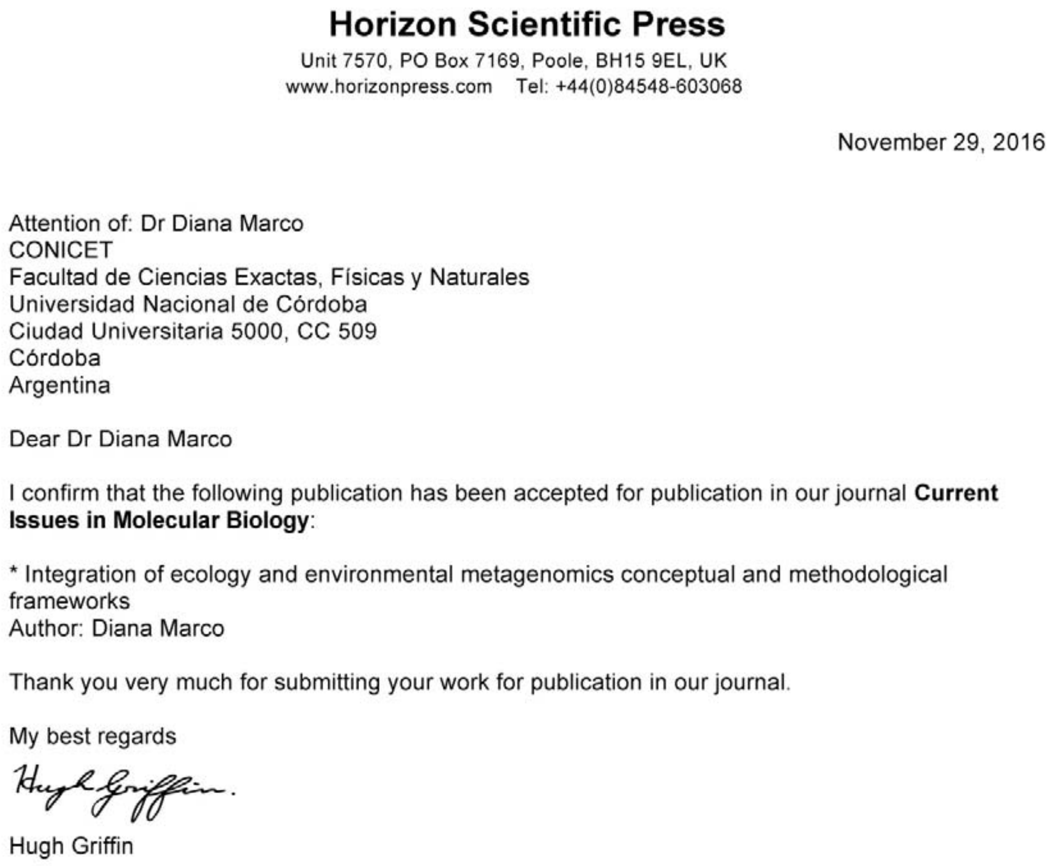

Horizon Scientific Press
Unit 7570, PO Box 7169, Poole, BH15 9EL, UK
www.horizonpress.com Tel: +44(0)84548-603068

November 29, 2016

Attention of: Dr Diana Marco
CONICET
Facultad de Ciencias Exactas, Físicas y Naturales
Universidad Nacional de Córdoba
Ciudad Universitaria 5000, CC 509
Córdoba
Argentina

Dear Dr Diana Marco

I confirm that the following publication has been accepted for publication in our journal **Current Issues in Molecular Biology**:

* Integration of ecology and environmental metagenomics conceptual and methodological frameworks
Author: Diana Marco

Thank you very much for submitting your work for publication in our journal.

My best regards

Hugh Griffin.

Hugh Griffin

Integration of ecology and environmental metagenomics conceptual and methodological frameworks

Diana Marco

Faculty of Biological Sciences, Cordoba National University, and CONICET, Av. Velez Sarsfield 1611, CP 5000, Córdoba, Argentina

Running title: Integrating ecology and metagenomics concepts and methodology

***E-mail:** dmarco@agro.unc.edu.ar

Abstract

Although from its origin metagenomics was concerned with composition of communities of microbial OTUs (Operational Taxonomic Units) living in a given habitat and their diversity and functional heterogeneity (concepts already well rooted in ecology), the new field was more "environmentally" than "ecologically" oriented. Probably by circumstantial reasons, metagenomics and ecology followed rather independent trajectories and conceptual and methodological gaps appeared. Recently, calls for the need of integrating the theoretical basis and methodologies coming from metagenomics (and other meta-omics) and ecology have been made. Here I will address some of the principles and methods of field ecology that, although useful in the context of environmental metagenomic studies, have been rather disregarded. In particular, I will emphasize the contribution of some well established concepts and methods of field ecology to a an appropriate field sampling and experimental design of environmental metagenomic studies

Introduction

The early beginning of metagenomics can be traced back to 1980 decade. The pioneering work of Pace et al. using ribosomal RNA to study natural microbial populations without cultivation (Pace et al. 1985), the work from Woese and Fox (1977), proposing the usage of ribosomal RNA as a tool for establishing phylogenetic relationships among microbial kingdoms, and the formal definition of metagenome by Handelsman and colleagues in 1998 (Handelsman et al., 1998) are the main milestones in the development of one of the most innovative scientific fields in the last decades. Since then, the number of works involving metagenomics and other meta-omics has grown exponentially, as a direct consequence of the advantages arising from the new ability of accessing microbial information bypassing cultivation, and the development of new and cheaper sequencing techniques. The variety of habitats explored with metagenomics and other meta-omics has also increased exponentially, with virtually no limits in the diversity of samples taken, from field to microcosm experiments through organism-specific microbiomes. In parallel, the number of useful applications of metagenomics and meta-omics studies has also increased greatly, from agriculture to medicine passing through obtention of bioproducts like new enzymes.

A typical pipeline for metagenomic studies (Figure 1) begins with the delimitation of the microbiome of interest, the field and/or experimental sampling design, and the extraction of the genetic material.

(Fig. 1)

Extracting the desired material from an ever increasing range of metagenomic samples involves developing new and suitable methodologies. In the case of metagenomics, obtaining the metagenomic data requires more and more sophisticated and cheaper sequencing methods, and even sequencing-free strategies are being developed. But at present, one of the biggest challenges lies on the next steps of organising, classifying, analysing and interpreting the vast amount of data generated by metagenomis and meta-omics. The other great challenge, and perhaps the most disregarded, is however at the very beginning of the pipeline. While new statistical and bioinformatic techniques to treat the increasing amount of data produced are continuously appearing (see Odintsova et al. and Sudarikov et al. in this volume), the matter of how to get reliable data from an adequate sampling either from field, microcosm or other types of habitats is still largely overlooked. Several authors have drawn attention to this aspect during the last years, especially on the need of statistically adequate replicates for metagenomic studies (Prosser, 2010; Fierer et al., 2012; Knight, 2013; Creer et al., 2016), although with some controversy (Lennon, 2011).

This call for adequate, replicated sampling designs and the subsequent controversy may be related to the very beginning of the metagenomics approach, from the molecular biology field applied to microbial genomics (Pace et al., 1985; Woese and Fox, 1997) to the challenge of linking the genomic information with the organism or ecosystem from which the DNA was isolated (Handelsman, 2004). Although from its origin metagenomics was concerned with composition of communities of microbial OTUs (Operational Taxonomic Units) living in a given habitat and their diversity and functional heterogeneity (concepts already well established in ecology), the new field was more "environmentally" than "ecologically" oriented (O´Malley and Dupré, 2009). By the time metagenomics emerged as a new field, ecology was an already well established discipline with a high degree of formalisation and a powerful theoretical and methodological background. However, probably by circumstantial reasons, the two disciplines followed rather independent trajectories and conceptual and methodological gaps appeared. "Environmental genomics", "microbial population genomics", "ecogenomics", were some of the new terms coined to refer to metagenomics (DeLong, 2004), all of them resembling or alluding to ecological concepts. Some well-defined ecological concepts like "biodiversity" or "niche" were adopted in metagenomics studies, although with different meanings or interpretations to those already established in ecology. In particular, the term "niche" began to be used in metagenomic and environmental genomics studies, in spite of its original definition in ecology, centred on the traditional concept of species (Marco, 2008), and it is still being applied without further revision. In another, methodological example, the usage of statistical multivariate analyses, appropriate for multivariate data common in metagenomics and ecology, was still in its infancy in metagenomics as late as the beginning of 21st the century (Ramette, 2007), while they constituted one of the most used statistical methods in ecology since the middle of 20th century (Goodall, 1954). Thus, among others, the before mentioned controversy over the matter of taking replicates for metagenomic studies clearly appears as a consequence of the parallel trajectories followed by environmental genomics and ecological fields.

Recently, a call to environmental sequencing studies to adhere to robust ecological study design, allowing for an adequate number of sites/replicates to provide statistical power, as well as ensuring the collection of a robust set of environmental metadata (e.g. climate variables, soil pH) has been made (Creer 2016). Clearly, there is a strong need of integrating the theoretical basis and methodologies coming from metagenomics (and other meta-omics) and ecology, and early it was recognised that metagenomics´ power would be realized when it is integrated with classical ecological approaches (Reisenfield et al., 2004).

Here I will address some of the principles and methods of field ecology that, although useful in the context of environmental metagenomic studies, have been in my opinion rather disregarded. In particular, I will emphasize the contribution of some well established concepts and methods of field ecology to a an appropriate field sampling and experimental design of environmental metagenomic studies. For space reasons I will not extensively address here other "meta-omics" approaches, although clearly for metatranscriptomics, metaproteomics, metametabolomics, lipidomics, and other emerging approaches (Meiring et al., 2011), the considerations made here about metagenomics studies are, with some caveats, amply valid. I will refer mainly to soil studies for examples since soil is one of the habitats where environmental metagenomics is firstly showing integration with ecological concepts and methods.

Some concepts and methods of field ecology useful in the context of environmental metagenomic studies

Looking for composition and function

Metagenomic studies usually have two main purposes, asking who is there? (composition approach), or what are they doing? (functional approach). The first approach aim to answer questions about OTUs/genes like phylogenetic relationships, community structure (composition and relative abundances), diversity, etc. The second approach is oriented to the study of genes performing specific functions. The two approaches may be assimilated to the proposed classification of metagenomic studies into "open" and "closed" formats (Zhou et al. 2015). The "open" format does not require a previous knowledge of the metagenomic community, and is more used in exploratory studies of composition and diversity, but allowing for gene discoveries (that may be later related to functions). Massive sequencing techniques are the most conspicuous methodologies used in this approach. The "closed" format, on the contrary, is focused on already known genes performing functions of interest, and their detection is for example performed by functional gene arrays.

In the same way, in ecological studies the focus may be on community structure (species composition and abundance) and diversity (commonly, α, within community and β, between communities), or on function, through the definition of functional guilds (groups of species performing similar functions) (Simberloff and Dayan, 1991). The functional approach in ecology is mainly based on functional traits of the species in a community allowing to group them in guilds or functional groups (Wilson, 1999), for example, birds with similar beak morphology are expected to feed on the same resources. In metagenomics, from the beginning, studies focused on community composition and diversity. However, by quantification of particular genes intervening in a given metabolic route, functional metagenomic studies allow to infer the existence of specific microbial functional guilds in the metagenomic community. One well known example is the determination of genes intervening in denitrification pathways from soil microbiomes (Demanèche et al, 2009). Recently, fungi functional diversity has began to be investigated through a bioinformatic tool, FUNGuild, that allows to taxonomically parse fungal OTUs by ecological guild from high-throughput sequencing data (Nguyen et al., 2016). In the last years, a comprehensive approach combining metagenomics and other meta-omics like metatranscriptomics and metaprotreomics has allowed to understand the functioning of the methylotroph guilds, to discover new pathways and new players in the methane and other methylated compounds cycle, and to understand its relations with the N cycle (Chistoserdova 2014; this volume).

However, although functional diversity is increasingly recognised as an important component of biodiversity, in comparison to taxonomic diversity, methods of quantifying functional diversity are less well developed. Petchey and Gaston (2002) proposed a measure of functional diversity (FD), defined as the total branch length of a functional dendrogram, constructed using species functional traits. Various characteristics of FD make it preferable to other measures of functional diversity, such as the number of functional groups in a community. This method has began to be used recently with metagenomic functional data as well. For example, Salles et al. (2015) found that for functions such as denitrification, the diversity of functional, *nir* gene sequences are better predictors of functioning than the diversity of sequences of phylogenetic markers. A unified, flexible and multifaceted framework to estimate microbial diversity based on taxonomic, phylogenetic or functional data and across temporal and spatial scales has been recently proposed (Escalas et al., 2013).

Spatial and temporal scales

Spatial scaling issues have been recognised since early in ecology, mainly because the spatial scale chosen for sampling may have profound effects on the patterns found (Wiens, 1989). Two interesting concepts to deal with the scaling problem are the *extent* and the *grain* of a study (O'Neill *et al.,* 1986). Extent is the overall area encompassed by a study to be described by sampling. Grain is the size of the individual units of observation, for example the size of the grids used to count species in a plant community. Both extent and grain of a study should be defined by our knowledge of the system to study, for example discerning the effects of physical processes that could act at broader scales from more local, edaphic or biological interactions. Thus, while vegetation patterns at biogeographical scales are mainly determined by climatic variables, the extent of a distinctive grassland may be determined by local, edaphic variables.

Finding (or not) a pattern will depend on the homogeneity or heterogeneity of the extent considered, and on the grain size. As grain increases, a greater proportion of the spatial heterogeneity of the system is contained within a sample or grain and is lost to the study resolution, while between-grain heterogeneity decreases (Wiens 1989). If the occurrence of species in quadrats is recorded, rare species will be less likely to be recorded as grain size increases; this effect is more pronounced if the species are widely scattered in small patches than if they are highly aggregated (Levin, 1989). Figure 2 shows the effect of choosing a given grain size when the variable of interest is distributed in patches (grey) in an homogeneous matrix (white).

(Fig. 2)

Given an extent (large, black outer quadrat encompassing the study area), a given grain size (small red sampling quadrats), will reflect for example the smaller patchiness but it will miss the heterogeneity at a broader scale (larger patches and matrix). Conversely, choosing a larger grain (larger red sampling quadrats) will result in missing the smaller patch heterogeneity, since now the sampling quadrat will encompass more spatial heterogeneity, while variance between sampling quadrats will decrease. In more technical terms, the variance (the degree of spatial autocorrelation among sampling points) will change with the extent and grain size chosen for the study. Of course, the election of the extent and the grain size (sampling quadrats for example) should depend on the hypothesis and aim of the study. Choosing the relevant scale, extent and grain size for a study requires some previous knowledge about the spatial distribution of the variable under study and the habitat variables that could influence its distribution.

At field, there may be domains of scale, regions of the spectrum over which, for a particular phenomenon in a particular ecological system, patterns either do not change or change monotonically with changes in scale. Domains are separated by relatively sharp transitions from dominance by one set of factors to dominance by other sets. If the focus is on phenomena at a particular scale domain, studies conducted at finer scales will fail to include important features of pattern or causal controls; studies restricted to broader scales will fail to reveal the pattern or mechanistic relationships because such linkages are averaged out or are characteristic only of the particular domain (Wiens, 1989). Different methods have been early used in ecology to assess spatial heterogeneity and to detect scale domains. For a series of point samples, the average squared difference (semivariance) or the spatial autocorrelation between two points may be expressed in semivariograms as a function of the distance between them to estimate the scale of patchiness in a system (Sokal and Oden, 1978). Other methods used are spectral analysis (Legendre and Demers, 1984, Legendre and Gauthier, 2014), dimensional analysis (Lewis and Platt, 1982), and fractal geometry (Burrough, 1983). All these early developed methods, although with some refinements, are still in use in field ecology, while new methods, like graph theory are beginning to be used (Fortin et al., 2012).

Intimately related with the spatial heterogeneity of many ecological systems in nature, there is the problem of spatial pseudoreplication. Pseudoreplication is defined as the use of inferential statistics to test for treatment effects with data from experiments where either treatments are not replicated (though samples may be) or replicates are not statistically independent (Hurlbert, 1984). In statistical terms, depending on the type of pseudoreplication incurred, two effects may arise, increase the probability of rejecting our null hypothesis when it is true (inflated Type I error), or increase the probability of accepting the null hypothesis when it is false (inflated Type II error) (for a detailed explanation see Odintsova et al., this volume). In "simple" pseudoreplication, there are no true replicates of treatment, while in "sacrificial pseudoreplication", there is true replication of treatments but data from replicates are pooled prior to statistical analysis, or two or more samples or measurements taken from each experimental unit are treated as independent replicates. Information on the variance among treatment replicates exists in the original data, but is confounded with the variance among samples (within replicates) or else is effectively thrown away when the samples from the two or more replicates are pooled (hence “sacrificial”) (Hurlbert, 1984). Without entering into technical details, replication reduces the effects of “noise” or random variation or error, thereby increasing the precision of an estimate of, e.g., the mean of a treatment (or field variable) or the difference between two treatments (or field variables) (Hurlbert, 1984). Thus, coming back to the example in Fig. 2, to detect any spatial pattern of a given field variable, like for example, a soil contaminat that could be conditioning the presence and abundance of metagenomic communities of microbes able of metabolising the contaminant, not only the extent and grain of the study must be taken into account but also an appropriate replicated sampling design is needed. A random sampling design, with a high number of sampling quadrats of the right grain covering a great part of the extent may be adequate, but a systematic design may be more convenient to reflect the spatial pattern. However, systematic designs run the risk that the spacing interval may coincide with the period of some periodically varying property of the experimental area (Hurlbert, 1984), taking back to the scale issue.

The ecological principles and methods above mentioned are entirely valid for choosing the spatial scale, extent and grain in environmental metagenomic studies.

Moreover, as metagenomic communities are increasingly being recognised as spatially heterogeneous, special care should be taken when choosing the spatial scale for a study. Although only from recently, the soil microbiome is one of the most studied at different spatial scales, from biogeographical extent to scales smaller than 1 m. At each length scale different drivers of microbiome community organisation are expected to act. The soil main drivers acting at ecosystem (regional and biogeographic) scales (> m) are factors like climatic patterns and biogeochemical processes, at meta-community scales (cm to m) environmental gradients (pH, soil moisture, etc.) are the main factors, while at microbiome community level (10-10^3 µm) very local ecological interactions shape the pattern and functioning of microbial aggregations characteristic of such small scales (Cordero and Datta, 2016). While some evidence of defined distribution patterns have been found at regional and continental scales, examples of clear patterns for smaller scales are scarce (O'Brien et al. 2016). However, as the issue of the grain size election in general has not clearly been addressed, it is not surprising that many studies have not been able to detect significant patterns either in OTUs or genes distribution, nor significant correlations among metagenomic community variables and habitat variables like soil pH, moisture, and other factors assumed to be potentially relevant in shaping soil microbiome distributions.

Besides, but related to the issue of the small spatial scales typical of the microbiome communities, their highly patchy distribution, attributable to different factors in each habitat, complicates even more the election of the grain size for sampling. For example, the soil appear to be a rather homogeneous habitat at cm scales, but it is extremely heterogeneous and patchy at smaller scales of µm, more relevant to the microbiome. As described by Vos et al. (2016), at these small scales, the soil is composed by micro-aggregates (at 10 µm scale) with micro-pores filled with water, clustered into macroaggregates with meso- and macro-pores (at 100 μm scale) filled with water or air, depending on the moisture status of the soil. Thus, the patchy distribution of resources, large distances between bacterial cells and incomplete connectivity often restrict nutrient access and the ability to interact with other cells. Cell division also result in a short distance dispersal, and thus many bacteria remain in micro-aggregates where micro-pores offer refuge against predators and dehydration, contributing to the micro-scale patchiness of microbial communities. These small-scale patchiness appears to be inherent to the widely extended microbial activity of creating biofilms. Biofilms are ubiquitous, spatially heterogeneous systems that have high cell densities, and typically comprise many microbial species. Biofilm heterogeneity may arise through local conditions of the substrate. Further sources of heterogeneity are the ability of cells in biofilms to undergo differentiation, and ecological interactions (competition, facilitation) among microbes in the biofilm, sometimes creating heterogeneity from homogenous initial conditions (Nadell et al., 2016; Flemming et al., 2016).

The same principles behind extent and grain selection, and replication for a classical ecological study should be taken into account when formulating the hypothesis and designing a spatial sampling for a metagenomic study (Cordero and Datta, 2016). Scales of domain and spatial heterogeneity can be assessed at field in environmental metagenomics, using the same methods already used in field ecology for decades (Gonzalez et al., 2012). In the last years an increasing number of environmental metagenomic studies on spatial distribution of metagenomic communities at different extents and grains have appeared, from cm to hundreds of km (Correa-Galeote et al., 2013; Shi et al., 2015). On smaller scales, using a microcosm approach, Reim et al. (2012), sub-sampled the top 3-mm of a water-saturated soil at near in situ conditions in 100-µm steps, focusing on *pmo*A as a functional and phylogenetic marker in methane-oxidizing bacteria.

Unfortunately, the lack of adequate replication in environmental metagenomic studies is still very common, either by "simple pseudorelication" (no true replicates), but in many cases by "sacrificial pseudoreplication" (by pooling samples from true replicates). However, an increasing number of researchers are taking into account the necessity of design experiments and field studies with adequate replication. A global initiative, the Earth Microbiome Project (EMP; www.earthmicrobiome.org), seeks to systematically characterize microbial taxonomic and functional biodiversity across global ecosystems through an standardization of the protocols used to generate and analyze the data between studies. EMP is fully aware of the problem of pseudoreplication and is working towards a standardised protocol for sampling design to be adopted by all the research groups contributing samples (Knight et al., 2013).

Temporal scales are inherently connected with spatial scaling in ecology, and the tendency is to integrate both scales in ecological studies (Legendre and Gauthier, 2014). Increasing the spatial scale, the time scale of important processes also increases because processes operate at slower rates, time lags increase, and indirect effects become increasingly important (Wiens, 1989). The dynamics of different ecological phenomena in different systems follow different trajectories in space and time. For example, relevant processes to perennial plants in grasslands, like species competition and grazing, may occur in hundreds of square metres and through decades, while processes relevant to soil arthropods, restricted to smaller, local spaces and with much more shorter lives, may be defined in days and hours. In soil characteristic short timescales occur over hours to seasons. Soil microbes greatly vary their abundance and activity over timescales of hours to days (Bardgett et al., 2005). This variation is related to factors such as predation of microbes by bacteriophages, soil animals, the action of abiotic stresses (e.g. wet–dry and freeze–thaw cycles) (Mikola et al., 2002), and importantly, temporal variation in the supply of carbon and other nutrients from roots to soil (Bardgett et al., 2005). Such variations also occur at seasonal time scales. There is a general idea that soil microbes are inactive during the winter. However, Schadt et al. (2003) found in alpine soils that the biomass of microbes reached its annual maximum when soil is still frozen in late winter, and showed a significant decay thereafter. Between winter and summer there is an almost complete turnover of the microbial community, with many novel DNA sequences (Schadt et al., 2003) with different functional attributes (Lipson and Schmidt, 2004). Thus, and at least in alpine soils, one, snapshot sampling in a given time of the year may underestimate microbial diversity. Following temporal microbiome dynamics recently allowed to address the important role in community diversity of taxa that are typically in very low abundance but occasionally achieve prevalence (Shade and Gilbert, 2015).

In ecological studies, often a series of observations on the abundance of the species or variable of interest is made at equal intervals over a period of time, to detect any hidden temporal pattern through statistical procedures. Most of these statistical methods are based on time-series analysis, which allow to extract information and to identify scales of temporal patterns. One of the essential tools in time-series analysis is the periodogram or spectrum, in what is called spectral analysis. The signal (the time series) is decomposed into harmonic components based on Fourier analysis, similarly to a partition of the variance of the series, into its different oscillating components with different frequencies (periods). Peaks in the periodogram or in the spectrum indicate which periods contribute most to the variance of the series (Cazelles et al., 2008). Spectral analysis has a long way in ecology, back to the work from Bartlett (1954) that analyzed lynx temporal abundances using periodograms, and since then, amply used in ecology and population dynamics.

Although the analysis of temporal variability has an old tradition in ecology only recently has it began to be implemented with metagenomic data. Classical time series and other related techniques are increasingly used to study microbiome data obtained by metagenomics and other meta-omics approaches to assess diversity, function and ecological interactions (exhaustively reviewed in Faust et al. (2015)). These techniques have some specific requirements, that should be taken into account at the time of planning the field or experimental design. Increasing sampling frequencies in general provide higher resolution on metagenomic community dynamics although at an increased costs, thus a compromise should be reached. Sampling regularity is another important requirement for analysis techniques involving autocorrelation. Estimates for time points missing in samplings with irregular intervals can be used, but this technique can mislead conclusions if specific statistical modelling assumptions are not met. Another issue is that, although most of the time series analysis require long time records with short and regular sampling intervals, in general metagenomic time series tend to have few time points, with many sampling point gaps and many records with zero values, characteristics that create challenges for statistical analyses. Besides these problems, in metagenomic studies, just as in many ecological systems, linear correlation analyses are difficult to justify since non-linear dynamics seems to be the norm and not the exception. Rapidly variable relationships between variables in microbial community dynamics cause transient correlations that may result in spurious patterns. To overcome this problem, techniques like convergent cross-mapping can be applied to time-series data by examining the degree to which temporal components of a given variable are useful to predict the state of another variable (Sugihara et al., 2012).

Another important issue in temporal metagenomic studies is again pseudoreplication. However, as replicates in time are not easily available for temporal metagenomic studies, combining information across replicate, multiple time series can improve the inference of interactions from observations, and help to distinguish stochastic fluctuations from real temporal patterns (Hekstra et al., 2012).

On more point on the temporal scales framework in ecology and metagenomics. Classically, evolutionary time and ecological time have been differentiated. Evolutionary time operates on a longer time scale, over which changes in gene frequencies in species populations can be described as trends. Ecological time operates on a shorter time scale, over which changes in populations occur with little or no gene frequency changes (Schneider, 1994). These concepts have been developed in the context of plant and animal ecology and evolution, based on general and well known mechanisms of changes in gene frequencies: mutation, migration, genetic drift, and natural selection. However, it is not clear if this distinction can directly be extrapolated to microbial ecology and evolutionary time scales. Bacteria and fungi acquire genetic heterogeneity through other mechanisms besides mutation, like horizontal gene transfer by plasmids, transport of genetic material by phage, and capture of nucleic acids from the environment (Zaneveld et al., 2008; Fitzpatrick, 2011). The horizontally (not genealogically) acquired genes in general contribute to the adaptation of bacteria to local competitive or environmental pressures (Cohan, 2002), encoding for antibiotic resistance, novel metabolic functions, toxin production, symbiotic abilities, and other functions. Thus, this horizontally acquired genetic material confer fitness advantage to receipt bacteria in the appropriate circumstances (Dobrindt et al., 2004), acting as a true evolutionary force. The horizontal transfer or acquisition of this extra genetic material occurs over very short times and may establish a new lineage with new functional abilities in few years (Sullivan et al., 1995). This creates a conflict with the classical distinction between ecological and evolutionary time, that should be taken into account when considering the issue of time scales in metagenomic studies.

The spatial and temporal scales of a study thus determine the range of patterns and processes that can be detected. If we study a system at an inappropriate scale, we may not detect its actual dynamics and patterns but may instead no detect any pattern at all or identify patterns that are artifacts of scale. One interesting concept, used in ecology for long, is multiscale analysis: performing an analysis with respect to multiples of a unit of measurements (Schneider, 1994). By changing the unit of analysis, and thus changing the resolution, it is expected to find different patterns of the variable of interest. For example, changing the sampling quadrat size (the grain) and recording soil microbiome diversity in nested quadrats of 1 cm^2, 10 cm^2 and 100 cm2, probably diversity indexes or other metagenomic community variables will change. This is different from simply spanning many quadrats of any of this sizes in a greater space (changing the extent of the study). For example, Shi et al. (2015), in a study mentioned as multiscaled, investigated the biogeographical patterns of microbial functional genes in 24 heath soils from across the Arctic using GeoChip-based metagenomics. Principal coordinates of neighbour matrices (PCNM)-based analysis was used to analyse data across several spatial scales. However, although the sampling locations were scattered around the Canadian, Alaskan and European Arctic in a very broad extent, the grain used was the same (sampling quadrats of 12 x 12 cm). Thus, this approach can be interpreted as not truly multiscaled, since the correlations were measured between quadrats similar in size at different distances. Multiscale analysis can be used to assess changes in time as well, although studies with a temporal multiscale approach in metagenomics are only beginning to appear (Stempfhuber, 2016). Multiscale approaches, in combination with unified spatial and temporal frameworks for metagenomic studies, will soon allow to improve our understanding of the variability of microbial communities (Gonzalez et al., 2012; Gilbert and Henry, 2015).

Finally, it should be stressed here that all the considerations made about sampling metagenomic data should be taken into account for the collection of environmental metadata (climate variables, soil parameters, etc.). There is an interesting tendency to integrate metadata information in integrative workflows for processing and analysing metagenomic data on most of the currently available platforms (Ladoukakis et al., 2015). The issue of metadata collection is an important and urgent problem, that should be taken into account by the metagenomics research community, to elaborate standardised samplings protocols and share them.

Mathematical modelling

An entire paper would be needed to address in detail the issue of mathematical modelling in ecology and its influences on the recent surge of microbial community modelling. However, being mathematical modelling an increasingly important topic in metagenomics, a I will give a brief account here.

In a broad sense, a model is any abstraction of a system, built using a conceptual, mathematical, or logical, alone or combined, frameworks. In particular, mathematical modelling has been used since early in ecology. The origins of modern population ecology models can be traced back to the end of 18th century, with the model describing human population exponential growth built by Thomas Malthus (1798), and to the middle of 19th century, with the logistic growth model formulated by Pierre-François Verlhust (1845), also for human populations. In the first decades of the 20th century, these models were rediscovered by the first population ecologists, like John Gray McKendrick (for bacterial growth) and Alfred J. Lotka who, together with Vito Volterra, are considered the founders of population ecology. Since then, mathematical modelling has been implemented in every ecological field and organization level, from population to ecosystem ecology.

An exhaustive review of the huge variety of mathematical models used in ecology (deterministic, stochastic; discrete, continuous; mean-field, individually-based; etc.) (Müller and Kuttler, 2015), is out of the scope of this work, but perhaps one of the most helpful classifications of ecological models is in phenomenological and mechanistic models. Phenomenological (also called statistical) models are based on observed patterns in the data, while mechanistic models are built addressing directly the mechanisms generating observed processes and patterns. Phenomenological models provide no information about the underlying ecological mechanisms, since there is no a unique relationship between statistical patterns and mechanisms, and their predictive power is somewhat restricted to conditions comparable to those from the data to build the model were taken. On the other hand, since mechanistic models attempt to understand the phenomenon modelled, they are usually regarded as enclosing more explanatory and predictive powers than phenomenological models. For example, building a phenomenological model for a species dispersal distance using a regression model based on actual dispersal records taken at field does not tell much about the mechanisms underlying the dispersal pattern found, and the model would be applicable only to a similar scenario and within the ranges of dispersal actually recorded. Building a mechanistic model, however, including the main mechanisms involved in dispersal of, for example, wind dispersed seeds, like seed morphology, wind direction and velocity, and elevation and topographic landscape, would inform about more general features of the dispersal, like interactions among the variables included, and allow for greater and more extrapolating predictive ability. However, in some cases, both kinds of models can be complementary, since some parts of a mechanistic model, not suitable for being backed by an explicit mechanism, may contain statistical relationships (Kendall et al., 1999).

From some time to now, the tendency in ecology has been to move on from purely phenomenological models to more explanatory and predictive mechanistic models. This tendency is also beginning to permeate the work in microbial systems, thus contributing to the foundation of a modern microbial ecology (Gonzalez et al., 2012, Liberles et al., 2013). The development of mathematical models with a basis on mechanistic understanding, integrated with controlled experiments will allow to convert the huge empirical knowledge gained through microbial metagenomics and other meta-omics into fundamental insights and testable predictions about microbione composition, function and dynamics (Widder et al., 2016). In a succint but informative review, Widder and colleagues show how metagenomic and other meta-omics data can be integrated with different modelling approaches. Dynamical models of deterministic and mechanistic nature (like difference equations and flux balance analysis), stochastic dynamical systems (like Markov chains, random walks), individual-based models, and other approaches can be used to find patterns at different spatial and temporal scales, and at different ecological organization levels (from single cells to microbiomes at community and ecosystem levels), and to generate explanations and predictions about microbiome structure and function. Some modelling approaches, although essentially phenomenological, may however contribute to the generation of new hypotheses on microbiome structure and function. Network analysis has been used in ecology to study co-occurrence networks established by calculating correlations between the abundance of individual species to detect interactions among them in the community for long (Jordano, 1987). This approach has recently began to be used in microbial ecology. For example, Barberán and colleagues calculated associations between microbial taxa and applied network analysis approaches to a 16S rRNA gene barcoded pyrosequencing dataset containing 4,160,000 bacterial and archaeal sequences from 151 soil samples from a broad range of ecosystem types. The analysis revealed habitat generalists and specialists, co-occurrence patterns including general non-random association, common life history strategies at broad taxonomic levels and unexpected relationships between community members. Thus, although regarded as not purely mechanistic, network analysis has the potential of exploring inter-taxa correlations to gain a more integrated understanding of microbial community structure and the ecological rules guiding community assembly (Barberán et al., 2012). New modelling approaches tend to integrate different modelling tools to integrate information from different sources. For example, Noecker et al. (2015), in a systems biology approach, propose a comprehensive framework to systematically link variation in metabolomic data with community composition by utilizing taxonomic, genomic, and metabolic information. Their approach integrate available and inferred genomic data, metabolic network modelling, and a method for predicting community-wide metabolite turnover to estimate the biosynthetic and degradation potential of a given community.

Conclusions

Metagenomics and other meta-omics constitute, due to their inherent nature, a complex field placed at the intersection of many disciplines, like molecular biology, microbiology, ecology, chemistry, bioinformatics, among others, and new ones are hastily being implicated. The theoretical and methodological complexity arising from this multifaceted and dynamic field requires the integration of useful theoretical basis and methodologies coming from already well established disciplines like ecology, and dealing with change of paradigms like the traditional organism-centred approach to a new, organism- and species-free context. Thus, while some concepts and methodologies coming from ecology should be revised for application on metagenomics and meta-omics fields, like niche theory, other do not require great changes and it is predicted to be increasingly adopted by environmental metagenomics. Ecological principles behind spatial and temporal scales should be taken into account when formulating the hypothesis and sampling design for metagenomic and meta-omics studies, and the wealth of modelling approaches developed through decades by ecologists is being proven extremely useful in the context of metagenomics.

Acknowledgements

DEM is a research member of the Consejo Nacional de Investigaciones Científicas y Tecnológicas (CONICET), Argentina.

References

Barberán, A., Bates, S.T., Casamayor, E.O., and Fierer, N. (2012). Using network analysis to explore co-occurrence patterns in soil microbial communities. The ISME Journal *6*, 343-351.

Bardgett, R.D., Bowman, W.D., Kaufmann, R., and Schmidt, S.K. (2005). A temporal approach to linking aboveground and belowground ecology. Trends in Ecology & Evolution *20*, 634-641.

Bartlett, M.S. (1954) Problèmes de l'analyse spectrale des sèries temporelles stationnaires. Publ. Inst. Stat. Univ. Paris *3*,119–134.

Burrough, P.A. (1983) Multiscale sources of spatial variation in soil. I. The application of fractal concepts to nested levels of soil variation. Journal of Soil Science *34***,** 577-597.

Cazelles, B., Chavez, M., Berteaux, D., Ménard, F., Vik, J.O., Jenouvrier, S., and Stenseth, N.C. (2008). Wavelet analysis of ecological time series. Oecologia *156*, 287-304.

Chistoserdova, L. (2014). Functional metagenomics of the nitrogen cycle in freshwater lakes with focus on methylotrophic bacteria. In Metagenomics of the microbial nitrogen cycle. Theory, methods and applications, D. Marco, ed. (Norfolk, UK: Caister Academic Press), pp. 195-208.

Cohan, F. M. (2002). What are bacterial species?. Annual Reviews in Microbiology *56*, 457-487.

Cordero, O.X., and Datta, M.S. (2016). Microbial interactions and community assembly at microscales. Current Opinion in Microbiology, *31*, 227-234.

Correa-Galeote, D., Marco, D.E., Tortosa, G., Bru, David, Philippot, L., and Bedmar, E.J. (2013). Spatial distribution of N-cycling microbial communities showed complex patterns in constructed wetland sediments. FEMS Microbiology Ecology *83*, 340-351.

Creer, S., Deiner, K., Frey, S., Porazinska, D., Taberlet, P., Thomas, W.K., Potter, C. and Bik, H.M. (2016). The ecologist's field guide to sequence-based identification of biodiversity. Methods in Ecology and Evolution *7*, 1008–1018.

DeLong, E. F. (2004). Microbial population genomics and ecology: the road ahead. Environmental Microbiology, *6*, 875-878.

Demanèche, S., Philippot, L., David, M.M., Navarro, E., Vogel, T.M., and Simonet, P. (2009). Characterization of denitrification gene clusters of soil bacteria via a metagenomic approach. Applied and Environmental Microbiology *75*, 534-537.

Dobrindt, U., Hochhut, B., Hentschel, U., and Hacker, J. (2004). Genomic islands in pathogenic and environmental microorganisms. Nature Reviews Microbiology *2*, 414-424.

Escalas, A., Bouvier, T., Mouchet, M.A., Leprieur, F., Bouvier, C., Troussellier, M., and Mouillot, D. (2013). A unifying quantitative framework for exploring the multiple facets of microbial biodiversity across diverse scales. Environmental Microbiology *15*, 2642-2657.

Faust, K., Lahti, L., Gonze, D., de Vos, W.M., and Raes, J. (2015). Metagenomics meets time series analysis: unraveling microbial community dynamics. Current Opinion in Microbiology *25*, 56-66.

Fierer, N., Lauber, C.L., Ramirez, K.S., Zaneveld, J., Bradford, M.A., and Knight, R. (2012). Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients. The ISME Journal *6*, 1007-1017.

Fitzpatrick, D.A. (2012). Horizontal gene transfer in fungi. FEMS Microbiology Letters *329*, 1-8.

Flemming, H.C., Wingender, J., Szewzyk, U., Steinberg, P., Rice, S.A., and Kjelleberg, S. (2016). Biofilms: an emergent form of bacterial life. Nature Reviews Microbiology *14*, 563-575.

Fortin, M.J., James, P.M., MacKenzie, A., Melles, S.J., and Rayfield, B. (2012). Spatial statistics, spatial regression, and graph theory in ecology. Spatial Statistics *1*, 100-109.

Gilbert, J.A., and Henry, C. (2015). Predicting ecosystem emergent properties at multiple scales. Environmental Microbiology Reports *7*, 20-22.

Gonzalez, A., King, A., Robeson II, M.S., Song, S., Shade, A., Metcalf, J.L., and Knight, R. (2012). Characterizing microbial communities through space and time. Current Opinion in Biotechnology *23*, 431-436.

Goodall, D.W. (1954). Vegetational classification and vegetational continua. Angew. Pflanzensoziologie, Wien. Festchrift Aich. *1*,168-182.

Handelsman, J., Rondon, M.R., Brady, S.F., Clardy, J., and Goodman, R.M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chemistry & Biology *5*, R245-R249.

Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. Microbiology and Molecular Biology Reviews *68*, 669-685.

Hekstra, D.R., Leibler, S. (2012). Contingency and statistical laws in replicate microbial closed ecosystems. Cell *149*,1164-1173.

Hurlbert, S.H. (1984). Pseudoreplication and the design of ecological field experiments. Ecological Monographs, *54*, 187-211.

Jordano, P. (1987). Patterns of mutualistic interactions in pollination and seed dispersal: Connectance, dependence asymmetries, and coevolution. American Naturalist *129*, 657–677.

Ju, F., and Zhang, T. (2015). Experimental design and bioinformatics analysis for the application of metagenomics in environmental sciences and biotechnology. Environmental Science and Technology *49*, 12628-12640.

Kendall, B.E., Briggs, C.J., Murdoch, W.W., Turchin, P., Ellner, S.P., McCauley, E., Nisbet, R.M. and Wood, S.N. (1999). Why do populations cycle? A synthesis of statistical and mechanistic modeling approaches. Ecology *80*, 1789-1805.

Knight, R., Jansson, J., Field, D., Fierer, N., Desai, N., Fuhrman, J. A., Hugenholtz, P., van der Lelie, D., Meyer, F., Stevens, R., et al. (2013). Unlocking the potential of metagenomics through replicated experimental design. Nature Biotechnology *30*, 513-520.

Kurtz, Z.D., Müller, C.L., Miraldi, E.R., Littman, D.R., Blaser, M.J., Bonneau, and R.A. (2015). Sparse and compositionally robust inference of microbial ecological networks. PLoS Comput Biol. *11*, e1004226.

Ladoukakis, E., Kolisis, F.N., and Chatziioannou, A A. (2015). Integrative workflows for metagenomic analysis. In Multi-omic Data Integration, Frontiers Research Topics, P. Tieri, Ch. Nardini and J.E. Dent, eds. pp. 115-125.

Legendre, L., and Demers, S. (1984). Towards dynamic biological oceanography and limnology. Canadian Journal of Fishery and Aquatic Science *41*, 2-19.

Legendre, P., and Gauthier, O. (2014). Statistical methods for temporal and space–time analysis of community composition data. Proceedings of the Royal Society of London B: Biological Sciences *281*, 20132728.

Lennon, J.T. (2011). Replication, lies and lesser-known truths regarding experimental design in environmental microbiology. Environmental Microbiology *13*, 1383-1386.

Levin, S.A. (1989) Challenges in the development of a theory of ecosystem structure and function. In Perspectives in Ecologicol Theory, J. Roughgarden, R.M. May and S.A. Levin, eds. (Princeton, N.J: Princeton University Press), pp. 242-255.

Lewis, M.R., and Platt, T. (1982) Scales of variation in estuarine ecosystems. In Estuarine Comparisons, V.S. Kennedy, ed. (New York: Academic Press), pp. 3-20.

Liberles, D.A., Teufel, A.I., Liu, L., and Stadler, T. (2013). On the need for mechanistic models in computational genomics and metagenomics. Genome biology and evolution *5*, 2008-2018.

Lipson, D.A., and Schmidt, S.K. (2004). Seasonal changes in an alpine soil bacterial community in the Colorado Rocky Mountains. Applied and environmental microbiology *70*, 2867-2879.

Linquist, S., Cottenie, K., Elliott, T.A., Saylor, B., Kremer, S.C., and Gregory, T.R. (2015). Applying ecological models to communities of genetic elements: the case of Neutral Theory. Molecular ecology *24*, 3232-3242.

Malthus, T. (1798) An Essay on the Principle of Population, as it affects the future improvement of society with remarks on the speculations of Mr. Godwin, M. Condorcet, and other writers. Anonymously published.

Marco, D. (2008). Metagenomics and the niche concept. Theory in Biosciences *127*, 241-247.

Meiring, T.L., Bauer, R., Scheepers, I., Ohloff, C., Tuffin, I.M., and Cowan, D.A. (2011). Metagenomics and beyond: current approaches and integration with complementary technologies. In Metagenomics: current innovations and future trends, D. Marco, ed. (Norfolk, UK: Caister Academic Press), pp. 1-19.

Mikola, J., Bardgett, R.D., and Hedlund, K. (2002). Biodiversity, ecosystem functioning and soil decomposer food webs. In Biodiversity and ecosystem functioning: synthesis and perspectives, M. Loreau, S. Naeem, and P. Inchausti, eds. (Oxford, UK: Oxford University Press), pp. 169-180.

Müller, J., and Kuttler, C. (2015). Methods and Models in Mathematical Biology (Berlin Heidlerberg: Springer-Verlag).

Nadell, C.D., Drescher, K., and Foster, K.R. (2016). Spatial structure, cooperation and competition in biofilms. Nature Reviews Microbiology *14*, 589–600.

Nguyen, N.H., Song, Z., Bates, S.T., Branco, S., Tedersoo, L., Menke, J., Schilling, J.S., and Kennedy, P. G. (2016). FUNGuild: an open annotation tool for parsing fungal community datasets by ecological guild. Fungal Ecology *20*, 241-248.

Noecker, C., Eng, A., Srinivasan, S., Theriot, C.M., Young, V.B., Jansson, J. K., Fredricks, D.N., and Borenstein, E. (2016). Metabolic model-based integration of microbiome taxonomic and metabolomic profiles elucidates mechanistic links between ecological and metabolic variation. mSystems *1*, e00013-15.

O'Brien, S.L., Gibbons, S.M., Owens, S.M., Hampton-Marcell, J., Johnston, E.R., Jastrow, J.D., Gilbert, J.A., Meyer, F., and Antonopoulos, D. A. (2016). Spatial scale drives patterns in soil bacterial diversity. Environmental Microbiology *18*, 2039–2051.

O'Malley, M. A., & Dupré, J. (2009). Philosophical themes in metagenomics. In Metagenomics: Theory, methods and applications, D. Marco, ed. (Norfolk, UK: Caister Academic Press), pp. 183-208.

O'Neill, R.V., DeAngelis, D.L., Waide, J.B., and Allen, T.F.H. (1986). A hierarchical concept of ecosystems (Princeton, New Jersey, USA: Princeton University Press).

Pace, N.R., Stahl, D.A., Lane, D.J., and Olsen, G.J. (1985). Analyzing natural microbial populations by rRNA sequences. ASM News *51*, 4-12.

Petchey, O.L., and Gaston, K.J. (2002). Functional diversity (FD), species richness and community composition. Ecology Letters *5*, 402-411.

Prosser, J.I. (2010). Replicate or lie. Environmental Microbiology *12*, 1806-1810.

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

Reim, A., Lüke, C., Krause, S., Pratscher, J., and Frenzel, P. (2012). One millimetre makes the difference: high-resolution analysis of methane-oxidizing bacteria and their specific activity at the oxic–anoxic interface in a flooded paddy soil. The ISME Journal *6*, 2128-2139.

Riesenfeld, C.S., Schloss, P.D., and Handelsman, J. (2004). Metagenomics: genomic analysis of microbial communities. Annual Review Genetics *38*, 525-552.

Salles, J.F., Le Roux, X., and Poly, F. (2015). Relating phylogenetic and functional diversity among denitrifiers and quantifying their capacity to predict community functioning. In The causes and consequences of microbial community structure, D.R. Nemergut, A. Shade, and C. Violle, eds. Frontiers Research Topics 140-153.

Shade, A., and Gilbert, J.A. (2015). Temporal patterns of rarity provide a more complete view of microbial diversity. Trends in Microbiology *23*, 335-340.

Schadt, C.W., Martin, A.P., Lipson, D.A., and Schmidt, S.K. (2003). Seasonal dynamics of previously unknown fungal lineages in tundra soils. Science *301*, 1359-1361.

Schneider, D.C. (1994). Quantitative ecology: spatial and temporal scaling (San Diego, California, USA: Academic Press).

Shi, Y., Grogan, P., Sun, H., Xiong, J., Yang, Y., Zhou, J., and Chu, H. (2015). Multi-scale variability analysis reveals the importance of spatial distance in shaping Arctic soil microbial functional communities. Soil Biology and Biochemistry *86*, 126-134.

Simberloff, D., and Dayan, T. (1991). The guild concept and the structure of ecological communities. Annual Review of Rcology and Rystematics *22*, 115-143.

Sokal, R.R., and Oden, N.L. (1978) Spatial autocorrelation in biology. 2. Some biological implications and four applications of evolutionary and ecological interest. Biological Journal of the Linnean Society *10*, 229- 249.

Stempfhuber, B.H.J. (2016). Drivers for the performance of nitrifying organisms and their temporal and spatial interaction in grassland and forest ecosystems (Doctoral dissertation, Dissertation, München, Technische Universität München).

Sugihara, G., May, R., Ye, H., Hsieh, C.-H., Deyle, E., Fogarty, M., and Munch, S. (2012) Detecting causality in complex ecosystems. Science *338,* 496–500.

Sullivan, J.T., Patrick, H.N., Lowther, W.L., Scott, D.B., and Ronson, C.W. (1995). Nodulating strains of *Rhizobium loti* arise through chromosomal symbiotic gene transfer in the environment. Proceedings of the National Academy of Sciences *92*, 8985-8989.

Verhulst, P.-F. (1845). Recherches mathématiques sur la loi d'accroissement de la population. Nouveaux Mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles *18*, 1–42.

Vos, M., Wolf, A.B., Jennings, S.J., and Kowalchuk, G.A. (2013). Micro-scale determinants of bacterial diversity in soil. FEMS Microbiology Reviews, *37*, 936-954.

Widder, S., Allen, R.J., Pfeiffer, T., Curtis, T.P., Wiuf, C., Sloan, W.T., Cordero, O.X., Brown, S.P., Momeni, B., Shou, W. et al. (2016). Challenges in microbial ecology: building predictive understanding of community function and dynamics. The ISME Journal 1-12.

Wiens, J.A. (1989). Spatial scaling in ecology. Functional Ecology *3*, 385-397.

Wilson, J.B. (1999). Guilds, functional types and ecological groups. Oikos *86*, 507-522.

Woese, C.R., and Fox, G.E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proceedings of the National Academy of Sciences, *74*, 5088-5090.

Zaneveld, J.R., Nemergut, D.R., and Knight, R. (2008). Are all horizontal gene transfers created equal? Prospects for mechanism-based studies of HGT patterns. Microbiology 154, 1-15.

Zhou, J., He, Z., Yang, Y., Deng, Y., Tringe, S.G., and Alvarez-Cohen, L. (2015). High-throughput metagenomic technologies for complex microbial community analysis: open and closed formats. MBio *6*, e02288-14.

Figure Legends

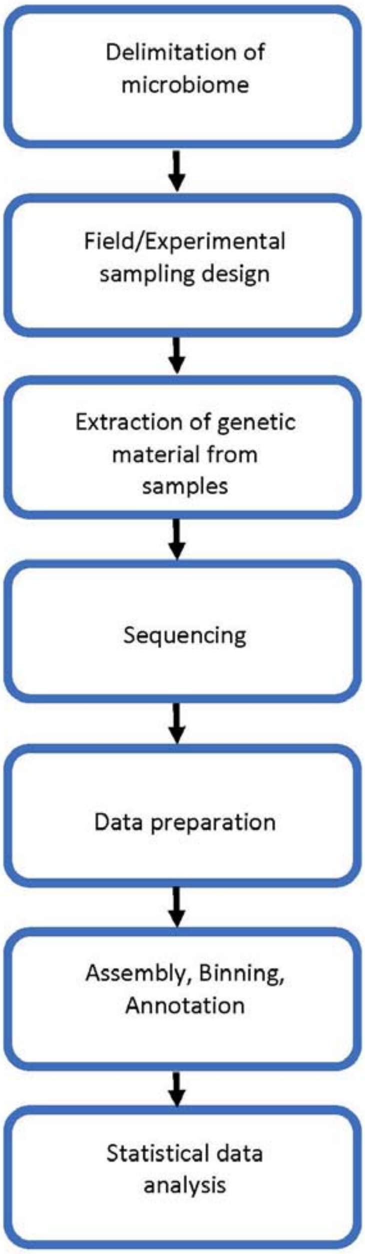
Delimitation of microbiome
Field/Experimental sampling design
Extraction of genetic material from samples
Sequencing
Data preparation
Assembly, Binning, Annotation
Statistical data analysis

Fig. 1. A typical pipeline for metagenomic studies.

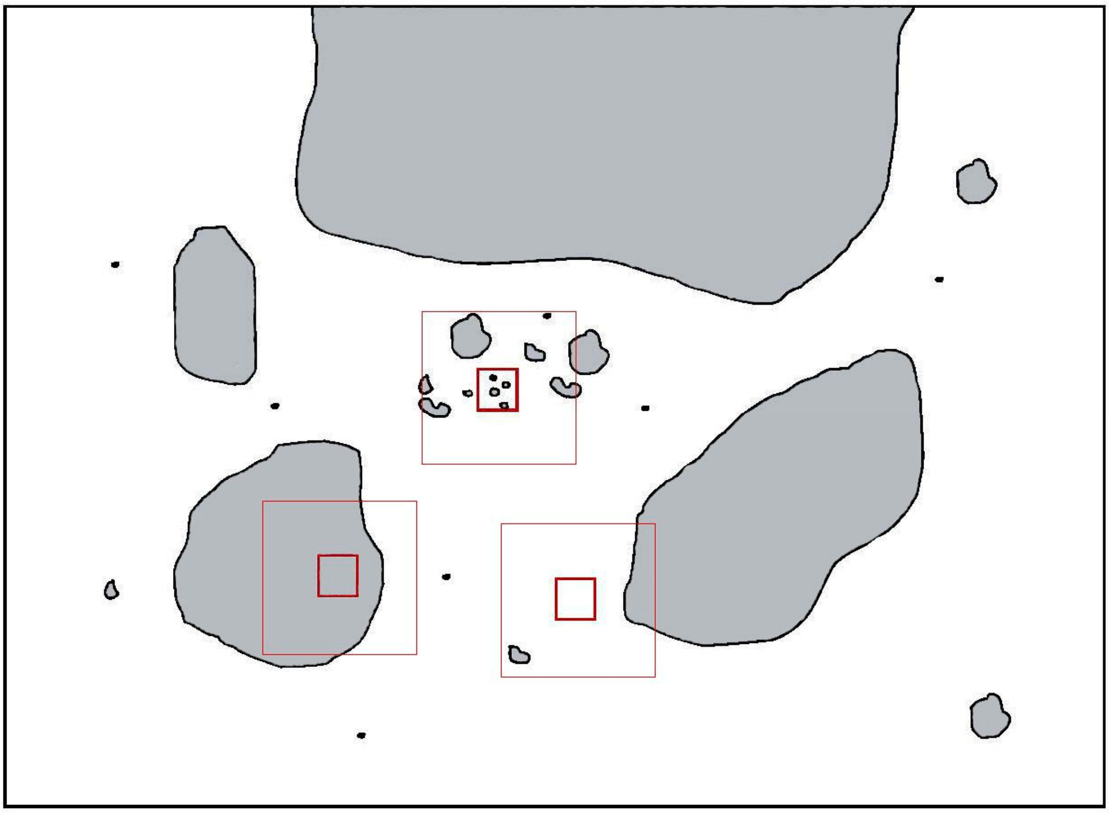

Fig. 2 The effect of choosing a given grain (sampling unit) size when the variable of interest is distributed in patches (grey) in an homogeneous matrix (white). Given an extent (black outer quadrat encompassing the study area), a given grain size (small red sampling quadrats), will reflect for example the smaller patchiness but it will miss the heterogenity at a broader scale (larger patches and matrix). Conversely, choosing a larger grain (larger red sampling quadrats) will result in missing the smaller patch heterogeneity, and variance between sampling quadrats will decrease.