






Minireview

Modelling spatial patterns in host-associated microbial communities

Matti O. Ruuskanen ^{1,2*†} Guilhem Sommeria-Klein ^{3†}
Aki S. Havulinna ^{2,4} Teemu J. Niiranen ^{1,2,5} and
Leo Lahti ³

¹Department of Internal Medicine, University of Turku, Turku, Finland.

²Finnish Institute for Health and Welfare, Helsinki, Finland.

³Department of Computing, University of Turku, Turku, Finland.

⁴Institute for Molecular Medicine Finland, FIMM-HiLIFE, Helsinki, Finland.

⁵Division of Medicine, Turku University Hospital, Turku, Finland.

Summary

Microbial communities exhibit spatial structure at different scales, due to constant interactions with their environment and dispersal limitation. While this spatial structure is often considered in studies focusing on free-living environmental communities, it has received less attention in the context of host-associated microbial communities or microbiota. The wider adoption of methods accounting for spatial variation in these communities will help to address open questions in basic microbial ecology as well as realize the full potential of microbiome-aided medicine. Here, we first overview known factors affecting the composition of microbiota across diverse host types and at different scales, with a focus on the human gut as one of the most actively studied microbiota. We outline a number of topical open questions in the field related to spatial variation and patterns. We then review the existing methodology for the spatial modelling of microbiota. We suggest that methodology from related fields, such as systems biology and macro-organismal ecology, could be adapted to

obtain more accurate models of spatial structure. We further posit that methodological developments in the spatial modelling and analysis of microbiota could in turn broadly benefit theoretical and applied ecology and contribute to the development of novel industrial and clinical applications.

Introduction

In addition to playing a key role in global biogeochemical cycles, microbial communities occur in and on multicellular eukaryote hosts, such as plants and animals. Such host-associated microbial communities, which may include bacteria, protists, fungi and archaea, are referred to as ‘microbiota’ - or when taken together with their genomes, metabolites, viruses and physico-chemical environment, as ‘microbiomes’ (Berg *et al.*, 2020). Microbiota contribute to the adaptation of the host to varying environments, for example by breaking down compounds for easier absorption and preventing the growth of pathogens. These symbiotic relationships are an important evolutionary force both for the hosts (McFall-Ngai *et al.*, 2013) and their associated microbes (Garcia and Gerardo, 2014). Multicellular hosts and their persistent microbial symbionts are in fact increasingly recognized as unified biological entities (or ‘holobionts’) from an evolutionary perspective (Simon *et al.*, 2019). Microbiota composition is, however, not solely dependent on the identity and phylogeny of the host organism but is shaped by many factors that boil down to four fundamental ecological processes: selection, drift, dispersal and speciation (Vellend, 2010). Together, these factors lead to temporally varying spatial patterns at different scales, from the level of individual cells to biogeographic patterns. Spatial patterns in microbiota composition have been observed, for example on comparable tissue types across host individuals in plants (Bakker *et al.*, 2014), and in animals, such as marine invertebrates (van de Water *et al.*, 2018), insects (Wang *et al.*, 2020), amphibians (Griffiths *et al.*, 2018), ray-finned fish (Smith *et al.*, 2015) and mammals (Tung *et al.*, 2015).

Received 9 December, 2020; revised 2 March, 2021; accepted 11 March, 2021. *For correspondence. E-mail matti.ruuskanen@utu.fi.
†These authors contributed equally to the work.

Microbiota composition varies dynamically as individual microbes interact with their abiotic and biotic environment, such as physical forces, acidity, redox potential and a multitude of chemical compounds. Many of these factors are directly related to the host, such as the local physicochemical conditions, diet and degree of exposure to the host's environment, but interactions between microbes are also highly important, as are stochastic drift and dispersal (Harris *et al.*, 2017). Spatial structure arises from these processes for two broad reasons. First, most environmental factors affecting the host and its microbiota are unevenly spatially distributed. Second, spatial distance combined with random drift can itself generate spatial heterogeneity in the microbiota by limiting the dispersal of microbes between hosts or between different parts of the host's body. Therefore, spatial patterns contain precious information on how host-associated microbial communities establish themselves, persist and change over time.

It has only recently become possible to reveal the complete taxonomic composition of microbial communities through high-throughput sequencing (Caporaso *et al.*, 2011; Quince *et al.*, 2017). The analysis and interpretation of these data, however, poses a series of technical challenges related to the compositional nature of the data (i.e. only relative microbe abundances are measured in each sample), the discretization of molecular observations into discrete taxa, the accuracy of taxonomic and functional assignments and the large number of rare taxa. Analytical methods addressing these problems have been actively developed (Knight *et al.*, 2018), but they often do not explicitly account for spatial variation, especially in the case of host-associated communities (Björk *et al.*, 2018). Yet the development and wider application of spatial modelling techniques is a key to providing answers to fundamental questions on the ecology of microbiota, such as the relative influence of dispersal and local environmental conditions on community composition, or how and why communities shift between alternative states (Gonze *et al.*, 2018). Accounting for spatial variation also represents an important remaining challenge in medical microbiology, because its confounding effects can reduce the applicability of human microbiota analyses in diagnostics (Gaulke and Sharpton, 2018; He *et al.*, 2018). Microbiota composition and function have been shown to substantially contribute to chronic diseases such as irritable bowel syndrome, colorectal cancer, fatty liver disease, asthma and dementia (Feng *et al.*, 2018). Thus, the study and diagnosis of these and other diseases could greatly benefit from a better understanding of spatial variability in the microbiota, and of its causes and consequences.

In this review, we first provide an overview of factors influencing spatial patterns in host-associated microbial

communities, outline practical considerations in their analysis and present topical questions whose investigation would benefit from spatial modelling techniques and tools. We then cover recent methodological developments in spatial machine learning and probabilistic modelling, which provide new means to harness the spatial information in the data. Finally, we detail several extensions and modifications that could significantly improve the applicability of such approaches in microbiome research. While the human gut is often used as an example in this review as one of the most studied microbiota, the spatial modelling approaches that we discuss have broad applicability, and we also provide illustrating examples from various other host organisms.

Axes of variation in microbiota

Host-associated microbial communities generally differ in composition from free-living environmental communities (Adair and Douglas, 2017). Various factors are known to exert selective pressures on the resident microbes, leading eventually to the establishment of communities with relatively stable compositions. Known stabilizing selective pressures include (i) the host immune system (Hooper *et al.*, 2012) and other compounds produced by the host (Fischbach and Segre, 2016); (ii) metabolic products of other microbes such as antimicrobial toxins (Wexler *et al.*, 2016), enzymes (Rakoff-Nahoum *et al.*, 2016) and signalling compounds (Garcia, 2018); (iii) physicochemical constraints such as temperature (Sepulveda and Moeller, 2020), pH (Sylvain *et al.*, 2016), oxygen (Albenberg *et al.*, 2014) and particularly in gastrointestinal communities, host diet (O'Keefe *et al.*, 2015; Riaz Rajoka *et al.*, 2017). The relative strength of the different selective pressures on microbial communities depends strongly on the host and on the specific site (Adair and Douglas, 2017).

Microbial communities are considered to have a relatively stable composition within a single host at a specific host site (Coyte *et al.*, 2015). However, they exhibit large variations along the following axes: (i) between host species at comparable host sites (i.e. on the same tissue type or at the same body site), (ii) across the different surfaces and compartments of a single host species, (iii) between individuals of the same host species at comparable host sites and (iv) along time in a given microbial community when the spatial location of the host individual changes (Fig. 1). A large part of this variation is spatially structured, and we briefly review below the spatial processes and spatially correlated factors that are known to contribute to these patterns. Because of the methodological focus of this review, only a small number of examples from a variety of microbiota are discussed here to outline possible spatially relevant factors in the context of the different axes of variation.

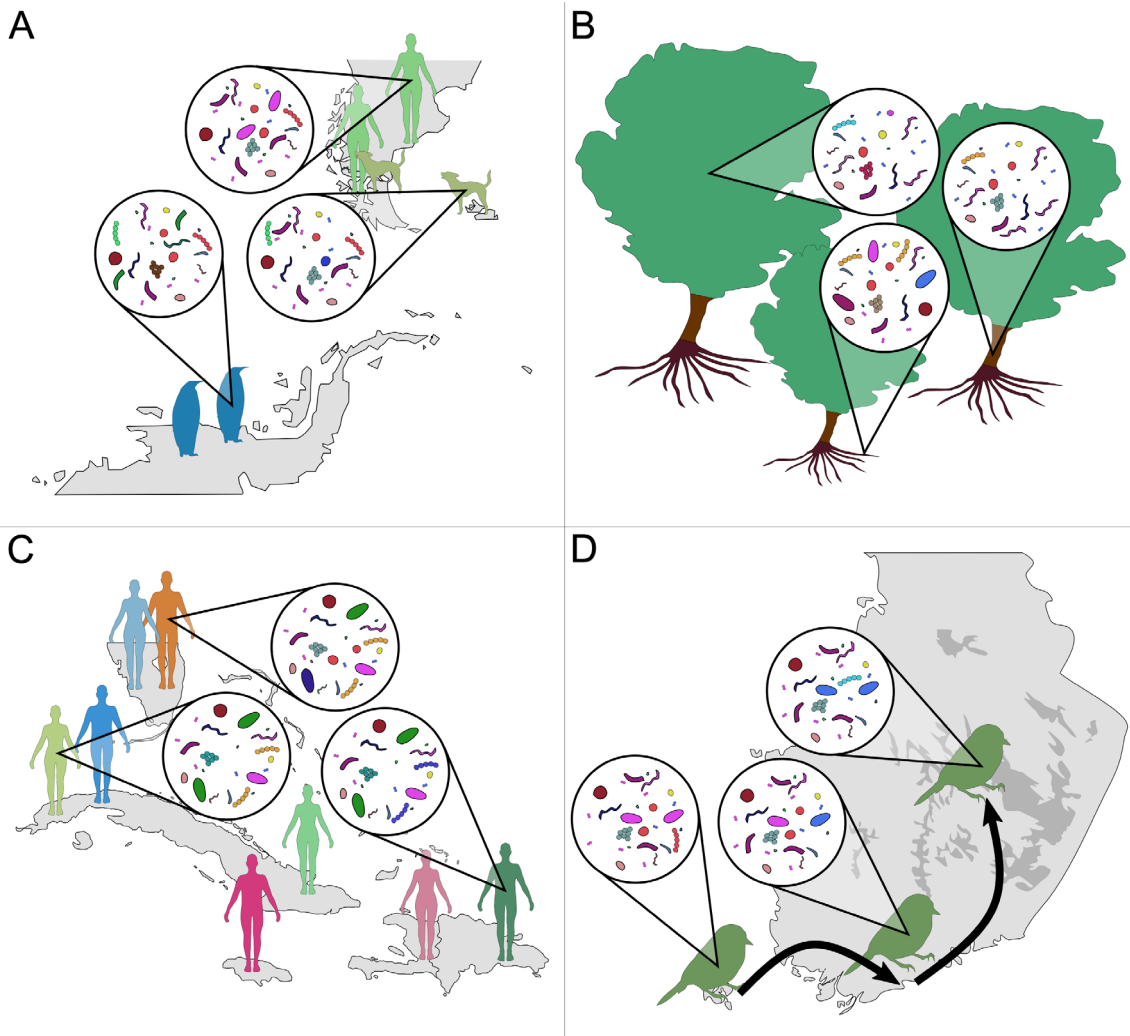


Fig 1. Axes of variation in host-associated microbial communities with examples of related spatial dimensions.

A. Variation between different host species at a comparable host site, i.e. on the same tissue type or at the same body site.

B. Variation across the different surfaces and compartments in a single host species, like the leaves, trunk and roots of a tree species.

C. Variation between individuals of the same host species at a comparable host site.

D. Variation of the microbiota over time at the same host site of the same individual depending on the individual's location. Parts of this figure were adapted from the open source material (NuclearVacuum, 2008; Nordwestern, 2015; Silar, 2016; Rashidi, 2017; DataBase Center for Life Science, 2018). [Color figure can be viewed at wileyonlinelibrary.com]

Variation between host species

One widely studied subject in microbial ecology has been the establishment and maintenance of unique microbial communities in different host species at roughly comparable host sites, like in the gastrointestinal tract of animals (Fig. 1A). In a single host individual of any one species, the development of the microbiota often proceeds through time in a predictable fashion through the primary succession of introduced microbes (Ortiz-Álvarez *et al.*, 2018). The vertical transmission of specific microbes from the parent to the offspring is important in most multicellular eukaryotes in providing the offspring with an initial inoculum (Bright and Bulgheresi, 2010).

However, free-living environmental microbes can also establish stable communities in or on various hosts, often gaining fitness benefits in the mutualistic association (Garcia and Gerardo, 2014). Consequently, host species sharing the same spatial distribution tend to have more similar microbiota due to the incorporation of the same environmental microbial taxa. The direct horizontal transfer of microbes between host species further increases this similarity, for example between distantly related ground-dwelling mammals (Perofsky *et al.*, 2019). Hence, the microbiota of spatially coexisting host species are linked together into a wider metacommunity (Adair and Douglas, 2017). Host-microbe associations can nevertheless also be extremely specific, such as when

environmental *Vibrio fischeri* strains colonize the light organs of the squid *Euprymna scolopes* (McFall-Ngai and Ruby, 1991).

Selective pressures on the microbiota vary between hosts, and these variations exhibit a phylogenetic signal. For example, the phylogenetic relatedness of mammalian hosts correlates with the similarity of their gut microbiota (Song *et al.*, 2020). However, this is not a universal rule, as distantly related birds and bats have surprisingly similar gut communities. This similarity is likely caused by the reduced immune regulation in these hosts, perhaps attributable to flight physiology (see Song *et al.*, 2020), which results in relaxed constraints on microbiota composition. We have only begun to chart how various mechanisms affect the interspecies differences in the microbiota, and disentangling the relative effects of the horizontal transfer of microbes, vertical transfer and host-specific selective pressure on microbiota composition across host species is an active area of research (Perez-Lamarque and Morlon, 2019; Leftwich *et al.*, 2020). Our understanding of these mechanisms could greatly benefit from incorporating spatial information explicitly in the modelling frameworks.

Variation across surfaces and compartments in a single host species

The composition of microbial communities varies greatly between external surfaces and internal sites in individual host species (Fig. 1B). Indeed, most often the communities on external surfaces appear to be regulated by environmental variables such as temperature, and internal communities by host-related factors like the immune system and diet (Woodhams *et al.*, 2020). Community composition also varies among the external or internal sites. For instance, communities demonstrate distinct spatial distributions along the mammalian gastrointestinal tract to the scale of specific microhabitats, such as the lumen of the large intestine, mucus layers and colonic crypts (Zhang *et al.*, 2014; Donaldson *et al.*, 2016). The communities can be highly organized down to the micrometre scale on surfaces such as the human tongue dorsum (Wilbert *et al.*, 2020). Distinct communities are also observed between the different compartments of plants such as the rhizosphere, phyllosphere, and leaf and root endospheres (Hacquard, 2016). Understanding the processes at play at these finest spatial scales would benefit from the sampling of communities along spatial gradients using a dedicated methodology, rather than at distinct sites in and on the host organism (see 'Accounting for scale' below). These types of analyses would represent a shift from thinking in terms of categorical host sites to a continuous landscape of host-associated

microbiota (Proctor and Relman, 2017) and would require spatially explicit modelling approaches.

Variation between individuals of the same species

Another axis of variation can be observed in microbiota composition between individuals of the same host species, when sampling a community at the same host site (Fig. 1C). This type of variation is currently receiving much attention in humans due to its medical relevance. For example, comparing the gut microbiota of patients suffering from a range of diseases to those of healthy controls has led to a number of discoveries on the role played by the gut microbiota in disease pathogenesis (Feng *et al.*, 2018). The differences in host site-specific microbiota in both diseased and healthy hosts relate back to the individual life histories of the hosts and include, for example their genetic background (Benson *et al.*, 2010), initial colonization with microbes (Callens *et al.*, 2018) and related founder effects in the microbial community (Litvak and Bäuml, 2019), environmental exposures (Chiu *et al.*, 2020), diet (Riaz Rajoka *et al.*, 2017), aging-related changes (Langille *et al.*, 2014), medication (Falony *et al.*, 2016) and the diseases themselves (Malla *et al.*, 2019). Many of these factors are unevenly distributed across space, thus producing also spatial patterns in microbiota compositions. The factors affecting inter-individual differences are also often unknown or unmeasured due to practical constraints. Because spatial information captures at least part of this variation, incorporating it in the analysis would be beneficial even when the source of the variation is unknown. Furthermore, as in the case of variations between host species, incorporating spatial information is an efficient means to account for the introduction of microbes from the environment or through horizontal transfer from other individuals.

Variation in the same community over time

The fourth axis of variation often examined in microbiota studies is between states of the microbial community in the same host individual at the same host site over time (Fig. 1D). The current state of a microbial community depends on its past states and on the influence of factors with uneven spatial distribution, which are described above. Thus, it is impossible to completely separate spatial patterns from temporal variation in the communities.

While host site-specific communities in the same individual can exhibit stable composition over time (Coyle *et al.*, 2015), hourly to daily variations are common, for example in the mammalian gastrointestinal tract (David *et al.*, 2014; Maurice *et al.*, 2015; Voigt *et al.*, 2016). It is likely that most host-associated communities have multiple stable configurations, which can provide the hosts

with similar necessary functions and between which they can 'switch'. Functional redundancy between two communities does not mean that they are equivalent, however, and communities with different initial taxonomic compositions can be expected to react in different ways to stressors (Moya and Ferrer, 2016). Disease-associated (or 'dysbiotic') states of the microbiota are reported to be especially unstable through time, likely due to a reduction in the host's regulation ability (Zaneveld *et al.*, 2017).

Longitudinal studies of the human gut microbiota have shown that the (geographical) relocation of individuals can have measurable long- and short-term impacts on their microbiota (David *et al.*, 2014; Kaplan *et al.*, 2019). Relocation-associated changes have also been observed over time in the microbiota of migratory birds (Wu *et al.*, 2018) and stingless bee colonies (Hall *et al.*, 2021). While true temporal models are beyond the scope of this review, spatial approaches should thus not overlook the possible temporal aspects of the data. Microbiota time-series data with enough spatial coverage to allow the simultaneous investigation of spatial and temporal patterns are currently rare, but such studies will be crucial to establish a mechanistic understanding of the communities.

Spatial patterns and the importance of scale

As seen in the previous section, spatial structures are thought to stem from the horizontal dispersion of microbes between hosts (Antwis *et al.*, 2018) and from the environmental filtering of communities by spatially correlated factors. In humans, for instance, such factors include diet for the gut microbiota (Filippo *et al.*, 2010) or lifestyle and environment for the skin microbiota (Lehtimäki *et al.*, 2018). However, the relative importance of the different spatially correlated factors is likely scale-dependent (Ladau and Eloe-Fadrosh, 2019). Hence, when designing studies on host-associated microbiota (along any axis of variation), one should carefully consider the spatial scale of the sampling and the possible processes affecting community composition at that scale.

Spatial patterns across scales

The scale-dependence of spatial patterns in microbiota composition is well illustrated by the known patterns of inter-individual differences in the human gut microbiota. Within a household, the horizontal dispersal of microbes increases the similarity between cohabiting individuals (Finnicum *et al.*, 2019). At the neighbourhood scale, the effect of vegetation cover in the living environment affects inter-individual differences, likely due to the dispersal of environmental microbes (Parajuli *et al.*, 2020). At the regional to country scale, spatial patterns in microbiota composition can be attributed to differences in ethnicity

(Deschasaux *et al.*, 2018) and lifestyle (Gupta *et al.*, 2017), which both likely affect selection, through genetics and diet for example. At the global scale, the observed patterns are likely due to geographically variable microbial inputs from the environment and to selection through diet and cultural traditions (Gupta *et al.*, 2017; Senghor *et al.*, 2018). Although data on dispersal limitation are sparse for the human gut microbiota, this might play an important role in amplifying geospatial differences. Indeed, dispersal rates appear to differ between bacterial taxa in the human gut (Harris *et al.*, 2017), and in other mammals, dispersal limitation has been shown to contribute to interspecies differences in gut microbiota composition (Moeller *et al.*, 2017). Finally, if the differences in lifestyles and diets between industrialized and rural populations (O'Keefe *et al.*, 2015) are maintained over the timescales of microbial evolution, speciation through adaptation of the gut microbiota to an industrialized lifestyle (Sonnenburg and Sonnenburg, 2019) might also amplify geospatial community differences.

In addition to the scale-dependent relative importance of different processes, the scale of sampling also likely affects the phylogenetic or taxonomic scale of the observed differences (Ladau and Eloe-Fadrosh, 2019). For example, global variability in human gut microbiota composition can be reduced to broad community types separable at the genus level (Costea *et al.*, 2018), but diverging functional traits within populations may only be observable at the species level (Vieira-Silva *et al.*, 2016; Tett *et al.*, 2019). Furthermore, cohabiting individuals can share microbial species at the strain level (Truong *et al.*, 2017), and specific microbial strains can be stably present in the gut community of a host individual for decades (Koo *et al.*, 2019).

Accounting for scale

A consequence of the above is that the spatial grain of the study should guide the choice of its design, sampling and taxonomic resolution (Fig. 2), and the possible integration of information on the function and metabolic activity from 'omics data' (Knight *et al.*, 2018; Ladau and Eloe-Fadrosh, 2019). Indeed, while proper modelling techniques are instrumental in addressing the spatial aspects of microbiome research, a prerequisite is that the data enable these analyses. This point comes down to one of the basic principles of computer science, 'garbage in, garbage out', first noted well over a century ago (Babbage, 1864).

Identification of community members should be performed at the most accurate practically available resolution, as the lower units can always be hierarchically grouped at higher levels, for example to reduce the computational burden of the analysis. Strain-level

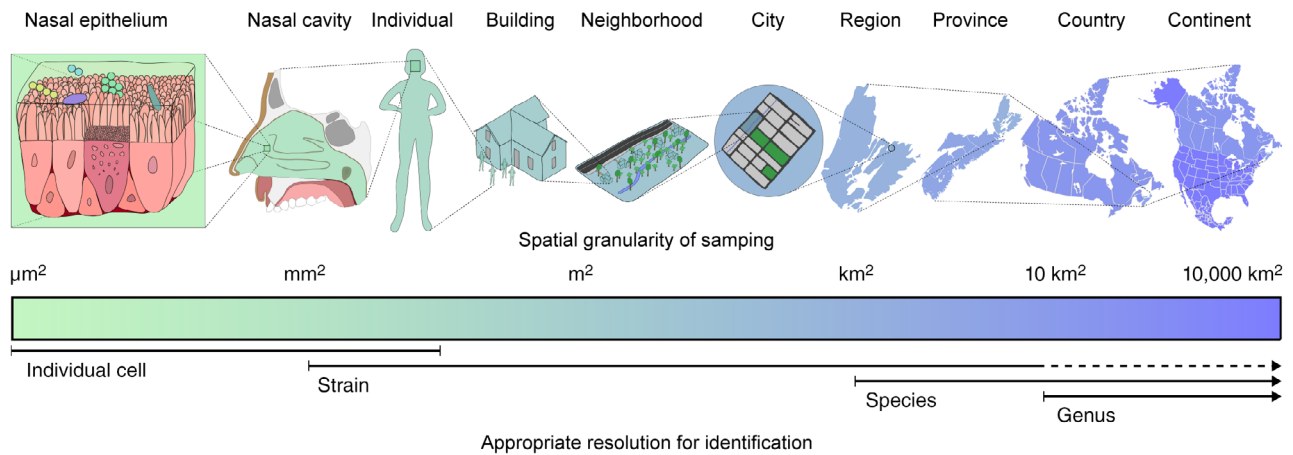


Fig 2. An example of spatial scales in the context of the human nasal microbiota and of the potentially appropriate taxonomic resolutions for their study. Patterns of variation in microbiota are scale-dependent, which should be considered in study design and data analysis. For example, species or genus level identification might not suffice to observe the patterns occurring at finer spatial scales. [Color figure can be viewed at wileyonlinelibrary.com]

identification of microbes is currently possible even from (deep) metagenomic sequencing of bulk samples (Anyansi *et al.*, 2020). For species-level identification, shallow shotgun sequencing has emerged as a viable alternative to 16S rRNA gene metabarcoding, with a higher taxonomic coverage and accuracy at only slightly increased sequencing cost (Hillmann *et al.*, 2018). Recent developments in long-read sequencing might facilitate the use of the full-length 16S rRNA gene in identifying microbes down to the species or even strain level (Johnson *et al.*, 2019). Finally, single-cell isolation, amplification and sequencing of either DNA (Xu and Zhao, 2018) or RNA (Ma *et al.*, 2019), which enables further functional characterization, can be applied to identify individual cells at high resolution.

Most microbiota studies currently use bulk samples, obtained for instance from faeces or by swabbing, to compare communities. While these approaches have proved useful in elucidating large-scale differences in microbiota composition between different hosts and even between sites within individual hosts, both faecal samples (Ingala *et al.*, 2018) and swabs (Prast-Nielsen *et al.*, 2019) are slightly biased proxies for the total community composition at the focal host sites. Thus, new minimally invasive methods and accurate sampling tools and technologies would be highly beneficial for this field of study (Tang *et al.*, 2020). New methods with a high spatial resolution at the micrometre to millimetre scales (Fig. 2) are required to better understand the composition, function and organization of the communities on the host surfaces. Sampling communities at these scales while preserving their spatial organization is inherently difficult, but methods such as fluorescence spectral imaging (Wilbert *et al.*, 2020) and sampling techniques such

as cryofracturing (Sheth *et al.*, 2019) have previously shown to be highly applicable.

Open questions in spatial microbiota ecology

While microbiota have been intensively studied over the past decade, our understanding of the ecological processes governing their composition (i.e. selection, drift, dispersal and speciation) and of their relative importance in different contexts and at different scales is still limited (Woodhams *et al.*, 2020). Key open questions in the field regard relate in particular to the importance of horizontal transfer between individuals, of functional redundancy between communities, of founder effects and stochasticity in community dynamics, and of rapid evolution within the host lifetime (Table 1). While much research effort currently focuses on humans, these are general questions that can (and should) also be addressed in other hosts, if only because they are easier to study experimentally. In addition to these general questions, it is of major interest for medical research to better understand geographical variation in the human microbiota and its significance to human health. Because these questions all involve a spatial aspect, the use of models that explicitly take spatial structure into account represents an important step towards addressing these goals.

Spatial modelling frameworks

Much of the methodological development in statistical microbiology has focused on between-sample comparisons assessing the effect of different conditions or treatments, disregarding the spatial and ecological contexts of

Table 1. Key open questions related to spatial variation in host-associated microbial communities.

<i>Processes shaping community composition:</i>	How does the relative importance of the four fundamental processes governing community composition, that is, selection, drift, speciation and dispersal, vary across spatial contexts (e.g. between host species, tissue types or body sites, environments, spatial scales)?
<i>Horizontal dispersal:</i>	How important is the horizontal dispersal of microbes between host individuals and species, and how does it depend on the characteristics of the microbes (e.g. physiology, relative abundance and activity) and on environmental conditions?
<i>Functional redundancy:</i>	How does the host selectively filter newly acquired microbes from the environment and how functionally similar are comparable microbiota (i.e. same host species and same host site) in different geographical locations?
<i>Founder effects and ecological succession:</i>	How do founder effects and interactions between pre-established and introduced microbes affect community assembly?
<i>Rapid microbial evolution:</i>	To which extent does microbial evolution taking place within the host influence its microbiota over short time scales (i.e. over the lifetime of the host or over a few host generations)?
<i>Medical relevance of spatial patterns:</i>	What is the extent of geographical variation in dysbiotic (i.e. disease-associated) human microbiota, and how could this variation affect the pathophysiology, diagnosis, prognosis and treatment of different diseases?

the microbial communities (Fernandes *et al.*, 2014). Machine learning methods, for example based on regression trees, are increasingly used for the ecological interpretation of microbial data (Marcos-Zambrano *et al.*, 2021; Moreno-Indias *et al.*, 2021), and so are network-based approaches inferring potential interactions between taxa (Kurtz *et al.*, 2015) but often without accounting for spatial structure. Macroecology (macro-organismal) community ecology and macroecology represent promising sources of inspiration for spatial models in microbial ecology. Community ecology and macroecology differ in that the former is concerned with the spatial and temporal scales associated with a community of locally coexisting organisms (these scales depend on the organisms considered), while macroecology is concerned with ecological patterns across scales - from the community scale to the global scale, and this for both macrobes and microbes. Despite the significant differences between micro- and macrobial ecology (Ladau and Eløe-Fadros, 2019), many of the methods developed to study communities of macrobes are potentially also applicable to microbial communities.

According to the conceptual synthesis in community ecology, 'species are added to communities via speciation and dispersal, and the relative abundances of these species are then shaped by drift and selection, as well as ongoing dispersal, to drive community dynamics' (Vellend, 2010). These processes are inherently the same regardless of the identity and size of the biological organisms. High-throughput sequencing methods now enable comprehensively assessing the composition of

microbial communities, which provides microbial ecologists with community composition data increasingly similar to those analysed in traditional community ecology. A number of technical limitations remain, such as uneven DNA extraction efficiency, PCR and sequencing errors, uneven taxonomic resolution, incomplete reference databases and the compositionality of the data (Knight *et al.*, 2018). While this may bias the ecological interpretation of the data (Sommeria-Klein *et al.*, 2016), the uncertainty thus introduced has been steadily decreasing and has now become, in the case of host-associated microbial communities, comparable in magnitude to that of traditional community ecology data (Rocchini *et al.*, 2011). Finally, the increasing use of DNA metabarcoding in plant and animal ecology further contributes to a convergence in data types and methodological approaches between microbial and macrobial ecology (Deiner *et al.*, 2017). This provides the opportunity for microbial ecologists to tap into the rich body of models accounting for a spatial structure that has been developed for macrobial community ecology.

We first review below the classical statistical methods used in macrobial ecology to account for spatial structure and their limitations. In addition to these methods mainly based on dissimilarity metrics and linear models, both macrobial and microbial ecology have seen a rising use of 'predictive modelling' approaches over the last decade. These approaches can be divided into two broad categories, which we review in the subsequent two sections in a spatial context: classical machine learning approaches, for instance, based on decision/regression trees or neural networks, and probabilistic modelling approaches, sometimes referred to as 'probabilistic machine learning' (Ghahramani, 2015). Both types of approaches rely on optimizing, or fitting, a potentially high-dimensional model to the data, however, in machine learning the inference is based on a learning algorithm, while in probabilistic modelling it is based on an explicit probabilistic model (i.e. a mathematical model that predicts the probability distribution of outcomes), which can be more easily constrained by assumptions about the data. Both types of approaches have in common the ability to readily reveal non-linear dependencies and interactions in the data and to make predictions to new data.

Classical statistical ecology

A common approach for the analysis of spatial community composition data in both macrobial and microbial ecology is to normalize taxa abundances per sample, compute pairwise dissimilarities in composition between samples (β -diversity) and perform analyses on the resulting dissimilarity matrix. The advantage of this approach in a spatial setting is that it easily enables

investigating the effect of spatial distance on the pairwise dissimilarities between samples. Classical analyses include simple statistical tests (e.g. Mantel tests against spatial distance or environmental dissimilarity), clustering (e.g. Hierarchical Clustering, Partitioning around Medoids) and ordination (e.g. Multidimensional Scaling; Legendre and Legendre, 2012). Until now, these methods have proved widely useful in analysing the high-dimensional data produced by high-throughput sequencing in microbial ecology (Paliy and Shankar, 2016).

Despite their established usefulness, dissimilarity-based approaches can also produce misleading results and obscure data interpretation (Warton *et al.*, 2012). It is particularly true in the case of microbial community data, which are characterized by compositionality, a high number of rare taxa leading to sparse composition matrices (i.e. with many zeros), and a strong heterogeneity in total read count across samples (Knight *et al.*, 2018). Moreover, most dissimilarity-based statistical methods do not allow incorporating additional data after analysis or making predictions on new samples. This limits their use, for instance, in medical diagnostics and environmental monitoring (Cullen *et al.*, 2020). Fully multivariate statistical approaches, in which the original composition of all samples is jointly analysed, are an alternative to dissimilarity-based methods. They are both less biased and more statistically powerful, especially when the samples are spatially distributed (Legendre *et al.*, 2005). In a fully multivariate approach, the spatial structure of the data can be accounted for by decomposing the matrix of between-sample spatial distances into a set of eigenvectors, called Moran's Eigenvector Maps or Principal Coordinates of Neighbour Matrices, to be used as explanatory variables representing the possible patterns of spatial autocorrelation associated with the sample layout (Legendre and Legendre, 2012). Standard multivariate statistical methods nevertheless assume linear relationships, which makes them inappropriate to model taxa distributions along spatial gradients when taxa abundances exhibit non-linear or even non-monotonous spatial trends (Austin, 2007; Paliy and Shankar, 2016). Furthermore, the commonly used multivariate methods cannot account for the multiple levels of spatial organization of host-associated communities, forming a nested hierarchy (Björk *et al.*, 2018).

Machine learning

Data-intensive research in microbial ecology often takes advantage of popular machine learning methods such as neural networks, decision trees, support vector machines, gradient boosting and ensembles of learners. These techniques have become increasingly popular due to their relatively easy adoption and the limited need for

human intervention during the analysis (Cordier *et al.*, 2019; Qu *et al.*, 2019). They are highly flexible and require little prior parameterization, which makes them well suited for studies with a limited understanding of the mechanisms at play and of the relative importance of the different variables, as is often the case in microbial ecosystems. They are also well suited to data sets with complex structure that exhibit non-linear dependencies and interactions between many variables. They can be used to identify useful properties from the data, such as the dependency between the abundances of taxonomic or functional groups and biometric, environmental, and spatial variables. These properties can then be used, for instance, to optimize model performance for diagnostic or prognostic in medical applications, or for environmental monitoring.

Machine learning methods often feature a large number of parameters with respect to the number of data points, which makes them highly flexible but also prone to overfitting the training data and generalizing poorly. To remediate this, model performance and accuracy are typically evaluated through cross-validation, that is, by quantifying how well the model generalizes to new observations (known as 'out-of-sample' data), rather than through goodness-of-fit to a single dataset (as measured by R^2 or a P -value in classical statistics). Care should nevertheless be exercised when dealing with small sample sizes (Vabalas *et al.*, 2019), or in the case of spatially autocorrelated data, in which case spatially disjoint training and test (validation) sets should be used to avoid overestimating model performance (Meyer *et al.*, 2018; Schratz *et al.*, 2019).

Despite their high performance in classification and regression tasks, the main limitation of these methods is that they function as 'black boxes': the fitted model has limited interpretability, and it does not usually account for the underlying mechanisms. The structure learned by the model can nevertheless be investigated. For instance, the relationship between input and output variables can be visualized through partial dependence plots, obtained by varying the input variables one at a time within the trained model (Greenwell, 2017). Such *a posteriori* investigations may help understand how specific variables and their interactions contribute to the final predictions. Importantly, they enable estimating effect sizes for individual or multiple interacting variables.

Few studies on microbiota using machine learning have so far incorporated spatial information, which can often be attributed to an insufficient number of samples for reliably detecting spatial patterns. Yet, including spatial data and analyses in studies with an adequate sample size can lead to remarkable performance gains. A study using random forests for disease diagnosis in a Chinese province showed, for instance, that the

classification accuracy improved as finer spatial scales were considered, from the regional to the neighbourhood scale (He *et al.*, 2018). It also found that extrapolating locally trained models to larger geographic areas led to poorer performance. Variable selection with random forests and gradient boosting trees was also used in the analysis of gut microbiota from a Finnish population cohort to predict fatty liver disease across geographical regions (Ruuskanen *et al.*, 2021), and ensemble logistic regression was used to trace the geographical origin of clams based on 16S rRNA metabarcoding data on their microbiota (Milan *et al.*, 2019). These studies, however, incorporate spatial information as discrete location information rather than as continuous variables, and accounting for the spatial structure of the data more explicitly could further improve model performance.

Probabilistic modelling

The main limitations of classical machine learning methods is that they poorly estimate uncertainty and that the underlying models are either implicit or difficult to interpret. An alternative is to rely on an explicit probabilistic model, associated with a likelihood function. Inferences can then be made on the data through likelihood maximization, or through Bayesian inference provided that prior distributions have been specified for the inferred parameters. Probabilistic modelling allows providing rigorous uncertainty estimates but also guiding inference with *a priori* knowledge on data structure or on the mechanisms at play, and thus giving a clearer biological interpretation to the inferred parameters. While classical statistics also relies on fitting a (sometimes implicit) probabilistic model to data, increasing computing power is now allowing for more and more complex models, which may rely on non-normal distributions, accommodate non-linear relationships between variables and be hierarchically structured (Gelman, 2014). As in non-probabilistic machine learning, it has become a common approach to fit highly flexible models and to assess their generalizability through cross-validation (Ghahramani, 2015). An alternative is to fit models that are more strongly constrained by hypotheses about the data, and to then compare either their goodness-of-fit or their out-of-sample predictive performance to reveal the hypothesis most consistent with the data.

The explicit probabilistic modelling of the spatial variation in host-associated microbial communities, and of their scale-dependent relationship with the host and the environment, can be achieved through Species Distribution Models (SDMs) borrowed from macrobial ecology. SDMs have long been used to predict the spatial distribution of species based on observed species occurrences and bioclimatic variables (Miller, 2010). Nevertheless,

simple bioclimatic models cannot capture the effect of many factors affecting species distributions, such as biotic interactions and dispersal limitation (Pearson and Dawson, 2003). This has led to the introduction of hierarchical models able to incorporate these factors in a scale-dependent way while accounting for multiple sources of uncertainty in the data (Hefley and Hooten, 2016). One of the latest developments of this line of research is Joint Species Distribution Models (JSDMs), which enable the joint estimation of the distribution of multiple species based on both abiotic conditions and biotic interactions (Latimer *et al.*, 2009; Ovaskainen and Abrego, 2020). From a technical standpoint, JSDMs are generalized linear mixed models, in which the spatial structure of the data can be accounted for through the covariance matrix of the residuals. JSDMs can be applied to both count and presence–absence data. They can account for environmental covariates, functional traits and phylogenetic relationships between the organisms, and produce model-based variance partitions, ordinations and co-occurrence networks as output.

While the computational costs of the earlier JSDMs were intractable for microbial data, it is now possible to handle hundreds of taxa and samples in a reasonable time (Tikhonov *et al.*, 2020a; Tikhonov *et al.*, 2020b). A few recent studies applied JSDMs to investigate spatial patterns in microbiota (Björk *et al.*, 2018; Aivelo *et al.*, 2019; Minard *et al.*, 2019). In particular, a recent study adapted JSDMs to microbiota by incorporating host phylogeny and traits and illustrated this development on bird and sponge microbiota (Björk *et al.*, 2018). JSDMs were also used to show that variation in the abundance of microbial taxa in tick-associated microbiota is mostly associated with host-specific factors, although environmental effects can be large for individual microbes, including human pathogens (Aivelo *et al.*, 2019). This study demonstrates the use of JSDMs to partition variance between spatial effects and host-related factors and to obtain co-occurrence networks. A study conducted on caterpillar microbiota revealed phylogenetic structuring in the communities, with related microbial taxa exhibiting similar patterns (Minard *et al.*, 2019). The communities displayed high variation between individual caterpillars, on which neither the host- and host plant-related factors nor spatial structure appeared to have significant influence. These studies illustrate the potential of JSDMs to model microbiota, including microbe-to-microbe interactions and the relative effect of different processes on the occurrence or abundance of microbial taxa at different scales.

The use of JSDMs for host-associated microbial ecology has nevertheless a number of limitations. First, microbial communities tend to comprise a higher share of rare taxa than communities of macrobes. Although latent

variables and inter-taxa associations can be used to improve predictions on rare taxa (Tikhonov *et al.*, 2017), most taxa are likely to occur too sparsely to be amenable to analysis with JSDM unless it is performed at a coarse enough taxonomic resolution (by grouping at higher levels). Second, current JSDMs do not explicitly account for the compositionality of microbial community data, which may bias the inference (Björk *et al.*, 2018). Third and finally, they only allow host-associated factors to influence the microbiota but not the other way around (Aivelo *et al.*, 2019), and evolutionary processes are not accounted for, which can limit their predictive potential (Cotto *et al.*, 2020). The latter limitation is a stronger concern when dealing with microbes compared with macrobes, as the timescale of their evolutionary adaptation is much shorter (Ferreiro *et al.*, 2018).

Other probabilistic modelling approaches have been developed to account for the specificities of microbial data (compositionality, a highly heterogeneous read count across samples and many rare taxa), although they do not yet explicitly account for spatial structure. They usually model the sampling process explicitly using probability distributions belonging to the Dirichlet-multinomial family (La Rosa *et al.*, 2012). This forms the mathematical foundation for various model-based analyses of microbiota: reconstruction of association networks (Kurtz *et al.*, 2015), classification of microbiota into discrete categories based on their composition (Holmes *et al.*, 2012; Ding and Schloss, 2014) and construction of assemblages of taxa based on their co-occurrence and covariance across samples (Hosoda *et al.*, 2020). Assemblage models are a particularly interesting alternative to taxon-centric models for modelling high-diversity microbial datasets (Sommeria-Klein *et al.*, 2020). The resulting assemblages may be interpreted as groups of microbes with the same ecological niche, and the decomposition into assemblages strongly reduces the dimensionality of the data for downstream analyses. Finally, neutral ecological models describe the stochastic dynamics of ecological communities - including dispersal, drift and speciation - under the assumption that all taxa are equivalent in their competitive abilities. They yield stationary distributions belonging to the Dirichlet-multinomial family for the composition of communities, and they have been used for the ecological interpretation of human gut microbiota data (Harris *et al.*, 2017).

Perspectives

In the light of advances in other research fields, the potential of predictive modelling for the analysis of spatial data still appears largely underexploited in microbiota studies. For example, random forest approaches have been accurate in predicting regional lithology in Australia

using continuous spatial information (Cracknell and Reading, 2014), or the spread of a forest disease (*Sphaeropsis* blight) in Spain using spatial cross-validation (Schratz *et al.*, 2019). In another example, a geographically weighted ensemble of deep neural networks, gradient boosting trees and random forests accurately predicted temporal wind speeds over mainland China (Li, 2019). The application of frameworks such as these could possibly elucidate the drivers behind the spatial distribution of host-associated microbial community diversity or of individual taxa in these communities. In disease models where microbiota composition is used as a diagnostic tool, we posit that spatial structure should be better accounted for in study design. Merely incorporating spatial data in the current machine learning frameworks as a proxy for unmeasured spatially correlated variables could already improve their performance.

Likewise, a number of extensions and modifications to JSDMs could likely improve their performance for the high number of taxa that characterizes microbial studies. While a generalized linear modelling framework is usually at the core of JSDM models, their performance can be further improved by using Gaussian processes instead (Ingram *et al.*, 2020; Vanhatalo *et al.*, 2020). Advanced computational techniques such as Integrated Nested Laplace Approximation (Blangiardo *et al.*, 2013) could also be used to enhance their computational efficiency. Furthermore, the use of log-ratio transforms to accommodate compositional data in an unbiased way (Gloor *et al.*, 2017), and of Gaussian processes to quantify autocorrelation between hosts (as suggested by Björk *et al.*, 2018) would increase the suitability of JSDMs to the study of microbiota. Other types of models, such as source tracking models aiming at identifying the origin of contaminants in microbial samples (Knights *et al.*, 2011), could be used to model the effect of dispersal between hosts in a spatial context. Finally, further use of these models to assess spatial effects would rely on and benefit from more even and intensive sampling of communities at spatial scales relevant to the study questions, similarly to macrobial ecology studies (see, e.g. Tikhonov *et al.*, 2020a).

Concluding remarks

Host-associated microbial communities vary greatly in space (and time), even at a single host site in a single host species. This variability can now be observed with the use of various high-throughput sequencing and single-cell sampling and imaging methods, but its causes remain largely unclear. It is likely that patterns in these communities could be better understood if their spatial structure were properly incorporated in the analyses. Spatial data in microbiota studies can both reflect the

varying ability of the organisms to disperse between and within hosts and serve as a proxy for unknown or unmeasured spatially correlated variables. Recent developments in spatial analysis enable accounting for the scale-dependent hierarchical structure of microbiota and for non-linear interactions between variables, but these approaches are still greatly underused in microbial ecology. Indeed, the complexity of microbial community data, the limited scalability of the methods, and the lack of openly available implementations and benchmark case studies are slowing down the development of the field. Further development of computational efficiency, adjustment to the specific properties of microbiota profiling data and the incorporation of evolutionary processes would facilitate the use of these methods in the spatial modelling of microbiota. Their growing use in microbial ecology could in return spur new methodological development and applications in macrobial ecology, as well as industrial and clinical applications.

Author Contributions

M.O.R., G.S.-K., L.L. and T.J.N. designed the work. M.O.R. and G.S.-K. wrote the manuscript. M.O.R. drafted the figures. A.S.H., T.J.N. and L.L. provided critical feedback and suggestions on the paper. L.L. and T.J.N. supervised the work. All authors gave final approval of the version to be published.

References

Adair, K.L., and Douglas, A.E. (2017) Making a microbiome: the many determinants of host-associated microbial community composition. *Curr Opin Microbiol* **35**: 23–29.

Avelo, T., Norberg, A., and Tschirren, B. (2019) Bacterial microbiota composition of Ixodes ricinus ticks: the role of environmental variation, tick characteristics and microbial interactions. *PeerJ* **7**: e8217.

Albenberg, L., Esipova, T.V., Judge, C.P., Bittinger, K., Chen, J., Laughlin, A., *et al.* (2014) Correlation between intraluminal oxygen gradient and radial partitioning of intestinal microbiota. *Gastroenterology* **147**: 1055–1063.e8.

Antwis, R.E., Lea, J.M.D., Unwin, B., and Shultz, S. (2018) Gut microbiome composition is associated with spatial structuring and social interactions in semi-feral Welsh Mountain ponies. *Microbiome* **6**: 207.

Anyansi, C., Straub, T.J., Manson, A.L., Earl, A.M., and Abeel, T. (2020) Computational methods for strain-level microbial detection in colony and metagenome sequencing data. *Front Microbiol* **11**: 1925.

Austin, M. (2007) Species distribution models and ecological theory: a critical assessment and some possible new approaches. *Ecol Model* **200**: 1–19.

Babbage, C. (1864) *Passages from the Life of a Philosopher*. Cambridge: Cambridge University Press.

Bakker, M.G., Schlatter, D.C., Otto-Hanson, L., and Kinkel, L.L. (2014) Diffuse symbioses: roles of plant–plant, plant–microbe and microbe–microbe interactions in structuring the soil microbiome. *Mol Ecol* **23**: 1571–1583.

Benson, A.K., Kelly, S.A., Legge, R., Ma, F., Low, S.J., Kim, J., *et al.* (2010) Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. *Proc Natl Acad Sci U S A* **107**: 18933–18938.

Berg, G., Rybakova, D., Fischer, D., Cernava, T., Vergès, M.-C.C., Charles, T., *et al.* (2020) Microbiome definition re-visited: old concepts and new challenges. *Microbiome* **8**: 103.

Björk, J.R., Hui, F.K.C., O'Hara, R.B., and Montoya, J.M. (2018) Uncovering the drivers of host-associated microbiota with joint species distribution modelling. *Mol Ecol* **27**: 2714–2724.

Blangiardo, M., Cameletti, M., Baio, G., and Rue, H. (2013) Spatial and spatio-temporal models with R-INLA. *Spat Spatio-temporal Epidemiol* **4**: 33–49.

Bright, M., and Bulgheresi, S. (2010) A complex journey: transmission of microbial symbionts. *Nat Rev Microbiol* **8**: 218–230.

Callens, M., Watanabe, H., Kato, Y., Miura, J., and Decaestecker, E. (2018) Microbiota inoculum composition affects holobiont assembly and host growth in *Daphnia*. *Microbiome* **6**: 56.

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Lozupone, C.A., Turnbaugh, P.J., *et al.* (2011) Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci U S A* **108**: 4516–4522.

Chiu, K., Warner, G., Nowak, R.A., Flaws, J.A., and Mei, W. (2020) The impact of environmental chemicals on the gut microbiome. *Toxicol Sci* **176**: 253–284.

Cordier, T., Lanzén, A., Apothéloz-Perret-Gentil, L., Stoeck, T., and Pawlowski, J. (2019) Embracing environmental genomics and machine learning for routine bio-monitoring. *Trends Microbiol* **27**: 387–397.

Costea, P.I., Hildebrand, F., Arumugam, M., Bäckhed, F., Blaser, M.J., Bushman, F.D., *et al.* (2018) Enterotypes in the landscape of gut microbial community composition. *Nat Microbiol* **3**: 8–16.

Cotto, O., Schmid, M., and Guillaume, F. (2020) Nemo-age: spatially explicit simulations of eco-evolutionary dynamics in stage-structured populations under changing environments. *Methods Ecol Evol* **11**: 1227–1236.

Coyte, K.Z., Schluter, J., and Foster, K.R. (2015) The ecology of the microbiome: networks, competition, and stability. *Science* **350**: 663–666.

Cracknell, M.J., and Reading, A.M. (2014) Geological mapping using remote sensing data: a comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. *Comput Geosci* **63**: 22–33.

Cullen, C.M., Aneja, K.K., Beyhan, S., Cho, C.E., Woloszynek, S., Convertino, M., *et al.* (2020) Emerging priorities for microbiome research. *Front Microbiol* **11**: 136.

DataBase Center for Life Science (2018) *Human (female)* [WWW document]. URL <https://commons.wikimedia.org/w/index.php?curid=70908949>.

- David, L.A., Materna, A.C., Friedman, J., Campos-Baptista, M.I., Blackburn, M.C., Perrotta, A., et al. (2014) Host lifestyle affects human microbiota on daily time-scales. *Genome Biol* **15**: R89.
- Deiner, K., Bik, H.M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., et al. (2017) Environmental DNA metabarcoding: transforming how we survey animal and plant communities. *Mol Ecol* **26**: 5872–5895.
- Deschasaux, M., Bouter, K.E., Prodan, A., Levin, E., Groen, A.K., Herrema, H., et al. (2018) Depicting the composition of gut microbiota in a population with varied ethnic origins but shared geography. *Nat Med* **24**: 1526–1531.
- Ding, T., and Schloss, P.D. (2014) Dynamics and associations of microbial community types across the human body. *Nature* **509**: 357–360.
- Donaldson, G.P., Lee, S.M., and Mazmanian, S.K. (2016) Gut biogeography of the bacterial microbiota. *Nat Rev Microbiol* **14**: 20–32.
- Falony, G., Joossens, M., Vieira-Silva, S., Wang, J., Darzi, Y., Faust, K., et al. (2016) Population-level analysis of gut microbiome variation. *Science* **352**: 560–564.
- Feng, Q., Chen, W.-D., and Wang, Y.-D. (2018) Gut microbiota: an integral moderator in health and disease. *Front Microbiol* **9**: 151.
- Fernandes, A.D., Reid, J.N., Macklaim, J.M., McMurrrough, T.A., Edgell, D.R., and Gloor, G.B. (2014) Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome* **2**: 15.
- Ferreiro, A., Crook, N., Gasparri, A.J., and Dantas, G. (2018) Multiscale evolutionary dynamics of host-associated microbiomes. *Cell* **172**: 1216–1227.
- Filippo, C.D., Cavalieri, D., Paola, M.D., Ramazzotti, M., Poullet, J.B., Massart, S., et al. (2010) Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc Natl Acad Sci U S A* **107**: 14691–14696.
- Finnicum, C.T., Beck, J.J., Dolan, C.V., Davis, C., Willemsen, G., Ehli, E.A., et al. (2019) Cohabitation is associated with a greater resemblance in gut microbiota which can impact cardiometabolic and inflammatory risk. *BMC Microbiol* **19**: 230.
- Fischbach, M.A., and Segre, J.A. (2016) Signaling in host-associated microbial communities. *Cell* **164**: 1288–1300.
- Garcia, E.C. (2018) Contact-dependent interbacterial toxins deliver a message. *Curr Opin Microbiol* **42**: 40–46.
- Garcia, J.R., and Gerardo, N.M. (2014) The symbiont side of symbiosis: do microbes really benefit? *Front Microbiol* **5**: 510.
- Gaulke, C.A., and Sharpton, T.J. (2018) The influence of ethnicity and geography on human gut microbiome composition. *Nat Med* **24**: 1495–1496.
- Gelman, A. (2014) *Bayesian data analysis*, 3rd ed. Boca Raton: CRC Press.
- Ghahramani, Z. (2015) Probabilistic machine learning and artificial intelligence. *Nature* **521**: 452–459.
- Gloor, G.B., Macklaim, J.M., Pawlowsky-Glahn, V., and Egozcue, J.J. (2017) Microbiome datasets are compositional: and this is not optional. *Front Microbiol* **8**: 2224.
- Gonze, D., Coyte, K.Z., Lahti, L., and Faust, K. (2018) Microbial communities as dynamical systems. *Curr Opin Microbiol* **44**: 41–49.
- Greenwell, B.M. (2017) Pdp: an R package for constructing partial dependence plots. *R J* **9**: 421–436.
- Griffiths, S.M., Harrison, X.A., Weldon, C., Wood, M.D., Pretorius, A., Hopkins, K., et al. (2018) Genetic variability and ontogeny predict microbiome structure in a disease-challenged montane amphibian. *ISME J* **12**: 2506–2517.
- Gupta, V.K., Paul, S., and Dutta, C. (2017) Geography, ethnicity or subsistence-specific variations in human microbiome composition and diversity. *Front Microbiol* **8**: 1162.
- Hacquard, S. (2016) Disentangling the factors shaping microbiota composition across the plant holobiont. *New Phytol* **209**: 454–457.
- Hall, M.A., Brettell, L.E., Liu, H., Nacko, S., Spooner-Hart, R., Riegler, M., and Cook, J.M. (2021) Temporal changes in the microbiome of stingless bee foragers following colony relocation. *FEMS Microbiol Ecol* **97**: fiae236.
- Harris, K., Parsons, T.L., Ijaz, U.Z., Lahti, L., Holmes, I., and Quince, C. (2017) Linking statistical and ecological theory: Hubbell's unified neutral theory of biodiversity as a hierarchical Dirichlet process. *Proc IEEE* **105**: 516–529.
- He, Y., Wu, W., Zheng, H.-M., Li, P., McDonald, D., Sheng, H.-F., et al. (2018) Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nat Med* **24**: 1532–1535.
- Hefley, T.J., and Hooten, M.B. (2016) Hierarchical species distribution models. *Curr Landscape Ecol Rep* **1**: 87–97.
- Hillmann, B., Al-Ghalith, G.A., Shields-Cutler, R.R., Zhu, Q., Gohl, D.M., Beckman, K.B., et al. (2018) Evaluating the information content of shallow shotgun metagenomics. *mSystems* **3**: e00069-18.
- Holmes, I., Harris, K., and Quince, C. (2012) Dirichlet multinomial mixtures: generative models for microbial Metagenomics. *PLoS One* **7**: e30126.
- Hooper, L.V., Littman, D.R., and Macpherson, A.J. (2012) Interactions between the microbiota and the immune system. *Science* **336**: 1268–1273.
- Hosoda, S., Nishijima, S., Fukunaga, T., Hattori, M., and Hamada, M. (2020) Revealing the microbial assemblage structure in the human gut microbiome using latent Dirichlet allocation. *Microbiome* **8**: 95.
- Ingala, M.R., Simmons, N.B., Wultsch, C., Krampis, K., Speer, K.A., and Perkins, S.L. (2018) Comparing microbiome sampling methods in a wild mammal: fecal and intestinal samples record different signals of host ecology, evolution. *Front Microbiol* **9**: 803.
- Ingram, M., Vukcevic, D., and Golding, N. (2020) Multi-output Gaussian processes for species distribution modelling. *Methods Ecol Evol* **11**: 1587–1598.
- Johnson, J.S., Spakowicz, D.J., Hong, B.-Y., Petersen, L.M., Demkowicz, P., Chen, L., et al. (2019) Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat Commun* **10**: 5029.
- Kaplan, R.C., Wang, Z., Usyk, M., Sotres-Alvarez, D., Daviglius, M.L., Schneiderman, N., et al. (2019) Gut microbiome composition in the Hispanic community health study/study of Latinos is shaped by geographic relocation, environmental factors, and obesity. *Genome Biol* **20**: 219.

- Knight, R., Vrbanac, A., Taylor, B.C., Aksenov, A., Callewaert, C., Debelius, J., *et al.* (2018) Best practices for analysing microbiomes. *Nat Rev Microbiol* **16**: 410–422.
- Knight, D., Kuczynski, J., Charlson, E.S., Zaneveld, J., Mozer, M.C., Collman, R.G., *et al.* (2011) Bayesian community-wide culture-independent microbial source tracking. *Nat Methods* **8**: 761–763.
- Koo, H., Hakim, J.A., Crossman, D.K., Lefkowitz, E.J., and Morrow, C.D. (2019) Sharing of gut microbial strains between selected individual sets of twins cohabitating for decades. *PLOS One* **14**: e0226111.
- Kurtz, Z.D., Müller, C.L., Miraldi, E.R., Littman, D.R., Blaser, M.J., and Bonneau, R.A. (2015) Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput Biol* **11**: e1004226.
- La Rosa, P.S., Brooks, J.P., Deych, E., Boone, E.L., Edwards, D.J., Wang, Q., *et al.* (2012) Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PLOS One* **7**: e52078.
- Ladau, J., and Eloe-Fadrosh, E.A. (2019) Spatial, temporal, and phylogenetic scales of microbial ecology. *Trends Microbiol* **27**: 662–669.
- Langille, M.G., Meehan, C.J., Koenig, J.E., Dhanani, A.S., Rose, R.A., Howlett, S.E., and Beiko, R.G. (2014) Microbial shifts in the aging mouse gut. *Microbiome* **2**: 50.
- Latimer, A.M., Banerjee, S., Sang, H., Jr., Mosher, E.S., and Jr, J.A.S. (2009) Hierarchical models facilitate spatial analysis of large data sets: a case study on invasive plant species in the northeastern United States. *Ecol Lett* **12**: 144–154.
- Leftwich, P.T., Edgington, M.P., and Chapman, T. (2020) Transmission efficiency drives host–microbe associations. *Proc R Soc B: Biol Sci* **287**: 20200820.
- Legendre, P., Borcard, D., and Peres-Neto, P.R. (2005) Analyzing beta diversity: partitioning the spatial variation of community composition data. *Ecol Monogr* **75**: 435–450.
- Legendre, P., and Legendre, L. (2012) *Numerical Ecology, Third English Edition*. Amsterdam: Elsevier.
- Lehtimäki, J., Sinkko, H., Hielm-Björkman, A., Salmela, E., Tiira, K., Laatikainen, T., *et al.* (2018) Skin microbiota and allergic symptoms associate with exposure to environmental microbes. *Proc Natl Acad Sci U S A* **115**: 4897–4902.
- Li, L. (2019) Geographically weighted machine learning and downscaling for high-resolution spatiotemporal estimations of wind speed. *Remote Sens (Basel)* **11**: 1378.
- Litvak, Y., and Bäuml, A.J. (2019) The founder hypothesis: a basis for microbiota resistance, diversity in taxa carriage, and colonization resistance against pathogens. *PLoS Pathog* **15**: e1007563.
- Ma, Q., Bücking, H., Gonzalez Hernandez, J.L., and Subramanian, S. (2019) Single-cell RNA sequencing of plant-associated bacterial communities. *Front Microbiol* **10**: 2452.
- Malla, M.A., Dubey, A., Kumar, A., Yadav, S., Hashem, A., and Abd Allah, E.F. (2019) Exploring the human microbiome: the potential future role of next-generation sequencing in disease diagnosis and treatment. *Front Immunol* **9**: 9.
- Marcos-Zambrano, L.J., Karadzovic-Hadziabdic, K., Loncar Turukalo, T., Przymus, P., Trajkovic, V., Aasmets, O., *et al.* (2021) Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment. *Front Microbiol* **12**: 313.
- Maurice, C.F., CL Knowles, S., Ladau, J., Pollard, K.S., Fenton, A., Pedersen, A.B., and Turnbaugh, P.J. (2015) Marked seasonal variation in the wild mouse gut microbiota. *ISME J* **9**: 2423–2434.
- McFall-Ngai, M., Hadfield, M.G., Bosch, T.C.G., Carey, H.V., Domazet-Lošo, T., Douglas, A.E., *et al.* (2013) Animals in a bacterial world, a new imperative for the life sciences. *Proc Natl Acad Sci U S A* **110**: 3229–3236.
- McFall-Ngai, M.J., and Ruby, E.G. (1991) Symbiont recognition and subsequent morphogenesis as early events in an animal-bacterial mutualism. *Science* **254**: 1491–1494.
- Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., and Naus, T. (2018) Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. *Environ Model Software* **101**: 1–9.
- Milan, M., Maroso, F., Dalla Rovere, G., Carraro, L., Ferrareso, S., Patarnello, T., *et al.* (2019) Tracing seafood at high spatial resolution using NGS-generated data and machine learning: comparing microbiome versus SNPs. *Food Chem* **286**: 413–420.
- Miller, J. (2010) Species distribution modeling. *Geogr Compass* **4**: 490–509.
- Minard, G., Tikhonov, G., Ovaskainen, O., and Saastamoinen, M. (2019) Variation in *Melitaea cinxia* gut microbiota is phylogenetically highly structured but only mildly driven by host plant microbiota, sex or parasitism. *bioRxiv* 510446.
- Moeller, A.H., Suzuki, T.A., Lin, D., Lacey, E.A., Wasser, S. K., and Nachman, M.W. (2017) Dispersal limitation promotes the diversification of the mammalian gut microbiota. *Proc Natl Acad Sci U S A* **114**: 13768–13773.
- Moreno-Indias, I., Lahti, L., Nedyalkova, M., Elbere, I., Roshchupkin, G., Adilovic, M., *et al.* (2021) Statistical and machine learning techniques in human microbiome studies: contemporary challenges and solutions. *Front Microbiol* **12**: 635781.
- Moya, A., and Ferrer, M. (2016) Functional redundancy-induced stability of gut microbiota subjected to disturbance. *Trends Microbiol* **24**: 402–413.
- Nordwestern (2015) *A political Map of Europe in SVG format without disputed areas and conflict regions* [WWW document]. URL <https://commons.wikimedia.org/w/index.php?curid=39655163>.
- NuclearVacuum (2008) *Blank map of Antarctica* [WWW document]. URL <https://commons.wikimedia.org/w/index.php?curid=8799641>.
- O’Keefe, S.J.D., Li, J.V., Lahti, L., Ou, J., Carbonero, F., Mohammed, K., *et al.* (2015) Fat, fibre and cancer risk in African Americans and rural Africans. *Nat Commun* **6**: 6342.
- Ortiz-Álvarez, R., Fierer, N., de los Ríos, A., Casamayor, E.O., and Barberán, A. (2018) Consistent changes in the taxonomic structure and functional attributes of bacterial communities during primary succession. *ISME J* **12**: 1658–1667.
- Ovaskainen, O., and Abrego, N. (2020) *Joint species distribution modelling: With applications in R*. Cambridge: Cambridge University Press.

- Paliy, O., and Shankar, V. (2016) Application of multivariate statistical techniques in microbial ecology. *Mol Ecol* **25**: 1032–1057.
- Parajuli, A., Hui, N., Puhakka, R., Oikarinen, S., Grönroos, M., Selonen, V.A.O., et al. (2020) Yard vegetation is associated with gut microbiota composition. *Sci Total Environ* **713**: 136707.
- Pearson, R.G., and Dawson, T.P. (2003) Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Glob Ecol Biogeogr* **12**: 361–371.
- Perez-Lamarque, B., and Morlon, H. (2019) Characterizing symbiont inheritance during host–microbiota evolution: application to the great apes gut microbiota. *Mol Ecol Resour* **19**: 1659–1671.
- Perofsky, A.C., Lewis, R.J., and Meyers, L.A. (2019) Terrestriality and bacterial transfer: a comparative study of gut microbiomes in sympatric Malagasy mammals. *ISME J* **13**: 50–63.
- Prast-Nielsen, S., Tobin, A.-M., Adamzik, K., Powles, A., Hugerth, L.W., Sweeney, C., et al. (2019) Investigation of the skin microbiome: swabs vs. biopsies. *Br J Dermatol* **181**: 572–579.
- Proctor, D.M., and Relman, D.A. (2017) The landscape ecology and microbiota of the human nose, mouth, and throat. *Cell Host Microbe* **21**: 421–432.
- Qu, K., Guo, F., Liu, X., Lin, Y., and Zou, Q. (2019) Application of machine learning in microbiology. *Front Microbiol* **10**: 827.
- Quince, C., Walker, A.W., Simpson, J.T., Loman, N.J., and Segata, N. (2017) Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol* **35**: 833–844.
- Rakoff-Nahoum, S., Foster, K.R., and Comstock, L.E. (2016) The evolution of cooperation within the gut microbiota. *Nature* **533**: 255–259.
- Rashidi, P. (2017) *Alone tree* [WWW document]. URL <https://commons.wikimedia.org/w/index.php?curid=71468868>.
- Riaz Rajoka, M.S., Shi, J., Mehwish, H.M., Zhu, J., Li, Q., Shao, D., et al. (2017) Interaction between diet composition and gut microbiota and its impact on gastrointestinal tract health. *Food Sci Human Wellness* **6**: 121–130.
- Rocchini, D., Hortal, J., Lengyel, S., Lobo, J.M., Jiménez-Valverde, A., Ricotta, C., et al. (2011) Accounting for uncertainty when mapping species distributions: the need for maps of ignorance. *Prog Phys Geogr: Earth Environ* **35**: 211–226.
- Ruuskanen, M.O., Åberg, F., Männistö, V., Havulinna, A.S., Méric, G., Liu, Y., et al. (2021). Links between gut microbiome composition and fatty liver disease in a large population sample. *Gut Microbes* **13**: 1–22. <https://doi.org/10.1080/19490976.2021.1888673>.
- Schratz, P., Muenchow, J., Iturriza, E., Richter, J., and Brenning, A. (2019) Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecol Model* **406**: 109–120.
- Senghor, B., Sokhna, C., Ruimy, R., and Lagier, J.-C. (2018) Gut microbiota diversity according to dietary habits and geographical provenance. *Human Microbiome J* **7–8**: 1–9.
- Sepulveda, J., and Moeller, A.H. (2020) The effects of temperature on animal gut microbiomes. *Front Microbiol* **11**: 384.
- Sheth, R.U., Li, M., Jiang, W., Sims, P.A., Leong, K.W., and Wang, H.H. (2019) Spatial metagenomic characterization of microbial biogeography in the gut. *Nat Biotechnol* **37**: 877–883.
- Silar (2016) *Hundesport in Beskiden* [WWW document]. URL <https://commons.wikimedia.org/w/index.php?curid=49121882>.
- Simon, J.-C., Marchesi, J.R., Mougel, C., and Selosse, M.-A. (2019) Host-microbiota interactions: from holobiont theory to analysis. *Microbiome* **7**: 5.
- Smith, C.C., Snowberg, L.K., Gregory Caporaso, J., Knight, R., and Bolnick, D.I. (2015) Dietary input of microbes and host genetic variation shape among-population differences in stickleback gut microbiota. *ISME J* **9**: 2515–2526.
- Sommeria-Klein, G., Zinger, L., Coissac, E., Iribar, A., Schimann, H., Taberlet, P., and Chave, J. (2020) Latent Dirichlet allocation reveals spatial and taxonomic structure in a DNA-based census of soil biodiversity from a tropical forest. *Mol Ecol Resour* **20**: 371–386.
- Sommeria-Klein, G., Zinger, L., Taberlet, P., Coissac, E., and Chave, J. (2016) Inferring neutral biodiversity parameters using environmental DNA data sets. *Sci Rep* **6**: 35644.
- Song, S.J., Sanders, J.G., Delsuc, F., Metcalf, J., Amato, K., Taylor, M.W., et al. (2020) Comparative analyses of vertebrate gut microbiomes reveal convergence between birds and bats. *MBio* **11**: e02901-19.
- Sonnenburg, E.D., and Sonnenburg, J.L. (2019) The ancestral and industrialized gut microbiota and implications for human health. *Nat Rev Microbiol* **17**: 383–390.
- Sylvain, F.-É., Cheaib, B., Llewellyn, M., Gabriel Correia, T., Barros Fagundes, D., Luis Val, A., and Derome, N. (2016) pH drop impacts differentially skin and gut microbiota of the Amazonian fish tambaqui (*Colossoma macropomum*). *Sci Rep* **6**: 32032.
- Tang, Q., Jin, G., Wang, G., Liu, T., Liu, X., Wang, B., and Cao, H. (2020) Current sampling methods for gut microbiota: a call for more precise devices. *Front Cell Infect Microbiol* **10**: 151.
- Tett, A., Huang, K.D., Asnicar, F., Fehlner-Peach, H., Pasolli, E., Karcher, N., et al. (2019) The *Prevotella* copri complex comprises four distinct clades underrepresented in westernized populations. *Cell Host Microbe* **26**: 666–679.e7.
- Tikhonov, G., Abrego, N., Dunson, D., and Ovaskainen, O. (2017) Using joint species distribution models for evaluating how species-to-species associations depend on the environmental context. *Methods Ecol Evol* **8**: 443–452.
- Tikhonov, G., Duan, L., Abrego, N., Newell, G., White, M., Dunson, D., and Ovaskainen, O. (2020a) Computationally efficient joint species distribution modeling of big spatial data. *Ecology* **101**: e02929.
- Tikhonov, G., Opedal, Ø.H., Abrego, N., Lehtikainen, A., Jonge, M.M.J., de Oksanen, J., and Ovaskainen, O. (2020b) Joint species distribution modelling with the r-package Hmsc. *Methods Ecol Evol* **11**: 442–447.
- Truong, D.T., Tett, A., Pasolli, E., Huttenhower, C., and Segata, N. (2017) Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res* **27**: 626–638.

- Tung, J., Barreiro, L.B., Burns, M.B., Grenier, J.-C., Lynch, J., Grieneisen, L.E., *et al.* (2015) Social networks predict gut microbiome composition in wild baboons. *Elife* **4**: e05224.
- Vabalas, A., Gowen, E., Poliakoff, E., and Casson, A.J. (2019) Machine learning algorithm validation with a limited sample size. *PLoS One* **14**: e0224365.
- van de Water, J.A.J.M., Allemand, D., and Ferrier-Pagès, C. (2018) Host-microbe interactions in octocoral holobionts - recent advances and perspectives. *Microbiome* **6**: 64.
- Vanhatalo, J., Hartmann, M., and Veneranta, L. (2020) Additive multivariate Gaussian processes for joint species distribution modeling with heterogeneous data. *Bayesian Anal* **15**: 415–447.
- Vellend, M. (2010) Conceptual synthesis in community ecology. *Q Rev Biol* **85**: 183–206.
- Vieira-Silva, S., Falony, G., Darzi, Y., Lima-Mendez, G., Garcia Yunta, R., Okuda, S., *et al.* (2016) Species–function relationships shape ecological properties of the human gut microbiome. *Nat Microbiol* **1**: 1–8.
- Voigt, R.M., Forsyth, C.B., Green, S.J., Engen, P.A., and Keshavarzian, A. (2016) Chapter nine - circadian rhythm and the gut microbiome. In *International Review of Neurobiology: Gut Microbiome and Behavior*, Cryan, J.F., and Clarke, G. (eds): New York: Academic Press, pp. 193–205.
- Wang, Y., Kapun, M., Waidele, L., Kuenzel, S., Bergland, A. O., and Staubach, F. (2020) Common structuring principles of the *Drosophila melanogaster* microbiome on a continental scale and between host and substrate. *Environ Microbiol Rep* **12**: 220–228.
- Warton, D.I., Wright, S.T., and Wang, Y. (2012) Distance-based multivariate analyses confound location and dispersion effects. *Methods Ecol Evol* **3**: 89–101.
- Wexler, A.G., Bao, Y., Whitney, J.C., Bobay, L.-M., Xavier, J.B., Schofield, W.B., *et al.* (2016) Human symbionts inject and neutralize antibacterial toxins to persist in the gut. *Proc Natl Acad Sci U S A* **113**: 3639–3644.
- Wilbert, S.A., Mark Welch, J.L., and Borisy, G.G. (2020) Spatial ecology of the human tongue dorsum microbiome. *Cell Rep* **30**: 4003–4015-e3.
- Woodhams, D.C., Bletz, M.C., Becker, C.G., Bender, H.A., Buitrago-Rosas, D., Diebboll, H., *et al.* (2020) Host-associated microbiomes are predicted by immune system complexity and climate. *Genome Biol* **21**: 23.
- Wu, Y., Yang, Y., Cao, L., Yin, H., Xu, M., Wang, Z., *et al.* (2018) Habitat environments impacted the gut microbiome of long-distance migratory swan geese but central species conserved. *Sci Rep* **8**: 13314.
- Xu, Y., and Zhao, F. (2018) Single-cell metagenomics: challenges and applications. *Protein Cell* **9**: 501–510.
- Zaneveld, J.R., McMinds, R., and Vega Thurber, R. (2017) Stress and stability: applying the Anna Karenina principle to animal microbiomes. *Nat Microbiol* **2**: 17121.
- Zhang, Z., Geng, J., Tang, X., Fan, H., Xu, J., Wen, X., *et al.* (2014) Spatial heterogeneity and co-occurrence patterns of human mucosal-associated intestinal microbiota. *ISME J* **8**: 881–893.