Harmonized Datasets of microbiological parameters from a French national-scale soil monitoring survey

Aurélien Cottin, Samuel Dequiedt, Christophe Djemiel, Nicolas Chemidlin Prévost-Bouré, Julie Tripied, Mélanie Lelièvre, Lucie Terreau, Tiffanie Régnier, Battle Karimi, Claudy Jolivet, et al.

scientific data

OPEN

DATA DESCRIPTOR

Harmonized Datasets of microbiological parameters from a French national-scale soil monitoring survey

Aurélien Cottin[1,3], Samuel Dequiedt[1], Christophe Djemiel[1], Nicolas Chemidlin Prévost-Bouré[1], Julie Tripied[1], Mélanie Lelièvre[1], Lucie Terreau[1], Tiffanie Régnier[1], Battle Karimi[1], Claudy Jolivet[2], Antonio Bispo[2], Nicolas Saby[2], Pierre-Alain Maron[1], Lionel Ranjard[1] & Sébastien Terrat[1,3] ✉

Microbiological datasets and associated environmental parameters from the French soil quality monitoring network (RMQS) offer an opportunity for long-term and large-scale soil quality monitoring. Soils supply important ecosystem services e.g. carbon dynamics/storage or mineral element recycling, supported by the soil microbial diversity (bacteria, archaea and fungi). Based on the 2,240 sites of the 2000–2015 RMQS, molecular tools were applied to characterize soil microbiota. Soil DNA analysis yielded molecular microbial biomass for 2,168 sites, bacterial and fungal qPCR for 2,073 sites, and high-throughput amplicon sequencing of targeted 16S rDNA bacterial and archaeal genes for 1,842 sites. All these datasets were partially or completely unavailable, so raw results files from RMQS microbiological studies were harmonized and published in a Dataverse repository to facilitate their reusability. Altogether, these datasets allow for in-depth studies of soil microbial ecology and biogeography, and will be updated with fungal datasets and the second currently ongoing monitoring campaign (2016–2027).

Background & Summary

Soils support many ecosystem services such as fertility, carbon storage, waste decomposition, pest and pathogen control, or water retention[1]. All these services are mainly supported by the huge reservoir of soil microbial biodiversity encompassing diverse taxa (bacteria, archaea or fungi). However, soil microbiota is constantly subjected to various natural or anthropogenic stresses associated to deforestation, land-use intensification and global warming[2–4]. These disturbances have a significant influence on these soil microbial communities and lead to an overall impact on soil functions. For an efficient conservation of soils, it is essential to be able to detect the emergence and trends of these changes at an early stage. For the past 15 years, the "Réseau de Mesures de la Qualité des Sols" (RMQS - French Soil Quality Monitoring Network) has been meeting these goals of long-term assessment and monitoring of soil quality in France[5,6].

The RMQS is based on the monitoring of 2,240 sites distributed across the whole French territory along a systematic square grid of 16 km × 16 km cells, to be representative of the different types of soils and their land uses[5]. Soil sampling, characterization and observations are made every 15 years at the center of each cell. The RMQS is probably one of the most intensive and extensive sampling strategy at a national scale in Europe[7]. The first sampling campaign took place from 2000 to 2009, while the second campaign is currently ongoing (2016 to 2027). The first campaign focused on the assessment of soil contamination and made it possible to map key soil parameters (28 variables such as pH, carbon organic content or texture) as well as several trace metal elements and persistent organic pollutants[6] https://www.gissol.fr/.

[1]Agroécologie, INRAE, Institut Agro, Univ. Bourgogne, Univ. Bourgogne Franche-Comté, F-21000, Dijon, France. [2]INRAE, InfoSol, F-45075, Orléans, France. [3]These authors contributed equally: Aurélien Cottin, Sébastien Terrat. ✉e-mail: sebastien.terrat@inrae.fr

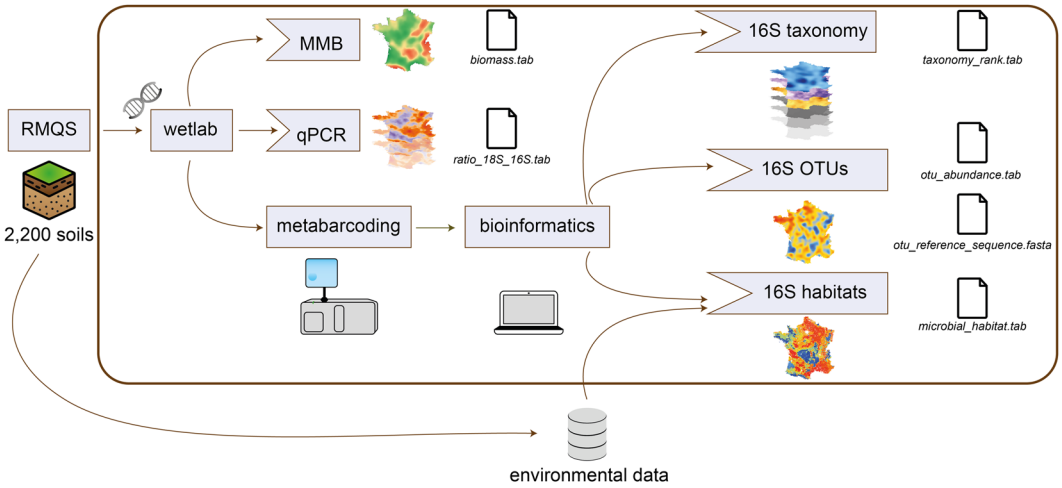

Fig. 1 Schematic overview of the RMQS1 microbial datasets.

Thanks to the use of various molecular tools, a substantial body of scientific knowledge has been produced on the RMQS soil microbiota (Fig. 1)[5,8–11]. Interestingly, each produced dataset (*e.g.* measures of bacterial 16S and fungal 18S gene abundances, the F:B ratio, or the diversity data) exhibited different and specific biogeographical patterns compared to each other taken separately, reflecting each one a complementary snapshot of soil microbial communities[11]. Moreover, several technical developments have been designed to standardize each applied method and the whole process[12–15]. Articles on the microbiological parameters of the RMQS samples were published since 2011, but the associated datasets were not necessarily published with their respective papers. To improve the reuse of these datasets, we republished and reorganized all available microbiological (i.e. molecular microbial biomass[16], fungal:bacterial ratio and gene abundance[17], bacterial and archaeal taxonomic characterization[18], habitat definition[19] and OTU matrix[20]) in a dataverse collection, with a focus on linking datasets with other RMQS environmental parameters more easily. The cleaned sequences itselves have been also deposited in a dedicated repository in the collection[21]. Some of the datasets (e.g. molecular microbial biomass, bacterial diversity using the newly published ReclustOR approach[14], or taxonomy obtained against a more recent version of the SILVA database) were also updated when necessary. Furthermore, this dataverse collection will be enriched with current and future analyses (e.g. fungal datasets, second sampling campaign). This dataset collection will provide information on the ecology of microbial communities at a territorial scale and describe the different microbial groups observed in French soils, their spatial distribution, their ecological requirements and their interactions. Altogether, this collection will help researchers to have a better understanding of soil microbial communities organization and dynamics across space.

Methods

French soil quality monitoring network. The detailed information given in the next paragraphs (French Soil Quality Monitoring Network and Soil Sampling) was already presented[8,22] and is being included in this work for ease. The French Soil Quality Monitoring Network (RMQS) was initially established to provide a national framework for observing changes in soil quality across France[22]. This network is part of a program of the soil scientific interest group "GIS Soil" and brings together representatives of the French ministries in charge of agriculture and the environment, the French Agency of the ecological transition (ADEME), the French National Research Institute for Agriculture, Food and Environment (INRAE), the French National Institute for Sustainable Development (IRD), and the French National Institute of Geographic and Forest Information (IGN)[6].

The survey consisted of 2,240 monitored sites located at the central nodes of 16-km grids that covered the French territory (5.5×10^5 km^2, 2,170 sites in mainland France and 70 sites in overseas territories)[6]. Each site was positioned with a precision of less than 0.5 m, and the soil profile, the local environment, climatic factors, and land cover were described.

Soil sampling. For each monitored site, 25 individual core samples were sampled from the topsoil (0–30 cm) following an unaligned sampling design within a 20 × 20 m area. The samples were pooled to obtain composite samples representative of each site that were air-dried under controlled temperature (30 °C) and humidity conditions, sieved to 2 mm and stored at −40 °C before analysis[8]. Many physicochemical parameters were measured for each soil, i.e. particle-size distribution, pH, organic carbon (C) content, nitrogen (N) content, the C/N ratio, the soluble phosphorus (P) content, the calcareous content, the cation exchange capacity (CEC) and exchangeable cations (Ca, Mg), for a total of 28 parameters[23]. Twelve trace metals and 70 persistent organic pollutants were also evaluated. They were performed by the Soil Analysis Laboratory of INRAE (Arras, France, https://las.hautsdefrance.hub.inrae.fr/).

Climatic data were obtained by interpolating observational data using the SAFRAN model[24]. Available climatic data were monthly rain, evapotranspiration and temperature at each node of a 12-km grid, averaged for the 1992–2004 period. Then, the RMQS site-specific data were linked to the climatic data by finding the closest node to each RMQS site within the 12-km climatic grid.

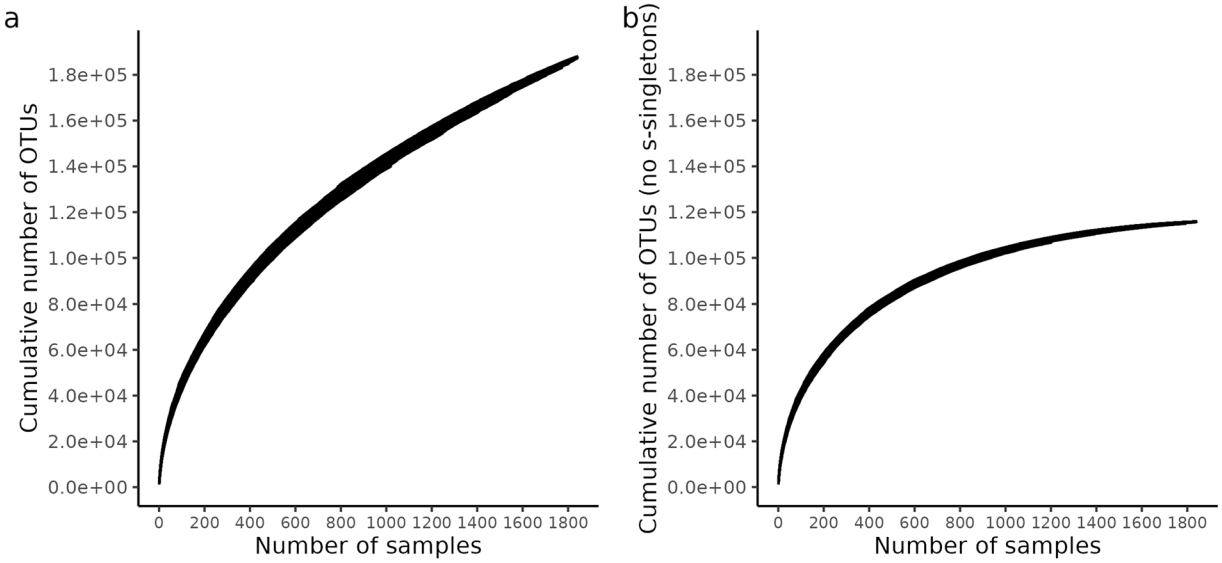

Fig. 2 Cumulative curves of 16S OTUs according to the number of samples using ReClustOR. (**a**) All OTUs considered for computation; (**b**) Computation after single-singleton deletion. One thousand cumulative curves were drawn with a random selection of soils. Based on Supplemental Fig. 2 from[9].

Land cover was recorded according to the CORINE land cover classification (https://land.copernicus.eu/en/technical-library/clc-2018-technical-guidelines/@@download/file). Both coarse and refined levels of land-cover classification were used.

Physical, chemical analyses and all other soil data (climatic and land cover) are available in a dataset collection in the French Public-research dataverse (https://entrepot.recherche.data.gouv.fr/) previously cited[23].

Datasets harmonization. The detailed biological information given in this subsection was already presented[5,8–11] and is being included in this work for ease. Each dataset was deposited in a global dataset collection in the French Public-research dataverse (https://entrepot.recherche.data.gouv.fr/), with a short descriptive name and linked to the original publication, to facilitate their harmonization. Each dataset, already published or not, was checked and updated if necessary (see below for details), and the site identifier was used as a key to link all measures across all datasets. Missing values were removed. Moreover, previously unpublished control data for microbial gene abundance and microbial community sequencing data were added in the collection.

Microbial DNA was extracted from one gram of each of the 2,240 composite soils sampled in each RMQS site, using the GnS-GII procedure, described previously[25]. Molecular microbial biomass values[16] were updated by extracting anew each data point from the Genosol platform database (https://www2.dijon.inrae.fr/plateforme_genosol/en), by selecting the latest measured value between 2012 and 2018 for each site, for a total of 2,168 values. Values lower than 2.5 $\mu g.g^{-1}$ DNA were considered out of range, but kept in the dataset (see[8,26] for details).

Microbial (bacterial and archaeal) habitat[5] were simply reformatted[19]. As microbial gene abundance was published recently[11], published dataset[17] correspond to the most recent version. The F:B ratio given in the file was computed with this formula: (number of 18S fungal copies) × 100 / (number of 16S bacterial copies). So the F:B given ratio can be summarized as the number of fungal cells for 100 bacterial cells.

OTUs clustering and taxonomy results were previously published[9,10] but were not made directly available. However, raw sequencing data were deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEB21351 (https://www.ebi.ac.uk/ena/browser/view/PRJEB21351). Bioinformatic analyses were ran again using the last version of BIOCOM-PIPE (V1.20)[15,27] with default parameters for pyrosequencing, updated SILVA database (version r132) and a new clustering tool called ReClustOR[14] (see the Data Records part for details). An average 39.27% increase of OTUs *per* sample was observed, but with the same global number of OTUs (188,030 OTUs), compared to the first analysis performed without ReClustOR[9,14]. These results were split in two datasets OTUs[20] and taxonomy[18]. Cumulative OTU curves from the analyzed samples (Fig. 2) indicated that the sequencing dataset represented efficiently the 16S microbial communities of the 1,842 soil samples when we ignored single-singleton sequences, as we reached a saturation. The high amount of different single-singleton sequences, representing rare microorganisms can be explained by the huge variability of soil environmental niches analyzed in the RMQS sampling survey. Taxonomic assessment results were given for all analyzed samples, at each taxonomic level (phylum, class, order, family, genus), against the r132 SILVA database.

Data Records

The following files are available at the French public research data repository (https://entrepot.recherche.data.gouv.fr/dataverse/rmqs_microbio). The physico-chemical dataset used in the "Usage Notes" section can also be found in the data repository previously cited[23]. The "`id_site`" column links all the data records with one another. More precisely, this "`id_site`" is the stripped-down identifier used by the RMQS project. See Fig. 1 for an overview of all available datasets.

Molecular microbial biomass (MMB). The biomass file[16] encompasses the results of the MMB quantification[8,26] for RMQS soil samples. Soil DNA was extracted using combined chemical, mechanical and thermal lysis methods to efficiently obtain the highest amount of DNA available[12]. This DNA extraction method was chosen because its comparison with commercial kits showed that it was the most reliable to recover important yield of DNA of sufficient quality for molecular analysis, with a specific and improved mechanical lysis step, and also because it is reliable and robust for the routine analysis of several hundreds of different soils (see[8,12,13]). Moreover, as the soil samples were previously sieved, we avoid any plant, detrital or root DNA extraction to extract DNA from microbial communities. The values range between 0 and more than 600 in microgram of DNA *per* gram of soil (μg DNA.g^{-1} soil). Quantifications lower than 2.5 $\mu.g^{-1}$ *DNA* are considered out of scale.

Diversity data (16S OTUs). The detailed methodological information given in this subsection was already presented[9,15,27] and is being included in this work for ease. A 16S rRNA gene fragment targeting the V3-V4 regions to characterize bacterial diversity was amplified using the primers F479 (CAGCMGCYGCNGTAANAC) and R888 (CCGYCAATTCMTTTRAGT)[9,13], and 2,132 soil samples were successfully amplified from the 2,173 DNA soil samples (obtained from the initially 2,240 soil samples available)[9]. The sequencing was done using a pyrosequencing method on a GS FLX Titanium (Roche 454 Sequencing System) by the Genoscope (Evry, France).

Bioinformatic analyses were realized using the last version of BIOCOM-PIPE (V1.20)[15,27] with default parameters. First, all raw sequences were sorted according to each multiplex identifier sequence and a preprocessing step was carried out to filter and delete low quality sequences based on (i) their length (fewer than 350 bases), (ii) their number of ambiguities (deletion of sequences with one or more N, or sequences with homopolymer of more than seven consecutive bases), and (iii) their primer sequence(s) (the proximal primer sequence had to be complete and without errors, but the distal primer can be incomplete, with a maximum of two mismatches tolerated). After strict dereplication, a specific step of chimera detection and filtering is launched called the "hunting-recovering"[15]. All remaining sequences (considered as cleaned) from the 2,140 samples have been deposited in a specific repository, in independent files for each sample[21]. The remaining high-quality sequences were normalized by random selection (10,000 high-quality sequences for each sample) to allow an efficient comparison of the datasets and avoid biased community comparisons. This normalization leads to the loss of some samples without sufficient sequencing depth, with only 1,842 soil samples kept. The high-quality sequences were clustered into OTUs at 95% of similarity after a global alignment, using the clustering tool called ReClustOR[14]. This level of clustering was chosen as it corresponds roughly to the *genus* level, particularly with our primer set (previous *in silico* evaluation)[10]. The clustering with ReClustOR was realized in two independent steps: all high-quality sequences were firstly clustered together using a classical clustering approach. Then, a post-clustering step with ReClustOR was done to improve the clustering using the RMQS OTUs as a reference database, to overcome problems (OTU stability and reliability) associated with classical clustering methods[14].

The first file is the abundance matrix of the whole dataset, with 1,842 samples as rows and 188,030 OTUs as columns[20]. Samples are named using the sampling site IDs and OTUs are ranked by abundance. Each row sums to 10,000; the rarefaction threshold used in the bioinformatics analysis. There was no filter on single sequences nor single-singleton values. The second file is the OTU core sequence (resulting from post-clustering at 95% similarity with ReClustOR), stored in a fasta format with 188,030 sequences. These files can be used directly in BIOCOM-PIPE with ReClustOR as a reference database. The third and last file is the taxonomic affiliation of each core OTU sequence using BIOCOM-PIPE, with names based on the R132 release of the Silva database[15].

Taxonomic data (16S). The taxonomic matrices (16S rDNA)[18] contain 1,842 sampling site IDs as rows, and 89 phyla, 194 classes, 495 orders, 739 families and 3,540 genera as columns for each file. The BIOCOM-PIPE taxonomic affiliation strategy used here is based on a complete classification of all high-quality reads, and not only representative OTU sequences, applied with USEARCH (v8.0.1623; www.drive5.com/usearch)[15]. All names are based on the R132 release of the Silva database[28]. Three columns were added for each file, "Unknown" (not matching any reference), "Unclassified" (missing taxa between genus and phylum) & "Environmental" (matched to sample from environmental study, generally with only a phylum name). Like OTU abundance, each row sums to the rarefaction threshold of 10,000. A metadata file is associated with each taxonomic abundance file, containing the highest taxonomic level of each taxon (e.g. the order metadata file contained the kingdom, phylum and class of each order).

Microbial habitats (16S). The detailed methodological information given in this subsection was already presented[5] and is being included in this work for ease. The microbial habitats were identified by fitting a multivariate regression tree $Y = f(X)$ where Y is the OTU matrix and X the set of environmental descriptors (land-use type, climatic factors, soil texture, pH, soil chemistry, elevation, etc.). The defined multivariate regression tree (MRT) analysis provided a 16-leaf tree explaining 35% of the distance-based variance, determined by five environmental variables: soil pH, land use, the C:N ratio, organic carbon content and average annual temperature. The habitats made up the leaves of the fitted tree: each contained a set of sites all similar in microbial composition and characterized by a conjunction of conditions on the explanatory X variables. The microbial habitat[19] file is a two-column table that associate each sampling site ID with one of the 16 defined microbial habitats (MH01 to MH16), with 1,798 rows that correspond to the 16S microbial habitat study[5]. The metadata deposited file contains the name and pH complex of each habitat.

Microbial gene abundances (16S, 18S) and 18S/16S ratio data. The methodological information given in this subsection was already presented[11] and is being included in this work for ease. Specific primers were used for the real-time PCR to quantify bacterial (16S; 341 F: CCTACGGGAGGCAGCAG and 515 R: ATTACCGCGGCTGCTGGCA) and fungal gene abundances (18S; FR1: ANCCATTCAATCGGTANT and FF390: CGATAACGAACGAGACCT)[29,30]. Conversely to the ITS region, the 18S gene contains conserved regions unaffected by length polymorphism. This polymorphism can lead to decrease the reproducibility and accuracy of the method with the ITS. This bias constitutes strong limitations in ecological studies of soil fungal communities.

The microbial (i.e. bacterial and fungal) abundances data files[17] contain the corrected gene abundances in terms of copies of 16S and 18S rDNA *per* gram of soil and the 18S/16S ratio from[11], with the corresponding site IDs (see Technical Validation section for details). The measured and corrected Ct values for 16S and 18S from each site are available in a second file. It is noteworthy that the F:B ratio given in the data file was computed with this formula: (number of 18S fungal copies) × 100 / (number of 16S bacterial copies). This F:B ratio is useful to evaluate the presence and the *equilibrium* of these two microbial groups in soil functioning, as fungi and bacteria display different metabolisms and life strategies but are key membres of many soil functions.

BIOCOM-PIPE Control samples, 16S sequencing library and parameters for 16S metabarcoding. Reference environmental DNA samples named "G4" in internal laboratory processes were added for each molecular analysis[31]. They were used for technical validation, but not published alongside the datasets. Each control sample is named after its sequencing library. The taxonomy and OTU abundance files for these control samples are built like the taxonomy and abundance file described above. As these internal control samples were clustered against the RMQS dataset in an open reference fashion, they contain new OTUs (noted as "OUT") that correspond to sequences that did not match any of 188,030 RMQS reference sequences. The library association files link each sample to its sequencing library, and each library is detailed (with a direct link to the EBI FTP server) in the metadata file.

The "`project.csv`" file contains information on multiplexing, with the library, associated primer and multiplex identifier of each sample. The "`Input.txt`" file is the parameter file for the BIOCOM-PIPE pipeline, with specific settings to be run on pyrosequencing data (the default was Illumina data). With these two files[31], the bioinformatic process used to obtain these results can re-ran directly with BIOCOM-PIPE.

Technical Validation

Technical validations for MMB quantification, habitat definition and qPCR analyses were previously described in their respective papers[5,8,11]. For example, to enhance the robustness of data comparison for qPCR analyses, a post-processing treatment was performed by calibration using the master curve method[32]. First, the mean of the reference DNA threshold cycles (Ct) of the whole dataset was computed. Differences in amplification efficiency between all PCR plates were estimated by computing derivation between the mean of the complete dataset reference Ct and the mean reference Ct of each PCR plate. Second, for each plate, the derivation was deducted from the Ct to obtain a corrected Ct. Third, the slope and intercept of a master calibration curve were calculated by using the values (corrected Ct and concentration) of all standards from all experiments (reference environmental DNA sample and of plasmid DNA standards). Finally, the number of rDNA copies of each environmental sample was defined based on the corresponding corrected Ct and the master calibration curve parameters.

For habitat definition, the optimal complexity of the tree was assessed as follows[5]. Given a size *s*, 10 trees of size *s* are fitted using a 10-fold cross validation. This provides a mean and standard error of the prediction error for *s*-size trees. Repeating this procedure for various sizes *s* help to select the shortest size within one standard error of the overall best size.

For the sequencing data (OTUs and taxonomy), internal controls were added to each library to ensure a robust comparison of the datasets across sequencing platforms and libraries, but were not described in the initial publication. These controls were built from a mixture of DNA extracted from 300 soil samples from a previous landscape study[33] for them to be as similar to the analyzed soil samples as possible. One internal control was added to each series of analyses during the first PCR step, and underwent the same analytical process as the samples did (PCR amplification, purification, quantification). These controls were used to assess most of the technical drift (benchtop molecular biology and sequencing) of each library. If a control highlighted one drift (i.e. low sequencing depth, microbial community composition drift or richness drift), the complete library was excluded from the analysis. Thus, 72 internal controls corresponding to 72 libraries were compared and checked in terms of richness and taxonomy reliability. For example, their richness was relatively consistent (around 2,600 OTUs on average), except four of them with highlighted drifts due to the technical process (Fig. 3a). These controls considered as "outliers" harbored a lower number of sequences, below the defined threshold of 10,000 high-quality sequences, explaining their low richness (Fig. 3a) and their drift (Fig. 3b). All samples in the same libraries as these "outliers" were therefore excluded (except for six of them), because the controls did not pass the complete validation process. Regarding the other controls not considered as "outliers", the observed differences regarding the biological variability of analyzed samples were clearly lower (see Fig. 3b), showing a good reliability of our validation process.

Usage Notes

The data described here were mainly collected to be used as references for soil microbial studies. Moreover, they can be matched with the data on soil physicochemical properties, site environment and land use available from another data collection[23]. A filtering step of this dataset is needed on two columns: "`site_officiel`"

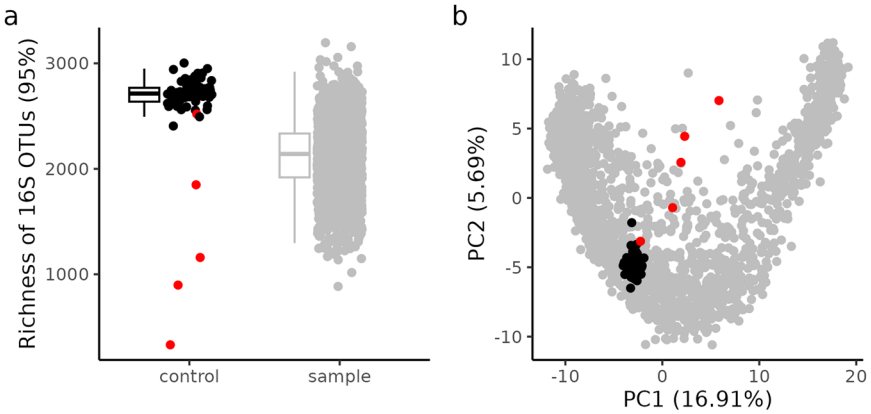

Fig. 3 Richness and *genus* composition of the RMQS and control samples. Black & red, control samples; gray, RMQS samples. Red control samples, "outliers" that did not pass the whole validation process. (**a**) RMQS and control samples richness of 16S OTUs. (**b**) PCoA analysis (first and second axis using robust Aitchison distance) of RMQS and control samples targeting the 16S *genus* taxonomic level.

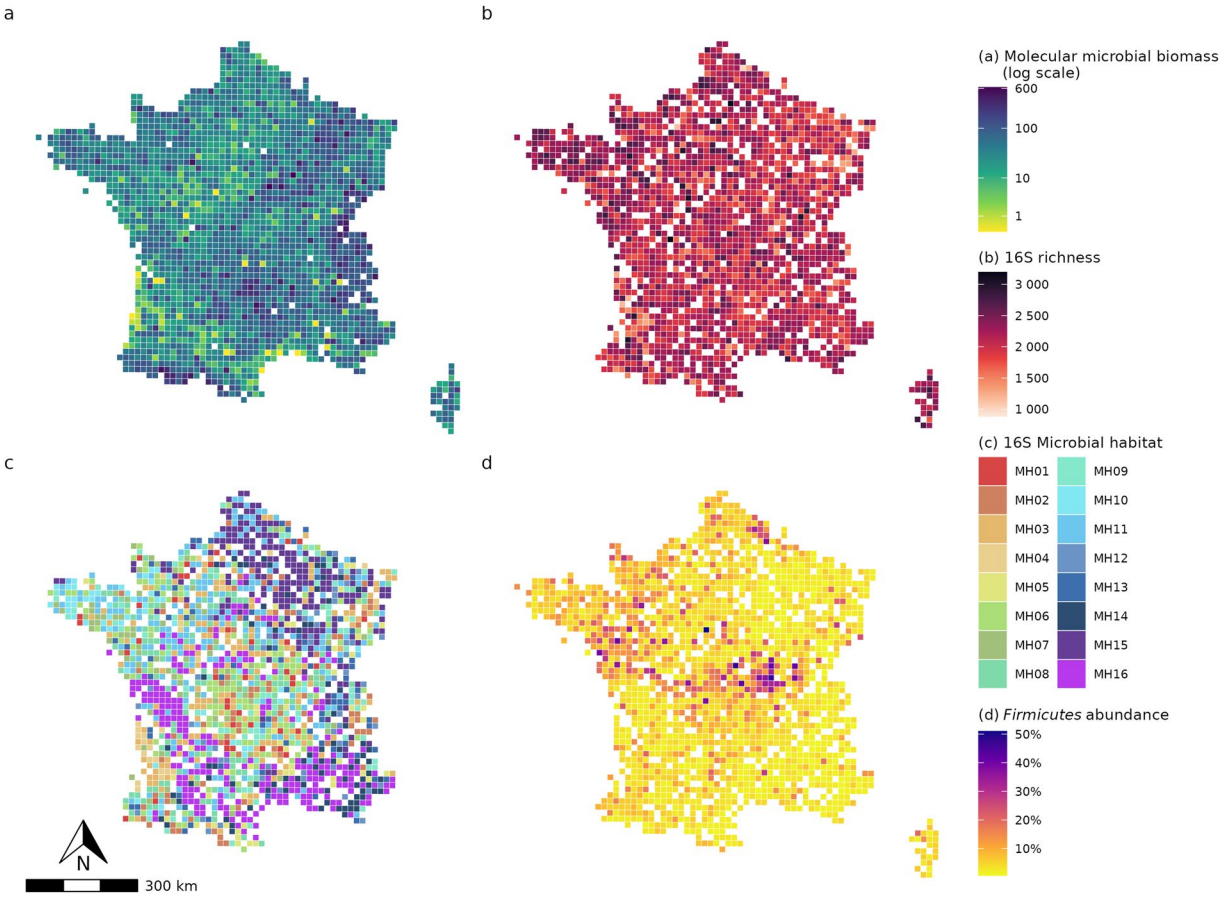

Fig. 4 Maps of some of the dataset variables based on theoretical positions in the RMQS systematic 16 × 16 km grid. (**a**) Molecular microbial biomass values from the `biomass.tsv` file. (**b**) Microbial richness (16S) values computed on the abundance matrix (`otu_abundance.tsv`). (**c**) Microbial habitat (16S) values from the `microbial_habitat.tsv` file. (**d**) *Firmicutes* abundance values (%) from `taxonomy_phylum.tsv`. White, missing data.

(official site) to `TRUE` and "`no_couche`" (soil layer number) to 1 to get matching dataset. Using the theoretical positions from the RMQS grid, maps can be produced for each of the data described in this paper (Fig. 4) to show molecular microbial biomass (a), 16S richness (b) from the 16S abundance matrix, 16S microbial habitat (c), and an example of a phylum from 16S taxonomic data (d).

Code availability

Bioinformatics analyses were performed with BIOCOM-PIPE software (v1.20), that can be downloaded at https://forgemia.inra.fr/biocom/biocom-pipe. The complete database and tools needed to run BIOCOM-PIPE can be downloaded from a dedicated Zenodo repository[27] matching its version. The sets of parameters are available in "pipeline config files parameters" from the Data Records section. For ease of access and reproducibility, the code used to load, filter and plot data to produce the figures of this paper is available at https://forgemia.inra.fr/biocom/data_paper_rmqs1_16s/-/tree/main/figures. The code gives access to R scripts[34] version 4, and uses metapackage tidyverse[35] and packages FactoMineR[36], hillR[37], janitor[38], patchwork[39] and sf[40].

Received: 19 February 2024; Accepted: 17 December 2024;
Published online: 08 January 2025

References

1. Guerra, C. A. *et al.* Global hotspots for soil nature conservation. *Nature* **610**, 693–698, https://doi.org/10.1038/s41586-022-05292-x (2022).
2. Bloor, J. M. G., Si-Moussi, S., Taberlet, P., Carrère, P. & Hedde, M. Analysis of complex trophic networks reveals the signature of land-use intensification on soil communities in agroecosystems. *Scientific Reports* **11**, 1–11, https://doi.org/10.1038/s41598-021-97300-9 (2021).
3. Delgado-Baquerizo, M. *et al.* Soil microbial communities drive the resistance of ecosystem multifunctionality to global change in drylands across the globe. *Ecology Letters* **20**, 1295–1305, https://doi.org/10.1111/ele.12826 (2017).
4. Trivedi, P., Delgado-Baquerizo, M., Anderson, I. & Singh, B. Response of soil properties and microbial communities to agriculture: Implications for primary productivity and soil health indicators. *Frontiers in Plant Science* **7**, 1–13, https://doi.org/10.3389/fpls.2016.00990 (2016).
5. Karimi, B. *et al.* Biogeography of soil microbial habitats across France. *Global Ecology and Biogeography* **29**, 1399–1411, https://doi.org/10.1111/geb.13118 (2020).
6. Jolivet, C. *et al. French Soil Quality Monitoring Network Manual RMQS2: second metropolitan campaign 2016–2027*. Publisher: INRAE (INRAE, 2022).
7. Gardi, C. *et al.* Soil biodiversity monitoring in Europe: ongoing activities and challenges. *European Journal of Soil Science* **60**, 807–819, https://doi.org/10.1111/j.1365-2389.2009.01177.x (2009).
8. Dequiedt, S. *et al.* Biogeographical patterns of soil molecular microbial biomass as influenced by soil characteristics and management: Biogeography of soil microbial biomass. *Global Ecology and Biogeography* **20**, 641–652, https://doi.org/10.1111/j.1466-8238.2010.00628.x (2011).
9. Terrat, S. *et al.* Mapping and predictive variations of soil bacterial richness across France. *PLOS ONE* **12**, e0186766, https://doi.org/10.1371/journal.pone.0186766 (2017).
10. Karimi, B. *et al.* Biogeography of soil bacteria and archaea across France. *Science Advances* **4**, eaat1808, https://doi.org/10.1126/sciadv.aat1808 (2018).
11. Djemiel, C. *et al.* Biogeographical patterns of the soil fungal:bacterial ratio across France. *mSphere* e00365–23, https://doi.org/10.1128/msphere.00365-23 (2023).
12. Terrat, S. *et al.* Molecular biomass and MetaTaxogenomic assessment of soil microbial communities as influenced by soil DNA extraction procedure: Soil DNA extraction impact on bacterial diversity. *Microbial Biotechnology* **5**, 135–141, https://doi.org/10.1111/j.1751-7915.2011.00307.x (2012).
13. Terrat, S. *et al.* Meta-barcoded evaluation of the ISO standard 11063 DNA extraction procedure to characterize soil bacterial and fungal community diversity and composition: Meta-barcoded evaluation of the ISO-11063 standard. *Microbial Biotechnology* **8**, 131–142, https://doi.org/10.1111/1751-7915.12162 (2015).
14. Terrat, S. *et al.* ReClustOR: a re–clustering tool using an open—reference method that improves operational taxonomic unit definition. *Methods in Ecology and Evolution* **11**, 168–180, https://doi.org/10.1111/2041-210X.13316 (2020).
15. Djemiel, C. *et al.* BIOCOM-PIPE: a new user-friendly metabarcoding pipeline for the characterization of microbial diversity from 16S, 18S and 23S rRNA gene amplicons. *BMC Bioinformatics* **21**, 492, https://doi.org/10.1186/s12859-020-03829-3 (2020).
16. Dequiedt, S. *et al.* RMQS1 molecular microbial biomass. *Research Data Gouv* https://doi.org/10.57745/VLWJ51 (2023).
17. Djemiel, C. *et al.* RMQS1 Microbial density (16S and 18S) and F:B ratio. *Research Data Gouv* https://doi.org/10.57745/1Z90HV (2023).
18. Karimi, B. *et al.* RMQS1 16S taxonomy. *Research Data Gouv* https://doi.org/10.57745/WIRXIC (2023).
19. Karimi, B. *et al.* RMQS1 16S microbial habitat. *Research Data Gouv* https://doi.org/10.57745/JK5WCD (2023).
20. Terrat, S. *et al.* RMQS1 16S operational taxonomic units. *Research Data Gouv* https://doi.org/10.57745/ZZWKGQ (2024).
21. Terrat, S. *et al.* RMQS1 16S unrarefied cleaned sequences. *Research Data Gouv* https://doi.org/10.57745/02XLWL (2024).
22. Arrouays, D. *et al.* A new projection in France: A multi-institutional soil quality monitoring network. In *Comptes Rendus de l'Academie d'Agriculture de France* **88**, 93–103 (2002).
23. Institut National De La Recherche Agronomique *et al.* Analyses physico-chimiques des sites du Réseau de Mesures de la Qualité des Sols (RMQS) du territoire métropolitain pour la 1ère campagne (2000-2009), avec coordonnées théoriques, https://doi.org/10.15454/QSXKGA (2021).
24. Quintana-Seguí, P. *et al.* Analysis of Near-Surface Atmospheric Variables: Validation of the SAFRAN Analysis over France. *Journal of Applied Meteorology and Climatology* **47**, 92–107, https://doi.org/10.1175/2007JAMC1636.1 (2008).
25. Terrat, S. *et al.* Improving soil bacterial taxa—area relationships assessment using DNA meta-barcoding. *Heredity* **114**, 468–475, https://doi.org/10.1038/hdy.2014.91 (2015).
26. Horrigue, W. *et al.* Predictive model of soil molecular microbial biomass. *Ecological Indicators* **64**, 203–211, https://doi.org/10.1016/j.ecolind.2015.12.004 (2016).
27. Terrat, S., Cottin, A. & Djemiel, C. BIOCOM-PIPE Databases-Tools V1.20. *Zenodo* https://doi.org/10.5281/ZENODO.4428756 (2021).
28. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* **41**(D1), D590–D596, https://doi.org/10.1093/nar/gks1219 (2012).
29. Plassart, P. *et al.* Evaluation of the ISO Standard 11063 DNA Extraction Procedure for Assessing Soil Microbial Abundance and Community Structure. *PLoS ONE* **7**, e44279, https://doi.org/10.1371/journal.pone.0044279 (2012).
30. Chemidlin Prévost-Bouré, N. *et al.* Validation and Application of a PCR Primer Set to Quantify Fungal Communities in the Soil Environment by Real-Time Quantitative PCR. *PLoS ONE* **6**, e24166, https://doi.org/10.1371/journal.pone.0024166 (2011).
31. Terrat, S. & Dequiedt, S. RMQS1 16S bioinformatic config files and control sample data. *Research Data Gouv* https://doi.org/10.57745/XBFOJP (2024).
32. Sivaganesan, M., Seifring, S., Varma, M., Haugland, R. A. & Shanks, O. C. A Bayesian method for calculating real-time quantitative PCR calibration curves using absolute plasmid DNA standards. *BMC Bioinformatics* **9**, 120, https://doi.org/10.1186/1471-2105-9-120 (2008).

33. Constancias, F. *et al*. Mapping and determinism of soil microbial community distribution across an agricultural landscape. *Microbiology Open* **4**, 505–517, https://doi.org/10.1002/mbo3.255 (2015).
34. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2022).
35. Wickham, H. *et al*. Welcome to the tidyverse. *Journal of Open Source Software* **4**, 1686, https://doi.org/10.21105/joss.01686 (2019).
36. Lê, S., Josse, J. & Husson, F. FactoMineR: A Package for Multivariate Analysis. *Journal of Statistical Software* **25**, 1–18, https://doi.org/10.18637/jss.v025.i01 (2008).
37. Li, D. hillR: taxonomic, functional, and phylogenetic diversity and similarity through Hill Numbers. *Journal of Open Source Software* **3**, 1041 (2018).
38. Firke, S. *janitor: Simple Tools for Examining and Cleaning Dirty Data* (2023).
39. Pedersen, T. L. *patchwork: The Composer of Plots* (2022).
40. Pebesma, E. Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal* **10**, 439–446, https://doi.org/10.32614/RJ-2018-009 (2018).

Acknowledgements

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 an Ecological Transition and Territorial Cohesion (MCT), the French Ministry of Agriculture and Food (MASA), the French Agency for Ecological Transition (ADEME), the French Biodiversity Agency (OFB), the National Institute of Geographic and Forest Information (IGN), the French Institute for Research and Development (IRD), the French Geological Survey (BRGM) and the National Research Institute for Agriculture, Food and Environment (INRAE). We thank all the soil surveyors and technical assistants involved in sampling the sites and the staff of the European Soil Samples Conservatory (INRAE Orléans) for preparing the soil samples. Calculations were performed using HPC resources from DNUM Centre de Calcul de l'Université de Bourgogne (CCUB). Due to the involvement of technical facilities of the GenoSol platform (DOI 10.15454/L7QN45) of the infrastructure ANAEE-Services, they received a grant from the French state through 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 Bourgogne Franche-Comté. The BRC GenoSol is also a part of BRC4Env (10.15454/TRBJTB), the "Environmental Resources" pillar of the Research Infrastructure AgroBRC-RARe. The microbial gene abundances study was granted by the project Agro-Eco Sol coordinated by Aurea AgroSciences in partnership with INRAE and ARVALIS. The Agro-Eco Sol program is supported by the French program "Investments for the Future," operated by ADEME (The French Environment and Energy Management Agency). We also thank the Regional Council of Bourgogne Franche-Comté as they funded a post-doctoral position through a seed project called "BD-RMQS-AgroEcoSol" (Number: 2021PRE00370), dedicated to this project. The funders had no role in study design, data interpretation, or the decision to submit the work for publication. Thanks are also extended to Annie Buchwalter for correcting and improving English language in the manuscript.

Author contributions

Conceptualization: A.C., C.D., S.D., L.R., S.T. Data compilation: A.C., C.D., S.D., S.T. Visualization: A.C., C.D. Supervision: S.T. Writing - original draft: A.C., S.T. Writing - review & editing: A.C., C.D., L.R., S.D., S.T. Data collection: S.D., C.D., N.C.P.-B., M.L., J.T., L.T., T.R., B.K., S.T. Data collection supervision: N.S., A.B., C.J., P.-A.M., L.R.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.T.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.