**Title:**

# MAUI-seq: Multiplexed, high-throughput amplicon diversity profiling using unique molecular identifiers

**Authors:**

Bryden Fields[1*] and Sara Moeskjær[2*], Ville-Petri Friman[1], Stig U. Andersen[2], and J. Peter W. Young[1]

*: These authors contributed equally to the work.

**Author affiliations:**

[1]Department of Biology, University of York, York, United Kingdom

[2]Department of Molecular Biology and Genetics, Aarhus University, Aarhus, Denmark

**Authors for correspondence:**

Peter Young, peter.young@york.ac.uk

Stig U. Andersen, sua@mbg.au.dk

**Keywords**:

Amplification bias, Environmental amplicon diversity, High-throughput amplicon sequencing, Rhizobium, Unique molecular identifier

## Abstract

Correcting for sequencing and PCR errors is a major challenge when characterising genetic diversity using high-throughput amplicon sequencing (HTAS). Clustering amplicons by sequence similarity is a robust and frequently used approach, but it reduces sensitivity and makes it more difficult to detect differences between closely related strains. We have developed a multiplexed HTAS method, MAUI-seq, that incorporates unique molecular identifiers (UMIs) to improve correction. We profile *Rhizobium leguminosarum* biovar *trifolii* (*Rlt*) synthetic DNA mixes and *Rlt* diversity in white clover (*Trifolium repens*) root nodules by multiplexed sequencing of two plasmid-borne nodulation genes and two chromosomal genes. We show that the two main advantages of UMIs are efficient elimination of chimeric reads and the ability to confidently distinguish alleles that differ at only a single position. In addition, multi-amplicon profiling of genes from different replicons enables increased diversity profiling resolution, allowing us to demonstrate limited strain overlap between geographically distinct locations. The method does not rely on commercial library preparation kits and provides cost-effective, sensitive and flexible profiling of intra-species diversity.

## Introduction

Evaluation of genetic diversity in environmental samples has become a fundamental approach in ecological studies (Birtel et al., 2015). If a core gene can be amplified from environmental samples with universal primers, the relative abundance of species in the community can be estimated from the proportions of species-specific variants among the amplicons. This is termed metabarcoding and, with the continued optimisation of high throughput sequencing, it has become a cost-effective way to detect multiple species simultaneously within a range of environmental samples (Fonseca, 2018; Krehenwinkel et al., 2018; Tessler et al., 2017; Elbrecht and Leese, 2015, Gohl et al., 2016; Poisot et al., 2013). Various methods have been developed to detect and accurately represent diversity within metagenomic samples, including high throughput amplicon sequencing (HTAS) and shotgun metagenomic sequencing. HTAS allows for more in-depth, higher resolution sequencing of a specific part of the genome for diversity estimates compared to shotgun sequencing, but at the cost of potential sequence errors (Gohl et al., 2016). Despite the extensive use of HTAS for interspecies ecological diversity studies, to our knowledge few investigations have utilised HTAS for intraspecies analysis (Poirier et al., 2018; Kinoti et al., 2017). As 16S rRNA amplicons are too highly conserved to estimate microbial within-species diversity, other target gene candidates would need to be considered in order to sufficiently discern intraspecies sequence variation.

Many studies have evaluated the extent of PCR based amplification errors and bias for HTAS diversity studies (Kebschull et al., 2015; Krehenwinkel et al., 2018; Elbrecht and Leese, 2015). Numerous known PCR biases reduce the accuracy of diversity and abundance estimations, with the major concern being the inability to confidently distinguish PCR error from natural allelic variation in environmental samples, and is an especially limiting factor from an intraspecies perspective.

Stochasticity of PCR amplification and polymerase mistakes were found to be the most significant causes for PCR errors in template representation (Kebschull et al., 2015). Polymerase errors introduce new sequences into the template population from amplification. These sequence errors include not only substitutions but also insertions and deletions. Using proof-reading polymerases, optimising DNA template concentration, and reducing PCR cycle number have been suggested to reduce occurrence of these errors (Oliver et al., 2015; Gohl et al., 2016; Kebschull et al., 2015). Additionally, production of chimeric sequences by template-switching can contribute significantly to PCR errors (Kebschull et al. 2015, Edgar et al., 2011, Edgar et al., 2016).

Bias is also observable during a multiplexed next-generation sequencing run, as it is difficult to achieve equal sequencing depth between samples. Therefore, read counts are usually quantified as a relative abundance, however recent studies have even suggested that diversity metrics should only be analysed on a presence-absence basis (Fonseca, 2018; Elbrecht and Leese, 2015; Palmer et al., 2018).

In order to account for introduction of sequence variants in PCR amplification several sequence classification approaches have been established to manage diversity estimates. The processes of either 'lumping' or 'splitting' sequences as strategies for diversity categorisation has long been debated. 'Lumping' involves grouping sequences together that share similarities. The most common method is the use of operational taxonomic units (OTUs) in microbial diversity studies which analyse target gene sequences and cluster based on an arbitrary fixed similarity threshold (QIIME, Rideout et al., 2018; DADA2, Callahan et al., 2016; UPARSE, Edgar, 2013; Huse et al., 2010; Lindahl et al., 2013: Callahan et al., 2016; Fierer et al., 2017; Poisot et al., 2013). Within species boundaries this technique could dramatically reduce the resolution of naturally occurring sequence variation. Conversely, 'splitting' aims to only cluster sequences that share the exact sequence. These form sequence groups called amplicon sequence variants (ASVs) (Callahan et al., 2016; Fierer et al., 2017). Using this strategy, biases from PCR errors are more evident and often variation induced by PCR error cannot be differentiated from natural allelic variation. Similarly, the increased taxonomic resolution of ASV clustering can contribute additional noise to datasets making analysis more challenging (Callahan et al., 2016). In many cases ASVs cannot rectify errors, for example some organisms contain several alleles of a gene in their genome, which could lead to overestimation of sample diversity. (Callahan et al., 2016. Fierer et al., 2017).

As a solution to this, assigning a unique molecular identifier (UMI) to every initial DNA template within a metagenomic sample enables evaluation of PCR amplification bias (Lundberg et al., 2013). Additionally, the UMI provides a potential route to address polymerase errors in metabarcoding studies. The UMI is generated from a set of random bases in the gene-specific forward inner primer, which will include a random DNA sequence into every initial DNA template upstream of the amplicon region during the first round of amplification. Once all original DNA template strands are assigned a unique UMI, the outer forward primer and the gene specific reverse primer can enable further amplification. Consequently, all subsequent DNA amplified from the original template will have the same UMI, so the number of reads amplified from the initial template can be calculated.

To our knowledge, UMIs have previously only been utilised for single amplicon interspecies investigations (Faith et al., 2013; Hoshino et al., 2017; Jabara et al., 2011; Kinde et al., 2011). Here, we show that multiplexing and UMI-labelling of amplicons improves error correction and resolution, allowing accurate intra-species profiling of bacterial genetic diversity.

## Results

### MAUI-Seq: multiplexing of amplicons labelled with UMIs

We present an HTAS method, MAUI-seq, suited for characterisation of intraspecies diversity that enables detection and correction of PCR amplification and polymerase errors. This is achieved by using a nested PCR method with primers that assign random UMIs to each initial template DNA at the first primer elongation PCR step (**Figure 1A-B**; Lundberg et al., 2013). The method enables amplification of multiple amplicons from different genes. This circumvents the issues of inaccurate base-calling on Illumina platforms, which arises when assessing single amplicons (Krueger et al., 2011). The protocol and data analysis pipeline can be applied to DNA extracted from any environmental sample. Here, we validate the method by assessing *Rhizobium leguminosarum* symbiovar *trifolii* (*Rlt*) diversity in DNA mixtures and white clover (*Trifolium repens*) root nodules. Within-species diversity estimates cannot be achieved using 16S rRNA amplicon reads, as the genomic region is too conserved to identify individual strains. In order to evaluate intraspecies genetic diversity, we instead amplify regions in four *Rlt* genes - the nodulation genes *nodA* and *nodD* that reside on an auxiliary plasmid and the chromosomal genes *rpoB* and *recA*.

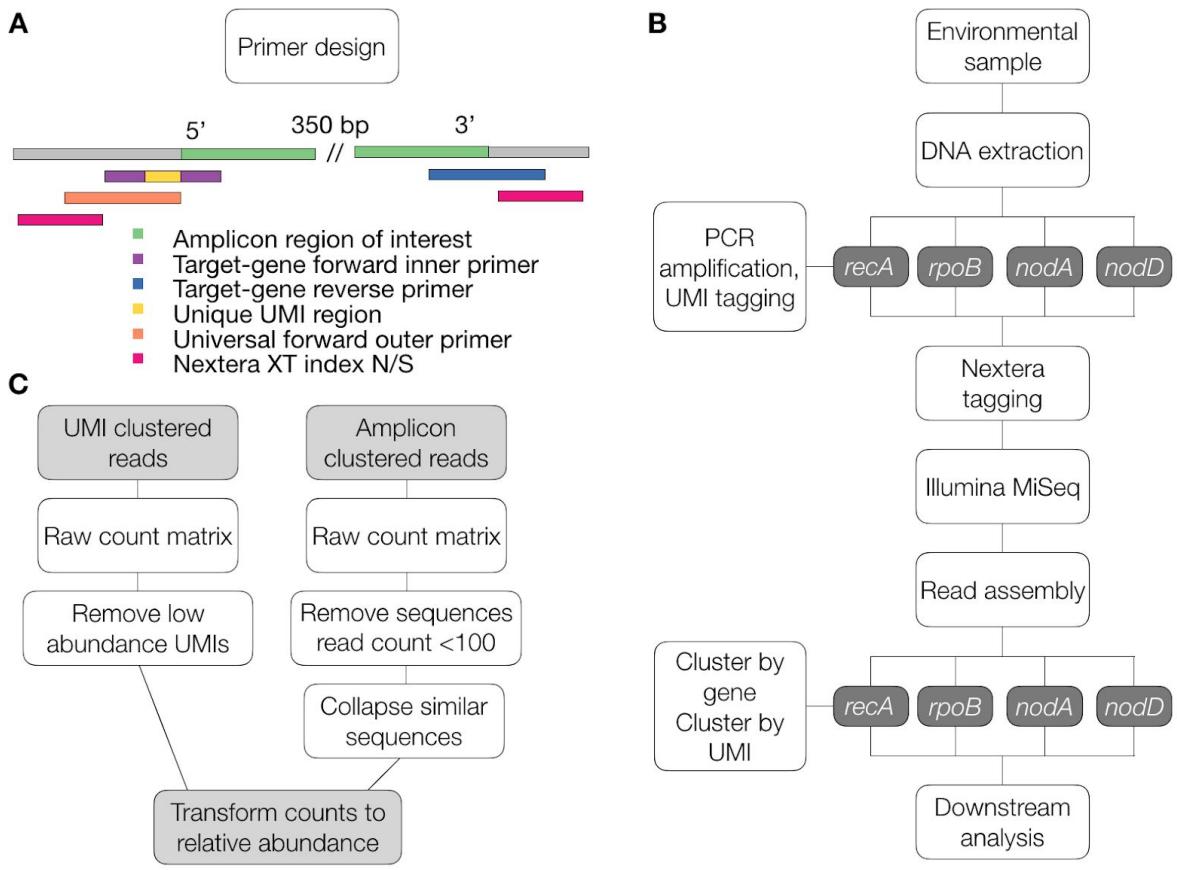


**Figure 1.** Primer design and method workflow. **A:** Primer design using the sense strand of the target DNA template as an example. The amplicon region of interest should be no longer than 500bp. The target-gene forward inner primer, universal forward outer primer and the target-gene reverse primer are all used in the initial PCR. The Nextera XT indices (Illumina, USA) are used in the Nextera barcoding PCR step. The Unique molecular identifier (UMI) region is shown in yellow on the target-gene forward inner primer. **B:** Method workflow. **C:** Data analysis workflow for UMI clustered and amplicon clustering of reads.

### Validation using purified DNA mixed in known proportions

Method accuracy was evaluated by profiling DNA mixtures with known strain DNA ratios. DNA was extracted from two *Rlt* strains differing by a minimum of 3bp in their *recA*, *rpoB*, *nodA,* and *nodD* amplicon sequences, and the extracted DNA was mixed in different ratios (**Table S1**). Sequencing reads were then filtered and clustered by two methods for comparison: UMI clustering and amplicon clustering. For UMI clustering, reads were first assigned to their target gene, and then clustered by UMI within each gene. The most abundant sequence variant associated with a specific UMI was chosen as the representative sequence for that cluster of sequences. The read abundance within each sequence cluster was then calculated for all samples. UMIs with an abundance of less than two across all

samples were removed from the analysis. Amplicon clustering involved clustering sequences purely by ASVs and calculating read abundance within these ASV clusters for all samples (**Figure 1C**). For each sample, the most abundant sequence associated with each *Rlt* strain was identified by UMI clustering and by amplicon sequence clustering. The total number of sequence clusters was reduced from 46 for amplicon clustering to 37 for UMI clustering (**Figure 2A**). The observed and expected strain ratios were compared for all four genes (**Figure 2C, Figure 2E, Figure S1, Figure S2**). The correlation ranged from 0.949 (*recA*) to 0.998 (*rpoB*) for the Phusion DNA polymerase and 0.963 (*rpoB*) to 0.996 (*nodD*) for the Platinum enzyme when the sequences were clustered by UMI. When clustered by amplicon, the correlation ranged from 0.943 (*recA*, Phusion) to 0.999 (*rpoB*, Phusion). Performance of the two polymerases was gene dependent, which could be due to differences in amplification efficiency for the four templates. During the amplification of short amplicons, chimeras can form due to template switching, but they can easily be identified using UMI clustering (**Figure 2B**). The proportion of chimeras formed in the simple two-strain mixture peaked at 6% and 13% for *rpoB* UMI clustering and amplicon clustering, respectively (**Figure 2D, Figure 2F**). For *nodA* the proportion of chimeras declined from 20% to 3.5% for amplicon clustering and UMI clustering, respectively. For amplicon clustered reads, no chimeras were detected for *recA*. Since the SM3 and SM170C only differ by three nucleotides, chimeras grouped with the more abundant sequences during the PCR error correction step (**Figure S2**). This could pose an issue for the relative abundance, since the chimeras were counted as real sequences.

**A**

| Method | Experiment | Total number of detected sequence clusters | | | | |
|---|---|---|---|---|---|---|
| | | *rpoB* | *recA* | *nodA* | *nodD* | **total** |
| **Synthetic mix** | Amplicon clustering | 4 | 2 | 7 | 33 | **46** |
| | UMI clustering | 14 | 4 | 5 | 14 | **37** |

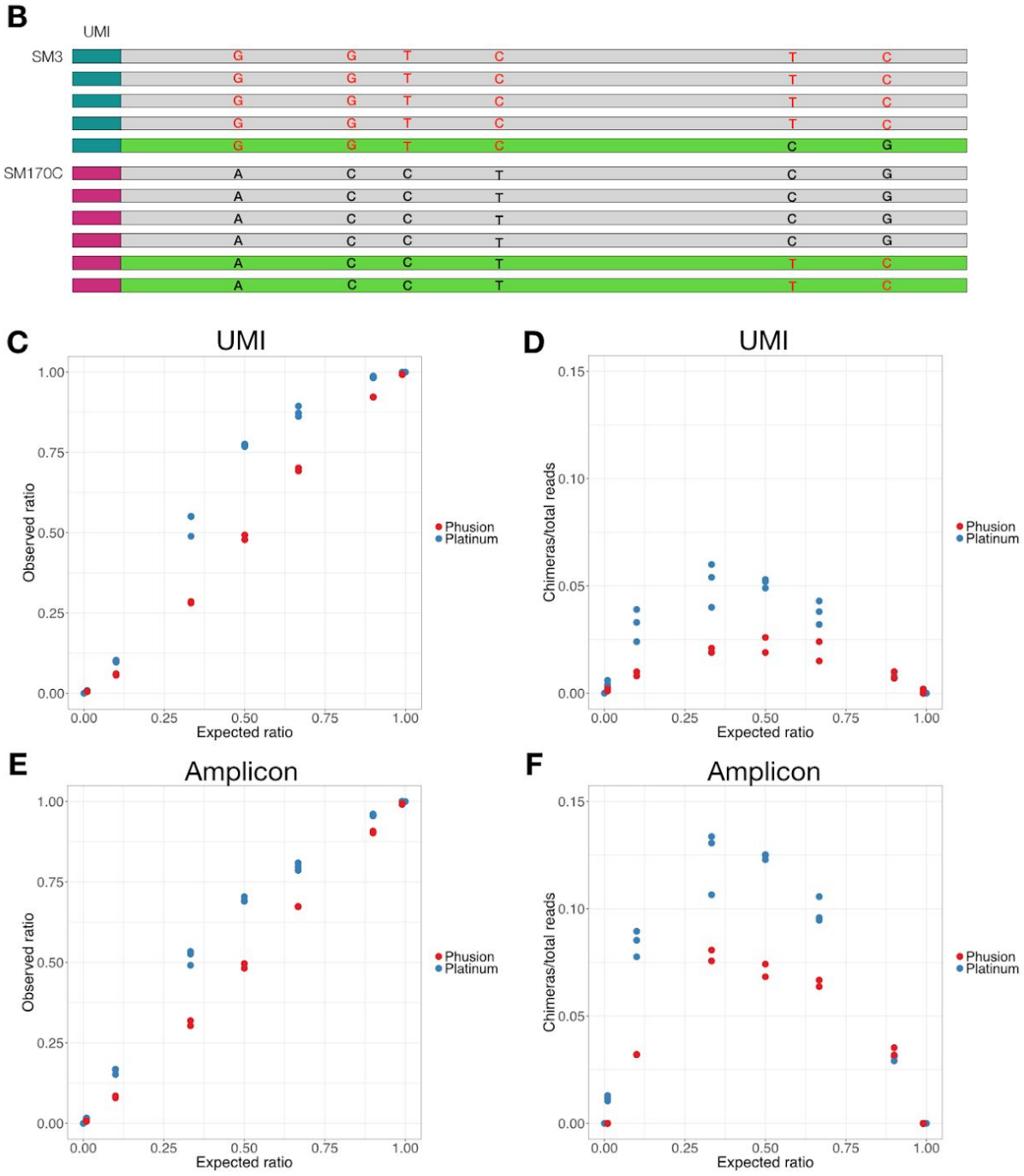


**Figure 2.** Performance on DNA mixtures. **A:** Total number of detected sequence clusters in the analysed dataset using amplicon clustering and UMI clustering. **B:** Schematic representation of type sequences for SM3 and SM170C (grey) and chimeric reads (green) identified for *rpoB*. **C and D:** UMI clustering of sequences. **C:** Observed ratio of SM3/SM170C versus the expected ratio of *rpoB* reads. Pearson correlation for Phusion=0.998 and Platinum=0.963. **D:** Proportion of chimeras compared to

total reads. **E and F:** Amplicon clustering of sequences. **E:** Observed ratio of SM3/SM170C versus the expected ratio of *rpoB* reads. Pearson correlation for Phusion=0.999 and Platinum=0.980. **F:** Proportion of chimeras compared to total reads.

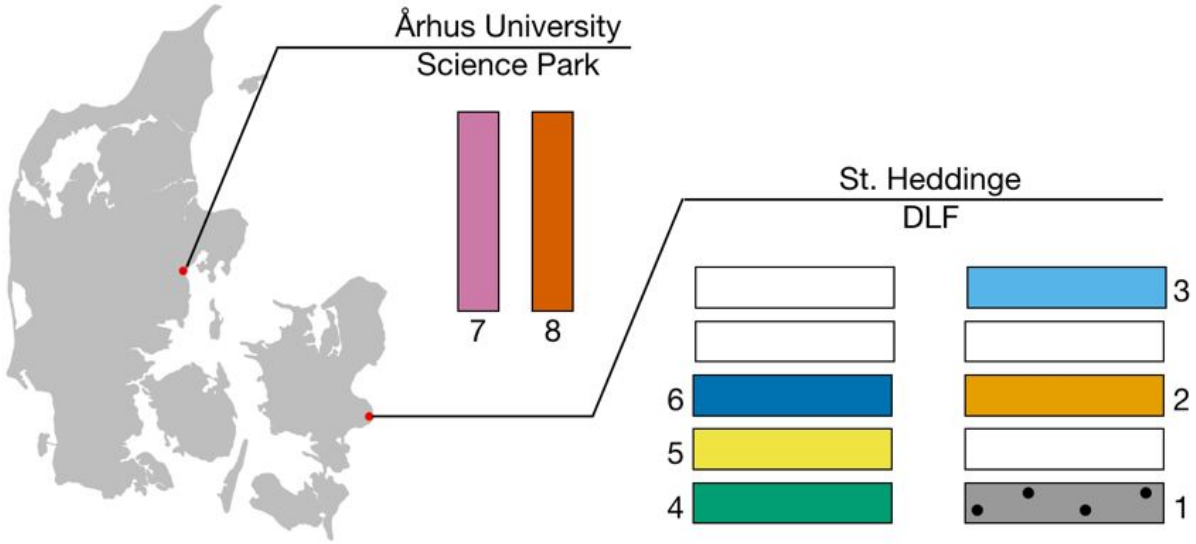


**Figure 3.** Analysis of *Rlt* diversity in root nodule samples. Sampling locations in the Aarhus University Science Park and DLF trial station in Store Heddinge. DNA was isolated from 100 nodules from four points (black dots) on each individual plot.

**UMIs facilitate error correction and improve sensitivity**

To test the method on more complex samples, we compared root nodules from two locations in Denmark, a DLF trial station at Store Heddinge on Sealand and a lawn at Aarhus University in Jutland (**Figure 3**). 100 nodules were pooled for each sample and each plot was sampled in four replicates. Clover root nodules are usually colonised by a single rhizobium strain, providing an easily assessed dataset where a maximum of 100 unique sequences per gene is expected per sample.

After UMI clustering, the total number of distinct sequences was 287. Using cumulative filtering the bottom 1% of UMI clusters were removed, leaving a total of 79 sequences when using the UMI workflow. Amplicon clustering retained 2 million unique sequences. The majority of these were low read-count variants and were removed after filtering for read counts over 100. To correct for PCR errors, sequences differing by only one base pair were collapsed into the most closest relative with the highest abundance. The final number of sequences when using the amplicon clustering workflow was reduced to 323 (**Table 1, Figure 4**). In the sequences of 196 natural *Rlt* isolates (Cavassim et al., 2019) the diversity observed in the amplicon regions was within the same range as was observed when the reads were clustered by UMI (**Table 1**).

The binning of sequences that differ by only a single base, which is necessary in order to reduce the error noise in amplicon sequences, can lead to loss of important information about population diversity. For example, the two most abundant *rpoB* alleles in our populations (**Figure 4B**) differ by just one nucleotide. Both these alleles are common in *Rlt* genome sequences (Cavassim et al., 2019), so we believe they are genuine. The UMI approach detects them separately, and shows that their relative abundance varies across samples, whereas amplicon clustering conflates them into a single sequence, substantially underestimating the true diversity within and between these samples.

**Table 1.** Total number of detected sequence clusters in the root nodule samples using amplicon clustering and UMI clustering. The datasets in bold will be used for all downstream analysis.

| **Method** | **Filtering** | **Total number of detected sequence clusters** | | | | |
|---|---|---|---|---|---|---|
| | | *rpoB* | *recA* | *nodA* | *nodD* | total |
| Total reads | | 4.60M | 2.09M | 3.83M | 4.51M | 15.1M |
| Unique reads pre-filtering | None | 539k | 230k | 552k | 772k | 2039k |
| Amplicon clustering | Remove reads with read count<100 | 249 | 134 | 325 | 242 | 950 |
| | **Collapse similar sequences** | **76** | **27** | **110** | **110** | **323** |
| UMI clustering | UMI clustering | 63 | 98 | 82 | 44 | 287 |
| | **Remove rarest 1% of UMIs** | **27** | **56** | **38** | **28** | **79** |
| Unique sequences from 200 genomes | | 13 | 17 | 14 | 10 | 54 |

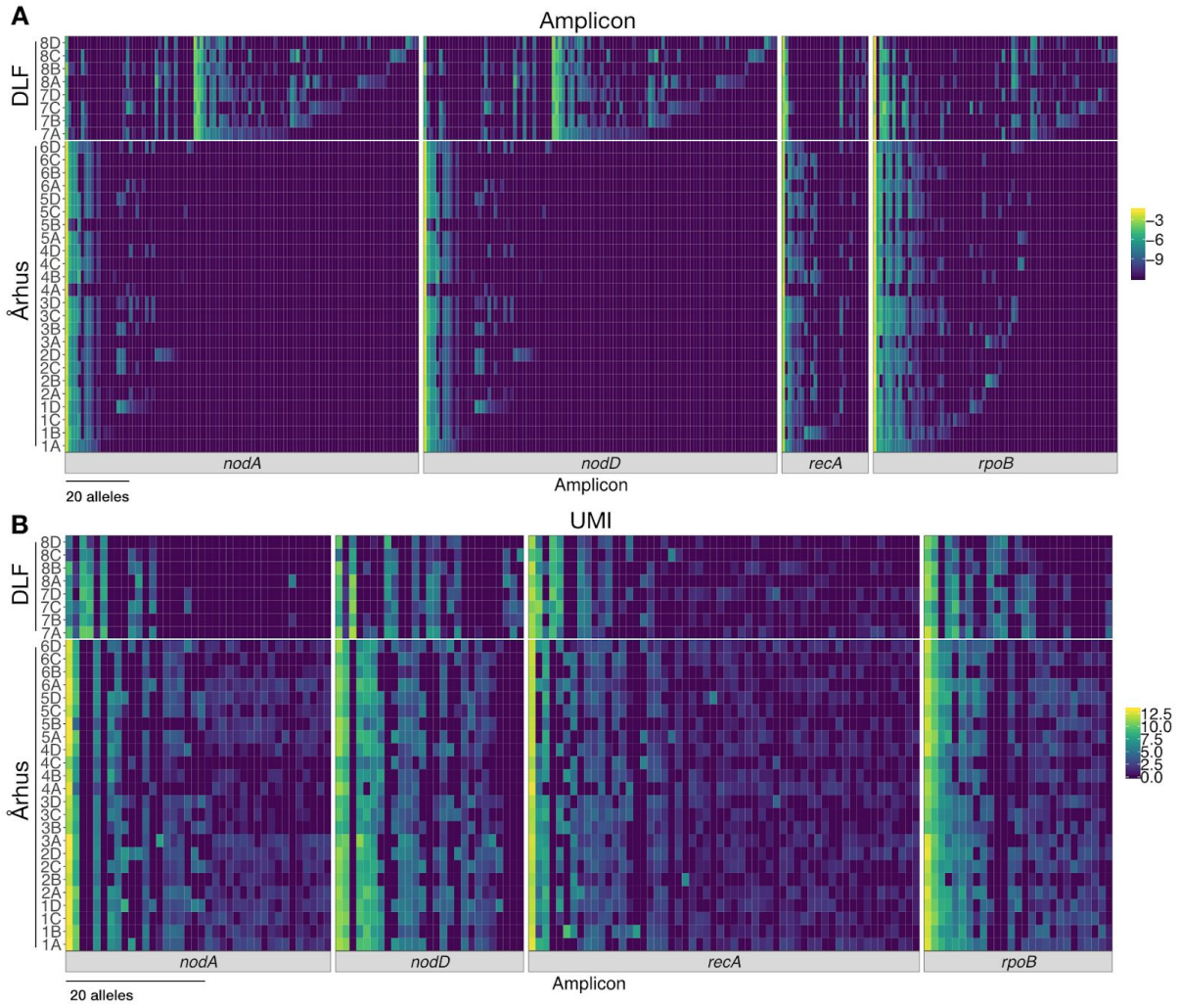


**Figure 4.** Amplicon diversity in nodule samples. **A:** Heatmap of the log2 transformed relative allele abundance of sequence clusters for individual genes when clustered by ampliconI. **B:** Heatmap of the log2 transformed relative allele abundance of sequence clusters for individual genes when clustered by UMI.

### Multi-amplicon analysis improves resolution and classification

We have profiled four amplicons, comprising two chromosomal genes (*rpoB* and *recA*) and two nodulation genes residing on a symbiosis plasmid (*nodA* and *nodD*). Examining the number of shared alleles between the DLF and Aarhus sites for the four amplicons, the chromosomal genes show a large overlap and the same alleles are most abundant at both sites. In the absence of additional evidence, the chromosomal gene data could lead to the conclusion that the two geographically distinct sites share a large proportion of strains. However, the nodulation genes show a different allele profile with more alleles unique to each site and markedly different allele abundance patterns between sites (**Figure 4B**). Considering the information from all four amplicons, it becomes clear that the root nodules at

the two geographically distinct sites share only a very limited number of closely related strains, demonstrating that there is a substantial advantage to multi-amplicon profiling.

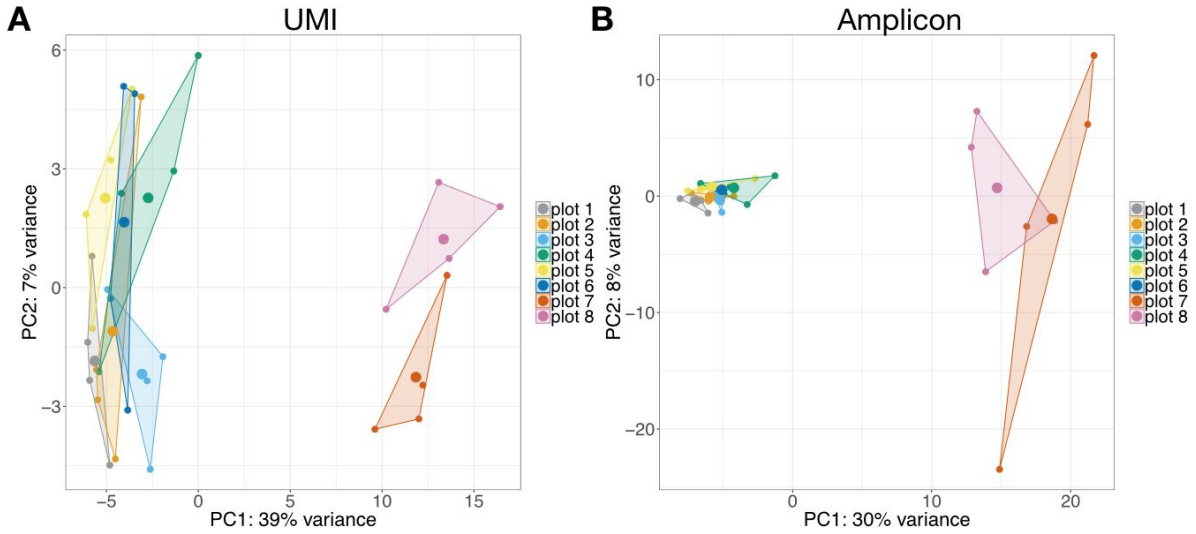


**Figure 5.** Principal Component Analysis of *Rlt* diversity in root nodule samples. Reads from all four genes (*recA*, *rpoB*, *nodA*, and *nodD*) were included in the analysis. **A**: Reads clustered by UMI. **B:** Reads clustered by amplicon.

**Joint assessment of four amplicons robustly distinguishes geographically distinct sites**

When clustering the sequences by UMI (PC1: 39% variance, PC2: 7% variance) or amplicon (PC1: 30% variance, PC2: 8% variance), nodule samples clustered strongly by geographic location (**Figure 5**). However, within the Aarhus and DLF sites, samples from different plots did not show a clear separation (**Figure 5**). The DLF samples cluster more tightly than the Aarhus samples, which is possibly because *Rlt* DNA from DLF plots was isolated from root nodules of the same white clover variety (Klondike), whereas Aarhus *Rhizobium* DNA was isolated from wild white clover root nodules. The variance explained by PC1, which separates the DLF and Jutland sites, was substantially higher for UMI clustering than for amplicon clustering for all four genes considered individually (**Figure S3, Figure S4**).

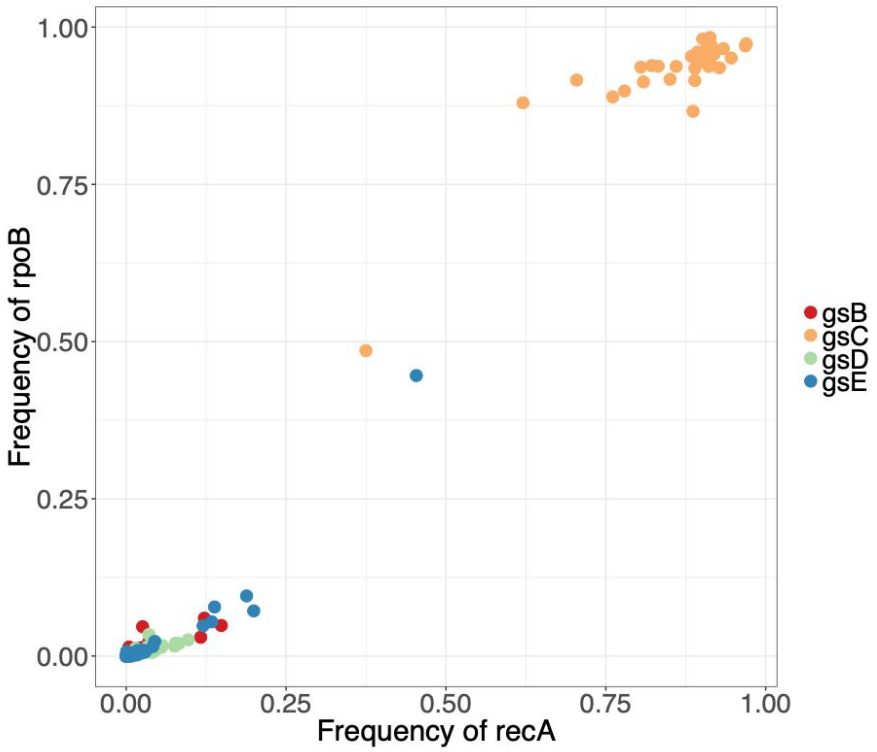


**Figure 6.** Genospecies frequency of *recA* reads plotted against the genospecies frequency of *rpoB* reads for UMI clustered reads from environmental samples. Each data point represents the fraction of the amplicons belonging to a genospecies within each sample.

The simultaneous use of more than one marker gene can also provide more robust estimates of the species composition of a sample. Although *R. leguminosarum* is traditionally regarded as a single species, it can in fact be divided into at least five distinct genospecies (Kumar et al., 2015), and each genospecies has a distinctive set of alleles for core genes such as *rpoB* and *recA* (Cavassim et al., 2019). Hence, we could use these two genes to obtain independent estimates the relative abundance of these genospecies in each sample by assigning each allele to its corresponding genospecies based on these previous publications. The two genes produce highly correlated estimates of genospecies abundance within each sample (Pearson correlation=0.994) (**Figure 6**).

## Discussion

We propose a new HTAS method (MAUI-seq) designed to assess genetic diversity within species. Unlike the myriad of methods targeted at between-species variation to investigate community ecology, this method provides a robust and sensitive way of studying the diversity within a single species. Several amplicons are monitored simultaneously to increase resolution and circumvent the Illumina base-calling issue observed in single-amplicon studies (Krueger et al., 2011).

We tested the method on a synthetic mix of DNA from two *Rlt* strains. The method reliably reproduced the expected ratio in each sample. When clustering the reads by Unique Molecular Identifier (UMI) as opposed to amplicon clustering, the total number of detected sequence clusters was 20%lower (**Figure 2**). A similar reduction of sample complexity was observed when root nodule samples from white clover were tested (**Table 1, Figure 4**). This reduction did not impair the sensitivity of the method. On the contrary, when applied to environmental samples from two geographically distinct locations, we could differentiate samples from the two locations more clearly when reads were clustered by UMI. Increasing the number of monitored amplicons from one to four further increased our ability to robustly distinguish samples from two locations (**Figure 4**). The within-species diversity is reliably reproduced between core gene amplicons, and this highlights the versatility of the method, since it allows a great degree of freedom when choosing target amplicons (**Figure 6**). A proof-reading and a non-proof-reading polymerase were tested (Phusion and Platinum, respectively). Their relative accuracy in reproducing known allele frequencies differed among genes, but Phusion reliably produced a lower proportion of chimeric reads.

The method reduces the financial restrictions of processing large numbers of samples, while accounting for potential PCR errors during downstream sequence analysis. Primers were designed to amplify specifically *Rlt* housekeeping and symbiosis genes, and also provide linker sequences for Nextera index adapters, obviating the need for costly library preparation kits (Wen et al., 2017).

The UMI assignment to each initial DNA template allows users to identify all reads amplified from the same initial template, and so reads were firstly clustered by identical UMI sequences. The advantage of this method is to avoid clustering purely by amplicon sequence which could be subjected to polymerase errors, chimeric sequence generation, and tag-switching (Nguyen et al., 2014), greatly increasing the unique allele false positive rate. For that reason, the UMI assignment from the target gene primer can produce a more accurate estimate of allelic diversity within populations. Clustering by UMI reduced the overall number of detected sequence clusters for both the synthetic mix and the

environmental samples, compared to clustering by amplicon sequence similarity alone. This is due to the UMI clustering identifying not only biases in PCR amplification, but also highlighting polymerase transcription errors within the amplicon sequence and merging these with the original template sequence.

Our method is based on the well-established metabarcoding gene amplicon sequencing method, which commonly uses 16S rRNA amplicon reads from soil DNA samples to characterise soil microbial community diversity and meso- and macro-fauna biodiversity (Orgiazzi et al., 2015; Vasileiadis et al., 2012; Sinclair et al., 2015). While our method validation evaluated *Rlt* intraspecies allelic diversity, this method can also be used for interspecies studies, depending on the target genes chosen, achieving higher resolution when describing prokaryotic populations. Potential further applications of this method include snapshot community analysis through to community dynamics time series experiments. It will allow high-resolution studies of single bacterial species, enabling more detailed, culture-independent population genetics, which could further our understanding of natural bacterial communities. Environmental samples could similarly range from an entirely prokaryotic perspective, as with our investigation, to exclusively eukaryotic and even multitrophic communities, depending on the primer specificity and target genes selected. However, care must be taken with eukaryotic organisms to consider the intragenomic heterogeneity and copy number of samples during analysis (Poisot et al., 2013).

Our analysis highlights the sensitivity of amplicon diversity methods and its dependency on the target gene. We found that the method is able to clearly differentiate samples from two distinct geographic locations with increased sensitivity when considering all four amplicons in combination compared to single amplicon assessment.

Similarly, we observed that polymerase efficiency was also gene dependent. Phusion and Platinum are both prokaryotic polymerases, and the amplification errors produced are most often single-base substitutions, though the error rate is reportedly lower for Phusion (Kindle et al., 2011). A trade-off between fidelity and amplification efficiency is observed for samples with low DNA concentrations and when the amplification rate is low, since Platinum more readily amplifies low-concentration templates, but has a higher error rate (Brodin et al., 2015). Both enzymes displayed persistent template preferences, regardless of clustering method. Using UMI clustering removes most PCR errors unless they are introduced early in the amplification process. When analysing the presence of chimeric reads, Phusion reduced the proportion of chimeras to total reads from 6% to 3% when the initial DNA template ratio was 0.5, where most chimeras are expected.

Some common errors arise when using UMIs. Previous studies reported that the most common error observed was a single base pair substitution mutation resulting in a "parent" and "offspring" primer ID. They calculated that an error in their 10bp primer ID was expected every 100 reads (Brodin et al., 2015). For our analysis, the most commonly detected error in UMIs was a frameshift mutation shifting the 12 bp UMI sequence. In the read processing step, the UMI detection window was therefore increased to 13bp allowing detection of the UMI despite a one base pair insertion. Another issue is the "birthday problem", stating that the probability of assigning the same UMI to more than one sequence by chance is higher than expected, and will only increase as sequencing technology improves and read depth increases (Sheward et al., 2012). Previous studies have used primer IDs that were 8-10bp long (Brodin et al., 2015; Jabara et al., 2011). One solution to this problem is to increase the length of the random base pair sequence section. In this study we have used a 12bp UMI to increase the number of possible sequences.

Advances in the use of synthetic community spike-in controls to enable reliable quantification of genetic diversity are still under development (Kebschull et al., 2015; Edgar et al., 2017; Palmer et al., 2018). Other approaches to estimating abundance include adding UMIs to sequences and simply counting the occurence of each UMI only one time (Kivioja et al., 2012) and using qSeq where the random tag is added to each template in a single primer extension step (Hoshino et al., 2017). The proposed method is currently only validated for relative abundance diversity estimation only, with and without a proof-reading enzyme for PCR amplification. However, the protocol could be optimised with a spike-in control or an adaptation of UMI based quantification methods for absolute abundance estimation in the future.

## Materials and methods

### Preparation of DNA mixtures

Two *Rlt* strains (SM3 and SM170C) were chosen based on their *recA*, *rpoB*, *nodA* and *nodD* amplicon region sequence divergence, with a minimum of 3 base pair differences required for each gene. Strains were grown on Tryptone Yeast agar (28°C, 48hrs). Culture was resuspended in 750ul of the DNeasy Powerlyzer PowerSoil DNA isolation kit (QIAGEN, USA) and DNA was extracted following the manufacturer's instructions. DNA sample concentrations were calculated using QuBit (Thermofisher Scientific Inc., USA). DNA

samples of the two strains were diluted to the same concentration and mixed in various ratios (Supplementary Table 1).

**Preparation of environmental samples**

White clover (*Trifolium repens*) roots were collected from two locations: DLF Store Heddinge, Denmark (6 plots) and Aarhus University Science Park, Aarhus, Denmark (2 plots) (Supplementary Figure 1). The clover variety sampled was Klondike (DLF, Denmark) and wild white clover, respectively. 100 large pink nodules were collected from 4 points on each plot. Therefore, in total 32 samples were collected. Nodules were stored at -20°C until DNA extraction.

Nodule samples were thawed at room temperature and crushed using a sterile homogeniser stick. Crushed nodules were mixed with 750µl Bead Solution from the DNeasy PowerLyzer PowerSoil DNA isolation kit (QIAGEN, USA) and DNA was extracted following the manufacturer's instructions. Nanodrop 3300 (Thermofisher Scientific Inc., USA) calculated the DNA sample concentrations.

**PCR and purification**

Primer sequences were designed for two *Rlt* housekeeping genes, recombinase A (*recA*) and RNA polymerase B (*rpoB*) and for two *Rlt* specific symbiosis genes, *nodA* and *nodD* (Supplementary Table 1).

Primers include: a target-gene forward inner primer, a universal forward outer primer, and a target-gene reverse primer. The concentration of the inner forward primer is 100-fold lower than the universal forward outer primer and the reverse primer (Figure 1). This was to reduce the primer competition in the PCR, and to stop the inner forward primer assigning new unique UMIs to all DNA molecules after every round of amplification. The PCR reaction mixture and thermocycler programme are provided (Supplementary Tables 3 and 4).

PCRs were undertaken individually for each primer set using Platinum Taq DNA polymerase (Thermofisher Scientific Inc., USA) to ensure equal PCR product quantities were produced for each gene (Supplementary Table 2) (expected band size: 381bp). Therefore, four PCR products were produced for each sample, and were subsequently pooled and purified using AMPure XP Beads following the manufacturer instructions (Beckman Coulter, USA). Successful PCR amplification was confirmed by a 0.5X TBE 2% agarose gel run at 90V for 2 hours.

For the DNA mixture samples, PCRs were run in triplicate. DNA from only one strain was also processed as a control to determine the level of cross contamination between samples.

Some samples were also amplified using the proof-reading polymerase, Phusion High-Fidelity Taq (Thermofisher Scientific Inc., USA), to evaluate the improved sequencing quality and sequencing correction by using a proof-reading polymerase.

**Nextera indexing for multiplexing and MiSeq sequencing**

Samples were indexed for multiplexing sequencing libraries with Nextera XT DNA Library Preparation Kit v2 set A (Illumina, USA) using the Phusion High-Fidelity DNA polymerase (Thermofisher Scientific Inc., USA). PCR reaction mixture and programme are detailed respectively (Supplementary Tables 7 and 8). Indices were added in unique combinations as specified in the manufacturer instructions (Illumina, USA).

The PCR product was purified on a 0.5X TBE 1.5% agarose gel and extracted with the QIAQuick gel extraction kit (QIAGEN, USA) (expected band length: ~454bp). PCR amplicon concentrations quantified were using GelAnalyzer2010a and normalised to 10nM (Lazar, 2010). A pooled sample was quantified, Bioanalyzer quality checked and MiSeq sequenced (2x300bp paired end reads) by the University of York Technology Facility. Detailed protocol is available in Supplementary File 1.

**Read processing and data analysis**

PEAR assembler was used to merge paired ends (Zhang et al. 2014). Python scripts used to cluster reads by UMI are provided (Supplementary Script 1 and 2). Additionally, reads can be clustered purely by amplicon sequence, with a maximum of 2 base pairs difference within each cluster (Supplementary Scripts 3). For amplicon sequence-based clustering, only clusters with more than 74 reads were considered for downstream analysis. However, for UMI sequence-based clustering, clusters with less than 32 sequences to each UMI were omitted. The number of clusters used for analysis after processing by both clustering methods are provided (Supplementary Table 8). Output abundance data is then processed for statistical analysis and figure generation using various R packages (Supplementary Script 4; R Core Team, 2017; Wickham, 2009). To cluster sequences initially, reference sequences for all four *Rlt* genes and genospecies 'type' sequences for *recA* and *rpoB* were collected from known *Rhizobium* genospecies classified strains from NCBI (Kumar et al., 2015). Principal components were calculated with the R 'prcomp' package using singular value decomposition to explain the *Rhizobium* diversity and abundance within each sub-plot sample. Differences in allele frequencies between samples was quantified using Bray-Curtis beta-diversity estimation using the R package 'vegdist.' PERMANOVA tests were performed

using the R package ‘adonis’. All scripts are available in the Github repository: https://github.com/brydenfields/MAUI-seq

## Acknowledgements

We thank David Sherlock for his experimental expertise, while developing this method, the York Technology Facility for sequencing and Asger Bachmann, Terry Mun, and Maria Izabel Cavassim for preliminary data analysis, and DLF for access to their clover field trials. This work was funded by grant no. 4105-00007A from Innovation Fund Denmark (S.U.A.). Initial development of the method was funded by the EU FP7-KBBE project LEGATO (J.P.W.Y).

## Author contributions

Conceptualization: JPWY; Methodology: JPWY; Software: BF, SM, JPWY; Validation: BF, SM, JPWY, SUA; Formal analysis: BF, SM, JPWY; Investigation: BF, SM; Resources: JPWY, SUA, VPF; Data curation: BF, SM, JPWY; Writing - original draft: BF, SM; Writing - review and editing: BF, SM, JPWY, SUA, VPF; Visualisation: BF, SM; Supervision: JPWY, SUA, VPF; Project administration: JPWY, SUA; Funding acquisition: JPWY, SUA.

## References

Birtel, J., Walser, J. C., Pichon, S., Bürgmann, H., & Matthews, B. (2015). Estimating bacterial diversity for ecological studies: methods, metrics, and assumptions. *PloS one*, *10*(4), e0125356.

Björn, D, Lindahl, R, Nilsson, H, Tedersoo, L, Abarenkov, K, Carlsen, T, Kjøller, R, Kõljalg, U, Pennanen, T, Rosendahl, S, Stenlid, J, and Kauserud, H. (2013). Fungal community analysis by high-throughput sequencing of amplified markers – a user's guide. *New Phytologist*. 199(1); 288-299.

Bokulich, NA, Kaehler, BD, Rideout, JR, Dillon, M, Bolyen, E, Knight, R, Huttley, GA, and Caporaso, JG. (2018). Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2’s q2-feature-classifier plugin. *Microbiome*. 6; DOI: 10.1186/s40168-018-0470-z.

Botnen, SS, Davey, ML, Halvorsen, R, Kauserud H. (2018). Sequence clustering threshold has little effect on the recovery of microbial community structure. *Mol Ecol Resour*.18(5); 1064-1076.

Brodin, J., Hedskog, C., Heddini, A., Benard, E., Neher, R. A., Mild, M., & Albert, J. (2015). Challenges with using primer IDs to improve accuracy of next generation sequencing. *PLoS One*, *10*(3), e0119123.

Callahan, BJ, McMurdie, PJ, Rosen, MJ, Han, AW, Johnson, AJA, and Holmes, SP. (2016). DADA2: High resolution sample inference from Illumina amplicon data. *Nat Methods*. 13(7); 581–583.

Caporaso, JG, Kuczynski, J, Stombaugh, J, Bittinger, K, Bushman, FD, Costello, EK, Fierer, N, Pen˜a, AG, Goodrich, JK, Gordon, JI, Huttley, GA, Kelley, ST, Knights, D, Koenig, JE, Ley, RE, Lozupone, CA, McDonald, D, Muegge, BD, Pirrung, M, Reeder, J, Sevinsky, JR, Turnbaugh, PJ, Walters, WA, Widmann, J, Yatsunenko, T, Zaneveld, J, and Knight, R. (2010). QIIME allows analysis of highthroughput community sequencing data. *Nature Methods*. 7(5); 335–336.

Cavassim, MIA, Moeskjaer, S, Moslemi, C, Fields, B, Bachmann, A, Vilhjalmsson, B, Schierup, MH, Young,JPW, Andersen, SU. (2019) bioRxiv 526707; doi: https://doi.org/10.1101/526707

Edgar, R.C. et al., 2011. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* , 27(16), pp.2194–2200.

Edgar, RC. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*. 10(10); 996–998. DOI 10.1038/nmeth.2604.

Edgar, RC. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Methods*. 10(10); 996-8. DOI: 10.1038/nmeth.2604.

Edgar, RC. (2016). UCHIME2: improved chimera prediction for amplicon sequencing. bioRxiv 074252; doi: https://doi.org/10.1101/074252

Edgar, RC. (2017). UNBIAS: an attempt to correct abundance bias in 16S sequencing, with limited success. *bioRxiv*. 124149. DOI 10.1101/124149.

Elbrecht, V, and Leese, F. (2015). Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass–sequence relationships with an innovative metabarcoding protocol. *PLoS ONE*. 10; e0130324. DOI: 10.1371/journal.pone. 0130324

Faith, J.J. et al., 2013. The long-term stability of the human gut microbiota. *Science*, 341(6141), p.1237439.

Fierer, N, Brewer, T, and Choudoir, M. (2017). Lumping versus splitting – is it time for microbial ecologists to abandon OTUs? http://fiererlab.org/2017/05/02/lumping-versus-splitting-is-it-time-for-microbial-ecologists-to-abandon-otus/ [Accessed: 5/12/2018]

Fonseca, VG. (2018). Pitfalls in relative abundance estimation using eDNA metabarcoding. *Mol Ecol Resour*. 18; 923–926.

Fu, L, Niu, B, Zhu, Z, Wu, S, and Li, W. (2012). Cd-hit: accelerated for clustering the next generation sequencing data. *Bioinformatics*. 28(23); 3150–3152.

Gohl, DM, MacLean, A, Hauge, A, Becker, A, Walek, D, and Beckman, KB. (2016). An optimized protocol for high-throughput amplicon-based microbiome profiling. *Protocol Exchange*. DOI:10.1038/protex.2016.030.

Hoshino T, Inagaki F (2017) Correction: Application of Stochastic Labeling with Random-Sequence Barcodes for Simultaneous Quantification and Sequencing of Environmental 16S rRNA Genes. PLOS ONE 12(3): e0173546.

Huse, SM, Welch, DM, Morrison, HG, and Sogin, ML. (2010). Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ Microbiol*. 12(7); 1889–1898.

Jabara, C.B. et al., 2011. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proceedings of the National Academy of Sciences of the United States of America*, 108(50), pp.20166–20171.

Jollife, IT. (2002). Principal Component Analysis. 2nd ed. New York: Springer

Kebschull, JM, and Zador, AM. (2015). Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Research*. 43(21); e143. DOI 10.1093/nar/gkv717.

Kinde, I. et al., 2011. Detection and quantification of rare mutations with massively parallel sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, 108(23), pp.9530–9535.

Kinoti, W.M. et al., 2017. Generic Amplicon Deep Sequencing to Determine Ilarvirus Species Diversity in Australian Prunus. *Frontiers in microbiology*, 8, p.1219.

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., & Taipale, J. (2012). Counting absolute numbers of molecules using unique molecular identifiers. *Nature methods*, *9*(1), 72.

Koeppel, AF, and Wu, M. (2013). Surprisingly extensive mixed phylogenetic and ecological signals among bacterial Operational Taxonomic Units. *Nucleic Acids Research*. 41(10); 5175–5188. DOI: 10.1093/nar/gkt241.

Krehenwinkel, H, Kennedy, SR, Rueda, A, Lam, A, Gillespie, RG. (2018). Scaling up DNA barcoding – Primer sets for simple and cost efficient arthropod systematics by multiplex PCR and Illumina amplicon sequencing. *Methods in Ecology and Evolution*. 9(11); 2181-2193. DOI: 10.1111/2041-210X.13064

Krueger F, Andrews SR, Osborne CS (2011) Large Scale Loss of Data in Low-Diversity Illumina Sequencing Libraries Can Be Recovered by Deferred Cluster Calling. PLoS ONE 6(1): e16607. https://doi.org/10.1371/journal.pone.0016607

Kumar, N, Lad, G, Giuntini, E, et al. (2015) Bacterial genospecies that are not ecologically coherent: population genomics of Rhizobium leguminosarum. *Open Biology*. 5; 140133.

Lazar, I. (2010). Gelanalyzer 2010a: Freeware 1d gel electrophoresis image analysis software. Available at: http://www.gelanalyzer.com/.

Li, W, and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 22;1658–9.

Lundberg, D. S., Yourstone, S., Mieczkowski, P., Jones, C. D., & Dangl, J. L. (2013). Practical innovations for high-throughput amplicon sequencing. Nature Methods, 10(10), 999–1002. http://doi.org/10.1038/nmeth.2634

McWilliam, H, Li, W, Uludag, M, Squizzato, S, Park, YM, Buso, N, Cow- ley, AP, and Lopez, R. (2013). Analysis tool web services from the embl-ebi. *Nucleic acids research*. 41(W1); W597–W600.

Nguyen, NH, Smith, D, Peay, K, and Kennedy, P. (2014). Parsing ecological signal from noise in next generation amplicon sequencing. *New Phytologist*. DOI: 10.1111/nph.12923.

Oliver, AK, Brown, SP, Callaham, MA, Jumpponen, A. (2015). Polymerase matters: nonproofreading enzymes inflate fungal community richness estimates by up to 15%. *Fungal Ecology*. 15; 86–89. DOI 10.1016/j.funeco.2015.03.003.

Orgiazzi, A., Dunbar, M. B., Panagos, P., de Groot, G. A., & Lemanceau, P. (2015). Soil biodiversity and DNA barcodes: opportunities and challenges. *Soil Biology and Biochemistry*, *80*, 244-250.

Palmer, JM, Jusino, MA, Banik, MT, and Lindner, DL. (2018). Non-biological synthetic spike-in controls and the AMPtk software pipeline improve mycobiome data. *PeerJ*. 6; e4925.

Peres-Neto, PR, Jackson, DA, and Somers, KM. (2005). How Many Principal Components? Stopping Rules for Determining the Number of Non-Trivial Axes Revisited. *British Journal of Statistical Psychology*. 49; 974–97.

Poirier, S. et al., 2018. Deciphering intra-species bacterial diversity of meat and seafood spoilage microbiota using gyrB amplicon sequencing: A comparative analysis with 16S rDNA V3-V4 amplicon sequencing. *PloS one*, 13(9), p.e0204629.

Poisot, T, Péquin, B, and Gravel, D. (2013). High-Throughput Sequencing: A Roadmap Toward Community Ecology. *Ecology and Evolution*. 3(4); 1125-1139.

R Core Team. (2017). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Schloss, PD, and Westcott, SL. (2011). Assessing and Improving Methods Used in Operational Taxonomic Unit-Based Approaches for 16S rRNA Gene Sequence Analysis. *Appl Environ Microbiol*. 77(10); 3219–3226.

Sheward, D. J., Murrell, B., & Williamson, C. (2012). Degenerate Primer IDs and the birthday problem. *Proceedings of the National Academy of Sciences*, *109*(21), E1330-E1330.

Sinclair, L., Osman, O. A., Bertilsson, S., & Eiler, A. (2015). Microbial community composition and diversity via 16S rRNA gene amplicons: evaluating the illumina platform. *PloS one*, *10*(2), e0116955.

Tessler, M, Neumann, JS, Afshinnekoo, E, Pineda, M, Hersch, R, Velho, LFM, Segovia, BT, Lansac-Toha, FA, Lemke, M, DeSalle, R, Mason CE, and Brugler, MR. (2017). Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. *Scientific Reports*. 7; 6589.

Vasileiadis, S., Puglisi, E., Arena, M., Cappa, F., Cocconcelli, P. S., & Trevisan, M. (2012). Soil bacterial diversity screening using single 16S rRNA gene V regions coupled with multi-million read generating sequencing technologies. *PloS one*, *7*(8), e42671.

Verlag. https://goo.gl/SB86SR.

Wen, C, Wu, L, Qin, Y, Van Nostrand, JD, Ning, D, Sun, B, Xue, K, Liu, F, Deng Y, Liang, Y, and Zhou, J. (2017). Evaluation of the reproducibility of amplicon sequencing with Illumina MiSeq platform. *PLoS one*. 12(4); p.e0176716.

Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Zhang, J, Kobert, K, Flouri, T, and Stamatakis, A. (2014). Pear: a fast and accurate Illumina paired-end read merger. *Bioinformatics*. 30(5); 614.