Bioassessment of groundwater ecosystems

III. A comparison of eDNA metabarcoding and metagenomics for assessing groundwater communities

On 15 December 2023, the *Nature Repair (Consequential Amendments) Act 2023* amended the EPBC Act to expand the IESC's remit to all unconventional gas developments. This publication was developed prior to these amendments of the EPBC Act commencing.

This publication is funded by the Australian Government Department of Climate Change, Energy, the Environment and Water. The views and opinions expressed in this publication are those of the authors and do not necessarily reflect those of the Australian Government or the Minister for the Environment and Water.

Citation

Hose GC, McKnight K, Greenfield P, Chariton A and Korbel K 2024. Bioassessment of groundwater ecosystems III. A comparison of eDNA metabarcoding and metagenomics for assessing groundwater communities. Report prepared for the Independent Expert Scientific Committee on Unconventional Gas Development and Large Coal Mining Development through the Department of Climate Change, Energy, the Environment and Water. Commonwealth of Australia.

Contact details

For information about this report or about the work of the IESC, please contact:

IESC Secretariat

Office of Water Science
Department of Climate Change, Energy, the Environment and Water
GPO Box 787
CANBERRA ACT 2601

The report can be accessed at http://www.iesc.environment.gov.au/

Images

Front cover: DNA extraction under clean laboratory conditions | Location: North Ryde, NSW | Photo credit: K Korbel

Back cover: PCR preparation under clean laboratory conditions | Location: North Ryde, NSW | Photo credit: G Hose

All other images © Department of Climate Change, Energy, the Environment and Water unless specified otherwise.

Bioassessment of groundwater ecosystems

III.A comparison of eDNA metabarcoding and metagenomics for assessing groundwater communities

Sampling approaches to understand the impacts of coal seam gas and large coal mining development on stygofaunal and microbial assemblages in groundwater

Contents

Executive summary

The analysis of DNA in environmental samples provides an unprecedented ability to explore the diversity and function of ecosystems. This is particularly the case for microbial assemblages, which, prior to this revolution, have been difficult to characterise. Analysis of environmental DNA (eDNA) has been applied across all ecosystem types but has seen relatively little application in groundwaters (Saccò et al. 2022), despite its huge potential to characterise biological communities that are poorly described and understood, such as those in the subsurface.

Given the past challenges in characterising microbial communities, it is not surprising that bacteria and other microorganisms have rarely been considered in environmental impact assessments, even though microbes in groundwater are responsible for important ecosystem services, including removing or degrading contaminants so that groundwater is clean and fit for human use (Griebler et al. 2019). Microbial communities and the services that they provide are at risk from human activities such as large coal mining (LCM) and coal seam gas (CSG) development that can lead to changes in the hydrology and quality of water within aquifers. It is important, therefore, that groundwater microbial communities are considered in impact assessments for such activities, and that robust and sensitive methods for detecting microbial community change are developed.

Stages 1 and 2 of this project compared sampling approaches to characterise the microbes and stygofauna in shallow alluvial and fractured sandstone aquifers, including assessing the suitability of eDNA-based approaches for use in routine environmental monitoring and assessment (Korbel et al. 2022a; Korbel et al. 2023). These studies focused on metabarcoding, which uses small fragments of taxonomically informative DNA shed by organisms in the environment to characterise the diversity and structure of the community in which they exist, and highlighted the utility of eDNA approaches for monitoring and assessing microbes and invertebrates in groundwater. The aim of this study is to extend that work by exploring the more technical aspects of eDNA analysis for characterising the biota in shallow aquifers. Part 1 of this study explores methods to improve the quantitative comparison of eDNA samples by adding 'spikes' during laboratory analysis. Part 2 compares the metabarcoding approach, used in Stages 1 and 2, with metagenome analysis that analyses all DNA in a sample to characterise both the functional genes and taxonomic structure.

The multi-step pipeline for processing eDNA samples includes steps, such as DNA amplification by polymerase chain reaction (PCR) and sequencing, that have potential to bias outcomes because DNA in different samples may amplify or be sequenced differently. The outcome is that similar samples may yield a different number of sequence reads, making quantitative comparisons of sequence data difficult. We spiked our eDNA samples with short, artificial DNA fragments as an internal standard during both the PCR and the sequencing stages of the processing pipeline and used these to standardise the sequence reads in each sample.

The addition of spikes required a large amount of preparation to optimise the spike process and concentration. Rather than making sequence reads more similar, the standardisation of sequence read numbers by the spikes increased divergence among the samples. Without standardisation by spike read number, ordinations showed that samples grouped together based on aquifer type (sample source). The grouping by aquifer became less clear in the standardised samples, and the abundance of sequence reads, influenced by the standardisation, had a strong influence on the pattern of samples in the ordination plot. While some studies promote the use of spikes to improve the quantitative comparison of metabarcode data, there was no clear evidence from this study that spiking aids the ability to quantify sequence abundances in groundwater samples. We do not recommend this approach as part of routine analyses until methods for doing so are further refined.

Analysis of metagenomes compiled from several different gene databases showed differences in the functional assemblages of samples from different aquifer types and from pre- and post-purge bore samples. Overall,

metagenomes derived from different databases all responded to a similar suite of environmental variables. Taxonomic assemblage information derived from the metagenomes also showed separation of samples by aquifer types and pre- and post-purge bore samples, consistent with the patterns derived from metabarcoding of the same samples. Metagenome analysis provides a very large amount of information on the functional genes present within a sample, which is more detailed than can be inferred from metabarcoding analyses. However, although metagenome analysis provides a greater depth of information on microbial functions within aquifers, this comes at considerably higher cost and greater complexity of analysis. Metagenome and metabarcoding data were similar in their ability to discriminate samples based on taxonomic composition. We recommend that metabarcoding is sufficient for the analysis of microbes for the purpose of routine monitoring, and that the benefits of metagenome analysis for monitoring are not yet outweighed by the additional cost and analytical effort. However, it is likely that as the cost and complexity of metagenome analysis decrease, such analysis may be more accessible for routine monitoring and will be a very valuable tool for monitoring and assessment of groundwater ecosystems in the future.

1. Introduction

The Independent Expert Scientific Committee on Unconventional Gas Development and Large Coal Mining Development (IESC) is a statutory body under the *Environment Protection and Biodiversity Conservation Act 1999* (EPBC Act) and provides advice to the Australian Government Environment Minister on priorities for research. One identified priority is improving the understanding of potential risks associated with large coal mining (LCM) and coal seam gas (CSG) developments to groundwater as a water resource and to its associated health. This report is the third stage of a project commissioned by the IESC on the suitability of various sampling methods for completing groundwater biota surveys and biomonitoring within aquifers.

Stages 1 and 2 of this project (Korbel et al. 2022a; Korbel et al. 2023) provided details on the assessment and monitoring methods for microbial assemblages and stygofauna in shallow alluvial (Stage 1) and fractured sandstone (Stage 2) aquifers. The Stage 1 and 2 studies compared numerous sampling strategies (e.g., unpurged versus purged samples, sample volumes) and sampling methods (e.g., nets, bailers, pumps, environmental DNA (eDNA), environmental RNA (eRNA) (Stage 1 only), primers) within alluvial and fractured sandstone aquifers to help inform the IESC of the most practical and cost-effective strategies by which proponents of LCM and CSG developments might accurately characterise the biodiversity within groundwater ecosystems.

Stage 3 of the study builds on the earlier stages and compares the methods for eDNA analysis in groundwaters. Specifically, this project compares the metabarcoding analyses for groundwater communities collected in Stages 1 and 2 with metagenome analysis of those same samples. Further, this project also examines the potential approaches for quantitative comparisons of sequence data arising from eDNA samples. Robust analysis of eDNA and guidance for its use in groundwaters are critical if eDNA is to be more widely applied for groundwater bioassessments.

1.1 Spiking for quantitation of DNA sequence reads

The analysis of DNA residing within the environment (e.g., directly from the organism, shed by the organism or extracted from a matrix such as sediment), termed eDNA, is a rapid, non-invasive and potentially cost-effective tool for characterising biodiversity (Taberlet et al. 2018) and is a potentially powerful tool for monitoring and assessing groundwater ecosystems (Korbel et al. 2022a; Korbel et al. 2023; Saccò et al. 2022). However, a current limitation for biomonitoring using eDNA is that such data are typically qualitative (presence/absence) or semi-quantitative (relative abundance), and reliable, fully quantitative data are rare. This limitation arises because data from each metabarcoding sample contain different numbers of sequence reads. For example, one replicate sample may have 10,000 reads, while another has 20,000. This may be a consequence of several factors, such as differences in the amount of DNA in the sample, taxon-specific amplification biases (i.e., differences in the way individual gene fragments behave under polymerase chain reaction (PCR) conditions), the stochastic nature of PCR, or biases in the analysis pipeline (Kanagawa 2003; Weiss et al. 2017; Cameron et al. 2021), and may not reflect the actual differences in abundances of taxa in the underlying communities (Nichols et al. 2018). There is often a correlation between the number of sequence reads in a sample and the taxa richness of that sample (e.g., Weiss et al. 2017), which has important implications for diversity assessments. However, the number of reads between samples can vary markedly. A common approach to address this is by standardisation, either by a sample attribute (such as sample read number, also known as 'scaling' (Weiss et al. 2017)) or via rarefaction, in which all samples are randomly resampled to a consistent number of reads (also referred to as 'normalisation' (Weiss et al. 2017)). However, rarefaction has some significant artefacts and can result in a substantial loss of data as the read number in every sample is reduced to the lowest common denominator or corrected to inflate samples with lower numbers of reads (McMurdie and Holmes 2014; Cameron et al. 2021).

Spiking DNA is a technique used in metabarcoding studies to enable the quantitation of target DNA sequences following sequencing. This technique involves adding known amounts of an artificial or obscure (unlikely to occur naturally in the sample) DNA fragments, known as spike-in controls or reference sequences, to the environmental sample before DNA extraction and subsequent processing (Smets et al. 2016; Deagle et al. 2018; Tkacz et al. 2018). As each sample receives the same amount of spike, the resulting number of sequence reads should be the same between samples. However, this rarely occurs, because of biases and stochasticity in the amplification and sequencing processes, and the read number of the spiked sequence can be used to correct for biases between samples. The correction is done by dividing the read numbers for each operational taxonomic unit (OTU) by the read number for the spiked sequence in that sample (Ji et al. 2020). While a spike-in sequence can improve the quantitation of eDNA data sequences (Ji et al. 2020), the approach does not correct for bias due to DNA from different species being preferentially amplified (Ji et al. 2020) or weak relationships between DNA concentration in environmental samples and the abundance of organisms in the environment (but see Tillotson et al. 2018; Spear et al. 2021).

The aim of spiking eDNA is to enable within-species quantitation – i.e., being able to conclude that 'species 1 is more abundant in sample A than in sample B' (Ji et al. 2020; Garrido-Sanz et al. 2022). This is important for bioassessment because quantifying and comparing absolute changes in taxa abundance is typically more statistically powerful than comparing changes in relative abundances or presence/absence of taxa. The extraction of reliable species-abundance information from eDNA can also enhance diet and quantitative food web analysis (Thomas et al. 2016; Deagle et al. 2019; Peel et al. 2019) and modelling of species distributions and population dynamics (Carraro et al. 2020; Carraro et al. 2021; Abrego et al. 2021; Rojahn et al. 2021), which may provide additional sensitive metrics for detecting environmental change.

1.2 Metabarcoding and metagenomes

Microbial assemblages form the basis of food webs in most aquifers and play crucial roles in mediating and facilitating biogeochemical reactions that underpin many ecosystem services (e.g., Griebler and Lueders 2009; Sang et al. 2018; Griebler et al. 2019). Groundwater microbes respond rapidly to environmental impacts such as contamination (Tischer et al. 2013) and land use changes (Baily et al. 2011; Korbel et al. 2013) that can lead to reductions in biodiversity and capacity for biogeochemical processing and provision of ecosystem services (Hemme et al. 2015; Korbel et al. 2022b). Activities associated with coal mining and CSG development typically influence groundwater hydrology and the physical and chemical properties of groundwater (Hose et al. 2015). However, the implications of these alterations of groundwater hydrology and groundwater quality for microbial communities, their ecological functions and associated ecosystem services are poorly known.

Environmental impact statements (EISs) rarely consider and assess the risks of coal mining or CSG activities to groundwater microbial assemblages. This is largely because of the currently limited knowledge of groundwater microbial assemblages in Australia, particularly the spatial and temporal variability in their function and composition, and how microbial composition and activity respond to environmental change. Furthermore, there are practical challenges to sampling and characterising microbial assemblages in aquifers as part of environmental impact assessment and routine monitoring.

Recent advances in eDNA technology have yielded rapid and cost-effective approaches for assessing structure and biodiversity of groundwater biota (Saccò et al. 2022). Coupling eDNA techniques with high-throughput sequencing enables the detection of numerous (often hundreds of) different species within a single environmental sample. This approach allows the characterisation of entire ecological communities (eukaryotes and prokaryotes) and greatly increases our ability to comprehensively determine the true biodiversity within ecosystems (Deiner et al. 2017; Ruppert et al. 2019). The ability to detect trace amounts of eDNA (i.e., sensitivity) is constantly improving while costs are decreasing, making eDNA analyses increasingly feasible for routine monitoring and assessment.

The application of eDNA-based approaches for assessing groundwater microbial assemblages in Australia, and elsewhere, is in its infancy (Saccò et al. 2022). However, work so far (e.g., Korbel et al. 2017; Korbel et al. 2022c; Korbel et al. 2023) demonstrates the capacity of these approaches to detect diverse assemblages of Bacteria and Archaea, infer their metabolic functions and identify how these might be impaired by groundwater contamination. Improved understanding of functional responses of groundwater microbial assemblages to environmental change would help refine conceptual models of the pathways by which coal mining and CSG activities could impact groundwater ecosystems.

The functional and taxonomic composition of microbial assemblages can be determined or inferred using two different molecular-based approaches: metagenomics (also known as 'shotgun' sequencing) and metabarcoding. Metabarcoding uses small fragments of taxonomically informative DNA shed by organisms in the environment to characterise the diversity and structure of the community in which they exist. The small fragments are extracted from the sample, copied many times using PCR and sequenced, which generates a list of the gene fragments identified and a count of those fragments. In contrast, metagenomics involves sequencing and analysing the entire genetic material (DNA or RNA) present in a sample, without targeting specific genes or regions. The DNA or RNA in the sample is fragmented ('shotgunned') and sequenced, and the sequenced fragments reconstructed into longer read sequences that provide both functional and taxonomic information (e.g., Hemme et al. 2015). This approach provides a comprehensive view of the organisms and genetic elements present in the community, including both known and unknown species. The main advantage of metagenomics is that it does not require amplification using PCR, which can introduce considerable bias (e.g., Krehenwinkel et al. 2017), thereby allowing composition and function to be quantified and reliably compared between samples. Conversely, the drawbacks of metagenomic methods are the higher sequencing costs, the intensive computational requirements, and the potential underestimation of biodiversity because the genes associated with taxonomic assignment generally make up less than 0.1% of the reads in the case of prokaryotes (Greenfield et al. 2019).

Functional information on microbial communities can also be inferred from the taxonomy of prokaryotes, which is determined using metabarcoding of taxonomically informative gene regions (i.e., 16S rDNA) (e.g., Korbel et al. 2022a,b). Functional inference can be performed using packages such as PICRUSt (Langille et al. 2013), Tax4Fun (Aβhauer et al. 2015) or FAPROTAX (Louca et al. 2016) and is possible because of the relationship between phylogeny and biomolecular function (Chaffron et al. 2010) – i.e., phylogenetically related microbes can typically perform the same functions. The main advantage of the functional inference approach is that it is based on compositional data that are obtained from metabarcoding, which means that the only additional cost is associated with interpretation. However, because it is based on metabarcoding, it has biases related to PCR amplification and is further limited by its dependence on reference genomes (which are not always well known), the region targeted by the primers, and the limited knowledge of the function of a large proportion of the global microbiome. Fortunately the same DNA extract that is used for metabarcoding can also be used for shotgun sequencing, allowing direct comparison of approaches, as has been done in this study. The comparison of these two approaches is one of the main aims of this study.

1.3 Project aims and report structure

1.3.1 Aim and core research questions

This project is Stage 3 of an investigation into approaches for biomonitoring groundwater ecosystems and follows Stage 1 (Korbel et al. 2022a) and Stage 2 (Korbel et al. 2023). This report describes further analyses of samples collected in the Stage 1 and Stage 2 reports. In this study, samples collected from alluvial aquifers (Stage 1) and fractured rock aquifers (Stage 2) and analysed using metabarcoding were re-analysed using metagenomics. The ability of metabarcoding and metagenomics to discriminate differences among sample groups is compared.

A challenge of metabarcoding is the (in)ability to quantitatively compare samples based on the number of sequence reads. In this study, we explore the use of spikes added to samples prior to sequencing to enable quantitative comparisons of sequencing outputs.

The project scope includes four core components, rephrased as research questions:

- Are there differences in the taxonomic and functional composition of biotic assemblages in alluvial and sandstone aquifers as characterised by metagenomics and metabarcoding?
- Are there associations between water quality parameters and microbial function and composition in both aquifers?
- Can spiking be used to enable quantitative comparisons of eDNA data?
- Is it feasible for consultants to collect and process groundwater samples for either shotgun sequencing or metabarcoding approaches to assess potential impacts of coal mining and CSG development on microbial composition and activity?

1.3.2 Report structure

This report is divided into four sections. This first section provides background to the project and its aims. The second section details the sampling regime and provides a summary of the field, laboratory and data analysis methods used in the study (for full details, see the Stage 1 and 2 reports (Korbel et al. 2022a; Korbel et al. 2023)). The third section presents the results and discussion in the context of the research questions. The fourth section provides a summary of project findings and recommendations on their feasibility for routine groundwater monitoring.

2. Field, laboratory and analysis methods

2.1 Project study area and sample sites

Field sampling of sites was undertaken in May 2021 in alluvial aquifers of the Namoi River catchment, New South Wales (NSW), and in February 2022 in the shallow fractured sandstone aquifers of the Hawkesbury-Nepean catchment and Central Coast region, NSW (Figure 1). Detailed descriptions of these study regions and specific site locations are given in Korbel et al. (2022a, 2023). Fifteen monitoring bores were sampled in each aquifer type. All bores were constructed of PVC casings that were completely enclosed except for discrete sections with vertical slots 10 m to 45 m below ground, allowing the entrance of groundwater from the aquifer.

Figure 1. Map showing sampling locations of fractured rock aquifer samples (blue symbols) in the Hawkesbury-Nepean catchment and Central Coast region, and alluvial aquifer samples (green symbols) in the Namoi River catchment

Map source: Google Earth.

2.2 Sample collection and analysis

Field methods and procedures for sample collection are described in detail in Korbel et al. (2022a, 2023). Pre-purge samples were collected from bores using a sterile bailer into a sterile plastic container. After pre-purge samples were collected, 30 L of groundwater was extracted using a motorised inertia pump and a further sample was collected. A further 150 L of groundwater was then pumped (total 180 L) and a third groundwater sample was then collected in a sterile plastic container. Water samples were stored on ice in the dark after collection and were processed within seven hours of collection. Water samples were filtered onto sterile cellulose membrane filters (0.2 µm) and frozen at -

20°C until analysed. The pre-purge and post-purge (180 L) samples were analysed for the comparison of metabarcoding and metagenomics. The samples collected after 30 L was pumped were used for the DNA spiking experiment.

2.3 Laboratory methods

2.3.1 Groundwater quality analyses

Pre- and post-purge water samples were collected and preserved as required. All water quality analyses were conducted at CSIRO (Lucas Heights, NSW) and ALS (Sydney), as detailed in Korbel et al. (2022a, 2023). Physico-chemical water quality properties of pre- and post-purge samples were recorded in the field using a YSI ProPlus multiprobe meter (YSI Inc.).

2.3.2 DNA extraction

Methods for DNA extraction were as described in detail in Korbel et al. (2022a, 2023). DNA extraction was performed on 0.25 g (filter paper and sediment) samples using a DNeasy PowerSoil Pro Kit (QIAGEN) following the manufacturer's protocol with slight modifications. The quality and purity of isolated DNA in all samples were then checked using a spectrophotometer.

2.3.3 Metabarcode analysis (eDNA)

Methods for metabarcode analysis, including PCR conditions, were as described in detail in Korbel et al. (2022a, 2023). DNA was amplified using PCR targeting the 16S rDNA (prokaryote) and the 18S rDNA (eukaryote) genes, using primers and PCR conditions as listed in Appendix 1. Primers and PCR conditions for eDNA.

Following PCR, DNA was pooled at equimolar concentrations (final concentration 50 ng/μL to 60 ng/μL) and purified using AMPure beads (Beckman Coulter Inc., CA, USA). Samples were sequenced by the Ramaciotti Centre, UNSW, using Illumina MiSeq (PE 250) after passing quality assurance checks that included screening DNA quality and quantity.

2.3.4 Metagenomics

DNA extracted from the pre- and post-purge samples was analysed using shotgun sequencing by the Ramaciotti Centre, UNSW, using NovaSeq 6000 S4 with two 150 base pair (sequence read length) lanes after passing quality assurance checks that included screening DNA quality and quantity.

2.3.5 Metabarcode analysis (eDNA) – spiking experiment

Extracted DNA samples were used to investigate a method of spiking synthetic 16S bacterial rDNA gBlocks® Gene Fragments to determine bias in the PCR and sequencing process. Thirty eDNA samples (30 L pump volume from 15 sites for both Stage 1 (Namoi River catchment) and Stage 2 (Hawkesbury River catchment) were used in the experiment. All samples were amplified by PCR with negative (DNA-free H2O) and positive controls (16S rDNA positive control).

To investigate the effectiveness of sample spiking to improve quantification of sequence reads, we spiked the extracted DNA samples with fixed amounts of known synthetic DNA sequences. We used two 16S synthetic bacterial rDNA gBlocks® Gene Fragments as the spikes, selected at random. Synthetic spike sequences were sourced from Integrated DNA Technologies (IDTdna.com). Details of the synthetic sequences are provided in Appendix 1. Primers and PCR conditions for eDNA.

The experimental design included three treatments:

1. Samples spiked with Spike 1: added before PCR to investigate PCR amplification bias
2. Samples spiked with Spike 2: added after PCR to investigate sequencing bias
3. Samples spiked with Spike 1 and Spike 2: to investigate the effects of both PCR and sequencing bias.

Synthetic Spikes 1 and 2 were received as 500 ng of dried product and kept at 4° C until resuspension in sterile TE (Tris + EDTA) buffer (Sigma Aldrich) to achieve a 10 ng/µL solution. These solutions were immediately vortexed and incubated at 50° C for 20 minutes following the manufacturer's protocol. After incubation, the synthetic rDNA solution was vortexed and centrifuged. Initial 1:50 stock dilutions were prepared in sterile tubes with DNA-free water and these, along with the stocks, were immediately stored at -30° C.

Spike 1 was added to individual eDNA samples prior to PCR to enable comparisons of the relative amplification of samples. Spike 1 was diluted to 1:500,000,000; 2 µL of spike was added into 20 µL of the final PCR reaction volume and amplified with the sample by PCR using primers targeting the 16S rDNA gene (see Appendix 1. Primers and PCR conditions for eDNA). Spike 1 was added to achieve a concentration of 4 ng/µL to 5 ng/µL after amplification.

Spike 2 was diluted to 1:2,000,000. Aliquots of Spike 2 were labelled with tags that matched those added to the individual eDNA samples. Aliquots (2 µL) of diluted Spike 2 stock were added into 20 µL final PCR reaction volumes and amplified with 16S rDNA primer barcodes that matched those used to amplify the sample eDNA. The amplified concentrations were 35 ng/µL to 40 ng/µL. The final concentration of Spike 2 added to amplified eDNA was 10 ng of DNA into 10 µL of PCR product containing amplified sample eDNA.

Spike 2 was used to investigate any bias introduced during sequencing. The synthetic sequences have the same properties as the amplicon products; however, as they are artificial, they have no match in sequence databases. PCR products were determined by gel electrophoresis. Samples were checked for concentration, purified, pooled and sequenced as above.

A series of quantitative PCRs (qPCRs) were performed for both gBlocks® Gene Fragments spike fragments prior to use to determine the dilution factor required to achieve a low concentration of amplified synthetic 'spike' DNA. This aimed at preventing the eDNA samples being swamped by Spike 2 when amplified. qPCR conditions followed the 16S rDNA PCR conditions. All dilutions were performed using the 1:50 stock solutions and were prepared in sterile tubes with DNA-free H2O. Dilution factors ranged from 1:1,000 to 1:1,000,000,000 (final PCR dilution factor). qPCRs were performed on the synthetic rDNA alone, and the Spike 1 sequence was added to an eDNA sample to determine if a difference in the concentration could be detected when comparing an eDNA sample and an eDNA sample with the addition of the spike. Qubit fluorometric assays were used to determine the concentration of DNA in the spike samples. The Qubit fluorometer was calibrated before each use, and quantification followed the instrument protocol.

All amplified PCR sample concentrations were measured using the Quant-iT™ PicoGreen™ dsDNA Assay Kit (Thermo Fisher Scientific) and PHERAstar Microplate Reader. Preparation of the reagents was performed using sterile solutions and PicoGreen™ dye was added to each sample in light-sensitive conditions. All samples were analysed with blanks and a standard curve using Lambda DNA.

Spiked samples were pooled in equimolar concentration. The pooled tube of samples was purified using AMPure XP Reagent (Beckman Coulter), which uses paramagnetic beads to selectively bind DNA. The protocol also incorporates a wash procedure to remove excess primers, nucleotides, salts and enzymes. Following purification, Qubit analysis was used to determine the final concentration of the pooled and purified sample. This was then pooled in equimolar concentrations and sequenced by Ramaciotti Centre, UNSW, using Illumina MiSeq (2 x 250 base pair paired end).

2.3.6 Bioinformatics

All metabarcode sequence data were processed using custom software designed by Paul Greenfield (CSIRO/Macquarie University) (see Korbel et al. (2017) and Sutcliffe et al. (2017) for details). For full details of bioinformatics, see Appendix 2. 16S rDNA amplicon bioinformatic methods. Additional cleaning of datasets included the removal of samples that had a total sequence count of less than 5,000 reads, and adjustments of counts in samples based on the number of positive controls found in individual samples.

For 16S rDNA (prokaryote) compositional data, inferred functional profile data were obtained using the FAPROTAX program (Louca et al. 2016), which assigns bacterial OTUs to functional groups.

Metagenome data were analysed using KBase (https://www.kbase.us), developed by the US Department of Energy (Arkin et al. 2018). Within KBase, assemblies were done using SPAdes (3.15.5) (Bankevich et al. 2012; Prjibelski et al. 2020) using the metagenomic assembly option, and the annotations were done using Prokka (Seemann 2014). After quality assurance to remove erroneous and poor-quality reads, metagenomes were reconstructed. From the reconstructed metagenomes, functional genes were identified using the Clusters of Orthologous Genes (COG), Gene Ontology (GO) and Enzyme Commission (EC) databases separately, as well as a hybridised compilation of annotated genes across all databases. The taxonomy of microbes in the assemblages was identified by targeting the 16S rDNA and 18s rDNA genes using Kelpie 2.0 (Greenfield et al. 2019).

2.3.7 Data analysis

Differences in assemblage data (stygofauna, eDNA assemblages) were visualised using non-metric multidimensional scaling (nMDS). Sequence reads for individual operational taxonomic units from 16S rDNA metabarcoding (OTUs) and gene sequence counts from metagenome analysis were transformed to relative abundance and square root transformed (Hellinger transformation). Read numbers for 18S rDNA sequences from metabarcode and metagenome analyses were presence/absence transformed prior to analysis. Similarity among samples was estimated using the Bray-Curtis similarity.

Differences among functional and compositional assemblages were analysed using permutational multivariate analysis of variance (PERMANOVA) (Anderson 2001) with a nested design (sampling bore as a random factor nested within aquifer type and pre-purge/post-purge sampling as a fixed factor). Analysis of similarity (ANOSIM) (Clarke and Green 1988) was also used to compare differences between sample groups.

Relationships between environmental variables (including water chemistry) and biotic communities were modelled using distance-based linear models (DistLM) (Anderson et al. 2008). Environmental data were normalised before analysis, and strongly correlated ($r>0.90$) variables were removed prior to analysis, based on draftsman plots (Clarke and Ainsworth 1993). PRIMER-e version 6.1.11 (PRIMER-e Ltd, Plymouth, UK) was used to complete all multivariate analyses, with univariate analyses completed in Minitab version 17 (Minitab Inc., Pennsylvania, USA). The significance level (α) for univariate and multivariate inferential tests was set at 0.05.

The effects of standardising sequence abundances by the spike read numbers were analysed by comparing the standardised and unstandardised data. 'Unstandardised' sequence assemblage data were transformed using the Hellinger transformation, in which data were converted to relative abundance (dividing by sample read number) and applying square root transformation. This transformation uses the square root transformation to account for the additivity of the relative abundance-transformed sequence numbers (Legendre and Legendre 2012). Taxa contributing less than 1% to the abundance of any one sample in the dataset were excluded (Chariton et al. 2015). In 'spike-standardised' data, sequence read numbers for a taxon were divided by the spike sequence read number in that sample. Samples were square root transformed prior to calculating similarities to down-weight the importance of highly abundant taxa (Legendre and Legendre 2012). Differences between unstandardised and standardised data were visualised using nMDS and compared using PERMANOVA.

All data are publicly available on completion of this study via the Macquarie University Research Data Repository (metagenome DOI: 10.25949/22825832; spiking experiment 16S rDNA DOI: 10.25949/22825871).

3. Results and discussion

3.1 Metabarcode spiking experiment

3.1.1 Detection of spikes in samples

The read number for Spike 1 (added to the raw DNA extracts prior to PCR) was less than 25 reads in 29 of the 60 samples in which it was detected (i.e., in the Spike 1 and Spike 1 + 2 treatments). This modest rate of detection suggests that a slightly higher concentration of the spike DNA could be added in future trials. Of the samples to which Spike 2 was added (post-PCR), five had less than 10 reads and were not included in the analysis. All others had more than 100 reads of Spike 2. Overall, there were sufficient reads for all spikes in the three spiked samples from 11 of the 30 samples analysed. The amplicons of the synthetic spikes (1 and 2) were slightly longer (280 base pair) compared to the 16S rDNA amplicons.

Read numbers for Spike 1 were below 10,000 reads and mostly less than 4,000 (Figure 2a). There was a significant negative correlation between Spike 1 read number and total read number of a sample (excluding the spike) ($r=-0.59$, $p<0.001$) (Figure 2a). Read numbers for Spike 2 were typically below 5,000 reads, with the exception of one sample (bore 30298) that had over 26,000 (Figure 2b). With the outlying value removed (Figure 2c), there was a significant negative correlation between Spike 2 read number and total read number of a sample (excluding the spike) ($r=-0.55$, $p<0.001$).

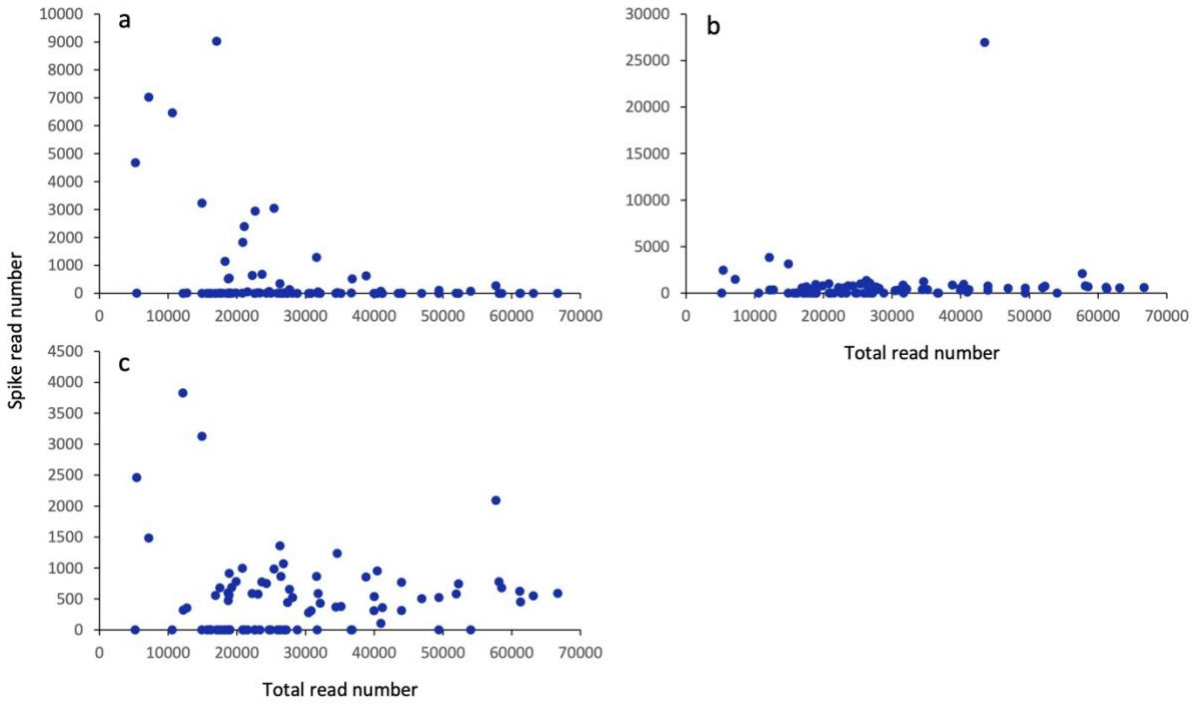

Figure 2. Plot of spike read number against total read number (excluding spike reads) for a) pre-PCR Spike 1, b) pre-sequencing Spike 2 and c) pre-sequencing Spike 2 with outlying value (spike read = 26,936) removed

3.1.2 Spike 1 (pre-PCR)

The ordination of the microbial assemblages standardised by total read number showed a clear separation of sites by aquifer type (Figure 3a). A similar separation was evident among the data standardised by Spike 1 read number, with the exception of bore 36567, which appeared more like samples from the fractured rock than the other alluvial aquifer samples (Figure 3b). Alluvial samples had, on average, a greater total read number than sandstone aquifer samples. Bore 36567 had the highest read number of Spike 1 (9,036 reads). The standardisation by this relatively large number effectively reduced the read numbers for all taxa in that sample (relative to other samples), while lower spike read numbers of other samples effectively 'increased' the abundance of taxa in the samples from the sandstone aquifer, making those samples more similar.

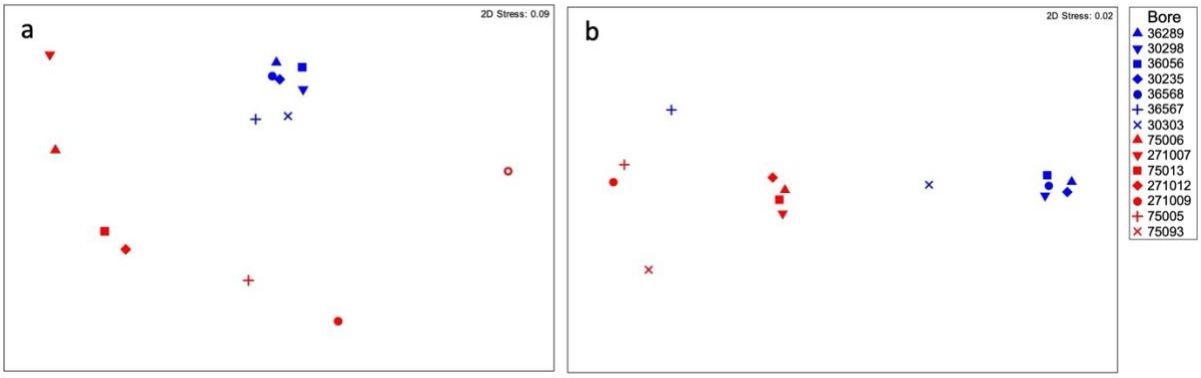

Figure 3. nMDS ordination of 16S rDNA metabarcoding assemblages of groundwater microbial assemblages from bores in alluvial (blue symbols) and fractured rock (red symbols) aquifers: a) relative abundance based on sample read number; b) abundance standardised based on read number of a pre-PCR spike (Spike 1)

There was only a weak negative correlation (r=-0.51) between the total sequence read number and that for Spike 1 (Figure 4). Samples from the alluvial aquifer that formed the cluster to the right of Figure 3b all had relatively low Spike 1 read numbers (25 to 110) and standardisation values (Spike 1 reads/total reads = 0.0009 to 0.002), which may explain why those samples remain relatively similar to each other after standardisation by Spike 1 read number.

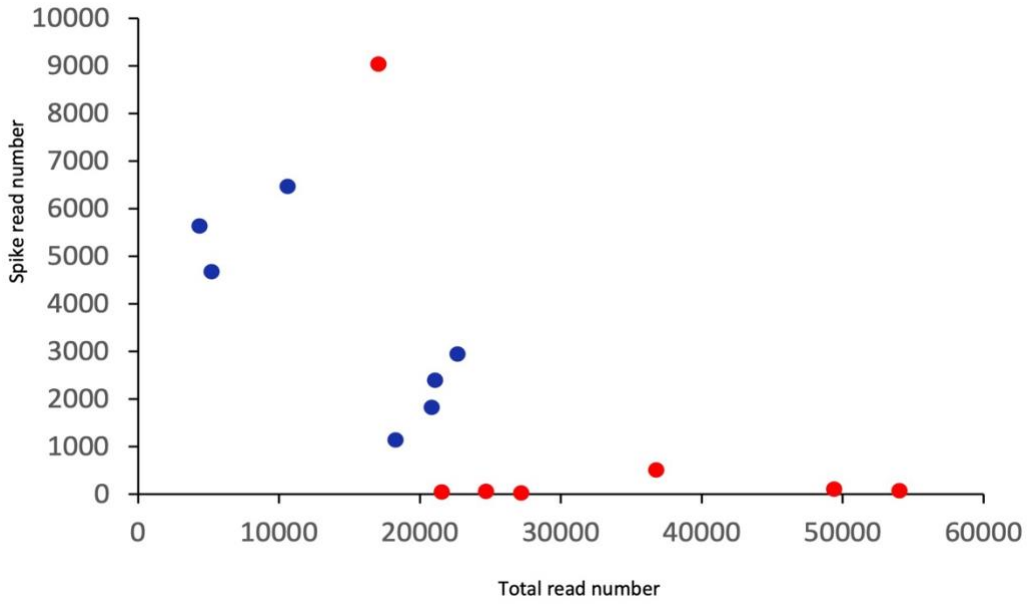

Figure 4. Plot of total read number (excluding spike reads) against Spike 1 read number in eDNA samples of microbial 16S rDNA assemblages from alluvial (blue symbols) and fractured rock (red symbols) aquifers

Despite the apparent shift in the similarities among some samples in the nMDS (Figure 3), there was a significant correlation (RELATE, r=0.445, p=0.001) between the Bray-Curtis similarity matrices that underpin those ordinations.

3.1.3 Spike 2 (post-PCR)

When standardised by total read number, there was a clear separation of samples by catchment type (Figure 5a). When standardised by Spike 2 read number, the separation by aquifer type was evident within the main cluster of samples to the right of Figure 5b, but several samples from both aquifer types formed a separate cluster to the left. These five samples (bores 271009, 75093, 75005, 36567 and 36056) were those with the highest Spike 2 read number (Figure 6) and in which the ratio of Spike 2 reads to total reads was greatest (>0.03). Consequently, the abundance of taxa in those samples after standardisation by the Spike 2 read number was one to two orders of magnitude lower than in other samples.

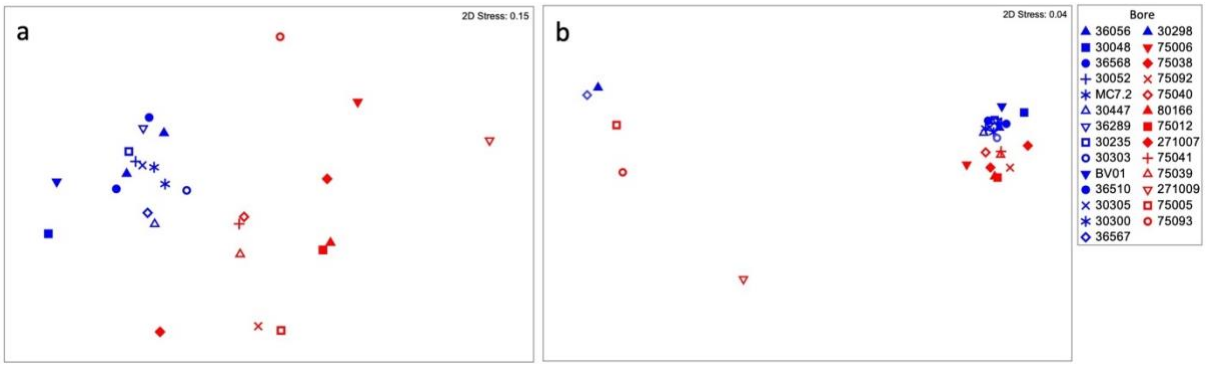

Figure 5. nMDS ordination of 16S rDNA metabarcoding assemblages of groundwater microbial assemblages from bores in alluvial (blue symbols) and fractured rock (red symbols) aquifers: a) relative abundance based on sample read number; b) abundance standardised based on read number of a post-PCR spike (Spike 2)

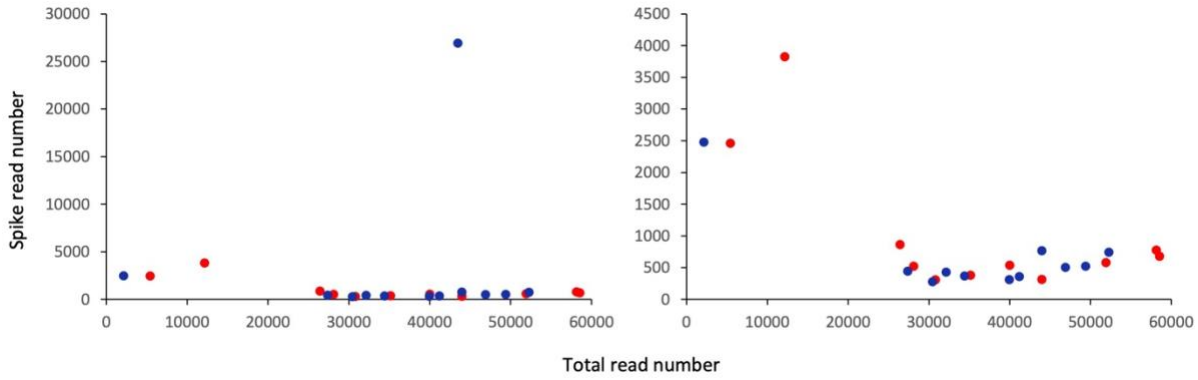

Figure 6. Plot of total read number (excluding spike reads) against Spike 2 read number in eDNA samples of microbial 16S rDNA assemblages from alluvial (blue symbols) and fractured rock (red symbols) aquifers, with a) all samples and b) outlier removed

Despite the apparent shift in the similarities among the samples in the nMDS (Figure 5), there was a significant correlation (RELATE, r=0.604, p=0.001) between the similarity matrices that underpin those ordinations. This correlation suggests that there was little impact of the standardisation method on the similarity among the samples.

3.1.4 Spike 1 + 2 (pre- and post-PCR)

The ordination of all samples standardised by total read number shows a clear pattern of separation among samples based on aquifer type (Figure 7a), with the exception of the sample from bore 36567, which was separated from other alluvial aquifer samples in the ordination plot. Bores 36567 and 271012 group together in the ordination plot, reflecting their relatively low total sequence read numbers. Standardising the taxa read numbers in each sample by the Spike 1 and then the Spike 2 read numbers effectively reduced total read numbers for some samples by between one and four orders of magnitude more than the standardisation by total read number. This effectively increased the variability between the samples to such an extent that the sample abundance becomes the driving factor in the nMDS. In Figure 7b, the sample from bore 36567 was again remote from other alluvial aquifer samples and grouped close to the sample from bore 271012. The distribution of samples along the horizontal axis in Figure 7b reflects the rank order of denominator (Spike 1 x Spike 2 read abundance) used in the standardisation (shown on the y axis in Figure 8).

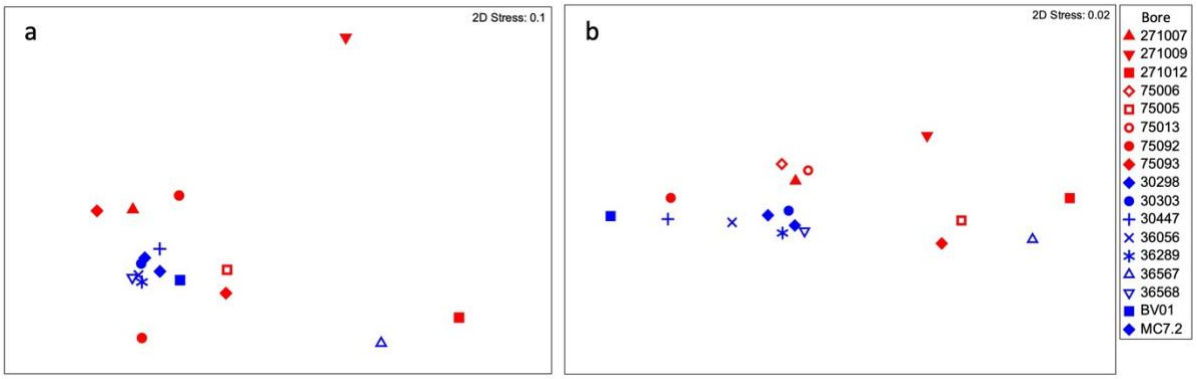

Figure 7. nMDS ordination of 16S rDNA metabarcoding assemblages of groundwater microbial assemblages from bores in alluvial (blue symbols) and fractured rock (red symbols) aquifers: a) relative abundance based on sample read number with all samples; b) abundance standardised based on read number of a pre-PCR spike (Spike 1) and then a post-PCR spike (Spike 2) with all samples

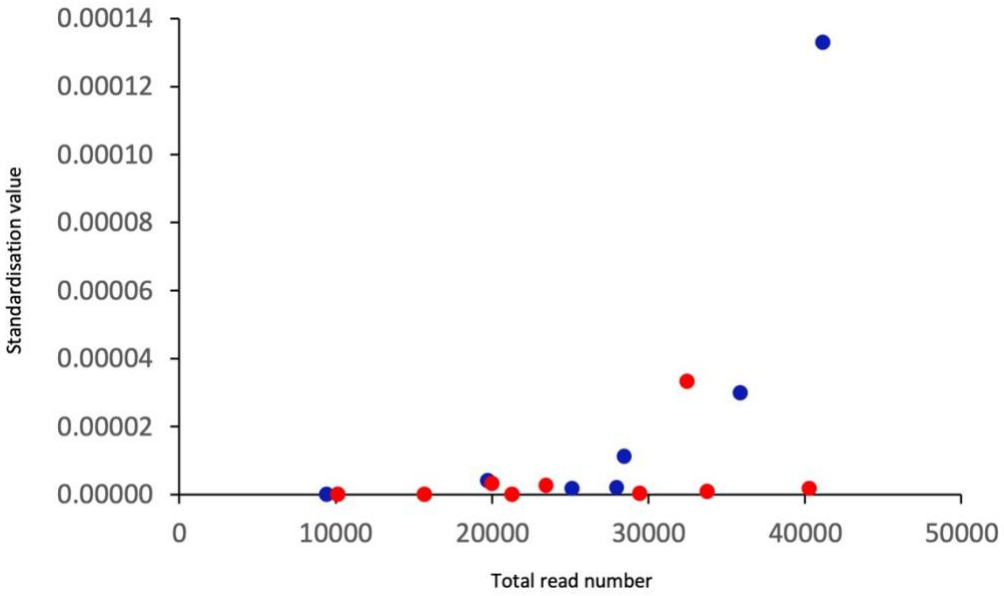

Figure 8. Plot of total read number (excluding spike reads) against Spike 1 + 2 standardisation value (1/Spike 1 read number x 1/Spike 2 read number) applied to eDNA samples of microbial 16S rDNA assemblages from alluvial (blue symbols) and fractured rock (red symbols) aquifers

Given the clear shift in the ordinations (Figure 7), it is not surprising that there was a change in the pattern of similarity among samples, such that there was no significant correlation (RELATE, r=0.199, p=0.100) between the rank similarity matrices that underpin those ordinations.

3.1.5 Comparison of spiking treatments

There were 11 sites from which samples from all spiking treatments met quality assurance requirements. Among those samples, there was a significant difference in the assemblages between catchments (p=0.036) and data treatments (p=0.001), and the differences between catchments were not consistent across data treatments, leading to a significant interaction term (p=0.025). The ordination below (Figure 9) shows the separation of samples by data treatment along the x axis and catchment along the y axis. Partitioning the variance of the PERMANOVA indicated

a large amount of unexplained variance among samples. Spiking method accounted for around 25% of the variation in the assemblage data.

As in the ordinations (Figure 3, Figure 5 and Figure 7), the variation among spike-standardised samples was driven primarily by the magnitude of the standardisation; samples in Figure 9 are distributed along the horizontal axis (from left to right) in the rank order of the magnitude of the standardisation (i.e., samples to the left in Figure 9 are those with the smallest standardisation factors).

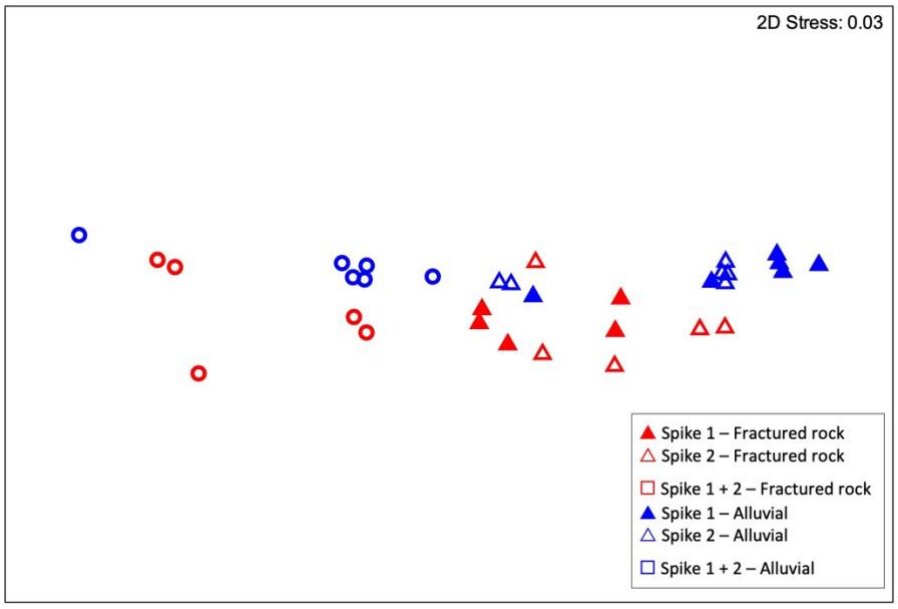

Figure 9. nMDS ordination of 16S rDNA metabarcoding assemblages of groundwater microbial assemblages in alluvial and fractured rock aquifers

Read abundances of assemblages were standardised based on spiked sequence abundances. Spike 1 = pre-PCR spike, Spike 2 = post-PCR spike, Spike 1 + 2 = both Spike 1 and Spike 2.

RELATE analysis, which compares the similarities between all samples with and without standardisation using the spikes, was significant ($r=0.140$, $p=0.039$), suggesting that the spike-standardised and read-total-standardised data showed similar patterns among samples. Although the standardisation had a strong impact on the similarity between samples, samples were more similar to those from the same aquifer type and spike treatment than to those of different aquifer types and spike treatments – i.e., were more similar to samples within the same group than between groups.

The standardisations based on the spike read number of Spike 1 or Spike 2 were similar, and typically only one or two orders of magnitude different to the standardisation by total read number. As a result, samples primarily clustered by aquifer type in all ordinations and rank similarities among samples were retained, such that RELATE analyses between datasets standardised by total read number and either Spike 1 or Spike 2 remained significant ($p<0.05$). However, the standardisation by Spike 1 and Spike 2 changed taxa read numbers by up to several orders of magnitude.

It remains unclear whether a single spike can effectively account for PCR or sequencing bias, or indeed has any advantages over using relative abundances based on total read counts. There are numerous steps in the metabarcoding pipeline, from DNA extraction to bioinformatics, that preferentially measure some taxa over others within a sample (Hugerth and Andersson 2017; Pollock et al. 2018; McLaren et al. 2019). PCR amplification efficiency is related to G+C content of the amplicon (Suzuki and Giovanni 1996; Aird et al. 2011; Mallona et al. 2011), suggesting that (1) amplification efficiency will vary between taxa, and (2) the pre-PCR Spike 1 will only reflect

amplification efficiency of taxa with similar G+C content and therefore the pre-PCR spike sequence will not reflect the bias of all sequences in a sample or library. Consequently, the standardisation by a pre-PCR spike may be introducing additional 'noise' or error into the analysis.

Illumina sequencing has inherent biases such that not all amplicons are sequenced with the same efficiency (Ross et al. 2013). As for PCR bias, sequencer bias is also related to G+C content of the amplicon (Ross et al. 2013). Illumina platforms may also introduce biases when sequencing DNA libraries with low gene diversity (Krueger et al. 2011; Lear et al. 2018), which is a common attribute of groundwater microbial communities (Griebler and Lueders 2009). Together, these also suggest that standardisation by a single sequence may not adequately correct the sequencing bias, and that the sequences used for this purpose should be chosen carefully. The artificial sequences used as spikes in this study were randomly generated. Future studies may wish to use multiple spikes, perhaps with high and low G+C content, which could be used to correct specific sequences based on G+C content.

It is evident from the ordinations that the standardisation using spikes changes the similarities among samples. In the ordinations presented above, when data were standardised by total read number, the primary division among samples was based on aquifer types, such that samples from the same aquifer type mostly clustered together. In ordinations based on data standardised by spikes, the distribution of samples in ordination space often reflected the magnitude of the standardisation, suggesting that the abundances of taxa had a greater influence on the between-sample similarities in those ordinations, leading to a breakdown in the sample groupings based on aquifer type. This may be contrary to expectation, because the purpose of the spike is to remove issues associated with bias within the pipeline and restore parity of sequence read numbers between samples. While it is difficult to know the true relationships between samples, the breakdown of the expected patterns of similarity among samples from the same aquifer suggests that the desired result is not being achieved. Lou et al. (2023) suggest that using spikes to address the noise or bias in the processing pipeline is enough to improve the quantification and analysis of eDNA data. Our results suggest that this might not be the case. However, it is currently difficult to know if the standardisation by spike read numbers actually improves the representativeness of the eDNA data of the microbial community. There remain a number of unknown and unquantified biases or sources of error in the entire eDNA analysis approach, including the differential amplification and sequencing of species (Ji et al. 2020), on which standardising by spike number may have little effect, or indeed magnify.

A challenge with the uneven read numbers between samples in sequence data is that low read samples typically have lower diversity. While standardisation can increase read number, it does not proportionally increase the diversity of the sample, so the issue of how representative a sample is of diversity remains. In terms of assessing groundwater ecosystems in which α-diversity is typically low, reliably assessing diversity is perhaps a more important issue than reliably assessing abundance. This may be exacerbated by spiking if the spike sequence 'swamps' the sample and reduces the depth of sequencing of other amplicons in the sample. Efforts to improve eDNA-based analyses for environmental impact assessment should, in the first instance, address the uncertainties of the methods to measure diversity.

Luo et al. (2023) argue that DNA spikes do enable effective cross-species quantification, and that this is sufficient to address many of the key challenges currently associated with eDNA metabarcoding. However, while there is sound reasoning for including a spike in metabarcode samples, the relative benefits of the spiking process are unclear because of the many remaining uncertainties. The spiking process is time-consuming and provides a further potential source of error and contamination such that the current benefits do not outweigh the additional costs and effort required to spike eDNA samples.

Our results from spiking samples indicate that …

Spikes added to samples prior to PCR require a large amount of work to estimate suitable concentration so as to not 'swamp' the sample DNA.

The concentrations of DNA to use in spikes added after PCR were easier to determine and yielded higher numbers of sequences than those added prior to PCR.

At this stage, there is no clear indication that adding spikes aids the ability to quantify total or relative abundances of microbes within environmental DNA samples.

3.2 Metagenomics

Metagenome sequences were obtained for 34 of the 60 samples submitted for sequencing. Overall, there were fewer metagenomes produced for post-purge samples compared to pre-purge samples, which reflects the relatively low concentration of DNA in those samples.

3.2.1 All identified genes/products

A total of 10,900 functional genes were identified from the metagenomes across all samples. Over 45% (4,962) of the genes were present in all samples. Samples contained between 7,294 and 9,226 genes.

There was a significant difference in the metagenomes by aquifer type ($p=0.036$), as evident in Figure 10. There was also a significant difference between pre- and post-purge samples ($p=0.006$). The significant difference between pre- and post-purge samples is not clearly evident in Figure 10 because the analysis is based on differences between samples from within the same bore, so although significant, it is not apparent as a clustering among groups of pre- and post-purge samples in the nMDS (cf. Figure 3). There was no significant difference between bores ($p=0.166$), and the catchment x pre-/post-purge interaction was not significant ($p=0.279$). When examined collectively (sequential test), four variables (temperature, pH, total nitrogen (TN) and dissolved oxygen (DO)) were found to explain a significant proportion (25.2%) of the total variation in the functional genes derived from the metagenomic data (Table 1).

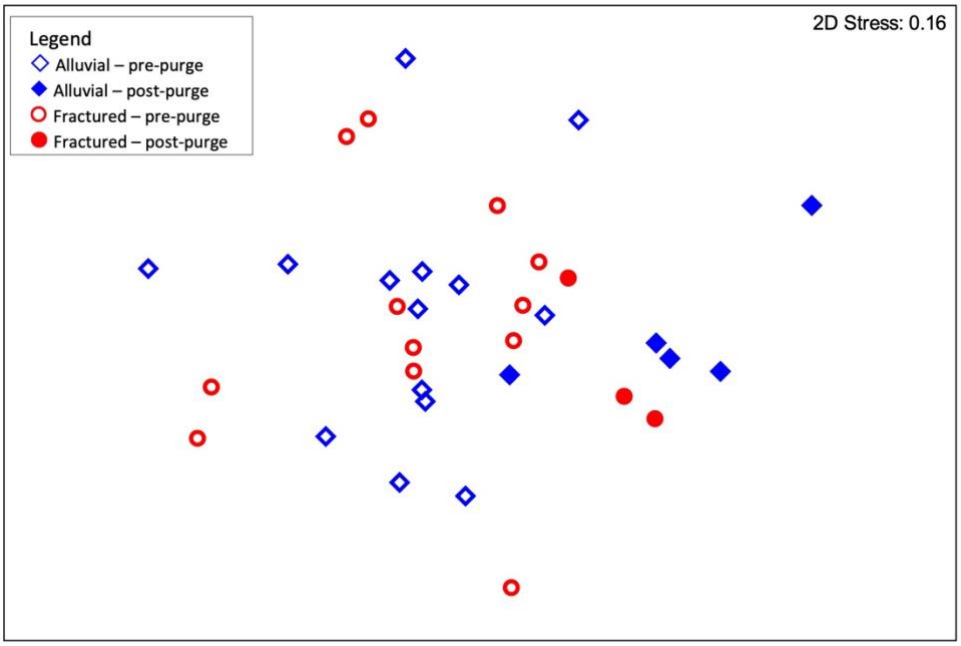

Figure 10. nMDS ordination of 34 metagenomes based on functional genes compiled from available databases

Samples collected from pre- and post-purge groundwater samples in sandstone and alluvial aquifers.

Samples with metagenomes annotated using all databases combined were significantly correlated with sulfate, DOC, DO and TN concentrations, as well as pH and temperature as independent variables (Table 1). The stepwise DistLM model included sulfate and dissolved organic carbon (DOC) concentrations, pH and temperature as variables contributing significantly to explaining 35% of the variation in the metagenomes between samples (Table 1).

Table 1. Proportion of variation (r^2) in functional gene assemblages based on metagenome data using all available databases explained by environmental variables individually and cumulatively in stepwise selection

Variable	Individual r^2	Individual p	Cumulative r^2	Cumulative Cumulative r^2	Cumulative p
Temperature	0.071	**0.005**	0.071	0.071	**0.007**
pH	0.064	**0.015**	0.065	0.137	**0.008**
Total nitrogen	0.064	**0.010**	0.064	0.200	**0.008**
Dissolved oxygen	0.063	**0.017**	0.052	0.252	**0.024**
Water level	0.052	**0.048**			
Mean slot depth	0.051	**0.043**			
Electrical conductivity	0.043	0.125			
Sulfate	0.042	0.133			

Variable	Individual r^2	Individual p	Cumulative r^2	Cumulative Cumulative r^2	Cumulative p
Total phosphorus	0.040	0.166			
Oxidation-reduction potential	0.034	0.351			
Dissolved organic carbon	0.032	0.458			

Values in bold were significant (p<0.05).

3.2.2 Genes identified in the COG database

In total, 2,748 genes were identified from the COG database. More than 70% (1,980) of these genes were present in all samples. There was a significant difference in the metagenomes by aquifer type (p=0.041) and between pre- and post-purge samples (p=0.013) (Figure 11). There was no significant difference between bores (p=0.321), and the catchment x pre-/post-purge interaction was not significant (p=0.324).

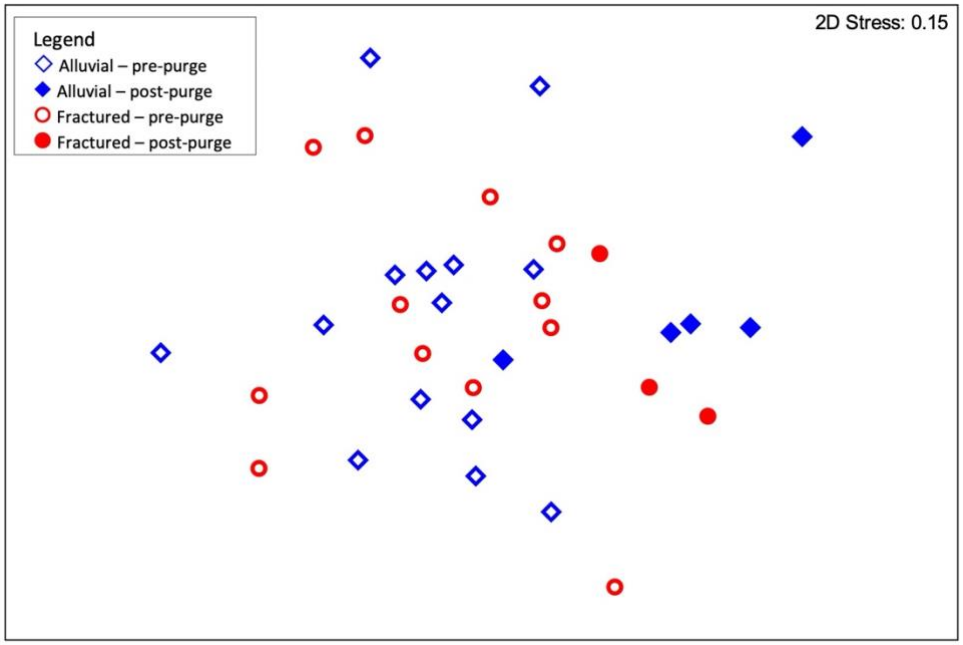

Figure 11. nMDS ordination of 34 metagenomes based on functional genes compiled from the Cluster of Orthologous Genes database

Samples collected from pre- and post-purge groundwater samples in sandstone and alluvial aquifers.

Samples with genomes annotated using the COG database were significantly correlated with sulfate, DOC, TN and DO concentrations, as well as pH and temperature when tested individually (Table 2). Stepwise DistLM analysis included sulfate concentration, pH, DOC concentration and temperature as the variables that accounted for a unique and significant portion of the variation in metagenomes between samples. Together these variables explained 37.3% of the variation in the metagenomes between samples (Table 2).

Table 2. Proportion of variation (r^2) in functional gene assemblages based on metagenome data using functions identified in the Cluster of Orthologous Genes database explained by environmental variables individually and in stepwise selection

Variable	Individual r^2	Individual p	Cumulative r^2	Cumulative Cumulative r^2	Cumulative p
Sulfate	0.103	**0.001**	0.103	0.103	**0.002**
pH	0.082	**0.013**	0.096	0.199	**0.002**
Dissolved organic carbon	0.098	**0.003**	0.096	0.295	**0.001**
Temperature	0.088	**0.007**	0.077	0.373	**0.002**
Total nitrogen	0.085	**0.007**			
Dissolved oxygen	0.080	**0.014**			
Water level	0.053	0.129			
Electrical conductivity	0.052	0.116			
Total phosphorus	0.051	0.126			
Mean slot depth	0.049	0.158			
Oxidation-reduction potential	0.040	0.372			

Values in bold were significant ($p<0.05$).

3.2.1 Genes identified in the GO functional database

A total of 3,256 genes were identified in the samples that were present in the GO functional database. Over 75% of the genes (2,472) were present in all samples. The number of functional genes identified in samples ranged from 2,858 to 3,109. Samples were separated by aquifer type, but not clearly by pre- and post-purge (Figure 12). Nevertheless, gene composition of samples differed significantly with catchment type ($p=0.031$) and sample ($p=0.013$), but not between bores ($p=0.168$); the catchment x pre-/post-purge interaction was also not significant ($p=0.253$). Stepwise DistLM analysis found that four variables (TN, temperature, pH and DO) collectively explained 26.3% of the variation in the GO-derived functional assemblages (Table 3).

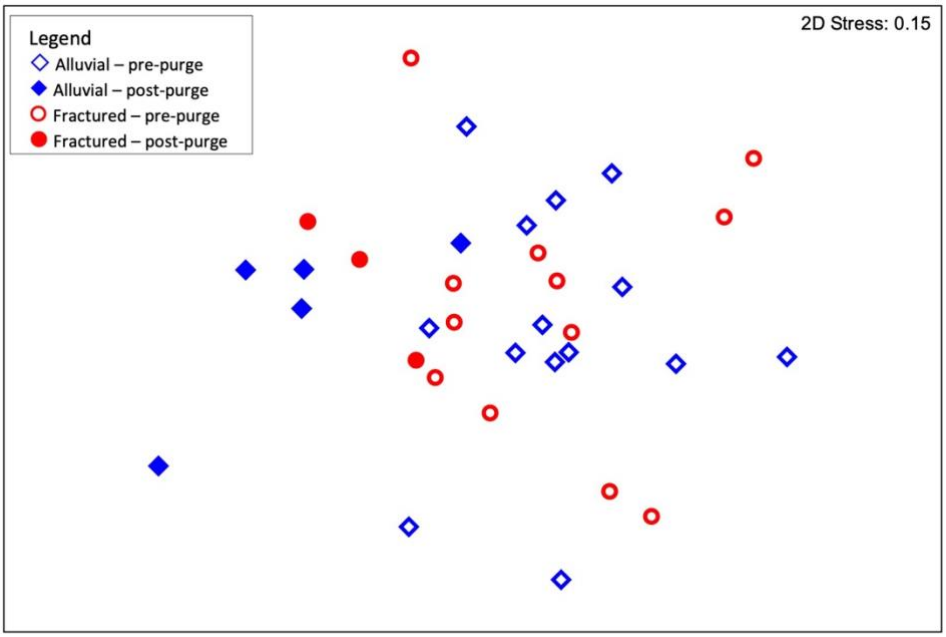

Figure 12. nMDS ordination of 34 metagenomes based on functional genes compiled from the Gene Ontology database

Samples collected from pre- and post-purge groundwater samples in sandstone and alluvial aquifers.

Table 3. Proportion of variation (r^2) in functional gene assemblages based on metagenome data using functions identified in the Gene Ontology database explained by environmental variables individually and in stepwise selection

Variable	Individual r^2	Individual p	Cumulative r^2	Cumulative Cumulative r^2	Cumulative p
Total nitrogen	0.071	**0.005**	0.071	0.071	**0.008**
Temperature	0.069	**0.009**	0.071	0.141	**0.010**
pH	0.065	**0.017**	0.065	0.206	**0.007**
Dissolved oxygen	0.067	**0.009**	0.057	0.263	**0.019**
Water level	0.052	0.075			
Mean slot depth	0.046	0.108			
Sulfate	0.044	0.140			
Electrical conductivity	0.041	0.179			
Total phosphorus	0.039	0.205			
Oxidation-reduction potential	0.036	0.332			
Dissolved organic carbon	0.033	0.401			

Values in bold were significant ($p<0.05$).

3.2.2 Genes identified in the EC database

Across all samples, 2,530 genes were identified from the EC database. Approximately 66% of genes (1,670) were present in all samples. The number of functional genes per sample, as identified from the EC database, varied from 2,084 to 2,335.

The nMDS ordination shows a clear separation by aquifer type, but not between pre- and post-purge samples (Figure 13). Nevertheless, differences in gene composition of samples between both aquifer type and pre- and post-purge collection were significant (p=0.040 and p=0.008, respectively). There was no significant difference between bores (p=0.226), and the aquifer type x pre-/post-purge interaction was not significant (p=0.250). Stepwise DistLM analysis found that 25.2% of the variation in the EC-derived functional metagenomes was explained by TN, temperature, pH and DO (Table 4).

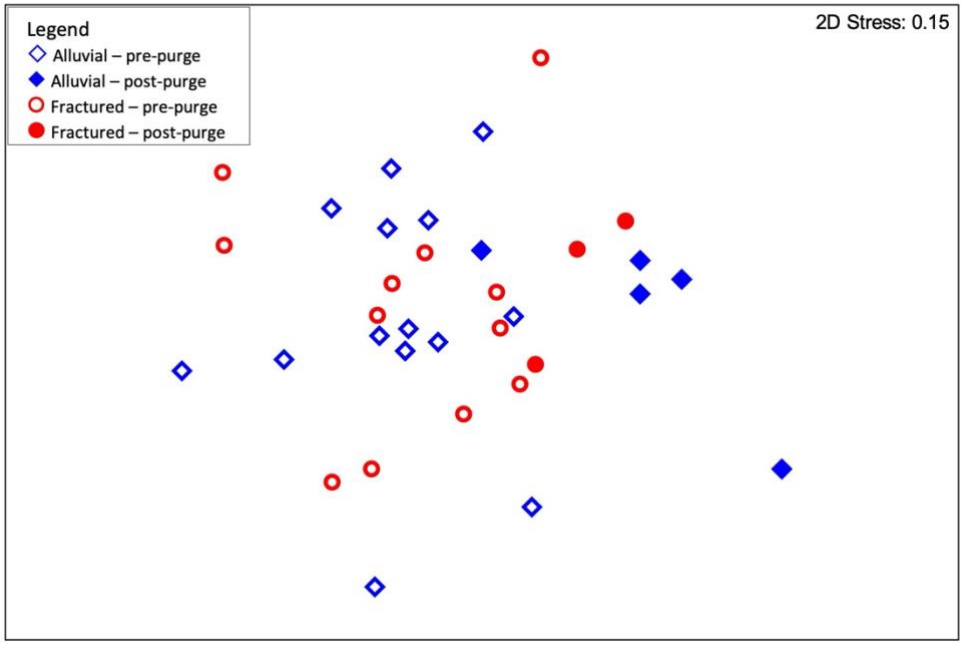

Figure 13. nMDS ordination of 34 metagenomes based on functional genes compiled from the Enzyme Commission database

Samples collected from pre- and post-purge groundwater samples in sandstone and alluvial aquifers.

Table 4. Proportion of variation (r^2) in functional gene assemblages based on metagenome data using functions identified in the Enzyme Commission database explained by environmental variables individually and in stepwise selection

Variable	Individual r^2	Individual p	Cumulative r^2	Cumulative Cumulative r^2	Cumulative p
Total nitrogen	0.067	**0.018**	0.067	0.067	**0.007**
Temperature	0.066	**0.009**	0.068	0.135	**0.009**
pH	0.063	**0.021**	0.064	0.199	**0.008**
Dissolved oxygen	0.064	**0.015**	0.053	0.252	**0.026**
Water level	0.050	0.071			

Variable	Individual r^2	Individual p	Cumulative r^2	Cumulative Cumulative r^2	Cumulative p
Mean slot depth	0.048	0.077			
Sulfate	0.041	0.158			
Electrical conductivity	0.040	0.180			
Oxidation-reduction potential	0.037	0.310			
Total phosphorus	0.036	0.237			
Dissolved organic carbon	0.033	0.428			

Values in bold were significant (p<0.05).

3.2.3 16S rDNA (prokaryote) simulated metabarcoding (Kelpie)

There was a significant difference in the simulated 16S rDNA metabarcode assemblages compiled using Kelpie 2.0 by aquifer type (p=0.001) and between pre- and post-purge samples (p=0.011) (Figure 14). There was also a significant difference between bores (p=0.007); however, the catchment x pre-/post-purge interaction was not significant (p=0.079).

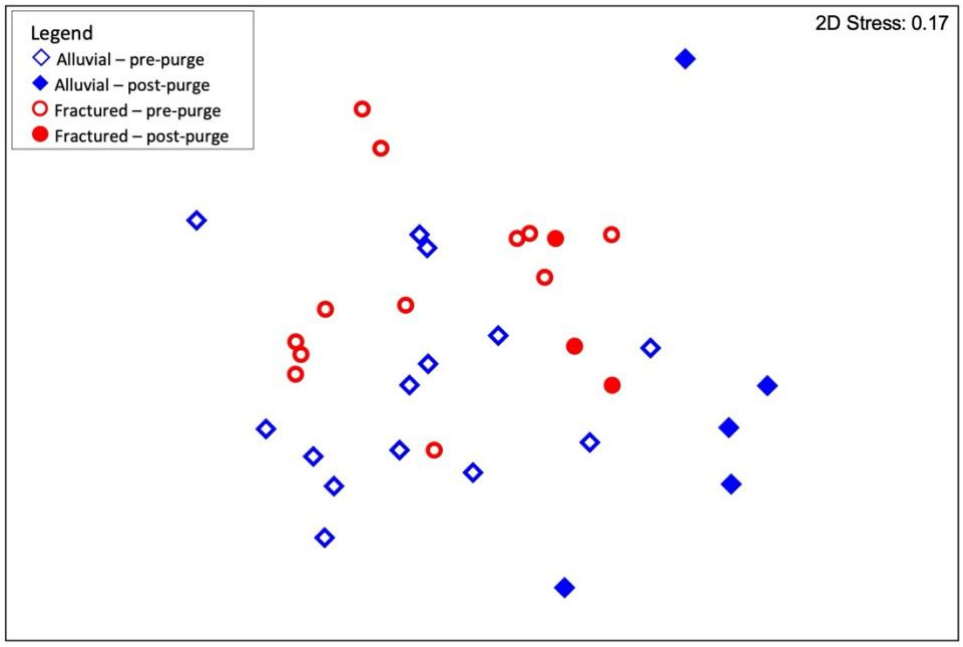

Figure 14. nMDS ordination of prokaryote assemblages in groundwater based on 16S rDNA genes identified within metagenome data using Kelpie

Samples represent pre- and post-purge groundwater samples in sandstone and alluvial aquifers.

Simulated 16S rDNA metabarcode assemblages based on Kelpie were significantly correlated with all variables except total phosphorus, oxygen-reduction potential and water level when tested individually (Table 5). When added sequentially in a stepwise model, only sulfate, DOC, pH and temperature contributed significantly to explaining 33.8% of the variability in the 16S metabarcode assemblages (Table 5).

Table 5. Proportion of variation (r^2) in prokaryote assemblages in groundwater based on 16S rDNA genes identified within metagenome data explained by environmental variables individually and in stepwise selection

Variable	Individual r^2	Individual p	Cumulative r^2	Cumulative Cumulative r^2	Cumulative p
Sulfate	0.097	**0.001**	0.097	0.097	**0.001**
Dissolved organic carbon	0.095	**0.001**	0.097	0.194	**0.001**
pH	0.090	**0.006**	0.087	0.281	**0.003**
Temperature	0.077	**0.007**	0.057	0.338	**0.020**
Electrical conductivity	0.090	**0.006**			
Total nitrogen	0.088	**0.002**			
Dissolved oxygen	0.084	**0.007**			
Mean slot depth	0.063	**0.043**			
Total phosphorus	0.051	0.135			
Oxidation-reduction potential	0.048	0.154			
Water level	0.041	0.297			

Values in bold were significant ($p<0.05$).

3.2.4 18S rDNA (eukaryote) simulated metabarcoding (Kelpie)

There was a significant difference in the simulated 18S rDNA metabarcode assemblages compiled using Kelpie 2.0 by aquifer type (p=0.001) and between pre- and post-purge samples (p=0.023) (Figure 15). There was also a significant difference between bores (p=0.012); however, the catchment x pre-/post-purge interaction was not significant (p=0.324). When examined collectively, 23.3% of the variation in the 18S rDNA simulated eukaryote data (derived from Kelpie) could be explained by EC, TN, mean slot depth and DO (Table 6).

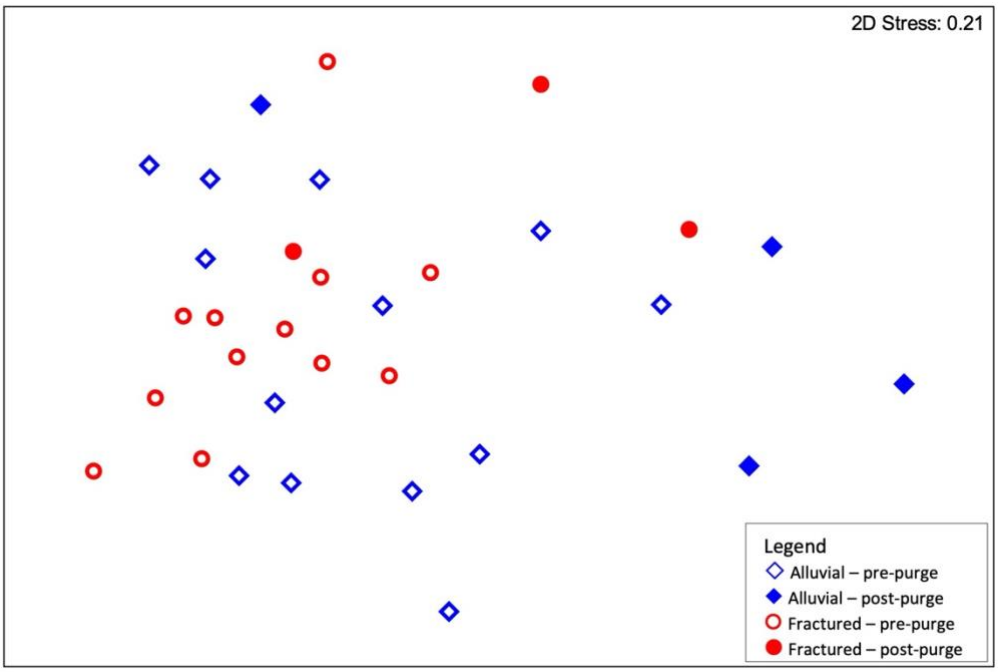

Figure 15. nMDS ordination of eukaryote assemblages in groundwater based on 18S rDNA genes identified within metagenome data using Kelpie

Samples represent pre- and post-purge groundwater samples in sandstone and alluvial aquifers.

Table 6. Proportion of variation (r^2) in eukaryote assemblages in groundwater based on 18S rDNA genes identified within metagenome data explained by environmental variables individually and in stepwise selection

Variable	Individual r^2	Individual p	Cumulative r^2	Cumulative Cumulative r^2	Cumulative p
Electrical conductivity	0.072	**0.008**	0.072	0.072	**0.008**
Total nitrogen	0.049	0.070	0.059	0.130	**0.019**
Mean slot depth	0.051	0.072	0.053	0.184	**0.032**
Dissolved oxygen	0.067	**0.014**	0.049	0.233	**0.048**
pH	0.065	**0.020**			
Oxidation-reduction potential	0.061	**0.024**			
Temperature	0.051	0.061			
Water level (depth to water –	0.049	0.097			
Sulfate	0.043	0.107			
Dissolved organic carbon	0.028	0.600			
Total phosphorus	0.028	0.621			

Values in bold were significant ($p<0.05$).

3.3 Metabarcode analysis

3.3.1 16S rDNA (prokaryote) metabarcode samples

Analysis of the 16S rDNA metabarcode data was limited to those samples from which metagenomes were retrieved. Within the main cluster of samples, there was a clear separation of samples by aquifer type and by pre- and post-purge (Figure 16). Of note in this analysis were the two outlying samples from wells 75040 and 75041-2 (fractured rock), which were much less diverse than other samples from the same aquifer (Korbel et al. 2022b). Samples from different aquifer types (p=0.001) and pre- and post-purge collections (p=0.041) were significantly different. Differences between bores (p=0.568) and the aquifer type x pre-/post-purge interaction (p=0.098) were not significant.

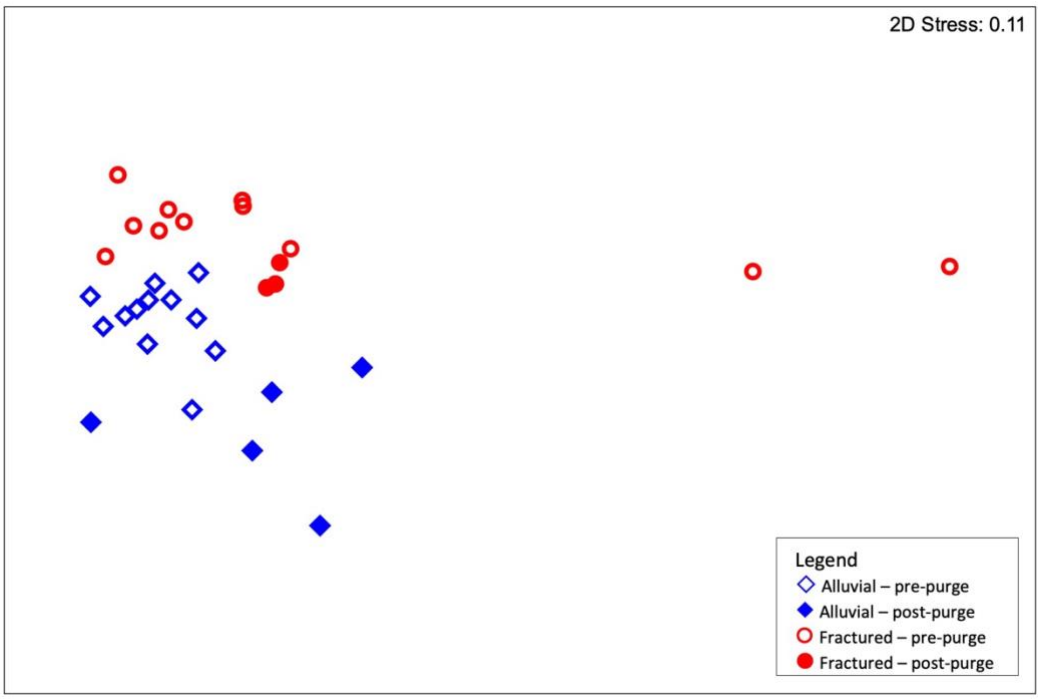

Figure 16. nMDS ordination of prokaryote assemblages in groundwater based on 16S rDNA genes identified using metabarcoding

Samples represent pre- and post-purge groundwater samples in sandstone and alluvial aquifers.

DistLM analysis identified pH, temperature, EC, DO, water level and mean slot depth as being independently significantly correlated with 16S rDNA metabarcode assemblages (Table 7). Stepwise (sequential) analysis identified pH, temperature, DOC and sulfate as variables contributing significantly to community structure, and together explaining 36.3% of the variation in 16S rDNA composition (Table 7).

Table 7. Proportion of variation (r^2) in prokaryote community structure based on 16S rDNA metabarcoding explained by environmental variables individually and in stepwise selection

Variable	Individual r^2	Individual p	Cumulative r^2	Cumulative Cumulative r^2	Cumulative p
pH	0.144	**0.001**	0.144	0.144	**0.001**
Temperature	0.099	**0.005**	0.102	0.245	**0.004**

Variable	Individual r^2	Individual p	Cumulative r^2	Cumulative Cumulative r^2	Cumulative p
Dissolved organic carbon	0.063	0.076	0.061	0.306	**0.026**
Sulfate	0.078	0.068	0.057	0.364	**0.036**
Electrical conductivity	0.096	**0.010**			
Water level (depth to water – casing height)	0.068	**0.041**			
Dissolved oxygen	0.068	**0.018**			
Mean slot depth	0.066	**0.042**			
Total nitrogen	0.053	0.097			
Oxidation-reduction potential	0.038	0.178			
Total phosphorus	0.034	0.378			

Values in bold were significant (p<0.05).

3.3.2 16S rDNA FAPROTAX

In the inferred functional assemblages derived from applying FAPROTAX to the 16S rDNA metabarcode data, there was no clear separation among samples by aquifer type and pre-/post-purge collection (Figure 17). However, there was a significant difference in the functional assemblages between aquifer types (p=0.002), but not between pre- and post-purge samples (p=0.061). There was a significant difference between bores (p=0.024), but the aquifer type x pre-/post-purge interaction was not significant (p=0.449). DistLM analysis found that 38.4% of the variation in the inferred functional assemblages could be explained by electrical conductivity, mean slot depth, pH, TN and DO (Table 8).

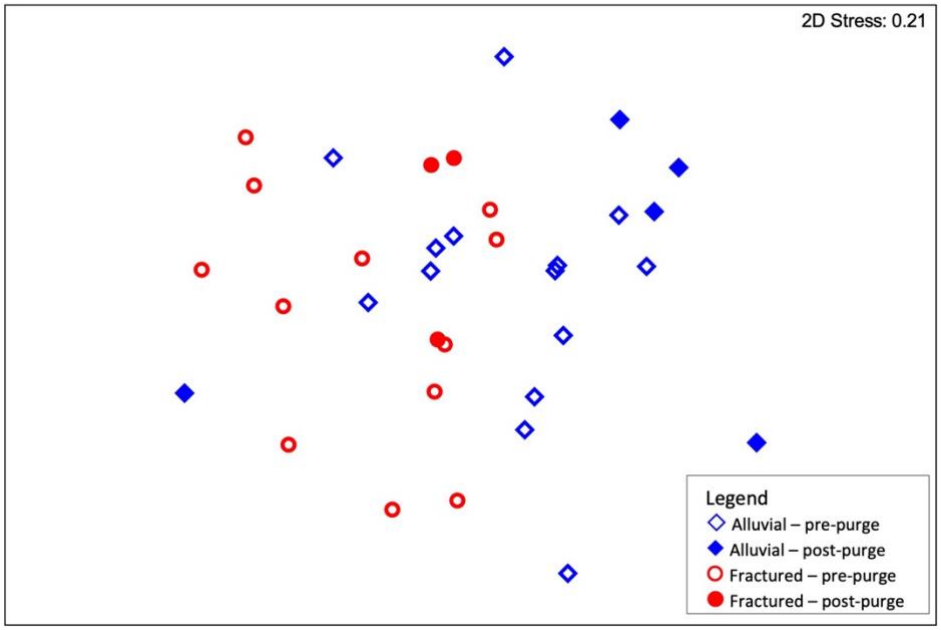

Figure 17. nMDS ordination of prokaryote functional assemblages in groundwater based on FAPROTAX analysis of 16S rDNA genes identified using metabarcoding

Samples represent pre- and post-purge groundwater samples in sandstone and alluvial aquifers.

Table 8. Proportion of variation (r^2) in functional prokaryote community structure based on FAPROTAX analysis of 16S rDNA explained by environmental variables individually and in stepwise selection

Variable	Individual r^2	Individual p	Cumulative r^2	Cumulative Cumulative r^2	Cumulative p
Electrical conductivity	0.120	**0.001**	0.120	0.120	**0.001**
Mean slot depth	0.091	**0.003**	0.081	0.201	**0.004**
pH	0.111	**0.002**	0.073	0.274	**0.002**
Total nitrogen	0.063	**0.025**	0.063	0.338	**0.008**
Dissolved oxygen	0.111	**0.001**	0.046	0.384	**0.027**
Sulfate	0.077	**0.004**			
Temperature	0.068	**0.022**			
Water level	0.056	0.052			
Total phosphorus	0.053	0.070			
Oxidation-reduction potential	0.041	0.225			
Dissolved organic carbon	0.035	0.358			

Values in bold were significant ($p<0.05$).

3.3.3 18S rDNA metabarcode

There was a clear separation in 18S metabarcode assemblages by aquifer type, and between pre- and post-purge samples within each aquifer type (Figure 18). As seen for the simulated 18S rDNA metabarcode data (Figure 15), the differences between aquifer types (p=0.001) and between pre- and post-purge samples (p=0.035) were significant. There was also a significant difference between bores (p=0.007); however, the catchment x pre-/post-purge interaction was not significant (p=0.219). In this analysis, the samples from bore 30303 grouped most closely with other samples from the alluvial aquifer. When examined collectively, only two environmental variables, pH and DO, were significantly correlated with 18S rDNA composition, cumulatively explaining 16.8% of the total variation in the eukaryote data (Table 9).

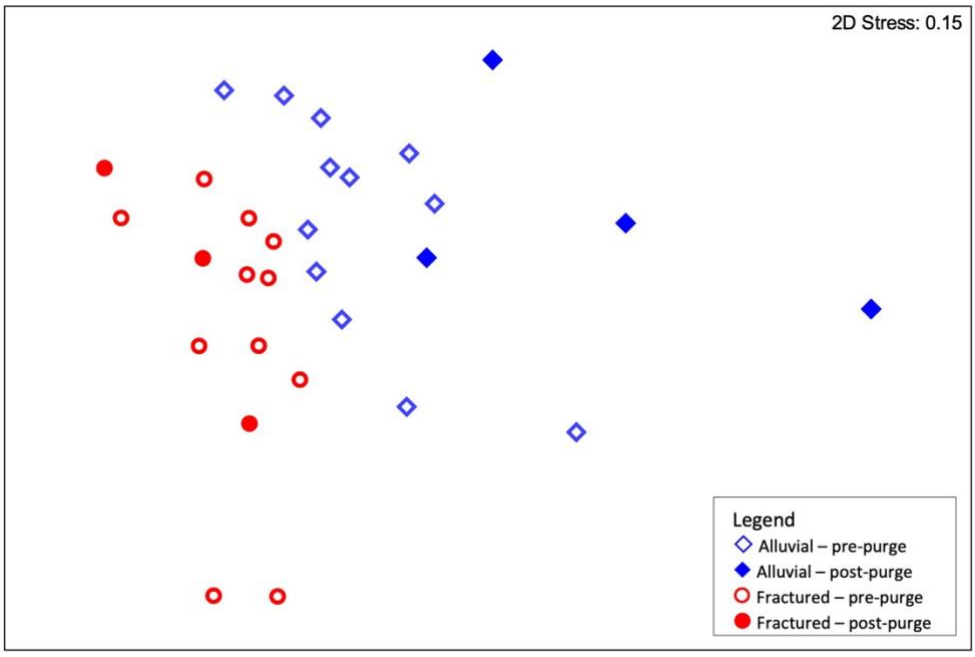

Figure 18. nMDS ordination of eukaryote assemblages in groundwater based on 18S rDNA genes identified using metabarcoding

Samples represent pre- and post-purge groundwater samples in sandstone and alluvial aquifers.

Table 9. Proportion of variation (r^2) in eukaryote community structure based on 18S rDNA metabarcodes explained by individual environmental variables individually and in stepwise selection

Variable	Individual r^2	Individual p	Cumulative r^2	Cumulative Cumulative r^2	Cumulative p
pH	0.107	**0.001**	0.107	0.107	**0.001**
Dissolved oxygen	0.082	**0.006**	0.061	0.168	**0.021**
Electrical conductivity	0.086	**0.006**			
Sulfate	0.069	**0.028**			
Oxidation-reduction potential	0.055	0.146			
Total nitrogen	0.042	0.173			

Variable	Individual r^2	Individual p	Cumulative r^2	Cumulative Cumulative r^2	Cumulative p
Total phosphorus	0.037	0.281			
Water level	0.028	0.556			
Mean slot depth	0.027	0.595			
Dissolved organic carbon	0.025	0.586			
Temperature	0.021	0.795			

Values in bold were significant (p<0.05).

3.4 Comparison of metagenome and metabarcode data

The analyses of the functional metagenomes, described in detail in the previous section, showed, overall, similar trends to the metabarcoding results here and in the Stage 1 and Stage 2 reports (Korbel et al. 2022a; Korbel et al. 2023). Ordinations typically showed separation of samples among catchment types, and all metagenome and metabarcoding datasets differed significantly between aquifer types (Table 10). Although not always evident in the nMDS ordinations, all metagenome datasets and all but the 16S rDNA FAPROTAX (based on metabarcoding) differed significantly between pre- and post-purge samples (Table 10). None of the functional genome datasets differed between bores (Table 10). In contrast, the taxonomic assemblages based on 16S rDNA or 18S rDNA metagenome sequences differed between bores, which was generally consistent with the metabarcoding data, except for the 16S rDNA metabarcodes, in which the difference between bores was not significant (p=0.568). These findings suggest that metagenome data reflect differences in environmental conditions as successfully as metabarcoding. The lack of significant differences between bores in the functional metagenomes compared to the taxonomic assessments may reflect differences in composition between bores, which might be expected given the high degree of endemism at small spatial scales that is common in groundwater organisms (Hahn and Matzke 2005; Hancock and Boulton 2008), but shows potentially a similarity among bores in terms of their function, despite different taxa undertaking those functions in different bores.

Table 10. Summary of statistically significant differences (p<0.05) between main effects and their interaction in assemblages derived from functional and taxonomic metagenome (MG) analysis and metabarcoding (MB)

Factors	All genes	GO	COG	EC	16S MG	16S MB	16S MB FAPROTAX	18S MG	18S MB
Aquifer type	✓	✓	✓	✓	✓	✓	✓	✓	✓
Pre-/post-purge	✓	✓	✓	✓	✓	✓	×	✓	✓
Bores (nested within aquifer type)	×	×	×	×	✓	×	✓	✓	✓
Aquifer type x pre-/post-purge	×	×	×	×	×	×	×	×	×

COG = Cluster of Orthologous Genes database, EC = Enzyme Commission database, GO = Gene Ontology database, 16S = 16S rDNA gene, 18S = 18S rDNA gene, MG = derived from metagenome, MB = derived from metabarcoding. FAPROTAX refers to functional assignment (Louca et al. 2016).

Consistent with the similar responses to PERMANOVA, functional metagenome datasets had a high degree of similarity, and relationships among samples were highly correlated (r>0.98) between datasets (Table 11), which is not surprising given the overlap among the individual databases, as well as the data compiled across those databases. Functional and taxonomic metagenome assemblages were less strongly correlated between datasets (Table 11) but were generally significantly correlated (p=0.001), except for COG and 18S rDNA (p=0.201). However, this is not surprising because COG functions are based on proteins from Bacteria and Archaea and do not include eukaryotes (Galperin et al. 2021). The microbial (16S rDNA) metagenome data were more strongly correlated with the functional metagenome data than were the eukaryote (18S rDNA) data, because functional processes are driven by microbial genes and because the prokaryote biomass is likely to be greater than the eukaryote biomass in these samples (Hose et al. 2022).

Table 11. Spearman rank correlation of Bray-Curtis similarities among samples characterised using 16S metabarcoding and 16S metagenome data

	COG	EC	GO	16S MG	18S MG
All genes	r=0.983, p=0.001	r=0.984, p=0.001	r=0.987, p=0.001	r=0.675, p=0.001	r=0.353, p=0.001
COG		r=0.983, p=0.001	r=0.986, p=0.001	r=0.652, p=0.001	r=0.079, p=0.201
EC			r=0.994, p=0.001	r=0.642, p=0.001	r=0.335, p=0.001
GO				r=0.656, p=0.001	r=0.343, p=0.001
16S					r=0.538, p=0.001

COG = Cluster of Orthologous Genes database, EC = Enzyme Commission database, GO = Gene Ontology database, 16S = 16S rDNA gene, 18S = 18S rDNA gene, MG = derived from metagenome.

The functional metagenome datasets were significantly correlated with the 16S rDNA and FAPROTAX metabarcode data (Table 12, r>0.20, p<0.045), but not with the 18S rDNA metabarcode data (Table 12, r<0.004, p>0.46). As for the metagenome data, the lack of correlation between the 18S rDNA (eukaryote) metabarcode data and the functional metagenome data is not surprising because the metagenomes are focused on prokaryote functions. Importantly, there was a significant (p≤0.002) correlation between the 16S rDNA and 18S rDNA metagenome datasets and the respective 16S rDNA and 18S rDNA metabarcode assemblages (Table 12).

Functional and taxonomic metagenome assemblages were less strongly correlated between datasets (Table 11) but were generally significantly correlated (p=0.001), except for COG and 18S rDNA (p=0.201). However, this is not surprising because eukaryote communities are not linked to the functions associated with COG. The microbial (16S rDNA) metagenome data were more strongly correlated with the functional metagenome data than were the eukaryote (18S rDNA) data, because functional processes are driven by microbial genes, and because the prokaryote biomass is likely to be greater than the eukaryote biomass in these samples (Hose et al. 2022).

Table 12. Spearman rank correlation of Bray-Curtis similarities among samples characterised using metabarcoding and functional metagenomics approaches

	All genes	COG	EC	GO	16S MG	18S MG
16S rDNA metabarcode	r=0.209, p=0.040	r=0.235, p=0.019	r=0.210, p=0.045	r=0.217, p=0.035	r=0.478, p=0.001	
16S rDNA FAPROTAX	r=0.514, p=0.001	r=0.529, p=0.001	r=0.524, p=0.001	r=0.531, p=0.001	r=0.560, p=0.001	
18S rDNA metabarcode	r=0.003, p=0.467	r=0.001, p=0.504	r=-0.012, p=0.522	r=0.002, p=0.489		r=0.306, p=0.002

COG = Cluster of Orthologous Genes database, EC = Enzyme Committee database, GO = Gene Ontology database, 16S = 16S rDNA gene, 18S = 18S rDNA gene, MG = derived from metagenome.

DistLM analyses across all variants of the metagenome data and the 16S rDNA metabarcode data identified a similar suite of variables as contributing most to explaining variation in the assemblage composition across samples (Table 13). Functional assemblages from all genes and from the EC and GO databases were influenced by temperature, pH, and TN and DO concentrations. Variables that best explained variation in COG functions were similar but included DOC and sulfate. The different response of COG compared to the other functional assemblages is consistent with other analyses above. The 16S rDNA assemblages from metabarcoding and metagenome analyses responded to a similar suite of variables (Table 13), with sulfate, DOC concentration, pH and temperature all explaining a significant portion of the variation in the prokaryote assemblages. The 16S rDNA metabarcode assemblages were also correlated with TN concentrations, which were significantly correlated with the 16S metagenome data alone (Table 13) but were not included in the final stepwise model.

Table 13. Summary of significant environmental variables identified by stepwise DistLM analysis for each dataset

Order	All genes	COG	EC	GO	16S MG	16S MB	16S MB FAPROTAX	18S MG	18S MB
1	Temp	Sulfate	Total N	Total N	Sulfate	pH	EC	EC	pH
2	pH	pH	Temp	Temp	DOC	Temp	Depth	Total	DO
3	Total N	DOC	pH	pH	pH	DOC	pH	Depth	
4	DO	Temp	DO	DO	Temp	Sulfate	Total N	DO	

EC = electrical conductivity, DO = dissolved oxygen, temp = temperature, DOC = dissolved organic carbon, depth = mean slot depth, TN = total nitrogen, COG = Cluster of Orthologous Genes database, EC = Enzyme Commission database, GO = Gene Ontology database, 16S = 16S rDNA gene, 18S = 18S rDNA gene, MG = derived from metagenome, MB = derived from metabarcoding. Full details of DistLM analyses are provided in sections 3.2 and 3.3 of this report.

RELATE analysis, which correlates the similarity among the 18S rDNA simulated metabarcode data compiled using Kelpie 2.0 with that among the 18S rDNA metabarcode samples based on presence/absence of OTUs, was significant (r=0.306, p=0.004), indicating that there was a significant rank correlation between the two matrices. Comparisons of similarities based on relative abundance data also indicated a significant correlation among similarities between samples (r=0.621, p=0.001). The suite of variables that best explained variation in the 16S rDNA FAPROTAX data (i.e., primarily EC and depth) was different to those that best explained other functional and 16S rDNA datasets. However, FAPROTAX infers function based on taxonomy, in contrast to the metagenomics approaches, which are based on the presence of genes with known function. Interestingly, there was

not a strong congruence between the variables that best explained the differences in the 18S rDNA metabarcode and metagenome assemblages. However, it is important to note that in all of these DistLM analyses, there were no strong environmental gradients across the samples. This was because sampling sites within an aquifer type were chosen to be similar and, given the clear differences in each dataset by aquifer type, the DistLM outcomes typically identified those variables that reflect aquifer types.

Results from this study show that metabarcode and metagenome data, overall, perform similarly in their ability to discriminate microbial assemblages from different aquifers and in response to environmental conditions. We have focused on community and functional composition, rather than diversity *per se*, and note that metabarcoding and metagenome approaches may differ (e.g., Becker and Pushkareva 2023) or not (e.g., Rubiola et al. 2022) in their ability to characterise microbial diversity. Metagenome data have the advantage of providing functional and taxonomic information (such as using Kelpie (Greenfield et al. 2019)), but doing so requires additional, specialist bioinformatics support.

This study has shown differences in the metagenomes of pre- and post-purge samples. Metagenomes were characterised for relatively few post-purge samples, which is likely due to a low mass of DNA in those samples. This is consistent with the findings of Stages 1 and 2 of this project, which reported that post-purge groundwater samples contain lower concentrations of DNA compared to pre-purge samples and differ in their microbial composition (Korbel et al. 2022a; Korbel et al. 2023) and metagenomes, as shown here. Post-purge sampling is essential to characterise aquifer communities; based on this, further development of the sampling approaches used in this study is needed to increase DNA yield and ensure that sufficient high-quality DNA is available for analysis. Given the relatively high cost of metagenome sequencing, it is prudent that all possible steps are taken to maximise the likelihood of successful sequencing.

Compared to metabarcoding, metagenome analysis provides similar depth of information on the taxonomic composition of the groundwater microbial communities and provides potentially greater depth of information on the function and composition of the communities but comes at greater cost and complexity. Indicative costs of sequencing the metagenomes of 60 samples (of which 34 were successfully sequenced) was $20,000. In comparison, metabarcode preparation and sequencing of the same 60 samples cost in the order of $8,000 (for 16S rDNA only). These indicative costs of analysis may be greater if using a commercial analytical service. The depth of information provided by metagenome analysis is valuable for assessing environmental impact and, where possible, this is an ideal means to assess groundwater microbial assemblages. However, metabarcoding still provides a detailed characterisation of the microbial assemblage, is able to show changes in assemblages over space and time and can be used to infer functional changes based on that taxonomy, either using tools such as FAPROTAX (Louca et al. 2016) or the abundance of specific taxa (e.g., Korbel et al. 2022b). At present, the knowledge gained from metagenomes that is of specific benefit to routine biomonitoring does not outweigh the greater cost over metabarcoding analysis. Indeed, metabarcoding is likely to be sufficient for groundwater monitoring for the foreseeable future, but as cost decreases and access to metagenome analysis increases, metagenome analysis will likely become the gold standard.

Our results for metagenome analysis indicate that …

This approach requires a large volume of DNA that was difficult to obtain from post-purge samples.

Pre-purge samples typically had higher DNA volumes, making them more amenable to metagenome analysis. However, pre-purge samples represent the community specifically associated with the environment provided by the bore and may not represent the community in the aquifer (see Stage 1 and 2 reports)

Analysis of 16S and 18S metabarcode sequences from the metagenome data using Kelpie (software) indicated similar patterns of differences among samples from different aquifer types and pre- and post-purge samples to those from the respective metabarcode analysis.

Results from this study show that metabarcode and metagenome data, overall, perform similarly in their ability to discriminate microbial assemblages from different aquifers and in response to environmental conditions.

Metagenome analysis provides greater depth of information on community composition and function than does metabarcoding.

Metabarcoding is less expensive than metagenome analysis and is likely to be sufficient for groundwater monitoring until the costs of metagenome analyses decline.

4. Summary and recommendations

The aims of this report were to investigate the differences between metabarcoding and metagenome analysis of eDNA within groundwater, and comment on their pros and cons. In this section we provide a summary of outcomes and recommendations based on the four core research questions (Section 1.3.1).

Are there differences in the taxonomic and functional composition of biotic assemblages in alluvial and sandstone aquifers as characterised by metagenomics and metabarcoding?

- Both metagenomics and metabarcode analysis identified differences in the functional and composition structure of microbial assemblages between sandstone and alluvial aquifers.

Are there associations between water quality parameters and microbial function and composition in both aquifers?

- The taxonomic structure and function of microbial communities were associated with water quality and environmental conditions. Communities characterised using metagenome and metabarcoding analyses responded to a similar suite of water quality variables.

Can spiking be used to enable quantitative comparisons of eDNA data?

- The use of spikes to quantify abundance of taxa in eDNA samples is technically challenging, is time-consuming, and introduces additional uncertainties into the quantitation.
- Spikes added to samples prior to PCR require a large amount of work to estimate suitable concentration so as to not 'swamp' the sample DNA.
- There is no clear evidence that spiking aids the ability to quantify abundances or improve analyses based on relative abundance of microbes within environmental samples.

Is it feasible for consultants to collect and process groundwater samples for either shotgun sequencing or metabarcoding approaches to assess potential impacts of coal mining and CSG development on microbial composition and activity?

- Metagenome analysis indicated:
  - biological community differences between pre-purge and post-purge samples in terms of functional genes and taxonomic composition
  - biological community differences between aquifer types in terms of functional genes and taxonomic composition and in response to environmental conditions.
- Compared to metabarcoding, metagenome analysis:
  - provides a similar depth of information on the taxonomic composition of the groundwater microbial communities
  - provides a similar analysis of microbial responses to environmental conditions
  - provides greater detail and depth of information on the functions of microbial assemblages
  - currently costs more than 2.5 times the cost of metabarcoding analysis.

- Currently metabarcoding is sufficient for the analysis of microbes and stygofauna for the purpose of monitoring; however, it is likely that as metagenome analyses improve and costs reduce, metagenome analysis may be more accessible for routine monitoring of groundwater in the future.

Acknowledgments

Sample collection and analysis was undertaken by Grant Hose (Macquarie University (MU)), Kathryn Korbel (MU), Kitty McKnight (MU), Tess Nelson (MU), Zac Whetters (MU) and Harry Anderson (MU). This report was prepared by Grant Hose, Kathryn Korbel and Kitty McKnight. All bioinformatics analyses were performed by Paul Greenfield (CSIRO). Anthony Chariton provided guidance on molecular methods.

We acknowledge the people of Darkinjung, Dharug and Dharawal Nations as the traditional owners and custodians of the lands on which this work was undertaken. We pay respects to all Elders past and present and to the Indigenous leaders of tomorrow.

Reference list

Abrego N, Roslin T, Huotari T, Ji Y, Schmidt NM, Wang J, Yu DW and Ovaskainen O 2021. Accounting for species interactions is necessary for predicting how arctic arthropod communities respond to climate change. *Ecography*, 44, 885–896.

Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C and Gnirke A 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, 12, R18.

Anderson MJ 2001. A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26, 32–46.

Anderson MJ, Gorley RN and Clarke KR 2008. *PERMANOVA+ for PRIMER: guide to software and statistical methods*. PRIMER-E, Plymouth.

Apprill A, McNally S, Parsons R and Weber L 2015. Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. *Aquatic Microbial Ecology*, 75, 129–137.

Arkin AP, Cottingham RW, Henry CS, Harris NL, Stevens RL, Maslov S, Dehal P, Ware D, Perez F, Canon S, Sneddon MW, Henderson ML, Riehl WJ, Murphy-Olson D, Chan SY, Kamimura RT, Kumari S, Drake MM, Brettin TS, Glass EM, Chivian D, Gunter D, Weston DJ, Allen BH, Baumohl J, Best AA, Bowen B, Brenner SE, Bun CC, Chandonia J-M, Chia J-M, Colasanti R, Conrad N, Davis JJ, Davison BH, Dejongh M, Devoid S, Dietrich E, Dubchak I, Edirisinghe JN, Fang G, Faria JP, Frybarger PM, Gerlach W, Gerstein M, Greiner A, Gurtowski J, Haun HL, He F, Jain R, Joachimiak MP, Keegan KP, Kondo S, Kumar V, Land ML, Meyer F, Mills M, Novichkov PS, Oh T, Olsen GJ, Olson R, Parrello B, Pasternak S, Pearson E, Poon SS, Price GA, Ramakrishnan S, Ranjan P, Ronald PC, Schatz MC, Seaver SMD, Shukla M, Sutormin RA, Syed MH, Thomason J, Tintle NL, Wang D, Xia F, Yoo H, Yoo S and Yu D 2018. KBase: United States Department of Energy Systems Biology Knowledgebase. *Nature Biotechnology*, 36, 566–569.

Aßhauer KP, Wemheuer B, Daniel R and Meinicke P 2015. Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. *Bioinformatics*, 31, 2882–2884

Baily A, Rock L, Watson C and Fenton O 2011. Spatial and temporal variations in groundwater nitrate at an intensive dairy farm in south-east Ireland: insights from stable isotope data. *Agricultural Ecosystems and Environments*, 144, 308–318.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA and Pevzner PA 2019. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, 19, 455–477.

Becker B and Pushkareva E 2023. Metagenomics provides a deeper assessment of the diversity of bacterial communities in polar soils than metabarcoding. *Genes*, 14, 812.

Cameron ES, Schmidt PJ, Tremblay BJ, Emelko MB and Müller KM 2021. Enhancing diversity analysis by repeatedly rarefying next generation sequencing data describing microbial communities. *Scientific Reports*, 11, 22302.

Carraro L, Mächler E, Wüthrich R and Altermatt F 2020. Environmental DNA allows upscaling spatial patterns of biodiversity in freshwater ecosystems. *Nature Communications*, 11, 3585.

Carraro L, Stauffer JB and Altermatt F 2021. How to design optimal eDNA sampling strategies for biomonitoring in river networks 2021. *Environmental DNA*, 3, 157–172.

Chaffron S, Rehrauer H, Pernthaler J and von Mering C 2010. A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Research*, 20, 947–959.

Chariton AA, Stephenson S, Morgan MJ, Steven ADL, Colloff MJ, Court LN and Hardy CM 2015. Metabarcoding of benthic eukaryote communities predicts the ecological condition of estuaries. *Environmental Pollution*, 203, 165¬–174.

Clarke KR and Ainsworth M 1993. A method of linking multivariate community structure to environmental variables. *Marine Ecology Progress Series*, 92, 205–219.

Clarke KR and Green RH 1988. Statistical design and analysis for a 'biological effects' study. *Marine Ecology Progress Series*, 46, 213–226.

Deagle BE, Clarke LJ, Kitchener JA, Polanowski AM and Davidson AT 2018. Genetic monitoring of open ocean biodiversity: an evaluation of DNA metabarcoding for processing continuous plankton recorder samples. *Molecular Ecology Resources*, 18, 391–406.

Deagle BE, Thomas AC, McInnes JC, Clarke LJ, Vesterinen EJ, Clare EL, Kartzinel TR and Eveson JP 2019. Counting with DNA in metabarcoding studies: how should we convert sequence reads to dietary data? *Molecular Ecology*, 28, 391–406.

Deiner K, Bik HM, Mächler E, Seymour M, Lacoursière-Roussel A, Altermatt F, Creer S, Bista I, Lodge DM, de Vere N, Pfrender ME and Bernatchez L 2017. Environmental DNA metabarcoding: transforming how we survey animal and plant communities. *Molecular Ecology*, 26, 5872–5895.

Galperin MY, Wolf YI, Makarova KS, Alvarez RV, Landsman D and Koonin EV 2021. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Research*, 49, D274–D281.

Garrido-Sanz L, Senar MÀ and Piñol J 2022. Relative species abundance estimation in artificial mixtures of insects using mito-metagenomics and a correction factor for the mitochondrial DNA copy number. *Molecular Ecology Resources*, 22, 153–167.

Greenfield P, Tran-Dinh N and Midgley D 2019. Kelpie: generating full-length 'amplicons' from whole-metagenome datasets. *PeerJ*, 6, e6174.

Griebler C, Avramov M and Hose G 2019. Groundwater ecosystems and their services – current status and potential risks. In Schröter M, Bonn A, Klotz S, Seppelt R and Baessler C (eds), *Atlas of ecosystem services – drivers, risks, and societal responses*. Springer, Cham, 197–203.

Griebler C and Lueders T 2009. Microbial biodiversity in groundwater ecosystems. *Freshwater Biology*, 54, 649–677.

Hahn HJ and Matzke D 2005. A comparison of stygofauna communities inside and outside groundwater bores. *Limnologica*, 35, 31–44.

Hancock PJ and Boulton AJ 2008. Stygofauna biodiversity and endemism in four alluvial aquifers in eastern Australia. *Invertebrate Systematics*, 22, 117–126.

Hardy CM, Krul ES, Hartley DM and Oliver RL 2010. Carbon source accounting for fish using combined DNA and stable isotope analyses in a regulated lowland river weir pool. *Molecular Ecology*, 19, 197–212.

Hemme CL, Tu Q, Shi Z, Qin Y, Gao W, Deng Y, Nostrand JDV, Wu L, He Z, Chain PSG, Tringe SG, Fields MW, Rubin EM, Tiedje JM, Hazen TC, Arkin AP and Zhou J 2015. Comparative metagenomics reveals impact of contaminants on groundwater microbiomes. *Frontiers in Microbiology*, 6, 1205.

Hose GC, Chariton A, Daam M, Di Lorenzo T, Galassi DMP, Halse SA, Reboleira ASPS, Robertson AL, Schmidt SI, Korbel K 2022. Invertebrate traits, diversity and the vulnerability of groundwater ecosystems. *Functional Ecology*. 36, 2200-2214.

Hose GC, Sreekanth J, Barron O and Pollino C 2015. *Stygofauna in Australian groundwater systems: extent of knowledge*. Report to Australian Coal Association Research Program. Macquarie University and CSIRO.

Hugerth LW and Andersson AF 2017 Analysing microbial community composition through amplicon sequencing: from sampling to hypothesis testing. *Frontiers in Microbiology*, 8, 1561.

Ji Y, Huotari T, Roslin T, Schmidt NM, Wang J, Yu DW and Ovaskainen O 2020. SPIKEPIPE: a metagenomic pipeline for the accurate quantification of eukaryotic species occurrences and intraspecific abundance change using DNA barcodes or mitogenomes. *Molecular Ecology Resources*, 20, 256–267

Kanagawa T 2003. Bias and artifacts in multitemplate polymerase chain reactions (PCR). *Journal of Bioscience and Bioengineering*, 96, 317–323.

Korbel K, Chariton A, Stephenson S, Greenfield P and Hose GC 2017. Wells provide a distorted view of life in the aquifer: implications for sampling, monitoring and assessment of groundwater ecosystems. *Scientific Reports*, 7, 40702.

Korbel K, McKnight K, Greenfield P, Angel B, Adams M, Chariton A and Hose G 2022a. *Bioassessment of groundwater ecosystems. I. Sampling methods and analysis of eDNA for microbes and stygofauna in shallow alluvial aquifers*. Report prepared for the Independent Expert Scientific Committee on Coal Seam Gas and Large Coal Mining Development through the Department of Climate Change, Energy, the Environment and Water, Commonwealth of Australia.

Korbel K, McKnight K, Greenfield P, Angel B, Adams M, Chariton A and Hose GC 2023. *Bioassessment of groundwater ecosystems. II. Sampling methods and analysis of eDNA for microbes and stygofauna in shallow sandstone aquifers*. Report prepared for the Independent Expert Scientific Committee on Coal Seam Gas and Large Coal Mining Development through the Department of Climate Change, Energy, the Environment and Water, Commonwealth of Australia.

Korbel KL and Hose GC 2017. The Weighted Groundwater Health Index: improving the monitoring and management of groundwater resources. *Ecological Indicators*, 75, 64–181.

Korbel KL, Greenfield P and Hose GC 2022b. Agricultural practices linked to shifts in groundwater microbial structure and denitrifying bacteria. *Science of the Total Environment*, 807, 150870

Korbel KL, Hancock PJ, Serov P, Lim RP and Hose GC 2013. Groundwater ecosystems vary with land use across a mixed agricultural landscape. *Journal of Environmental Quality*, 42, 380–390.

Korbel KL, Rutlidge H, Hose GC, Eberhard SM and Andersen MS 2022c. Dynamics of microbiotic patterns reveal surface water groundwater interactions in intermittent and perennial streams. *Science of the Total Environment*, 811, 152380.

Krehenwinkel H, Wolf M, Lim JY, Rominger AJ, Simison WB and Gillespie RG 2017. Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. *Scientific Reports*, 7, 17668.

Krueger F, Andrews SR and Osborne CS 2011. Large scale loss of data in low-diversity Illumina sequencing libraries can be recovered by deferred cluster calling. *PLoS One*, 6, e16607.

Langille MGI, Zaneveld J, Caporaso JG, Mcdonald D, Knights D, Reyes JA, Clemente JC, Burkepile DE, Vega Thurber RL, Knight R, Beiko RG and Huttenhower C 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology*, 31, 814–821.

Lear G, Dickie I, Banks JC, Boyer S, Buckley HL, Buckley TR, Cruickshank R, Dopheide, A, Handley KM, Hermans S, Kamke J, Lee CK, MacDiarmid R, Morales SE, Orlovich DA, Smissen R, Wood J and Holdaway R 2018. Methods for the extraction, storage, amplification and sequencing of DNA from environmental samples. *New Zealand Journal of Ecology*, 42, 10A–50A.

Legendre P and Legendre L 2012. *Numerical ecology*. 3rd Edition. Elsevier, Amsterdam.

Louca S, Parfrey LW and Doebeli M 2016. Decoupling function and taxonomy in the global ocean microbiome. *Science*, 353, 1272–1277.

Luo M, Ji Y, Warton D and Yu DW 2023. Extracting abundance information from DNA-based data. *Molecular Ecology Resources*, 23, 174–189.

Mallona I, Weiss J and Egea-Cortines M 2011. pcrEfficiency: a Web tool for PCR amplification efficiency prediction. *BMC Bioinformatics*, 12404.

McLaren MR, Willis AD and Callahan BJ 2019. Consistent and correctable bias in metagenomic sequencing experiments. *Elife*, 8, e46923.

McMurdie PJ and Holmes S 2014. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Computational Biology*, 10, e1003531.

Nichols RV, Vollmers C, Newsom LA, Wang Y, Heintzman PD, Leighton M, Green RE and Shapiro B 2018. Minimizing polymerase biases in metabarcoding. *Molecular Ecology Resources*, 18, 927–939.

Parada AE, Needham DM and Fuhrman JA 2016. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environmental Microbiology*, 18, 1403–1414.

Pavlovska M, Prekrasna I, Parnikoza I and Dykyi E 2021. Soil sample preservation strategy affects the microbial community structure. *Microbes and Environments*, 36, ME20134.

Peel N, Dicks LV, Clark MD, Heavens D, Percival-Alwyn L, Cooper C, Davies RG, Leggett RM and Yu DW 2019. Semi-quantitative characterisation of mixed pollen samples using MinION sequencing and Reverse Metagenomics (RevMet). *Methods in Ecology and Evolution*, 10, 1690–1701.

Pollock J, Glendinning L, Wisedchanwet T and Watson M 2018. The madness of microbiome: attempting to find consensus 'best practice' for 16S microbiome studies. *Applied and Environmental Microbiology*, 84, e02627-17.

Prjibelski A, Antipov D, Meleshko D, Lapidus A and Korobeynikov A 2020. Using SPAdes de novo assembler. *Current Protocols in Bioinformatics*, 70, e102.

Rojahn J, Pearce L, Gleeson DM, Duncan RP, Gilligan DM and Bylemans J 2021. Semi-quantitative characterisation of mixed pollen samples using MinION sequencing and Reverse Metagenomics. *Methods in Ecology and Evolution*, 10, 1690–1701.

Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C and Jaffe DB 2013. Characterizing and measuring bias in sequence data. *Genome Biology*, 14, R51.

Rubiola S, Macori G, Civera T, Fanning S, Mitchell M and Chiesa F 2022. Comparison between full-length 16S rRNA metabarcoding and whole metagenome sequencing suggests the use of either is suitable for large-scale microbiome studies. *Foodborne Pathogens and Disease*, 19, 495–504.

Ruppert KM, Kline RJ and Rahman MS 2019. Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: a systematic review in methods, monitoring, and applications of global eDNA. *Global Ecology and Conservation*, 17, e00547.

Saccò M, Guzik MT, van der Heyde M, Nevill P, Cooper SJB, Austin AD, Coates PJ, Allentoft ME and White NE 2022. eDNA in subterranean ecosystems: applications, technical aspects, and future prospects. *Science of the Total Environment*, 820, 153223.

Sang S, Zhang X, Dai H, Hu BX, Ou H and Sun L 2018. Diversity and predictive metabolic pathways of the prokaryotic microbial community along a groundwater salinity gradient of the Pearl River Delta, China. *Scientific Reports*, 8, 17317.

Seemann T 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30, 2068-9.

Smets W, Leff JW, Bradford MA, McCulley RL, Lebeer S and Fierer N 2016. A method for simultaneous measurement of soil bacterial abundances and community composition via 16S rRNA gene sequencing. *Soil Biology and Biochemistry*, 96, 145–151.

Spear MJ, Embke HS, Krysan PJ and Vander Zanden MJ 2021. Application of eDNA as a tool for assessing fish population abundance. *Environmental DNA*, 3, 83–91.

Sutcliffe B, Chariton AA, Harford AJ, Hose GC, Stephenson S, Greenfield P, Midgley DJ and Paulsen IT 2017. Insights from the genomes of microbes thriving in uranium-enriched sediments. *Microbial Ecology*, 75, 970–984.

Suzuki MT and Giovannoni SJ 1996. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Applied Environmental Microbiology*, 62, 625–630.

Taberlet P, Bonin A, Zinger L and Coissac E 2018. *Environmental DNA: for biodiversity research and monitoring*. Oxford University Press.

Thomas AC, Deagle BE, Eveson JP, Harsch CH and Trites AW 2016. Quantitative DNA metabarcoding: improved estimates of species proportional biomass using correction factors derived from control material. *Molecular Ecology Resources*, 16, 714–726.

Tillotson MD, Kelly RP, Duda JJ, Hoy M, Kralj J and Quinn TP 2018. Concentrations of environmental DNA (eDNA) reflect spawning salmon abundance at fine spatial and temporal scales. *Biological Conservation*, 220, 1–11.

Tischer K, Kleinsteuber S, Schleinitz KM, Fetzer I, Spott O, Stange F, Lohse U, Franz J, Neumann F, Gerling S, Schmidt C, Hasselwander E, Harms H and Wendeberg A 2013. Microbial communities along biogeochemical gradients in a hydrocarbon-contaminated aquifer. *Environmental Microbiology*, 15, 2603–2615.

Tkacz A, Hortala M and Poole PS 2018. Absolute quantitation of microbiota abundance in environmental samples. *Microbiome*, 6, 110

Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, Lozupone C, Zaneveld JR, Vázquez-Baeza Y, Birmingham A, Hyde ER and Knight R 2017. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, 5, 27.

Appendices

Appendix 1. Primers and PCR conditions for eDNA

Table A1.1. Primer sequences used in PCR

Gene	Size (bp)	Primer ID	Sequence	Reference
16S rDNA	350	515FB	GTGYCAGCMGCCGCGGTAA	Parada et al. 2016
		806FB	GGACTACNVGGGTWTCTAAT	Apprill et al. 2015
18S rDNA	200–500	All18SF	5'-TGGTGCATGGCCGTTCTTAGT-3'	Hardy et al. 2010
		All18SR	5'-CATCTAAGGGCATCACAGACC-3'	

bp = base pair

Table A1.2. Synthetic 16S rDNA gBlocks® Gene Fragments

Gene	Size (bp)	Sequence
Spike 1: **16S rDNA Bacteria_Syn2**	319	5'- GTG CCA GCA GCC GCG GTA ACT TGT AGG CTC GAC CCT GAT GGA TTG CTG TCA CAT TAC GAC GCA TCC CAA ATA AGC CGC TCT CAG AAG AAA GCC ATG CCC TTA TTA CTT GTG CGC GCG GCG GCC ATC CTG AAT GCC TCG TAC GTT TCA AAG GCT TAC TCC GGT GCT GGC CGA TTC GAC TTA ACA CTC TGT GCT TCT CTG GTC GCT CCG TCC TTT GCA GTG ATC AAA TCC GAC GGG ACT CGC ACC TGT GGG CAA CGA AAC GAG GTG ATT ACC CTA CGG AAG CCT GCT CCC ACT AGA TAA TCA TTA GAT ACC CTA GTA GTC C -3'
Spike 2: **16S rDNA Bacteria_Syn3**	319	5'- GTG CCA GCA GCC GCG GTA ATA TAG ATT CAC TTC TGG AGG TCT GCC AAC CAC CCA GGA CCC ATC TCA CAA TCC TAT ACA GAC ATC GAG GCA CGC AGC TGG AAA GAC TGC TAA AGC ATC ACA CGT ACC CAA CTT ATG GCC AAG TCC TAG TCT GTC CGT TGG ACT GGC GTG GCC TCC CGC CAC GTC GAC GTA ATA GTC CCG GAG TAT CAA GTT ACC TAG CAC GGC GAA ATA TGA AAC TTT CTG ATA TTA ATG TAG ATC ATT TGA CTC TCT CCG CAT AAG GCC AGC GTG AGG CGC GTG TCA GGA TTA GAT ACC CTA GTA GTC C -3'

bp = base pair

Table A1.3. PCR cycle details for 16S rDNA metabarcoding

Primer		Temperature (°C)	Time
16S			
Initial denaturation		95	10 min
35 x PCR cycles	Denaturation	94	45 sec
	Hybridisation	50	60 sec
	Elongation	72	90 sec
Final elongation		72	10 min

Appendix 2. 16S rDNA amplicon bioinformatic methods

The Illumina MiSeq 16S amplicon data were processed using an in-house custom pipeline based on USearch tools and Ribosomal Database Project (RDP). This hybrid pipeline takes files of reads and generates a single operational taxonomic unit (OTU) table covering all of the samples in the study. Each OTU is classified both by using RDP and by matching the sequence to a curated set of 16S reference sequences. The use of two independent classification techniques is done to provide some insight into the reliability of the taxonomic assignments.

The pipeline first demultiplexed the data to produce a pair of read files for each sample. These paired reads were then merged, trimmed and dereplicated, and then clustered at 97% similarity to generate a set of representative OTU sequences. The merging, dereplicating and clustering steps were done using USearch v8.1.1812 tools (fastq_mergepairs, derep_fulllength and cluster_otus). The merging step excluded any merged reads with greater than 1 expected error (fastq_merge_maxee 1.0). The clustering step also checked for chimeras, running each sequence through UParse-ref using the current set of OTUs as a reference database. If the optimal model is chimeric, the sequence is discarded. Each of these OTU sequences was then classified in two different ways: by using the RDP Classifier (v2.10.2) to determine a taxonomic classification for each sequence, down to best level of genus; and by using usearch_global to find the best match for each sequence within a curated set of 16S reference sequences, giving a species-level classification for each OTU sequence. The 16S reference set used for the species-level classification was built from the RDP Classifier's training set (v14), augmented with additional sequences from the Genomes OnLine Database (GOLD). The pipeline then used usearchglobal to map the merged reads from each sample back onto the OTU sequences to get accurate read counts for each OTU/sample pairing. The classified OTUs and the counts for each sample were then used to generate OTU tables in both text and .biom (v1) formats, complete with taxonomic classifications, species assignments and counts for each sample. Summaries of the OTU classifications were also produced at taxonomic levels from phylum to genus and species. All singleton reads were removed prior to the OTU formation step. The datasets were then filtered by removing OTUs with <10 counts across samples. Counts within individual samples were then adjusted based on the number of counts of positive controls that had jumped between samples. Rare species were removed for 16S datasets.

While we recognise that there are issues with using the number of amplicon sequence reads as a surrogate for taxon abundance, there is currently no consensus on the most appropriate strategy for the analysis of such data. Although commonly practised, we have chosen not to rarefy these data (i.e., randomly resample to standardise all samples to a minimum read number) prior to analysis because of the loss of important biological information that this process mandates (e.g., McMurdie and Holmes 2014), and because we have already removed rare taxa that are potential erroneous sequences in our earlier data screening processes (see previous paragraph). Instead we have normalised read numbers for each taxon by dividing by the total read number for the sample, thereby expressing each taxon in terms of its relative read abundance.

Australian Government | This initiative is funded by the Australian Government