nature microbiology

Perspective

https://doi.org/10.1038/s41564-022-01228-3

Defending Earth's terrestrial microbiome

In the format provided by the authors and unedited

Estimating Global Soil Fungal Microbiome Exploration Priorities using the Global Fungi dataset

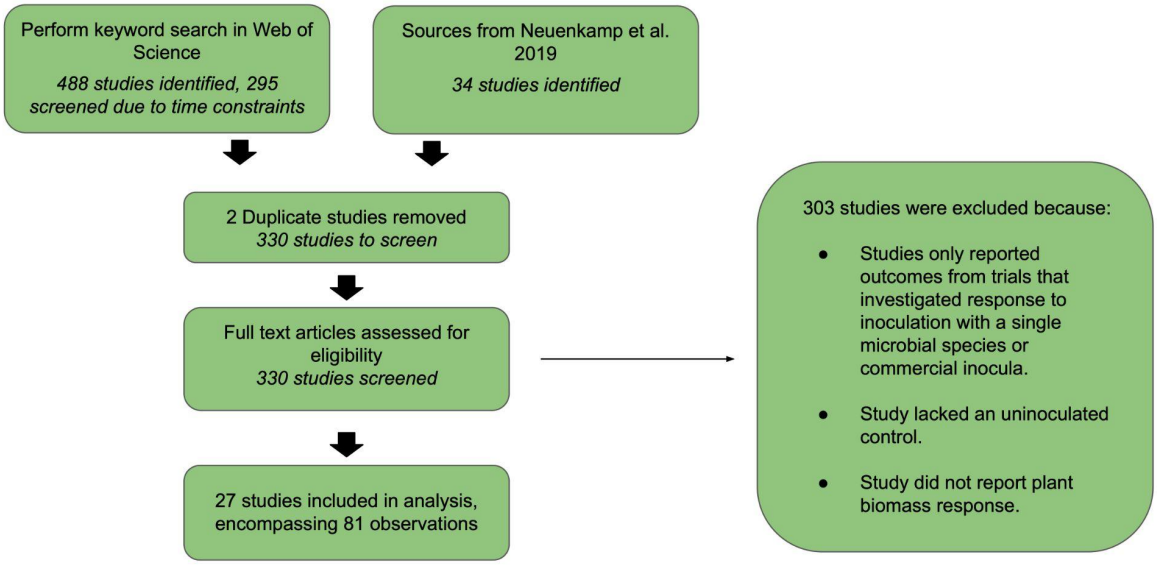
We used Global Fungi[1], a global meta-dataset of soil fungal DNA sequencing observations, to understand which areas of the Earth are most under-studied with respect to soil fungal biodiversity. As our focus is on soil microbial communities, we subsetted to samples taken from any soil environment. In addition, we restricted the sample set to those that sequenced the ITS region (ITS1, ITS2 or full ITS), using an Illumina, PacBio, Roche-454 or IonTorrent sequencing platform. This left ~10,000 observations for our under-sampling analysis.

We described under-sampling in two ways. First, we considered environmental novelty - which combinations of climate, soil, vegetation, topography, and anthropogenic influence have never been sampled? To do so we first extracted environmental covariate data from 68 global layers, providing information on soil physio-chemical properties, climate, vegetation, topography and anthropogenic variation. A complete list of global covariate layers used, with their associated references, is presented in Supplementary Data File 1. Next, we transformed each observation's 68 covariate layers into Principal Component space using the PCA function within the sklearn[2] package for Python 3.8.10[3], and using the eigenvectors we transformed the global covariate data into the same PC space. For each of the 136 bivariate combinations of the first 17 PC axes, collectively explaining 90% of observed variation in environmental space, we assessed whether pixel values of the global covariate data fell within or outside the convex hulls of the sampled data (Figure 2A). Additional detail in this method can be found in van den Hoogen et al. 2019[4]. Second, we considered absolute geographic distance from the current set of geo-referenced observations, with areas further from sampled locations receiving higher under-sampling scores (Figure 2B). To generate the final image (Figure 2C), we overlaid figures 2A using a weighted averaging approach. We present both maps individually, as well as a combined map that is a 2:1 averaging of the spatial distance and environmental distance visualizations. The maps are ultimately independent of one another, and so the choice of averaging ratio is arbitrary. We selected a 2:1 ratio as we felt it best captured the most striking visual elements of both maps. We present readers each map individually as well to further increase transparency.

Data Collection for Microbiome Restoration Meta Analysis

We began our data collection using a recent meta-analysis on soil mycorrhizal restoration (Neuenkamp et al. 2019)[5]. We expanded this analysis by searching Web of Science using the key words: *mycorrhiz*, seedling* and different expressions for soil inoculation (*soil inocul**, *whole soil inocul**, *soil transplant**, *whole community inocul**) in order to reach a large range of

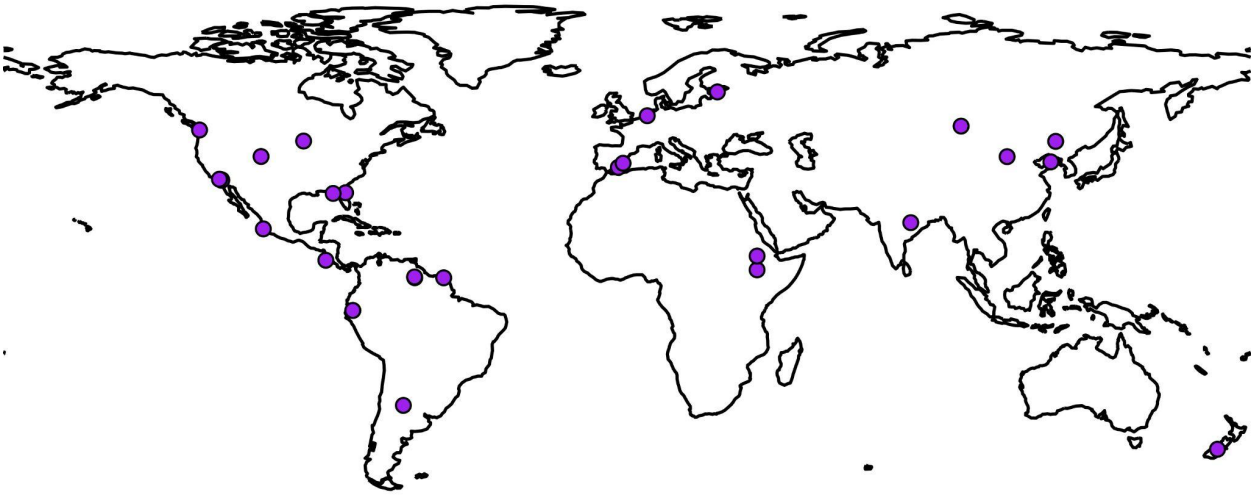
inoculation experiments. Studies must have met the following criteria in order to be included. First, only studies were taken into account which presented raw data for plant biomass responses. Second, studies which used single fungal strains for inoculation or commercial inoculum were not considered. Finally, a study must consist of at least one treatment (inoculation with living organisms) and a corresponding control for which either sterilized field soil was added or no addition at all. Our literature search was concluded on February 15, 2022. A flow chart of our data collection procedure is presented in Supplementary Methods Figure 1.



Supplementary Methods Figure 1. Flow chart of meta-analytic literature search and screening.

To investigate the effect of soil microbiome inoculation on plant growth, we calculated log response ratios comparing the biomass of plants inoculated with whole soil, wild spores, or root samples, to uninoculated controls. We did not include studies examining commercial inoculum products. Our dataset included both field and laboratory studies. For data filtering, we followed Neuenkamp et al.[5] to ensure that our datasets would be compatible. As such, for studies with multiple time points, we only examined the last time point. For studies in which multiple treatment groups meeting our criteria were compared with a single control, we calculated and used composite means and variances following Borenstein et al[6]. Similarly, if multiple control groups were present (for example, a control with no soil inoculation as well as a control with a sterilized soil inoculation), we also calculated and made use of composite means and variances. We were unable to extract a measure of variance from five studies, so we imputed standard deviation using

the methodology of Sinclair & Bracken 1994 as implemented in R package metagear[7,8]. We excluded studies from which sample sizes could not be extracted. Then, we calculated log response ratios and variances using escalc from R package metafor[9]. We combined our dataset with that compiled by Neuenkamp et al, using their already-calculated log response ratios and variances. However, several log response ratios reported by Neuenkamp et al were listed with variances of zero, which caused problems because responses were weighted by the inverse of their variance in our downstream modeling. Accordingly, we instead assigned variances equal to the smallest non-zero variance value observed in our dataset to these log response ratios. After filtering and dataset harmonization, we retained 81 comparisons sourced from 27 different studies. A complete list of studies used, and observed effect sizes can be found in Supplementary Data File 2. A map of the spatial distribution of observations is presented here (Supplementary Methods Figure 2).



Supplementary Methods Figure 2. Spatial distribution of soil inoculation studies used in this analysis.

Analysis of plant biomass response to microbiome restoration

To estimate effect size across all studies, we used a linear mixed effects model as implemented in metafor command rma.mv[9]. In this model, we included a categorical random effect for data source to account for the correlation of multiple different comparisons sourced from the same study and weighted each datapoint by the inverse of its variance. Output from this model was exponentiated for generation of figures and reporting of effect sizes on an untransformed scale.

References

1. Větrovský, T. *et al.* GlobalFungi, a global database of fungal occurrences from high-throughput-sequencing metabarcoding studies. *Sci Data* **7**, 228 (2020).
2. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
3. Rossum, G. van & Drake, F. L. *The Python language reference*. (Python Software Foundation, 2010).
4. van den Hoogen, J. *et al.* Soil nematode abundance and functional group composition at a global scale. *Nature* **572**, 194–198 (2019).
5. Neuenkamp, L., Prober, S. M., Price, J. N., Zobel, M. & Standish, R. J. Benefits of mycorrhizal inoculation to ecological restoration depend on plant functional type, restoration context and time. *Fungal Ecology* **40**, 140–149 (2019).
6. *Introduction to meta-analysis*. (John Wiley & Sons, 2009).
7. Lajeunesse, M. J. Facilitating systematic reviews, data extraction and meta-analysis with the metagear package for r. *Methods Ecol Evol* **7**, 323–330 (2016).
8. Sinclair, J. C. & Bracken, M. B. Clinically useful measures of effect in binary analyses of randomized trials. *Journal of Clinical Epidemiology* **47**, 881–889 (1994).
9. Viechtbauer, W. Conducting Meta-Analyses in *R* with the **metafor** Package. *J. Stat. Soft.* **36**, (2010).