RESEARCH ARTICLE

Ecological inference on bacterial succession using curve-based community fingerprint data analysis, demonstrated with rhizoremediation experiment

Anu Mikkonen, Kaisa Lappi, Kaisa Wallenius, Kristina Lindström & Leena Suominen

Department of Food and Environmental Sciences, University of Helsinki, Helsinki, Finland

Correspondence: Anu Mikkonen, Department of Food and Environmental Sciences, Division of Microbiology, Viikinkaari 9, PO Box 56, FIN-00014 University of Helsinki, Helsinki, Finland. Tel.: + 358 9191 59281; fax: + 358 9191 59322; e-mail: anu.s.mikkonen@helsinki.fi

Received 31 March 2011; revised 20 July 2011; accepted 16 August 2011.
Final version published online 12 September 2011.

DOI: 10.1111/j.1574-6941.2011.01187.x

Editor: Christoph Tebbe

Keywords
length heterogeneity PCR (LH-PCR); hydrocarbon contamination; soil biomonitoring; *Aquabacterium*; *Galega orientalis*.

Abstract

Nucleic acid-based community fingerprinting methods are valuable tools in microbial ecology, as they offer rapid and robust means to compare large series of replicates and references. To avoid the time-consuming and potentially subjective procedures of peak-based examination, we assessed the possibility to apply direct curve-based data analysis on community fingerprints produced with bacterial length heterogeneity PCR (LH-PCR). The dataset comprised 180 profiles from a 21-week rhizoremediation greenhouse experiment with three treatments and 10 sampling times. Curve-based analysis quantified the progressive effect of the plant (*Galega orientalis*) and the reversible effect of the contaminant (fuel oil) on bacterial succession. The major observed community shifts were assigned to changes in plant biomass and contamination level by canonical correlation analysis. A novel method to extract relative abundance data from the fingerprint curves for Shannon diversity index revealed contamination to reversibly decrease community complexity. By cloning and sequencing the fragment lengths, recognized to change in time in the averaged LH-PCR profiles, we identified *Aquabacterium* (*Betaproteobacteria*) as the putative r-strategic fuel oil degrader, and K-strategic *Alphaproteobacteria* growing in abundance later in succession. Curve-based community fingerprint analysis can be used for rapid data prescreening or as a robust alternative for the more heavily parameterized peak-based analysis.

Introduction

Nucleic acid-based community fingerprinting methods provide microbial ecologists with a rapid, reproducible and economic means to analyse large sample numbers required in comparative studies. The first fingerprinting method, denaturing gradient gel electrophoresis (Muyzer *et al.*, 1993), is slowly being replaced by terminal restriction fragment length polymorphism (T-RFLP; Liu *et al.*, 1997), which benefits from easier optimization and better reproducibility of PCR without the GC clamp (Oros-Sichler *et al.*, 2007; Rettedal *et al.*, 2010), as well as sensitivity and more accurate standardization with the capillary electrophoresis and fluorescence-based detection (Kent & Triplett, 2002; Kirk *et al.*, 2004). The disadvantages of T-RFLP arise from the digestion step: more labour, time and money are required in comparison with fingerprinting methods that separate the amplicons directly. In addition, incomplete digestion may produce several peaks from a single sequence (Osborn *et al.*, 2000; Egert & Friedrich, 2003; Mills *et al.*, 2003). Accordingly, when comparing T-RFLP with length heterogeneity PCR (LH-PCR; Suzuki *et al.*, 1998), Mills *et al.* (2003) recommended the latter due to the better reproducibility and lower complexity of the analysis. With LH-PCR, the drawback is the low separation power, as differentiation is based solely on the natural length variation (approx. 460–560 bp) of the first third of the 16S rRNA gene sequence (Tiirola *et al.*, 2003). Thus, the theoretical maximum number of peaks is small, especially compared with automated ribosomal intergenic spacer analysis (ARISA), where the PCR products vary between 200 and 1200 bp

(Ranjard *et al.*, 2003). The ARISA may be especially sensitive in differentiating complex soil communities, but its downside is the significantly poorer reference database for internal transcribed spacer sequences when compared with ribosomal genes.

Although the technical details of these community analyses have been discussed widely, less attention has been given to the subsequent data processing step, which is necessary to refine the numerous fingerprint images into ecologically significant knowledge on the studied phenomena (Blackwood *et al.*, 2003; Thies, 2007; Culman *et al.*, 2008). Most microbial ecological studies analyse fingerprint data as (semi)quantitative 'species abundances' derived from the relative contribution of the different peaks to the total fingerprint signal intensity (Oros-Sichler *et al.*, 2007). This peak-based data analysis approach requires that each putative peak (or band, separated amplicon or ribotype) in each fingerprint profile be (i) separated from noise or background signal (*peak presence*), (ii) assigned to a specific size class or bin (*peak identity*), and (iii) quantified based on presence/absence, absolute height or area, or relative height or area (*peak quantity*) (Osborn *et al.*, 2000; Abdo *et al.*, 2006; Osborne *et al.*, 2006; Thies, 2007). Even if many of these steps can be automated (e.g. Joossens *et al.*, 2011), the results are commonly checked manually (Culman *et al.*, 2008), which is laborious with large data sets and may introduce subjective bias. Another approach is to use the total fingerprint signal (electropherogram of the fluorescence as a function of run time) without peak assignment (Oros-Sichler *et al.*, 2007). This curve-based fingerprint analysis starts with pairwise comparisons of all the fingerprints with Pearson correlation coefficient. The resultant similarity matrix is most often used for descriptive data exploration like clustering or ordination, but hypothesis-driven approaches have also been reported, including group significance testing (Kropf *et al.*, 2004), and multivariate analysis with reference data using Mantel test (Brad *et al.*, 2008). In addition to rapidity and reproducibility with fewer parameters to be optimized, the curve-based approach may also be more sensitive, as more signal data points are included in the analysis. Furthermore, as the called length of a DNA fragment is seldom a whole number, but a decimal affected for example by sequence purine content (Kaplan & Kitts, 2003), this additional information between the integer length bins may bear biological significance.

The objective of this study was to investigate whether automated curve-based community fingerprint data analysis can be used for statistically sound and ecologically meaningful inference on soil bacterial dynamics. To assess community succession and effects of different treatments, we analysed samples from a 21-week greenhouse rhizoremediation experiment with fuel oil as the soil contaminant and legume *Galega orientalis* as the plant (Mikkonen *et al.*, 2011). The rhizoremediation treatment (oil and plant) was monitored in parallel with uncontaminated (plant, but no oil) and unvegetated (oil, but no plant) references at 10 destructive samplings. The LH-PCR was selected as the community fingerprinting method due to its reproducibility and technical simplicity. Six replicates resulted in 180 bacterial community fingerprints, which we examined to address whether curve-based data analysis could facilitate (i) quantifying treatment effects and changes in community complexity as a function of time, and (ii) linking major community shifts to specific changes in explanatory variables and bacterial taxa.

Materials and methods

Experimental setup

The details of the rhizoremediation monitoring greenhouse study were described in detail by Mikkonen *et al.* (2011), and are only summarized here. The experimental setup was designed for destructive sampling and included three biological replicates (pots) per treatment per sampling week. The three treatments were Contaminated Vegetated (ContVeg, i.e. rhizoremediation treatment), Contaminated (Cont, i.e. unvegetated reference) and Vegetated (Veg, i.e. uncontaminated reference). The biological nature of the fuel oil dissipation was confirmed with sterile reference [$AgNO_3$ 3 g (kg freshwt soil)$^{-1}$]. Each independent replicate pot was placed on a different block (greenhouse table divided to three adjacent sectors), and the positions of all the pots within the block were randomized weekly. The experimental soil (final composition sandy loam with total C% 1.8, N% 0.7, pH_{CaCl2} 5.4) was prepared separately for each 2 L pot. For the contaminated treatments, soil was contaminated to 3 g (kg drywt soil)$^{-1}$ with fuel oil using sand as a carrier matrix to ensure homogenous mixing. Equal amounts of sand, but no oil was added to the uncontaminated reference. The soils were allowed to stabilize after the preparation and contamination of the pots for 4 days before sowing. For the vegetated treatments, 25 seeds of the legume *G. orientalis* cultivar Gale (fodder galega) were then sowed per pot and inoculated with *Rhizobium galegae* HAMBI540 by spraying diluted broth culture on the soil surface. On week 5, vegetation was thinned to five shoots per pot. Soil moisture was maintained at approximately 30% of the water holding capacity; soil dry weight correlated with no other variable in any of the treatments, which confirmed that the normalization of moisture level was successful. First, samples (week 0) were taken immediately

after sowing, after which sampling was performed at increasing intervals (weeks 1, 2, 3, 4, 6, 8, 12, 16 and 21). Upon sampling, the entire content of the pot was sieved (< 2 mm) and mixed manually. Subsamples for DNA extraction were stored at −70 °C.

Bacterial community profiling by length heterogeneity PCR

Soil DNA extraction

The DNA was extracted using FastDNA SPIN Kit for Soil (Qbiogene) from 0.50 g fresh soil samples according to the manufacturer's instructions, except for prolongation of the lysing matrix tube centrifugation time (5 min) and increase of the final elution volume (120 μL). Duplicate DNA extractions from three replicates of three treatments at 10 sampling points resulted in 180 DNA extracts. The quantity of the extracted DNA was measured fluorometrically (PicoGreen dsDNA Quantitation Reagent Kit; Molecular Probes).

PCR and capillary electrophoresis

The LH-PCR protocol for bacterial community fingerprinting was modified from Tiirola *et al.* (2003) using the general bacterial primers fD1 (AGAGTTTGATCCTGGCTCAG) and 5′FAM-labelled PRUN518r (ATTACCGCG GCTGCTGG). One microlitre of 1/10 water-dilution of the DNA extract (1–5 ng of DNA) was used as a template, so that the concentration of the co-extracted humic PCR inhibitors (indicated by the yellow colour of the extract) was standardized instead of the exact DNA amount. PCR was executed in the final reaction volume of 50 μL containing 0.2 mM of each dNTP (Finnzymes, Finland), 0.3 mM of both primers (Oligomer, Finland), 0.05% of bovine serum albumin (BSA acetylated; Promega), 1 × Biotools reaction buffer with 2 mM $MgCl_2$ and 1 U of DNA polymerase (Biotools, Spain). Peltier Thermal Cycler DNA Engine (MJ Research) was used for the amplification with the following programme: initial denaturation at 95 °C for 5 min, followed by 30 cycles of: 94 °C for 45 s, 55 °C for 1 min and 72 °C for 2 min. The quantity and quality of the PCR products were checked on a 2% agarose gel. No PCR amplification replicates were included, as PCR and capillary electrophoresis were tested to be highly reproducible (profile similarity of replicate PCR products and capillary runs from the same DNA extract > 99.8% with Pearson correlation).

The amplicons were separated based on their length with polyacrylamide capillary electrophoresis using ABI PRISM 310 Genetic Analyser with a 47-cm capillary and POP-6 Polymer (Applied Biosystems). Between 0.5 and 3 μL of the PCR product or its water-dilution was mixed with Hi-Di Formamide (Applied Biosystems) and self-made molecular size standard comprising similarly amplified HEX-labelled PCR products of known length (Tiirola *et al.*, 2003), to give the final sample volume of 15 μL. The amount of loaded PCR product was optimized to increase sensitivity by high signal/noise ratio, but signal saturation by overloading was avoided (highest sample peak above ~ 1000 but below ~ 8000 relative fluorescence units). The mixture was denatured at 98 °C for 2 min (Peltier Thermal Cycler DNA Engine; MJ Research) and ran using the Genetic Analyser with the following conditions: injection seconds 10 s, injection and run voltage 15.0 kV, run temperature 60 °C and run time 70 min. The different fluorophore labels bound to the reverse primer have slightly different effect on the mobility of the single-stranded DNA, and this dye shift should be determined (Sutton *et al.*, 2011), especially if the called lengths are compared with sequence lengths *in silico*. With our analytical system, using HEX-labelled standards and FAM-labelled samples, the called sizes of the amplicons were consistently 0.7 bp shorter than the actual sequence lengths.

Fingerprint data processing and profile similarity calculation

The fingerprint electropherograms for HEX- and FAM-traces of each sample run were imported from GENESCAN 3.7 (Applied Biosystems) to BIONUMERICS 5.0 (Applied Maths, Sint-Martens-Latem, Belgium) as 12-bit densitometric curves with the Curve Conventer operation. Sample FAM-curves were normalized (aligned) with the internal HEX-labelled standards. The fingerprint active area of analysis was restricted to the expected amplicon size 460–565 bp, which corresponded exactly to 1605 data points (resolution 15.3 points per bp). Pairwise profile similarities were calculated with Pearson correlation coefficient, which accounts for the whole range of the 1605 data points, and is unaffected by differences in the total fingerprint intensity. No signal detection threshold was applied, but if completely empty areas in the profiles would exist, their exclusion might be recommended to avoid overestimation of similarity with Pearson correlation due to shared zero counts. Optimization of 0.30%, corresponding to $\pm$1 bp shift, was allowed to optimize the alignment of the two profiles compared. Average fingerprints (curves) for specific treatment and week were created from the replicates with Create averaged fingerprint script.

Cloning, clone library screening and sequencing

To identify the bacterial taxa that caused significant community shifts, 16S rRNA genes from selected weeks were

cloned and sequenced. The six replicate DNA extracts from the specific treatment and week were mixed in equal ratios, diluted 1/10 in water and used as a template in a new PCR reaction. This PCR was performed with the same LH-PCR primers fD1 and PRUN518r, but neither of the primers was labelled, and a proofreading DNA polymerase was used to minimize amplification errors. The reaction mixture was optimized for the proofreading polymerase and contained 0.2 mM of each dNTP (Finnzymes), 0.5 mM of both primers (Oligomer), 0.05% of BSA (Promega), 1 × Phusion HF buffer (Finnzymes), 1 U of Phusion DNA Polymerase (Finnzymes) and 1 μL of template in the final reaction volume of 50 μL. The PCR cycling conditions were: initial denaturation at 98 °C for 3 min, 30 cycles of 98 °C for 20 s, 57 °C for 45 s, 72 °C for 30 s and final elongation at 72 °C for 5 min. The PCR products were purified with MinElute PCR Purification Kit (Qiagen, Germany) to remove the proofreading enzyme activity, after which the A overhangs required in TA cloning were added to them. The reaction volume of 20 μL [1 × Biotools reaction buffer with 2 mM $MgCl_2$ (Biotools), 0.2 mM dATP (Finnzymes), 0.5 U of DNA Polymerase (Biotools) and 9 μL of the PCR product] was incubated at 72 °C for 20 min. The 16S rRNA gene fragments of approximately 500 bp were cloned into TOP10 cells using TOPO-TA Cloning kit (Invitrogen) according to the manufacturer's instructions. The clone libraries were screened with the LH-PCR protocol described above according to the pooling technique of Grant & Ogilvie (2004) to identify and select the clones containing the amplicon lengths of interest. The plasmids were extracted using GeneJET Plasmid Miniprep Kit (Fermentas), and both strands of the inserts were sequenced at the DNA Sequencing and Genomics Laboratory, Institute of Biotechnology, University of Helsinki using the cloning vector primers UP (M13 -20) and RP.

Fingerprint and sequence data analysis

Unconstrained ordination of the LH-PCR community profiles with multidimensional scaling (MDS) was performed in BIONUMERICS 5.0 using Pearson correlation as the similarity coefficient. BIONUMERICS optimizes the positions of the entries in the three-dimensional MDS ordination space using a gradient descent method by iteratively recalculating them until the improvement between the new solution and the previous one is smaller than 0.1%.

Aligned fingerprint curves were exported from BIONUMERICS 5.0 to MS Excel 2007 and normalized by the total fluorescence (sum of fluorescence intensity across all 1605 data points) over the entire series of 180 profiles for comparative visualization. The MS Excel 2007, SPSS 15 for Windows and freely downloadable FORTRAN programme CAP (Canonical analysis of principal coordinates; Anderson, 2004) were for used for calculations and statistical testing. Due to the innate skewed distribution of Pearson correlation coefficients, all calculations with the fingerprint profile similarities were performed after Fisher (z) transformation of the similarity values (Wallenius *et al.*, 2010). After that the data were roughly normal, as checked with Normal Q–Q plots, and allowed calculation of mean, SD and 95% confidence intervals. All the six replicate DNA extracts were considered as equal replicates, as they are systematically clustered by the three blocks (Ward dendrograms of Pearson correlations, calculated in BIONUMERICS 5.0) in only five of 30 week × treatment combinations, suggesting that the duplicate extracts from the same pot were not consistently more similar than extracts from other replicate pots. An alternative solution would have been to use average profiles calculated from the two replicate extractions, but we specifically wanted to evaluate the suitability of the curve-based analysis for unmodified fingerprint raw data. The effects of contamination and vegetation were calculated as weekly average similarities between Contaminated Vegetated treatment and reference treatments. The significance of the effect was evaluated by permutation test in the CAP program (generalized discriminant analysis, Anderson & Robinson, 2003) with 9999 permutations. For the dissimilarity matrix, the fingerprint distances were calculated by subtracting the pairwise Pearson correlation similarity values from 1 (Ramette, 2007). Effect significance was also evaluated against weekly estimates for a 'dissimilarity baseline' (1 – lower 95% confidence limit for within-treatment similarities).

The CAP program was used also for canonical correlation analysis to assess links between bacterial community structure and environmental variables. According to Anderson & Willis (2003), the advantages of CAP over previously proposed methods for ecological constrained multivariate analysis are that any distance measure can be used, and no unrealistic assumptions about the species or environmental data sets (multivariate normality and independence) or their relationship (fixed unimodal or linear relationship) are made. The fingerprint dissimilarities (1 – Pearson correlation) comprised the distance matrix of the 'response data'. Additional variables reported by Mikkonen *et al.* (2011) were used as 'explanatory data'; however, the calculation itself treats both multivariate data sets symmetrically (Anderson & Willis, 2003). Only those six sampling weeks, on which the full set of the explanatory variables were measured, were included in the analysis. Significance testing of the canonical correlation analysis was based on 9999 permutations. Correlations of principal coordinate scores printed by the CAP program, as well as normalized fingerprint fluorescence

intensities at specific data points (in bp), with other variables were calculated using nonparametric Spearman rank-order correlation.

Shannon diversity index H' was calculated automatically from the normalized fingerprint curve data without traditional peak assignment, by regarding at least ½ bp continuous rises in the curve as a unit of diversity; such a rise corresponds to a peak 1 bp wide, the empirically observed minimum width of a discrete peak from a single cloned sequence (data not shown). More specifically, if the fluorescence intensity increased for at least seven of eight consecutive data points (1 bp was 15.2 data points, and incontinuity or noise of 1/8 data point was allowed), this was considered as unit of diversity comparable to the traditional peak concept. The quantity was the height of the consecutive rise (in normalized relative fluorescence units). The calculations were performed in MS Excel 2007 with basic mathematical and logical functions. The relative proportions of these consecutive increments were used as p_i in the Shannon diversity index formula $H' = -\sum p_i \ln p_i$. According to Magurran (2004), parametric statistics can be used on Shannon index data, and so mean and standard deviation (SD) of H' were calculated directly from the six replicates without any transformation. Due to the small replicate numbers, nonparametric Kruskal–Wallis test was used to test if Shannon H' differed in the three treatments.

The sequences were analysed with BIONUMERICS 6.0 (Applied Maths, Sint-Martens-Latem, Belgium), where the quality of the chromatograms was checked, the consensus sequence constructed and primer regions excluded. The phylogenetic affiliation of the trimmed 16S rRNA gene sequences (length 431–486 bp) were retrieved by searching for its nearest neighbours with SINA Webaligner of the SILVA rRNA database project (Pruesse *et al.*, 2007) and recording their taxonomic positions according to the SILVA taxonomy, which utilizes the current reference ARB/SILVA rRNA alignment (Schloss, 2009). The similarity of the sequences with their nearest neighbours was calculated in BIONUMERICS 6.0 using the Standard Pairwise alignment algorithm with default settings. Sequences were deposited in European Nucleotide Archive (EMBL-Bank) under accession numbers FR878041–FR878058 with full minimum information about a marker gene sequence (MIMARKS; Yilmaz *et al.*, 2011).

Results and discussion

Quantification of the specific effects of contamination and vegetation on bacterial community succession

Bacterial community fingerprints can hardly reproduce the complete picture of the target community, but they are suitable for monitoring community responses to alterations (Hartmann & Widmer, 2008). One of the major aims of our rhizoremediation experiment was to follow up the specific effects of contamination and vegetation on soil chemical and microbiological indicators (Mikkonen *et al.*, 2011). As negative reference, treatments for both factors were monitored in parallel, and we were, in the current work, able to quantify their relative contributions to the succession of the bacterial community in the Contaminated Vegetated soil. Figure 1 shows the effects of the oil and the plant on bacterial community structure as a function of time, visualized as curve-based community profile dissimilarity (1 – Pearson correlation) between the rhizoremediation treatment and the reference. Even if the curve-based analysis showed consistently lower dissimilarities than equivalent peak-based analysis (Supporting Information, Fig. S1), the trends and statistical conclusions of the effects were nearly identical.

The effect of oil on bacterial community structure was immediate, major and by and large reversible: a massive decrease in fingerprint similarity between Contaminated Vegetated treatment and Vegetated reference was seen already on week 1 (Fig. 1). The differentiation of the

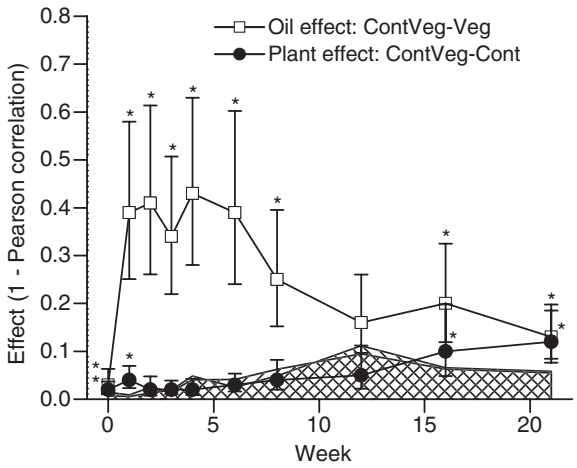

Fig. 1. The specific effects of vegetation and contamination on bacterial community succession during rhizoremediation process. The effect size is shown as dissimilarity of community structures (1 – curve-based Pearson correlation coefficient for similarity) between the rhizoremediation treatment and the reference treatment. Mean and SD were calculated with Fisher-transformed Pearson correlations crossed between six replicates of each treatment ($n = 36$ for each figure point). Significant effect is depicted with an asterisk and stands for significant difference ($P < 0.05$) between the treatments, calculated with generalized discriminant analysis with 9999 permutations of the dissimilarity matrix. In addition, nonsignificant 'baseline dissimilarity' is visualized as hatched area, calculated based on weekly lower 95% confidence limits of within-treatment similarities ($n = 30$): (/) for ContVeg and Veg (baseline for oil effect) and (\) for ContVeg and Cont (baseline for plant effect).

community structure was parallel to the observed increases in oil degrader counts and soil microbial biomass (Mikkonen *et al.*, 2011). In addition, the bacterial community response mirrored the typical hockey stick-shaped hydrocarbon dissipation curve (fig. 2a in Mikkonen *et al.*, 2011). Although the oil effect was notably reduced along the contaminant biodegradation process, it was still significant on week 21, when the soil hydrocarbon content had reached the legislative clean soil threshold level (0.3 g kg^{-1}, i.e. 10% of the initial contamination load). Margesin *et al.* (2007), who monitored a diesel biodegradation process in microcosms with soil PLFA profiling, found no significant community effects of contamination or incubation time at the rather low hydrocarbon concentration comparable with our initial contamination load (2.5 g kg^{-1}). Bacterial community fingerprinting with LH-PCR thus appears to be a more sensitive microbiological indicator of soil contamination and restoration.

The effect of plant on bacterial community structure was smaller than the effect of oil (Fig. 1). An increasing trend of the impact parallel with the growth of the plants was observed, but the rhizosphere effect was significant only from week 16 (curve-based approach) or week 12 (peak-based approach, Fig. S1). The sampling strategy to sieve and mix the whole content of the pot probably delayed the observation of a rhizosphere effect, which was now seen practically only when the entire soil volume was in the close proximity of galega roots. On week 21, the effect of the plant *per se* may have been enhanced also by the slightly, but significantly lower contamination level reached in the rhizoremediation treatment (90% of fuel oil dissipated vs. 87% dissipation in Contaminated reference; Mikkonen *et al.*, 2011). In addition to the effect of the grown plants on the late weeks of the experiment, a significant, but transient difference, between Contaminated Vegetated treatment and Contaminated reference was seen on weeks 0–1. Possible explanation for this was the rhizobial inoculant sprayed only on the galega seeds in the former treatment, emphasizing the sensitivity of DNA-based monitoring.

Grouping pattern of the bacterial community profiles

Visualization of community structure grouping patterns with MDS 'lets the data speak for itself' in explorative analysis (Anderson & Willis, 2003; Ramette, 2007) and, as a pre-screening method, helps to identify those representative exemplars best suitable for more in-depth analysis (Abdo *et al.*, 2006). The MDS plot in Fig. 2 visualizes the similarities (curve-based Pearson correlation) of the 180 LH-PCR profiles of the rhizoremediation experiment. Three major groups were observed, corresponding to treatment (oil or no oil) and sampling week. The succession of the Vegetated reference proceeded directly from group I (early communities) to group III (developed communities). Both contaminated treatments jumped from group I to group II (oil-degrading communities) immediately on week 1, and as the oil was degraded, developed again in the same direction with the uncontaminated soils to group III. The soil bacterial resilience to contamination was thus not explained by the community structure returning to the precontamination state, but was associated with the community developing in the same direction as in the uncontaminated soil. Peak-based MDS ordination (Fig. S2), in the other hand, showed no clear groups, but depicted separate circular paths of succession for contaminated treatments and Vegetated reference. Applying different (dis)similarity measures in ecological studies may reveal different insights (Anderson & Robinson, 2003), but with the current dataset, the conclusions were similar with both fingerprint data analysis approaches: contaminated soils differed from uncontaminated soil, plant effect was minor even at the end of the experiment, and clear time-dependent changes were seen. With regard to bioprocess monitoring, our findings emphasize the need for sufficient references to discriminate between treatment-derived and uncontrolled or time-related effects.

Correspondence between changes in bacterial community structure and other soil parameters

Canonical correlation analysis of principal coordinates was performed with the CAP program (Anderson, 2004) to analyse the dynamic relationship between LH-PCR profiles and a set of other environmental variables monitored throughout the experiment: chemical soil parameters (pH, moisture, C, N, hydrocarbon content) and plant and microbial biomass (galega shoot dryweight, soil DNA content) (detailed data in Mikkonen *et al.*, 2011). In addition to exploring the multivariate associations, we wanted to assess how strictly the controlled parameters (contamination, vegetation and sampling week) determined the bacterial community structure. One advantage of the CAP program is that it allows input of the response data as dissimilarity matrix, which could be derived from the fingerprint Pearson correlations without peak binning.

The canonical analysis with the CAP program starts with an unconstrained ordination, a principal coordinate analysis (PCO) of the community profiles. The PCO was in essence parallel to the MDS ordination (Fig. 2): the fingerprint scores on the first PCO axis correlated with the *x*-axis scores of MDS at $r = 0.997$. As this first PCO

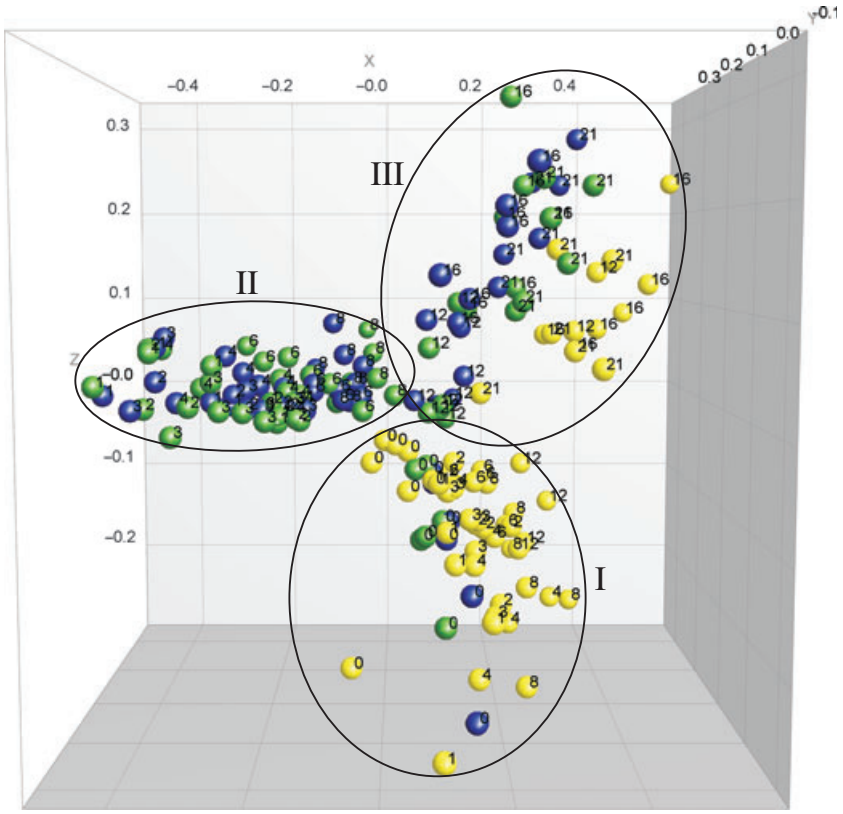

Fig. 2. Three-dimensional multidimensional scaling ordination for the 180 LH-PCR profiles of Contaminated Vegetated treatment (green/medium grey), Contaminated reference (blue/dark grey) and Vegetated reference (yellow/light grey) from sampling weeks 0 to 21 (Arabic number next to the symbol). The observed groups (Roman numbers) contain: (I) Early group – all treatments on week 0 and Veg on weeks 1–12. (II) Oil-degrading group – ContVeg and Cont on weeks 1–8. (III) Developed group – all treatments on weeks 12–21.

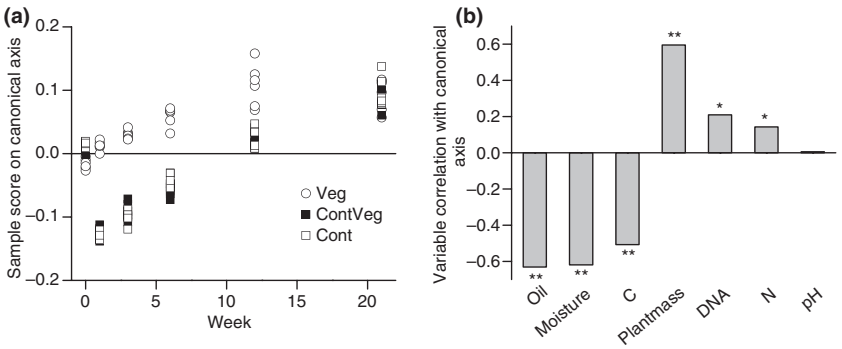

Fig. 3. Canonical correlation analysis of principal coordinates showing the relationships between the development of bacterial community structure and a set of other monitored variables. The two multivariate data sets are reduced to univariate correlations: 79% of community variation is expressed on the single canonical axis that relates the fingerprint profiles to the explanatory variables. (a) Visualizes the distribution of bacterial community profiles on the canonical axis as a function of time, whereas (b) shows the size and direction of correlation of the explanatory variables with the same canonical axis. Asterisks next to the column mark significant correlation between variable values and community PCO scores (*$P < 0.05$ and **$P < 0.01$, Spearman correlation).

axis alone sufficed to explain 98% of the variation in the community structure dissimilarities, only a single canonical axis was formed in the following constrained ordination with the explanatory variables. Figure 3a illustrates the bacterial community development in the three different treatments by showing the scores of the bacterial

community profiles (response data) on this single canonical axis as a function of the sampling week. The correlations of the explanatory variables with the same canonical axis are visualized in Fig. 3b. The statistics of the constrained analysis – the high eigenvalue 0.79 of the single obtained canonical axis and permutated *P*-value 0.0001 – suggest that the observed relationships between the bacterial community structure and other variables were not due to chance.

The controlled experimental parameters oil (soil total petroleum hydrocarbon content) and plant (galega shoot dry weight) were the variables that had the single highest negative and positive correlation with the canonical axis respectively. The ability of the constrained analysis to explain distribution of the community structure data was nearly equally high if only these two explanatory variables were used instead of all seven variables [canonical axis eigenvalue(2) 0.73 vs. eigenvalue(7) 0.79, canonical analysis trace statistic $P = 0.0001$ for both]. Both of these variables were also time dependent (Spearman correlation with sampling week $P < 0.01$ with intrinsic value of rho > 0.4). The experimental setup of the greenhouse experiment was thus successful in determining bacterial community structure development, with little unidentified factors (resulting for example from poorly mixed soil material or different conditions in different blocks) confounding the succession.

Soil moisture and carbon were strongly correlated with soil hydrocarbon content and negative canonical axis scores characteristic of the oil-degrading bacterial communities (group II in Fig. 2). Mikkonen *et al.* (2011) showed that soil C/N ratio increased with fuel oil addition and remained elevated even after the contaminant had been degraded, probably due to assimilation of the oil-derived C by the degrader microorganisms. Soil DNA (used as a proxy of microbial biomass) and N contents were more loosely associated to plant size and later phases of the bacterial community development (group III in Fig. 2). Indeed, putative positive effects on soil microbial numbers and nutrient status are the main arguments for the use of legumes in restoration of contaminated soils (Suominen *et al.*, 2000). In contrast to earlier reports (e.g. Kennedy *et al.*, 2004; Fierer & Jackson, 2006), soil pH explained very little of the soil bacterial community structure variation, although there were statistically significant changes in pH during the experiment. A possible explanation for the lack of effect is that the observed changes in pH were small (all values within 0.8 pH units).

Identification of the changing taxa

In peak-based fingerprint analysis, the specific changes in profiles are observed by comparing the relative abundances of binned peak sizes, commonly shown with pie charts or 100% stacked columns. Inference based on such a presentation is very sensitive to inaccuracies in binning, which should be either checked and corrected manually (Culman *et al.*, 2008) or performed with extensive and carefully justified automatic algorithms (Ramette, 2009). The curve-based alternative is to visualize the changes with profiles normalized by total fluorescence intensity. To identify the specific responding regions in the LH-PCR profiles responsible for the community shifts observed during the rhizoremediation experiment (formation of the groups I-III in Fig. 2), we constructed averaged fingerprints from the six replicates per treatment per week. Figure 4 shows the averaged community profiles of the three different treatments from week 0 (beginning, group I), week 3 (during rapid oil degradation, group II for contaminated and group I for uncontaminated soil) and week 21 (end, group III).

The region around called fragment length of 521 bp responded to contamination: it rose to dominate the entire bacterial community during the intensive oil degradation phase (weeks 1–12) in the contaminated soils

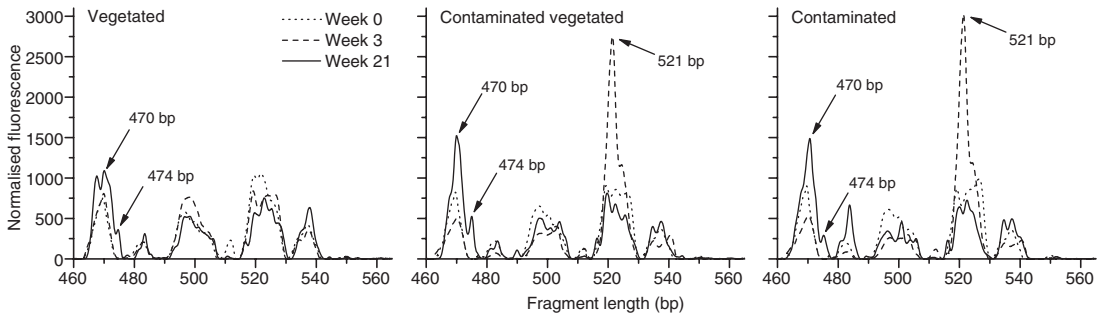

Fig. 4. Bacterial community profiles on weeks 0, 3 and 21 in Vegetated, Contaminated Vegetated and Contaminated soil. Each profile is the arithmetic average of six replicate fingerprints, normalized by the total fluorescence intensity. The regions corresponding to the major bacterial community shifts are indicated.

(Fig. 4). As the increase was associated to doubling of soil microbial biomass (Mikkonen *et al.*, 2011), the growth was not merely relative, but also absolute. The relative fluorescence at 521 bp decreased once the oil degradation levelled off, and this region never peaked in the vegetated reference soil. Besides visual observation, this region could also be quantitatively associated to fuel oil degradation: the Spearman correlation for normalized fluorescence at 521 bp and square root of total petroleum hydrocarbon dissipation per week was 0.729 ($P < 0.001$, $n = 36$). These observations indicate that the amplicons of approx. 521 bp were derived from r-strategic bacteria (Fierer *et al.*, 2007; Shrestha *et al.*, 2007), able to rapidly react to newly available carbon source and to gain competitive advantage from the degradation of the easily consumable fuel oil hydrocarbons, simultaneously cleaning up the soil.

Although the effect of the plant on bacterial community structure during the rhizoremediation process was significant on the late weeks of the experiment (Fig. 1 and S1), this dissimilarity could not be associated to distinct differences in the averaged LH-PCR profiles of the vegetated soils vs. Contaminated reference on week 21 (Fig. 4). The analysis of community fingerprints as multivariate data (based on either curves or peaks) could be expected to detect differences more sensitively and robustly than the alternative the strategy of focussing on narrow regions or single peaks.

On the other hand, a relative increase in the shortest fragments towards the end of the experiment was observed in all three soils. Especially, the amplicons of size 470 and 474 bp peaked on week 21 (Fig. 4). As this was a general trend not restricted to contaminated or vegetated soils, these fragments probably represented K-strategists (Fierer *et al.*, 2007; Shrestha *et al.*, 2007) competitive in stabilized conditions of later successional stages.

To identify these changing taxa, we used the LH-PCR primers to construct clone libraries of each treatment on weeks 3 and 21 (six 96-well plates). All libraries were screened for the r-strategic 521 bp and K-strategic 467–474 bp with LH-PCR. Each cloned fragment that represented these amplicon sizes of interest $\pm$2 bp was sequenced.

Altogether, 33 clones with amplicon size 465–474 bp were found in the three libraries from week 21: nine from the rhizoremediation treatment, 11 from Vegetated reference and 13 from the Contaminated reference. The taxonomic affiliations of the closest SILVA database matches of these K-strategists were *Alphaproteobacteria* (Rhizobiales, Sphingomonadales, Caulobacterales and Rhodospirillales, minimum two clones per order), Cyanobacteria (Leptolyngbya and Nostoc, one each) and several uncertain phyla (two TM7, two OD1 and one OP11). More specifically, the clones with the exact sequence length of 470 bp were always *Alphaproteobacteria* (all four abovementioned orders). *Alphaproteobacteria* are among the most common bacterial taxa in soil (Janssen, 2006) and were also in the work of Shrestha *et al.* (2007) found to prevail in late succession, typical of K-strategists. The clones with the length of 474 bp were the representatives of Cyanobacteria and TM7.

Ten clones representing the oil degradation peak (519–521 bp) were found from the clone libraries of the two contaminated soils from week 3. Eight clones had higher than 97.5% sequence similarity with each other, the two outliers showing minimum 91% and 84% similarity with the other nine sequences. However, the taxonomic affiliation of the closest SILVA database match was the genus *Aquabacterium* (*Betaproteobacteria*) for all the 10 oil degradation-related clones (Table 1). Clones of the size 519–520 bp were not found in the clone libraries of the Vegetated reference or either contaminated soils from week 21. All clones of observed size 516–518 and 521–522 bp from these libraries were sequenced and identified to belong to various genera of Proteobacteria and Firmicutes, but never to the genus *Aquabacterium*.

Our results suggest that the genus *Aquabacterium* had a significant role in the fuel oil degradation process and may have been responsible for the rapid consumption of the easily degradable hydrocarbons. This genus was first characterized from biofilms of drinking water distribution system (Kalmbach *et al.*, 1999), but

Table 1. The closest database matches of the cloned 16S rRNA genes representing the oil degradation-related peak (~520 bp, found solely in week 3 clone libraries of the contaminated treatments), and their taxonomic affiliation according to SILVA taxonomy

Called/real clone amplicon size (bp)	SILVA taxonomy (no. clones from libraries)	Nearest neighbour (% sequence similarity), sequence source
520/521	*Aquabacterium* (3 ContVeg, 2 Cont)	AF089858 (99.8–100), isolate from drinking water biofilm
520/521	*Aquabacterium* (1 ContVeg, 1 Cont)	AB362826 (99.8), environmental isolate grown at elevated CO_2
520/521	*Aquabacterium* (1 Cont)	AF523022 (98.5), clone from natural mineral water
519/520	*Aquabacterium* (1 Cont)	EU567044 (94.8), clone from soil
521/520	*Aquabacterium* (1 ContVeg)	AY662015 (93.0), clone from uranium-contaminated groundwater

has also been found in soils, especially abundant in a Canadian mildly acidic boreal forest soil (up to 2.7% of all bacterial V9 amplicons; Fulthorpe *et al.*, 2008). *Betaproteobacteria*, including *Aquabacterium*, have recently been suggested to have a significant role in the *in situ* degradation of hydrocarbons like benzene (Alfreider & Vogt, 2007) and naphthalene (Yu & Chu, 2005). Along with these earlier findings, our results indicate that *Betaproteobacteria* can in certain cases be a more important active biodegraders of readily utilizable hydrocarbons in soil than *Gammaproteobacteria*, the dominant role of which in contaminated environments (so-called bacterial community 'gamma-shift') has been reported in most former studies (Militon *et al.*, 2010). In addition, Fierer *et al.* (2007) reported that *Betaproteobacteria* are often more abundant in soil than *Gammaproteobacteria*, and exhibit copiotrophic (corresponding to r-strategic) attributes. However, there is little literature data on the ability of betaproteobacterial isolates to grow on alkanes (Parales, 2010), the major and most readily degradable constituent of fuel oil (70% in the oil product used in this experiment; Mikkonen *et al.*, 2011). To our knowledge, this is the first report of a dynamic connection between oil degradation rate and relative abundance of a betaproteobacterial genus (*Aquabacterium*) in soil. Further studies with stable isotope probing as well as isolation of the degraders as pure cultures are required before the putative role of *Aquabacterium* in bioremediation of aliphatic hydrocarbon contaminated soil can be verified.

Curve-based Shannon index calculation to monitor changes in bacterial community complexity

Although one peak or band in a community fingerprint profile seldom corresponds exactly to single species, microbial ecologists have commonly used fingerprint data-based diversity indices for comparisons and ecological inference on community complexity and dynamics (Hill *et al.*, 2003; Hartmann & Widmer, 2008). To avoid the debated procedures of peak calling and quantification (Abdo *et al.*, 2006; Blackwood *et al.*, 2007), we tested the possibility to extract the relative abundance data required in calculation of Shannon diversity index H' directly from the fingerprint curves. Profiles were normalized by the total fluorescence intensity, after which each rise in the curve that continued for at least ½ bp was considered a unit of diversity. Its quantity was the height of the continuous rise (in relative fluorescence units), and its contribution to the sum of all the rises in that profile was used for the Shannon diversity index calculation.

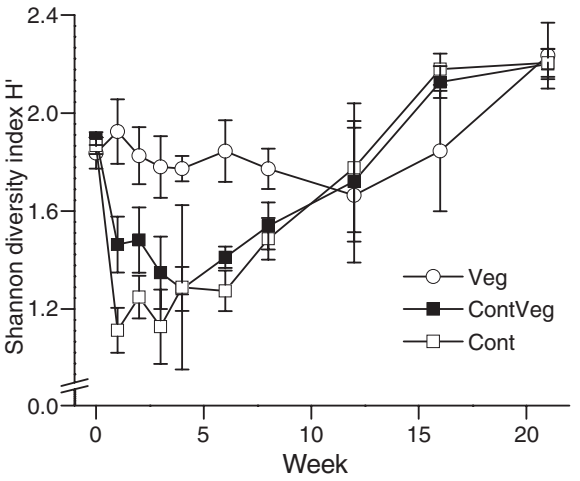

Fig. 5. Changes in the bacterial community complexity in Contaminated Vegetated treatment and Vegetated and Contaminated references. Shannon index H' was calculated directly from the normalized fingerprint profiles without traditional peak assignment. The theoretical scale of Shannon index varies from 0 (one rise) to 3.18 (the maximum observed 24 rises all equally high). The community complexity was significantly reduced by contamination from week 1 to week 8, calculated with Kruskal–Wallis test ($n = 6$ each, $P < 0.05$).

With this curve-based diversity index we could observe a significant, but reversible reduction in Shannon H' due to soil contamination (Fig. 5). The index thus successfully visualized the massive, but temporary dominance of *Aquabacterium* (approx. 521 bp) during the weeks of rapid hydrocarbon degradation (see section Identification of the changing taxa). The same pattern was observed with Shannon H' calculated with the traditional peak-based technique (Fig. S3). Ecologically deleterious effects of soil pollution are usually regarded to be associated with decrease in biodiversity due to disturbance of the most sensitive species. In this experiment, the contamination-associated decrease in Shannon index coincided not with decrease, but with increase of soil microbial biomass during the initial rapid degradation of the fuel oil (fig. 2c in Mikkonen *et al.*, 2011). When anthropogenic environmental disturbance is evaluated, it may also be useful to consider such stimulating ecological effects of contaminants in addition to the deleterious ones (Kefford *et al.*, 2008).

No vegetation-associated effect on the community complexity was observed: Shannon index was indifferent in all soils at the end of the experiment (Fig. 5), regardless of profile similarity comparison showing significant effects of both plant and oil on week 21 (Fig. 1 and S1). In this regard, our results are in agreement with earlier reports (e.g. Hartmann & Widmer, 2006) on higher

sensitivity of (dis)similarity-based fingerprint data analysis methods in detecting differences in bacterial communities.

Conclusion

Curve-based fingerprint analysis has earlier been commonly applied to cluster complex soil community profiles; for example Smalla *et al.* (2007) and Wallenius *et al.* (2011) found that Pearson correlation reliably and sensitively grouped bacterial communities based on soil type or land use, respectively. However, hierarchical clustering alone gives a very narrow view on community data. According to Anderson & Willis (2003), analysis of multivariate ecological data should ideally always include (explorative) unconstrained ordination to observe general patterns, (hypothesis-driven) constrained ordination, statistical testing of *a priori* hypotheses, and identification of species responsible for the patterns. The present work demonstrates how curve-based fingerprint analysis can be applied to all of these data analysis approaches. In addition, we present an alternative method for the extraction of relative abundance data for calculation of Shannon diversity index from the curves. The curve-based approach can be used for rapid screening of large community fingerprint datasets required in ecological studies, as it is more straightforward than peak-based analysis and requires much fewer parameters (related to peak detection, binning and quantification as well as data transformation and analysis) to be optimized. The results of this work also suggest that ecological conclusions on a microbiological process (based on comparisons between different treatments and time-points) may be largely equivalent with curve and peak data, making peak analysis possibly redundant.

More specifically, curve-based fingerprint data analysis was successfully utilized for analysis of 180 LH-PCR bacterial community profiles from a 21-week rhizoremediation experiment. The largely reversible effect of the contaminant (fuel oil) and the progressive effect of the plant (galega) on the soil bacterial populations were not only validated, but their relative contributions could also be quantified by comparison to reference treatments. These controlled parameters sufficed to explain most of the variability in the community structure. In addition, curve-based fingerprint analysis combined to cloning and sequencing facilitated identification of K and r-strategic bacterial taxa. The proportion of *Alphaproteobacteria* increased in the later phases of succession regardless of contamination or vegetation, designating them as K-strategists. *Aquabacterium* dominated the community fingerprints during the intensive oil degradation and its relative abundance correlated with the hydrocarbon degradation rate, suggesting a new r-strategic betaproteobacterial oil degrader genus.

Acknowledgements

We thank Marja Tiirola for advice in setting up the LH-PCR analysis and for the template DNA for LH-PCR standards, Leena A. Räsänen and Aneta Dresler-Nurmi for their advice at the initiation of capillary electrophoresis and BioNumerics analyses, and Hanna Sinkko for helpful discussions on CAP. Aregu Aserse is thanked for technical assistance in screening the clone libraries. Elina Kondo is acknowledged for kindly permitting the use of the oil analysis data (originally presented in Mikkonen *et al.*, 2011) as explanatory variables. This work was funded by the Academy of Finland, Ekokem Oy, University of Helsinki and Maa- ja vesitekniikan tuki ry.

References

Abdo Z, Schüette UME, Bent SJ, Williams CJ, Forney LJ & Joyce P (2006) Statistical methods for characterizing diversity of microbial communities by analysis of terminal restriction fragment length polymorphisms of 16S rRNA genes. *Environ Microbiol* **8**: 929–938.

Alfreider A & Vogt C (2007) Bacterial diversity and aerobic biodegradation potential in a BTEX-contaminated aquifer. *Water Air Soil Pollut* **183**: 415–426.

Anderson MJ (2004) *CAP: a FORTRAN Computer Program for Canonical Analysis of Principal Coordinates*. Department of Statistics, University of Auckland, New Zealand. Available at: http://www.stat.auckland.ac.nz/~mja/Programs.htm, accessed 9 February 2010.

Anderson MJ & Robinson J (2003) Generalised discriminant analysis based on distances. *Aust N Z J Stat* **45**: 301–318.

Anderson MJ & Willis TJ (2003) Canonical analysis of principal coordinates: a useful method of constrained ordination for ecology. *Ecology* **84**: 511–525.

Blackwood CB, Marsh T, Kim SH & Paul EA (2003) Terminal restriction fragment length polymorphism data analysis for quantitative comparison of microbial communities. *Appl Environ Microbiol* **69**: 926–932.

Blackwood CB, Hudleston D, Zak DR & Buyer JS (2007) Interpreting ecological diversity indices applied to terminal restriction fragment length polymorphism data: insights from simulated microbial communities. *Appl Environ Microbiol* **73**: 5276–5283.

Brad T, Van Breukelen BM, Braster M, Van Straalen NM & Röling WF (2008) Spatial heterogeneity in sediment-associated bacterial and eukaryotic communities in a landfill leachate-contaminated aquifer. *FEMS Microbiol Ecol* **65**: 534–543.

Culman SW, Gauch HG, Blackwood CB & Thies JE (2008) Analysis of T-RFLP data using analysis of variance and

ordination methods: a comparative study. *J Microbiol Methods* **75**: 55–63.

Egert M & Friedrich MW (2003) Formation of pseudo-terminal restriction fragments, a PCR-related bias affecting terminal restriction fragment length polymorphism analysis of microbial community structure. *Appl Environ Microbiol* **69**: 2555–2562.

Fierer N & Jackson RB (2006) The diversity and biogeography of soil bacterial communities. *Proc Natl Acad Sci USA* **103**: 626–631.

Fierer N, Bradford MA & Jackson RB (2007) Toward an ecological classification of soil bacteria. *Ecology* **88**: 1354–1364.

Fulthorpe R, Roesch LFW, Riva A & Tripplet EW (2008) Distantly sampled soils carry few species in common. *ISME J* **2**: 901–910.

Grant A & Ogilvie LA (2004) Name that microbe: rapid identification of taxa responsible for individual fragments in fingerprints of microbial community structure. *Mol Ecol Notes* **4**: 133–136.

Hartmann M & Widmer F (2006) Community structure analyses are more sensitive to differences in soil bacterial communities than anonymous diversity indices. *Appl Environ Microbiol* **72**: 7804–7812.

Hartmann M & Widmer F (2008) Reliability of detecting composition and changes of microbial communities by T-RFLP genetic profiling. *FEMS Microbiol Ecol* **63**: 248–260.

Hill TCJ, Walsh KA, Harris JA & Moffett BF (2003) Using ecological diversity measures with bacterial communities. *FEMS Microbiol Ecol* **43**: 1–11.

Janssen PH (2006) Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes. *Appl Environ Microbiol* **72**: 1719–1728.

Joossens M, Huys G, Van Steen K, Cnockaert M, Vermeire S, Rutgeerts P, Verbeke K, Vandamme P & De Preter V (2011) High-throughput method for comparative analysis of denaturing gradient gel electrophoresis profiles from human fecal samples reveals significant increases in two bifidobacterial species after inulin-type prebiotic intake. *FEMS Microbiol Ecol* **75**: 343–349.

Kalmbach S, Manz W, Wecke J & Szewzyk U (1999) *Aquabacterium* gen. nov., with description of *Aquabacterium citratiphilum* sp. nov., *Aquabacterium parvum* sp. nov. and *Aquabacterium commune* sp. nov., three *in situ* dominant bacterial species from the Berlin drinking water system. *Int J Syst Bacteriol* **49**: 769–777.

Kaplan CW & Kitts CL (2003) Variation between observed and true terminal restriction fragment length is dependent on true TRF length and purine content. *J Microbiol Methods* **54**: 121–125.

Kefford BJ, Zalizniak L, Warne MSJ & Nugegoda D (2008) Is the integration of hormesis and essentiality into ecotoxicology now opening Pandora's Box? *Environ Pollut* **151**: 516–523.

Kennedy N, Brodie E, Connolly J & Clipson N (2004) Impact of lime, nitrogen and plant species on bacterial community structure in grassland microcosms. *Environ Microbiol* **6**: 1070–1080.

Kent AD & Triplett EW (2002) Microbial communities and their interactions in soil and rhizosphere ecosystems. *Annu Rev Microbiol* **56**: 211–236.

Kirk JL, Beaudette LA, Hart M, Moutoglis P, Klironomos JN, Lee H & Trevors JT (2004) Methods of studying soil microbial diversity. *J Microbiol Methods* **58**: 169–188.

Kropf S, Heuer H, Gruning M & Smalla K (2004) Significance test for comparing complex microbial community fingerprints using pairwise similarity measures. *J Microbiol Methods* **57**: 187–195.

Liu W-T, Marsh TL, Cheng H & Forney LJ (1997) Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. *Appl Environ Microbiol* **63**: 4516–4522.

Magurran AE (2004) *Measuring Biological Diversity*. Blackwell Publishing, Oxford, UK.

Margesin R, Hämmerle M & Tscherko D (2007) Microbial activity and community composition during bioremediation of diesel-oil-contaminated soil: effects of hydrocarbon concentration, fertilizers, and incubation time. *Microb Ecol* **53**: 259–269.

Mikkonen A, Kondo E, Lappi K, Wallenius K, Lindström K, Hartikainen H & Suominen L (2011) Contaminant and plant-derived changes in soil chemical and microbiological indicators during fuel oil rhizoremediation with *Galega orientalis*. *Geoderma* **160**: 336–346.

Militon C, Boucher D, Vachelard C, Perchet G, Barra V, Troquet J, Peyretaillade E & Peyret P (2010) Bacterial community changes during bioremediation of aliphatic hydrocarbon-contaminated soil. *FEMS Microbiol Ecol* **74**: 669–681.

Mills DEK, Fitzgerald K, Litchfield CD & Gillevet PM (2003) A comparison of DNA profiling techniques for monitoring nutrient impact on microbial community composition during bioremediation of petroleum-contaminated soils. *J Microbiol Methods* **54**: 57–74.

Muyzer G, de Wall E & Uitterlinden A (1993) Profiling of complex microbial populations by DGGE of PCR-amplified genes coding for 16S rRNA. *Appl Environ Microbiol* **59**: 695–700.

Oros-Sichler M, Costa R, Heuer H & Smalla K (2007) Molecular fingerprinting techniques to analyze soil microbial communities. *Modern Soil Microbiology* (van Elsas JD, Jansson JK & Trevors JT, eds), pp. 355–386. CRC Press, Oxon, UK.

Osborn AM, Moore ERB & Timmis KN (2000) An evaluation of terminal-restriction fragment length polymorphism (T-RFLP) analysis for the study of microbial community structure and dynamics. *Environ Microbiol* **2**: 39–50.

Osborne CA, Rees GN, Bernstein Y & Janssen PH (2006) New threshold and confidence estimates for terminal restriction fragment length polymorphism analysis of complex bacterial communities. *Appl Environ Microbiol* **72**: 1270–1278.

Parales RE (2010) Hydrocarbon degradation by Betaproteobacteria. *Handbook of Hydrocarbon and Lipid Microbiology* (Timmis KN, ed.), pp. 1715–1724. Springer, Berlin, Germany.

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J & Glöckner FO (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* **35**: 7188–7196.

Ramette A (2007) Multivariate analyses in microbial ecology. *FEMS Microbiol Ecol* **62**: 142–160.

Ramette A (2009) Quantitative community fingerprinting methods for estimating the abundance of operational taxonomic units in natural microbial communities. *Appl Environ Microbiol* **75**: 2495–2505.

Ranjard L, Lejon DPH, Mougel C, Schehrer L, Merdinoglu D & Chaussod R (2003) Sampling strategy in molecular microbial ecology: influence of soil sample size on DNA fingerprinting analysis of fungal and bacterial communities. *Environ Microbiol* **5**: 1111–1120.

Rettedal EA, Clay S & Brözel VS (2010) GC-Clamp primer batches yield 16S rRNA amplicon pools with variable GC clamps, affecting denaturing gradient gel electrophoresis profiles. *FEMS Microbiol Lett* **312**: 55–62.

Schloss PD (2009) A high-throughput DNA sequence aligner for microbial ecology studies. *PLoS ONE* **4**: e8230.

Shrestha PM, Noll M & Liesack W (2007) Phylogenetic identity, growth-response time and rRNA operon copy number of soil bacteria indicate different stages of community succession. *Environ Microbiol* **9**: 2464–2474.

Smalla K, Oros-Sichler M, Milling A, Heuer H, Baumgarte S, Becker R, Neuberd G, Kropf S, Ulrich A & Tebbe CC (2007) Bacterial diversity of soils assessed by DGGE, T-RFLP and SSCP fingerprints of PCR-amplified 16S rRNA gene fragments: Do the different methods provide similar results? *J Microbiol Methods* **69**: 470–479.

Suominen L, Jussila MM, Mäkeläinen K, Romantschuk M & Lindström K (2000) Evaluation of the *Galega–Rhizobium galegae* system for the bioremediation of oil-contaminated soil. *Environ Pollut* **107**: 239–244.

Sutton JT, Robertson BC & Jamieson IG (2011) Dye shift: a neglected source of genotyping error in molecular ecology. *Mol Ecol Resour* **11**: 514–520.

Suzuki M, Rappe MS & Giovannoni SJ (1998) Kinetic bias in estimates of coastal picoplankton community structure obtained by measurements of small-subunit rRNA gene PCR amplicon length heterogeneity. *Appl Environ Microbiol* **64**: 4522–4529.

Thies JE (2007) Soil microbial community analysis using terminal restriction fragment length polymorphisms. *Soil Sci Soc Am J* **71**: 579–591.

Tiirola MA, Suvilampi JE, Kulomaa MS & Rintala JA (2003) Microbial diversity in a thermophilic aerobic biofilm process: analysis by length heterogeneity PCR (LH-PCR). *Water Res* **37**: 2259–2268.

Wallenius K, Rita H, Simpanen S, Mikkonen A & Niemi RM (2010) Sample storage for soil enzyme activity and bacterial community profiles. *J Microbiol Methods* **81**: 48–55.

Wallenius K, Rita H, Mikkonen A, Lappi K, Lindström K, Hartikainen K, Raateland A & Niemi RM (2011) Effects of land use on the level, variation and spatial structure of soil enzyme activities and bacterial communities. *Soil Biol Biochem* **43**: 1464–1473.

Yilmaz P, Kottmann R, Field D *et al.* (2011) Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat Biotechnol* **29**: 415–420.

Yu C-P & Chu K-H (2005) A quantitative assay for linking microbial community function and structure of a naphthalene-degrading microbial consortium. *Environ Sci Technol* **39**: 9611–9619.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1. The specific effects of vegetation and contamination on bacterial community succession during rhizoremediation process.

Fig. S2. Three-dimensional peak-based multidimensional scaling (MDS) ordination (area-sensitive Jaccard similarity) for the 180 LH-PCR profiles of Contaminated Vegetated treatment (green), Contaminated reference (blue) and Vegetated reference (yellow) from sampling weeks 0–21 (Arabic number next to the symbol).

Fig. S3. Changes in the bacterial community complexity in Contaminated Vegetated treatment and Vegetated and Contaminated references.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.