# Analysis of soil eDNA functional genes: potential to increase profitability and sustainability of pastoral agriculture

SA Wakelin, VM Cave, BE Dignam, C D'Ath, M Tourna, LM Condron, J Zhou, JD Van Nostrand & M O'Callaghan

View supplementary material

Published online: 25 Jul 2016.

Submit your article to this journal

Article views: 371

View related articles

View Crossmark data

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

# Analysis of soil eDNA functional genes: potential to increase profitability and sustainability of pastoral agriculture

SA Wakelin[a] , VM Cave[a], BE Dignam[a,b], C D'Ath[a], M Tourna[a,c], LM Condron[b,d], J Zhou[e,f,g], JD Van Nostrand[e] and M O'Callaghan[a]

[a]Lincoln Science Centre, AgResearch Ltd, Christchurch, New Zealand; [b]Bio-Protection Research Centre, Lincoln University, Canterbury, New Zealand; [c]Department of Biochemistry and Biotechnology, University of Thessaly, Larisa, Greece; [d]Faculty of Agriculture and Life Sciences, Lincoln University, Christchurch, New Zealand; [e]Department of Microbiology and Plant Sciences, Institute for Environmental Genomics, University of Oklahoma, Norman, Oklahoma, USA; [f]Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA; [g]School of Environment, Tsinghau University, Beijing, China

**ABSTRACT**
Management of soil biological resources to optimise plant production, efficiency of nutrient inputs, and system sustainability is an emerging opportunity for pastoral agriculture. To achieve these goals, suitable tools that can assess the functional state of the soil ecosystem must be developed and standardised approaches to their application adopted. Towards this end, we have undertaken comprehensive, high-density functional-gene microarray analysis (GeoChip5) of environmental DNA (eDNA) extracted from 50 pastoral soils. When combined with soil, environmental and management metadata, the information can be used to provide insights into soil biological processes spanning greenhouse gas emissions, through to natural suppression of plant root diseases. To provide an example of a structured workflow of analysis in a pastoral system context, we analysed the GeoChip data using a combination of approaches spanning routine univariate methods through to more complex multivariate and network-based analysis. Analyses were restricted to comparing effects of land-use (dairy or 'other' farming systems), and exploring relationships of the GeoChip data with the soil properties from each sample. These exemplar analyses present a pathway for the application of eDNA approaches (GeoChip or others) to deliver outcomes for pastoral agricultural in New Zealand.

## Introduction

Biology and biological processes occurring in soils have many direct links to plant productivity. Most importantly, the diversity of soil microbiology constitutes the primary reservoir of beneficial and deleterious organisms with which plants interact. These associations can be so important that introduction of a key microbial group to natural and managed soil ecosystems can have major impacts on plant production and

above-ground ecosystem productivity (van der Heijden et al. 1998; Anderson et al. 2004; Herridge et al. 2008).

Soil biology not only directly affects plant production, but regulates soil formation, cycling of important nutrients (carbon [C], nitrogen [N], phosphorus [P], sulphur [S]), provides a source and sink of greenhouse gases and affects water quality (e.g. nitrate-N, enteric pathogens) (Roper & Gupta 1995). In New Zealand, approximately 30% of the land area is used for high-producing pastoral agriculture (Statistics New Zealand 2012). As such, there is not only potential to increase the productive capacity of pastures through management of soil biology, but as a consequence of the large extent of land area under pasture, there are strong opportunities to achieve environmental outcomes linked to the range of ecosystem services supported by soils (Coleman & Whitman 2005; Ogunseitan 2005). Given this, it is bewildering that soil ecosystem services work to date has not explicitly studied 'soil biology'.

Productive gains in farming systems have been achieved through genetic gain in pasture and livestock breeding (Barrett et al. 2015; Santos et al. 2015). In both cases, advances have been underpinned by fundamental knowledge of organisms' genetics, physiology and phenotypes, which collectively allow for breeding and selection across multiple phenotypes. For the other major biological component of farming systems, i.e. the soil biology, the same primary requirements need to be met, but gaining this knowledge is vastly more difficult. The diversity and number of organisms in soils is massive ($>10^4$ to $10^5$ species/g soil; Curtis et al. 2002; Torsvik et al. 2002), traits are not usually visually scorable and phenotypes (e.g. a soils with enhanced disease suppression or greater nitrogen-use efficiency) are the result of the interaction of a multitude of species, often spanning phyla and/or trophic groups. There is often an idiosyncratic relationship between species and function (Nielsen et al. 2011) that needs to be considered alongside issues of size and scale. Furthermore, the soil system is not static; functions such as disease suppression, nutrient cycling and ecosystem stability (resilience and resistance) are emergent properties (Cardinale et al. 2003; Konopka 2009). Thus, an ecosystem approach is clearly required to identify opportunities for achieving gains in soil function. Given the unique issues associated with soil biology (species and diversity, organisms' size, often unknown associations between biological elements [e.g. species] and function), ecological genomics is being as an important integrative tool (Nesme et al. 2016; Myrold et al. 2014; Nannipieri et al. 2014). In this context, ecological genomics refers to analysis of the soil environmental DNA (eDNA) as a 'metagenome' (harbouring the pooled DNA from many thousands of species), and understanding relationships to both the wider abiotic and biotic environment.

The comprehensive assessment of soil ecosystems—and, more importantly, functional genes associated with ecosystem processes—is rapidly becoming cost effective. This is principally through the development of a range of high-throughput genomics-era tools such as next generation DNA sequencing (NGS) and the development of various microarray platforms (Nannipieri et al. 2014). One such example is the GeoChip functional microarray system (He et al. 2007). This platform allows for the simultaneous detection of genes spanning a broad range of functions associated with biogeochemical cycling of nutrients, through antibiotic resistance and plant growth regulation (He et al. 2007). Thus, an assessment of the functional capacity of the soil ecosystem is obtained and this can be compared with samples collected over time, space and/or following change in management (and so forth). Collectively, this information can be used to understand

how soil ecosystems vary with soil properties, environment (e.g. climate change), management and land-use alteration (He et al. 2007; Yergeau et al. 2009; Wakelin et al. 2013a; Paula et al. 2014). The standardised application of GeoChip (or one of a number of other ecological genomics tools) can aid in providing a functional understanding of soil ecosystems and deliver the productive and environmental potential underlying pastoral faming systems.

The aim of this investigation is to demonstrate a pipeline of analysis of soil environmental genomics for New Zealand's pastoral agricultural sector based on the GeoChip5 functional microarray (Tu et al. 2014). The sample data set is comprised of eDNA from 50 well-characterised New Zealand pastoral soils. The pastoral soils were collected from across varying soil types, geographic zones and under a range of system intensities, spanning high stocking rate dairy through to expansive sheep grazing systems. For simplicity, relationships between the functional gene data and farm management were compared only for dairy vs 'other' grazing systems (see later definition) in this work. A comprehensive set of soil physicochemical properties and long-term environmental conditions from the sites (Wakelin et al. 2013b) enables the assessment of the soil ecological genomics against a set of appropriate metadata, and these can be used in future studies to assess, for example, influences of intensification more precisely. Overall, the two sets enable changes in functional soil biology to be interpreted alongside soil, climatic and farm-management influences.

## Materials and methods

### Sample sites and eDNA

Environmental DNA was extracted from 50 representative samples of pasture soil collected from across New Zealand. Soils were collected from 11 major soil groups and spanned 10 geographic zones. There were sufficient numbers of samples taken from Brown, Pallic and Recent soil groups to allow for structured analysis of 'soil type' effects. Soils were also collected from under high- and low-intensity land use. These were arbitrarily defined as pasture being used for dairy or 'other' grazing purposes, and this is supported by changes in soil nutrient status (Wakelin et al. 2013b). A full description of the sites, soil sampling, analysis of soil physicochemical properties, and DNA extraction and purification are given in Wakelin et al. (2013b). The eDNA samples are archived (−80 °C) at AgResearch, Lincoln.

### GeoChip functional microarray analysis

A subsample of eDNA from each of the pasture sites was transported on dry-ice to the Institute for Environmental Genomics, University of Oklahoma for analysis on GeoChip5 (Yan et al. 2015). The quantity of eDNA in samples from soil sites 2, 18, 29, 33 and 35 were below minimum requirements for direct analysis (500 ng). In order to increase the total amount of DNA in these samples, the eDNA was amplified using whole community genome amplification (WCGA) using the phi29 enzyme system (Wu et al. 2006; Wang et al. 2011). Processing of the DNA then followed the standard system: labelling, hybridisation, scanning and noise reduction/quality control of the

data. These steps have been described in full detail elsewhere (Paula et al. 2014; Tu et al. 2014; Yan et al. 2015). Briefly, DNA for all samples was labelled with Cy-3 (using random primers) and the Klenow fragment, purified, dried and rehydrated in hybridisation solution (scaled up to 125 uL) along with Cy5-labelled universal standards. The mixtures were hybridised to the GeoChip5 array, washed and scanned on a NimbleGen MS200 array system (Roche) and the spot intensity data recorded. Probe spots with a signal to noise ratio <2 were removed.

## Data handling and pre-treatment

The complete data set contained 87,678 variables (arrays-spot intensity values) for each of the 50 soil samples (4.83 million values). These were log-transformed and preliminary groupings of the samples according to similarity (Euclidean distances) were made with hierarchical clustering (group average method). The significance of cluster groupings was tested using the similarity profile test (SIMPROF; Clarke et al. 2008) at $\alpha = 0.05$. Samples from sites 2, 18, 29, 33 and 35 distinctly out-grouped with respect to the wider data set. As these were the only samples pre-processed for WCGA, it is highly likely that this was due to amplification bias when using phi29 (Wang et al. 2011). As there was strong evidence for experimentally added bias into these samples, they were removed from further analysis, reducing the sample size from 50 to 45.

The GeoChip5 array embodies a range of gene categories spanning nutrient cycling to metal resistance. The underlying microbial processes affecting the distribution and abundances of the different categories in soils may be dissimilar. As such, combined (total array) analysis may result in a high degree of total 'noise'; i.e. the responses within any one category may possibly obfuscate others. As such, analysis was performed on six different 'gene set' groupings. These included: (1) the total array data (all probe sets); and then broad groups of functions associated with (2) nutrient cycling (C, N, P, S, etc); (3) organic remediation; (4) metal homeostasis; (5) stress and secondary metabolism; and (6) others (virulence, movement, phylogenetic markers, clustered regularly interspaced short palindromic repeats [CRISPR], electron transfer and any others).

In addition, for each gene set, the data were considered at three 'aggregation levels': the probe level (e.g. probe #365896421; this is for the nirK gene from *Bradyrhizobium* sp. strain STM 3843); up to the gene level (e.g. all probe spots covering detection of nirK genes); and subcategory 1 (all metabolic genes associated with denitrification).

The aggregation of data to higher levels may potentially result in unintended bias. For example, for some processes (usually not well understood at a DNA level), only a few probe sets are represented on the array. The aggregation of these data has little total effect, and adds bias when compared with the aggregation of processes for which many probe sets are represented. To deal with this, the data were put on equivalent scales using normalisation ($x = 0$, $\sigma_x = 1$).

The relative effects of grouping data to gene sets and at different levels of aggregation were investigated. For the data at the 'probe set' aggregation level, the data were compared with and without normalisation of the variables. For each of the 24 data sets, a distance matrix of functional gene similarity between the soil samples was generated using Euclidean distances. Correlations were made among these distance matrices using Spearman's rank method, and a 2nd stage resemblance matrix generated (Clarke et al. 2006). Distances

among samples (2nd stage) were visualised by non-metric multidimensional scaling (nMDS) ordination. The importance of 'gene set' and aggregation were formally tested by two-way analysis of similarities (ANOSIM; Clarke 1993), with effects tested for significance against a permutation generated null-distribution (999 times).

In subsequent analysis, the data for the 45 soil samples were investigated at probe, gene and subcategory aggregation levels for each of the six gene sets. When testing was conducted with the associated soil edaphic and environmental properties, this was conducted on the entire set of 43 variables described by Wakelin et al. (2013b). The following variables were corrected for skewing by ln-transformed prior to analysis: Olsen P, sulphate-S, Al (CaCl extractable), C:N ratio, eMn, eZn, eCu, eCo, total P, total Ca, total Cu, total Co, K, Mg, Na, rainfall and elevation (elev). The variables were normalised as before.

The generation of resemblance matrices, clustering, SIMPROF testing, 2nd stage analysis, ANOSIM and nMDS ordination were conducted in PRIMERv7 (PRIMER-E).

## Analysis approach

The latest versions of the GeoChip array (generation 5) have 161,962 probe sets covering more than 150,000 genes and spanning 410 gene categories from nutrient cycling to antibiotic resistance (Tu et al. 2014). As such, the information provided through analysis of various subsets of the GeoChip data can provide information for a wide range of different studies, several of which are under way. In order to describe a structured approach for data analysis, spanning multiple levels (probe, gene and family), analysis of GeoChip data within this study was deliberately generalised for the analysis of the alpha and beta community diversity. Then, a specific subset of genes associated with microbial production of plant growth regulatory compounds (e.g. hormones) was used to model abundance-based and network analysis.

## Alpha diversity

Alpha diversity effectively reduces the entire data for each sample (many thousands of data points) into a single 'diversity' value, generally indicative of ecosystem richness, evenness or a composite value. The resulting statistic can then be evaluated using univariate approaches. To demonstrate this, alpha diversity for each sample was measured through calculation of the Shannon ($H'$; log e) and Simpson's ($1-\lambda$) indices; these were derived from the total GeoChip data sets. Unbalanced ANOVAs were used to test whether alpha diversity varied between levels of land-use intensification and soil type. For land-use intensification, there were 18 dairy and 27 'other' samples. Testing for soil-type influences was conducted only on soils that were represented with a reasonable level of replication: Brown (13), Pallic (7) and Recent (16) (Hewitt 1998).

Environmental and edaphic variables that explain the most variation in the Shannon index were selected using stepwise linear regression, for a maximum of 10 steps. At each step, variables were added or dropped based on the ratio of residual mean squares. Stepwise regressions were performed for each of the gene-set groupings and at each level of aggregation, and the results synthesised into a summary probabilities table. These analyses were conducted in Genstat 17 (VSN International 2014).

### Beta diversity

In comparison with alpha diversity, beta diversity considers the multivariate composition (assemblage of species or functions) within the ecosystem. Resemblance matrices (Euclidean distance on normalised variable data) were generated for each of the gene sets across the aggregation levels. Permutational, multivariate analysis of variation (PERMANOVA) was used to determine the influence of land use and soil type on beta diversity (i.e. compositional variation) in functional genes across the samples. The model included land use and soil type as fixed factors, and tested for main and interactive effects of these under a reduced model with type III sums of squares and fixed effects sum set to zero for mixed terms. A full description of the approach is outlined by Anderson et al. (2008).

Links between the GeoChip data sets to soil and environmental variables were made using multivariate stepwise analysis, BVSTEP (Clarke & Warwick 1998). BVSTEP uses a combination of forward selection and backward elimination to stepwise select combinations of explanatory variables that maximise correlation (Spearman's rho) to the fixed GeoChip resemblance data matrix. As for the alpha-diversity stepwise regression, summary output tables were generated by identifying variables contributing to each of the final models. Analyses were conducted in PRIMERv7 (PRIMER-E).

### Gene abundance analysis: plant growth regulatory genes

The abundance of the plant growth regulatory genes (PGRGs) was used as an example gene set for analysis. The genes comprising these include sped (spermidinie synthase), Ipya (indole-3-pyruvate decarboxylase), spe (spermine), Eth (ethylene cycling), Nep (necrosis- and ethylene-inducing protein-inducing proteins) and Cks (xanthine dehydrogenase). The 257 probe intensity values for the individual genes were aggregated (summed) to a single PGRG value from each soil.

Assessment of soil group and land-use effects on PGRGs was conducted in GenStat 17 using unbalanced ANOVAs. For the soil group test, only comparisons between Pallic, Brown and Recent soils were made as the other groups had low representation. Associations between plant hormone genes with individual soil and environmental variables were made using linear regresssion (as before).

### Molecular ecological network analyses

Ecological networks are an interpretation of biological interactions where the nodes (species/biological component) are connected according to strength of association. These are often used for analysis of among-species relationships, but are increasingly used for functional interpretation of ecosystem networks, including those generated by molecular approaches (summarised in Deng et al. 2012). As a demonstration of functional molecular ecological network (fMEN) analysis, the gene set associated with plant growth regulatory genes was analysed (see previous section) and comparison between the two farming systems (dairy vs 'other') undertaken.

For analysis, genes were excluded that were not present in >50% of the samples. An adjacency matrix was created based on random matrix theory (Deng et al. 2012) with threshold (cut off) of 0.81 for 'other' and 0.89 for dairy. These thresholds were then

used to cut off over the correlation matrices. The network plots were produced in Cytoscape (Shannon et al. 2003) and spaced according to the Spring-Electric algorithm of the AllegroLayout app.

## Results

We demonstrate a range of methods for the analysis of functional molecular ecology data with the focus on application to pastoral (and other agricultural) farming systems. The analysis approach spans simple reduction of the complex data sets to a single (univariate) statistic (e.g. a diversity index) for each sample (and the relatively straightforward analysis of this datum), through to more complex analysis that preserves the inherent multivariate nature of the data and explores relationships among the various functional elements (probes to gene families) of the system. Combined with a thorough understanding of the farming system and monitoring of the desired functions (e.g. disease suppression or nutrient-use efficiency), the approach provides a structured pathway in which functional soil molecular ecology can increase on-farm profitability and sustainability (Figure 1).

To demonstrate the approach, we undertook exemplar analysis of GeoChip5 data from 50 New Zealand pasture soils (Wakelin et al. 2013b). Given the importance of ecosystem-based approaches in translating soil functional ecology to impacts on ecosystem processes, analysis pipelines that can preserve the complexity of the ecosystem and map this against soil and environmental variables will have far greater power than less complex analysis. A conceptual approach to ecological genomics for pasture farming systems is presented in Figure 1. This needs to be underpinned by appropriate considerations of data handling and pre-treatment.

### *Data handling and pre-treatment*

Clustering of the entire GeoChip5 data set showed the samples pre-treated for WCGA significantly out-grouped from the wider data set ($P < 0.05$; Figure 2A). As the influence of the phi29 amplification was highly significant ($P < 0.01$) there was justification for removal of these five samples from ongoing analysis.

Gene sets and aggregation level were both significant factors affecting similarity in the GeoChip data sets. PERMANOVA analysis indicated that the aggregation of the data from probe through to subcategory level was a stronger driver ($\sqrt{CV} = 0.127$, $P = 0.001$) than partitioning the entire data set into separate gene sets ($\sqrt{CV} = 0.033$, $P = 0.032$). The influences of these can be seen in the 2nd stage nMDS ordination (Figure 2B). The influence of aggregation separates the data across the 'x-axis', and the separation of data into different gene sets has an influence within each aggregation group (diagonal separation).

The order of separation among the gene sets was conserved among the 'gene' and 'subcategory' aggregation groups; it is likely that analysis of the functional array data at either of these levels will result in similar outcomes. At the probe level of aggregation, there was little influence of separation of the data into gene groups. An approach based on analysis of the data at gene and/or subcategory levels, with separation into the different gene
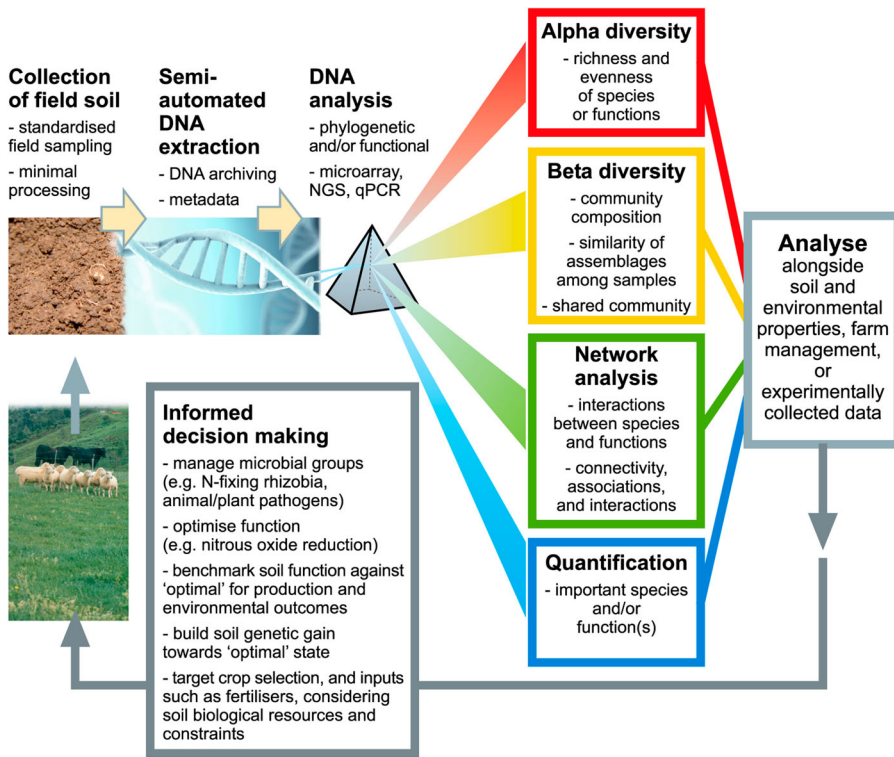
**Figure 1.** Pipeline of analysis of environmental genomics for applied pastoral productive and environmental outcomes. Collection of soil for environmental genomics can (and should) be made alongside analysis of other soil properties (pH, C, N, P, S, etc). Standard field sampling (corer with numerous sub-samples across the field) and homogenisation of the soil sample allows for representative 'DNA' to be extracted. Extraction of eDNA from soils is now routine, with commercially available kits and service providers. Analysis of eDNA can be conducted using a range of different methods, most of which are accessible through service providers. Routine analysis, such as qPCR, can be conducted by most biologically focused research institutes. The approach for data will be highly dependent on the intended use for the researcher or farmer. For example, understanding how farm management pathways result in emergence of soil-borne disease suppression; this will require highly complex analysis based on ecosystem reconstruction. Many farmers, however, maybe interested in detection of key genes of interest for their farming system; e.g. are there sufficient rhizobia in the soil to enable clover nodulation, or what is the disease and pathogen load of a soil before pasture renovation (allowing for targeted pesticide use)? These analyses, therefore, enable 'informed decision making' that takes into consideration soil biological resources. The use of this approach builds, over time, an understanding of the normal operating range of farming systems, identifies new opportunities to alter the trajectory of emergence of functions, and provides a basis to benchmark performance and ideally build genetic gain.

families, provides the greatest separation among samples and allows for analysis with the most power to discriminate effects.

## *Alpha diversity*

For all six gene set groupings, there was a very high degree of correlation (Pearson) between Simpson's and Shannon diversity indices. The mean correlation (*r*) among the
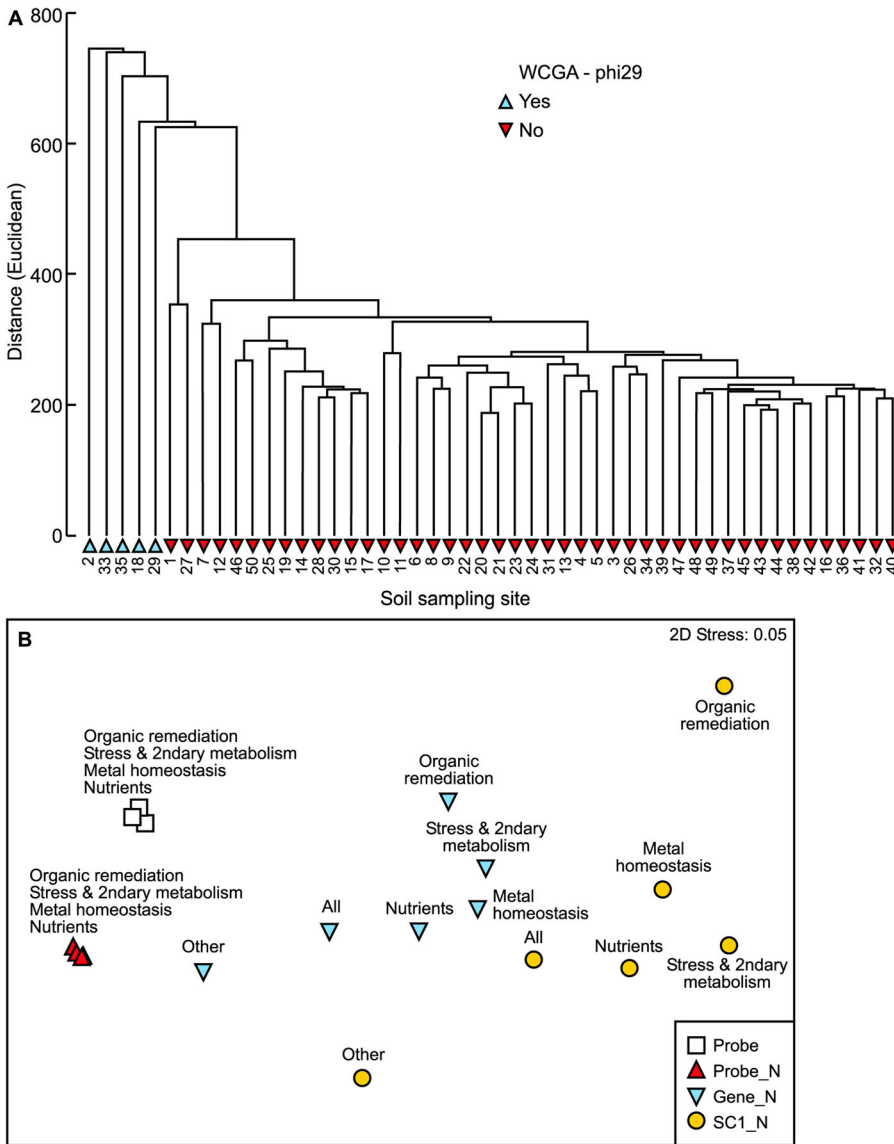
**Figure 2. A,** Clustering (group average, hierarchical agglomerative method) of samples showing the very strong effect of whole community genome amplification on the similarity and grouping of samples; **B,** second-stage nMDS ordination showing effects of separation of the data into different gene sets, and aggregation from probe values to subcategories. Treatments with an 'N' suffix indicate the data have been normalised. SC1, subcategory level 1 aggregation of the data.

comparisons was 0.972 and all were highly significant ($P < 0.0001$). As such, only the data for the Shannon index ($H'$) were considered.

The effect of land-use intensification (dairy vs 'other') on Shannon's diversity was compared across the six gene sets at the probe levels. There was no evidence of land-use effects on diversity for 17 of the 18 tests. The exception was stress and secondary metabolism at the subcategory level, where a $P$ value of 0.077 was calculated. However, the differences in

mean diversity between the dairy ($H' = 2.634$) and 'other' ($H' = 2.633$) were trivial. Similarly, there was no evidence for soil-type influences on Shannon's diversity. Indeed, the average $P$ value across the 18 tests was 0.839.

The summary results for the regressions of the Shannon index to soil and environmental variables are given in Table S1. Of the 43 variables, there was a highly consistent relationship with total cobalt (Co) and solar radiation to diversity ($H'$) across the gene sets and aggregation levels (Table S1). These relationships were positive; i.e. diversity increased with total Co and sunlight (a driver of plant growth and below ground potential productivity).

## Beta diversity

The summary influences of land use, soil type and their interaction on GeoChip array data are given in Table 1. There was strong support for the influence of land use type (dairy compared with other input systems) on functional genes (Table 1). This effect was weakest when the entire data set was considered, however when partitioned into different

**Table 1.** Summary PERMANOVA results table testing the influence of land use and soil group on similarity in function (GeoChip5 data) of pasture soil ecosystems.

| All | Subcategory $\sqrt{CV}$ | $P$(perm) | Gene $\sqrt{CV}$ | $P$(perm) | Probe $\sqrt{CV}$ | $P$(perm) |
|---|---|---|---|---|---|---|
| **Land use** | 5.239 | 0.116 | 13.586 | 0.109 | 80.272 | **0.069** |
| Soil group | −1.200 | 0.525 | 1.487 | 0.457 | 32.008 | 0.314 |
| Interaction | 5.670 | **0.012** | 13.883 | 0.116 | 77.853 | **0.099** |
| Residual variance | 10.812 | | 30.131 | | 225.030 | |
| **Nutrients** | | | | | | |
| Land use | 2.547 | **0.079** | 6.227 | **0.088** | 40.984 | **0.067** |
| Soil group | −0.595 | 0.492 | −1.813 | 0.529 | 14.360 | 0.358 |
| Interaction | 2.737 | 0.111 | 6.500 | 0.118 | 39.505 | 0.103 |
| Residual variance | 4.848 | | 13.228 | | 114.940 | |
| **Metal homeostasis** | | | | | | |
| Land use | 2.528 | **0.094** | 5.115 | **0.091** | 41.533 | **0.069** |
| Soil group | −0.632 | 0.483 | −1.505 | 0.531 | 16.557 | 0.303 |
| Interaction | 2.797 | 0.117 | 5.473 | 0.108 | 40.228 | **0.090** |
| Residual variance | 4.916 | | 10.465 | | 116.750 | |
| **Organic remediation** | | | | | | |
| Land use | 1.612 | 0.102 | 4.791 | **0.081** | 24.065 | **0.072** |
| Soil group | −0.487 | 0.544 | −0.579 | 0.463 | 8.414 | 0.333 |
| Interaction | 1.424 | 0.145 | 4.561 | 0.118 | 22.967 | 0.113 |
| Residual variance | 2.901 | | 9.624 | | 65.184 | |
| **Stress and 2nd metabolism** | | | | | | |
| Land use | 2.485 | 0.113 | 5.866 | **0.087** | 33.733 | **0.075** |
| Soil group | −0.384 | 0.461 | 0.443 | 0.431 | 14.483 | 0.317 |
| Interaction | 2.737 | 0.108 | 6.263 | **0.086** | 33.063 | **0.091** |
| Residual variance | 4.947 | | 11.526 | | 94.457 | |
| **Other** | | | | | | |
| Land use | 2.408 | 0.126 | 7.892 | **0.090** | 36.358 | **0.077** |
| Soil group | −0.549 | 0.469 | 2.811 | 0.360 | 16.237 | 0.263 |
| Interaction | 2.746 | 0.112 | 7.779 | 0.104 | 35.518 | 0.104 |
| Residual variance | 6.028 | | 19.948 | | 103.090 | |

($\sqrt{}$)CV is the square root of the component of variation, which is a dataset-dependent measure of the effect of size in units of the community dissimilarities (i.e. increasing positive values); negative values indicate zero components (Anderson et al. 2008)
$P$(perm) is the probability estimate ($P$) of the effect occurring by chance. Probabilities were estimated by permutation (perm); unique permutations exceed 998 in all tests.
Bold values are $P < 0.1$

gene sets, a consistent influence of land-use intensification was observed across the soil functional groups (Table 1). Furthermore, the likelihood of land use having a significant effect on the soil gene groups ($P_{perm}$ values) was generally greater when tested at probe or gene level compared with data aggregated to subcategory level (Table 1). Across all data sets analysed, the treatment factor 'soil type' did not contribute to explaining a significant component of the variation (Table 1).

A subset of soil and environmental variables were found to be correlated with variation in GeoChip data (Table 2). The most frequently contributing variables were total calcium, total cobalt and sulphate sulphur. The total correlations declined as the data were aggregated from probe to subcategory level, indicating loss of connection of the data with environmental properties.

## Gene abundance analysis: plant growth regulatory genes

The abundance of PGRGs genes did not vary between the major soil types (Pallic, Brown or Recent; $P = 0.91$), nor between high- and low-intensity land uses ($P = 0.419$)

**Table 2.** Summary of multivariate stepwise selection of environmental and soil properties that best explain (Spearman correlation) variation in the GeoChip5 data among gene groups and with increasing level of aggregation.

| | All data | Nutrient cycling | Metal homeostasis | Organic remediation | Stress and 2nd metabolism[a] | Other | Total |
|---|---|---|---|---|---|---|---|
| **Subcategory 1** | | | | | | | |
| Total Calcium | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 6 |
| Total cobalt | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 6 |
| Sulphate-S | ✓ | ✓ | ✓ | | ✓ | ✓ | 5 |
| AMN:Total N[b] | ✓ | | ✓ | ✓ | ✓ | ✓ | 5 |
| Potassium | ✓ | | ✓ | ✓ | ✓ | ✓ | 5 |
| Solar radiation | | | ✓ | | | | 1 |
| Correlation ($\rho$) | 0.204 | 0.200 | 0.196 | 0.202 | 0.186 | 0.192 | |
| **Gene** | | | | | | | |
| Total Calcium | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 6 |
| Total cobalt | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 6 |
| Sulphate-S | ✓ | ✓ | ✓ | ✓ | | | 4 |
| AMN:Total N | ✓ | ✓ | | ✓ | | | 3 |
| Potassium | ✓ | ✓ | | ✓ | | | 3 |
| C:N ratio[c] | | ✓ | | ✓ | | | 1 |
| Correlation ($\rho$) | 0.214 | 0.214 | 0.205 | 0.187 | 0.188 | 0.225 | |
| **Probe** | | | | | | | |
| Total Calcium | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 6 |
| Sulphate-S | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 6 |
| TBS | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 6 |
| Elevation | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 6 |
| C:N ratio | ✓ | ✓ | ✓ | | ✓ | ✓ | 5 |
| Total cobalt | ✓ | ✓ | | ✓ | ✓ | ✓ | 5 |
| AMN:Total N | | | ✓ | | | ✓ | 2 |
| Correlation ($\rho$) | 0.250 | 0.260 | 0.257 | 0.238 | 0.247 | 0.24 | |

[a]Stress and secondary metabolism related genes.
[b]Ratio of anaerobically mineralisable nitrogen (organic N) and total N.
[c]Carbon to nitrogen ratio.
TBS, total base saturation.
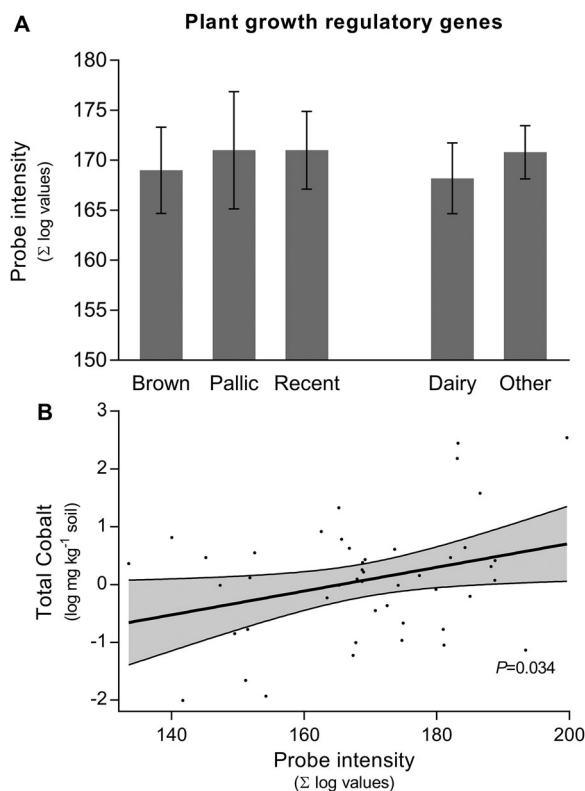✓ indicates that the associated variable was included as part of the explanatory selection.

**Figure 3. A,** Influence of soil type and land use on abundance of genes associated with production and catabolism of microbially produced plant growth regulatory genes (PGRGs). Error bars represent SEM; **B,** linear regression between soil total cobalt concentration and PGRGs. The 95% confidence interval is given in the shaded area.

(Figure 3A). However, there was a significant ($P = 0.034$), albeit weak, positive relationship between PGRG abundances and the total cobalt content of soils ($P = 0.034$; Figure 3B).

## Molecular ecological network analysis

Analysis of the functional molecular ecological networks showed strong differences in connectivity (association) of the plant growth regulatory genes between the two land uses. The genes in pastoral farming systems other than dairy ('other') co-occurred in samples to a much greater degree than those in dairy soils (Figure 4). These differences in degree of fMEN associations are supported by various metrics, such as average connectivity and total links among each network.

## Discussion

We describe a pathway of analysis that may be adopted to provide structured analysis of functional soil molecular data, and identify important points of consideration when conducting the analysis. The exemplar data were based on a high-density, DNA microarray
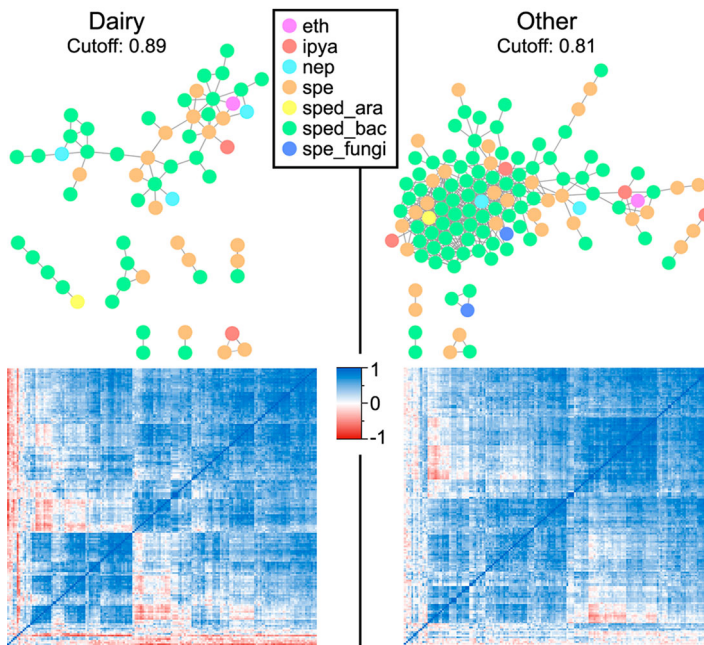
**Figure 4.** Network analysis of plant growth regulatory genes (PGRGs) in soils from under dairy or 'other' land use. Each node signifies a PGRG. The correlation (heat maps) underpinning the network graphs, with hierarchical clustering of the genes, are given below.

system; these data provide a wealth of information on the functional properties of New Zealand pastoral soil ecosystems. The analysis of this data set can be used to provide insights into soil functions associated with productivity and environmental sustainability of pastoral agriculture. In addition to understanding the responses of individual gene sets of interest to a research project, the approach can be used to benchmark the health of soil ecosystems, and following from this build genetic gain in soil ecosystems. Given the extent of pastoral agriculture as a percentage of New Zealand's total land area (c. 30%), even minor gains will translate to large influences if widely adopted. Furthermore, benchmarking and managing soil ecosystems for optimal outcomes can provide the evidence base to support the use of environmental provenance as a value-enhancing attribute for export markets.

While GeoChip is used as the research tool in this study, there are a range of ecological genomics methods available. These include various next generation sequencing platforms through to high-density oligonucleotide arrays for phylogenetic or other functional analysis (van Straalen & Roelofs 2007; Nannipieri et al. 2014). The various advantages, limitations and opportunities afforded by these have been widely described (Thomas et al. 2012; Myrold et al. 2014; Nesme et al. 2016). These include technology accessibility, coverage (e.g. number or range of genes detectable), various bias (e.g. detection limited to gene present on an array platform), through to generated data quality, and ability to manage, analyse, interpret and share data.

The GeoChip, for example, is currently not well suited for high-throughput or routine diagnostic use; it has a single provider of the technology and the per-array costs are

relatively high. However, as a research tool it is excellent as a single provider ensures high consistency of data quality generated among samples, allowing for valid comparisons both within and among studies. Furthermore, being an array-based technology, the GeoChip is restricted to detection of the range of genes present on the array; detection of novel functions is not possible and care must be taken to statistically control for bias associated with over-representation of well-characterised or highly common genes. However, using an array-based system expedites analysis of the data as the probes, genes and functions are already characterised. Accordingly, the bioinformatic analysis has a clear path, is fast and produces consistent results. Overall, the most appropriate platform for pasture soil ecological genomics will depend on multiple factors and there are significant trade-offs among the various technology platforms. Regardless of the platform used, a highly standardised approach to the analysis and processing of data within an experimental study is critical. We provide an example of this (Figure 1); however, it requires validation across other data sets and may require adaptation based on application of different technology platforms (e.g. NGS).

The analysis of large multivariate data sets can be daunting. However, we show that a range of methods can be used, from simple reduction of the complex data sets to a single (univariate) statistic (e.g. a diversity index) for each sample and the relatively straightforward analysis of this, through to more complex analysis that preserves the inherent multivariate nature of the data and explores relationships among the various functional elements (probes to gene families) of the system. The data analysis in this paper is aimed at providing examples of the approach in a structured manner. However, while not a specific goal of this work, the results from the indicative analysis pathways reveal some interesting insights into soil ecosystem function. In particular, we demonstrate the importance of farm system intensification (land use) as a key factor linked with shifts in soil functional ecology. The groupings of samples by land use were simply based on dairy or 'other' grazing systems (Wakelin et al. 2013b). This is a reasonably arbitrary surrogate for high- and low-input systems, but is reflected by a general trend for increased nutrients (primarily N and P) in dairy classified soils (Wakelin et al. 2013b). There is, however, a continuum of intensification between low- and high-input systems, where high rates of stock grazing and inputs occur outside of dairy and vice versa. As such, the emergence of consistent 'land-use' effects from this preliminary analysis is important, as it demonstrates how farm management directly impacts on below-ground ecosystem functional composition. The challenge is to link these data to processes within the soil ecosystem and then develop models based on these relationships within a farming-systems context.

Another key outcome was the complete lack of association of 'soil group' as a driver of soil ecosystems (although not formally investigated for the fMEN analysis). Although soil functions did not consistently vary between groups, the results should be interpreted with caution. Only three major soil groups were assessed (Brown, Pallic and Recent), as these had suitable levels of representation among the total samples which comprised 11 soil groups. Although soil functions were similar when tested among the three soil groups, differences are likely to be evident when comparisons are made between soils that strongly differ in pedogenesis, such as Allophanic, Pumice or organic soils.

The results showed a consistent association between a few soil properties and variation in ecosystem function. Of interest was the link with sulphate-S. When comparing the influence

of grassland intensification on soil ecosystem function (using GeoChipV3), Wakelin et al. (2013c) also identified sulphate-S as one of the key edaphic properties linked with alteration of the soil ecosystem. The consistency of findings across these two studies builds strength that these associative findings may have a true mechanistic basis.

The association between the functional genes data and sulphate-S maybe a general indicator of land-use intensification and concomitant increases in fertiliser inputs. However, to support this we would expect strong associations between sulphate-S and other indicators of fertiliser use, particularly phosphorus. Indeed, while we find a reasonable association between total P and total S ($R^2 = 0.45$), the link between total P and sulphate-S is very weak ($R^2 = 0.08$), as is the association between total S and sulphate-S ($R^2 = 0.28$). As such, although considerable total sulphur has been added to farming soils through anthropogenic inputs, and is therefore a useful indicator of system intensification, these inputs have not linearly translated to increased soil sulphate-S, the biologically available form that is associatively linked with wider functional genomics of the soils. As such, the association with the GeoChip data and sulphate-S indicates that soil biogeochemical processes affecting sulphur availability may affect wider nutrient cycling processes.

Environmental properties had little associative links with the soil biology under pastures. However, key variables associated with rainfall (e.g. soil moisture deficit) confounded inputs of irrigation into many of the farms. Detailed information on these inputs, along with stocking rate, botanical composition, pasture age and rotation, and frequency and types of other farm inputs will add essential knowledge to support further studies. Indeed, the inclusion of as much metadata as practicable will be key to increasing interpretability of future eDNA data from soils. In order to provide opportunities for the pastoral sector, the ecologists, bioinformaticians and statisticians must use farming system knowledge and work alongside farmers.

The coupling of GeoChip data with network analysis provides a powerful means to assess functional interactions with soil ecosystems. In the example given, analysis was made within a single set of genes associated with production of plant growth regulating hormones. Strong changes in network topology, along with differences in degree of network connectivity, were evident between soils from dairy and 'other' land uses. These results show the importance of farming management (intensification) on this single set of genes potentially influencing plant growth. However, the wider analysis of the data set will provide much deeper insights into soil ecosystem function. A particularly important goal is understanding interactions among a number of gene families (e.g. all nutrient cycling genes) to assess how a wider ecosystem state, such as reduced emissions of $N_2O$, develops as an emergent property of many interactions. Currently, for example, quantification of individual N-cycling genes is used to infer the potential of soils to denitrify and release $N_2O$ (e.g. Cui et al. 2016). However, the activity of organisms responsible for these N-cycling processes is also directly integrated with the wider flux of C, P and other elements in the soil ecosystem, and may even be particularly sensitive to microelements such as copper (Cu) (Sullivan et al. 2013) that are essential components of several N-cycling metalloenzymes. Understanding the ecology of N-cycling taxa using N-cycle specific genes thus provides only a relatively narrow view of the wider ecology associated with the N-cycle. It is how these organisms exist and function within a wider ecosystem context that is important. By understanding interactions at this level, critical links (genes or gene families) within networks can be identified that are sensitive to farm

management (Figure 1). Given the level of complexity within soil ecosystems, and the low current level of understanding that has focused on individual components, such critical interactions may currently be entirely unpredictable without application of such approaches.

## Conclusion

Analysis of eDNA offers many opportunities for farmers to monitor and then manage soils for increased system productivity and/or sustainability. The adoption of such tools requires careful application of standardised approaches, extending from the gene-assessment technology, analytical approach, through to interpretation of the genetic data as a predictor of ecosystem function. An approach based on DNA microarrays, such as the GeoChip tool addresses several of these issues, in particular the standardisation of the gene-assessment technology. Based on this, a range of analytical approaches can be used for analysis of the data (as described) and, given the standardised technology basis, a generic pipeline developed. This will enable robust comparison of results across samples, thereby increasing the ability to interpret results against a large database. Currently the array technology described is prohibitively expensive for routine assessment of farm soils. However, given the potential benefits of the approach, the development of an inexpensive 'open-array' system for New Zealand would be of much benefit.

## Supplementary data

Table S1. Association between soil and environmental variables and the Shannon Index (GeoChip5 dataset).

## Acknowledgements

## Disclosure statement

## Funding

## ORCiD

SA Wakelin ⬤ http://orcid.org/0000-0002-1167-8699

# References

Anderson PK, Cunningham AA, Patel NG, Morales FL, Epstein PR, Daszak P. 2004. Emerging infectious diseases of plants: pathogen pollution, climate change and agrotechnology drivers. Trends Ecol Evol. 19:535–544.

Anderson MJ, Gorley RN, Clarke KR. 2008. PERMANOVA+ for PRIMER – Guide to software and statistical methods. Plymouth, UK: Primer-E Ltd.

Barrett BA, Faville MJ, Nichols SN, Simpson WR, Bryan GT, Conner AJ. 2015. Breaking through the feed barrier: Options for improving forage genetics. Anim Prod Sci. 55:883–892.

Cardinale BJ, Harvey CT, Gross K, Ives AR. 2003. Biodiversity and biocontrol: Emergent impacts of a multi-enemy assemblage on pest suppression and crop yield in an agroecosystem. Ecol Lett. 6:857–865.

Clarke KR. 1993. Non-parametric multivariate analysis of changes in community structure. Aust. J. Ecol. 18:117–143.

Clarke KR, Warwick RM. 1998. Quantifying structural redundancy in ecological communities. Oecologia. 113:278–289.

Clarke KR, Somerfield PJ, Airoldi L, Warwick RM. 2006. Exploring interactions by second-stage analysis. J. Exp. Mar. Bio. Ecol. 338:179–192.

Clarke K, Somerfield P, Gorley R. 2008. Testing of null hypotheses in exploratory community analyses: similarity profiles and biota-environment linkage. J Exp Mar Biol Ecol. 366:56–69.

Coleman DC, Whitman WB. 2005. Linking species richness, biodiversity and ecosystem function in soil systems. Pedobiologia. 49:479–497.

Cui P, Fan F, Yin C, Song A, Huang P, Tang Y, Zhu P, Peng C, Li T, Wakelin SA, Liang Y. 2016. Long-term organic and inorganic fertilization alters temperature sensitivity of potential $N_2O$ emissions and associated microbes. Soil Biology and Biochemistry. 93:131–141.

Curtis TP, Sloan WT, Scannell JW. 2002. Estimating prokaryotic diversity and its limits. Proceedings of the National Academy of Sciences USA. 99:10494–10499.

Deng Y, Jiang Y-H, Yang Y, He Z, Luo F, Zhou J. 2012. Molecular ecological network analyses. BMC Bioinformatics. 13:113. http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-113

He Z, Gentry TJ, Schadt CW, Wu L, Liebich J, Chong SC, Huang Z, Wu W, Gu B, Jardine P, et al. 2007. GeoChip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes. The ISME Journal. 1:67–77.

van der Heijden MGA, Boller T, Wiemken A, Sanders IR. 1998. Different arbuscular mycorrhizal fungal species are potential determinants of plant community structure. Ecology. 79:2082–2091.

Herridge DF, Peoples MB, Boddey RM. 2008. Global inputs of biological fixation in agricultural systems. Plant Soil. 311:1–18.

Hewitt AE. 1998. New Zealand soil classification. Landcare Research Science Series. 2nd ed. Lincoln, New Zealand: Manaaki Whenua Press.

Konopka A. 2009. What is microbial community ecology?. The ISME Journal. 3:1223–1230.

Myrold DD, Zeglin LH, Jansson JK. 2014. The potential of metagenomic approaches for understanding soil microbial processes. Soil Sci Soc Am J. 78:3–10.

Nannipieri P, Pietramellara G, Renella G. 2014. Omics in Soil Science. Norfolk, UK: Caister Academic Press, 198 pp.

Nesme J, Achouak W, Agathos S, Bailey M, Baldrian P, Brunel D, Frostegard A, Heulin T, Jansson JK, Jurkevitch E, et al. 2016. Back to the future of soil metagenomics. Frontiers in Microbiology. 7:533. http://dx.doi.org/10.3389/fmicb.2016.00073

Nielsen UN, Ayres E, Wall DH, Bardgett RD. 2011. Soil biodiversity and carbon cycling: A review and synthesis of studies examining diversity-function relationships. Eur J Soil Sci. 62:105–116.

Ogunseitan O. 2005. Microbial Diversity. Oxford, UK: Blackwell Science Ltd, 292 pp.

Paula FS, Rodrigues JL, Zhou J, Wu L, Mueller RC, Mirza BS, Bohannan BJ, Nüsslein K, Deng Y, Tiedje JM, Pellizari VH. 2014. Land use change alters functional gene diversity, composition and abundance in Amazon forest soil microbial communities. Mol Ecol. 23:2988–99. doi:10.1111/mec.12786. Epub 2014 Jun 6.

Roper MM, Gupta VVSR. 1995. Management practices and soil biota. Aust. J. Soil Res. 33:321–339.

Santos BFS, Mchugh N, Byrne TJ, Berry DP, Amer PR. 2015. Comparison of breeding objectives across countries with application to sheep indexes in New Zealand and Ireland. J Anim Breed Genet. 132:144–154.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13:2498–504.

Statistics New Zealand. 2012. Agricultural areas in hectares, by usage and region, at 30 June 2012. http://www.stats.govt.nz/browse_for_stats/industry_sectors/agriculture-horticulture-forestry/2012-agricultural-census-tables/land-use.aspx. 2012 Agricultural Census tables.

van Straalen NM, Roelofs D. 2007. An introduction to ecological genomics. Oxford, UK: Oxford University Press.

Sullivan MJ, Gates AJ, Appia-Ayme C, Rowley G, Richardson DJ. 2013. Copper control of bacterial nitrous oxide emission and its impact on vitamin $B_{12}$-dependent metabolism. P Natl A Sci. 110:19926–19931.

Thomas T, Gilbert J, Meyer F. 2012. Metagenomics – a guide from sampling to data analysis. Microbial Informatics and Experimentation. 2:3. doi:10.1186/2042-5783-2-3

Torsvik V, Ovreas L, Thingstad TF. 2002. Prokaryotic diversity – magnitude, dynamics, and controlling factors. Science. 296:1064–1066.

Tu Q, Yu H, He Z, et al. 2014. GeoChip 4: a functional gene array-based high throughput environmental technology for microbial community analysis. Mol Ecol Resour. 14:914–928.

VSN International. 2014. GenStat for Windows 17th Edition. Hemel Hempstead, UK: VSN International.

Wakelin SA, Barratt BIP, Gerard E, Gregg AL, Brodie EL, Andersen GL, DeSantis TZ, Zhou J, He Z, Kowalchuk GA, O'Callaghan M. 2013a. Shifts in the phylogenetic structure and functional capacity of soil microbial communities follow alteration of native tussock grassland ecosystems. Soil Biol Biochem. 83:568–584.

Wakelin SA, van Koten C, O'Callaghan M, Brown M. 2013b. Physicochemical properties of 50 New Zealand pasture soils: a starting point for assessing and managing soil microbial resources. New Zeal J Agr Res. 56:248–260.

Wakelin SA, Barratt BIP, Gerard E, Gregg AL, Brodie EL, Andersen GL, DeSantis TZ, Zhou J, He Z, Kowalchuk GA, O'Callaghan M. 2013c. Shifts in the phylogenetic structure and functional capacity of soil microbial communities follow alteration of native tussock grassland ecosystems. Soil Biol Biochem. 83:568–584.

Wang J, Van Nostrand JD, Wu L, He Z, Li G, Zhou J. 2011. Microarray-Based Evaluation of Whole-Community Genome DNA Amplification Methods. Appl. Environ. Microb. 77:4241–4245.

Wu L, Liu X, Schadt CW, Zhou J. 2006. Microarray-based analysis of subnanogram quantities of microbial community DNAs by using whole-community genome amplification. Appl. Environ. Microb. 72:4931–4941.

Yan Q, Bi Y, Deng Y, He Z, Wu L, Van Nostrand JD, Shi Z, Li J, Wang X, Hu Z, Yu Y, Zhou J. 2015. Impacts of the Three Gorges Dam on microbial structure and potential function. Sci Rep. 5:8605. http://www.nature.com/articles/srep08605

Yergeau E, Schoondermark-Stolk SA, Brodie EL, Déjean S, DeSantis TZ, Gonçalves O, Piceno YM, Andersen GL, Kowalchuk GA. 2009. Environmental microarray analyses of Antarctic soil microbial communities. The ISME Journal. 3:340–351.