

**An investigation into the prokaryotic diversity of UK forest
soils using high throughput sequencing**

Submitted by Theresa Mary Hudson, to the University of Exeter as a thesis for the degree of
Doctor of Philosophy in September 2010.

This thesis is available for Library use on the understanding that it is copyright material and that
no quotation from the thesis may be published without proper acknowledgement.

I certify that all material in this thesis which is not my own work has been identified and that no
material has previously been submitted and approved for the award of a degree by this or any
other University.

(Signature).....

Abstract

The development of high throughput sequencing techniques has allowed for greater investigations into environmental microbial communities in this thesis. Using the Illumina Solexa sequencing technology we investigated prokaryote communities in replicate soil samples from four soil types: both the organic and mineral layers of two UK forest sites, an oak stand and a Corsican pine stand. Prokaryote diversity was assessed using the 16S SSU rRNA encoding genes from DNA and rRNA (reverse transcribed to cDNA) extracted simultaneously from the same sample. From a total of > 4.5 million sequence reads, only 0.6 % of the reads overlapped between RNA and DNA samples. Marked differences were seen between the distribution of Phyla from RNA and DNA. The community structure showed a small number of Operational Taxonomic Units (OTUs) having high relative abundance and many OTUs being detected in low or single numbers, representing a 'rare biosphere' community composition. Taxonomy of RNA was further described at phylum and family level. Only 30 % of OTUs could be assigned taxonomy at phylum level and, of those, only 46 % at family or genus level (bootstrap support >70). At phylum level, the RNA samples were homogeneous across all soil types, dominated by β/γ -Proteobacteria, α -Proteobacteria and Acidobacteria. 7 phyla were detected in all samples and rarely detected phyla showed high diversity between samples. At family level the samples were much more heterogeneous, different phyla showed different family taxonomic distribution patterns, with Firmicutes and Actinobacteria the most diverse. We used principal component analysis comparing 27 environmental variables against the Shannon-Weaver diversity index of each soil sample. Each soil type was seen to have its own unique fingerprint of diversity. pH was observed as the most influential parameter to diversity, and C:N ratio, zirconium and zinc levels also considered important. 24 of the 27 parameters explained the majority of the variation in the diversity of the soil samples, concluding that prokaryote diversity in these soils is determined by complex interactions between a wide range of environmental conditions. Together this demonstrates that the prokaryote soil community are extremely diverse and dynamic.

List of contents

Chapter 1	General Introduction	14
1.1	There are microbes in all domains of life	14
1.1.1	The discovery of single celled organisms	14
1.1.2	Whittaker's five kingdom classification	15
1.1.3	The three domains of life	16
1.1.3.1	The inaccuracy of the prokaryotic and eukaryotic division	16
1.1.3.2	The use of phylogenetics to identify the three domains	17
1.1.4	Describing Bacteria	19
1.1.4.1	Cytological differences between Bacteria and Eukaryotes	19
1.1.4.2	Comparing Bacterial species	20
1.1.4.3	Bacterial genome size	21
1.1.4.4	Bacterial genome structure	22
1.1.4.5	Defining a Bacterial 'species'	24
1.1.4.6	Describing the taxonomy of Bacteria	26
1.1.5	Describing Archaea	28
1.1.5.1	Comparing Archaea with Bacteria and Eukaryotes	28
1.1.5.2	Taxonomy of Archaea	31
1.2	Measuring microbial diversity	32
1.2.1	Cultivation methods	32
1.2.2	Molecular fingerprinting techniques	33
1.2.3	Sequencing technologies	34

1.2.3.1	Using a sequence approach to identify operational taxonomic units and assess diversity	34
1.2.3.1.1	Use of the small subunit rRNA encoding gene to investigate diversity	34
1.2.3.1.2	Comparing variation across the 16S SSU rRNA gene	36
1.2.3.1.3	Limitations of investigating the SSU rRNA encoding gene	36
1.2.3.1.4	Controlling for environmental DNA survey bias	37
1.2.3.2	Creation of environmental clone libraries	38
1.2.3.3	Using a metagenomic approach to assess diversity	39
1.2.3.4	Next generation sequencing technology	40
1.2.3.4.1	Overview of next generation sequencing technology	40
1.2.3.4.2	454 genome sequencer	41
1.2.3.4.3	ABI SOLiD genome analyzer	42
1.2.3.4.4	Illumina (Solexa) genome analyzer	42
1.2.3.4.5	Limitations of next generation sequencing technologies	43
1.3	Global microbial diversity	44
1.3.1	How vast is diversity?	44
1.3.2	Biogeography	46
1.3.2.1	Patterns of biogeography	46
1.3.2.2	Processes determining biogeography	47
1.3.3	The ubiquitous distribution hypothesis	49
1.3.4	The endemic model	50
1.3.5	Balancing the UDH and endemic models	52
1.3.6	The rare biosphere hypothesis	53

1.3.7	Transitions in the rare biosphere	55
1.3.8	Rare biosphere and next generation sequencing techniques	56
1.3.8.1	The development of the 'rare biosphere' hypothesis using next generation sequencing	56
1.3.8.2	Limitations of using next generation sequencing to investigate the 'rare biosphere'	58
1.4	Soil Microbial Diversity	60
1.4.1	Describing the soil environment	60
1.4.2	Prokaryotic diversity in the soil	61
1.4.3	Variation between soil clone libraries	62
1.4.4	Environmental influences on microbial community composition in soil	65
1.5	Present Study	69
Chapter 2	Materials and Methods/ Method Development/ Preliminary Data	71
2.1	Aims	71
2.2	Description of forest sample sites	71
2.2.1	Vegetation composition and soil types of sample sites	71
2.2.2	Soil sampling	72
2.2.3	Physical and chemical properties of both sample sites	76
2.2.3.1	Measuring pH	76
2.2.3.2	Measuring soil moisture content	76
2.2.4	Chemical properties of both sample sites	76
2.2.4.1	Total carbon and nitrogen	76
2.2.4.2	Elemental analysis	77

2.3	Molecular analysis of both sample sites	77
2.3.1	Extraction of nucleic acids from forest soils	77
2.3.1.1	Extraction of RNA from forest soils	77
2.3.1.2	Reverse transcription of the RNA	79
2.3.1.3	Extraction of DNA from the same soil sample used for RNA extraction	80
2.3.2	PCR amplification	81
2.3.3	Using Denaturing Gradient Gel Electrophoresis (DGGE)	84
2.3.3.1	General denaturing gradient gel electrophoresis method	84
2.3.3.2	Denaturing gradient gel electrophoresis targeting 16S SSU rDNA	84
2.3.4	Clone libraries to analyse complexity across taxonomic groups	86
2.3.4.1	General cloning method	86
2.3.4.2	Construction of the clone library	87
2.3.4.3	Prokaryotic clone library	87
2.3.5	Using Illumina Solexa next generation sequencing to investigate prokaryotic diversity and relative abundances of prokaryotic groups in four different soil environments	89
2.3.5.1	Preparation of PCR products for Solexa sequencing	89
2.3.5.2	Illumina Solexa technique	91
Chapter 3	The use of Illumina Solexa sequencing to assess prokaryotic diversity of UK forest soils using DNA and RNA	93
3.1	Introduction	93
3.1.1	Use of the 16S SSU rRNA encoding gene to investigate prokaryotic diversity	93
3.1.2	Comparing the use of RNA and DNA to investigate microbial diversity	93

3.1.2.1	Overview of comparison between RNA and DNA in diversity studies	93
3.1.2.2	Studies showing RNA and DNA based communities to be similar	95
3.1.2.3	Studies showing RNA and DNA based communities to be significantly different	96
3.1.2.4	Is it important to use both RNA and DNA based diversity analyses?	97
3.1.3	Why use the Illumina Solexa sequencing platform?	98
3.1.4	The objectives of this study	100
3.2	Materials and Methods	100
3.2.1	Sample preparation	100
3.2.2	Analysis of the sequence data	101
3.3	Results	103
3.3.1	Number of sequences per sample, initial and after screening	103
3.3.2	Errors during the sequencing process	106
3.3.3	Rarefaction and rank abundance curves	107
3.3.4	Expanded oak organic sampling and a linear increase in complexity	112
3.3.5	Comparison of DNA and RNA samples	113
3.3.5.1	Distribution of sequences between RNA and DNA samples	113
3.3.5.2	Comparison of RNA/DNA samples at Phyla level	113
3.4	Discussion	117
3.4.1	The effectiveness of the Solexa sequencing technology in diversity studies	117
3.4.1.1	The number of reads and distribution of reads across samples	117
3.4.1.2	Overview of community structure of both DNA and RNA libraries	120

3.4.2	Comparison of the libraries generated by DNA/RNA samples	121
3.5	Conclusion	123
Chapter 4	Taxonomic analysis of prokaryote communities of UK forest soils	125
4.1	Introduction	125
4.1.1	Bacterial community structure in soil at different taxonomic levels	125
4.1.2	The use of replicates in the investigation of bacterial diversity in soil	126
4.1.3	The use of RNA in investigating diversity	127
4.1.4	The objectives of this study	127
4.2	Materials and Methods	128
4.2.1	Sample preparation	128
4.2.2	Taxonomic analysis	128
4.3	Results	129
4.3.1	Analysing taxonomic diversity across multiple RNA sequence replicates in each soil type	129
4.3.1.1	Relationship between the prokaryotic community of 20 RNA soil replicates as described using a cladogram	129
4.3.1.2	The distribution of Phyla in 20 soil replicates from four soil types	131
4.3.2	Distribution of reads across all four soil types	135
4.3.2.1	Discussion of inclusion of oak organic sample number 6	135
4.3.2.2	Differences between inclusion and exclusion of oak organic sample number 6	135
4.3.2.3	Distribution of OTUs for both Venn diagrams including and excluding oak organic sample number 6	135
4.3.3	Taxonomic distribution at family level of all four soil types	138
4.3.3.1	Overall taxonomic distribution	138

4.3.3.2	Phyla showing similar family distributions across samples	140
4.3.3.3	Phyla showing different family distributions across samples	145
4.3.3.4	Identifying possible members of the 'rare biosphere'	146
4.4	Discussion	147
4.4.1	Comparison of RNA replicates between the four soil types	147
4.4.2	Taxonomic distribution of the four soil types using RNA libraries	148
4.4.3	Identifying the 'rare biosphere' in the RNA samples	150
4.5	Conclusion	151
Chapter 5	Using multivariate statistics to assess the influence of environmental parameters on prokaryote diversity across UK forest soils	154
5.1	Introduction	154
5.1.1	The effect of environmental parameters on prokaryote community composition and diversity	154
5.1.2	Using multivariate statistics to investigate environmental influences on prokaryote diversity	155
5.1.2.1	How multivariate statistics are used in microbial ecology	155
5.1.2.2	Statistical limitations of PCA	155
5.1.2.3	Measuring diversity in microbial communities	156
5.2	Materials and methods	157
5.2.1	Sample preparation and soil characterisation	157
5.2.2	Calculating diversity	158
5.2.3	Principal components analysis (PCA)	158
5.3	Results	159
5.3.1	Chemistry descriptions of the soil samples	159

5.3.2	Diversity measures	162
5.3.3	Description of the principal components analysis	164
5.3.3.1	Overview of the principal components analysis	164
5.3.3.2	PCA describing the influence of trace elements on prokaryote diversity	164
5.3.3.3	PCA describing the influence of soil parameters (pH, soil moisture, C, N, C:N ratio, S, Al, Fe, Mn, K, Mg, P, Ca) on prokaryote diversity	167
5.4	Discussion	172
5.4.1	Diversity measures of the prokaryote community of these forest soils	172
5.5	Conclusion	174
Chapter 6	Conclusions	175
6.1	General conclusions from this study	175
6.2	Future work	178
Chapter 7	Bibliography	181