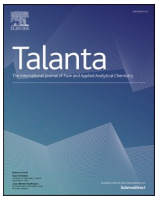
Talanta 274 (2024) 125954

Contents lists available at ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta



Integration and holistic analysis of multiple multidimensional soil data sets

Lisa I. Pilkington [a,b,*], William Kerner [a], Daniela Bertoldi [c], Roberto Larcher [c], Soon A. Lee [d], Matthew R. Goddard [d,e], Davide Albanese [f], Pietro Franceschi [f,**], Bruno Fedrizzi [a,***]

[a] School of Chemical Sciences, University of Auckland, Auckland, 1010, New Zealand
[b] Te Pūnaha Matatini, Auckland, 1142, New Zealand
[c] Food Characterisation and Processing Department, Technology Transfer Centre, Fondazione Edmund Mach, San Michele all'Adige, 38098, Italy
[d] School of Biological Sciences, University of Auckland, Auckland, 1010, New Zealand
[e] School of Life and Environmental Sciences, Joseph Banks Laboratories, University of Lincoln, LN6 7DL, UK
[f] Unit of Computational Biology, Research and Innovation Centre, Fondazione Edmund Mach, San Michele all'Adige, 38098, Italy

ARTICLE INFO

Keywords:
Soil analysis
Statistical workflow
Confounding variables
Variable association
Compositional data
Variable transformation

ABSTRACT

Complex matrices such as soil have a range of measurable characteristics, and thus data to describe them can be considered multidimensional. These characteristics can be strongly influenced by factors that introduce confounding effects that hinder analyses. Traditional statistical approaches lack the flexibility and granularity required to adequately evaluate such matrices, particularly those with large dataset of varying data types (i.e. quantitative non-compositional, quantitative compositional). We present a statistical workflow designed to effectively analyse complex, multidimensional systems, even in the presence of confounding variables. The developed methodology involves exploratory analysis to identify the presence of confounding variables, followed by data decomposition (including strategies for both compositional and non-compositional quantitative data) to minimise the influence of these confounding factors such as sampling site/location. These data processing methods then allow for common patterns to be highlighted in the data, including the identification of biomarkers and determination of non-trivial associations between variables. We demonstrate the utility of this statistical workflow by jointly analysing the chemical composition and fungal biodiversity of New Zealand vineyard soils that have been managed with either organic low-input or conventional input approaches. By applying this pipeline, we were able to identify biomarkers that distinguish viticultural soil from both approaches and also unearth links and associations between the chemical and metagenomic profiles. While soil is an example of a system that can require this type of statistical methodology, there are a range of biological and ecological systems that are challenging to analyse due to the complex interplay of global and local effects. Utilising our developed pipeline will greatly enhance the way that these systems can be studied and the quality and impact of insight gained from their analysis.

1. Introduction

Analysis of complex, multifaceted systems are a constant challenge in the natural sciences. One example of such a multidimensional system is soil. Soils underpin productivity in agricultural ecosystems as they provide and control the cycling of essential nutrients, govern water availability and play roles in carbon sequestration [1,2], with these functions driven by soil microorganisms [3–5]. Two main soil components that are interdependent but comprise complex multivariate measures are microbial and chemical composition. Modern DNA metagenomics sequencing and analytical chemistry approaches are now able to reveal extremely complex and detailed multivariate information on soils, providing multilevel characterisation of this system.

One central area that requires a better understanding is the effect of different agricultural management approaches on soils, particularly the use of synthetic pesticides. The health of agricultural soils has been degraded by intensive farming methods over the last few decades, and changes in practices that increase soil health but still produce sufficient

* Corresponding author. University of Auckland, Private Bag 92019, Auckland, 1142, New Zealand.
** Corresponding author.
*** Corresponding author.
E-mail addresses: lisa.pilkington@auckland.ac.nz (L.I. Pilkington), pietro.franceschi@fmach.it (P. Franceschi), b.fedrizzi@auckland.ac.nz (B. Fedrizzi).

https://doi.org/10.1016/j.talanta.2024.125954
Received 16 June 2023; Received in revised form 9 March 2024; Accepted 18 March 2024
Available online 4 April 2024

yields are urgently required [6]. However, we are still unaware of the long-term sustainability of organic and low-input farming as well as the other challenges and limitations that its use may present. Soils harbour one quarter of the world's biodiversity and approximately 40% of the globe's land-area is dedicated to agriculture. The effects of agrochemicals on the vast array of microbes and how this modulates nutrient availability to affect productivity is poorly characterised generally [1, 7–10]. We are aware of no studies or methods that integrate the effects of agrochemicals on soil biology and chemistry simultaneously, and we address that here.

The lack of these integrated analyses of soil is due to a number of challenges, for example, chemical and metagenomic data have inherently different properties. While chemical data is quantitative, a metagenomic dataset is compositional and this disparity in structure leads to difficulties in relating and integrating these datasets. Additionally, soil, like many matrices, are strongly influenced by a vast array of factors, including climate, topography, weathering and land-use that potentially represent confounding effects, complicating even straightforward "one-layer" analyses. Lastly, if one wants to conduct "multi-layer" analyses to investigate the interplay between both chemical and microbial data, flexible measures of association other than traditionally-used approaches, such as deriving and analysing Pearson correlation coefficient or the Spearman rank coefficient, are required. Unfortunately these alternative measures tend to be sensitive to these aforementioned confounding variables.

In order to improve our analyses and understanding of complex, multilevel systems such as soil, and to answer questions such as the impact of viticultural management practice, new statistical approaches need to be developed. We propose a statistical workflow that enables the appropriate manipulation and analysis of chemical and metagenomic data. This approach removes the effect of confounding variables for subsequent analysis that highlights trends and common patterns in different datasets, investigates and identifies biomarkers in each of the datasets with regard to soil management, and reveals non-trivial associations between variables that goes beyond linear/monotonic associations.

Here we analyse the chemical concentration of elements, assessed through inductively coupled plasma mass spectrometry (ICP-MS) analysis, and fungal biodiversity by ITS barcode metagenomic information to derive fungal Operational Taxonomic Units (OTUs) of New Zealand vineyards soils that have been managed with either organic low-input or conventional input approaches. We use these data to demonstrate our proposed statistical strategy to more holistically analyse complex multidimensional systems.

2. Materials and methods

2.1. Soil samples

Samples were grouped by viticulture practice (conventional and organic) and subregion of Marlborough, New Zealand (see Table 1 for the number of vineyards matching each criteria). For all subregions there was at least one vineyard following each viticulture practice. It should be noted that commercially established vineyards with shared characteristics (i.e. irrigation regime, soil type, number of years certified organic (for the organic vineyards), method of pest control, etc.) were chosen so as to minimise influences of these factors on the analysis.

Table 1
Vineyard classification by subregion and viticulture practice.

	Viticulture practice	
Subregion	Conventional	Organic
Rapaura	3	1
Upper Wairau	3	3
Central Wairau	1	1
Southern Valleys	1	1

2.2. Sampling collection, handling and preparation

Soils were sampled from 14 vineyards from 15th February to 20th March 2013 (Table 1). At each vineyard, three roughly-equal rectangular blocks (sub-sections) were assigned based on soil type and/or topographical lay. After the three subsections were marked, five vines were randomly chosen and samples from under-vine and in the middle of the randomly chosen bay were selected, in each subsection, totalling fifteen soil sample sites per vineyard, five per sub-section. As traditional hand-held soil augers are unable to extract samples on some of the stonier sites in Marlborough, a stainless-steel crowbar and hammer were used to drive the crowbar down to a depth of 40 cm. At this point the crowbar was dug another 1–2 cm and the hole emptied to prevent contamination from soil deriving from shallower depths. The crowbar was then used to loosen some soil and approximately 100 g of non-sieved soil was collected and placed in a bag in a cooler that was lined with ice packs to slow down any biochemical reactions. At the end of the day the samples were frozen at −20 °C until soil preparation for analyses.

Prior to analysis, the soil samples were thawed, dried in an oven at 60 °C for three days and then sieved to 2 mm particles. The stainless-steel sieve was air-blown with compressed air between samples to prevent cross-contamination. The five samples from each subsection were homogenised north to south, thus incorporating the general soil variation in Marlborough, leaving three soil samples per vineyard. The weight of 2 mm soil ranged from 13.96 g to 64.87 g due to varied gravel content across soil types. The samples were then stored at 4 °C until analysis.

2.3. Soil pH analyses

Soil pH was analysed using two methods, one using 0.01 M calcium chloride solution and the other in water. A 1:5 ratio of soil to either (i) $CaCl_2$ (0.01 M) or (ii) ultrapure water (18.2 M Ω/cm, Millipore, Bedford, MA, USA) was used [11], using a modification of the reported method [12], but with a changed solution ratio due to limited soil, and only 1 g of soil as opposed to 10 g. Additionally, samples were stirred for 1 h [13], with 10 min rest prior to measurement.

2.4. ICP-MS analysis

Ultrapure water (18.2 M Ω/cm, Millipore, Bedford, MA, USA), nitric acid (67–69%; Superpure for trace analysis; Carlo Erba Reagents, Cornaredo, Italy), hydrochloric acid (37%), hydrogen peroxide (30%) and glutamic acid 99.5% (Merck, Darmstadt, Germany) were used. ICP Multielement or mono-element standard solutions were used for ICPMS calibration. In detail, ICP Multielement Standard Solution VI and mono-element standard solution of P, S, Cu and Mn (1 mg/mL) were purchased from Merck; Multielement Calibration Standard 1 and 3, Tuning solution (Li, Y, Ce, Tl 10 mg/L) and mono-element standard solution of Hg (10 mg/L) from Agilent Technologies (Santa Clara, CA, USA); ICP Multielement Standard Solution 4 and mono-element standard solutions of Rh, Sc, Tb (1 mg/mL) were from Aristar BDH (Poole, UK); mono-element standard solution of Cs (1 mg/mL) was from Ultra Scientific (Bologna, Italy), Fe 10 mg/mL from CPI international (Santa Rosa, CA, USA). All standard solutions were prepared in 1% HNO_3 and 0.2% HCl solution. All the materials used were previously washed with nitric acid (5%) and rinsed twice with ultrapure water. The soil used as reference material was provided by the "Wageningen evaluating programs for analytical laboratories" [14].

An aliquot (0.5 g) of air dried, sieved <2 mm and ground <0.2 mm soil was added with 1.5 mL of H_2O_2, 4.5 ml of HCl and 1.5 mL of HNO_3 and acid digested in a microwave system (MARS EXPRESS, CEM, Mathews, USA; max temperature 175 °C) using PTFE vessel. The *aqua regia*

extracted samples were diluted 50 times and analysed with an ICP-MS (Agilent 7500ce, Agilent Technologies, Tokyo, Japan) equipped with a collision/reaction chamber for the quantification of 57 mineral elements. In detail, ^{7}Li, ^{9}Be, ^{11}B, ^{27}Al, ^{31}P, ^{49}Ti, ^{55}Mn, ^{74}Ge, ^{85}Rb, ^{88}Sr, ^{89}Y, ^{98}Mo, ^{108}Pd, ^{109}Ag, ^{111}Cd, ^{115}In, ^{118}Sn, ^{121}Sb, ^{126}Te, ^{133}Cs, ^{137}Ba, ^{139}La, ^{140}Ce, ^{141}Pr, ^{143}Nd, ^{147}Sm, ^{157}Gd, ^{163}Dy, ^{165}Ho, ^{166}Er, ^{169}Tm, ^{171}Yb, ^{178}Hf, ^{185}Re, ^{193}Ir, ^{197}Au, ^{201}Hg, ^{205}Tl, $^{206+207+208}$Pb, ^{209}Bi, ^{232}Th and ^{238}U were quantified in "no gas" mode; ^{23}Na, ^{26}Mg, ^{39}K, ^{51}V, ^{52}Cr, ^{59}Co ^{56}Fe, ^{30}Ni, ^{63}Cu, ^{66}Zn, ^{75}As and ^{151}Eu in He mode and ^{40}Ca, ^{71}Ga and ^{78}Se in H_2 mode. In order to overcome possible drift during analytical sequence, an internal standard solution made of Sc, Rh and Tb 1 mg/L was added on-line. Instrumental parameters were optimised daily following the manufacturer's specifications with a Li, Y, Ce and Tl solution in order to maximise sensitivity and resolution and interferences due to double-charged and oxide ions.

The accuracy was verified using a soil provided by the "Wageningen evaluating programs for analytical laboratories" proficiency test. This soil was acid digested and analysed in each sample batch obtaining always Z-score results between ±3 (see Table S1 in the Supporting Information). Repeatability was verified preparing and analysing samples 5 times and obtaining average standard deviation % always below 10% for quantifiable elements except for trace elements Ge, Te, Pd, In and Hf below 20%. All analysed elements were quantifiable in all samples except for Re (<0.001 mg/kg in 2 samples) and Hg (<0.02 in 2 samples), whereas Au (<0.01 mg/kg) and Ir (<0.005 mg/kg) were always under the detection limit. The Limit of Quantification (LOQ) and %RSD for these analyses are provided in Table S2.

The dry combustion method was applied for total Carbon and Nitrogen quantification using a CN elemental analyser (Macro Vario CN, Elementar, Langenselbold Germany) and glutamic acid as reference material and weighting about 200 mg of sample. Method validation metrics, the Limit of Quantification (LOQ) and %RSD, for the analysis of these elements are provided in Table S3.

Prior to statistical analysis the chemical data was checked and cleaned, with elements in the chemical data that showed consistently-low counts, i.e. the lanthanides, removed.

2.5. DNA extraction, library preparation, and sequencing

DNA was extracted using the Zymo Research Soil Microbe DNA MiniPrepTM kits (Zymo Research, Irvine, CA, USA). Fungal communities were characterised and enumerated by 454-sequencing of the D1/D2 region of 26S ribosomal RNA, and amplified using NL1 and NL4 primers described with unique multiplex identifiers added as appropriate [15]. Sequencing this locus provides an effective method for taxonomic identification down to at least genus level as well as the quantification of the relative richness and abundances of fungal communities [16,17]. All PCR products were cleaned using AmpureXP beads (Beckman Coulter, Inc., Brea, CA, USA) and their quality checked by Agilent DNA1000 chips kit (Agilent Technologies, La Jolla, CA, USA). Sequencing was performed on a 454-junior instrument by New Zealand Genomics Limited, New Zealand. Negative controls were always included in the initial PCR steps to ensure no contamination.

2.6. Amplicon sequences analysis

Raw data FASTQ files were analysed using the software pipeline MICCA v. 1.2.0 [18]. After trimming forward and reverse primers, reads shorter than 300 bp and with an expected error rate higher than 0.5% were removed. Sequences longer than 300 bp were truncated. Filtered sequences were clustered into operational taxonomic units (OTUs) at 97% identity using the denovo greedy algorithm available in MICCA. OTUs were taxonomically classified using the Ribosomal Database Project (RDP) Classifier v2.11 [19]. Multiple sequence alignment (MSA) was performed on the denoised reads applying the Nearest Alignment Space Termination algorithm and the phylogenetic tree was inferred using FastTree [20,21]. To compensate for different sequencing depths, samples were rarefied to an even depth of 1150 (without replacement). Samples with less than 1150 reads were removed.

2.7. Statistical methodology

When data are collected there are potentially influential, local variables that may confound statistical analysis. Often, these variables add noise to the data and may confound any signal of the effects of interest. To mitigate the impacts of these confounding variables, various strategies can be used [22], but those utilised are often dictated by the nature of the data, and difficulties arise with certain types of data that are studied, including metagenomic data, which is intrinsically compositional [23–25]. Compositional data exists in a sub-space, the simplex, rather than real Euclidean space, and as such commonly-used metrics, including the aforementioned measures, are not appropriate nor valid [26,27]. This is because, in compositional data, the distance between variables is sensitive and dependent on the presence and absence of other components/variables [26]. As a result, correlation analysis can yield false positive results and multivariate analysis can provide erroneous and incorrect conclusions [27,28]. New strategies are required that appropriately transform data, remove the impact of confounding variables to unveil the impact of global effects in situations where confounding, local effects, are unavoidable.

The statistical strategy described here was designed to be applied to these situations to appropriately analyse complex multidimensional systems. We have developed a data analysis scheme that allows for the removal of influential confounding variables from multilevel chemical and metagenomic data that limit the analysis of factors of interest and in many cases, would otherwise render any efforts to analyse underlying data, futile. By controlling for these potentially confounding variables, the data can then be analysed to identify any patterns and biomarkers that highlight variable associations present with the effect of interest despite any interference from confounding variables. Details of the statistical pipeline are as follows (Fig. 1).

i) *Unsupervised exploratory analysis:* conducted to observe notable trends, outliers and indicate any confounding variables.
ii) *Data decomposition:* data are then decomposed to remove the influence of this confounding factor. If the data is quantitative and non-compositional in nature, this involves deduction of the median value for each level of the factor for each variable. For quantitative compositional data, first the data is offset by 1 and a total sum of squares (TSS) normalisation is carried out. The data is then filtered in order to focus on the OTUs present in a large fraction of the samples (discarding the OTUs which are present in <30% of samples) [29] and a centred log ratio (CLR) transformation [26] is applied. Once transformed, the data can then undergo the decomposition process as described for the non-compositional data.
iii) *Biomarker identification:* build classifying models that provide information on variable importance being strongly desired as these can be identified as biomarkers.
iv) *Variable association:* the biomarkers can be investigated for their any correlations/interrelationships.
v) *Trend identification:* mutual information can identify any variables that have a global association, regardless of their importance for classification.

The proposed statistical methodology allows for the analysis of general trends in the data, the discovery of associations and correlations and the analysis of global effects, in the presence of confounding factors in the original samples. The general approach of this method involves the creation of subgroups according to characteristics that are influenced by local factors. Once these subgroups have been created, the effects of these local factors are able to be removed, allowing for any

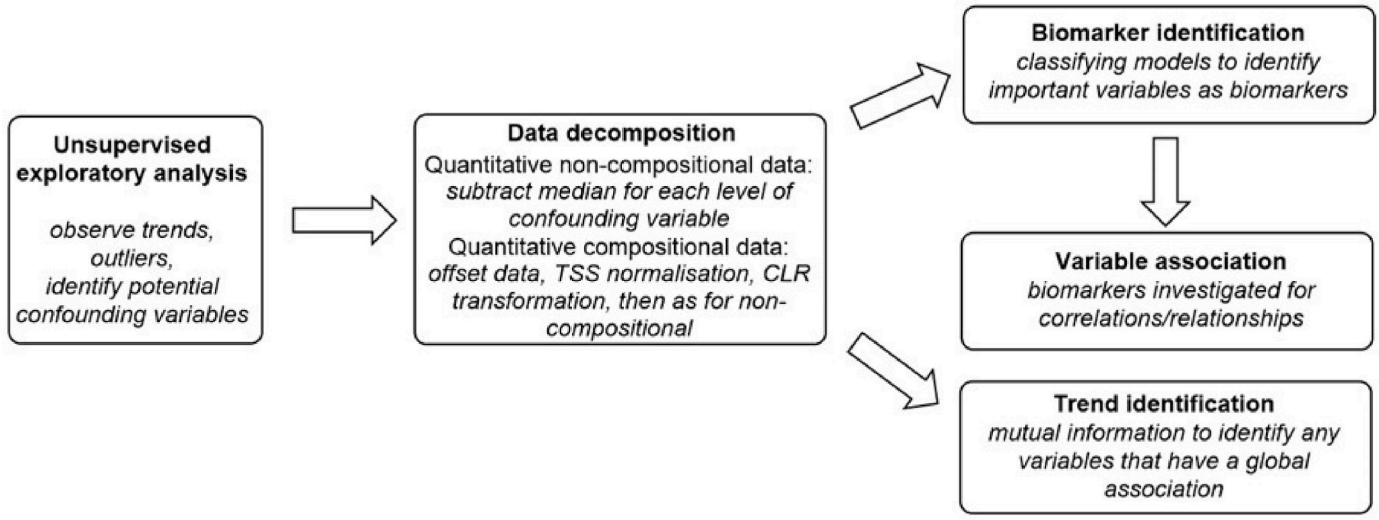


Fig. 1. Depiction of the developed methodology. In the first instance, the collected data will undergo unsupervised exploratory analysis to assess any overall trends, outliers and any observable impacts of a confounding variable. If a confounding variable is identified, its influence can be removed using a data decomposition step, depending on if it is quantitative compositional or non-compositional data. Once the data has been transformed, models can be built to identify variables of importance, biomarkers which can be explored further through association analysis. Mutual information can also be used to identify variables that have a global association.

remaining differences to be attributed to the global effect of interest.

While soil is an example of a complex system that requires this type of statistical methodology, there are a range of biological and ecological systems, including applications in soil, air, water and multiomic analysis, that have properties that are challenging to analyse due to the complex interplay of global and local effects and would also benefit from such an approach. The utility, benefit and validity of using this statistical pipeline is exemplified herein.

3. Results and discussion

To evaluate the utility of the proposed novel analytical approach we analysed soil chemistry and microbiome data to assess the effect of conventional and organic viticultural management practices on soils and test if there are any correlations between the responses of soil chemistry and biology to management approach. A priori we were aware of the possibility potentially complicating this analysis: soil was sampled from different geographical subregions within the Marlborough region in New Zealand, and there are reports that the influence of subregion may be a confounding factor for both soil chemistry and biology that could mask the underlying impact of viticultural management practice [3,4,7,16].

3.1. Unsupervised exploratory analysis

Exploratory analysis of the data was first carried out in the form of principal coordinate analysis, PCoA, using a binary dissimilarity index and principal component analysis, PCA, for the metagenomic and chemical data, respectively (Fig. 2). This unsupervised exploratory analysis was performed to allow for visualisation of the two data sets, particularly to ascertain the major factors which contribute to the dataset variability.

The variability of the metagenomic data was distributed over a large number of components (two of which are shown in Fig. 2A; A1 accounting for 6% of the variability in the data and A3 accounting for 4% of the variability), while the chemical data was largely represented by two principal components (principal component 1 and 2 respectively accounted for 40% and 24% of the total variability in the data; Fig. 2B).

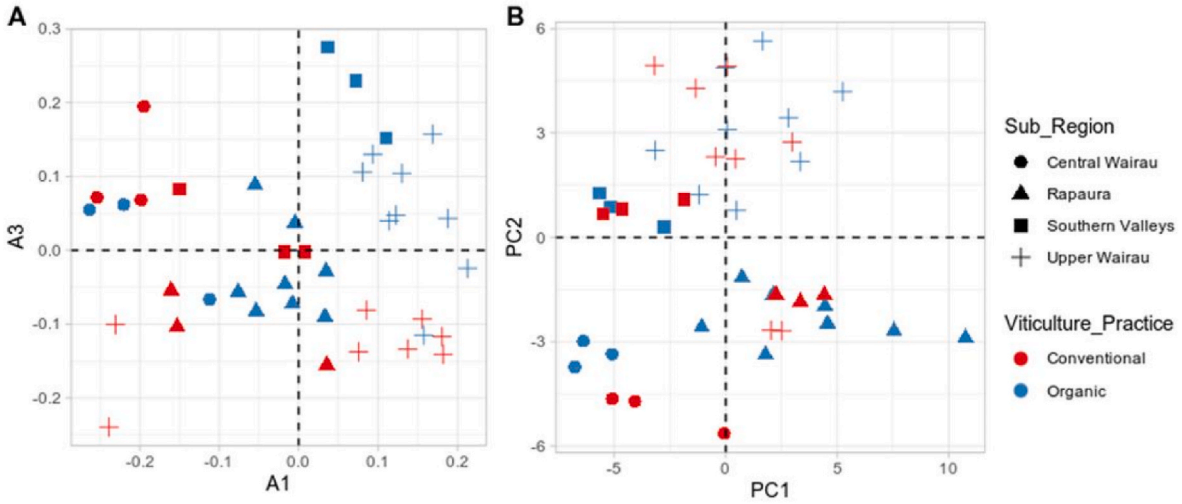


Fig. 2. Results of the exploratory analysis for the metagenomic (PCoA, A) and chemical (PCA, B) data.

While the variability in the metagenomic data is much more complex than the chemical data, it is visually apparent that there were indications of correlations with viticultural management practice in both data sets (particularly for the metagenomic data). However, as suspected, it appears that subregion (symbol in Fig. 2) has an influence in both biology and chemistry data sets, as there is a very clear clustering by sampling location.

3.2. Data decomposition

The next stage in the workflow was to decompose the data sets to remove the influence of this confounding sub-region factor evident in the exploratory analysis.

For the chemical data, the median of each sub-region was deducted for each variable, following the idea underlying ASCA (Analysis of variance – Simultaneous Component Analysis) [30]. For the quantitative and compositional metagenomic data, first the data was offset by 1 and a total sum of squares (TSS) normalisation performed. Subsequent to filtering, a centred log ratio (CLR) transformation [26] was applied. By calculating the TSS, the compositional data are restricted to a space where, in the case of metagenomic data, the sum of all OTU proportions for a given sample equals to 1 – this approach accommodates varying sampling and sequencing depth [31]. As this normalisation approach is representative of relative information and provides a bounded rather than Euclidean space, a CLR transformation enables subsequent multivariate methods to be applied [26,32,33]. Once transformed, the data was then decomposed as described for the chemical data. Following the removal of the subregion effect, the transformed data underwent exploratory analysis using PCA which indicated the subregional effect was removed (Figure S1).

3.3. Biomarker identification

In order to test if soils could be effectively classified by viticultural management based upon the untransformed (Fig. 3, red and blue) and transformed data (Fig. 3, green and purple) data, random forest [34] classifier models were employed. While not part of the proposed workflow, analysis of the untransformed data was included in this instance to demonstrate the effect of removing the confounding factor, particularly on biomarker identification.

Analysis of the performance of 100 such models for each data set, with correct labels (red, green) and random labels (blue, purple) are shown using the Matthew's correlation coefficient (mcc, Fig. 3) [35]. Models exhibited a distinctly enhanced classification efficiency compared to when groups were randomly assigned (median $\approx$ 0, i.e. random prediction, blue and purple; p-value $\approx$0 for Wilcoxon signed-rank tests). This indicates that the random forest models are effective in classifying management practice and that there are sufficient differences in the chemical and metagenomic data between viticultural management practices that can be exploited to build these models. Overall, the classification models constructed using the transformed metagenomic data (Fig. 3 right, green, median $\approx$ 0.8) were better than their counterpart based on the transformed chemical data (Fig. 3, left, green, median $\approx$ 0.3), suggesting that the differences between management practices in the soil fungal communities was more pronounced and greater than the chemical profiles.

While the random forest models based on the untransformed metagenomic data did perform well (Fig. 3, right, red, median $\approx$ 0.7), transforming the data to remove the influence of the subregion confounding factor resulted in a larger classification efficiency for viticultural management practice. Interestingly, this was not seen for the chemical data, where the classification model constructed using the transformed data (Fig. 3, left, green, median $\approx$ 0.3) has a slightly lower classification ability than the untransformed data (Fig. 3, left, red, median $\approx$ 0.5). This may be due to the fact that in transforming the data - removing the subregion median from each value, for each variable - variability in the data is being removed. While this does result in a slight drop in performance of this classification model, the model is still able to make predictions and transforming the data means that conflicting data is removed or minimised in the analysis, something that is of great benefit when identifying biomarkers and variables of interest.

Random forest model performance for the classifiers based on the transformed and untransformed data is, overall, not markedly different. However, it can be seen that through removal of subregion transformation, the classifier becomes slightly less efficient, suggesting that the inclusion of the subregion confounding factor was biasing the analysis.

Random forest models allow specific chemical and metagenomic

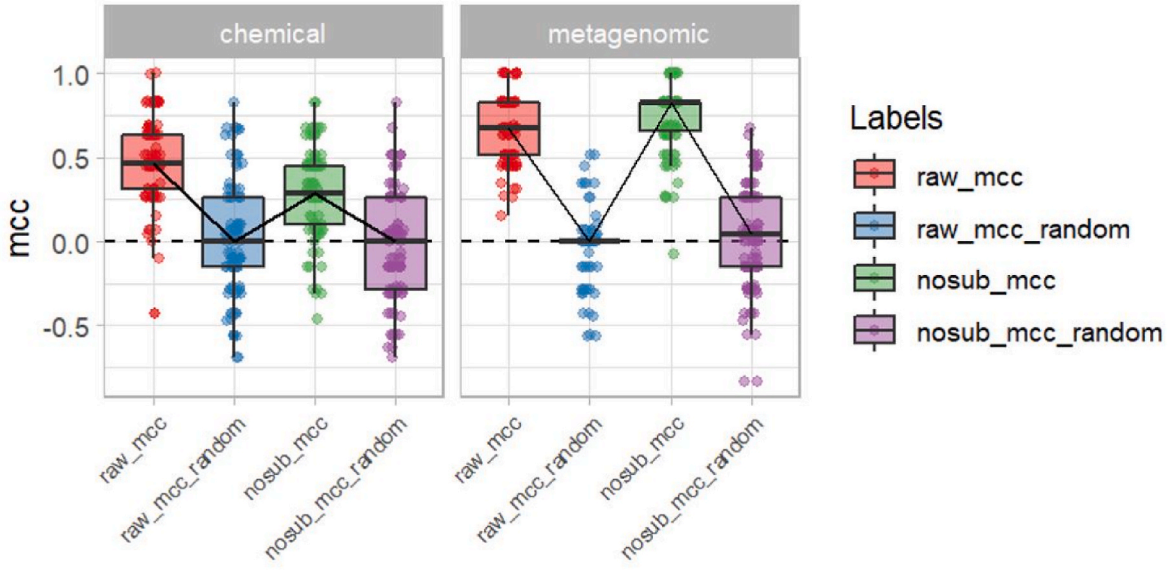


Fig. 3. Mathews correlation coefficients of the 100 random forest classifier models built using the chemical data (left) and metagenomic data (right) with correct viticultural practice labels (red for untransformed data, green for transformed data) and random labels (as the control, blue for untransformed data, purple for transformed data). Matthews correlation coefficient is a measure of classification efficiency, with a value of +1 indicating perfect prediction, 0 indicating random/an ineffectual model and −1 indicating complete disagreement between the observed and predicted class.

variables that differ according to viticultural management practice to be identified, and those with the greatest differences for the transformed data are shown in Figs. 4A and 5A. The benefit and influence of transforming the data is apparent when identifying and analysing the variables deemed most important for the random forest models to classify viticultural practices: the biomarkers (Figs. 4 and 5). While some of the most important variables from the untransformed data (see Figures S2 and S3) are the same for the transformed data (i.e. Se, Sr, Mn, Soil_pH_H20, DENOVO021, DENOVO010, DENOVO053), other influential variables come to light following data transformation (i.e. Mo, Li, DENOVO037, DENOVO001, DENOVO003).

It can also be noted that the variables that had particularly varied and incoherent trends in the untransformed data (i.e. Soil_pH_CaCl, DENOVO03) were not found to be important variables in the classification models built using the transformed data. This indicates that this methodology was useful for identifying true "global" biomarkers with similar trends across the levels of the confounding variable.

Furthermore, the trends between organic and conventional practices for the important variables are shown to be more similar and coherent across the subregions following transformation. This provides confidence in the statistical approach to identify these variables as biomarkers. This is particularly true for the metagenomic data, with all of the important variables except for DENOVO032 showing consistency in trend across all subregions (Fig. 5B), although it is also apparent that the trends are far more conserved for the transformed chemical data (Fig. 4B) than the untransformed chemical data (Figure S2B).

3.4. Variable association

In addition to identifying biomarkers that distinguish viticultural soil from organic and conventional practices, another goal of this analysis was to unearth any links and associations between the chemical and metagenomic profiles of the soil.

Using our workflow, one of the ways in which this can be done is to identify the most important variables in the classification model (i.e. Figs. 4A and 5A) and compare them to identify any potential correlations in a pairwise fashion. An example from this analysis is shown (Fig. 6) where the most influential chemical Se, was investigated for its correlation with some of the most discerning metagenomic variables (i.e. DENOVO021 and DENOVO138) and their potential to distinguish viticultural management practices were correlated. It can be noted that between Se and DENOVO021 there is an overall trend, with two observable clusters present, grouped by viticultural management practice (Fig. 6 left). Results of this investigation suggest a relationship between selenium and this metagenomic marker. This, coupled with naturally low levels of selenium in New Zealand soils [36], is an interesting observation, indicating that changes in selenium levels in New Zealand soil may be manipulated by viticultural practice, and that this is correlated with a change in the fungal composition of soils in vineyards. This is but one example of what can be achieved following this approach to identify links between chemical and biological variables.

That being said, it would not be expected that such a strong association between all pairs of important variables be apparent, as can be seen in relating Se and DENOVO138 (Fig. 6, right).

3.5. Trend identification

The above strategy of investigating links between biomarkers precludes that the variables are biomarkers that are important for the classification of soil as being from either organic or conventional management practices that provides information on what differentiates these two groups. It is also of interest to discover underlying trends and associations between chemical and metagenomic data that hold across all of the samples, regardless of viticultural practice. A greater understanding of this complex system can be attained by having this added capability in the analysis.

Another method to formally identify if there are common associations between any of the metagenomic and chemical measures when the subregion is removed, is by conducting an association analysis using mutual information. A methodology to identify such associations and determine the strength of the relationships was recently reported using the Python-based open-source software, MICtools [37].

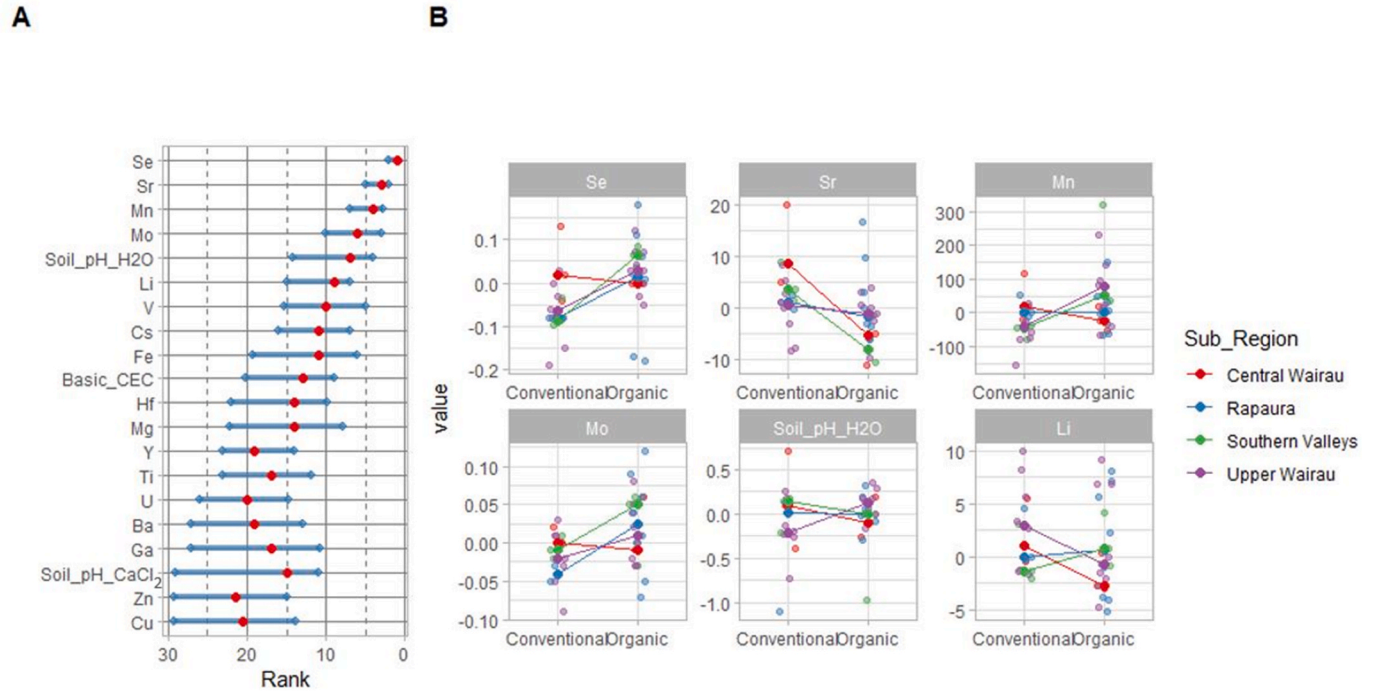


Fig. 4. (A) Median (red) and confidence intervals (blue, 95%) for the ranks of the most important transformed chemical variables in the 100 random forest models, and (B) values for each of the most important transformed chemical variables (according to the random forest models) for the different viticultural management practices, separated by subregion.

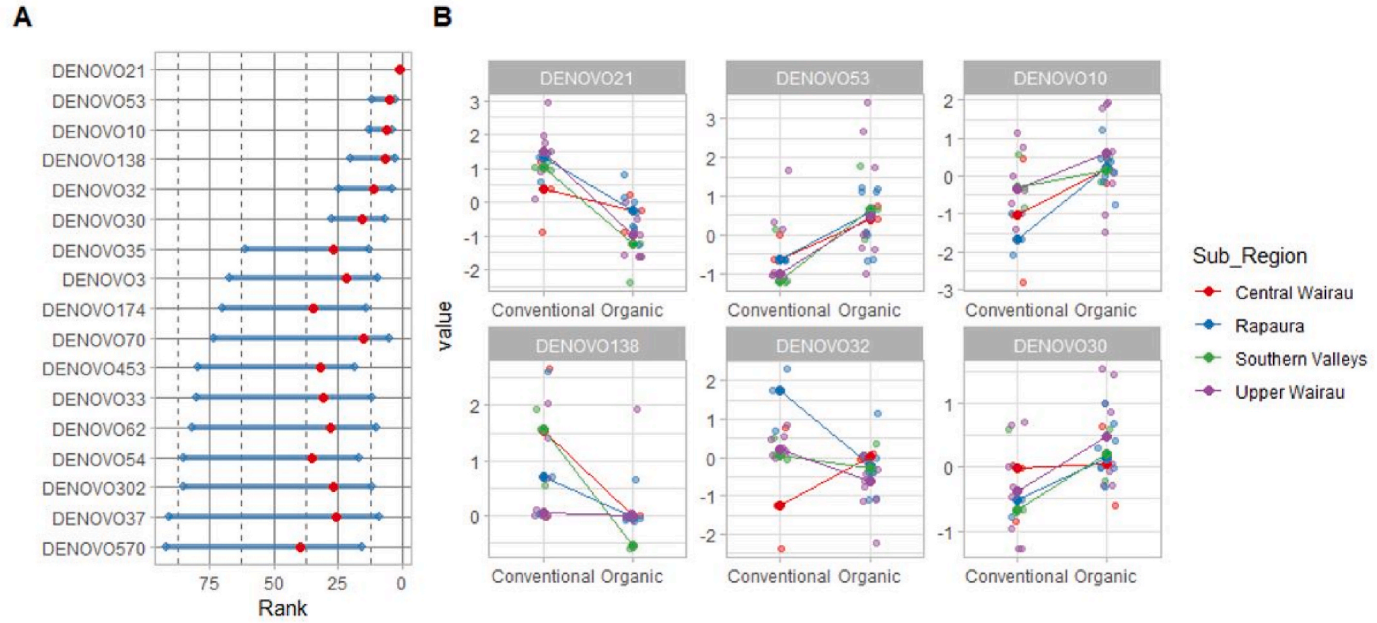


Fig. 5. (A) Median (red) and confidence intervals (blue, 95%) for the ranks of the most important transformed metagenomic variables in the 100 random forest models, and (B) values for each of the most important transformed metagenomic variables (according to the random forest models) for the different viticultural management practices, separated by subregion.

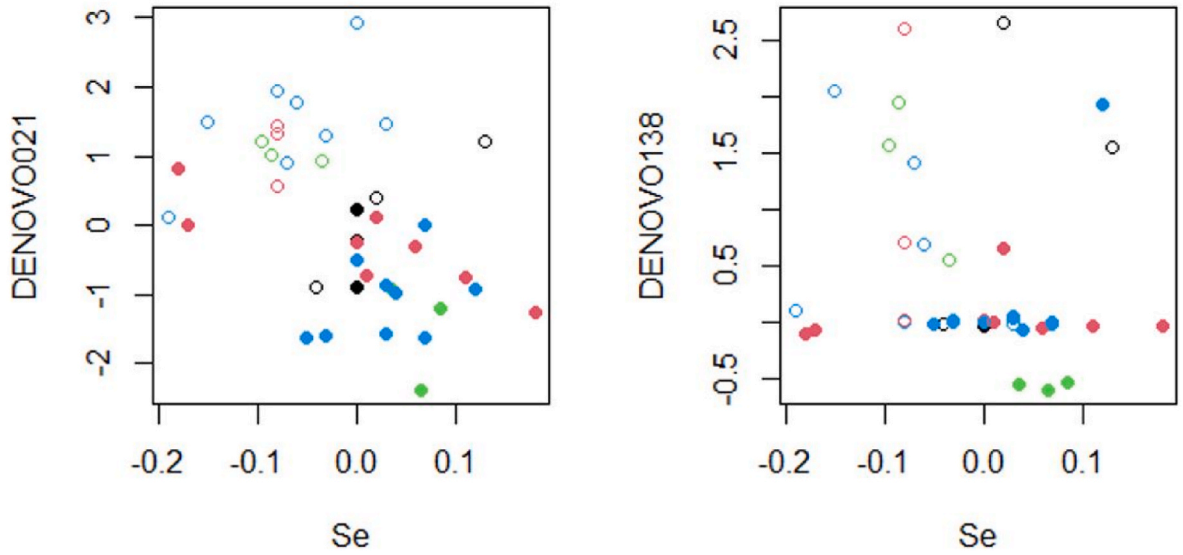


Fig. 6. Exploration of the relationship between the most important chemical- and metagenomic-related variable(s), selenium and DENOVO021 (Fungi; Ascomycota; Dothideomycetes; Pleosporales; Montagnulaceae, left) and DENOVO138 (Fungi; Basidiomycota; Agaricomycetes; Geastrales; Geastraceae; Geastrum, right), respectively, and the impact of conventional (circle) and organic (filled) management practices on their relationship.

Applying this analysis to the soil data demonstrated that this approach was able to identify previously-unknown relationships between variables from the chemical and metagenomic profiles of the soil. An example of variables that had a high value of global association in all samples was C and N and DENOVO02, where one can see a clear decreasing trend between DENOVO002 and C and N levels (Fig. 7A). This trend, however, would have been masked had the effect of subregion not been removed (Fig. 7B). This clearly exhibits how confounding effects, i.e. subregion, can nullify the outcomes of association analysis. This also demonstrates that this analysis strategy allows one to analyse data without the influence of the confounding variables to reveal any links and associations between variables. To otherwise identify and model such a multi-level relationship would require the use of a multi-level analysis (i.e. use of a mixed effect model).

The variables C, N and DENOVO02 are not biomarkers (i.e. their levels are not clearly distinguished between soils of the two management practices). Using mutual information enables discovery of variables with high levels of association, despite not being organic/conventional biomarkers.

4. Conclusion

Here we present an effective statistical pipeline to analyse and study complex systems that are strongly influenced by confounding variables that would otherwise render classification inference very difficult. In this work, we have applied this statistical workflow to soil data from

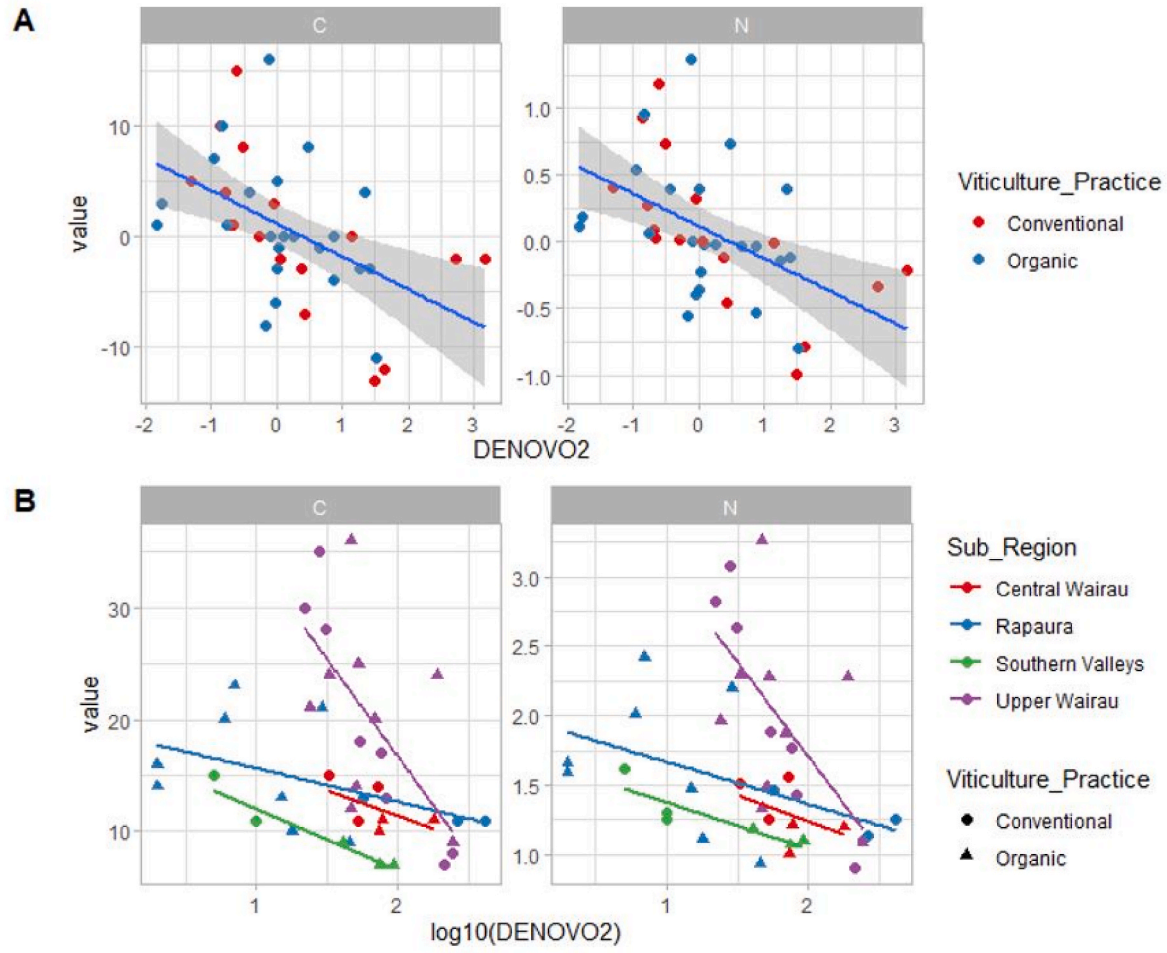


Fig. 7. Exploration of the relationship between strongly-associated chemical- and metagenomic-related variable(s), carbon (left) and nitrogen (right) and DENOVO2, respectively, on the transformed (A) and untransformed (B) data, showing a clear relationship when the data is transformed, that is a very good representation of the relationships observed in the raw data.

vineyards that follow either organic or conventional viticultural management practices. The workflow involves exploratory analysis to identify confounding variables followed by data decomposition (including strategies for both compositional and non-compositional quantitative data) to remove influence of this confounding factor. These data processing methods then allow for common patterns to be highlighted in these datasets, identification of biomarkers and determination of non-trivial associations between variables. While soil is an example of a multidimensional system that requires this type of statistical methodology, there are a range of similar natural science systems that have properties that are challenging to analyse due to the complex interplay of global and local effects and would also benefit from such an approach.

CRediT authorship contribution statement

Lisa I. Pilkington: Formal analysis, Methodology, Writing – original draft, Investigation, Visualization. **William Kerner:** Investigation, Writing – review & editing, Methodology. **Daniela Bertoldi:** Investigation, Validation, Writing – review & editing, Methodology. **Roberto Larcher:** Investigation, Methodology, Resources, Writing – review & editing. **Soon A. Lee:** Investigation, Methodology, Writing – review & editing. **Matthew R. Goddard:** Investigation, Methodology, Resources, Supervision, Writing – review & editing. **Davide Albanese:** Formal analysis, Methodology, Writing – review & editing. **Pietro Franceschi:** Conceptualization, Formal analysis, Investigation, Methodology, Resources, Visualization, Writing – review & editing. **Bruno Fedrizzi:** Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

William Kerner reports financial support was provided by Callaghan Innovation.

Data availability

Data will be made available on request.

Acknowledgments

This project was funded by Callaghan Innovation through an R&D Student Fellowship grant. WK and BF would like to thank Darling Wines for the continued support of this project. The completion of this research would not have been possible without the enthusiasm, cooperation and assistance of the many collaborating participants who allowed access to their land: Kerner Estate, Darling Wines, Churton, Crawford, Delegats, Huia, Pernod Ricard, and Seresin as well as Bruce Miller and Jane Casey, Paul and Jackie Irwin, Tim and Sally Crawford.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.talanta.2024.125954.

References

[1] M. Hartmann, B. Frey, J. Mayer, P. Mäder, F. Widmer, Distinct soil microbial diversity under long-term organic and conventional farming, ISME J. 9 (2015) 1177.

[2] J.W. Doran, M.R. Zeiss, Soil health and sustainability: managing the biotic component of soil quality, Appl. Soil Ecol. 15 (2000) 3.

[3] E. Coller, A. Cestaro, R. Zanzotti, D. Bertoldi, M. Pindo, S. Larger, D. Albanese, E. Mescalchin, C. Donati, Microbiome of vineyard soils is shaped by geography and management, Microbiome 7 (2019) 140.

[4] K.N. Burns, D.A. Kluepfel, S.L. Strauss, N.A. Bokulich, D. Cantu, K.L. Steenwerth, Vineyard soil bacterial diversity and composition revealed by 16S rRNA genes: differentiation by geographic features, Soil Biol. Biochem. 91 (2015) 232.

[5] F. Widmer, F. Rasche, M. Hartmann, A. Fliessbach, Community structures and substrate utilization of bacteria in soils from organic and conventional farming systems of the DOK long-term field experiment, Appl. Soil Ecol. 33 (2006) 294–307.

[6] R.J. Rickson, L.K. Deeks, A. Graves, J.A.H. Harris, M.G. Kibblewhite, R. Sakrabani, Input constraints to food production: the impact of soil degradation, Food Secur. 7 (2015) 351.

[7] P. Giraldo-Perez, V. Raw, M. Greven, M.R. Goddard, A small effect of conservation agriculture on soil biodiversity that differs between biological kingdoms and geographic locations, iScience 24 (2021) 102280.

[8] C.A. Guerra, A. Heintz-Buschart, J. Sikorski, et al., Blind spots in global soil biodiversity and ecosystem function research, Nat. Commun. 11 (2020) 3870.

[9] P. Harkes, A.K.A. Suleiman, S.J.J. van den Elsen, J.J. Haan, M. Holterman, E. E. Kuramae, J. Helder, Conventional and organic soil management as divergent drivers of resident and active fractions of major soil food web constituents, Sci. Rep. 9 (2019) 13521.

[10] P. Morrison-Whittle, S.A. Lee, M.R. Goddard, Fungal communities are differentially affected by conventional and biodynamic agricultural management approaches in vineyard ecosystems, Agric. Ecosyst. Environ. 246 (2017) 306.

[11] A. Al-Busaidi, P. Cookson, T. Yamamoto, Methods of pH determination in calcareous soils: use of electrolytes and suspension effect *Aus*, J. Soil Res. 43 (2005) 541.

[12] L.C.S. Blakemore, P.L. Philip Lee, B.K. Daly, Methods for Chemical Analysis of Soils; Lower Hutt, N.Z, 1987.

[13] B. Lake, Understanding Soil pH. New South Wales Acid, Soil Action Program, Australia, 2000.

[14] D. van Dijk, Wageningen evaluating programmes for analytical laboratories (wepal): a world of experience, Commun. Soil Sci. Plant Anal. 33 (2002) 2457.

[15] C.P. Kurtzman, C.J. Robnett, Phylogenetic relationships among yeasts of the 'Saccharomyces complex' determined from multigene sequence analyses, FEMS Yeast Res. 3 (2003) 417.

[16] P. Morrison-Whittle, M.R. Goddard, Quantifying the relative roles of selective and neutral processes in defining eukaryotic microbial communities, ISME J. 9 (2015) 2003.

[17] M.W. Taylor, P. Tsai, N. Anfang, H.A. Ross, M.R. Goddard, Pyrosequencing reveals regional differences in fruit-associated fungal communities, Environ. Microbiol. 6 (2014) 2848.

[18] D. Albanese, P. Fontana, C. De Filippo, et al., MICCA: a complete and accurate software for taxonomic profiling of metagenomic data, Sci. Rep. 5 (2015) 9743.

[19] Q. Wang, G.M. Garrity, J.M. Tiedje, J.R. Cole, Naïve bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy, Appl. Environ. Microbiol. 73 (2007) 5261.

[20] Jr DeSantis, T. Z, P. Hugenholtz, K. Keller, E.L. Brodie, N. Larsen, Y.M. Piceno, R. Phan, A. Andersen, NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes, Nucleic Acids Res. 34 (2006) W394–W399.

[21] M.N. Price, P.S. Dehal, A.P. Arkin, FastTree 2 – approximately maximum-likelihood trees for large alignments, PLoS One 5 (2010) e9490.

[22] J. Westfall, T. Yarkoni, Statistically controlling for confounding constructs is harder than you think, PLoS One 11 (2016) e0152719.

[23] H. Li, Microbiome, metagenomics, and high-dimensional compositional data analysis, Annu. Rev. Stat. Appl. 2 (2015) 73.

[24] M.S. Kumar, E.V. Slud, K. Okrah, S.C. Hicks, S. Hannenhalli, H. Corrada Bravo, Analysis and correction of compositional bias in sparse sequencing count data, BMC Genom. 19 (2018) 799.

[25] G.B. Gloor, J.M. Macklaim, V. Pawlowsky-Glahn, J.J. Egozcue, Microbiome datasets are compositional: and this is not optional, Front. Microbiol. 8 (2017) 2224.

[26] J. Aitchison, The statistical analysis of compositional data, J. R. Stat. Soc. Ser. B Methodol. 44 (1982) 139.

[27] T.P. Quinn, I. Erb, M.F. Richardson, T.M. Crowley, Understanding sequencing data as compositions: an outlook and review, Bioinformatics 34 (2018) 2870.

[28] K. Pearson, O.M.F.E. Henrici, VII. Mathematical contributions to the theory of evolution III. Regression, heredity, and panmixia. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 187 (1896) 253.

[29] M. Arumugam, J. Raes, E. Pelletier, D. Le Paslier, et al., Enterotypes of the human gut microbiome, Nature 473 (2011) 174.

[30] A.K. Smilde, J.J. Jansen, H.C.J. Hoefsloot, R.-J.A.N. Lamers, J. van der Greef, M. E. Timmerman, ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data, Bioinformatics 21 (2005) 3043.

[31] K.-A. Lê Cao, F.R. Ignacio Gonzalez, S. Dejean, B. with key contributors Gautier, F. Bartolo, P. contributions from Monget, J. Coquery, F.-Z. Yao, Liquet, B. mixOmics, Omics data integration project, R package version 6.1.1. https://CRAN.R-project.org/package=mixOmics, , 2016.

[32] K.-A. Lê Cao, M.-E. Costello, V.A. Lakis, F. Bartolo, X.-Y. Chua, R. Brazeilles, P. Rondeau, MixMC: a multivariate statistical framework to gain insight into microbial communities, PLoS One 11 (2016) e0160169.

[33] P. Filzmoser, K. Hron, C. Reimann, Principal component analysis for compositional data with outliers, Environmetrics 20 (2009) 621.

[34] H. Tin Kam, The random subspace method for constructing decision forests, IEEE Trans. Pattern Anal. Mach. Intell. 20 (1998) 832.

[35] B.W. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochim. Biophys. Acta Protein Struct. 405 (1975) 442.

[36] A. Haug, R.D. Graham, O.A. Christophersen, G.H. Lyons, How to use the world's scarce selenium resources efficiently to increase the selenium concentration in food, Microb. Ecol. Health Dis. 19 (2007) 209.

[37] D. Albanese, S. Riccadonna, C. Donati, P. Franceschi, A practical tool for maximal information coefficient analysis, GigaScience 7 (2018) 1.