Establishing optimum DNA annotation methods to investigate the impacts of flooding on microbial communities and functions

Richard James Randle-Boggis

PhD

University of York

Biology

June 2016

Abstract

Environmental change will have significant impacts on microbial ecosystems. Microorganisms dominate most biogeochemical pathways, and environmental perturbations may alter these functions. Such functions include nutrient cycling, pollution abatement and greenhouse gas emission, and it is paramount that the impact of environmental change on ecosystems is understood. High throughput DNA sequencing provides a window into complex microbial communities and their functional potential, thus allowing us to empirically study how such communities respond to predicted future environments. There are, however, caveats and challenges associated with such technologies, particularly with converting billions of sequencing base calls into species and function counts. This thesis firstly quantifies the performances of sequence annotation tools and parameters using a simulated metagenome. It is found that tools differ in performance, and that parameter selection can significantly reduce annotation accuracy e.g. One Codex correctly annotated many sequences at the genus level, whereas MG-RAST RefSeq produced many false positive annotations. The results provide a guideline to quantitatively inform researchers about the impacts of certain choices on annotation performance, and show that some published studies may be drawing incorrect conclusions.

This thesis also investigates the impacts of increased flooding frequency and duration on soil microbial ecosystems, in line with predicted climate change. Increased frequency has significant impacts on biodiversity, community composition and potential function. SkyLine, a novel, continuous gas flux measuring system, was used to record CO_2 and CH_4 fluxes. Increased flooding duration significantly reduced CH_4 oxidation and increased CO_2 assimilation, with the combined global warming potential of these gasses reduced.

Contents

List of Tables

List of Figures

Acknowledgements

First I would like to express thanks to Peter Ashton and Thorunn Helgason for their help, advice and support throughout my project. Peter provided excellent support and he helped me get over many hurdles. Thorunn provided a wealth of knowledge and wisdom to my project. It has been a pleasure to work with them both.

My Thesis Advisory Panel members Phil Ineson and James Chong contributed valuable suggestions and scientific knowledge. For this I thank them. I would also like to thank Neil Boonham, Mel Sapp, and Richard Thwaites from Fera Science Ltd. for their supervision. I would like to thank James Stockdale, and Phil Ineson again, for allowing me to use their equipment and providing valuable knowledge.

For teaching me programming skills, I would like to thank Sandy Macdonald and Toby Hodges. Sandy in particular provided advice in practically all aspects of my PhD. It was a pleasure sharing an office him and he taught me many computational techniques.

I made many great friends in York, who all made this experience even more enjoyable. Thank you to all of you. I would also like to thank my parents for their continued support over the years, and Jacques and Tina Maurice too. My gran's continued interest in my research, and her knowledge on whatever topic I am studying, is both appreciated and perplexing. She probably knows more about my thesis than I do.

Iain Croall definitely helped with the "downtime" aspects of a PhD, and he was very supportive and motivational. I shall say this here: you were right.

Finally, I would like to say a huge thanks to Joëlle Maurice. She made living in York all the more enjoyable, and supported me through the good times and the challenging. The frequent holidays certainly helped focus my mind on my research.

Author's declaration

I, Richard James Randle-Boggis, declare that this thesis and the work presented in it are my own and have been generated by me as the result of my own original research. Raw gas flux data generated for the fourth chapter were processed by James Stockdale, who provided me with fluxes over the time cycles recorded by SkyLine. I conducted all further analyses of this data. Where others contributed to work, it is stated accordingly.

This work was done wholly while in candidature for a research degree at the University of York. Where I have consulted the published work of others, this is always clearly attributed.

The second chapter has been published in FEMS Microbiology Ecology (Randle-Boggis *et al.*, 2016), co-authored by Peter Ashton, Thorunn Helgason and Melanie Sapp. The third and fourth chapters are in preparation for submission into Environmental Microbiology and The ISME Journal, respectively.

This PhD was funded by the University of York and Fera Science Ltd. (Formally the Food and Environment Research Agency) Seedcorn funding. The experiment conducted in Chapter 4 was partly funded by NERC Macronutrient Cycles research programme.

1 Introduction

1.1 Environmental change and flooding

1.1.1 Climate change

It is widely accepted that predicted climate change would increase the frequency and severity of extreme precipitation events in the UK. These events include greater precipitation in winter and more severe droughts in summer that are interspersed with sporadic heavy precipitation (Blenkinsop and Fowler, 2007; Collins *et al.*, 2013; Houghton, 2001; Kirtman *et al.*, 2013; Kleinen and Petschel-Held, 2007; Min *et al.*, 2011; Murphy *et al.*, 2009; Patz *et al.*, 2005; Trenberth, 1999). Evaporation rates will rise with higher atmospheric temperatures and the warmer air will hold more water. The period between precipitation events and the amount of precipitation that falls will increase (Trenberth *et al.*, 2003). The Climate Change Risk Assessment (CCRA) recognised that the risk of flooding is likely to increase in the UK and it has identified flooding as the most serious of 100 challenges facing the UK's economy, society and environment arising from climate change (DEFRA, 2012; Morse, 2010). Growing populations will likely cause changes in land use (e.g. increased agricultural land development and increased paved/concrete areas); certain land use changes are associated with greater flooding frequency (Poff, 2002).

The European Environment Agency predicts that the frequency of winter and spring flooding will increase in the UK (Kurnik *et al.*, 2012). The predicted changes in precipitation are supported by analysis of historical data (Jones *et al.*, 2013; Min *et al.*, 2011; Osborn *et al.*, 2000). Murphy *et al.* (2009) describe the predictions for changes in precipitation rates regionally and nationally for the UK; in summary, with the most probable situation under the medium emissions scenario (see Nakićenović *et al.*, 2000), the greatest increase in winter precipitation will be 33% by 2050. This figure is expected to intensify further by 2080.

Flooding affects greenhouse gas fluxes (e.g. CO_2 and CH_4) from soil and plants, which will further affects climate change (Conrad and Rothfuss, 1991; Kelly *et al.*, 1997; Miyata *et al.*, 2000) through radiative forcing (Myhre *et al.*, 2013). Investigating the responses of ecosystems to flooding is crucial to understanding climate feedback cycles associated with ecosystem functions.

1.1.2 Microbial ecosystems

Anoxic conditions resulting from flooding will impact soil properties and ecosystems (Ponnamperuma, 1984; Stams and Plugge, 2010). For example, Zhou *et al.* (2002) found that soils saturated in water have reduced bacterial diversities. Microorganisms dominate most biogeochemical cycles (Falkowski *et al.*, 2008), and alterations to community structure and function will result in changes in these cycles. As the frequency of extreme weather conditions is predicted to increase, it is necessary to understand how these changes will affect ecosystem structures and functions.

One response to flooding is the induction of anoxia in bulk soil, which reduces aerobic respiration while allowing anaerobic organisms to thrive. Such organisms include methanogens (Conrad, 2007) and denitrifiers (Zumft, 1997), which both impact greenhouse gas fluxes. Methanogens are strict anaerobes, producing methane from acetate and/or hydrogen. Methane has a 100-year global warming potential (GWP) 34 times greater than CO_2 (Myhre *et al.*, 2013), thus an increase in emissions from more flooding events could play a role in a positive feedback cycle of climatic warming. Denitrifiers convert nitrate to nitrous oxide and inorganic nitrogen under anaerobic conditions, although they may also function under aerobic conditions.

Some studies have investigated the effects of flooding on microbial ecosystems using targeted approaches. Studying four sites with varying flooding patterns along a river, Bodelier *et al.* (2012) used denaturing gel gradient electrophoresis (DGGE) and phospholipid fatty acid analysis (PLFA) and found that the abundance of methanotrophs increased as flooding increased. Kemnitz *et al.* (2004) identified an increase in methanogen diversity in areas with greater flooding from the same river using terminal-restriction fragment length polymorphism (T-RFLP). Unger *et al.* (2009) found that flooding decreased the bacteria:fungi ratio using PLFA over five weeks. These studies provide a useful insight into the effects of flooding on microbial diversity and community composition, however it is clear that a deeper understanding of the effects of environmental change on the whole community, and function is required.

1.1.3 Methanogenesis and methane oxidation

One possible consequence of flooding is an increase in CH_4 emissions, due to an increase in methanogen abundance. Methanogens (phylum: Euryachaeota, domain: Archaea) produce CH_4 under anaerobic conditions. There are two main pathways methanogens use to obtain their energy: substrates such as H_2, CO, formate,

isopropanol and ethanol act as electron donors that allow the reduction of CO_2 to CH_4; alternatively, substrates such as acetate, methanol, trimethylamine and dimethylsulfide can be cleaved, with the carboxyl group being oxidised to CO_2 and the methyl group being reduced to CH_4. Two major groups of methanogens are acetotrophic and hydrogenotropic methanogens. Acetotrophic methanogens either belong to the *Methanosarcina* or the *Methanosaeta* genera, and they convert acetic acid to CH_4 and CO_2 as shown below:

$$CH_3COOH \rightarrow CH_4 + CO_2$$

Hydrogenotrophic methanogens reduce CO_2 with H_2 to produce CH_4 and water:

$$CO_2 + 4H_2 \rightarrow CH_4 + 2H_2O$$

When soils flood, methane is produced after a lag phase, or reduction phase, during which inorganic electron receptors such as nitrate, sulphate and ferric iron are reduced and CH_4 production is supressed. During this phase, saccharolysis and fermentation occur, and H_2 and acetate are produced. These substrates are used by the methanogens during the next phase, the methanogenic phase, where the production rate of CH_4 rapidly increases. Production fluctuates as sulphate and iron reducers deplete H_2 levels (Ratering and Conrad, 1998); once the sulphate and ferric iron substrates have been depleted, methenogenesis continues as H_2 levels rise again. Eventually the production and consumption of H_2 and acetate reach a steady phase. Conrad (2007) provides a detailed review of the processes that occur during methanogenesis.

Methanotrophs metabolise methane to produce their energy and they are found both aerobically and anaerobically. Anaerobic methanotrophs (domain: Archaea) mostly occur in marine sediments within syntrophic microbial consortia. They oxidise CH_4 using sulphate as an electron acceptor:

$$CH_4 + SO_4^{2-} \rightarrow HCO_3^- + HS^- + H_2O$$

Anaerobic oxidation of methane (AOM) with nitrate has also been found in an anaerobic sewage digester (Raghoebarsing *et al.*, 2006).

Aerobic methanotrophs oxidise methane with O_2 and reducing agents using a methane monooxygenase (MMO), producing methanol and water:

$$CH_4 + O_2 + 2[H] \rightarrow CH_3OH + H_2O$$

During a flood, once trapped O_2 has been depleted via aerobic reactions, the bulk soil is anoxic and becomes suitable for anaerobic methanogens but not aerobic methanotrophs. The top 3 mm of soil, however, may be oxic and therefore suitable for aerobic methanotrophs but not anaerobic methanogens (Gilbert and Frenzel, 1998). Despite occupying such a small volume of the soil ecosystem, relative to the anoxic environments, according to Conrad and Rothfuss (1991) this layer acts as a sink for >80 % of the CH_4 that diffuses into the flood water. Any partially oxic layers around plant roots also provide a habitat for aerobic methanotrophs, although root exudates and other organic materials associated with the rhizosphere provide substrates for methanogenesis under anaerobic conditions (Conrad, 2007). As floodwater is drained, CH_4 oxidation extends into the deeper layers (Henckel *et al.*, 2001). As well as oxidising freshly produced CH_4 from flood conditions, methanotrophs can consume atmospheric CH_4 (Conrad, 2007).

There is a multitude of papers that have studied methanogenesis, methane emissions, methanogen populations and microbial ecosystems in paddy fields (e.g. Takai, 1970; Holzapfel-Pschorn, Conrad & Seiler, 1985; Aselmann & Crutzen, 1989; Schütz, Seiler & Conrad, 1989; Mayer & Conrad, 1990; Cai *et al.*, 1997; Conrad *et al.*, 2008; Han *et al.*, 2013; Zhang *et al.*, 2013), however these are focussed on fields located in tropical regions that are either permanently flooded or flooded seasonally, rather than pasture soil that only floods during extreme weather events. Paddy fields also contain plants that allow for a greater exchange of gasses between the rhizosphere and the atmosphere via the aerenchyma (Jackson and Armstrong, 1999). There are no studies that investigate the effects of increased flood duration on microbial ecosystems and methane emissions in pasture soil.

1.2 Metagenomics

1.2.1 Overview

The study of microbial ecosystems provides a valuable insight into the biodiversity of a site and what biotic functions are associated with the community. Relatively little is known about environmental microorganisms however, and studying such ecosystems accurately can prove challenging due to the difficulties associated with culturing species (Tringe *et al.*, 2005); it is estimated that 99.8 % of microbial species in some environments cannot be cultured in the lab (Streit and Schmitz, 2004). A culture-

independent method is therefore required to study such ecosystems reliably, and metagenomics may provide the solution.

Metagenomics is the genetic study of microbial communities sampled directly from the environment (Hugenholtz, 2002; Shah *et al.*, 2011; Thomas *et al.*, 2012). It may provide the answer to two typical questions of a microbial ecology study: "who's there?"; i.e. what species are present, and "what are they doing?"; i.e. what functions are being carried out (Desai *et al.*, 2012; Scholz *et al.*, 2012; Tringe *et al.*, 2005). The term was originally coined by Handelsman *et al.* in 1998. Other terms used to describe the method include: environmental DNA libraries (Stein *et al.*, 1996), recombinant environmental libraries (Courtois *et al.*, 2003), community genome (Tyson *et al.*, 2004), environmental whole genome shotgun sequencing (Venter *et al.*, 2004), plus other less common terms (Riesenfeld *et al.*, 2004). Shotgun metagenomics aims to attain an unbiased sample of microbial community genomes, with as much genomic material sequenced as possible. Obtaining metagenomic sequence data uses DNA sequencing. As it bypasses the culturing procedure, metagenomic studies can reveal species and their functions that were originally unobtainable through culturing (Handelsman, 2004; Hugenholtz, 2002; Riesenfeld *et al.*, 2004). This relatively new insight into microbial ecosystems will develop knowledge and understanding of how communities interact with their environment, and, furthermore, how changes in the environment will affect ecosystems.

As metagenomics moved from being a novel approach for gaining an insight into microbial communities to a fundamental tool for studying ecosystems, there is a requirement for research to incorporate the statistical rigor of ecological studies. Sample replicates should be used to increase the statistical confidence of conclusions drawn from metagenomic studies (Knight *et al.*, 2012; Prosser, 2010). However, this has limitations: microbial communities can be so complex that significant variations exist between communities taken just centimetres apart (Teeling and Glöckner, 2012). Therefore, several replicates are required to reduce the effects of such variation. Cost used to be a prohibitive factor for many users in addressing this issue, however new DNA sequencing technologies have reduced this.

One limit to the application of metagenomics is that only potential function is observed from DNA. To understand actual functional responses from genetic data, metatranscriptomics is required. Metatranscriptomics quantifies the abundances of mRNA and therefore gene expression, allowing for actual microbial functions to be

observed. There are, however, challenges associated with using metatranscriptomics to study complex environmental samples, such as the reliable extraction and storage of mRNA when sampling in remote locations. Carvalhais *et al.* (2012) discuss challenges facing metatranscriptomics as well as highlighting the potential benefits of the field.

The Genomic Standards Consortium (GSC) was established to standardise the description of genomic data (Field *et al.*, 2008). This is because the vast amounts of data produced in genomic and metagenomic projects, the diversity of environments and conditions that the data relate to, and the wide variety of methods available to analyse data can all lead to challenges in analyses and comparing different datasets. Metadata is typically included in metagenomic analyses (Di Bella *et al.*, 2013). This 'data about the data' can include a breadth of information such as sampling time and date, longitude and latitude, depth, temperature and pH. It provides a concise summary of a sample's origin, it allows other users to search for metagenomes based on a variety of factors and it helps determine whether different samples are suitable for comparison.

1.2.2 DNA sequencing

DNA sequencing is the process of determining the sequence of nucleotides in a DNA strand. The first major development in DNA sequencing was made by Sanger *et al.* in 1977, which greatly advanced the capabilities of genetic studies. Sanger sequencing was the major procedure of choice for the next 25 years. With demand for faster and cheaper methods to sequence DNA, new technologies entered the market in the early 2000s, known collectively as High throughput sequencing. The pace that technologies are developed is rapid (Glenn, 2011). High throughput sequencing technologies produce more data at a greater speed and a lower cost than traditional sequencing methods, although each technique has its own advantages and disadvantages. The higher sequencing throughput means a greater coverage of genetic material in environments can be obtained. Rarer species can therefore be revealed and subtler differences between samples can be detected, allowing more accurate and reliable conclusions to be made.

The most popular techniques are summarised in the following sub-sections. Table A.1, extracted from Glenn (2011) and updated with statistics for 2016, provides a comparison of the throughputs and the sequencing costs associated with a number of sequencing technologies available.

The output of several sequencing technologies is a chromatogram with peaks that represent the bases. Base-calling computer programs such as phred (Ewing and Green, 1998; Ewing *et al.*, 1998) convert these chromatograms into files such as FASTQ and SFF that contain read IDs, quality scores, and the genetic sequence for each read. The quality scores refer to phred scores, depending on the base-calling program used; these scores indicate the accuracy of base calling. The equation for calculating phred scores is:

$$Q = -10 \log E$$

where E is the error probability.

1.2.2.1 Sanger sequencing

Sanger sequencing uses dideoxynucleotide triphosphates (ddNTP). These contain a hydrogen group on the 3' carbon, rather than a hydroxyl group (OH) as is found on deoxynucleotide triphosphates (dNTP), to terminate sequencing.

When added to a sequence, ddNTP prevents the addition of further nucleotides because a phosphodiester bond cannot be formed between the hydrogen group and the incoming nucleotide, thus the sequence is terminated. Typical read lengths from Sanger sequencing are approximately 800 base pairs (Tringe and Rubin, 2005; Tringe *et al.*, 2005). Two methods of termination can be used:

Chain-termination:

The DNA is denatured into single strands using heat and amplified by PCR. A primer is annealed to one of the template strands and the strands are added to four tubes: 'A', 'T', 'G' and 'C'. The tubes contain DNA polymerase, all four dNTPs (dATP, dTTP, dGTP, dCTP) and the tubes' respective ddNTP (ddATP, ddTTP, ddGTP and ddCTP). The ddNTPs are at a lower concentration than the dNTPs to promote sequencing. The DNA polymerase binds to the primer and sequencing begins. The dNTPs bind to the template strands and eventually a ddNTP will bind, terminating sequencing for that strand. The result is template strands with sequence strands of different lengths. The DNA strands are denatured into single strands (ssDNA) and run through a gel via electrophoresis to separate the different strand lengths; shorter lengths travel further as they can manoeuvre through the pores more easily. Comparing the reads of the four tubes side-by-side displays the sequence of nucleotides that the strand comprises.

Dye-termination:

In dye-termination, the ddNTPs are fluorescently labelled and all added to the same tube. The same process for chain-termination occurs, except that once the process is complete the gel is exposed to ultra-violet light. This reveals which ddNTP was added at each position via the differences in emitted light wavelengths, and thus which dNTP is complementary for that position.

1.2.2.2 454 Pyrosequencing

Released in 2005 by 454 Life Sciences (Roche Diagnostics Corporation, Basel, Switzerland), 454 pyrosequencing uses light signals released when a nucleotide binds to a template strand to identify the base.

DNA adaptors are ligated to single strand DNA and then attached to beads, one strand per bead. The strands are amplified via emulsion PCR (ePCR) and the beads placed into wells on a plate, one bead per well. A primer is hybridised to the strand and incubated with four enzymes: DNA polymerase, ATP sulfurylase, luciferase, and apyrase; and with two substrates: adenosine 5´ phosphosulfate and luciferin.

A solution containing identical nucleotides is added and if the nucleotide complements the next available base on the template DNA strand then the DNA polymerase binds it, releasing pyrophosphate (PPi). In the presence of adenosine 5´ phosphosulfate, ATP sulfurylase converts the PPi into ATP. The ATP provides energy for luciferase to convert luciferin to oxyluciferin, generating light that is monitored by a charge-coupled device (CCD) camera. The amount of ATP produced, and therefore light generated, is relative to the number of nucleotides bound to the template strand, i.e. the signal for 'AA' would be approximately twice as strong as for 'A'. The apyrase degrades unincorporated nucleotides and remaining ATP, and the process continues using a different nucleotide. If the added nucleotide does not complement the next available nucleotide on the template strand then no light is produced and the free nucleotides are degraded by the apyrase.

Typical read lengths are 400 to 700 bases (454 Life Sciences, a Roche Company, 2012; Ansorge, 2009; Glenn, 2011), making 454 pyrosequencing a suitable platform if long reads are desired. However, the error rates are greater than those from most newer techniques (Di Bella *et al.*, 2013), particularly from homopolymeric sequences.

1.2.2.3 Ion Torrent

Ion Torrent (Thermo Fisher Scientific Inc., Waltham, MA, USA), uses semiconductor technology to detect changes in pH resulting from the sequencing process.

As with 454 pyrosequencing, DNA adaptors are ligated to strands of DNA that are then attached to beads and amplified via ePCR. Primers and DNA polymerase are attached to the strands and the beads are inserted into wells on a chip, one bead per well. Nucleotides are sequentially added to the solution. If a nucleotide is incorporated into a template strand of DNA, a hydrogen ion is released as a by-product, changing the pH of the solution. Beneath the wells is an ion-sensitive layer and beneath that a proprietary ion sensor; the change in pH is detected by the ion sensor, converted to a voltage and subsequently a digital signal that indicates the DNA sequence. If the base is not incorporated then no voltage change is detected. If two bases are incorporated then the voltage change is doubled, and so forth for multiple bases. After each flow of nucleotides, a wash is used to remove remaining nucleotides from the solution before the process continues (Glenn, 2011; Life Technologies, 2012; Rothberg *et al.*, 2011). Read lengths are typically between 150 and 300 base pairs; Table 1.1 displays the published performance figures for the PGM and Proton platforms. The main advantage of the Ion Torrent platform compared to other platforms is short runtime.

1.2.2.4 Illumina sequencing

The Illumina sequencing platform was commercialised in 2006 under Solexa, who were acquired by Illumina (San Diago, CA, USA) in 2007. The technique is based on sequence by synthesis and utilises novel nucleotides with fluorescently labelled terminator groups at the 3'-end of the base, as well as a DNA polymerase capable of binding these to a template strand.

DNA adaptors are ligated to both ends of DNA strands, which are then attached at one end to a surface coated with adaptors and complementary adaptors. The free ends of the immobilised strands hybridise with the complementary adaptors forming a bridge. The template strands are amplified via ePCR to enhance reliability and the complimentary strands of DNA are washed away. Sequencing occurs in both the forward and reverse directions of the template DNA strands, producing paired ends. All four of the fluorescently labelled nucleotides are added to the solution simultaneously and the DNA polymerase binds the complementary base to the template strand. The competition between the bases produced by adding them simultaneously minimises

incorporation bias and associated errors. Once incorporated, a CCD camera detects the terminator group from its fluorescent dye. The group is removed from the base and sequencing continues.

When the paired end reads are paired together, read lengths are increased, accuracy is increased and insertion and deletion (indel) errors are reduced; indels involve either adding an extra nucleotide or skipping a present one during sequencing (Illumina, 2014). Read lengths and throughputs vary depending on the platform and library preparation kit used. The low error rate provides a great benefit for Illumina sequencing, however the short reads produced may produce annotation challenges in downstream analyses, and the runtime is greater than the Ion Torrent platform. The published performance figures for the MiSeq and HiSeq platforms are displayed in Table 1.1.

Table 1.1. Ion Torrent, SOLiD and Illumina performance figures. The performance figures for the Ion Torrent, SOLiD and Illumina platforms stated by the manufactures for various technologies available as of 17th February 2014 (Illumina, 2014; Life Technologies, 2014).

Platform	Platform technology	Runtime (hours)	Throughput (bp)	Sequence count	Read length (bp)
Ion Torrent PGM	Ion 314™ Chip v2 200 reads	2.3	30-50 Mb	400,000-550,000	200
Ion Torrent PGM	Ion 314™ Chip v2 400 reads	3.7	60-100 Mb	400,000-550,000	400
Ion Torrent PGM	Ion 316™ Chip v2 200 reads	3.0	300-600 Mb	2-3 million	200
Ion Torrent PGM	Ion 316™ Chip v2 400 reads	4.9	600 Mb-1 Gb	2-3 million	400
Ion Torrent PGM	Ion 318™ Chip v2 200 reads	4.4	600 Mb-1 Gb	4-5.5 million	200
Ion Torrent PGM	Ion 318™ Chip v2 400 reads	7.3	1.2-2 Gb	4-5.5 million	400
Ion Torrent Proton	Ion PI™ Chip	4.0	10 Gb	60-80 million	200
Illumina MiSeq	V2 Reagent kit	39	7.5-8.5 Gb	12-15 million	2 x 250
Illumina MiSeq	V3 Reagent kit	65	13.2-15 Gb	22-25 million	2 x 300
Illumina HiSeq	V4 Reagent kit	144	1 Tb	4 billion	2 x 125

1.2.2.5 Nanopore sequencing

Nanopore sequencing, first commercialised by Oxford Nanopore Technologies Ltd (Oxford, UK) in 2012, sequences whole DNA strands directly as they pass through a protein nanopore. This vastly increases the read lengths produced, potentially up to 10^4 to 10^6 bases (Feng *et al.*, 2015). Error rates are still worse than Illumina sequencing, although they are rapidly decreasing (Loman and Watson, 2015).

An electrical potential is applied to an electrically resistant membrane that contains a protein nanopore, creating a current of flowing ions through the nanopore. DNA polymerase binds to the DNA and to the nanopore, splitting the strand. The enzyme ratchets the ssDNA through the nanopore one base at a time. The nucleotides disrupt the ion flow, characteristically altering the electrical current and allowing for the identification of the base. An annealed hairpin structure at the end of the DNA allows the complimentary strand to be sequenced immediately after the first strand, improving read accuracy.

The Oxford Nanopore strand sequencing device, the MinION™, is portable in size, comparable to a large USB drive. This removes the need for large desktop machinery and allows for sequencing at field sites.

1.2.3 Amplicons, metagenomes or whole genomes?

DNA sequencing typically takes three forms: targeting specific genes (gene-based amplicons), amplifying untargeted DNA fragments obtained from the metagenome (metagenomics), or shotgun sequencing the sample (Whole Genome Shotgun (WGS) sequencing) to generate whole genomes; each technique has its advantages and limitations.

Gene-based approaches, such as detecting 16S rRNA genes, can accurately identify if target species are present in a sample and give an indication of their abundance (Hugenholtz, 2002; Lane *et al.*, 1985; Woese and Fox, 1977). Conserved genes such as the 16S rRNA gene are used because they are ubiquitous in prokaryotes and they contain both highly conserved regions and variable regions (Hugenholtz, 2002), allowing many species to be studied with a relatively high level of taxonomic accuracy. 16S rRNA gene amplicon sequencing is a popular method for determining the taxonomic composition of a community; it does not require as great of a sequencing coverage than metagenomic sequencing. However, it does have its limitations: the 16S

rRNA gene is approximately 1,550 nucleotides long, therefore most NGS platform will not sequence the whole gene in one fragment and annotation accuracy may be reduced. To overcome this, certain regions of the gene, such as V1, V3 and V4 are often selected for sequencing (Di Bella *et al.*, 2013). Another limitation is the variation in the conservation of the 16S rRNA gene. Some different species, such as *Aeromonas salmonicida* and *A. hydrophila,* contain 16S rRNA genes that are approximately 99 % identical, while other species, such as *Desulfitobacterium hafniense,* contain multiple copies of the 16S rRNA gene that are intragenomically variable by as much as 5 % (Mende *et al.*, 2013). This makes reliably identifying species more challenging.

Other genes, for example Environmental Gene Tags (EGTs) that represent certain functions (Scholz *et al.*, 2012; Tringe and Rubin, 2005; Tringe *et al.*, 2005), can provide an indication of what functions may be occurring in an environment. For example, the methyl coenzyme-M reductase (mcrA) gene is used to identify methanogen species (Luton *et al.*, 2002). This gene-based technique may, however, be biased towards some species. This is due to the unequal amplification of certain genes based on the primers used (Kröber *et al.*, 2009; Shah *et al.*, 2011). Hong *et al.* (2009) reported that PCR primers miss up to half of the rRNA diversity of a highly diverse bacterial community. This bias negates the ability to accurately determine community composition for most environmental samples. Many functions, such as denitrification, are spread across several taxonomic groups, thus by selecting a gene-based approach, specific functions can be identified more accurately than by interrogating taxonomic data.

In contrast to gene-based amplicon sequencing, metagenomics and WGS detects the majority of the genetic material without the same bias towards selected genes. It also ascertains which species and functional genes are present (Allen and Banfield, 2005; Fuhrman, 2012). Using metagenomic sequences provides a multitude of genes that can be used to identify taxa, as opposed to using targeted gene sequences such as 16S rRNA gene sequencing; Shakya *et al.* (2013) found that rRNA marker genes extracted from metagenomic sequences provided a poor determinant of community structure.

While there are several advantages to being able to obtain most of the genetic material from a sample, there are caveats associated with current technologies and methodologies. Currently, most commercially available (and cost/time efficient) sequencing technologies produce short reads, typically ranging between 300 to 1,000 base pairs in length. These short reads can limit the annotation resolution between closely related organisms and functions; for conserved genes this can even occur

across high-level taxa (Shakya *et al.*, 2013). The assignment of genetic material to known species and genes is limited by the extent of reference information available in the databases. This may result in a large portion of unassigned data (Mavromatis *et al.*, 2007). Furthering this, because of the challenges associated with cultivation of the vast majority of microbial species (Streit and Schmitz, 2004), there is a poor representation of many microbial species in the databases (Hugenholtz, 2002). The data available in the databases may also be incorrect. Mavromatis *et al.* (2007) estimate that 10 – 20 % of genes in non-curated metagenomic data sets are inaccurate due to low quality sequences and sequencing errors. The broad sequencing involved in metagenomic projects may also not be deep enough to detect rare species in complex communities (Desai *et al.*, 2012; Fuhrman, 2012; Shah *et al.*, 2011).

Another factor to consider is the cost of sequencing; if the topic of study is identifying the community structure and abundances of target organisms then sequencing for 16S rRNA genes will be cheaper than sequencing whole genomes. Moreover, alternative methods from DNA sequencing may suffice for the needs of the study at a small fraction of the cost. However, with the advances in technology and a greater demand for whole community sequencing, the cost of metagenomic sequencing is now becoming affordable for even small research projects.

1.2.4 Analysis

With the advances in NGS and the reductions in costs, DNA sequencing is no longer the limiting factor in metagenomic studies. Rather, the computational analysis of the vast amounts of data produced is proving to be the limiting factor (Desai *et al.*, 2012; Di Bella *et al.*, 2013; Teeling and Glöckner, 2012). The standard method for deciphering a genetic code is to compare it to a reference nucleotide library such as the National Center for Biotechnology Institute (NCBI) database (NCBI, 2012) using a BLAST (Basic Local Alignment Search Tool) search tool. With the genomic library growing rapidly, BLAST searches are returning more and more results; however, computationally they are intensive and conducting such types of BLAST searches are not practical for analysing large scale metagenomic data without using large, dedicated servers (Desai *et al.*, 2012).

To meet the growing demand for techniques to analyse the large quantities of complex data produced from metagenomic studies, software have been developed that use large external servers. These servers are more computationally powerful and capable of

analysing large datasets quickly and efficiently, compared to most individual computers or laboratory servers. Two popular systems available publicly to analyse metagenomic data are the MG-RAST server (MetaGenomics-Rapid Annotations using Subsystems Technology) (Meyer *et al.*, 2008) and QIIME (Quantitative Insight Into Microbial Ecology) (Caporaso *et al.*, 2010), although others are available such as MEGAN (MEtaGenome Analyser) (Huson *et al.*, 2007), RAPSearch2 (Zhao *et al.*, 2012), PAUDA (Huson and Xie, 2014) and DIAMOND (Buchfink *et al.*, 2015).

The MG-RAST server is web-based and provides an easy-to-use resource that annotates sequence data based on a variety of databases. It allows for taxonomic and functional analyses. The system is modular, meaning that it can be rapidly modified and updated to meet the new demands of such a fluid area of study.

Sequence files (fasta, fastq or SSF) are uploaded to the server along with a Metadata spreadsheet and a text file containing the sample barcodes if necessary; single files containing barcoded sequences from different samples can be demultiplexed and split into separate files using the barcode text provided. Illumina paired-end files can be merged, with unmerged reads either retained or discarded. Users define the maximum number of bases in a sequence with a phred score below a user-determined value for passing a quality filter; sequences with a frequency of bases above this number are discarded. The data is annotated (the process of assigning taxonomic and functional information to sequences), against a selection of databases. MG-RAST includes the M5NR database (M5 non-redundant protein database), a GSC initiative that incorporates a selection of databases into its search (see Chapter 1.3.5).

Users are able to specify the E-value cut-off, the minimum identity cut-off, and the minimum alignment length cut-off to be used for annotation. Respectively, this refers to the probability that sequences have been annotated by chance, the minimum percentage of a sequence required to be matched with a reference sequence, and the minimum base length required for a match. The minimum identity cut-off value will vary depending on the sequences that the user is investigating. For example, if using the highly conserved 16S rRNA gene sequence to identify species, a high identity cut-off would be desirable as much of the sequence is conserved across species, with only a small section being variable. The typical identity cut-off for 16S rRNA gene studies is 97 % (Quince *et al.*, 2008). When studying metagenomes, much of the DNA will be variable between species, and a lower identity cut-off would be desirable. MG-RAST's default identity cut-off is 60 %.

MG-RAST normalises abundance counts for the samples to reduce biases and generate a relative abundance value for comparisons of taxa within and between samples. The phylogenetic structure and relative abundances of samples can be graphically viewed and analysed with a variety of techniques, including producing pie charts, abundance tables, PCoA graphs and other statistical analyses to compare samples. In addition to phylogenetic analyses, functional analyses can be conducted, including producing KEGG maps (Kyoto Encyclopaedia of Genes and Genomes) (Ogata *et al.*, 1999). KEGG is a database used for functional analyses of genes and genomes and one output is a graphical map displaying biochemical pathways. MG-RAST data may be made publicly available if the user wishes, allowing others to analyse them and use them in their studies. It is a requirement of the GSC that the data in publications is publically available, and that the accession numbers used to identify the data are stated.

QIIME analyses amplicon-sequencing data and, unlike MG-RAST, it is a command-line-based program downloaded to a computer or internal server, rather than web-based. It provides users with more control of setting parameters, although it is not suited for metagenomic investigations.

While MG-RAST provides a user-friendly approach to analyse large quantities of data, the method of identifying sequences using a per cent sequence alignment match can, like with BLAST, take several days to complete for a metagenome. Two new programs, One Codex (https://onecodex.com/) and Kraken (Wood and Salzberg, 2014), work by comparing *k*-mers from a sequence, which are sequences of a set length, to a reference database of *k*-mers; the greatest number of 100 % *k*-mer matches determines the classification. According to Wood and Salzberg (2014), Kraken is over 900 times faster than Megablast with similar genus-level sensitivity and precision results; sensitivity refers to the percentage of genera classified and precision refers to the percentage of attempted classifications that are correct. Unlike MG-RAST, One Codex and Kraken only classify the taxonomy of sequences; they do not provide functional information.

One Codex is web-based and users upload sequence files to One Codex servers for annotation. The results are displayed online with a variety of graphs and abundance data can be downloaded for further analyses. Kraken is open-sourced and is available for users to download to their servers.

In WGS, sequences are assembled to produce whole genomes. As NGS techniques produce shorter reads than Sanger sequencing, it may be necessary to assemble

sequences into longer reads to improve annotation accuracy. One *caveat* of assembly is the generation of a bias for more abundant organisms. This is particularly problematic for complex communities (Mavromatis *et al.*, 2007; Teeling and Glöckner, 2012) and thus assembly is not necessarily suitable for many environmental metagenomic projects. Mavromatis *et al.* (2007) found that assembling sequences from a highly complex simulated metagenome had little value, producing only short contigs, and that low-abundance organisms were mostly represented by unassembled reads. Chimeric contigs can also be produced from sequences derived from different organisms (Di Bella *et al.*, 2013), especially from those with closely related genomes (Hallam *et al.*, 2006). Nevertheless, assembled contigs may allow for the annotation of metagenomic sequences that, unassembled, would not be annotated or would be annotated incorrectly. Many programs are available that assemble sequences using a variety of different algorithms, see Miller, Koren & Sutton, (2010) and Teeling and Glöckner (2012) for examples and details.

Being a relatively new field of study, several caveats and pitfalls exist in metagenomic analysis and bioinformatics. These ultimately revolve around sequence annotation accuracy. As mentioned, vast amounts of sequence data can now be produced, but confidence in the ecological inferences is still questionable; while NGS overcomes the coverage issue of previous technologies, the short reads are prone to misidentification. New technologies, such as nanopore sequencing, aim to overcome this. In its current state, however, ecological conclusions drawn in peer-reviewed journals may still need to be considered with some scepticism.

1.2.5 Environmental applications

Microorganisms are critical to the functioning of all ecosystems and they dominate most biogeochemical cycles (e.g. carbon, nitrogen, sulphur) (Falkowski *et al.*, 2008). Knowledge of community structure and functional capacity is crucial to understanding how changes in the environment will impact these communities, functions, and biogeochemical cycles.

Metagenomics can be used to determine how microbial communities and their functions differ between samples. This provides a quantitative insight into how microbial communities vary in different environments. How environmental changes affect communities can also be investigated. Through manipulative experiments, such as applying a pesticide to a sample, the impact of treatment on microbial communities

can be assessed. The application of metagenomics can be used to assess a wide variety of changes, from small-scale localised changes, such as the addition of a specific pollutant, to global scale issues such as climate change.

In their chapter in "Metagenomics: Theory, Methods and Applications", Georga *et al.* (2010) discuss another application of metagenomics: discovering new approaches to bioremediation. This takes the use of metagenomics from quantifying the effects of environmental change to developing techniques for remediating negative effects such as soil contamination. Mokili *et al.* (2012) discuss the future perspectives of using metagenomics to discover viruses, providing yet another useful application of metagenomics.

In addition to surpassing issues related to culturing microorganisms, metagenomics provides a time-effective and accurate method for biomonitoring (Hajibabaei *et al.*, 2011), i.e. using the observation of species richness and abundance to determine the health of an environment. Indices of microbial diversity and species composition provide a sensitive measure of the health of an ecosystem (Kisand *et al.*, 2012), so metagenomic analyses can be used to accurately assess microbial biodiversity. Not only does this provide an indication of ecosystem health, it also provides information about the potential functions conducted by the microorganisms. For example, pristine, healthy sites may be characterised by oligotrophic organisms that are adapted to relatively low nutrient levels, whereas polluted sites may be characterised by specialist organisms genetically adapted to cope with stress factors and abnormally high levels of nutrients (Kisand *et al.*, 2012). Traditional approaches to biomonitoring using eukaryotes (see Rosenberg and Resh, 1993) rely on the morphological identification of species; however, this is time consuming and it is difficult to identify certain species at different stages of their life-cycle, reducing the reliability of the technique. Metagenomics negates this problem and provides a timely method to assess the taxonomic composition of an ecosystem, regardless of life-cycle stages.

Care should still be taken when using metagenomics to study environments. Microbial communities can vary at minute spatial and temporal scales (Farley and Fitter, 1999), thus an adjustment of just a couple millimetres in sample location could significantly affect conclusions drawn. While ensuring an adequate number of replicates are used to help reduce variation in a study, understanding spatial and temporal fluctuations themselves is paramount to developing holistic knowledge of an ecosystem and any responses to environmental change. Once future technologies have addressed the short

read length issue, and annotation databases have been further expanded, the question will not be "is metagenomics suitable for this study?", but "how do we use metagenomics correctly in this study?"

1.2.6 Alternative methods

Metagenomics provides an insight into whole communities, however it is currently an expensive research option and the depth of information produced may not be necessary for certain studies. Other, partial community analyses can provide less detailed indications of the taxonomy and functional capacity of a community, or more specific information about targeted groups of organisms; these methods are cheaper than metagenomic techniques and they may provide more suitable data if a complete insight is not required. Such methods include Denaturing-Gradient Gel Electrophoresis (DGGE), Single-Strand Conformation Polymorphism (SSCP), Random Amplified Polymorphic DNA (RAPD), Amplified Ribosomal DNA Restriction Analysis (ARDRA), Terminal-Restriction Fragment Length Polymorphism (T-RFLP), DNA microarrays, Quantitative-PCR (Q-PCR), Fluorescence *In-Situ* Hybridisation (FISH), and Phospholipid Fatty Acid Analysis (PFAA).

1.3 Overview and aims

The aims of this thesis are to a) evaluate DNA sequence annotation methods for metagenomes and establish a pipeline for studying the effects of environmental change on ecosystems, b) investigate the impacts of flooding frequency on soil microbial communities and their potential functions, and c) investigate the impacts of flood duration on microbial communities, their potential functions, and CO_2 and CH_4 fluxes. Quantifying the abundance of genes in a sample provides an insight into the functional capabilities of a community. By measuring gas fluxes, actual functions can be recorded and correlated with genetic data. The results from a) will be considered when analysing the data produced for b) and c).

Unlike the studies mentioned in section 1.1, which use partial community analysis techniques, this project will use high throughput sequencing of metagenomes to analyse the whole community structure and function. This, along with continuous CO_2 and CH_4 flux measurements, will provide a much deeper insight into the effects of flooding on ecosystems, while also determining the benefits and pitfalls of using metagenomics to assess environmental change. As mentioned in Knight *et al.* (2012), microbial ecology research is shifting from sequencing for taxonomic identification

purposes to using rigorous experimental designs for studying community functions. The aims of this thesis will be achieved by using a simulated metagenome to establish optimum analysis pipelines, which will then be used to study how flooding affects microbial communities and their functions. Carbon dioxide and CH_4 measurements will determine actual functional responses to flooding.

2 Evaluating techniques for metagenome annotation using simulated sequence data.

2.1 Abstract

The advent of high throughput sequencing has allowed huge amounts of DNA sequence data to be produced, advancing the capabilities of microbial ecosystem studies. The current challenge is identifying from which microorganisms and genes the DNA originated. Several tools and databases are available for annotating DNA sequences. The tools, databases and parameters used can have a significant impact on the results: naïve choice of these factors can result in a false representation of community composition and function. We used a simulated metagenome to show how different parameters affect annotation accuracy by evaluating the sequence annotation performances of MEGAN, MG-RAST, One Codex and Megablast. This simulated metagenome allowed the recovery of known organism and function abundances to be quantitatively evaluated, which is not possible for environmental metagenomes. The performance of each program and database varied, e.g. One Codex correctly annotated many sequences at the genus level, whereas MG-RAST RefSeq produced many false positive annotations. This effect decreased as the taxonomic level investigated increased. Selecting more stringent parameters decreases the annotation sensitivity, but increases precision. Ultimately, there is a trade-off between taxonomic resolution and annotation accuracy. These results should be considered when annotating metagenomes and interpreting results from previous studies.

2.2 Introduction

The advent of high throughput sequencing and metagenomics has resulted in increasing numbers of ever-larger datasets describing the community structure and function of a variety of different environments, from the human gut (Arumugam *et al.*, 2011; David *et al.*, 2014) to arctic peat soils (Lipson *et al.*, 2013) and deep-sea vents (Anderson *et al.*, 2015; Xie *et al.*, 2011), to name a few. High throughput sequencing technologies have greatly reduced sequencing costs and speed, and researchers can now affordably study whole microbial communities and functions. Prior to this, the focus was on community species composition, studied using 16S rRNA targeted gene amplicon sequencing. Amplicon sequencing does not require the DNA coverage that metagenomic studies require and can accurately identify which species are present in a sample (Hugenholtz, 2002; Lane *et al.*, 1985; Woese and Fox, 1977), but it does not

provide the depth of information, such as gene function, that full metagenome sequencing and annotation provides. Cost is no longer the primary limiting factor for undertaking metagenomic studies; rather it is now bioinformatics and the processing power required to process the data produced. Illumina's HiSeq platform, for example, can affordably sequence the most complex of microbial communities, and the challenge now is to interpret the data produced.

Henry *et al.* (2014) provide an extensive directory of tools available for different tasks involved in a metagenomic project pipeline, related to a range of 'omics' studies. These may include bespoke bioinformatic pipelines, downloadable programs and web-based services. MEGAN (Huson *et al.*, 2007) is a popular Graphical User Interface program for analysing and visualising BLAST results to study the taxonomy of microbial communities. While MEGAN typically analyses BLAST results in a few minutes, running BLAST searches against reference sequences in a database is computationally intensive and slow for metagenomes. Web-based servers are increasingly popular for processing large amounts of data. With an intuitive web interface and a variety of analytical tools to choose from, MG-RAST (Meyer *et al.*, 2008) is increasingly cited. MG-RAST allows users to upload raw sequence files that are processed through quality filters and annotated using a selection of user-defined parameters, such as reference databases, minimum identity cut-off values, maximum E-values, or expect-values, and minimum alignment lengths. Details of the processing procedure can be found in the MG-RAST Technical Report (Wilke *et al.*, 2013).

In response to the growing size of data sequenced, faster alignment methods are being produced. RAPSearch2 (Zhao *et al.*, 2012) translates nucleotide sequences and aligns them with annotated protein sequences, reporting to be c. 100 fold faster than BLASTX with only a 1.3-3.2 % reduction in sensitivity (the proportion of sequences annotated). With "accelerate" mode, the speed increase is up to 1,000 fold. PAUDA (Huson and Xie, 2014) uses a similar approach and claims to be 10,000 fold faster than BLASTX, although with a significant reduction in sensitivity. DIAMOND (Buchfink *et al.*, 2015) purports to be both fast and accurate, with a 20,000 fold increase in processing speed compared to BLASTX. In sensitive mode, 99 % of sequences are aligned, with a speed increase of 2,000 fold compared to BLASTX. Like BLAST with Megablast, RAPSearch2 and DIAMOND offer fast and sensitive modes, each coming at the cost of the other. The outputs from both programs can be viewed and analysed using MEGAN.

One Codex is a web-based program that uses a different technique to BLAST and MG-RAST to classify sequences (https://onecodex.com/). The program designers report that it runs 900 times faster than BLAST while maintaining similar genus-level sensitivity and precision (the proportion of annotated sequences that are correctly identified), taking hours rather than days to classify most metagenomes. One Codex works by comparing *k*-mers (sequences of a set length) from a sequence to a reference database of *k*-mers; the greatest number of 100 % *k*-mer matches determines the classification. BLAST and MG-RAST classify sequences by matching them with the most similar sequences in a database. Unlike MG-RAST, One Codex does not annotate genes for function.

The choice of database, minimum identity cut-off value (i.e. sequence match stringency), minimum alignment length cut-off value and minimum E-value limit (the probability a match has occurred by chance) all influence sequence annotation accuracy, which, in turn, affect the reproducibility and interpretation of the data. An inherent issue with metagenomic studies is that establishing the accuracy of sequence annotation for environmental samples is practically impossible, given that the quantities of organisms and genes are unknown. Therefore, determining the most effective annotation method is fundamental to investigating environmental communities with confidence.

2.2.1 Databases

There are a variety of different reference nucleotide and amino acid databases available for annotating gene or protein sequences (Table A.2). The M5NR database (Wilke *et al.*, 2012) incorporates information from a selection of different databases (Table A.2), increasing the amount of reference data available for annotation. Using a single reference database may be the best option in some cases, for example where gene amplicons are used as a method to identify taxa, rather than other genes.

Whereas taxonomic nomenclature is universal, governed by international conventions, there are multiple approaches for functional classification. Two popular methods include Clusters of Orthologous Groups (COG) (Tatusov *et al.*, 1997) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000). COGs comprise orthologous functions that allow for functional description of poorly characterised genomes based on protein orthologs. KEGG provides a reference database of sequences with functional pathway annotations. Both methods include a

hierarchy of functional descriptions. At the highest level, COG descriptions are characterised under: Cellular processes, information storage and processing, metabolism, and poorly characterised. KEGG descriptions are characterised under: cellular processes, environmental information processing, genetic information processing, human diseases, and metabolism. Due to the differences in characterisation approaches, COG and KEGG annotations cannot be compared directly. COG is currently freely available. KEGG operates on a subscription basis, and MG-RAST uses the latest freely available version (updated in 2008).

2.2.2 Parameters

Selecting a minimum identity cut-off value for metagenome analysis is challenging because interspecific sequence identity varies among genes. Too high a value will accurately identify genes with highly conserved regions, such as the 16S rRNA gene or highly conserved coding genes with little synonymous substitution, but may fail to identify genes or non-coding regions that are highly variable. Conversely, a value too low will allow for highly variable genes to be identified, but may also incorrectly identify an organism/function, thus providing false community/function profiles.

The optimum identity cut-off point for species identification using the 16S rRNA gene is widely accepted as 97 % (Stackebrandt and Goebel, 1994; Rosselló-Mora and Amann, 2001; Chun *et al.*, 2007; Richter and Rosselló-Móra, 2009; Větrovský and Baldrian, 2013; Mende *et al.*, 2013), although this value has its limitations. Some species, such as certain *Rickettsia spp.*, have a 16S rRNA gene similarity greater than 97 %, thus a cut-off value at this level would not differentiate between the species (Fournier *et al.*, 2003). Stackebrandt and Goebel (1994) suggest that a higher value may be more appropriate, but fewer sequences would be annotated due to sequencing errors and sequence mutations. Typically, lower cut-off values are suitable for metagenomic studies as the multitude of genes that contain varying degrees of conservation are sequenced. The default value used by MG-RAST, and used in many metagenomic studies (e.g. Tatusov, Koonin and Lipman 1997; Lipson *et al.* 2013), is 60 %, as this allows for identification using less conserved genes and non-coding regions.

Minimum alignment lengths set the minimum length of sequence considered for annotation. A lower value allows shorter sequences to be annotated, although the chance of incorrectly annotating a shorter sequence is higher. A higher value will reduce this chance, but may also reduce the number of annotations overall. Combining

a low minimum alignment length with a strict minimum identity cut-off value allows shorter sequences to be annotated but with a high match criteria.

Setting maximum E-values and minimum alignment lengths allows stringency of annotations to be controlled. E-values denote the maximum probability that a sequence annotation has occurred by chance. Lower maximum E-values will reduce the number of possible incorrect annotations, although this also reduces number of annotations retained for analysis. The default maximum E-value used by MG-RAST is $1\text{-}e^{-5}$.

2.2.3 Aims

The aim of this study is to evaluate the accuracy of MEGAN, MG-RAST and One Codex annotation methods while investigating how using different databases and parameters impact the annotation of metagenomes. To do this, a novel simulated metagenome was generated using the NCBI whole bacterial genome database and annotated using each pipeline and, for MG-RAST, with different reference databases, minimum identity cut-off values, minimum alignment lengths and maximum E-values.

Using a simulated metagenome comprising known genome abundances allows the accuracy of annotation to be quantified. The simulated metagenome was also annotated using Megablast, a faster variation of BLAST, to provide a control and so that MEGAN, MG-RAST and One Codex could be compared to a standard in sequence annotation. Comparing the MEGAN, MG-RAST and One Codex annotations to the Megablast annotations will quantify the accuracy of these programs for annotating sequences from organisms whose genomes are stored in the NCBI databases.

2.3 Methodology

2.3.1 Metagenome simulation

A simulated metagenome, hereafter Simmet, was created using NeSSM (Jia *et al.*, 2013), comprising the complete NCBI bacterial genome database (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/all.fna.tar.gz, May 2013 collection, accessed on 29/04/14). NeSSM creates synthetic metagenomes from input genomes based on user-defined parameters (e.g. sequence count, length and abundance distributions) that aim to simulate real sequencing data, including expected sequencing errors (i.e. substitutions, insertions, and deletions) based on the chosen sequencing

technology simulated (see "Step II: error models and sequencing coverage bias estimation in Jia *et al.* (2013)). 2,400,000 sequences with a read length of 450 base pairs were designated for simulation, based on 454 pyrosequencing.

One strain for each of the 1,505 species in the NCBI bacterial genome database was randomly selected to be included in the simulation because certain species, e.g. model organisms and human pathogens such as *Escherichia coli*, *Salmonella enterica*, *Mycobacterium tuberculosis*, *Bacillus cereus* and *Staphylococcus aureus*, have been extensively studied and are over-represented in the databases. The resulting genus richness was 688. The species abundance distribution used for simulation was derived from the abundance distribution of a pasture soil metagenome (sequence count: 2,378,586, MG-RAST ID: 4554767.3) (See Equation 1).

Equation 1.

$$y = -2490ln(x) + 19748$$

where x is the randomly selected species rank.

The sequences were processed with Sickle (Joshi and Fass, 2011) to trim low quality ends, with the average threshold phred score set at 20 (a base call error rate of 1 %).

2.3.2 Analysis

The Simmet metagenome file was annotated with Megablast (available from: http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastDocs&DOC_TYPE=Download) as a control, using a reference database of the genomes used to create Simmet. This quantifies the effect that the simulated sequencing errors have on the annotations. The NCBI nucleotide database (updated 17/11/14) (ftp://ftp.ncbi.nlm.nih.gov/blast/db/) was also used to assess the annotation performance of Megablast. The maximum E-value selected was $1\text{-}e^{-5}$ and the minimum alignment length, 15 bases. Megablast annotations using the Simmet database will be referred to as "control" and those using the NCBI nucleotide database will be referred to as "Megablast". The BLAST results were uploaded to MEGAN (version 5.2.3) and analysed using the same parameters used in the BLAST.

Simmet was uploaded to MG-RAST and One Codex. The databases investigated within MG-RAST were: GenBank, GreenGenes, RDP, RefSeq, SEED, SwissProt and TrEMBL. The M5NR and M5RNA databases were excluded from individual sequence analysis, as

individual sequence annotations were not available for download from MG-RAST for these databases. RefSeq was used for One Codex. For both the Megablast and MG-RAST annotations, which use a minimum sequence alignment match to annotate sequences, the minimum identity cut-off values tested were: 40 %, 50 %, 60 %, 70 %, 80 %, 90 %, 95 % and 97 %. The minimum alignment lengths tested were: 10, 15, 20, 25, 30, 25, 40, 45, 50, 55 and 60 bp. The maximum E-values tested were: $1\text{-}e^{-1}$, $1\text{-}e^{-5}$, $1\text{-}e^{-10}$ and $1\text{-}e^{-15}$. Aside from testing, default parameters were used: 60 %, 15 bp and $1\text{-}e^{-5}$, respectively, for minimum identify cut-off, minimum alignment length and maximum E-value.

The sequence IDs and annotations were extracted from the Megablast results (https://github.com/sandyjmacdonald/blast_parser) and full taxonomic lineages were generated for each sequence using the NCBI taxonomy database (available from: ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/; NCBI database version generated 13/05/2013). Species level was excluded from analysis due to the high variation in annotated species nomenclature and the accepted caveats associated with microbial species classification (Achtman and Wagner, 2008; Gevers *et al.*, 2005), e.g. horizontal gene transfer (Bapteste and Boucher, 2009; Gogarten and Townsend, 2005). Discrepancies identified between databases for organism names were corrected for, such as NCBI using the old name Chloroflexia (as of 17/11/14) and MG-RAST using the new name Chloroflexi for the same class. Those that were not annotated were named "Unidentified" and those that were annotated but were either ambiguously annotated or not annotated at all taxonomic levels had the corresponding levels in the lineage replaced with "Unclassified". For MEGAN and One Codex, NCBI taxa IDs were used to generate the lineages.

The taxonomic lineage for each annotated sequence was compared to the lineage for the corresponding source sequence in Simmet to determine the annotation sensitivity and precision at each taxonomic level. The effect that minimum identity cut-off values, minimum alignment lengths and maximum E-values had on annotation sensitivity and precision were established using Megablast and MG-RAST. The correlations between the relative abundances for each taxon in Simmet and in the annotations were calculated using Pearson's product moment correlation coefficient. Domain was excluded due to the small number of taxa. The natural logarithms of the relative abundance values were calculated for plotting, as the original distributions would not visually convey the variations in low abundance taxa. The taxa richness values for each taxonomic level were calculated.

Unlike investigating taxa, correct functional annotations cannot be ascertained with 100 % confidence. To investigate functional annotation performance, protein sequences associated with the sequences in Simmet were extracted from GenBank records and annotated using the KEGG Automatic Annotation Server (KAAS) (Moriya *et al.*, 2007) and WebMGA (Wu *et al.*, 2011). Both are web-based functional annotation tools independent of those investigated in this study. They did not contain sequencing errors and thus provided the best possible indication of the functional annotation accuracy, although the caveats associated with sequence annotation (e.g. possibly incorrectly assigning a function) are present.

KEGG Orthology (KO) and COG IDs were extracted from the KASS and WebMGA results, respectively, for each sequence annotated and compared with the IDs assigned by MG-RAST. The parameters set for the taxonomic investigation were used, and the minimum identity cut-off values investigated were: 40 %, 50 %, 60 %, 70 %, 80 %, 90 % and 95 %. The minimum alignment lengths tested were: 10, 15, 20, 25, 30, 25, 40, 45, 50, 55 and 60 base pairs. The maximum E-values tested were: $1\text{-}e^{-1}$, $1\text{-}e^{-5}$, $1\text{-}e^{-10}$ and $1\text{-}e^{-15}$.

2.4 Results

2.4.1 Simulation and annotation

NeSSM produced 2,399,077 sequences (length range: 195 to 459 bp, median length 377 bp). The average phred quality scores remained above 20 until beyond 400 bases (Figure A.1) and 98.7 % of sequences are between 300 and 400 base pairs long (Figure A.2). Of the 2,399,077 sequences, KASS annotated 1,341,362 (55.9 %) sequences and WebMGA annotated 1,945,674 sequences (81.1 %).

2.4.2 Parameters (Blast and MG-RAST)

More stringent parameter values resulted in fewer sequence annotations but had a greater precision; lower values resulted in more annotations being made, but these comprised increases in both correct and incorrect annotations. For example, with cut-off values of 95 % and 40 %, MG-RAST RefSeq annotated 40.2 % and 90.3 % sequences, respectively with incorrect annotation rates of 2.9 % and 34.5 % at the genus level. This was observed for all parameters tested and for both taxonomic and functional annotations (Figures 2.1-2.6). As the taxonomic level moved up the taxonomic hierarchy, more sequences were correctly annotated (e.g. 0.5 % and 10.2 % for MG-

RAST RefSeq with cut-off values of 95 % and 40 %, respectively, at the class level). Note that sensitivity is unaffected by the taxonomic level investigated.

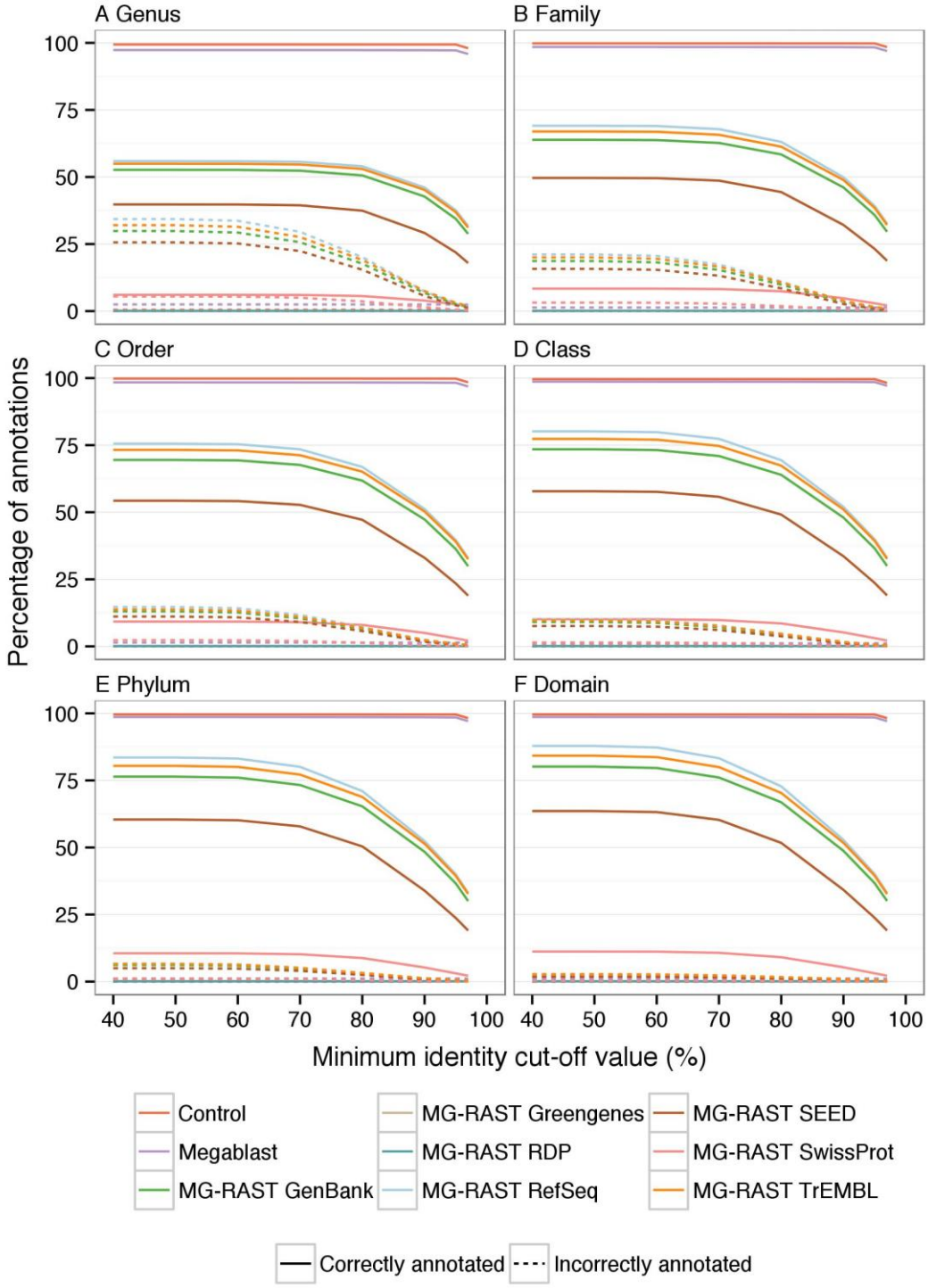


Figure 2.1. Effect of minimum identity cut-off values on taxonomic annotation. The effect of changing minimum identity cut-off value on the number of sequences correctly and incorrectly annotated across the taxonomic levels.

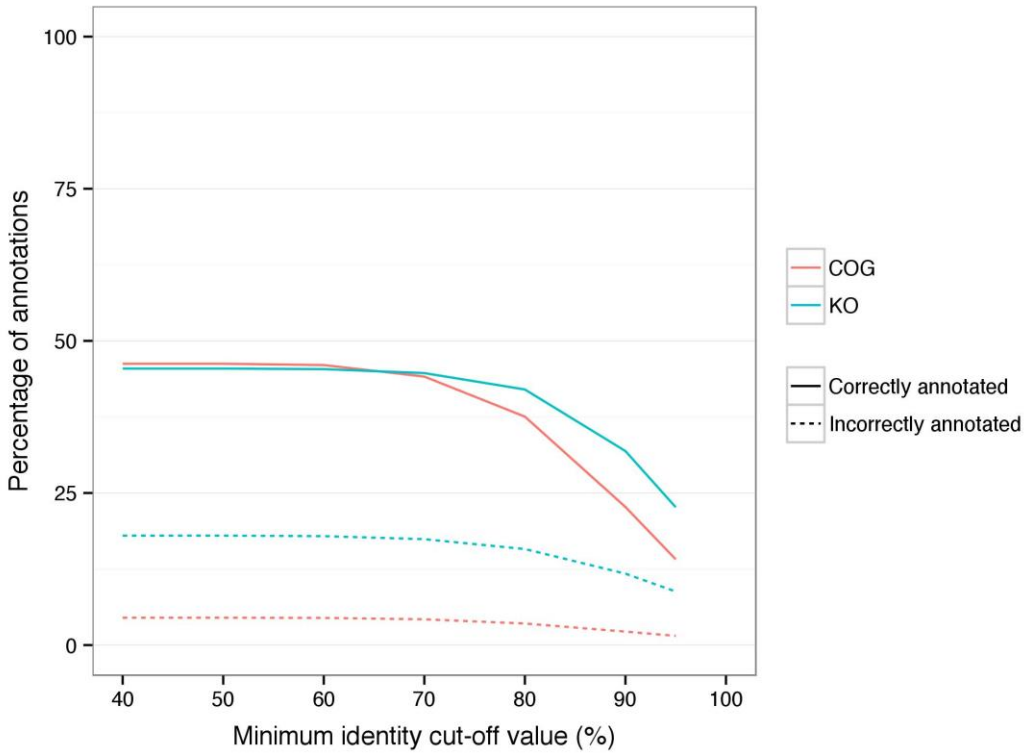


Figure 2.2. Effect of minimum identity cut-off values on functional annotation. The effect of changing minimum identity cut-off value on the number of sequences correctly and incorrectly annotated for functions.

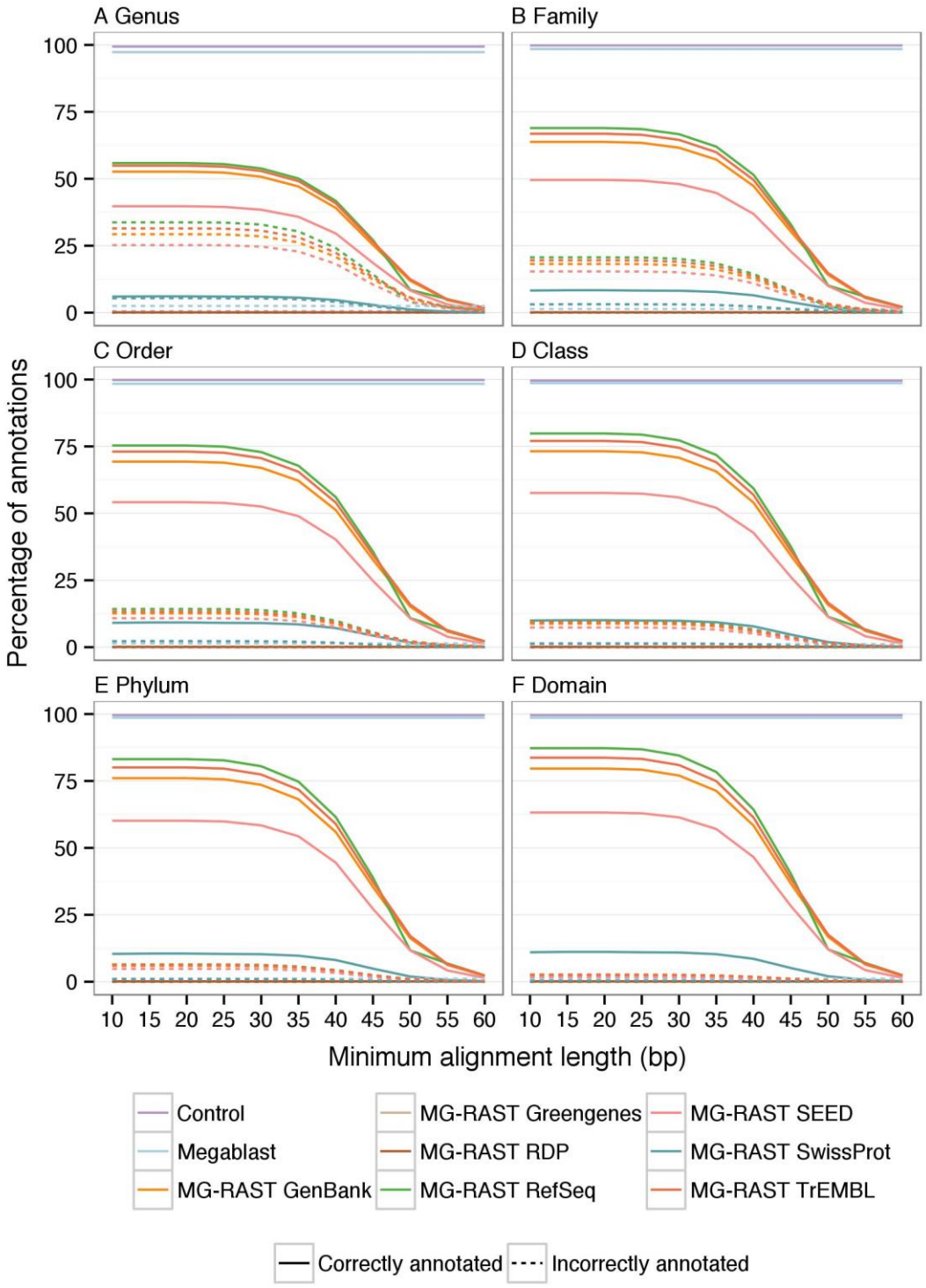


Figure 2.3. Effect of minimum alignment length on taxonomic annotation. The effect of changing minimum alignment length on the number of sequences correctly and incorrectly annotated across the taxonomic levels.

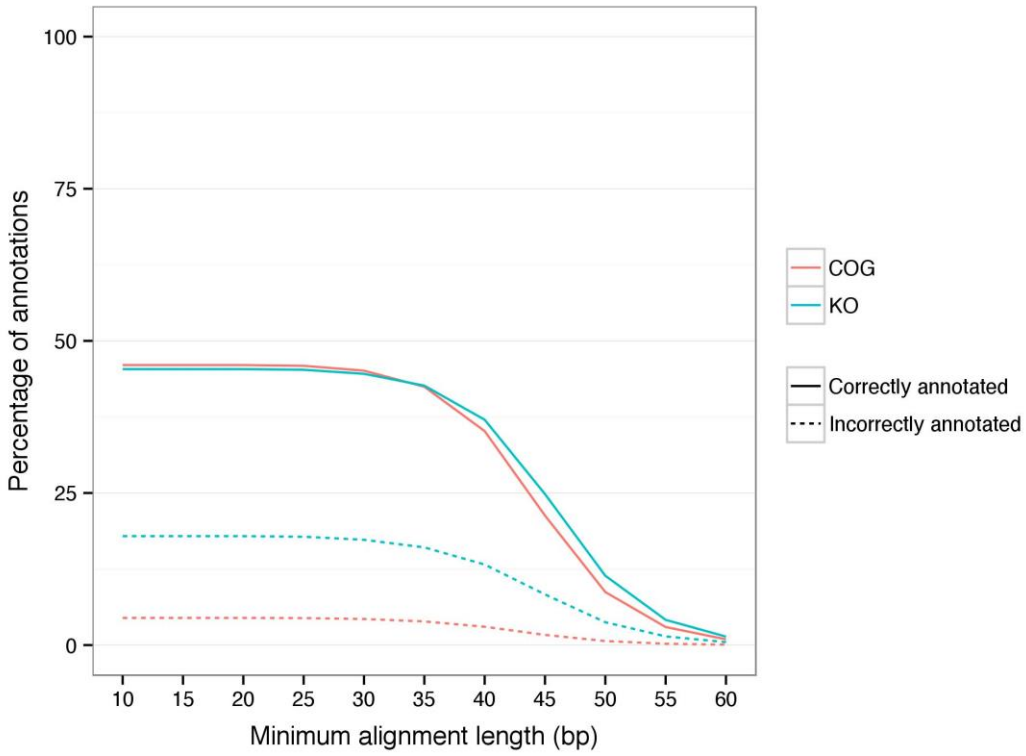


Figure 2.4. Effect of minimum alignment length on functional annotation. The effect of changing minimum alignment length on the number of sequences correctly and incorrectly annotated for functions.

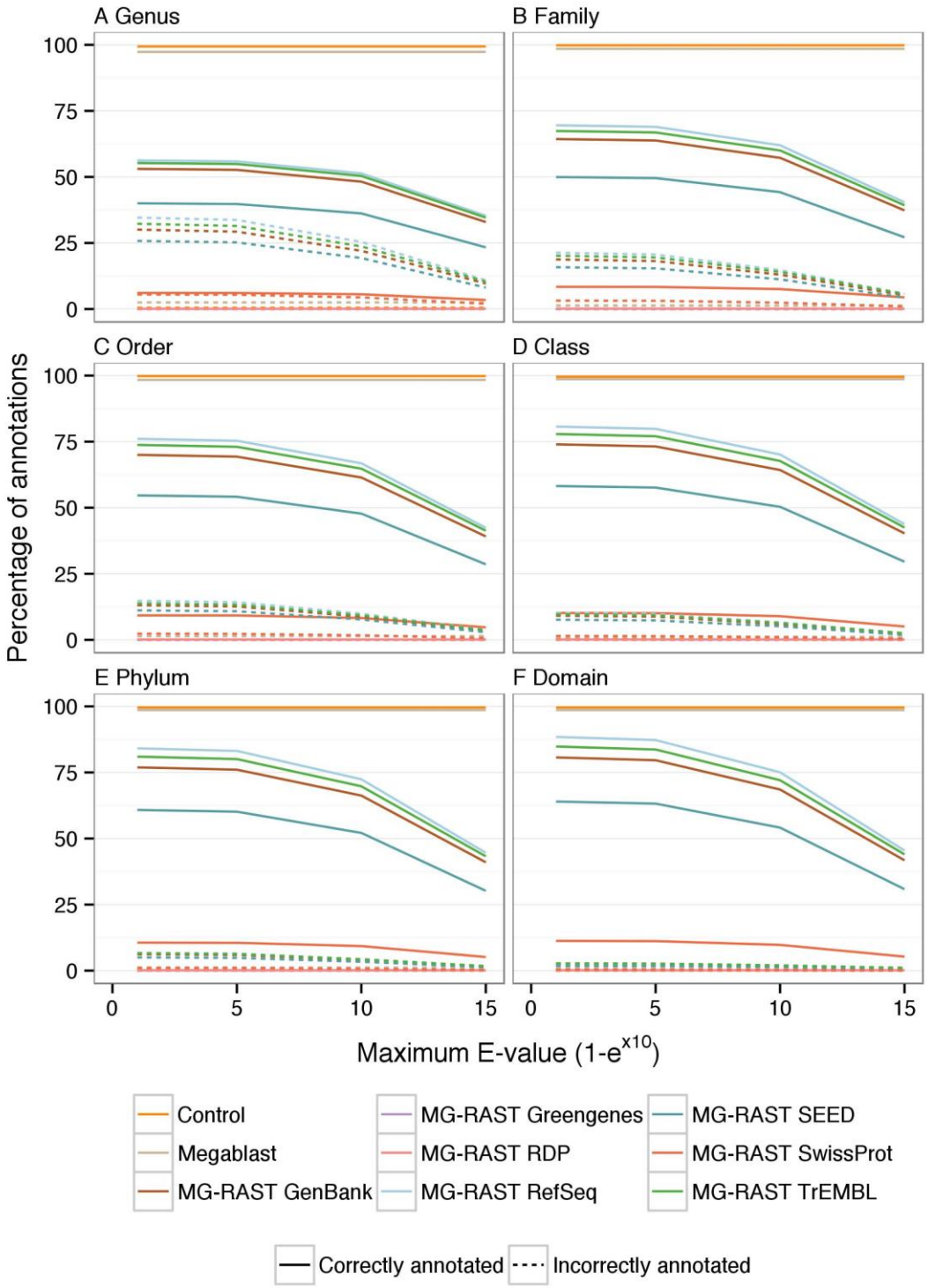


Figure 2.5. Effect of maximum E-value on taxonomic annotation. The effect of changing maximum E-value on the number of sequences correctly and incorrectly annotated across the taxonomic levels.

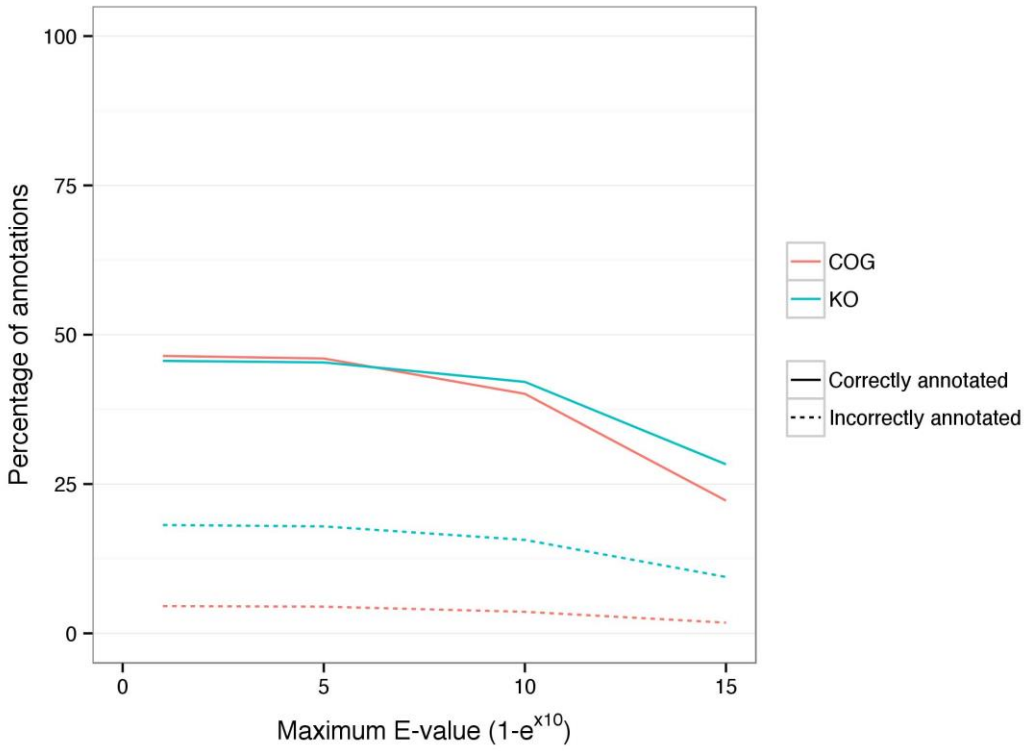


Figure 2.6. Effect of maximum E-value on functional annotation. The effect of changing maximum E-value value on the number of sequences correctly and incorrectly annotated for functions.

The correlation coefficients between the taxa relative abundances in Simmet and in the annotations decreased as parameter stringency increased (Figure 2.7, associated scatter plots in Figures A.3-A.5). Most databases achieved maximum correlations with a minimum identity cut-off value of 50 %, a minimum alignment length of 30 bp and a maximum E-value of $1\text{-}e^{-1}$. Greater decreases in correlation coefficients occurred with a minimum identity cut-off value above 70 % and a minimum alignment length greater than 40 bp.

Figure 2.7. Abundance correlations for different parameter values. The Pearson's product-moment correlation coefficients for the correlations between the Genus relative abundances from Simmet and those from various annotation methods using different A) minimum identity cut-off values, B) Minimum alignment lengths, and C) maximum E-values.

2.4.3 Annotation sensitivity and precision

The control taxonomically annotated 99.9 % of sequences and had a genus precision of 99.5 %. This produced the greatest number of correct annotations (99.4 %) (Table 2.1, Figure 2.8, Table A.3 for all taxonomic levels). Megablast annotated 99.8 % of sequences and had a genus precision of 97.5 %. One Codex annotated all of the sequences, but incorrectly annotated more sequences (5.8 %) than MEGAN (2.9 %), Megablast (2.5 %) and the control (0.5 %). Megablast, MEGAN and One Codex correctly annotated 97.3 %, 95.7 % and 94.2 % sequences respectively, significantly more than the next most successful methods: MG-RAST RefSeq (55.9 %), MG-RAST TrEMBL (54.9 %) and MG-RAST GenBank (52.7 %). MG-RAST RDP and MG-RAST Greengenes, both rRNA databases, annotated less than 1 % of the sequences. This is consistent with the expected frequency of rRNA genes within bacterial genomes (Větrovský and Baldrian, 2013). As the taxonomic level increases, precision increases and becomes more similar across the different databases.

MG-RAST KEGG annotated 63.3 % of the sequences and had a precision of 71.7 %, with 45.4 % of sequences correctly assigned a function and 17.9 % incorrectly assigned a function. MG-RAST COG annotated 50.5 % of the sequences and had a precision of 91.1 %, resulting in 46.0 % of sequence being correctly assigned a function and 4.5 % being incorrectly assigned a function (Tables A.4-A.6). The portions of sequences correctly annotated by both methods were 81.5 % for MG-RAST KEGG and 55.4 % for MG-RAST COG.

Table 2.1. Taxonomic annotation statistics. The Simmet taxonomic annotation statistics for each method and database at the Genus level using default parameters.

Method	Database	Sensitivity (%)	Correctly annotated (%)	Incorrectly annotated (%)
Megablast	Control	99.88	99.39	0.49
Megablast	NCBI	99.81	97.32	2.49
MEGAN	MEGAN	98.56	95.65	2.91
MG-RAST	GenBank	81.94	52.65	29.30
MG-RAST	Greengenes	0.11	0.08	0.03
MG-RAST	RDP	0.13	0.10	0.03
MG-RAST	RefSeq	89.58	55.90	33.68
MG-RAST	SEED	64.97	39.75	25.22
MG-RAST	SwissProt	11.49	6.08	5.42
MG-RAST	TrEMBL	86.37	54.93	31.44
One Codex	One Codex	100.00	94.18	5.82

Figure 2.8. Annotation performance. The annotation sensitivity and number of sequences correctly annotated from a variety of methods and databases across the taxonomic levels investigated.

2.4.4 Taxa abundance correlations

The control showed the greatest genus-level correlation ($r^2 = 0.98$). Megablast had the greatest genus-level correlation with Simmet after the control ($r^2 = 0.95$), MG-RAST SEED had the weakest ($r^2 = 0.49$). MEGAN and One Codex had genus-level correlations of $r^2 = 0.90$ and $r^2 = 0.93$, respectively. The greatest correlation achieved, aside from the control, was by Megablast at the phylum level ($r^2 > 0.99$) (Figures 2.9 & A.6).

MG-RAST M5NR and MG-RAST RefSeq generated 87 and 56 false positive class identifications, respectively (Table 2.2). MEGAN had only two false positive class identifications ("Unidentified" and Insecta) and one false negative identification (Solibacteres). One Codex also had a low abundance of false positive class identifications (eight) and no false negative class identifications. Classes with many false positive identifications include eukaryotes, particularly fungi, and bacteria such as Spartobacteria. The greatest fold differences for classes can be found in Table A.7.

Table 2.2. False positive and negative Class relative abundances. The false positive and negative Classes from MG-RAST M5NR, RefSeq, One Codex and MEGAN annotations.

Class	Simmet	Annotation
A.1 The top 10 false positive Classes for MG-RAST M5NR		
Erysipelotrichi	0.0	0.00109
Dehalococcoidetes	0.0	0.00100
Ktedonobacteria	0.0	0.00014
Spartobacteria	0.0	0.00013
Mammalia	0.0	0.00012
Insecta	0.0	0.00011
Eurotiomycetes	0.0	0.00009
Sordariomycetes	0.0	0.00009
Saccharomycetes	0.0	0.00008
Liliopsida	0.0	0.00007
A.2 The top 10 false positive Classes for MG-RAST RefSeq		
Spartobacteria	0.0	0.00011
Ktedonobacteria	0.0	0.00010
Insecta	0.0	0.00009
Eurotiomycetes	0.0	0.00008
Mammalia	0.0	0.00007

Class	Simmet	Annotation
Saccharomycetes	0.0	0.00007
Lentisphaeria	0.0	0.00007
Anthozoa	0.0	0.00005
Amphibia	0.0	0.00005
Zetaproteobacteria	0.0	0.00005
A.3 The false positive Classes for One Codex		
Sordariomycetes	0.0	0.00001
Holophagae	0.0	<0.00000
Ktedonobacteria	0.0	<0.00000
Eurotiomycetes	0.0	<0.00000
Leotiomycetes	0.0	<0.00000
Dothideomycetes	0.0	<0.00000
Nitrospinia	0.0	<0.00000
Saccharomycetes	0.0	<0.00000
A.4 The false positive Classes for MEGAN		
Insecta	0.0	0.00042
B.1 The false negative Classes for MG-RAST M5NR		
Dehalococcoidia	0.00122	0.0
Ignavibacteria	0.00059	0.0
Erysipelotrichia	0.00057	0.0
Chthonomonadetes	0.00055	0.0
Phycisphaerae	0.00049	0.0
Caldilineae	0.00045	0.0
Caldisericia	0.00019	0.0
B.2 The false negative Classes for MG-RAST RefSeq		
Anaerolineae	0.00127	0.0
Ignavibacteria	0.00059	0.0
Chthonomonadetes	0.00055	0.0
Phycisphaerae	0.00049	0.0
Caldilineae	0.00045	0.0
Thermodesulfobacteria	0.00040	0.0
Caldisericia	0.00019	0.0
B.3 The false negative Classes for One Codex		
NA		

Class	Simmet	Annotation
B.4 The false negative Classes for MEGAN		
Solibacteres	0.00111	0.0

Figure 2.9. Abundance correlations for different taxonomic levels. The Pearson's product-moment correlation coefficients for the correlations between the relative abundances from Simmet and those from the annotation methods.

2.4.5 Taxa richness

Six of the annotation methods underestimated the genus richness and six overestimated it (Table 2.3, Figure 2.10). The control perfectly estimated the genus richness. The next closest estimate was achieved by MEGAN (97.7 %), followed by MG-RAST SwissProt (95.5 %), MG-RAST M5RNA (95.2 %), Megablast (110.2 %) and MG-RAST RefSeq (118.2 %). MG-RAST M5NR produced the most incorrect richness value at 1,244 genera (180.8 %). One Codex overstated the genus richness by 26.7 %. The methods were inconsistent in response to taxonomic level. With increasing taxonomic level some estimates increased in accuracy while others decreased (Figure 2.10, Table A.8). Excluding the control and the domain level, where the number of taxa is low, MEGAN achieved the most accurate richness value (101.2 %) at the family level. MG-RAST M5NR achieved the most inaccurate richness value (253.3 %) at the order level. Megablast and One Codex achieved accurate results relative to other methods, but they still overstated taxa richness at every taxonomic level.

Table 2.3. Genus richness. The genus richness estimates and the differences from Simmet for each annotation method. Due to the low numbers, Domain is excluded from comparisons. Richness values at all taxonomic levels can be found in Table A.8.

Method	Database	Richness	Difference (%)
Simmet	N/A	688	N/A
Megablast	Control	688	100.00
Megablast	Megablast	758	110.17
MEGAN	MEGAN	672	97.67
MG-RAST	GenBank	1,090	158.43
MG-RAST	Greengenes	404	58.72
MG-RAST	M5NR	1,245	180.96
MG-RAST	M5RNA	655	95.20
MG-RAST	RDP	469	68.17
MG-RAST	RefSeq	813	118.17
MG-RAST	SEED	445	64.68
MG-RAST	SwissProt	657	95.49
MG-RAST	TrEMBL	1,094	159.01
One Codex	One Codex	872	126.74

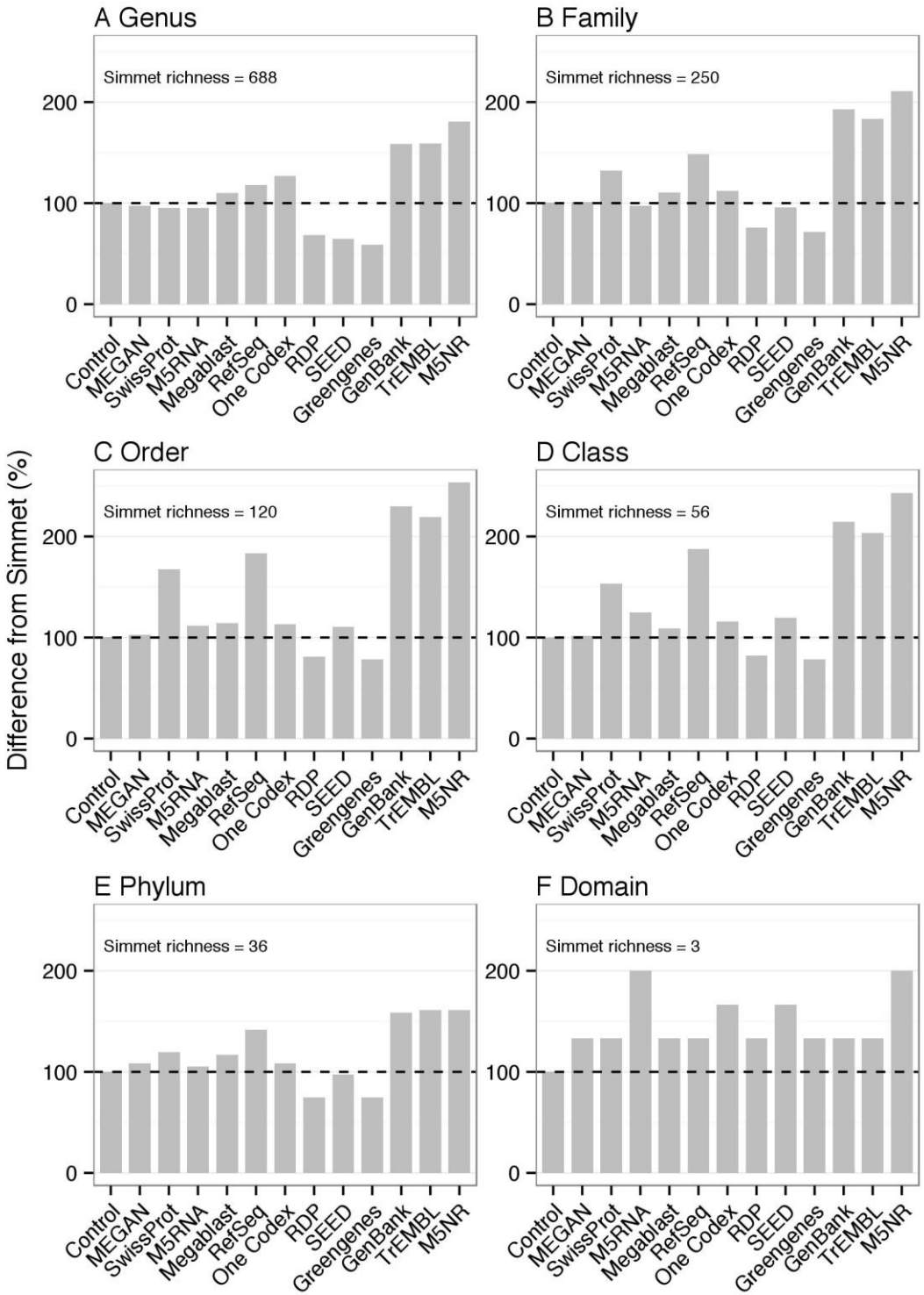


Figure 2.10. Taxa richness. The differences between annotated richness values and the actual richness value (dashed line) for each taxonomic level.

2.5 Discussion

In this study the performances of MEGAN, MG-RAST, One Codex and Megablast were evaluated by determining their sequence annotation accuracies. All common taxonomic levels above species are studied, building on the work by Lindgreen *et al.*, (2016) who study several tools at the genus and phylum levels. A guideline for researchers to establish the annotation accuracy costs of investigating different taxonomic levels is provided, allowing them to optimise their investigations depending on their requirements for taxonomic resolution. MG-RAST and Megablast use a selection of parameters to determine the stringency of matching a sequence with a reference sequence in a database. Less stringent parameters (i.e. lower minimum identity cut-off values, lower minimum alignment lengths and higher maximum E-values) annotate more sequences, but more incorrect annotations are made, thus producing an incorrect community profile. More stringent parameters reduce the number of incorrect annotations, but many fewer annotations are made, resulting in much of the data being rejected. Shakya *et al.*, (2013) drew similar conclusions for varying minimum identity cut-off values. Decreases in sensitivity generally occur from minimum identify cut-off values above 60 %, minimum alignment lengths greater than 30 bp or maximum E-values below $1\text{-}e^{-5}$; therefore the default values used by MG-RAST maximise sensitivity. According to Carr and Borenstein (2014), the impact of parameters such as E-value will vary depending on read-length, something that should be considered in future evaluations as newer sequencing technologies produce longer reads (e.g. nanopore sequencing (Branton *et al.*, 2008)). The sensitivities and the number of sequences correctly annotated are relatively low for MG-RAST at the genus and family levels. At the order level the values are higher, suggesting that this would be the optimum taxonomic level to study, which maximises the amount of data used without producing too many incorrect annotations. Ultimately, there is a trade-off between taxonomic resolution and annotation accuracy, and this must be considered when determining methods for metagenomic studies.

A marginal number of sequences were not annotated by the control and an even smaller number were incorrectly annotated. These discrepancies are due to the sequencing errors inserted into the simulation. It is therefore concluded that 0.5 % of inter-sample difference at the genus level may be attributed to sequencing error, an important consideration when interpreting data obtained from environmental samples using these methods. This is supported by Hoff (2009) and Carr and Borenstein (2014),

who found that increasing error rates decrease gene prediction accuracy. As the error rates of high throughput sequencing technologies improve, this effect will reduce.

One Codex had the greatest annotation sensitivity and the fourth highest annotation precision. This is probably due to a combination of the kmer-based annotation method that it uses and the simulated metagenome being created using the NCBI genome database, the primary reference source for One Codex. Other than the control, Megablast correctly annotated the most sequences at the genus level (97.3 %), although the sensitivity of this method was 0.2 % less than One Codex. MEGAN had the second highest precision, annotating 98.6 % of sequences, with 95.7 % correct annotations. This suggests that Megablast is the most reliable method for annotating sequences, and indicates that it is more conservative than One Codex when assigning a sequence hit but also less likely to misidentify a sequence. MEGAN's performance was similar to Megablast, which is expected as MEGAN processed the Megablast output. Discrepancies between the two are therefore derived from MEGAN's processing.

MG-RAST RefSeq had the fifth greatest annotation sensitivity and the greatest of the MG-RAST annotations (excluding MG-RAST M5NR, for which sequence-specific annotation data were unavailable), although it also achieved the greatest number of misidentifications. At the genus level, 33.7 % of sequences were misidentified and 55.9 % were correctly identified, leaving the remainder unassigned despite the fact that all taxa in Simmet are fully sequenced. This would suggest that investigating metagenomes at the genus level would be unreliable, generating many false positives and implying an incorrect community structure and composition. This supports Garcia-Etxebarria *et al*, (2014), who found that more annotations are made at higher taxonomic levels and that discrepancies between known frequencies and annotations increase at lower taxonomic levels, and Lindgreen *et al.* (2016), who report decreases in community annotation accuracy at the genus level compared to phylum. At the class level, the proportion of incorrect annotations is reduced to fewer than 10 % for MG-RAST RefSeq, with 80 % being annotated correctly. While taxonomic resolution is reduced, it ensures that the confidence in the annotations remains high.

MG-RAST KEGG correctly annotated a similar number of sequences to MG-RAST COG, but incorrectly annotated many more. KEGG offers a more descriptive annotation as it comprises specific gene and pathway annotations, whereas COG provides descriptions based on orthologous sequences. However, the specificity of KEGG classifications may be the cause of the incorrect annotations as there are more annotations to be selected

from and there may be more closely related functions, increasing the chance of misidentification. Because KEGG is now subscription based, and MG-RAST uses the last free version (2008), it will not contain information added after that date. Our results are in line with those produced by Lindgreen *et al.* (2016), who also conclude that MG-RAST's functional annotation was accurate.

The control, One Codex, Megablast and MEGAN achieved the greatest correlation coefficients between Simmet and annotation abundances at the genus level, all above 0.9. For all MG-RAST annotations the correlation coefficients were less than 0.8. For MG-RAST, the greatest correlation of all abundances was achieved at the order level by the M5NR database, closely followed by TrEMBL and RefSeq. These correlations inform us about community-wide analyses, but not sequence sensitivity and precision as correlating abundances values may occur from coincidental incorrect annotations.

MG-RAST over-annotated many more classes than MEGAN and One Codex, for which the most abundant feature was the unidentified group. This supports the sensitivity and precision data in suggesting that One Codex is more likely to categorise unknown sequences as unidentified, rather than incorrectly identifying them.

The genus richness estimated by MG-RAST M5NR was 81.0 % greater that Simmet's actual richness, the highest overstatement, while MEGAN achieved the most accurate genus richness value (2.3 % lower) after the control (100.2 %). This overstatement could be due to the greater number of sequences present in MG-RAST M5NR. MG-RAST M5RNA produced a relatively accurate estimate of genus richness (95.2 %); as M5RNA is a 16S rRNA gene database, it is unlikely to annotate non-16S rRNA gene sequences, reducing the number of incorrect identifications. However, the taxa abundance correlations show that MG-RAST M5RNA achieved the second lowest correlation with Simmet at the genus level, and the lowest at all other taxonomic levels. MG-RAST RefSeq generated the fifth most accurate richness value, greater than One Codex, although not as accurate as Megablast and MEGAN. Combined with its high abundance correlation with Simmet, this suggests that MG-RAST RefSeq provides a relatively accurate representation of both the richness of a community and the abundance of organisms present. MEGAN and One Codex achieve more accurate taxa richness values and taxa abundance correlations than MG-RAST RefSeq at the family level and above, suggesting they would be a viable alternative to MG-RAST RefSeq.

One limitation with evaluating annotations using organism nomenclature, rather than taxon IDs (which were unavailable for MG-RAST sequence-specific annotation data), is the lack of taxonomic metadata curation in some databases. Some genomes in the NCBI database are stored with the abbreviated species name rather than complete name, thus *A.mediterranei* would not automatically be identified as an *Amycolatopsis* species. Furthermore, as names are updated, disparities can form between different databases. For example, the class Chloroflexia has been renamed to Chloroflexi, and is called this by MG-RAST. However, NCBI is using the old name Chloroflexia (as of 17/11/14), thus sequences identified as Chloroflexi would not be correctly matched in Simmet. These issues were corrected for during data processing; however there may be other cases of disparities in the plethora of organisms present in the analysis. A solution to this would be to use the taxon IDs instead, however these were not available for sequence-specific annotations downloaded from MG-RAST.

In conclusion, One Codex, Megablast and MEGAN are suitable methods for annotating DNA sequences that are located in the reference databases that they use for annotation, with One Codex offering fast, web-based analyses and MEGAN providing a user-friendly Graphical User Interface to analyse BLAST results. Results appear to vary significantly depending on the program and parameters used, a conclusion also drawn by Lindgreen *et al.* (2016). While MG-RAST appears to have a greater rate of incorrect assignments, this is reduced when investigating higher taxonomic levels (e.g. with RefSeq: over 33 % at the genus level compared to less than 15 % and 10 % at the order and class levels). The correlations between the annotated taxa abundances are greatest for MG-RAST at the order level, using M5NR, TrEMBL or RefSeq. In many of the tests, MG-RAST M5NR proved to be a reliable database, but the diversity indices suggest that it is less reliable than MG-RAST RefSeq; at the class, order and family levels MG-RAST M5NR estimates more the double the actual richness values. Therefore, it it hypothesised that MG-RAST M5NR would generate more false positive sequence annotations than MG-RAST RefSeq. A simulated metagenome allows for the quantification of annotation errors. This study complements the work by Mavromatis *et al.* (2007), who evaluated different metagenomic processing methods using a simulated metagenome developed from 113 isolated genomes, and by Pignatelli and Moya (2011), who used simulated data to study the performances of *de novo* short-read assembly programs. It should be noted that the performances of the methods discussed in this study are likely to differ from the reported results when annotating environmental sequence data; a greater number of sequences are likely to be

unidentified due to the multitude of uncultured microorganisms (Streit and Schmitz, 2004) and non-sequenced microbial genomes (Tringe *et al.*, 2005) that are currently absent from the NCBI whole bacterial genome database. While this research focussed on a selection of annotation methods, the overall conclusions drawn should be considered for any pipeline.

In this study the annotation errors for a selection of parameters and databases are quantified. Analysis pipelines are not equivalent and certain parameters can significantly reduce the confidence in results. The findings from this chapter were used to construct the analytical pipelines for chapters 3 and 4. MG-RAST was selected to annotate the sequences due to its annotation speed, particularly for genetic functions; RefSeq was chosen as the reference database in MG-RAST based on the combined accuracy of taxa richness estimates, annotation sensitivity and annotation precision found in this chapter. A minimum identify cut off value of 60% was selected as this chapter concludes that, while higher values achieve a greater precision, they suffer a significant loss in sensitivity. A maximum E-value of $1e^{-15}$ was selected as lower values reduce the number of incorrect annotations. A minimum alignment length of 20 bp was selected as it increases sensitivity without significantly reducing precision, whereas higher values (e.g. over 35 bp) result in a strong reduction in sensitivity.

This chapter should be used as a guideline when determining methods for annotating metagenomic sequences and considered when interpreting metagenomic results. Ultimately, the most appropriate balance between taxonomic resolution, annotation sensitivity and annotation precision needs to be identified for each study conducted.

3 The effects of increased flooding frequency on a laboratory controlled microbial ecosystem.

3.1 Abstract

The impacts of increased flooding frequency on soil microbial communities and potential functions, in line with predicted environmental changes, were investigated in a laboratory-controlled environment. More frequent flooding events altered microbial community composition and increased diversity. Significant changes in taxa and functional gene abundances were identified and quantified. These changes include shifts in abundances of taxa and functions involved in biochemical cycles, such as nitrogen and sulphur cycling.

3.2 Introduction

3.2.1 Climate change and flooding

It is predicted that climatic changes will increase the frequency of extreme precipitation events in the UK, particularly in winter, and that this will result in an increase in flooding frequency (Trenberth, 1999; Houghton, 2001; Kleinen and Petschel-Held, 2007; Murphy *et al.*, 2009; Min *et al.*, 2011; Collins *et al.*, 2013; Kirtman *et al.*, 2013). This will alter soil microbial ecosystems and biogeochemical cycles (e.g. N, C, Fe and S), at least transiently. Complex microbial communities, such as those found in soil, can be highly responsive to environmental changes (Schmidt *et al.*, 2000; Waldrop and Firestone, 2006; Rinnan *et al.*, 2007). These cycles are fundamental to many areas of society, the environment and the economy, for example recycling nutrients for crop growth, producing or sequestering greenhouse gases, and degrading pollutants. It is important to understand how an increase in flooding frequency will spatially and temporally change soil microbial ecosystems and their functions.

3.2.2 Flooding and microbial ecosystems

Alternating flooding and draining will perturb microbial communities as the anoxia will kill some populations and allow others to develop (Denef *et al.*, 2001; Holling, 1973). Cycling between the two states will inhibit the community from stabilising with a predominantly aerobic or anaerobic population, and those that thrive will be able to tolerate both conditions. Flood duration will impact the community as redox potentials take time to decrease during anoxia (Wang *et al.*, 1993; Mohanty *et al.*, 2013), with

denitrification occurring, then iron and sulphur reduction, then finally methanogenesis once the iron and sulphur compounds have been reduced (Reddy and Patrick, 1975; Patrick and Jugsujinda, 1992). Drainage oxidises these compounds again, increasing the redox potential and inhibiting downstream reduction processes. Baldwin and Mitchell (2000) found that nitrification and denitrification decreased after periods of desiccation but increased again after rewetting, and Morillas *et al.* (2015) found that increased dry/wetting frequency decreased nitrification.

Anaerobic soils may contain methanogens, archaea that produce CH_4 under strictly anaerobic conditions, and flooding could increase their populations (Conrad, 2007). Methanotrophs, found both aerobically and anaerobically, metabolise CH_4. Methane has a 100-year global warming potential 32 times greater than CO_2 (Myhre *et al.*, 2013), thus studying the factors that increase CH_4 flux is essential for understanding climate change risks. Studies of rice paddies (Yagi *et al.*, 1996; Sigren *et al.*, 1997; Ratering and Conrad, 1998) found that short-term drainage of floods resulted in a sharp decrease in CH_4 emissions. This is expected because methanogens are intolerant to even low levels of oxygen (Conrad, 2007). However, once being flooded again, CH_4 emissions are still supressed. This may be caused by the oxidation of reduced sulphate and ferric iron during drainage (Patrick and Jugsujinda, 1992) providing a fresh source of substrates for sulphate/iron reducing bacteria. These would outcompete methanogens for H_2 and acetate (Conrad, 2007). How microbial communities will respond to frequent flooding and drainage on pasture soil is yet to be investigated.

While flooding induces anoxia in the bulk soil, the oxic state present during and after drainage may restore the community to its previous state. Ponnamperuma (1984) stated that most of the changes to the physical, chemical and biological processes of soil in response the flooding are reversed with draining and drying. The rate at which this occurs depends on many factors, such as the proliferation rates of species, redox potentials, the quantities of metabolic substrates present, and the rate at which floods subside. Obligate aerobic and facultative anaerobic bacteria grow best in aerobic conditions, but some can survive periods of hypoxia or anoxia, e.g. *Methylosinus trichosporium* (Roslev and King, 1994) and *Mycobacterium smegmatis* (Berney *et al.*, 2014). Frequent flooding interspersed with drainage will therefore only inhibit the growth of many bacteria species, rather than kill them. Furthermore, as a moist environment is preferable for many aerobic bacterial species (Heller, 1941; Roberson *et al.*, 1993; Potts, 1994; Fredrickson *et al.*, 2008), occasional flooding will provide a suitable environment for these species during drained periods.

3.2.3 Hypotheses

This chapter investigates the impacts of increased flooding frequency on laboratory-controlled microbial communities and their functions. It is hypothesised that increased flooding frequency will significantly change the composition and decrease alpha-diversity of microbial communities and their potential functions. Significant increases in abundances of genes involved in methane production and sulphate reduction are predicted following greater flooding frequencies, with decreases in methane oxidation genes.

3.3 Methodology

3.3.1 Experimental design

Soil was collected in summer 2013 from a pasture field in Wiltshire located next to the confluence of the River Sem and the River Nadder (Lat. 51.044770, Long. -2.111945) (Figures 3.1 & 3.2). The soil association is Wickham 2: fine loamy over clayey soil (Supporting Information A.3.1) (National Soil Resources Institute (NSRI), 2013).

Figure 3.1. The river confluence from where the soil was extracted. The photo was taken from the point of extraction.

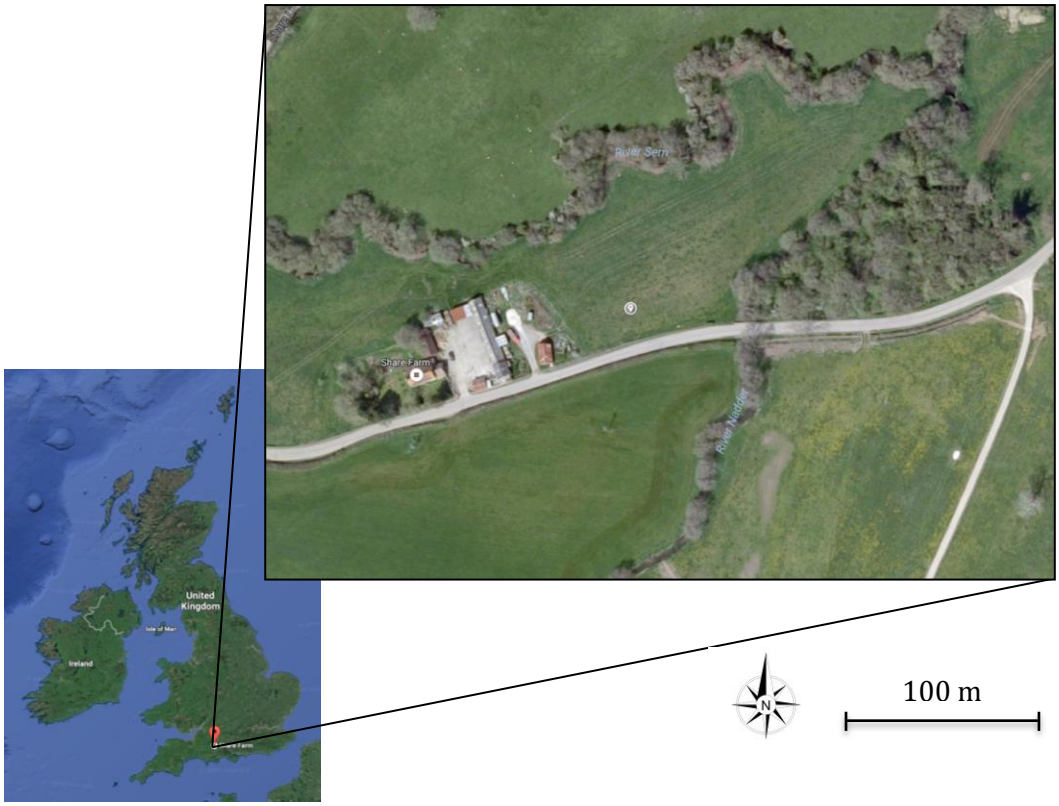


Figure 3.2. Extraction location. The location of soil extraction from pastureland adjacent to the confluence of the River Sem and the River Nadder. The white marker depicts the extraction location (Infoterra Ltd. & Bluesky, Map data, Google 2016).

The soil was passed through a 6 mm sieve and left to dry for seven days at room temperature. It was then homogenised and placed in six 8 (h) x 10 (d) cm plastic pots, 700 g per pot.

A soil sample was wetted and allowed to drain to measure the gravimetric content (GWC) of the soil at field capacity. 50.0 g of the sample was placed in at oven at 95 °C and left overnight. The sample was then placed in a desiccator and left to cool before being weighed to establish how much water had been lost. The mean GWC for the soil at field capacity was 0.37 g/g (Supporting Information A.3.2). The soil was under drought condition at the start of the experiment with a GWC of 0.04 g/g.

3.3.2 Treatment

All replicates were subjected to an initial flood for two weeks. The pots were placed in open 1.8 l containers (18 (h) x 12 (d) cm) and filled with deionised water to a soil-surface depth of 20 mm (Figure 3.3). After two weeks, all replicates were drained and their GWC brought to field capacity (Supporting Information A.3.2) using a Büchner

flask. For the remainder of the experiment, the 1 x flood treatments were not flooded again. The 3 x flood treatments were left drained for two weeks, then subjected to two further two-week flooding treatments, with a two-week period in between and at the end where they were left to drain freely (Table 3.1).

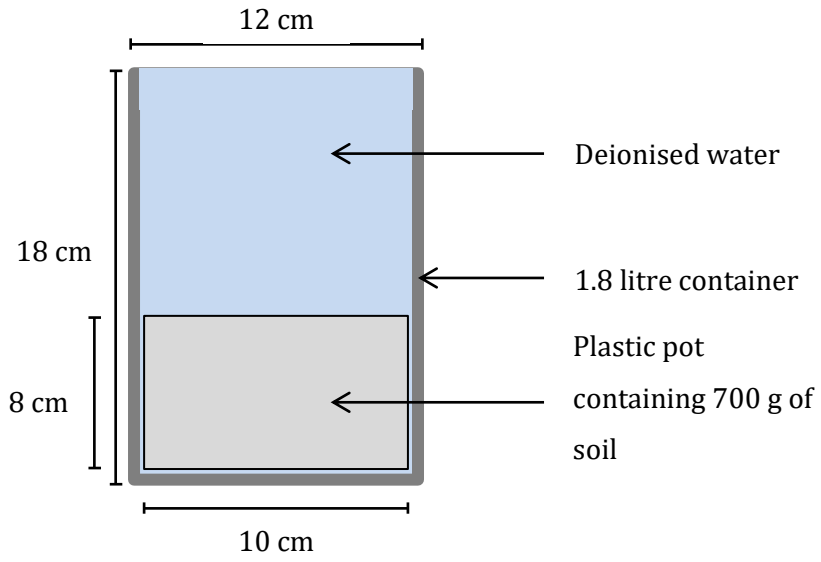


Figure 3.3. Experimental setup. A diagram depicting the experimental setup of the flooded pots of soil in open containers.

Table 3.1. The treatment regime for the laboratory experiment.

Time period (weeks)	1 x flood	3 x floods
1-2	Saturation	Saturation
3-4	Drained	Drained
5-6	Drained	Saturation
7-8	Drained	Drained
9-10	Drained	Saturation
11-12	Drained	Drained

3.3.3 DNA sampling

Three randomly selected soil samples (2 g) were extracted from the homogenised soil prior to filling the treatment containers, representing the starting soil community. Before each treatment manipulation, a soil sample was randomly extracted from each container from a depth of 5 cm using a 2 cm corer. DNA was extracted within two hours of sample collection using a PowerSoil® DNA Isolation kit (250 mg) (Mo Bio Laboratories Inc., Carlsbad, CA, USA) following the manufacturer's protocol.

The concentration and purity of DNA was established using a NanoDrop™ 8000 UV-Vis Spectrophotometer (Thermo Scientific, Waltham, MA, USA). The DNA was stored at -80 °C. After each extraction, the holes from the corer were filled using soil from additional pots that had been subjected to the same treatments as the experiments. They were marked with a cocktail stick to ensure no direct sampling contamination.

3.3.4 Sequencing

DNA quantities were determined using a Qubit® Fluorometer (Life Technologies Corporation, Carlsbad, CA, USA) and the technical extraction replicates were pooled together in equal quantities to form biological replicates. Samples were purified using an Agencourt AMPure XP bead clean-up kit following the manufacture's protocol (Beckman Coulter (UK) Ltd., High Wycombe, UK). Samples with concentrations greater than 10 ng/µl were diluted 1:10 using RNase-free water to ensure that the quantities were appropriate for use with the Nextera XT DNA sample preparation kit (Illumina UK, Little Chesterfield, UK); concentrations too high result in fragment lengths that are too long for sequencing. The samples were further diluted with RNase-free water to make 5 µl of solution with approximately 10 ng of DNA. DNA libraries were produced using the Nextera XT DNA sample preparation kit following the manufacture's protocol (barcodes listed in Table A.9). The samples were pooled, resulting in a DNA concentration of 17.5 ng/µl. The libraries were sequenced using a MiSeq Personal Sequencer (Illumina UK, Little Chesterfield, UK), with the assistance of Ummey Hany (Fera Science Ltd.), following the manufacturer's protocol. The v3 reagent kit was used, generating paired-end reads of 600 bp.

3.3.5 Analyses

The paired-end reads were merged with PEAR (Zhang *et al.*, 2014). Unmerged forward reads were trimmed with Sickle (https://github.com/najoshi/sickle) using a mean

phred score threshold of 25 to remove low quality ends without removing large amounts of data. The unmerged reverse reads were discarded to remove abundance bias when included with merged reads. The merged and trimmed forward read files were concatenated and uploaded to MG-RAST (Meyer *et al.*, 2008).

Sequences were annotated with a representative hit annotation technique, which selects a single, unambiguous annotation for each feature. The RefSeq database was used for taxonomic identification and Subsystems for functional assignment. The maximum E-value was $1e^{-15}$, providing a strict search parameter. The minimum sequence identity was 60 %, and the minimum alignment length was 20 bases. Taxa and functions with a total abundance below five across all samples were removed, as confident conclusions cannot be drawn for such low representations. Relative abundance values were generated and arc-sin square root transformed. Raw abundance values were square root transformed to calculate Bacteria:Archaea ratios. Rarefaction curves display the taxa richness per sequence count, visualising the effectiveness of sequence coverage. The lowest taxonomic level studied was the order level. Below this, the proportion of potential incorrect annotations was considered unacceptable (Randle-Boggis *et al.*, 2016).

The α-diversity of each sample was calculated using the Shannon index, an abundance-weighted average of the logarithm of the relative abundances of taxa. Treatment dissimilarities were tested with Analysis Of Similarity (ANOSIM, 100,000 permutations), Principal Coordinates Analysis (PCoA) and hierarchical clustering, all using the Bray-Curtis dissimilarity method. Taxa and function PCoA weightings were ranked and plotted (Figure A.7); those with a weighting > 0.02 or < -0.02 were considered for further analysis as this is where the curves begin to plateau. Significant changes to the relative abundances of orders and functions were tested for using ANOVA. Multiple comparison corrections were made using the Benjamini-Hochberg procedure. Significant differences in the abundances of methanogenesis, CH_4 oxidation and sulphur reduction genes were selectively tested for using ANOVA.

3.4 Results

3.4.1 Sequencing

8,408,535 paired-end sequences were generated with a mean sample sequence count of 934,300 ± 664,308 (Table 3.2). PEAR merged 78.98 ± 4.45 % of reads. All samples maintain a mean phred score greater than 30 (Figure 3.4). The mean sequence length

after merging and trimming was 231 ± 131 bases (Figure 3.5). The rarefaction curves suggest that sequence coverage was sufficient in all samples to represent the microbial community at the genus level; an enhanced sampling effort would yield only a few additional genera (Figure 3.6). Three x Floods replicate 1 would benefit the most from enhanced sampling.

Table 3.2. Sequence counts. The sequence counts for the raw sequences, processed sequences and sequences that passed MG-RAST's quality filter.

Treatment	Replicate	Raw sequence count	Sequence count (merged and trimmed forward)	Sequences count (passed MG-RAST quality filter)
Start	1	374,034	373,154	363,743
Start	2	2,386,787	2,378,586	2,291,611
Start	3	811,185	808,005	779,132
Mean		*1,190,668*	*1,186,582*	*1,144,829*
1 x flood	1	1,555,451	1,545,693	1,462,741
1 x flood	2	542,972	539,732	514,611
1 x flood	3	731,075	725,710	685,044
Mean		*943,166*	*937,045*	*887,465*
3 x floods	1	265,695	265,430	262,255
3 x floods	2	1,038,396	1,035,089	1,003,472
3 x floods	3	702,940	698,585	668,241
Mean		*669,010*	*666,368*	*644,656*

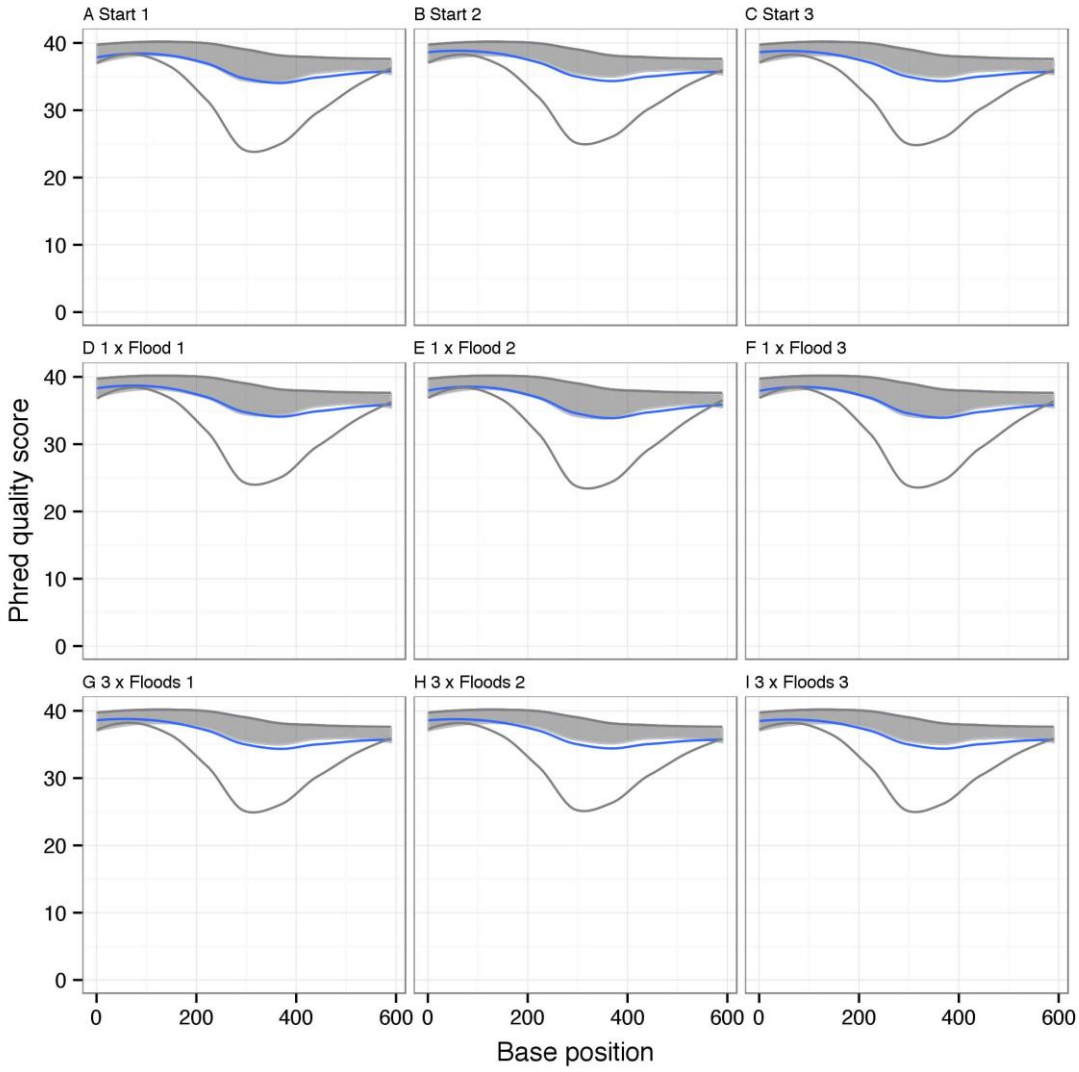


Figure 3.4. Phred quality scores. The phred quality score statistics for each base position after sequences were merged or trimmed. The dip around the 300 base-pair mark represents the end of the trimmed forward reads that were appended to the merged paired-end reads. The blue line shows the mean, the shaded grey area represents the interquartile range and the grey lines represent the 10th and 90th percentiles.

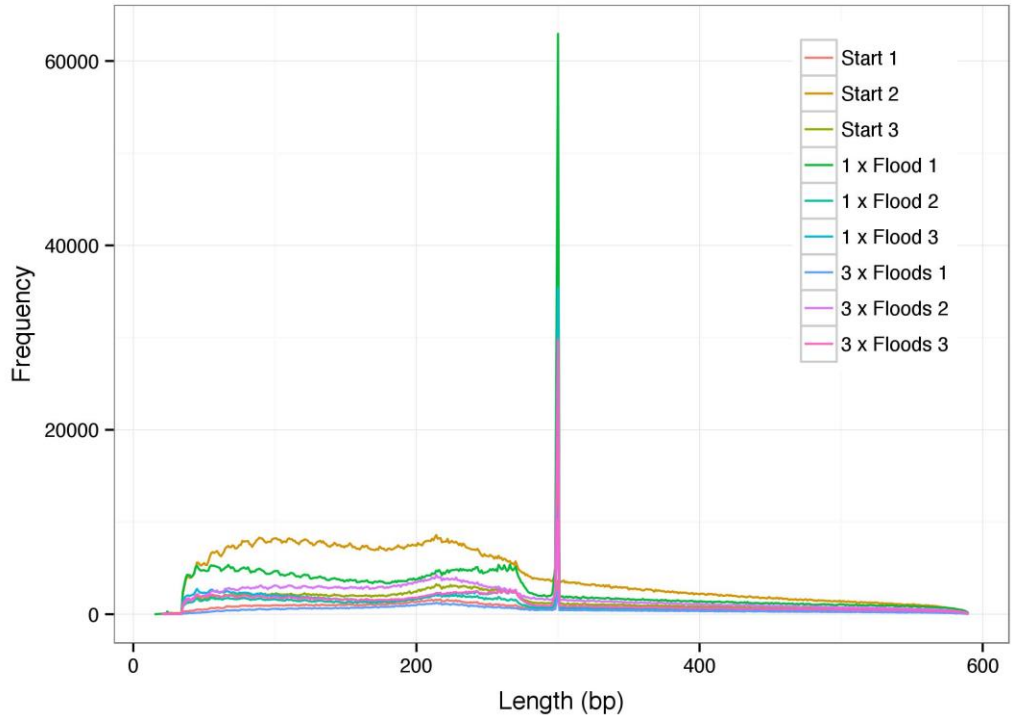


Figure 3.5. Sequence length distributions.

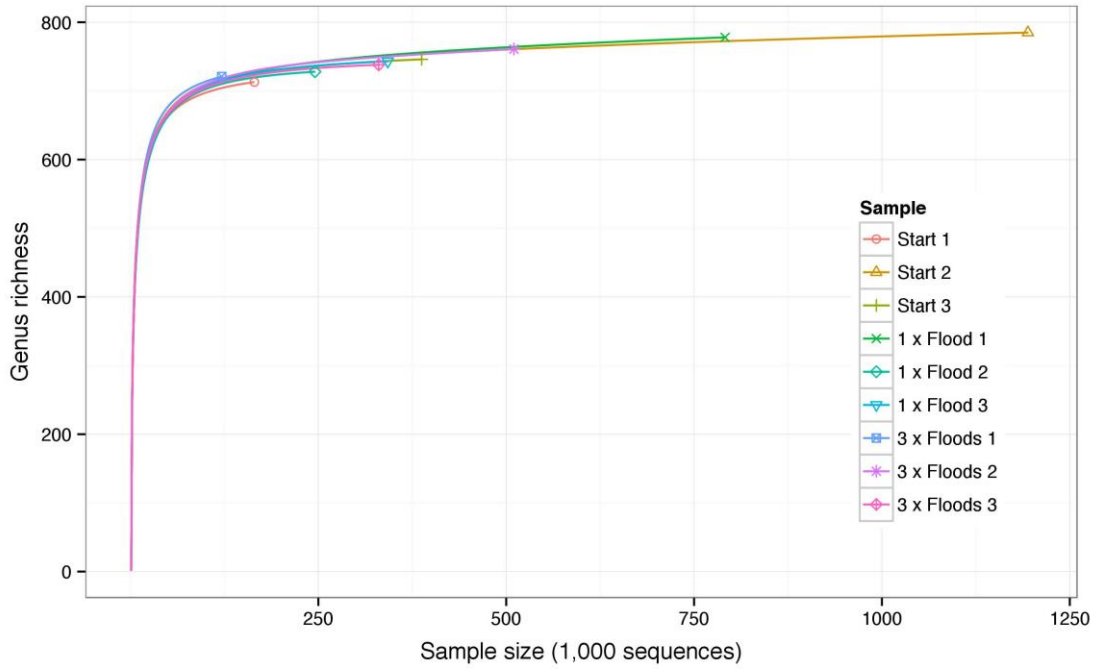


Figure 3.6. Genus rarefaction. Genus rarefaction curves displaying estimates of genus richness observed per sequence. A plateauing curve signifies sufficient community coverage, where an enhanced sampling effort would not yield many additional genera.

3.4.2 Diversity and Bacteria:Archaea ratio

There was a significant difference between the order α-diversities of the samples (Start: 4.478 ± 0.010, 1 x Flood: 4.465 ± 0.005, 3 x Floods: 4.492 ± 0.007; ANOVA, F = 8.486, df = 2, p = 0.018). Post-hoc testing revealed that the 3 x Floods treatment is significantly different from the 1 x Flood treatment (Tukey's HSD, p = 0.015).

The Bacteria:Archaea ratio significantly increased in response to flooding (√ transformed (n:1): Start: 12.01 ± 0.15, 1 x Flood: 12.26 ± 0.31, 3 x Floods: 12.74 ± 0.11; ANOVA, F = 26.85, df = 2, p = 0.001; Tukey's HSD, Start & 1 x Flood: p = 0.001, Start & 3 x Floods, p = 0.012).

3.4.3 Sample dissimilarities

Flood frequency had a significant effect on the microbial community taxonomic composition (ANOSIM, R: 0.679, *p* = 0.023) and function (ANOSIM, R: 0.251, *p* = 0.003). Both treatments were taxonomically dissimilar from each other and the Start samples (Figures 3.7-3.10). The 3 x Floods samples were functionally dissimilar from the Start and from 1 x Flood samples, both of which were not dissimilar from each other (Figures 3.11 & 3.12, Tables 3.3 & 3.4).

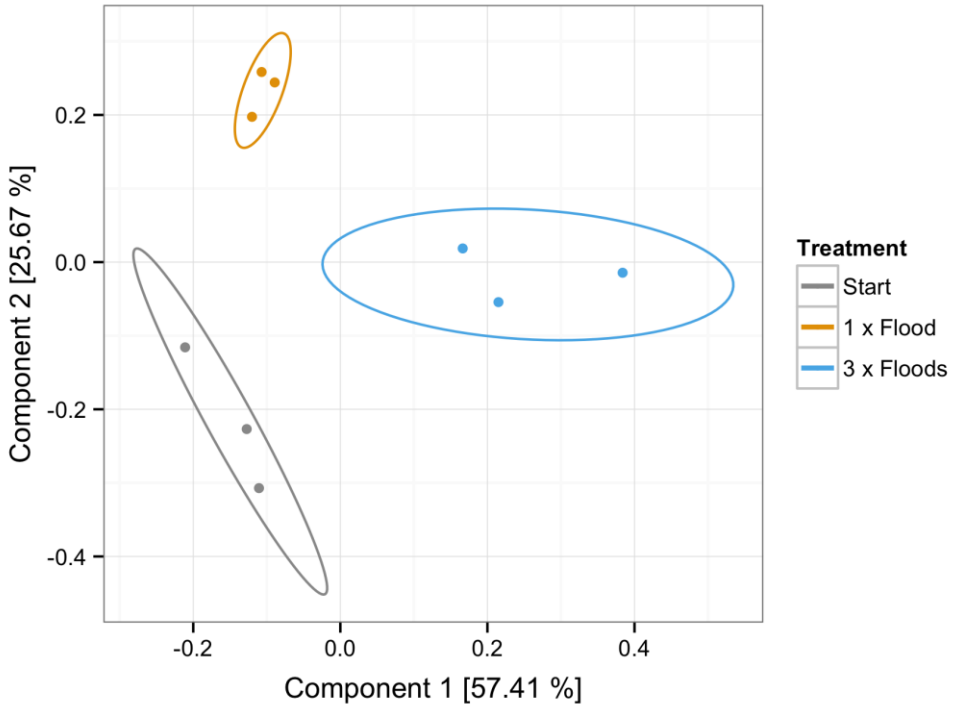


Figure 3.7. Order PCoA. A PCoA of the relative abundance of orders (Bray-Curtis distance method). Ellipses display 95 % confidence intervals.

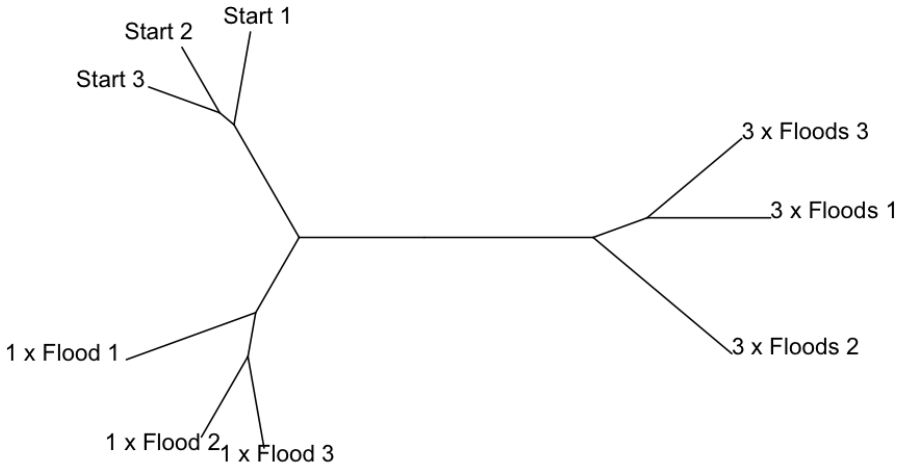


Figure 3.8. Order hierarchical clustering. A hierarchical clustering analysis of the community composition at the order level (Bray-Curtis distance method).

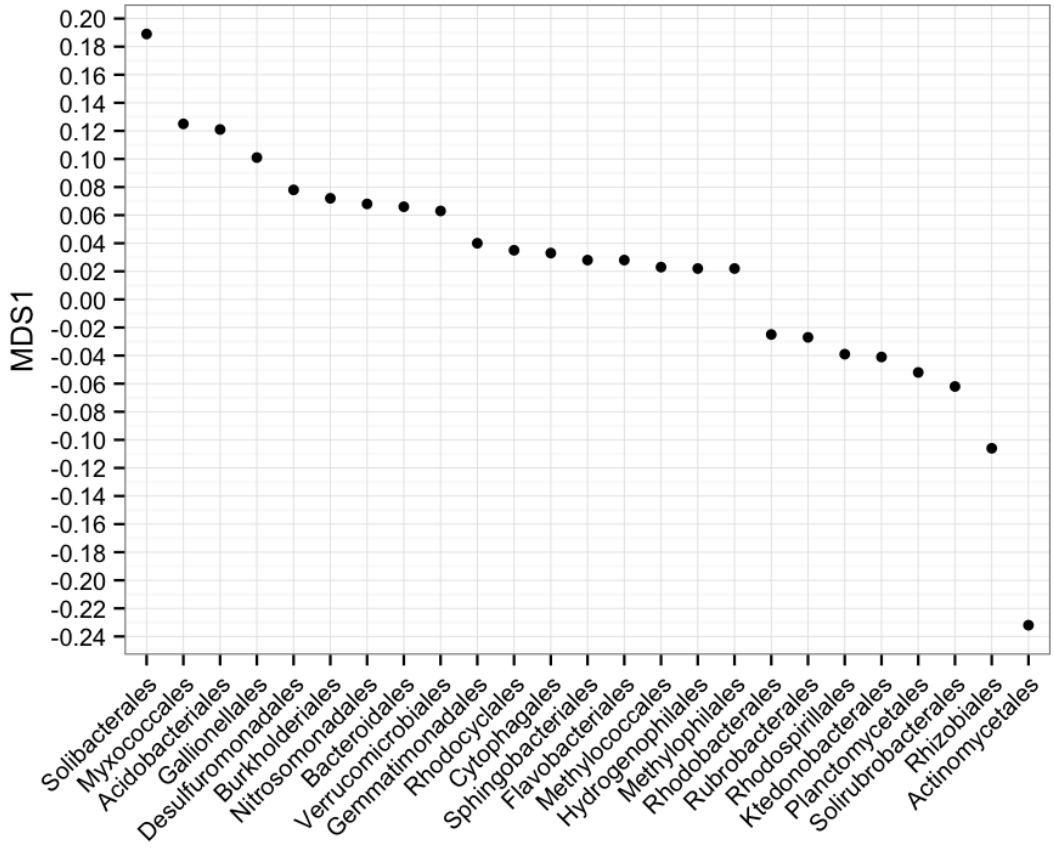


Figure 3.9. Order PCoA component 1 weightings. Orders with PCoA weightings > 0.02 or < -0.02 for component 1 in the taxonomic PCoA.

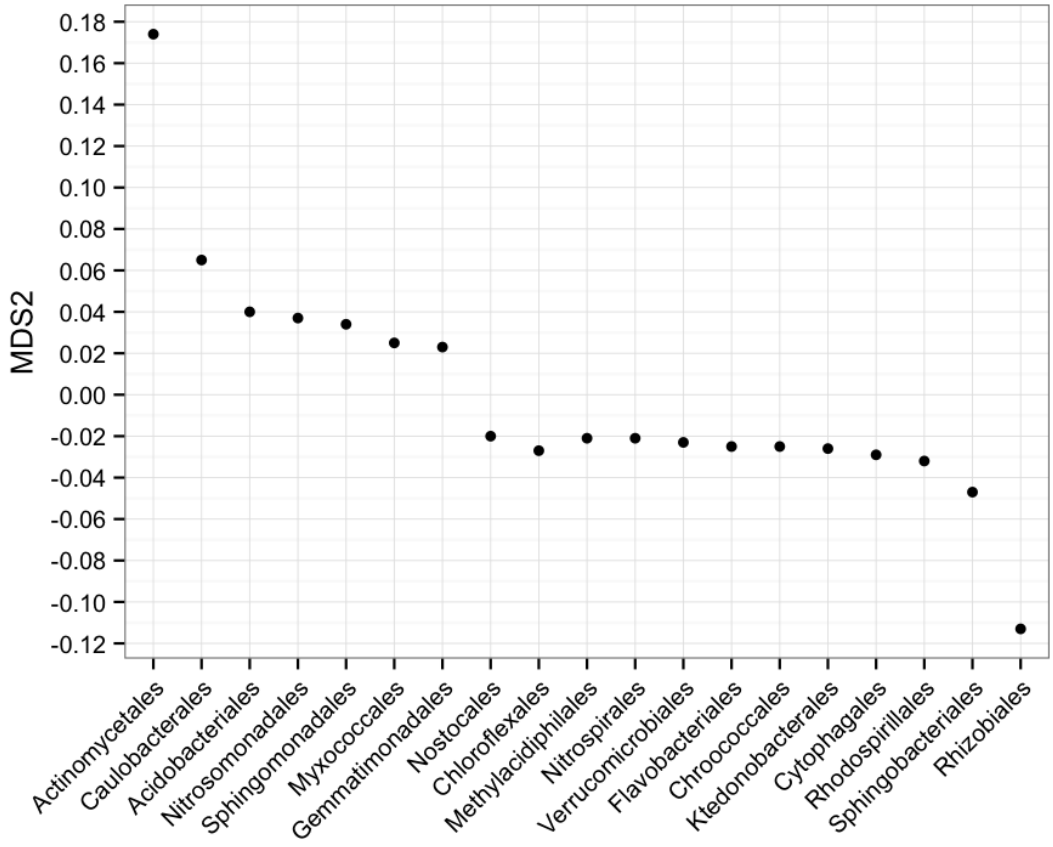


Figure 3.10. Order PCoA component 2 weightings. Orders with PCoA weightings > 0.02 or < -0.02 for component 2 in the taxonomic PCoA.

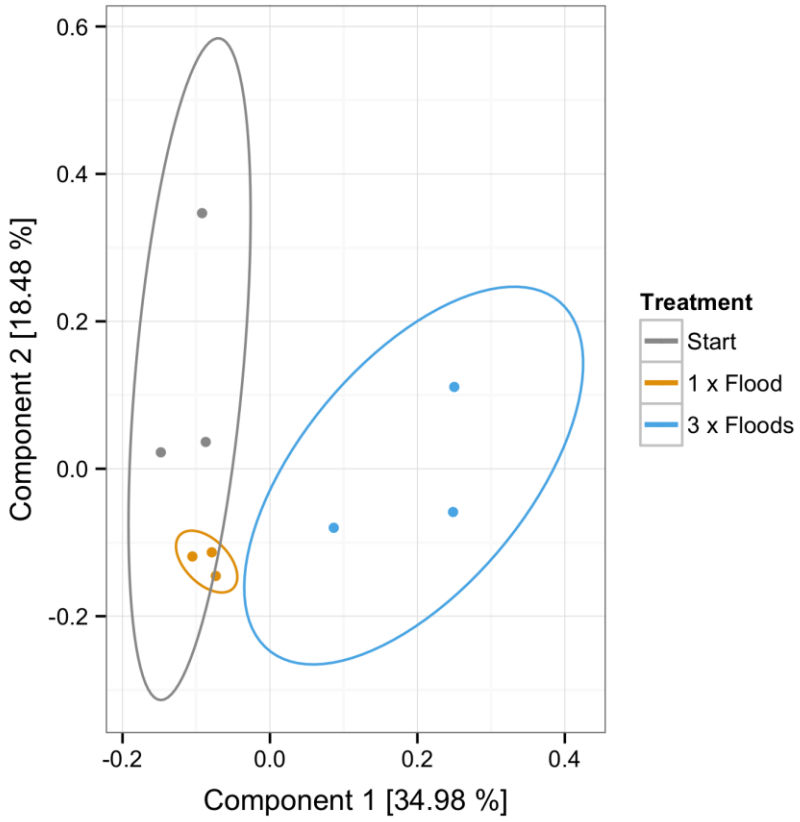


Figure 3.11. Functional PCoA. A PCoA of potential Subsystems level 3 functions (Bray-Curtis distance method). Ellipses display 95 % confidence intervals.

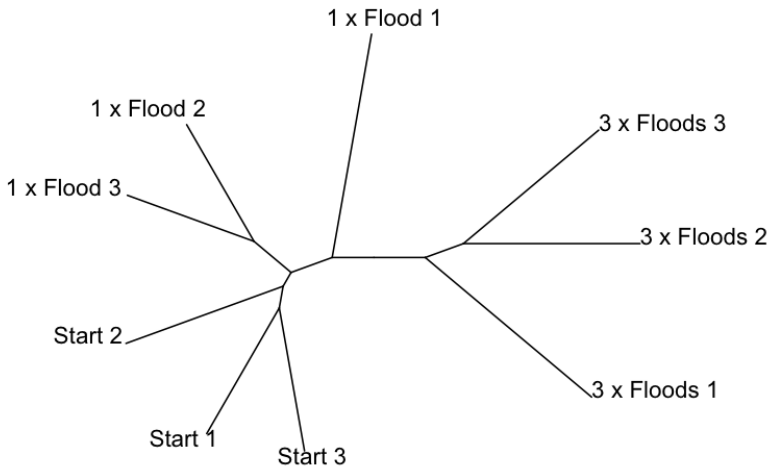


Figure 3.12. Functional hierarchical clustering. A hierarchical clustering analysis of potential level 2 functions (Bray-Curtis distance method).

Table 3.3. Functional PCoA component 1 weightings. Functions with PCoA weightings > 0.02 or < -0.02 for component 1 in the functional PCoA.

Function	MDS1
Cobalt-zinc-cadmium resistance	0.042
Ton and Tol transport systems	0.040
Flagellar motility	0.036
Bacterial Chemotaxis	0.035
Hydrogenases	0.030
Iron acquisition in Vibrio	0.029
Sugar utilization in Thermotogales	0.026
Lactose and Galactose Uptake and Utilization	0.026
C jejuni colonization of chick caeca	0.025
Lactose utilization	0.024
Zinc resistance	0.022
Two-component regulatory systems in Campylobacter	0.021
Nitrosative stress	0.021
Outer membrane	0.021
Respiratory Complex I	0.021
Major Outer Membrane Proteins	0.020
General Secretion Pathway	0.020
Siderophore Pyoverdine	0.020
Phospholipid and Fatty acid biosynthesis related cluster	0.020
Niacin-Choline transport and metabolism	-0.020
Coenzyme PQQ synthesis	-0.020
Cobalamin synthesis	-0.021
Proline, 4-hydroxyproline uptake and utilization	-0.021
Amidase clustered with urea and nitrile hydratase functions	-0.021
Glutathione analogs: mycothiol	-0.021
Iojap	-0.024
Creatine and Creatinine Degradation	-0.027
Phage integration and excision	-0.028
cAMP signaling in bacteria	-0.033
CBSS-222523.1.peg.1311	-0.033
CO Dehydrogenase	-0.043
CBSS-314269.3.peg.1840 (CO Dehydrogenase proteins)	-0.044

Table 3.4. Functional PCoA component 2 weightings. Functions with PCoA weightings > 0.02 or < -0.02 for component 2 in the functional PCoA.

Function	MDS2
CBSS-222523.1.peg.1311	0.034
cAMP_signaling_in_bacteria	0.034
Iojap	0.034
Cluster_with_phosphopentomutase_paralog	-0.020
SigmaB_stress_responce_regulation	-0.024

3.4.4 Taxonomic and functional abundances

The most abundant phyla across the samples were Proteobacteria, Actinobacteria, Acidobacteria and Verrucomicrobia (Figure 3.13), which are often dominant phyla (Janssen, 2006).

The relative abundances of 29 orders (out of 223) were significantly different among the treatments, after correcting *p*-values for multiple comparison corrections (ANOVA and Benjamini Hochberg) (Table 3.5). Most significant differences occur between the 1 x Flood treatment and the 3 x Floods (Tukey's HSD). There were no significant differences between Subsystems level 3 functions, however at level 2, 14 out of 166 functions were significantly different (Table 3.6).

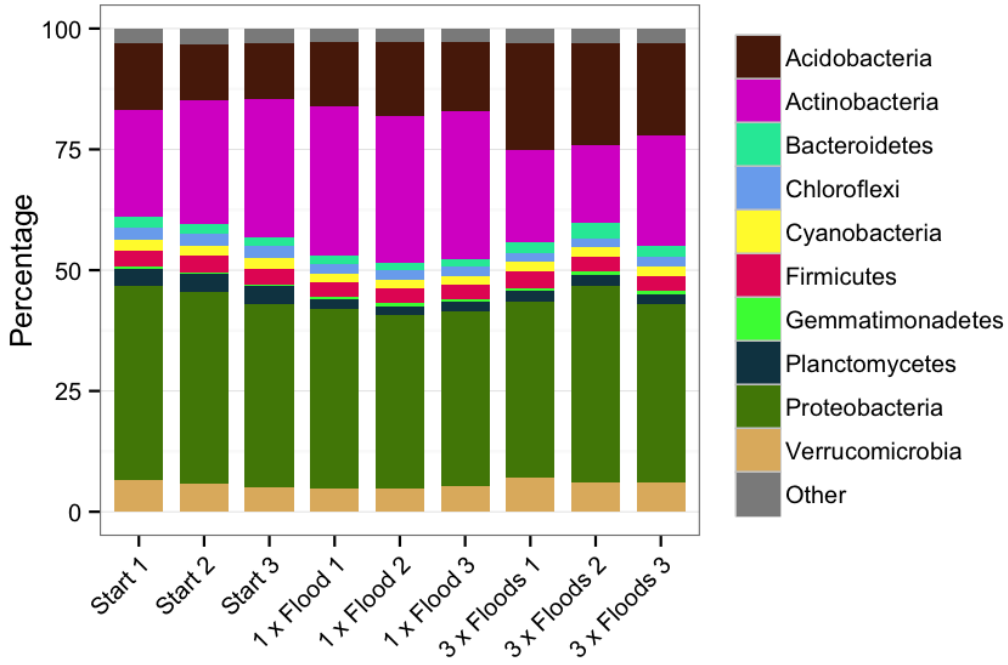


Figure 3.13. Phyla relative abundances per sample.

Of the 206 orders detected at the start, 66 population abundances increased, 122 decreased and 18 populations were undetected after receiving one flood (Table A.10). 107 increased, 78 decreased and 21 were undetected after receiving three floods (Table A.11). 17 orders were undetected in the starting soil but were detected at the end of the experiment. Figures 3.14 & 3.15 show the fold changes between orders (Tables A.12 & A.13). Of the 1,080 level 3 functions detected at the start, 537 relative abundances increased, 471 relative abundances decreased and 46 relative abundances were undetected after receiving one flood (Table A.14). 512 level 3 functions increased, 483 decreased and 46 were undetected after receiving three floods (Table A.15). 39 level 3 functions were undetected in the starting soil but were detected at the end of the experiment.

Table 3.5. Significantly different orders. The orders with significantly different relative abundances between the samples (ANOVA). The *p*-values were adjusted for multiple comparisons (Benjamini-Hochberg) and post-hoc tests were performed (Tukey's HSD). "*N.S.*" = not significant.

Order	Start ($\bar{x}$)	1 x F ($\bar{x}$)	3 x F ($\bar{x}$)	Corrected *p*-value	Tukey's HSD		
					Start / 1F	Start / 3F	1 x F / 3 x F
Rhodospirillales	0.014	0.013	0.012	<0.000	<0.000	<0.000	<0.000
Planctomycetales	0.020	0.015	0.016	<0.000	<0.000	<0.000	0.002
Myxococcales	0.016	0.018	0.022	0.001	0.035	<0.000	<0.000
Sphingomonadales	0.008	0.009	0.008	0.008	<0.000	*N.S.*	<0.000
Rhizobiales	0.046	0.041	0.039	0.010	0.002	<0.000	*N.S.*
Enterobacteriales	0.007	0.007	0.007	0.011	0.001	*N.S.*	0.002
Caulobacterales	0.008	0.011	0.009	0.012	<0.000	0.014	0.015
Rhodobacterales	0.013	0.012	0.012	0.012	0.006	<0.000	0.025
Solibacterales	0.026	0.028	0.034	0.013	*N.S.*	0.001	0.003
Nitrospirales	0.006	0.005	0.005	0.013	0.001	*N.S.*	0.003
Acidobacteriales	0.016	0.019	0.022	0.014	0.033	<0.000	0.005
Nostocales	0.007	0.007	0.007	0.015	0.001	*N.S.*	0.003
Fibrobacterales	0.001	0.001	0.002	0.017	*N.S.*	0.001	0.004
Desulfobacterales	0.005	0.005	0.006	0.017	*N.S.*	0.016	0.001
Chroococcales	0.011	0.010	0.010	0.020	0.001	*N.S.*	0.012
Burkholderiales	0.023	0.024	0.026	0.022	*N.S.*	0.002	0.009
Syntrophobacterales	0.006	0.006	0.007	0.023	0.026	*N.S.*	0.002
Sphaerobacterales	0.006	0.006	0.005	0.025	0.039	0.002	*N.S.*
Cytophagales	0.008	0.007	0.008	0.031	0.039	*N.S.*	0.003
Desulfovibrionales	0.007	0.006	0.007	0.034	0.037	*N.S.*	0.003
Chlorobiales	0.006	0.006	0.006	0.039	0.032	*N.S.*	0.005
Chloroflexales	0.010	0.009	0.009	0.040	0.007	0.010	*N.S.*
Oscillatoriales	0.006	0.005	0.006	0.040	0.005	*N.S.*	0.021
Alteromonadales	0.007	0.006	0.007	0.040	0.041	*N.S.*	0.005
Nitrosomonadales	0.006	0.008	0.009	0.041	*N.S.*	0.005	*N.S.*
Lentisphaerales	0.002	0.002	0.002	0.041	*N.S.*	0.018	0.007
Rubrobacterales	0.007	0.007	0.006	0.041	*N.S.*	0.005	0.017

Order	Start ($\bar{x}$)	1 x F ($\bar{x}$)	3 x F ($\bar{x}$)	Corrected *p*-value	Tukey's HSD		
					Start / 1F	Start / 3F	1 x F / 3 x F
Gemmatimonadales	0.006	0.007	0.008	0.043	*N.S.*	0.006	*N.S.*
Actiniaria	0.001	0.001	0.001	0.049	*N.S.*	*N.S.*	0.007
Chromatiales	0.008	0.008	0.008	<0.050	*N.S.*	0.049	0.008

Table 3.6. Significantly different functions. The Subsystems level 2 functions with significantly different relative abundances between the samples (ANOVA). The *p*-values were adjusted for multiple comparisons (Benjamini-Hochberg) and post-hoc tests were performed (Tukey's HSD). "*N.S.*" = not significant.

Function	Start ($\bar{x}$)	1 x F ($\bar{x}$)	3 x F ($\bar{x}$)	Corrected *p*-value	Tukey's HSD		
					Start / 1F	Start / 3F	1 x F / 3 x F
Monosaccharides	0.012	0.012	0.013	0.007	0.019	<0.000	<0.000
ABC transporters	0.008	0.008	0.008	0.008	0.001	<0.000	*N.S.*
Resistance to antibiotics and toxic compounds	0.017	0.016	0.018	0.010	0.043	0.002	<0.000
Peripheral pathways for catabolism of aromatic compounds	0.010	0.010	0.010	0.010	*N.S.*	0.001	<0.000
Bacterial cytostatics differentiation factors and antibiotics	0.001	0.002	0.001	0.011	0.008	0.012	<0.000
Phages Prophages	0.011	0.010	0.010	0.020	0.025	0.001	0.019
Nucleotidyl phosphate metabolic cluster	0.009	0.008	0.008	0.022	0.004	0.001	*N.S.*
Cytochrome biogenesis	0.005	0.005	0.005	0.033	*N.S.*	0.006	0.002
Molybdopterin oxidoreductase	0.003	0.003	0.003	0.035	*N.S.*	0.005	0.003
Capsular and extracellular polysacchrides	0.011	0.011	0.011	0.036	*N.S.*	0.030	0.002
Toxins and superantigens	0.001	0.001	0.001	0.037	*N.S.*	0.002	0.013
Organic sulfur assimilation	0.008	0.007	0.007	0.038	0.021	0.003	*N.S.*
Tricarboxylate	0.003	0.003	0.003	0.040	0.002	*N.S.*	*N.S.*

Function	Start ($\bar{x}$)	1 x F ($\bar{x}$)	3 x F ($\bar{x}$)	Corrected *p*-value	Tukey's HSD		
					Start / 1F	Start / 3F	1 x F / 3 x F
transporter							
Metabolism of central aromatic intermediates	0.008	0.008	0.008	0.040	*N.S.*	0.005	0.007
Alpha proteobacterial cluster of hypotheticals	0.002	0.001	0.001	0.049	*N.S.*	0.004	*N.S.*

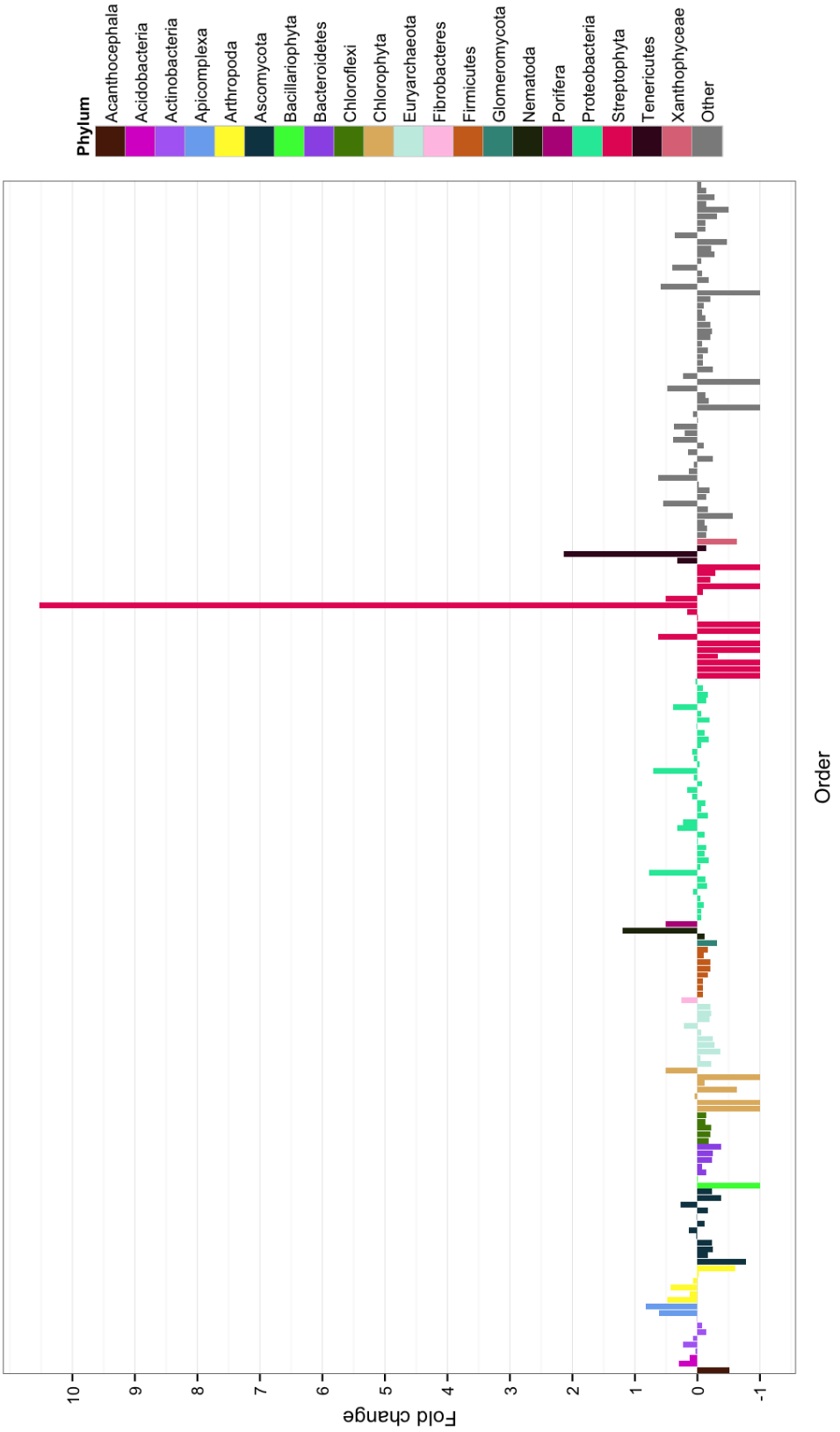


Figure 3.14. Order fold changes after one flood. The fold changes of orders, coloured by phyla, between the Start and 1 x Flood treatment.

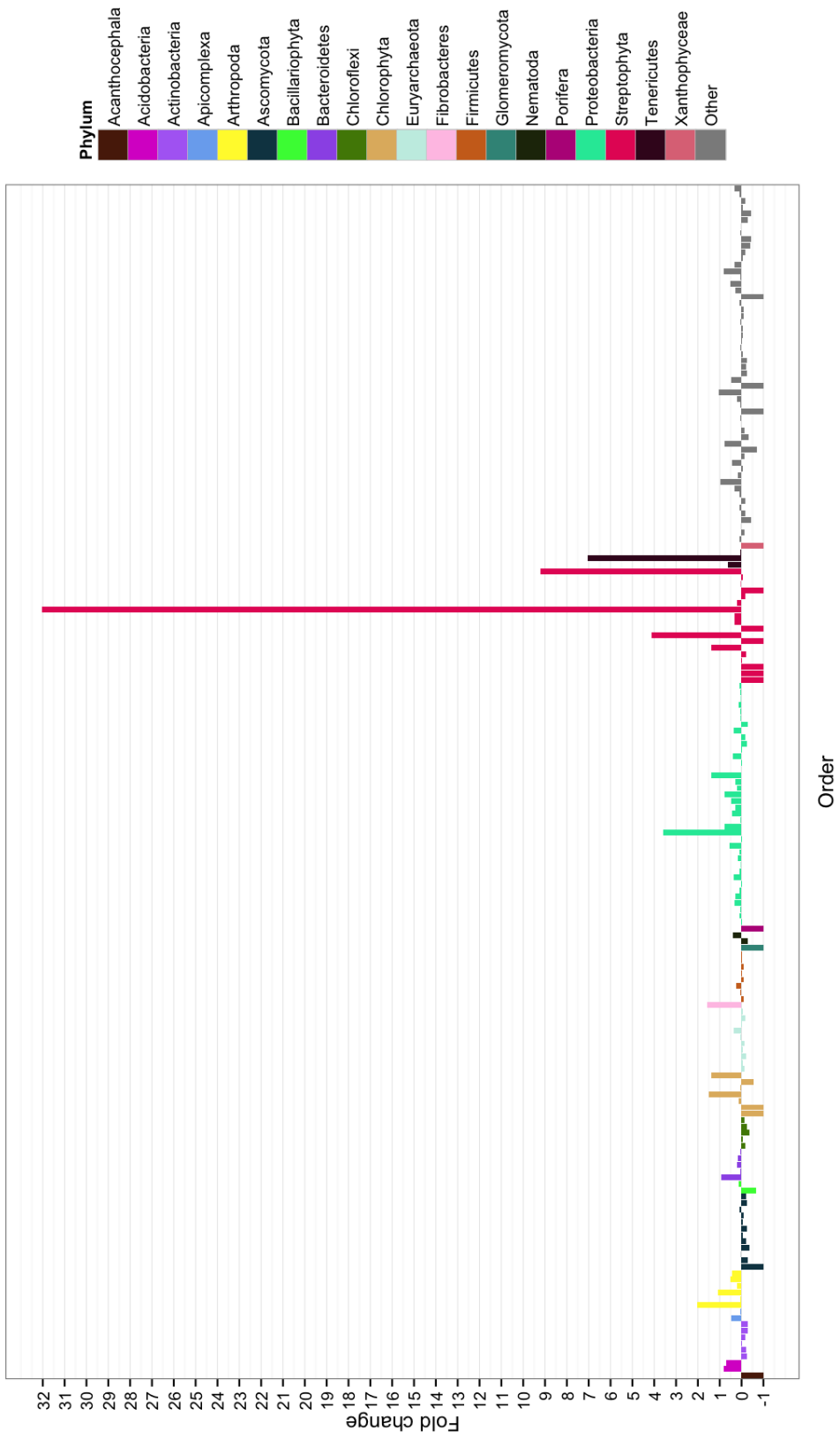


Figure 3.15. Order fold changes after three floods. The fold changes of orders, coloured by phyla, between the Start and 3 x Floods treatment.

3.4.5 Relative abundance of selected functional groups

The relative abundances of genes involved in methanogenesis and CH_4 oxidation were not significantly different (ANOVA, methanogenesis: F = 1.681, df = 2, p = 0.263; CH_4 oxidation: F = 2.535, df = 2, p = 0.159). There was a significant difference in the relative abundances of genes involved in Sulphate reduction (ANOVA, F = 11.07, df = 2, p = 0.001, Tukey's HSD: Start & 3 x Floods: p = 0.014, 1 x Flood & 3 x Flood: p = 0.017) (Figure 3.16).

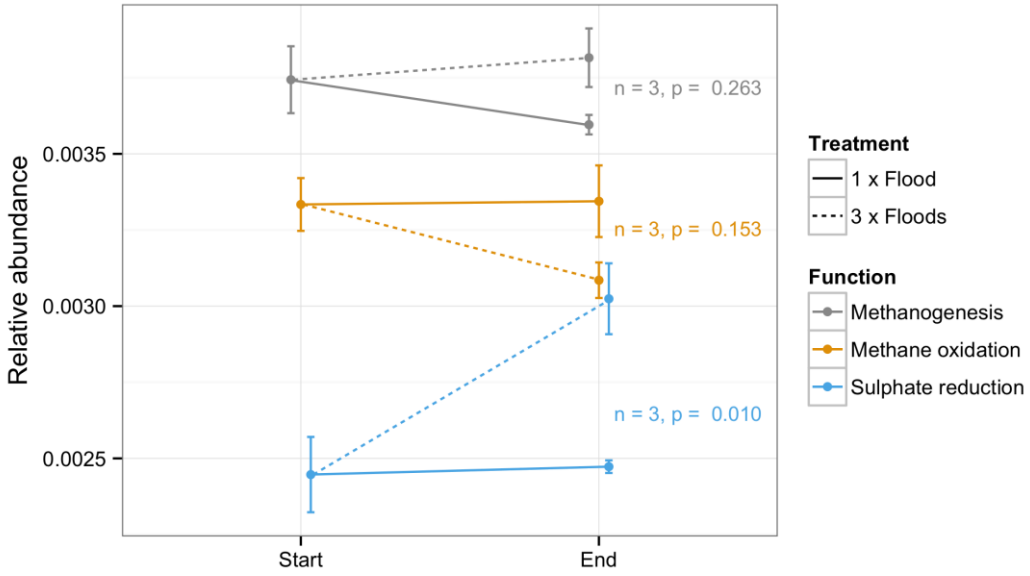


Figure 3.16. Selected functional responses. The differences in relative abundances of genes involved in methanogenesis, methane oxidation and sulphur reduction (ANOVA). Error bars show standard deviation.

3.5 Discussion

3.5.1 Diversity and Bacteria:Archaea ratio

The order α-diversities were significantly greater in the samples that received three floods. While it is hypothesised that anaerobic environments would tend towards a lower α-diversity over time, the short-term repeated shifts between aerobic and anaerobic conditions would inhibit a community from stabilising, allowing populations of facultative anaerobic organisms to develop and aerobic populations to recover after flooding.

The Bacteria:Archaea ratio increased in response to flooding, with the greatest ratio being observed in the 1 x Flood treatment. Decreases in archaea relative abundance were observed. Archaeal RNA polymerase initiation factors and archaeal thermosomes, involved respectively in transcription and protein structuring, also decreased. Most archaea are either strict anaerobes or can only tolerate low levels of oxygen (Berg *et al.*, 2010), thus the drainage period in between the floods would kill several of the strict anaerobes. Some bacteria, on the other hand, can survive periods of hypoxia or anoxia (Roslev and King, 1994; Berney *et al.*, 2014) and some would thrive in the moist environment provided by the initial flood (Heller, 1941; Roberson *et al.*, 1993; Potts, 1994; Fredrickson *et al.*, 2008).

3.5.2 Sample dissimilarities

The community compositions were all dissimilar, revealing that flood frequency has a strong impact on community structure; the 3 x Flooded communities were the most distinct. This was expected, as soil microbial communities can be highly responsive to environmental changes (Schmidt *et al.*, 2000; Waldrop and Firestone, 2006; Rinnan *et al.*, 2007). While the 3 x Flooded communities were functionally different from the other two, the start and 1 x Flood communities were not dissimilar. The discrepancy between taxonomic and functional results is most likely due to functional complementarity among taxonomically different microbial communities; this is a reminder that taxonomic fluctuations do not necessarily imply functional shifts. The 3 x Flood samples are less aerobically stable than the other samples, thus communities will undergo a much greater shift that would likely include significant functional differences.

3.5.3 Taxonomic and functional shifts

Orders that decreased in response to both treatments include several eukaryotes such as fungi (Capnodiales, Mucorales and Polyporales) and algae (Cyanidiales). As this was a controlled laboratory experiment using homogenised soil, the loss of free organic matter due to consumption may cause the populations of many fungi and algae to initially decrease. Algae orders Chroococcales and Oscillatoriales both decreased after receiving one flood, but not three. The Repeated flooding would limit the effects of desiccation between floods, allowing organisms that prefer moist environments to survive.

Most bacteria involved in the nitrogen cycle that declined in abundance decreased in response to one flood and not to three floods. These include Enterobacteriales, which are largely facultative anaerobes and nitrate reducers (Imhoff, 2005), and Nitrospirales and Nostocales, both aerobic nitrifying bacteria. Furthermore, the relative abundances of RNA polymerase sigma-54 factor RpoN, a function involved in nitrogen assimilation and fixation (Powell *et al.*, 1995; Gardner *et al.*, 2003), its response regulator, and nitrosative stress genes all increased in response to three floods but not one. Populations of nitrifying Nitrosomonadales increased in response to both treatments. Initial wetting releases nitrogen that becomes available for nitrification, however after long periods of desiccation several bacteria die off (De Groot and Van Wijck, 1993). As the initial influx of nitrites is assimilated there will be less available for nitrifying bacteria to oxidise. The abundance of genomes containing heterocyst (nitrogen-fixing cells) formation genes in cyanobacteria increased in response to one flood; heterocysts are formed during nitrogen stress, supporting most of our taxonomic findings (with the exception of Nitrosomonadales). Rewetting allows oxidised material to be reduced again, continuing the cycle, and the additional periods of anoxia will permit denitrification (Baldwin and Mitchell, 2000). Verhoeven *et al.* (2014) discovered a decrease in nitrification and denitrification in mangroves after increased flooding frequency, opposing these findings. Nutrient cycling is influenced by a variety of factors such as nutrient availability, redox potential, microbial community composition, temperature, and many others. Therefore, discrepancies between results are expected due to differing experimental conditions, for example saline mangroves versus terrestrial pasture soils.

Rhizobiales abundance declined in response to both treatments. Rhizobiales includes four nitrogen-fixing families, supporting the results discussed in the previous paragraph. The order also includes the methanotrophic family Methylococcaceae, which are important oxidisers of CH_4 in flooded soils (Conrad, 1996); each of these families' abundances declined in response to both treatments.

It is hypothesised that methanogen and methanotroph populations would increase in response to a greater flooding frequency, contradicting the Rhizobiales data. Indeed, methanotrophic Methylococcales populations did increase after three floods. Methanotrophic Rhizobiales can survive anaerobic conditions, so it is expected that their populations would also have increased in response to flooding and associated CH_4 emissions. A significant increase in methanogen populations was not observed however. Therefore, an increase in CH_4 production may not have occurred, meaning

that methanotroph populations such as Methylococcaceae will not have grown. In fact, genes involved in the serine-glyoxylate cycle, a part of methylotrophic metabolism (Ensign, 2006), decreased in response to three floods. The lack of developing methanogen populations could be explained by the increase in sulphate-reducing bacteria after three floods, as these initially out-compete methanogens for substrates to metabolise (Conrad, 2007). The greater taxonomic resolution achievable by NGS, compared to other methods such as DGGE and T-RFLP, allows for more detailed understandings of ecosystems to be made. However, the complexity of interactions and responses means that environmental data such as nutrient content and gas fluxes are necessary to make reliable conclusions. To further understand the methanogen/methanotroph results discussed above, sulphur compound content, hydrogen content and CH_4 fluxes need to be measured. This would verify the potential functional responses observed in the DNA.

Populations of strict anaerobic organisms decreased after one flood followed by oxygenation, and many increased in response to three floods. Syntrophobacterales, Chlorobiales, Clostridiales, and Desulfovibrionales are all obligate anaerobes that decreased after the one flood treatment and increased after three floods. Chlorobiales oxidise sulphur compounds, H_2 or Fe(II) (Bryant and Frigaard, 2006), and Desulfovibrionales reduce sulphates, thus are important in mineral cycling. Alkanesulfonate assimilation, involved in sulphur assimilation during limited sulphur availability (Ellis, 2011), decreased after three floods; this supports our taxonomic findings. Genes involved in organic sulphur assimilation decreased overall in response to both treatments.

The reduction of Fe(III) during the floods would likely have caused the increase in the Fe(II) oxidising bacteria Gallionellales observed after both treatments, due to the spike in substrate availability (Conrad, 2007). Both treatments resulted in an increase in abundance of genes involved in iron acquisition, transport and metabolism, with Ton and Tol transport systems (iron transport, (Noinaj *et al.*, 2010)) increasing after three floods only. The increase in reduced metals and other substrates would explain the increase observed in Cobalt-zinc-cadmium resistance genes and substrate uptake regulation (e.g. Ton and Tol transport systems). These increases were not observed in the one-flood samples, probably due to the resulting oxidation after drainage. Hydrogenase genes, largely involved in anaerobic metabolism (Vignais and Billoud, 2007), also increased after three floods. To further understand these interactions, the chemical components of the soil need to be measured.

Not all anaerobes decreased after the one-flood treatment followed by desiccation; Rhodocyclales, which are anaerobic oligotrophs, increased after both treatments. Fibrobacteres, which include many anaerobic rumen bacteria (Ransom-Jones *et al.*, 2012), increased after three floods but did not change significantly after one flood.

Other orders that increased after both treatments include Euglyphida, Gemmatimonadetes, and Myxococcales. Euglyphida are amoebae common in soils, marshes and organic-rich environments that feed on bacteria (Lamentowicz *et al.*, 2011). DeBruyn *et al.* (2011), with evidence supported by a meta-analysis, suggest that Gemmatimonadetes are adapted to arid conditions, inferring this result is unexpected. However, Gemmatimonadetes typically make up 2.2 % of soil bacteria (Janssen, 2006), and the only characterised species was isolated from wastewater (Zhang, 2003), thus presence in moist soils is not to be unexpected. The increase in Myxococcales hints at one of the current caveats of metagenomics. Myxococcales has an exceptionally long genome (ca. 13 mb) (Schneiker *et al.*, 2007), so variations in relative abundances will be disproportionate and give a false impression of community structure. This could be accounted for using the genome sizes of all organisms present, however currently this information is not available for complex communities. This issue is exacerbated in eukaryotes, where not only are genomes typically much longer, but the frequency of genes and the functional complexity are not correlated with genome length; this is known as the C-value paradox (Thomas, 1971).

Planctomycetes, Rhodobacterales and Rhodospirillales decreased after both treatments. These are typically aquatic bacteria, and Rhodospirillales can use sulphide or hydrogen as an electron donor (sulphide is produced by sulphate reducing bacteria typically under anaerobic conditions (Barton, 1995), although they can function aerobically (Kjeldsen *et al.*, 2004; Muyzer and Stams, 2008)). It could therefore be expected that Planctomycetes, Rhodobacterales and Rhodospirillales populations would increase in response to flooding due to the anoxic conditions and availability of reduced substrates. To gain a better understanding of these results, the chemical properties of the soil need to be studied throughout the experiment.

Many of the greater fold changes in relative abundances were attributed to mammals and insects, for example: Carnivora, Lagomorpha, Coleoptera, Hemiptera and Phthiraptera. In the case of the mammals, this is likely to be from residual DNA in the soil, such as from skin. The soil was sieved and homogenised prior to the experiment, and no insects were observed. While some invertebrates are microscopic, caution

should be taken before conclusions are made about these orders; the DNA observed could, like the mammals, be residual rather than actual reflections of populations responding to the treatments.

Genes involved in cell growth (RNA polymerase sigma-70 factor), cell signalling (bacterial cAMP signalling) and membrane transport (ABC transporters) decreased after both treatments. The reduction in carbon input due to the removal of plants would restrict growth. Cell signalling is most beneficial when bacterial cell densities are at their highest (Darch *et al.*, 2012), so a reduction is expected as the carbon reduction and water stresses perturb populations. The reduction in membrane transport genes could due to the the sieving, homogenisation and removal of plants reducing the amount of extracellular compounds available for cell uptake, thus favouring species adapted to relatively lower nutrient environments (than *in situ* pasture soils).

Genes involved in flagellum motility and bacterial chemotaxis increased in response to three floods, but not one flood, suggesting a possible link between flooding frequency and bacterial mobility. Flooding changes the chemical composition of soil, prompting chemotaxis (Bren and Eisenbach, 2000). Transcriptomics would be advantages here to determine which genes are being expressed, rather than just observing which are present. As technology develops, studying mRNA allows for more accurate conclusions to be made. For example, the changes mentioned above may be caused by factors unrelated to mobility that decrease the abundance of organisms, and thus DNA, that utilise flagellar. Our data alludes to this complication, as the abundance of DNA involved in transcription regulation and gene expression appears to decrease in response to either one flood or both treatments. Observing the abundance of mRNA would allow us to determine if gene expression is actually decreasing.

Several genes involved in broader functions, i.e. metabolism, fatty acid metabolism, anaerobic carbon monoxide metabolism, pathogenesis and protection, have varied results, thus broad conclusions cannot be made for these functions. Instead, our results indicate more specific responses to varying flooding frequencies that could be used as a basis for future, more targeted studies.

3.5.4 Conclusion

Some impacts of increasing flooding frequency on microbial communities, and their functions, were quantified. Communities appear to significantly differ when they have received additional floods, and functional changes reflect this. Many differences identified relate to the reduction and oxidation of substances associated with anoxia. Changes were not observed in methanogen populations, therefore as long as water drains between floods, an increase in flooding frequency is not expected to increase CH_4 emissions.

Conducting a laboratory experiment allows for variables to be controlled and specific mechanisms tested acutely. To more accurately represent environmental applications, further experiments in the field need to be conducted to investigate the impacts of flooding on *in-situ* communities. Some key advantages of this would be 1) the lack of additional anthropogenic soil disturbance, 2) the inclusion of plants that act as a carbon source (among many other things), and 3) the inclusion of diurnal variations in environmental factors such as temperature and light irradiance.

4 The effects of increased flood duration on pasture microbial ecosystems, carbon dioxide fluxes and methane fluxes.

4.1 Abstract

The impacts of flooding duration on microbial communities, CO_2 fluxes and CH_4 fluxes in a pasture were investigated. Carbon dioxide and CH_4 fluxes were measured continuously using novel techniques to ascertain some of the actual functional responses. Significant changes were identified in the gas fluxes shortly after the flooding periods, with CO_2 fluxes decreasing and CH_4 fluxes increasing after receiving a longer flood. These results suggest that increased flood durations ultimately decrease the global warming potential of CO_2 and CH_4; the greater sequestration of CO_2 by recovering plants in moist soil outweighs the reduction in CH_4 uptake due to increased methanogenesis and/or decreased CH_4 oxidation. A microbial response to increased flooding duration was not observed, likely due to experimental limitations; namely practical limits inhibiting the number of replicates required to observe responses in *in situ* complex communities.

4.2 Introduction

4.2.1 Microbial ecosystems and flooding

The predicted increased in extreme winter and spring precipitation events in the UK (Collins *et al.*, 2013; Houghton, 2001; Kirtman *et al.*, 2013; Kleinen and Petschel-Held, 2007; Min *et al.*, 2011; Murphy *et al.*, 2009; Trenberth, 1999) will result in an increase in the frequency and duration of floods (Kurnik *et al.*, 2012). These predictions have been supported by the analysis of historical data (Jones *et al.*, 2013; Min *et al.*, 2011; Osborn *et al.*, 2000). For the medium emissions scenario (see Nakićenović *et al.*, 2000), the most likely greatest increase in winter precipitation will be 33% by 2050. This is expected to intensify further by 2080.

Along with other biotic and abiotic perturbations, anoxic conditions resulting from flooding will affect soil properties and ecosystems (Ponnamperuma, 1984; Stams and Plugge, 2010). Zhou *et al.* (2002) reported that soils saturated in water have reduced bacterial diversities. Microorganisms dominate most biogeochemical cycles, and alterations to community structure and function may result in changes to these cycles.

As the frequencies of extreme weather conditions are predicted to increase, it is necessary to understand how these changes will affect ecosystems and their functions.

Anoxia in bulk soil resulting from flooding reduces aerobic respiration while allowing anaerobic organisms to thrive. Such organisms include methanogens (Conrad, 2007) and denitrifiers (Zumft, 1997), which both affect greenhouse gas fluxes. Methanogens are strict anaerobes, producing methane from acetate and/or hydrogen. Methane is the second most important greenhouse gas after CO_2 in terms of global warming effects; the 100-year global warming potential (GWP) is 34 times greater than CO_2 (Myhre *et al.*, 2013), therefore an increase in emissions from more flooding events could play a role in a positive feedback cycle of climatic warming. Pall *et al.* (2011) estimated that twentieth century anthropogenic greenhouse gas emissions increased the risk of flooding that occurred in England and Wales in Autumn 2000. Nine out of 10 climate model simulations suggested a 20 % increase and two out of three simulations suggested a 90 % increase.

Some studies research the effects of flooding on microbial ecosystems using targeted approaches. Studying four sites with varying flooding patterns along a river, Bodelier *et al.* (2012) discovered that the abundance of methanotrophs increased with the increase in flooding using denaturing gel gradient electrophoresis (DGGE) and phospholipid fatty acid analysis (PLFA). Kemnitz *et al.* (2004) identified an increase in methanogen diversity in samples from the same river using terminal-restriction fragment length polymorphism (T-RFLP). Unger *et al.* (2009) found that flooding decreased the bacteria:fungi ratio using PLFA. These studies provide a useful insight into the effects of flooding on microbial diversity and community composition; however it is clear that a deeper understanding of the impacts of environmental stressors on the whole community is required. Furthermore, a gene-orientated analysis is required to understand the functional responses to flooding in a pasture field.

4.2.2 Metagenomics

The advent of high throughput sequencing allows for a much deeper study of microorganisms, which is not limited to targeting specific organisms or genes. Using new sequencing technology, DNA fragments obtained from an environmental sample can be sequenced and either annotated using a DNA sequence database, or clustered together to identify similar sequences. This approach allows for the discovery of which organisms are present, which functions they are capable of performing, and which

biochemical pathways are used in these functions. With these technologies, new questions about how environmental change will affect microbial communities can be investigated.

Certain limitations need to be considered for metagenomic studies. Organisms or genes can only be reliably annotated if they are present in reference databases, meaning that sequences not matching a reference will either be unidentified, or possibly incorrectly annotated. Furthermore, sequencing errors such as base substitution can increase the chance of incorrectly annotating a sequence. These caveats will be reduced as more sequences are added to the databases and as sequencing technologies improve. Organism or function abundances are calculated from the number of sequences annotated, thus genome lengths can bias abundance values. Prokaryotic genome length typically range from 0.6 to 10 megabases (Saint-Girons and Cole, 1999), and Nayfach and Pollard (2014) found that genome lengths in the human gut produced abundance biases. Metagenomics provides a powerful insight into complex microbial diversities, and as long as these limitations are considered, hypotheses can be tested that would not be possible using previous technologies.

4.2.3 Carbon dioxide fluxes and flooding

Carbon dioxide fluxes in pasture fields are predominantly regulated by plant photosynthesis and plant/microbial aerobic respiration, and to a small extent the oxidation of carbon compounds (Raich and Schlesinger, 1992). Photosynthesis uptakes CO_2 and produces O_2, and aerobic respiration is *visa versa*. Many factors affect these processes, including temperature (Berry and Bjorkman, 1980; Lloyd and Taylor, 1994; Liu *et al.*, 2006), humidity (Rawson *et al.*, 1977; Leach, 1979; Bunce, 1984), species type and biomass (Raich and Schlesinger, 1992; Wang *et al.*, 2003), glucose/H_2O supply and O_2/CO_2 supply (Wang *et al.*, 2003; Peterson and Lajtha, 2013; Wei *et al.*, 2015). Photosynthetically active radiation (PAR) (Caldwell, 1981) and nutrient supply (e.g. N, P, K, Mn) (Tissue *et al.*, 1993; Rengel, 1999) also affect photosynthesis. As PAR and temperature vary diurnally, so does photosynthesis (D R Geiger and Servaites, 1994; Parkin and Kaspar, 2003; Bernacchi *et al.*, 2006) and respiration (Tang *et al.*, 2005). During the day, photosynthetic rates increase as PAR and temperature increase. At night, photosynthesis decreases when PAR and temperature decrease. Seasonal variations are also observed for the same reasons (Field *et al.*, 1998; Lavigne *et al.*, 2004), with greater photosynthetic rates in summer than in winter.

Flooding affects both respiration and photosynthesis. Once all the O_2 is consumed within saturated bulk soil, respiration switches from aerobic to anaerobic, decreasing the emission of CO_2. Drainage and aeration of the soil reverts respiration back to aerobic. Most non-aquatic plants and plants lacking aerenchyma will begin to die after prolonged submersion in a flood (Pereira *et al.*, 1986; Kennedy *et al.*, 1992). Not only does this decrease photosynthesis and therefore CO_2 assimilation, the plant material will be decomposed, producing CO_2 and CH_4. Kelly *et al.* (1997) observed an experimentally flooded reservoir shift from a carbon sink to a carbon source. Studying gas fluxes in response to flooding is complicated by temporal delays in observations of fluxes. Miyata *et al.* (2000) attribute their observed increase in CO_2 emissions from drained paddy soil to the removal of the gas diffusion barrier that restricts emission.

See Chapter 1.1.3 for an introduction to methane fluxes and flooding.

4.2.4 Aim and hypotheses

The aim of this study is to investigate how an increase in flood duration on pasture soil affects microbial communities, their potential functions, and CO_2 and CH_4 fluxes. Due to the large quantity of taxonomic and gene abundance data generated, and continuous gas fluxes measurements, several hypotheses can be tested. It is hypothesised that increased flooding duration on pasture soils will significantly change a) CO_2 and CH_4 fluxes, b) the biodiversity of microbial communities, c) the Bacteria:Archaea abundance ratio, d) microbial community composition and potential function, e) the abundances of taxa and functional genes in microbial communities, and f) the abundance of genes involved in methane production, methane oxidation, and sulphate reduction.

4.3 Methodology

4.3.1 Experimental design and treatment

Eight gravity-fed lysimeters (30 cm height x 20 cm diameter) containing soil cores, taken from a single homogenous area of unimproved pasture at the experiment site, were placed linearly in a pasture field one year before the experiment began (Figures 4.1 & 4.2). The short and long flood duration lysimeters were randomly paired in four groups to reduce any site gradient effects. The site is located at the confluence of two rivers (lat 51.044770 lon -2.111945) in Wiltshire, UK. The field is known to typically flood once a year, at least partially by the river confluence. The soil association is

Wickham 2: fine loamy over clayey soil (Figure 4.3, Supporting Information A.3.1 (National Soil Resources Institute (NSRI), 2013). The vegetation in the field is predominantly Graminoid, along with Trifolium and Taraxacum. A hose and tap controlled drainage (Figure 4.4). Lysimeters were placed 30 cm apart in the ground. Access shafts for the hoses and taps were drilled to a depth of 60 cm, 25 cm away from each lysimeter (Figure 4.5).

Flooding was induced by filling the lysimeters with natural spring water. After eight days, the short flood lysimeters were drained and left free draining. After 42 days, the long flood lysimeters were drained.

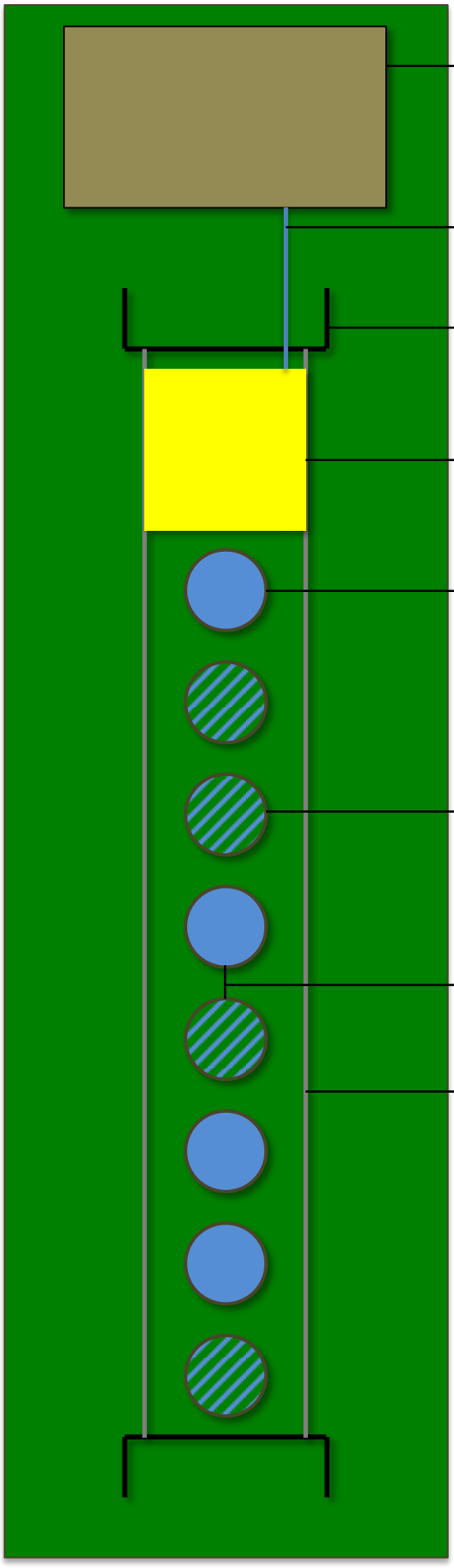


Figure 4.1. The experimental plot design. Note that there were other lysimeters present along the line, not drawn here, which were part of another experiment.

Not to scale.

Figure 4.2. The field site.

Figure 4.3. The soil profile. The top 7 cm is the AO horizon, a soil layer with high organic content, and this is followed by the A horizon where organic materials and minerals are mixed.

Figure 4.4. *In situ* lysimeter. A lysimeter and access shaft with the cover in place.

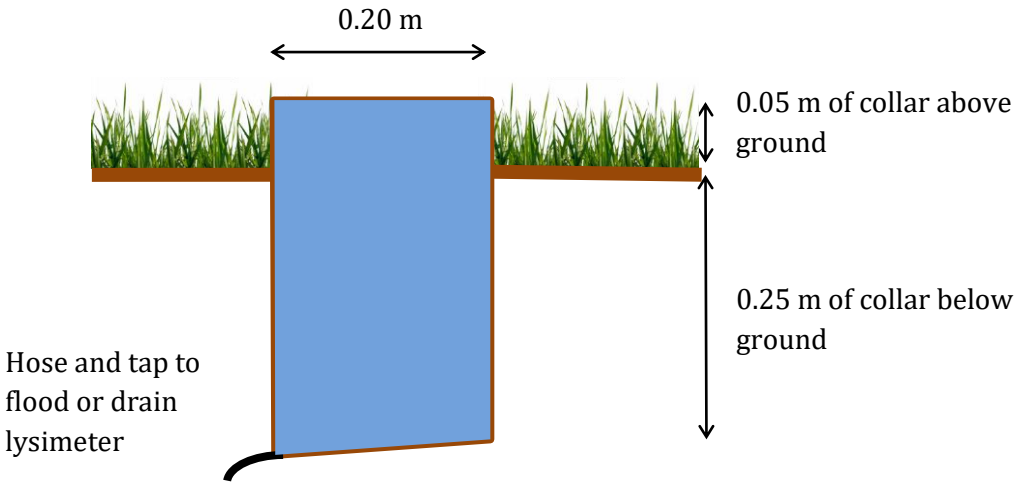


Figure 4.5. A graphical representation of a lysimeter under flooded conditions.

4.3.2 CO_2 and CH_4 flux measurements and analysis

Carbon dioxide and CH_4 flux measurements were recorded using a LiCor LI-8100 infra-red gas analyser and a Los Gatos Research Fast Greenhouse Gas Analyser (Model number 901-0010), respectively, deployed on a SkyLine system (Figure 4.6). Measurements were automatically taken for five minutes per plot every three hours, from the beginning of the experiment until ca. six weeks after drainage. A transparent measurement chamber was used to allow light to pass and reduce interference on photosynthesis. Due to a technical error, a large gap occurred between 23rd March and 28th April. Fluxes were therefore analysed separately as two distinct date groups, 4th March to 23rd March and 28th April to 13th May.

Carbon dioxide measurements were converted from µmol m^2 sec^{-1} to mg m^2 hr^{-1}. Fluxes were tested for changes over time using repeated measures ANOVAs and responses to flooding treatments using t-tests. Differences between cumulative fluxes were tested using t-tests. The impacts of treatment on the 100-year global warming potential (GWP) were calculated by comparing the differences in total gas flux; a 100-year GWP of 34 CO_2 equivalent (CO_2e) was used for methane (Myhre *et al.*, 2013)

Figure 4.6. SkyLine. The SkyLine system used to measure CO_2 and CH_4 flux. The unit carrying the gas analyser moves along the cables and stops above each lysimeter, where the chamber descends and attaches to the lysimeter to record a measurement.

4.3.3 Soil sampling and DNA extraction

At the start (03/03/2015) and the end (13/04/2015) of the experiment, three soil samples were extracted randomly from each lysimeter at a depth of 10 cm using a 2 cm soil corer. Randomisation was achieved using a co-ordinated 2 cm grid and a random number generator. DNA was extracted from the soil using a PowerSoil® DNA Isolation kit (250 mg) (Mo Bio Laboratories Inc., Carlsbad, CA, USA) following the manufacturer's protocol. DNA concentration and purity was established using a

NanoDrop™ 1000 UV-Vis Spectrophotometer (Thermo Scientific, Waltham, MA, USA). The DNA was stored at -80 °C until DNA library preparation.

4.3.4 DNA library preparation

The DNA replicates for each lysimeter were pooled and quantified using a Qubit dsDNA HS assay kit (Life Technologies Corporation, Carlsbad, CA, USA). DNA libraries were prepared by Sally James (University of York) using a TruSeq Nano DNA LT library prep kit (Illumina UK, Little Chesterfield, UK), following the manufacturer's guidelines and recommended protocol for library insert size of ~550 bp. 200 ng gDNA in a total volume of 52.5 µl was fragmented using a M220 sonicator (Covaris Ltd., Brighton, UK) with the following settings: Duty factor: 20 %; Peak Incident power: 50W; Cycles / burst: 20; Duration: 45 Seconds; Temperature – 20 °C.

Library quality was assessed by running 1 µl on a Bioanalyzer DNA 7500 chip (Agilent Technologies, Inc., Santa Clara, CA, USA) and concentrations determined using the Qubit dsDNA HS assay kit. Samples were diluted to 10 mM concentrations and pooled at equimolar ratios. The libraries were sequenced at the Bauer Core Facility, Harvard University (MA, USA) using a HiSeq 2500 Sequencer (Illumina UK) with the V3 kit (125 bp).

4.3.5 Sequence processing and analysis

Residual adapter sequences and sequences shorter than 30 bp were removed from the raw sequences with Cutadapt (Martin, 2011). Remaining paired-ends were assembled with Megahit (Li *et al.*, 2015), using a k-mer length of 27. The minimum contig length was 200 bp. The Burrows-Wheeler Alignment tool (BWA) (Li and Durbin, 2009) was used to map sequences to the contigs and SAMtools (Li *et al.*, 2009) was used to extract mapped and unmapped sequence identity, and abundances in contigs. Megahit does not provide this information, thus BWA and SAMtools provide the best indication of contig composition and abundance. Unassembled paired-ends were merged using PEAR (Zhang *et al.*, 2014). Remaining singleton sequences that were neither assembled nor merged, and singletons generated from Cutadapt, were concatenated with the contigs and the merged paired-ends to produce a single fasta file. Sequence abundances were appended to the sequence IDs. Sequence length and quality statistics were established using Biopython (Cock *et al.*, 2009). For contigs, the N50 length was calculated as quality information was not available. The N50 length is the length at

which all contigs of that length or greater contain at least half of the sum of all the contigs' lengths.

The processed sequence files were upload to MG-RAST (Meyer *et al.*, 2008) for annotation. Low quality sequences that have more than five bases with a phred score lower than 15 were excluded from analysis. The remaining sequences were annotated with a "representative hit" annotation technique, which selects a single, unambiguous annotation for each feature, using the RefSeq database for taxonomic identification and Subsystems for function assignment. The maximum e-value cut-off was $1e^{-15}$, providing a strict search parameter, the minimum identity cut-off was 60 %, and the minimum alignment length cut-off was 20 bases. Taxonomic and functional annotation abundance data were downloaded from MG-RAST for further processing. Taxa and functions with a total abundance below five across all samples were removed as confident conclusions cannot be drawn for such low representations. Relative abundance values were generated and arc-sin square root transformed. Raw abundance values were square root transformed to calculate Bacteria:Archaea ratios. Any means are reported with standard deviation.

Order was the lowest taxonomic level investigated, as the taxa annotation error rate may be too high below this (Randle-Boggis *et al.*, 2016). The α-diversity of each sample was calculated using the Shannon index, an abundance-weighted average of the logarithm of the relative abundances of species. Treatment dissimilarities were tested with Analysis Of Similarity (ANOSIM, 100,000 permutations), Principal Coordinates Analysis (PCoA) and hierarchical clustering, all using the Bray-Curtis dissimilarity method. Significant differences in order and functional relative abundances were determined using independent t-tests. P-values were corrected for multiple comparisons using the Benjamini-Hochberg procedure.

4.4 Results

4.4.1 CO_2 and CH_4 fluxes

CO_2 and CH_4 fluxes did not change significantly over time during flooding, but they did after drainage (Table 4.1, Figures 4.7 & 4.8). The fluxes were not significantly different between treatments during the first date period, but they were during the second (Table 4.1).

The total quantities of CO_2 and CH_4 emitted did not differ significantly between the treatments after the first recorded time period, but they did after the second (Table 4.1, Figures 4.9 & 4.10). This resulted in a GWP of CO_2 319.3 ± 221.3 CO_2e lower after the longer flood for the second time period, compared to the shorter flood. The GWP of CH_4 increased by 0.111 ± 0.098 CO_2e.

Table 4.1. CO_2 and CH_4 statistical tests results. Significant p values (< 0.05) are highlighted in bold.

Date range	Gas	Test	Results
04/03/2015 – 23/03/2015	CO_2	Flux, Repeated measures ANOVA	F = 0.49, df = 1, p = 0.483
28/04/2015 – 13/05/2015	CO_2	Flux, Repeated measures ANOVA	F = 9.17, df = 1, **p = 0.003**
04/03/2015 – 23/03/2015	CH_4	Flux, Repeated measures ANOVA	F = 0.47, df = 1, p = 0.493
28/04/2015 – 13/05/2015	CH_4	Flux, Repeated measures ANOVA	F = 23.38, df = 1, **p < 0.000**
04/03/2015 – 23/03/2015	CO_2	Flux vs. treatment, t-test	Short flood μ = 0.149 g CO_2 m^2 hr^{-1}, Long flood μ = 0. 140 g CO_2 m^2 hr^{-1}, t = -0.41, df = 1049, p = 0.680
28/04/2015 – 13/05/2015	CO_2	Flux vs. treatment, t-test	Short flood μ = 0.170 g CO_2 m^2 hr^{-1}, Long flood μ = -0.076 g CO_2 m^2 hr^{-1}, t = -4.79, df = 998, **p < 0.000**
04/03/2015 – 23/03/2015	CH_4	Flux vs. treatment, t-test	Short flood μ = -0.016 mg CH_4 m^2 hr^{-1}, Long flood μ = -0.015 mg CH_4 m^2 hr^{-1}, t = 0.36, df = 994, p = 0.720
28/04/2015 – 13/05/2015	CH_4	Flux vs. treatment, t-test	Short flood μ = -0.045 mg CH_4 m^2 hr^{-1}, Long flood μ = -0.020 mg CH_4 m^2 hr^{-1}, t = 7.76, df = 984, **p < 0.000**

Date range	Gas	Test	Results
04/03/2015 – 23/03/2015	CO_2	Cumulative flux, t-test	Short flood μ = 241.7 ± 122.1 kg CO_2 m^2, Long flood μ = 213.7 ± 52.7 kg CO_2 m^2, t = 0.33, p = 0.756
28/04/2015 – 13/05/2015	CO_2	Cumulative flux, t-test	Short flood μ = 229.9 ± 80.6 kg CO_2 m^2, Long flood μ = -90.6 ± 141.3 kg CO_2 m^2, t = 5.62, **p = 0.001**
04/03/2015 – 23/03/2015	CH_4	Cumulative flux, t-test	Short flood μ = -25.3 ± 14.0 g CH_4 m^2, Long flood μ = -24.7 ± 3.5 g CH_4 m^2, t = 0.06, p = 0.952
28/04/2015 – 13/05/2015	CH_4	Cumulative flux, t-test	Short flood μ = -57.2 ± 8.8 g CH_4 m^2, Long flood μ = -24.7 ± 19.9 g CH_4 m^2, t = -5.22, **p = 0.002**

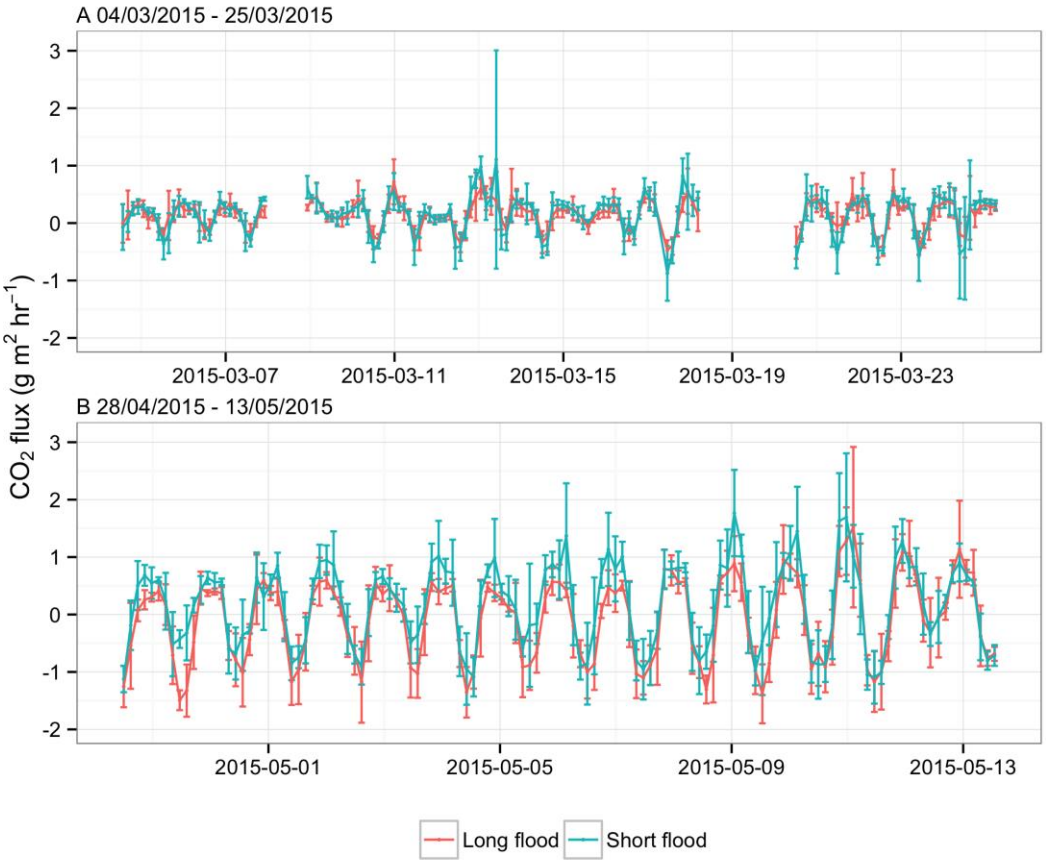


Figure 4.7. CO_2 fluxes. CO_2 fluxes for the two treatments between a) 04/03/2015 and 23/03/2015, and b) 28/04/2015 and 13/05/2015. The gaps between the 8th and 9th, and the 18th and 21st March are due to technical problems with the SkyLine device.

Figure 4.8. CH_4 fluxes. CH_4 fluxes for the two treatments between a) 04/03/2015 and 23/03/2015, and b) 28/04/2015 and 13/05/2015. The gaps between the 8th and 9th, and the 18th and 21st March are due to technical problems with the SkyLine device.

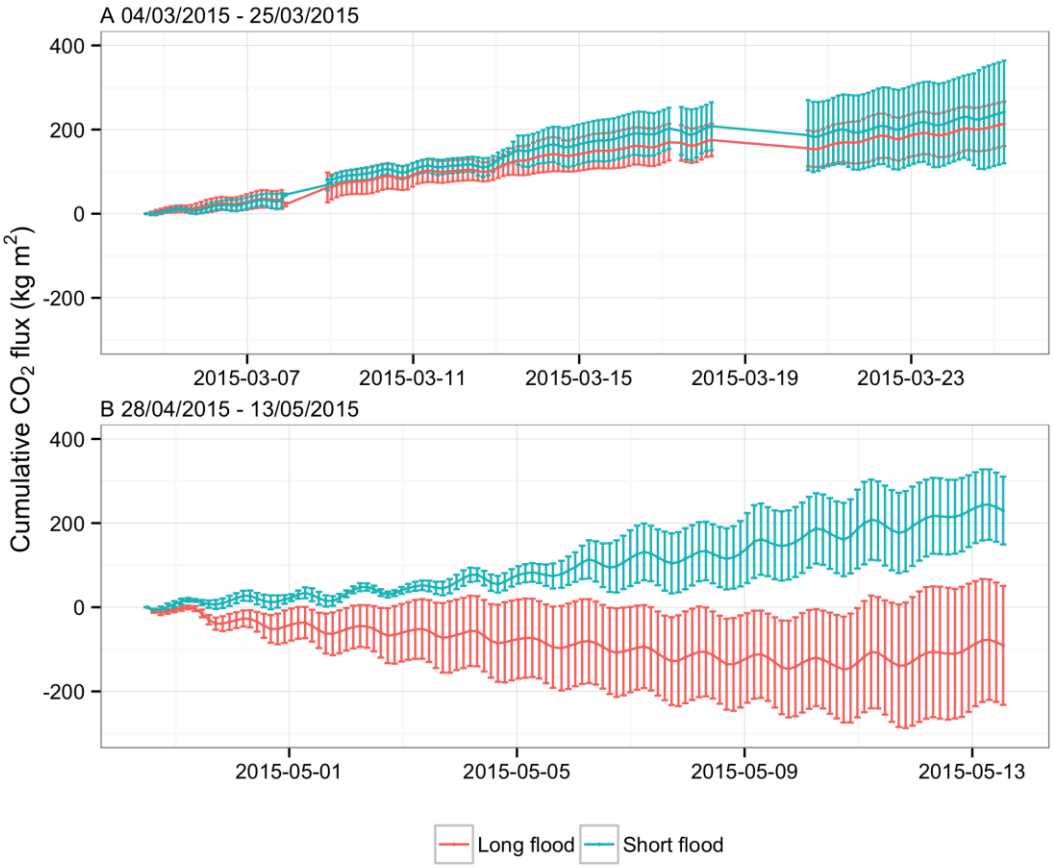


Figure 4.9. Cumulative CO_2 flux. The cumulative flux between a) 04/03/2015 and 23/03/2015, and b) 28/04/2015 and 13/05/2015. The gaps between the 8th and 9th, and the 18th and 21st March are due to technical problems with the SkyLine device.

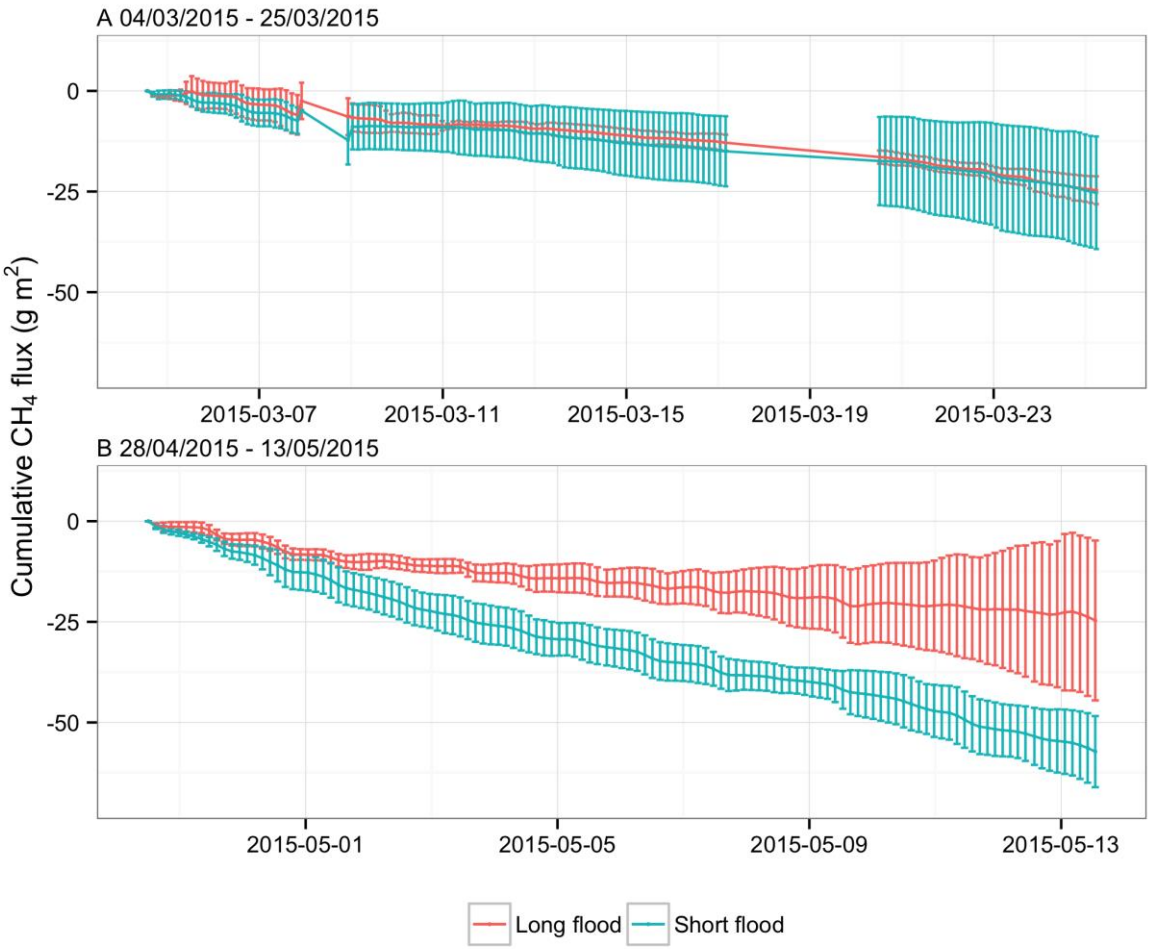


Figure 4.10. Cumulative CH_4 flux. The cumulative flux between a) 04/03/2015 and 23/03/2015, and b) 28/04/2015 and 13/05/2015. The gaps between the 8th and 9th, and the 18th and 21st March are due to technical problems with the SkyLine device.

4.4.2 Sequencing

239,854,566 paired-end sequences were generated, totalling 6.0 x 10^{10} bases. The mean sample paired-end sequence count was 14,990,910 ± 6,564,511 (Table A.16). The mean phred quality score was 34.51 and 87.72 % of bases had a phred score of 30 or greater. 239,115,348 sequences remained after removing residual adapter sequences and sequences shorter than 30 bp, resulting in a mean median length of 125, a mean phred score of 34.55 and 87.84 % of bases with a phred score of 30 or greater (Table A.17). 4,364,713 contigs were generated, with 138,491,929 sequences being mapped to contigs. The mean contig N50 length was 626 bp. The maximum contig length was 71.6 kb (Table A.18). A mean of 313,104 (± 115,375) paired end sequences were merged (Table A.19), with a mean phred score of 37.8 and 97.72 % of bases

having a phred score of 30 or greater. A mean of 10,625,195 (± 3,425,901) singleton sequences remained, with a mean phred score of 34.45 and 87.52 % of bases having a phred score of 30 or greater (Table A.20). The final concatenated fasta files contained a mean of 21,836,288 (± 7,220,504) sequences, with a mean median length of 125 (Table A.21).

4.4.3 Diversity and Bacteria:Archaea ratio

There were no significant differences in the order α-diversities between treatments at the start (Short Flood: 4.78 ± 0.02, Long Flood: 4.79 ± 0.02; t-test, t = 0.72, df = 5.56, p = 0.499) and at the end (Short Flood: 4.77 ± 0.02, Long Flood: 4.78 ± 0.01; t-test, t = 0.92, df = 4.86, p = 0.404) of the experiment. Time did not significantly affect α-diversity (Start: 4.79 ± 0.02, End: 4.77 ± 0.1; t-test, t = -1.48, df = 12.34, p = 0.163).

The Bacteria:Archaea ratio was not significantly different between treatments at the start (√ transformed (n:1): Short Flood: 10.21 ± 0.40, Long Flood: 10.30 ± 0.29; t-test, t = 0.35, df = 5.52, p = 0.739) nor at the end (√ transformed (n:1): Short Flood: 10.42 ± 0.38, Long Flood: 10.64 ± 0.17; t-test, t = 1.03, df = 4.13, p = 0.360 of the experiment, and time did not significantly affect it either (√ transformed (n:1): Start: 10.25 ± 0.33, End: 10.53 ± 0.30; t-test, t = 1.77, df = 13.87, p = 0.099).

4.4.4 Sample dissimilarity

The PCoA and hierarchical clustering suggest that flooding duration does not have an impact on the microbial community taxonomic composition (ANOSIM, R: -0.058, p = 0.732) and function (ANOSIM, R: 0.004, p = 0.451) (Figures 4.11-4.14).

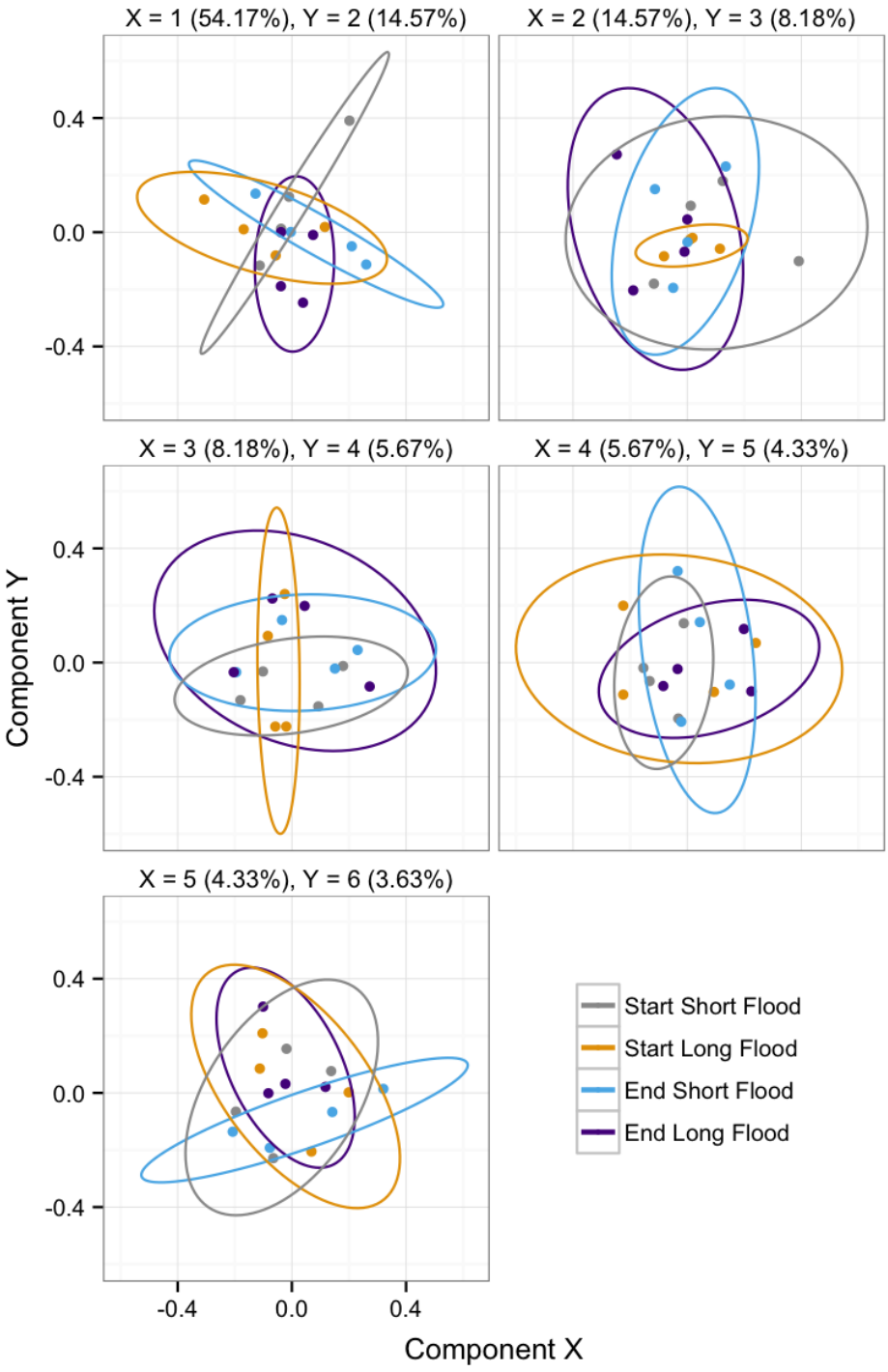


Figure 4.11. Order PCoA. A PCoA of the relative abundance of orders (Bray-Curtis distance method). Each panel plots different components and the proportion of variations explained by each component is displayed in parentheses in the panel title.

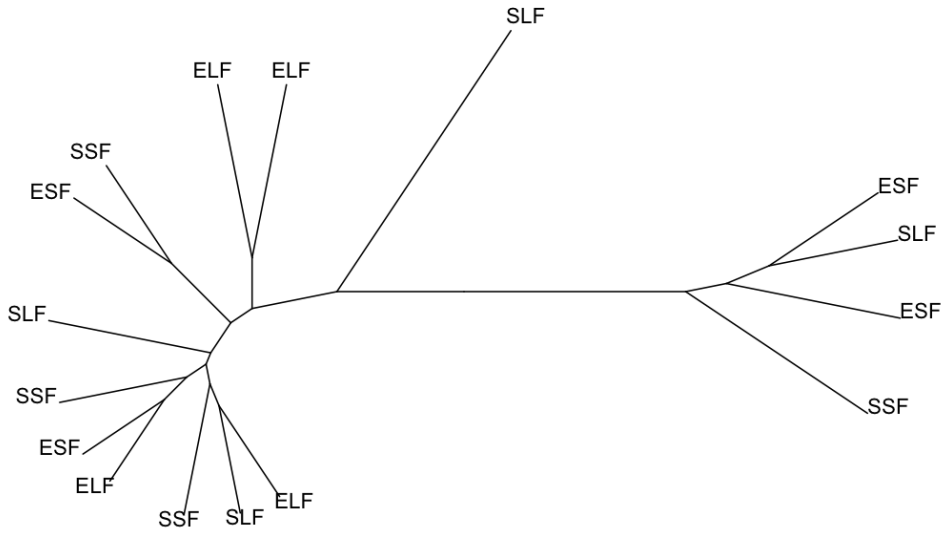


Figure 4.12. Order hierarchical clustering. A hierarchical clustering analysis of the community composition at the order level (Bray-Curtis distance method). Key: S = Start, E = End, SF = Short flood, LF = Long flood.

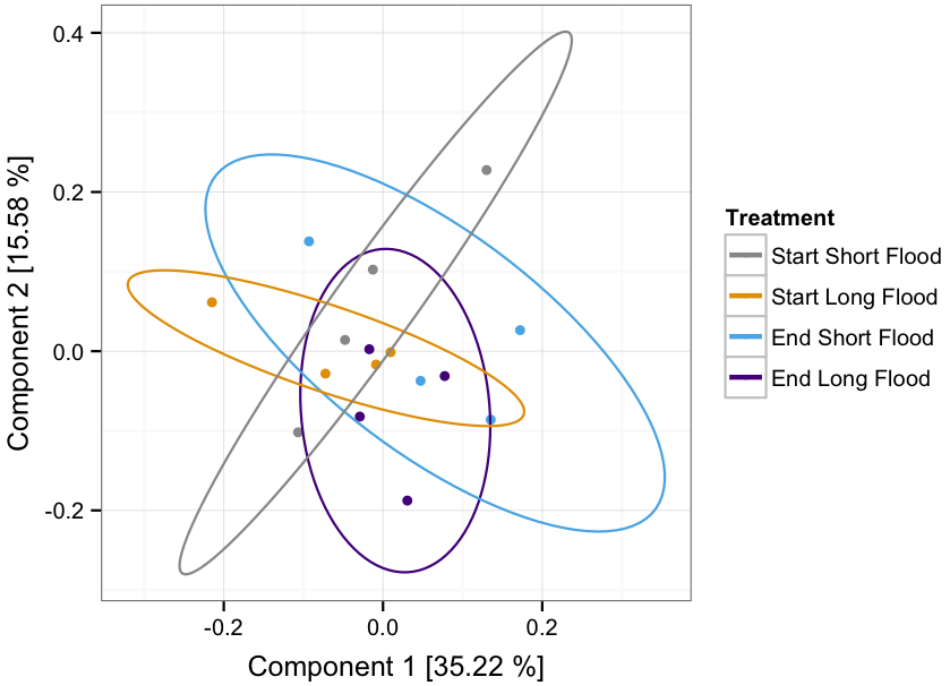


Figure 4.13. Functional PCoA. A PCoA of potential functions (Bray-Curtis distance method).

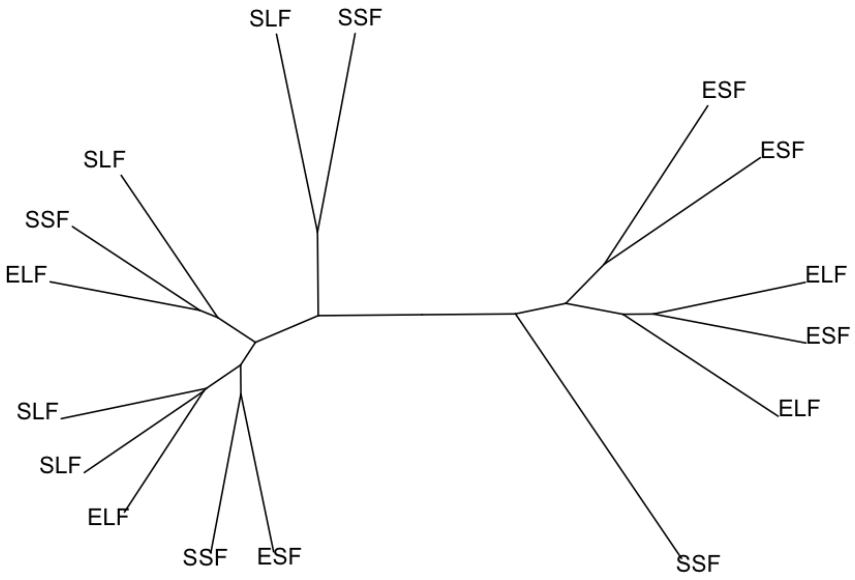


Figure 4.14. Functional hierarchical clustering. A hierarchical clustering analysis of potential functions (Bray-Curtis distance method). Key: S = Start, E = End, SF = Short flood, LF = Long flood.

4.4.5 Taxonomic and functional abundances

Sixty-one phyla were identified. The most abundant phyla across the samples were Proteobacteria, Actinobacteria and Acidobacteria (Figure 4.15), all dominant soil phyla (Janssen, 2006).

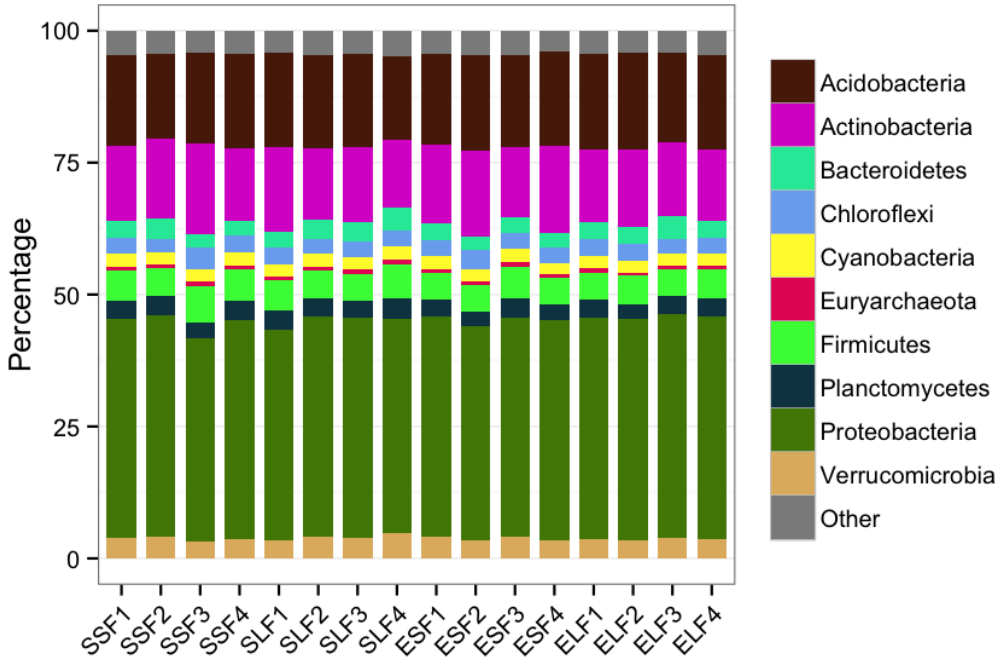


Figure 4.15. Phyla relative abundances. The relative abundances of phyla in each sample. The legend displays the 10 most abundance phyla. Key: S = Start, E = End, SF = Short flood, LF = Long flood.

There were no significant differences between the order relative abundances of bacteria, archaea, eukaryota or level 2 functions when subjected to the different flood durations.

Of the 110 bacteria orders detected in the samples, 84 populations decreased in relative abundance in response to the short floods and 95 decreased in response to the long floods. Sphingomonadales, Rhodobacterales, Caulobacterales and Xanthomonadales increased in response to the long floods but decreased after the short floods. Of the 10 greatest variations between order relative abundance fold changes, one was due to an increase in response to a longer flood (Sphingomonadales, 12.96 % greater), and the rest were due to a decrease (Table A.22).

Nineteen archaea orders were detected, with nine and seven mean populations decreasing in response to short and long flooding, respectively (Table A.23). Facultative anaerobic Sulfolobales, which reduce sulphur under anaerobic conditions, increased in response to both treatments. Desulfurococcales, also facultative anaerobes that reduce sulphur, nitrate and nitrite under anaerobic conditions, increased in response to the long flood but decreased with the short flood. Methanogenic orders appear to vary in response to the treatments, with Methanopyrales and Methanocellales increasing under the two conditions and Methanobacteriales, Methanomicrobiales, Methanococcales and Methanosarcinales decreasing.

For eukaryota, 204 were detected, with 146 and 126 mean populations decreasing in response to short and long flooding, respectively. Of the 167 level 2 functions detected, 92 and 96 decreased in relative abundance in response to the short floods and the long floods, respectively.

4.5 Discussion

The impacts of increased flooding duration on microbial communities and CO_2 and CH_4 fluxes were investigated. Using a novel method, SkyLine, gas fluxes were recorded continually, 24 hours a day in approximately three-hour cycles. This provided a high temporal resolution of flux responses to treatments. Relatively large quantities of DNA material were obtained using high throughput sequencing, allowing a confident observation of the community.

4.5.1 CO_2 and CH_4 fluxes

Carbon dioxide and CH_4 fluxes did not change significantly in response to flooding manipulation, aside from diurnal shifts. However, after the flooding manipulation ended, fluxes altered significantly, and they varied between treatments too. This delayed response could be due to several factors, such as the time taken for the plants and microbial communities to respond to treatment and for changes in gas fluxes within the soil to become detectable. Water saturation produces a diffusion barrier that restricts gas emission (Ponnamperuma, 1984; Miyata *et al.*, 2000); drainage removes this barrier and releases gasses built up during the flood. Grable (1966) found gas movement to be 10^4 fold slower in flooded soils than in aerated soils. Further, shifts in methane metabolism may be delayed due to the redox cascade time period (Wang *et al.*, 1993; Mohanty *et al.*, 2013). As these responses occurred later than anticipated, there is no microbial or genetic data to discern the cause. These findings highlight the

importance of considering long-range temporal scales when modelling gas flux responses.

The diurnal variation in CO_2 flux is typical and due to changes in photosynthetic rates throughout the day; this is greater during the later months when hours of daylight increase. During daylight, photosynthetic rates increase beyond the rate of respiration and thus sequester CO_2. At night, when light levels are low, photosynthesis is inhibited and respiration produces a net emission of CO_2. Cumulatively, both treatments were net sources of CO_2 during the flooding, however the long flood plots became net sinks approximately two months after treatments were applied; between 28th April and 13th May 2015, 229.9 kg m^2 of CO_2 were emitted from the short flood plots and 90.6 kg m^2 were sequestered in the long flood plots. Plants may become damaged or die during the flooding (Kennedy *et al.*, 1992; Pereira *et al.*, 1986), more so under longer floods. Kelly *et al.* (1997) attribute the conversion of an experimentally flooded reservoir from a carbon sink to a carbon source to the death of plants reducing photosynthesis and increasing decomposition; their flood period lasted over four months and over a much greater spatial scale than studied here, allowing more death and decomposition. After drainage, plant growth and recovery would assimilate more CO_2 (Smith and Stitt, 2007). Soil moisture would also be greater after the longer floods, enhancing growth and CO_2 uptake. To investigate this further, regular soil moisture content measurements should be made; due to the location of the site, regular manual measurements could not be taken, and automated measurements within the lysimeters would disturb the sample site or interfere with SkyLine if placed externally.

A net consumption of CH_4 was observed in the second time period, with the short flood plots consuming more than the long floods (-57.2 g m^2 and -24.7 g m^2, respectively). Methane is produced via anaerobic methanogenesis and oxidised via aerobic CH_4 oxidation. The longer floods may increase methanogenesis and/or reduce methanotrophy (Ratering and Conrad, 1998), therefore reducing net CH_4 uptake during and after the floods. Longer floods than those studied may see a net emission of CH_4, as observed in flooded rice paddies (Schütz *et al.*, 1989; Conrad *et al.*, 2008; Zhang *et al.*, 2013). The value of these experiments, however, is learning the impacts of climate change consequences on pasture ecosystems; the duration of flood studied should therefore be realistic to predicted changes. These results show that CO_2 sequestration is greater after longer floods on pasture soil, resulting in a net consumption rather than emission. In terms of global warming potential, the reduction in overall CH_4 assimilation after longer floods is negligible compared to CO_2

sequestration. To expand these results, the impacts of increased flooding duration on other GHGs, such as N_2O, should also be studied.

4.5.2 Microbial communities and functions

Little difference is observed in the microbial community. Background variation in the microbial communities could be caused by many factors, such as minute spatial variation in minerals and nutrients in the soil (Farley and Fitter, 1999), variations in hydrology and leachate flows (Kachanoski *et al.*, 1988), variations in oxygen concentrations and redox potentials (Eriksson *et al.*, 2003; Niedermeier and Robinson, 2007), influxes of nutrients such as nitrogen compounds from animal faeces (Cain *et al.*, 1999; Powell *et al.*, 1999), variations in plant species and the rhizosphere that affect carbon supply and nutrient concentrations (Jobbágy and Jackson, 2004), and disturbance from invertebrates such as earthworms (which in turn affect many of the aforementioned factors) (Lavelle *et al.*, 1998; Duboisset *et al.*, 1999). The complex spatial and temporal heterogeneity (Ettema and Wardle, 2002), both biotic and abiotic, within pasture soils presents a challenging ecosystem for empirical studies. Understanding the mechanisms behind variations in microbial communities will enable an enhanced understanding of large-scale ecosystems to be developed.

Technical and practical constraints limited the number of replicates required to identify treatment responses beyond background variation. These constraints include physical limitations of SkyLine, such as the maximum feasible length of the experimental plot for the machine to operate safely and effectively. Once the success of SkyLine is established, new engineering solutions could allow for larger plots, and thus more replicates, to be studied. For example, a modified version of SkyLine, SkyGas, is currently in development that allows for two-dimensional plots to be studied, rather than one-dimensional linear plots. This increase in spatial scale, combined with the high temporal resolution, would make SkyGas a highly valuable asset for studying ecosystems. Cost can also be a constraint with metagenomic studies, however current technologies already provide a large genetic coverage of complex environmental samples, and advances in technology will see well-replicated projects becoming affordable. Further, greater temporal scales could be studied, extending projects from looking at “start” and “end” communities, to observing how communities respond over time. This is particularly important given the delayed response in gas flux shown here.

Despite these constraints, the experimental design followed standard practices. The number of replicates was sufficient for statistical testing and a random block design removed any location biases. Therefore, the lack of observed microbial response may be due to ecological resilience to change (Holling, 1973). Enhanced spatial and temporal sampling would prove this hypothesis correct or incorrect. Targeted experiments, such as looking at micro-scale variation in methanogen populations, should also be conducted to determine the biological responses underpinning the observed changes in gas flux, without being masked by the vast heterogeneity within the whole community. Combining these two broad groups of study, e.g. impacts of change at micro scales and at macro scales, allows for a holistic understanding of environmental change ecology.

4.5.3 Summary

Increased flood duration affects both CO_2 and CH_4 fluxes in pasture fields, with longer floods sequestering more CO_2 and assimilating less CH_4 after floods have drained. In the context of climate change, this reduces the overall GWP for these gasses, with the increase in CH_4 GWP being negligible compared to the decrease in CO_2 GWP. DNA coverage per sample was sufficient, thus more samples are required to identify the mechanistic responses in microbial communities, and the treatment responses against the myriad of background environmental noise. Further, increased temporal sampling will allow us to discern how taxa, and functions, respond differently over time. In the future, this will be achievable with more spatially capable automated gas measuring systems and cheaper DNA sequencing methods facilitating a greater number of replicates and thus sample for extraction over time. These results reveal a functional response to increased flood duration and further studies should be conducted to identify the mechanisms behind this.

5 Discussion

5.1 Thesis summary

High throughput sequencing and metagenomics provide in-depth observations of microbial communities and their potential functions, allowing researchers to develop complex understandings of the interactions between microbial communities and their environments. As with all new technology however, there are caveats and challenges. This thesis quantifies the performances of current metagenome annotation methods, highlighting pitfalls and caveats associated with certain techniques. Two experimental chapters investigate the impacts of flooding on microbial ecosystems. With increased flooding being an imminent threat to the United Kingdom, and little research published on the impacts of flooding on pastureland microbial ecosystems, it is an important topic to study.

5.1.1 Chapter 2 summary

Due to the vast amounts of complex data available in metagenomes, annotating DNA sequencing accurately and with confidence can be a challenge. The first data chapter in this thesis used a simulated metagenome to quantify the annotation performances of various programs and parameters. Pitfalls associated with certain annotation choices were identified and quantified, providing a guideline of annotation error rates for researchers to consider when designing metagenomic analysis pipelines or interpreting results of published studies. The findings from this chapter were used in the analyses of the two subsequent data chapters.

5.1.2 Chapter 3 summary

The second data chapter investigated the impact of increased flooding frequency on soil microbial communities and potential functions, in line with predicted environmental changes. Additional flooding events altered microbial community composition and diversity, and significant differences between taxonomic groups and functional genes were identified. This laboratory experiment used homogenised soil absent of higher plants. Therefore, the results show the microbial response to increased flooding frequency. To practically apply these results to environments outside of the lab, *in situ* experiments should be conducted. Further, to extend the functional data from DNA, i.e. potential function, to actual functional responses, chemical analyses should be undertaken. The third data chapter does just this.

5.1.3 Chapter 4 summary

The third and final data chapter investigated the impacts of flooding duration on microbial communities in a pasture. Carbon dioxide and CH_4 fluxes were measured continuously using novel techniques to ascertain some of the actual functional responses. Significant changes were identified in the gas fluxes shortly after the flooding periods, with CO_2 fluxes decreasing and CH_4 fluxes increasing after receiving a longer flood. These results suggest that increased flood duration ultimately decreases the global warming potential of CO_2 and CH_4; the greater sequestration of CO_2 by recovering plants in moist soil outweighs the reduction in CH_4 uptake due to increased methanogenesis and/or decreased CH_4 oxidation.

A microbial response to increased flooding duration was not observed. As functional responses were observed in gas fluxes, and many taxonomic and functional gene responses were observed in the laboratory experiment, is it likely that the lack of observed microbial response was due to experimental limitations; namely practical limits with the SkyLine system inhibiting the number of replicates obtainable. Due to the multitude of environmental variables that need to be accounted for in field experiments, several replicates are required to statistically identify ecological responses. Rarefaction curves plotted for both the laboratory and field experiments show that the DNA coverage attained by both the Illumina MiSeq and Illumina HiSeq are great enough to confidently observe the respective microbial communities. Therefore, enhanced sequencing would not produce significantly more accurate results and more biological replicates are required to identify responses *in situ*.

5.2 DNA sequencing and environmental change

The error rates identified in Chapter 2 should not be taken lightly. They suggest that over a third of genera in metagenomic studies may be incorrectly annotated. One large factor contributing to annotation errors is the short read length produced by current NGS technologies. While base call error rates have improved drastically, read lengths under one kilobase are not long enough to cover most genes, thus there is an element of prediction that sequence annotations programs must make. Future NGS technologies, such as nanopore sequencing, aim to remove this *caveat* by sequencing whole strands of DNA. Current error rates remain high, however they are decreasing rapidly (Loman and Watson, 2015). Based on trends within NGS, within a few years handheld DNA sequencers will be able to sequence whole DNA strands *in situ*.

Sequences of these lengths should theoretically be capable of perfect annotation, and the largest limitation will become the annotation methods themselves. Issues include the lack of species and genes within reference databases and the ability of the algorithms to address biological variations within species' genomes, such as interspecies gene transfer. It is likely that there will be a shift away from taxonomic identification and towards functional classification as our knowledge of gene analysis increases. Furthermore, as molecular techniques improve, RNA sequencing of metatranscriptomes will allow researchers to quantify gene expression, moving the field from identifying potential functional activities to actual functional activities.

Current DNA sequencing technologies provide coverage deep enough to confidently observe microbial community compositions. Experimental designs, however, need to include several replicates to identify changes in complex environmental samples. The cost of DNA library preparation and sequencing is still a limiting factor for many researchers. However, this will decrease with advances in sequencing technology. Sequencing the human genome, for example, cost $100 million in 2001 and now costs $1,000 (Wetterstrand, 2016).

5.3 Limitations

This thesis uses relatively new DNA sequencing technologies combined with completely novel and custom gas analysis systems to study the impacts of flooding on microbial ecosystems. Several mechanical and technological issues were encountered during the research. SkyLine was custom built to continuously monitor CO_2 and CH_4 fluxes over several months. Developing a fully functional prototype was delayed by over a year, removing the feasibility of other planned experiments with the site (e.g. targeted mRNA sequencing to identify gene expression for observed functional responses). Furthermore, the experiment site was approximately 250 miles away from the University of York. Any failures could take days to notice and, depending on the failure, weeks to fix and redeploy. This led to gaps in the gas flux data. Failures range from mechanical issues with the chamber sensing when it is located above a lysimeter, to land owners accidentally turning off the power to the system. Due to the vast amounts of continuous data produced however, the experiment is still considered successful. Future experiments using this system will hopefully avoid such failures as they have now been identified.

Due to the structure of SkyLine, a maximum length for which the system could operate safely restricted the number of lysimeters available. A three-dimensional plot, opposed to the linear plot used, would allow for many more biological replicates. This would increase the statistical power for identifying responses to treatments.

The laboratory experiment also suffered a mechanical failure. A CH_4 analysis machine was set up to monitor CH_4 fluxes throughout the experiment, however it broke a week into the experiment and it was not fixed until after the experiment finished. The flooding treatments and microbial studies were completely successful however, and the gas flux analysis was emitted from the manuscript.

5.4 Future work

To expand the applications for Chapter 2, the performances of other metagenomic annotation programs should be tested; such as IMG-ER (Markowitz *et al.*, 2009), RAPSearch2 (Zhao *et al.*, 2012), PAUDA (Huson and Xie, 2014), DIAMOND (Buchfink *et al.*, 2015). The performances of these programs are reported to differ in speed and sensitivity, thus Simmet would provide a valuable quantitative comparison between them. Computational running time may also be a limiting factor for researchers, so Simmet could be used to fairly compare the running speeds of different programs.

Simmet contained error rates and DNA coverage likely to be generated from 454 pyrosequencing. This produced conservative results, as 454 pyrosequencing is relatively erroneous compared to newer technologies such as Illumina paired-end sequencing. Analysing variations of Simmet to investigate the performances of annotation tools and parameters for different sequencing technologies would prove valuable; it would determine if different tools are optimum depending on the sequencing technology, rather than concluding overall performances. For example, some tools may perform better for short reads whereas others may handle base-calling errors better.

To further understand the mechanisms and responses to flooding treatments, additional biogeochemical measurements should be made. For example, measuring redox potential, pH, gravimetric water content (more frequently), and particular chemical concentrations (e.g. N, C, P). Other gas fluxes should also be monitored, such as N_2O, another greenhouse gas with a high global warming potential (228 CO_2e over 100 years (Myhre *et al.*, 2013)). This would support any conclusions on functional responses based on gene abundances. Plant species should also be quantified in field

experiments to test for effects of species on microbial responses (e.g. *Trifolium* and nitrogen fixation).

As sequencing costs decrease, more frequent observations of microbial communities and functions could be made. Responses over time could therefore be measured, including any fluctuations that are missed when only sampling at the start and end time points. The mechanism behind the delayed response to CO_2 and CH_4 fluxes observed in Chapter 4 could also be investigated. Communities from different depths could also be studied, adding a spatial understanding to microbial responses to flooding on pasture soils. Several studies show variations in microbial communities at different depths (Fang and Moncrieff, 2005; Lipson *et al.*, 2013; McDonald *et al.*, 1999). Larger lysimeters would be required to take more frequent soil samples and from different depths. This would also reduce any edge effects of the lysimeters, such as increased drainage and leaching. To achieve this, SkyLine would need a larger gas measuring chamber and larger, more durable supporting structures. Combined with a greater number of replicates and cheaper sequencing technologies, several hypotheses with reduced spatial and temporal constraints could be tested. The mechanisms behind responses should be studied at micro scales as well as macro scales to develop a holistic understanding of ecosystems.

Metagenomics allows for potential functional responses to be observed. In Chapter 4, this is combined with gas flux measurements to observe actual functional responses. To fully understand the mechanism underpinning the functional responses, metatranscriptomes should be studied to quantify gene expression. Metagenomics only quantifies gene abundance, not expression; metatranscriptomics would bridge the gap between quantifying potential functions and quantifying expressed functions.

While metatranscriptomics produces actual expression data, caveats with current sequencing technologies would still apply. One major *caveat* being annotation error rates resulting from short reads. New sequencing technologies such as nanopore sequencing will achieve longer read lengths, drastically reducing these errors and allowing greater confidence in sequence annotations. These technologies may also provide *in situ* sequencing capabilities, eliminating the need for sample processing and library preparation in a laboratory prior to sequencing.

5.5 Concluding statement

This thesis evaluates DNA sequence annotation tools and investigates how microbial communities, their functions, and CO_2 and CH_4 fluxes respond to increased flooding frequency and duration. Caveats and pitfalls were identified in annotation performances, and potential error rates quantified. These errors may already be present in published data and this thesis highlights the importance of questioning annotation reliability when both selecting parameters for one's own research and when reading published data.

The laboratory experiment yielded many statistically positive results for the microbial data, providing an insight into how controlled communities respond to environmental changes. However, few such results were observed in the field experiment, even though significant gas flux responses were observed. This is most likely due to the low signal to noise ratio present in the wealth of data produced. Combined with the error rates identified in Chapter 2, this thesis shows that the field of metagenomics is still in its infancy, and it provides an insight into the confidence limits of these results and limitations that prevent biological responses and mechanisms from being observed. Solely using metagenomics provides a noisy overview. Until future technologies reduce error rates and provide more practical data, it is suggested that targeted approaches are also used, such as amplicon sequencing and qPCR, which answer more precise biological questions.

List of Appendixes

A.1 Chapter 1 Supporting information

Table A.1. NGS platforms. List of NGS platforms and their expected throughputs, read lengths and costs per gigabase. Extracted from Glenn (2011) and updated in 2016 by http://www.molecularecologist.com/next-gen-fieldguide-2016/.

Instrument	Run time	Millions of Reads/run	Bases / read	cost/Gb
Illumina MiniSeq - Mid	17 hrs.	8	300	$229.17
Illumina MiniSeq - High	7 hrs.	25	75	$426.67
Illumina MiniSeq - High	13 hrs.	25	150	$249.33
Illumina MiniSeq - High	24 hrs.	25	300	$200.00
Illumina MiSeq v2 Nano	17 hrs.	1	300	$1,866.67
Illumina MiSeq v2 Nano	28 hrs.	1	500	$1,360.00
Illumina MiSeq v2 Micro	19 hrs.	4	300	$708.33
Illumina MiSeq v2	4 hrs.	15	50	$1,060.00
Illumina MiSeq v2	24 hrs.	15	300	$225.56
Illumina MiSeq v2	39 hrs.	15	500	$151.33
Illumina MiSeq v3	21 hrs.	25	150	$233.33
Illumina MiSeq v3	56 hrs.	25	600	$102.00
Illumina NextSeq 500 - Mid v2	15 hrs.	130	150	$52.82
Illumina NextSeq 500 - Mid v2	26 hrs.	130	300	$42.31
Illumina NextSeq 500 - High v2	11 hrs.	400	75	$46.00
Illumina NextSeq 500 - High v2	18 hrs.	400	150	$44.17

Instrument	Run time	Millions of Reads/run	Bases / read	cost/Gb
Illumina NextSeq 500 - High v2	29 hrs.	400	300	$35.33
Illumina HiSeq 2500 - rapid run	10 hrs.	300	50	$95.33
Illumina HiSeq 2500 - rapid run	40 hrs.	300	200	$55.33
Illumina HiSeq 2500 - rapid run v2	10 hrs.	300	50	$95.33
Illumina HiSeq 2500 - rapid run v2	27 hrs.	300	200	$55.33
Illumina HiSeq 2500 - rapid run v2	60 hrs.	300	500	$37.77
Illumina HiSeq 2500 - high output v3	2 days	1500	50	$83.20
Illumina HiSeq 2500 - high output v3	11 days	1500	200	$52.87
Illumina HiSeq 2500 - high output v4	40 hrs.	2000	50	$62.40
Illumina HiSeq 2500 - high output v4	6 days	2000	250	$28.82
Illumina HiSeq 4000	1 day	2500	50	$48.08
Illumina HiSeq 4000	2 days	2500	150	$29.36
Illumina HiSeq 4000	3.5 days	2500	300	$20.53
Illumina HiSeq X - Five	< 3 days	3000	300	$10.63
Illumina HiSeq X - Ten	< 3 days	3000	300	$7.08
Ion Torrent – PGM 314 chip v2	4 hrs.	0.55	400	$2,154.55
Ion Torrent – PGM 316 chip v2	5 hrs.	3	400	$561.67
Ion Torrent – PGM 318 chip v2	7 hrs.	5.5	400	$397.27
Ion Torrent - Proton I	4 hrs.	80	200	$62.50
Ion Torrent - S5 520 chip	4 hrs.	5	400	$476.50

Instrument	Run time	Millions of Reads/run	Bases / read	cost/Gb
Ion Torrent - S5 530 chip	4 hrs.	20	400	$139.13
Ion Torrent - S5 540 chip	2.5 hrs.	80	200	$79.69
Oxford Nanopore MinION (std. speed; low volume user)	varies	0.6	10000	$150.00
Oxford Nanopore MinION (fast mode; high volume user)	varies	4.4	10000	$6.14
Oxford Nanopore PromethION (single flow cell; fast mode; claimed)	varies	26	10000	?
Oxford Nanopore PromethION (48 flow cells; fast mode; claimed)	varies	1250	10000	?
Pacific Biosciences RS II	≤6 hrs.	0.055	12000	$303.03
Pacific Biosciences Sequel	≤6 hrs.	0.385	10000	$181.82
SOLiD – 5500 (PI)	8 days	700	110	$79.23
SOLiD – 5500xl (4hq)	8 days	1410	110	$67.72

A.2 Chapter 2 Supporting information

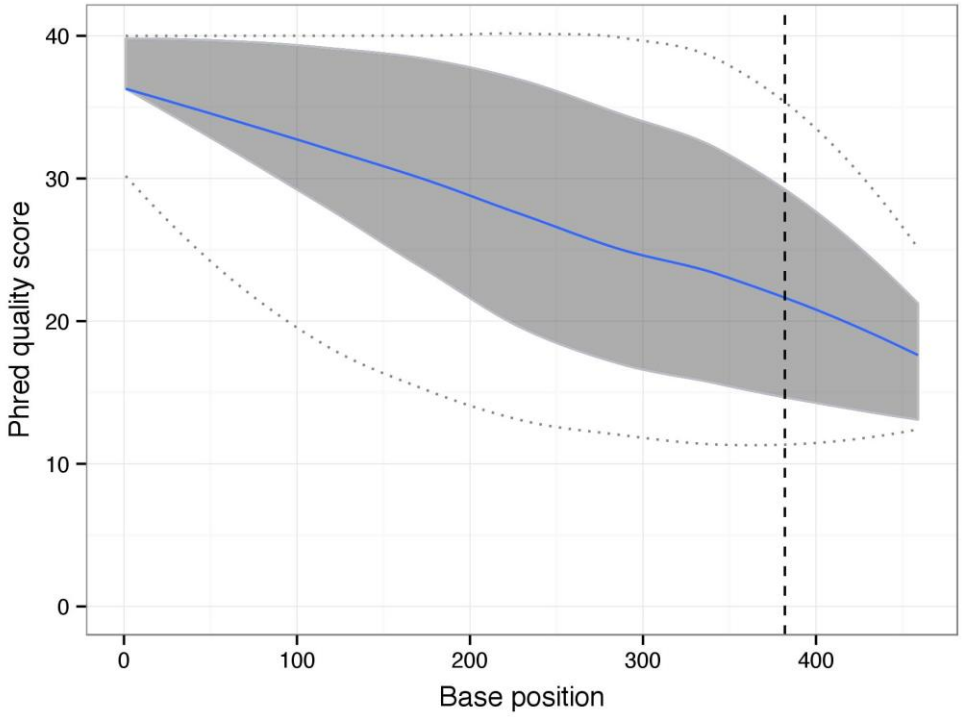


Figure A.1. Phred quality scores. The phred quality scores for the bases along the sequences in the simulated metagenome. The blue line shows the mean, the shaded grey area represents the interquartile range and the grey lines represent the 10th and 90th percentiles. The area to the left of the dashed line represents approximately 95 % of the sequences.

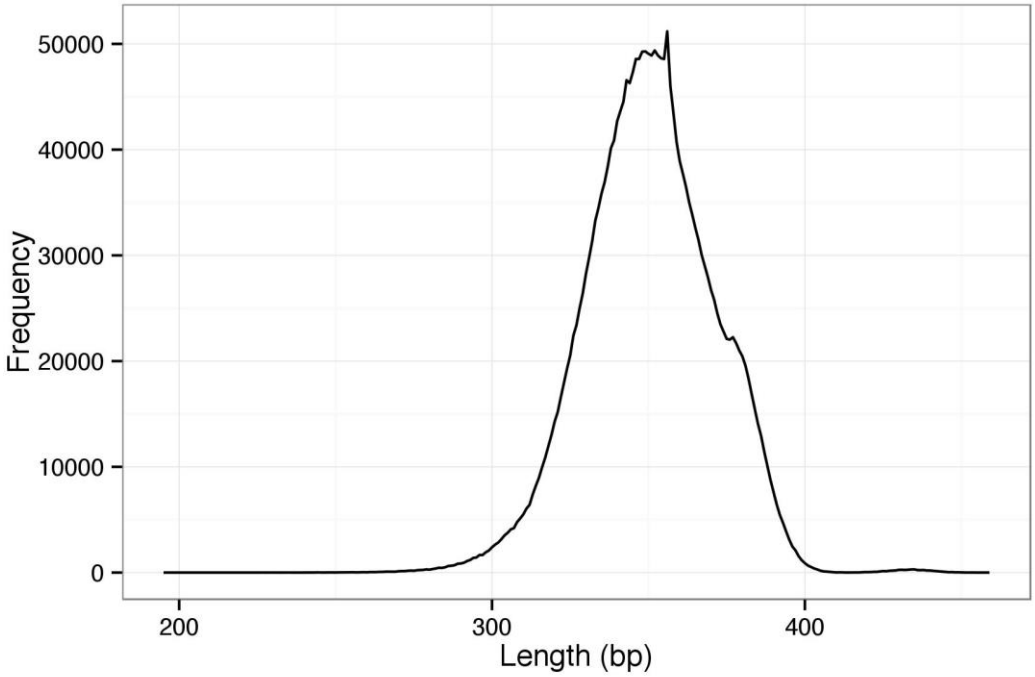


Figure A.2. The sequence length distribution for the simulated metagenome.

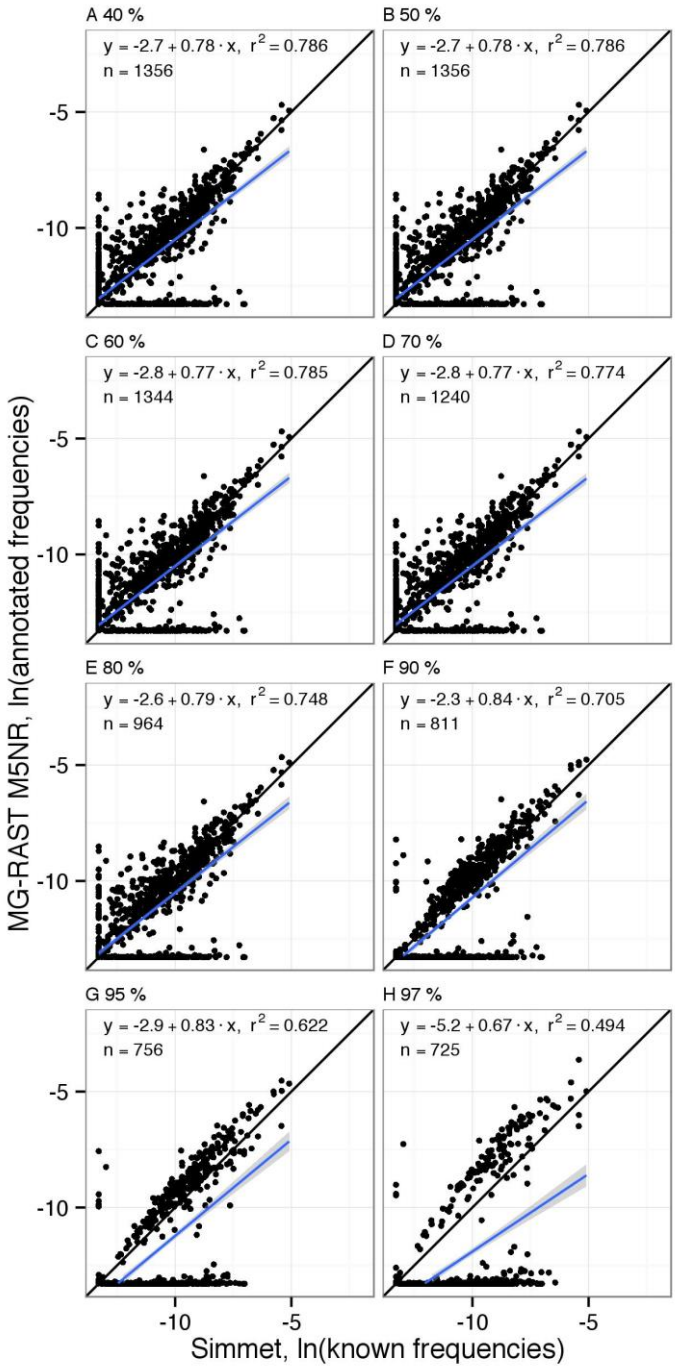


Figure A.3. Genus relative abundances. The genus relative abundances in Simmet and annotated by MG-RAST, using the M5NR database, with various minimum identity cut-off values. The black line represents a complete positive correlation and the blue line shows the line of best fit, with the 95 % confidence interval shaded in grey.

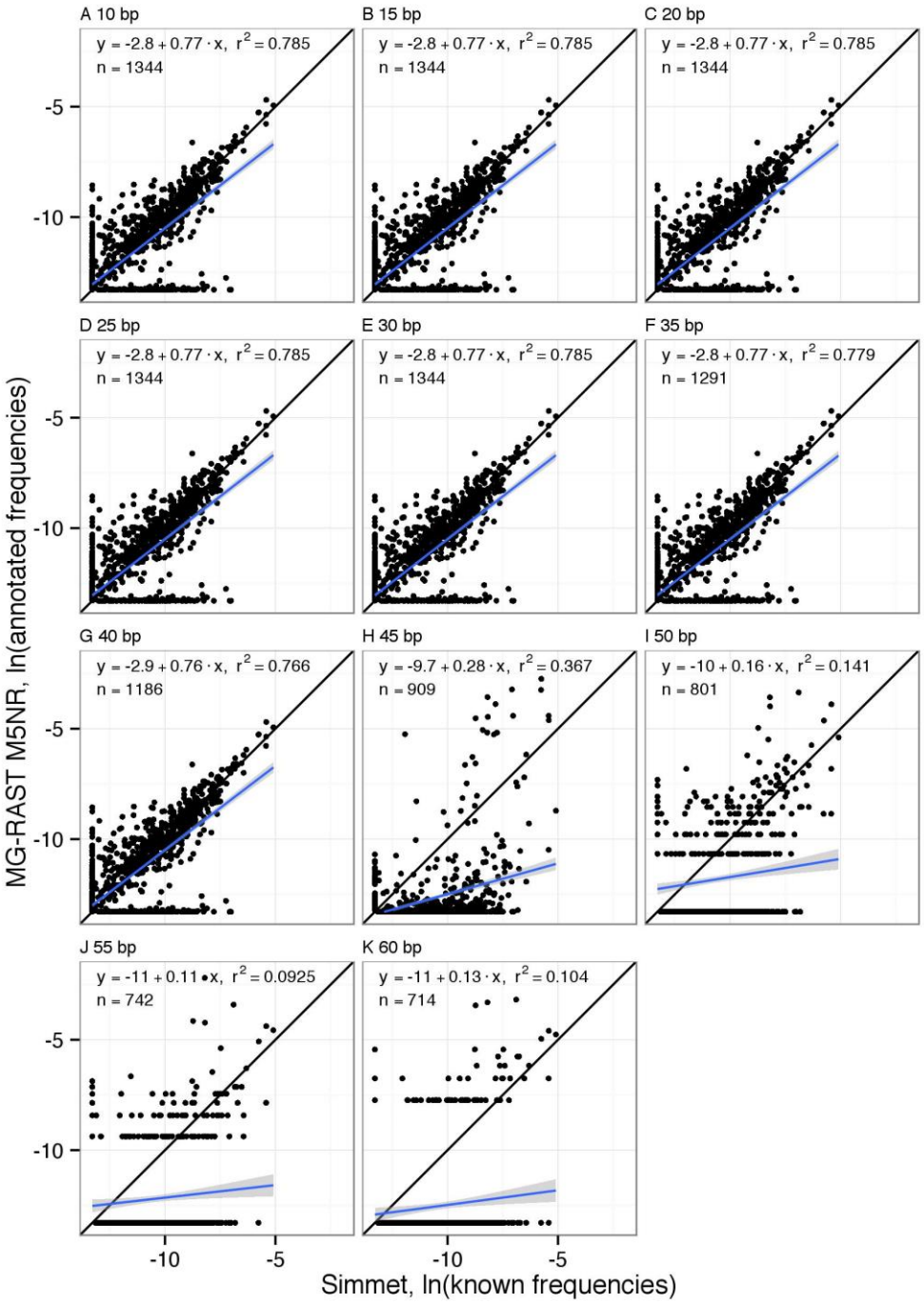


Figure A.4. Genus relative abundances. The genus relative abundances in Simmet and annotated by MG-RAST, using the M5NR database, with various minimum alignment length. The black line represents a complete positive correlation and the blue line shows the line of best fit, with the 95 % confidence interval shaded in grey.

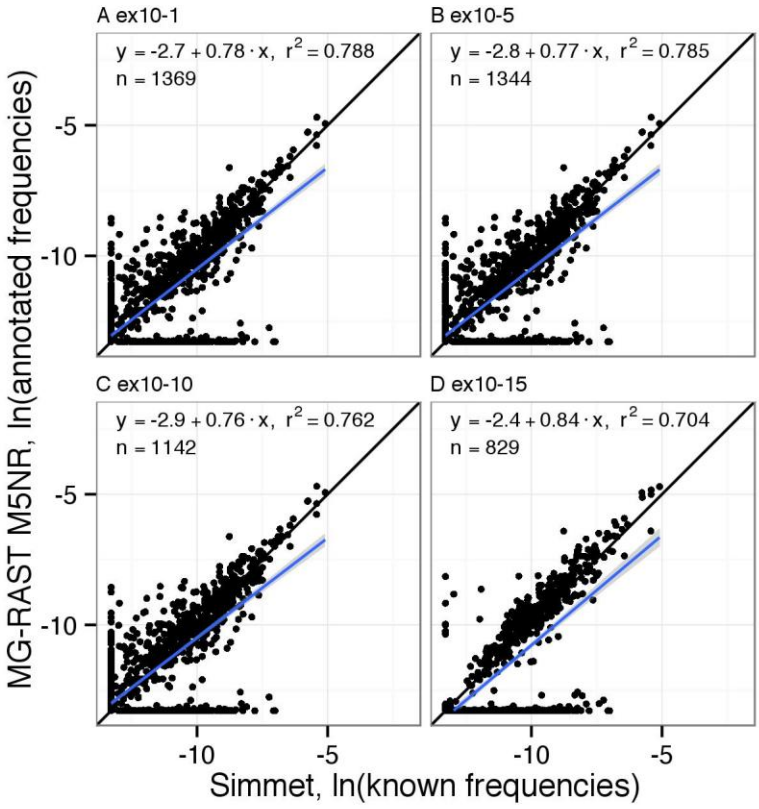


Figure A.5. Genus relative abundances. The genus relative abundances in Simmet and annotated by MG-RAST, using the M5NR database, with various maximum E-values. The black line represents a complete positive correlation and the blue line shows the line of best fit, with the 95 % confidence interval shaded in grey.

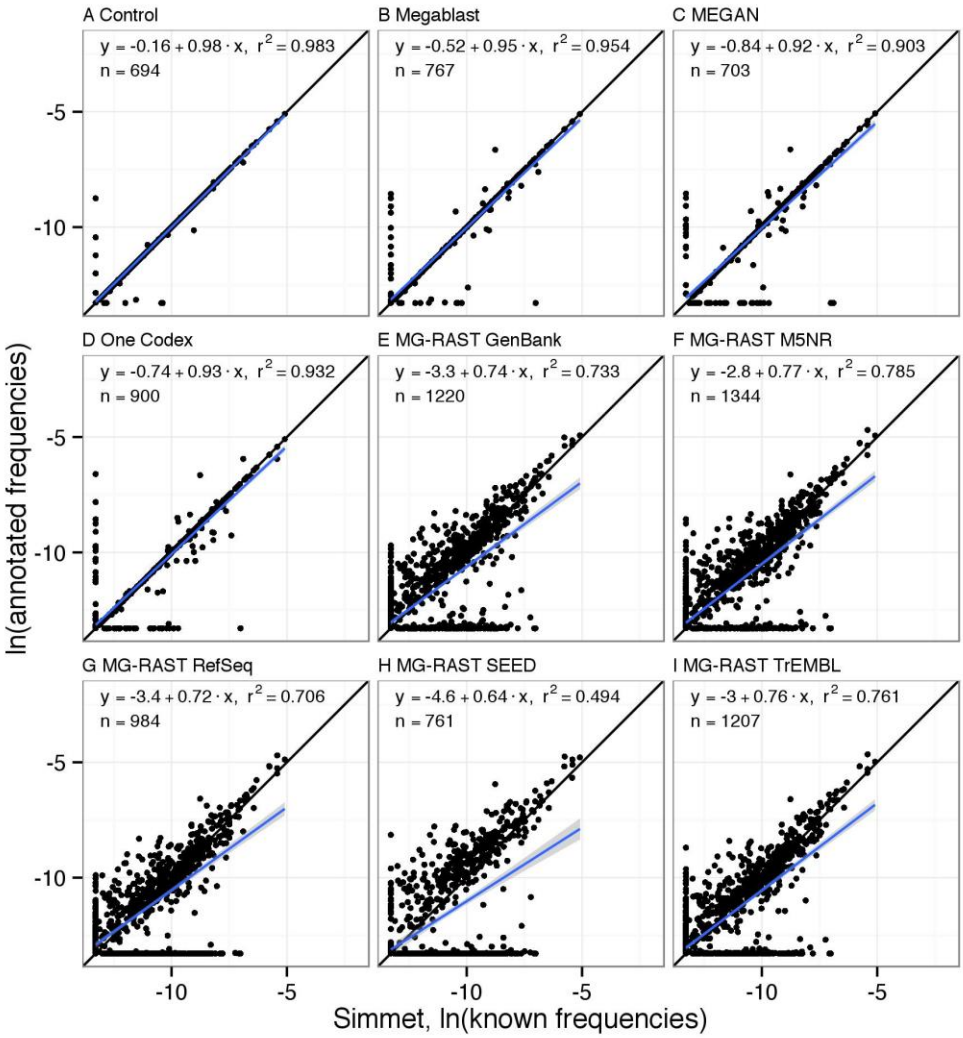


Figure A.6. Genus relative abundances. The correlations between known relative abundances of genera in Simmet and the relative abundances of genera from the annotation methods. The natural logarithm of the relative abundances was used. The black line represents a correlation coefficient of 1, the blue line shows the line of best fit, and the 95 % confidence intervals are shaded grey.

Table A.2. Descriptions for commonly used sequence databases.

Database	Description
Greengenes (http://greengenes.lbl.gov/)	Bacterial and archaeal 16S rRNA gene sequences.
M5 Non-Redundant protein database (M5NR) (http://tools.metagenomics.anl.gov/)	Incorporates several databases: European Bioinformatics Institute, Gene Ontology, Joint Genome Institute, KEGG, NCBI, Phage Annotation Tools and Methods, SEED, UniProt, Virginia Bioinformatics Institute, and the Evolutionary genealogy of genes: Non-supervised Orthologous Groups (EggNOG).
M5 Non-Redundant ribosomal database (M5RNA) (http://metagenomics.anl.gov/)	Incorporates SILVA, Greengenes, and RDP.
NCBI GenBank (http://www.ncbi.nlm.nih.gov/genbank/)	A non-curated database of sequences that have been submitted by individual laboratories and part of the International Nucleotide Sequence Database Collaboration (INSDC).
NCBI RefSeq (http://www.ncbi.nlm.nih.gov/refseq/)	Derived from the INSDC to produce a non-redundant, curated database of sequences from multiple sources.
NCBI Nucleotide (http://www.ncbi.nlm.nih.gov/nuccore)	Sequences from several sources, including GenBank and RefSeq.
Ribosomal Database Project (RDP) (https://rdp.cme.msu.edu/)	Bacterial and archaeal 16S rRNA gene sequences and fungal 28S rRNA fungal sequences.
SEED (http://www.theseed.org/)	Curated genomic data for archaea, bacteria, eukaryotes and viruses.
SILVA (http://www.arb-silva.de/)	Quality checked small (16S/18S, SSU) and large subunit (23S/28S, LSU) ribosomal RNA (rRNA) sequences for Bacteria, Archaea and Eukarya.
UniProt Swiss-Prot	Manually annotated and reviewed

Database	Description
(http://www.uniprot.org/)	sequence database.
UniProt TrEMBL (http://www.uniprot.org/)	Automatically annotated and non-reviewed sequence database.

Table A.3. Taxonomic annotation statistics. The Simmet annotation statistics for each method and database at all taxonomic levels excluding species.

Method	Database	Taxonomic level	Sensitivity (%)	Correctly annotated (%)	Incorrectly annotated (%)
Megablast	Control	Genus	99.88	99.39	0.49
Megablast	NCBI	Genus	99.81	97.32	2.49
MEGAN	MEGAN	Genus	98.56	95.65	2.91
MG-RAST	GenBank	Genus	81.94	52.65	29.30
MG-RAST	Greengenes	Genus	0.11	0.08	0.03
MG-RAST	RDP	Genus	0.13	0.10	0.03
MG-RAST	RefSeq	Genus	89.58	55.90	33.68
MG-RAST	SEED	Genus	64.97	39.75	25.22
MG-RAST	SwissProt	Genus	11.49	6.08	5.42
MG-RAST	TrEMBL	Genus	86.37	54.93	31.44
One Codex	One Codex	Genus	100.00	94.18	5.82
Megablast	Control	Family	99.88	99.82	0.07
Megablast	NCBI	Family	99.81	98.47	1.34
MEGAN	MEGAN	Family	98.56	97.02	1.54
MG-RAST	GenBank	Family	81.94	63.76	18.18
MG-RAST	Greengenes	Family	0.11	0.10	0.02
MG-RAST	RDP	Family	0.13	0.12	0.02
MG-RAST	RefSeq	Family	89.58	69.01	20.58
MG-RAST	SEED	Family	64.97	49.56	15.41
MG-RAST	SwissProt	Family	11.49	8.37	3.12
MG-RAST	TrEMBL	Family	86.37	66.85	19.51
One Codex	One Codex	Family	100.00	96.34	3.66
Megablast	Control	Order	99.88	99.82	0.06
Megablast	NCBI	Order	99.81	98.40	1.41

Method	Database	Taxonomic level	Sensitivity (%)	Correctly annotated (%)	Incorrectly annotated (%)
MEGAN	MEGAN	Order	98.56	97.21	1.35
MG-RAST	GenBank	Order	81.94	69.33	12.62
MG-RAST	Greengenes	Order	0.11	0.10	0.01
MG-RAST	RDP	Order	0.13	0.12	0.01
MG-RAST	RefSeq	Order	89.58	75.35	14.23
MG-RAST	SEED	Order	64.97	54.16	10.82
MG-RAST	SwissProt	Order	11.49	9.23	2.27
MG-RAST	TrEMBL	Order	86.37	73.06	13.31
One Codex	One Codex	Order	100.00	97.25	2.75
Megablast	Control	Class	99.88	99.60	0.28
Megablast	NCBI	Class	99.81	98.66	1.15
MEGAN	MEGAN	Class	98.56	97.53	1.03
MG-RAST	GenBank	Class	81.94	73.19	8.75
MG-RAST	Greengenes	Class	0.11	0.10	0.01
MG-RAST	RDP	Class	0.13	0.13	0.00
MG-RAST	RefSeq	Class	89.58	79.86	9.73
MG-RAST	SEED	Class	64.97	57.62	7.35
MG-RAST	SwissProt	Class	11.49	10.08	1.41
MG-RAST	TrEMBL	Class	86.37	77.06	9.31
One Codex	One Codex	Class	100.00	97.38	2.62
Megablast	Control	Phylum	99.88	99.60	0.28
Megablast	NCBI	Phylum	99.81	98.64	1.18
MEGAN	MEGAN	Phylum	98.56	97.45	1.11
MG-RAST	GenBank	Phylum	81.94	76.04	5.90
MG-RAST	Greengenes	Phylum	0.11	0.11	0.00
MG-RAST	RDP	Phylum	0.13	0.13	0.00
MG-RAST	RefSeq	Phylum	89.58	83.15	6.43
MG-RAST	SEED	Phylum	64.97	60.15	4.83
MG-RAST	SwissProt	Phylum	11.49	10.52	0.97
MG-RAST	TrEMBL	Phylum	86.37	80.05	6.31
One Codex	One Codex	Phylum	100.00	97.54	2.46
Megablast	Control	Domain	99.88	99.60	0.28
Megablast	NCBI	Domain	99.81	98.64	1.17

Method	Database	Taxonomic level	Sensitivity (%)	Correctly annotated (%)	Incorrectly annotated (%)
MEGAN	MEGAN	Domain	98.56	97.45	1.11
MG-RAST	GenBank	Domain	81.94	79.63	2.32
MG-RAST	Greengenes	Domain	0.11	0.11	0.00
MG-RAST	RDP	Domain	0.13	0.13	0.00
MG-RAST	RefSeq	Domain	89.58	87.28	2.31
MG-RAST	SEED	Domain	64.97	63.20	1.78
MG-RAST	SwissProt	Domain	11.49	11.16	0.34
MG-RAST	TrEMBL	Domain	86.37	83.68	2.68
One Codex	One Codex	Domain	100.00	97.72	2.28

Table A.4. Functional annotation statistics. The Simmet function annotation statistics for MG-RAST COG and MG-RAST KO using different minimum identity cut-off values.

Database	Minimum identity cut-off	Sensitivity (%)	Correctly annotated (%)	Incorrectly annotated (%)
MG-RAST COG	40 %	50.74	46.24	4.50
MG-RAST COG	50 %	50.74	46.24	4.50
MG-RAST COG	60 %	50.49	46.02	4.47
MG-RAST COG	70 %	48.37	44.13	4.24
MG-RAST COG	80 %	41.10	37.55	3.55
MG-RAST COG	90 %	24.92	22.70	2.22
MG-RAST COG	95 %	15.62	14.10	1.52
MG-RAST KO	40 %	63.44	45.45	17.99
MG-RAST KO	50 %	63.44	45.45	17.99
MG-RAST KO	60 %	63.28	45.36	17.91
MG-RAST KO	70 %	62.11	44.70	17.41
MG-RAST KO	80 %	57.80	42.01	15.79
MG-RAST KO	90 %	43.63	31.89	11.74
MG-RAST KO	95 %	31.47	22.66	8.81

Table A.5. Functional annotation statistics. The Simmet function annotation statistics for MG-RAST COG and MG-RAST KO using different minimum alignment lengths.

Database	Minimum alignment length	Sensitivity (%)	Correctly annotated (%)	Incorrectly annotated (%)
MG-RAST COG	10 bp	50.49	46.02	4.47
MG-RAST COG	15 bp	50.49	46.02	4.47
MG-RAST COG	20 bp	50.49	46.02	4.47
MG-RAST COG	25 bp	50.35	45.90	4.44
MG-RAST COG	30 bp	49.39	45.10	4.29
MG-RAST COG	35 bp	46.35	42.46	3.89
MG-RAST COG	40 bp	38.19	35.17	3.02
MG-RAST COG	45 bp	22.97	21.29	1.68
MG-RAST COG	50 bp	9.39	8.72	0.67
MG-RAST COG	55 bp	3.20	2.97	0.23
MG-RAST COG	60 bp	1.03	0.95	0.07
MG-RAST KO	10 bp	63.28	45.36	17.91
MG-RAST KO	15 bp	63.28	45.36	17.91
MG-RAST KO	20 bp	63.28	45.36	17.91
MG-RAST KO	25 bp	63.06	45.25	17.81
MG-RAST KO	30 bp	61.90	44.59	17.31
MG-RAST KO	35 bp	58.70	42.64	16.06
MG-RAST KO	40 bp	50.28	37.04	13.24
MG-RAST KO	45 bp	33.15	24.82	8.33
MG-RAST KO	50 bp	15.16	11.40	3.76
MG-RAST KO	55 bp	5.58	4.15	1.43
MG-RAST KO	60 bp	1.88	1.39	0.49

Table A.6. Functional annotation statistics. The Simmet function annotation statistics for MG-RAST COG and MG-RAST KO using different minimum alignment lengths.

Database	Maximum E-value	Sensitivity (%)	Correctly annotated (%)	Incorrectly annotated (%)
MG-RAST COG	1-e^{1}	51.01	46.45	4.56
MG-RAST COG	1-e^{5}	50.49	46.02	4.47
MG-RAST COG	1-e^{10}	43.71	40.10	3.61
MG-RAST COG	1-e^{15}	24.00	22.22	1.79
MG-RAST KO	1-e^{1}	63.76	45.62	18.14
MG-RAST KO	1-e^{5}	63.28	45.36	17.91
MG-RAST KO	1-e^{10}	57.74	42.10	15.65
MG-RAST KO	1-e^{15}	37.74	28.30	9.44

Table A.7. Class fold differences. The fold differences in relative abundance for classes present in Simmet and in the annotations for MG-RAST M5NR, MG-RAST RefSeq, MEGAN and One Codex. Note that there are more than five perfectly accurate class annotations for One Codex.

Class	Simmet	Annotation	Fold difference
A.1 The five most over-annotated classes for M5NR			
Fibrobacteria	0.00002	0.00164	66.790
Verrucomicrobiae	0.00004	0.00032	7.825
Solibacteres	0.00111	0.00149	1.340
Nitrospira	0.00046	0.00058	1.251
Betaproteobacteria	0.07203	0.08942	1.241
A.2 The five most over-annotated classes for RefSeq			
Verrucomicrobiae	0.00004	0.00019	4.665
Fibrobacteria	0.00002	0.00007	2.899
Fusobacteriia	0.00332	0.00461	1.390
Erysipelotrichia	0.00057	0.00075	1.325
Betaproteobacteria	0.07203	0.09174	1.274
A.3 The five most over-annotated classes for MEGAN			
Nitrospira	0.00046	0.00086	1.865
Alphaproteobacteria	0.09676	0.10418	1.077
Mollicutes	0.00853	0.00900	1.054

Class	Simmet	Annotation	Fold difference
Betaproteobacteria	0.07203	0.07399	1.027
Cytophagia	0.00906	0.00924	1.019
A.4 The five most over-annotated classes for One Codex			
Nitrospira	0.00046	0.00085	1.838
Erysipelotrichia	0.00057	0.00074	1.305
Synergistia	0.00250	0.00318	1.272
Alphaproteobacteria	0.09676	0.10434	1.078
Betaproteobacteria	0.07203	0.07447	1.034
B.1 The five most accurately annotated classes for M5NR			
Bacilli	0.07507	0.07835	1.044
Chloroflexi	0.00424	0.00430	1.013
Epsilonproteobacteria	0.01166	0.01157	0.992
Mollicutes	0.00853	0.00809	0.948
Flavobacteriia	0.02413	0.02252	0.934
B.2 The five most accurately annotated classes for RefSeq			
Dehalococcoidia	0.00122	0.00126	1.026
Deferribacteres	0.00212	0.00211	0.998
Halobacteria	0.01788	0.01776	0.993
Bacteroidia	0.02094	0.02023	0.966
Deinococci	0.00958	0.00921	0.961
B.3 The five most accurately annotated classes for MEGAN			
Aquificae	0.00367	0.00370	1.009
Synergistia	0.00250	0.00252	1.009
Gammaproteobacteria	0.17852	0.17981	1.007
Bacteroidia	0.02094	0.02102	1.004
Spirochaetia	0.01588	0.01589	1.001
B.4 The five most accurately annotated classes for One Codex			
Acidobacteriia	0.00282	0.00282	1.000
Anaerolineae	0.00127	0.00127	1.000
Archaeoglobi	0.00101	0.00101	1.000
Caldilineae	0.00045	0.00045	1.000
Caldisericia	0.00019	0.00019	1.000
C.1 The five most under-annotated classes for M5NR			
Synergistia	0.00250	0.00134	0.535

Class	Simmet	Annotation	Fold difference
Methanobacteria	0.00599	0.00299	0.498
Planctomycetia	0.01154	0.00537	0.465
Sphingobacteriia	0.01159	0.00449	0.388
Thermodesulfobacteria	0.00040	0.00000	0.002
C.2 The five most under-annotated classes for RefSeq			
Spirochaetia	0.01588	0.00947	0.596
Methanobacteria	0.00599	0.00327	0.546
Synergistia	0.00250	0.00132	0.530
Sphingobacteriia	0.01159	0.00410	0.354
Planctomycetia	0.01154	0.00386	0.334
C.3 The five most under-annotated classes for MEGAN			
Spirochaetia	0.01588	0.01589	1.001
Clostridia	0.06934	0.06804	0.981
Flavobacteriia	0.02413	0.02266	0.939
Thermoplasmata	0.00140	0.00117	0.838
Unclassified	0.06768	0.05259	0.777
C.4 The five most under-annotated classes for One Codex			
Bacilli	0.07507	0.07404	0.986
Flavobacteriia	0.02413	0.02361	0.978
Epsilonproteobacteria	0.01166	0.01075	0.921
Unclassified	0.06704	0.04893	0.723
Deferribacteres	0.00212	0.00103	0.488

Table A.8. Taxa richness. The taxa richness estimates and the differences from Simmet for each annotation method. Due to the low numbers, domain is excluded from comparisons.

Method	Database	Taxonomic level	Richness	Difference (%)
Simmet	N/A	Genus	688	N/A
Megablast	Control	Genus	688	100.00
Megablast	NCBI	Genus	758	110.17
MEGAN	MEGAN	Genus	672	97.67
MG-RAST	GenBank	Genus	1090	158.43
MG-RAST	Greengenes	Genus	404	58.72

Method	Database	Taxonomic level	Richness	Difference (%)
MG-RAST	M5NR	Genus	1245	180.96
MG-RAST	M5RNA	Genus	655	95.20
MG-RAST	RDP	Genus	469	68.17
MG-RAST	RefSeq	Genus	813	118.17
MG-RAST	SEED	Genus	445	64.68
MG-RAST	SwissProt	Genus	657	95.49
MG-RAST	TrEMBL	Genus	1094	159.01
One Codex	One Codex	Genus	872	126.74
Simmet	N/A	Family	250	N/A
Megablast	Control	Family	251	100.40
Megablast	NCBI	Family	277	110.80
MEGAN	MEGAN	Family	253	101.20
MG-RAST	GenBank	Family	482	192.80
MG-RAST	Greengenes	Family	179	71.60
MG-RAST	M5NR	Family	527	210.80
MG-RAST	M5RNA	Family	244	97.60
MG-RAST	RDP	Family	190	76.00
MG-RAST	RefSeq	Family	372	148.80
MG-RAST	SEED	Family	240	96.00
MG-RAST	SwissProt	Family	331	132.40
MG-RAST	TrEMBL	Family	459	183.60
One Codex	One Codex	Family	281	112.40
Simmet	N/A	Order	120	N/A
Megablast	Control	Order	121	100.83
Megablast	NCBI	Order	137	114.17
MEGAN	MEGAN	Order	123	102.50
MG-RAST	GenBank	Order	276	230.00
MG-RAST	Greengenes	Order	94	78.33
MG-RAST	M5NR	Order	304	253.33
MG-RAST	M5RNA	Order	134	111.67
MG-RAST	RDP	Order	97	80.83
MG-RAST	RefSeq	Order	220	183.33
MG-RAST	SEED	Order	133	110.83
MG-RAST	SwissProt	Order	201	167.50
MG-RAST	TrEMBL	Order	263	219.17

Method	Database	Taxonomic level	Richness	Difference (%)
One Codex	One Codex	Order	136	113.33
Simmet	N/A	Class	56	N/A
Megablast	Control	Class	56	100.00
Megablast	NCBI	Class	61	108.93
MEGAN	MEGAN	Class	57	101.79
MG-RAST	GenBank	Class	120	214.29
MG-RAST	Greengenes	Class	44	78.57
MG-RAST	M5NR	Class	136	242.86
MG-RAST	M5RNA	Class	70	125.00
MG-RAST	RDP	Class	46	82.14
MG-RAST	RefSeq	Class	105	187.50
MG-RAST	SEED	Class	67	119.64
MG-RAST	SwissProt	Class	86	153.57
MG-RAST	TrEMBL	Class	114	203.57
One Codex	One Codex	Class	65	116.07
Simmet	N/A	Phylum	36	N/A
Megablast	Control	Phylum	36	100.00
Megablast	NCBI	Phylum	42	116.67
MEGAN	MEGAN	Phylum	39	108.33
MG-RAST	GenBank	Phylum	57	158.33
MG-RAST	Greengenes	Phylum	27	75.00
MG-RAST	M5NR	Phylum	58	161.11
MG-RAST	M5RNA	Phylum	38	105.56
MG-RAST	RDP	Phylum	27	75.00
MG-RAST	RefSeq	Phylum	51	141.67
MG-RAST	SEED	Phylum	35	97.22
MG-RAST	SwissProt	Phylum	43	119.44
MG-RAST	TrEMBL	Phylum	58	161.11
One Codex	One Codex	Phylum	39	108.33

A.3 Chapter 3 Supporting information

Supporting Information A.3.1. Soil association.

Derived from the Academic Soils Site Report for location 392413E, 127330N, 1km x 1km (National Soil Resources Institute (NSRI), 2013).

Soil Association: Wickham 2 (711f)

a. General Description

Slowly permeable seasonally waterlogged fine loamy over clayey, fine silty over clayey and clayey soils. Small areas of slowly permeable calcareous soils on steeper slopes.

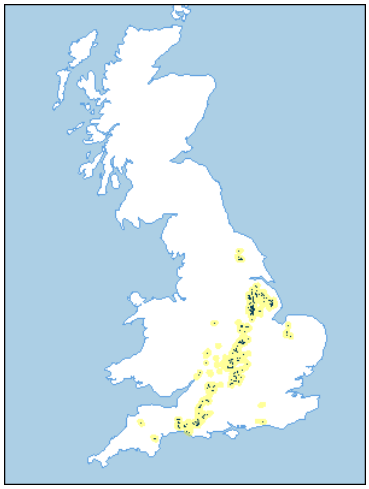
The major landuse on this association is defined as winter cereals and grassland in the midlands; cereals in the eastern region dairying in the South West.

b. Distribution (England & Wales)

The Wickham 2 association covers 1485km2 of England and Wales, which accounts for 0.98% of the landmass. The distribution of this association is shown in Supporting Information Figure A.3.1.1. Note that the yellow shading represents a buffer to highlight the location of very small areas of the association.

c. Comprising Soil Series

Multiple soil series comprise a soil association. The soil series of the Wickham 2 association are outlined in Supporting Information Table A.3.1.1 below. In some cases other minor soil series are present at a particular site, and these have been grouped together under the heading “Other”.



Supporting Information Figure A.3.1.1. Wickham 2 association distribution. The component soil series of the WICKHAM 2 soil association. Because absolute proportions of the comprising series in this association vary from location to location, the national proportions are provided.

Soil Series	Description	Area %
Wickham (Wh)	Medium loamy or medium silty drift over clayey material passing to clay or soft mudstone swelling clayey material passing to clay or soft mudstone	50
Denchworth (Da)	Swelling clayey material passing to clay or soft mudstone	15
Oxpasture (Ox)	Medium loamy or medium silty drift over clayey material passing to clay or soft mudstone	15
Evesham (Ea)	Swelling clayey material passing to clay or soft mudstone	10
Other	Other minor soils	10

Supporting Information A.3.2. Gravimetric water content

The Gravimetric water contents (GWC) of the samples were significantly different between treatments both after two weeks of drainage and after two weeks of flooding (independent two-sample t-test: drainage: t = -11.172, df 2.605, p = 0.003; flooding: t = -9.958, df 4, p = 0.001, Supporting Information Tables 6.3.2.1 and 7.3.2.2). To identify if this had an impact on the DNA extraction quality, the dry soil weights were recorded after drainage (week 8) and after treatment flooding (week 10), and tested for correlations with DNA concentration and purity. There was no significant correlation between GWC and DNA concentration after drainage, but there was after flooding (Pearson's product-moment correlation: drainage: p-value = 0.426, r -0.151; flooding: p-value = 0.024, r -0.412). DNA concentrations were still sufficient for library preparation. There were no significant correlations between GWC and 260:280 ratio (Pearson's product-moment correlation: drainage: p-value = 0.612, r -0.0964; flooding: p-value = 0.665, r 0.082), and GWC and 260:230 ratio (Pearson's product-moment correlation: drainage: p-value = 0.160, r 0.263; flooding: p-value = 0.261, r 0.212).

Supporting Information Table A.3.2.1. The gravimetric water content (GWC) of the soil at field capacity. μ = 0.369.

Replicate	Dish (g)	Start (g)	End (g)	Water (g)	GWC g/g
1	42.9	92.9	74.1	18.8	0.376
2	42.9	92.9	74.6	18.3	0.366
3	42.9	92.9	74.7	18.2	0.364

Supporting Information Table A.3.2.2. The gravimetric water content (GWC) of the soil prior to starting the experiment. μ = 0.039.

Replicate	Dish (g)	Start (g)	End (g)	Water (g)	GWC g/g
1	42.9	92.9	91.0	1.9	0.038
2	36.8	86.8	84.8	2.0	0.040
3	73.9	123.9	121.9	2.0	0.040

Supporting Information A.3.3. Comparing DNA isolation kits.

Mo Bio PowerSoil® versus Mo Bio PowerLyzer™ PowerSoil® DNA isolation kits.

Introduction

The PowerSoil® DNA isolation kit (Mo Bio Laboratories Inc., Carlsbad, CA, USA) is a popular kit for extracting DNA from a wide variety of soil samples (Kennedy *et al.*, 2013). However, certain samples still prove to be challenging to work with. Clay soils, for example, are particularly difficult to isolate from; the DNA strands bind tightly to the soil particles (Cai *et al.*, 2006; Crecchio and Stotzky, 1998) and humic substances co-purify the DNA due to their similar molecular structures (Dong *et al.*, 2006). Furthermore, humic acid inhibits the actions of enzymes involved in DNA isolation, such as restriction endonuclease and DNA polymerase (Dong *et al.*, 2006; Tebbe and Vahjen, 1993; Yankson and Steck, 2009). These factors can result in low yields of DNA and contaminated samples.

The PowerLyzer™ PowerSoil® kit (Mo Bio Laboratories Inc.) claims to extract more DNA from challenging soils (Kennedy *et al.*, 2013). This kit features 0.1 mm glass beads rather than 0.7 mm garnet beads as the PowerSoil® kit uses, thus it is more effective at mechanically breaking up the soil and lysing cells. The disadvantage of this kit is that it is also more aggressive may fragment the DNA to a greater extent than the PowerSoil® kit would.

To test which kit would be most appropriate to use with clay soil from Share Farm, Wiltshire, the two kits were compared alongside each other using the same soil sample (see Supporting Information A.3.1 for soil association).

Detailed information about each kit is available at www.mobio.com.

Methodology

Homogenised soil was sieved using a 6 mm sieve and ground to a fine material. The DNA was then extracted from four replicates for each kit following the manufacturers' protocols.

To establish the quantity and purity of the DNA extracted, the final extraction solutions were tested using a NanoDrop™ 8000 UV-Vis Spectrophotometer (Thermo Scientific, Waltham, MA, USA) and run on a 1.5 % agarose gel at 5 V / cm.

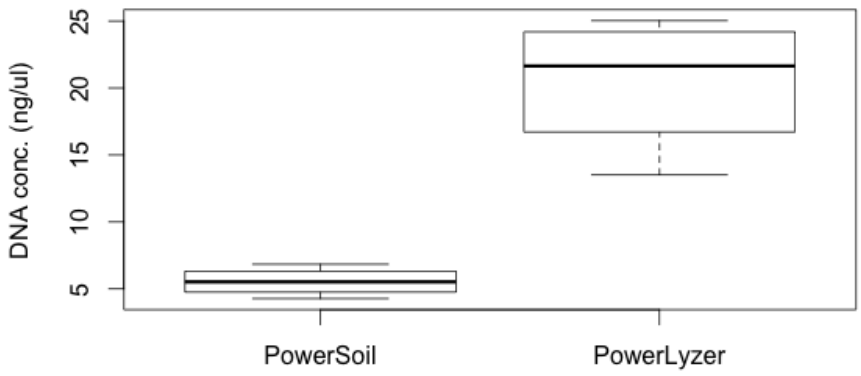
Results

The NanoDrop™ spectrophotometer results indicate a significant difference between the DNA concentrations and the 260:280 absorbance ratios achieved between the two kits, but not the 260:230 absorbance ratios (two-sample t-tests: concentration: t = -5.728, df = 3.264, p-value = 0.008; 260:280 ratio: t = 3.15, df = 3.38, p-value = 0.043; 260:230 ratio: t = -1.369, df = 5.631, p-value = 0.223) (Supporting Information Table A.3.3.1, Supporting Information Figures A.3.3.1 – A.3.3.4). A 260:280 absorbance ratio lower than 1.8 typically indicates contamination with proteins and a 260:230 absorbance ration lower that 1.5 typically indicates contamination with salts, carbohydrates or solvents such as phenol.

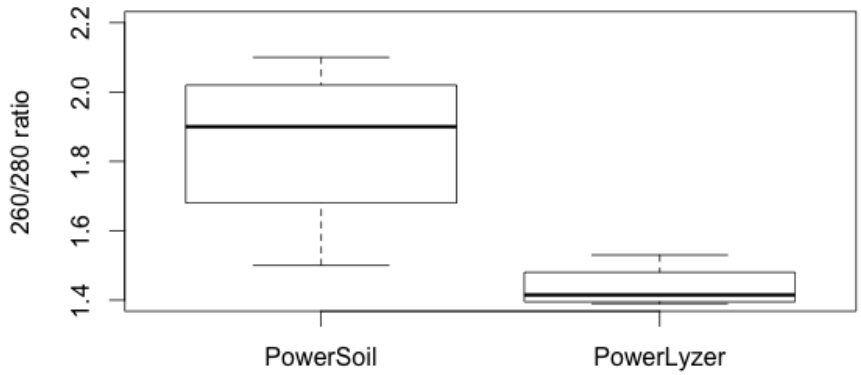
The results of gel electrophoresis indicate that the DNA strand lengths are similar between the kits, but the PowerLyzer™ PowerSoil® DNA has achieved a greater yield (Supporting Information Figure A.3.3.4).

Supporting Information Table A.3.3.1. DNA concentrations. Mean DNA concentrations, 260:280 absorbance ratios and 260:230 absorbance ratios for the PowerSoil® and the PowerLyzer™ PowerSoil® DNA isolation kits, determined using a NanoDrop™ 8000 UV-Vis Spectrophotometer.

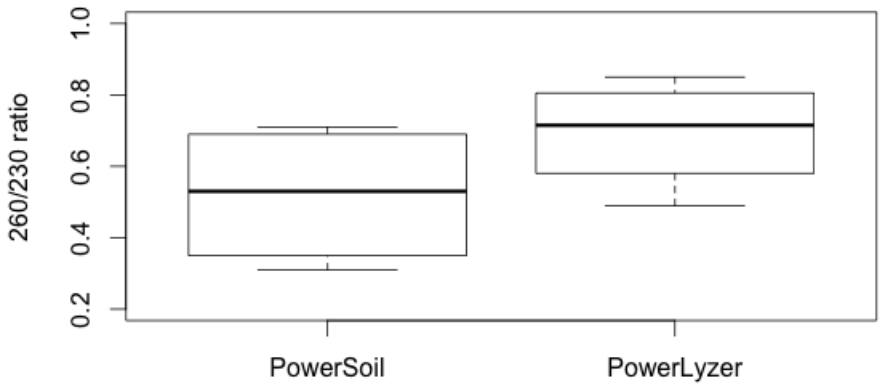
Kit	DNA conc. (ng/µl)	260:280 ratio	260:230 ratio
PowerSoil® (4 reps)	5.54	1.85	0.52
PowerLyzer™ (4 reps)	20.46	1.44	0.69



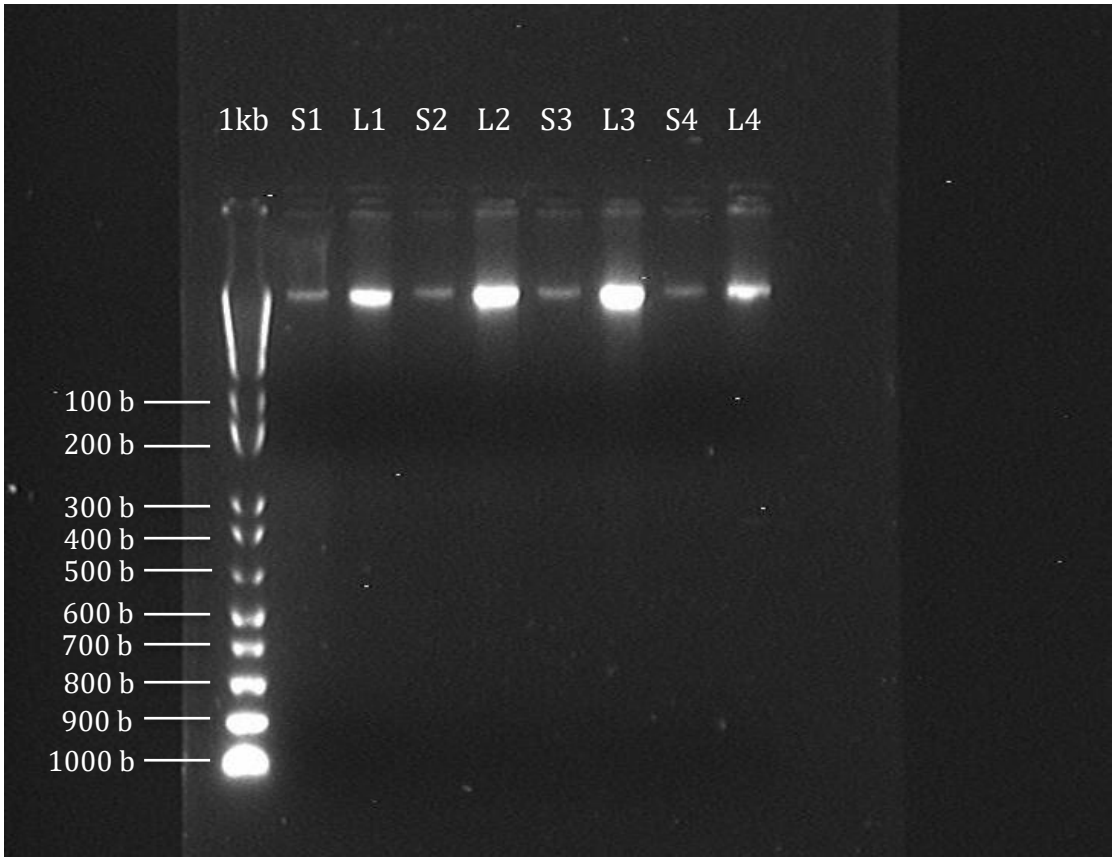
Supporting Information Figure A.3.3.1. DNA concentrations. Box plot depicting the ranges, quartiles and means for the DNA concentrations achieved from the PowerSoil® and the PowerLyzer™ PowerSoil® DNA isolation kits.



Supporting Information Figure A.3.3.2. DNA concentrations. Box plot depicting the ranges, quartiles and means for the DNA 260:280 absorbance ratios achieved from the PowerSoil® and the PowerLyzer™ PowerSoil® DNA isolation kits.



Supporting Information Figure A.3.3.3. Box plot depicting the ranges, quartiles and means for the DNA 260:230 absorbance ratios achieved from the PowerSoil® and the PowerLyzer™ PowerSoil® DNA isolation kits.



Supporting Information Figure A.3.3.4. Gel electrophoresis. The gel electrophoresis results displaying the DNA isolated using the PowerSoil® (S) and the PowerLyzer™ PowerSoil® (L) DNA isolation kits. The left lane is a 1kb commercially available marker.

Conclusion

The PowerLyzer™ PowerSoil® DNA isolation kit extracts more DNA from the loamy, clayey soils that the PowerSoil® kit, however the purity of DNA isolated with the PowerSoil® is greater than that of the PowerLyzer™ PowerSoil® kit. The quantity of DNA isolated from the PowerSoil® kit is still sufficient enough for metagenomic sequencing.

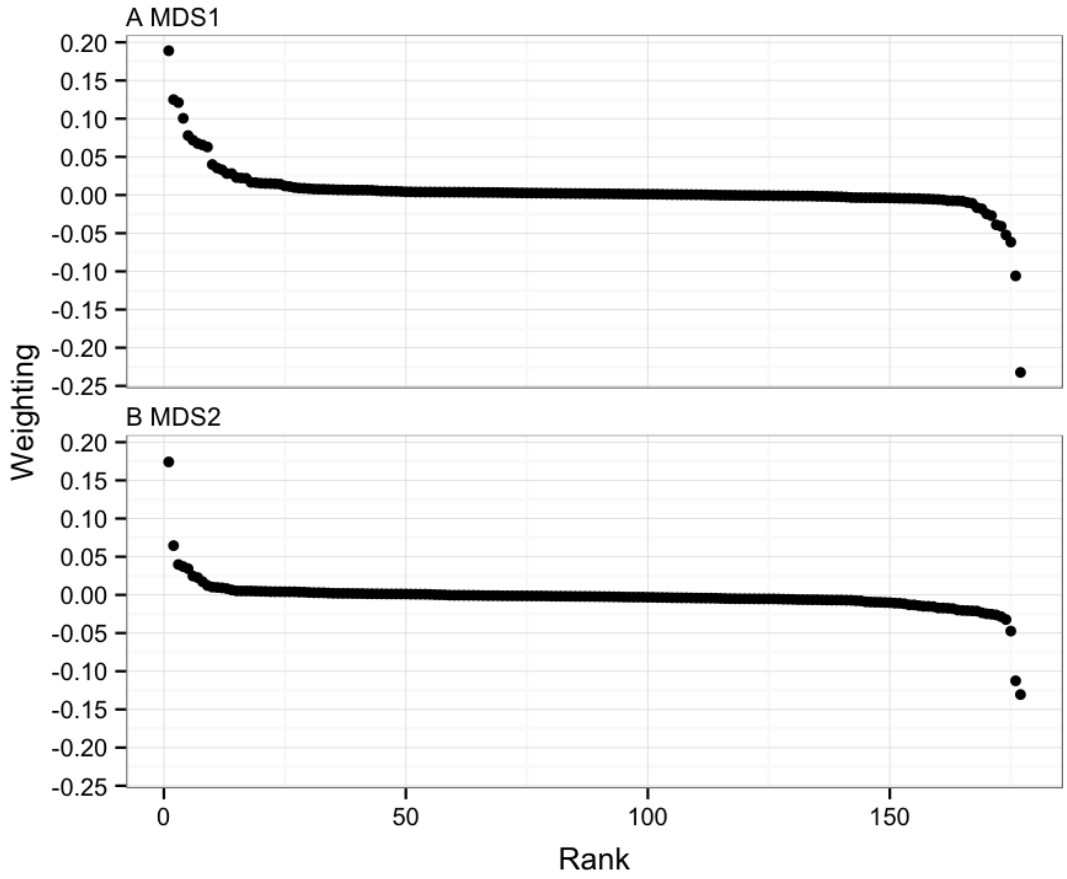


Figure A.7. Order PCoA component weightings. Ranked PCoA component weightings for components (MDS) 1 and 2 at the order level.

Table A.9. The barcodes used in the duel multiplexing system.

Sample	Index 1 title	Index 1 sequence	Index 2 title	Index 2 sequence
Start 1	N701	TAAGGCGA	S501	TAGATCGC
Start 2	N702	CGTACTAG	S501	TAGATCGC
Start 3	N703	AGGCAGAA	S501	TAGATCGC
1 x Flood 1	N707	CTCTCTAC	S502	CTCTCTAT
1 x Flood 2	N708	CAGAGAGG	S502	CTCTCTAT
1 x Flood 3	N709	GCTACGCT	S502	CTCTCTAT
3 x Floods 1	N704	TCCTGAGC	S501	TAGATCGC
3 x Floods 2	N705	GGACTCCT	S501	TAGATCGC
3 x Floods 3	N706	TAGGCATG	S502	CTCTCTAT

Table A.10. Start vs. 1 Flood order absolute change. The greatest absolute changes in order relative abundances between the starting soil and soil that received one flood. Data for the three-flooded soil is included for comparison.

Order	Start ($\bar{x}$)	1 x Flood ($\bar{x}$)	3 x Floods ($\bar{x}$)	1 x Flood Absolute change	3 x Floods Absolute change
Actinomycetales	2.50^{E-01}	3.07^{E-01}	1.98^{E-01}	5.70^{E-02}	-5.29^{E-02}
Solibacterales	7.00^{E-02}	7.84^{E-02}	1.18^{E-01}	8.46^{E-03}	4.79^{E-02}
Acidobacteriales	2.73^{E-02}	3.52^{E-02}	4.94^{E-02}	7.91^{E-03}	2.21^{E-02}
Caulobacterales	6.29^{E-03}	1.12^{E-02}	8.58^{E-03}	4.87^{E-03}	2.29^{E-03}
Myxococcales	2.71^{E-02}	3.16^{E-02}	4.79^{E-02}	4.51^{E-03}	2.08^{E-02}
Burkholderiales	5.20^{E-02}	5.56^{E-02}	6.63^{E-02}	3.57^{E-03}	1.43^{E-02}
Nitrosomonadales	3.39^{E-03}	5.81^{E-03}	8.09^{E-03}	2.42^{E-03}	4.70^{E-03}
Sphingomonadales	5.92^{E-03}	8.21^{E-03}	6.18^{E-03}	2.29^{E-03}	2.60^{E-04}
Gemmatimonadales	3.66^{E-03}	5.12^{E-03}	6.63^{E-03}	1.47^{E-03}	2.97^{E-03}
Gallionellales	1.14^{E-03}	1.51^{E-03}	5.23^{E-03}	3.70^{E-04}	4.09^{E-03}
Cytophagales	5.98^{E-03}	4.57^{E-03}	7.19^{E-03}	-1.41^{E-03}	1.21^{E-03}
Verrucomicrobiales	2.56^{E-02}	2.40^{E-02}	3.35^{E-02}	-1.61^{E-03}	7.95^{E-03}
Chloroflexales	9.04^{E-03}	7.41^{E-03}	7.54^{E-03}	-1.63^{E-03}	-1.50^{E-03}
Rhodobacterales	1.71^{E-02}	1.52^{E-02}	1.39^{E-02}	-1.91^{E-03}	-3.22^{E-03}
Chroococcales	1.12^{E-02}	9.22^{E-03}	1.05^{E-02}	-1.93^{E-03}	-6.95^{E-04}
Ktedonobacterales	1.02^{E-02}	7.93^{E-03}	6.55^{E-03}	-2.23^{E-03}	-3.60^{E-03}
Sphingobacteriales	6.29^{E-03}	3.91^{E-03}	6.53^{E-03}	-2.38^{E-03}	2.34^{E-04}
Rhodospirillales	2.09^{E-02}	1.69^{E-02}	1.50^{E-02}	-4.01^{E-03}	-5.88^{E-03}
Planctomycetales	3.99^{E-02}	2.11^{E-02}	2.41^{E-02}	-1.88^{E-02}	-1.58^{E-02}
Rhizobiales	2.07^{E-01}	1.69^{E-01}	1.54^{E-01}	-3.85^{E-02}	-5.28^{E-02}

Table A.11. Start vs. 3 Floods order absolute change. The greatest absolute changes in order relative abundances between the starting soil and soil that received three floods. Data for the single-flood soil is included for comparison.

Order	Start ($\bar{x}$)	1 x Flood ($\bar{x}$)	3 x Floods ($\bar{x}$)	1 x Flood Absolute change	3 x Floods Absolute change
Solibacterales	7.00^{E-02}	7.84^{E-02}	1.18^{E-01}	8.46^{E-03}	4.79^{E-02}
Acidobacteriales	2.73^{E-02}	3.52^{E-02}	4.94^{E-02}	7.91^{E-03}	2.21^{E-02}
Myxococcales	2.71^{E-02}	3.16^{E-02}	4.79^{E-02}	4.51^{E-03}	2.08^{E-02}
Burkholderiales	5.20^{E-02}	5.56^{E-02}	6.63^{E-02}	3.57^{E-03}	1.43^{E-02}
Verrucomicrobiales	2.56^{E-02}	2.40^{E-02}	3.35^{E-02}	-1.61^{E-03}	7.95^{E-03}
Desulfuromonadales	1.25^{E-02}	1.24^{E-02}	1.92^{E-02}	-1.27^{E-04}	6.68^{E-03}
Nitrosomonadales	3.39^{E-03}	5.81^{E-03}	8.09^{E-03}	2.42^{E-03}	4.70^{E-03}
Gallionellales	1.14^{E-03}	1.51^{E-03}	5.23^{E-03}	3.70^{E-04}	4.09^{E-03}
Gemmatimonadales	3.66^{E-03}	5.12^{E-03}	6.63^{E-03}	1.47^{E-03}	2.97^{E-03}
Bacteroidales	3.21^{E-03}	2.74^{E-03}	6.14^{E-03}	-4.71^{E-04}	2.93^{E-03}
Bacillales	1.53^{E-02}	1.40^{E-02}	1.41^{E-02}	-1.27^{E-03}	-1.17^{E-03}
Rubrobacterales	4.79^{E-03}	4.44^{E-03}	3.38^{E-03}	-3.44^{E-04}	-1.40^{E-03}
Chloroflexales	9.04^{E-03}	7.41^{E-03}	7.54^{E-03}	-1.63^{E-03}	-1.50^{E-03}
Rhodobacterales	1.71^{E-02}	1.52^{E-02}	1.39^{E-02}	-1.91^{E-03}	-3.22^{E-03}
Ktedonobacterales	1.02^{E-02}	7.93^{E-03}	6.55^{E-03}	-2.23^{E-03}	-3.60^{E-03}
Solirubrobacterales	1.66^{E-02}	1.67^{E-02}	1.16^{E-02}	5.64^{E-05}	-5.03^{E-03}
Rhodospirillales	2.09^{E-02}	1.69^{E-02}	1.50^{E-02}	-4.01^{E-03}	-5.88^{E-03}
Planctomycetales	3.99^{E-02}	2.11^{E-02}	2.41^{E-02}	-1.88^{E-02}	-1.58^{E-02}
Rhizobiales	2.07^{E-01}	1.69^{E-01}	1.54^{E-01}	-3.85^{E-02}	-5.28^{E-02}
Actinomycetales	2.50^{E-01}	3.07^{E-01}	1.98^{E-01}	5.70^{E-02}	-5.29^{E-02}

Table A.12. Start vs. 1 Flood order fold change. The greatest fold changes in order relative abundances between the starting soil and soil that received one flood. Data for the three-flooded soil is included for comparison.

Order	Start ($\bar{x}$)	1 x Flood ($\bar{x}$)	3 x Floods ($\bar{x}$)	1 x Flood Fold change	3 x Floods Fold change
Marchantiales	3.00^{E-07}	3.46^{E-06}	9.90^{E-06}	11.535	33.056
Bangiales	3.00^{E-07}	2.82^{E-06}	0.00	9.416	0.000
Euglyphida	2.40^{E-06}	9.71^{E-06}	1.54^{E-05}	4.052	6.424
Entomoplasmatales	8.99^{E-07}	2.82^{E-06}	7.24^{E-06}	3.139	8.049
Spirurida	5.19^{E-05}	1.14^{E-04}	7.24^{E-05}	2.200	1.395
Hymenostomatida	2.22^{E-05}	4.57^{E-05}	5.26^{E-05}	2.061	2.372
Haemosporida	3.78^{E-05}	6.88^{E-05}	3.98^{E-05}	1.820	1.053
Caulobacterales	6.29^{E-03}	1.12^{E-02}	8.58^{E-03}	1.774	1.364
Nitrosomonadales	3.39^{E-03}	5.81^{E-03}	8.09^{E-03}	1.713	2.383
Peniculida	3.65^{E-05}	6.21^{E-05}	7.53^{E-05}	1.703	2.066
Coleochaetales	5.99^{E-07}	0.00	3.06^{E-06}	0.000	5.109
Asparagales	9.20^{E-07}	0.00	0.00	0.000	0.000
Caryophyllales	9.20^{E-07}	0.00	7.19^{E-07}	0.000	0.782
Cyrtocrinida	9.20^{E-07}	0.00	0.00	0.000	0.000
Synurales	9.20^{E-07}	0.00	0.00	0.000	0.000
Eupodiscales	2.20^{E-06}	0.00	7.19^{E-07}	0.000	0.327
Polypodiales	2.20^{E-06}	0.00	0.00	0.000	0.000
Pseudoscourfieldiales	2.44^{E-06}	0.00	1.11^{E-06}	0.000	0.456
Fabales	3.10^{E-06}	0.00	0.00	0.000	0.000

Table A.13. Start vs. 3 Floods order fold change. The greatest fold changes in order relative abundances between the starting soil and soil that received three floods. Data for the single-flood soil is included for comparison.

Order	Start ($\bar{x}$)	1 x Flood ($\bar{x}$)	3 x Floods ($\bar{x}$)	1 x Flood Fold change	3 x Floods Fold change
Marchantiales	3.00^{E-07}	3.46^{E-06}	9.90^{E-06}	11.535	33.056
Zygnematales	3.00^{E-07}	0.00	3.06^{E-06}	0.000	10.218
Entomoplasmatales	8.99^{E-07}	2.82^{E-06}	7.24^{E-06}	3.139	8.049
Euglyphida	2.40^{E-06}	9.71^{E-06}	1.54^{E-05}	4.052	6.424
Coleochaetales	5.99^{E-07}	0.00	3.06^{E-06}	0.000	5.109
Gallionellales	1.14^{E-03}	1.51^{E-03}	5.23^{E-03}	1.326	4.602
Coleoptera	1.70^{E-05}	2.52^{E-05}	5.12^{E-05}	1.488	3.019
Pyrenomonadales	3.00^{E-06}	3.56^{E-06}	7.84^{E-06}	1.189	2.615
Fibrobacterales	1.03^{E-04}	1.29^{E-04}	2.66^{E-04}	1.252	2.576
Chlorellales	1.22^{E-06}	4.53^{E-07}	3.06^{E-06}	0.371	2.510
Chlorokybales	9.20^{E-07}	1.50^{E-06}	0.00	1.632	0.000
Echinorhynchida	9.20^{E-07}	4.53^{E-07}	0.00	0.492	0.000
Asparagales	9.20^{E-07}	0.00	0.00	0.000	0.000
Cyrtocrinida	9.20^{E-07}	0.00	0.00	0.000	0.000
Synurales	9.20^{E-07}	0.00	0.00	0.000	0.000
Glomerales	1.52^{E-06}	1.05^{E-06}	0.00	0.691	0.000
Polypodiales	2.20^{E-06}	0.00	0.00	0.000	0.000
Vaucheriales	2.50^{E-06}	9.05^{E-07}	0.00	0.362	0.000
Fabales	3.10^{E-06}	0.00	0.00	0.000	0.000
Capnodiales	6.54^{E-06}	1.46^{E-06}	0.00	0.224	0.000

Table A.14. Start vs. 1 Flood function absolute change. The greatest absolute changes in Subsystems level 3 function relative abundances between the starting soil and soil that received one flood. Data for the three-flooded soil is included for comparison.

Function	Start ($\bar{x}$)	1 x Flood ($\bar{x}$)	3 x Floods ($\bar{x}$)	1 x Flood Absolute change	3 x Floods Absolute change
Ton and Tol transport systems	3.20^{E-03}	3.72^{E-03}	4.59^{E-03}	5.25^{E-04}	1.39^{E-03}
Fatty acid degradation regulons	7.08^{E-03}	7.55^{E-03}	6.68^{E-03}	4.74^{E-04}	-3.93^{E-04}
Iron acquisition in Vibrio	1.96^{E-03}	2.38^{E-03}	2.81^{E-03}	4.19^{E-04}	8.48^{E-04}
Biotin biosynthesis	4.96^{E-03}	5.37^{E-03}	4.75^{E-03}	4.04^{E-04}	-2.10^{E-04}
CBSS-316057.3.peg.1308	2.61^{E-03}	3.01^{E-03}	2.56^{E-03}	4.02^{E-04}	-5.35^{E-05}
Fatty acid metabolism cluster	6.04^{E-03}	6.40^{E-03}	5.66^{E-03}	3.62^{E-04}	-3.74^{E-04}
n-Phenylalkanoic acid degradation	6.04^{E-03}	6.40^{E-03}	5.66^{E-03}	3.61^{E-04}	-3.76^{E-04}
Sugar utilization in Thermotogales	1.29^{E-02}	1.32^{E-02}	1.45^{E-02}	3.48^{E-04}	1.63^{E-03}
Lysine fermentation	3.60^{E-03}	3.94^{E-03}	3.63^{E-03}	3.41^{E-04}	2.96^{E-05}
Leucine Degradation and HMG-CoA Metabolism	3.39^{E-03}	3.69^{E-03}	3.38^{E-03}	2.99^{E-04}	-9.14^{E-06}
Universal GTPases	5.20^{E-03}	4.90^{E-03}	5.22^{E-03}	-2.98^{E-04}	1.98^{E-05}
Sulfatases and sulfatase modifying factor 1	3.72^{E-03}	3.42^{E-03}	3.40^{E-03}	-3.00^{E-04}	-3.20^{E-04}
RNA polymerase bacterial	3.06^{E-03}	2.76^{E-03}	2.95^{E-03}	-3.02^{E-04}	-1.10^{E-04}
Respiratory Complex I	4.87^{E-03}	4.54^{E-03}	5.34^{E-03}	-3.29^{E-04}	4.67^{E-04}
Alkanesulfonate assimilation	3.80^{E-03}	3.38^{E-03}	3.28^{E-03}	-4.28^{E-04}	-5.22^{E-04}

Function	Start ($\bar{x}$)	1 x Flood ($\bar{x}$)	3 x Floods ($\bar{x}$)	1 x Flood Absolute change	3 x Floods Absolute change
Phage integration and excision	5.23^{E-03}	4.79^{E-03}	4.17^{E-03}	-4.34^{E-04}	-1.05^{E-03}
ABC transporter branched-chain amino acid (TC 3.A.1.4.1)	3.73^{E-03}	3.18^{E-03}	3.14^{E-03}	-5.48^{E-04}	-5.90^{E-04}
cAMP signaling in bacteria	8.19^{E-03}	6.78^{E-03}	6.22^{E-03}	-1.42^{E-03}	-1.97^{E-03}
CBSS-222523.1.peg.1311	7.83^{E-03}	6.36^{E-03}	5.91^{E-03}	-1.46^{E-03}	-1.91^{E-03}
Iojap	1.00^{E-02}	8.29^{E-03}	8.14^{E-03}	-1.71^{E-03}	-1.86^{E-03}

Table A.15. Start vs. 3 Floods function absolute change. The greatest absolute changes in Subsystems level 3 function relative abundances between the starting soil and soil that received three floods. Data for the single-flood soil is included for comparison.

Function	Start ($\bar{x}$)	1 x Flood ($\bar{x}$)	3 x Floods ($\bar{x}$)	1 x Flood Absolute change	3 x Floods Absolute change
Sugar utilization in Thermotogales	1.29^{E-02}	1.32^{E-02}	1.45^{E-02}	3.48^{E-04}	1.63^{E-03}
Ton and Tol transport systems	3.20^{E-03}	3.72^{E-03}	4.59^{E-03}	5.25^{E-04}	1.39^{E-03}
Cobalt-zinc-cadmium resistance	5.34^{E-03}	5.05^{E-03}	6.67^{E-03}	-2.95^{E-04}	1.33^{E-03}
Bacterial Chemotaxis	2.26^{E-03}	2.32^{E-03}	3.15^{E-03}	6.08^{E-05}	8.89^{E-04}
Flagellar motility	1.74^{E-03}	2.00^{E-03}	2.61^{E-03}	2.61^{E-04}	8.64^{E-04}
Iron acquisition in Vibrio	1.96^{E-03}	2.38^{E-03}	2.81^{E-03}	4.19^{E-04}	8.48^{E-04}
Lactose and Galactose Uptake and Utilization	2.67^{E-03}	2.69^{E-03}	3.33^{E-03}	1.75^{E-05}	6.60^{E-04}

Function	Start ($\bar{x}$)	1 x Flood ($\bar{x}$)	3 x Floods ($\bar{x}$)	1 x Flood Absolute change	3 x Floods Absolute change
Hydrogenases	1.34^{E-03}	1.45^{E-03}	1.99^{E-03}	1.09^{E-04}	6.56^{E-04}
Flagellum	3.05^{E-03}	3.29^{E-03}	3.66^{E-03}	2.41^{E-04}	6.10^{E-04}
C jejuni colonization of chick caeca	1.79^{E-03}	1.92^{E-03}	2.31^{E-03}	1.30^{E-04}	5.18^{E-04}
Serine-glyoxylate cycle	1.53^{E-02}	1.51^{E-02}	1.48^{E-02}	-2.01^{E-04}	-5.05^{E-04}
Alkanesulfonate assimilation	3.80^{E-03}	3.38^{E-03}	3.28^{E-03}	-4.28^{E-04}	-5.22^{E-04}
ABC transporter branched-chain amino acid (TC 3.A.1.4.1)	3.73^{E-03}	3.18^{E-03}	3.14^{E-03}	-5.48^{E-04}	-5.90^{E-04}
Trehalose Biosynthesis	5.52^{E-03}	5.48^{E-03}	4.85^{E-03}	-4.17^{E-05}	-6.78^{E-04}
Phage integration and excision	5.23^{E-03}	4.79^{E-03}	4.17^{E-03}	-4.34^{E-04}	-1.05^{E-03}
CBSS-314269.3.peg.1840	3.07^{E-03}	2.86^{E-03}	1.98^{E-03}	-2.07^{E-04}	-1.09^{E-03}
CO Dehydrogenase	3.19^{E-03}	3.02^{E-03}	2.10^{E-03}	-1.68^{E-04}	-1.09^{E-03}
Iojap	1.00^{E-02}	8.29^{E-03}	8.14^{E-03}	-1.71^{E-03}	-1.86^{E-03}
CBSS-222523.1.peg.1311	7.83^{E-03}	6.36^{E-03}	5.91^{E-03}	-1.46^{E-03}	-1.91^{E-03}
cAMP signaling in bacteria	8.19^{E-03}	6.78^{E-03}	6.22^{E-03}	-1.42^{E-03}	-1.97^{E-03}

A.4 Chapter 4 Supporting information

Table A.16. Sequence counts and phred scores. Sequence count and quality statistics for the raw sequence data.

Time	Treatment	Sequence direction	Replicate	Sequence count	Mean phred score	% phred >= 30
Start	Short flood	Forward	1	13,178,010	34.89	88.87
Start	Short flood	Reverse	1	13,178,010	33.86	85.10
Start	Short flood	Forward	2	13,519,539	35.26	90.37
Start	Short flood	Reverse	2	13,519,539	32.99	82.51
Start	Short flood	Forward	3	8,096,419	35.36	90.70
Start	Short flood	Reverse	3	8,096,419	33.63	84.54
Start	Short flood	Forward	4	14,955,948	35.36	90.69
Start	Short flood	Reverse	4	14,955,948	33.80	85.30
Start	Long flood	Forward	1	13,667,294	35.27	90.24
Start	Long flood	Reverse	1	13,667,294	33.16	83.10
Start	Long flood	Forward	2	21,382,061	35.49	91.13
Start	Long flood	Reverse	2	21,382,061	33.89	85.49
Start	Long flood	Forward	3	9,246,737	35.20	90.13
Start	Long flood	Reverse	3	9,246,737	32.95	82.58
Start	Long flood	Forward	4	17,236,374	35.53	91.29
Start	Long flood	Reverse	4	17,236,374	33.65	84.79
End	Short flood	Forward	1	14,218,936	35.26	90.19
End	Short flood	Reverse	1	14,218,936	33.13	82.69
End	Short flood	Forward	2	18,937,461	35.06	89.73
End	Short flood	Reverse	2	18,937,461	32.63	81.36
End	Short flood	Forward	3	10,306,388	35.66	91.94
End	Short flood	Reverse	3	10,306,388	33.64	84.62
End	Short flood	Forward	4	7,315,247	35.64	91.78
End	Short flood	Reverse	4	7,315,247	34.40	87.34
End	Long flood	Forward	1	16,737,273	35.55	91.35
End	Long flood	Reverse	1	16,737,273	33.94	85.70
End	Long flood	Forward	2	34,963,404	35.58	91.51
End	Long flood	Reverse	2	34,963,404	34.23	86.65
End	Long flood	Forward	3	14,811,291	35.57	91.55

Time	Treatment	Sequence direction	Replicate	Sequence count	Mean phred score	% phred >= 30
End	Long flood	Reverse	3	14,811,291	33.84	85.40
End	Long flood	Forward	4	11,282,184	35.30	90.42
End	Long flood	Reverse	4	11,282,184	33.71	84.97

Table A.17. Sequence counts and phred scores statistics. Sequence counts and phred scores statistics for the sequences with residual adapters and sequences shorter than 30 bp removed.

Time	Treatment	Sequence direction	Replicate	Sequence count	Mean phred score	% phred >= 30
Start	Short flood	Forward	1	13,167,459	34.91	88.91
Start	Short flood	Reverse	1	13,178,010	33.86	85.10
Start	Short flood	Forward	2	13,348,293	35.44	90.87
Start	Short flood	Reverse	2	13,519,532	33.00	82.52
Start	Short flood	Forward	3	8,041,376	35.46	90.99
Start	Short flood	Reverse	3	8,096,418	33.63	84.55
Start	Short flood	Forward	4	14,870,998	35.44	90.92
Start	Short flood	Reverse	4	14,955,948	33.81	85.31
Start	Long flood	Forward	1	13,641,879	35.30	90.33
Start	Long flood	Reverse	1	13,667,294	33.16	83.10
Start	Long flood	Forward	2	21,353,113	35.51	91.19
Start	Long flood	Reverse	2	21,382,061	33.89	85.50
Start	Long flood	Forward	3	9,134,143	35.37	90.63
Start	Long flood	Reverse	3	9,246,737	32.95	82.59
Start	Long flood	Forward	4	17,188,578	35.57	91.41
Start	Long flood	Reverse	4	17,236,372	33.65	84.79
End	Short flood	Forward	1	14,178,596	35.30	90.31
End	Short flood	Reverse	1	14,218,935	33.13	82.70
End	Short flood	Forward	2	18,517,944	35.37	90.61
End	Short flood	Reverse	2	18,937,427	32.63	81.37
End	Short flood	Forward	3	10,179,397	35.84	92.44
End	Short flood	Reverse	3	10,306,384	33.64	84.63
End	Short flood	Forward	4	7,283,479	35.71	91.96

Time	Treatment	Sequence direction	Replicate	Sequence count	Mean phred score	% phred >= 30
End	Short flood	Reverse	4	7,315,247	34.40	87.35
End	Long flood	Forward	1	16,727,417	35.56	91.38
End	Long flood	Reverse	1	16,737,272	33.94	85.70
End	Long flood	Forward	2	34,819,041	35.64	91.69
End	Long flood	Reverse	2	34,963,404	34.24	86.66
End	Long flood	Forward	3	14,681,262	35.70	91.91
End	Long flood	Reverse	3	14,811,290	33.85	85.41
End	Long flood	Forward	4	11,243,207	35.35	90.57
End	Long flood	Reverse	4	11,282,184	33.71	84.98

Table A.18. The counts, lengths and mapped sequence statistics. Key to sample code characters: 1st/Time: S = Start, E = End; 2nd/Treatment: S = Short flood, L = Long flood; 3rd/Sequence direction: F = Forward, R = Reverse; 4th/Replicate: # = replicate number.

Sample	Contig count	Mapped sequence count	Max seq length	Median seq length	N50
SS1	180,243	4,972,649	16,139	400	528
SS2	226,579	6,693,856	17,916	414	582
SS3	154,695	4,094,516	13,568	418	590
SS4	272,214	8,005,390	15,075	420	608
SL1	266,851	7,706,878	19,308	422	606
Sl2	371,180	11,557,709	14,903	411	587
SL3	138,106	4,000,078	17,680	409	559
SL4	338,877	10,308,764	17,525	426	633
ES1	255,144	7,077,006	22,142	409	546
ES2	325,203	10,176,589	21,584	412	592
ES3	146,316	3,850,882	10,309	409	538
ES4	103,085	2,560,460	9,825	408	531
EL1	353,486	11,038,378	21,384	436	688
EL2	782,791	29,766,205	71,600	434	775
EL3	278,866	8,588,956	15,863	417	608
EL4	171,077	4,621,702	13,897	409	543

Table A.19. Merged sequence counts and phred scores statistics. Key to sample code characters: 1st/Time: S = Start, E = End; 2nd/Treatment: S = Short flood, L = Long flood; 3rd/Sequence direction: F = Forward, R = Reverse; 4th/Replicate: # = replicate number.

Sample	Sequence count	Median seq length	Mean phred score	% phred >= 30
SS1	251,235	212	37.66	97.44
SS2	268,808	208	37.76	97.57
SS3	166,112	209	37.76	97.62
SS4	390,275	201	37.91	97.93
SL1	248,511	211	37.66	97.36
Sl2	422,734	212	37.71	97.60
SL3	199,068	210	37.71	97.49
SL4	350,674	208	37.79	97.70
ES1	262,519	208	37.75	97.49
ES2	348,790	209	37.73	97.49
ES3	251,109	208	37.84	97.89
ES4	198,899	207	37.86	98.00
EL1	337,765	206	37.83	97.77
EL2	649,013	205	37.89	97.96
EL3	352,470	203	37.91	97.97
EL4	311,681	202	37.89	97.85

Table A.20. Singleton sequence counts and phred scores statistics.

Time	Treatment	Sequence direction	Replicate	Sequence count	Mean phred score	% phred >= 30
Start	Short flood	Forward	1	10,636,323	34.90	88.87
Start	Short flood	Reverse	1	10,664,965	33.78	84.84
Start	Short flood	Forward	2	9,965,021	35.43	90.83
Start	Short flood	Reverse	2	9,951,675	32.87	82.13
Start	Short flood	Forward	3	5,970,745	35.45	90.92
Start	Short flood	Reverse	3	5,977,166	33.52	84.20
Start	Short flood	Forward	4	10,806,900	35.44	90.89
Start	Short flood	Reverse	4	10,833,621	33.66	84.86

Time	Treatment	Sequence direction	Replicate	Sequence count	Mean phred score	% phred >= 30
Start	Long flood	Forward	1	9,681,863	35.28	90.23
Start	Long flood	Reverse	1	9,800,102	32.78	81.97
Start	Long flood	Forward	2	15,451,662	35.51	91.18
Start	Long flood	Reverse	2	15,549,337	33.70	84.92
Start	Long flood	Forward	3	7,111,400	35.37	90.61
Start	Long flood	Reverse	3	7,105,220	32.83	82.25
Start	Long flood	Forward	4	11,915,528	35.57	91.38
Start	Long flood	Reverse	4	12,020,533	33.37	83.95
End	Short flood	Forward	1	10,558,122	35.29	90.27
End	Short flood	Reverse	1	10,628,150	32.91	82.03
End	Short flood	Forward	2	13,394,177	35.36	90.57
End	Short flood	Reverse	2	13,293,904	32.59	81.26
End	Short flood	Forward	3	8,261,141	35.84	92.44
End	Short flood	Reverse	3	8,206,358	33.70	84.79
End	Short flood	Forward	4	5,990,804	35.71	91.95
End	Short flood	Reverse	4	5,986,205	34.37	87.25
End	Long flood	Forward	1	11,087,300	35.55	91.31
End	Long flood	Reverse	1	11,197,462	33.63	84.78
End	Long flood	Forward	2	19,678,772	35.63	91.63
End	Long flood	Reverse	2	19,807,989	33.99	85.93
End	Long flood	Forward	3	10,326,854	35.69	91.87
End	Long flood	Reverse	3	10,333,408	33.73	85.07
End	Long flood	Forward	4	8,894,423	35.35	90.53
End	Long flood	Reverse	4	8,919,104	33.58	84.58

Table A.21. The processed sequence/contig counts and lengths.

Time	Treatment	Replicate	Sequence count	Median seq length	Max seq length
Start	Short flood	1	21,732,766	125	16,139
Start	Short flood	2	20,412,083	125	17,916
Start	Short flood	3	12,268,718	125	13,568
Start	Short flood	4	22,303,010	125	15,075
Start	Long flood	1	19,997,327	125	19,308
Start	Long flood	2	31,794,913	125	14,903
Start	Long flood	3	14,553,794	125	17,680
Start	Long flood	4	24,625,612	125	17,525
End	Short flood	1	21,703,935	125	22,142
End	Short flood	2	27,362,074	125	21,584
End	Short flood	3	16,864,924	125	10,309
End	Short flood	4	12,278,993	125	9,825
End	Long flood	1	22,976,013	125	21,384
End	Long flood	2	40,918,565	125	71,600
End	Long flood	3	21,291,598	125	15,863
End	Long flood	4	18,296,285	125	13,897

Table A.22. Bacterial abundance variations. The 10 greatest bacterial order relative abundance variations between the treatments.

Order	Start Short Flood ($\bar{x}$)	Start Long Flood ($\bar{x}$)	End Short Flood ($\bar{x}$)	End Long Flood ($\bar{x}$)	Short flood fold change (%)	Long flood fold change (%)	Diff. (Long vs. Short flood) (%)
Acholeplasmatales	$5.0^{E-05} \pm 6.7^{E-06}$	$5.3^{E-05} \pm 1.0^{E-05}$	$4.7^{E-05} \pm 5.3^{E-06}$	$4.2^{E-05} \pm 4.1^{E-06}$	94.00	78.66	83.68
Sphingomonadales	$9.5^{E-03} \pm 1.0^{E-03}$	$8.8^{E-03} \pm 6.2^{E-04}$	$9.1^{E-03} \pm 2.8^{E-04}$	$9.6^{E-03} \pm 1.1^{E-03}$	96.21	108.69	112.96
Gemmatimonadales	$3.9^{E-03} \pm 3.2^{E-04}$	$4.4^{E-03} \pm 3.9^{E-04}$	$4.1^{E-03} \pm 4.7^{E-04}$	$4.0^{E-03} \pm 1.8^{E-04}$	104.30	91.71	87.93
Verrucomicrobiales	$1.9^{E-02} \pm 2.7^{E-03}$	$2.1^{E-02} \pm 3.8^{E-03}$	$1.9^{E-02} \pm 2.4^{E-03}$	$1.8^{E-02} \pm 8.2^{E-04}$	99.23	87.98	88.66
unclassified (Poribacteria)	$3.5^{E-04} \pm 3.1^{E-05}$	$3.8^{E-04} \pm 7.1^{E-05}$	$3.4^{E-04} \pm 6.3^{E-05}$	$3.3^{E-04} \pm 3.5^{E-05}$	96.77	85.81	88.68
Nitrospirales	$3.1^{E-03} \pm 3.6^{E-04}$	$3.4^{E-03} \pm 6.8^{E-04}$	$3.0^{E-03} \pm 4.8^{E-04}$	$3.0^{E-03} \pm 3.4^{E-04}$	97.96	87.20	89.02
Methylacidiphilales	$1.3^{E-03} \pm 7.6^{E-05}$	$1.4^{E-03} \pm 1.3^{E-04}$	$1.4^{E-03} \pm 1.4^{E-04}$	$1.3^{E-03} \pm 2.0^{E-05}$	107.09	95.35	89.04
Puniceicoccales	$7.5^{E-04} \pm 1.3^{E-04}$	$8.6^{E-04} \pm 1.5^{E-04}$	$7.4^{E-04} \pm 1.1^{E-04}$	$7.6^{E-04} \pm 5.3^{E-05}$	98.45	87.92	89.31
unclassified (Opitutae)	$7.0^{E-03} \pm 9.3^{E-04}$	$8.0^{E-03} \pm 1.5^{E-03}$	$7.0^{E-03} \pm 9.7^{E-04}$	$7.1^{E-03} \pm 4.8^{E-04}$	100.25	89.71	89.49

Order	Start Short Flood ($\bar{x}$)	Start Long Flood ($\bar{x}$)	End Short Flood ($\bar{x}$)	End Long Flood ($\bar{x}$)	Short flood fold change (%)	Long flood fold change (%)	Diff. (Long vs. Short flood) (%)
Elusimicrobiales	$1.8^{E-04} \pm 1.8^{E-05}$	$2.0^{E-04} \pm 4.4^{E-05}$	$1.8^{E-04} \pm 2.9^{E-05}$	$1.7^{E-04} \pm 7.9^{E-06}$	97.27	87.24	89.69

Table A.23. Archaeal abundance variations. The 10 greatest archaeal order relative abundance variations between the treatments.

Order	Start Short Flood ($\bar{x}$)	Start Long Flood ($\bar{x}$)	End Short Flood ($\bar{x}$)	End Long Flood ($\bar{x}$)	Short flood fold change (%)	Long flood fold change (%)	Diff. (Long vs. Short flood) (%)
Nitrosopumilales	$5.3^{E-02} \pm 1.7^{E-02}$	$4.1^{E-02} \pm 1.4^{E-02}$	$6.5^{E-02} \pm 1.2^{E-02}$	$5.4^{E-02} \pm 7.7^{E-03}$	122.99	133.21	108.31
Cenarchaeales	$2.5^{E-02} \pm 7.5^{E-03}$	$2.0^{E-02} \pm 5.7^{E-03}$	$3.1^{E-02} \pm 5.4^{E-03}$	$2.6^{E-02} \pm 3.3^{E-03}$	122.24	130.93	107.11
Thermoplasmatales	$2.2^{E-02} \pm 2.5^{E-03}$	$2.1^{E-02} \pm 3.6^{E-03}$	$2.0^{E-02} \pm 6.7^{E-04}$	$2.0^{E-02} \pm 1.1^{E-03}$	92.04	96.13	104.44
Desulfurococcales	$3.3^{E-02} \pm 1.2^{E-03}$	$3.3^{E-02} \pm 1.8^{E-03}$	$3.3^{E-02} \pm 1.3^{E-03}$	$3.3^{E-02} \pm 7.4^{E-04}$	99.69	100.68	100.99
Methanomicrobiales	$1.4^{E-01} \pm 1.2^{E-02}$	$1.4^{E-01} \pm 3.9^{E-03}$	$1.4^{E-01} \pm 1.4^{E-03}$	$1.4^{E-01} \pm 6.2^{E-03}$	98.55	99.42	100.88
Halobacteriales	1.4^{E-01}	1.5^{E-01}	1.4^{E-01}	1.4^{E-01}	96.29	97.07	100.82

Order	Start Short Flood (x̄)	Start Long Flood (x̄)	End Short Flood (x̄)	End Long Flood (x̄)	Short flood fold change (%)	Long flood fold change (%)	Diff. (Long vs. Short flood) (%)
	± 9.1E-03	± 3.4E-03	± 3.2E-03	± 4.5E-03			
Methanosarcinales	2.1E-01 ± 6.2E-03	2.1E-01 ± 5.9E-03	2.0E-01 ± 4.5E-03	2.1E-01 ± 4.5E-03	97.73	98.36	100.65
Methanobacteriales	3.9E-02 ± 3.4E-03	4.2E-02 ± 2.1E-03	3.9E-02 ± 1.7E-03	4.1E-02 ± 9.2E-04	98.09	98.47	100.38
Methanococcales	4.0E-02 ± 2.0E-03	4.2E-02 ± 2.6E-03	3.7E-02 ± 3.1E-03	3.8E-02 ± 2.5E-03	91.76	91.76	99.99
unclassified (Korarchaeota)	1.1E-02 ± 5.3E-04	1.1E-02 ± 4.4E-04	1.1E-02 ± 1.0E-03	1.1E-02 ± 5.8E-04	100.62	100.10	99.48
Methanopyrales	9.6E-03 ± 3.3E-04	9.8E-03 ± 6.6E-04	9.9E-03 ± 5.6E-04	1.0E-02 ± 3.7E-04	102.37	101.82	99.47
Methanocellales	2.4E-02 ± 1.7E-03	2.4E-02 ± 4.0E-04	2.4E-02 ± 5.6E-04	2.5E-02 ± 2.5E-04	101.35	100.64	99.30
Thermococcales	6.3E-02 ± 1.5E-03	6.4E-02 ± 3.3E-03	6.2E-02 ± 5.1E-03	6.2E-02 ± 2.4E-03	97.52	96.53	98.99
Sulfolobales	4.9E-02 ± 4.8E-03	4.8E-02 ± 3.5E-04	5.0E-02 ± 1.5E-03	4.8E-02 ± 1.7E-03	102.77	101.18	98.45
Acidilobales	3.1E-03 ± 3.4E-04	2.8E-03 ± 2.3E-04	3.1E-03 ± 9.2E-05	2.8E-03 ± 1.4E-04	101.11	99.52	98.43
unclassified	5.0E-02	5.1E-02	5.0E-02	5.0E-02	98.86	97.21	98.33

Order	Start Short Flood ($\bar{x}$)	Start Long Flood ($\bar{x}$)	End Short Flood ($\bar{x}$)	End Long Flood ($\bar{x}$)	Short flood fold change (%)	Long flood fold change (%)	Diff. (Long vs. Short flood) (%)
(Euryarchaeota)	± 1.3E-03	± 2.6E-03	± 1.5E-03	± 7.4E-04			
Thermoproteales	4.6E-02 ± 2.0E-03	4.6E-02 ± 1.8E-03	4.6E-02 ± 1.0E-03	4.5E-02 ± 1.8E-03	99.48	97.55	98.06
Archaeoglobales	3.9E-02 ± 1.3E-03	4.1E-02 ± 1.5E-03	4.0E-02 ± 9.9E-04	3.9E-02 ± 1.0E-03	101.62	94.01	92.52
unclassified (Nanoarchaeota)	4.1E-04 ± 1.6E-04	5.2E-04 ± 8.6E-05	4.8E-04 ± 1.1E-04	4.8E-04 ± 8.0E-05	117.53	94.09	80.06

List of abbreviations

API	Application programming interface
BLAST	Basic Local Alignment Search Tool
CCD	Charge-coupled device
CH_4	Methane
CO_2	Carbon dioxide
CO_2e	Carbon dioxide equivalent
ePCR	Emulsion Polymerase chain reaction
DDBJ	DNA DataBank of Japan
ddNTP	Dideoxynucleotide triphosphate
DGGE	Denaturing gradient gel electrophoresis
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleotide triphosphate
EBI	European Bioinformatics Institute
EMBL	European Molecular Biology Laboratory
GSC	Genomics Standards Consortium
GWC	Gravimetric water content
GWP	Global warming potential
INSDC	International Nucleotide Sequence Database Collaboration
M5NR	M5 non-redundant protein database
MG-RAST	Metagenomics – Rapid Annotations using Sub-systems Technology
NeSSM	Next-Generation Sequencing Simulator for Metagenomics

OTU	Operational taxonomic unit
PAR	Photosynthetically active radiation
PCoA	Principal coordinates analysis
PCR	Polymerase chain reaction
PEAR	Paired-end read merger
PLFA	Phospholipid fatty acid
PPi	Pyrophosphate
QIIME	Quantitative Insight Into Microbial Ecology
RDP	Ribosomal Database Project
RNA	Ribonucleic acid
Simmet	Simulated metagenome
ssDNA	Single-strand DNA
T-RFLP	Terminal-restriction fragment length polymorphism
WGS	Whole genome shotgun sequencing

References

454 Life Sciences, a Roche Company (2012). Products - Technology: 454 Life Sciences, a Roche Company. *Available at http://454.com/products/technology.asp*, accessed 12/03/2012.

Achtman, M., and Wagner, M. (2008). Microbial diversity and the genetic nature of microbial species. Nat Rev Micro *6*, 431–440.

Allen, E.E., and Banfield, J.F. (2005). Community genomics in microbial ecology and evolution. Nature Reviews Microbiology *3*, 489–498.

Anderson, R.E., Sogin, M.L., and Baross, J.A. (2015). Biogeography and ecology of the rare and abundant microbial lineages in deep-sea hydrothermal vents. FEMS Microbiology Ecology *91*, 1–11.

Ansorge, W.J. (2009). Next-generation DNA sequencing techniques. New Biotechnology *25*, 195–203.

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D.R., Fernandes, G.R., Tap, J., Bruls, T., Batto, J.-M., *et al.* (2011). Enterotypes of the human gut microbiome. Nature *473*, 174–180.

Aselmann, I., and Crutzen, P.J. (1989). Global distribution of natural freshwater wetlands and rice paddies, their net primary productivity, seasonality and possible methane emissions. Journal of Atmospheric Chemistry *8*, 307–358.

Baldwin, D. s., and Mitchell, A. m. (2000). The effects of drying and re-flooding on the sediment and soil nutrient dynamics of lowland river–floodplain systems: a synthesis. Regul. Rivers: Res. Mgmt. *16*, 457–467.

Bapteste, E., and Boucher, Y. (2009). Epistemological Impacts of Horizontal Gene Transfer on Classification in Microbiology. In Horizontal Gene Transfer, D.M.B. Gogarten, D.J.P. Gogarten, and D.L.C. Olendzenski, eds. (New York City, USA: Humana Press), pp. 55–72.

Barton, L. (1995). Biotechnology Handbooks: Sulfate-Reducing Bacteria. (New York, USA: Plenum Press).

Berg, I.A., Kockelkorn, D., Ramos-Vera, W.H., Say, R.F., Zarzycki, J., Hügler, M., Alber, B.E., and Fuchs, G. (2010). Autotrophic carbon fixation in archaea. Nat Rev Micro *8*, 447–460.

Bernacchi, C.J., Leakey, A.D.B., Heady, L.E., Morgan, P.B., Dohleman, F.G., McGrath, J.M., Gillespie, K.M., Wittig, V.E., Rogers, A., Long, S.P., *et al.* (2006). Hourly and seasonal variation in photosynthesis and stomatal conductance of soybean grown at future CO(2) and ozone concentrations for 3 years under fully open-air field conditions. Plant Cell Environ. *29*, 2077–2090.

Berney, M., Greening, C., Conrad, R., Jacobs, W.R., and Cook, G.M. (2014). An obligately aerobic soil bacterium activates fermentative hydrogen production to survive reductive stress during hypoxia. Proc. Natl. Acad. Sci. U.S.A. *111*, 11479–11484.

Berry, J., and Bjorkman, O. (1980). Photosynthetic Response and Adaptation to Temperature in Higher Plants. Annual Review of Plant Physiology *31*, 491–543.

Blenkinsop, S., and Fowler, H.J. (2007). Changes in drought frequency, severity and duration for the British Isles projected by the PRUDENCE regional climate models. Journal of Hydrology *342*, 50–71.

Bodelier, P.L., Bar-Gilissen, M.-J., Meima-Franke, M., and Hordijk, K. (2012). Structural and functional response of methane-consuming microbial communities to different flooding regimes in riparian soils. Ecol Evol *2*, 106–127.

Branton, D., Deamer, D.W., Marziali, A., Bayley, H., Benner, S.A., Butler, T., Di Ventra, M., Garaj, S., Hibbs, A., Huang, X., *et al.* (2008). The potential and challenges of nanopore sequencing. Nat Biotech *26*, 1146–1153.

Bren, A., and Eisenbach, M. (2000). How Signals Are Heard during Bacterial Chemotaxis: Protein-Protein Interactions in Sensory Signal Propagation. J. Bacteriol. *182*, 6865–6873.

Bryant, D.A., and Frigaard, N.-U. (2006). Prokaryotic photosynthesis and phototrophy illuminated. Trends in Microbiology *14*, 488–496.

Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat Meth *12*, 59–60.

Bunce, J.A. (1984). Effects of Humidity on Photosynthesis. J. Exp. Bot. *35*, 1245–1251.

Cai, P., Huang, Q., Zhang, X., and Chen, H. (2006). Adsorption of DNA on clay minerals and various colloidal particles from an Alfisol. Soil Biology and Biochemistry *38*, 471–476.

Cai, Z., Xing, G., Yan, X., Xu, H., Tsuruta, H., Yagi, K., and Minami, K. (1997). Methane and nitrous oxide emissions from rice paddy fields as affected by nitrogen fertilisers and water management. Plant and Soil *196*, 7–14.

Cain, M.L., Subler, S., Evans, J.P., and Fortin, M.-J. (1999). Sampling spatial and temporal variation in soil nitrogen availability. Oecologia *118*, 397–404.

Caldwell, M.M. (1981). Plant Response to Solar Ultraviolet Radiation. In Physiological Plant Ecology I, P.D.O.L. Lange, P.P.S. Nobel, P.C.B. Osmond, and P.D.H. Ziegler, eds. (Springer Berlin Heidelberg), pp. 169–197.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., *et al.* (2010). QIIME allows analysis of high-throughput community sequencing data. Nature Methods *7*, 335.

Carr, R., and Borenstein, E. (2014). Comparative Analysis of Functional Metagenomic Annotation and the Mappability of Short Reads. PLoS ONE *9*, e105776.

Carvalhais, L.C., Dennis, P.G., Tyson, G.W., and Schenk, P.M. (2012). Application of metatranscriptomics to soil environments. Journal of Microbiological Methods *91*, 246–251.

Chun, J., Lee, J.-H., Jung, Y., Kim, M., Kim, S., Kim, B.K., and Lim, Y.-W. (2007). EzTaxon: a web-based tool for the identification of prokaryotes based on 16S ribosomal RNA gene sequences. Int J Syst Evol Microbiol *57*, 2259–2261.

Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., *et al.* (2009). Biopython: freely available Python tools for computational molecular biology and. Bioinformatics *25*, 1422–1423.

Collins, M., Knutti, R., Arblaster, J., Dufresne, J.-L., Fichefet, T., Friedlingstein, P., Gao, X., Gutowski Jr., W.J., Johns, T., Krinner, G., *et al.* (2013). Long-term Climate Change: Projections, Commitments and Irreversibility. In Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, T. Stocker, D. Qin, G.-K. Plattner, M.

Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P.M. Midgley, eds. (Cambridge University Press),.

Conrad, R. (1996). Soil microorganisms as controllers of atmospheric trace gases (H2, CO, CH4, OCS, N2O, and NO). Microbiol Rev *60*, 609–640.

Conrad, R. (2007). Microbial Ecology of Methanogens and Methanotrophs. In Advances in Agronomy, Donald L. Sparks, ed. (Academic Press), pp. 1–63.

Conrad, R., and Rothfuss, F. (1991). Methane oxidation in the soil surface layer of a flooded rice field and the effect of ammonium. Biology and Fertility of Soils *12*, 28–32.

Conrad, R., Klose, M., Noll, M., Kemnitz, D., and Bodelier, P.L.E. (2008). Soil type links microbial colonization of rice roots to methane emission. Global Change Biology *14*, 657–669.

Courtois, S., Cappellano, C.M., Ball, M., Francou, F.-X., Normand, P., Helynck, G., Martinez, A., Kolvek, S.J., Hopke, J., Osburne, M.S., *et al.* (2003). Recombinant environmental libraries provide access to microbial diversity for drug discovery from natural products. Appl. Environ. Microbiol. *69*, 49–55.

Crecchio, C., and Stotzky, G. (1998). Binding of DNA on humic acids: effect on transformation of Bacillus subtilis and resistance to DNase. Soil Biology and Biochemistry *30*, 1061–1067.

Darch, S.E., West, S.A., Winzer, K., and Diggle, S.P. (2012). Density-dependent fitness benefits in quorum-sensing bacterial populations. Proc Natl Acad Sci U S A *109*, 8259–8263.

David, L.A., Maurice, C.F., Carmody, R.N., Gootenberg, D.B., Button, J.E., Wolfe, B.E., Ling, A.V., Devlin, A.S., Varma, Y., Fischbach, M.A., *et al.* (2014). Diet rapidly and reproducibly alters the human gut microbiome. Nature *505*, 559–563.

DeBruyn, J.M., Nixon, L.T., Fawaz, M.N., Johnson, A.M., and Radosevich, M. (2011). Global Biogeography and Quantitative Seasonal Dynamics of Gemmatimonadetes in Soil. Appl. Environ. Microbiol. *77*, 6295–6300.

DEFRA (2012). UK climate change risk assessment: Government report, *Department for Environment, Food & Rural Affairs* (London: Stationery Office).

De Groot, C.-J., and Van Wijck, C. (1993). The impact of desiccation of a freshwater marsh (Garcines Nord, Camargue, France) on sediment-water-vegetation interactions. Hydrobiologia *252*, 83–94.

Denef, K., Six, J., Bossuyt, H., Frey, S.D., Elliott, E.T., Merckx, R., and Paustian, K. (2001). Influence of dry–wet cycles on the interrelationship between aggregate, particulate organic matter, and microbial community dynamics. Soil Biology and Biochemistry *33*, 1599–1611.

Desai, N., Antonopoulos, D., Gilbert, J.A., Glass, E.M., and Meyer, F. (2012). From genomics to metagenomics. Current Opinion in Biotechnology *23*, 72–76.

Di Bella, J.M., Bao, Y., Gloor, G.B., Burton, J.P., and Reid, G. (2013). High throughput sequencing methods and analysis for microbiome research. Journal of Microbiological Methods *95*, 401–414.

Dong, D., Yan, A., Liu, H., Zhang, X., and Xu, Y. (2006). Removal of humic substances from soil DNA using aluminium sulfate. Journal of Microbiological Methods *66*, 217–222.

D R Geiger, and Servaites, and J.C. (1994). Diurnal Regulation of Photosynthetic Carbon Metabolism in C3 Plants. Annual Review of Plant Physiology and Plant Molecular Biology *45*, 235–256.

Duboisset, A., Gilot, C., Pashanasi, B., Lavelle, P., and Brussaard, L. (1999). Effects of Earthworms on Soil Structure and Physical Properties Eric Blanchart1, Alain Albrecht2, Julio Alegre3. Earthworm Management in Tropical Agroecosystems 149.

Ellis, H.R. (2011). Mechanism for sulfur acquisition by the alkanesulfonate monooxygenase system. Bioorganic Chemistry *39*, 178–184.

Ensign, S.A. (2006). Revisiting the glyoxylate cycle: alternate pathways for microbial acetate assimilation. Molecular Microbiology *61*, 274–276.

Eriksson, P.G., Svensson, J.M., and Carrer, G.M. (2003). Temporal changes and spatial variation of soil oxygen consumption, nitrification and denitrification rates in a tidal salt marsh of the Lagoon of Venice, Italy. Estuarine, Coastal and Shelf Science *58*, 861–871.

Ettema, C.H., and Wardle, D.A. (2002). Spatial soil ecology. Trends in Ecology & Evolution *17*, 177–183.

Ewing, B., and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. *8*, 186–194.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. *8*, 175–185.

Falkowski, P.G., Fenchel, T., and Delong, E.F. (2008). The microbial engines that drive Earth's biogeochemical cycles. Science *320*, 1034–1039.

Fang, C., and Moncrieff, J.B. (2005). The variation of soil microbial respiration with depth in relation to soil carbon composition. Plant and Soil *268*, 243–253.

Farley, R.A., and Fitter, A.H. (1999). Temporal and spatial variation in soil resources in a deciduous woodland. Journal of Ecology *87*, 688–696.

Feng, Y., Zhang, Y., Ying, C., Wang, D., and Du, C. (2015). Nanopore-based Fourth-generation DNA Sequencing Technology. Genomics, Proteomics & Bioinformatics *13*, 4–16.

Field, C.B., Behrenfeld, M.J., Randerson, J.T., and Falkowski, P. (1998). Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. Science *281*, 237–240.

Field, D., Garrity, G., Gray, T., Morrison, N., Selengut, J., Sterk, P., Tatusova, T., Thomson, N., Allen, M.J., Angiuoli, S.V., *et al.* (2008). The minimum information about a genome sequence (MIGS) specification. Nature Biotechnology *26*, 541.

Fournier, P.-E., Dumler, J.S., Greub, G., Zhang, J., Wu, Y., and Raoult, D. (2003). Gene sequence-based criteria for identification of new rickettsia isolates and description of Rickettsia heilongjiangensis sp. nov. J. Clin. Microbiol. *41*, 5456–5465.

Fredrickson, J.K., Li, S.W., Gaidamakova, E.K., Matrosova, V.Y., Zhai, M., Sulloway, H.M., Scholten, J.C., Brown, M.G., Balkwill, D.L., and Daly, M.J. (2008). Protein oxidation: key to bacterial desiccation resistance? ISME J *2*, 393–403.

Fuhrman, J.A. (2012). Metagenomics and its connection to microbial community organization. F1000 Biol Rep *4*.

Garcia-Etxebarria, K., Garcia-Garcerà, M., and Calafell, F. (2014). Consistency of metagenomic assignment programs in simulated and real data. BMC Bioinformatics *15*, 90.

Gardner, A.M., Gessner, C.R., and Gardner, P.R. (2003). Regulation of the Nitric Oxide Reduction Operon (norRVW) in Escherichia coli Role of NorR and ç54 in the nitric oxide stress response. J. Biol. Chem. *278*, 10081–10086.

Georga, I., Stenuit, B., and Agathos, S. (2010). Application of Metagenomics to Bioremediation. In Metagenomics: Theory, Methods and Applications, D. Marco, ed. (Horizon Scientific Press), pp. 119–133.

Gevers, D., Cohan, F.M., Lawrence, J.G., Spratt, B.G., Coenye, T., Feil, E.J., Stackebrandt, E., de Peer, Y.V., Vandamme, P., Thompson, F.L., *et al.* (2005). Re-evaluating prokaryotic species. Nat Rev Micro *3*, 733–739.

Gilbert, B., and Frenzel, P. (1998). Rice roots and CH4 oxidation: the activity of bacteria, their distribution and the microenvironment. Soil Biology and Biochemistry *30*, 1903–1916.

Glenn, T.C. (2011). Field guide to next-generation DNA sequencers. Molecular Ecology Resources *11*, 759–769.

Gogarten, J.P., and Townsend, J.P. (2005). Horizontal gene transfer, genome innovation and evolution. Nat Rev Micro *3*, 679–687.

Grable, A.R. (1966). Soil Aeration and Plant Growth. In Advances in Agronomy, (Elsevier), pp. 57–106.

Hajibabaei, M., Shokralla, S., Zhou, X., Singer, G.A.C., and Baird, D.J. (2011). Environmental Barcoding: A Next-Generation Sequencing Approach for Biomonitoring Applications Using River Benthos. PLoS ONE *6*, e17497.

Hallam, S.J., Konstantinidis, K.T., Putnam, N., Schleper, C., Watanabe, Y., Sugahara, J., Preston, C., Torre, J. de la, Richardson, P.M., and DeLong, E.F. (2006). Genomic analysis of the uncultivated marine crenarchaeote Cenarchaeum symbiosum. PNAS *103*, 18296–18301.

Han, C., Zhong, W., Shen, W., Cai, Z., and Liu, B. (2013). Transgenic Bt rice has adverse impacts on CH4 flux and rhizospheric methanogenic archaeal and methanotrophic bacterial communities. Plant and Soil *369*, 297–316.

Handelsman, J. (2004). Metagenomics: Application of Genomics to Uncultured Microorganisms. Microbiol. Mol. Biol. Rev. *68*, 669–685.

Handelsman, J., Rondon, M.R., Brady, S.F., Clardy, J., and Goodman, R.M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem. Biol. *5*, R245–R249.

Heller, G. (1941). A quantitative study of environmental factors involved in survival and death of bacteria in the desiccated state. Journal of Bacteriology *41*, 109.

Henckel, T., Jäckel, U., and Conrad, R. (2001). Vertical distribution of the methanotrophic community after drainage of rice field soil. FEMS Microbiology Ecology *34*, 279–291.

Henry, V.J., Bandrowski, A.E., Pepin, A.-S., Gonzalez, B.J., and Desfeux, A. (2014). OMICtools: an informative directory for multi-omic data analysis. Database *2014*, bau069–bau069.

Hoff, K.J. (2009). The effect of sequencing errors on metagenomic gene prediction. BMC Genomics *10*, 520.

Holling, C.S. (1973). Resilience and Stability of Ecological Systems. Annual Review of Ecology and Systematics *4*, 1–23.

Holzapfel-Pschorn, A., Conrad, R., and Seiler, W. (1985). Production, oxidation and emission of methane in rice paddies. FEMS Microbiology Letters *31*, 343–351.

Hong, S., Bunge, J., Leslin, C., Jeon, S., and Epstein, S.S. (2009). Polymerase chain reaction primers miss half of rRNA microbial diversity. The ISME Journal *3*, 1365–1373.

Houghton, J. (2001). The science of global warming. Interdisciplinary Science Reviews *26*, 247–257.

Hugenholtz, P. (2002). Exploring prokaryotic diversity in the genomic era. Genome Biology *3*, reviews0003.

Huson, D.H., and Xie, C. (2014). A poor man's BLASTX--high-throughput metagenomic protein database search using PAUDA. Bioinformatics *30*, 38–39.

Huson, D.H., Auch, A.F., Qi, J., and Schuster, S.C. (2007). MEGAN analysis of metagenomic data. Genome Res. *17*, 000–000.

Illumina (2014). Illumina - Sequencing Technology. *Available at http://www.illumina.com/technology/next-generation-sequencing/sequencing-technology.html*, accessed 12/04/2012.

Imhoff, J.F. (2005). "Enterobacteriales." In Bergey's Manual® of Systematic Bacteriology, D.J. Brenner, N.R. Krieg, J.T. Staley, G.M.G. Sc.D, D.R. Boone, P.D. Vos, M. Goodfellow, F.A. Rainey, and K.-H. Schleifer, eds. (Springer US), pp. 587–850.

Jackson, M.B., and Armstrong, W. (1999). Formation of Aerenchyma and the Processes of Plant Ventilation in Relation to Soil Flooding and Submergence. Plant Biology *1*, 274–287.

Janssen, P.H. (2006). Identifying the Dominant Soil Bacterial Taxa in Libraries of 16S rRNA and 16S rRNA Genes. Appl. Environ. Microbiol. *72*, 1719–1728.

Jia, B., Xuan, L., Cai, K., Hu, Z., Ma, L., and Wei, C. (2013). NeSSM: A Next-Generation Sequencing Simulator for Metagenomics. PLoS ONE *8*, e75448.

Jobbágy, E.G., and Jackson, R.B. (2004). The Uplift of Soil Nutrients by Plants: Biogeochemical Consequences Across Scales. Ecology *85*, 2380–2389.

Jones, M.R., Fowler, H.J., Kilsby, C.G., and Blenkinsop, S. (2013). An assessment of changes in seasonal and annual extreme rainfall in the UK between 1961 and 2009. International Journal of Climatology *33*, 1178–1194.

Joshi, N., and Fass, J. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software].

Kachanoski, R.G., Wesenbeeck, I.J.V., and Gregorich, E.G. (1988). Estimating Spatial Variations of Soil Water Content Using Noncontacting Electromagnetic Inductive Methods. Can. J. Soil. Sci. *68*, 715–722.

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucl. Acids Res. *28*, 27–30.

Kelly, C.A., Rudd, J.W.M., Bodaly, R.A., Roulet, N.P., St.Louis, V.L., Heyes, A., Moore, T.R., Schiff, S., Aravena, R., Scott, K.J., *et al.* (1997). Increases in Fluxes of Greenhouse Gases and Methyl Mercury following Flooding of an Experimental Reservoir. Environ. Sci. Technol. *31*, 1334–1344.

Kemnitz, D., Chin, K.-J., Bodelier, P., and Conrad, R. (2004). Community analysis of methanogenic archaea within a riparian flooding gradient. Environmental Microbiology *6*, 449–461.

Kennedy, R.A., Rumpho, M.E., and Fox, T.C. (1992). Anaerobic Metabolism in Plants. Plant Physiol. *100*, 1–6.

Kennedy, S., Callahan, H., and Carlson, M. (2013). Tips and Tricks for isolation of DNA & RNA from challenging samples (MO BIO Laboratories, Inc.).

Kirtman, B., Power, S.B., Adedoyin, A.J., Boer, G.J., Bojariu, R., Camilloni, I., Doblas-Reyes, F., Fiore, A.M., Kimoto, M., Meehl, G., *et al.* (2013). Near-term Climate Change: Projections and Predictability. In Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, T. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P.M. Midgley, eds. (Cambridge University Press).

Kisand, V., Valente, A., Lahm, A., Tanet, G., and Lettieri, T. (2012). Phylogenetic and Functional Metagenomic Profiling for Assessing Microbial Biodiversity in Environmental Monitoring. PLoS ONE *7*, e43630.

Kjeldsen, K.U., Joulian, C., and Ingvorsen, K. (2004). Oxygen Tolerance of Sulfate-Reducing Bacteria in Activated Sludge. Environ. Sci. Technol. *38*, 2038–2043.

Kleinen, T., and Petschel-Held, G. (2007). Integrated assessment of changes in flooding probabilities due to climate change. Climatic Change *81*, 283–312.

Knight, R., Jansson, J., Field, D., Fierer, N., Desai, N., Fuhrman, J.A., Hugenholtz, P., van der Lelie, D., Meyer, F., Stevens, R., *et al.* (2012). Unlocking the potential of metagenomics through replicated experimental design. Nat Biotech *30*, 513–520.

Kröber, M., Bekel, T., Diaz, N.N., Goesmann, A., Jaenicke, S., Krause, L., Miller, D., Runte, K.J., Viehöver, P., Pühler, A., *et al.* (2009). Phylogenetic characterization of a biogas

plant microbial community integrating clone library 16S-rDNA sequences and metagenome sequence data obtained by 454-pyrosequencing. J. Biotechnol. *142*, 38–49.

Kurnik, B., Hemming, D., and Hartley, A. (2012). Section 2.2: Key climate variables. In Climate Change, Impacts and Vulnerability in Europe 2012, (Office for Official Publications of the European Union), pp. 54–72.

Lamentowicz, Ł., Gąbka, M., Rusińska, A., Sobczyński, T., Owsianny, P.M., and Lamentowicz, M. (2011). Testate Amoeba (Arcellinida, Euglyphida) Ecology along a Poor-Rich Gradient in Fens of Western Poland. International Review of Hydrobiology *96*, 356–380.

Lane, D.J., Pace, B., Olsen, G.J., Stahl, D.A., Sogin, M.L., and Pace, N.R. (1985). Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. PNAS *82*, 6955–6959.

Lavelle, P., Pashanasi, B., Charpentier, F., Gilot, C., Rossi, J.-P., Derouard, L., André, J., Ponge, J.-F., and Bernier, N. (1998). Large-scale effects of earthworms on soil organic matter and nutrient dynamics. Earthworm Ecology *10*.

Lavigne, M.B., Foster, R.J., and Goodine, G. (2004). Seasonal and annual changes in soil respiration in relation to soil temperature, water potential and trenching. Tree Physiology *24*, 415–424.

Leach, J.E. (1979). Some effects of air temperature and humidity on crop and leaf photosynthesis, transpiration and resistance to gas transfer. Annals of Applied Biology *92*, 287–297.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics *25*, 1754–1760.

Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics btv033.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, 1000 Genome Project Data Processing (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Life Technologies (2012). How Does Semiconductor Sequencing Work? | Life Technologies. *Available at https://www.thermofisher.com/uk/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-technology.html*, accessed 04/12/2012.

Life Technologies (2014). Next Generation Sequencing. *Available at https://www.thermofisher.com/uk/en/home/life-science/sequencing/next-generation-sequencing.html*, accessed 17/02/2014.

Lindgreen, S., Adair, K.L., and Gardner, P.P. (2016). An evaluation of the accuracy and speed of metagenome analysis tools. Scientific Reports *6*, 19233.

Lipson, D.A., Haggerty, J.M., Srinivas, A., Raab, T.K., Sathe, S., and Dinsdale, E.A. (2013). Metagenomic Insights into Anaerobic Metabolism along an Arctic Peat Soil Profile. PLoS ONE *8*, e64659.

Liu, Q., Edwards, N.T., Post, W.M., Gu, L., Ledford, J., and Lenhart, S. (2006). Temperature-independent diel variation in soil respiration observed from a temperate deciduous forest. Global Change Biology *12*, 2136–2145.

Lloyd, J., and Taylor, J.A. (1994). On the Temperature Dependence of Soil Respiration. Functional Ecology *8*, 315.

Loman, N.J., and Watson, M. (2015). Successful test launch for nanopore sequencing. Nat Meth *12*, 303–304.

Luton, P.E., Wayne, J.M., Sharp, R.J., and Riley, P.W. (2002). The mcrA gene as an alternative to 16S rRNA in the phylogenetic analysis of methanogen populations in landfill. Microbiology *148*, 3521–3530.

Markowitz, V.M., Mavromatis, K., Ivanova, N.N., Chen, I.-M.A., Chu, K., and Kyrpides, N.C. (2009). IMG ER: a system for microbial genome annotation expert review and curation. Bioinformatics *25*, 2271–2278.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal *17*, pp. 10–12.

Mavromatis, K., Ivanova, N., Barry, K., Shapiro, H., Goltsman, E., McHardy, A.C., Rigoutsos, I., Salamov, A., Korzeniewski, F., Land, M., *et al.* (2007). Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. Nat Meth *4*, 495–500.

Mayer, P.H., and Conrad, R. (1990). Factors influencing the population of methanogenic bacteria and the initiation of methane production upon flooding of paddy soil. FEMS Microbiology Letters *73*, 103–111.

McDonald, I.R., Upton, M., Hall, G., Pickup, R.W., Edwards, C., Saunders, J.R., Ritchie, D.A., and Murrell, J.C. (1999). Molecular Ecological Analysis of Methanogens and Methanotrophs in Blanket Bog Peat. Microb Ecol *38*, 225–233.

Mende, D.R., Sunagawa, S., Zeller, G., and Bork, P. (2013). Accurate and universal delineation of prokaryotic species. Nat Meth *10*, 881–884.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E.M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., *et al.* (2008). The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics *9*, 386.

Miller, J.R., Koren, S., and Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. Genomics *95*, 315–327.

Min, S.-K., Zhang, X., Zwiers, F.W., and Hegerl, G.C. (2011). Human contribution to more-intense precipitation extremes. Nature *470*, 378–381.

Miyata, A., Leuning, R., Denmead, O.T., Kim, J., and Harazono, Y. (2000). Carbon dioxide and methane fluxes from an intermittently flooded paddy field. Agricultural and Forest Meteorology *102*, 287–303.

Mohanty, S.R., Kollah, B., Sharma, V.K., Singh, A.B., Singh, M., and Rao, A.S. (2013). Methane oxidation and methane driven redox process during sequential reduction of a flooded soil ecosystem. Ann Microbiol *64*, 65–74.

Mokili, J.L., Rohwer, F., and Dutilh, B.E. (2012). Metagenomics and future perspectives in virus discovery. Curr Opin Virol *2*, 63–77.

Morillas, L., Durán, J., Rodríguez, A., Roales, J., Gallardo, A., Lovett, G.M., and Groffman, P.M. (2015). Nitrogen supply modulates the effect of changes in drying–rewetting frequency on soil C and N cycling and greenhouse gas exchange. Glob Change Biol *21*, 3854–3863.

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C., and Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. Nucl. Acids Res. *35*, W182–W185.

Morse, D. (2010). Future Flooding volume 1: Future risks and their drivers | Our work | BIS. UK Government.

Murphy, J.M., Sexton, D.M.H., Jenkins, G.J., Boorman, P.M., Brown, C.C., Clark, R.T., Collins, M., Harris, G.R., Kendon, E.J., Betts, R.A., *et al.* (2009). Probabilistic projections of seasonal climate changes. In UK Climate Projections Science Report: Climate Change Projections, (Exeter: Met Office Hadley Centre), pp. 90–123.

Muyzer, G., and Stams, A.J.M. (2008). The ecology and biotechnology of sulphate-reducing bacteria. Nature Reviews Microbiology *6,* 441–454.

Myhre, G., Shindell, D., Bréon, F.-M., Collins, W.D., Fuglestvedt, J., Huang, J., Koch, D., Lamarque, J.-F., Lee, D., Mendoza, B., *et al.* (2013). Anthropogenic and Natural Radiative Forcing. In Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, T. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P.M. Midgley, eds. (Cambridge University Press).

Nakićenović, N., Davidson, O., Davis, G., Grübler, A., Kram, T., Lebre La Rovere, E., Metz, B., Morita, T., Pepper, W., Pitcher, H., *et al.* (2000). Emissions scenarios. a special report of IPCC Working Group III ([Geneva]: Intergovernmental Panel on Climate Change).

National Soil Resources Institute (NSRI) (2013). Academic Soils Site Report for location 392413E, 127330N, 1km x 1km (National Soil Resources Institute, Cranfield University).

Nayfach, S., and Pollard, K.S. (2014). Average genome size estimation enables accurate quantification of gene family abundance and sheds light on the functional ecology of the human microbiome. bioRxiv 009001.

NCBI (2012). National Center for Biotechnology Information. *Available at https://www.ncbi.nlm.nih.gov/,* accessed 07/12/2012.

Niedermeier, A., and Robinson, J.S. (2007). Hydrological controls on soil redox dynamics in a peat-based, restored wetland. Geoderma *137*, 318–326.

Noinaj, N., Guillier, M., Barnard, T.J., and Buchanan, S.K. (2010). TonB-dependent transporters: regulation, structure, and function. Annu Rev Microbiol *64*, 43–60.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucl. Acids Res. *27*, 29–34.

Osborn, T.J., Hulme, M., Jones, P.D., and Basnett, T.A. (2000). Observed trends in the daily intensity of United Kingdom precipitation. International Journal of Climatology *20*, 347–364.

Pall, P., Aina, T., Stone, D.A., Stott, P.A., Nozawa, T., Hilberts, A.G.J., Lohmann, D., and Allen, M.R. (2011). Anthropogenic greenhouse gas contribution to flood risk in England and Wales in autumn 2000. Nature *470*, 382–385.

Parkin, T.B., and Kaspar, T.C. (2003). Temperature Controls on Diurnal Carbon Dioxide Flux. Soil Science Society of America Journal *67*, 1763.

Patrick, W.H., and Jugsujinda, A. (1992). Sequential Reduction and Oxidation of Inorganic Nitrogen, Manganese, and Iron in Flooded Soil. Soil Science Society of America Journal *56*, 1071.

Patz, J.A., Campbell-Lendrum, D., Holloway, T., and Foley, J.A. (2005). Impact of regional climate change on human health. Nature *438*, 310–317.

Pereira, J.S., Tenhunen, J.D., Lange, O.L., Beyschlag, W., Meyer, A., and David, M.M. (1986). Seasonal and diurnal patterns in leaf gas exchange of Eucalyptusglobulus trees growing in Portugal. Can. J. For. Res. *16*, 177–184.

Peterson, F.S., and Lajtha, K.J. (2013). Linking aboveground net primary productivity to soil carbon and dissolved organic carbon in complex terrain: LINKING ANPP, SOC, AND DOC. Journal of Geophysical Research: Biogeosciences *118*, 1225–1236.

Pignatelli, M., and Moya, A. (2011). Evaluating the Fidelity of De Novo Short Read Metagenomic Assembly Using Simulated Data. PLoS ONE *6*, e19984.

Poff, N.L. (2002). Ecological response to and management of increased flooding caused by climate change. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences *360*, 1497–1510.

Ponnamperuma, F.N. (1984). Effects of flooding on soils. In Flooding and Plant Growth, G. Meurant, *eds.* (Academic Press).

Potts, M. (1994). Desiccation tolerance. Microbiological Reviews *58*, 755–805.

Powell, B.S., Court, D.L., Inada, T., Nakamura, Y., Michotey, V., Cui, X., Reizer, A., Saier, M.H., and Reizer, J. (1995). Novel Proteins of the Phosphotransferase System Encoded within the rpoN Operon of Escherichia coli. J. Biol. Chem. *270*, 4822–4839.

Powell, J.M., Ikpe, F.N., and Somda, Z.C. (1999). Crop yield and the fate of nitrogen and phosphorus following application of plant material and feces to soil. Nutrient Cycling in Agroecosystems *54*, 215–226.

Prosser, J.I. (2010). Replicate or lie: The need for replication. Environmental Microbiology *12*, 1806–1810.

Quince, C., Curtis, T.P., and Sloan, W.T. (2008). The rational exploration of microbial diversity. ISME J *2*, 997–1006.

Raghoebarsing, A.A., Pol, A., van de Pas-Schoonen, K.T., Smolders, A.J.P., Ettwig, K.F., Rijpstra, W.I.C., Schouten, S., Damsté, J.S.S., Op den Camp, H.J.M., Jetten, M.S.M., *et al.* (2006). A microbial consortium couples anaerobic methane oxidation to denitrification. Nature *440*, 918–921.

Raich, J.W., and Schlesinger, W.H. (1992). The global carbon dioxide flux in soil respiration and its relationship to vegetation and climate. Tellus B *44*, 81–99.

Randle-Boggis, R.J., Helgason, T., Sapp, M., and Ashton, P.D. (2016). Evaluating techniques for metagenome annotation using simulated sequence data. FEMS Microbiol Ecol *92*.

Ransom-Jones, E., Jones, D.L., McCarthy, A.J., and McDonald, J.E. (2012). The Fibrobacteres: an Important Phylum of Cellulose-Degrading Bacteria. Microbial Ecology *63*, 267–281.

Ratering, S., and Conrad, R. (1998). Effects of short-term drainage and aeration on the production of methane in submerged rice soil. Global Change Biology *4*, 397–407.

Rawson, H.M., Begg, J.E., and Woodward, R.G. (1977). The effect of atmospheric humidity on photosynthesis, transpiration and water use efficiency of leaves of several plant species. Planta *134*, 5–10.

Reddy, K.R., and Patrick, W.H. (1975). Effect of alternate aerobic and anaerobic conditions on redox potential, organic matter decomposition and nitrogen loss in a flooded soil. Soil Biology and Biochemistry *7*, 87–94.

Rengel, Z. (1999). Mineral Nutrition of Crops: Fundamental Mechanisms and Implications (CRC Press).

Richter, M., and Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. PNAS *106*, 19126–19131.

Riesenfeld, C.S., Schloss, P.D., and Handelsman, J. (2004). METAGENOMICS: Genomic Analysis of Microbial Communities. Annual Review of Genetics *38*, 525–552.

Rinnan, R., Michelsen, A., Bååth, E., and Jonasson, S. (2007). Fifteen years of climate change manipulations alter soil microbial communities in a subarctic heath ecosystem. Global Change Biology *13*, 28–39.

Roberson, E.B., Chenu, C., and Firestone, M.K. (1993). Microstructural changes in bacterial exopolysaccharides during desiccation. Soil Biology and Biochemistry *25*, 1299–1301.

Rosenberg, D.M., and Resh, V.H. (1993). Freshwater biomonitoring and benthic macroinvertebrates. (Chapman & Hall).

Roslev, P., and King, G.M. (1994). Survival and recovery of methanotrophic bacteria starved under oxic and anoxic conditions. Applied and Environmental Microbiology *60*, 2602–2608.

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

Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., Leamon, J.H., Johnson, K., Milgrew, M.J., Edwards, M., *et al.* (2011). An integrated semiconductor device enabling non-optical genome sequencing. Nature *475*, 348–352.

Saint-Girons, I., and Cole, S.T. (1999). Bacterial Genomes—All Shapes and Sizes. In Organization of the Prokaryotic Genome, R.L. Charlebois, ed. (American Society of Microbiology), pp. 35–62.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. PNAS *74*, 5463–5467.

Schmidt, I.K., Ruess, L., Bååth, E., Michelsen, A., Ekelund, F., and Jonasson, S. (2000). Long-term manipulation of the microbes and microfauna of two subarctic heaths by addition of fungicide, bactericide, carbon and fertilizer. Soil Biology and Biochemistry *32*, 707–720.

Schneiker, S., Perlova, O., Kaiser, O., Gerth, K., Alici, A., Altmeyer, M.O., Bartels, D., Bekel, T., Beyer, S., Bode, E., *et al.* (2007). Complete genome sequence of the myxobacterium Sorangium cellulosum. Nat Biotech *25*, 1281–1289.

Scholz, M.B., Lo, C.-C., and Chain, P.S. (2012). Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. Current Opinion in Biotechnology *23*, 9–15.

Schütz, H., Seiler, W., and Conrad, R. (1989). Processes involved in formation and emission of methane in rice paddies. Biogeochemistry *7*, 33–53.

Shah, N., Tang, H., Doak, T.G., and Ye, Y. (2011). Comparing bacterial communities inferred from 16S rRNA gene sequencing and shotgun metagenomics. Pac Symp Biocomput 165–176.

Shakya, M., Quince, C., Campbell, J.H., Yang, Z.K., Schadt, C.W., and Podar, M. (2013). Comparative metagenomic and rRNA microbial diversity characterization using Archaeal and Bacterial synthetic communities. Environ Microbiol *15*, 1882–1899.

Sigren, L.K., Lewis, S.T., Fisher, F.M., and Sass, R.L. (1997). Effects of field drainage on soil parameters related to methane production and emission from rice paddies. Global Biogeochemical Cycles *11*, 151–162.

Smith, A.M., and Stitt, M. (2007). Coordination of carbon supply and plant growth. Plant, Cell & Environment *30*, 1126–1149.

Stackebrandt, E., and Goebel, B.M. (1994). Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. Int J Syst Bacteriol *44*, 846–849.

Stams, A.J.M., and Plugge, P. (2010). The microbiology of methanogenesis. In Methane and Climate Change, (Earthscan), pp. 14–26.

Stein, J.L., Marsh, T.L., Wu, K.Y., Shizuya, H., and DeLong, E.F. (1996). Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. J Bacteriol *178*, 591–599.

Streit, W.R., and Schmitz, R.A. (2004). Metagenomics--the key to the uncultured microbes. Curr. Opin. Microbiol. *7*, 492–498.

Takai, Y. (1970). The mechanism of methane fermentation in flooded paddy soil. Soil Science and Plant Nutrition *16*, 238–244.

Tang, J., Baldocchi, D.D., and Xu, L. (2005). Tree photosynthesis modulates soil respiration on a diurnal time scale. Global Change Biology *11*, 1298–1304.

Tatusov, R.L., Koonin, E.V., and Lipman, D.J. (1997). A Genomic Perspective on Protein Families. Science *278*, 631–637.

Tebbe, C.C., and Vahjen, W. (1993). Interference of humic acids and DNA extracted directly from soil in detection and transformation of recombinant DNA from bacteria and a yeast. Appl Environ Microbiol *59*, 2657–2665.

Teeling, H., and Glöckner, F.O. (2012). Current opportunities and challenges in microbial metagenome analysis—a bioinformatic perspective. Brief Bioinform *13*, 728–742.

Thomas, C.A. (1971). The Genetic Organization of Chromosomes. Annual Review of Genetics *5*, 237–256.

Thomas, T., Gilbert, J., and Meyer, F. (2012). Metagenomics - a guide from sampling to data analysis. Microbial Informatics and Experimentation *2*, 3.

Tissue, D.T., Thomas, R.B., and Strain, B.R. (1993). Long-term effects of elevated CO2 and nutrients on photosynthesis and rubisco in loblolly pine seedlings. Plant, Cell & Environment *16*, 859–865.

Trenberth, K.E. (1999). Conceptual Framework for Changes of Extremes of the Hydrological Cycle with Climate Change. Climatic Change *42*, 327–339.

Trenberth, K.E., Dai, A., Rasmussen, R.M., and Parsons, D.B. (2003). The Changing Character of Precipitation. Bulletin of the American Meteorological Society *84*, 1205–1217.

Tringe, S.G., and Rubin, E.M. (2005). Metagenomics: DNA sequencing of environmental samples. Nature Reviews Genetics *6*, 805–814.

Tringe, S.G., Mering, C. von, Kobayashi, A., Salamov, A.A., Chen, K., Chang, H.W., Podar, M., Short, J.M., Mathur, E.J., Detter, J.C., *et al.* (2005). Comparative Metagenomics of Microbial Communities. Science *308*, 554–557.

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

Unger, I.M., Kennedy, A.C., and Muzika, R.M. (2009). Flooding effects on soil microbial communities. Applied Soil Ecology *42*, 1–8.

Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W., *et al.* (2004). Environmental Genome Shotgun Sequencing of the Sargasso Sea. Science *304*, 66–74.

Verhoeven, J.T.A., Laanbroek, H.J., Rains, M.C., and Whigham, D.F. (2014). Effects of increased summer flooding on nitrogen dynamics in impounded mangroves. J. Environ. Manage. *139*, 217–226.

Větrovský, T., and Baldrian, P. (2013). The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses. PLoS ONE *8*, e57923.

Vignais, P.M., and Billoud, B. (2007). Occurrence, Classification, and Biological Function of Hydrogenases: An Overview. Chem. Rev. *107*, 4206–4272.

Waldrop, M.P., and Firestone, M.K. (2006). Response of Microbial Community Composition and Function to Soil Climate Change. Microb Ecol *52*, 716–724.

Wang, W.., Dalal, R.., Moody, P.., and Smith, C.. (2003). Relationships of soil respiration to microbial biomass, substrate availability and clay content. Soil Biology and Biochemistry *35*, 273–284.

Wang, Z.P., DeLaune, R.D., Patrick, W.H., and Masscheleyn, P.H. (1993). Soil Redox and pH Effects on Methane Production in a Flooded Rice Soil. Soil Science Society of America Journal *57*, 382.

Wei, H., Chen, X., Xiao, G., Guenet, B., Vicca, S., and Shen, W. (2015). Are variations in heterotrophic soil respiration related to changes in substrate availability and microbial biomass carbon in the subtropical forests? Scientific Reports *5*, 18370.

Wetterstrand, K. (2016). DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). *Available at https://www.genome.gov/sequencingcosts/*, accessed 10/05/2016.

Wilke, A., Harrison, T., Wilkening, J., Field, D., Glass, E.M., Kyrpides, N., Mavrommatis, K., and Meyer, F. (2012). The M5nr: a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools. BMC Bioinformatics *13*, 141.

Wilke, A., Glass, E.M., Bischof, J., Braithwaite, D., DSouza, M., Gerlach, W., Harrison, T., Keegan, K., Matthews, H., Paczian, T., *et al.* (2013). MG-RAST Technical report and manual v3.3.6 r1. (Argonne National Laboratory: University of Chicago).

Woese, C.R., and Fox, G.E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. PNAS *74*, 5088–5090.

Wood, D.E., and Salzberg, S.L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology *15*, R46.

Wu, S., Zhu, Z., Fu, L., Niu, B., and Li, W. (2011). WebMGA: a customizable web server for fast metagenomic sequence analysis. BMC Genomics *12*, 444.

Xie, W., Wang, F., Guo, L., Chen, Z., Sievert, S.M., Meng, J., Huang, G., Li, Y., Yan, Q., Wu, S., *et al.* (2011). Comparative metagenomics of microbial communities inhabiting deep-sea hydrothermal vent chimneys with contrasting chemistries. ISME J *5*, 414–426.

Yagi, K., Tsuruta, H., Kanda, K., and Minami, K. (1996). Effect of water management on methane emission from a Japanese rice paddy field: Automated methane monitoring. Global Biogeochemical Cycles *10*, 255–267.

Yankson, K.K., and Steck, T.R. (2009). Strategy for Extracting DNA from Clay Soil and Detecting a Specific Target Sequence via Selective Enrichment and Real-Time (Quantitative) PCR Amplification. Appl Environ Microbiol *75*, 6017–6021.

Zhang, H. (2003). Gemmatimonas aurantiaca gen. nov., sp. nov., a Gram-negative, aerobic, polyphosphate-accumulating micro-organism, the first cultured representative of the new bacterial phylum Gemmatimonadetes phyl. nov. International Journal Of Systematic And Evolutionary Microbiology *53*, 1155–1163.

Zhang, H.-L., Bai, X.-L., Xue, J.-F., Chen, Z.-D., Tang, H.-M., and Chen, F. (2013). Emissions of CH4 and N2O under Different Tillage Systems from Double-Cropped Paddy Fields in Southern China. PLoS ONE *8*, e65277.

Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics *30*, 614–620.

Zhao, Y., Tang, H., and Ye, Y. (2012). RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. Bioinformatics *28*, 125–126.

Zhou, J., Xia, B., Treves, D.S., Wu, L.-Y., Marsh, T.L., O'Neill, R.V., Palumbo, A.V., and Tiedje, J.M. (2002). Spatial and Resource Factors Influencing High Microbial Diversity in Soil. Appl. Environ. Microbiol. *68*, 326–334.

Zumft, W.G. (1997). Cell biology and molecular basis of denitrification. Microbiol. Mol. Biol. Rev. *61*, 533–616.