microorganisms

MDPI

Review

Beyond Basic Diversity Estimates—Analytical Tools for Mechanistic Interpretations of Amplicon Sequencing Data

Anna Trego [1], Ciara Keating [2], Corine Nzeteu [1], Alison Graham [1], Vincent O'Flaherty [1] and Umer Zeeshan Ijaz [3,*]

[1] Microbial Ecology Laboratory, School of Biological and Chemical Sciences and the Ryan Institute, University of Galway, University Road, H91 TK33 Galway, Ireland
[2] Institute of Biodiversity, Animal Health & Comparative Medicine, The University of Glasgow, Oakfield Avenue, Glasgow G12 8LT, UK
[3] Water Engineering Group, School of Engineering, The University of Glasgow, Oakfield Avenue, Glasgow G12 8LT, UK
* Correspondence: umer.Ijaz@glasgow.ac.uk; Tel.: +44-(0)141-330-6458

Abstract: Understanding microbial ecology through amplifying short read regions, typically 16S rRNA for prokaryotic species or 18S rRNA for eukaryotic species, remains a popular, economical choice. These methods provide relative abundances of key microbial taxa, which, depending on the experimental design, can be used to infer mechanistic ecological underpinnings. In this review, we discuss recent advancements in in situ analytical tools that have the power to elucidate ecological phenomena, unveil the metabolic potential of microbial communities, identify complex multidimensional interactions between species, and compare stability and complexity under different conditions. Additionally, we highlight methods that incorporate various modalities and additional information, which in combination with abundance data, can help us understand how microbial communities respond to change in a typical ecosystem. Whilst the field of microbial informatics continues to progress substantially, our emphasis is on popular methods that are applicable to a broad range of study designs. The application of these methods can increase our mechanistic understanding of the ongoing dynamics of complex microbial communities.

Keywords: 16S rRNA; amplicons; ecology; microbiome; sequence analysis

Citation: Trego, A.; Keating, C.; Nzeteu, C.; Graham, A.; O'Flaherty, V.; Ijaz, U.Z. Beyond Basic Diversity Estimates—Analytical Tools for Mechanistic Interpretations of Amplicon Sequencing Data. *Microorganisms* **2022**, *10*, 1961. https://doi.org/10.3390/microorganisms10101961

Academic Editor: Juan M. Gonzalez

Received: 17 August 2022
Accepted: 30 September 2022
Published: 1 October 2022

1. High-Throughput Sequencing: Widely Used, but Under-Explored

The past several decades have seen strategic advancements in sequencing technologies [1,2], which have shaped our fundamental understanding of the human genome [3], marine and terrestrial biogeochemical cycling [4,5], and global biodiversity [6]. This, in turn, has had significant impacts in practice for clinical medicine, forensics, environmental engineering and biotechnology. For example, sequencing has enabled achievements such as the study of the 'unculturable majority'—all the organisms that cannot be successfully cultured in the lab [6,7]. This has improved our understanding of biodiversity, ecology and evolution and identified previously unknown organisms which have revolutionized biotechnology and medicine [8]. Furthermore, the application of sequencing technologies to cancer research has facilitated the identification of disease-specific drivers, mutational signatures, tumor mutational burden and neo-antigens, offering the promise of personalized patient care [9]. Moreover, the recent (2020) complete, telomere-to-telomere sequencing of the human genome [10] surely signifies that in terms of sequencing, the best is yet to come.

Such advancements are often the result of cutting-edge sequencing approaches such as whole genome sequencing (WGS) [11,12], sequencing of transcriptomes (RNA-seq) [13,14], long-read sequencing (Oxford NanoPore, MinION, SMRT-seq) [15], or single cell sequencing (scRNA-seq) [16]. For microbial ecologists, shotgun metagenomics (the untargeted WGS of all the microbial genomes present in a sample) has enabled large-scale investigations

into complex microbiomes [12]. It is a powerful tool which can provide taxonomic profiles, recover novel genomes and investigate the functional potential of a community. While the past 15 years have resulted in substantial reductions in terms of sequencing costs, the downstream data processing of metagenomes remains a significant challenge, particularly in view of co-assembly of sequence reads from all samples—a step required to bin the contigs into metagenomic assembled genomes (MAGs) later on. The co-assembly process typically requires holding debruijn graphs in memory, and with the enormous amount of sequencing data generated from the latest sequencing technologies, such as NovaSeq from Illumina, the memory footprint (RAM) often becomes a bottleneck. In addition, to capture accurate diversity, sequencing depth becomes an important factor, particularly for rarer genomes. Therefore, in terms of economics and computational requirements, WGS is often used on a small subset of samples, typically informed by a larger short-read amplicons dataset.

The choice between WGS and short-read amplicons also comes down to the aim of the proposed study, whether the goal is to explore microbial ecology [17–19], or to discover novel genomes [20]. Not only is sequencing short-read amplicons cheaper [21], but the downstream analysis is more accessible, faster and can access ever-improving reference databases [22–25]. Combined, this results in a more economical way to test hypotheses. Indeed, the widespread application of amplicon sequencing has resulted in a wealth of taxonomic and phylogenetic information about a variety of complex microbiomes [26,27]. It should, however, be noted that amplicon sequencing has its own limitations including short read lengths, sequencing errors, and relative abundances that will always be biased towards certain species due to primer-based amplification and different numbers of the 16S gene [28].

Whilst there is a preponderance of amplicon-based studies, published across a diverse range of fields, more often than not, they include only basic analyses, including: (i) diversity estimates (within samples and between samples); and (ii) how microbial species differ or remain persistent in either a case–control or spatial/temporal gradient. These analyses may serve the basic ambitions of a given study, however, to gain a mechanistic understanding with ecological relevance, we need to go beyond basic analyses.

Almost daily, novel methods and models are being reported for analyzing amplicon sequencing data [29–33]. Often, such methods are applied in the context of the human gut microbiome, but they are easily adaptable for other fields within microbial ecology—as long as the study design and datasets are suitable. These new analytical tools generally fall into one (or two) of several categories: methods which (i) quantify microbial community assembly; (ii) map network inferences; (iii) monitor spatial/temporal dynamics; (iv) integrate various types of datasets (integrative 'omics); (v) identify discriminant or differential taxa; (vi) find correlations between species abundance and environmental variables; or (vii) predict functional patterns. Given that high-throughput sequencing is still being frequently employed, but that new analytical tools are slow to be utilized, our aim in this review is two-fold: (i) to highlight trends in experimental design and data processing for studies utilizing short-read amplicon sequencing; and (ii) to highlight the utility of more sophisticated sequence analyses, showcasing several easily applied, new analytical techniques for enhanced mechanistic understanding of complex microbiomes.

2. Study Design Considerations: Planning for Statistics

The type of downstream microbiome analysis that can be applied is entirely dependent on the original hypotheses being tested and study design parameters. Therefore, it is always wise to plan ahead for the type of analysis that will generate meaningful data and best test the experimental hypotheses. Microbiome analysis is hugely affected by a lack of harmonized protocols, including sample collection methods, storage, processing and downstream analysis [12,34,35]. Furthermore, variation in sample handling, sampling size, controls [36], the choice of extraction method [37], sequencing blank/mock communities, choice of sequencing platform and the downstream analysis [38,39] can contribute to the

introduction of various biases leading to inconsistent and incomparable results. Therefore, in microbiome research, all these parameters must be carefully chosen and defined [34,35]. Whichever approach we use, we want it to be minimally biased against error.

At the bare minimum, microbial community samples are obtained in a case–control relationship (changing physico-chemical parameters in environmental datasets, or, pathology versus healthy controls in medical datasets). In some cases, the sampling is done over spatial or temporal gradients, particularly in intervention studies where the goal is to modulate the microbiome through some sort of treatment. This becomes convoluted for human microbiome datasets where there is an additional complication of 'pairedness' by virtue of multiple samples collected from the same subject over the course of the treatment [40]. Moreover, in other cases, there is a multifactorial or nested design [41]. For all these different study designs, there is no unified statistical framework, and each type will require a different set of statistical tools.

Finally, an important consideration for any microbiome study is the number of replicate samples sequenced per experimental condition. Statistically, the required number of samples depends on the effect size—the magnitude of difference between categories [36]. Tools such as "Evident" [42] have been developed to help estimate sample size based on projected effect size and records from similar studies [36]. In addition to thinking about the statistical power of the number of replicates, various types of downstream statistical models have their own requirements. While the number of replicates required per category is under constant debate, from our experience, replicate numbers can be chosen based on the type of analysis intended. For example, in terms of very basic statistics, usually at least three replicates per category are required. This will allow for basic diversity statistics, the identification of discriminant taxa and a core microbiome (e.g., [43]). However, for more advanced ecological modelling to be applied, usually a minimum of five to six replicates are required; this will usually satisfy criteria for the null models which often underpin community assembly analysis (e.g., [18,44,45]). Finally, for even more involved methods such as network inferences, upwards of 35 samples per category are often necessary to obtain reliable results (e.g., [46,47]).

When planning a microbiome study, it is often advisable to plan backwards—thinking about what types of statistical analyses are best going to help answer the experimental questions. Although it may seem counterintuitive, there is often more power and impact in designing studies that test fewer conditions but include more replicates. In this way we can go beyond basic diversity estimates and identify and pinpoint specific ecological patterns from the data.

3. Current Trends in Data Processing

Several processing pipelines have been developed which include a series of steps and programs employed to align, denoise, and remove spurious sequences: e.g., MOTHUR [48], QIIME 2 [49], and KRAKEN2 [50]. While these pipelines are flexibly designed and updated for quality assurance, users generally follow a series of steps without significant deviation. Such pipelines typically resolve species through an operational taxonomic unit (OTU) approach (often clustered at 97% similarity), or an amplicon sequencing variant (ASV) approach [51]. While ASVs have gained popularity in recent years, ongoing research has shown that they often yield similar diversity trends as OTUs [52], and that such an approach can artificially split bacterial genomes into clusters [53]. Conversely, it has recently been proposed that an ASV-based approach better represent the sequence diversity of functional genes, where it can be difficult to identify an appropriate threshold [54].

Within processing pipelines, tools such as VSEARCH [55] or USEARCH [56] were traditionally used for denoising, dereplicating and clustering into OTUs. The current release of USEARCH (v.11) includes a denoising option which will generate ZOTUs (zero-radius OTU). These are considered suitable for diversity analysis, although where traditional OTUs at 97% similarity may include more than one species, one species may have more than one ZOTU [57,58]. Conversely, DADA2 [59] and Deblur [60] are useful for resolving

ASVs. Deblur, in particular uses error profiles to obtain putative error-free sequences from sequencing data with better sensitivity and specificity than other available tools [60]. The product of these processing steps is to produce a feature table (Figure 1), wherein ASVs or OTUs are grouped, but taxonomy has yet to be assigned.

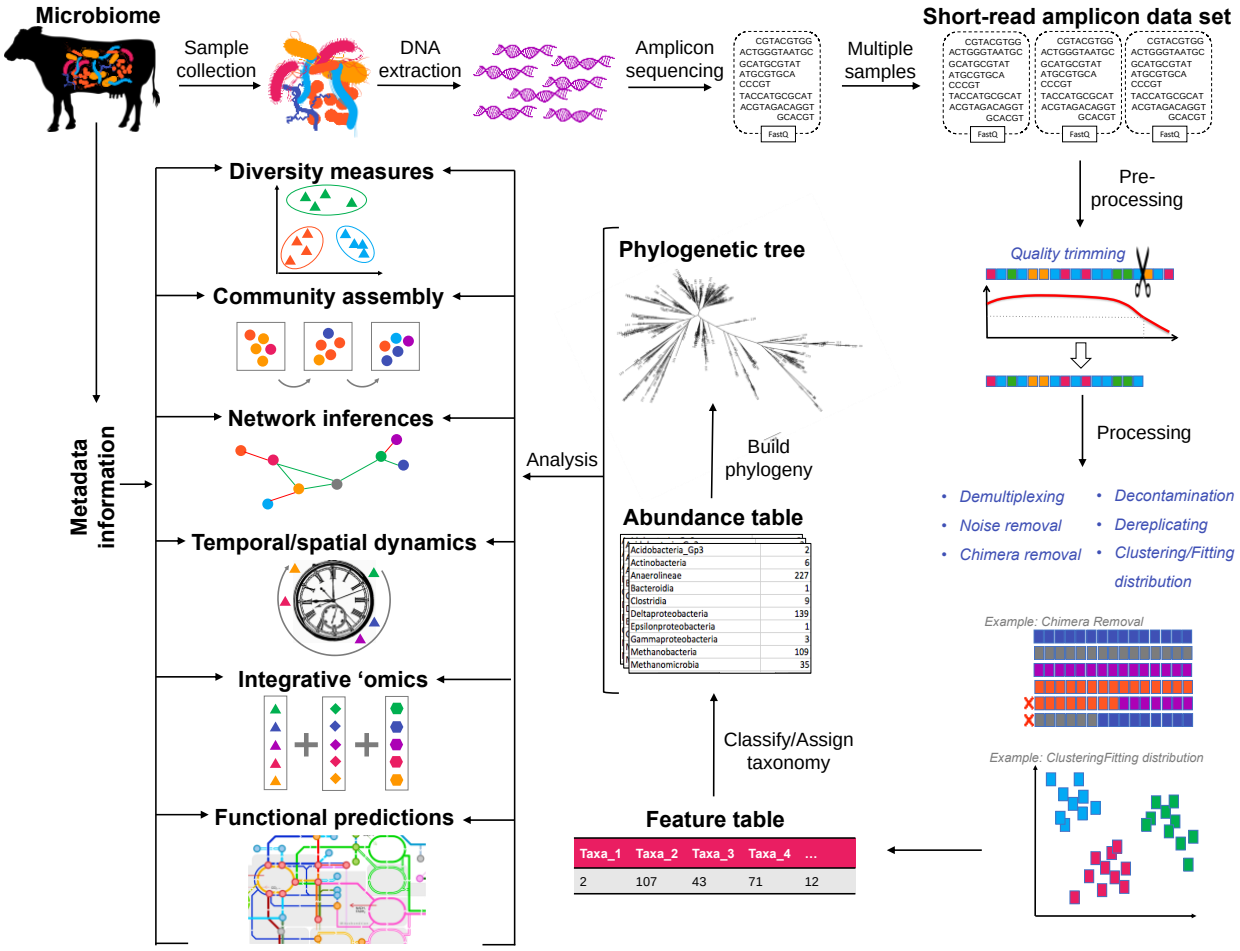

Figure 1. Summary of amplicon sequencing workflow. Basic diversity analysis are only a small fraction of available types of analyses available for microbiome analysis.

In either an ASV- or OTU-based approach, taxonomy is resolved by aligning the sequences against a reference database up to species level, where possible. The alignment process is most often implemented as a classifier (Figure 1), such as the traditional naïve Bayesian classifier (NBC). Recently however, the Bayesian lowest common ancestor (BLCA) algorithm has been proposed to give shades-of-grey assignments (with confidences) by considering the phylogeny of the reference 'hits' [61]. This approach, as opposed to NBC, is able to provide greater taxonomic resolution.

Several taxonomic databases are available, including Greengenes [62], RDP [63], AutoTax [64] and SILVA [22]. Currently, SILVA ribosomal RNA gene database (release 138) is the most comprehensive (containing 436,680 sequences) for 16S rRNA data. However, species-level taxonomy is not curated within this database with species classifications often falling into the category of 'uncultured' or 'metagenome'. In contrast, environment-specific and highly curated databases have also been developed, such as MiDAS (specific to wastewater treatment microbiomes) [24], TaxAss (for freshwater microbiomes) [25] or RefSoil (for soil microbiomes) [65], which help facilitate enhanced species-level assignments. Once taxonomy has been assigned and classified (yielding the final abundance table), a phylogenetic tree can be constructed. Both the abundance table and phylogenetic tree are required for downstream statistical analysis.

4. Traditional Measures of Diversity: Revisiting the Basics

The most fundamental ecological questions center around biodiversity, which is vital in terms of ecosystem function. Recently, for example, the idea of whether or not diversity begets diversity was explored. The authors concluded that for low-diversity systems, diversity does promote more diversity, but that eventually ecosystem diversity will reach a plateau [66]. Indeed, substantial efforts have been made in both macro- and microbiology to quantify and monitor diversity patterns. This is typically accomplished using alpha and beta diversity measures, by highlighting taxa that are dominant, differential, or persistent, and analyzing them within the context of similarity or dissimilarity between samples. These analyses are often performed in R using *vegan* [67] and the *phyloseq* package [68].

Alpha diversity is one of the most common and well-established metrics for measuring diversity. It is a means of calculating, at a local scale, the richness (number of observed taxa) evenness (distribution of abundances of the observed taxa), or both [69]. There are multiple approaches to calculating the alpha diversity of a sample, but the three most widely-accepted methods are the Shannon entropy [70], the rarefied richness [71], and/or the Simpson index [72]. Rarefied richness (rarefied to the lowest number of sequencing reads) is a useful means of comparing the number of observed taxa within a sample, without giving any consideration to abundances. It is often employed to track how the overall quantity of species changes over time, space or experimental conditions. Alternatively, Shannon entropy is a measure of the balance of taxa within a sample in terms of abundances. A high value indicates that all taxa within the community are equally abundant, while reduced values generally suggest that a sub-group of taxa are becoming more dominant within the community [70]. Additionally, a unified family of diversity indices called Hill Numbers exist. These are typically alpha diversity-generating formulas (^{q}D), which are often parameterized with q, for example, q = 0 gives richness estimate, q = 1 gives Shannon entropy, and q = 2 gives inverse Simpson index. Depending on the value of q used, the returned diversity can result in more or less emphasis on rare/abundant species [73]. Notably, there are several biases and assumptions associated with the estimation of alpha diversity and the application of measurement error models has been suggested to be useful in adjusting for any uncertainty [74].

While alpha diversity is the diversity within a sample, beta diversity compares diversity between samples. There are three common beta diversity distance metrics: (i) Bray–Curtis; (ii) UniFrac; and (iii) the weighted UniFrac [75–77]. Bray–Curtis distances are calculated based on abundances and work very well for ecological datasets [78]. UniFrac, however, is based on phylogenetic distances (thus requiring a phylogenetic tree) and considers how closely related the taxa within a community are by looking for the unique fraction of the phylogenetic tree. Notably, it is based on presence/absence data and therefore emphasizes rare community members [75]. Weighted UniFrac, however takes both counts and phylogeny into account and therefore is biased toward the dominant groups [76]. The beta diversity for any given distance metric is generally visualized on either a principal coordinate analysis plot (PCoA), or a non-metric multi-dimensional scaling (NMDS) plot [79]. PCA-type methods rely on eigen value/eigen vector decomposition, often of the covariance matrix. This is accomplished by generating a new abundance table, which has the same number of dimensions as the original table, but all the variability is shifted to the first dimension. They typically preserve distances between the samples in a reduced representation, but at the expense of loss of variability. NMDS, on the other hand, transforms the data (by optimizing a stress function) to however many dimensions the data is to be visualized for. Whilst it captures better variability, it comes at the cost of conserving distances in the reduced space, mainly maintaining their monotonic relationships. In general, PCoA plots are more accurate than NMDS for studies with fewer sample numbers and many features (ASVs/OTUs) [80]. In all cases, however, we look for clustering patterns between the samples, where similar communities will cluster together. Finally, a simple tool allows us to assess if any of the samples diverge too much from the average beta diversity of the ecosystem. This technique is called the local contribution to beta diversity (LCBD) [81],

is available in R as a part of the *adespatial* package [82], and can enable us to have a very simple, quantifiable measure for "microbial dysbiosis".

The next most common types of diversity analysis include: (i) highlighting the abundant fraction of the microbiome, usually with a heatmap or a taxa-bar plot; and (ii) looking for taxa that are differential (changing between categories); (iii) persistent (core); or (iv) conditionally rare and abundant (CRAT). Differential taxa, or discriminant taxa can be calculated in several ways using a variety of thresholds [17,83]. Notably, sPLS-DA analysis, available through the *mixOmics* package (discussed later in Section 8) uses an advanced algorithm to identify discriminant taxa within microbiomes [84]. Core microbiome analysis is equally versatile in terms of how a core microbiome is defined. Notably, the core microbiome has previously been defined as any feature present in at least 85% of the samples [85]. Furthermore, several detection thresholds can be applied in terms of abundance to sort the core microbiome from the low-abundant core, to the high-abundant core yielding a 2-dimensional representation of the taxa that persist within the system. This model has now been applied to several ecosystems including wastewater treatment facilities [44], chicken [17], and fish microbiomes [83]. Finally, an emerging way to look at both the abundant and rare fractions of the microbiome is to focus on those taxa whose abundance is conditional, CRAT taxa [86].

These few analyses constitute the main bulk of data analytical techniques typically used in published manuscripts. However, as 'omics modalities are becoming cheaper, there is a world-over shift to incorporate additional modalities (flow-cytometry, transcriptomics, metabolomics, proteomics) to fill in the gaps that arise with static nature of 16S rRNA datasets as some species may be active but in low abundance or vise versa [43,87,88]. These demand a new way of consolidating all the information by developing methodologies that not only give correlations between the datasets, but also have a discrimination component which reduce the features to set of absolute minimal features that capture the main patterns.

Additionally, we often also want to capture underlying ecological principles particularly in terms of the environmental habitat where these microbes are observed and to assign them specific roles, i.e., whether they are generalists of specialists; whether they are influenced by the environment; whether they share a niche with other microbes; whether they have symbiotic relationships or compete for resources; or whether they are predominately active (throughout the spatial or temporal gradients) or are transient and proliferate sporadically in response to a biotic or abiotic influence. These explorations require bespoke analytical techniques, the majority of which have been explored in macro-ecology literature and are slowly being adapted to microbiome research.

5. Identifying Mechanisms Driving Microbial Community Assembly

Despite decades of research, development, and process optimization, several outstanding questions remain regarding the ecology of engineered systems. Paramount among these is how dynamic communities assemble both during system start-up and throughout all subsequent operational phases. If we can determine whether assembly mechanisms are predictable, and pinpoint precisely when and how we can manipulate these processes, we will be one step closer to being able to manage and control these communities to serve particular functions. Key to this is understanding the driving forces shaping the microbiome [89].

Ecologists have been working on identifying distributions of species patterns among sites, trying to come up with simple quantifiable metrics on incidence tables (presence/absence) such as *Coherence* (degree to which spatial patterns could be collapsed into a single dimension); *Species Turnover* (which describes the number of species replacements after the collapse); and *Boundary Clumping* (how the edges of species are distributed across the dimensions) [90]. These three measures, adopted to microbiome research [91] are then useful to understand the metacommunity structure. Relatedly, an additional framework describes species distributions along environmental gradients, manifesting themselves as metacommunity structures with the following discernable patterns: *Random*

(no gradients or patterns in species found); *Checkerboards* (species pairs have mutually exclusive distrbutions and such pairs occur independently); *Nested* (nested subsets are observed); *Evenly spaced Gradients* (species ranges are arranged more evenly); *Gleasonian* (gradients results in species turnover, but species ranges are random); and *Clementsian* (discrete communities that replace each other as a group) [92]. Whilst these are useful methods, they do not take into account species' abundance, nor phylogenetic relatedness—a limitation that has resulted in minimal uptake.

While we still lack a general theory to explain how communities are assembled, community ecologists have, over time, converged towards a framework which acknowledges that these processes are either stochastic or deterministic. Deterministic influences can either be biotic or abiotic which leave their influence on the observed community structure, whether on the composition, or on the phylogeny. "Stochasticity", on the other hand, refers to completely random effects shaping the microbiome [93]. Elucidating these community assembly processes, if deterministic, then provides a means to look for patterns in the surrounding environment that are responsible (e.g., some physico-chemical characteristics). This is highly relevant for experimental designs where the goal is to alter the community structure to make it optimal in view of certain performance criteria (for example, optimal functioning of wastewater treatment, or reducing microbial dysbiosis in pathological conditions).

To this end, new models, usually requiring only sequencing abundance data (such as an ASV- or OTU-table) and sometimes a phylogenetic tree, are continuously being developed to help identify and quantify ecological mechanisms driving change in these communities. Most often, these are based on a null modelling procedure, where community structure (composition or phylogeny) is perturbed by creating in situ artificial variations (typically 999), by conserving a certain property of the original samples (typically, alpha diversity). Depending on the choice of the mathematical metric used, its deviation from the original community structure to the average of when applied to altered structures, has the power to reveal ecological phenomena.

Notably, Quantitative Process Estimates (QPE) uses null modelling to quantify assembly within an ecological framework based on selection and dispersal mechanisms [94]. Specifically, it classifies assembly mechanisms as (i) *variable selection*, when selective environmental conditions result in high compositional turnover; (ii) *homogenous selection*, when static environmental conditions result in consistent selective pressure; (iii) *dispersal limitation*, when low rates of dispersal (movement of microorganisms in space) result in high community turnover (this drives ecological drift); (iv) *homogenizing dispersal*, when high compositional turnover is fueled by high dispersal rates; or (v) "*undominated*", when compositional turnover is neither the result of selection nor dispersal [18,95]. QPE holds several advantages over other types of assembly quantification tools: (i) it incorporates both phylogenetically informed, and phylogenetically uninformed data making it better suited for understanding ecological phenomena—whereas several other available methods do not take phylogeny into account (such as the stochasticity ratio (SR), discussed next); (ii) with QPE we can measure the relative contribution of different processes acting simultaneously upon a community; and (iii) QPE allows for the identification and comparison of dominant assembly mechanisms under different conditions. A recent extension to this framework is by incorporating phylogenetic bin-based null model analysis which is essentially the same procedure but at finer bin-levels with sufficient phylogenetic signal [96].

An alternative to divulging stochastic and deterministic assembly mechanisms is through the normalized stochasticity ratio (NST) approach [97]. This method, while useful, is not nearly as informative as QPE, especially in terms of exact mechanisms. For each community, it yields a percentage value called the stochasticity ratio (SR), indicating the quantity of stochastic processes that shaped that community—implying that all other remaining processes were deterministic. It is, however, a straightforward way of tracking the contribution of stochasticity.

Quantifying and assessing changes in biological diversity are central aspects of many ecological studies, yet accurate methods of estimating biological diversity from microbial samples remains a challenging problem. Although Hill numbers were first proposed several decades ago as alpha diversity-generating formulas, their extension and parameterization to beta diversity space is only recent, providing a means of identifying ecological assembly mechanisms [98], particularly to see if the differences between samples are driven by stochastic or deterministic processes. Additionally, it has been suggested that Hill numbers can also be used to examine the relationships between different community members by incorporating phylogenetic information. This could provide additional information about functional differences between communities [98].

Another common framework for discussing community assembly mechanisms is in terms of niche vs. neutral processes [99,100]. Although sometimes used interchangeably, these concepts are not equivalent to deterministic and stochastic processes. While neutrality reflects the ecological equivalence of species (when demographic rates such as birth, death and dispersal are identical), stochasticity suggests random, probabilistic, variation in species' demographic rates. Likewise, niche processes suggest differences between species' mean demographic rates, while determinism implies an absence of random, probabilistic, changes in species' demographic rates [101]. Using simulated stochastic and deterministic metacommunities, Tucker et al. (2016) were able to use the abundance-based ß-null deviation measures to differentiate between niche and neutral community assembly. This has since been applied to understand the community assembly of fish (Atlantic cod) microbiomes [45] and marine bioplankton [19].

An elegant, and easily applied tool for examining microbial ecosystems is the competitive lottery model for clade-based assembly [29]. The model assumes that phylogenetically related organisms are functionally similar, sharing similar gene content, preferential niche space and metabolisms—all giving rise to intense clade-level competition. Within a given clade/group, 'winners' would arise, being more 'fit' to the given conditions, or simply having arrived to the community and established first (priority effects). These 'winner' ASVs make up a majority of the abundance within their clade—capturing >90% of the groups abundance [29]. Application of this model to anaerobic digestion (AD) systems helped to describe not only how overall abundances shifted, but identified specific 'winners' which either adapted to the given environment, becoming 'winners', or were unsuited and lost their 'winner-status' [18,44]. The power of the lottery model is that it identifies specific 'winning' ASVs for each clade/group, which can be tracked across conditions, allowing for a deeper understanding of how different trophic groups are behaving, responding and functioning.

Continuing to think about how niche spaces influence microbial communities, recent work has advanced our ability to use microbiome sequence data to identify *generalists* and *specialists* within complex communities [102,103]. While generalists are taxa that inhabit a broad range of environments, or environmental gradients, specialists occupy a narrow and discriminatory niche space. Using the publicly available package in R, *MicroNiche*, we can use abundance tables and corresponding data to differentiate between generalist and specialist species, but also to identify species that are positively or negatively correlated with environmental data of interest [103]. Similarly, we can consider the specificity of microorganisms to a given environmental gradient. Using the R package *Specificity*, the quantity of specificity can be calculated and correlated across different environmental covariates. Additionally, the package will identify taxa that are specific to a given environmental covariate [31].

6. Network Inferences: Identifying Relationships and Revealing Complexity

At the very basic level, co-occurrence networks are a useful way to interpret microbial community dynamics. These networks can handle the scale and diversity of microbiome data, and have the added advantages of being able to identify microbial interactions/associations [104], identify network re-organizations under varying condi-

tions [105], identify 'hub' species [106] or estimate diversity [107]. Recently, networks were even used to predict a 'die score'—identifying organisms within the community that were most likely to be eliminated under various conditions [108]. While compositional approaches to network modelling are varied and continue to evolve [109], here we will highlight three different approaches, beyond simple correlation-based approaches.

First, SPIEC-EASI (**SP**arse **I**nvers**E** **C**ovariance Estimation for **E**cological **AS**sociation **I**nference) remains a robust and popular choice for network inference [110] which is derived from Graphical Lasso method, a sparse penalized maximum likelihood estimator for the concentration or precision when the species abundances are modelled as a Gaussian distribution. SPIEC-EASI was developed specifically to address two issues: (i) that amplicon-based studies yield ASV abundances that are compositional (abundances are relative to one another) and (ii) that results of such studies are often biased by poor sampling depth. It yields network visualizations (at a specified taxonomic rank), that indicate relative abundances, network clustering within the community and positive and negative associations between groups. It is a flexible tool for evaluating the adaptation, development and interactions within a microbial community. Recently, Boolean logic has been applied to microbial communities, allowing abundance patterns to emerge in new ways, via Boolean multi-dimensional arrangements [104]. The idea revolves around compartmentalizing interactions between two or more species by finding thresholds at which the species in an interacting ecosystem can be deemed as present or absent to infer *co-presence*, *one-way*, and *co-exclusion* relationships by calculating the probability scores on the inferred compartments. Moreover, these various relationships can be identified in multiple dimensions, where 2D relationships are between two interacting species, and 3D relationships are identified between three interacting species.

Any one change in community dynamics can cause a cascade of additional disturbances within the system. In many cases understanding how to maintain stability would be beneficial [111,112]. The complexity-stability paradigm predicts that stable ecosystems have an upper limit to their overall complexity [113]. Therefore, it would be unlikely to observe communities comprising both a large number of species and a high degree of interspecific interactions. Instead, to maintain stability, microbial communities have evolved to be either small and highly interacting, or large but weakly interacting. This trade-off between community complexity and stability emphasizes the role of interspecific interactions in governing overall species richness of a community [47]. Generally, calculating the complexity-stability across different communities has been challenging because while the species richness can, in theory, be measured, the degree of interspecific interactions is more challenging.

However, identifying relationships of interest between microbial species is useful to reveal symbiosis and antagonism, as microbial species rarely act alone, striving for resources with a cascade of pathways where multiple species enable conversion of substrates. These relationships can be associations as well as causal relationships, and May's stability criteria [113] models interactions between species as a Generalised Lotka Volterra Model. Herein one can use the interaction matrix A_{ij} (effect of species j on species i) to reveal the stability-complexity relationship of the ecosystem which can be derived as a curve that satisfies $\alpha\sqrt{nC} < 1$ where α^2 and C are the variance and density ("*connectance*") of the non-zero off-diagnoal elements of A_{ij}. However, this all depends on accurate network inference, limited by concise framework to incorporate ecological understanding of interacting species.

Recently, Yonatan et al. (2022) has addressed the May's stability criteria by calculating "effective connectance" (degree of interactions) without the need to infer the networks explicitly. This is done by analysing the beta diversity measures such as 'dissimilarity', and "overlap" for N(N-1)/2 sample pairs for given N samples in a category. The authors suggest that in addition to evaluating complexity-stability relationships, the effective connectance may give insights into the degree to which local perturbations, which cause shifts in the abundance of one or a few species, will propagate and affect the whole community.

Communities with high effective connectance will be more substantially affected by a few changes in abundance than communities with low effective connectance [47].

7. Over Space and Time: Measuring Temporal/Spatial Dynamics

Space and time are arguably two of the most important variables in many microbiome studies. Sampling communities over time or space allows for the assessment of community development [18,44], microbial evolution [114], biogeography [115], global distributions [116], dispersal rates [95], functional robustness [30], and/or responses to treatments or perturbations [117]. Such studies benefit from analytical methods which evaluate (i) trends over time according to variables of interest; (ii) magnitude of change within individual categories; and/or (iii) the inherent variation in complex biological systems [118]. Many analytical options are continuously developed; here we highlight some interesting/useful options.

For assessing beta diversity on temporal or spatial scales, zeta diversity has been recently proposed. This can better detect shifts and transitions, and emphasizes the roles of rare versus abundant species in the microbial consortia [119,120]. Next, built directly into QIIME 2, *q2-longitudinal* is a software plugin which offers multiple methods for the streamlined analysis and visualization of longitudinal data, providing valuable information on temporal trends [118]. Several functions are available including analysis of (i) volatility; (ii) feature-volatility; (iii) linear mixed effects (LME); (iv) first differences and (v) first distances, among several others. Microbial volatility is best described as the variance in microbial abundance, diversity or other metric over time. Changes in volatility can be indicative of system disturbances and thus can be a useful parameter in the context of engineered bioprocesses. The most common way to examine change over time is to compare the average relative abundances of taxa over time, usually via a box plot or a heat map. Unfortunately, such approaches usually work by targeting the most abundant microorganisms, ignoring features that are actually associated with specific time points. Using the feature-volatility function, machine learning regressors (random forests by default) learn the structure of the data and identify all features that are predictive of different categories (time points). Next, LME is able to test whether a taxa's relative abundances are impacted by time, or other available parameters. Finally, the first differences and first distances actions allow for the assessment of the rate of change between time points [118].

Another interesting tool is the phylogenetic recruitment model, which examines microbial species succession over temporal scales [121]. The method works by detecting the order in which new species are detected and is linked to phylogenetic diversity (PD) estimates. The authors use the dispersion parameter (D), which is calculated based on the probability of detection of new species by fitting a logistic error model on temporal changes in phylogenetic diversity (PD) estimates. The value of D determines the primary recruitment mechanisms. If $D = 0$, then all species have an equiprobable chance of recruitment (neutral). If $D > 0$, then phylogenetically divergent taxa (to the taxa detected in the previous time-points) are preferentially added to the community (overdispersed). In contrast, if $D < 0$, then phylogenetically similar taxa are preferentially added to the community (underdispersed—or *nepotistic*). This model has been implemented to understand recruitment trends in microbiomes [45,121] where the communities were generally nepotistic and recruitment was disrupted in 'perturbed' communities.

Recently, an R package called *splinectomeR* was introduced, allowing the assessment of statistical changes within a longitudinal dataset using three key functions: *permuspliner*, *sliding_spliner* and *trendyspliner* [122]. These functions can test hypotheses regarding observations over time without transforming or collapsing the data points, as is done in many existing longitudinal studies. The first function, permuspliner, tests overall variability and noise between two groups in longitudinal studies. With this we can identify taxa that change longitudinally/spatially. The second function, sliding_spliner, gains information on the specific time when change occurred, identifying periods/spatial ranges when there is change in taxa abundances between multiple categories. Importantly, this function requires

a large dataset (50–100 data points). Finally, the third function, trendyspliner, tests for a significant non-zero overall trends in a single population over time, identifying whether an abundance profile is linear or non-linear. *splinectomeR* achieved these statistical analyses using summary splines and randomly permuted distribution to assess the significance of the observed magnitude of change between groups or trends over time.

An alternative approach has made advances in integrating temporal datasets of different types. If, for example, a timeseries of major taxa is observed, then on temporal basis we can find a subset of these taxa that cluster together under the 'time-omics' framework [123]. This is especially useful when a study contains both microbial community data and other types of heterogenous biological, environmental, metabolome, chemical or phenotypic data—which are reduced to subset of series, significant in the context of temporal dynamics of microbiome experiments. The method selects key temporal features with strong associations within an "automated" clustering driven by principal coordinate analysis (PCA), where each dimension comprises of two clusters. This way, for a given category, species are clustered that have similar temporal evolution (using Silhouette metric). Each species is modeled as a function of time, considering all the variabilities of the different replicates in a linear-mixed model spline framework. Moreover, the fitted splines enable us to predict/interpolate time points that might be missing. One can then use sparse PCA to retain the most important features (taxa). The method is not only useful for microbiome data, but can also be used for any other modality (flowcytometry, metabolomics, etc.) where there is temporal data acquisition, the only difference being the normalization procedure. Finally, if a study includes matching longitudinal data (i.e., multi-modalities: metagenomics + metabolomics), then we can also get clusters with features across multiple datasets—useful in giving a mechanistic understanding.

One of the most compelling questions for microbial ecologists is, whether, or to what degree a microbial community is deterministically dependent on its initial/previous composition. To address this question, a new model MTV-LMM (Microbial Temporal Variability Linear Mixed Model) has been proposed [124]. MTV-LMM is a linear mixed model that can be applied to predict the temporal dynamics of microbial communities in longitudinal studies. Application of this model will identify time-dependent taxa—those affected by the past composition of the microbiome—these microbes can then be used to describe the temporal trajectories of the microbiome. Moreover, MTV-LMM introduces a concept termed 'time-explainability', a measure of the fraction of temporal-variance that can be explained by the community profile at previous time points. The application MTV-LMM to engineered systems, for example, would yield valuable information on time-dependent compositional patterns allowing for the prediction and modulation of these microbial communities.

Other notable efforts have been made to give a more mechanistic understanding of species richness across different spatial scales [125]. Species richness is one of the most common measures of biodiversity, but the specific ecological processes driving change are difficult to disentangle. To this end, a framework was developed wherein variation in species richness can be decomposed into three components: (i) species abundance distribution; (ii) species density; and (iii) species spatial aggregation/distribution. By constructing several types of rarefaction curves, the relative contribution of the three components of species richness can be measured across scales. These methods are available in the R package *mobr*. Such tools will help ecologists move beyond single-scale analyses such as species richness alone.

8. Integrative 'Omics: Combining Multiple Datasets

With advances in computational biology we are now in a position to go beyond simply observing microbial communities, and can rather identify meaningful connections by integrating several 'omics technologies (targeted sequencing strategies, proteomics and/or metabolomics) through sophisticated dimensionality reduction algorithms [126]. These not only optimize for correlation between multiple datasets but also give the dis-

criminatory power to filter out features that do not change in a case–control relationship (multiple-category comparisons). While simply merging data sets might lead to false positive hypotheses, three types of integrative approaches have been developed: (i) data complexity reductions; and (ii) supervised or (iii) unsupervised integration. Several packages have been developed to computationally handle multivariate analysis of such large biological data sets.

Recently, *STATegra* was designed as a conceptual framework, designed to be as user-friendly as possible for the identification of biological features and pathways within large data sets [127]. It is available in R through the *STATegRa* Bioconductor package. There are four primary steps with various tools available at each. In the first instance, each 'omic data set is analysed separately (univariate analysis). Following this, the data sets are jointly analysed for component analysis, feature identification and exploratory analysis.

Notably, *mixOmics* is a user-friendly R package which takes a systems-biology approach, focusing on probing relationships through data exploration, dimension reduction, integration and visualization [84]. It uses multivariate projection-based methodologies in order to handle large data sets and highlight variation in biological systems; and their relaxed assumptions about data distribution make the methods highly flexible, fitting a wide range of study designs. Importantly, the methods allow for the identification of discriminant groups within biological data sets, using several approaches. Firstly, the sparse projection to latent structure discriminant analysis (sPLS-DA) can be applied as a supervised analysis of one data set, yielding discriminant features at a given taxonomic level. Additionally, when a study consists of several independent data sets measured on the same predictors, MINT can be used to integrate these studies and identify common discriminants. Both sPLS-DA and MINT can be easily applied to 16S rRNA data sets [17,83,128]. Finally, DIABLO (Data Integration Analysis for Biomarker discovery using Latent variable approaches for Omics studies) allows for the integration of data sets using the same biological samples measured on different 'omics platforms (i.e., metagenomics and proteomics) [84,129]. Notably, a common bottleneck to experimental design is that parallel measurements from various 'omics technologies are required. However, the ongoing development of such integrative frameworks will continue to be essential to the discovery of new biology as data sets become increasingly complex.

9. Meta-Data Integration: Using Regressions to Identify Key Taxa

Typically studies in microbial ecology use a case–control, multi-factorial design, seeking out spatial or temporal relationships between microbiome data and other sources of variation from the environment. Often, the aim of these controlled experiments is to manipulate the environment to modulate the microbiome. Sources of variation can take the form of continuous or categorical variables including clinical, socio-economic, chemical, or other parameters. Identifying relationships between these covariates and the microbiome allows us to understand and predict how complex communities respond to stress, environmental perturbations, or other disturbances. This is usually achieved using various types of regression analysis [130].

Recent years have seen a growing interest in the development of new methods for multivariate analysis of microbiome data [33,131]. In particular, joint species distribution models, which use random effects to identify correlations between environmental variables and predictions of species abundances, have garnered increased interest [33,132]. Such models assume that species will respond jointly to the environment as well as to each other, and thus have the potential to pinpoint the causes of species co-occurrence patterns and identify microorganisms that are positively or negatively associated with a given covariate [132,133]. A key approach to this class of statistical modeling is the generalized linear latent variable model (GLLVM) [134], which is capable of handling datasets containing thousands of species—a common challenge when the number of observed species greatly outnumbers the number of samples [135,136]. Previously, the application of GLLVMs has been computationally slow and impractical for large datasets. However, *gllvm*, a new

package for R, is able to quickly apply GLLVMs to large multivariate datasets, with features allowing for model selection and high-quality visualizations [33,136].

There are several challenges to the statistical analysis of microbiome data. As mentioned previously, the size and multidimensional nature of such data are a significant challenge. Additionally, the compositional nature of microbiome data presents another significant challenge, i.e., the fact that changes in the abundance of one species induce changes in the observed abundances of the other species (that abundances are relative to one another). Compositional data analysis (CoDA) is an alternative framework which seeks to provide methods to appropriately handle complex compositional data [137]. While there are many CoDA approaches, a subset are formulated from generalized linear models, each having specific constraints [130,138]. Among these are *selbal*, which relies on a selection of microbial balances [139]; clr-lasso, which is a simple penalized regression [140–142] on centered log-ratio (clr)-transformed data [138]; and *coda-lasso*, which performs penalized regression on a log-contrast regression model [32]. In particular, *coda-lasso* works particularly well when the aim is the identification of taxa most associated with a given environmental variable [138]. Notably it is now available within *coda4microbiome*, a new R package for the analysis of microbiome data within the CoDA framework [32].

10. Predictive Functional Modelling: Using Structure to Infer Function

The advantage of functional profiling over taxonomic profiling is that it assesses what a microbial community can do, rather than simply who is present [143,144]. Advances in sequencing technologies have been accompanied by progress in the comprehensiveness of the associated databases. Better databases means that the predictive modelling which can be accomplished using targeted gene sequencing (usually the 16S rRNA gene as a cheap alternative to shotgun metagenomics [145]) now offers somewhat better exposition on putative functional behavior. This is especially true when combined with new tools that are able to greatly reduce or eliminate amplification biases and/or errors based on amplicon-genome linkages [146].

PICRUSt2 [147] is a prime example of an accessible and much-improved database. It boasts ~20,000 genomes while is predecessor, PICRUSt1, only included 2011 genomes. Using the QIIME 2 plugin KEGG enzymes and MetaCyc pathway predictions can be found. Such predictions are highly dependent on the number of pathways available for the reference genomes. The algorithm consistently predicts pathways that yield greater than 0.8 correlation with the actual pathways observed using shotgun metagenomic equivalents as highlighted by the authors [147]. While Picrust2 has been shown to be more accurate in its predictive capacity, care should still be taken during interpretation as they are still only putative estimates.

Similarly, Tax4Fun2 [148], FAPROTAX [149] and BugBase [150] have all been developed to predict function from taxonomic profiles. *Tax4Fun2* is an R package that makes such predictions based on 16S rRNA gene sequence data regardless of sequencing platform. The algorithm yields a list of KEGG Orthologs (KOs) relating to specific functions and is additionally able to calculate functional gene redundancies [148]. FAPROTAX is a manually constructed database that maps prokaryotic taxa to function. It converts abundance profiles into putative functional group abundance profiles. Notably, an individual taxon may be affiliated with multiple functions within FAPROTAX [149]. BugBase is an algorithm which predicts organism-level coverage of functional pathways but also provides biologically interpretable phenotypes such as oxygen tolerance, biofilm formation and pathogenic potential. Notably, BugBase can also be customized to identify particular traits of interest [150]. All four approaches provide putative functional mapping when used with 16S rRNA sequence data.

Recently, there have also been advances in how to correctly estimate beta diversity between functions. Traditionally, Bray–Curtis distances (or any other count measure) were applied on functional abundance (KEGG Orthologs: KOs) tables obtained from PICRUSt2 or any other metabolic prediction software. These mathematical measures assume each

feature (typically microbes) to be independent, which does not hold true for functional enzymes. This is mainly because there is redundancy in KOs spaces with multiple KOs serving the same function, some available to some species, and others to a different set of species. Therefore, to capture the hierarchy and inherent dependences between KOs, a new beta diversity measure called Hierarchical Meta-Storms (HMS) [151] has been introduced. This takes into account the information of which KOs are implicated in the definition of pathways (as a hierarchical structure), and then starts at the bottom level of these hierarchies and then propagates the abundances upwards to give a weighted dissimilarity measure. This then provides higher sensitivity for detecting variations in upper-level metabolic pathways between samples.

Finally, a simple approach to assessing microbiome stability measures functional robustness by linking changes in the structure of the community to specific functional shifts [30]. Robustness is an important factor of the microbial community, especially for engineered ecosystems, which often depend on community resilience for process stability [152]. While functional dynamics are best examined via metagenomic or metatranscriptomic datasets, the Taxa Function Robustness measures the degree to which a shift in a community's structural (taxonomic) profile (using amplicons) will result in a change in its functional capacity, particularly two parameters, "*Attenuation*" and "*Buffering*" summarizes the functional robustness of communities in each samples, that can reveal not only overall functional robustness, but can also help look at specific pathways of interest. It is based on the underlying assumption that a community's functional capacity is directly related to the genes available within that community [30]. When the structural profile shifts, certain genes (and therefore functions) may be lost entirely, while others may remain—redundant within other, more stable species. With this tool we can track changes in functional capacity across time, according to environmental conditions, or between different systems. While all of these methods for predicting functionality of microbiomes are useful and highly informative, it is worth reiterating that they only give putative functionalities and should therefore be interpreted with caution. Indeed, they maybe be highly useful for generating hypotheses rather than drawing strong conclusions.

11. Conclusions

Microbial ecology is not just about simply describing which microbes increase or decrease in abundance based on study design. Advances in informatics have provided at our fingertips the ability to identify "unknown" factors that come into play in assembling microbial communities; reveal spatial and temporal patterns; and give a snapshot of how stable or complex a system is, and whether the microbial community is able to withstand environmental perturbations. Simply relying on genomic based identification and proliferation of microbial species is not sufficient to infer mechanistic patterns. For this, we need to incorporate other modalities, and meta-data, including physico-chemical parameters, which is only recently possible with the advancement in integrated 'omics techniques. Shotgun metagenomics remains very popular by virtue of recovering microbial genomes and can give metabolic potential of microbial species, however, suffers from accurately estimating the diversity, by virtue of depth of resolution. Therefore, studies that focus on ecology and taxonomic expanse of the microbial communities are best suited to use amplicon-based surveys, which now are well advanced in terms of databases. Here, we have highlighted methods that are in routine use and serve as a guideline to perform microbial community analyses.

Author Contributions: U.Z.I. directed this study. A.T. and U.Z.I. drafted the manuscript with contributions from C.K., C.N., A.G. and V.O. All authors approve the paper and agree on accountability of the work therein. All authors have read and agreed to the published version of the manuscript.

Funding: The manuscript is supported by a NERC Independent Research Fellowship (NE/L011956/1) and EPSRC (EP/P029329/1 and EP/V030515/1). VOF was additionally supported by grants from the Higher Education Authority (HEA) of Ireland through: the Programme for Research at Third Level

Institutions, Cycle 5 (PRTLI-5), co-funded by the European Regional Development Fund (ERDF); the Enterprise Ireland Technology Centres Programme (TC/2014/0016) and Science Foundation Ireland (14/IA/2371 and 16/RC/3889).

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank the funders for supporting this work.

Conflicts of Interest: The authors declare no competing interest.

References

1. Shendure, J.; Balasubramanian, S.; Church, G.M.; Gilbert, W.; Rogers, J.; Schloss, J.A.; Waterston, R.H. DNA sequencing at 40: Past, present and future. *Nature* **2017**, *550*, 345–353. [CrossRef] [PubMed]
2. Shendure, J.; Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **2008**, *26*, 1135. [CrossRef] [PubMed]
3. Lander, E.S.; Linton, L.M.; Birren, B.; Nusbaum, C.; Zody, M.C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; et al. Initial sequencing and analysis of the human genome. *Nature* **2001**, *409*, 860–921. [CrossRef] [PubMed]
4. Tyson, G.W.; Chapman, J.; Hugenholtz, P.; Allen, E.E.; Ram, R.J.; Richardson, P.M.; Solovyev, V.V.; Rubin, E.M.; Rokhsar, D.S.; Banfield, J.F. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **2004**, *428*, 37–43. [CrossRef] [PubMed]
5. Venter, J.C.; Remington, K.; Heidelberg, J.F.; Halpern, A.L.; Rusch, D.; Eisen, J.A.; Wu, D.; Paulsen, I.; Nelson, K.E.; Nelson, W. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **2004**, *304*, 66–74. [CrossRef]
6. Schloss, P.D.; Handelsman, J. Metagenomics for studying unculturable microorganisms: Cutting the Gordian knot. *Genome Biol.* **2005**, *6*, 229. [CrossRef]
7. Liu, S.; Moon, C.D.; Zheng, N.; Huws, S.; Zhao, S.; Wang, J. Opportunities and challenges of using metagenomic data to bring uncultured microbes into cultivation. *Microbiome* **2022**, *10*, 76. [CrossRef]
8. Lightbody, G.; Haberland, V.; Browne, F.; Taggart, L.; Zheng, H.; Parkes, E.; Blayney, J.K. Review of applications of high-throughput sequencing in personalized medicine: Barriers and facilitators of future progress in research and clinical application. *Brief. Bioinform.* **2019**, *20*, 1795–1811. [CrossRef]
9. Ley, T.J.; Mardis, E.R.; Ding, L.; Fulton, B.; McLellan, M.D.; Chen, K.; Dooling, D.; Dunford-Shore, B.H.; McGrath, S.; Hickenbotham, M.; et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **2008**, *456*, 66–72. [CrossRef]
10. Miga, K.H.; Koren, S.; Rhie, A.; Vollger, M.R.; Gershman, A.; Bzikadze, A.; Brooks, S.; Howe, E.; Porubsky, D.; Logsdon, G.A. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **2020**, *585*, 79–84. [CrossRef]
11. Lam, H.Y.K.; Clark, M.J.; Chen, R.; Chen, R.; Natsoulis, G.; O'Huallachain, M.; Dewey, F.E.; Habegger, L.; Ashley, E.A.; Gerstein, M.B.; et al. Performance comparison of whole-genome sequencing platforms. *Nat. Biotechnol.* **2012**, *30*, 78–82. [CrossRef] [PubMed]
12. Quince, C.; Walker, A.W.; Simpson, J.T.; Loman, N.J.; Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **2017**, *35*, 833–844. [CrossRef] [PubMed]
13. Wang, Z.; Gerstein, M.; Snyder, M. RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **2009**, *10*, 57–63. [CrossRef] [PubMed]
14. Marguerat, S.; Bähler, J. RNA-seq: From technology to biology. *Cell. Mol. Life Sci.* **2010**, *67*, 569–579. [CrossRef] [PubMed]
15. Amarasinghe, S.L.; Su, S.; Dong, X.; Zappia, L.; Ritchie, M.E.; Gouil, Q. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* **2020**, *21*, 30. [CrossRef]
16. Saliba, A.-E.; Westermann, A.J.; Gorski, S.A.; Vogel, J. Single-cell RNA-seq: Advances and future challenges. *Nucleic Acids Res.* **2014**, *42*, 8845–8860. [CrossRef]
17. McKenna, A.; Ijaz, U.Z.; Kelly, C.; Linton, M.; Sloan, W.T.; Green, B.D.; Lavery, U.; Dorrell, N.; Wren, B.W.; Richmond, A.; et al. Impact of industrial production system parameters on chicken microbiomes: Mechanisms to improve performance and reduce Campylobacter. *Microbiome* **2020**, *8*, 128. [CrossRef]
18. Trego, A.C.; McAteer, P.G.; Nzeteu, C.; Mahony, T.; Abram, F.; Ijaz, U.Z.; O'Flaherty, V. Combined Stochastic and Deterministic Processes Drive Community Assembly of Anaerobic Microbiomes during Granule Flotation. *Front. Microbiol.* **2021**, *12*, 1165. [CrossRef]
19. Nikolova, C.; Ijaz, U.Z.; Gutierrez, T. Exploration of marine bacterioplankton community assembly mechanisms during chemical dispersant and surfactant-assisted oil biodegradation. *Ecol. Evol.* **2021**, *11*, 13862–13874. [CrossRef]
20. Liu, R.; Wei, X.; Song, W.; Wang, L.; Cao, J.; Wu, J.; Thomas, T.; Jin, T.; Wang, Z.; Wei, W.; et al. Novel Chloroflexi genomes from the deepest ocean reveal metabolic strategies for the adaptation to deep-sea habitats. *Microbiome* **2022**, *10*, 75. [CrossRef]
21. Meek, M.H.; Larson, W.A. The future is now: Amplicon sequencing and sequence capture usher in the conservation genomics era. *Mol. Ecol. Resour.* **2019**, *19*, 795–803. [CrossRef] [PubMed]
22. Quast, C.; Pruesse, E.; Yilmaz, P.; Gerken, J.; Schweer, T.; Yarza, P.; Peplies, J.; Glöckner, F.O. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res.* **2013**, *41*, D590–D596. [CrossRef] [PubMed]

23. Yilmaz, P.; Parfrey, L.W.; Yarza, P.; Gerken, J.; Pruesse, E.; Quast, C.; Schweer, T.; Peplies, J.; Ludwig, W.; Glöckner, F.O. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Res.* **2014**, *42*, D643–D648. [CrossRef] [PubMed]
24. Dueholm, M.K.D.; Nierychlo, M.; Andersen, K.S.; Rudkjøbing, V.; Knutsson, S.; Albertsen, M.; Nielsen, P.H. MiDAS 4: A global catalogue of full-length 16S rRNA gene sequences and taxonomy for studies of bacterial communities in wastewater treatment plants. *Nat. Commun.* **2022**, *13*, 1908. [CrossRef] [PubMed]
25. Rohwer, R.; Hamilton, J.; Newton, R.; McMahon, K.; Rodrigues, J. TaxAss: Leveraging a Custom Freshwater Database Achieves Fine-Scale Taxonomic Resolution. *mSphere* **2022**, *3*, e00327-18. [CrossRef] [PubMed]
26. Prodan, A.; Tremaroli, V.; Brolin, H.; Zwinderman, A.H.; Nieuwdorp, M.; Levin, E. Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. *PLoS ONE* **2020**, *15*, e0227434. [CrossRef]
27. van Dijk, E.L.; Jaszczyszyn, Y.; Naquin, D.; Thermes, C. The Third Revolution in Sequencing Technology. *Trends Genet.* **2018**, *34*, 666–681. [CrossRef]
28. Poretsky, R.; Rodriguez-R, L.M.; Luo, C.; Tsementzi, D.; Konstantinidis, K.T. Strengths and Limitations of 16S rRNA Gene Amplicon Sequencing in Revealing Temporal Microbial Community Dynamics. *PLoS ONE* **2014**, *9*, e93827. [CrossRef]
29. Verster, A.J.; Borenstein, E. Competitive lottery-based assembly of selected clades in the human gut microbiome. *Microbiome* **2018**, *6*, 186. [CrossRef]
30. Eng, A.; Borenstein, E. Taxa-function robustness in microbial communities. *Microbiome* **2018**, *6*, 45. [CrossRef]
31. Darcy, J.L.; Amend, A.S.; Swift, S.O.I.; Sommers, P.S.; Lozupone, C.A. specificity: An R package for analysis of feature specificity to environmental and higher dimensional variables, applied to microbiome species data. *Environ. Microbiome* **2022**, *17*, 34. [CrossRef] [PubMed]
32. Calle, M.L.; Susin, A. coda4microbiome: Compositional data analysis for microbiome studies. *bioRxiv* **2022**. [CrossRef]
33. Niku, J.; Hui, F.K.C.; Taskinen, S.; Warton, D.I. gllvm: Fast analysis of multivariate abundance data with generalized linear latent variable models in R. *Methods Ecol. Evol.* **2019**, *10*, 2173–2182. [CrossRef]
34. Bharti, R.; Grimm, D.G. Current challenges and best-practice protocols for microbiome analysis. *Brief. Bioinform.* **2021**, *22*, 178–193. [CrossRef] [PubMed]
35. Maki, K.A.; Diallo, A.F.; Lockwood, M.B.; Franks, A.T.; Green, S.J.; Joseph, P. V Considerations when designing a microbiome study: Implications for nursing science. *Biol. Res. Nurs.* **2019**, *21*, 125–141. [CrossRef]
36. Goodrich, J.K.; Di Rienzi, S.C.; Poole, A.C.; Koren, O.; Walters, W.A.; Caporaso, J.G.; Knight, R.; Ley, R.E. Conducting a Microbiome Study. *Cell* **2014**, *158*, 250–262. [CrossRef]
37. Salter, S.J.; Cox, M.J.; Turek, E.M.; Calus, S.T.; Cookson, W.O.; Moffatt, M.F.; Turner, P.; Parkhill, J.; Loman, N.J.; Walker, A.W. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **2014**, *12*, 87. [CrossRef]
38. D'Amore, R.; Ijaz, U.Z.; Schirmer, M.; Kenny, J.G.; Gregory, R.; Darby, A.C.; Shakya, M.; Podar, M.; Quince, C.; Hall, N. A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. *BMC Genom.* **2016**, *17*, 55. [CrossRef]
39. Schirmer, M.; Ijaz, U.Z.; D'Amore, R.; Hall, N.; Sloan, W.T.; Quince, C. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res.* **2015**, *43*, gku1341. [CrossRef]
40. Quince, C.; Ijaz, U.Z.; Loman, N.; Eren, A.M.; Saulnier, D.; Russell, J.; Haig, S.J.; Calus, S.T.; Quick, J.; Barclay, A. Extensive modulation of the fecal metagenome in children with Crohn's disease during exclusive enteral nutrition. *Am. J. Gastroenterol.* **2015**, *110*, 1718. [CrossRef]
41. Schielzeth, H.; Nakagawa, S. Nested by design: Model fitting and interpretation in a mixed model era. *Methods Ecol. Evol.* **2013**, *4*, 14–24. [CrossRef]
42. Rahman, G.; McDonald, D.; Gonzalez, A.; Vázquez-Baeza, Y.; Jiang, L.; Casals-Pascual, C.; Peddada, S.; Hakim, D.; Dilmore, A.H.; Nowinski, B.; et al. Scalable power analysis and effect size exploration of microbiome community differences with Evident. *bioRxiv* **2022**. [CrossRef]
43. Trego, A.C.; O'Sullivan, S.; Quince, C.; Mills, S.; Ijaz, U.Z.; Collins, G. Size Shapes the Active Microbiome of the Methanogenic Granules, Corroborating a Biofilm Life Cycle. *mSystems* **2020**, *5*, e00323-20. [CrossRef] [PubMed]
44. Trego, A.C.; Holohan, B.C.; Keating, C.; Graham, A.; O'Connor, S.; Gerardo, M.; Hughes, D.; Zeeshan Ijaz, U.; O'Flaherty, V. First Proof of Concept for Full-Scale, Direct, Low-Temperature Anaerobic Treatment of Municipal Wastewater. *Bioresour. Technol.* **2021**, *341*, 125786. [CrossRef]
45. Keating, C.; Bolton-Warberg, M.; Hinchcliffe, J.; Davies, R.; Whelan, S.; Wan, A.H.L.; Fitzgerald, R.D.; Davies, S.J.; Smith, C.J.; Ijaz, U.Z. Key Drivers of Ecological Assembly in the Hindgut of Atlantic Cod (*Gadus morhua*) when Fed with a Macroalgal Supplemented diet—How Robust Is the Gut to Taxonomic Perturbation? *bioRxiv* **2021**. [CrossRef]
46. Thom, C.; Smith, C.J.; Moore, G.; Weir, P.; Ijaz, U.Z. Microbiomes in drinking water treatment and distribution: A meta-analysis from source to tap. *Water Res.* **2022**, *212*, 118106. [CrossRef]
47. Yonatan, Y.; Amit, G.; Friedman, J.; Bashan, A. Complexity–stability trade-off in empirical microbial ecosystems. *Nat. Ecol. Evol.* **2022**, *6*, 693–700. [CrossRef]
48. Schloss, P.D.; Westcott, S.L.; Ryabin, T.; Hall, J.R.; Hartmann, M.; Hollister, E.B.; Lesniewski, R.A.; Oakley, B.B.; Parks, D.H.; Robinson, C.J. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **2009**, *75*, 7537–7541. [CrossRef]

49. Bolyen, E.; Rideout, J.R.; Dillon, M.R.; Bokulich, N.A.; Abnet, C.C.; Al-Ghalith, G.A.; Alexander, H.; Alm, E.J.; Arumugam, M.; Asnicar, F. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **2019**, *37*, 852–857. [CrossRef]
50. Lu, J.; Salzberg, S.L. Ultrafast and accurate 16S rRNA microbial community analysis using Kraken 2. *Microbiome* **2020**, *8*, 124. [CrossRef]
51. Callahan, B.J.; McMurdie, P.J.; Holmes, S.P. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* **2017**, *11*, 2639–2643. [CrossRef] [PubMed]
52. Glassman, S.I.; Martiny, J.B.H. Broadscale ecological patterns are robust to use of exact sequence variants versus operational taxonomic units. *mSphere* **2018**, *3*, e00148-18. [CrossRef] [PubMed]
53. Schloss, P.D.; McMahon, K. Amplicon Sequence Variants Artificially Split Bacterial Genomes into Separate Clusters. *mSphere* **2021**, *6*, e00191-21. [CrossRef] [PubMed]
54. Cholet, F.; Lisik, A.; Agogue, H.; Ijaz, U.Z.; Pineau, P.; Lachaussée, N.; Smith, C.J. Ecological Observations Based on Functional Gene Sequencing Are Sensitive to the Amplicon Processing Method. *bioRxiv* **2022**. [CrossRef] [PubMed]
55. Rognes, T.; Flouri, T.; Nichols, B.; Quince, C.; Mahé, F. VSEARCH: A versatile open source tool for metagenomics. *PeerJ* **2016**, *2016*, 2584. [CrossRef] [PubMed]
56. Edgar, R.C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **2010**, *26*, 2460–2461. [CrossRef] [PubMed]
57. Edgar, R.C. UNOISE2: Improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv* **2016**. [CrossRef]
58. Abellan-Schneyder, I.; Matchado, M.S.; Reitmeier, S.; Sommer, A.; Sewald, Z.; Baumbach, J.; List, M.; Neuhaus, K. Primer, Pipelines, Parameters: Issues in 16S rRNA Gene Sequencing. *mSphere* **2021**, *6*, e01202-20. [CrossRef]
59. Callahan, B.J.; McMurdie, P.J.; Rosen, M.J.; Han, A.W.; Johnson, A.J.A.; Holmes, S.P. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **2016**, *13*, 581–583. [CrossRef]
60. Amir, A.; McDonald, D.; Navas-Molina, J.A.; Kopylova, E.; Morton, J.T.; Zech Xu, Z.; Kightley, E.P.; Thompson, L.R.; Hyde, E.R.; Gonzalez, A.; et al. Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *mSystems* **2017**, *2*, e00191-16. [CrossRef]
61. Gao, X.; Lin, H.; Revanna, K.; Dong, Q. A Bayesian taxonomic classification method for 16S rRNA gene sequences with improved species-level accuracy. *BMC Bioinformatics* **2017**, *18*, 247. [CrossRef] [PubMed]
62. McDonald, D.; Price, M.N.; Goodrich, J.; Nawrocki, E.P.; DeSantis, T.Z.; Probst, A.; Andersen, G.L.; Knight, R.; Hugenholtz, P. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* **2012**, *6*, 610–618. [CrossRef] [PubMed]
63. Wang, Q.; Garrity, G.M.; Tiedje, J.M.; Cole, J.R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **2007**, *73*, 5261–5267. [CrossRef] [PubMed]
64. Simonsen, D.M.; Skytte, A.K.; Jon, M.S.; Munk, K.J.; Erika, Y.; Michael, K.S.; Mads, A.; Halkjær, N.P.; Nicole, D. Generation of Comprehensive Ecosystem-Specific Reference Databases with Species-Level Resolution by High-Throughput Full-Length 16S rRNA Gene Sequencing and Automated Taxonomy Assignment (AutoTax). *MBio* **2020**, *11*, e01557-20. [CrossRef]
65. Choi, J.; Yang, F.; Stepanauskas, R.; Cardenas, E.; Garoutte, A.; Williams, R.; Flater, J.; Tiedje, J.M.; Hofmockel, K.S.; Gelder, B.; et al. Strategies to improve reference databases for soil microbiomes. *ISME J.* **2017**, *11*, 829–834. [CrossRef] [PubMed]
66. Madi, N.; Vos, M.; Murall, C.L.; Legendre, P.; Shapiro, B.J. Does diversity beget diversity in microbiomes? *eLife* **2020**, *9*, e58999. [CrossRef] [PubMed]
67. Oksanen, J.; Blanchet, F.G.; Friendly, M.; Kindt, R.; Legendre, P.; McGlinn, D.; Minchin, P.R.; O'hara, R.B.; Simpson, G.L.; Solymos, P. *Vegan: Community Ecology Package. R package, v. 2.4–6*; R Core Team: Vienna, Austria, 2018.
68. McMurdie, P.J.; Holmes, S. Phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* **2013**, *8*, e61217. [CrossRef]
69. Lande, R. Statistics and Partitioning of Species Diversity, and Similarity among Multiple Communities. *Oikos* **1996**, *76*, 5–13. [CrossRef]
70. Shannon, C.E.; Weaver, W. *The Mathematical Theory of Communication*; University of Illinois: Champaign, IL, USA, 1949.
71. Fisher, R.A.; Corbet, A.S.; Williams, C.B. The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.* **1943**, *12*, 42–58. [CrossRef]
72. Simpson, E.H. Measurement of Diversity. *Nature* **1949**, *163*, 688. [CrossRef]
73. Hill, M.O. Diversity and Evenness: A Unifying Notation and Its Consequences. *Ecology* **1973**, *54*, 427–432. [CrossRef]
74. Willis, A.D. Rarefaction, Alpha Diversity, and Statistics. *Front. Microbiol.* 2019. [CrossRef] [PubMed]
75. Lozupone, C.; Lladser, M.E.; Knights, D.; Stombaugh, J.; Knight, R. UniFrac: An effective distance metric for microbial community comparison. *ISME J.* **2011**, *5*, 169–172. [CrossRef] [PubMed]
76. Lozupone, C.A.; Micah, H.; Kelley, S.T.; Knight, R. Quantitative and Qualitative β Diversity Measures Lead to Different Insights into Factors That Structure Microbial Communities. *Appl. Environ. Microbiol.* **2007**, *73*, 1576–1585. [CrossRef] [PubMed]
77. Bray, J.R.; Curtis, J.T. An ordination of the upland forest communities of southern Wisconsin. *Ecol. Monogr.* **1957**, *27*, 326–349. [CrossRef]
78. Beals, E.W. Bray-Curtis Ordination: An Effective Strategy for Analysis of Multivariate Ecological Data. In *Advances in Ecological Research*; MacFadyen, A., Ford, E.D., Eds.; Academic Press: Cambridge, MA, USA, 1984; Volume 14, pp. 1–55. ISBN 0065-2504.
79. Ramette, A. Multivariate analyses in microbial ecology. *FEMS Microbiol. Ecol.* **2007**, *62*, 142–160. [CrossRef]

80. Paliy, O.; Shankar, V. Application of multivariate statistical techniques in microbial ecology. *Mol. Ecol.* **2016**, *25*, 1032–1057. [CrossRef]
81. Pierre, L.; Miquel, C.; Hélène, M. Beta diversity as the variance of community data: Dissimilarity coefficients and partitioning. *Ecol. Lett.* **2013**, *16*, 951–963. [CrossRef]
82. Dray, S.; Blanchet, G.; Borcard, D.; Guenard, G.; Jombart, T.; Larocque, G.; Legendre, P.; Madi, N.; Wagner, H.H. Package ' adespatial'. 2017.
83. Keating, C.; Bolton-Warberg, M.; Hinchcliffe, J.; Davies, R.; Whelan, S.; Wan, A.H.L.; Fitzgerald, R.D.; Davies, S.J.; Ijaz, U.Z.; Smith, C.J. Temporal changes in the gut microbiota in farmed Atlantic cod (*Gadus morhua*) outweigh the response to diet supplementation with macroalgae. *Anim. Microbiome* **2021**, *3*, 7. [CrossRef]
84. Rohart, F.; Gautier, B.; Singh, A.; Lê Cao, K.-A. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLOS Comput. Biol.* **2017**, *13*, e1005752. [CrossRef]
85. Jalanka-Tuovinen, J.; Salonen, A.; Nikkilä, J.; Immonen, O.; Kekkonen, R.; Lahti, L.; Palva, A.; de Vos, W.M. Intestinal Microbiota in Healthy Adults: Temporal Analysis Reveals Individual and Common Core and Relation to Intestinal Symptoms. *PLoS ONE* **2011**, *6*, e23035. [CrossRef] [PubMed]
86. Dai, T.; Zhang, Y.; Tang, Y.; Bai, Y.; Tao, Y.; Huang, B.; Wen, D. Identifying the key taxonomic categories that characterize microbial community diversity using full-scale classification: A case study of microbial communities in the sediments of Hangzhou Bay. *FEMS Microbiol. Ecol.* **2016**, *92*, fiw150. [CrossRef] [PubMed]
87. De Vrieze, J.; Pinto, A.J.; Sloan, W.T.; Ijaz, U.Z. The active microbial community more accurately reflects the anaerobic digestion process: 16S rRNA (gene) sequencing as a predictive tool. *Microbiome* **2018**, *6*, 63. [CrossRef]
88. Props, R.; Kerckhof, F.-M.; Rubbens, P.; De Vrieze, J.; Hernandez Sanabria, E.; Waegeman, W.; Monsieurs, P.; Hammes, F.; Boon, N. Absolute quantification of microbial taxon abundances. *ISME J.* **2017**, *11*, 584–587. [CrossRef] [PubMed]
89. Furman, O.; Shenhav, L.; Sasson, G.; Kokou, F.; Honig, H.; Jacoby, S.; Hertz, T.; Cordero, O.X.; Halperin, E.; Mizrahi, I. Stochasticity constrained by deterministic effects of diet and age drive rumen microbiome assembly dynamics. *Nat. Commun.* **2020**, *11*, 1904. [CrossRef] [PubMed]
90. Leibold, M.A.; Mikkelson, G.M. Coherence, species turnover, and boundary clumping: Elements of meta-community structure. *Oikos* **2002**, *97*, 237–250. [CrossRef]
91. Vass, M.; Székely, A.J.; Lindström, E.S.; Langenheder, S. Using null models to compare bacterial and microeukaryotic metacommunity assembly under shifting environmental conditions. *Sci. Rep.* **2020**, *10*, 2455. [CrossRef] [PubMed]
92. Presley, S.J.; Higgins, C.L.; Willig, M.R. A comprehensive framework for the evaluation of metacommunity structure. *Oikos* **2010**, *119*, 908–917. [CrossRef]
93. Zhou, J.; Ning, D. Stochastic Community Assembly: Does It Matter in Microbial Ecology? *Microbiol. Mol. Biol. Rev.* **2017**, *81*, e00002-17. [CrossRef]
94. Stegen, J.C.; Lin, X.; Fredrickson, J.K.; Chen, X.; Kennedy, D.W.; Murray, C.J.; Rockhold, M.L.; Konopka, A. Quantifying community assembly processes and identifying features that impose them. *ISME J.* **2013**, *7*, 2069–2079. [CrossRef]
95. Stegen, J.C.; Lin, X.; Fredrickson, J.K.; Konopka, A.E. Estimating and mapping ecological processes influencing microbial community assembly. *Front. Microbiol.* **2015**, *6*, 370. [CrossRef] [PubMed]
96. Ning, D.; Yuan, M.; Wu, L.; Zhang, Y.; Guo, X.; Zhou, X.; Yang, Y.; Arkin, A.P.; Firestone, M.K.; Zhou, J. A quantitative framework reveals ecological drivers of grassland microbial community assembly in response to warming. *Nat. Commun.* **2020**, *11*, 4717. [CrossRef] [PubMed]
97. Ning, D.; Deng, Y.; Tiedje, J.M.; Zhou, J. A general framework for quantitatively assessing ecological stochasticity. *Proc. Natl. Acad. Sci.* **2019**, *116*, 16892–16898. [CrossRef]
98. Modin, O.; Liébana, R.; Saheb-Alam, S.; Wilén, B.-M.; Suarez, C.; Hermansson, M.; Persson, F. Hill-based dissimilarity indices and null models for analysis of microbial community assembly. *Microbiome* **2020**, *8*, 132. [CrossRef]
99. Hubbell, S.P. The unified neutral theory of biodiversity and biogeography (MPB-32). In *The Unified Neutral Theory of Biodiversity and Biogeography (MPB-32)*; Princeton University Press: Princeton, NJ, USA, 2011; ISBN 1400837529.
100. Sloan, W.T.; Lunn, M.; Woodcock, S.; Head, I.M.; Nee, S.; Curtis, T.P. Quantifying the roles of immigration and chance in shaping prokaryote community structure. *Environ. Microbiol.* **2006**, *8*, 732–740. [CrossRef] [PubMed]
101. Tucker, C.M.; Shoemaker, L.G.; Davies, K.F.; Nemergut, D.R.; Melbourne, B.A. Differentiating between niche and neutral assembly in metacommunities using null models of β-diversity. *Oikos* **2016**, *125*, 778–789. [CrossRef]
102. Kokou, F.; Sasson, G.; Friedman, J.; Eyal, S.; Ovadia, O.; Harpaz, S.; Cnaani, A.; Mizrahi, I. Core gut microbial communities are maintained by beneficial interactions and strain variability in fish. *Nat. Microbiol.* **2019**, *4*, 2456–2465. [CrossRef]
103. Finn, D.R.; Yu, J.; Ilhan, Z.E.; Fernandes, V.M.C.; Penton, C.R.; Krajmalnik-Brown, R.; Garcia-Pichel, F.; Vogel, T.M. MicroNiche: An R package for assessing microbial niche breadth and overlap from amplicon sequencing data. *FEMS Microbiol. Ecol.* **2020**, *96*, fiaa131. [CrossRef] [PubMed]
104. Golovko, G.; Kamil, K.; Albayrak, L.; Nia, A.M.; Duarte, R.S.A.; Chumakov, S.; Fofanov, Y. Identification of multidimensional Boolean patterns in microbial communities. *Microbiome* **2020**, *8*, 131. [CrossRef]
105. Durán, C.; Ciucci, S.; Palladini, A.; Ijaz, U.Z.; Zippo, A.G.; Sterbini, F.P.; Masucci, L.; Cammarota, G.; Ianiro, G.; Spuul, P.; et al. Nonlinear machine learning pattern recognition and bacteria-metabolite multilayer network analysis of perturbed gastric microbiome. *Nat. Commun.* **2021**, *12*, 1926. [CrossRef]

106. Röttjers, L.; Faust, K. From hairballs to hypotheses–biological insights from microbial networks. *FEMS Microbiol. Rev.* **2018**, *42*, 761–780. [CrossRef] [PubMed]
107. Willis, A.D.; Martin, B.D. Estimating diversity in networked ecological communities. *Biostatistics* **2022**, *23*, 207–222. [CrossRef] [PubMed]
108. Oulas, A.; Zachariou, M.; Chasapis, C.T.; Tomazou, M.; Ijaz, U.Z.; Schmartz, G.P.; Spyrou, G.M.; Vlamis-Gardikas, A. Putative Antimicrobial Peptides within Bacterial Proteomes Affect Bacterial Predominance: A Network Analysis Perspective. *Front. Microbiol.* **2021**, *12*, 752674. [CrossRef]
109. Kumar, M.; Ji, B.; Zengler, K.; Nielsen, J. Modelling approaches for studying the microbiome. *Nat. Microbiol.* **2019**, *4*, 1253–1267. [CrossRef] [PubMed]
110. Kurtz, Z.D.; Müller, C.L.; Miraldi, E.R.; Littman, D.R.; Blaser, M.J.; Bonneau, R.A. Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLOS Comput. Biol.* **2015**, *11*, e1004226. [CrossRef] [PubMed]
111. Donohue, I.; Hillebrand, H.; Montoya, J.M.; Petchey, O.L.; Pimm, S.L.; Fowler, M.S.; Healy, K.; Jackson, A.L.; Lurgi, M.; McClean, D.; et al. Navigating the complexity of ecological stability. *Ecol. Lett.* **2016**, *19*, 1172–1185. [CrossRef] [PubMed]
112. Kéfi, S.; Domínguez-García, V.; Donohue, I.; Fontaine, C.; Thébault, E.; Dakos, V. Advancing our understanding of ecological stability. *Ecol. Lett.* **2019**, *22*, 1349–1356. [CrossRef] [PubMed]
113. May, R.M. Will a Large Complex System be Stable? *Nature* **1972**, *238*, 413–414. [CrossRef] [PubMed]
114. Becraft, E.D.; Vetter, M.C.Y.L.; Bezuidt, O.K.I.; Brown, J.M.; Labonté, J.M.; Kauneckaite-Griguole, K.; Salkauskaite, R.; Alzbutas, G.; Sackett, J.D.; Kruger, B.R. Evolutionary stasis of a deep subsurface microbial lineage. *ISME J.* **2021**, *15*, 2830–2842. [CrossRef]
115. Hamdan, L.J.; Hampel, J.J.; Moseley, R.D.; Mugge, R.L.; Ray, A.; Salerno, J.L.; Damour, M. Deep-sea shipwrecks represent island-like ecosystems for marine microbiomes. *ISME J.* **2021**, 1–9. [CrossRef]
116. Ramirez, K.S.; Knight, C.G.; De Hollander, M.; Brearley, F.Q.; Constantinides, B.; Cotton, A.; Creer, S.; Crowther, T.W.; Davison, J.; Delgado-Baquerizo, M. Detecting macroecological patterns in bacterial communities across independent studies of global soils. *Nat. Microbiol.* **2018**, *3*, 189–196. [CrossRef] [PubMed]
117. McAteer, P.G.; Trego, A.C.; Thorn, C.; Mahony, T.; Abram, F.; O'Flaherty, V. Reactor configuration influences microbial community structure during high-rate, low-temperature anaerobic treatment of dairy wastewater. *Bioresour. Technol.* **2020**, *307*, 123221. [CrossRef] [PubMed]
118. Bokulich, N.A.; Dillon, M.R.; Yilong, Z.; Ram, R.J.; Evan, B.; Huilin, L.; Albert, P.S.; Gregory, C.J.; Mani, A. q2-longitudinal: Longitudinal and Paired-Sample Analyses of Microbiome Data. *mSystems* **2021**, *3*, e00219-18. [CrossRef] [PubMed]
119. Riva, F.; Mammola, S. Rarity facets of biodiversity: Integrating Zeta diversity and Dark diversity to understand the nature of commonness and rarity. *Ecol. Evol.* **2021**, *11*, 13912–13919. [CrossRef]
120. Buckley, H.L.; Day, N.J.; Case, B.S.; Lear, G. Measuring change in biological communities: Multivariate analysis approaches for temporal datasets with low sample size. *PeerJ* **2021**, *9*, e11096. [CrossRef]
121. Darcy, J.L.; Washburne, A.D.; Robeson, M.S.; Prest, T.; Schmidt, S.K.; Lozupone, C.A. A phylogenetic model for the recruitment of species into microbial communities and application to studies of the human microbiome. *ISME J.* **2020**, *14*, 1359–1368. [CrossRef]
122. Shields-Cutler, R.R.; Al-Ghalith, G.A.; Yassour, M.; Knights, D. SplinectomeR Enables Group Comparisons in Longitudinal Microbiome Studies. *Front. Microbiol.* **2018**, *9*, 785. [CrossRef]
123. Bodein, A.; Chapleur, O.; Droit, A.; Lê Cao, K.-A. A Generic Multivariate Framework for the Integration of Microbiome Longitudinal Studies With Other Data Types. *Front. Genet.* **2019**, *10*, 963. [CrossRef]
124. Shenhav, L.; Furman, O.; Briscoe, L.; Thompson, M.; Silverman, J.D.; Mizrahi, I.; Halperin, E. Modeling the temporal dynamics of the gut microbial community in adults and infants. *PLOS Comput. Biol.* **2019**, *15*, e1006960. [CrossRef]
125. McGlinn, D.J.; Xiao, X.; May, F.; Gotelli, N.J.; Engel, T.; Blowes, S.A.; Knight, T.M.; Purschke, O.; Chase, J.M.; McGill, B.J. Measurement of Biodiversity (MoB): A method to separate the scale-dependent effects of species abundance distribution, density, and aggregation on diversity change. *Methods Ecol. Evol.* **2019**, *10*, 258–269. [CrossRef]
126. Karczewski, K.J.; Snyder, M.P. Integrative omics for health and disease. *Nat. Rev. Genet.* **2018**, *19*, 299–310. [CrossRef] [PubMed]
127. Planell, N.; Lagani, V.; Sebastian-Leon, P.; van der Kloet, F.; Ewing, E.; Karathanasis, N.; Urdangarin, A.; Arozarena, I.; Jagodic, M.; Tsamardinos, I.; et al. STATegra: Multi-Omics Data Integration—A Conceptual Scheme with a Bioinformatics Pipeline. *Front. Genet.* **2021**, *12*, 620453. [CrossRef] [PubMed]
128. Mills, S.; Trego, A.C.; Lens, P.N.L.; Ijaz, U.Z.; Collins, G. A Distinct, Flocculent, Acidogenic Microbial Community Accompanies Methanogenic Granules in Anaerobic Digesters. *Microbiol. Spectr.* **2021**, *9*, e00784-21. [CrossRef] [PubMed]
129. Frau, A.; Ijaz, U.Z.; Slater, R.; Jonkers, D.; Penders, J.; Campbell, B.J.; Kenny, J.G.; Hall, N.; Lenzi, L.; Burkitt, M.D.; et al. Inter-kingdom relationships in Crohn's disease explored using a multi-omics approach. *Gut Microbes* **2021**, *13*, 1930871. [CrossRef] [PubMed]
130. Lu, J.; Shi, P.; Li, H. Generalized linear models with linear constraints for microbiome compositional data. *Biometrics* **2019**, *75*, 235–244. [CrossRef] [PubMed]
131. Ovaskainen, O.; Tikhonov, G.; Norberg, A.; Guillaume Blanchet, F.; Duan, L.; Dunson, D.; Roslin, T.; Abrego, N. How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecol. Lett.* **2017**, *20*, 561–576. [CrossRef]
132. Ovaskainen, O.; Abrego, N. *Joint Species Distribution Modelling: With Applications in R*; Cambridge University Press: Cambridge, UK, 2020; ISBN 1108674151.

133. Tikhonov, G.; Opedal, Ø.H.; Abrego, N.; Lehikoinen, A.; de Jonge, M.M.J.; Oksanen, J.; Ovaskainen, O. Joint species distribution modelling with the r-package Hmsc. *Methods Ecol. Evol.* **2020**, *11*, 442–447. [CrossRef]
134. Skrondal, A.; Rabe-Hesketh, S. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2004; ISBN 042920549X.
135. Warton, D.I.; Blanchet, F.G.; O'Hara, R.B.; Ovaskainen, O.; Taskinen, S.; Walker, S.C.; Hui, F.K.C. So many variables: Joint modeling in community ecology. *Trends Ecol. Evol.* **2015**, *30*, 766–779. [CrossRef]
136. Niku, J.; Warton, D.I.; Hui, F.K.C.; Taskinen, S. Generalized linear latent variable models for multivariate count and biomass data in ecology. *J. Agric. Biol. Environ. Stat.* **2017**, *22*, 498–522. [CrossRef]
137. Aitchison, J. The statistical analysis of compositional data. *J. R. Stat. Soc. Ser. B* **1982**, *44*, 139–160. [CrossRef]
138. Susin, A.; Wang, Y.; Lê Cao, K.-A.; Calle, M.L. Variable selection in microbiome compositional data analysis. *NAR Genom. Bioinform.* **2020**, *2*, lqaa029. [CrossRef] [PubMed]
139. Rivera-Pinto, J.; Egozcue, J.J.; Pawlowsky-Glahn, V.; Paredes, R.; Noguera-Julian, M.; Calle, M.L. Balances: A new perspective for microbiome analysis. *mSystems* **2018**, *3*, e00053-18. [CrossRef]
140. le Cessie, S.; van Houwelingen, J.C. Ridge Estimators in Logistic Regression. *J. R. Stat. Soc. Ser. C Appl. Stat.* **1992**, *41*, 191–201. [CrossRef]
141. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2005**, *67*, 301–320. [CrossRef]
142. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267–288. [CrossRef]
143. Knight, R.; Vrbanac, A.; Taylor, B.C.; Aksenov, A.; Callewaert, C.; Debelius, J.; Gonzalez, A.; Kosciolek, T.; McCall, L.-I.; McDonald, D. Best practices for analysing microbiomes. *Nat. Rev. Microbiol.* **2018**, *16*, 410–422. [CrossRef]
144. Langille, M.G.I. Exploring linkages between taxonomic and functional profiles of the human microbiome. *mSystems* **2018**, *3*, e00163-17. [CrossRef]
145. Huttenhower, C.; Gevers, D.; Knight, R.; Abubucker, S.; Badger, J.H.; Chinwalla, A.T.; Giglio, M.G. Structure, function and diversity of the healthy human microbiome. *Nature* **2012**, *486*, 207.
146. Jing, G.; Zhang, Y.; Cui, W.; Liu, L.; Xu, J.; Su, X. Meta-Apo improves accuracy of 16S-amplicon-based prediction of microbiome function. *BMC Genom.* **2021**, *22*, 9. [CrossRef]
147. Douglas, G.M.; Maffei, V.J.; Zaneveld, J.R.; Yurgel, S.N.; Brown, J.R.; Taylor, C.M.; Huttenhower, C.; Langille, M.G.I. PICRUSt2 for prediction of metagenome functions. *Nat. Biotechnol.* **2020**, *38*, 685–688. [CrossRef]
148. Wemheuer, F.; Taylor, J.A.; Daniel, R.; Johnston, E.; Meinicke, P.; Thomas, T.; Wemheuer, B. Tax4Fun2: Prediction of habitat-specific functional profiles and functional redundancy based on 16S rRNA gene sequences. *Environ. Microbiome* **2020**, *15*, 11. [CrossRef] [PubMed]
149. Louca, S.; Parfrey, L.W.; Doebeli, M. Decoupling function and taxonomy in the global ocean microbiome. *Science* **2016**, *353*, 1272–1277. [CrossRef] [PubMed]
150. Ward, T.; Larson, J.; Meulemans, J.; Hillmann, B.; Lynch, J.; Sidiropoulos, D.; Spear, J.R.; Caporaso, G.; Blekhman, R.; Knight, R.; et al. BugBase predicts organism-level microbiome phenotypes. *bioRxiv* **2017**. [CrossRef]
151. Zhang, Y.; Jing, G.; Chen, Y.; Li, J.; Su, X. Hierarchical Meta-Storms enables comprehensive and rapid comparison of microbiome functional profiles on a large scale using hierarchical dissimilarity metrics and parallel computing. *Bioinforma. Adv.* **2021**, *1*, vbab003. [CrossRef]
152. Campanaro, S.; Treu, L.; Rodriguez-R, L.M.; Kovalovszki, A.; Ziels, R.M.; Maus, I.; Zhu, X.; Kougias, P.G.; Basile, A.; Luo, G.; et al. The anaerobic digestion microbiome: A collection of 1600 metagenome-assembled genomes shows high species diversity related to methane production. *bioRxiv* **2019**. [CrossRef]