Mauchline, T. H., Hayat, R., Clark, I. M. and Hirsch, P. R. 2018. Old meets new: most probable number validation of metagenomic and metatranscriptomic datasets in soil. *Letters in Applied Microbiology.* 66 (1), pp. 14-18.

Letters in Applied Microbiology

Letters in Applied Microbiology ISSN 0266-8254

ORIGINAL ARTICLE

Old meets new: most probable number validation of metagenomic and metatranscriptomic datasets in soil

T.H. Mauchline[1], R. Hayat[2], I.M. Clark[1] and P.R. Hirsch[1]

1 Sustainable Agriculture Sciences Department, Rothamsted Research, Harpenden, Hertfordshire, UK
2 Departmet of Soil Science and SWC, PMAS Arid Agriculture University, Rawalpindi, Pakistan

Significance and Impact of the Study: This study has demonstrated for the first time a functional assay validation of metagenomic and metatranscriptomic datasets by utilizing the clover and *Rhizobium leguminosarum* sv. *trifolii* mutualism. The results show that the Most Probable Number results corroborate the results of the 'omics approaches and gives confidence to the study of other biological systems where such a cross-check is not available.

Keywords
metagenomics, metatranscriptomics, most probable number, rhizobia, soil.

Correspondence
Tim H. Mauchline, Sustainable Agriculture Sciences Department, Rothamsted Research, Harpenden, Herts, UK. AL52JQ.
E-mail: tim.mauchline@rothamsted.ac.uk

2017/0966: received 18 May 2017, revised 17 October 2017 and accepted 1 November 2017

doi:10.1111/lam.12821

Abstract

Metagenomics and metatranscriptomics provide insights into biological processes in complex substrates such as soil, but linking the presence and expression of genes with functions can be difficult. Here, we obtain traditional most probable number estimates (MPN) of *Rhizobium* abundance in soil as a form of sample validation. Our work shows that in the Highfield experiment at Rothamsted, which has three contrasting conditions (>50 years continual bare fallow, wheat and grassland), MPN based on host plant nodulation assays corroborate metagenomic and metatranscriptomic estimates for *Rhizobium leguminosarum* sv. *trifolii* abundance. This validation is important to legitimize soil metagenomics and metatranscriptomics for the study of complex relationships between gene function and phylogeny.

Introduction

Recent advances in metagenomics and transcriptomics with next generation sequencing have empowered researchers with the ability to study biological systems in an unbiased fashion. However, the use of these technologies in complex environments such as soil are not free of problems as they often yield short length reads at low copy number and at resolutions insufficient to reconstruct entire metagenomes, as such they are often limited to informing on the abundance and expression of single genes (Lombard *et al.* 2011). It is likely that a combination of 'omics and conventional methods will prove to be more powerful than any single approach. In this study, we compare abundance and activity estimates of the important nitrogen fixing bacterium *Rhizobium leguminosarum leguminosarum* sv. *trifolii* in the soil metagenome and metatranscriptome at the Rothamsted Highfield experiment (Hirsch *et al.* 2009) with a Most Probable Number (MPN) bioassay. Highfield is located on a site that had been grassland for centuries, and compares the effect of three long-term treatments which were established between 1949 and 1959: bare fallow (where plant growth has been suppressed by tilling), continuous wheat and a continuous grassland sward, the latter possessing the clover host plant for *R. leguminosarum* sv. *trifolii* symbiosis. Previous work has shown that the organic reserves as well as microbial abundance have declined in the bare fallow soil due to a drastic reduction in fresh carbon inputs (Hirsch *et al.* 2009). Due to the specific mutualistic interaction between clover and *R. leguminosarum* sv. *trifolii* we anticipate that the grassland soil samples support the largest population of this bacterium, and as these bacteria are also generally rhizosphere competent that the arable treatment will also support a larger population of *R. leguminosarum* sv. *trifolii* than in bare fallow soil where there are no plant inputs. We hypothesize that the MPN and

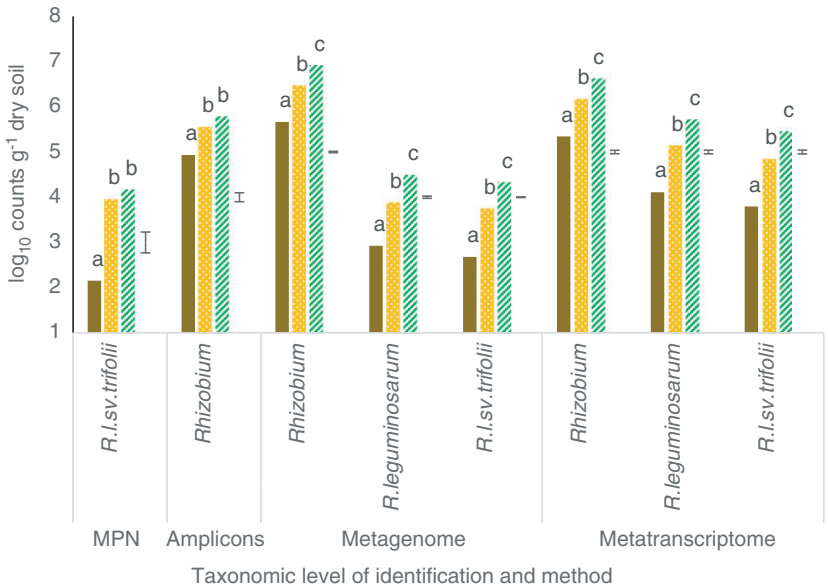

Figure 1 Comparison of functional and molecular estimations of rhizobial abundance on the Highfield experiment permanent treatments. Means were compared using Tukey's *post hoc* method. Statistically significant differences are recognized by use of different letters above bars. (■) bare fallow; (▩) arable and (▨) grass. [Colour figure can be viewed at wileyonlinelibrary.com]

'omics data will corroborate each other and support OTU assignment of metagenomic and metatranscriptomic datasets.

Results and discussion

MPN assays indicated that the *R. leguminosarum* sv. *trifolii* population was lowest in bare fallow soil, highest in the grassland, with the arable plots at an intermediate level (Fig. 1). Analysis of metagenomic DNA using 16S rRNA gene amplicons, which identified reads only to the genus level, showed the same trend, bare fallow <arable <grassland (Fig. 1). The number and expression of *Rhizobium* genes per gram soil was measured across all plots at three taxonomic levels. We found that reads assigned to *Rhizobium* came from many species but those from *R. leguminosarum* were dominated by *R. leguminosarum* sv. *trifolii,* also showing an increasing trend: bare fallow <arable <grassland (Fig. 1). We also found strong positive correlations of *R. leguminosarum* sv. *trifolii* abundance as estimated by MPN with *Rhizobium* amplicon abundance as well as number of *R. leguminosarum* sv. *trifolii* genes and transcripts per g soil (Fig. 2). Therefore, the amplicon, metagenomics and metatranscriptomics analyses all strongly agree with the MPN counts of *R. leguminosarum* sv. *trifolii* revealed in the clover nodulation assay (Figs 1 and 2). To investigate a host-specific nodulation gene required for symbiotic interactions with the host, we also enumerated the constitutive master regulator *nodD* in *R. leguminosarum* sv. *trifolii.* Copies were least abundant in bare fallow soil for both metagenomic and metatranscriptomic datasets, but the total number of reads assigned to this gene across all replicate plots (21, 65, 72 in bare fallow, arable and grassland soil respectively) were too low for statistical analysis, highlighting the limitation of metagenomics and metatranscriptomics for quantifying relatively low copy genes and transcripts. A small number of copies of the *nodA* gene which is known to be induced during nodulation were detected in the arable and grassland transcriptomes (3 and 6 respectively), again too low to draw inference. Nevertheless, the broad trend indicating fewer *nodD* and *nodA* genes and transcripts in bare fallow soil in comparison with the other plots agrees with the MPN estimates. We have previously used a qPCR assay based on *nodD* to estimate *R. leguminosarum* sv. *trifolii* numbers in a range of soils, finding correlation with MPN estimates (Macdonald *et al.* 2011).

The most striking observation from this work is that bare fallow soil has the lowest *R. leguminosarum* sv. *trifolli* population, followed by an intermediate population level in the arable plots, and higher numbers in the grassland plots. This is explained as the bare fallow soil is least abundant in nutrients and organic matter, with poor physical structure, supporting the smallest bacterial population of the three treatments (Hirsch *et al.* 2009). The arable plots receive mineral fertilizer to stimulate the growth of wheat, and rhizobia are known to be competent soil and rhizosphere saprophytes with large genomes and functional plasticity (Young *et al.* 2006), and so compete well in this environment. The highest population in the grassland plots is presumably due to the clover

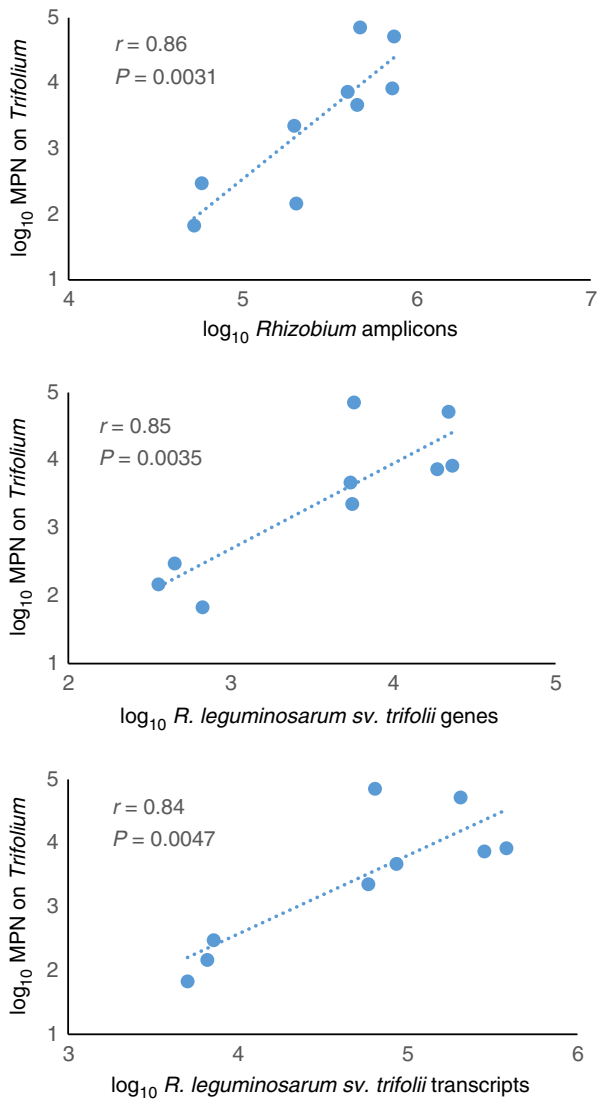

Figure 2 Correlation of most probable number estimates of *Rhizobium leguminosarum* sv. *trifolii* with the abundance of amplicons assigned to *Rhizobium* and DNA or RNA reads assigned to *R. leguminosarum* sv. *trifolii*. Pearson's linear correlation coefficient *r* is given, followed by the probability of no correlation. [Colour figure can be viewed at wileyonlinelibrary.com]

legume host existing among the sward, and in addition non-host grasses can provide nutrients and organic matter to the system which ultimately supports a larger microbial population than the other treatments.

In summary, this work demonstrates that land use has a clear effect on the abundance of rhizobia in soil. It also demonstrates that metagenomics can be used to reliably estimate the population of *R. leguminosarum* sv. *trifolii* in field soil as it agrees with the conventional MPN 'gold standard' of quantification for this bacterium, which makes use of the mutualistic life style with the clover host. In order for clover nodule formation to occur, a large number of rhizobial genes must be present and expressed in a coordinated way within a single bacterium, something that soil nucleic acid extraction 'omics approaches cannot readily confirm. This validation is important to legitimize soil metagenomics and metatranscriptomics for the study of complex interactions between gene function and phylogeny, not only in rhizobia but also for other soil microbes where gene and transcript titre is relatively low and a validation cross-check of function is not possible. Future studies could investigate more subtle changes in land management, such as levels of N fertilization, to ascertain to what level congruence of method comparison occurs.

Materials and methods

Soil sampling

The Highfield experiment is located on the Rothamsted Research farm in Harpenden, Hertfordshire, UK. The site had been under pasture for centuries when, in 1949, sections were switched to continuous arable (wheat) cultivation. In 1959 further areas of grassland were converted to a bare fallow treatment in which plants are regularly removed by tilling. Highfield has a random block design consisting of replicated plots as described by Jensen *et al.* 2017. Soil was collected from the Highfield permanent plots in October 2011 to 10 cm depth using a 3 cm diameter corer; the top 2 cm containing root mats and other plant detritus was discarded. Three replicate plots per treatment were sampled. Ten cores per plot were pooled, thoroughly mixed whilst sieving through 2 mm mesh; then samples were frozen at −80°C, soil for MPN assays was stored at 4°C. All implements were cleaned with 70% ethanol between sampling/sieving soil from each grid area.

Clover nodulation assays

Assays were prepared as described by Hirsch and Skinner (1992). This involved surface sterilization of clover seeds with 100% ethanol and 2% bleach followed by washing with sterile water. Germinated seeds were aseptically transferred to medical tube slopes containing 30 ml of Jenson medium (Jensen 1942) and inoculated with 1 ml of five to 78 125-fold soil diluted in sterile water. Five technical replicates were prepared for each dilution as well as a control sterile water inoculum. The seedlings were incubated for 4 weeks with 16 h light at 18°C. After this time, root systems were examined for presence or absence of nodules. Results were adjusted according to soil dry weight and the effective *R. leguminosarum* sv.

trifolii most probable number in each soil was calculated as described by Brockwell (1963).

Nucleic acid extraction and analysis

Community DNA and RNA was extracted from a minimum of 2 g soil using the MoBio RNA PowerSoil® (Carlsbad, CA, USA) Total RNA isolation kit followed by the RNA PowerSoil®DNA Elution Accessory kit, with three replicates for each soil treatment. All RNA samples were DNAase treated with Ambion Turbo DNA-free™ (Hemel Hempstead, UK). When necessary, extracts were pooled to provide sufficient material for sequencing. For the 3 years prior to this study, soil 16S rRNA gene copy numbers were estimated by qPCR with universal primers to allow normalization of results (Clark *et al.* 2012).

Bacterial and archaeal 16S rRNA genes were amplified and sequenced at the High-throughput Genome Analysis Core (HGAC), Argonne National Laboratory (USA) using an Illumina® MiSeq sequencer (Pylro *et al.* 2014). Sequenced amplicons were assigned to taxa using Qiime as described by Caporaso *et al.* 2010. Full metagenomic sequencing of >10 μg DNA from each replicate soil treatment was provided by Illumina®, Cambridge, UK using a HiSeq 2000, generating 150 bp paired end reads. RNA was subjected to ribodepletion and sequenced by The Earlham Institute, Norwich, UK using a HiSeq 2000, generating 100 bp paired end reads. Sequences were quality checked using the FASTX-Toolkit (ver. 0.0.13.2, http://hannonlab.cshl.edu/fastx_toolkit/index.html) with a quality threshold of 25, minimum length 100 bp for DNA and 80 bp for RNA. Reads were assigned to taxa using DIAMOND (Buchfink *et al.* 2015) and analysed in MEGAN5 (Huson *et al.* 2007).

For amplicon data analysis, the abundance of reads assigned as *Rhizobium* in each soil sample was estimated as a proportion of the total number of reads g^{-1} dry soil. For the metagenome and metatranscriptome analysis: *Rhizobium*, *R. leguminosarum* and *R. leguminosarum* sv. *trifolii* populations were estimated as a proportion of the total reads g^{-1} dry soil in each sample which were assigned to the three-respective taxonomic levels.

To estimate the number of *nodA* and *nodD* genes and transcripts we used Decypher BLAST (Time Logic®) to interrogate the metagenomes and metatranscriptomes with full-length *R. leguminosarum* sv. *trifolii* WSM 1689 gene sequences derived from NCBI (https://www.ncbi.nlm.nih.gov/).

Statistical analyses

Statistical analysis was performed using one-factor ANOVA in GenStat 17th Edition (VSN International Ltd., Hemel Hempstead, UK). To check that each set of measured values met the assumptions of ANOVA and were normally distributed, residuals were plotted. If they did not show normal distribution, data was $\log_{10}$-transformed and again checked for normal distribution of residuals. Treatment comparisons with F statistics with $P < 0{\cdot}01$ were considered significant, $P < 0{\cdot}001$ highly significant. Means were compared using Tukey's *post hoc* method in the GenStat multiple comparison menu with 95% confidence; means are considered significantly different are indicated with different letters.

Acknowledgements

Rothamsted Research receives strategic funding from the Biotechnology and Biological Research Council of the UK, this work was supported by BBS/E/C/00005196 and BB/P01268X/1.

Conflict of Interest

No conflict of interest declared.

References

Brockwell, J. (1963) Accuracy of a plant-infection technique for counting populations of *Rhizobium trifolii*. *Appl Microbiol* **11**, 377–383.

Buchfink, B., Xie, C. and Huson, D.H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59–60.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G. *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**, 335–336.

Clark, I.M., Buchkina, N., Jhurreea, D., Goulding, K.W.T. and Hirsch, P.R. (2012) Impacts of nitrogen application rates on the activity and diversity of denitrifying bacteria in the Broadbalk wheat experiment. *Philiso Trans R Soc B* **367**, 1235–1244.

Hirsch, P.R. and Skinner, F.A. (1992) The identification and classification of Rhizobium and Bradyrhizobium. In *Identification Methods in Applied and Environmental Microbiology*, eds. Board, R.G., Jones, D. and Skinner, F.A. pp. 45–65. Oxford: Blackwell Scientific Publications.

Hirsch, P.R., Gilliam, L.M., Sohi, S.P., Williams, J.K., Clark, I.M. and Murray, P.J. (2009) Starving the soil of plant inputs for 50 years reduces abundance but not diversity of soil bacterial communities. *Soil Biol Biochem* **41**, 2021–2024.

Huson, D.H., Auch, A.F., Qi, J. and Schuster, S.C. (2007) MEGAN analysis of metagenomic data. *Genome Res* **17**, 377–386.

Jensen, H.L. (1942) Nitrogen fixation in leguminous plants. II. Is symbiotic nitrogen fixation influenced by Azotobacter?. *Proc Linn Soc NSW* **57**, 205–212.

Jensen, J.L., Schjønning, P., Watts, C.W., Christensen, B.T. and Munkholm, L.J. (2017) Soil texture analysis revisited: removal of organic matter matters more than ever. *PLoS ONE* **12**, e017803.

Lombard, N., Prestat, E., van Elsas, J.D. and Simonet, P. (2011) Soil-specific limitations for access and analysis of soil microbial communities by metagenomics. *FEMS Microbiol Ecol* **78**, 31–49.

Macdonald, C.A., Clark, I.M., Hirsch, P.R., Zhao, F.J. and McGrath, S.P. (2011) Development of a real-time PCR assay for detection and quantification of *Rhizobium leguminosarum* bacteria and discrimination between different biovars in zinc-contaminated soil. *Appl Environ Microbiol* **77**, 4626–4633.

Pylro, V.S., Roesch, L.F.W., Morais, D.K., Clark, I.M., Hirsch, P.R. and Totola, M.R. (2014) Data analysis for 16S microbial profiling from different benchtop sequencing platforms. *J Microbiol Methods* **107**, 30–37.

Young, J.P., Crossman, L.C., Johnston, A.W., Thomson, N.R., Ghazoui, Z.F., Hull, K.H., Wexler, M., Curson, A.R. *et al.* (2006) The genome of *Rhizobium leguminosarum* has recognizable core and accessory components. *Genome Biol* **7**, R34.

