**Title:** A Meta-Analysis of the Microbial Distance-Decay Relationship

**Authors:** Dave R. Clark, Graham J.C. Underwood, Terry J. McGenity, and Alex J. Dumbrell

**Addresses:** School of Biological Sciences, University of Essex, Wivenhoe Park, Colchester, Essex, CO4 3SQ, UK.

**Keywords:** Bacteria, Archaea, spatial ecology, Mantel test, macroecology, dispersal limitation

**Running title:** Microbial Distance-Decay Relationships

**Abstract**

Ecological communities closer together in space are generally more similar than those further apart, according to the distance-decay (d-d) of similarity relationship. Historically, microorganisms were assumed to defy this relationship due to their capacity for long distance, passive dispersal. Yet, molecular studies have recorded highly variable d-d relationships in a range of microbial communities from different environments, using very different methods. Therefore, we sought to understand whether this variability in microbial d-d relationships is caused by different study methodologies or biological contexts. To do this, we conducted a meta-analysis and gathered information on 287 microbial d-d relationships, using the Mantel correlation coefficient as a measure of effect size. We found that d-d relationships were only weakly related to sequencing depth, and differences between community quantification methods (e.g. fingerprinting, high-throughput sequencing, morphological) were only observed for statistically significant d-d relationships. Distance-decay relationships derived using phylogenetic measures of community dissimilarity (e.g. Unifrac) were generally weaker than those using compositional dissimilarity metrics (e.g. Jaccard, Bray-Curtis). Significant differences were also observed between different study systems, with soils showing weaker d-d relationships, and host-associated systems showing stronger relationships. Stronger d-d relationships were also observed over large spatial scales, but no difference was observed between different taxa. We conclude that the microbial d-d relationship may be highly context-dependent, changing with study system and spatial scale, but could also be influenced by the methods used. We provide suggestions for selecting methods that will enhance ecological signals.

**Introduction**

The distance-decay (d-d) relationship is one of the most commonly studied relationships in macroecology (Nekola & White 1999; Condit *et al.* 2002; Soininen *et al.* 2007). The relationship quantifies how the similarity of community composition decays with increasing geographic distance between communities, such that communities close together contain more similar species assemblages than those further apart. In studies of “macro-organisms”, distance-decay relationships are frequently observed, and may occur over relatively short geographic distances, dependent on the dispersal abilities of the organisms present.

Microorganisms are thought to be capable of frequent and/or long-distance dispersal. One of the earliest hypotheses about the nature of the microbial distance-decay relationship was formulated by Dutch microbiologist, Baas-Becking. Within this hypothesis, Baas Becking postulated that “everything is everywhere but the environment selects” (Baas Becking 1934). The rationale behind this hypothesis was that microorganisms should be efficient long distance dispersers due to their small size and often high population densities. Therefore, according to EiE, microorganisms should not exhibit a distance-decay relationship because communities separated by large geographic distances will contain the same, globally dispersed species (Finlay 2002; Finlay & Fenchel 2004).

However, the rapid development of molecular methods to study microbial communities, coupled with the provocative nature of EiE has resulted in an explosion of studies testing the distance-decay relationship in microbial communities, revealing mixed results. Many studies have reported relationships, varying in steepness, between microbial community composition and geographic distance for a range of scales and taxa (Dumbrell *et al.* 2010; Martiny *et al.* 2011; Barreto *et al.* 2014). Yet, a great number of studies have also found no correlation between microbial community composition and distance (Hazard *et al.* 2013; Kivlin *et al.* 2014). Therefore, the universality of this relationship for microbial communities remains unclear.

Over the past two decades, the tools available to ecologists have allowed the generation of datasets several orders of magnitude larger. From serially sequencing clone libraries, through community fingerprinting methods, and most recently high-throughput sequencing, previous research into the microbial d-d relationship is based upon a plethora of methods with varying degrees of taxonomic resolution and community coverage (Muyzer 1999; Glenn 2011). In addition to the varying methods used to study microbial communities, there are now a bewildering array of indices available to quantify the (dis)similarity between microbial communities, including qualitative (based on presence/absence of species), quantitative (based on composition and abundance of species), and phylogenetic (based on relatedness of communities). Such indices have different properties in terms of how they weight rare or common species, and in how they are influenced by sample sizes or species richness (Baselga 2012; Beck *et al.* 2013). Therefore, such methodological differences between studies may significantly influence our measurement of the d-d relationship in microbial ecology.

On the other hand various aspects of biological context may also play a role. Here, context could be considered to be the organisms studied (e.g. Bacteria, Archaea, Fungi), the study system in question (soil, freshwater, extreme systems), or the spatial scale of the study. Taxonomic disparities between studies may cause different results if dispersal abilities are different between microorganisms. For example, cell size has been shown to influence dispersal properties in microbial organisms (Wilkinson *et al.* 2012), and varies between microbial taxa, with Eukarya tending to have larger cells than Bacteria or Archaea (Soininen *et al.* 2013). Study system may also influence the rate of the d-d relationship as previously reported (Soininen *et al.* 2007). Different environments may have different degrees of connectivity, thus facilitating or hindering long-distance dispersal, e.g. . Additionally, different environments maintain environmental gradients over different distances. For instance, soils and sediments may support strong environmental gradients over small spatial scales (e.g. Dumbrell *et al*. 2010), whereas the well mixed surface waters of oceans may facilitate more diffuse environmental gradients. Finally, the spatial scale over which the study occurs could influence the outcome. Larger spatial scales may result in a decrease in dispersal between communities, and greater environmental heterogeneity, both of which may result in stronger d-d relationships (Martiny *et al.* 2011).

Given the variability in microbial distance-decay relationships reported in the literature, we sought to understand whether methodological or contextual differences between studies may influence measured distance-decay relationships. To do this, we synthesise available data on the microbial distance-decay relationship in a meta-analysis, and use factors relating to methodological or contextual aspects of each study to predict this relationship. If methodological differences between studies explain the distance-decay relationship, we predict that:

* Studies based on metagenetic sequencing will find stronger distance-decay relationships than those based on community fingerprinting, or low throughput methods.
* More intense sequencing depth approaches will result in stronger distance-decay relationships.
* Different similarity indices will result in different distance-decay relationships.

Whereas, if factors relating to the context of each study influence the nature of the distance-decay relationship, we hypothesise that:

* The distance-decay relationship will vary between different study taxa.
* The d-d relationship will vary between study systems.
* Studies covering larger spatial scales will result in stronger distance-decay relationships.

**Methods**

*Meta-Analysis*

We assembled a dataset on microbial distance-decay relationships by conducting a systematic literature search using the Web of Science search portal. To do this, we designed five different search terms in order to maximise the size of the resulting dataset, whilst minimising irrelevant or unsuitable studies (Table 1). All five searches were conducted on 08/06/2017, and results between 1900-2017 were saved. We downloaded all search results from Web of Science and used the “metagear” package (version 0.4) in R (version 3.4.1) to manually screen abstracts for suitability for inclusion in our study (R Development Core Team 2016; Lajeunesse 2016). We focussed on studies that had tested the distance-decay relationship using the Mantel correlation test. The Mantel test is used to test for correlation between two matrices (i.e. community dissimilarity and geographic distance), and is a common method for testing this relationship (Legendre *et al.* 2015). The Mantel correlation coefficient varies between -1 and 1, and we used this as our effect size. To standardise correlation coefficients between studies that had used similarity matrices, rather than dissimilarity matrices, we multiplied the former by -1, so that all correlation coefficients reflect the correlation between dissimilarity and geographic distance. Partial Mantel statistics (which are able to test for correlation between two matrices whilst controlling for a third) were excluded as they may be heavily influenced by other variables included in the test, and are therefore not easily comparable between studies. We extracted information relating to the methodology of each study as well as the biological context. Specific details of the information we extracted are in Table 2.

*Statistical Analyses*

We used ANOVA to test whether the Mantel coefficients were different between community characterisation methods, different dissimilarity indices, study taxa, and different environments. We also tested whether continuous variables such as study scale or sequencing depth were related to the Mantel coefficients. We first log transformed both study scale and sequencing depth as these variables spanned several orders of magnitude.

**Results**

Our initial Web of Science searches resulted in 2,250 search hits (Table 1). After removing duplicate hits (i.e. studies that appeared in multiple searches), this number was reduced to 2,031 hits. Manual screening of the abstracts yielded 547 studies that were deemed to be potentially suitable for use in this analysis. Data were obtained from 108 studies represented in 33 journals (Fig. S1), resulting in 287 recorded d-d relationships (Fig. 1). Reported Mantel correlation coefficients ranged from -0.24 to 0.95, with a mean of 0.27 (std. error = 0.014).

*Influence of methodological factors on D-D relationship*

To determine whether the microbial distance-decay relationship may be influenced by methodological factors, we tested whether the method of community characterisation, sampling depth, or choice of community similarity index influence the Mantel correlation coefficient. In contrast to our prediction, high-throughput sequencing methods (HTS) did not result in significantly higher Mantel coefficients compared to fingerprinting methods, or other low-resolution methods (Figure 2A; F2, 284 = 0.19, *P* = 0.83). However, when we examined only statistically significant (alpha = 0.05) Mantel coefficients (Fig. 2B), studies that used HTS had higher Mantel coefficients, approaching statistical significance (F2, 175 = 2.73, *P* = 0.07).

Sampling depth was also significantly and positively related to the Mantel coefficient, albeit with a small effect size (Fig. S2; slope = 0.02, *P* < 0.05, adj-R2 = 0.02). Significant Mantel coefficients were not related to sequencing depth (slope = 0.01, *P* = 0.28). Sampling depth was not correlated to sampling effort (Pearson’s R = 0.03, *P* = 0.64), and was also unrelated to Mantel test statistics (slope = -0.02, *P* = 0.10).

In line with our expectation, significant differences were also detected between dissimilarity indexes (F14, 271 = 4.96, *P* < 0.001). Several indices were excluded from this analysis as they had too few occurrences to calculate a reliable estimate of the mean (indices with < 4 occurrences were excluded). Tukey HSD tests showed Mantel coefficients from Raup-Crick and Unifrac indices were significantly lower than Bray-Curtis (*P* < 0.01 in each case, Fig. 3A), and Sørensen-based coefficients were higher than Euclidean, Raup-Crick, and Unifrac indices (*P* < 0.01 in all cases, Fig. 3A). Furthermore, Mantel coefficients were significantly different between index types (Fig. 3B; F2, 284 = 5.41, *P* < 0.01), with Tukey HSD test showing that Mantel coefficients based on phylogenetic distances were significantly lower than those based on abundance (*P* < 0.01) and binary indices (*P* < 0.05).

*Influence of biological context on D-D relationship*

As well as methodological considerations, we also tested for the influence of several biological contextual aspects including study taxa, scale, and biome. By far the most commonly studied taxa were Bacteria, followed by Fungi and microbial Eukaryotes, yet no difference was found in the Mantel coefficients associated with each taxon (F5, 281 = 1.39, *P* = 0.23). Examining only statistically significant Mantel coefficients revealed marginally significant differences between taxa (F5, 172 = 2.51, *P* < 0.05) with studies incorporating Bacteria and Fungi (n = 3) being significantly lower than studies on Archaea (Tukey HSD; *P* < 0.05). Removing these three data points resulted in no differences between taxa (F4, 170 = 1.45, *P* = 0.22).

For those biomes with more than three data points, significant differences in Mantel coefficients were found (F8, 262 = 8.80, *P* < 0.001). Specifically, sponge associated communities displayed higher coefficients than all other biomes (Tukey HSD; *P* < 0.05 in all cases), and grassland communities had lower coefficients than most other biomes (forest, lake, ocean, river, sediment, and sponge. Tukey HSD; *P* < 0.05 in all cases). Furthermore, the different types of environmental materials sampled showed significant differences in Mantel coefficients (Fig. S3; F4, 280 = 7.35, *P* < 0.001). Surprisingly, soils showed significantly lower coefficients than host, sediment, or water-associated coefficients (Tukey HSD; *P* < 0.01 in all cases).

Finally, concordant with our expectation, there was a significant, positive relationship between the (log) study scale and the Mantel coefficient (slope = 0.016, *P* < 0.001, adj-R2 = 0.12), showing that studies with larger spatial extents tend to find stronger correlations between community dissimilarity and geographic distance (Fig. 4). This relationship held when we examined only significant Mantel coefficients, and after accounting for sampling effort (slope = 0.016, *P <* 0.001, adj-R2 = 0.13). Sampling effort was not correlated with scale (Pearson’s R = 0.03, *P* = 0.64), showing that larger scale studies do not necessarily sample more.

**Discussion**

Two decades of research into the spatial ecology of microbial communities has resulted in a highly variable impression of the microbial distance-decay (d-d) relationship. Our meta-analysis of 287 microbial d-d relationships has revealed two main findings. Firstly, d-d relationships may be influenced by methodological choices, including the sampling depth used and the type of dissimilarity index. Secondly, as expected, the d-d relationship also appears to be dependent on various aspects of biological context, with different d-d relationships observed between different biomes and spatial extents.

The rapid development of methods in microbial ecology has improved our ability to detect and characterise ecological patterns in microbial communities, with high-throughput sequencing (HTS) platforms able to quantify microbial communities in ever increasing detail (Roesch *et al.* 2007; Caporaso *et al.* 2012). The tremendous sequencing depth of HTS platforms allows them to illuminate the “rare biosphere” (Sogin *et al.* 2006), thus elevating them over other approaches such as “fingerprinting”, which tend to capture a smaller proportion of the community. Initially, our results suggested that HTS-based approaches yielded similar strength d-d relationships to lower-resolution methods, such as fingerprinting, and lower throughput methods, such as Sanger sequencing, suggesting that the massive sequencing depths offered by HTS platforms are not necessary to capture these ecological patterns (van Dorst *et al.* 2014). However, when we examined only statistically significant d-d relationships, the relationships derived from HTS approaches were stronger than other approaches. The ability of different methods to alter the strength of the d-d relationship is expected for two reasons. Firstly, fingerprinting and HTS approaches capture microbial diversity at different taxonomic resolutions. Comparative approaches have shown that fingerprinting approaches such as ARISA may be comparable to HTS data at the phylum level for instance (Gobet *et al.* 2014). Fingerprinting methods are therefore limited in that they may not detect compositional differences between communities at increasingly fine taxonomic resolutions (Ramette & Tiedje 2007; Bissett *et al.* 2010). This may weaken the d-d relationship in instances where communities are similar at the family level, but dissimilar at finer taxonomic levels. Secondly, fingerprinting methods are less able to sample from the “rare biosphere”, unlike HTS approaches. This is significant as microbial communities often follow an occupancy-abundance relationship in which the most common organisms are also the most widespread, and the rarer organisms are the most restricted (Soininen & Heino 2005; Liu *et al.* 2015). Therefore, sampling only the most common, widespread organisms would flatten the d-d relationship by making communities appear artificially similar in composition. This is in contrast to recent studies which show that spatial turnover in communities is adequately reflected by “common species” (Heino & Soininen 2010). However, microbial communities are often enormously diverse and exhibit extremely “long tailed” species abundance distributions, such that the vast majority of microbial species in a community are “rare” (Hong *et al.* 2006; Galand *et al.* 2009; Locey & Lennon 2016). Therefore, it is likely that in microbial communities, common species alone may not adequately reflect patterns in spatial turnover (Galand *et al.* 2009).

Another methodological choice that was found to influence the strength of the microbial d-d relationship is the choice of dissimilarity index. Dissimilarity indices can vary in the type of data they consider (quantitative vs qualitative), the type of distance they quantify (compositional vs. phylogenetic), and the weight they place on common, rare, or absent species (Anderson *et al.* 2011). Within our study, we found significant differences in the d-d relationship between different indices, and between different index types. In particular, d-d relationship using phylogenetic indices were significantly flatter than compositional indices, whereas there was no difference between binary (presence/absence) and abundance based indices. Phylogenetic dissimilarity metrics may result in lower Mantel correlation coefficients for the same reason that fingerprinting methods do; because communities predominantly differ at fine taxonomic resolutions. This means that communities may share large proportions of phylogenetic branch lengths, but will differ in exact species or OTU composition. In contrast, community composition metrics give no weight to how related communities are at broader taxonomic levels. The result of this is that communities appear more similar than when compared with compositional dissimilarity metrics, resulting in flatter d-d relationships (and therefore lower Mantel coefficients). This effect might be exacerbated when all sampled communities are from environmentally similar sites, which select for particular taxonomic groups. For example, extremophilic habitats such as solar salterns, may be dominated by relatively few families, yet may be distinct at the OTU/species level (Zhaxybayeva *et al.* 2013).

Surprisingly, we observed no difference between quantitative and qualitative dissimilarity indices. This suggests that qualitative compositional differences between communities drive d-d relationships rather than quantitative changes in species composition and abundance. Previous studies that have applied both presence/absence-based and abundance-based indices suggest that they may be highly correlated in microbial communities (Martiny *et al.* 2011), and result in similar estimations of d-d relationships (e.g. Green *et al*. 2004, Glassman *et al*. 2015). Our analysis also revealed an overwhelming bias in the use of dissimilarity metrics to classic indices such as Bray-Curtis or Jaccard. Whilst these indices are undoubtedly amongst the most frequently used, not only in microbial ecology, but also in other areas of ecology, we wish to highlight several contemporary indices that may better suit the types of questions microbial ecologists ask. For instance, classic metrics do not take into consideration co-occurrence information present within the data. To this end, a new family of metrics have been defined which account for species co-occurrence as well as shared taxa (Schmidt *et al.* 2017). Additionally, many indices rely on equal sample sizes, and are sensitive to differences in species richness (Green & Bohannan 2006), with potentially confounding effects on d-d relationships (Baselga 2007). Chao *et al*. (2005) therefore extended classic indices such as Jaccard and Sørensen to take into account unseen species, and also to make them less sensitive to differential sample sizes, thus reducing the need for problematic post-sequencing sample normalisation (McMurdie & Holmes 2014). Finally, it is known that many indices merge true compositional turnover (replacement of species) and nestedness (whereby communities are subsets of one another). Fortunately, modified versions of classic indices such as Jaccard, Sorensen, and Bray-Curtis have been developed that allow partitioning of these processes, which should enable a more mechanistic understanding of the processes behind d-d relationships (Baselga 2010, 2013; Podani & Schmera 2011). We echo the call of Green and Bohanan (2006) for microbial ecologists to more carefully consider their use of dissimilarity metrics, especially now that many are implemented in popular analysis software such as R (e.g. Baselga and Orme 2012).

Whilst significant differences were found between different methodological approaches, we also found differences relating to the biological context of each study. Against our expectation, soil-based studies had weaker d-d relationships than studies using other environmental materials. Soils are relatively stable habitats, capable of maintaining significant environmental gradients over relatively small spatial scales. Therefore, we expected the combination of high habitat heterogeneity coupled with limited opportunity for dispersal to result in stronger d-d relationships. It is possible that the environmental gradients present in soils do not change linearly over geographic distance, for example if the similar environmental conditions are patchily distributed. Alternatively, microorganisms may be able to disperse through soils more effectively than previously thought, perhaps via association with other soil organisms (Warmink *et al.* 2011), or by dispersal of soil itself (Kellogg & Griffin 2006; Favet *et al.* 2013), or other passive means (Joung *et al.* 2017).

Originally, we expected that studies of aquatic microbial communities may show the weakest d-d relationships as riverine or oceanic hydrology may provide an effective dispersal mechanism, thus homogenising microbial communities and presenting more diffuse environmental gradients over larger spatial scales. Contrarily, we found that aquatic communities actually showed stronger d-d relationships indicating increased spatial turnover in aquatic microbial communities. Soininen *et al*. (2007) recorded similar distance-decay rates between terrestrial, marine and aquatic ecosystems, showing that biome-dependent d-d relationships may be a feature of microbial communities. Host-associated communities showed relatively strong, but variable d-d relationships. We suggest that this is caused jointly by the ecology of the host species, in combination with the degree of host specificity with the associated microbial community. For example, if the host is not dispersal limited, and associates with a large variety of microorganisms, then the d-d relationship may be relatively flat. However, if the host is dispersal limited, and associates with a very specific microbiome, the d-d relationship might be steeper. To develop our understanding of the macroecology of host-associated microbial communities, an interesting approach would be to compare microbial d-d relationships of sessile and motile hosts (which were excluded from this analysis), as incorporating the ecology of the host (e.g. movement, interactions, range size) would likely provide further explanatory power.

Finally, we also found a relationship between the strength of the d-d relationship and the spatial scale over which the study was conducted. Scale-dependent d-d relationships have previously been reported (Bissett *et al.* 2010; Martiny *et al.* 2011; Soininen *et al.* 2011), albeit with contrasting results. Our results are comparable to those of Martiny *et al*. (2011) and Soininen *et al*. (2011), who reported that d-d relationships for various microbial communities were generally steeper as greater spatial scales were incorporated. The scale dependence of this relationship may be explained by greater environmental heterogeneity in large scale studies, thus communities are subjected to different environmental filters, resulting in dissimilar communities. In combination with this, communities separated by very large geographic distances should have minimal dispersal between them, assuming microbial dispersal is linearly related to geographic distance. Alternatively, this observation may be a statistical artefact, caused by studies with very large spatial extents incorporating many zero similarity community comparisons (i.e. communities with no species in common), therefore biasing our quantification of the d-d relationship (Millar *et al.* 2011; Steinbauer *et al.* 2012). This point highlights that careful consideration is required in the statistical analysis of d-d relationships, especially when incorporating large geographic extents or highly dissimilar communities.

Despite its common use in the literature as evidence for neutral processes in microbial ecology, the d-d relationship alone does not provide evidence for neutral processes acting on microbial communities. As discussed previously, d-d relationships can arise from spatially autocorrelated environmental gradients as well as dispersal limitation (Nekola & White 1999). Furthermore, dispersal limitation itself is not solely a property of ecological neutrality. Dispersal limitation may be stochastic as predicted by neutral theory (Chave 2004), but also by asymmetric dispersal abilities between organisms (Salomon *et al.* 2010; Liu & Zhou 2011), thus violating the central tenet of neutral theory; that organisms are ecologically equivalent (Hubbell 2001). Thus, we suggest caution in attributing distance-decay relationships to either niche or neutral processes without further evidence, for example from examining species-abundance distributions (e.g. Dumbrell *et al*. 2010). However, this is not to say that examining distance-decay relationships is futile, as the relationship jointly reflects species turnover due to historical, environmental, and spatial factors, all of which are important factors to consider in studying biodiversity (Nekola & White 1999).

Moving beyond distance-decay relationships, focussing on other factors that influence the compositional similarity of microbial communities should provide interesting results. For example, quantifying the extent to which microorganisms differ in their dispersal abilities, and what traits are responsible for these differences may help to provide information on the biogeography of microorganisms at the population level, and given appropriate statistical approaches may allow us to predict the range size and habitat occupancy of different microbes. Furthermore, it is commonly assumed that the rate of dispersal (connectivity) is linearly related to the spatial distance between communities. However, given that array dispersal vectors that microbes are potentially able to move between habitat patches by, making this assumption is risky. Therefore, we are optimistic that the growing movement towards examining the role of connectivity *per se* (Declerck et al. 2013; Vannette et al. 2016), rather than using geographical distance as a proxy, will provide a fruitful direction for spatial microbial ecology. By modeling the dispersal process itself and accounting for topographical factors such as connectivity, these suggestions would lead to a more mechanistic understanding of the spatial ecology of microbial communities.

**References**

Anderson MJ, Crist TO, Chase JM *et al.* (2011) Navigating the multiple meanings of β diversity: A roadmap for the practicing ecologist. *Ecology Letters*, **14**, 19–28.

Baas Becking LGM (1934) *Geobiologie of inleiding tot de milieukunde*. W.P. Van Stockum & Zoon, The Hague, Netherlands.

Barreto DP, Conrad R, Klose M, Claus P, Enrich-Prast A (2014) Distance-decay and taxa-area relationships for bacteria, archaea and methanogenic archaea in a tropical lake sediment (J Waldenström, Ed,). *PLoS ONE*, **9**, e110128.

Baselga A (2007) Disentangling Distance Decay of Similarity from Richness Gradients: Response to Soininen *et al*. 2007. *Ecography*, **30**, 838–841.

Baselga A (2010) Partitioning the turnover and nestedness components of beta diversity. *Global Ecology and Biogeography*, **19**, 134–143.

Baselga A (2012) The relationship between species replacement, dissimilarity derived from nestedness, and nestedness. *Global Ecology and Biogeography*, **21**, 1223–1232.

Baselga A (2013) Separating the two components of abundance-based dissimilarity: Balanced changes in abundance vs. abundance. *Methods in Ecology and Evolution*, **4**, 552–557.

Baselga A, Orme CDL (2012) betapart: an R package for the study of beta diversity. *Methods in Ecology and Evolution*, **3**, 808–812.

Beck J, Holloway JD, Schwanghart W (2013) Undersampling and the measurement of beta diversity. *Methods in Ecology and Evolution*, **4**, 370–382.

Bissett A, Richardson AE, Baker G, Wakelin S, Thrall PH (2010) Life history determines biogeographical patterns of soil bacterial communities over multiple spatial scales. *Molecular Ecology*, **19**, 4315–4327.

Caporaso JG, Lauber CL, Walters W a *et al.* (2012) Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The ISME Journal*, **6**, 1621–1624.

Chao A, Chazdon RL, Colwell RK, Shen T-J (2005) A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecology Letters*, **8**, 148–159.

Chave J (2004) Neutral theory and community ecology. *Ecology letters*, **7**, 1–39.

Condit R, Pitman N, Leigh EG *et al.* (2002) Beta-diversity in tropical forest trees. *Science*, **295**, 666–669.

Declerck SAJ, Winter C, Shurin JB, Suttle CA, Matthews B (2013) Effects of patch connectivity and heterogeneity on metacommunity structure of planktonic bacteria and viruses. *The ISME journal*, **7**, 533–42.

van Dorst J, Bissett A, Palmer AS *et al.* (2014) Community fingerprinting in a sequencing world. *FEMS Microbiology Ecology*, **89**, 316–330.

Dumbrell AJ, Nelson M, Helgason T, Dytham C, Fitter AH (2010) Relative roles of niche and neutral processes in structuring a soil microbial community. *The ISME journal*, **4**, 337–345.

Favet J, Lapanje A, Giongo A *et al.* (2013) Microbial hitchhikers on intercontinental dust: catching a lift in Chad. *The ISME journal*, **7**, 850–867.

Finlay BJ (2002) Global dispersal of free-living microbial eukaryote species. *Science*, **296**, 1061–1063.

Finlay BJ, Fenchel T (2004) Cosmopolitan Metapopulations of Free-Living Microbial Eukaryotes. *Protist*, **155**, 237–244.

Galand PE, Casamayor EO, Kirchman DL, Lovejoy C (2009) Ecology of the rare microbial biosphere of the Arctic Ocean. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 22427–22432.

Glassman SI, Peay KG, Talbot JM *et al.* (2015) A continental view of pine-associated ectomycorrhizal fungal spore banks: A quiescent functional guild with a strong biogeographic pattern. *New Phytologist*, **205**, 1619–1631.

Glenn TC (2011) Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*, **11**, 759–769.

Gobet A, Boetius A, Ramette A (2014) Ecological coherence of diversity patterns derived from classical fingerprinting and Next Generation Sequencing techniques. *Environmental Microbiology*, **16**, 2672–2681.

Green J, Bohannan BJM (2006) Spatial scaling of microbial biodiversity. *Trends in Ecology and Evolution*, **21**, 501–507.

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

Hazard C, Gosling P, van der Gast CJ *et al.* (2013) The role of local environment and geographical distance in determining community composition of arbuscular mycorrhizal fungi at the landscape scale. *The ISME journal*, **7**, 498–508.

Heino J, Soininen J (2010) Are common species sufficient in describing turnover in aquatic metacommunities along environmental and spatial gradients? *Limnology and Oceanography*, **55**, 2397–2402.

Hong SH, Bunge J, Jeon SO, Epstein SS (2006) Predicting microbial species richness. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 117–122.

Hubbell SP (2001) *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton University Press.

Joung YS, Ge Z, Buie CR (2017) Bioaerosol generation by raindrops on soil. *Nature Communications*, **8**, 14668.

Kellogg CA, Griffin DW (2006) Aerobiology and the global transport of desert dust. *Trends in Ecology and Evolution*, **21**, 638–644.

Kivlin SN, Winston GC, Goulden ML, Treseder KK (2014) Environmental filtering affects soil fungal community composition more than dispersal limitation at regional scales. *Fungal Ecology*, **12**, 14–25.

Lajeunesse MJ (2016) Facilitating systematic reviews, data extraction and meta-analysis with the metagear package for R. *Methods in Ecology and Evolution*, **7**, 323–330.

Legendre P, Fortin M-J, Borcard D (2015) Should the Mantel test be used in spatial analysis?. *Methods in Ecology and Evolution*, **6**, 1239–1247.

Liu L, Yang J, Yu Z, Wilkinson DM (2015) The biogeography of abundant and rare bacterioplankton in lakes and reservoirs of China. *The ISME Journal*, **9**, 2068–2077.

Liu J, Zhou S (2011) Asymmetry in species regional dispersal ability and the neutral theory. *PLoS ONE*, **6**, e24128.

Locey KJ, Lennon JT (2016) Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences*, 201521291.

Martiny JBH, Eisen JA, Penn K, Allison SD, Horner-Devine MC (2011) Drivers of bacterial β-diversity depend on spatial scale. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 7850–7854.

McMurdie PJ, Holmes S (2014) Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS computational biology*, **10**, e1003531.

Millar RB, Anderson MJ, Tolimieri N (2011) Much ado about nothings: Using zero similarity points in distance-decay curves. *Ecology*, **92**, 1717–1722.

Muyzer G (1999) DGGE/TGGE a method for identifying genes from natural ecosystems. *Current Opinion in Microbiology*, **2**, 317–322.

Nekola JC, White PS (1999) The distance decay of similarity in biogeography and ecology. *Journal of Biogeography*, **26**, 867–878.

Podani J, Schmera D (2011) A new conceptual and methodological framework for exploring and explaining pattern in presence - absence data. *Oikos*, **120**, 1625–1638.

R Developement Core Team (2016) R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing, Vienna, Austria*.

Ramette A, Tiedje JM (2007) Biogeography: an emerging cornerstone for understanding prokaryotic diversity, ecology, and evolution. *Microbial ecology*, **53**, 197–207.

Roesch LFW, Fulthorpe RR, Riva A *et al.* (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. *The ISME journal*, **1**, 283–90.

Salomon Y, Connolly SR, Bode L (2010) Effects of asymmetric dispersal on the coexistence of competing species. *Ecology Letters*, **13**, 432–441.

Schmidt TSB, Matias Rodrigues JF, von Mering C (2017) A family of interaction-adjusted indices of community similarity. *The ISME Journal*, **11**, 791–807.

Sogin ML, Morrison HG, Huber JA *et al.* (2006) Microbial diversity in the deep sea and the underexplored “rare biosphere.” *Proceedings of the National Academy of Sciences of the USA*, **103**, 12115–12120.

Soininen J, Heino J (2005) Relationships between local population persistence, local abundance and regional occupancy of species: Distribution patterns of diatoms in boreal streams. *Journal of Biogeography*, **32**, 1971–1978.

Soininen J, Korhonen JJ, Karhu J, Vetterli A (2011) Disentangling the spatial patterns in community composition of prokaryotic and eukaryotic lake plankton. *Limnology and Oceanography*, **56**, 508–520.

Soininen J, Korhonen JJ, Luoto M (2013) Stochastic species distributions are driven by organism size. *Ecology*, **94**, 660–670.

Soininen J, McDonald R, Hillebrand H (2007) The distance decay of similarity in ecological communities. *Ecography*, **30**, 3–12.

Steinbauer MJ, Dolos K, Reineking B, Beierkuhnlein C (2012) Current measures for distance decay in similarity of species composition are influenced by study extent and grain size. *Global Ecology and Biogeography*, **21**, 1203–1212.

Vannette RL, Leopold DR, Fukami T (2016) Forest area and connectivity influence root-associated fungal communities in a fragmented landscape. *Ecology*, **97**, 2374–2383.

Warmink JA, Nazir R, Corten B, van Elsas . JD (2011) Hitchhikers on the fungal highway: The helper effect for bacterial migration via fungal hyphae. *Soil Biology and Biochemistry*, **43**, 760–765.

Wilkinson DM, Koumoutsaris S, Mitchell EAD, Bey I (2012) Modelling the effect of size on the aerial dispersal of microorganisms. *Journal of Biogeography*, **39**, 89–97.

Zhaxybayeva O, Stepanauskas R, Mohan NR, Papke RT (2013) Cell sorting analysis of geographically separated hypersaline environments. *Extremophiles*, **17**, 265–275.

**Tables and Boxes**

Table 1. Details of the five Web of Science search terms and, the number of hits. A Web of Science search history file is provided in the Supplementary Material.

|  |  |  |
| --- | --- | --- |
| Search | Search terms | Number of hits |
| 1 | TS = (biogeograph\*) AND TS = (bacteria\* OR archaea\* OR microb\* OR microorganism\*) | 1,872 |
| 2 | TS = (macroecolog\*) AND TS = (bacteria\* OR archaea\* OR microb\* OR microorganism\*) | 85 |
| 3 | TS = ("everything is everywhere") AND TS = (bacteria\* OR archaea\* OR microb\* OR microorganism\*) | 53 |
| 4 | TS = ("geographic distance") AND TS = (bacteria\* OR archaea\* OR microb\* OR microorganism\*) | 133 |
| 5 | TS = ("distance decay") AND TS = (bacteria\* OR archaea\* OR microb\* OR microorganism\*) | 107 |
| \* is used as a wildcard to allow searches to match multiple terms, e.g. microb\* could match “microbiome”, “microbial”, and “microbe” | | |

Box 1. Details of the explanatory variables extracted from each study.

|  |
| --- |
| *Community characterisation method*  This refers to the method used to quantify the species present in their sample and their abundances (if applicable). Each d-d relationship was categorised into either high-throughput sequencing (HTS; Pyrosequencing, Illumina, Ion Torrent, Pac-Bio), community fingerprinting (ARISA, TRFLP, DGGE, PhyloChip), or other (Sanger sequencing of clone libraries, morphological identification).  *Sampling depth*  This refers to the sequencing depth in sequence-based studies, or number of individuals counted in morphology-based studies. For sequencing studies, we recorded the number of sequences after rarefaction, or if this was not given, the average number of sequences per sample. As it is hard to quantify the sampling depth of fingerprinting approaches, we recorded these as NA and excluded them from analyses involving sampling depth.  *Sampling effort*  This variable represents the number of individual communities/samples used to formulate the d-d relationship.  *Dissimilarity index*  We recorded the dissimilarity index from which each d-d relationship was calculated. After these had been recorded, we categorised them as abundance based (Bray-Curtis, Horn-Morisita, Euclidean, Hellinger, Theta), binary (Jaccard, Raup-Crick, Sørensen, Simpson, βsim), or phylogenetic (Unifrac, Rao, β-mean nearest taxon distance, β-mean pairwise distance).  *Study taxon*  We categorised d-d relationships into broad taxonomic categories (Archaea, Bacteria, Eukarya, Fungi). If a d-d relationship was based on multiple taxa, then an appropriate category was added as necessary (i.e. Bacteria + Archaea).  *Scale*  We recorded scale as the maximum distance separating communities (in km). If this was not stated in the text or provided in supplementary material (e.g. in a geographic distance matrix), it was calculated from given geographic coordinates, or estimated from the d-d graph itself or from scaled maps, if no coordinates were provided.  *Biome*  We categorised d-d relationships based on their biome (agriculture, air, aquifer, indoor, coral, desert, dune, flower, forest, grassland, ice, lake, marsh, mine, ocean, paddy, river, sediment, sewer, sponge), reflecting the type of environment the communities occupied.  *Environmental material*  This variable represents the type of material that the sampled communities occupied. We categorised d-d relationships as air, host, sediment, soil, or water.  *P value*  As an additional comparison, we also recorded *P* values for d-d relationships where possible. We recorded unadjusted *P* values, and here use a global alpha value of 0.05 for simplicity, regardless of multiple tests conducted by each study. |

**Figures**

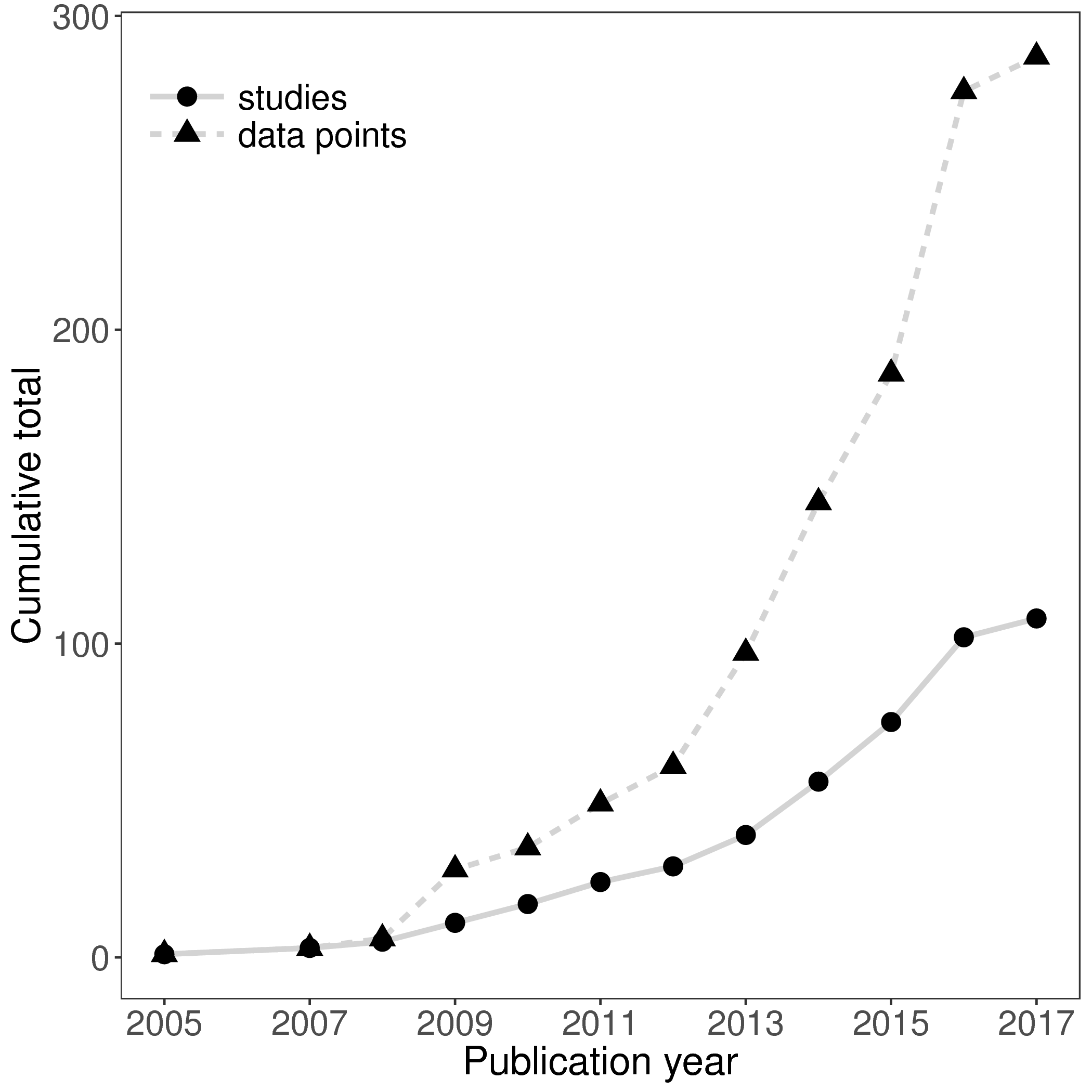
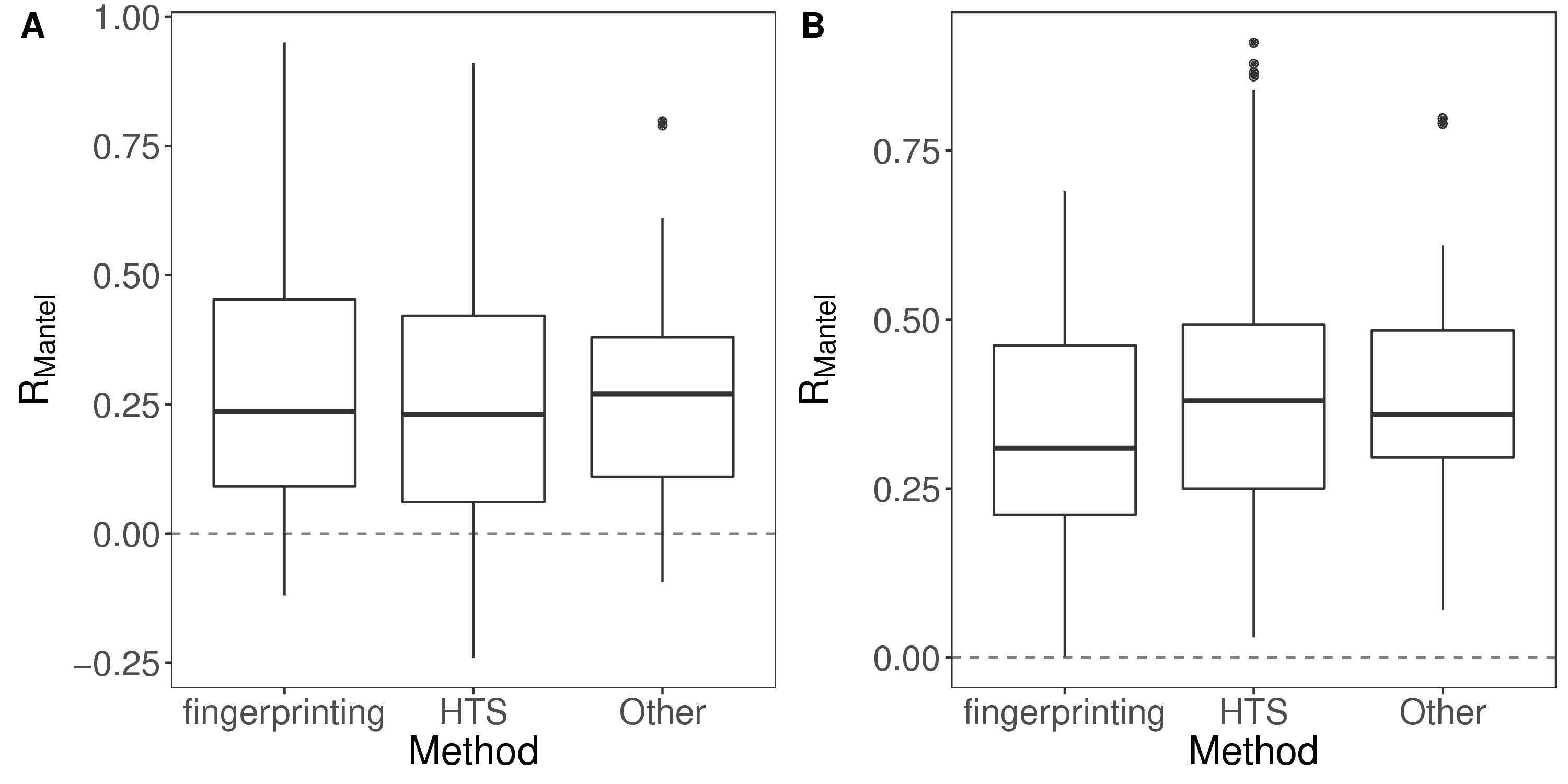


Figure 1. The cumulative number of distance-decay relationships and publications included in this study, through time.

Figure 2. (A) All, and (B) only statistically significant, Mantel correlation coefficients (RMantel) from studies based on high-throughput sequencing (HTS), community fingerprinting approaches (such as DGGE or TRFLP), or other low resolution/throughput methods (morphological identification, Sanger sequencing of clone libraries). Larger Mantel coefficients indicate stronger correlation between community dissimilarity and geographic distance.

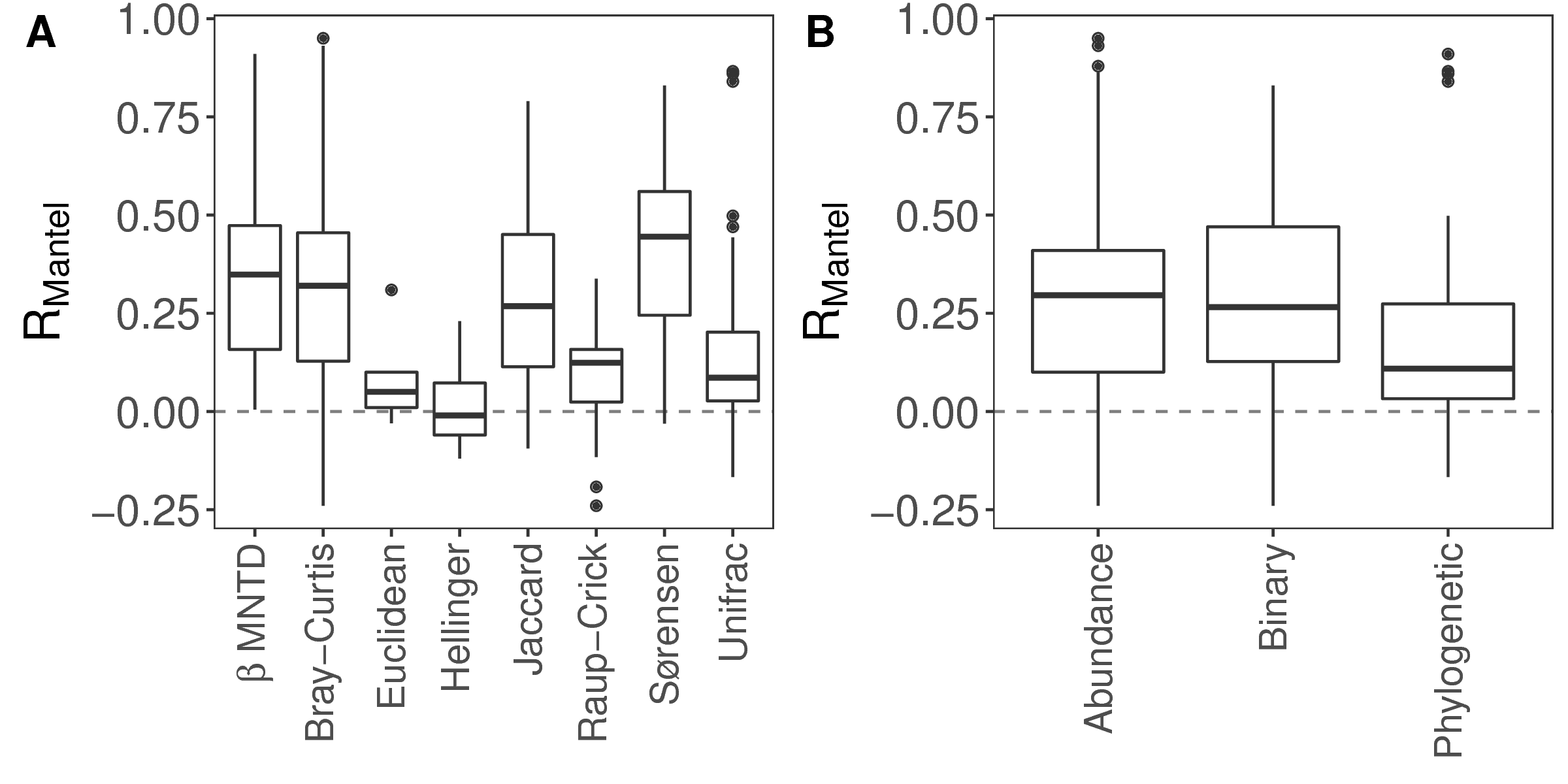


Figure 3. Mantel correlation coefficients from distance-decay relationships based on (A) different dissimilarity indices and, (B) different types of dissimilarity index. Index types reflect the different data requirements and type of distance (e.g. community composition or phylogenetic relatedness). Larger Mantel coefficients indicate stronger correlation between community dissimilarity and geographic distance.

Figure 4. The relationship between Mantel correlation coefficients and the geographic extent over which the distance-decay relationship was measured. The solid line shows the fit of a linear model (slope = 0.016, *P* < 0.001, adj-R2 = 0.12). The positive relationship indicates that larger scale studies tend to record stronger distance-decay relationships.