**Title:** Spatial Scaling Patterns of Microbial Community Composition are Context-Dependent

**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, Eukarya, distance-decay of similarity, Mantel test, macroecology, dispersal limitation, meta-analysis

**Running title:** Meta-Analysis of Microbial Distance-Decay Relationships

**Abstract**

According to the distance-decay (d-d) of similarity relationship, communities close together in space, are more similar in composition than those further apart. Recent molecular evidence suggests that microbial communities may show similar spatial patterns in β-diversity. However, observed microbial d-d relationships are highly variable, having been recorded from a variety of ecological settings, with a plethora of different methods. Therefore, we conducted a meta-analysis to understand whether the variability in microbial d-d relationships is caused by different study methodologies, or ecological contexts. We gathered data on 287 microbial d-d relationships and found that, both ecological context and methodology had significant impacts on the strength of microbial d-d relationships. Specifically, larger spatial extents increased the strength of d-d relationships, whilst differences also emerged between biomes and micro-habitats, with soils showing significantly weaker relationships than other micro-habitats. Predictably, community coverage was positively related to the strength of d-d relationships, whilst relationships based on phylogenetic indices of community similarity were weaker than those based on binary or abundance based indices. We conclude that microbial biogeographic patterns are dependent on ecological context, but may also be obscured or enhanced according to methodological choices. Finally, we offer methodological and conceptual guidance that may advance our understanding of microbial biogeographic patterns, such as the d-d relationship.

**Introduction**

The distance-decay (d-d) of community similarity is one of the most studied macroecological relationships (Nekola & White 1999; Soininen *et al.,* 2007). The relationship quantifies the decrease in compositional similarity between communities with increasing geographic distance, such that proximate communities are more similar than distant communities. D-d relationships arise through several different, but often interacting ecological processes and consequently, are of considerable interest to ecologists (Nekola & White 1999; Soininen *et al.,* 2007; Hanson *et al.,* 2012). Firstly, d-d relationships can form through spatially structured niche processes. Communities are often structured by the shared environmental niches of their component species. Consequently, in habitats where spatially structured environmental gradients occur, communities close together in space experience similar environmental conditions, thus selecting for similar species. Alternatively, d-d relationships may occur through neutral processes. Dispersal limitation enhances d-d relationships by limiting the connectivity between communities, whilst drift contributes to d-d through stochastic speciation and extinction, thereby facilitating more compositionally dissimilar communities.

Distance-decay relationships are well documented in a multitude of “macro-organisms”, yet they are of particular interest to microbial ecologists, as microorganisms posess several characteristics that may defy classic d-d relationships. Firstly, their small size facilitates passive dispersal over large geographic distances by vectors such as wind, bio-aerosolisation, or ocean currents (e.g. Bisson *et al.,* 2007; Favet *et al.,* 2013; Joung *et al.,* 2017). Additionally, microorganisms often maintain high population densities in the environment leading to dispersal by “mass effects”, whereby high dispersal rates from areas of increased population density maintain populations in less optimal environments (Shmida & Wilson, 1985). Finally, some microorganisms are able to enter vegetative states, such as cysts or spores, allowing them to survive and disperse through suboptimal environments. Combined, these properties make microorganisms effective dispersers, leading to the assumption that microorganisms are globally dispersed, and that microbial communities only show d-d relationships under niche processes in spatially autocorrelated environments (Baas Becking 1934; Finlay 2002). However, emprical tests of microbial d-d relationships have yielded mixed results. Many studies have detected little or no evidence of d-d relationships in microbial communities (Hazard *et al.,* 2013; Kivlin *et al.,* 2014), whilst others report relationships of varying steepness, across a range of spatial extents, study systems, and taxa (e.g. Dumbrell *et al.,* 2010; Martiny *et al.,* 2011; Clark *et al*., 2017). As a result, the generality of spatial patterns in microbial communities remains unclear.

The variation in d-d relationships could be due to different ecological contexts incorporated by studies. The study systems commonly of interest to microbial ecologists vary in terms of connectivity, thereby facilitating or hindering dispersal between communities. In well connected systems where dispersal is possible, such as oceanic waters, d-d relationships should be weaker than systems in which dispersal is limited, such as host-associated systems. Moreover, study systems differ in the environmental gradients they support. Soils for example, can support strong environmental gradients over distances of a few meters, leading to steep d-d relationships (e.g. Dumbrell *et al.,* 2010). In other systems, environmental conditions are less strongly spatially autocorrelated, leading to weaker d-d relationships. Additionally, different study organisms are likely to yield variable d-d relationships, particularly if we consider dispersal to be a trait mediated process. For example, small cell sizes lead to more efficient long distance dispersal (Wilkinson *et al.,* 2012; Norros *et al.,* 2014), thus organisms with smaller cell sizes, such as Bacteria and Archaea, should disperse further than microbial Eukarya, leading to weaker d-d relationships. Finally, it is known that spatial extent can influence our perception of ecological relationships, and may contribute to variable d-d relationships (Steinbauer *et al.,* 2012). Studies incorporating larger spatial extents may find stronger d-d relationships as dispersal is assumed to be negatively related to geographic distance. Furthermore, large scale studies will likely incorporate greater environmental heterogeneity, and thus niche based processes may enhance d-d relationships (Martiny *et al.,* 2011).

Whilst ecological context may contribute to variability in microbial d-d relationships, so too could differing methodologies. The rapid development of molecular tools to study microbial communities has revolutionised microbial ecology (Clark *et al*., 2018). However, as a result, our perception of microbial d-d relationships is based on methods that vary substantially in the coverage (ability to detect rare taxa) and resolution (ability to resolve closely related taxa) they offer (Muyzer 1999; Glenn 2011). Early methods such as clone library sequencing and community fingerprinting methods (e.g. DGGE and TRFLP) are limited in their ability to detect rare taxa (Bartram *et al.,* 2011), potentially missing endemic species. In turn, this could result in artificially similar community compositions, affecting the d-d relationship (Hanson *et al.,* 2012). In contrast, high-throughout sequencing (HTS) platforms offer massively improved community coverage, enabling the detection of rarer, and potentially endemic taxa, more accurately quantifying the compositional similarity between communities. Furthermore, the ability to resolve closely related taxa is an important factor in being able to accurately quantify similarity in community composition. Methods such as fingerprinting and morphological surveys may “lump” closely related taxa or are vulnerable to “cryptic” species respectively, both of which may result taxa appearing to have artificially wide distributions, increasing community similarity (Hanson *et al*., 2012; Bruns & Taylor 2016). Whereas, methods yielding DNA/RNA sequence data, such as clone library sequencing or high-throughput sequencing, are able to resolve taxa providing appropriate bioinformatic analyses are applied. Therefore, endemic species can be identified, and the compositional similarity between communities can be more accurately quantified. In addition to the varying quantification methods used by microbial ecologists, analytical methods could also influence d-d relationships. In particular, the choice of similarity index may be important. An array of indices are available to quantify the similarity in operational taxonomic unit (OTU; a sequence-similarity based pseudo-species definition) composition between communities, including qualitative (based on presence/absence of species e.g. Jaccard’s index), quantitative (based on composition and abundance of species e.g. Bray-Curtis), and phylogenetic indices (based on relatedness of communities e.g. Unifrac). These indices have different properties in terms of how they are influenced by sample sizes or species richness (Baselga 2012; Beck *et al.,* 2013), and especially in terms of what they quantify (e.g. phylogenetic similarity versus compositional similarity). For example, phylogenetic indices would be expected to yield weaker distance-decay relationships than other metrics, because communities can be phylogenetically closely related, yet dissimilar in OTU composition (e.g. Bryant *et al.,* 2008). On the other hand, quantitative indices are able to reflect more fine scale changes in community structure as they account for changes in the abundances of species, and thus should result in stronger d-d relationships as accounting for abundance changes should yield more dissimilar community compositions.

Our perception of fundamental ecological relationships in microbial communities, such as the d-d relationship, are therefore vulnerable to several potential sources of variability that may be ecological or artefactual. Yet, the influence of such factors on ecological relationships in microbial communities remains poorly quantified. Therefore, we sought to understand whether methodological or contextual differences between studies influence reported d-d relationships in microbial communities. To do this, we conduct a meta-analysis to synthesise available data on microbial d-d relationships, and test the effect of factors relating to methodology or ecological context on the strength of d-d relationships. Specifically, we test the following hypotheses:

* H1 Bacteria and Archaea will show weaker d-d relationships than other microbial taxa due to their smaller size and higher population densities in most environments.
* H2 Soils and host-associated study systems will show stronger d-d relationships than other systems, due to their ability to maintain steep physicochemical gradients or limited host range size respectively, whilst aquatic systems will have weaker d-d relationships due to the potential for increased connectivity between communities.
* H3 The spatial extent of a study will be positively related to the strength of a d-d relationship, as larger extent studies incorporate greater environmental heterogeneity and lower dispersal rates between communities.
* H4 Higher resolution community quantification methods, such as high-throughput sequencing, will yield stronger d-d relationships due to their ability to resolve closely related taxa.
* H5 Coverage (e.g. number of sequences, or number of individuals counted) will be positively related to the strength of d-d relationships as higher community coverage methods will capture more of the rare endemic taxa, thus decreasing community similarity.
* H6 Phylogenetic similarity metrics will result in weaker d-d relationships than other metrics as communities can be phylogenetically similar, yet different at the OTU level.

**Methods**

*Meta-Analysis*

In order to test our hypotheses, we first gathered available data on microbial d-d relationships via a systematic literature search. To do this, five search terms were designed to detect relevant studies (Table 1). All literature searches were conducted using the Web of Science search portal on 08/06/2017, and all results published between 1900-2017 were retained. To further filter the dataset to studies suitable for testing our hypotheses, search results were downloaded and manually screened using the “metagear” (version 0.4; Lajeunesse 2016) package in R (version 3.4.1; R Development Core Team 2016). Here, “suitable studies” were those that tested the relationship between community similarity and geographic distance in microbial communities, and not studies of “macroorganisms”, or studies of strain-level genetic distance (e.g. using multi-locus sequence typing).

From these studies, we extracted Mantel correlation coefficients as an effect-size measure of the d-d relationship. The Mantel test is used to test for correlation between two distance matrices (i.e. community dissimilarity and geographic distance), and the resulting correlation coefficient is an ideal effect size measure for several reasons. Firstly, the Mantel correlation test is the most frequently used method for testing the statistical significance of d-d relationships in microbial ecology (e.g. Ramette, 2007; Franklin & Mills, 2007). Secondly, as the Mantel coefficient is a standard correlation coefficient (i.e. is bound by -1 and 1), it provides an easily interpretable and comparable measure of effect size (Harrison, 2010).

Within the literature, community similarity is often quantified as a distance (e.g. dissimilarity, or 1 - similarity) and therefore, any correlation coefficients based on similarity, instead of dissimilarity, were multiplied by -1 so that the direction of correlation was consistent across studies. For clarity, here a Mantel correlation coefficient of 1 indicates a strong d-d relationship, 0 indicates a lack of correlation between community (dis)similarity, and -1 indicates a strong negative correlation. It is possible that Mantel correlation coefficients can be negative in cases where more distant communities are more similar to each other than neighbouring communities. There are ecological reasons why this might occur (e.g. frequent long distance-dispersal, or environmental heterogeneity), and so they are included in our analysis. Partial Mantel statistics (which test for correlation between two matrices whilst controlling for a third) were excluded as they are influenced by other variables included in the test, and are therefore not comparable between studies.

In order to test our hypotheses, several variables relating to the ecological context and methodology of each d-d relationship were recorded. Details of these variables are described in Box 1.

*Statistical Analyses*

In order to determine whether d-d relationships varied between categoric variables (as in hypotheses 1, 2, 4, and 6), we used ANOVA tests. In tests where significant differences between groups were found, Tukey’s Honest Significant Difference (HSD) tests were used to determine which groups were different. To test hypotheses 3 and 5, linear regressions were used to test relationships. The variables spatial extent and coverage were log transformed prior to aid model fitting, as they spanned several orders of magnitude.

**Results**

The 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 decreased to 2,031 hits. Manual screening of the abstracts yielded 547 studies that were deemed to be potentially suitable for use in this analysis. A total of 287 Mantel correlation coefficients were successfully obtained from 108 studies represented in 33 journals (Fig. 1). Of the 439 studies that were unsuitable for inclusion within this analysis, most dod not test for correlation between geographic distance and community (dis)similarity (although the abstract still contained the search terms), whilst others had used alternate methods (e.g. multilocus sequence typing on individual species, or spatial eigenvector analysis). Reported Mantel correlation coefficients ranged from -0.24 to 0.95, with a mean of 0.27 (std. error = 0.014).

*Influence of Biological Context on the Distance-Decay Relationship*

In order to determine whether different ecological contexts can influence the strength of d-d relationships, the influence of ecological factors including study taxa, study system, and spatial scale were tested. Within the dataset, the most commonly studied taxa were Bacteria, followed by Fungi, micro-Eukaryotes, and Archaea. No significant difference was found in the Mantel coefficients associated with each taxa (*F*5, 281 = 1.39, *P* = 0.23), in disagreement with our hypothesis (H1). Examining only statistically significant Mantel coefficients revealed marginally significant differences between taxa (*F*5, 172 = 2.51, *P* < 0.05) with studies incorporating both bacteria and fungi (*n* = 3) simultaneously, being significantly lower than studies on Archaea (Tukey HSD; *P* < 0.05).

The d-d relationships in our dataset originated from 20 different biomes. Of these, 11 had fewer than three d-d relationships, and so were excluded from biome analyses. The most frequently studied biomes were grasslands (*n* = 62), forest (*n* = 57), and lakes (*n =* 44). In partial agreement with our hypothesis that biomes would show different strength d-d relationships (H2), Mantel coefficients differed significantly between biomes (*F*8, 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. 3.2; *F*4, 280 = 7.35, *P* < 0.001). In disagreement with our hypothesis, soils showed significantly lower coefficients than host-associated, sediment, or water d-d coefficients (Tukey HSD; *P* < 0.01 in all cases), in contrast with H2.

Finally, concordant with our hypothesis that studies over larger spatial extents would show stronger d-d relationships (H3), there was a significant, positive relationship between the (log) spatial extent and Mantel coefficients (Fig. 3, slope = 0.016, *P* < 0.001, adj-*R2* = 0.12). To examine whether this relationship was due to larger spatial extents having greater sampling effort, we tested for correlation between spatial extent and sampling effort, and found no correlation (Pearson’s ρ = 0.03, *P* = 0.64). Additionally, when we included sampling effort as a model covariate alongside spatial extent, the relationship between Mantel coefficients and spatial extent was unaltered (slope = 0.016, *P <* 0.001, adj-*R2* = 0.13), indicating that the positive relationship found was not due to increased sampling effort in large spatial extent studies.

*Influence of Methodological Factors on the Distance-Decay 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. Our hypothesis that high-resolution methods would result in stronger d-d relationships was unsupported (Fig. 4A), as we found no difference between different resolution methods (F2, 284 = 0.47, *P* = 0.62). Even when only statistically significant d-d relationships were examined, the difference between different resolution methods only approached statistical significance (F2, 175 = 2.73, *P* = 0.07).

Community coverage was also significantly and positively related to the Mantel coefficient, albeit with a small effect size (slope = 0.02, *P* < 0.05, adj-*R2* = 0.02), supporting our hypothesis (H5) that greater coverage would result in stronger d-d relationships. Furthermore, coverage was not correlated to sampling effort (Pearson’s ρ = 0.03, *P* = 0.64), showing that studies with greater coverage did not necessarily incorporate more samples.

In support of our hypothesis that different (dis)similarity indices would result in different strength d-d relationships (H6), we found significant differences were detected between dissimilarity indices (*F*14, 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 central tendency (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. 5A), whilst Sørensen based coefficients were higher than Euclidean, Raup-Crick, and Unifrac indices (*P* < 0.01 in all cases, Fig. 5A). Furthermore, Mantel coefficients were significantly different between index types (Fig. 5B; *F*2, 284 = 5.41, *P* < 0.01), and Tukey HSD tests showed that Mantel coefficients based on phylogenetic distances were significantly lower than both abundance (*P* < 0.01) and binary based indices (*P* < 0.05), further supporting H6.

**Discussion**

Two decades of research into the spatial ecology of microbial communities has not yielded a consistent distance-decay relationship. Our meta-analysis of 287 microbial d-d relationships suggests the reasons for this lack of consistency are two-fold. Firstly, the disparity in ecological contexts between studies may explain the variability in d-d relationships. Here, d-d relationships were found to signficantly vary between study systems and spatial extent, although study taxon had no effect, suggesting that the ability of different study systems and spatial scales to alter the ecological processes that underlie d-d should not be disregarded. Secondly, methodological differences between studies, including dissimilarity index, data resolution, and sample coverage, were all found to significantly relate to observed d-d relationships. Therefore, these findings indicate that methodological differences between studies have the power to obscure general patterns in microbial ecology, inhibiting the detection of “macroecological” laws that may be general to microbial life.

The development of molecular methods 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). These methods differ in their resolution and coverage, which we hypothesised could obscure or enhance the detection of ecological patterns such as d-d relationships. Suprisingly, we detected no differences in d-d relationships between methods of differing resolutions. The ability to resolve closely related or cryptic taxa afforded by sequencing approaches allows researchers to quantify microbial diversity at fine taxonomic resolutions, which has previously been found to be an important determinant of our ability to detect biogeographical patterns (Hanson *et al.,* 2012). This is supported by empirical studies in which spatial patterns emerge only when taxa are defined at higher resolutions (e.g. Martiny *et al.,* 2008). Yet, other studies record biogeographical patterns that are robust to different taxonomic resolutions. For example, Clark *et al.* (2017) found that communities of halophilic Archaea showed no evidence of biogeographic regionalisation. Within the study, OTUs were defined at two different similarity thresholds (97 and 99% similarity), representing different taxonomic resolutions. Yet, d-d relationships and biogeographic patterns were highly similar for both OTU definitions. Therefore, the relative importance of resolution may depend on ecological context. For example, the benefits of high-resolution methods depend on the relatedness of the species present within the communities. If species are sufficiently different from each other, the differences between low- and high-resolution methods are minimised. Therefore, ecological context and methodology are unavoidably linked, meaning that methodological choice should be informed by ecology (Hanson *et al*., 2012).

Additionally, we found a positive relationship between coverage and the strength of microbial d-d relationships, albeit weaker than expected. Unlike HTS approaches, fingerprinting, clone library sequencing, and morphological surveys are unable to sample the “rare biosphere”. This is significant in microbial communities due to the extremely “long tailed” species abundance distribution they often exhibit (Hong *et al.,* 2006; Galand *et al.,* 2009; Locey & Lennon 2016). In practise, this means that the vast majority of microbial species in a community are rare. Furthermore, microbial communities often follow an “occupancy-abundance” relationship, whereby the most abundant species are also the most widespread and rare species tend to be restricted (Soininen & Heino 2005; Liu *et al.,* 2015). Consequently, missing the rare species in a community will miss the endemic species. As a result, only the widespread species will be quantified, meaning that communities will appear more similar to each other, resulting in a weaker d-d relationship. The ability of common species to reflect ecological patterns of the wider community is debated (Galand *et al.,* 2009; Heino & Soininen 2010, van Dorst *et al.,* 2014), but our results suggest that ecological signal may be enhanced with greater community coverage. Researchers should therefore ensure that their chosen methods are sufficiently able to detect rare, and possibly endemic, species in their chosen study communities.

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 weaker than compositional indices, whereas there was no difference between binary (presence/absence) and abundance based indices. Phylogenetic dissimilarity metrics may result in weaker d-d relationships because they do not adequately reflect the community similarity at fine enough taxonomic resolutions. For example, consider two microbial communities. At most taxonomic resolutions, these communities may be highly similar, being composed of similar microbial families or genera. Yet, it is entirely possible that they share very few OTUs in common. Quantitative and qualitative metrics would reflect this, yet phylogenetic metrics would be weighted towards the similarity between the two communities caused by their similarity at more coarse taxonomic resolutions. The result of this is that communities appear more similar when phylogenetic indices are used (Bryant *et al*., 2008), potentially 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, can be highly similar at broad taxonomic levels, yet distinct at the OTU/species level (Clark *et al*., 2017).

Against expectation, no difference was observed 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. In agreement with previous studies that have applied both binary and abundance based indices, these two measures of community similarity are likely to be highly correlated (Martiny *et al.,* 2011), and result in similar estimations of d-d relationships (e.g. Green *et al*., 2004, Glassman *et al*., 2015). This analysis also revealed that classic dissimilarity metrics, such as Bray-Curtis or Jaccard's index, are overwhelmingly the most frequently used in studies of microbial d-d relationships. These indices are undoubtedly the most frequently used, not only in microbial ecology, but also the wider field of ecology. However, we draw attention to several contemporary indices that may better suit the types of questions microbial ecologists ask, as well as the types of data we generate. 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 that account for species co-occurrences as well as shared taxa (Schmidt *et al.,* 2017), thereby informing us of the role of co-occurrence in β-diversity patterns. 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 account for unobserved species, and to make them less sensitive to variable sample sizes, reducing the need for post-sequencing normalisation of sample sizes (McMurdie & Holmes 2014). Furthermore, “fuzzy logic” has been applied to similarity indices in order to reduce the impact of false absence in community data (Barbosa 2015), which may be helpful in lower coverage studies. Finally, many indices are known to merge compositional turnover (replacement of species) and nestedness (whereby communities are subsets of one another), thereby blurring the underlying ecological processes. To combat this, modified versions of classic indices such as Jaccard, Sorensen, and Bray-Curtis have been developed, allowing the partitioning of community similarity metrics into their turnover and nestedness components (Baselga 2010, 2013; Podani & Schmera 2011). We echo the call of Green and Bohanan (2006) for microbial ecologists to exercise more care in their choice of dissimilarity metrics, especially now that many are implemented in popular and freely accessible 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 ecological context of each study. Against our hypothesis, soil-based studies had weaker d-d relationships than studies of other microhabitats. Soils are relatively stable habitats, in that they maintain physical structure and are therefore capable of maintaining significant environmental gradients over relatively small spatial scales (Dumbrell *et al*., 2010). Therefore, we expected the combination of high habitat heterogeneity coupled with limited opportunity for dispersal to result in stronger d-d relationships than for example, oceanic waters, where physicochemical gradients are more diffuse. 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, soil microorganisms may be able to disperse more effectively than previously thought, perhaps via association with other soil organisms (Warmink *et al.,* 2011), migratory species such as birds (Bisson *et al*., 2007), wind blown soil particles (Favet *et al.,* 2013), or via bioaerosols (Joung *et al.,* 2017). Furthermore, soils may harbour a more extensive microbial “seed bank” of dormant organisms that could weaken the d-d relationship (Lennon & Jones, 2011). This is because dormant cells are not subject to environmental selection. Thus in systems where the spatial structure of the environment is causing the d-d relationship, the microbial seed bank may diminish the effects of environmental selection on microbial communities, especially as distinguishing dormant from active cells in complex environmental samples can be challenging (Emerson *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, homogenising microbial communities and presenting more diffuse environmental gradients over larger spatial scales. In contrast, 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 specific microbiome, the d-d relationship will 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 (motile host-associated d-d relationships were excluded in this analysis), thereby elucidating the role of the host’s ecology (e.g. movement, interactions, range size) in determining microbial biogeographic patterns.

Finally, we also found a relationship between the strength of the d-d relationship and the spatial extent at 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 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 more dissimilar communities. In combination with this, communities separated by very large geographic distances should have minimal dispersal between them, assuming connectivity is negatively related to geographic distance. Alternatively, this result may be a statistical artefact, caused by studies with 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.

The d-d relationship is frequently interpreted as evidence for neutrality in the microbial literature. However, 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 can also be a trait-based process, (Salomon *et al.,* 2010; Liu & Zhou 2011), thus violating the central tenet of neutral theory; that organisms are ecologically equivalent (Hubbell 2001). Therefore, it is worth reiterating the tenet of Hanson *et al*. (2010); that a d-d relationship alone is insufficient to assign microbial community assembly to either niche or neutral processes without further evidence, for example from examining species-abundance distributions (e.g. Dumbrell *et al*., 2010). That is not to say that examining d-d relationships is futile. The d-d relationship jointly reflects species turnover due to historical, environmental, and spatial factors, all of which are important to consider in studying the spatial-scaling of 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 the traits responsible for these differences may help to provide information on the biogeography of microorganisms at the population level. Given appropriate statistical approaches, this may allow us to predict the range size and habitat occupancy of specific microbes. Furthermore, it is commonly assumed that the connectivity between communities is linearly related to the spatial distance between communities. However, given that the relevant dispersal vectors may differ profoundly in the spatial scales they operate over, this assumption should be tested. Therefore, 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 likely provide a fruitful direction for spatial microbial ecology. By modeling the dispersal process itself and accounting for connectivity, a more mechanistic understanding of the spatial ecology of microbial communities could be gained.

**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.

Barbosa AM (2015) fuzzySim: applying fuzzy logic to binary similarity indices in ecology. *Methods in Ecology and Evolution*, **6**, 853-858.

Bartram AK, Lynch MD, Stearns JC, Moreno-Hagelsieb G, Neufeld JD (2011) Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end Illumina reads. *Applied and Environmental Microbiology*, **77**, 3846-3852.

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.

Bisson IA, Marra PP, Burtt EH, Sikaroodi M, Gillevet PM (2007) A molecular comparison of plumage and soil bacteria across biogeographic, ecological, and taxonomic scales. *Microbial Ecology*, **54**, 65-81.

Bruns TD, Taylor JW (2016) Comment on “Global assessment of arbuscular mycorrhizal fungus diversity reveals very low endemism”. *Science*, **351**, 826-826.

Bryant JA, Lamanna C, Morlon H, Kerkhoff AJ, Enquist BJ, Green JL (2008) Microbes on mountainsides: contrasting elevational patterns of bacterial and plant diversity. *Proceedings of the National Academy of Sciences USA*, **105** (Supplement 1), 11505-11511.

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.

Clark DR, Mathieu M, Mourot L, Dufossé L, Underwood JC, Dumbrell AJ, McGenity TJ (2017) Biogeography at the Limits of Life: Do Extremophilic Microbial Communities Show Biogeographic Regionalisation? *Global Ecology and Biogeography,* **26**, 1435–1446.

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**, 533–542.

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.

Emerson JB, Adams RI, Román CMB *et al.* (2017) Schrödinger’s microbes: Tools for distinguishing the living from the dead in microbial ecosystems. *Microbiome*, **5**, 86.

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.

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 USA*, **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.

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.

Hanson CA, Fuhrman JA, Horner-Devine MC, Martiny JB (2012) Beyond biogeographic patterns: processes shaping the microbial landscape. *Nature Reviews Microbiology*, **10**, 497-506.

Harrison F (2011) Getting started with meta‐analysis. *Methods in Ecology and Evolution*, **2**, 1-10.

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 USA*, **103**, 117–122.

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

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

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.

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 USA*, 201521291.

Martiny AC, Tai AP, Veneziano D, Primeau F, Chisholm SW (2009) Taxonomic resolution, ecotypes and the biogeography of Prochlorococcus. *Environmental Microbiology*, **11**, 823-832.

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 USA*, **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.

Norros V, Rannik Ü, Hussein T, Petäjä T, Vesala T, Ovaskainen O (2014) Do small spores disperse further than large spores?. *Ecology*, **95**, 1612-1621.

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 (2007) Multivariate analyses in microbial ecology. *FEMS Microbiology Ecology*, **62**, 142-160.

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.

Shmida AVI, Wilson MV (1985) Biological determinants of species diversity. *Journal of Biogeography*, **12** 1-20.

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

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.

**Tables and Boxes**

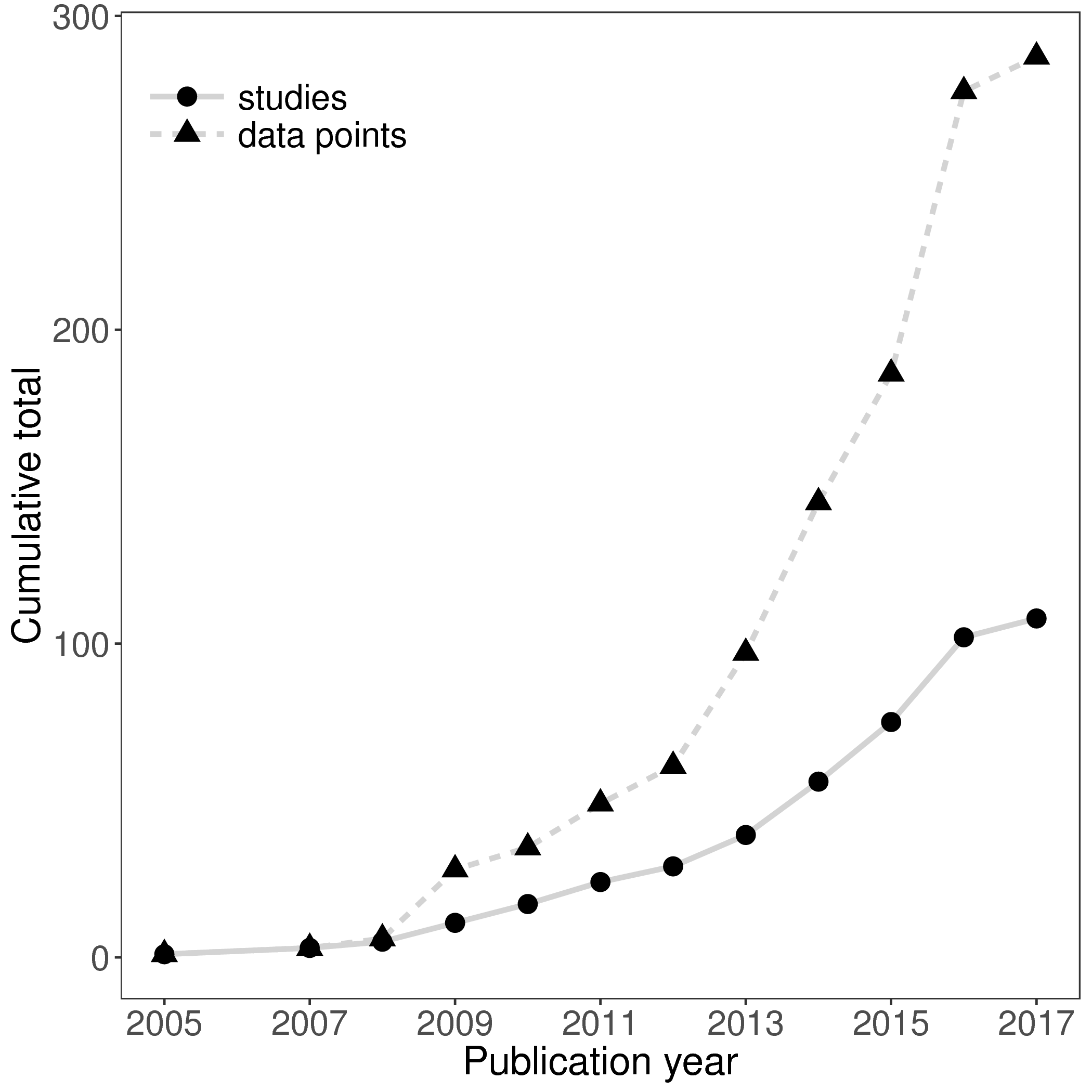
Table 1. Details of the five Web of Science search terms and, the number of hits. A Web of Science search history file is provided in the Supplementary Material.

|  |  |  |
| --- | --- | --- |
| 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” | | |

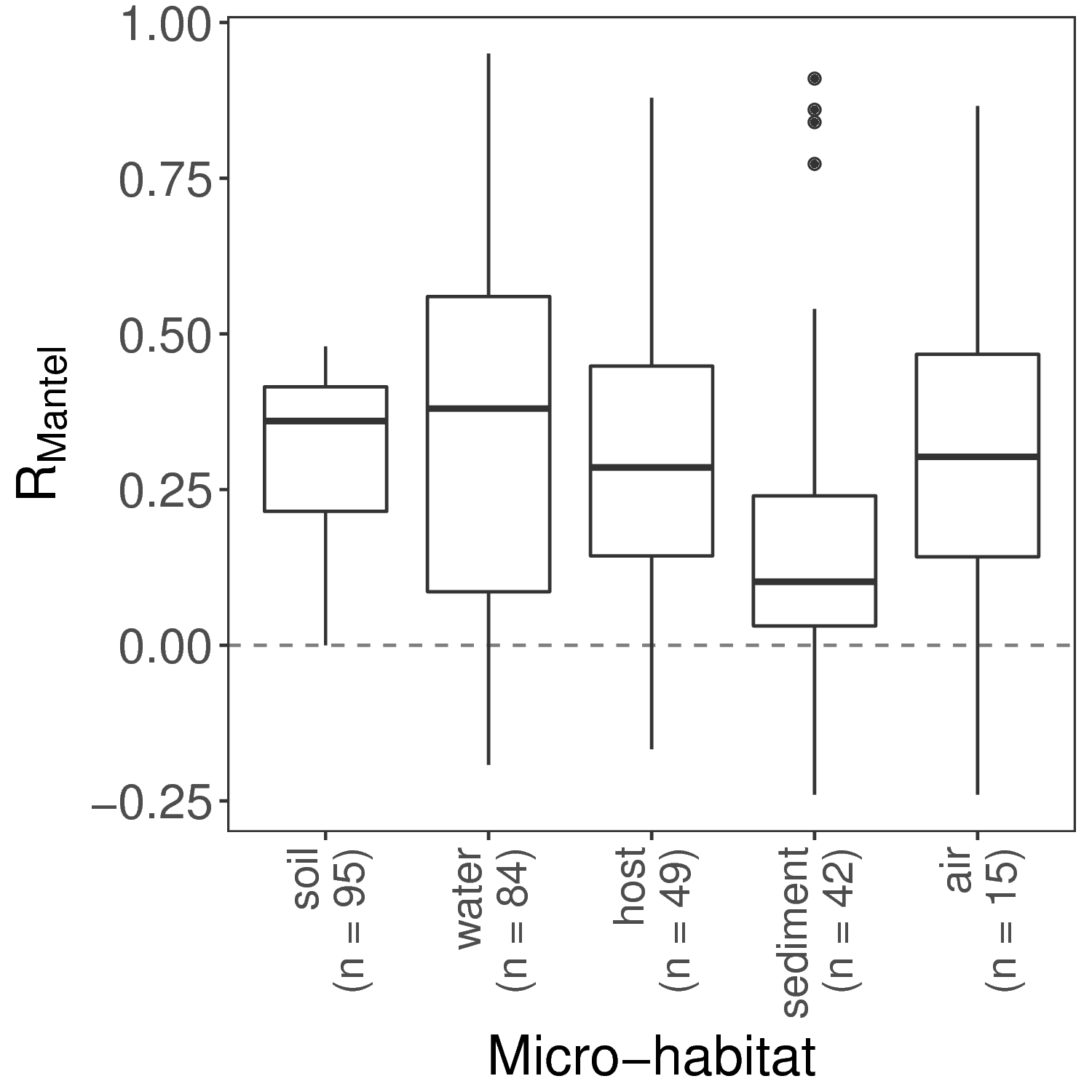
Box 1. Details of the explanatory variables extracted from each study.

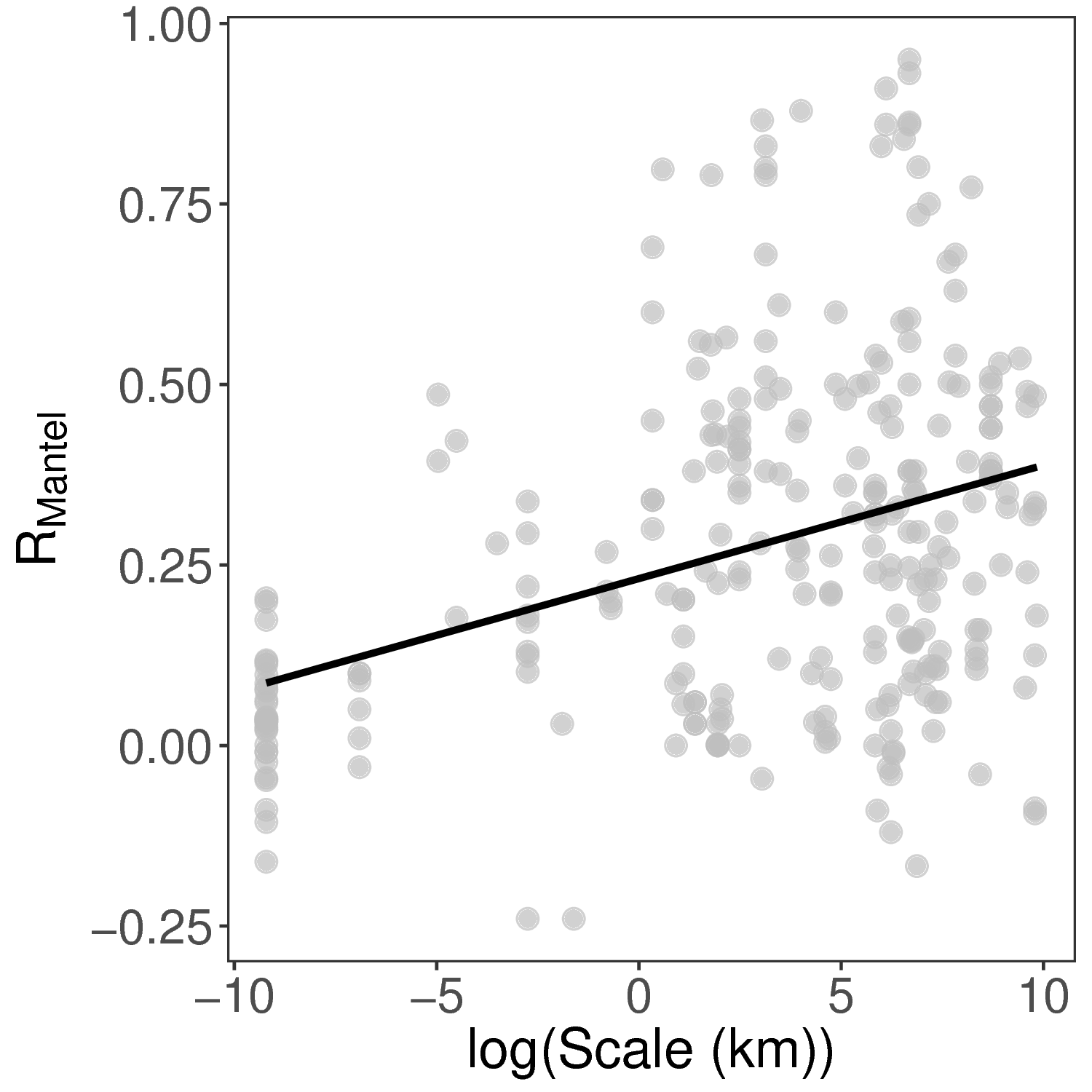
|  |
| --- |
| *Resolution*  This refers to the resolution of the method used to quantify the community. Each d-d relationship was categorised into either high-resolution (high-throughput or Sanger sequencing), intermediate (community fingerprinting e.g. ARISA, TRFLP, DGGE, PhyloChip), or low resolution (morphological identification).  *Coverage*  This refers to the sequencing depth in sequencing based studies, or number of individuals counted in morphological based studies, per sample. 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 there is no comparable measure of coverage for fingerprinting studies, we recorded these as NA and excluded them from analyses involving coverage.  *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 used to calculate each d-d relationship. Recorded dissimilarity indices were then categorised as quantitative (Bray-Curtis, Horn-Morisita, Euclidean, Hellinger, Theta), qualitative (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).  *Spatial extent*  We recorded extent as the maximum distance separating communities (in km). If this was not stated in text or provided in supplementary material (e.g. in a geographic distance matrix), it was calculated from given geographic coordinates, estimated from the d-d graph itself, or estimated from scaled maps.  *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.  *Micro-habitat*  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 significance threshold of 0.05 for simplicity, regardless of multiple tests conducted by each study.  *Mantel coefficient*  Mantel coefficients quanitfy the correlation between community dissimilarity and geographic distance. Values of 1 indicate a strong d-d relationship with dissimilarity increasing with distance, whereas values close to 0 indicate a lack of correlation. |

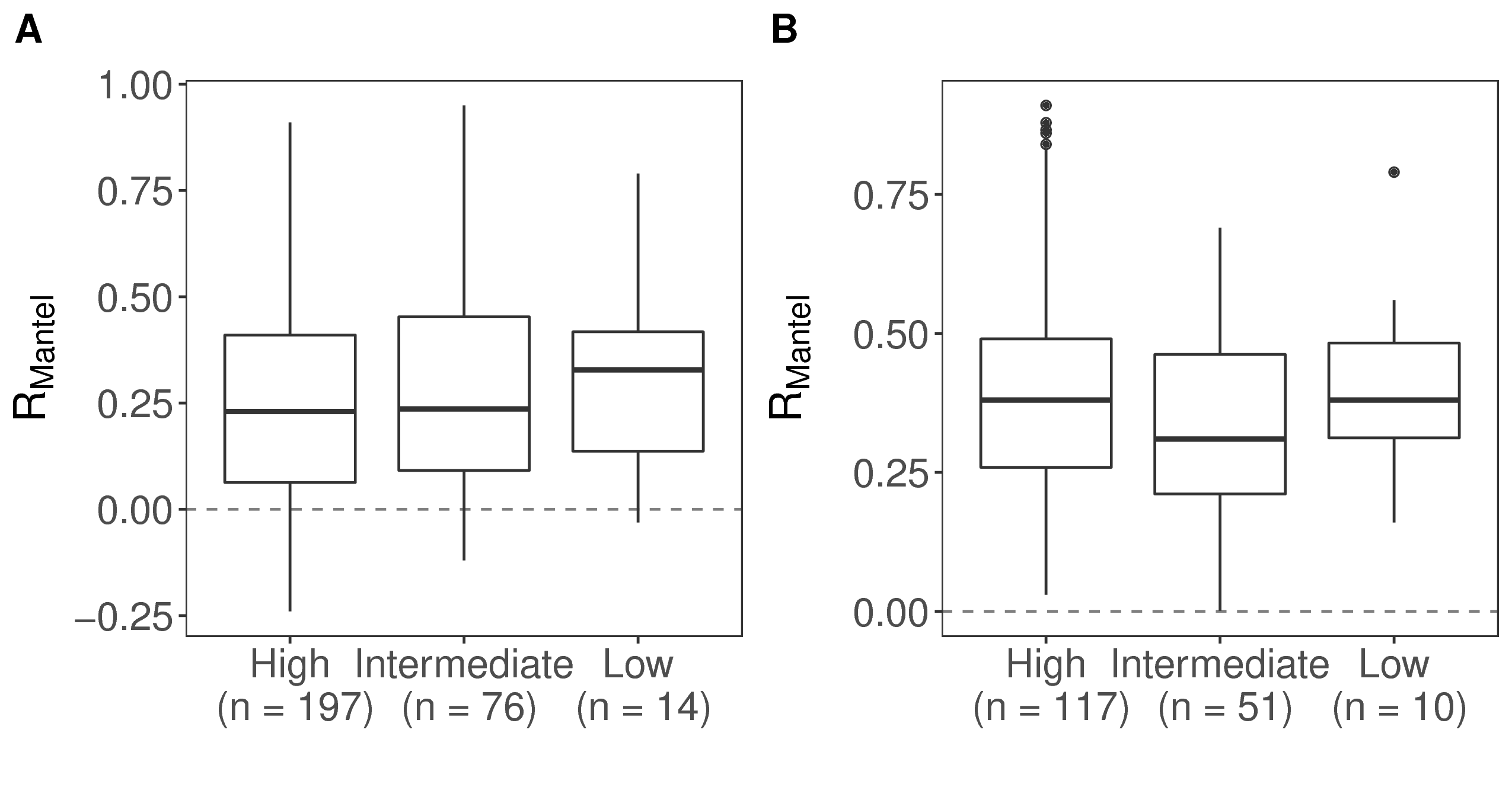
**Figures**

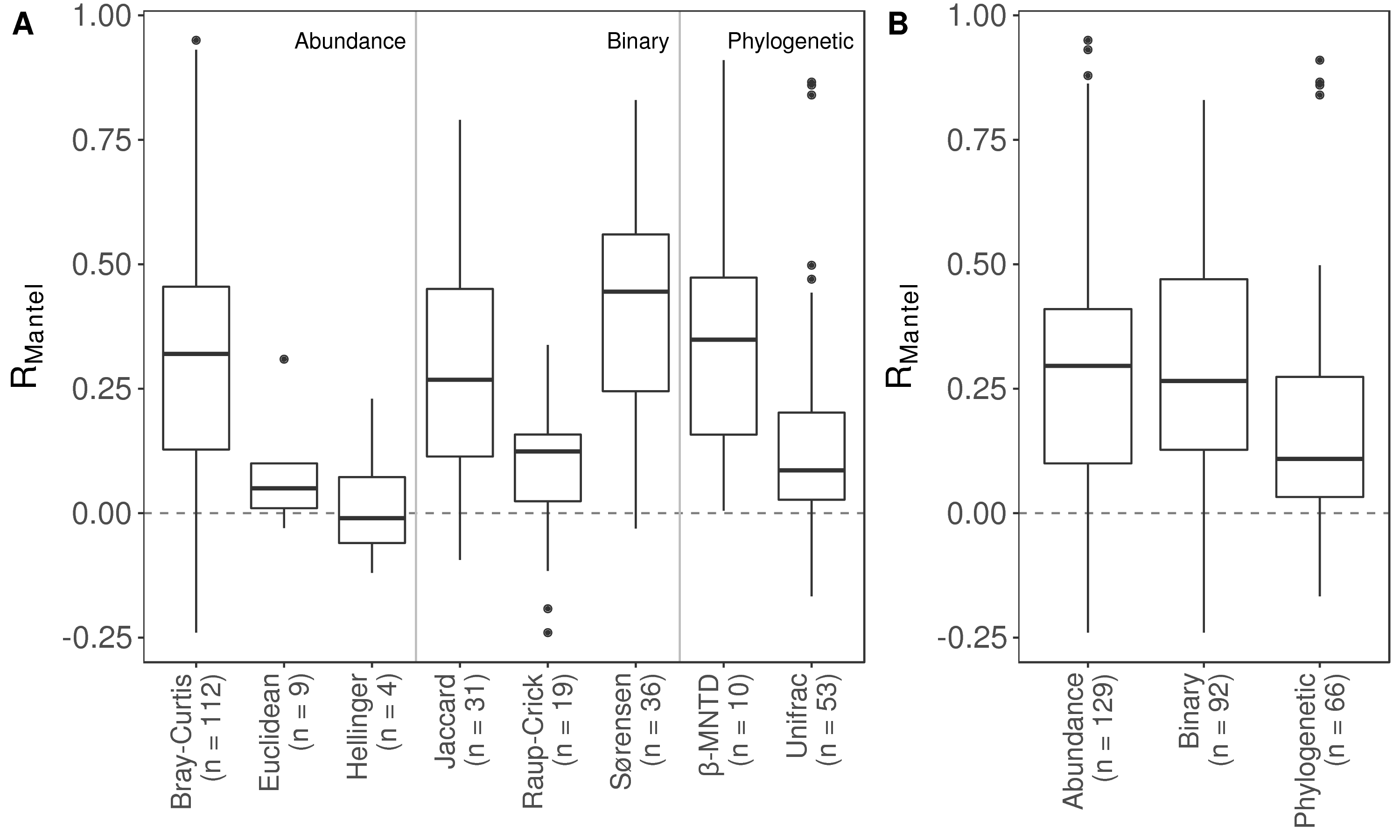
****

**Figure 1** The cumulative number of distance-decay (d-d) relationships and publications included in this study, through time. Studies referes to the number of publications in which microbial d-d relationships are tested, whilst data points refers to the number of d-d relationships reported.

**Figure 2** Mantel correlation coefficients from microbial communities sampled from different micro-habitats. Larger Mantel coefficients indicate stronger distance-decay relationships.

**Figure 3** 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.

**Figure 4** (A) All, and (B) only statistically significant, Mantel correlation coefficients (RMantel) from studies based on high-resolution (high-throughput or Sanger sequencing), intermediate (community fingerprinting approaches such as DGGE or TRFLP), or low resolution methods (morphological identification). Larger Mantel coefficients indicate stronger distance-decay relationships.

**Figure 5** 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.