.75
AAA22307	3.2.1.4	AAQ21383	3.2.1.4	ADD61911	3.2.1.4	ABF50867	3.2.1.75
AAA22408	3.2.1.4	AAQ24883	3.2.1.4	ADD71777	3.2.1.4	ABK27195	3.2.1.75
AAA22496	3.2.1.4	AAQ31832	3.2.1.4	ADD73709	3.2.1.4	ABK27199	3.2.1.75
AAA22631	3.2.1.4	AAQ31833	3.2.1.4	ADE83057	3.2.1.4	ABV71387	3.2.1.75
AAA22909	3.2.1.4	AAO63626	3.2.1.4	ADH51728	3.2.1.4	ACM42428	3.2.1.75
AAA20893	3.2.1.4	BAD01163	3.2.1.4	EFI96731	3.2.1.4	AAA34208	3.2.1.78
AAA23089	3.2.1.4	BAD01164	3.2.1.4	ADJ93836	3.2.1.4	AAA67426	3.2.1.78
AAC37035	3.2.1.4	CAE81955	3.2.1.4	ADK66823	3.2.1.4	CAA90423	3.2.1.78
AAA23220	3.2.1.4	AAR29981	3.2.1.4	ADL25000	3.2.1.4	BAA25188	3.2.1.78
AAA23221	3.2.1.4	AAR60515	3.2.1.4	ADL25356	3.2.1.4	BAA25878	3.2.1.78
AAA51444	3.2.1.4	CAF02232	3.2.1.4	ADL26743	3.2.1.4	AAC71692	3.2.1.78
AAA23230	3.2.1.4	EAA62395	3.2.1.4	ADL27061	3.2.1.4	CAA06924	3.2.1.78
AAA24893	3.2.1.4	EAA65878	3.2.1.4	ADM12805	3.2.1.4	AAD09354	3.2.1.78
AAA61980	3.2.1.4	CAF05574	3.2.1.4	ADM89627	3.2.1.4	AAA26710	3.2.1.78
AAA26467	3.2.1.4	AAS58467	3.2.1.4	ADM99099	3.2.1.4	AAD36302	3.2.1.78
AAA26469	3.2.1.4	CAD61244	3.2.1.4	ADN02392	3.2.1.4	CAB56854	3.2.1.78
AAA27612	3.2.1.4	AAU40977	3.2.1.4	ADP05286	3.2.1.4	CAB56856	3.2.1.78
AAA34213	3.2.1.4	AAU27988	3.2.1.4	ADR64663	3.2.1.4	AAF22274	3.2.1.78
BAA00045	3.2.1.4	AAC02964	3.2.1.4	ADU21608	3.2.1.4	CAB76904	3.2.1.78
BAA00859	3.2.1.4	AAV25061	3.2.1.4	ADU28719	3.2.1.4	AAF06110	3.2.1.78
BAA00793	3.2.1.4	BAD67544	3.2.1.4	ADU28720	3.2.1.4	AAG00883	3.2.1.78
BAA14354	3.2.1.4	BAD72778	3.2.1.4	ADU31612	3.2.1.4	CAC08208	3.2.1.78
BAA03070	3.2.1.4	CAE82178	3.2.1.4	ADU86901	3.2.1.4	CAC08442	3.2.1.78
AAB19708	3.2.1.4	CAH69214	3.2.1.4	ADU86902	3.2.1.4	AAG45159	3.2.1.78
AAC60541	3.2.1.4	AAX18655	3.2.1.4	ADX05684	3.2.1.4	AAL01213	3.2.1.78
AAA73189	3.2.1.4	BAD90558	3.2.1.4	ADX05688	3.2.1.4	NP_229032	3.2.1.78
BAA32286	3.2.1.4	AAY03292	3.2.1.4	ADX05696	3.2.1.4	CAC81056	3.2.1.78
CAA53592	3.2.1.4	CAJ00038	3.2.1.4	ADX05697	3.2.1.4	AAK53459	3.2.1.78
CAA83238	3.2.1.4	CAJ00039	3.2.1.4	ADX05703	3.2.1.4	AAL91241	3.2.1.78
CAA83942	3.2.1.4	AAZ22322	3.2.1.4	ADX05705	3.2.1.4	AAM41068	3.2.1.78
CAA55823	3.2.1.4	AAZ54939	3.2.1.4	ADX05717	3.2.1.4	NP_637144	3.2.1.78
CAA53631	3.2.1.4	AAZ56745	3.2.1.4	ADX05718	3.2.1.4	AAM56792	3.2.1.78
CAA82317	3.2.1.4	CAJ28076	3.2.1.4	ADX05725	3.2.1.4	AAM56796	3.2.1.78
AAA91966	3.2.1.4	CAJ28077	3.2.1.4	ADX05729	3.2.1.4	AAN27517	3.2.1.78
AAC37033	3.2.1.4	ABA42184	3.2.1.4	ADX05732	3.2.1.4	AAN27518	3.2.1.78
AAB03889	3.2.1.4	ABA42185	3.2.1.4	ADX05734	3.2.1.4	AAN34823	3.2.1.78
CAA60493	3.2.1.4	ABA64553	3.2.1.4	ADX05739	3.2.1.4	AAN72165	3.2.1.78
AAC43478	3.2.1.4	BAE46390	3.2.1.4	ADX05742	3.2.1.4	AAO31759	3.2.1.78
CAA01934	3.2.1.4	ABB51612	3.2.1.4	ADX05751	3.2.1.4	AAO31760	3.2.1.78

CAA01935	3.2.1.4	CAJ19151	3.2.1.4	ADX05760	3.2.1.4	AAO31761	3.2.1.78
CAA61740	3.2.1.4	ABB92850	3.2.1.4	ADX78144	3.2.1.4	CAC51690	3.2.1.78
AAA75477	3.2.1.4	ABC30636	3.2.1.4	ADX78145	3.2.1.4	NP_171733	3.2.1.78
AAB38548	3.2.1.4	ABD80834	3.2.1.4	ADZ44606	3.2.1.4	AAQ31837	3.2.1.78
BAA12676	3.2.1.4	ABD81750	3.2.1.4	AEB00655	3.2.1.4	AAQ79152	3.2.1.78
BAA12744	3.2.1.4	ABD81754	3.2.1.4	AEB53062	3.2.1.4	AAQ79153	3.2.1.78
CAA97610	3.2.1.4	ABD81896	3.2.1.4	AEE46054	3.2.1.4	AAM26920	3.2.1.78
CAB01405	3.2.1.4	ABD82186	3.2.1.4	AEG76944	3.2.1.4	EAA58449	3.2.1.78
AAD04193	3.2.1.4	ABD82494	3.2.1.4	AEJ76923	3.2.1.4	EAA63265	3.2.1.78
AAB40891	3.2.1.4	ABD82496	3.2.1.4	AEM23896	3.2.1.4	EAA63326	3.2.1.78
CAB06784	3.2.1.4	ABE22094	3.2.1.4	AEM23898	3.2.1.4	AAT06599	3.2.1.78
CAB05881	3.2.1.4	ABE60666	3.2.1.4	AEM45646	3.2.1.4	AAT39478	3.2.1.78
AAB51451	3.2.1.4	ABE60714	3.2.1.4	AEO53769	3.2.1.4	AAB87859	3.2.1.78
CAA73113	3.2.1.4	ABF50848	3.2.1.4	AEQ58914	3.2.1.4	CAH68693	3.2.1.78
AAC97596	3.2.1.4	ABF50872	3.2.1.4	AEV59725	3.2.1.4	AAX01860	3.2.1.78
AAC49731	3.2.1.4	BAA12826	3.2.1.4	AEV59734	3.2.1.4	AAX87002	3.2.1.78
AAB61461	3.2.1.4	ABG46712	3.2.1.4	AEV59736	3.2.1.4	AAX87003	3.2.1.78
AAC48327	3.2.1.4	ABG58383	3.2.1.4	AFC68970	3.2.1.4	EAL85463	3.2.1.78
AAC48326	3.2.1.4	ABG59366	3.2.1.4	AFG25592	3.2.1.4	BAD99527	3.2.1.78
AAB69347	3.2.1.4	ABG78039	3.2.1.4	AFJ05146	3.2.1.4	AAZ54938	3.2.1.78
AAB69348	3.2.1.4	ABH71811	3.2.1.4	AFJ44728	3.2.1.4	ABB88954	3.2.1.78
AAC48325	3.2.1.4	ABH84883	3.2.1.4	BAM21527	3.2.1.4	ABC13790	3.2.1.78
AAC48341	3.2.1.4	ABI94085	3.2.1.4	AFN89566	3.2.1.4	ABC13791	3.2.1.78
AAC02536	3.2.1.4	ABI94086	3.2.1.4	AFX88666	3.2.1.4	ABC59553	3.2.1.78
AAC05164	3.2.1.4	ABK52387	3.2.1.4	AFX88668	3.2.1.4	BAE78456	3.2.1.78
AAC06196	3.2.1.4	CAL94853	3.2.1.4	AFX88671	3.2.1.4	ABC87082	3.2.1.78
AAC06197	3.2.1.4	ABN51772	3.2.1.4	AFX88673	3.2.1.4	ABF50861	3.2.1.78
AAC08587	3.2.1.4	ABN54070	3.2.1.4	AFY97404	3.2.1.4	ABF50863	3.2.1.78
AAC09379	3.2.1.4	CAK45103	3.2.1.4	AGH53362	3.2.1.4	ABF50878	3.2.1.78
AAC15707	3.2.1.4	ABP66297	3.2.1.4	AGL50932	3.2.1.4	ABG79370	3.2.1.78
AAC15708	3.2.1.4	ABP66692	3.2.1.4	AHA42547	3.2.1.4	ABG88068	3.2.1.78
AAC19169	3.2.1.4	AAU23613	3.2.1.4	AHF23845	3.2.1.4	ABJ41262	3.2.1.78
BAA29030	3.2.1.4	CAM98473	3.2.1.4	AHF24998	3.2.1.4	ABJ41263	3.2.1.78
BAA30271	3.2.1.4	ABS61403	3.2.1.4	AID57617	3.2.1.4	ABJ41266	3.2.1.78
BAA31712	3.2.1.4	ABU45500	3.2.1.4	AIY93123	3.2.1.4	ABJ41267	3.2.1.78
AAC33848	3.2.1.4	ABV08875	3.2.1.4	CAA43597	3.2.1.74	ABJ41268	3.2.1.78
AAC33860	3.2.1.4	ABV08876	3.2.1.4	CAA36207	3.2.1.74	XP_001262744	3.2.1.78
CAA76775	3.2.1.4	ABV45393	3.2.1.4	AAA50210	3.2.1.74	ABN52056	3.2.1.78
AAC63094	3.2.1.4	CAP07661	3.2.1.4	ABE60715	3.2.1.74	CAK96471	3.2.1.78
CAA03653	3.2.1.4	BAF87299	3.2.1.4	ACE82870	3.2.1.74	ABQ47550	3.2.1.78
CAA03658	3.2.1.4	ABX41541	3.2.1.4	ACX75179	3.2.1.74	AAG00315	3.2.1.78
AAC63988	3.2.1.4	ABX42426	3.2.1.4	ADL26975	3.2.1.74	ABV68808	3.2.1.78
AAC63989	3.2.1.4	ABX76045	3.2.1.4	CAB76938	3.2.1.91	CAP71606	3.2.1.78
CAA11965	3.2.1.4	ABX76046	3.2.1.4	AFR99035	3.2.1.104	ACE82655	3.2.1.78
BAA36216	3.2.1.4	ABX76048	3.2.1.4	AAB67050	3.2.1.123	ACE84673	3.2.1.78
BAA74515	3.2.1.4	ABX76050	3.2.1.4	BAB16369	3.2.1.123	ACE84941	3.2.1.78

BAA76394	3.2.1.4	ABY28340	3.2.1.4	BAB16370	3.2.1.123	ACH56965	3.2.1.78
CAB42449	3.2.1.4	ABY52965	3.2.1.4	BAB17317	3.2.1.123	ACH58410	3.2.1.78
CAB42450	3.2.1.4	ABZ29259	3.2.1.4	BAD20464	3.2.1.123	ACH58411	3.2.1.78
AAD39739	3.2.1.4	CAQ03244	3.2.1.4	BAF56440	3.2.1.123	ACJ06979	3.2.1.78
AAD43818	3.2.1.4	ACA61149	3.2.1.4	BAC65342	3.2.1.132	CAT81455	3.2.1.78
CAB49854	3.2.1.4	ACA61152	3.2.1.4	BAG70961	3.2.1.149	ACL75115	3.2.1.78
BAA82592	3.2.1.4	ACA61168	3.2.1.4	AAR65335	3.2.1.151	ACM94273	3.2.1.78
AAD45868	3.2.1.4	ACB06750	3.2.1.4	AAR65336	3.2.1.151	BAG69482	3.2.1.78
CAB58698	3.2.1.4	ACD36972	3.2.1.4	BAE44526	3.2.1.151	ACU52526	3.2.1.78
CAB59143	3.2.1.4	ACE06751	3.2.1.4	ACZ54907	3.2.1.151	ACU52527	3.2.1.78
CAB59144	3.2.1.4	ACE10214	3.2.1.4	BAF42338	3.2.1.164	BAI52931	3.2.1.78
CAB59165	3.2.1.4	ACE14923	3.2.1.4	CAK38078	3.2.1.164	ADF28533	3.2.1.78
AAF18152	3.2.1.4	ACE14924	3.2.1.4	AAS19695	3.2.1.25	ADK91085	3.2.1.78
AAF00074	3.2.1.4	ACE14925	3.2.1.4	AEH51033	3.2.1.25	ADL52514	3.2.1.78
BAA90480	3.2.1.4	ACE84076	3.2.1.4	CAJ75961	3.2.1.45	ADL52789	3.2.1.78
BAA92146	3.2.1.4	ACH63253	3.2.1.4	BAL46040	3.2.1.45	ADN93457	3.2.1.78
BAA92430	3.2.1.4	ACH67609	3.2.1.4	BAL46041	3.2.1.45	ADO14134	3.2.1.78
CAB92326	3.2.1.4	ACH69873	3.2.1.4	EIE79467	3.2.1.45	BAJ60954	3.2.1.78
CAA44467	3.2.1.4	AAD48494	3.2.1.4	AFR92751	3.2.1.45	ADW82104	3.2.1.78
AAF83628	3.2.1.4	ACI15227	3.2.1.4	AHV83755	3.2.1.45	BAK05001	3.2.1.78
AAF85505	3.2.1.4	ACI18520	3.2.1.4	AAB24895	3.2.1.58	ADZ99027	3.2.1.78
BAB04322	3.2.1.4	ACI63223	3.2.1.4	CAA41952	3.2.1.58	ADZ99301	3.2.1.78
CAC18529	3.2.1.4	CAR96693	3.2.1.4	CAA39908	3.2.1.58	AEE27414	3.2.1.78
BAB19360	3.2.1.4	ACJ12786	3.2.1.4	CAA86950	3.2.1.58	AEE43708	3.2.1.78
AAG44364	3.2.1.4	ACJ12787	3.2.1.4	CAA63536	3.2.1.58	BAK26781	3.2.1.78
AAG45162	3.2.1.4	ACJ12788	3.2.1.4	CAA92719	3.2.1.58	AEP84473	3.2.1.78
AAG50051	3.2.1.4	ACJ12789	3.2.1.4	CAA86948	3.2.1.58	AEV40667	3.2.1.78
AAG59832	3.2.1.4	ACJ12790	3.2.1.4	CAA86949	3.2.1.58	AEV41143	3.2.1.78
CAC27410	3.2.1.4	ACJ12791	3.2.1.4	CAA86951	3.2.1.58	AEY76082	3.2.1.78
AAK16222	3.2.1.4	ACJ12792	3.2.1.4	CAA86952	3.2.1.58	AFC38441	3.2.1.78
AAK21881	3.2.1.4	ACJ60856	3.2.1.4	CAA94100	3.2.1.58	AFJ59924	3.2.1.78
AAK21882	3.2.1.4	CAT02251	3.2.1.4	CAA99399	3.2.1.58	AFJ68087	3.2.1.78
AAK39540	3.2.1.4	ACJ71329	3.2.1.4	CAA21969	3.2.1.58	BAM62868	3.2.1.78
AAE59925	3.2.1.4	ACK38261	3.2.1.4	CAA11018	3.2.1.58	AFX59322	3.2.1.78
AAE59927	3.2.1.4	ACK41955	3.2.1.4	NP_010547	3.2.1.58	AGA35556	3.2.1.78
AAE60102	3.2.1.4	CAT16607	3.2.1.4	AAF65310	3.2.1.58	AGC24277	3.2.1.78
AAK60011	3.2.1.4	CAT16610	3.2.1.4	CAC07551	3.2.1.58	AGG69666	3.2.1.78
NP_126623	3.2.1.4	CAL91975	3.2.1.4	AAM08614	3.2.1.58	AGG69667	3.2.1.78
NP_143072	3.2.1.4	CAV28462	3.2.1.4	AAM08821	3.2.1.58	AGH62580	3.2.1.78
BAB62295	3.2.1.4	ACL75118	3.2.1.4	AAM21469	3.2.1.58	AGL50158	3.2.1.78
BAB62317	3.2.1.4	ACL75216	3.2.1.4	AAP53379	3.2.1.58	AGV01048	3.2.1.78
BAB62319	3.2.1.4	ACL75458	3.2.1.4	CAD97460	3.2.1.58	AGW24296	3.2.1.78
AAK94871	3.2.1.4	ACL76673	3.2.1.4	EAA62113	3.2.1.58	AGU71466	3.2.1.78
AAK85303	3.2.1.4	ACM59720	3.2.1.4	AAV05307	3.2.1.58	AHB89702	3.2.1.78
AAL16412	3.2.1.4	ACM60954	3.2.1.4	BAD97445	3.2.1.58	AHB89703	3.2.1.78
CAD17313	3.2.1.4	ACN43345	3.2.1.4	BAD97446	3.2.1.58	AHD18866	3.2.1.78

AAE84292	3.2.1.4	ACN62172	3.2.1.4	AAY28969	3.2.1.58	BAP19029	3.2.1.78
AAB61462	3.2.1.4	ACO55737	3.2.1.4	BAE58099	3.2.1.58	CDP31001	3.2.1.78
AAL83749	3.2.1.4	ACR23656	3.2.1.4	ABF50886	3.2.1.58	EDV05070	3.2.1.8
BAB86867	3.2.1.4	ACR23659	3.2.1.4	BAF26372	3.2.1.58	EEC54456	3.2.1.8
AAM23649	3.2.1.4	ACR59602	3.2.1.4	CAK43212	3.2.1.58	ADI70667	3.2.1.8
AAK21883	3.2.1.4	ACR82487	3.2.1.4	EDU47467	3.2.1.58	ADI70668	3.2.1.8
AAM42791	3.2.1.4	ACR87895	3.2.1.4	BAG89316	3.2.1.58		
AAM42791.1	3.2.1.4	CAZ67882	3.2.1.4	CAY69081	3.2.1.58		

Table A3

Table A3: Shown are the mean values of measured soil characteristics of the long-term tillage experiment for soil under conventional (CT), medium (MT) or reduced (RT) tillage (T) and under high (HF), medium (MF) and low (LF) fertilization (F), when appropriate. Furthermore, *P*-values for differences between tillage- or fertilization-treatment means are included, calculated using split-plot ANOVA. Different letters (in *Italics*) behind the mean values indicate a significant different mean. Values given per gram soil are calculated based on soil dry weight.

	Means			*P*-values		
	Grand	T	F	T	F	TxF
pH	6.285			0.444	0.593	0.763
Temperature	6.689			0.341	0.325	0.837
Gravimetric water content (%)		RT : 22.8 *a* MT: 21.9 *b* CT: 19.2 *c*		0.000	0.106	0.055
C content (%)		RT: 1.64 *a* MT: 1.37 *ab* CT: 1.10 *b*		0.022	0.776	0.505
N content (%)		RT: 0.17 *a* MT: 0.15 *ab* CT: 0.12 *b*		0.019	0.584	0.740
C_{mic} (mg g^{-1} soil)		RT: 0.303 *a* MT: 0.213 *b* CT: 0.152 *c*		0.003	0.396	0.207
N_{mic} (mg g^{-1} soil)		RT: 0.038 *a* MT: 0.030 *b* CT: 0.020 *c*		0.004	0.264	0.202
DOC (µg g^{-1} soil)		RT: 7.516 *a* MT: 1.953 *b* CT: 0.805 *b*		0.013	0.857	0.571
TDN (µg g^{-1} soil)		RT: 4.794 *a* MT: 2.090 *b* CT: 2.134 *b*		0.056	0.581	0.591
Nitrate (µg g^{-1} soil)	3.20			0.184	0.742	0.335
Ammonium (µg g^{-1} soil)	0.60			0.439	0.376	0.648
Xylosidase (pmol MU $hour^{-1}$ g^{-1} soil)		RT: 493,596 *a* MT: 374,997 *b*	HF: 433,252 *a* MF: 374,645 *b*	0.001	0.000	0.424

		CT: 297,153 *c*	LF: 357,849 *b*			
Cellobiohydrolase (pmol MU hour^{-1} g^{-1} soil)		RT: 26,620 *a* MT: 17,395 *b*		0.003	0.673	0.914
		CT: 6,437 *c*				
β-glucosidase (pmol MU hour^{-1} g^{-1} soil)		RT: 479,006 *a* MT: 370,931 *ab*		0.015	0.145	0.717
		CT: 266,171 *b*				
Xylosidase (pmol MU hour^{-1} g^{-1} soil µg^{-1} Cmic)			HF: 2,099 *a* MF: 1,703 *b* LF: 1,643 *b*	0.197	0.036	0.445
Cellobiohydrolase (pmol MU hour^{-1} g^{-1} soil µg^{-1} Cmic)		RT: 88.6 *a* MT: 83.3 *a* CT: 42.3 *b*		0.014	0.149	0.308
β-glucosidase (pmol MU hour^{-1} g^{-1} soil µg^{-1} Cmic)	1,747			0.694	0.179	0.349

Table A4

Table A4: Statistics of the metagenome sequencing of the long-term tillage experiment, shown per replicate of conventional (CT) and reduced (RT) tillage in the top 10 cm of soil.

Raw reads	**CT**	**CT**	**CT**	**RT**	**RT**	**RT**
Number of reads	130,726	129,916	184,890	226,876	226,771	199,436
Average length (bps)	638	641	640	633	626	630
Clean reads	**CT**	**CT**	**CT**	**RT**	**RT**	**RT**
Number of reads	112,985	112,585	159,373	193,285	193,988	170,423
Average length (bps)	412	413	421	408	405	403

Table A5

Table A5: Microbial groups or protein-coding genes affected by tillage treatment. Shown are averages (AVG), standard deviations (SD) and the total sum (Total) of relative abundances of reads (in % of all metagenome reads x 10^{-3}) annotated to microbial groups or protein-coding genes for the metagenomes of soil under conventional (CT) or reduced (RT) tillage from the long-term tillage experiment. Only the microbial groups or protein-coding genes that significantly differ between tillage treatments (paired t-test *P*-value <0.05) and of which the total relative abundance is higher than 0.01% of all metagenome reads are shown. An exception has been made for the cellulase-annotations and the taxonomic annotations of the cellulolytic enzymes and of the six most-abundant cellulase domain families (AA8, GH1, GH3, GH94, CBM2 and CBM6), which are less abundant than 0.01% of all metagenome reads. The shaded *P*-values are those which are not valid anymore after *P*-value correction (see **section 2**: Materials and Methods). If a microbial group or protein-coding gene is not abundant enough and neither has a valid *P*-value, it is not shown.

Microbial groups or protein-coding genes	CT (AVG)	CT (SD)	RT (AVG)	RT (SD)	Total	paired t-test *P*-value
Phylum						
Chloroflexi	1210.2	158.4	927.8	130.5	1037.8	0.009
Armatimonadetes	109.3	17.1	95.6	16.0	101.2	0.006

Crenarchaeota	23.4	4.5	16.7	4.3	19.5	0.042
Class						
Deltaproteobacteria	2966.4	222.3	3291.8	305.6	3154.5	0.027
Solibacteres	1120.1	52.4	1059.7	67.3	1083.4	0.048
Chloroflexia	416.2	47.3	314.6	44.2	355.8	0.012
Ktedonobacteria	106.9	3.7	77.0	7.1	89.5	0.041
Fimbriimonadia	66.0	11.3	53.4	11.2	58.3	0.045
Stigonematales	26.7	1.4	19.2	1.0	22.2	0.027
Order						
Solibacterales	1120.1	52.4	1059.7	67.3	1083.4	0.048
Chloroflexales	318.1	26.2	247.8	33.7	276.8	0.023
Ktedonobacterales	106.9	3.7	77.0	7.1	89.5	0.041
Fimbriimonadales	66.0	11.3	53.4	11.2	58.3	0.045
Stigonematales	26.7	1.4	19.2	1.0	22.2	0.027
KEGG Level 3-pathways						
Carbohydrate Metabolism	4604.2	15.0	4438.3	55.7	4521.3	0.049
Xenobiotics Biodegradation and Metabolism	1659.3	4.6	1531.7	22.4	1595.5	0.015
KEGG Level 4-pathways						
Glyoxylate and dicarboxylate metabolism	671.9	27.6	642.6	17.0	657.3	0.042
Valine, leucine and isoleucine degradation	625.0	23.8	572.1	26.5	598.6	0.030
Nitrogen metabolism	1039.8	19.6	994.1	26.5	1017.0	0.042
Methane metabolism	781.9	29.1	734.3	19.4	758.1	0.021
Fatty acid metabolism	547.8	18.3	492.4	16.7	520.1	0.002
Selenocompound metabolism	303.1	11.4	315.6	8.9	309.4	0.029
beta-Alanine metabolism	300.4	24.3	270.0	14.7	285.2	0.032
Toluene degradation	177.2	11.5	153.8	17.4	165.5	0.024
Drug metabolism - other enzymes	134.6	1.5	117.9	2.6	126.3	0.004
Styrene degradation	108.5	5.8	91.1	4.6	99.8	0.022
Ethylbenzene degradation	72.9	7.2	58.0	5.3	65.5	0.046
Arachidonic acid metabolism	66.7	1.7	58.2	2.1	62.4	0.044
Nonribosomal peptide structures	34.1	4.8	38.9	4.8	36.5	0.017
Cellulase domain family AA8						
Solirubrobacterales	0.8	0.1	0.3	0.3	0.5	0.034
Cellulase domain family GH3						
Chloroflexi	1.0	0.9	1.6	1.0	1.4	0.027

Table A6

Table A6: Shown are the mean values of measured soil characteristics of the organic tillage experiment for soil under different farming systems (F, organic farming with plough tillage (CT), organic farming with reduced tillage (RT), conventional farming with plough tillage (PT) and conventional farming under no-tillage (NT)), under different cover crop treatments (CC, legume (L) or no cover crop (NO)) and in two different depths (D, in the top 0-6 cm (Layer 0) and 10-16 cm (Layer 1) from the soil surface), when appropriate. Furthermore, *P*-values for factor effects are included, calculated using split-split-plot ANOVA. Different letters (in *Italics*) behind the mean values indicate a significant different mean. Values given per gram soil are calculated based

on soil dry weight.

	Means		*P*-values						
			F	CC	D	FxCC	FxD	DxCC	FxCCxD
pH	Layer 1: 7.912 *a*		0.258	0.831	0.000	0.131	0.105	0.810	0.683
	Layer 0: 7.762 *b*								
Gravimetric water content (%)	RT0: 19.7 *a*		0.358	0.188	0.115	0.512	0.001	0.747	0.299
	PT1: 19.7 *a*								
	NT0: 19.7 *ab*								
	NT1: 19.3 *ab*								
	CT1: 19,2 *ab*								
	RT1: 19.0 *bc*								
	CT0: 18.4 *c*								
	PT0: 18.3 *c*								
C content (%)	Layer 0: 1.77 *a*		0.766	0.177	0.008	0.065	0.630	0.164	0.186
	Layer 1: 1.45 *b*								
N content (%)	Layer 0: 0.17 *a*		0.229	0.377	0.006	0.781	0.529	0.118	0.702
	Layer 1: 0.15 *b*								
C_{mic} (mg g^{-1} soil)	RT0: 0.988 *a*		0.331	0.259	0.013	0.875	0.001	0.777	0.818
	NT0: 0.767 *b*								
	PT1: 0.754 *b*								
	CT1: 0.665 *b*								
	CT0: 0.663 *b*								
	RT1: 0.629 *bc*								
	PT0: 0.585 *bc*								
	NT1: 0.448 *c*								
N_{mic} (mg g^{-1} soil)	RT0: 0.144 *a*		0.161	0.246	0.000	0.068	0.000	0.838	0.228
	NT0: 0.103 *b*								
	CT0: 0.103 *bc*								
	CT1: 0.097 *bc*								
	PT0: 0.094 *bc*								
	PT1: 0.091 *bcd*								
	RT1: 0.091 *cd*								
	NT1: 0.080 *d*								
DOC (µg g^{-1} soil)	CT1_L: 208.84	*a*	0.621	0.274	0.798	0.654	0.820	0.204	0.007
	CT0_NO: 205.01	*a*							
	NT1_ NO: 154.7	*ab*							
	NT0_L: 120.31	*abc*							
	RT0_L: 115.60	*abc*							
	RT1_L: 107.13	*abc*							
	PT0_ NO: 84.93	*abc*							
	PT1_L: 62.88	*abc*							
	RT1_ NO: 61.59	*abc*							
	NT1_L: 59.83	*abc*							
	PT0_L: 59.82	*abc*							
	NT0_ NO: 48.70	*bc*							
	CT0_L: 31.65	*bc*							
	RT0_ NO: 29.70	*bc*							
	PT1_ NO: -0.66	*c*							

	CT1_ NO: -13.66 *c*							
TDN (µg g^{-1} soil)	4.735	0.492	0.519	0.636	0.533	0.779	0.867	0.392
Nitrate (µg g^{-1} soil)	NT: 5.40 *a* RT: 3.81 *ab* PT: 3.61 *b* CT: 2.43 *b*	0.020	0.378	0.454	0.354	0.926	0.659	0.171
Ammonium (µg g^{-1} soil)	Layer 0: 0.28 *a* Layer 1: 0.10 *b*	0.773	0.588	0.000	0.444	0.630	0.887	0.188

Table A7

Table A7: Statistics of the metagenome sequencing of the organic tillage experiment, shown per replicate of conventional (CT) and reduced (RT) tillage in the top 0-6 cm and 10-16 cm from the soil surface. NC= negative sequencing control.

Sample	**CT 0-6cm**	**CT 0-6cm**	**CT 0-6cm**	**CT 0-6cm**	**RT 0-6cm**	**RT 0-6cm**	**RT 0-6cm**	**RT 0-6cm**	**NC 0-6cm**
number of raw reads	3.789 x 10^6	2.530 x 10^6	3.265 x 10^6	3.559 x 10^6	2.606 x 10^6	3.234 x 10^6	3.671 x 10^6	3.502 x 10^6	382
Average length (bps)	300	300	300	300	300	300	300	300	300
number of clean reads	3.694 x 10^6	2.483 x 10^6	3.239 x 10^6	3.490 x 10^6	2.553 x 10^6	3.145 x 10^6	3.626 x 10^6	3.398 x 10^6	161
Average length (bps)	230	214	206	216	216	232	235	242	215
Sample	**CT 10-16cm**	**CT 10-16cm**	**CT 10-16cm**	**CT 10-16cm**	**RT 10-16cm**	**RT 10-16cm**	**RT 10-16cm**	**RT 10-16cm**	**NC 10-16cm**
number of raw reads	3.533 x 10^6	3.431 x 10^6	3.302 x 10^6	3.932 x 10^6	3.548 x 10^6	3.387 x 10^6	3.114 x 10^6	3.477 x 10^6	634
Average length (bps)	300	300	300	300	300	300	300	300	300
number of clean reads	3.389 x 10^6	3.286 x 10^6	3.100 x 10^6	3.732 x 10^6	3.392 x 10^6	3.233 x 10^6	2.931 x 10^6	3.344 x 10^6	447
Average length (bps)	287	291	305	297	276	300	298	280	273

Table A8

Table A8: Microbial groups or protein-coding genes affected by tillage treatment. Shown are averages (AVG), standard deviations (SD) and the total sum (Total) of relative abundances of reads (in % of all metagenome reads x 10^{-3}) annotated to microbial groups or protein-coding genes for the metagenomes of the top 0-6 cm or top 10-16 cm soil under conventional (CT) or reduced (RT) tillage from the organic tillage experiment. Only the microbial groups or protein-coding genes that significantly differ (ANOVA *P*-value <0.05) between tillage treatments (Effect=T) or soil depths (Effect=D) or both (Effect=I) and of which the total relative abundance is higher than 0.01% of all metagenome reads. An exception has been made for the cellulase-annotations and the taxonomic assignments of the cellulolytic enzymes and of the six most-abundant cellulase domain families (AA8, GH1, GH3, GH94, CBM2 and CBM6), which are less abundant than 0.01% of all metagenome reads. The

shaded *P*-values are those which are not valid anymore after *P*-value correction (see **section 2:** Materials and Methods). If a microbial group or protein-coding gene is not abundant enough and neither has a valid *P*-value, it is not shown.

	AVG 0-6 cm		AVG 10-16 cm		SD 0-6 cm		SD 10-16 cm				
	CT	RT	CT	RT	CT	RT	CT	RT	Total	*P*-value	Effect
Phylum											
Acidobacteria	2282.4	2235.6	3184.4	3397.1	332.7	221.8	213.8	266.3	2784.4	0.000	D
Armatimonadetes	50.9	49.2	74.3	71.9	4.4	8.4	4.8	7.4	62.0	0.000	D
Basidiomycota	8.2	10.5	10.7	16.3	3.0	1.1	2.3	1.3	11.4	0.007	D
Chordata	11.3	11.5	14.8	17.4	3.8	0.8	0.6	0.8	13.7	0.002	D
Crenarchaeota	32.7	33.0	53.9	58.6	6.1	7.6	6.4	2.0	44.6	0.001	D
Deinococcus-Thermus	79.1	78.6	98.1	98.6	1.6	2.0	1.8	3.4	88.9	0.000	D
Euryarchaeota	106.2	103.3	137.9	143.4	4.9	3.0	5.5	2.9	123.1	0.000	D
Gemmatimonadetes	687.5	668.0	984.7	996.2	50.8	142.5	66.3	102.2	838.6	0.000	D
Microgenomates	26.4	26.0	40.3	44.2	3.7	1.9	1.4	1.2	34.3	0.000	D
Nitrospirae	300.7	288.1	437.1	440.4	33.7	12.0	31.6	14.4	368.0	0.000	D
Parcubacteria	58.0	57.3	86.1	95.3	7.3	1.3	5.5	3.1	74.4	0.000	D
Planctomycetes	1501.9	1492.1	1783.7	1777.1	94.3	85.3	127.2	62.3	1640.3	0.000	D
Spirochaetes	43.9	43.1	59.8	59.7	1.9	1.9	2.6	2.0	51.8	0.000	D
Thaumarchaeota	336.8	330.2	510.5	537.0	68.3	72.7	34.9	18.2	428.9	0.001	D
Verrucomicrobia	1233.2	1403.2	1554.0	1628.5	201.9	103.4	182.9	114.3	1459.1	0.001	D
Actinobacteria	8905.8	9731.6	8147.6	7384.3	399.6	535.0	540.7	740.0	8542.0	0.004	I
Ascomycota	61.3	77.7	71.2	61.5	10.6	7.3	11.4	9.9	67.8	0.016	I
Chlamydiae	12.0	11.4	14.7	17.3	1.6	0.8	1.1	1.7	13.8	**0.046**	I
Chlorobi	19.7	18.6	25.4	28.3	1.3	0.5	1.3	1.5	23.0	0.007	I
Chloroflexi	559.1	543.2	744.5	808.9	12.1	61.4	11.7	46.7	666.7	0.002	I
Cyanobacteria	596.8	539.2	661.7	677.3	26.1	7.0	21.4	18.5	620.1	0.009	I
Firmicutes	593.5	583.3	718.3	747.4	25.0	8.2	10.4	13.3	661.6	0.017	I
Ignavibacteriae	16.2	15.8	23.9	27.6	2.2	1.5	2.3	1.7	21.0	0.003	I
Proteobacteria	18499.3	19877.8	21035.5	19575.9	265.9	897.4	453.2	414.3	19770.0	0.001	I
Family											
Acidobacteriaceae	688.6	684.8	943.8	1011.6	57.9	69.6	52.7	111.8	834.7	0.000	D
Holophagaceae	18.9	19.3	27.8	29.4	1.5	1.4	2.0	2.1	23.9	0.000	D
Rubrobacteraceae	83.7	78.5	76.8	66.4	15.3	15.2	13.2	4.1	76.3	**0.021**	D
Patulibacteraceae	36.6	36.0	34.2	32.1	4.3	1.6	2.4	1.6	34.8	**0.023**	D
Nitriliruptoraceae	22.2	21.9	25.7	23.9	1.4	2.5	2.8	2.2	23.4	0.016	D
Hyphomicrobiaceae	148.0	145.4	172.3	161.9	7.3	9.8	3.0	14.7	157.2	0.000	D
Rhodobacteraceae	158.8	156.3	149.8	131.9	15.9	24.7	18.5	11.1	148.8	0.005	D

Methylocystaceae	25.3	26.4	27.9	27.8	1.0	1.1	0.3	1.4	26.9	0.003	D
Burkholderiaceae	242.5	257.7	288.4	278.8	6.6	12.0	15.8	5.1	267.1	0.001	D
Nitrosomonadaceae	79.0	77.4	89.5	90.8	7.1	3.8	7.6	10.6	84.0	**0.042**	D
Chromobacteriaceae	17.5	18.4	21.4	20.9	1.0	1.3	1.1	1.0	19.6	0.001	D
Methylophilaceae	12.1	13.1	15.9	14.7	1.6	1.0	0.9	1.9	13.9	0.003	D
Methylococcaceae	74.6	75.1	97.9	99.3	2.7	3.9	2.6	4.6	86.9	0.000	D
Chromatiaceae	49.4	49.5	63.6	62.8	1.5	1.2	4.5	2.3	56.4	0.000	D
Alteromonadaceae	36.1	37.3	48.5	49.0	1.0	2.7	2.9	2.3	42.9	0.000	D
Thiotrichaceae	28.1	27.9	36.8	36.2	1.0	2.0	1.3	1.8	32.4	0.000	D
Competibacteraceae	18.8	19.3	24.1	24.6	1.7	1.3	1.3	2.5	21.7	0.001	D
Vibrionaceae	17.4	18.2	21.7	22.8	0.9	1.2	1.0	1.4	20.1	0.001	D
Legionellaceae	14.1	14.7	18.5	19.7	0.5	1.0	0.7	1.8	16.8	0.000	D
Halomonadaceae	10.8	10.8	12.8	12.1	0.2	0.5	0.9	0.9	11.6	0.003	D
Kofleriaceae	139.5	137.2	202.0	182.1	16.3	19.4	13.9	5.7	166.1	0.000	D
Anaeromyxobacteraceae	81.7	82.5	106.1	100.5	4.4	17.3	10.8	5.3	93.0	0.001	D
Geobacteraceae	62.7	65.5	91.9	98.3	3.2	12.2	9.8	17.9	79.8	0.000	D
Desulfobacteraceae	62.1	61.1	84.1	89.6	2.9	4.6	3.4	6.2	74.4	0.000	D
Desulfovibrionaceae	35.2	35.6	44.9	47.1	1.5	1.2	2.6	0.8	40.8	0.000	D
Desulfobulbaceae	16.3	14.9	21.4	21.9	0.8	1.4	0.9	0.7	18.6	0.000	D
Syntrophobacteraceae	13.2	14.0	20.1	22.0	0.8	0.8	1.0	1.5	17.4	0.000	D
Vulgatibacteraceae	13.1	13.6	18.1	17.7	0.6	0.9	0.7	0.6	15.7	0.000	D
Pelobacteraceae	9.0	10.0	12.6	13.6	1.0	2.0	1.6	2.7	11.3	0.000	D
Chthonomonadaceae	14.8	14.4	21.0	22.0	1.7	1.3	2.2	1.2	18.2	0.000	D
Flammeovirgaceae	26.6	31.1	35.9	33.7	2.9	4.7	2.6	1.4	31.9	0.013	D
Rhodothermaceae	23.7	23.3	34.0	35.3	1.0	1.1	1.0	1.7	29.1	0.000	D
Cyclobacteriaceae	23.5	25.7	30.5	29.4	2.6	2.5	1.5	1.3	27.3	0.000	D
Prolixibacteraceae	9.0	9.4	11.7	12.6	0.7	0.8	0.4	0.6	10.7	0.000	D
Roseiflexaceae	53.6	53.8	68.5	71.6	7.4	5.6	7.9	6.5	62.2	0.000	D
Caldilineaceae	54.1	53.8	64.9	65.3	0.9	5.5	4.7	3.2	59.5	0.000	D
Sphaerobacteraceae	53.8	53.0	65.8	65.2	2.5	4.7	2.5	4.7	59.5	0.000	D
Ktedonobacteraceae	39.4	38.5	49.0	51.2	2.5	1.2	2.4	1.9	44.7	0.000	D

Herpetosiphonaceae	22.9	22.9	31.9	29.4	1.8	0.6	4.1	1.8	26.9	0.000	D
Chloroflexaceae	20.8	20.8	27.4	29.1	1.5	1.0	1.4	1.2	24.6	0.000	D
Oscillochloridaceae	10.6	10.9	13.8	14.8	0.8	0.6	1.7	0.6	12.6	0.000	D
Thermogemmatisporaceae	8.8	8.4	11.1	11.7	0.8	0.6	1.3	0.3	10.1	0.000	D
Thermaceae	36.9	36.3	46.1	47.9	1.7	1.4	2.1	0.9	41.9	0.000	D
Deinococcaceae	30.1	30.5	37.1	36.1	1.7	1.3	0.2	0.8	33.6	0.000	D
Halobacteriaceae	29.1	27.9	37.7	38.2	1.2	1.8	1.6	1.1	33.3	0.000	D
Methanosarcinaceae	21.1	19.7	25.7	27.0	1.2	1.1	1.2	1.2	23.4	0.000	D
Paenibacillaceae	66.8	65.2	81.9	82.4	1.7	1.5	1.3	1.7	74.2	0.000	D
Peptococcaceae	35.5	35.9	44.4	45.2	1.6	1.0	1.5	1.5	40.2	0.000	D
Clostridiaceae	20.0	20.6	24.1	26.4	1.1	1.0	1.6	2.0	22.8	0.000	D
Ruminococcaceae	11.2	11.0	14.2	14.1	0.3	1.1	1.2	1.8	12.6	0.001	D
Thermoanaerobacteraceae	10.7	10.3	13.4	14.1	0.6	0.5	1.1	1.1	12.1	0.000	D
Alicyclobacillaceae	10.2	10.2	11.6	12.6	0.5	0.4	0.8	1.2	11.2	0.001	D
Gemmatimonadaceae	177.9	179.0	239.9	233.3	17.4	16.1	21.6	8.0	208.5	0.000	D
Nitrospiraceae	300.6	288.0	437.0	440.2	14.5	31.6	12.1	33.8	367.9	0.000	D
Planctomycetaceae	1383.8	1373.0	1610.8	1603.8	59.5	127.3	79.0	87.4	1494.1	0.000	D
Candidatus Brocadiaceae	33.4	32.1	47.9	51.2	2.0	2.3	1.4	2.9	41.3	0.000	D
Phycisphaeraceae	26.5	29.0	42.6	38.8	0.7	2.1	3.3	2.3	34.3	0.000	D
Leptospiraceae	25.4	25.3	35.4	34.7	1.6	1.8	0.9	1.0	30.3	0.000	D
Spirochaetaceae	17.2	16.4	22.8	23.3	0.3	1.0	1.3	0.8	20.0	0.000	D
Nitrososphaeraceae	217.6	211.9	330.9	346.7	11.1	22.2	50.1	45.8	277.0	0.001	D
Verrucomicrobia subdivision 3	357.0	377.5	525.3	621.9	51.4	76.4	53.6	98.8	472.9	0.000	D
Solibacteraceae	329.2	323.8	438.9	470.6	41.8	41.9	38.5	56.8	391.9	**0.034**	I
Mycobacteriaceae	769.3	841.6	636.5	596.9	41.0	36.4	23.9	22.7	709.9	**0.016**	I
Nocardioidaceae	359.0	420.3	269.6	232.2	42.9	45.2	18.5	30.2	320.7	0.008	I
Solirubrobacteraceae	314.0	353.6	285.0	265.4	37.4	11.3	13.9	10.7	305.0	**0.016**	I
Micromonosporaceae	263.6	303.9	291.6	274.4	7.1	33.1	21.3	19.2	283.3	0.012	I
Pseudonocardiaceae	238.9	256.4	228.0	206.7	20.4	9.4	14.6	13.8	232.8	0.002	I
Acidimicrobiaceae	187.7	219.1	166.4	135.8	22.4	33.5	18.2	21.5	176.8	0.001	I
Microbacteriaceae	149.1	199.3	117.7	85.8	24.3	34.3	28.0	14.2	137.1	0.002	I
Propionibacteriaceae	129.9	151.8	128.2	114.8	14.7	17.5	20.6	10.9	131.1	0.008	I

Geodermatophilaceae	145.0	142.9	125.9	97.4	18.2	11.3	16.9	4.7	127.7	**0.038**	I
Nocardiaceae	114.9	128.8	110.0	104.4	6.4	6.7	7.7	2.8	114.5	**0.015**	I
Intrasporangiaceae	113.2	141.8	93.6	81.3	9.0	16.4	8.0	6.2	107.3	0.004	I
Micrococcaceae	127.1	135.5	84.2	70.4	7.7	4.4	2.9	5.0	104.0	0.005	I
Frankiaceae	85.5	90.2	85.9	79.4	8.5	3.3	3.6	5.9	85.5	**0.027**	I
Cellulomonadaceae	42.9	60.9	38.8	29.3	4.6	8.4	8.6	4.4	42.9	0.002	I
Nakamurellaceae	34.5	58.8	26.7	23.1	2.3	2.9	1.9	2.6	35.6	0.000	I
Sphingomonadaceae	612.4	677.5	636.1	514.1	52.9	25.9	38.8	11.1	611.2	0.000	I
Rhodospirillaceae	300.0	308.9	312.4	253.6	75.8	29.9	91.2	15.1	292.5	0.003	I
Methylobacteriaceae	271.1	311.3	246.5	209.2	23.2	27.4	36.3	7.6	258.6	0.003	I
Phyllobacteriaceae	226.7	241.5	224.8	207.7	18.1	37.7	19.4	19.0	224.8	**0.036**	I
Rhizobiaceae	174.4	189.7	183.4	167.2	7.0	8.2	10.6	4.4	178.9	0.007	I
Caulobacteraceae	99.9	115.4	105.9	83.2	6.5	24.2	5.0	4.5	101.5	**0.018**	I
Acetobacteraceae	74.8	79.4	66.9	62.3	6.1	4.7	7.0	2.8	70.7	**0.020**	I
Erythrobacteraceae	36.7	44.8	37.9	30.0	6.6	5.3	5.5	2.8	37.3	0.003	I
Hyphomonadaceae	20.9	21.7	27.0	24.8	1.0	1.1	0.9	0.8	23.7	0.001	I
Comamonadaceae	360.1	437.6	381.9	300.7	25.0	57.3	16.4	15.8	370.3	0.000	I
Oxalobacteraceae	96.8	111.4	111.8	102.9	4.1	9.7	2.2	3.1	105.9	0.001	I
Rhodocyclaceae	74.3	83.1	96.8	97.4	3.1	7.5	5.9	3.4	88.0	**0.030**	I
Alcaligenaceae	56.2	60.9	65.9	61.7	1.0	1.7	3.4	0.8	61.2	0.007	I
Xanthomonadaceae	396.5	445.9	416.4	326.4	29.1	58.4	38.5	11.3	396.7	0.003	I
Pseudomonadaceae	132.2	157.6	154.4	147.0	12.7	13.4	13.0	10.0	148.0	**0.027**	I
Ectothiorhodospiraceae	33.9	36.6	46.2	42.7	2.2	1.9	3.9	1.9	39.9	0.011	I
Polyangiaceae	434.3	474.6	529.1	473.6	19.1	50.5	21.8	38.9	479.8	0.006	I
Cystobacteraceae	198.0	216.8	228.0	199.3	14.5	13.9	11.5	8.9	211.5	0.001	I
Labilitrichaceae	157.1	176.4	188.8	152.5	11.3	18.7	6.2	13.1	169.1	0.000	I
Sandaracinaceae	88.2	100.5	113.7	100.3	4.3	8.1	7.7	6.2	100.9	0.009	I
Myxococcaceae	85.2	92.9	105.3	96.7	3.7	5.4	6.9	5.0	95.2	0.005	I
Syntrophaceae	25.7	23.5	35.3	38.9	2.1	2.1	1.6	2.3	30.9	0.003	I
Nannocystaceae	18.5	21.1	24.9	22.5	0.7	1.5	2.3	0.9	21.8	0.003	I
Bdellovibrionaceae	15.6	17.6	17.0	13.9	0.5	1.4	2.3	0.4	16.0	0.012	I
Saprospiraceae	22.9	28.0	29.8	29.1	2.8	4.7	1.4	2.0	27.5	**0.044**	I
Chlorobiaceae	19.6	18.6	25.4	28.2	1.5	1.3	0.5	1.3	23.0	0.007	I

Anaerolineaceae	84.6	76.3	126.5	162.3	14.0	6.3	25.4	9.9	113.1	0.001	I
Ardenticatenaceae	15.9	15.3	21.3	23.6	0.9	0.2	1.3	0.5	19.1	0.001	I
Microchaetaceae	45.9	58.6	61.6	53.4	5.5	9.2	4.9	1.2	54.9	0.012	I
Nostocaceae	32.6	24.6	25.0	24.4	6.1	2.7	1.8	0.9	26.6	**0.023**	I
Scytonemataceae	16.3	14.9	20.0	20.9	1.2	1.1	0.8	1.8	18.1	**0.029**	I
Bacillaceae	64.8	62.6	75.2	79.9	2.2	1.8	1.9	3.4	70.7	**0.025**	I
Veillonellaceae	12.0	11.2	13.3	14.8	1.2	0.8	0.6	0.9	12.8	0.007	I
Clostridiales Family XVII. Incertae Sedis	10.4	9.5	11.8	12.5	0.7	0.2	0.5	0.3	11.1	**0.020**	I
Ignavibacteriaceae	8.2	7.8	12.0	13.8	0.8	1.0	1.2	1.0	10.5	0.004	I
Verrucomicrobiaceae	182.6	214.7	201.4	188.4	11.9	18.2	23.1	18.3	197.0	**0.040**	I
Opitutaceae	152.1	178.3	162.6	146.3	16.0	20.5	10.2	14.6	159.9	0.009	I
Level 4-pathways											
Pyruvate metabolism	794.7	828.7	855.8	826.1	25.4	9.4	20.0	11.1	827.2	0.009	I
Glycolysis / Gluconeogenesis	753.5	786.0	831.7	808.2	18.6	4.6	12.8	6.2	795.7	0.003	I
Amino sugar and nucleotide sugar metabolism	586.1	616.5	684.8	671.0	11.8	12.3	9.7	7.4	640.6	0.012	I
Butanoate metabolism	666.0	696.2	738.6	704.9	18.9	6.7	15.6	5.3	702.3	0.001	I
Starch and sucrose metabolism	581.6	621.8	668.5	645.3	9.3	5.7	17.1	8.0	629.9	0.002	I
Propanoate metabolism	668.7	695.6	727.5	705.6	14.7	8.4	10.4	6.6	700.2	0.003	I
Glyoxylate and dicarboxylate metabolism	653.5	680.2	713.8	688.2	16.2	6.5	10.7	8.5	684.6	0.004	I
Citrate cycle (TCA cycle)	612.2	635.0	664.9	642.7	13.4	4.2	10.1	10.9	639.3	0.003	I
Pentose phosphate pathway	461.9	473.7	522.5	518.9	13.6	8.7	2.4	10.6	494.9	0.000	D
Fructose and mannose metabolism	335.6	351.7	392.9	378.6	9.7	7.6	8.6	6.9	365.4	0.011	I
Galactose metabolism	258.0	273.2	308.8	301.1	6.8	7.2	2.8	6.8	285.8	0.009	I
Pentose and glucuronate interconversions	253.0	271.6	285.2	274.7	7.4	7.8	10.1	3.2	271.5	0.010	I
C5-Branched dibasic acid metabolism	160.0	162.6	177.9	169.6	3.3	2.1	2.2	1.8	167.7	0.005	I
Ascorbate and	142.1	149.8	156.9	153.3	3.9	3.7	2.4	1.9	150.8	0.020	I

aldarate metabolism											
Inositol phosphate metabolism	110.3	115.9	124.6	117.4	3.5	0.4	3.3	2.9	117.2	0.001	I
Arginine and proline metabolism	713.7	741.6	801.0	773.3	14.3	10.5	12.7	10.5	758.4	0.004	I
Glycine. serine and threonine metabolism	600.0	622.4	674.2	635.8	18.8	11.0	9.6	9.8	634.1	0.014	I
Alanine. aspartate and glutamate metabolism	656.0	682.6	723.2	705.9	9.8	8.6	4.7	3.7	692.7	0.002	I
Valine. leucine and isoleucine degradation	592.8	619.4	632.7	606.8	19.7	11.0	13.8	14.5	613.7	0.008	I
Cysteine and methionine metabolism	450.0	469.4	517.6	499.8	8.9	6.0	7.0	6.0	484.9	0.004	I
Tyrosine metabolism	415.2	441.4	487.9	461.8	12.3	3.9	10.6	3.6	452.4	0.001	I
Phenylalanine metabolism	415.1	437.2	479.9	450.0	12.8	8.9	11.8	5.9	446.3	0.004	I
Histidine metabolism	371.9	391.6	426.1	403.0	14.4	8.7	10.2	5.2	398.8	0.020	I
Tryptophan metabolism	382.2	400.3	420.2	400.5	11.0	10.5	5.1	2.2	401.3	0.010	I
Phenylalanine. tyrosine and tryptophan biosynthesis	332.6	347.3	383.1	364.2	11.2	5.8	10.4	2.6	357.3	0.015	I
Valine. leucine and isoleucine biosynthesis	300.2	310.3	331.6	317.6	4.2	5.6	6.6	2.5	315.0	0.009	I
Lysine degradation	300.0	314.8	326.8	311.9	7.1	9.4	6.1	5.3	313.7	0.006	I
Lysine biosynthesis	209.8	221.7	240.2	225.9	4.7	8.3	4.5	0.9	224.7	0.005	I
Nitrogen metabolism	932.2	980.9	1081.7	1065.6	17.8	14.0	12.4	17.3	1016.5	0.015	I
Oxidative phosphorylation	779.8	823.6	911.5	896.9	19.0	10.0	12.4	9.9	854.2	0.013	I
Carbon fixation pathways in prokaryotes	773.8	797.9	839.2	819.8	16.5	8.5	11.9	7.2	808.6	0.011	I
Methane metabolism	732.4	752.3	809.2	783.9	26.1	5.3	14.4	9.1	770.6	0.001	D
Carbon fixation in photosynthetic organisms	291.8	301.2	313.1	304.5	8.3	3.8	7.0	4.8	302.9	0.024	I
Sulfur metabolism	159.8	168.2	174.1	169.0	4.3	2.4	2.2	0.4	168.0	0.007	I

Photosynthesis	100.6	104.9	115.6	113.1	3.6	1.9	2.2	1.6	108.7	0.021	I
Purine metabolism	1241.6	1297.5	1427.2	1379.5	20.6	19.3	29.6	7.7	1338.2	0.004	I
Pyrimidine metabolism	840.1	875.1	945.9	916.8	17.6	15.9	9.8	10.5	895.6	0.006	I
Fatty acid metabolism	511.2	535.5	554.1	528.7	16.2	9.2	11.3	7.0	533.0	0.003	I
Fatty acid biosynthesis	267.9	277.7	308.6	298.4	5.1	8.6	6.6	1.4	288.6	0.023	I
Glycerophospholipid metabolism	176.3	188.5	215.6	205.9	5.5	5.2	3.6	4.0	197.1	0.004	I
Glycerolipid metabolism	163.6	173.1	179.3	171.6	4.7	4.0	2.4	2.4	172.1	0.003	I
Biosynthesis of unsaturated fatty acids	151.1	156.4	178.5	174.3	4.7	8.2	7.5	0.4	165.5	0.001	D
Sphingolipid metabolism	132.2	137.4	150.4	149.2	5.4	3.7	6.3	3.1	142.6	0.000	D
Steroid hormone biosynthesis	106.3	108.4	119.1	117.5	3.7	3.2	5.1	2.2	113.0	0.000	D
Synthesis and degradation of ketone bodies	98.4	103.9	102.5	98.1	3.3	2.8	4.3	2.6	100.8	0.015	I
Linoleic acid metabolism	77.5	83.8	93.6	87.8	2.3	2.7	2.4	2.1	85.9	0.001	I
Arachidonic acid metabolism	55.3	57.1	64.9	63.0	2.7	1.4	1.9	1.7	60.2	0.001	D
Ether lipid metabolism	15.6	16.3	18.8	18.4	1.1	1.8	1.3	1.5	17.4	0.000	D
Aminobenzoate degradation	308.6	327.1	369.5	352.8	11.0	8.3	5.2	4.5	340.3	0.003	I
Benzoate degradation	310.3	327.0	361.0	342.9	10.2	8.8	5.4	0.9	336.0	0.003	I
Chloroalkane and chloroalkene degradation	297.0	316.4	349.1	338.2	7.9	7.7	3.5	5.3	325.8	0.001	I
Bisphenol degradation	152.1	165.8	180.8	171.4	5.5	3.5	4.8	3.3	167.9	0.001	I
Toluene degradation	152.6	158.3	174.6	170.5	4.4	3.9	2.9	1.5	164.3	**0.030**	I
Polycyclic aromatic hydrocarbon degradation	129.6	140.9	160.0	152.2	5.1	2.4	3.5	3.0	146.0	0.001	I
Metabolism of xenobiotics by cytochrome P450	121.6	131.3	142.6	135.7	6.6	2.5	6.1	1.9	133.1	0.003	I
Nitrotoluene degradation	126.9	121.7	137.6	138.1	7.3	3.3	3.6	2.0	131.4	0.000	D
Drug metabolism - cytochrome P450	117.2	126.8	137.4	129.6	6.2	1.7	4.5	2.1	128.0	0.003	I

Naphthalene degradation	114.7	125.0	131.5	125.1	5.9	1.2	3.4	2.8	124.3	0.003	I
Drug metabolism - other enzymes	112.2	119.5	127.0	123.7	3.7	2.3	4.5	2.6	120.7	0.019	I
Chlorocyclohexane and chlorobenzene degradation	102.9	108.7	120.7	116.9	3.2	3.3	2.2	2.0	112.5	0.014	I
Caprolactam degradation	102.7	109.1	115.2	109.7	4.3	2.4	2.2	1.9	109.4	0.010	I
Styrene degradation	97.9	100.6	111.1	105.0	1.9	3.8	3.2	1.9	103.9	0.009	I
Ethylbenzene degradation	54.3	59.5	64.2	59.7	2.2	3.1	2.1	0.7	59.5	0.007	I
Fluorobenzoate degradation	49.2	49.2	58.0	56.5	2.0	2.2	1.8	1.0	53.3	0.000	D
Dioxin degradation	41.9	44.6	50.5	46.8	1.9	1.5	0.9	1.4	46.1	0.007	I
Xylene degradation	36.5	38.3	42.3	39.6	1.1	1.6	1.0	0.6	39.3	0.006	I
Atrazine degradation	31.5	34.5	32.8	30.6	1.3	0.9	1.6	1.4	32.4	0.008	I
Steroid degradation	26.1	27.8	32.1	30.7	2.1	1.1	1.2	0.4	29.3	0.002	D
Two-component system	1272.6	1358.5	1587.5	1555.9	48.1	48.0	10.4	50.2	1446.9	0.010	I
PI3K-Akt signaling pathway	31.1	35.1	37.9	36.0	1.1	1.7	2.5	1.1	35.1	0.007	I
Phosphatidylinositol signaling system	24.2	25.1	28.6	27.1	1.2	1.3	1.0	0.9	26.3	**0.032**	I
MAPK signaling pathway - yeast	11.3	12.1	14.7	14.1	0.5	1.3	0.4	0.8	13.1	0.001	D
ABC transporters	1192.5	1258.6	1351.4	1312.1	30.3	7.9	38.1	5.5	1280.4	0.008	I
Bacterial secretion system	266.8	289.9	330.6	319.8	10.9	6.2	10.0	6.1	302.3	0.018	I
Phosphotransferase system (PTS)	38.2	40.6	42.9	41.4	0.7	1.9	0.8	1.4	40.9	0.018	I
Aminoacyl-tRNA biosynthesis	648.1	669.5	726.7	706.1	9.9	6.7	9.3	15.9	688.4	0.014	I
Ribosome	357.3	372.7	407.1	394.9	5.9	8.3	9.8	7.8	383.5	**0.043**	I
RNA transport	21.0	22.4	29.0	28.9	0.7	1.5	1.3	2.0	25.4	0.000	D
Ribosome biogenesis in eukaryotes	15.6	17.1	18.2	16.6	0.6	1.6	0.4	1.0	16.9	**0.045**	I
RNA polymerase	166.3	171.5	183.3	181.1	3.1	0.8	5.1	3.1	175.6	0.001	D
Peroxisome	196.2	208.5	222.0	210.3	3.9	3.1	3.5	1.6	209.5	0.001	I
Lysosome	43.7	48.3	56.6	55.8	2.1	1.7	1.2	1.8	51.2	0.019	I
Mismatch repair	376.9	398.5	436.6	420.1	12.7	8.0	6.8	5.7	408.7	0.012	I
Nucleotide excision repair	416.7	435.9	466.4	449.3	10.9	7.0	8.3	6.0	442.6	0.013	I

Homologous recombination	304.1	324.0	346.3	329.0	9.4	6.0	5.3	3.9	326.4	0.001	I
Base excision repair	278.4	289.8	318.1	307.1	13.0	7.1	5.6	3.9	299.0	**0.047**	I
DNA replication	254.3	268.8	290.0	277.4	5.7	3.8	4.7	5.3	272.9	0.004	I
Non-homologous end-joining	114.7	119.3	133.6	129.2	4.5	3.0	4.4	2.6	124.4	0.001	D
Limonene and pinene degradation	208.9	227.7	234.3	221.6	7.4	7.6	5.1	4.3	223.4	0.005	I
Terpenoid backbone biosynthesis	196.3	204.1	214.1	205.6	5.5	6.6	4.7	2.2	205.2	0.023	I
Geraniol degradation	176.4	191.5	194.2	181.1	7.5	5.5	3.9	5.6	186.1	0.003	I
Polyketide sugar unit biosynthesis	78.7	81.2	89.0	88.6	1.3	1.3	1.5	3.3	84.4	0.000	D
Biosynthesis of ansamycins	55.8	56.5	58.4	58.3	1.6	1.0	1.7	1.5	57.3	**0.032**	D
Tetracycline biosynthesis	49.9	50.2	56.5	56.7	1.8	1.4	2.0	1.1	53.4	0.000	D
Biosynthesis of vancomycin group antibiotics	39.6	40.4	44.4	43.2	1.1	0.9	1.0	2.0	41.9	0.004	D
Nonribosomal peptide structures	22.8	23.3	28.1	30.2	1.5	1.6	0.8	1.8	26.2	0.000	D
Biosynthesis of siderophore group nonribosomal peptides	22.7	23.3	26.3	25.0	1.0	1.0	1.6	1.4	24.4	0.003	D
Carotenoid biosynthesis	19.3	20.2	22.2	20.8	0.8	1.8	1.4	0.6	20.7	**0.037**	D
Sesquiterpenoid and triterpenoid biosynthesis	11.0	11.2	12.5	13.5	0.4	0.4	0.7	0.4	12.1	0.000	D
Selenocompound metabolism	302.6	307.5	339.2	329.4	7.9	2.9	4.0	2.4	320.1	0.019	I
beta-Alanine metabolism	287.1	301.5	313.1	301.4	9.2	5.9	5.4	4.3	301.1	0.004	I
Glutathione metabolism	269.6	287.6	309.0	294.6	8.2	4.2	9.2	2.2	290.6	0.008	I
Taurine and hypotaurine metabolism	142.1	149.1	155.5	149.0	4.2	1.0	4.4	2.8	149.1	0.006	I
Cyanoamino acid metabolism	134.4	142.1	152.5	145.4	5.0	4.8	3.3	2.8	143.9	0.005	I
D-Glutamine and D-glutamate metabolism	59.9	63.7	70.9	68.2	1.9	2.0	1.7	1.0	65.8	**0.027**	I
D-Alanine metabolism	41.4	44.1	51.6	49.0	1.3	2.9	1.0	1.6	46.6	**0.027**	I

Phosphonate and phosphinate metabolism	28.8	30.5	33.7	32.3	1.4	1.5	0.9	1.4	31.4	0.003	D
Porphyrin and chlorophyll metabolism	268.8	277.6	306.9	296.0	7.7	6.7	7.0	2.5	287.8	**0.035**	I
One carbon pool by folate	217.0	223.5	238.7	229.6	7.5	2.8	4.2	5.6	227.6	0.018	I
Nicotinate and nicotinamide metabolism	178.2	183.9	205.5	191.6	6.5	4.5	4.0	4.4	190.1	0.001	I
Ubiquinone and other terpenoid-quinone biosynthesis	123.5	130.6	149.0	143.9	3.4	4.4	3.6	2.5	137.0	0.010	I
Folate biosynthesis	120.3	123.7	141.5	138.1	4.9	2.8	3.0	3.7	131.2	0.000	D
Thiamine metabolism	99.5	100.8	117.1	113.1	2.9	3.5	2.6	2.8	107.9	0.005	I
Vitamin B6 metabolism	80.4	82.5	88.8	84.3	3.1	3.5	0.8	2.3	84.2	**0.033**	I
Biotin metabolism	65.1	67.8	82.0	77.8	2.2	3.4	0.7	1.9	73.4	**0.025**	I
Retinol metabolism	67.9	72.4	75.8	72.7	3.6	0.7	3.4	1.8	72.4	0.004	I
Peptidoglycan biosynthesis	279.0	294.7	335.5	314.2	9.6	9.3	4.2	3.1	306.6	0.002	I
Lipopolysaccharide biosynthesis	104.2	110.8	137.6	132.3	6.3	10.0	4.8	4.4	121.7	0.000	D
Other glycan degradation	51.9	56.6	61.4	61.9	3.5	5.6	2.0	4.7	58.1	**0.025**	I
N-Glycan biosynthesis	34.6	36.7	46.4	47.4	1.5	2.4	1.0	1.7	41.4	0.000	D
Glycosaminoglycan degradation	25.1	28.3	33.2	33.4	1.1	0.9	0.2	1.4	30.0	0.017	I
Glycosphingolipid biosynthesis - globo series	16.9	19.4	19.7	20.6	0.9	1.5	0.3	1.0	19.2	0.015	I
RNA degradation	360.4	375.7	409.3	397.4	5.8	3.4	8.3	3.4	386.0	0.005	I
Protein export	199.5	212.5	239.3	232.6	8.8	2.4	7.6	3.0	221.3	0.000	D
Sulfur relay system	107.5	109.1	127.4	121.2	2.2	1.7	1.9	3.3	116.5	**0.036**	I
Protein processing in endoplasmic reticulum	35.6	38.1	45.7	45.0	1.1	0.8	3.7	2.4	41.2	0.001	D
Plant-pathogen I	80.9	84.9	88.4	85.1	0.7	2.8	2.8	1.5	84.9	**0.028**	I
Bacterial chemotaxis	175.3	192.8	217.2	209.9	8.1	11.0	6.9	11.5	199.3	0.019	I
Flagellar assembly	71.4	80.5	83.7	77.2	3.4	5.0	4.2	3.4	78.4	0.007	I
Cell cycle - Caulobacter	284.7	297.5	336.3	329.8	8.5	3.1	5.0	2.6	312.6	**0.030**	I
Meiosis - yeast	149.9	156.9	204.5	203.2	8.0	8.4	4.5	6.4	179.3	0.000	D

Apoptosis	10.9	11.0	13.6	12.8	0.4	0.9	0.5	0.6	12.1	0.000	D
Tropane. piperidine and pyridine alkaloid biosynthesis	175.7	185.5	207.6	195.7	4.7	2.9	7.0	2.2	191.4	0.002	I
Streptomycin biosynthesis	154.1	159.6	177.7	174.3	2.3	1.8	2.5	4.3	166.6	**0.028**	I
Novobiocin biosynthesis	144.7	151.4	172.9	162.7	4.8	1.8	5.9	2.1	158.2	0.003	I
Phenylpropanoid biosynthesis	75.2	81.7	80.2	75.4	1.8	3.1	3.1	3.1	78.2	0.002	I
Isoquinoline alkaloid biosynthesis	53.8	55.7	64.0	62.7	1.4	1.3	1.4	1.3	59.2	0.012	I
Penicillin and cephalosporin biosynthesis	44.7	48.8	62.4	59.5	2.5	3.0	0.3	2.5	54.0	0.007	I
Stilbenoid. diarylheptanoid and gingerol biosynthesis	46.8	51.1	54.3	51.5	1.8	0.7	1.8	0.3	51.0	0.000	I
beta-Lactam resistance	22.3	24.9	32.5	31.6	2.1	2.4	1.3	2.0	27.9	0.000	D
Butirosin and neomycin biosynthesis	14.3	14.7	17.5	16.3	0.7	0.8	0.7	0.6	15.7	0.000	D
Cellulase functions											
EC321.21	49.3	55.2	54.6	51.4	1.9	3.8	1.7	2.7	52.7	0.003	I
EC321.4	21.2	23.6	29.8	29.9	1.7	1.8	1.3	1.9	26.2	0.000	D
Cellulase function EC321.21											
Anaerolineaceae	0.2	0.2	0.3	0.3	0.0	0.1	0.1	0.1	4.1	0.000	D
Actinobacteria	11.0	12.6	10.4	8.9	1.4	1.1	0.9	0.4	10.7	**0.034**	I
Acidobacteria	1.6	1.4	2.4	2.5	0.4	0.4	0.4	0.4	2.0	0.000	D
Firmicutes	0.4	0.6	0.6	0.6	0.1	0.1	0.1	0.1	0.6	0.001	I
Gemmatimonadetes	0.5	0.5	0.7	0.8	0.2	0.1	0.1	0.2	0.6	0.002	D
Proteobacteria	7.6	8.3	8.7	7.1	0.3	0.7	0.9	0.7	7.9	0.003	I
Cellulase function EC321.4											
Cyanobacteria	0.4	0.4	0.9	0.8	0.1	0.2	0.1	0.1	0.6	0.000	D
Firmicutes	0.4	0.4	0.7	0.6	0.1	0.1	0.1	0.1	0.5	0.009	D
Gemmatimonadetes	0.3	0.2	0.5	0.5	0.2	0.1	0.1	0.1	0.4	0.020	D
Cellulase function EC321.91											
Planctomycetaceae	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.018	D

Cellulase domain families											
CBM2	18.7	19.7	26.3	26.8	1.1	1.6	0.7	2.1	23.0	0.000	D
CBM3	1.0	1.6	1.4	1.3	0.2	0.1	0.2	0.2	1.3	0.021	I
CBM4	1.7	2.3	2.3	1.9	0.2	0.2	0.2	0.3	2.0	0.001	I
CBM6	3.5	4.1	5.6	5.4	0.4	0.7	0.2	0.9	4.7	0.016	I
CBM30	1.4	1.5	2.7	2.4	0.1	0.2	0.4	0.4	2.0	0.000	D
CBM32	3.7	4.1	7.8	8.7	0.5	0.9	0.3	1.4	6.1	0.000	D
CBM44	4.8	5.3	7.4	6.9	0.3	0.6	0.3	0.6	6.1	0.000	D
CBM65	0.3	0.3	0.5	0.2	0.0	0.1	0.1	0.1	0.3	0.005	I
AA3	2.1	2.4	3.6	3.4	0.4	0.1	0.4	0.3	2.9	0.000	D
AA8	7.9	9.4	11.5	10.5	0.7	0.6	1.2	0.6	9.8	**0.031**	I
GH1	13.6	15.8	14.6	13.9	1.1	0.5	0.5	0.3	14.5	0.001	I
GH3	16.7	18.1	20.7	19.7	1.3	1.2	0.8	1.1	18.9	0.003	D
GH5	3.2	3.3	4.4	4.2	0.4	0.5	0.3	0.8	3.8	0.003	D
GH6	1.2	1.5	1.5	1.1	0.2	0.2	0.3	0.2	1.3	0.026	I
GH9	1.2	1.4	2.0	2.0	0.2	0.1	0.1	0.3	1.6	0.000	D
GH26	2.3	2.9	4.3	3.9	0.2	0.5	0.4	0.5	3.4	0.021	I
GH44	2.5	3.0	3.2	3.3	0.3	0.3	0.4	0.5	3.0	0.009	D
GH51	1.7	1.9	3.1	3.1	0.3	0.4	0.3	0.3	2.5	0.000	D
GH74	1.8	2.2	3.6	3.6	0.4	0.5	0.2	0.6	2.8	0.000	D
GH94	14.9	16.8	23.4	23.6	1.6	2.4	1.1	1.7	19.8	0.000	D
Cellulase domain family AA8											
Acidobacteria	0.1	0.2	0.3	0.4	0.0	0.1	0.1	0.0	0.3	0.003	D
Planctomycetes	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.030	I
Cellulase domain family GH1											
Actinobacteria	3.1	4.1	3.3	2.7	0.6	0.7	0.2	0.1	3.3	0.007	I
Verrucomicrobia	0.1	0.1	0.1	0.0	0.0	0.1	0.1	0.0	0.1	0.017	I
Parcubacteria	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.007	I
Cellulase domain family GH3											
Streptosporangiaceae	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.010	D
Bacteroidetes	0.7	1.2	0.9	1.1	0.2	0.3	0.2	0.2	1.0	0.030	T
Verrucomicrobia	0.1	0.1	0.2	0.2	0.0	0.1	0.1	0.1	0.1	0.023	D
Chloroflexi	0.1	0.2	0.2	0.3	0.0	0.1	0.1	0.1	0.2	0.027	D
Acidobacteria	0.5	0.6	1.3	1.2	0.2	0.2	0.3	0.4	0.9	0.000	D
Gemmatimonadetes	0.3	0.3	0.4	0.5	0.1	0.0	0.1	0.1	0.4	0.012	D
Cellulase domain family GH94											
Anaeromyxobact	0.3	0.4	0.8	0.7	0.1	0.1	0.1	0.2	0.6	0.001	D

eraceae											
Bacteroidetes	0.1	0.1	0.1	0.1	0.0	0.1	0.1	0.1	0.1	0.038	T
Proteobacteria	5.5	6.4	9.6	9.1	0.5	0.8	0.6	0.9	7.7	0.000	D
Cellulase domain family CBM2											
Verrucomicrobia subdivision 3	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.1	0.0	0.016	D
Nostocaceae	0.1	0.1	0.2	0.1	0.0	0.0	0.0	0.0	0.1	0.013	D
Planctomycetaceae	0.2	0.2	0.4	0.4	0.1	0.1	0.1	0.0	0.3	0.012	D
Desulfobacteraceae	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.012	D
Sinobacteraceae	0.1	0.1	0.1	0.2	0.0	0.0	0.1	0.1	0.1	0.007	D
Xanthomonadaceae	1.2	1.2	1.5	1.6	0.2	0.3	0.2	0.2	1.4	0.010	D
Bacteroidetes	0.1	0.1	0.3	0.2	0.0	0.1	0.0	0.1	0.2	0.011	D
Verrucomicrobia	0.0	0.0	0.2	0.2	0.0	0.0	0.1	0.1	0.1	0.008	D
Cyanobacteria	1.4	1.6	2.4	2.2	0.2	0.2	0.2	0.4	1.9	0.000	D
Planctomycetes	0.2	0.2	0.4	0.4	0.1	0.1	0.1	0.0	0.3	0.012	D
Proteobacteria	2.9	3.3	4.0	4.0	0.1	0.6	0.3	0.4	3.6	0.001	D
Euryarchaeota	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.020	I
Cellulase domain family CBM6											
Bacteroidetes	0.2	0.3	0.4	0.4	0.1	0.1	0.1	0.0	0.4	0.035	D
Acidobacteria	0.2	0.1	0.4	0.3	0.1	0.1	0.1	0.1	0.3	0.007	D

Table A9

Table A9: Listed are the microbial families belonging to each co-occurrence community in the metagenome of the organic tillage experiment: 156 in the blue community, 80 in the green community and 19 in the red community. Of the 265 families with abundance high enough to include in the analysis (see Materials and Methods), 10 families did not positively correlate with other families.

Blue community	Green community	Red community
Aquificaceae	Acidimicrobiaceae	Bacteroidaceae
Hydrogenothermaceae	Actinomycetaceae	Marinilabiliaceae
Desulfurobacteriaceae	Actinopolysporaceae	Prevotellaceae
Chthonomonadaceae	Catenulisporaceae	Cytophagaceae
Rhodothermaceae	Corynebacteriaceae	Cryomorphaceae
Porphyromonadaceae	Gordoniaceae	Flavobacteriaceae
Prolixibacteraceae	Mycobacteriaceae	Chitinophagaceae
Cyclobacteriaceae	Nocardiaceae	Saprospiraceae
Flammeovirgaceae	Acidothermaceae	Sphingobacteriaceae
Chlorobiaceae	Cryptosporangiaceae	Microchaetaceae
Ignavibacteriaceae	Frankiaceae	Lactobacillaceae
Melioribacteraceae	Geodermatophilaceae	Kiloniellaceae
Chlamydiaceae	Nakamurellaceae	Sandaracinaceae

Criblamydiaceae	Sporichthyaceae	Enterobacteriaceae
Parachlamydiaceae	Glycomycetaceae	Moraxellaceae
Lentisphaeraceae	Jiangellaceae	Herpotrichiellaceae
Puniceicoccaceae	Kineosporiaceae	Aspergillaceae
Methylacidiphilaceae	Beutenbergiaceae	Saprolegniaceae
Verrucomicrobia subdivision 3	Bogoriellaceae	Siphoviridae
Anaerolineaceae	Brevibacteriaceae	
Ardenticatenaceae	Cellulomonadaceae	
Caldilineaceae	Demequinaceae	
Chloroflexaceae	Dermabacteraceae	
Oscillochloridaceae	Dermacoccaceae	
Roseiflexaceae	Dermatophilaceae	
Herpetosiphonaceae	Intrasporangiaceae	
Dehalococcoidaceae	Microbacteriaceae	
Ktedonobacteraceae	Micrococcaceae	
Thermogemmatisporaceae	Promicromonosporaceae	
Sphaerobacteraceae	Micromonosporaceae	
Thermomicrobiaceae	Nocardioidaceae	
Rivulariaceae	Propionibacteriaceae	
Scytonemataceae	Pseudonocardiaceae	
Prochlorococcaceae	Streptomycetaceae	
Deferribacteraceae	Nocardiopsaceae	
Deinococcaceae	Streptosporangiaceae	
Trueperaceae	Thermomonosporaceae	
Thermaceae	Bifidobacteriaceae	
Dictyoglomaceae	Candidatus Actinomarinaceae	
Acidobacteriaceae	Coriobacteriaceae	
Holophagaceae	Nitriliruptoraceae	
Solibacteraceae	Rubrobacteraceae	
Alicyclobacillaceae	Conexibacteraceae	
Bacillaceae	Patulibacteraceae	
Paenibacillaceae	Solirubrobacteraceae	
Planococcaceae	Opitutaceae	
Thermoactinomycetaceae	Verrucomicrobiaceae	
Clostridiaceae	Nostocaceae	
Clostridiales Family XVII. Incertae Sedis	Fibrobacteraceae	
Eubacteriaceae	Staphylococcaceae	
Heliobacteriaceae	Symbiobacteriaceae	
Lachnospiraceae	Caulobacteraceae	
Peptococcaceae	Aurantimonadaceae	
Ruminococcaceae	Bradyrhizobiaceae	
Syntrophomonadaceae	Brucellaceae	
Halanaerobiaceae	Hyphomicrobiaceae	
Halobacteroidaceae	Methylobacteriaceae	
Thermoanaerobacteraceae	Phyllobacteriaceae	
Thermoanaerobacterales Family III. Incertae Sedis	Rhizobiaceae	
Veillonellaceae	Rhodobiaceae	
Fusobacteriaceae	Xanthobacteraceae	
Gemmatimonadaceae	Rhodobacteraceae	
Nitrospinaceae	Acetobacteraceae	
Nitrospiraceae	Rhodospirillaceae	
Phycisphaeraceae	Erythrobacteraceae	
Candidatus Brocadiaceae	Sphingomonadaceae	
Planctomycetaceae	Comamonadaceae	

Kordiimonadaceae	Oxalobacteraceae
Magnetococcaceae	Neisseriaceae
Parvularculaceae	Bdellovibrionaceae
Beijerinckiaceae	Labilitrichaceae
Methylocystaceae	Polyangiaceae
Hyphomonadaceae	Aeromonadaceae
Alcaligenaceae	Pseudomonadaceae
Burkholderiaceae	Xanthomonadaceae
Sutterellaceae	Oxytrichidae
Ferrovaceae	Pleosporaceae
Gallionellaceae	Nectriaceae
Hydrogenophilaceae	Chlorellaceae
Methylophilaceae	Amaranthaceae
Chromobacteriaceae	
Nitrosomonadaceae	
Rhodocyclaceae	
Sulfuricellaceae	
Bacteriovoracaceae	
Desulfarculaceae	
Desulfobacteraceae	
Desulfobulbaceae	
Desulfohalobiaceae	
Desulfomicrobiaceae	
Desulfonatronumaceae	
Desulfovibrionaceae	
Desulfurellaceae	
Desulfuromonadaceae	
Geobacteraceae	
Pelobacteraceae	
Anaeromyxobacteraceae	
Myxococcaceae	
Vulgatibacteraceae	
Kofleriaceae	
Nannocystaceae	
Syntrophaceae	
Syntrophobacteraceae	
Syntrophorhabdaceae	
Campylobacteraceae	
Helicobacteraceae	
Acidithiobacillaceae	
Thermithiobacillaceae	
Alteromonadaceae	
Colwelliaceae	
Ferrimonadaceae	
Idiomarinaceae	
Pseudoalteromonadaceae	
Psychromonadaceae	
Shewanellaceae	
Chromatiaceae	
Ectothiorhodospiraceae	
Coxiellaceae	
Legionellaceae	
Methylococcaceae	
Methylothermaceae	

Alcanivoracaceae
Hahellaceae
Halomonadaceae
Oceanospirillaceae
Pasteurellaceae
Salinisphaeraceae
Piscirickettsiaceae
Thiotrichaceae
Competibacteraceae
Vibrionaceae
Algiphilaceae
Sinobacteraceae
Mariprofundaceae
Leptospiraceae
Spirochaetaceae
Synergistaceae
Thermodesulfobacteriaceae
Thermotogaceae
Sulfolobaceae
Thermoproteaceae
Archaeoglobaceae
Halobacteriaceae
Methanobacteriaceae
Methanocellaceae
Methanomicrobiaceae
Methanoregulaceae
Candidatus Methanoperedenaceae
Methanosaetaceae
Methanosarcinaceae
Thermococcaceae
Nitrosopumilaceae
Nitrososphaeraceae
Glomeraceae
Euphorbiaceae
Podoviridae

Table A10

Table A10: Number of GH5-cellulase amplicon sequences separated into subclusters using subclustering-cut-off expect-value of 10^{-11} and three different inflation values. Shown are only the subclusters containing taxonomically assigned amplicon sequences. Taxonomic assignments are shown using stringent (LCA-100) and less stringent (LCA-50) lowest common ancestor (LCA (123))-assignment cut-offs. Separation of sequences based on taxonomic assignments was most successful using inflation value 2.

		Inflation value 1.5		**Inflation value 2**		**Inflation value 3**	
Subcluster number	**Taxonomic assignment**	LCA-100	LCA-50	LCA-100	LCA-50	LCA-100	LCA-50
1	Actinobacteria	26	37				
	Proteobacteria	80	422	80	422	80	422
	uncultured protist	1	1	1	1	1	1
	Firmicutes		5		4		4

	Nematoda		65		64		64
	unassigned	479	56	463	53	462	52
2	Actinobacteria			26	37		
	Firmicutes				1		
	Nematoda				1		
	unassigned	10	10	16	3		
3	Actinobacteria					26	37
	Firmicutes						1
	Nematoda						1
	unassigned	8	8	10	10	17	4
11	Actinobacteria	1	1				
	unassigned	3	3				
12	Actinobacteria			1	1		
	unassigned			3	3		
13	Actinobacteria					1	1
	unassigned					3	3
15	Proteobacteria		3				
16	Proteobacteria				3		
17	Proteobacteria						3
37	Proteobacteria	1	1				
38	Proteobacteria			1	1		
39	Proteobacteria					1	1
54	Actinobacteria	1	1				
55	Actinobacteria	1	1	1	1		
56	Actinobacteria			1	1	1	1
57	Actinobacteria					1	1
68	Firmicutes	1	1				
69	Firmicutes			1	1		
70	Firmicutes					1	1
78	Basidiomycota		1				
79	Basidiomycota				1		
80	Basidiomycota						1
86	Proteobacteria	1	1				
87	Proteobacteria			1	1		
88	Proteobacteria					1	1

Figure A1

Figure A1: Amplified products recovered from genomic DNA from *C. japonicas* in triplicate (lane R1-3), showing an expected band of products of ~105 bp length. Lane 4 shows the included control amplification reaction without added template DNA (*neg*) and lane 5 shows the 1-kbp DNA ladder functioning as marker.

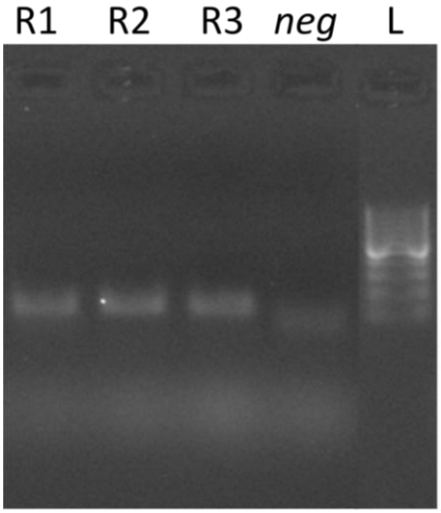


Figure A2

Figure A2 (next three pages): Maximum Likelihood-phylogenetic tree of cellulase-amplicon sequences and partial GH5-database amino acid sequences obtained from the CAZy database (60). The analysis was done using 669 amplicon sequences and 308 partial (amplicon region) sequences of experimentally validated GH5-cellulases (total 977 sequences) and included 64 alignment positions. The percentage of trees in which the associated sequences clustered together is shown next to the branches and was determined with 100 bootstrap replicates. Only bootstrap values higher than 25% are shown. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The taxonomic affiliation on phylum level of each sequence or collapsed cluster of sequences is indicated in the sequence-name and by a coloured marker; a filled marker for database sequences and an open marker for amplicon sequences. Taxonomic affiliations indicated with "(2)" in the sequence-name have been assigned using less stringent taxonomic annotation cut-offs (LCA=50%). Collapsed branches consist of sequences from the same taxonomic affiliation as indicated by the name of the collapsed cluster. The locations of the database sequences used for primer design are indicated by arrows. Database sequences belonging to clade 1, which are classified as GH5-subfamily 2, and clade 1.1 are outlined by a light- respectively dark-grey box. The location of amplicon sequences in the tree is indicated by the bar on the left side of the tree, with the colour corresponding to the subcluster number. An overview of the complete tree of database and amplicon sequences is given in the **Overview**-figure, whereas **part A** and **part B** show a larger version of the upper respectively lower half of the tree.

Figure A2

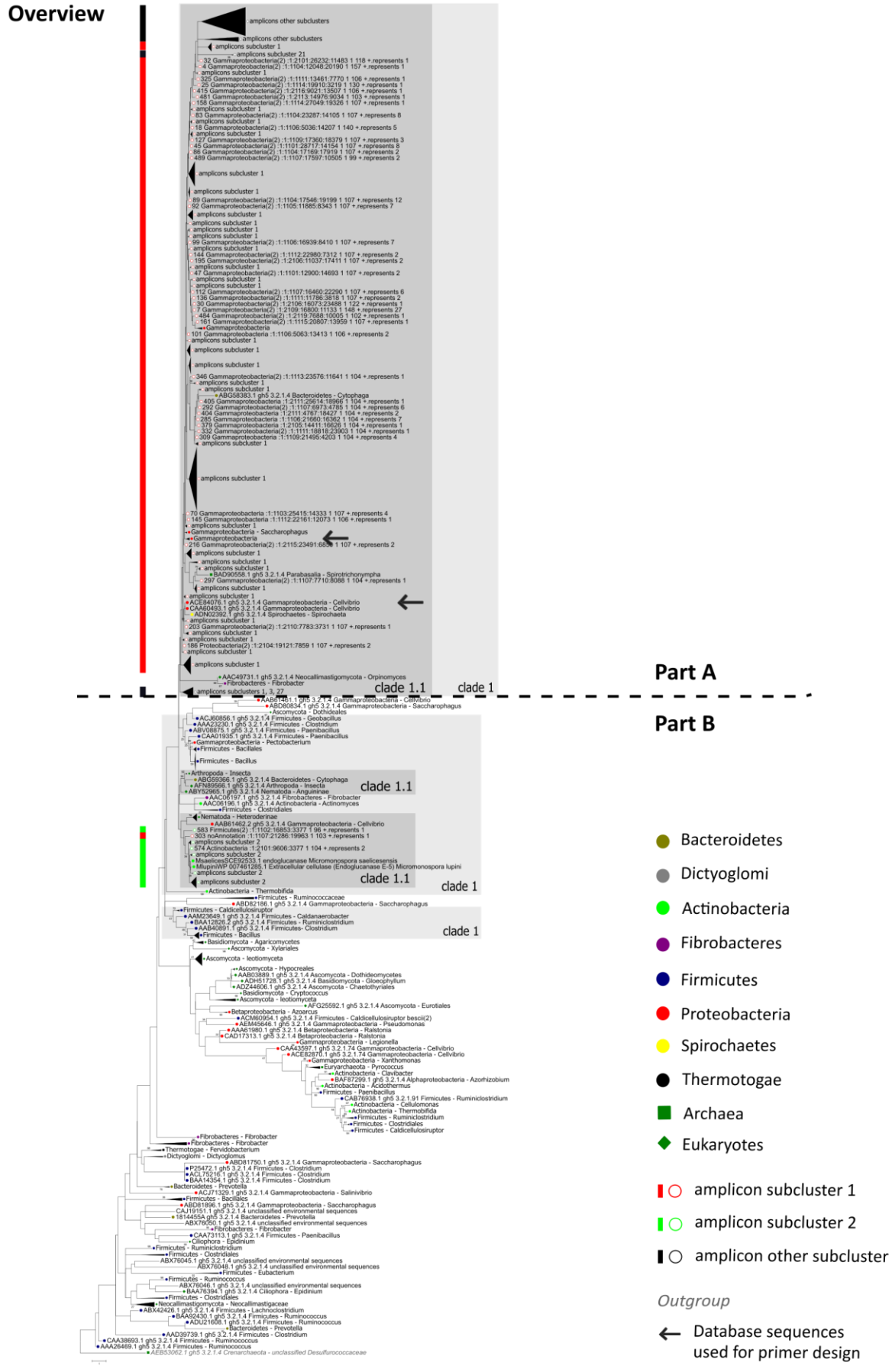

Overview
Part A
Part B
clade 1
clade 1.1
Bacteroidetes
Dictyoglomi
Actinobacteria
Fibrobacteres
Firmicutes
Proteobacteria
Spirochaetes
Thermotogae
Archaea
Eukaryotes
amplicon subcluster 1
amplicon subcluster 2
amplicon other subcluster
Outgroup
Database sequences used for primer design

Figure A2

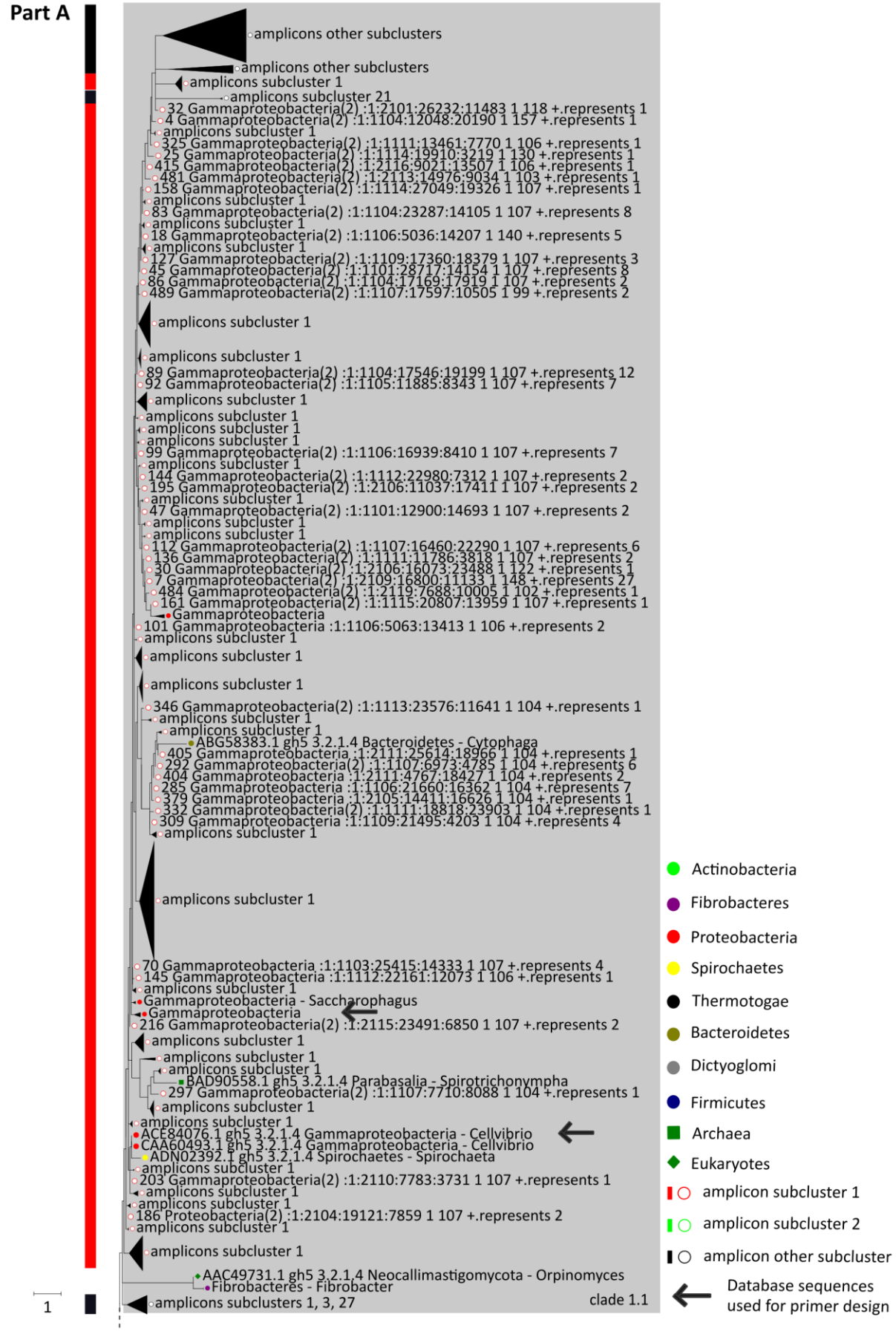

Part A
amplicons other subclusters
amplicons other subclusters
amplicons subcluster 1
amplicons subcluster 21
32 Gammaproteobacteria(2) :1:2101:26232:11483 1 118 +.represents 1
4 Gammaproteobacteria(2) :1:1104:12048:20190 1 157 +.represents 1
amplicons subcluster 1
325 Gammaproteobacteria(2) :1:1111:13461:7770 1 106 +.represents 1
25 Gammaproteobacteria(2) :1:1114:19910:3219 1 130 +.represents 1
415 Gammaproteobacteria(2) :1:2116:9021:13507 1 106 +.represents 1
481 Gammaproteobacteria(2) :1:2113:14976:9034 1 103 +.represents 1
158 Gammaproteobacteria(2) :1:1114:27049:19326 1 107 +.represents 1
amplicons subcluster 1
83 Gammaproteobacteria(2) :1:1104:23287:14105 1 107 +.represents 8
amplicons subcluster 1
18 Gammaproteobacteria(2) :1:1106:5036:14207 1 140 +.represents 5
amplicons subcluster 1
127 Gammaproteobacteria(2) :1:1109:17360:18379 1 107 +.represents 3
45 Gammaproteobacteria(2) :1:1101:28717:14154 1 107 +.represents 8
86 Gammaproteobacteria(2) :1:1104:17169:17919 1 107 +.represents 2
489 Gammaproteobacteria(2) :1:1107:17597:10505 1 99 +.represents 2
amplicons subcluster 1
amplicons subcluster 1
89 Gammaproteobacteria(2) :1:1104:17546:19199 1 107 +.represents 12
92 Gammaproteobacteria(2) :1:1105:11885:8343 1 107 +.represents 7
amplicons subcluster 1
amplicons subcluster 1
amplicons subcluster 1
amplicons subcluster 1
99 Gammaproteobacteria(2) :1:1106:16939:8410 1 107 +.represents 7
amplicons subcluster 1
144 Gammaproteobacteria(2) :1:1112:22980:7312 1 107 +.represents 2
195 Gammaproteobacteria(2) :1:2106:11037:17411 1 107 +.represents 2
amplicons subcluster 1
47 Gammaproteobacteria(2) :1:1101:12900:14693 1 107 +.represents 2
amplicons subcluster 1
amplicons subcluster 1
112 Gammaproteobacteria(2) :1:1107:16460:22290 1 107 +.represents 6
136 Gammaproteobacteria(2) :1:1111:11786:3818 1 107 +.represents 2
30 Gammaproteobacteria(2) :1:2106:16073:23488 1 122 +.represents 1
7 Gammaproteobacteria(2) :1:2109:16800:11133 1 148 +.represents 27
484 Gammaproteobacteria(2) :1:2119:7688:10005 1 102 +.represents 1
161 Gammaproteobacteria(2) :1:1115:20807:13959 1 107 +.represents 1
Gammaproteobacteria
101 Gammaproteobacteria :1:1106:5063:13413 1 106 +.represents 2
amplicons subcluster 1
amplicons subcluster 1
amplicons subcluster 1
346 Gammaproteobacteria(2) :1:1113:23576:11641 1 104 +.represents 1
amplicons subcluster 1
amplicons subcluster 1
ABG58383.1 gh5 3.2.1.4 Bacteroidetes - Cytophaga
405 Gammaproteobacteria :1:2111:25614:18966 1 104 +.represents 1
292 Gammaproteobacteria(2) :1:1107:6973:4785 1 104 +.represents 6
404 Gammaproteobacteria :1:2111:4767:18427 1 104 +.represents 2
285 Gammaproteobacteria :1:1106:21660:16362 1 104 +.represents 7
379 Gammaproteobacteria :1:2105:14411:16626 1 104 +.represents 1
332 Gammaproteobacteria(2) :1:1111:18818:23903 1 104 +.represents 1
309 Gammaproteobacteria :1:1109:21495:4203 1 104 +.represents 4
amplicons subcluster 1
amplicons subcluster 1
70 Gammaproteobacteria :1:1103:25415:14333 1 107 +.represents 4
145 Gammaproteobacteria :1:1112:22161:12073 1 106 +.represents 1
amplicons subcluster 1
Gammaproteobacteria - Saccharophagus
Gammaproteobacteria
216 Gammaproteobacteria(2) :1:2115:23491:6850 1 107 +.represents 2
amplicons subcluster 1
amplicons subcluster 1
amplicons subcluster 1
BAD90558.1 gh5 3.2.1.4 Parabasalia - Spirotrichonympha
297 Gammaproteobacteria(2) :1:1107:7710:8088 1 104 +.represents 1
amplicons subcluster 1
amplicons subcluster 1
ACE84076.1 gh5 3.2.1.4 Gammaproteobacteria - Cellvibrio
CAA60493.1 gh5 3.2.1.4 Gammaproteobacteria - Cellvibrio
ADN02392.1 gh5 3.2.1.4 Spirochaetes - Spirochaeta
amplicons subcluster 1
203 Gammaproteobacteria(2) :1:2110:7783:3731 1 107 +.represents 1
amplicons subcluster 1
amplicons subcluster 1
186 Proteobacteria(2) :1:2104:19121:7859 1 107 +.represents 2
amplicons subcluster 1
amplicons subcluster 1
AAC49731.1 gh5 3.2.1.4 Neocallimastigomycota - Orpinomyces
Fibrobacteres - Fibrobacter
amplicons subclusters 1, 3, 27
clade 1.1
1
Actinobacteria
Fibrobacteres
Proteobacteria
Spirochaetes
Thermotogae
Bacteroidetes
Dictyoglomi
Firmicutes
Archaea
Eukaryotes
amplicon subcluster 1
amplicon subcluster 2
amplicon other subcluster
Database sequences used for primer design

Figure A2

Part B

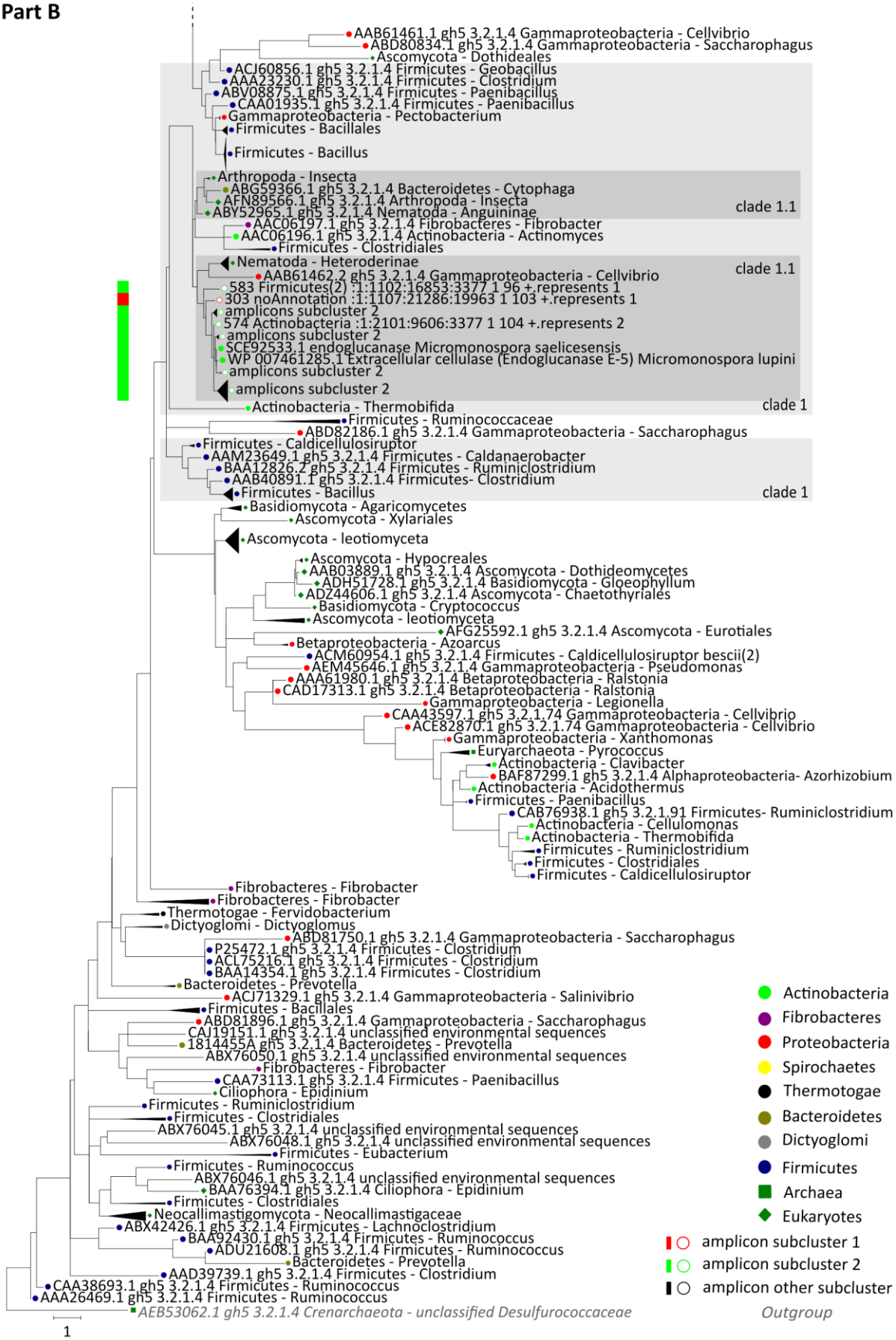
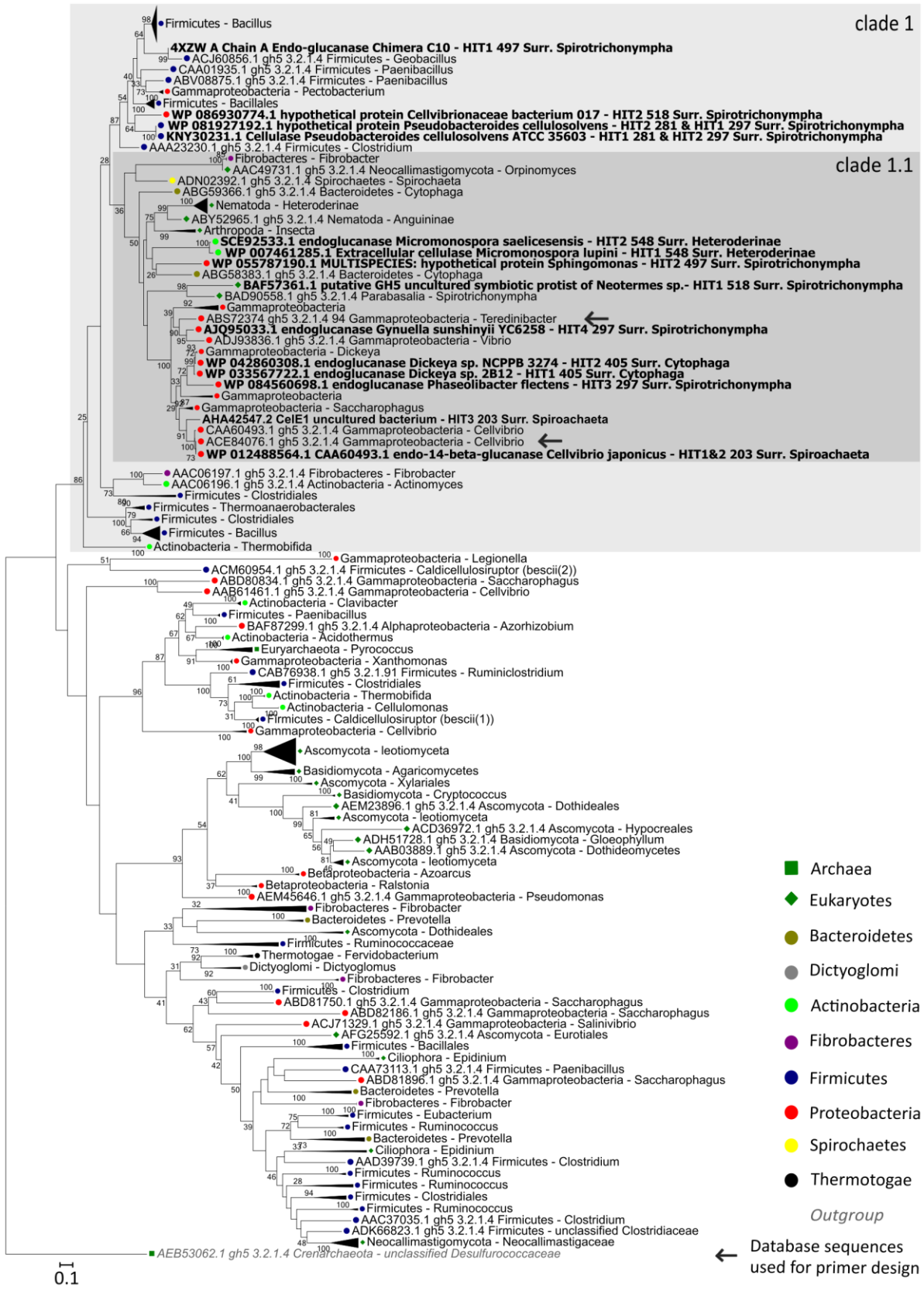


Figure A3

Figure A3 (next page): Maximum Likelihood-phylogenetic tree of 322 complete domain sequences of experimentally validated GH5-cellulases obtained from the CAZy database (60) and the protein sequences of top BLAST hits of selected amplicon sequences (indicated in bold). The analysis included 425 alignment positions. The percentage of trees in which the associated sequences clustered together is shown next to the branches and was determined with 100 bootstrap replicates. Only bootstrap values higher than 25% are shown. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The taxonomic affiliation on phylum level of each sequence or collapsed cluster of sequences is indicated by a colour- filled marker. Collapsed branches consist of sequences from the same taxonomic affiliation as indicated by the name of the collapsed cluster. Sequences belonging to clade 1.1 are outlined by a dark-grey box. The location of the database sequences used for primer design is indicated by an arrow.

Figure A3


clade 1
clade 1.1
Firmicutes - Bacillus
4XZW A Chain A Endo-glucanase Chimera C10 - HIT1 497 Surr. Spirotrichonympha
ACJ60856.1 gh5 3.2.1.4 Firmicutes - Geobacillus
CAA01935.1 gh5 3.2.1.4 Firmicutes - Paenibacillus
ABV08875.1 gh5 3.2.1.4 Firmicutes - Paenibacillus
Gammaproteobacteria - Pectobacterium
Firmicutes - Bacillales
WP 086930774.1 hypothetical protein Cellvibrionaceae bacterium 017 - HIT2 518 Surr. Spirotrichonympha
WP 081927192.1 hypothetical protein Pseudobacteroides cellulosolvens - HIT2 281 & HIT1 297 Surr. Spirotrichonympha
KNY30231.1 Cellulase Pseudobacteroides cellulosolvens ATCC 35603 - HIT1 281 & HIT2 297 Surr. Spirotrichonympha
AAA23230.1 gh5 3.2.1.4 Firmicutes - Clostridium
Fibrobacteres - Fibrobacter
AAC49731.1 gh5 3.2.1.4 Neocallimastigomycota - Orpinomyces
ADN02392.1 gh5 3.2.1.4 Spirochaetes - Spirochaeta
ABG59366.1 gh5 3.2.1.4 Bacteroidetes - Cytophaga
Nematoda - Heteroderinae
ABY52965.1 gh5 3.2.1.4 Nematoda - Anguininae
Arthropoda - Insecta
SCE92533.1 endoglucanase Micromonospora saelicesensis - HIT2 548 Surr. Heteroderinae
WP 007461285.1 Extracellular cellulase Micromonospora lupini - HIT1 548 Surr. Heteroderinae
WP 055787190.1 MULTISPECIES: hypothetical protein Sphingomonas - HIT2 497 Surr. Spirotrichonympha
ABG58383.1 gh5 3.2.1.4 Bacteroidetes - Cytophaga
BAF57361.1 putative GH5 uncultured symbiotic protist of Neotermes sp.- HIT1 518 Surr. Spirotrichonympha
BAD90558.1 gh5 3.2.1.4 Parabasalia - Spirotrichonympha
Gammaproteobacteria
ABS72374 gh5 3.2.1.4 94 Gammaproteobacteria - Teredinibacter
AJQ95033.1 endoglucanase Gynuella sunshinyii YC6258 - HIT4 297 Surr. Spirotrichonympha
ADJ93836.1 gh5 3.2.1.4 Gammaproteobacteria - Vibrio
Gammaproteobacteria - Dickeya
WP 042860308.1 endoglucanase Dickeya sp. NCPPB 3274 - HIT2 405 Surr. Cytophaga
WP 033567722.1 endoglucanase Dickeya sp. 2B12 - HIT1 405 Surr. Cytophaga
WP 084560698.1 endoglucanase Phaseolibacter flectens - HIT3 297 Surr. Spirotrichonympha
Gammaproteobacteria
Gammaproteobacteria - Saccharophagus
AHA42547.2 CelE1 uncultured bacterium - HIT3 203 Surr. Spiroachaeta
CAA60493.1 gh5 3.2.1.4 Gammaproteobacteria - Cellvibrio
ACE84076.1 gh5 3.2.1.4 Gammaproteobacteria - Cellvibrio
WP 012488564.1 CAA60493.1 endo-14-beta-glucanase Cellvibrio japonicus - HIT1&2 203 Surr. Spiroachaeta
AAC06197.1 gh5 3.2.1.4 Fibrobacteres - Fibrobacter
AAC06196.1 gh5 3.2.1.4 Actinobacteria - Actinomyces
Firmicutes - Clostridiales
Firmicutes - Thermoanaerobacterales
Firmicutes - Clostridiales
Firmicutes - Bacillus
Actinobacteria - Thermobifida
Gammaproteobacteria - Legionella
ACM60954.1 gh5 3.2.1.4 Firmicutes - Caldicellulosiruptor (bescii(2))
ABD80834.1 gh5 3.2.1.4 Gammaproteobacteria - Saccharophagus
AAB61461.1 gh5 3.2.1.4 Gammaproteobacteria - Cellvibrio
Actinobacteria - Clavibacter
Firmicutes - Paenibacillus
BAF87299.1 gh5 3.2.1.4 Alphaproteobacteria - Azorhizobium
Actinobacteria - Acidothermus
Euryarchaeota - Pyrococcus
Gammaproteobacteria - Xanthomonas
CAB76938.1 gh5 3.2.1.91 Firmicutes - Ruminiclostridium
Firmicutes - Clostridiales
Actinobacteria - Thermobifida
Actinobacteria - Cellulomonas
Firmicutes - Caldicellulosiruptor (bescii(1))
Gammaproteobacteria - Cellvibrio
Ascomycota - leotiomyceta
Basidiomycota - Agaricomycetes
Ascomycota - Xylariales
Basidiomycota - Cryptococcus
AEM23896.1 gh5 3.2.1.4 Ascomycota - Dothideales
Ascomycota - leotiomyceta
ACD36972.1 gh5 3.2.1.4 Ascomycota - Hypocreales
ADH51728.1 gh5 3.2.1.4 Basidiomycota - Gloeophyllum
AAB03889.1 gh5 3.2.1.4 Ascomycota - Dothideomycetes
Ascomycota - leotiomyceta
Betaproteobacteria - Azoarcus
Betaproteobacteria - Ralstonia
AEM45646.1 gh5 3.2.1.4 Gammaproteobacteria - Pseudomonas
Fibrobacteres - Fibrobacter
Bacteroidetes - Prevotella
Ascomycota - Dothideales
Firmicutes - Ruminococcaceae
Thermotogae - Fervidobacterium
Dictyoglomi - Dictyoglomus
Fibrobacteres - Fibrobacter
Firmicutes - Clostridium
ABD81750.1 gh5 3.2.1.4 Gammaproteobacteria - Saccharophagus
ABD82186.1 gh5 3.2.1.4 Gammaproteobacteria - Saccharophagus
ACJ71329.1 gh5 3.2.1.4 Gammaproteobacteria - Salinivibrio
AFG25592.1 gh5 3.2.1.4 Ascomycota - Eurotiales
Firmicutes - Bacillales
Ciliophora - Epidinium
CAA73113.1 gh5 3.2.1.4 Firmicutes - Paenibacillus
ABD81896.1 gh5 3.2.1.4 Gammaproteobacteria - Saccharophagus
Bacteroidetes - Prevotella
Fibrobacteres - Fibrobacter
Firmicutes - Eubacterium
Firmicutes - Ruminococcus
Bacteroidetes - Prevotella
Ciliophora - Epidinium
AAD39739.1 gh5 3.2.1.4 Firmicutes - Clostridium
Firmicutes - Ruminococcus
Firmicutes - Ruminococcus
Firmicutes - Clostridiales
Firmicutes - Ruminococcus
AAC37035.1 gh5 3.2.1.4 Firmicutes - Clostridium
ADK66823.1 gh5 3.2.1.4 Firmicutes - unclassified Clostridiaceae
Neocallimastigomycota - Neocallimastigaceae
AEB53062.1 gh5 3.2.1.4 Crenarchaeota - unclassified Desulfurococcaceae
0.1
Archaea
Eukaryotes
Bacteroidetes
Dictyoglomi
Actinobacteria
Fibrobacteres
Firmicutes
Proteobacteria
Spirochaetes
Thermotogae
Outgroup
Database sequences used for primer design

List of publications

First- or co-author publications that are included in the thesis or that are related to the topic:

- Stephanie Schulz, Fabian Bergekemper, Maria de Vries, Anne Schöler, Michael Schloter. **qPCR zur quantitativen Validierung von Metagenomdaten**. *BIOspektrum* 3-2016, 265-269;

- Anne Schoeler, Maria de Vries, Gisle Vestergaard, Michael Schloter. „**Chapter 12. Reconstruction of Transformation Processes Catalyzed by the Soil Microbiome Using Metagenomic Approaches**“ *Microbial Environmental Genomics*. Francis Martin & Stéphane Uroz. New York: Springer Science+Business Media, 197-206;

- Maria de Vries, Anne Schoeler, Julia Ertl, Zhuofei Xu, Michael Schloter. **Metagenomic analyses reveal no differences in genes involved in cellulose degradation under different tillage treatments.** *FEMS Microbiology Ecology* 91 (7): fiv069.

References

1. Schlesinger WH, Andrews JA. 2000. Soil respiration and the global carbon cycle. Biogeochemistry 48:7–20.
2. Lal R. 2004. Soil carbon sequestration to mitigate climate change. Geoderma 123:1–22.
3. Batjes NH. 2014. Total carbon and nitrogen in the soils of the world. Eur J Soil Sci 65:10–21.
4. Ruddiman WF. 2003. The Anthropogenic Greenhous Era Began Thousands of Years Ago. Clim Change 61:261–293.
5. Lal R. 2004. Soil carbon sequestration impacts on global climate change and food security. Science (80-) 304:1623–1627 ST–Soil carbon sequestration impacts.
6. Guo LB, Gifford RM. 2002. Soil carbon stocks and land use change: A meta analysis. Glob Chang Biol 8:345–360.
7. Obalum SE, Chibuike GU, Peth S, Ouyang Y. 2017. Soil organic matter as sole indicator of soil degradation. Environ Monit Assess 189.
8. Sanaullah M, Chabbi A, Maron PA, Baumann K, Tardy V, Blagodatskaya E, Kuzyakov Y, Rumpel C. 2016. How do microbial communities in top- and subsoil respond to root litter addition under field conditions? Soil Biol Biochem 103:28–38.
9. Elmajdoub B, Marschner P. 2013. Salinity reduces the ability of soil microbes to utilise cellulose. Biol Fertil Soils 49:379–386.
10. Simpson AJ, Simpson MJ, Smith E, Kelleher BP. 2007. Microbially derived inputs to soil organic matter: Are current estimates too low? Environ Sci Technol 41:8070–8076.
11. Donnelly PK, Entry JA, Crawford DL, Cromack K. 1990. Cellulose and lignin degradation in forest soils: Response to moisture, temperature, and acidity. Microb Ecol 20:289–295.
12. Berlemont R, Allison SD, Weihe C, Lu Y, Brodie EL, Martiny JBH, Martiny AC. 2014. Cellulolytic potential under environmental changes in microbial communities from grassland litter. Isme J 5:1–10.
13. Kainiemi V, Arvidsson J, Kätterer T. 2013. Short-term organic matter mineralisation following different types of tillage on a Swedish clay soil. Biol Fertil Soils.
14. Tebrügge F, Düring R-A. 1999. Reducing tillage intensity - A review of results from a long-term study in Germany. Soil Tillage Res 53:15–28.
15. Abdollahi L, Munkholm LJ, Garbout a. 2014. Tillage System and Cover Crop Effects on Soil Quality: II. Pore Characteristics. Soil Sci Soc Am J 78:271–279.
16. Mathew RP, Feng Y, Githinji L, Ankumah R, Balkcom KS. 2012. Impact of No-Tillage and Conventional Tillage Systems on Soil Microbial Communities. Appl Environ Soil Sci 2012:1–10.
17. Franzluebbers A. 2008. Linking soil and water quality in conservation agricultural systems. Electron J Integr Biosci 6:15–29.
18. van Groenigen KJ, Hastings A, Forristal D, Roth B, Jones M, Smith P. 2011. Soil C storage as affected by tillage and straw management: An assessment using field measurements and model predictions. Agric Ecosyst Environ 140:218–225.
19. Virto I, Barré P, Burlot A, Chenu C. 2012. Carbon input differences as the main factor explaining the variability in soil organic C storage in no-tilled compared to inversion tilled agrosystems. Biogeochemistry 108:17–26.
20. Alvarez R, Diaz RA, Barberob N, Santanatogliaa OJ, Blotta L. 1995. Soil organic carbon, microbial biomass and C02-C production from three tillage systems. Soil Tillage Res 33:17–28.
21. Ogle SM, Breidt FJ, Paustian K. 2005. Agricultural management impacts on soil organic carbon storage under moist and dry climatic conditions of temperate and tropical regions. Biogeochemistry 72:87–121.
22. Van den Putte A, Govers G, Diels J, Gillijns K, Demuzere M. 2010. Assessing the effect of soil tillage on crop growth: A meta-regression analysis on European crop yields under conservation agriculture. Eur J Agron 33:231–241.
23. Anken T, Weisskopf P, Zihlmann U, Forrer H, Jansa J, Perhacova K. 2004. Long-term tillage

system effects under moist cool conditions in Switzerland. Soil Tillage Res 78:171–183.
24. Küstermann B, Munch JC, Hülsbergen K-J. 2013. Effects of soil tillage and fertilization on resource efficiency and greenhouse gas emissions in a long-term field experiment in Southern Germany. Eur J Agron 49:61–73.
25. Vogeler I, Rogasik J, Funder U, Panten K, Schnug E. 2009. Effect of tillage systems and P-fertilization on soil physical and chemical properties, crop yield and nutrient uptake. Soil Tillage Res 103:137–143.
26. Brookes PC, Chen YF, Chen L, Qiu G, Luo Y, Xu J. 2017. Is the rate of mineralization of soil organic carbon under microbiological control? Soil Biol Biochem 112:127–139.
27. Stockmann U, Adams MA, Crawford JW, Field DJ, Henakaarchchi N, Jenkins M, Minasny B, McBratney AB, Courcelles V de R de, Singh K, Wheeler I, Abbott L, Angers DA, Baldock J, Bird M, Brookes PC, Chenu C, Jastrow JD, Lal R, Lehmann J, O'Donnell AG, Parton WJ, Whitehead D, Zimmermann M. 2013. The knowns, known unknowns and unknowns of sequestration of soil organic carbon. Agric Ecosyst Environ 164:80–99.
28. Kögel-Knabner I. 2017. The macromolecular organic composition of plant and microbial residues as inputs to soil organic matter: Fourteen years on. Soil Biol Biochem.
29. Cotrufo MF, Wallenstein MD, Boot CM, Denef K, Paul E. 2013. The Microbial Efficiency-Matrix Stabilization (MEMS) framework integrates plant litter decomposition with soil organic matter stabilization: Do labile plant inputs form stable soil organic matter? Glob Chang Biol 19:988–995.
30. Kosheleva YP, Trofimov SY. 2008. Characteristics of the biochemical composition of plant litter at different stages of decomposition (according to thermal analysis data). Biol Bull 35:64–69.
31. Bayer EA, Chanzy H, Lamed R, Shoham Y. 1998. Cellulose, cellulases and cellulosomes. Curr Opin Struct Biol.
32. Sharma A, Tewari R, Rana SS, Soni R, Soni SK. 2016. Cellulases: Classification, Methods of Determination and Industrial Applications. Appl Biochem Biotechnol.
33. Alexander JK. 1968. Purification and specificity of cellobiose phosphorylase from Clostridium thermocellum. J Biol Chem 243:2899–2904.
34. Duan C-J, Feng J-X. 2010. Mining metagenomes for novel cellulase genes. Biotechnol Lett 32:1765–75.
35. Nihira T, Saito Y, Nishimoto M, Kitaoka M, Igarashi K, Ohtsubo K, Nakai H. 2013. Discovery of cellobionic acid phosphorylase in cellulolytic bacteria and fungi. FEBS Lett 587:3556–3561.
36. Langston JA, Shaghasi T, Abbate E, Xu F, Vlasenko E, Sweeney MD. 2011. Oxidoreductive Cellulose Depolymerization by the Enzymes Cellobiose Dehydrogenase and Glycoside Hydrolase 61. Appl Environ Microbiol.
37. Lynd LR, Weimer PJ, van Zyl WH, Pretorius IS. 2002. Microbial cellulose utilization: fundamentals and biotechnology. Microbiol Mol Biol Rev 66:506–577, table of contents.
38. Schwarz WH. 2001. The cellulosome and cellulose degradation by anaerobic bacteria. Appl Microbiol Biotechnol 56:634–649.
39. Adney WS, Rivard CJ, Shiang M, Himmel ME. 1991. Anaerobic digestion of lignocellulosic biomass and wastes - Cellulases and related enzymes. Appl Biochem Biotechnol 30:165–183.
40. Lynd LR, Weimer PJ, Zyl WH Van, Isak S. 2002. Microbial Cellulose Utilization : Fundamentals and Biotechnology Microbial Cellulose Utilization : Fundamentals and Biotechnology Downloaded from http://mmbr.asm.org/ on February 6 , 2013 by INDIAN INSTITUTE OF TECHNOLOGY MADRAS 66.
41. Reichenbach H, Lang E, Schumann P, Spröer C. 2006. Byssovorax cruenta gen. nov., sp. nov., nom. rev., a cellulose-degrading myxobacterium: rediscovery of "Myxococcus cruentus" Thaxter 1897. Int J Syst Evol Microbiol 56:2357–63.
42. Baldrian P, Valášková V. 2008. Degradation of cellulose by basidiomycetous fungi. FEMS Microbiol Rev.

43. Fontes CMG a, Gilbert HJ. 2010. Cellulosomes: highly efficient nanomachines designed to deconstruct plant cell wall complex carbohydrates. Annu Rev Biochem 79:655–681.
44. J Baum T, S Hussey R, L Davis E. 2007. Root-Knot and Cyst Nematode Parasitism Genes: The Molecular Basis of Plant ParasitismGenetic engineering.
45. Yan Y, Smant G, Stokkermans J, Qin L, Helder J, Baum T, Schots a, Davis E. 1998. Genomic organization of four beta-1,4-endoglucanase genes in plant-parasitic cyst nematodes and its evolutionary implications. Gene 220:61–70.
46. Watanabe H, Noda H, Tokuda G, Lo N. 1998. A cellulase gene of termite origin. Nature 394:330–331.
47. Lo N, Tokuda G, Watanabe H. 2011. Evolution and function of endogenous termite cellulases, p. 51–67. *In* Biology of Termites: A Modern Synthesis.
48. Sugimura M, Watanabe H, Lo N, Saito H. 2003. Purification, characterization, cDNA cloning and nucleotide sequencing of a cellulase from the yellow-spotted longicorn beetle, Psacothea hilaris. Eur J Biochem 270:3455–3460.
49. Inoue T, Moriya S, Ohkuma M, Kudo T. 2005. Molecular cloning and characterization of a cellulase gene from a symbiotic protist of the lower termite, Coptotermes formosanus. Gene 349:67–75.
50. Todaka N, Lopez CM, Inoue T, Saita K, Maruyama JI, Arioka M, Kitamoto K, Kudo T, Moriya S. 2010. Heterologous expression and characterization of an endoglucanase from a symbiotic protist of the lower termite, reticulitermes speratus. Appl Biochem Biotechnol 160:1168–1178.
51. Talamantes D, Biabini N, Dang H, Abdoun K, Berlemont R. 2016. Natural diversity of cellulases, xylanases, and chitinases in bacteria. Biotechnol Biofuels 9.
52. Berlemont R, Martiny AC. 2016. Glycoside Hydrolases across Environmental Microbial Communities. PLoS Comput Biol 12.
53. Louime C, Abazinge M, Johnson E. 2006. Location, formation and biosynthetic regulation of cellulases in the gliding bacteria Cytophaga hutchinsonii. Int J Mol Sci 7:1–11.
54. Wood WE, Neubauer DG, Stutzenberger FJ. 1984. Cyclic AMP levels during induction and repression of cellulase biosynthesis in Thermomonospora curvata. J Bacteriol 160:1047–1054.
55. Ilmén M, Saloheimo ANU, Onnela ML, Penttilä ME, Ilme M. 1997. Regulation of cellulase gene expression in the filamentous fungus Trichoderma reesei. Appl Environ Microbiol 63:1298–1306.
56. Lee CC, Wong DWS, Robertson GH. 2001. Cloning and characterization of two cellulase genes from Lentinula edodes. FEMS Microbiol Lett 205:355–360.
57. Jia J, Dyer PS, Buswell JA, Peberdy JF. 1999. Cloning of the cbhI and cbhII genes involved in cellulose utilisation by the straw mushroom Volvariella volvacea. Mol Gen Genet 261:985–993.
58. Chow CM, Yague E, Raguz S, Wood DA, Thurston CF. 1994. The cel3 gene of Agaricus bisporus codes for a modular cellulase and is transcriptionally regulated by the carbon source. Appl Environ Microbiol 60:2779–2785.
59. Henrissat B. 1991. A classification of glycosyl hydrolases based on amino acid sequence similarities. Biochem J 280:309–316.
60. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. 2014. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res 42:490–495.
61. Levasseur A, Drula E, Lombard V, Coutinho PM, Henrissat B. 2013. Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes. Biotechnol Biofuels 6:41.
62. Davies G, Henrissat B. 1995. Structures and mechanisms of glycosyl hydrolases. Structure 3:853–859.
63. Henrissat B, Callebaut I, Fabrega S, Lehn P, Mornon JP, Davies G. 1995. Conserved catalytic machinery and the prediction of a common fold for several families of glycosyl hydrolases.

Proc Natl Acad Sci 92:7090–7094.
64. Ichikawa S, Yoshida M, Karita S, Kondo M, Goto M. 2015. Carbohydrate-binding modules influence substrate specificity of an endoglucanase from *Clostridium thermocellum*. Biosci Biotechnol Biochem 1–5.
65. Boraston AB, Bolam DN, Gilbert HJ, Davies GJ. 2004. Carbohydrate-binding modules: fine-tuning polysaccharide recognition. Biochem J 382:769–781.
66. Carrard G, Koivula a, Söderlund H, Béguin P. 2000. Cellulose-binding domains promote hydrolysis of different sites on crystalline cellulose. Proc Natl Acad Sci U S A 97:10342–10347.
67. Burstein T, Shulman M, Jindou S, Petkun S, Frolow F, Shoham Y, Bayer EA, Lamed R. 2009. Physical association of the catalytic and helper modules of a family-9 glycoside hydrolase is essential for activity. FEBS Lett 583:879–884.
68. Henriksson G, Nutt A, Henriksson H, Pettersson B, Ståhlberg J, Johansson G, Pettersson G. 1999. Endoglucanase 28 (Cel12A), a new Phanerochaete chrysosporium cellulase. Eur J Biochem 259:88–95.
69. Xia W, Bai Y, Cui Y, Xu X, Qian L, Shi P, Zhang W, Luo H, Zhan X, Yao B. 2016. Functional diversity of family 3 β-glucosidases from thermophilic cellulolytic fungus Humicola insolens Y1. Sci Rep 6:27062.
70. Sukharnikov LO, Alahuhta M, Brunecky R, Upadhyay a., Himmel ME, Lunin V V., Zhulin IB. 2012. Sequence, Structure, and Evolution of Cellulases in Glycoside Hydrolase Family 48. J Biol Chem 287:41068–41077.
71. Chen Z, Friedland GD, Pereira JH, Reveco SA, Chan R, Park JI, Thelen MP, Adams PD, Arkin AP, Keasling JD, Blanch HW, Simmons BA, Sale KL, Chivian D, Chhabra SR. 2012. Tracing determinants of dual substrate specificity in glycoside hydrolase family 5. J Biol Chem 287:25335–25343.
72. Tian L, Liu S, Wang S, Wang L. 2016. Ligand-binding specificity and promiscuity of the main lignocellulolytic enzyme families as revealed by active-site architecture analysis. Sci Rep 6:23605.
73. Berlemont R, Martiny AC. 2013. Phylogenetic distribution of potential cellulases in bacteria. Appl Environ Microbiol 79:1545–1554.
74. Berlemont R, Martiny AC. 2015. Genomic Potential for Polysaccharide Deconstruction in Bacteria. Appl Environ Microbiol 81:1513–1519.
75. Jiménez DJ, Korenblum E, Van Elsas JD. 2014. Novel multispecies microbial consortia involved in lignocellulose and 5-hydroxymethylfurfural bioconversion. Appl Microbiol Biotechnol 98:2789–2803.
76. Cortes-Tolalpa L, Salles JF, van Elsas JD. 2017. Bacterial Synergism in Lignocellulose Biomass Degradation – Complementary Roles of Degraders As Influenced by Complexity of the Carbon Source. Front Microbiol 8:1–14.
77. Barbi F, Bragalini C, Vallon L, Prudent E, Dubost A, Fraissinet-Tachet L, Marmeisse R, Luis P. 2014. PCR primers to study the diversity of expressed fungal genes encoding lignocellulolytic enzymes in soils using high-throughput sequencing. PLoS One 9.
78. Baldrian P, Kolařík M, Stursová M, Kopecký J, Valášková V, Větrovský T, Zifčáková L, Snajdr J, Rídl J, Vlček C, Voříšková J. 2012. Active and total microbial communities in forest soil are largely different and highly stratified during decomposition. ISME J 6:248–58.
79. Elisashvili V, Kachlishvili E, Asatiani M. 2015. Shiitake Medicinal Mushroom, Lentinus edodes (Higher Basidiomycetes) Productivity and Lignocellulolytic Enzyme Profiles during Wheat Straw and Tree Leaf Bioconversion. Int J Med Mushrooms 17:77–86.
80. Raman B, Pan C, Hurst GB, Rodriguez M, McKeown CK, Lankford PK, Samatova NF, Mielenz JR. 2009. Impact of Pretreated Switchgrass and Biomass Carbohydrates on Clostridium thermocellum ATCC 27405 Cellulosome Composition: A Quantitative Proteomic Analysis. PLoS One 4.
81. Blouzard JC, Coutinho PM, Fierobe HP, Henrissat B, Lignon S, Tardif C, Pagès S, De Philip P.

2010. Modulation of cellulosome composition in Clostridium cellulolyticum: Adaptation to the polysaccharide environment revealed by proteomic and carbohydrate-active enzyme analyses. Proteomics 10:541–554.
82. Harris P V., Welner D, McFarland KC, Re E, Navarro Poulsen JC, Brown K, Salbo R, Ding H, Vlasenko E, Merino S, Xu F, Cherry J, Larsen S, Lo Leggio L. 2010. Stimulation of lignocellulosic biomass hydrolysis by proteins of glycoside hydrolase family 61: Structure and function of a large, enigmatic family. Biochemistry 49:3305–3316.
83. Arfi Y, Shamshoum M, Rogachev I, Peleg Y, Bayer E a. 2014. Integration of bacterial lytic polysaccharide monooxygenases into designer cellulosomes promotes enhanced cellulose degradation. Proc Natl Acad Sci U S A 111:9109–14.
84. Gardner JG, Crouch L, Labourel A, Forsberg Z, Bukhman Y V., Vaaje-Kolstad G, Gilbert HJ, Keating DH. 2014. Systems biology defines the biological significance of redox-active proteins during cellulose degradation in an aerobic bacterium. Mol Microbiol 94:1121–1133.
85. Rensburg P Van, Zyl WH Van, Pretorius IS. 1996. Co-expression of a Phanerochaete chrysosporium gene and a Butyrivibrio fibrisolvens endo-beta-1,4-glucanase gene in Saccharomyces cerevisiae. Curr Genet 30:246–250.
86. Brunecky R, Alahuhta M, Xu Q, Donohoe BS, Crowley MF, Kataeva IA, Yang S-J, Resch MG, Adams MWW, Lunin V V., Himmel ME, Bomble YJ. 2013. Revealing Nature's Cellulase Diversity: The Digestion Mechanism of Caldicellulosiruptor bescii CelA. Science (80-) 342:1513–1516.
87. Wood TM, McCrae SI. 1986. The cellulase of Penicillium pinophilum. Synergism between enzyme components in solubilizing cellulose with special reference to the involvement of two immunologically distinct cellobiohydrolases. Biochem J 234:93–99.
88. Wilson DB. 2011. Microbial diversity of cellulose hydrolysis. Curr Opin Microbiol 14:259–263.
89. Baker JO, Ehrman CI, Adney WS, Thomas SR, Himmel ME. 1998. Hydrolysis of cellulose using ternary mixtures of purified cellulases. Appl Biochem Biotechnol 70–72:395–403.
90. Artzi L, Bayer EA, Moraïs S. 2016. Cellulosomes: bacterial nanomachines for dismantling plant polysaccharides. Nat Rev Microbiol 15:83–95.
91. Ravachol J, Borne R, Tardif C, De Philip P, Fierobe HP. 2014. Characterization of all family-9 glycoside hydrolases synthesized by the cellulosome-producing bacterium Clostridium cellulolyticum. J Biol Chem 289:7335–7348.
92. Bomble YJ, Beckham GT, Matthews JF, Nimlos MR, Himmel ME, Crowley MF. 2011. Modeling the self-assembly of the cellulosome enzyme complex. J Biol Chem 286:5614–5623.
93. Stern J, Kahn A, Vazana Y, Shamshoum M, Moraïs S, Lamed R, Bayer EA. 2015. Significance of relative position of cellulases in designer cellulosomes for optimized cellulolysis. PLoS One 10.
94. Boon E, Meehan CJ, Whidden C, Wong DHJ, Langille MGI, Beiko RG. 2014. Interactions in the microbiome: Communities of organisms and communities of genes. FEMS Microbiol Rev.
95. Danchin EGJ, Rosso M-N, Vieira P, de Almeida-Engler J, Coutinho PM, Henrissat B, Abad P. 2010. Multiple lateral gene transfers and duplications have promoted plant parasitism ability in nematodes. Proc Natl Acad Sci U S A 107:17651–17656.
96. Portillo MC, Leff JW, Lauber CL, Fierer N. 2013. Cell size distributions of soil bacterial and archaeal taxa. Appl Environ Microbiol 79:7610–7617.
97. Henneron L, Bernard L, Hedde M, Pelosi C, Villenave C, Chenu C, Bertrand M, Girardin C, Blanchart E. 2014. Fourteen years of evidence for positive effects of conservation agriculture and organic farming on soil life. Agron Sustain Dev 35:169–181.
98. Bengtsson J, Ahnström J, Weibull AC. 2005. The effects of organic agriculture on biodiversity and abundance: A meta-analysis. J Appl Ecol 42:261–269.
99. Oehl F, Sieverding E, Mäder P, Dubois D, Ineichen K, Boller T, Wiemken A. 2004. Impact of long-term conventional and organic farming on the diversity of arbuscular mycorrhizal fungi. Oecologia 138:574–583.
100. Aspeborg H, Coutinho PM, Wang Y, Brumer H, Henrissat B. 2012. Evolution, substrate

specificity and subfamily classification of glycoside hydrolase family 5 (GH5). BMC Evol Biol 12:186.
101. de Vries M, Schöler A, Ertl J, Xu Z, Schloter M. 2015. Metagenomic analyses reveal no differences in genes involved in cellulose degradation under different tillage treatments. FEMS Microbiol Ecol 1–10.
102. Uroz S, Ioannidis P, Lengelle J, Cébron A, Morin E, Buée M, Martin F. 2013. Functional Assays and Metagenomic Analyses Reveals Differences between the Microbial Communities Inhabiting the Soil Horizons of a Norway Spruce Plantation. PLoS One 8.
103. Wilhelm RC, Cardenas E, Leung H, Szeitz A, Jensen LD, Mohn WW. 2017. Long-term enrichment of stress-tolerant cellulolytic soil populations following timber harvesting evidenced by multi-Omic stable isotope probing. Front Microbiol.
104. Wang Q, Tull D, Meinke A, Gilkes NR, Warren RAJ, Aebersold R, Withers SG. 1993. Glu280 is the nucleophile in the active site of Clostridium thermocellum CelC, a family A endo-beta-1,4-glucanase. J Biol Chem 268:14096–14102.
105. Meyer-Aurich A, Gandorfer M, Gerl G, Kainz M. 2009. Tillage and Fertilizer Effects on Yield, Profitability, and Risk in a Corn-Wheat-Potato-Wheat Rotation. Agron J 106:1538–1547.
106. Wittwer R, Dorn B, Jossi W, Van Der Heijden MGA. 2017. Cover crops support ecological intensification of arable cropping systems. Sci Rep 7:1–12.
107. Joergensen RG. 1996. The fumigation-extraction method to estimate soil microbial biomass: calibration of the kEC-value. Soil Biol Biochem 28:25–31.
108. Pritsch K, Luedemann G, Matyssek R, Hartmann A, Schloter M, Scherb H, Grams TEE. 2005. Mycorrhizosphere responsiveness to atmospheric ozone and inoculation with Phytophthora citricola in a phytotron experiment with spruce/beech mixed cultures. Plant Biol 7:718–27.
109. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. 2011. Fast, scalable generation of high quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539.
110. Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI, Lanczycki CJ, Lu F, Marchler GH, Song JS, Thanki N, Wang Z, Yamashita RA, Zhang D, Zheng C, Bryant SH. 2015. CDD: NCBI's conserved domain database. Nucleic Acids Res 43:D222–D226.
111. Gulvik C a., Effler TC, Wilhelm SW, Buchan A. 2012. De-MetaST-BLAST: A Tool for the Validation of Degenerate Primer Sets and Data Mining of Publicly Available Metagenomes. PLoS One 7:1–7.
112. Klingenhoff A, Frech K, Quandt K, Werner T. 1999. Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity. Bioinformatics 15:180–186.
113. Griffiths RI, Whiteley AS, O'Donnell AG, Bailey MJ. 2000. Rapid method for coextraction of DNA and RNA from natural environments for analysis of ribosomal DNA- and rRNA-based microbial community composition. Appl Environ Microbiol 66:5488–5491.
114. Bach HJ, Tomanova J, Schloter M, Munch JC. 2002. Enumeration of total bacteria and bacteria with genes for proteolytic activity in pure cultures and in environmental samples by quantitative PCR mediated amplification. J Microbiol Methods 49:235–245.
115. White TJ, Bruns S, Lee S, Taylor J. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics, p. 315–322. *In* Innis, M, Gelfand, D, Sninsky, J, White, T (eds.), PCR Protocols: A Guide to Methods and Applications. Academic Press, Orlando, Florida.
116. Tange O. 2011. GNU Parallel: the command-line power tool. USENIX Mag 36:42–47.
117. Cox MP, Peterson DA, Biggs PJ. 2010. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics 11:485.
118. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards RA. 2008. The metagenomics RAST server - a public

resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9:386.
119. Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150–2.
120. Buchfink B, Xie C, Huson DH. 2014. Fast and sensitive protein alignment using DIAMOND. Nat Methods 12.
121. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–10.
122. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.
123. Huson D, Mitra S, Ruscheweyh H. 2011. Integrative analysis of environmental sequences using MEGAN4. Genome Res 21:1552–1560.
124. R Core Team. 2013. R: A Language and Environment for Statistical Computing. Vienna.
125. Rho M, Tang H, Ye Y. 2010. FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res 38:e191.
126. Eddy SR. 2011. Accelerated Profile HMM Searches. PLoS Comput Biol.
127. Punta M, Coggill P, Eberhardt R, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer E, Eddy S, Bateman A, Finn R. 2012. The Pfam protein families database. Nucleic Acids Res 40:D290-301.
128. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. 2012. DbCAN: A web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res 40:W445-51.
129. Schubert M, Lindgreen S, Orlando L. 2016. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res Notes 9:88.
130. Schmieder R, Edwards R. 2011. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLoS One 6.
131. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: A software Environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504.
132. Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, Raes J, Huttenhower C. 2012. Microbial co-occurrence relationships in the Human Microbiome. PLoS Comput Biol 8.
133. Schaub MT, Delvenne JC, Yaliraki SN, Barahona M. 2012. Markov dynamics as a zooming lens for multiscale community detection: Non clique-like communities and the field-of-view limit. PLoS One 7:e32210.
134. Oksanen J, Blanchet F, Kindt R, Legendre P, Minchin P, O'Hara R, Simpson G, Solymos P, Stevens M, Wagner H. 2013. vegan: Community Ecology Package. R package version 2.0-10. R Packag version.
135. Rodriguez-R LM, Konstantinidis KT. 2014. Nonpareil: A redundancy-based approach to assess the level of coverage in metagenomic datasets. Bioinformatics 30:629–635.
136. Magoc T, Salzberg SL. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27:2957–2963.
137. Pearson WR. 2000. Flexible sequence similarity searching with the FASTA3 program package. Methods Mol Biol 132:185–219.
138. Enright AJ, Dongen S Van, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30:1575–1584.
139. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Mol Biol Evol 30:2725–2729.
140. Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18:691–699.
141. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol 30:772–780.
142. Domínguez R, Souchon H, Lascombe M-B, Alzari PM. 1996. The Crystal Structure of a Family 5

Endoglucanase Mutant in Complexed and Uncomplexed Forms Reveals an Induced Fit Activation Mechanism. J Mol Biol 257:1042–1051.
143. Torsvik V, Øvreås L. 2002. Microbial diversity and function in soil: From genes to ecosystems. Curr Opin Microbiol 5:240–245.
144. Bergmann GT, Bates ST, Eilers KG, Lauber CL, Caporaso JG, Walters WA, Knight R, Fierer N. 2011. The under-recognized dominance of Verrucomicrobia in soil bacterial communities. Soil Biol Biochem 43:1450–1455.
145. Daniel R. 2005. The metagenomics of soil. Nat Rev Microbiol 3:470–478.
146. Li L-L, McCorkle SR, Monchy S, Taghavi S, van der Lelie D. 2009. Bioprospecting metagenomes: glycosyl hydrolases for converting biomass. Biotechnol Biofuels 2:10.
147. Zhou J, He Z, Yang Y, Deng Y, Tringe SG, Alvarez-Cohen L. 2015. High-throughput metagenomic technologies for complex microbial community analysis: Open and closed formats. MBio 6.
148. Myrold DD, Zeglin LH, Jansson JK. 2014. The Potential of Metagenomic Approaches for Understanding Soil Microbial Processes. Soil Sci Soc Am J 78:3.
149. Thomas T, Gilbert J, Meyer F. 2012. Metagenomics - a guide from sampling to data analysis. Microb Inform Exp 2:3.
150. Wooley JC, Godzik A, Friedberg I. 2010. A primer on metagenomics. PLoS Comput Biol.
151. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2013. GenBank. Nucleic Acids Res 41.
152. Pruitt KD, Tatusova T, Brown GR, Maglott DR. 2012. NCBI Reference Sequences (RefSeq): Current status, new features and genome annotation policy. Nucleic Acids Res 40.
153. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glöckner FO. 2007. SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res 35:7188–7196.
154. Amann RI, Ludwig W, Schleifer KH. 1995. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol Rev 59:143–69.
155. Curtis TP, Sloan WT, Scannell JW. 2002. Estimating prokaryotic diversity and its limits. Proc Natl Acad Sci 99:10494–10499.
156. Collado J, Platas G, Paulus B, Bills GF. 2007. High-throughput culturing of fungi from plant litter by a dilution-to-extinction technique. FEMS Microbiol Ecol 60:521–533.
157. Krüger M, Krüger C, Walker C, Stockinger H, Schüßler A. 2012. Phylogenetic reference data for systematics and phylotaxonomy of arbuscular mycorrhizal fungi from phylum to species level. New Phytol 193:970–984.
158. Buée M, Reich M, Murat C, Morin E, Nilsson RH, Uroz S, Martin F. 2009. 454 Pyrosequencing analyses of forest soils reveal an unexpectedly high fungal diversity. New Phytol 184:449–456.
159. Noronha MF, Lacerda Júnior GV, Gilbert JA, de Oliveira VM. 2017. Taxonomic and functional patterns across soil microbial communities of global biomes. Sci Total Environ 609:1064–1074.
160. Delmont TO, Prestat E, Keegan KP, Faubladier M, Robe P, Clark IM, Pelletier E, Hirsch PR, Meyer F, Gilbert JA, Le Paslier D, Simonet P, Vogel TM. 2012. Structure, fluctuation and magnitude of a natural grassland soil metagenome. ISME J 6:1677–1687.
161. Fierer N, Leff JW, Adams BJ, Nielsen UN, Bates ST, Lauber CL, Owens S, Gilbert JA, Wall DH, Caporaso JG. 2012. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. Proc Natl Acad Sci 109:21390–21395.
162. Andrade AC, Fróes A, Lopes FÁC, Thompson FL, Krüger RH, Dinsdale E, Bruce T. 2017. Diversity of Microbial Carbohydrate-Active enZYmes (CAZYmes) Associated with Freshwater and Soil Samples from Caatinga Biome. Microb Ecol 74:89–105.
163. Malik AA, Thomson BC, Whiteley AS, Bailey M, Griffiths RI. 2017. Bacterial physiological adaptations to contrasting edaphic conditions identified using landscape scale

metagenomics. MBio 8.
164. Manoharan L, Kushwaha SK, Hedlund K, Ahrén D. 2015. Captured metagenomics: Large-scale targeting of genes based on “sequence capture” reveals functional diversity in soils. DNA Res 22:451–460.
165. Land M, Hauser L, Jun S-R, Nookaew I, Leuze MR, Ahn T-H, Karpinets T, Lund O, Kora G, Wassenaar T, Poudel S, Ussery DW. 2015. Insights from 20 years of bacterial genome sequencing. Funct Integr Genomics 15:141–161.
166. Galperin MY, Makarova KS, Wolf YI, Koonin E V. 2015. Expanded Microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res 43:D261–D269.
167. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2016. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 44:D457–D462.
168. Bertrand H, Poly F, Van VT, Lombard N, Nalin R, Vogel TM, Simonet P. 2005. High molecular weight DNA recovery from soils prerequisite for biotechnological metagenomic library construction. J Microbiol Methods 62:1–11.
169. Marine R, Polson SW, Ravel J, Hatfull G, Russell D, Sullivan M, Syed F, Dumas M, Wommack KE. 2011. Evaluation of a transposase protocol for rapid generation of shotgun high-throughput sequencing libraries from nanogram quantities of DNA. Appl Environ Microbiol 77:8071–8079.
170. Bowers RM, Clum A, Tice H, Lim J, Singh K, Ciobanu D, Ngan CY, Cheng JF, Tringe SG, Woyke T. 2015. Impact of library preparation protocols and template quantity on the metagenomic reconstruction of a mock microbial community. BMC Genomics 16.
171. Schirmer M, Ijaz UZ, D’Amore R, Hall N, Sloan WT, Quince C. 2015. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Res 43.
172. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H, Altaf-Ul-Amin M, Ogasawara N, Kanaya S. 2011. Sequence-specific error profile of Illumina sequencers. Nucleic Acids Res 39.
173. Archer J, Baillie G, Watson SJ, Kellam P, Rambaut A, Robertson DL. 2012. Analysis of high-depth sequence data for studying viral diversity: A comparison of next generation sequencing platforms using Segminator II. BMC Bioinformatics 13.
174. Pearson WR. 2013. An introduction to sequence similarity (“homology”) searching. Curr Protoc Bioinforma.
175. Trimble WL, Keegan KP, D’Souza M, Wilke A, Wilkening J, Gilbert J, Meyer F. 2012. Short-read reading-frame predictors are not created equal: Sequence error causes loss of signal. BMC Bioinformatics 13.
176. Rodriguez-R LM, Konstantinidis KT. 2014. Estimating coverage in metagenomic data sets and why it matters. ISME J.
177. Schöler A, Jacquiod S, Vestergaard G, Schulz S, Schloter M. 2017. Analysis of soil microbial communities based on amplicon sequencing of marker genes. Biol Fertil Soils.
178. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, Bork P, Hugenholtz P, Rubin EM. 2005. Comparative metagenomics of microbial communities. Science 308:554–557.
179. Sharpton TJ. 2014. An introduction to the analysis of shotgun metagenomic data. Front Plant Sci 5:1–14.
180. Omelchenko M V., Galperin MY, Wolf YI, Koonin E V. 2010. Non-homologous isofunctional enzymes: A systematic analysis of alternative solutions in enzyme evolution. Biol Direct 5.
181. Zhang X, Wang S, Wu X, Liu S, Li D, Xu H, Gao P, Chen G, Wang L. 2015. Subsite-specific contributions of different aromatic residues in the active site architecture of glycoside hydrolase family 12. Sci Rep 5.
182. Busk PK, Lange L. 2013. Function-based classification of carbohydrate-active enzymes by

recognition of short, conserved peptide motifs. Appl Environ Microbiol 79:3380–3391.
183. Sukharnikov LO, Cantwell BJ, Podar M, Zhulin IB. 2011. Cellulases: Ambiguous nonhomologous enzymes in a genomic perspective. Trends Biotechnol.
184. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A. 2016. The Pfam protein families database: Towards a more sustainable future. Nucleic Acids Res 44:D279–D285.
185. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. 2012. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res 40:W445-51.
186. Hess M, Sczyrba A, Egan R, Kim T-W, Chokhawala H, Schroth G, Luo S, Clark DS, Chen F, Zhang T, Mackie RI, Pennacchio LA, Tringe SG, Visel A, Woyke T, Wang Z, Rubin EM. 2011. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science 331:463–467.
187. Xia Y, Ju F, Fang HHP, Zhang T. 2013. Mining of novel thermo-stable cellulolytic genes from a thermophilic cellulose-degrading consortium by metagenomics. PLoS One 8:e53779.
188. Güllert S, Fischer MA, Turaev D, Noebauer B, Ilmberger N, Wemheuer B, Alawi M, Rattei T, Daniel R, Schmitz RA, Grundhoff A, Streit WR. 2016. Deep metagenome and metatranscriptome analyses of microbial communities affiliated with an industrial biogas fermenter, a cow rumen, and elephant feces reveal major differences in carbohydrate hydrolysis strategies. Biotechnol Biofuels 9.
189. Cardenas E, Kranabetter JM, Hope G, Maas KR, Hallam S, Mohn WW. 2015. Forest harvesting reduces the soil metagenomic potential for biomass decomposition. ISME J 9:2465–2476.
190. Nagao C, Nagano N, Mizuguchi K. 2014. Prediction of detailed enzyme functions and identification of specificity determining residues by random forests. PLoS One 9.
191. Li L-L, Taghavi S, McCorkle SM, Zhang Y-B, Blewitt MG, Brunecky R, Adney WS, Himmel ME, Brumm P, Drinkwater C, Mead D a, Tringe SG, Lelie D Van Der. 2011. Bioprospecting metagenomics of decaying wood: mining for new glycoside hydrolases. Biotechnol Biofuels 4:23.
192. Beguin P. 1990. Molecular Biology of Cellulose Degradation. Annu Rev Microbiol 44:219–248.
193. Kalle E, Kubista M, Rensing C. 2014. Multi-template polymerase chain reaction. Biomol Detect Quantif.
194. Piao X, Sun L, Zhang T, Gan Y, Guan Y. 2008. Effects of mismatches and insertions on discrimination accuracy of nucleic acid probes. Acta Biochim Pol 55:713–720.
195. Pereyra LP, Hiibel SR, Prieto Riquelme M V., Reardon KF, Pruden a. 2010. Detection and quantification of functional genes of cellulosedegrading, fermentative, and sulfate-reducing bacteria and methanogenic archaea. Appl Environ Microbiol 76:2192–2202.
196. Elifantz H, Waidner LA, Michelou VK, Cottrell MT, Kirchman DL. 2008. Diversity and abundance of glycosyl hydrolase family 5 in the North Atlantic Ocean. FEMS Microbiol Ecol 63:316–327.
197. Nautiyal CS, Srivastava S, Mishra S, Asif MH, Chauhan PS, Singh PC, Nath P. 2013. Reduced cell wall degradation plays a role in cow dung-mediated management of wilt complex disease of chickpea. Biol Fertil Soils 49:881–891.
198. Wymelenberg A Vanden, Gaskell J, Mozuch M, BonDurant SS, Sabat G, Ralph J, Skyba O, Mansfield SD, Blanchette RA, Grigoriev I V., Kersten PJ, Cullen D. 2011. Significant alteration of gene expression in wood decay fungi Postia placenta and Phanerochaete chrysosporium by plant species. Appl Environ Microbiol 77:4499–4507.
199. Coradetti ST, Craig JP, Xiong Y, Shock T, Tian C, Glass NL. 2012. Conserved and essential transcription factors for cellulase gene expression in ascomycete fungi. Proc Natl Acad Sci 109:7397–7402.
200. de Menezes AB, Prendergast-Miller MT, Poonpatana P, Farrell M, Bissett A, Macdonald LM, Toscas P, Richardson AE, Thrall PH. 2015. C/N Ratio Drives Soil Actinobacterial Cellobiohydrolase Gene Diversity. Appl Environ Microbiol 81:3016–3028.

201. Brown JR. 2003. Ancient horizontal gene transfer. Nat Rev Genet 4:121–132.
202. Ravenhall M, Škunca N, Lassalle F, Dessimoz C. 2015. Inferring Horizontal Gene Transfer. PLOS Comput Biol 11:e1004095.
203. Trujillo ME, Kroppenstedt RM, Fernández-Molinero C, Schumann P, Martínez-Molina E. 2007. Micromonospora lupini sp. nov. and Micromonospora saelicesensis sp. nov., isolated from root nodules of Lupinus angustifolius. Int J Syst Evol Microbiol 57:2799–2804.
204. Kyndt T, Haegeman A, Gheysen G. 2008. Evolution of GHF5 endoglucanase gene structure in plant-parasitic nematodes: No evidence for an early domain shuffling event. BMC Evol Biol 8.
205. Xie G, Bruce DC, Challacombe JF, Chertkov O, Detter JC, Gilna P, Han CS, Lucas S, Misra M, Myers GL, Richardson P, Tapia R, Thayer N, Thompson LS, Brettin TS, Henrissat B, Wilson DB, McBride MJ. 2007. Genome sequence of the cellulolytic gliding bacterium Cytophaga hutchinsonii. Appl Environ Microbiol 73:3536–3546.
206. Trujillo ME, Riesco R, Benito P, Carro L. 2015. Endophytic actinobacteria and the interaction of Micromonospora and nitrogen fixing plants. Front Microbiol.
207. Lilley CJ, Atkinson HJ, Urwin PE. 2005. Molecular aspects of cyst nematodes. Mol Plant Pathol.
208. Haegeman A, Jones JT, Danchin EGJ. 2011. Horizontal Gene Transfer in Nematodes: A Catalyst for Plant Parasitism? Mol Plant-Microbe Interact 24:879–887.
209. Quillet L, Barray S, Labedan B, Petit F, Guespin-Michel J. 1995. The gene encoding the β-1,4-endoglucanase (CelA) from Myxococcus xanthus: evidence for independent acquisition by horizontal transfer of binding and catalytic domains from actinomycetes. Gene 158:23–29.
210. Bai Y, Müller DB, Srinivas G, Garrido-Oter R, Potthoff E, Rott M, Dombrowski N, Münch PC, Spaepen S, Remus-Emsermann M, Hüttel B, McHardy AC, Vorholt JA, Schulze-Lefert P. 2015. Functional overlap of the Arabidopsis leaf and root microbiota. Nature 528:364–369.
211. Hotopp JCD, Clark ME, Oliveira DCSG, Foster JM, Fischer P, Torres MCM, Giebel JD, Kumar N, Ishmael N, Wang S, Ingram J, Nene R V, Shepard J, Tomkins J, Richards S, Spiro DJ, Ghedin E, Slatko BE, Tettelin H, Werren JH. 2007. Widespread Lateral Gene Transfer from Intracellular Bacteria to Multicellular Eukaryotes. Science (80-) 317:1753–1756.
212. Glasner JD, Yang CH, Reverchon S, Hugouvieux-Cotte-Pattat N, Condemine G, Bohin JP, van Gijsegem F, Yang S, Franza T, Expert D, Plunkett G, Francisco MJS, Charkowski AO, Py B, Bell K, Rauscher L, Rodriguez-Palenzuela P, Toussaint A, Holeva MC, He SY, Douet V, Boccara M, Blanco C, Toth I, Anderson BD, Biehl BS, Mau B, Flynn SM, Barras F, Lindeberg M, Birch PRJ, Tsuyumu S, Shi X, Hibbing M, Yap MN, Carpentier M, Dassa E, Umehara M, Kim JF, Rusch M, Soni P, Mayhew GF, Fouts DE, Gill SR, Blattner FR, Keen NT, Perna NT. 2011. Genome sequence of the plant-pathogenic bacterium Dickeya dadantii 3937. J Bacteriol 193:2076–2077.
213. Endo K, Hakamada Y, Takizawa S, Kubota H, Sumitomo N, Kobayashi T, Ito S. 2001. A novel alkaline endoglucanase from an alkaliphilic Bacillus isolate: Enzymatic properties, and nucleotide and deduced amino acid sequences. Appl Microbiol Biotechnol 57:109–116.
214. Todaka N, Inoue T, Saita K, Ohkuma M, Nalepa CA, Lenz M, Kudo T, Moriya S. 2010. Phylogenetic analysis of cellulolytic enzyme genes from representative lineages of termites and a related cockroach. PLoS One 5.
215. Desai MS, Strassert JFH, Meuser K, Hertel H, Ikeda-Ohtsubo W, Radek R, Brune A. 2010. Strict cospeciation of devescovinid flagellates and Bacteroidales ectosymbionts in the gut of dry-wood termites (Kalotermitidae). Environ Microbiol 12:2120–2132.
216. Fujishima M, Kodama Y. 2012. Endosymbionts in Paramecium. Eur J Protistol 48:124–137.
217. Murugan R, Koch H-J, Joergensen RG. 2013. Long-term influence of different tillage intensities on soil microbial biomass, residues and community structure at different depths. Biol Fertil Soils 50:487–498.
218. D'Haene K, Sleutel S, De Neve S, Gabriels D, Hofman G. 2009. The effect of reduced tillage agriculture on carbon dynamics in silt loam soils. Nutr Cycl Agroecosystems 84:249–265.
219. Cooper J, Baranski M, Stewart G, Nobel-de Lange M, Bàrberi P, Fließbach A, Peigné J, Berner

A, Brock C, Casagrande M, Crowley O, David C, De Vliegher A, Döring TF, Dupont A, Entz M, Grosse M, Haase T, Halde C, Hammerl V, Huiting H, Leithold G, Messmer M, Schloter M, Sukkel W, van der Heijden MGA, Willekens K, Wittwer R, Mäder P. 2016. Shallow non-inversion tillage in organic farming maintains crop yields and increases soil C stocks: a meta-analysis. Agron Sustain Dev.

220. Dungait JAJ, Hopkins DW, Gregory AS, Whitmore AP. 2012. Soil organic matter turnover is governed by accessibility not recalcitrance. Glob Chang Biol.
221. Baldock JA, Skjemstad JO. 2000. Role of the soil matrix and minerals in protecting natural organic materials against biological attack, p. 697–710. *In* Organic Geochemistry.
222. Moyano FE, Vasilyeva N, Bouckaert L, Cook F, Craine J, Curiel Yuste J, Don A, Epron D, Formanek P, Franzluebbers A, Ilstedt U, Kätterer T, Orchard V, Reichstein M, Rey A, Ruamps L, Subke JA, Thomsen IK, Chenu C. 2012. The moisture response of soil heterotrophic respiration: Interaction with soil properties. Biogeosciences 9:1173–1182.
223. Dimassi B, Mary B, Fontaine S, Perveen N, Revaillot S, Cohan JP. 2014. Effect of nutrients availability and long-term tillage on priming effect and soil C mineralization. Soil Biol Biochem 78:332–339.
224. Kandeler E, Tscherko D, Spiegel H. 1999. Long-term monitoring of microbial biomass, N mineralisation and enzyme activities of a chernozem under different tillage management. Biol Fertil Soils 28:343–351.
225. Höflich G, Tauschke M, Kühn G, Werner K, Frielinghaus M, Höhn W. 1999. Influence of long-term conservation tillage on soil and rhizosphere microorganisms. Biol Fertil Soils 29:81–86.
226. Stockfisch N, Forstreuter T, Ehlers W. 1999. Ploughing effects on soil organic matter after twenty years of conservation tillage in Lower Saxony, Germany. Soil Tillage Res 52:91–101.
227. Wagg C, Dudenhöffer JH, Widmer F, van der Heijden MGA. 2018. Linking diversity, synchrony and stability in soil microbial communities. Funct Ecol.
228. Frey SD, Elliott ET, Paustian K. 1999. Bacterial and fungal abundance and biomass in conventional and no-tillage agroecosystems along two climatic gradients. Soil Biol Biochem 31:573–585.
229. Kandeler E, Tscherko D, Spiegel H. 1999. Long-term monitoring of microbial biomass, N mineralisation and enzyme activities of a chernozem under different tillage management. Biol Fertil Soils 28:343–351.
230. Chantigny MH. 2003. Dissolved and water-extractable organic matter in soils: A review on the influence of land use and management practices, p. 357–380. *In* Geoderma.
231. Straathof AL, Chincarini R, Comans RNJ, Hoffland E. 2014. Dynamics of soil dissolved organic carbon pools reveal both hydrophobic and hydrophilic compounds sustain microbial respiration. Soil Biol Biochem 79:109–116.
232. Boschker HTS, Cappenberg TE. 1998. Patterns of extracellular enzyme activities in littoral sediments of Lake Gooimeer, The Netherlands. FEMS Microbiol Ecol 25:79–86.
233. Wiermann C, Werner D, Horn R, Rostek J, Werner B. 2000. Stress/strain processes in a structured unsaturated silty loam Luvisol under different tillage treatments in Germany. Soil Tillage Res 53:117–128.
234. Bingeman CW, Varner JE, Martin WP. 1953. The effect of the addition of organic materials on the decomposition of an organic soil. Soil Sci Soc Am J 17:34–38.
235. Kuzyakov Y, Blagodatskaya E. 2015. Microbial hotspots and hot moments in soil: Concept & review. Soil Biol Biochem.
236. Fontaine S, Mariotti A, Abbadie L. 2003. The priming effect of organic matter: A question of microbial competition? Soil Biol Biochem 35:837–843.
237. Chen R, Senbayram M, Blagodatsky S, Myachina O, Dittert K, Xiangui L, Blagodatskaya E, Kuzyakov Y. 2014. Soil C and N availability determine the priming effect : microbial N mining and stoichiometric decomposition theories. Glob Chang Biol 20:2356–2367.
238. Bevivino A, Dalmastri C. 2017. Impact of Agricultural Land Management on Soil Bacterial

Community: A Case Study in the Mediterranean Area, p. 77–96. *In* Lukac, M, Grenni, P, Gamboni, M (eds.), Soil Biological Communities and Ecosystem Resilience. Springer.
239. Blagodatskaya E V., Blagodatsky SA, Anderson TH, Kuzyakov Y. 2009. Contrasting effects of glucose, living roots and maize straw on microbial growth kinetics and substrate availability in soil, p. 186–197. *In* European Journal of Soil Science.
240. Im WT, Hu ZY, Kim KH, Rhee SK, Meng H, Lee ST, Quan ZX. 2012. Description of fimbriimonas ginsengisoli gen. nov., sp. nov. within the fimbriimonadia class nov., of the phylum armatimonadetes. Antonie van Leeuwenhoek, Int J Gen Mol Microbiol 102:307–317.
241. Davis KER, Sangwan P, Janssen PH. 2011. Acidobacteria, Rubrobacteridae and Chloroflexi are abundant among very slow-growing and mini-colony-forming soil bacteria. Environ Microbiol 13:798–805.
242. Rughöft S, Herrmann M, Lazar CS, Cesarz S, Levick SR, Trumbore SE, Küsel K. 2016. Corrigendum: Community composition and abundance of bacterial, archaeal, and nitrifying populations in savanna soils on contrasting bedrock material in Kruger National Park, South Africa [Front. Microbiol, (2016), 7, (1638)]. doi: 10.3389/fmicb.2016.01638. Front Microbiol.
243. Sapp M, Harrison M, Hany U, Charlton A, Thwaites R. 2015. Comparing the effect of digestate and chemical fertiliser on soil bacteria. Appl Soil Ecol 86:1–9.
244. Ling N, Chen D, Guo H, Wei J, Bai Y, Shen Q, Hu S. 2017. Geoderma Differential responses of soil bacterial communities to long-term N and P inputs in a semi-arid steppe. Geoderma 292:25–33.
245. Sánchez-Marañón M, Miralles I, Aguirre-Garrido JF, Anguita-Maeso M, Millán V, Ortega R, García-Salcedo JA, Martínez-Abarca F, Soriano M. 2017. Changes in the soil bacterial community along a pedogenic gradient. Sci Rep 7.
246. Bernhardt P V. 2006. Enzyme electrochemistry - Biocatalysis on an electrode. Aust J Chem 59:233–256.
247. Liang Y, Gardner DR, Miller CD, Chen D, Anderson AJ, Weimer BC, Sims RC. 2006. Study of biochemical pathways and enzymes involved in pyrene degradation by Mycobacterium sp. strain KMS. Appl Environ Microbiol 72:7821–7828.
248. Baldock JA, Oades JM, Nelson PN, Skene TM, Golchin A, Clarke P. 1997. Assessing the extent of decomposition of natural organic materials using solid-state 13C NMR spectroscopy. Aust J Soil Res 35:1061–1084.
249. Baldock JA, Sanderman J, MacDonald LM, Puccini A, Hawke B, Szarvas S, McGowan J. 2013. Quantifying the allocation of soil organic carbon to biologically significant fractions. Soil Res 51:561–576.
250. Degrune F, Dufrêne M, Colinet G, Massart S, Taminiau B, Bodson B, Hiel MP, Daube G, Nezer C, Vandenbol M. 2015. A novel sub-phylum method discriminates better the impact of crop management on soil microbial community. Agron Sustain Dev 35:1157–1166.
251. Suto M, Tomita F. 2001. Induction and catabolite repression mechanisms of cellulase in fungi. J Biosci Bioeng.
252. Pordesimo LO, Hames BR, Sokhansanj S, Edens WC. 2005. Variation in corn stover composition and energy content with crop maturity. Biomass and Bioenergy 28:366–374.
253. Hanada S. 2014. The phylum chloroflexi, the family chloroflexaceae, and the related phototrophic families oscillochloridaceae and roseiflexaceae, p. 515–532. *In* The Prokaryotes: Other Major Lineages of Bacteria and The Archaea.
254. Loaiza Puerta V, Pujol Pereira EI, Wittwer R, van der Heijden M, Six J. 2018. Improvement of soil structure through organic crop management, conservation tillage and grass-clover ley. Soil Tillage Res 180:1–9.
255. Degrune F, Theodorakopoulos N, Dufrêne M, Colinet G, Bodson B, Hiel MP, Taminiau B, Nezer C, Daube G, Vandenbol M. 2016. No favorable effect of reduced tillage on microbial community diversity in a silty loam soil (Belgium). Agric Ecosyst Environ 224:12–21.
256. Ghimire R, Norton JB, Stahl PD, Norton U. 2014. Soil microbial substrate properties and

microbial community responses under irrigated organic and reduced-tillage crop and forage production systems. PLoS One 9.
257. Yao H, He Z, Wilson M, Campbell C. 2000. Microbial Biomass and Community Structure in a Sequence of Soils with Increasing Fertility and Changing Land Use. Microb Ecol 40:223–237.
258. Spedding TA, Hamel C, Mehuys GR, Madramootoo CA. 2004. Soil microbial dynamics in maize-growing soil under different tillage and residue management systems. Soil Biol Biochem 36:499–512.
259. Yang Q, Wang X, Shen Y. 2013. Comparison of soil microbial community catabolic diversity between Rhizosphere and bulk soil induced by tillage or residue retention. J Soil Sci Plant Nutr 13:187–199.
260. Huang M, Chen J, Cao F, Jiang L, Zou Y. 2016. Rhizosphere processes associated with the poor nutrient uptake in no-tillage rice (Oryza sativa L.) at tillering stage. Soil Tillage Res 163:10–13.
261. Bogužas V, Kairyte a, Jodaugiene D. 2010. Soil physical properties and earthworms as affected by soil tillage systems, straw and green manure management . Zemdirbyste 97:3–14.
262. Eylenbosch D, Fernández Pierna JA, Baeten V, Bodson B. 2015. Use of NIR hyperspectral imaging and chemometrics to quantify roots and crop residues in soil17th International Conference on Near Infrared Spectroscopy, organized by the International Council for Near Infrared Spectroscopy (ICNIRS) on 18-23 October 2015, Foz do Iguassu, Brasil.
263. Guan D, Zhang Y, Al-Kaisi MM, Wang Q, Zhang M, Li Z. 2015. Tillage practices effect on root distribution and water use efficiency of winter wheat under rain-fed condition in the North China Plain. Soil Tillage Res 146:286–295.
264. Barzegar AR, Mossavi MH, Asoodar MA, Herbert SJ. 2004. Root Mass Distribution of Winter Wheat as Influenced by Different Tillage Systems in Semi Arid Region. J Agron 3:223–228.
265. Dennis PG, Miller AJ, Hirsch PR. 2010. Are root exudates more important than other sources of rhizodeposits in structuring rhizosphere bacterial communities? FEMS Microbiol Ecol.
266. Bertin C, Yang X, Weston LA. 2003. The role of root exudates and allelochemicals in the rhizosphere. Plant Soil.
267. Fan K, Cardona C, Li Y, Shi Y, Xiang X, Shen C, Wang H, Gilbert JA, Chu H. 2017. Rhizosphere-associated bacterial network structure and spatial distribution differ significantly from bulk soil in wheat crop fields. Soil Biol Biochem 113:275–284.
268. Bérdy J. 2005. Bioactive Microbial Metabolites. J Antibiot (Tokyo) 58:1–26.
269. De Schrijver A, De Mot R. 1999. Degradation of pesticides by actinomycetes. Crit Rev Microbiol 25:85–119.
270. López-Mondéjar R, Zühlke D, Becher D, Riedel K, Baldrian P. 2016. Cellulose and hemicellulose decomposition by forest soil bacteria proceeds by the action of structurally variable enzymatic systems. Sci Rep 6.
271. Bloemberg G V., Lugtenberg BJJ. 2001. Molecular basis of plant growth promotion and biocontrol by rhizobacteria. Curr Opin Plant Biol.
272. Geddes BA, Oresnik IJ. 2014. Physiology, genetics, and biochemistry of carbon metabolism in the alphaproteobacterium *Sinorhizobium meliloti*. Can J Microbiol 60:491–507.
273. Moreira FS, Costa PB, Souza R De, Beneduzi A, Lisboa BB, Vargas LK, Passaglia LMP. 2016. Functional abilities of cultivable plant growth promoting bacteria associated with wheat (Triticum aestivum L .) crops. Genet Mol Biol 39:111–121.
274. Philippot L, Raaijmakers JM, Lemanceau P, Van Der Putten WH. 2013. Going back to the roots: The microbial ecology of the rhizosphere. Nat Rev Microbiol.
275. Vandenkoornhuyse P, Mahe S, Ineson P, Staddon P, Ostle N, Cliquet J-B, Francez A-J, Fitter AH, Young JPW. 2007. Active root-inhabiting microbes identified by rapid incorporation of plant-derived carbon into RNA. Proc Natl Acad Sci 104:16970–16975.
276. Peiffer JA, Spor A, Koren O, Jin Z, Tringe SG, Dangl JL, Buckler ES, Ley RE. 2013. Diversity and heritability of the maize rhizosphere microbiome under field conditions. Proc Natl Acad Sci 110:6548–6553.

277. Mendes R, Kruijt M, De Bruijn I, Dekkers E, Van Der Voort M, Schneider JHM, Piceno YM, DeSantis TZ, Andersen GL, Bakker PAHM, Raaijmakers JM. 2011. Deciphering the rhizosphere microbiome for disease-suppressive bacteria. Science (80-) 332:1097–1100.
278. Postma J, Stevens LH, Wiegers GL, Davelaar E, Nijhuis EH. 2009. Biological control of Pythium aphanidermatum in cucumber with a combined application of Lysobacter enzymogenes strain 3.1T8 and chitosan. Biol Control 48:301–309.
279. Boch J, Bonas U. 2010. *Xanthomonas* AvrBs3 Family-Type III Effectors: Discovery and Function. Annu Rev Phytopathol 48:419–436.
280. Wrótniak-Drzewiecka W, Brzezińska AJ, Dahm H, Ingle AP, Rai M. 2016. Current trends in myxobacteria research. Ann Microbiol.
281. Lupwayi NZ, Larney FJ, Blackshaw RE, Kanashiro DA, Pearson DC, Petri RM. 2017. Pyrosequencing reveals profiles of soil bacterial communities after 12 years of conservation management on irrigated crop rotations. Appl Soil Ecol 121:65–73.
282. Breidenbach B, Pump J, Dumont MG. 2015. Microbial Community Structure in the Rhizosphere of Rice Plants. Front Microbiol 6:1537.
283. Correa-Galeote D, Bedmar EJ, Fernández-González AJ, Fernández-López M, Arone GJ. 2016. Bacterial Communities in the Rhizosphere of Amilaceous Maize (Zea mays L.) as Assessed by Pyrosequencing. Front Plant Sci 7.
284. Kawasaki A, Donn S, Ryan PR, Mathesius U, Devilla R, Jones A, Watt M. 2016. Microbiome and exudates of the root and rhizosphere of brachypodium distachyon, a model for wheat. PLoS One 11.
285. Turner TR, Ramakrishnan K, Walshaw J, Heavens D, Alston M, Swarbreck D, Osbourn A, Grant A, Poole PS. 2013. Comparative metatranscriptomics reveals kingdom level changes in the rhizosphere microbiome of plants. ISME J 7:2248–2258.
286. Yun SC. 2014. Selection and a 3-year field trial of Sorangium cellulosum KYC 3262 against anthracnose in hot pepper. Plant Pathol J 30:279–287.
287. Garcia R, Müller R. 2014. The family polyangiaceae, p. 247–279. *In* Rosenberg, E, DeLong, EF, Lory, S, Stackebrandt, E, Thompson, F (eds.), The Prokaryotes. Springer, Berlin, Heidelberg.
288. Sharma G, Khatri I, Subramanian S. 2016. Complete Genome of the Starch-Degrading Myxobacteria Sandaracinus amylolyticus DSM 53668T. Genome Biol Evol 8:2520–2529.
289. Martinez-Garcia M, Brazel DM, Swan BK, Arnosti C, Chain PSG, Reitenga KG, Xie G, Poulton NJ, Gomez ML, Masland DED, Thompson B, Bellows WK, Ziervogel K, Lo CC, Ahmed S, Gleasner CD, Detter CJ, Stepanauskas R. 2012. Capturing single cell genomes of active polysaccharide degraders: An unexpected contribution of verrucomicrobia. PLoS One 7.
290. Soares FL, Melo IS, Dias ACF, Andreote FD. 2012. Cellulolytic bacteria from soils in harsh environments. World J Microbiol Biotechnol 28:2195–2203.
291. Burrell PC, O'Sullivan C, Song H, Clarke WP, Blackall LL. 2004. Identification, Detection, and Spatial Resolution of Clostridium Populations Responsible for Cellulose Degradation in a Methanogenic Landfill Leachate Bioreactor. Appl Environ Microbiol 70:2414–2419.
292. Trdá L, Boutrot F, Claverie J, Brulé D, Dorey S, Poinssot B. 2015. Perception of pathogenic or beneficial bacteria and their evasion of host immunity: pattern recognition receptors in the frontline. Front Plant Sci 6.
293. Kriaučiuniene Z, Velička R, Raudonius S. 2012. The influence of crop residues type on their decomposition rate in the soil: a litterbag study. Zemdirb 99:227–236.
294. Hewezi T, Howe P, Maier TR, Hussey RS, Mitchum MG, Davis EL, Baum TJ. 2008. Cellulose Binding Protein from the Parasitic Nematode Heterodera schachtii Interacts with Arabidopsis Pectin Methylesterase: Cooperative Cell Wall Modification during Parasitism. PLANT CELL ONLINE 20:3080–3093.
295. Abbott DW, Hrynuik S, Boraston AB. 2007. Identification and Characterization of a Novel Periplasmic Polygalacturonic Acid Binding Protein from Yersinia enterolitica. J Mol Biol 367:1023–1033.

296. Nataf Y, Bahari L, Kahel-Raifer H, Borovok I, Lamed R, Bayer EA, Sonenshein AL, Shoham Y. 2010. Clostridium thermocellum cellulosomal genes are regulated by extracytoplasmic polysaccharides via alternative sigma factors. Proc Natl Acad Sci 107:18646–18651.
297. Kahel-Raifer H, Jindou S, Bahari L, Nataf Y, Shoham Y, Bayer EA, Borovok I, Lamed R. 2010. The unique set of putative membrane-associated anti-σ factors in Clostridium thermocellum suggests a novel extracellular carbohydrate-sensing mechanism involved in gene regulation. FEMS Microbiol Lett 308:84–93.
298. Xu J, Bjursell MK, Himrod J, Deng S, Carmichael LK, Chiang HC, Hooper L V., Gordon JI. 2003. A genomic view of the human-Bacteroides thetaiotaomicron symbiosis. Science (80-) 299:2074–2076.
299. Bol R, Kandeler E, Amelung W, Glaser B, Marx MC, Preedy N, Lorenz K. 2003. Short-term effects of dairy slurry amendment on carbon sequestration and enzyme activities in a temperate grassland. Soil Biol Biochem 35:1411–1421.
300. Hungria M, Franchini JC, Brandão-Junior O, Kaschuk G, Souza RA. 2009. Soil microbial activity and crop sustainability in a long-term experiment with three soil-tillage and two crop-rotation systems. Appl Soil Ecol 42:288–296.
301. Carbonetto B, Rascovan N, Álvarez R, Mentaberry A, Vázquez MP. 2014. Structure, composition and metagenomic profile of soil microbiomes associated to agricultural land use and tillage systems in Argentine Pampas. PLoS One 9:11.
302. Souza RC, Cantão ME, Vasconcelos ATR, Nogueira MA, Hungria M. 2013. Soil metagenomics reveals differences under conventional and no-tillage with crop rotation or succession. Appl Soil Ecol 72:49–61.
303. Souza RC, Hungria M, Cantão ME, Vasconcelos ATR, Nogueira MA, Vicente VA. 2015. Metagenomic analysis reveals microbial functional redundancies and specificities in a soil under different tillage and crop-management regimes. Appl Soil Ecol 86:106–112.
304. DeAngelis KM, Gladden JM, Allgaier M, D'haeseleer P, Fortney JL, Reddy A, Hugenholtz P, Singer SW, Gheynst JS, Silver WL, Simmons BA, Hazen TC. 2010. Strategies for Enhancing the Effectiveness of Metagenomic-based Enzyme Discovery in Lignocellulolytic Microbial Communities. BioEnergy Res.
305. Orellana LH, Chee-Sanford JC, Sanford RA, Löffler FE, Konstantinidis KT. 2018. Year-round shotgun metagenomes reveal stable microbial communities in agricultural soils and novel ammonia oxidizers responding to fertilization. Appl Environ Microbiol 84.
306. Fierer N, Bradford MA, Jackson RB. 2007. Toward an ecological classification of soil bacteria. Ecology 88:1354–1364.
307. Beylich A, Graefe U, Elsner D. 2015. Response of microannelids to tillage at soil-monitoring sites in Schleswig-Holstein, Germany. Soil Org 87:121–135.
308. Rezácová V, Baldrian P, Hrselová H, Larsen J, Gryndler M. 2007. Influence of mineral and organic fertilization on soil fungi, enzyme activities and humic substances in a long-term field experiment. Folia Microbiol (Praha) 52:415–421.
309. Hartman K, van der Heijden MGA, Wittwer RA, Banerjee S, Walser J-C, Schlaeppi K. 2018. Cropping practices manipulate abundance patterns of root and soil microbiome members paving the way to smart farming. Microbiome 6:14.
310. Tedersoo L, Pärtel K, Jairus T, Gates G, Põldmaa K, Tamm H. 2009. Ascomycetes associated with ectomycorrhizas: Molecular diversity and ecology with particular reference to the Helotiales. Environ Microbiol 11:3166–3178.
311. Kuzyakov Y. 2010. Priming effects: Interactions between living and dead organic matter. Soil Biol Biochem 42:1363–1371.
312. Stott DE, Andrews SS, Liebig MA, Wienhold BJ, Karlen DL. 2010. Evaluation of β-Glucosidase Activity as a Soil Quality Indicator for the Soil Management Assessment Framework. Soil Sci Soc Am J 74:107.
313. Zhang L, Chen W, Burger M, Yang L, Gong P, Wu Z. 2015. Changes in soil carbon and enzyme

activity as a result of different long-term fertilization regimes in a greenhouse field. PLoS One 10.

314. Franchini JC, Crispino CC, Souza RA, Torres E, Hungria M. 2007. Microbiological parameters as indicators of soil quality under various soil management and crop rotation systems in southern Brazil. Soil Tillage Res 92:18–29.

List of abbreviations

AA	auxiliary activity
ANOVA	analysis of variance
BLAST(p)	Basic Local Alignment Search Tool (protein)
bp	base pairs
CAZy	Carbohydrate-Active enZYmes
CBM	carbohydrate binding module
CBP	cellulose binding protein
cfu	colony-forming units
CMC	carboxymethyl cellulose
C_{mic}	microbial biomass carbon
CT	conventional tillage
dbCAN	Database for automated Carbohydrate-active enzyme Annotation
DNA	deoxyribonucleic acid
DOC	dissolved organic carbon
EC	Enzyme Commission
e-value	expectation value
FAST-experiment	Farming System and Tillage experiment
GH	glycoside hydrolase
HF	high fertilization
HGT	horizontal gene transfer
HMM	Hidden Markov Model
ITS	Internal Transcribed Spacer
KEGG	Kyoto Encyclopedia of Genes and Genomes
L	legume cover crop
LF	low fertilization
LPMO	Lytic Polysaccharide Mono-oxygenase
LSD	least significant difference
M	mixture of cover crops
MEGAN	MetaGenome ANalyzer
MF	medium fertilization
MID	Molecular Identifier
MSA	multiple sequence alignment
MT	medium tillage
MU	methylumbeliferone
NCBI	National Center for Biotechnology Information
NL	non-legume cover crop
N_{mic}	microbial biomass nitrogen
NO	no cover crop
NT	no-tillage
ORF	open reading frame
PCR	polymerase chain reaction
Pfam (database)	Protein family (database)
PT	plough tillage
qPCR	quantitative real-time PCR
Refseq database	Reference Sequences database
RNA	ribonucleic acid
rRNA	ribosomal RNA
RT	reduced tillage
TC	total carbon
TN	total nitrogen
WEOC	water extractable organic carbon

Contributors to the data presented

All of the work described in this thesis was performed by me, except when explicitly noted otherwise. Furthermore, many bioinformatic analysis was done with help from or by adjusting scripts of Zhuofei Xu (at the time of cooperation: University of Copenhagen, Microbiology, Universitetsparken 15, 2100 København Ø, Denmark), Gisle Vestergaard (at the time of cooperation: Helmholtz Zentrum München, department of Environmental Genomics, Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany) and Julia Ertl (at the time of cooperation: Helmholtz Zentrum München, department of Environmental Genomics, Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany). Anne Schöler (at the time of cooperation: Helmholtz Zentrum München, department of Environmental Genomics, Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany) authored many scripts for the analysis of metagenome data with the statistical program R (124), some of which I adapted to analyse the data presented here. Furthermore, she was involved in data generation (soil sampling, chemical and biological analysis and metagenome library preparation) and bioinformatic analysis for data of the conventional farming experiment.

Acknowledgements

The last five years have been an exceptional learning experience in my life, both on a professional and on a personal level. I have experienced how to do science and how not to do science, I have familiarized myself with my own limits and potentials and that of other people. However, this would not have been possible without the contribution of several people and institutions.

First of all, I would like to thank the European Union and Marie Curie Actions for funding this project within the Trainbiodiverse training network. Participation in this network has provided me with excellent connections within the research area, insight into the ins and outs of scientific cooperations and a frame of reference for PhD-projects at different European universities. Therefore, I also thank Prof. Dr. Michael Schloter for giving me the opportunity to perform my PhD within this network and at the Environmental Genomics-research unit. Further financial support and facilities for this research were provided by the Helmholtz Zentrum München and the Helmholtz Graduate School HELENA, for which I give special thanks to Mrs. Renate Schlusen and Mrs. Monika Beer, who have given me the support and confidence I needed to finalize this work.

Second, I have greatly profited from the supervision by and scientific contributions of several people: Prof. Dr. Michael Schloter has provided much feedback, guidance and focus for my PhD-thesis, for which I am particularly greatful. Furthermore, I would like to thank Dr. Anne Schöler, who has been, besides an excellent supervisor, a great friend and discussion partner throughout the years who was always available for personal and professional advice. We have spent a great deal of time discussing fundamental questions in analysis methology and microbial ecology theory, also together with Dr. Joseph Nesme, who was always an enthusiastic discussion partner. High appre-ciation goes especially to Dr. Marion Engel, for taking much time and interest to provide me with critical but constructive and valuable comments on my thesis, for being open and supportive and for being part of my Ph.D. thesis committee. In this context, I also thank Prof. Dr. Timothy M. Vogel (Ecole Centrale de Lyon) for his strategic and objective input for the development of this research project as a Ph.D. thesis committee member. Furthermore, Dr. Gisle Vestergaard has given much conceptual and practical guidance in the topic of cellulase gene analysis and evolutionary theory, for which I am sincerely grateful. In addition, I would like to thank Dr. David Endesfelder and Dr. Gerd Welzl, who have contributed to this work with input for the co-occurrence respectively statistical analysis. My other colleagues at EGEN/COMI and those at the experimental farms in Scheyern and at Agroscope receive my thanks for their help with field sampling, lab work and friendly atmosphere.

Finally, I would like to express my gratitude to those who have been there for me during lunch and coffee breaks and who were open for a laugh, a cry or philosophical conversations. I can not start to mention everyone with whom I had the pleasure (friends at EGEN/COMI, IBÖE/BIOP, DINI, Trainbiodiverse) and who made my time more than worthwhile. These people include José and Nirav, whom I thank for their companionship and fun in the WG and Nirav for inspiration by being such a passionate researcher. They also include Irina, Anne, Matea, Antonis and Urska: we started out together, thank you for the great times during lunches, coffee breaks and especially for the good times after work. I am also grateful to Matea for showing courage and hands-on mentality by initiating the additional thesis supervision, which was critical for the finalization of this thesis. Much consolation and companionship I received as well from the IBÖE modelling guys: Christian (K.),

Florian, Christian (B.) and Christoph, thank you for your humour and friendship. Also Viviane and Karin I will always remember for their involvement, openness and honest conversations. Paramount during these last years was the continuing love I received from and faith of my partner and friend Christoph, which dragged me through times of disheartenment. Danke dafür, mein lieber Christoph. Auch deine Familie bin ich dankbar für ihre immer liebevolle Unterstützung. Nina und Andre, Erich und Jasmin, danke für die gute und lustige Zeit mit euch! Last but definitely not least, I am grateful for the patience, encouragement (Daf: “you are doing a great fucking job”) and genuine interest of all my family and friends in the Netherlands, who know how much I treasure them; pap en mam, Sandra, Ruth, Willem en Clarissa, Jan Willem, Keri, Brenda, Wendel en Lieke, Alice, Ryancke, Anke en Linda, Tessa, Miriam, Marije en Gerdien, Martijn, Agnieszka, Jaap en Karen, Menen, Sjoerd, Leks, Sebas en Biya, ik ben enorm dankbaar voor de lol die we samen hebben, voor jullie vriendschap en liefde, voor de tijd die jullie genomen hebben om mij op te zoeken in het verre München of om mij te ontvangen tijdens mijn vele reizen naar Nederland. Without you, I would have been totally Germinated ☺