Heredity (2015) 114, 468–475

www.nature.com/hdy

ORIGINAL ARTICLE

Improving soil bacterial taxa–area relationships assessment using DNA meta-barcoding

S Terrat[1], S Dequiedt[1], W Horrigue[1], M Lelievre[1], C Cruaud[2], NPA Saby[3], C Jolivet[3], D Arrouays[3], P-A Maron[1,4], L Ranjard[1,4,6] and N Chemidlin Prévost-Bouré[5,6]

The evaluation of the taxa–area relationship (TAR) with molecular fingerprinting data demonstrated the spatial structuration of soil microorganisms and provided insights into the processes shaping their diversity. The increasing use of massive sequencing technologies in biodiversity investigations has now raised the question of the advantages of such technologies over the fingerprinting approach for elucidation of the determinism of soil microbial community assembly in broad-scale biogeographic studies. Our objectives in this study were to compare DNA fingerprinting and meta-barcoding approaches for evaluating soil bacterial TAR and the determinism of soil bacterial community assembly on a broad scale. This comparison was performed on 392 soil samples from four French geographic regions with different levels of environmental heterogeneity. Both molecular approaches demonstrated a TAR with a significant slope but, because of its more sensitive description of soil bacterial community richness, meta-barcoding provided significantly higher and more accurate estimates of turnover rates. Both approaches were useful in evidencing the processes shaping bacterial diversity variations on a broad scale. When different taxonomic resolutions were considered for meta-barcoding data, they significantly influenced the estimation of turnover rates but not the relative importance of each component process. Altogether, DNA meta-barcoding provides a more accurate evaluation of the TAR and may lead to re-examination of the processes shaping soil bacterial community assembly. This should provide new insights into soil microbial ecology in the context of sustainable use of soil resources.
Heredity (2015) **114,** 468–475; doi:10.1038/hdy.2014.91; published online 8 October 2014

INTRODUCTION

Soils are highly complex ecosystems and are considered as one of the Earth's main reservoirs of biological diversity. Bacteria account for a major part of this biodiversity, and it is now clear that such microorganisms have a key role in soil functioning processes (for example, control of nutrient cycles, and directly influence plant, animal or human health; Nemergut *et al.*, 2011). However, many of the environmental factors regulating the diversity of below-ground bacteria, still need to be investigated, which limits our understanding of the distribution of such bacteria at various spatial scales (Hanson *et al.*, 2012).

Until recently, most biogeographic studies have been devoted to plants and animals, providing insights into the ecological processes (dispersal, selection, ecological drift and speciation), which shape the community assembly and dynamics of macroorganisms (Nemergut *et al.*, 2011, 2013; Hanson *et al.*, 2012). For microorganisms, the first biogeographic hypothesis was developed by Baas Becking in 1934: 'Everything is everywhere, *but*, the environment selects', implying that microbes would be homogeneously distributed on a broad scale and among various environments. Interestingly, the number of microbial biogeography studies has increased exponentially over the last decade because of progress with molecular tools for routine application and broad-scale sampling networks involving several hundreds of samples (Maron *et al.*, 2011; Hanson *et al.*, 2012). These studies are providing overwhelming evidence that microorganisms display biogeographic patterns, but that much remains to be described and understood about the ecological processes contributing to these biological distributions as well as their relative importance (Hanson *et al.*, 2012; Ranjard *et al.*, 2013).

The oldest and most relevant way to discriminate the spatial processing of microbial diversification is to evaluate the taxa–area relationship (TAR). The first TAR was reported by Arrhenius (1921) as a power–law relationship between species richness (S_A) in an area A and local species richness (S_0) and area (A):

$$S_A = S_0 A^z \quad (1)$$

In this equation, z represents the rate at which new species are sampled as the sampling area is increased. This has been extended to microorganisms by taking equation (1) and deriving a similarity distance–decay relationship between community similarity between sites and geographic distance between sites (equation 2):

$$\chi_d = \chi_D \left(\frac{d}{D}\right)^{-2z} \quad (2)$$

In this equation, z is the same parameter as in equation (1) and is commonly considered as a turnover rate. χ_d and χ_D are the

[1]INRA, UMR1347 Agroécologie-Plateforme GenoSol, BP 86510, Dijon, France; [2]Commissariat à l'Energie Atomique (CEA), Institut de Génomique (IG), Genoscope, Evry, France; [3]INRA, US1106 InfoSol, Orléans, France; [4]INRA, UMR1347 Agroécologie, BP 86510, Dijon, France and [5]AgroSup Dijon, UMR1347 Agroécologie, BP 86510, Dijon, France
[6]Co-senior authors.
Correspondence: Dr N Chemidlin Prévost-Bouré, AgroSup Dijon, UMR1347 Agroécologie, BP 86510, Dijon 21000, France.
E-mail: n.chemidlin@agrosupdijon.fr
Received 31 March 2014; revised 25 August 2014; accepted 1 September 2014; published online 8 October 2014

community similarities between sites located d meters and D meters apart from each other. This derivation is based on two assumptions: that community size is infinite and that z is steady (Rosindell *et al.*, 2011). For microorganisms, the infinite community size hypothesis may hold because soil microbial communities are commonly assumed to be very large and diverse, and the average abundance per microbial taxa is high (Harte *et al.*, 2009). On the contrary, the hypothesis of z remaining steady across scales, which assumes self-similarity as a probability rule for the spatial distribution of taxa abundance across spatial scales (Harte *et al.*, 1999), may not hold for soil microbes. Consequently: (i) it may be assumed that the similarity distance–decay relationship is equivalent to the TAR and (ii) the z estimates and subsequent conclusions may vary across scales but can be assumed constant for a given scale.

Nevertheless, beyond the debate concerning the form of the TAR equation, this relationship is assumed to result mainly from: (i) the accumulation of species as the sampling area is increased because of the increased number of different habitats sampled (corresponding to the selection process); (ii) population dynamics, with greater possibilities for colonization and speciation but lower extinction rates in larger areas, corresponding to dispersal limitations and ecological drift processes; and (iii) speciation processes within the considered organisms (Hubbell, 2001; Zinger *et al.*, 2014). Challenging the widespread idea that microorganisms exhibit a cosmopolitan distribution, TAR is now commonly used in a majority of microbial biogeographical studies to assess microbial community turnover rate and its relative potential dependence on 'dispersal' and 'selection' (Angel *et al.*, 2010; Martiny *et al.*, 2011; Ranjard *et al.*, 2013; Wang *et al.*, 2013; Zinger *et al.*, 2014). The estimated turnover rates for microbial communities in most studies range from 0.002 to 0.26 (Horner-Devine *et al.*, 2004; Green and Bohannan, 2006; Woodcock *et al.*, 2006), and are generally much lower than those estimated for macroorganisms (classical range: 0.1–0.25; Horner-Devine *et al.*, 2004). In addition, Ranjard *et al.* (2013) have shown that selection and limited dispersal are not mutually exclusive and that a non-negligible proportion of bacterial community variation on a broad scale might be explained by the latter.

Although these studies demonstrated a significant spatial structuring of bacterial communities into biogeographical patterns, they were mainly based on molecular approaches with limited resolution, such as fingerprinting methods (Angel *et al.*, 2010; Ranjard *et al.*, 2013), or low-depth sequencing (Martiny *et al.*, 2011). Nowadays, high-throughput sequencing technologies (for example, 454 pyrosequencing or Illumina) are readily available to assess microbial diversity with greater precision and provide huge amounts of taxonomic information, based on hundreds of thousands of ribosomal RNA (rRNA) gene sequences (here designated DNA meta-barcoding) from a single metagenomic DNA (Maron *et al.*, 2011). Increasing use of these technologies in biodiversity investigations raised methodological and conceptual insights to ecologists (Wang *et al.*, 2013; Zinger *et al.*, 2014). Regarding biogeography studies, it has raised the question of the potential gain offered by the greater resolution of the DNA meta-barcoding approach, as compared with fingerprinting, in providing a deeper understanding of the determinism of microbial community assembly on a broad scale (Terrat *et al.*, 2012; Lienhard *et al.*, 2013). Recently, Van Dorst *et al.* (2014) incorporated various spatial scales and demonstrated the similar capacities of DNA fingerprinting and meta-barcoding to discriminate bacterial communities and to correlate with environmental variables at a local scale, but the greater resolution of DNA meta-barcoding at a global scale. This underlines the importance of adopting DNA meta-barcoding to support studies on broad to global scales, and to reexamine the processes involved in community assembly.

Our objectives in this study were to compare soil bacterial TAR and the determinism of soil bacterial community assembly on a broad scale using both approaches, namely DNA fingerprinting (Automated RISA fingerprinting, ARISA data set in the following) and DNA meta-barcoding (454 pyrosequencing, NGS data set in the following), to characterize soil bacterial diversity. Four geographic regions in the RMQS data set ('Réseau de Mesures de la Qualité des Sols' = French Monitoring Network for Soil Quality, covering 2200 soils over the whole of France using a systematic grid 16 km × 16 km) were selected along a gradient of environmental heterogeneity, representing a total of 392 soils. As the meta-barcoding approach can provide taxonomic information at different resolutions, multiple operational taxonomic unit (OTU) clustering thresholds (80 to 97%) were used. The similarity between two communities was determined with the Sørensen index based on the amount of shared OTUs, irrespective of their relationship (Green *et al.*, 2004). In each region, the soil bacterial community turnover rates (z) were estimated using the above-described similarity distance–decay relationship (equation 2). A distance-based redundancy analysis was used to partition bacterial community variance according to pedo-climatic characteristics, land-use and spatial variables. Our main hypothesis was that DNA meta-barcoding would provide a more robust estimation of TAR and a better understanding of the processes involved in bacterial community assembly on a broad scale than molecular fingerprinting.

MATERIALS AND METHODS

Sampling design

Soil samples were provided by the soil genetic resource conservatory of the GenoSol platform (http://www2.dijon.inra.fr/plateforme_genosol/) and obtained from the soil storage facility of the RMQS. The RMQS database consists of soil samples obtained from a regular 16-km grid across the 550 000 km^2 of metropolitan France and was designed to monitor soil properties (Arrouays *et al.*, 2002). The baseline survey comprises 2200 sites (each corresponding to a composite soil sample obtained from 25 soil cores) and was started in 2001 and completed in 2009. No temporal effect has been observed (data not shown). The 392 sites analyzed in this study were organized into four geographic regions: Brittany (124 sites), Burgundy (109 sites), Landes (52 sites) and South-East (107 sites) with contrasting soil type, land-use (coarse level of the CORINE Land Cover classification; IFEN, http://www.ifen.fr; 7 classes: forest, crop systems, grasslands, particular natural ecosystems, vineyards/orchards, parkland and wild land), climate (Quintana-Segui *et al.*, 2008) and geomorphology (Supplementary Table S1). The sites within a region were at least 16-km apart. For each soil, the following pedo-climatic characteristics were examined: particle-size distribution, pH in water (pH_{water}), C:N ratio, organic carbon (C_{org}), N, soluble P, $CaCO_3$ and exchangeable cation (Ca, K and Mg) contents, sum of annual temperatures (°C) and annual rainfall (mm). Physical and chemical analyses were performed by the Soil Analysis Laboratory of INRA (Arras, France), which is accredited for such analyses by the French Ministry of Agriculture.

DNA molecular fingerprinting data (fingerprint data set)

The subset of 392 soil samples was selected from the DNA fingerprinting data and methods (DNA extraction, purification, quantification and automated ribosomal intergenic spacer analysis), originally described and analyzed in the study by Ranjard *et al.* (2013). After automated ribosomal intergenic spacer analysis, contingency tables were derived from the fingerprints with samples in lines and bands (referred to as OTU_{bin} in the following) in columns with a maximum of 100 bands per sample to avoid taking into account artefactual bands because of image analysis.

DNA meta-barcoding data (NGS data set)

Soil DNA extraction, purification and quantification. Microbial DNA was freshly extracted from soils using a procedure optimized by the GenoSol platform named GnS-GII (Plassart *et al.*, 2012; Terrat *et al.*, 2012). The main difference between the GnS-GII and the DNA extraction procedure used in Ranjard *et al.* (2013) is the grinding step. However, both of these DNA extraction methods can provide a representative picture of the community with DNA molecular fingerprinting approaches (Plassart *et al.*, 2012), but not if high-throughput sequencing technologies, with their greater resolution, are used (Terrat *et al.*, 2012). µl Aliquots of crude DNA extracts were loaded onto polyvinylpolypyrrolidone microbiospin minicolumns (BIO-RAD Laboratories, Marnes-la-Coquette, France) and centrifuged for 4 min at 1000 *g* and 10 °C. Eluates were then collected and purified for residual impurities using the Geneclean Turbo kit (MP-Biomedicals, New-York, NY, USA). Purified DNA extracts were then quantified using the PicoGreen staining Kit (Molecular Probes, Paris, France).

Pyrosequencing of 16S rRNA genes. Microbial diversity (*bacteria* and *archaea*) was determined for each soil by 454 pyrosequencing of ribosomal genes. A 16S rRNA gene fragment targeting the complete hypervariable regions V4 (576–682) and V5 (822–879) (numbering based on the *Escherichia coli* system of nomenclature (Brosius *et al.*, 1978)) with an appropriate size (about 450 bases) for 454-pyrosequencing was amplified using the primers F479 (5′-CAGC MGCYGCNGTAANAC-3′) and R888 (5′-CCGYCAATTCMTTTRAGT-3′). Homemade bioinformatic programs have been developed to search large DNA sequence databases for the presence of primers, including degeneracies, as coded by the IUPAC rules, and also additional mismatches in order to test primer improvement. The sequences investigated were SILVA, and direct extraction of every small subunit rRNA sequence from EMBL using acnuc, and also a dedicated reference database of 18S eukaryotic sequences, which had been thoroughly analyzed and annotated (Supplementary Table S2) for *in silico* match analysis. For each sample, 5 ng of DNA were used for a 25 µl PCR conducted under the following conditions: 94 °C for 2 min, 35 cycles of 30 s at 94 °C (denaturation), 30 s at 52 °C (hybridization) and 1 min at 72 °C (elongation), followed by 7 min at 72 °C (final elongation). The PCR products were purified using a MinElute gel extraction kit (Qiagen, Courtaboeuf, France) and quantified using the PicoGreen staining Kit (Molecular Probes). A second PCR of nine cycles was then conducted under similar PCR conditions with 5 ng of purified PCR products and 10-base pair multiplex identifiers, designed and validated by ROCHE (http://www.liv.ac.uk/media/livacuk/centreforgenomicresearch/The_GS_FLX_Titanium_Chemistry_Extended_MID_Set.pdf) and added before the 5′ position of the primers, and after the 3′ positions of the adapters to specifically identify each sample and avoid PCR bias. Finally, the PCR products were purified and quantified as described above. Pyrosequencing was then carried out on a GS FLX Titanium (Roche 454 Sequencing System) at Genoscope (Evry, France).

Bioinformatic analysis of 16S rRNA gene sequences. Bioinformatic analyses were done using the GnS-PIPE initially developed by the GenoSol platform (INRA, Dijon, France; Terrat *et al.*, 2012), and recently optimized. The parameters chosen for each bioinformatic step can be found in Supplementary Table S3. First, all the 16S raw reads were sorted according to the multiplex identifier sequences. The raw reads were then preprocessed (filtered and deleted) based on: (a) their minimum length, (b) their number of ambiguities (Ns) and (c) their primer sequences. A PERL program was then applied for rigorous de-replication (that is, clustering of strictly identical sequences). The de-replicated reads were then aligned using Infernal alignments, and clustered into OTUs using a PERL program that groups rare reads with abundant ones, and does not count differences in homopolymer lengths (here, a cluster is defined by the most abundant read, known as the centroid, and every read in the cluster must have similarity above the given identity threshold with the centroid). A filtering step was then carried out to check all single-singletons (reads detected only once and not clustered, which might be artifacts, such as PCR chimeras) based on the quality of their taxonomic assignments. More precisely, each single-singleton was compared with a dedicated reference database from the Silva curated database using similarity approaches (USEARCH), with sequences longer than 500 nucleotides, and kept only if their identity was higher than the defined threshold (Supplementary Table S3). When several reference sequences were found (defined maximum of 10), a taxonomic consensus was derived, that is, a read was assigned to a given taxonomy only if a majority of similar reference sequences had the same description. Finally, in order to compare the data sets efficiently and avoid biased community comparisons, a homogenization step of kept reads per sample was carried out, to a value close to the lowest observed among samples (9410 reads), by random selection (Gihring *et al.*, 2012). The global analysis of soil samples was then computed by merging all homogenized high-quality reads from each sample into one global file before subsequent analyses. As the global analysis of bacterial community structure and diversity relies on the construction of similarity clusters (or OTUs) of 16S rRNA gene PCR amplicons (Horner-Devine *et al.*, 2004), we chose to use OTUs to examine the distribution of 16S rRNA gene sequences in our data set. However, there is no single best definition of 'species', 'genus', etc… when this sequencing approach is used, (because of controversy about thresholds of similarity allowing clear differentiation of taxonomic units), so we applied the following thresholds of sequence similarity: 80, 85, 90, 95 and 97% (Rosselló-Mora and Amann, 2001; Nemergut *et al.*, 2011). Such multiple OTU definitions are analogous to comparing different taxonomic resolutions. The retained high-quality reads were then used to determine the OTU composition of samples at each level of similarity. Finally, contingency tables of OTUs were obtained with samples in lines and OTUs in columns, indicating the number of reads in each OTU for all samples. OTUs were also taxonomically assigned using the information from high-quality reads (Supplementary Table S4). The raw data sets are available in the EBI database system under project accession number PRJEB6290.

Data set post-processing. Two filtering steps were applied to the NGS contingency tables of OTUs to eliminate potentially artefactual data (because of sequencing errors or PCR chimeras for example). The first step consisted of removing OTUs that occurred only once in the overall data set, considered as experimental artifacts. The second filtering step was designed to avoid up-weighting the importance of rare OTUs in the data set (as the contingency tables would be converted into binary tables for statistical analyses). This filtering step consisted of changing the value for the sample in the global OTU to 0 if two conditions were verified: (i) for each OTU in the global contingency table, if the reads of one sample represented $<1\%$ of the total abundance of the OTU and (ii) if the reads from the given OTU for the analyzed sample represented $<0.1\%$ of the total number of cleaned sequences identified in the sample. This second filtering step made it possible to remove less information than a single filter, which removed OTU representing $<0.1\%$ of the total number of cleaned sequences identified in the sample (Supplementary Table S5). The contingency tables of OTUs (with samples in lines and OTUs in columns) were then converted into binary tables for subsequent analyses. OTU richness was compared between the raw data set and the filtered data set to evaluate the effects of the filters.

Statistical analyses

Evaluation of congruence between ARISA and 454 sequencing data sets. In order to evaluate if ARISA and 454 sequencing data sets were comparable, the correlation between the distance matrices derived from each data set (Sørensen index) was tested by means of Mantel test (*mantel* function in *vegan* package in R).

Environmental heterogeneity. The level of environmental heterogeneity between regions was determined by applying a multivariate analysis on mixed data using the ade4 package in R (http://cran.r-project.org/web/packages/ade4/index.html) with soil pedo-climatic characteristics, land-use and geomorphological data (elevation). Quantitative data were centered and scaled and qualitative data were converted into weighted binary variables (weight equal to $1/n$; n being the number of classes for the qualitative variables). Differences between the four regions were examined by between-group analysis and a Monte-Carlo permutation test (1000 permutations). The environmental heterogeneity in each region was determined from the site dispersion on the factorial map.

Evaluation of bacterial community composition turnover rate. The turnover rates (z) for bacterial community composition were derived from the slope of the TAR as described in Ranjard *et al.* (2013) following the method described in Harte *et al.* (1999) and by applying equation (3), which is the

log-transformation of equation (2):

$$\log_{10}(\chi_d) = (-2 * z) * \log_{10}(d) + b \quad (3)$$

where χ_d is the observed Sørensen's similarity between two soil samples that are d meters apart from each other; b is the intercept of the linear relationship and z the turnover rate of the community composition. In this study, z is assumed to be constant, that is, independent of d. The z estimate and its 95% confidence interval were derived from the slope (-2^*z) of the relationship between similarity and distance by weighted linear regression. The overlap of the 95% confidence intervals was used to test for significant differences in turnover rates between regions or methods.

Variance partitioning of community assembly according to environmental filters and space. The relative importance of residual spatial autocorrelation, pedo-climatic characteristics and land-use in determining community composition turnover was tested by db-RDA (Legendre *et al.*, 2005; Legendre and Fortin, 2010). Quantitative data were centered and scaled. Residual spatial autocorrelation was considered by introducing spatial variables into the analysis. These spatial variables were constructed from site coordinates (x, y, elevation) to reveal potential spatial trends at scales larger than the region, and from principal coordinates of neighbour matrices eigenfunctions in each region (pcnm function in vegan package, http://cran.r-project.org/web/packages/vegan/index.html). Only principal coordinates of neighbour matrices with a significant Moran index ($P<0.001$) were selected. Land-use corresponded to the CORINE Land Cover classes (IFEN, http://www.ifen.fr) recoded into dummy variables. Pedo-climatic characteristics consisted of climate and all the physico–chemical variables except sand. The most parsimonious model was obtained by forward selection from null to full model. The marginal effects of each set of filters were tested with an analysis of variance (ANOVA)-like permutation test for canonical analyses (anova.cca function in vegan package, http://cran.r-project.org/web/packages/vegan/index.html).

RESULTS

Environmental heterogeneity

The four regions were selected for their contrasting environmental heterogeneity as demonstrated by principal component analysis of mixed data (Figure 1a). This multivariate analysis resulted in the discrimination of the four regions on both axes (Monte-Carlo permutation test, $P<0.001$). On the first axis, Landes was significantly discriminated from Brittany, Burgundy and South-East, and these three regions were discriminated from each other on the second axis. The environmental heterogeneity differed strongly between regions, the Landes sites being less dispersed on the factorial map than the Brittany or Burgundy sites, which were less dispersed than those of the South-East. Figure 1b shows that the four regions could mainly be distinguished according to land-use (for example, 86% of the Landes sites are forest sites), a restricted set of soil physico–chemical characteristics (sand and silt contents, pHwater and CaCO3 content, P content and C:N ratio) and by differences in elevation. Climatic conditions did not have a significant role in discrimination between regions.

Post-processing and taxonomic resolution on NGS data

Post-processing steps were applied to the NGS data set in order to account for potentially artefactual data. The effects of the filters were assessed by comparing OTU richness in the raw and filtered data sets. OTU richness at the 80% similarity threshold ranged from 20 to 171 OTUs in the raw data set and from 10 to 126 OTUs in the filtered data set. It increased with the increasing similarity threshold in the clustering analysis to reach a maximum at 97% sequence similarity. At this level of sequence similarity, OTU richness ranged from 106 to 4687 OTUs in the raw data set and from 44 to 1641 OTUs in the filtered data set (Table 1). The filtering steps reduced the number of OTUs considered by 34 to 48%. This reduction was similar for each region and harbored only a slight increase as the similarity threshold increased, ranging from 34 to 45% at the 80% sequence similarity threshold, and from 39 to 48% at the 97% sequence similarity threshold.

Regarding the raw data set, the four regions could be ranked according to the median OTU richness across all sites:

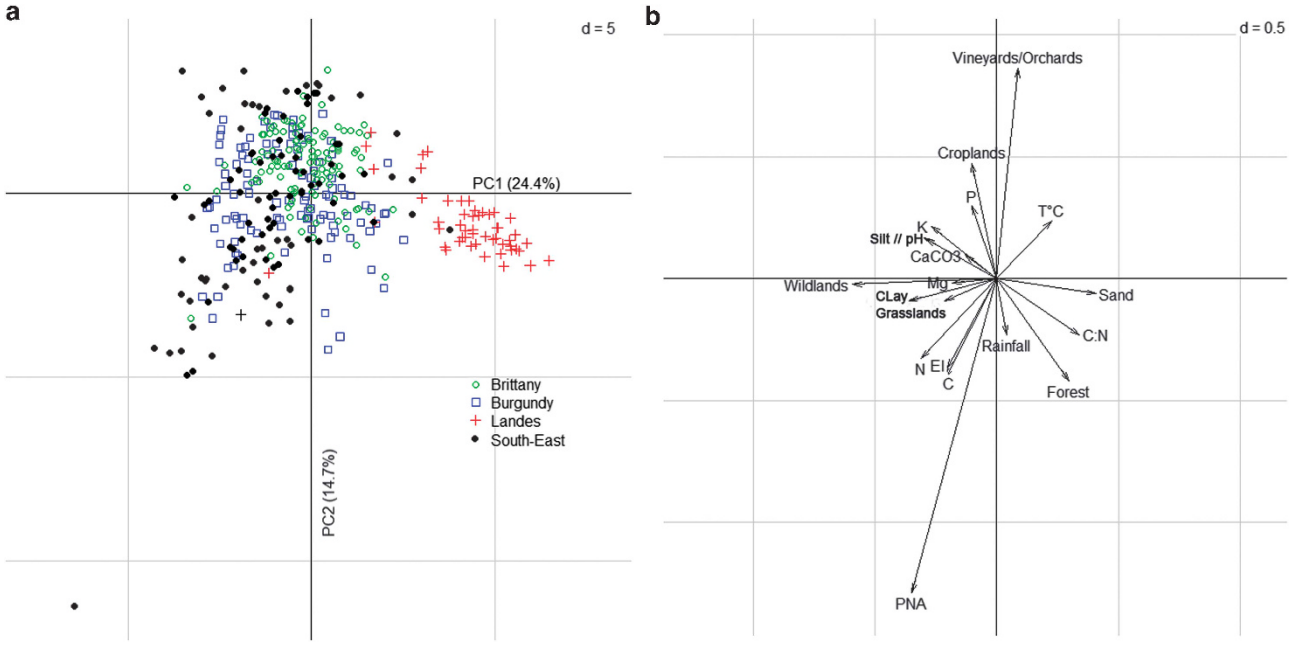

Figure 1 Environmental heterogeneity between the Brittany, Burgundy, Landes and South-East regions. (**a**) Factorial map representing the sites considered respectively to their region of origin in a principle component analysis on mixed data. Green open opened circles: Brittany; blue open opened squares: Burgundy; red crosses: Landes, black filled circles: South-East. (**b**) Correlation circle of the principle component analysis on mixed data. Monte-Carlo permutation test (1000 permutations) was significant ($P<0.001$).

Table 1 OTU numbers in the raw and filtered data sets

Clustering threshold	*Region* (n)	*OTU richness*						*Removed OTUs (%)*
		Raw data set			*Filtered data set*			
		Min	*Median (± s.d.)*	*Max*	*Min*	*Median (± s.d.)*	*Max*	
80%	Burgundy (109)	36	68 (±16.6)	115	20	39 (±11.5)	70	43
	Brittany (131)	45	92 (±17.9)	152	29	57 (±13.5)	125	38
	Landes (54)	43	60 (±19.3)	120	27	39 (±12.7)	90	34
	South-East (108)	20	86 (±21.0)	171	10	48 (±17.2)	126	45
85%	Burgundy (109)	83	156 (±37.4)	241	44	88 (±25.2)	153	44
	Brittany (131)	105	211 (±39.3)	360	52	123 (±29.9)	274	42
	Landes (54)	94	144 (±61.8)	435	62	90 (±43.8)	321	38
	South-East (108)	33	211 (±42.6)	377	15	120 (±33.6)	263	43
90%	Burgundy (109)	183	344 (±84.9)	577	91	181 (±55.0)	367	47
	Brittany (131)	217	488 (±99.9)	869	125	260 (±62.7)	478	47
	Landes (54)	201	343 (±188.5)	1361	136	223 (±133.8)	935	35
	South-East (108)	56	485 (±95.2)	835	22	277 (±71.6)	483	43
95%	Burgundy (109)	428	841 (±188.8)	1376	245	435 (±110.0)	828	48
	Brittany (131)	473	1171 (±227.0)	1862	316	628 (±127.2)	1106	46
	Landes (54)	475	875 (±496.1)	3423	356	570 (±247.4)	1622	35
	South-East (108)	93	1165 (±209.6)	1638	33	669 (±128.9)	997	43
97%	Burgundy (109)	669	1273 (±267.2)	2061	415	666 (±147.7)	1131	48
	Brittany (131)	758	1752 (±306.4)	2620	494	904 (±143.7)	1392	48
	Landes (54)	712	1360 (±691.1)	4687	541	832 (±249.4)	1641	39
	South-East (108)	106	1725 (±294.4)	2236	44	959 (±163.7)	1295	44

Abbreviation: OTU, operational taxonomic unit.
For the raw and the filtered data sets, the median number of OTUs identified per site, its minimum and its maximum were determined per region at each clustering threshold considered in this study. The s.d. of the mean is given in brackets. The percentage of OTUs removed by filtering steps was estimated by dividing the median number of OTUs in the filtered data set by the median number of OTUs in the raw data set.

Landes≈Burgundy<Brittany≈South-East ($P<0.05$, χ^2 test), whatever the clustering threshold. The same ranking of the different regions was obtained with the filtered data set, indicating that the overall trends in the diversity indices were not affected by the filtering steps adopted.

TAR evaluation

Mantel test comparisons between the distance matrices derived from the ARISA and 454 sequencing data sets (Sørensen index), highlighted significant correlations between these two data sets ($0.28<r<0.62$, $P<0.001$). These significant correlations showed that the two data sets were congruent and could be compared with one another (data not shown).

DNA fingerprinting (ARISA) and meta-barcoding were compared for their assessment of community turnover rate (z) using the Sørensen dissimilarity index on binary data. The estimated z ranged from 0.007 to 0.046, from 0.009 to 0.063, from 0.009 to 0.08 and from 0.013 to 0.09 in the Brittany, Burgundy, Landes and South-East regions, respectively (Figure 2). Except in Landes, the estimated z with the ARISA data set was always lower than the values obtained with the NGS data set, irrespective of the similarity thresholds used. In addition, z increased significantly with the sequence similarity threshold used for the clustering of OTUs, except between 95 and 97% similarity. Furthermore, the coefficients of variation of z were systematically smaller with the NGS data set than with the ARISA data set in every region and for every clustering level except in Landes at 80% similarity. Indeed, in Landes, the coefficients of variation for z ranged from 26 to 46% with NGS and were 40% with ARISA. For the three other regions, the coefficients of variation of z ranged from 8 to 12% with NGS and from 10 to 17% with ARISA. The same trend was observed when the four regions were compared with one another, except at the 85% similarity threshold where z was higher in the South-East than in Brittany, Landes and Burgundy (Figure 2). No significant differences were observed between these three regions. A similar trend was observed with the ARISA data set. Finally, no significant differences between regions were observed at a similarity threshold of 85%.

Variance partitioning of community assembly

The relative importance of the sets of spatial variables, land-use and pedo-climatic characteristics on variations in bacterial community assembly was used to compare the DNA meta-barcoding and fingerprinting approaches for their capacity to help understanding observed biological patterns.

The amount of variance in bacterial community assembly explained by spatial and environmental parameters ranged from 19.5% to 23.0%, 23.3% to 31.6%, 7.2% to 29.8% and 23.7% to 30.6% in Brittany, Burgundy, Landes and South-East, respectively, according to the molecular approach adopted (DNA fingerprinting vs DNA meta-barcoding) (Figure 3). In each region, slightly higher amounts of community variance were explained for the ARISA data set than for the NGS data set. In addition, similar amounts of community variance were explained between the different thresholds of sequence similarity used for OTUs clustering in the NGS approach with the Sørensen index (from 80 to 97% similarity levels, data not shown) in a given region.

All groups of explanatory variables (pedo-climatic characteristics, land-use or spatial variables) were selected with the ARISA and NGS data sets, independently of the Sørensen index (Figure 3). However, the land-use and spatial variables were not selected in Landes with NGS, whereas they were selected with ARISA. Comparison of variance partitioning of the ARISA and NGS data sets with the Sørensen index

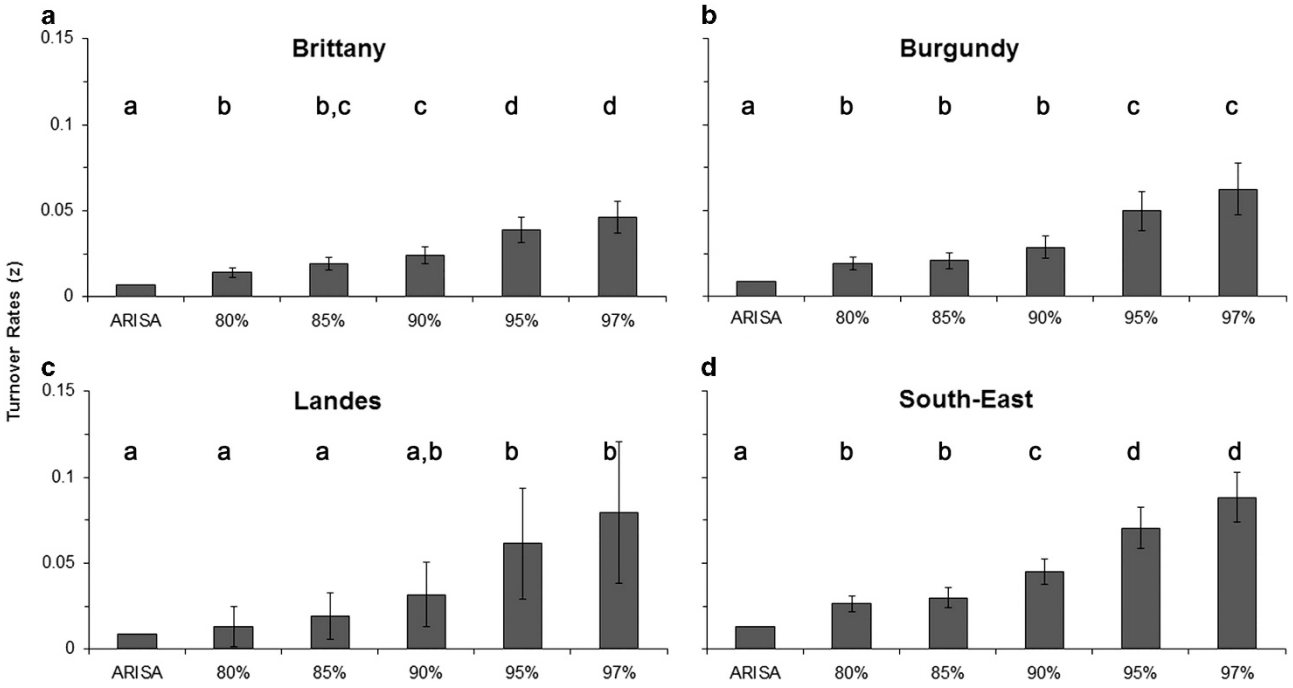

Figure 2 Comparison of ARISA and NGS approaches for the estimation of soil bacterial community turnover rate. Comparisons of each approach were performed within each region: (**a**) Brittany; (**b**) Burgundy; (**c**) Landes; (**d**) South-East. Percentages indicate the level of similarity considered in the NGS approach. Letters indicate significant differences between turnover rates in each region at the 5% probability level.

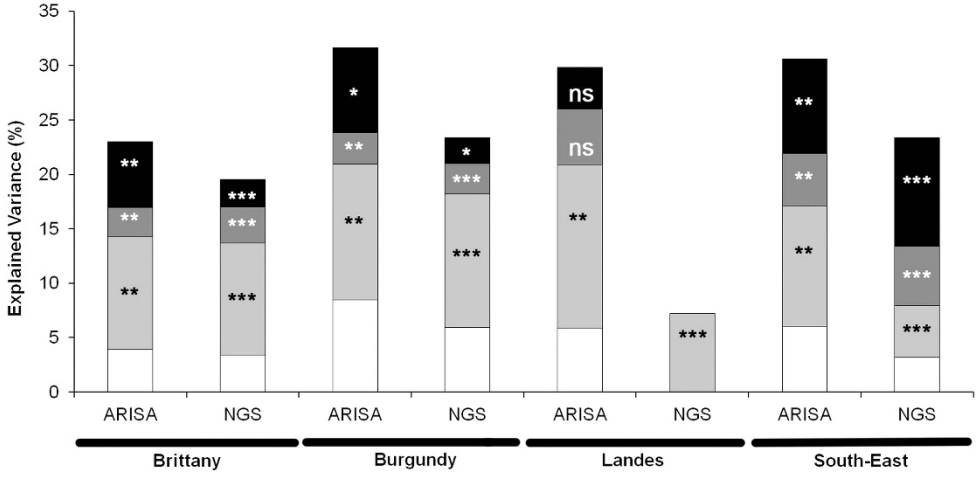

Figure 3 Variance partitioning of community composition with ARISA and NGS approaches in Brittany, Burgundy, Landes and South-East. Four groups of variables were considered: pedo-climatic characteristics (light grey), land-use (dark grey), spatial descriptors (black), which stand for residual spatial autocorrelation and interactions between the three sets of filters (white). Interactions between the three groups of variables were estimated but could not be tested for their significance. Percentages indicate the level of similarity considered in the NGS approach. Significance codes: ns: $P>0.05$; *: $P<0.05$; **: $P<0.01$; ***: $P<0.001$.

revealed slightly different results for some regions when groups of explanatory variables were ranked according to their marginal effect. In the Brittany, Burgundy and Landes regions, pedo-climatic characteristics explained the highest amount of variance (10%, 12.5% and 7–15%, respectively) whatever the molecular approach or dissimilarity index whereas spatial variables accounted for a systematically higher variance with the ARISA data set than with the NGS approach (Figure 3). For the South-East region, the use of ARISA or NGS data sets led to different hierarchies in the groups of explanatory variables. With the ARISA data set, pedo-climatic characteristics explained the highest amount of variance (11%), followed by spatial variables (8.7%) and land-use (4.8%). With the NGS data set, spatial variables were the most important (9.2%), followed by land-use and pedo-climatic characteristics (5.5% and 4.8%, respectively).

DISCUSSION

The four regions considered in this study followed a gradient of environmental heterogeneity: low level (Landes region), medium level (Brittany and Burgundy) and high level (South-East) as observed by means of the dispersion of sites according to regions in the multivariate analysis on mixed data. Most of these differences were related to variability of environmental parameters reported in the literature to be involved in shaping soil microbial diversity: land-use (Drenovsky *et al.*, 2010), soil characteristics (texture, pH, P content and C:N ratio;

Fierer and Jackson, 2006; Dequiedt *et al.*, 2011; Lienhard *et al.*, 2013) and to a lesser extent climate (Fierer and Jackson, 2006). This environmental heterogeneity had been shown to affect the soil bacterial community turnover rate (*z*, slope of the TAR in the similarity distance–decay relationship, Harte *et al.*, 2009; Ranjard *et al.*, 2013). These four regions therefore provided a valid sampling design for determining whether DNA meta-barcoding, as compared with molecular fingerprinting, would provide a more accurate estimation of TAR and a better understanding of the processes involved in bacterial community assembly on a broad scale.

A preliminary step in the comparison of DNA meta-barcoding and fingerprinting approaches was to post-process the data according to molecular analysis steps specific to each method. For DNA fingerprinting, this was handled by setting a fixed number of OTU_{bin} to be considered during the band profiles analysis (Ranjard *et al.*, 2001, 2013). For the NGS approach, methodological biases (for example, PCRs or sequencing errors), which might generate OTUs of low abundance and equally represented across samples, were removed by bioinformatic filters (Quince *et al.*, 2011) and by two post-processing steps. These steps allowed the preservation of information on the 'rare biosphere' while removing artifacts (Kunin *et al.*, 2010; Terrat *et al.*, 2012; Supplementary Table S5), conversely to the classical post-processing step (removal of all low abundant OTUs). As the data were analyzed as presence–absence data, these steps seemed a relevant consensus to avoid up-weighting the importance of rare and specific OTUs in the data set (Van Dorst *et al.*, 2014; Zinger *et al.*, 2014), while conserving a fine description of bacterial community assembly (Supplementary Table S3). These post-processing steps led to the conservation of at least 6000 sequences per sample, which was higher than those used in a recent study (*ca* 4000 reads per sample, Zinger *et al.*, 2014). This study demonstrated that community *z* was weakly affected by the sequencing depth per sample unless it was shallow (<500 sequences), and that it was independent of the number of reads between samples (Zinger *et al.*, 2014).

In this context, DNA fingerprint and meta-barcoding approaches were compared for their estimation of soil bacterial community *z*. Both approaches associated with the Sørensen index demonstrated significant *z* estimates, which were in accordance with those classically observed in the literature with fingerprinting or low-depth sequencing data (Green *et al.*, 2004, 2006; Horner-Devine *et al.*, 2004; Bell *et al.*, 2005; Woodcock *et al.*, 2006; Martiny *et al.*, 2011; Ranjard *et al.*, 2013; Zinger *et al.*, 2014). Interestingly, the *z* estimates observed with the DNA meta-barcoding approach were higher than those obtained with ARISA, although the OTU_{bin} richness in the ARISA data set was similar to the OTU richness in the NGS data set for low clustering thresholds (from 80 to 90% similarity). Moreover, the coefficients of variation of *z* estimated with the DNA meta-barcoding approach were systematically smaller than with the ARISA, thereby highlighting that the *z* estimates obtained by DNA meta-barcoding approach are more accurate. These differences in *z* estimates and coefficients of variation can easily be explained: ARISA involves the analysis of the length polymorphism of the intergenic spacer between the 16S and 23S ribosomal genes (IGS) that can be considered less informative than DNA meta-barcoding, which assesses the sequence (size and nucleotide composition) (Ranjard *et al.*, 2000; Terrat *et al.*, 2012). This accuracy of the NGS approach in describing community assembly may reduce the similarity between sites, leading to higher estimates of community turnover. From an ecological viewpoint, DNA fingerprint and meta-barcoding approaches displayed a similar capacity to discriminate samples and both demonstrated the significant spatial structuration of soil bacterial communities in the different regions considered. In addition, both methods revealed similar trends between the regions for the hierarchy of soil bacterial community *z*: Brittany $<$ (Burgundy, Landes) $<$ South-East, in agreement with the study by Van Dorst *et al.* (2014), which demonstrated the similar capacity of ARISA and DNA meta-barcoding to discriminate sites at a local spatial scale. This underlines the value of both methods in demonstrating ecological trends for soil bacterial communities. Nevertheless, the DNA meta-barcoding approach provides a finer description of soil bacterial community composition than ARISA, which supports the hypothesis that the turnover rates estimated derived from the latter approach would be underestimated. One advantage of the DNA meta-barcoding approach is that a description of community assembly can be drawn from the construction of similarity clusters (OTUs) at various thresholds of sequence similarity. Here, soil bacterial community *z* increased significantly with increasing clustering thresholds (80 to 97%). This increase was mainly related to an increase of OTU richness from low to high clustering thresholds, in agreement with Harte *et al.* (2009). Interestingly, the values of *z* obtained for high clustering thresholds were comparable to those observed for macroorganisms (MacArthur and Wilson, 2001; Horner-Devine *et al.*, 2004), high thresholds that are considered to reflect taxonomic levels classically used in macroecological studies (Rosselló-Mora and Amann, 2001). These findings raise insights for microbial ecologists studying the spatial structuring of soil microbial communities as they contradict the classically observed positive relationship between community turnover rate (*z*) and organism body size (Hillebrand *et al.*, 2001; Drakare *et al.*, 2006), and also suggest that soil microbial communities may display strong spatial structuration.

Meta-barcoding and fingerprint approaches were compared for their ability to correlate pedo-climatic, land-use and spatial variables with soil bacterial community distributions in a distance-based redundancy analysis (Sørensen index). These environmental sets of variables captured significant amounts of community variance for both data sets (8 to 35%), which were within the range of those reported in the literature for bacteria (Martiny *et al.*, 2011; Hanson *et al.*, 2012). Surprisingly, slightly higher amounts of community variance were explained for ARISA data. Van Dorst *et al.* (2014), in contrast, had reported the opposite trend. This might be related to differences in the spatial scales investigated (broad vs local scale) and in data processing (OTU_{bin} definition and bioinformatics steps for OTUs clustering).

The variance partitioning approach also allowed the comparison of the relative importance of pedo-climatic, land-use and spatial variables in shaping bacterial community assembly according to the molecular approach. The same hierarchy was observed for the different sets of variables in Burgundy and Brittany, but not in the Landes and South-East regions. Indeed, in the South-East region, spatial variables became the main driver of community assembly according to the DNA meta-barcoding approach. This result suggests that dispersal limitations may be as important as selection in shaping bacterial community assembly (Martiny *et al.*, 2011) because this process could lead to a spatial autocorrelation of bacterial communities between sites. This hypothesis was supported by Bryant *et al.* (2008) who highlighted the primary importance of elevation in limiting the dispersal of *Acidobacteria*. In the South-East region, elevation was the most important spatial variable explaining bacterial community assembly. Altogether, this suggests that use of the DNA meta-barcoding approach could lead to reexamination of the relative importance of the processes shaping soil microbial diversity on a broad scale.

CONCLUSION

Although DNA fingerprinting and meta-barcoding are both relevant to demonstrate the spatial structuration of soil microbial communities through significant TAR, the DNA meta-barcoding approach provides a finer description of soil bacterial community assembly. It also provides a more accurate estimation of community turnover rates. Considering the processes shaping soil bacterial diversity, both identical conclusions were not systematically obtained, suggesting that DNA meta-barcoding approach may lead to reexamine their relative importance. Nevertheless, this should be tested for other soil microbial communities like fungi. In addition, in a context of up-scaling studies in microbial biogeography, the meta-barcoding approach may help to identify not only the scales at which soil microbial communities are structured, but also the processes or the filters shaping their diversity at each spatial scale.

Data archiving

The raw data sets are available on the EBI database system under project accession number PRJEB6290.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This study was granted by the French National Research Agency (ANR) and the Genoscope (Evry, France). RMQS soil sampling and physico–chemical analyses were supported by a French Scientific Group of Interest on soils: the 'GIS Sol', involving the French Ministry for Ecology and Sustainable Development (MEDAD), the French Ministry of Agriculture (MAP), the French Institute for Environment (IFEN), the French Agency for Energy and Environment (ADEME), the French Institute for Research and Development (IRD), the French National Geographic Institute (IGN) and the National Institute for Agronomic Research (INRA). We thank all the soil surveyors and technical assistants involved in sampling the sites. This work, through the involvement of technical facilities at the GenoSol platform of the infrastructure ANAEE-France, received a grant from the French state via the National Agency for Research under the program 'Investments for the Future' (reference ANR-11-INBS-0001), as well as a grant from the Regional Council of Burgundy.

Angel R, Soares MIM, Ungar ED, Gillor O (2010). Biogeography of soil archaea and bacteria along a steep precipitation gradient. *ISME J* **4**: 553–563.

Arrhenius O (1921). Species and area. *J Ecol* **9**: 95–99.

Arrouays D, Jolivet C, Boulonne L, Bodineau G, Saby NPA, Grolleau E (2002). A new projection in France: a multi-institutional soil quality monitoring network. *Comptes Rendus l'Academie d'Agriculture Fr* **88**: 93–105.

Baas Becking LGM (1934). *Geobiologie of Inleiding tot de Milieukunde*. The Hague, The Netherlands.

Bell T, Ager D, Song J-I, Newman Ja, Thompson IP, Lilley AK *et al.* (2005). Larger islands house more bacterial taxa. *Science* **308**: 1884.

Brosius J, Palmer ML, Kennedy PJ, Noller HF (1978). Complete nucleotide sequence of a 16S ribosomal RNA gene from *Escherichia coli*. *Proc Natl Acad Sci USA* **75**: 4801–4805.

Bryant J, Lamanna C, Morlon H, Kerkhoff AJ, Enquist BJ, Green JL (2008). Colloquium paper: microbes on mountainsides: contrasting elevational patterns of bacterial and plant diversity. *Proc Natl Acad Sci USA* **105**: 11505–11511.

Dequiedt S, Saby NPA, Lelievre M, Jolivet C, Thioulouse J, Toutain B *et al.* (2011). Biogeographical patterns of soil molecular microbial biomass as influenced by soil characteristics and management. *Glob Ecol Biogeogr* **20**: 641–652.

Drakare S, Lennon JJ, Hillebrand H (2006). The imprint of the geographical, evolutionary and ecological context on species-area relationships. *Ecol Lett* **9**: 215–227.

Drenovsky RE, Steenwerth KL, Jackson LE, Scow KM (2010). Land use and climatic factors structure regional patterns in soil microbial communities. *Glob Ecol Biogeogr* **19**: 27–39.

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

Gihring TM, Green SJ, Schadt CW (2012). Massively parallel rRNA gene sequencing exacerbates the potential for biased community diversity comparisons due to variable library sizes. *Environ Microbiol* **14**: 285–290.

Green J, Bohannan BJM (2006). Spatial scaling of microbial biodiversity. *Trends Ecol Evol* **21**: 501–507.

Green JL, Holmes AJ, Westoby M, Oliver I, Briscoe D, Dangerfield M *et al.* (2004). Spatial scaling of microbial eukaryote diversity. *Nature* **432**: 747–750.

Hanson C, Fuhrman J, Horner-Devine MC, Martiny JBH (2012). Beyond biogeographic patterns: processes shaping the microbial landscape. *Nat Rev Microbiol* **10**: 497–506.

Harte J, Kinzig A, Green J (1999). Self-similarity in the distribution and abundance of species. *Science* **284**: 334–336.

Harte J, Smith AB, Storch D (2009). Biodiversity scales from plots to biomes with a universal species-area curve. *Ecol Lett* **12**: 789–797.

Hillebrand H, Watermann F, Karez R, Berninger U-G (2001). Differences in species richness patterns between unicellular and multicellular organisms. *Oecologia* **126**: 114–124.

Horner-Devine MC, Lage M, Hughes JB, Bohannan BJM (2004). A taxa-area relationship for bacteria. *Nature* **432**: 750–753.

Hubbell S (2001). *A Unified Neutral Theory of Biodiversity and Biogeography*. Princeton University Press: Princeton, NJ.

Kunin V, Engelbrektson A, Ochman H, Hugenholtz P (2010). Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ Microbiol* **12**: 118–123.

Legendre P, Borcard D, Peres-Neto PR (2005). Analyzing beta diversity: partitioning the spatial variation of community composition data. *Ecol Monogr* **75**: 435–450.

Legendre P, Fortin M-J (2010). Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. *Mol Ecol Resour* **10**: 831–844.

Lienhard P, Terrat S, Prévost-Bouré N, Nowak V, Régnier T, Sayphoummie S *et al.* (2013). Pyrosequencing evidences the impact of cropping on soil bacterial and fungal diversity in Laos tropical grassland. *Agron Sustain Dev* **34**: 525–533.

MacArthur RH, Wilson EO (1967). *The Theory of Island Biogeography*. Princeton Univ. Press, Princeton.

Maron P-A, Mougel C, Ranjard L (2011). Soil microbial diversity: methodological strategy, spatial overview and functional interest. *C R Biol* **334**: 403–411.

Martiny JBH, Eisen JA, Penn K, Allison SD, Horner-devine MC (2011). Drivers of bacterial β -diversity depend on spatial scale. *Proc Natl Acad Sci USA* **108**: 7850–7854.

Nemergut DR, Costello EK, Hamady M, Lozupone C, Jiang L, Schmidt SK *et al.* (2011). Global patterns in the biogeography of bacterial taxa. *Environ Microbiol* **13**: 135–144.

Nemergut DR, Schmidt SK, Fukami T, O'Neill SP, Bilinski TM, Stanish LF *et al.* (2013). Patterns and processes of microbial community assembly. *Microbiol Mol Biol Rev* **77**: 342–356.

Plassart P, Terrat S, Thomson B, Griffiths R, Dequiedt S, Lelievre M *et al.* (2012). Evaluation of the ISO standard 11063 DNA extraction procedure for assessing soil microbial abundance and community structure. *PLoS ONE* **7**: e44279.

Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ (2011). Removing noise from pyrosequenced amplicons. *BMC Bioinform* **12**: 38.

Quintana-Segui P, Le Moigne P, Durand Y, Martin E, Habets F, Baillon M *et al.* (2008). Analysis of Near-Surface Atmospheric Variables: Validation of the SAFRAN Analysis over France. *J App Meteorol* **47**: 92–107.

Ranjard L, Brothier E, Nazaret S (2000). Sequencing bands of ribosomal intergenic spacer analysis fingerprints for characterization and microscale distribution of soil bacterium populations responding to mercury spiking ch. *Appl Environ Microbiol* **66**: 5334–5339.

Ranjard L, Dequiedt S, Chemidlin Prévost-Bouré N, Thioulouse J, Saby NPA, Lelievre M *et al.* (2013). Turnover of soil bacterial diversity driven by wide-scale environmental heterogeneity. *Nat Commun* **4**: 1434.

Ranjard L, Poly F, Lata JC, Mougel C, Thioulouse J, Nazaret S (2001). Characterization of bacterial and fungal soil communities by automated ribosomal intergenic spacer analysis fingerprints: biological and methodological variability. *Appl Environ Microbiol* **67**: 4479–4487.

Rosindell J, Hubbell SP, Etienne RS (2011). The unified neutral theory of biodiversity and biogeography at age ten. *Trends Ecol Evol* **26**: 340–348.

Rosselló-Mora R, Amann R (2001). The species concept for prokaryotes. *FEMS Microbiol Rev* **25**: 39–67.

Terrat S, Christen R, Dequiedt S, Lelievre M, Nowak V, Regnier T *et al.* (2012). Molecular biomass and MetaTaxogenomic assessment of soil microbial communities as influenced by soil DNA extraction procedure. *Microb Biotechnol* **5**: 135–141.

Van Dorst J, Bissett A, Palmer AS, Brown M, Snape I, Stark JS *et al.* (2014). Community fingerprinting in a sequencing world. *FEMS Microbiol Ecol* **89**: 1–15.

Wang J, Shen J, Wu Y, Tu C, Soininen J, Stegen JC *et al.* (2013). Phylogenetic beta diversity in bacterial assemblages across ecosystems: deterministic versus stochastic processes. *ISME J* **7**: 1310–1321.

Woodcock S, Curtis TP, Head IM, Lunn M, Sloan WT (2006). Taxa-area relationships for microbes: the unsampled and the unseen. *Ecol Lett* **9**: 805–812.

Zinger L, Boetius A, Ramette A (2014). Bacterial taxa-area and distance-decay relationships in marine environments. *Mol Ecol* **23**: 954–964.

Supplementary Information accompanies this paper on Heredity website (http://www.nature.com/hdy)