SOIL METABARCODING **RESULTS**

Report number:	NM-TSV931
Order number:	102850
Company:	Sapperton Wilder
Contact:	
Project:	Sapperton Wilder project
Sample type:	Soil
Date of report:	20/04/2022
Number of samples:	78

*any sensitive information removed for general viewing purposes

Thank you for sending your samples for analysis by NatureMetrics. Your samples have been metabarcoded following our Soil Bacteria, Fungi and Fauna pipelines. Taxon-by-sample tables of your samples are attached to this report (Appendix A, Appendix B, and Appendix C, respectively). Each row in the table represents one taxon (OTU), shown with the lowest possible taxonomic assignment based on currently available reference data. Each column represents a sample, showing the number of sequence reads per detected OTU. Care should be taken in interpreting the numbers in terms of relative abundance, but a high sequence proportion can be interpreted as lending greater confidence to a detection.

Here we present an overview of the key results, followed by a more detailed report including an introduction to the project, the results for each target of interest, and details of the steps taken to amplify, sequence, and analyse your DNA. A glossary for terms in bold is provided at the end of the report to define key terms used within the report.

OVERVIEW OF YOUR RESULTS

- A total of 3,379 taxa were detected across the samples (1,384 bacteria, 1,676 fungi, and 319 fauna)
- The Northern, Central, and Southern Blocks were quite similar in their soil community composition but they had different composition compared to the "target" comparison habitat control areas
- Area-level (cumulative) richness for bacteria and fungi were higher in the Central Block compared to the Northern and Southern Blocks
- Area-level richness for soil fauna was lower in the Southern Block compared to the Central and Northern Blocks
- The sampling effort was sufficient to capture most of the soil bacterial, fungal, and faunal communities in the rewilding blocks which will provide a good baseline for future sampling surveys

Introduction

NatureMetrics was engaged by Sapperton Wilder to conduct DNA metabarcoding analysis on 78 soil samples. The aim of the works presented in this report is to provide a baseline survey for surface soil communities prior to rewilding management at a site in Sapperton, Gloucestershire. This will enable future surveys to monitor changes in these communities that occur as a result of the works. The sampled areas that will be subject to management change consisted of 19 fields across three blocks of land. Two additional fields were sampled to serve as agricultural controls for comparison in future sampling events. Five other areas were also sampled in the surrounding landscape which are representative of good condition habitats to inform future targets. Sampling locations are indicated in Figure 1.

The molecular techniques used in this study enable assessment of important components of biodiversity that are often neglected due to the difficulty of taxonomic identification using conventional morphological approaches. Three different DNA metabarcoding assays were applied to DNA extracted from each soil sample: bacteria, fungi, and fauna.

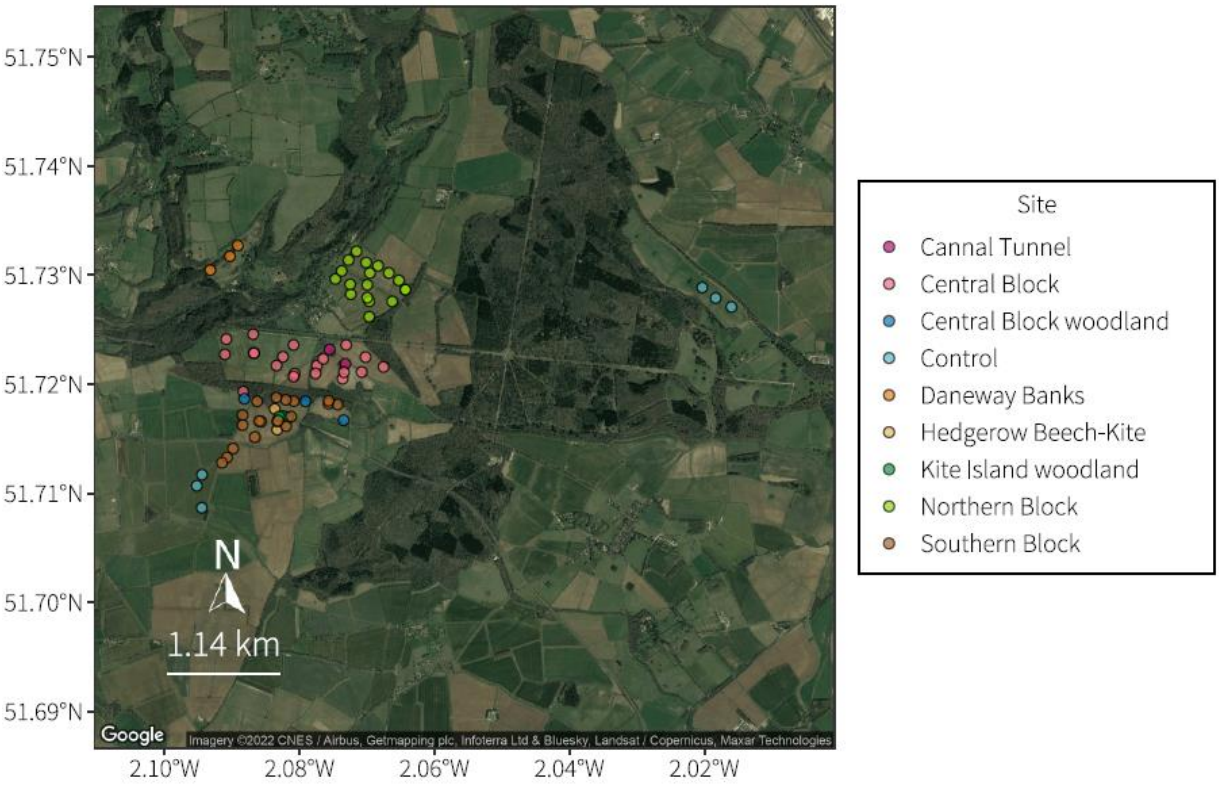

Figure 1. Soil sampling locations in Sapperton, Gloucestershire. Note that coordinates were not provided for two samples (211024OT1 and 211023CT3) so these are not shown on the map.

Results

Sequencing Data Summary

The final dataset contained a total of 3,379 OTUs across the samples: 1,384 bacteria, 1,676 fungi, and 319 fauna (Table 1). More fungal OTUs were identified at the species level compared to bacteria. This reflects differences in availability of reference sequences for different organisms within the reference databases and a higher proportion of assignment conflicts (100% matches to multiple species) in bacteria.

Table 1. Summary of the number of OTUs detected and the percentage of OTUs successfully classified at each taxonomic level for each target

Target	Number of OTUs	Phylum	Class	Order	Family	Genus	Species
Bacteria	1,384	80.3%	63.7%	50.1%	38.3%	17.6%	3.5%
Fungi	1,676	99.3%	91.9%	83.2%	67.4%	41.8%	17.8%
Fauna	319	99.4%	80.3%	86.8%	66.1%	27.9%	5.0%

Bacteria

In the bacterial dataset, OTUs were detected across 22 different phyla within the kingdom Bacteria. The average bacterial taxon richness per sample of rarefied data was 471.1 and ranged from 404 (211024OT1) to 532 (211023CT3). The overarching taxonomy of the detected OTUs is presented in Figure 2, which shows that the phylum with the highest richness of OTUs was Proteobacteria. The bacterial OTU with the most reads was from the order Burkholderiales. This OTU was detected in 78/78 samples. 132 bacterial OTUs were detected in every sample. There were 284 OTUs (20.5%) that were only detected in one sample each.

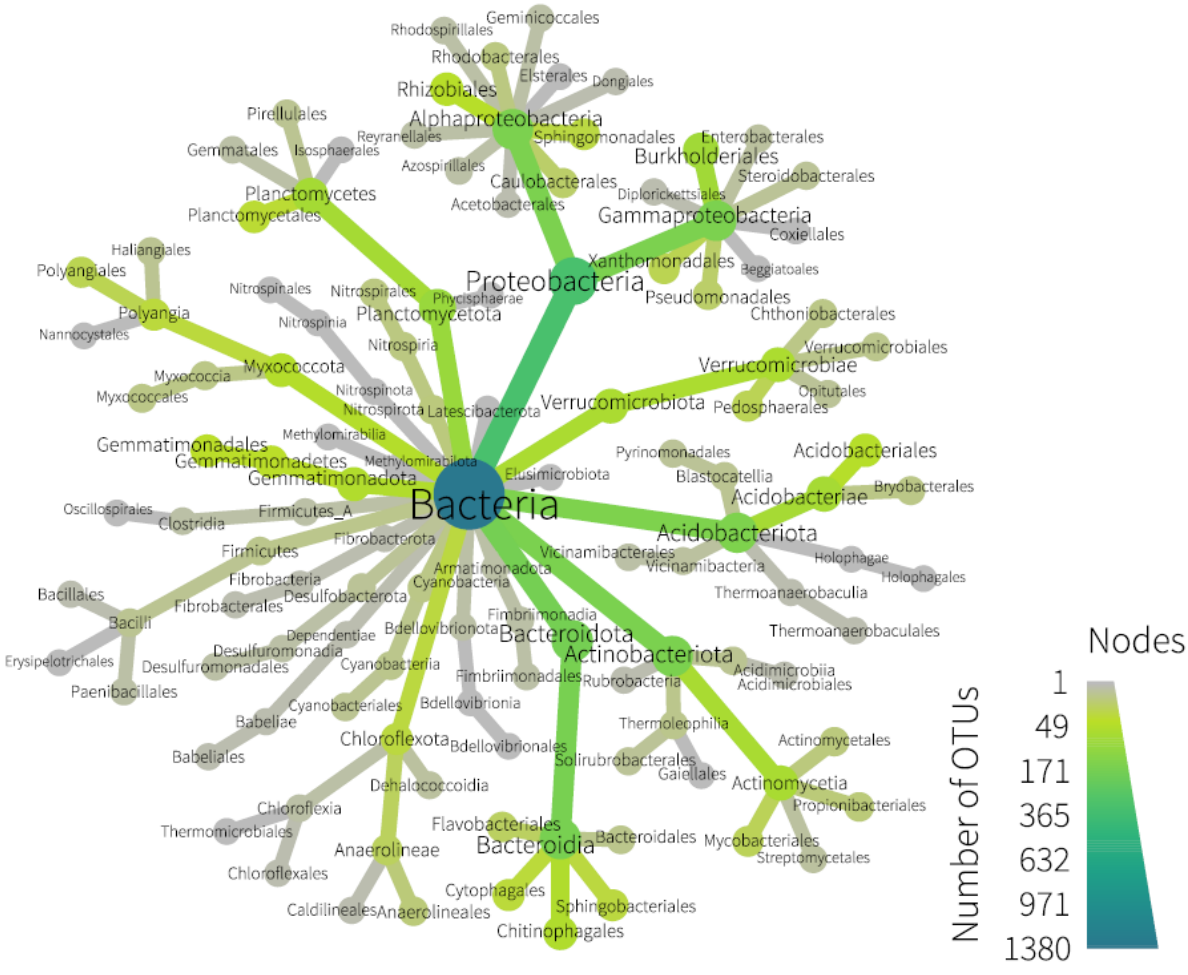

Figure 2. A taxonomic heat tree showing the number of OTUs across all samples for soil bacterial taxa down to the order rank. Each node (the circles) is a taxon and the edges (lines) show hierarchical relationships between taxa. The colour scale and the relative width of the node represent the number of taxa at each level.

The Northern, Central, and Southern Blocks were quite similar in bacterial community composition, as seen in Figure 3 where the points from those blocks cluster together. The Central Block was more variable in its composition compared to the Northern and Southern Blocks as indicated in the plot by a larger spread of points. The agricultural Control samples were quite similar to the rewilding blocks but clusters slightly off to the side indicating some differences. The "target" comparison habitat controls mostly clustered separately to the rewilding and agricultural control blocks, with the Hedgerow Beech-Kite being the most similar in composition to the rewilding blocks. Within the rewilding blocks, some fields clustered separately from each other but others had overlap (Figure 4).

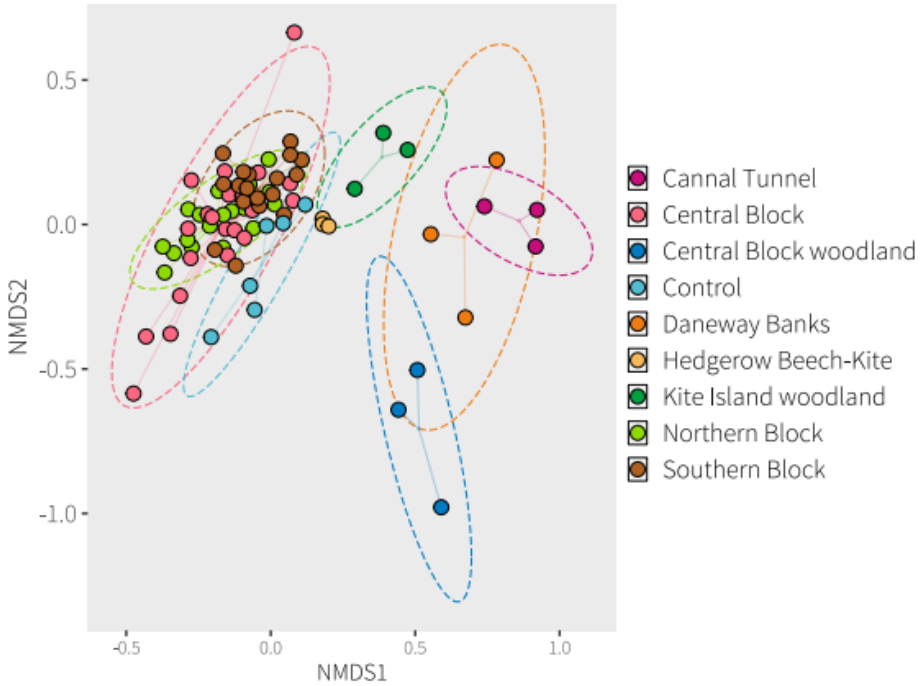

Figure 3. NMDS ordination plots based on Jaccard similarity index for soil bacterial communities. Points are coloured by area, with 95% confidence intervals for each area indicated by dashed ellipses.

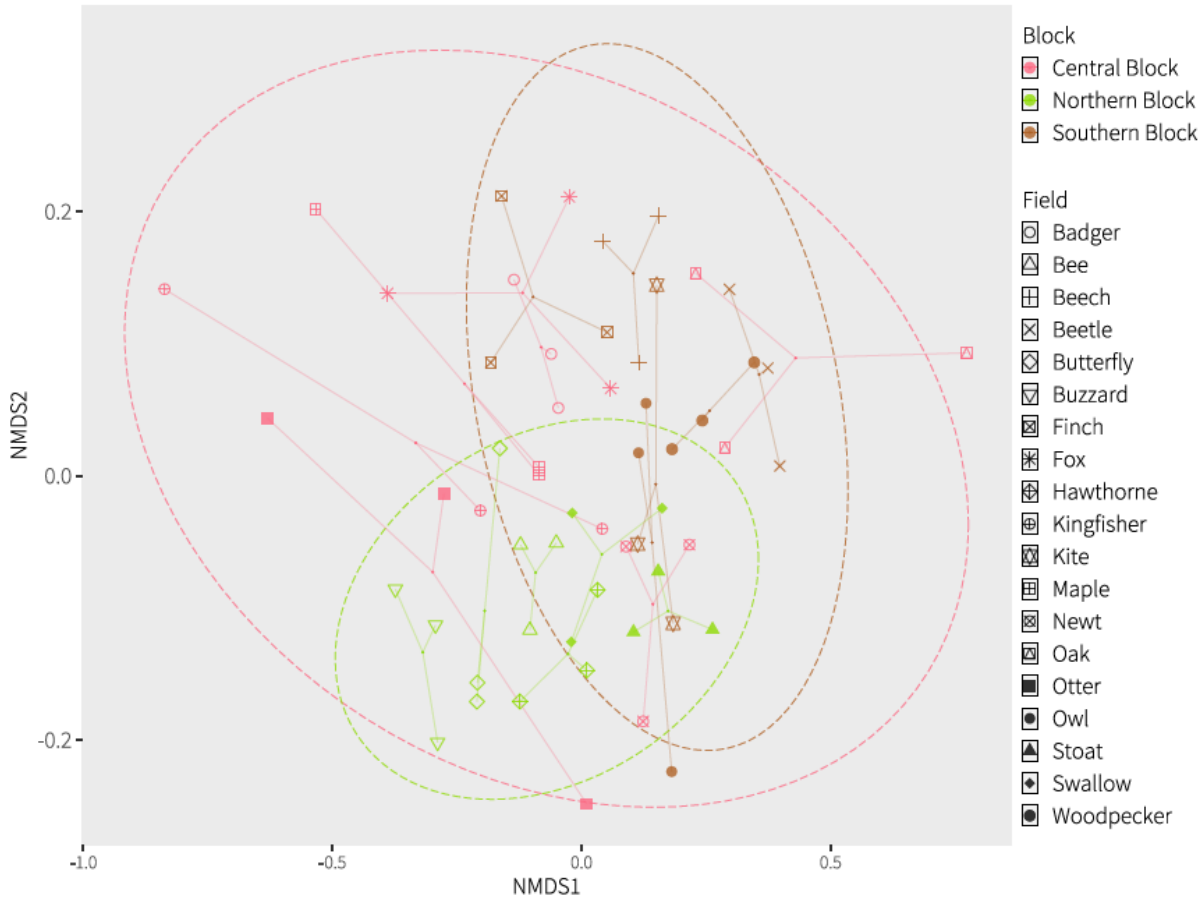

Figure 4. NMDS ordination plots based on Jaccard similarity index for soil bacterial communities in the rewilding fields. Points are coloured by block and shaped by field, with 95% confidence intervals for each block indicated by dashed ellipses.

Average bacterial taxon richness per sample was highest at the Cannal Tunnel control area and lowest at the Daneway Banks control area (Figure 5). The Northern, Central, and Southern Blocks were quite similar in terms of sample-level richness, with more variation in the Central Block. The species accumulation curves (Figure 6) show that the Central Block has higher area-level bacterial richness (cumulative richness) compared to the Northern and Southern Blocks. The curves for all blocks have started to plateau indicating that the sampling effort was sufficient to capture most of the soil bacterial community.

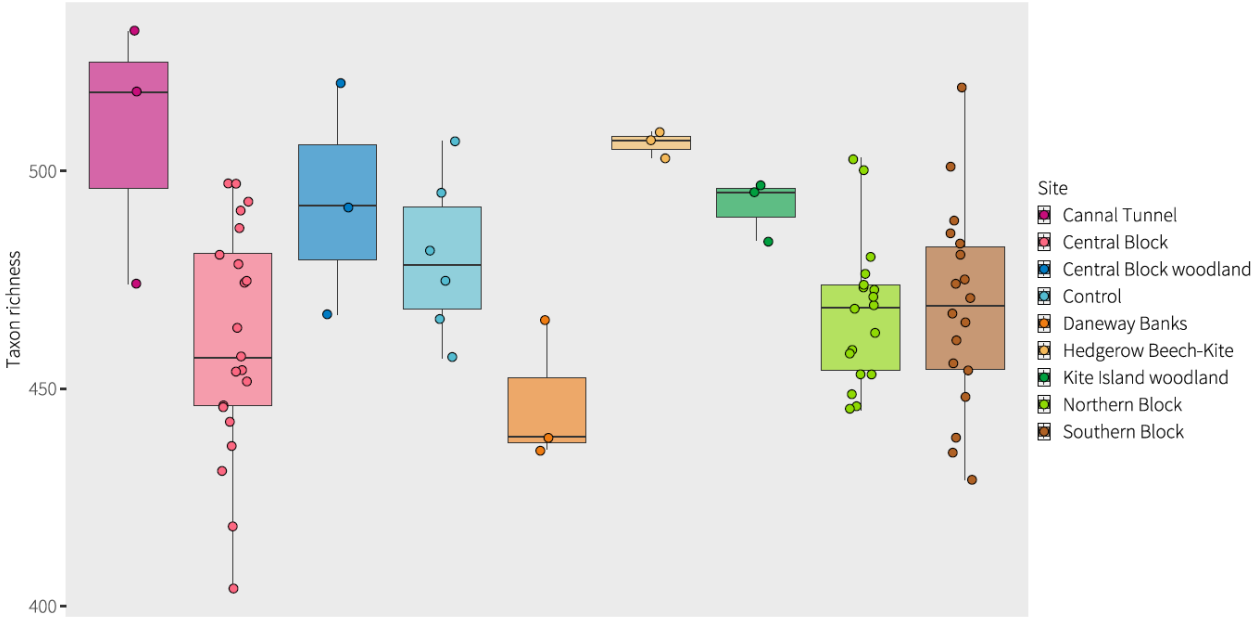

Figure 5. Taxon richness (number of OTUs) for soil bacterial communities within each area. The boxplot shows sample-level richness, with the box depicting the median between the upper/lower quantiles, the whiskers indicating minimum and maximum values, and dots showing richness values of each sample. Any samples beyond the whiskers are considered as outliers which are 1.5x the interquartile-range away from the upper or lower quartile.

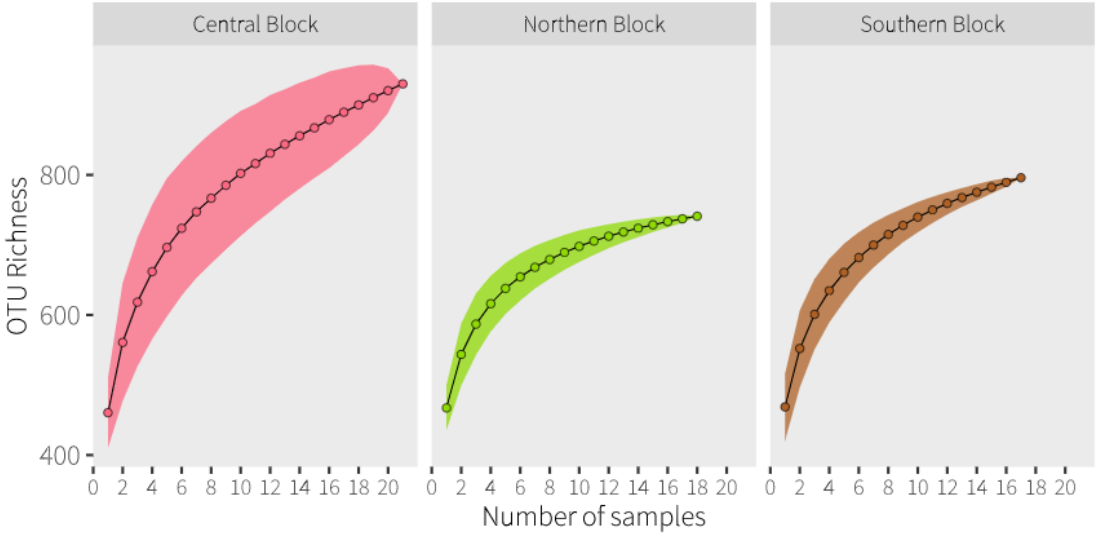

Figure 6. Species accumulation curves showing cumulative richness (number of OTUs) detected for bacteria across all samples within each rewilding block. Shading shows 2 x standard deviation.

Fungi

In the fungal dataset, OTUs were detected across 5 different phyla within the kingdom Fungi. The average fungal taxon richness per sample of rarefied data was 184.4 and ranged from 91 (211104BET2(M)) to 303 (211023CT3). The overarching taxonomy of the detected OTUs is presented in Figure 7, which shows that the phylum with the highest richness of OTUs was Ascomycota. The fungal OTU with the most reads was from the class Dothideomycetes. This OTU was detected in 77/78 of the samples. 3 fungal OTUs were detected in every sample. There were 618 OTUs (36.9%) that were only detected in one sample each.

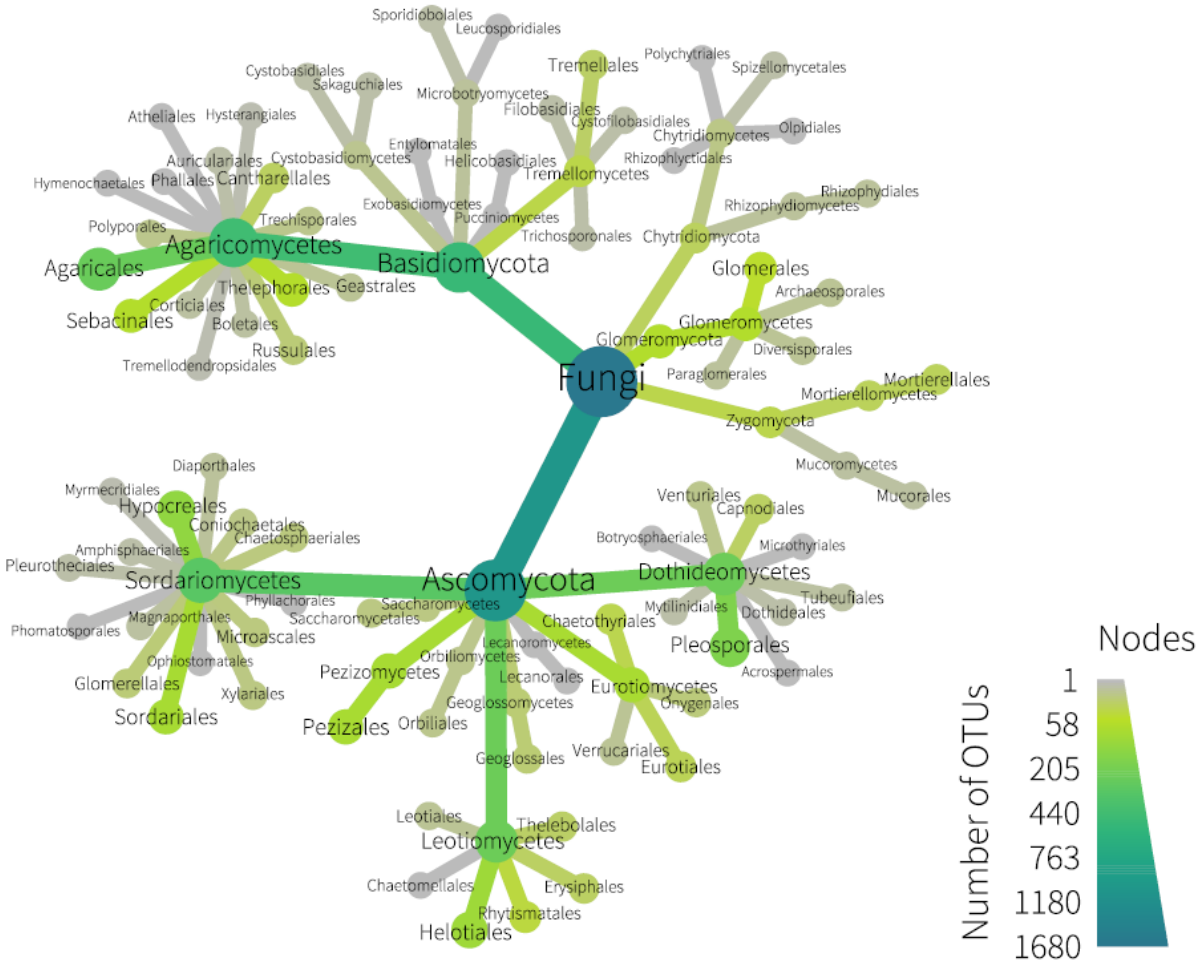

Figure 7. A taxonomic heat tree showing the number of OTUs across all samples for soil fungal taxa down to the order rank. Each node (the circles) is a taxon and the edges (lines) show hierarchical relationships between taxa. The colour scale and the relative width of the node represent the number of taxa at each level.

The “target” comparison habitat controls clustered separately to the rewilding and agricultural control blocks, with the Hedgerow Beech-Kite being the most similar in fungal community composition to the rewilding blocks (Figure 8). The Central Block woodland was the most variable in its fungal community composition. Within the rewilding blocks, The Northern Block had distinctly different fungal communities compared to the Central and Southern Blocks (Figure 9). Some fields were more distinct in their communities than others, for example, the Bee field clusters separately to the other Central Block fields.

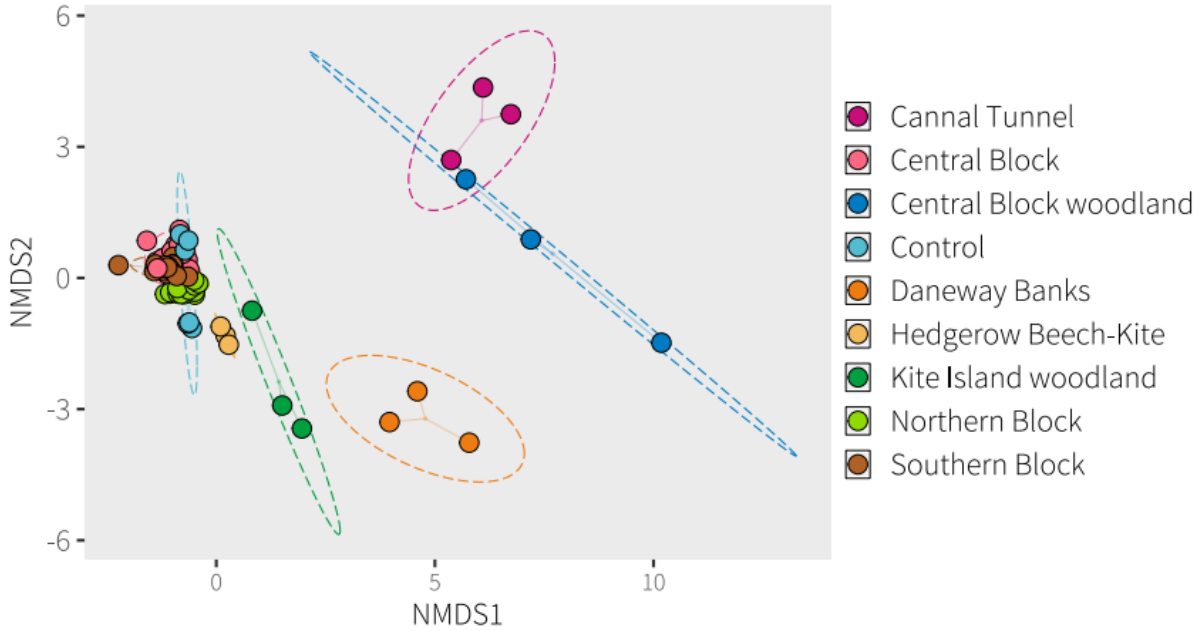

Figure 8. NMDS ordination plots based on Jaccard similarity index for soil fungal communities. Points are coloured by area, with 95% confidence intervals for each area indicated by dashed ellipses.

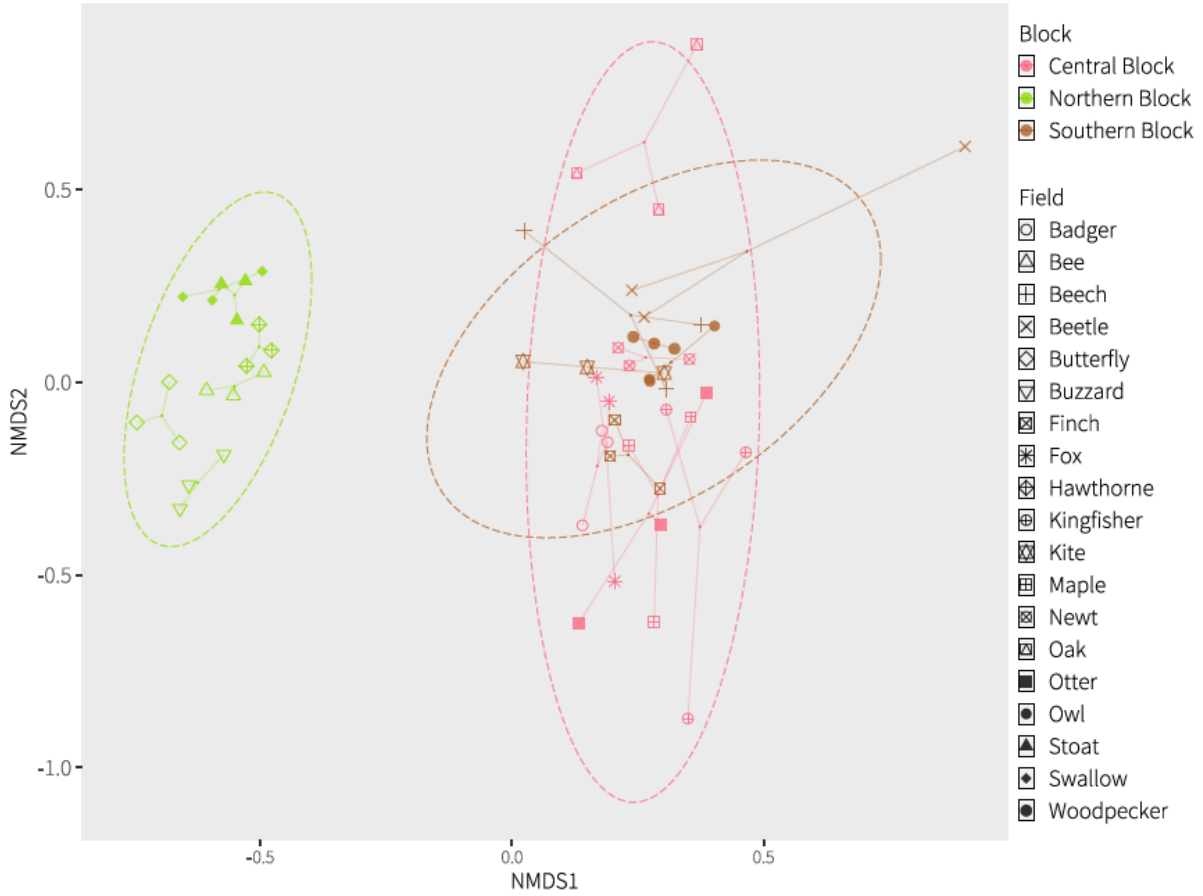

Figure 9. NMDS ordination plots based on Jaccard similarity index for soil fungal communities in the rewilding fields. Points are coloured by block and shaped by field, with 95% confidence intervals for each block indicated by dashed ellipses.

Average fungal taxon richness per sample was similarly high in the Cannal Tunnel, Daneway Banks, and agricultural Control areas (Figure 10). Sample-level fungal taxon richness was similar across the three rewilding blocks. The species accumulation curves (Figure 11) show that the Central Block has higher area-level fungal richness (cumulative richness) compared to the Northern and Southern Blocks. The curves for all blocks have started to plateau indicating that the sampling effort was sufficient to capture most of the soil fungal community.

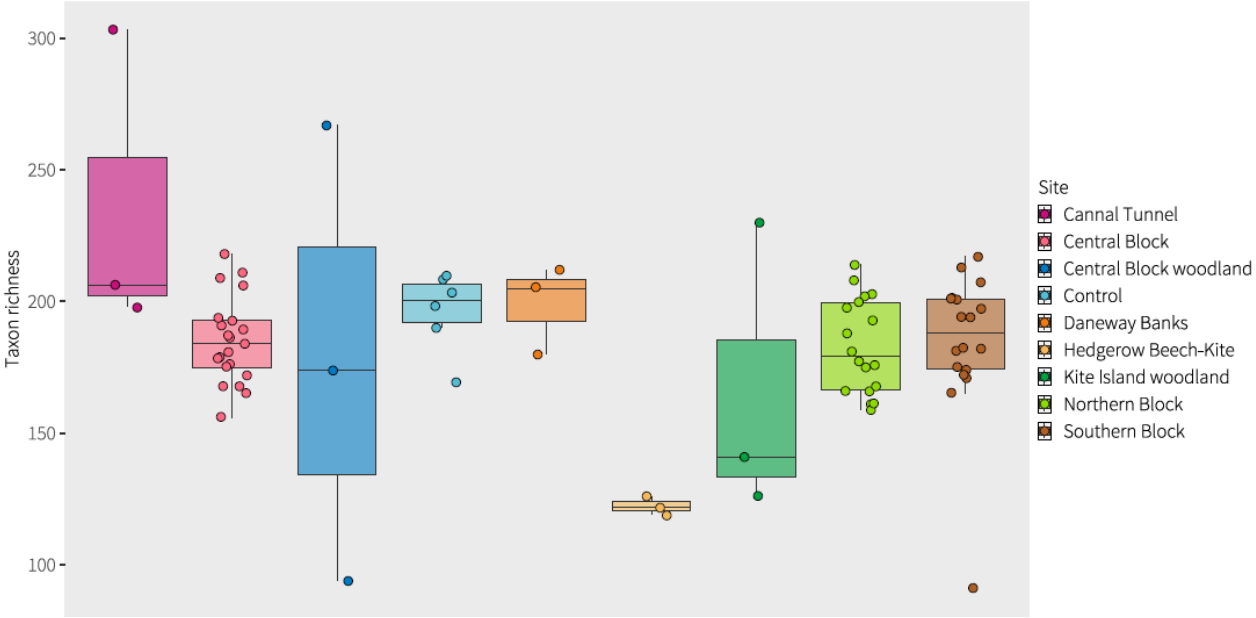

Figure 10. Taxon richness (number of OTUs) for soil fungal communities within each area. The boxplot shows sample-level richness, with the box depicting the median between the upper/lower quantiles, the whiskers indicating minimum and maximum values, and dots showing richness values of each sample. Any samples beyond the whiskers are considered as outliers which are 1.5x the interquartile-range away from the upper or lower quartile.

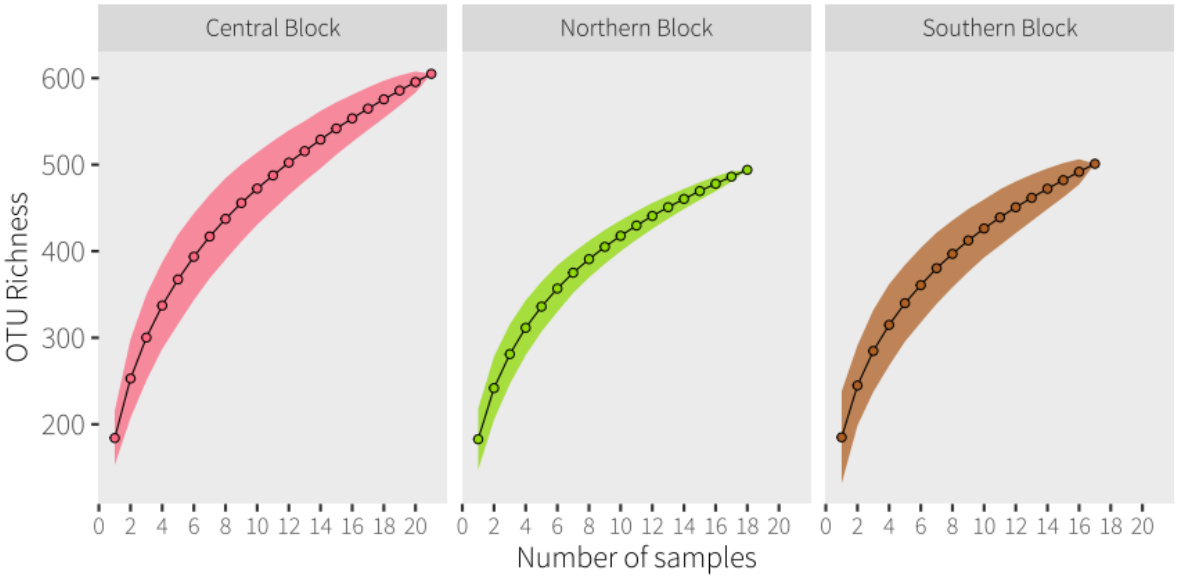

Figure 11. Species accumulation curves showing cumulative richness (number of OTUs) detected for fungi across all samples within each rewilding block. Shading shows 2 x standard deviation.

Fauna

In the faunal dataset, OTUs were detected across 8 different phyla within the kingdom Animalia. The average faunal taxon richness per sample of rarefied data was 40.5 and ranged from 7 (211022HA1) to 100 (211023CT2). The overarching taxonomy of the detected OTUs is presented in Figure 12, which shows that the phylum with the highest richness of OTUs was Arthropoda. The OTU with the highest proportion of reads was from the family Enchytraeidae, a group of annelids. This OTU was detected in 78/78 of the samples. This was the only faunal OTU detected in every sample. There were 104 OTUs (32.6%) that were only detected in one sample each.

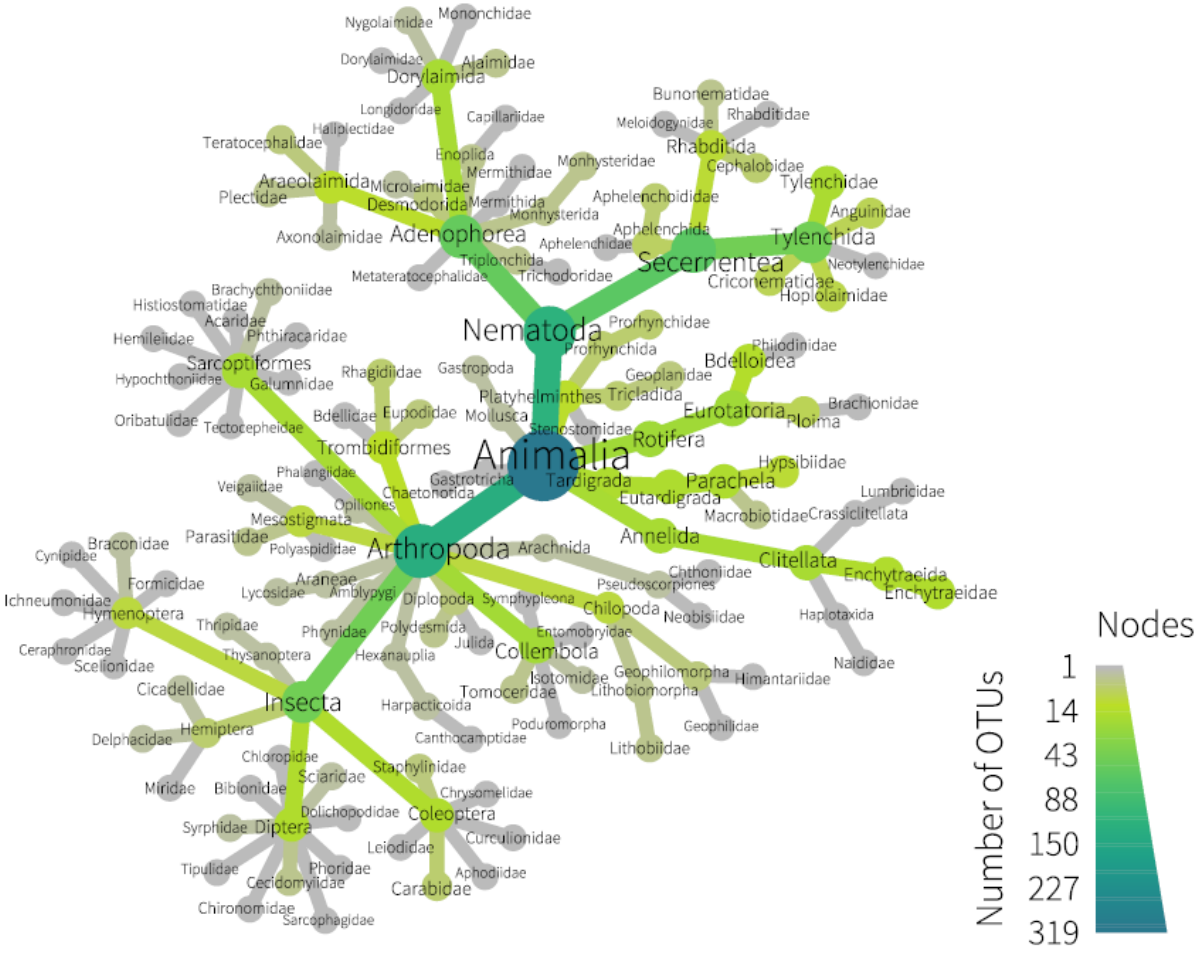

Figure 12. A taxonomic heat tree showing the number of OTUs across all samples for soil faunal taxa down to the family rank. Each node (the circles) is a taxon and the edges (lines) show hierarchical relationships between taxa. The colour scale and the relative width of the node represent the number of taxa at each level.

The rewilding blocks and the agricultural control area were all quite similar in their soil faunal community composition but the Northern Block was a bit more distinct with its samples clustering slightly apart (Figure 13). The "target" comparison habitat controls clustered separately to the rewilding and agricultural control blocks, with the Hedgerow Beech-Kite and Central Block Woodland being the most similar in faunal community composition to the rewilding blocks, and the Cannal Tunnel being the most different. Within each block, there was not much differentiation between fields, only the Bee field formed a separate cluster (Figure 14).

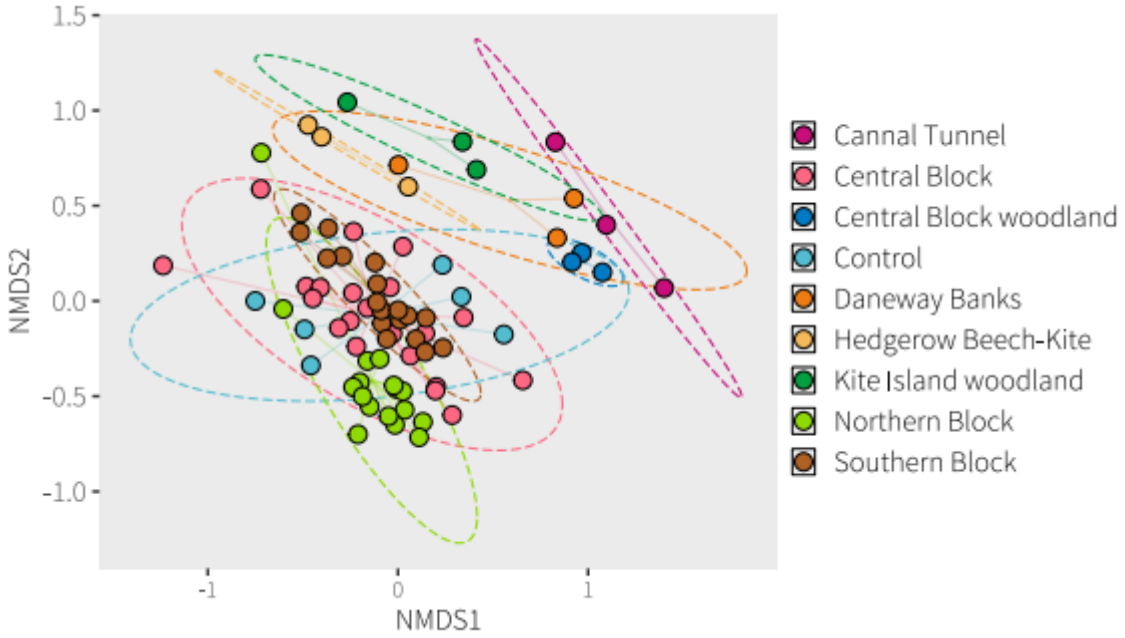

Figure 13. NMDS ordination plots based on Jaccard similarity index for soil faunal communities. Points are coloured by area, with 95% confidence intervals for each area indicated by dashed ellipses.

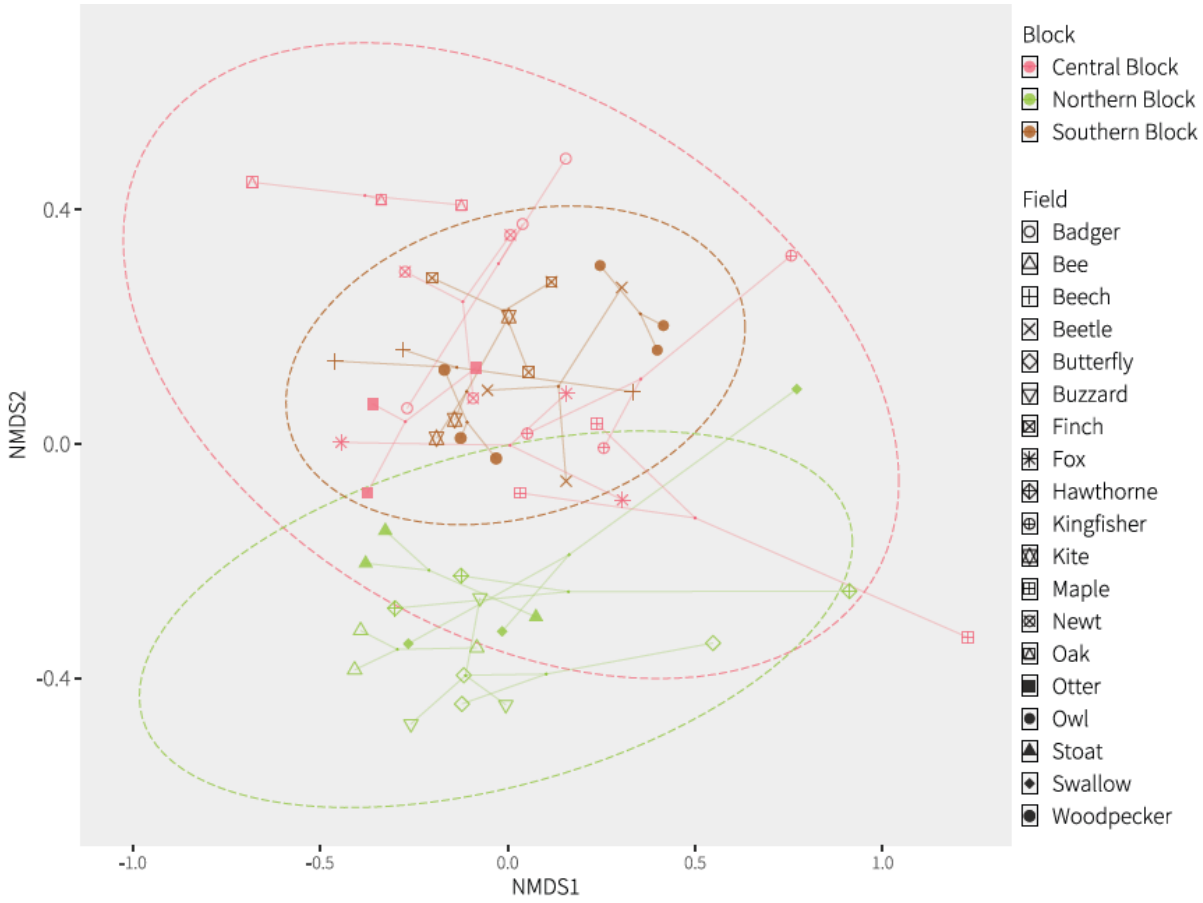

Figure 14. NMDS ordination plots based on Jaccard similarity index for soil faunal communities in the rewilding fields. Points are coloured by block and shaped by field, with 95% confidence intervals for each block indicated by dashed ellipses.

Average faunal taxon richness per sample was highest in the Cannal Tunnel area and lowest at the Hedgerow Beech-Kite area (Figure 15). Compared to soil bacteria and fungi, fauna showed more differentiation between the rewilding blocks, with the Northern Block having higher sample-level richness than the Central and Southern Blocks. However, area-level faunal richness (cumulative richness) in the Northern Block was similar to the Central Block, with the Southern Block having slightly lower area-level richness (Figure 16). The curves for all blocks have started to plateau indicating that the sampling effort was sufficient to capture most of the soil faunal community.

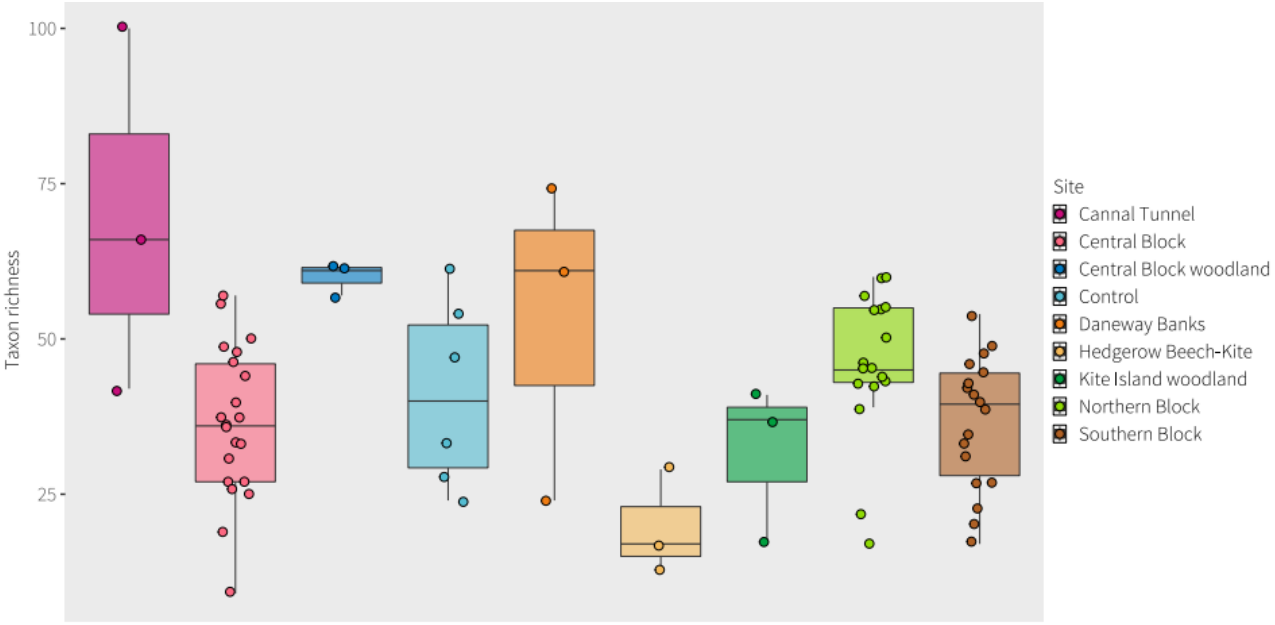

Figure 15. Taxon richness (number of OTUs) for soil faunal communities within each area. The boxplot shows sample-level richness, with the box depicting the median between the upper/lower quantiles, the whiskers indicating minimum and maximum values, and dots showing richness values of each sample. Any samples beyond the whiskers are considered as outliers which are 1.5x the interquartile-range away from the upper or lower quartile.

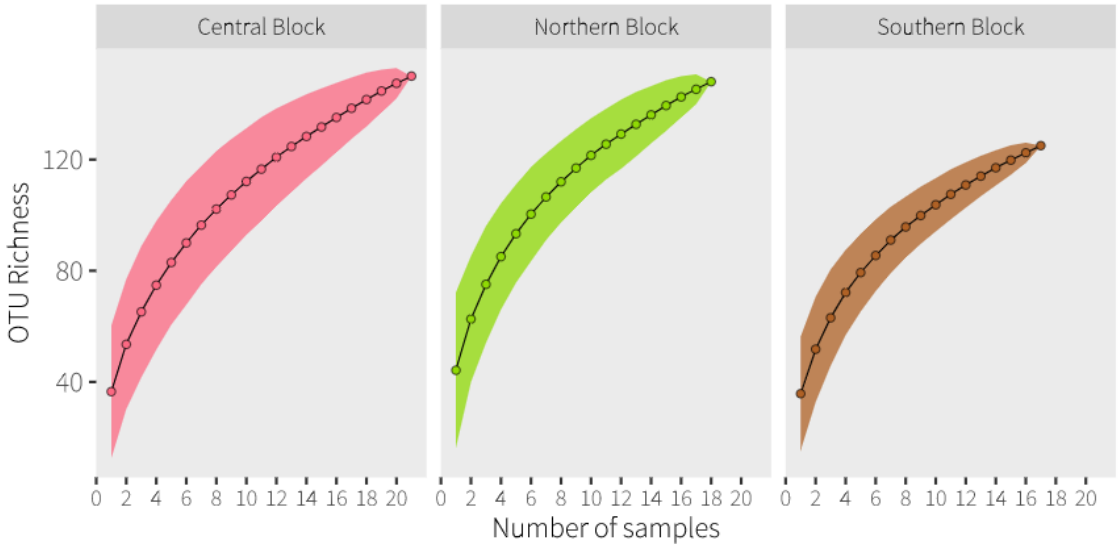

Figure 16. Species accumulation curves showing cumulative richness (number of OTUs) detected for soil fauna across all samples within each rewilding block. Shading shows 2 x standard deviation.

METHODS

DNA Extraction and Sequencing

Composite soil samples were collected by the client using a metal corer which was decontaminated between samples. Samples were received by NatureMetrics chilled in a cool box with ice packs. DNA was extracted from approximately 10 grams of each sample using a commercial DNA extraction kit. An extraction blank was also processed for each extraction batch. DNA was quantified using a Qubit DNA broad range kit according to the manufacturer's protocol.

Comment: The average yield of the DNA samples was 33.1 ng/µl and ranged from 6.3 ng/µl to 108 ng/µl (Appendix D). Any DNA in the extraction blank was below the detection limit.

DNAs were amplified in triplicate PCRs with primers targeting the V4 region of the 16S rRNA gene for bacteria, the internal transcribed spacer 2 (ITS2) region for fungi, and the V4-V5 region of the 18S rRNA gene for fauna. All PCRs were performed in the presence of a negative control, the extraction blank, and a positive control sample (a sample known to amplify with those primers). Amplification success was determined by gel electrophoresis.

Comment: PCR reactions were successful for all 78 samples (Appendix D). Electrophoresis bands were strong and of the expected size. No bands were observed on electrophoresis gels for the negative controls.

PCR replicates were pooled and purified, and sequencing indexes were added. Success was determined by gel electrophoresis.

Comment: All samples were successfully indexed, electrophoresis bands were strong and of the expected size. No repeat reactions were necessary.

Amplicons were purified and checked by gel electrophoresis, these were then quantified using a Qubit DNA broad range kit according to the manufacturer's protocol.

Comment: All amplicons were successfully purified and were of high yield (Appendix D).

All purified index PCRs were pooled into final libraries with each sample added in equal concentrations. The final libraries were sequenced using an Illumina MiSeq V3 kit at 10 pM with a 20% PhiX spike in.

Bioinformatics

Sequence data were processed using a custom bioinformatics pipeline for quality filtering, OTU clustering (97%) and taxonomic assignment.

Comment: Both negative and positive controls were as expected. A total of 21,716,507 (bacteria: 7,018,147, fungi: 6,374,961, and fauna: 8,323,399) high-quality sequences are represented in the results.

Consensus taxonomic assignments were made for each OTU using sequence similarity searches against two reference databases appropriate for the dataset, one generic (NCBI nt) and one specialised:

- Bacteria: SILVA 16S (v138.1)
- Fungi: UNITE (v8.2)
- Fauna: SILVA 18S (v138.1)

The GBIF taxonomic backbone was used for consistency between databases. Results from both searches were combined and assignments made to the lowest possible taxonomic level where there was consistency in the matches. Conflicts were flagged and resolved manually. Minimum similarity thresholds of 98%, 95%, and 92% were required for species-, genus-, and higher-level assignments, respectively. Identifications that were based on fewer than three reference matches have been flagged. Taxonomic identifications for fauna have been reported as tentative in cases where matches to reference sequences were sufficient for species- or genus-level identifications but the taxon is not known to be present in United Kingdom.

The OTU table was filtered to remove low abundance OTUs from each sample (<0.05% or <10 reads, whichever is the greater threshold for the sample). Results are presented for OTUs identified to target kingdom or below. Note that unidentified or misidentified taxa can result from incomplete or incorrect reference databases, and taxa may be missed due to low quality DNA, environmental contaminants, or the dominance of other species in the sample.

Sequencing depths were considered sufficient for all other samples except for 211022HA1 for soil fauna so this sample was removed from the faunal analyses. Samples with different numbers of reads are not directly comparable, therefore read counts obtained from the OTU tables were normalised by rarefying to the lowest sequencing depth prior to subsequent analyses.

Limitations

Methodologies have been chosen based on the state of the art, but these choices inherently introduce specific limitations and biases. For each of the target groups we have chosen primers that in our experience capture their targets well. Each of these primer sets will inherently miss taxa and this will be a systematic error. Unfortunately, there is no one primer set that captures all of the diversity, and the diversity present in soil makes it impossible to choose one primer set that balances specificity and resolution. This is only an issue if there is a particular target taxon (i.e. an indicator species) of interest.

Assigning taxonomic identities to the sequences is only possible through their comparison to reference databases, which are incomplete. This is not an issue if a taxonomy free approach is adopted – i.e. tracking changes over time by comparing datasets (as is advocated here), but it is a bigger concern if indicator species or functional groups (based on taxonomy) are required. It should be noted that multiple OTUs can be identified as belonging to the same species, which is most likely attributed to PCR

or sequencing artefacts but potentially intraspecific genomic variation or cryptic diversity. Also, it is possible for closely related species to have identical sequences in the targeted gene region and if the species present at your site is not in the database it could be identified as a different closely related species.

What is happening among the communities (i.e. functioning) may be driven more so by the dominant taxa as opposed to the breadth of diversity. The abundance of taxa cannot be directly inferred from the number of sequence reads. While the number of sequence reads is a consequence of abundance, it is also impacted by biomass, body type, activity, surface area, condition, primer bias, and species-specific variation in the genome.

END OF REPORT

Report issued by: **Hayley Craig**
Contact: **team@naturemetrics.co.uk**

GLOSSARY

bioinformatics	Refers to a data processing pipeline that takes the raw sequence data from high-throughput sequencing (often 20 million sequences or more) and transforms it into usable ecological data. Key steps for metabarcoding pipelines include quality filtering, trimming, merging paired ends, removal of sequencing errors such as chimeras, clustering of similar sequences into molecular operational taxonomic units (OTUs; each of which approximately represents a species), and matching one sequence from each cluster against a reference database. The output is a species-by-sample table showing how many sequences from each sample were identified as each species.
extraction blank	A DNA extraction with no soil added to assess potential contamination during the DNA extraction process.
gel electrophoresis	The process in which DNA is separated according to size and electrical charge via an electric current, while in a gel. The process is used to confirm the successful amplification of a specific size fragment of DNA.
high-throughput sequencing	Technology developed in the 2000s that produces millions of sequences in parallel. Enables thousands of different organisms from a mixture of species to be sequenced at once, so community DNA can be sequenced. Various different technologies exist to do this, but the most commonly used platform is Illumina's MiSeq. Also known as Next-Generation Sequencing (NGS) or parallel sequencing.
Jaccard similarity index	This index is a calculation that compares two samples to see which taxa are shared and which are distinct. The higher the percentage, the more similar two samples are in their community composition.
metabarcoding	Refers to identification of species assemblages from community DNA using barcode genes. PCR is carried out with non-specific primers, followed by high-throughput sequencing and bioinformatics processing. Can identify hundreds of species in each sample, and 100+ different samples can be processed in parallel to reduce sequencing cost.
negative control	Used to determine if PCR reactions are contaminated.
NMDS	Non-metric multidimensional scaling (NMDS) is a method that allows you to visualise the similarity of each sample to one another. The dissimilarity between each sample is calculated, taking into account shared taxa (Jaccard similarity index), and then configured into a 2D ordinal space that allows you to see the relationship of each sample to one another. Samples that are closer together are more similar to one another in terms of community composition, while samples that are further apart are less similar. This type of clustering analysis allows you to see if

	certain types of samples, for example, those from a particular habitat type, are more clustered together and therefore more similar to one another compared to other groups.
OTU	Short for Operational Taxonomic Unit. Similar sequences are clustered into OTUs at a defined similarity threshold. OTUs are approximately equivalent to species and are treated as such in our analyses. Species-level taxonomic assignments may or may not be possible, depending on the availability of reference sequences and the similarity between closely related species in the amplified marker. It may be possible to refine the taxonomic assignment for an OTU later as more sequences are added to reference databases.
PCR	Short for Polymerase chain reaction. A process by which millions of copies of a particular DNA segment are produced through a series of heating and cooling steps. Known as an 'amplification' process. One of the most common processes in molecular biology and a precursor to most sequencing-based analyses.
positive control	Used to determine whether the PCR is working correctly.
primers	Short sections of synthesised DNA that bind to either end of the DNA segment to be amplified by PCR. Can be designed to be totally specific to a particular species (so that only that species' DNA will be amplified from a community DNA sample), or to be very general so that a wide range of species' DNA will be amplified. Good design of primers is one of the critical factors in DNA-based monitoring.
rarefaction curve	A plot showing the number of taxa as a function of the sequencing depth (number of reads). Rarefaction curves grow rapidly at first as common species are found then reaches a plateau as only the rarest species remain to be detected.
rarefy	A normalisation technique which transforms the data to remove biases associated with uneven sampling depth (number of reads) across samples. The sampling depth of each sample is standardised to a specified number of reads (usually that of the sample with the lowest depth) by random resampling.
reference databases	Over time, the DNA sequences of many species have been compiled into publicly accessible databases by scientists from around the world. These databases serve as a reference against which unknown sequences can be queried to obtain a species identification. The most commonly accessed database is NCBI (National Center for Biotechnology Information), which is maintained by the US National Institute of Health. Anyone can search for DNA sequences at https://www.ncbi.nlm.nih.gov
richness	The total number of taxa within a sample.
sequence	A DNA sequence is made up of four nucleotide bases represented by the letters A, T, C & G. The precise order of these letters is used

	to compare genetic similarity among individuals or species and to identify species using reference databases. In high-throughput sequencing analyses (e.g. metabarcoding), many identical copies of the same sequence are obtained for each species in the sample. The number of copies obtained per species is known as the number of sequence reads, and this is often - although not always - related to the relative abundance of the species.
taxon (s.) / taxa (pl.)	Strictly, a taxonomic group. Here we use the term to describe groups of DNA sequences that are equivalent to species. We do not use the term species because we are unable to assign complete identifications to all of the groups at this time due to gaps in the available reference databases.
taxonomy species (s./pl.)	A group of genetically similar organisms that show a high degree of overall similarity in many independent characteristics. Related species are grouped together into progressively larger taxonomic units, from genus to kingdom. *Homo sapiens* (human) is an example of a species. genus (s.) / genera (pl.) - A group of closely related species. Each genus can include one or more species. Homo is an example of a genus. family (s.) / families (pl.) - A group of closely related genera. Homo sapiens is in the Family Hominidae (great apes). order (s.) / orders (pl.) - A group of closely related families. Homo sapiens is in the Order Primates. class (s.) / classes (pl.) - A group of closely related orders. Homo sapiens is in the Class Mammalia. phylum (s.) / phyla (pl.) - A group of closely related classes. Homo sapiens is in the Phylum Chordata.