environmental microbiology

Environmental Microbiology (2020) **00**(00), 00–00 doi:10.1111/1462-2920.15091

Mini Review

Applications of weighted association networks applied to compositional data in biology

Josh L. Espinoza[1,2] **Naisha Shah**[1] **Suren Singh**[2] **Karen E. Nelson**[1,2,3] **and Chris L. Dupont**[1*]

[1]*J. Craig Venter Institute, La Jolla, USA.*

[2]*Applied Sciences, Durban University of Technology, Durban, South Africa.*

[3]*J. Craig Venter Institute, Rockville, USA.*

Summary

Next-generation sequencing technologies have generated, and continue to produce, an increasingly large corpus of biological data. The data generated are inherently compositional as they convey only relative information dependent upon the capacity of the instrument, experimental design and technical bias. There is considerable information to be gained through network analysis by studying the interactions between components within a system. Network theory methods using compositional data are powerful approaches for quantifying relationships between biological components and their relevance to phenotype, environmental conditions or other external variables. However, many of the statistical assumptions used for network analysis are not designed for compositional data and can bias downstream results. In this mini-review, we illustrate the utility of network theory in biological systems and investigate modern techniques while introducing researchers to frameworks for implementation. We overview (1) compositional data analysis, (2) data transformations and (3) network theory along with insight on a battery of network types including static-, temporal-, sample-specific- and differential-networks. The intention of this mini-review is not to provide a comprehensive overview of network methods, rather to introduce microbiology researchers to (semi)-unsupervised data-driven approaches for inferring latent structures that may give insight into biological phenomena or abstract mechanics of complex systems.

Received 23 March, 2020; revised 15 May, 2020; accepted 18 May, 2020. *For correspondence. E-mail cdupont@jcvi.org; Tel.: 8582001886; Fax: 8582001800

Introduction

With the advent of next-generation sequencing (NGS) technologies, deep profiling of biological systems has become increasingly affordable and the collection of publicly available datasets is growing exponentially. NGS instruments estimate the relative abundance of discrete biological components (e.g. transcripts, 16/18SS rRNA, marker genes) within a community by sampling from a pool of nucleic acid fragments. A typical NGS 16SrRNA experiment consists of (i) sequencing 16S rRNA amplicons at a specified sequencing depth from environmental samples (Caporaso *et al*., 2011; Logares *et al*., 2020), human tissues/biofluids (Goodrich *et al*., 2016; Gomez *et al*., 2017; Voorhies *et al*., 2019) or built environments (Weiss *et al*., 2018; Checinska Sielaff *et al*., 2019); (ii) clustering highly similar sequences into operational taxonomic units (OTU) as a representative for closely related taxa (Sneath and Sokal, 1962; Schloss *et al*., 2009; Edgar, 2013) or using amplicon sequencing variants (ASV) (Callahan *et al*., 2016; Amir *et al*., 2017); (iii) generating abundance tables by counting NGS reads mapped to ecological units (e.g. OTU or ASV); and (iv) analysing the resulting abundance tables such as alpha/beta-diversity, differential abundance or network analysis.

Each of these subsampled fragments serve as a proxy for discrete biological components, and the percentage of each biological component is proportional to the true abundance in the sampled community. However, the measured abundance ultimately depends on the chemistry of the assay, not the input material (Quinn *et al*., 2018), and observations of these biological units are not entirely independent as the instruments can only measure reads up to the capacity of the machine. For example, if we imagine a sequencer as having a fixed number of slots, analogous to sequencing depth, then an increased abundance of one biological component may saturate the available slots leaving fewer slots for less abundant components and

© 2020 Society for Applied Microbiology and John Wiley & Sons Ltd.

potentially rendering low abundance components as undetected. NGS experiments often produce millions of reads, and after quality control followed by post-processing, the resulting product is an abundance table organized by quantification of biological components with respect to the total measured reads within a sample represents the sequencing depth.

Although observed abundances in macroscopic ecological context are typically independent events (Gloor *et al*., 2017), this is often not the case in microbial ecology as many organisms cannot be cultured in a controlled setting (Rinke *et al*., 2013) and can currently only be observed collectively through NGS technologies. As a result of this phenomena, NGS-derived datasets such as metagenomics and metatranscriptomics are inherently compositional. Compositional data are defined as a vector of strictly positive real numbers with an unknown or uninformative total (e.g. sequencing depth) as the abundance of each component represents parts of a whole and carries only relative information (Pawlowsky-Glahn *et al*., 2015). With this relativity, the information contained in the relationships between components is more essential than the information contained within an individual component (Rivera-Pinto *et al*., 2018). Exploration in this landscape requires vigilance and awareness of data characteristics, such as the lack of independence among compositional features, when applying statistical methods not designed for such assumptions as described in the literature (Gloor *et al*., 2017; Quinn *et al*., 2018).

The advantages of powerful analytical techniques such as machine-learning and network analysis on NGS datasets become increasingly attainable as sequencing costs continue dropping by several orders of magnitude. For instance, in 2001, the cost of sequencing the human genome was ~$100,000,000 compared to ~$1000 in 2019; exceeding far beyond the Moore's law predictions of ~$180,000 (DNA Sequencing Costs, 2020), machine-learning methods are routinely applied on two-dimensional data matrices represented by observations and features, where each feature represents an individual measurable property or characteristic of a phenomenon being observed (e.g. microbiome diversity) (Bishop, 2006). This terminology adapts to compositional datasets where an observation would represent a particular composition, a feature as an individual component and the numeric value as the count of the component within the composition. Extending the concepts into network theory, these discrete features can represent nodes within a network and continuous associations between features as weighted edges. In the context of a NGS-derived compositional dataset, a typical network instance could entail an individual sample as an observation/composition, OTU as a features/component and read counts where pairwise association between features would be the basis for the edge connectivity within the network. Later in this review, we elaborate on the methods behind generating such networks from compositional datasets.

With large datasets, researchers are not only investigating the abundance/depletion of features in relation to a specific condition, but also the (inferred) interactions between features. One way for such an investigation is by applying network theory. The versatility of graphical abstractions using nodes, edges and topological structure can be contextually applied to a wide array of problems. For instance, applications of network theory have been successful in several fields including studying plankton networks driving carbon export (Guidi *et al*., 2016), gene interactions related to weight physiology (Fuller *et al*., 2007), ecological shifts (Gomez *et al*., 2017) and metabolic potential (Espinoza *et al*., 2018) associated with carious lesions in children and regulatory metabolic interactions in marine diatoms (Levering *et al*., 2017) and bacterial soil communities (Mandakovic *et al*., 2018). Many biological networks are composed of molecules such as DNA, RNA, proteins and metabolites as the nodes, and edges between these nodes represent either curated or inferred interactions between them. Furthermore, advanced multi-omics approaches incorporating associations across modalities such as clinical tests, proteomics, amplicon, transcriptomics, cytokines, metabolomics and lipidomics have begun to pave the way towards precision health using systems biology (Schüssler-Fiorenza Rose *et al*., 2019; Zhou *et al*., 2019; Shomorony *et al*., 2020). There are several approaches for network analysis in systems biology that each have their advantages and caveats. The goal of this review is not to describe the landscape of network methods but to guide the reader through the process of implementing association networks from NGS-derived datasets which are inherently compositional.

Compositional data

Compositional data is defined as a *D*-part composition when all components are strictly positive real numbers that carry only relative information (Aitchison, 1982; Egozcue *et al*., 2003). The statistical methods for compositional data address the reality that these compositions do not exist in real Euclidean space but in a subset known as the simplex (Aitchison, 1982). Datasets generated from NGS technologies such as gene expression and 16/18S amplicons are compositional, sparse and have complex distributions such as negative binomial (Robinson *et al*., 2010; Love *et al*., 2014), zero-inflated gaussian (Paulson *et al*., 2013) and Dirichlet (Holmes *et al*., 2012; Chen and Li, 2013; Wadsworth *et al*., 2017;

© 2020 Society for Applied Microbiology and John Wiley & Sons Ltd.,

Harrison *et al*., 2019). Generating these datasets are often patient or sample limited making it difficult to produce a large number of sample observations without extensive resources and cooperation among collection agencies. This dilemma often subjects the dataset to 'the curse-of-dimensionality' in which the number of features vastly exceeds the number of observations potentially introducing statistical artefacts that can bias downstream analysis such as false positive correlations (Bellman, 2003). NGS-derived datasets are inherently incomplete as they are only parts of a complete system due to both biological and technical phenomena such as the capacity of a sequencing instrument to process reads (Gloor *et al*., 2017). This aspect presents a significant hurdle in analysis because a zero value may have different meanings in different datasets, and it is often difficult to distinguish the difference between true and false zeros (Kuhnert *et al*., 2005; Martin *et al*., 2005). Zeros can arise from many sources: (i) false zeros result from errors in experimental design or observational instrumentation; (ii) true zeros are either structural zeros and hypothesized in the statistical model or random zeros resulting from sampling variability (Parada *et al*., 2016; Blasco-Moreno *et al*., 2019). If not properly accounted for analytically, this missing information can introduce substantial artefacts in the downstream statistical analysis including comparing intra-sample patterns or association with other variables such as a phenotypic measurement. It is often useful to minimize excess sparsity to focus on core components within a system. The most common method for dealing with sparsity involves either filtering by prevalence, the addition of a minimal pseudocount or both respectively. However, incorporation of pseudocounts should be pursued with caution as haphazard usage can introduce statistical bias (Kumar *et al*., 2018). The handling of sparse data is an active area of research with interesting recent developments such as Robust Aitchison PCA (Martino *et al*., 2019). The balance between removing low prevalence features and retaining discriminative diversity depends on the research questions. For instance, if one is investigating community richness of soils in the context of potential natural products (Ling *et al*., 2015; Crits-Christoph *et al*., 2018), then it is reasonable to not remove any features assuming that the appropriate quality control and preprocessing were performed. In the scenario of inferred interactions in relation to a phenotype, it may be more beneficial to remove features with missing information to detail the relationships among core components. Most research questions will require a balance between the two extremes with some indication of when too many discriminative features are removed or when too many missing values have biased the data such as the notorious horseshoe effect (Diaconis *et al*., 2008; Morton *et al*., 2017).

The advantages and caveats of compositional data transformation

Relative abundance sensitivity and biased outcomes

Normalizations and transformations are standard approaches that are applied to compositional datasets when pursuing any type of weighted analysis beyond binary presence/absence and log-ratios. The most common normalization technique is total sum scaling (TSS), also referred to as relative abundance or closure (Aitchison, 1982), which divides each feature count by the sum of total counts in a sample. This technique removes technical bias that is related to differences in sequencing depth across samples. Despite the widespread adoption of this simple normalization, the abundances of specific components can drastically bias the results.

With TSS normalized data, the distance between variables is sensitive to the presence or absence of individual components and can reveal spurious relationships amongst unrelated variables resulting in false positive correlations (Pearson, 1896; Aitchison *et al*., 2000; Quinn *et al*., 2018). Consider the example illustrated in Fig. 1 involving a synthetic community of three OTUs with the following sample states: $sample_A$ (uniform abundances); $sample_B$ (doubling the abundances of OTU_1) and $sample_C$ (halving the abundances OTU_1). Notice the observed abundances of the community (Fig. 1A) and the TSS normalized abundances (Fig. 1B) show conflicting results when comparing between the samples. An increase in the abundance of OTU_1 within $sample_B$ introduces a false sense of depletion of OTU_2 and OTU_3, and the decrease in abundance in $sample_C$ suggests an enrichment of OTU_2 and OTU_3 when in reality their abundances did not change between samples. This artificial enrichment or depletion can lead to false positives in downstream analysis when investigating relationships between samples (e.g. network analysis, differential abundance, etc.). Using balance-trees is one method that bypasses the bias induced from TSS normalization as the ratios only reflect the values of the descendent nodes. A simplified example of balance trees is shown in Fig. 1C and D, where each internal node computes a ratio of the summed counts across both bifurcated paths. OTU_1 is a descendent of *y1* and not *y2*, and we illustrate that internal node *y1* is different between $sample_B$ and $sample_C$, whereas internal node *y2* is unchanged (Fig. 1C and D). This simplified concept of balance trees is vastly expanded in Aitchison geometry, in particular with isometric log ratio, and was first explored in geology (Egozcue and Pawlowsky-Glahn, 2005; Pawlowsky-Glahn and Egozcue, 2011). Since then, Aitchison geometry has recently been adopted by microbial-based datascience and is an effective alternative to address these

© 2020 Society for Applied Microbiology and John Wiley & Sons Ltd., *Environmental Microbiology*

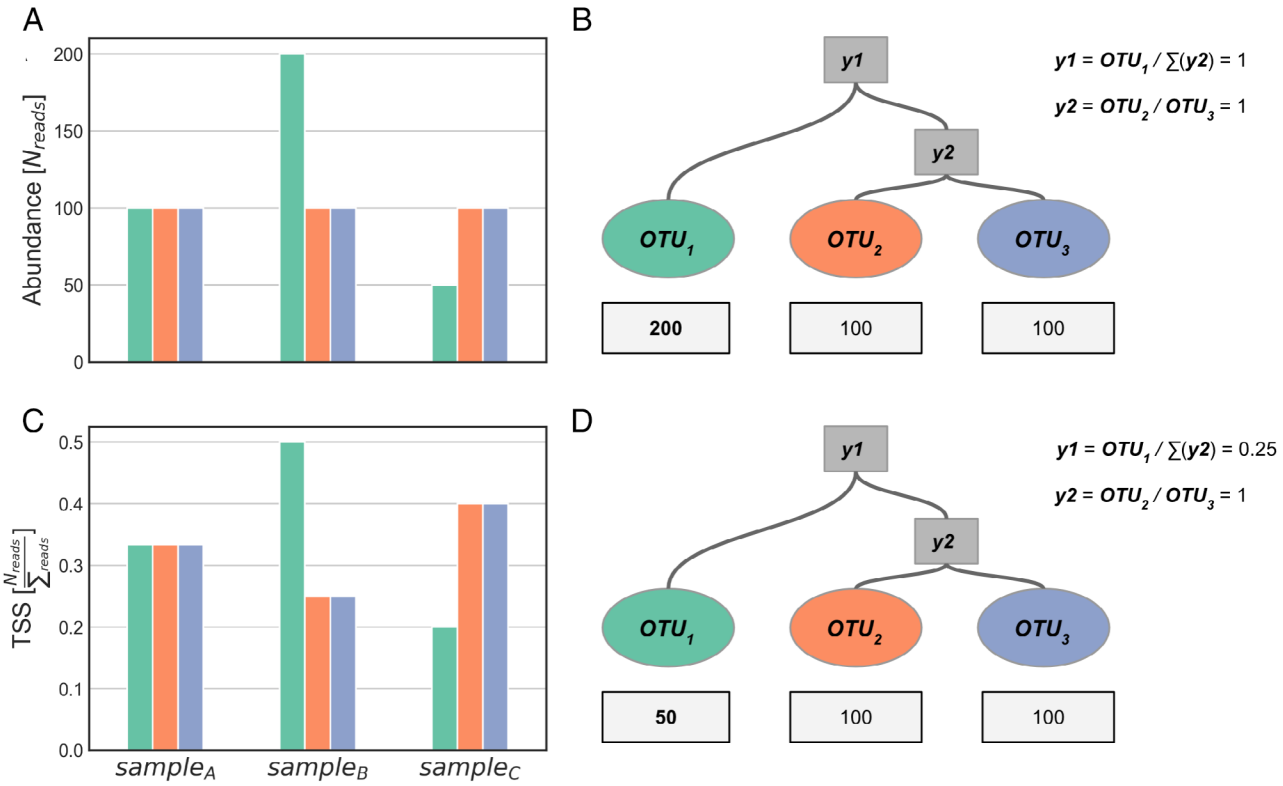

Fig 1. Comparison of balances and relative abundance with compositional data.A, B. Simple synthetic microbiome with three OTUs (OTU_{1-3}) and three samples ($sample_{A-C}$) represented as (A) absolute abundances and (B) relative abundances. C, D. Simplified examples of balance-trees where each internal node computes a ratio of the summed counts across both bifurcated paths for $sample_B$ and $sample_C$ respectively.

caveats (Silverman *et al*., 2017; Washburne *et al*., 2017; Rivera-Pinto *et al*., 2018; Morton *et al*., 2019).

Log-ratio transformations are a natural representation of compositional data

Log-ratio transformations perform equivalently on both the counts and proportions while capturing the relationships between features within the sample space (Pawlowsky-Glahn *et al*., 2015). The log-ratios are referred to as balances and are not sensitive to library-size or individual components. Aitchison geometry provides methods for analysing compositional datasets through log-ratios and isomorphisms that transform compositions from the Aitchison simplex to real space (Aitchison, 1982). One type of Aitchison geometry is the isometric log ratio (ILR) which operates on relative data in an unconstrained space with an orthogonal basis, thus, preserving all information in the original composition and is preferred when a non-singular covariance matrix is needed (Egozcue *et al*., 2003; Silverman *et al*., 2017). The ILR method uses a sequential binary partition to construct a new set of coordinates making it desirable in the field of microbiology where phylogenetic trees represent a natural coordinate system for vertical evolutionary relationships that distinguish taxonomy-derived components. As mentioned, ILR transformations have recently been repurposed by the microbiology community and have proven to evade many statistical artefacts introduced from an incorrectly represented sample space (Silverman *et al*., 2017; Washburne *et al*., 2017; Morton *et al*., 2017). The ILR transformation has been used to investigate taxonomic signatures in the human gut that are associated with obesity (Finucane *et al*., 2014), has been benchmarked for several supervised machine learning methods against popular normalization techniques (Knights *et al*., 2011; Silverman *et al*., 2017) and for linking external covariates to specific clades using regression methods (Washburne *et al*., 2017). ILR can also be used in a supervised setting for identifying taxa associated with a particular phenotype (i.e. feature selection) (Rivera-Pinto *et al*., 2018). However, one caveat of using balances is that the resulting data dimensionality is projected into a $D - 1$ dimensional space making interpretation difficult to directly associate specific components with particular trends. Additive log-ratio transformation (ALR) also projects the data into a $D - 1$ dimensional space and does not preserve distances as it is an isomorphism but not an isometry (Pawlowsky-Glahn *et al*., 2011). However, ALR is not as common in biological settings as it often requires a single unchanged referenced

© 2020 Society for Applied Microbiology and John Wiley & Sons Ltd., *Environmental Microbiology*

(Quinn *et al*., 2018), which is rarely available, and will not be described further in this review. Another type of Aitchison geometry is the centered log-ratio (CLR) transformation which preserves distances and is both an isomorphism and an isometry (Pawlowsky-Glahn *et al*., 2011). The CLR transformation is computed by taking the logarithm of each measurement and dividing by the geometric mean of the composition (i.e. arithmetic mean of logs) (Aitchison, 1982). An attractive feature of the CLR transformation is that the output retains the same dimensionality after transformation (Egozcue *et al*., 2003), which is not the case for ILR or ALR. This property allows for direct associations between a particular component and the transformed value without decomposing the balances amongst binary partitions as is required by ILR. CLR transformations have been applied to a wide range of biological topics including metagenomic binning using *k*-mer profiles (Laczny *et al*., 2015), the impact of gliadin in gluten-tolerant hosts (Zhang *et al*., 2017) and differential abundance (Fernandes *et al*., 2013; Mandal *et al*., 2015; Morton *et al*., 2019). However, it is important to note that the CLR transformation yields a coordinate system featuring a singular covariance matrix which may violate the assumptions of some statistical methods (Pawlowsky-Glahn *et al*., 2015).

CLR- and ILR-transformed data benefit from the following properties: (1) scale invariance, in that multiplying by a constant, such as library-size, will not influence the resulting transformation; (2) perturbation invariance, in that converting compositions between equivalent units does not affect the results; (3) permutation invariance, in that the order of components comprising the composition does not matter; and (4) sub-compositional dominance detailing that a subset of a complete composition contains less information than the whole composition (Quinn *et al*., 2018).

Recently, the CLR transformation has been modified by multi-additive log-ratio (MALR) transformations to use the geometric mean of a subset of components as a specialized reference (Quinn *et al*., 2018). *Fernandes et al*. introduced the interquartile log-ratio (IQLR) transformation which includes only components that exist within the interquartile range of total variance in the geometric mean calculation implemented in their *ALDEx2* package (Fernandes *et al*., 2013, 2014). Another variant is the robust centered log-ratio (RCLR) transformation which uses only the non-zero components introduced by *Martino et al*. in their *DEICODE* package (Martino *et al*., 2019). Every log-ratio transformation uses a unique reference and applications of such will be dependent on the hypothesis. For instance, if an analysis does not require analysis of individual components such as beta diversity analysis (e.g. clustering, ordination, manifold learning) then ILR may be the most effective. If the desired outcome is to analyse individual components, then perhaps CLR or a MALR variant may the preferred option over ILR which would require traversing the basis partitions. The selection of which method CLR-based method will again depend on the preferred interpretation. If the query dataset is relatively tame in terms of outliers, then CLR would be a practical option as the reference would be based on the sample's geometric mean. If outliers were an issue, then it may be more reasonable to use IQLR where the reference is based on the geometric mean of the components in the interquartile range. Understanding the concepts behind the reference sets will allow researchers to continue extending these methods to more specialized applications and expanding the paradigm of compositional data analysis.

Principal component-based correction

Compositional datasets in biology can often contain anomalies derived from latent technical or biological phenomena. When uncorrected, these confounding effects can lead to spurious associations in network-based analyses described in detail below (Parsana *et al*., 2019). Latent factor-based data correction has successfully been applied to address variation introduced from batch effects (Goh *et al*., 2017) and can be extended to other sources of variance. Recently, *Parsana et al*. have developed methods to regress out latent confounders captured within the top principal components (Parsana *et al*., 2019). Principal component-based correction can be described by the following: (1) compute $\boldsymbol{p}$ principal component loadings ($\boldsymbol{L}$) where $\boldsymbol{p}$ reflects the number of principal components used during singular value decomposition; (2) fitting a linear model on each feature $\boldsymbol{E_i = \mu_i + \beta_i \, x \, L_{1:p}}$ and (3) return the residuals. The assumptions of these methods posit that the network structure is scale-free (described below) and that the true topology is relatively sparse. The principal component-based correction implemented in *Parsana et al*. has been shown to reduce false positive associations when implemented as a preprocessing step before network construction using widely practiced methods such as weighted gene coexpression network analysis (WGCNA) (Langfelder and Horvath, 2008) and graphical lasso (Friedman *et al*., 2008) compared to instances with uncorrected input data. However, biological signals of interest may be removed by regressing out the top principal components in networks that violate the scale-free assumption. This technique could be adapted to incorporate non-Euclidean distances with alternative ordination methods and/or more complex manifold learning algorithms such *t*-Distributed Stochastic Neighbourhood Embeddings (t-SNE) (van der Maaten, 2013; Van Der Maaten, 2014) or Uniform Manifold Approximation and Projection (UMAP) (McInnes *et al*., 2018). Principal component-based correction, along with the proposed

© 2020 Society for Applied Microbiology and John Wiley & Sons Ltd., *Environmental Microbiology*

adaptations, have substantial potential in correcting for biological and technical variance. Although these methods have been in practice for at least a decade (Price *et al*., 2006), they have not been extensively benchmarked and explored on the wide array of modalities produced by NGS technologies and present a unique avenue in which to interrogate noisy datasets. Aside from principal component-based correction, noise reduction can be implemented in the experimental design with custom mixtures consisting of mock microbial communities (Bokulich *et al*., 2016; Parada *et al*., 2016) or spike-in internal controls such as External RNA Controls Consortium (ERCC) standards composed of synthetic RNA oligonucleotides spanning a range of nucleotide lengths and concentrations (Pine *et al*., 2016). However, these later methods must be embedded in experimental design and cannot be utilized *post hoc*.

Network theory, metrics and applications

Overview of network structure and terminology

A network is a graphical structure used to represent relations between discrete objects. It is a flexible abstract mathematical construct that can model systems with varying levels of complexity from simple binary networks to hierarchical networks. This versatility is alluring to researchers seeking to understand how discrete features are associated with each other, reflective of the inner mechanics of a system. However, the versatility of networks theory comes with a cost in that a network is highly sensitive to input data, thresholds, inference and transformations; therefore, implementation must be pursued strategically.

The discrete objects within a network (***G***) are referred to as nodes or vertices (***V***), and the connections between these nodes are referred to as edges (**E)**; formally, as ***G (V, E)*.** Edges are typically either weighted ($\boldsymbol{w} \in \mathbb{R}$) or unweighted ($\boldsymbol{w} \in \{0,1\}$), where the weights correspond with some numeric association value that represents the connection strength between the two nodes; although, many layout algorithms expect networks with positive real valued edge weights (Jacomy *et al*., 2014). This is formulated as ***(i, j, w)*** where ***i*** and ***j*** represent the source and target nodes respectively and ***w*** indicates edge weight. This representation of a static network can be extended to a dynamic network such as a temporal network, where edges are described by ***(i, j, w, t)*** as ***t*** represents time. A matrix consisting of these values is called an adjacency matrix (***A***) and is the core component of a network. ***A*** is often represented as a ($\boldsymbol{m} \times \boldsymbol{m}$) symmetric matrix where each $\boldsymbol{A_{ij}}$ represents a (weighted) connection defined by a real valued association function of two vectors of size ***n*** (***m*** = number of nodes and ***n*** = number of observations). Weighted networks are generally preferred over unweighted networks as they contain much more information than their binary counterpart. The weighted networks can either be symmetric or asymmetric, where the edges are undirected ($\boldsymbol{A_{ij}}$ equals $\boldsymbol{A_{ji}}$) or directed ($\boldsymbol{A_{ij}}$ may or may not equal $\boldsymbol{A_{ji}}$), respectively. Directed networks have been applied to modelling metabolic pathways, where nodes represent compounds and edges represent chemical reactions that transform metabolites into products (Levering *et al*., 2017). Undirected networks are more common when investigating compositional datasets and are covered in detail in the subsequent section of this review. Bayesian networks, a type of directed network, have been used to investigate longitudinal relationships within the infant gut microbiome (Mcgeachie *et al*., 2016) but take much longer to compute than undirected association networks and will not be further covered in this review. As mentioned, the flexibility of networks is essential for its powerful applications, whereas, often, interpretation is the limiting factor.

Association measures

Association networks are common among biological network analysis where each node typically represents a discrete feature, each edge represents an inferred interaction or association and the edge weight represents the strength of association between a pair of nodes. Association networks construction is highly modular and customizable from the selection of the association metric to edge detection (Fig. 2). Common association measures include correlation coefficient (Fuller *et al*., 2007), −log(*P*) (Shomorony *et al*., 2020), mutual information (Villaverde *et al*., 2014; Lachmann *et al*., 2016), Kullback–Leibler divergence (Lachmann *et al*., 2016) and proportionality (Lovell *et al*., 2015). The most common perhaps is the correlation coefficient which measures the relationship strength between a pair of nodes and exists within the interval [−1,1] where a value of 1 indicates an identical relationship among covariates. There are several types of correlation measures including Pearson, Spearman, Kendall rank and Biweight-midcorrelation. Pearson correlation measures linear relationships and is the most widely used correlation measure, albeit sensitive to outliers. Spearman correlation is a rank-based measure that is able to capture monotonic relationships, whereas Biweight-midcorrelation is a median-based correlation. Both Spearman and Biweight-midcorrelation tests are more robust to outliers than Pearson, whereas the latter is often more powerful (Hardin *et al*., 2007; Langfelder and Horvath, 2012; Song *et al*., 2012). Correlation coefficients as an edge association metric are desirable because they are easily calculated, are subject to several asymptotic statistical tests, scaled, and the sign of the measure can distinguish inverse

© 2020 Society for Applied Microbiology and John Wiley & Sons Ltd., *Environmental Microbiology*

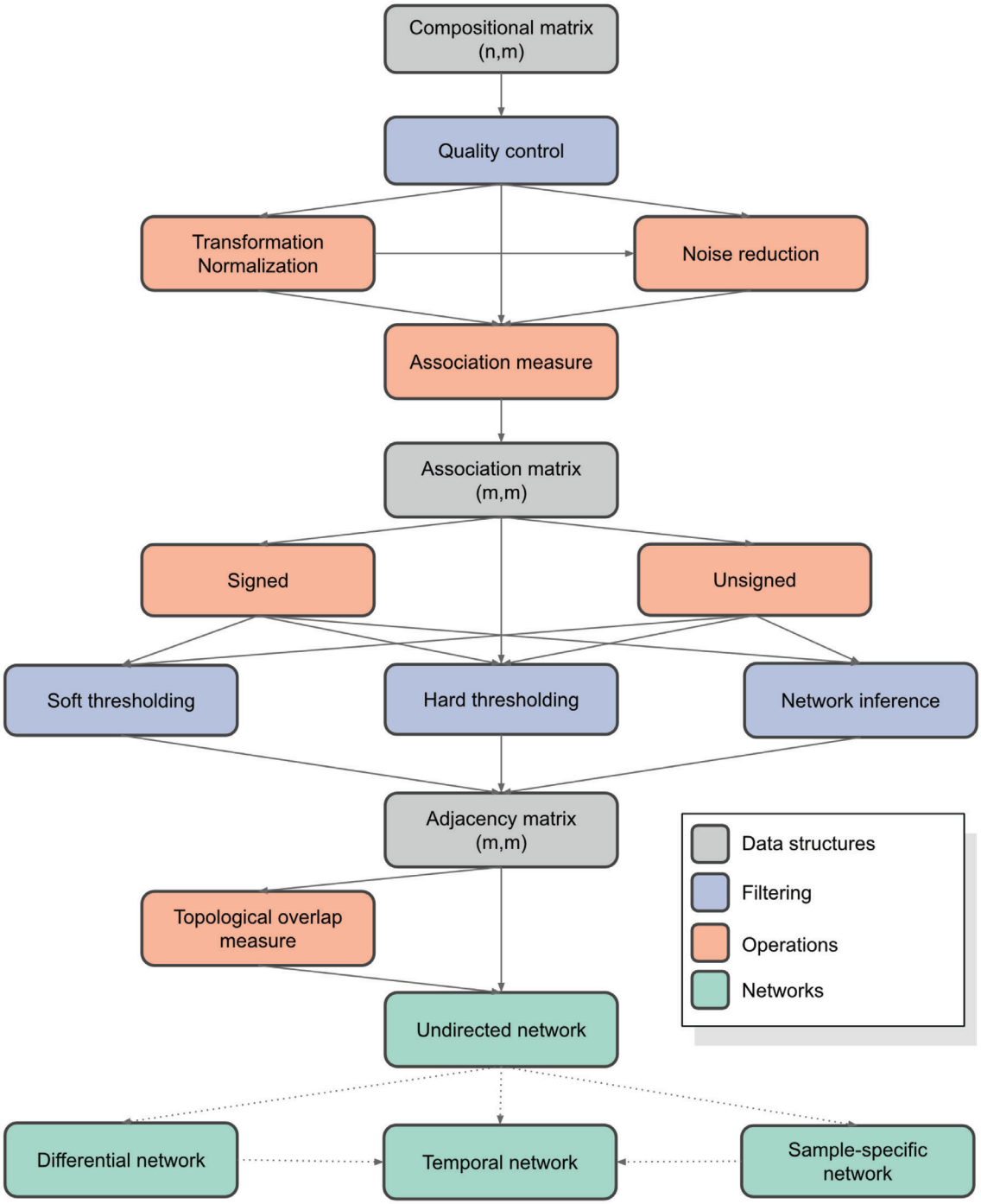

Fig 2. Schematic illustrating the modularity of network analysis.Modularity of network construction visualized by a network. The top of the visualization starts with compositional data as output from NGS methods. Traversing vertically goes further into the various operations detailed in the review to construct complex networks.

relationships (Song *et al*., 2012). However, correlation measures can be biased by compositionality (Friedman and Alm, 2012).

Many inference methods have been developed to leverage compositionality with sparse solutions to mitigate the effects of this bias including REBECCA (Ban *et al*., 2015), SparCC (Friedman and Alm, 2012), SPIEC-EASI (Kurtz *et al*., 2015) and CCLasso (Fang *et al*., 2015), whereas others rely on probabilistic graphical models (Tackmann *et al*., 2019) or permutation-based methods (Faust and

© 2020 Society for Applied Microbiology and John Wiley & Sons Ltd., *Environmental Microbiology*

Raes, 2016). Variance log-ratio (VLR) is another compositionally valid association metric that does not produce spurious results. VLR measures the concordance between two compositions (e.g. samples, observations or feature vectors) and computes the variance of the logarithm of one component as divided by a second component (Quinn *et al*., 2018). However, VLR has a substantial limitation in that it is unscaled with respect to the variances of the log components with the range [0, ∞), where zero indicates perfect coordination (Aitchison, 1982).

Proportionality is another compositionally valid association measure, implemented in the *propr* package, introduced by *Lovel et al*. and expanded upon by *Erb et al*. (Lovell *et al*., 2015; Erb and Notredame, 2016; Quinn *et al*., 2017). Proportionality can be thought of as a modified VLR that uses information about the variability of individual components to constrain VLR in a practical range. Proportionality also depends on the reference used for transformation and, unlike SparCC and SPIEC-EASI, does not assume the underlying structure of the associations are sparse (Quinn *et al*., 2018). There are typically three flavours of proportionality including ϕ, ϕ_s and ρ_p (Lovell *et al*., *2015*). The proportionality measures ϕ and ϕ_s both range between [0, ∞), similar to VLR, *as the asymmetric and symmetric versions respectively (*Quinn *et al*., 2017*). The proportionality measure* ρ_p *is the most akin to correlation as the pairwise application results in a symmetric matrix with values ranging from [−1,1] where a value of 1 indicates perfect proportionality amongst components (*Erb and Notredame, 2016*). A major advantage of proportionality measures is that they are robust when analysing relative data (*Lovell *et al*., 2015*) and tend not to produce spurious connections (*Quinn *et al*., 2017*); a stark contrast with Pearson's correlation coefficient which had considerable limitations when applied to compositional data. The properties of robustness to spurious results, scale invariance and interpretability positions proportionality as an effective association metric when inferring cooccurrence (*Bian *et al*., 2017*) and coexpression (*Lovell *et al*., 2015*) from NGS-derived datasets*.

Association measures often involve some type of data transformation in network analysis pipelines to achieve specific weight distributions. It is important to note that many algorithms require weights to be positive real numbers but some algorithms such as the Bellman–Ford shortest path algorithm can handle negative weights (Bang-Jensen and Gutin, 2009). Two common techniques to fulfil the weight assumptions are to investigate: (i) unsigned relationships where only the magnitude of the association is considered ($\boldsymbol{A_{ij}} = |\boldsymbol{\rho}|^{\boldsymbol{\beta}}$; or (ii) signed networks where the weights are forced into the interval [0,1] $\boldsymbol{A_{ij}} = (0.5 + 0.5\ \boldsymbol{\rho})^{\boldsymbol{\beta}}$; where $\boldsymbol{\rho}$ represents the association and $\boldsymbol{\beta}$ represents the soft thresholding power when $\boldsymbol{\beta} > 1$, or (iii) direction-specific relationships by analysing solely positive associations and masking negative associations (or the reverse). Each of these approaches produce unique network topologies and can be utilized to address different hypotheses depending on whether or not the type of interaction, either positive or negative, is relevant.

Determining network structure and spurious connections

Thresholding and inference are two techniques used to select which connections are represented in a network. Commonly used thresholding methods include: (i) hard thresholding and (ii) soft thresholding. Hard thresholding refers to a binary decision that defines two nodes to be connected (e.g. $\boldsymbol{w} \geq 0.7$). Such thresholds are often arbitrarily chosen without any statistical reasoning and could overlook potentially informative interactions which has been previously investigated (Connor *et al*., 2017). Another approach towards threshold selection is via permutation test to identify significant connections usually followed by multiple tests corrections such as false discovery rate (Unpingco, 2019; Pitman, 1937; Efron and Tibshirani, 1994). Soft thresholding refers to a method where the edge weights are shrunk towards zero and can be applied to networks with edge weights in the range of [0,1] by raising the weight to a power β emphasizing strong connections at the expense of weak connections (Langfelder and Horvath, 2008).

In most networks, the true topology of the network is not known *a priori* and must be inferred via computational methods. Therefore, the concept of false positive edges (i.e. spurious connections) based on true network topology is rather abstract and difficult to assess. In addition to this uncertainty, there exists the Simpson's paradox wherein associations can reverse or disappear when data sets are combined and analysed together (Kievit *et al*., 2013). In an effort to evaluate the presence of false positives, empirically derived functional pathways have been used as ground truth connections (true positives) and compared the associated gene sets using association-based network analysis with inferred structure (Parsana *et al*., 2019). In this paradigm, false positives were defined as edges that were observed between a pair of genes in the inferred network but absent in the list of curated connections. Spurious connections can often occur via the outlier effect (Heyer *et al*., 1999). For example, if the normalized abundance of two features are unrelated in all but a single observation, then the correlation coefficient may be much higher and can result in an inverse relationship (e.g. $\boldsymbol{\rho} = 0.87$ vs. −0.29) (Heyer *et al*., 1999). The topological overlap measure (TOM) is a powerful transformation for symmetric adjacency matrices, particularly for $\boldsymbol{w} \in [0,1]$, that considers pairs of nodes in relation to all nodes within the network instead of in isolation (Yip and Horvath, 2006; Ravasz

© 2020 Society for Applied Microbiology and John Wiley & Sons Ltd.,

et al., 2002). The TOM-based adjacency is particularly useful when the original adjacency matrix is sparse or susceptible to noise by replacing the isolated connections with weighted neighbourhood overlaps, thus, decreasing the effects of spurious or weak connections leading to more robust networks (Yip and Horvath, 2006; Dong and Horvath, 2007; Li and Horvath, 2007; Song *et al*., 2012). Association networks transformed via TOM often reduce or eliminate the number of false positive connections introduced by spurious correlations (Voigt and Almaas, 2019).

Some NGS-derived compositional datasets such as 16/18S amplicon sequencing are often sparse (Paulson *et al*., 2013; Kumar *et al*., 2018; Martino *et al*., 2019) depending on the diversity between samples (e.g. comparing multiple ecosystems or host body sites). Network analysis applied to sparse datasets are subject to false positives when using methods that are not designed for sparsity. *Sparse Correlations for Compositional data* (SparCC) is a technique for inferring correlations from compositional data often used for network analysis with the assumptions that the number of components is large and the true correlation network is sparse (Friedman and Alm, 2012). *Sparse Inverse Covariance Estimation for Ecological Association Inference* (SPIEC-EASI) is another statistical method for the inference of compositional networks, designed for ecologically derived datasets, seeking to address the 'curse-of-dimensionality' with a graphical model inference framework that relies on algorithms for sparse neighbourhood and inverse covariance selection (Kurtz *et al*., 2015). *Regularized estimation of the basis covariance based on compositional data* (REBECCA) identifies significant co-occurrence patterns by finding sparse solutions in a system with a deficient rank and estimating correlations between pairs of basis abundance using log-ratio transformation of counts (Ban *et al*., 2015). The network method used should be determined by the goal of the analysis. For instance, if the aim is to compare differential edges between a treatment system and a reference system, then it may be beneficial to use TOM for fully connected networks that are directly comparable. In contrast, if false positives are a critical concern or community detection is the aim, then it may be more appropriate to use one or a combination of the sparse methods described above.

Interpreting networks and evaluation metrics

Interpreting networks is often the limiting factor for applications beyond visualization. Fortunately, network theory offers several metrics that can be used to describe a particular network at varying levels of abstractions ranging from the network as a whole to objects including nodes and edges. One of the advantages of systems-wide analysis through networks is the ability identify and rank the most *important* nodes, or hubs, in a system (Layeghifard *et al*., 2017). Within a single organism gene expression network, the most important node could be a critical transcription factor, whereas in an environmental system, a hub node could be an organism essential for community stability. The most common metrics for static networks are degree, connectivity, flow and centrality. For simple networks (undirected and unweighted), the degree of a network is the number of connections a particular node contains. In an undirected weighted network, such as the association networks mentioned above, weighted-degree can be computed by summing the weighted connections for each node. The implementation of this weighted-degree is often referred to as connectivity such as in the intramodular connectivity calculations or advanced network visualization methods (Krzywinski *et al*., 2012). It should be noted that the aforementioned description of connectivity as analogous to weighted-degree is informal and the technical definition requires that node connectivity is equal to the minimum number of nodes that must be removed to disconnect the graph (Esfahanian, 2019). In a fully connected undirected network, weighted-degree is an extremely useful metric for measuring connectedness of a node within the network or a sub-network (e.g. intra-genus connectivity in microbial cooccurrence network). An unweighted-degree would yield a uniform distribution because each node is promiscuously connected to the rest in a fully connected network. In directed graphs, a node contains both an in- and out-degree which corresponds to connections into and outwards from a node, respectively. Extending the concepts of directed degree metrics, flow represents the difference between the out- and in-degree with positive and negative measures representing *sources* and *sinks* respectively (Krzywinski *et al*., 2012). In this paradigm, one may utilize Google's *PageRank* (Brin and Page, 1998; Page *et al*., 1999), a powerful link analysis algorithm, which is a variant of eigenvector centrality designed for quantifying the relative importance of a node within a network based on the directed flow of edges into and out of a node. Centrality comes in many different flavours such as eigenvector centrality (a measure of influence within a graph), closeness centrality (the average length of the shortest path between a node and all other nodes in the graph) and betweenness centrality (the frequency in which a node acts as a bridge along the shortest path between two other nodes). As with any analysis, the usefulness of each of these metrics is dependent on the research question, the structure of the graph and complexity of the edges. It is important to note that centrality measures have their caveats in that they underestimate

© 2020 Society for Applied Microbiology and John Wiley & Sons Ltd., *Environmental Microbiology*

the power of non-hub nodes due to heterogeneous topology of complex networks (Šikić *et al*., 2013; Layeghifard *et al*., 2017) and do not measure the difference between nodes (Bauer and Lizier, 2012).

Scale-free and heavy-tailed degree distribution topology

A network's organization is characterized by its structure; most notably, the distribution of node degrees. In a random network, the degree distribution is normally distributed. In complex systems, it is common for networks to self-organize into a scale-free state in that the probability $P(k)$ that a node in the network interacts with k other nodes decays as a power law following $P(k) \sim k^{\sim a}$ (Barabasi and Albert, 1999; Jeong *et al*., 2000; Barabási and Bonabeau, 2003). Scale-free networks are heterogenous, and their topology is dominated by a few highly connected nodes, referred to as hubs, which connect to the rest of the system (Zhang and Horvath, 2005). In network analysis, an examination of hubs typically represents influence within the system such as yeast protein–protein interaction networks and the relevance to proteins essential for survival (Jeong *et al*., 2001; Carter *et al*., 2004; Han *et al*., 2004). It should be noted that the prevalence of scale-free topologies in the natural world have been debated and suggested to be overestimated (Clauset *et al*., 2009; Mitchell, 2009; Broido and Clauset, 2019); therefore, assumptions on scale-free topologies should be properly assessed when exploring various preprocessing metrics, association measures and adjacency transformations.

Recent efforts have described the inconsistencies involving the applications of scale-free topologies with a meta-analysis using a large corpus of published networks, ranging from social networks to biological systems, by characterizing the extent of scale-free topologies including not scale-free, super-weak, weak, strong and strongest (Broido and Clauset, 2019). The findings from *Broido & Clauset et al*. revealed that scale-free structure is not universal, varies across domains and is often confounded as a generic stand-in for other heavy-tailed distributions such as log-normal. Regardless, compared to the entire *Index of Complex Networks* (ICON) corpus (N = 928 networks) with networks from biological, information, social, technological and transportation domains, biological networks were more likely to display the strongest level of direct evidence of scale-free structure.

Scale-free topology is well defined for simple networks (i.e. undirected, unweighted, and monoplex); although, the definition naturally generalizes to weighted networks where k takes on non-negative real numbers (Zhang and Horvath, 2005). In this generalization, scale-free topology can be approximated via the model fitting index R^2 of the linear regression modelled as $\log(p(k)) \sim \log(k)$ (Zhang and Horvath, 2005). Despite being better modelled using an exponentially truncated power law $p(k) \sim k^{-\gamma} \exp(-\alpha k)$ (Csanyi and Szendroi, 2003), *Zhang et al*. suggest the α and γ provide too much flexibility in curve fitting as R^2 is often more robust to adjacency parameters.

Advanced network analysis approaches

Differential networks

Comparing static networks, often referred to as cross-sectional networks, via differential network analysis (DiN [A]) is non-trivial and pertains to interrogating changes in feature interactions (i.e. edges) rather than the changes in the feature measurements (i.e. nodes). In particular, DiNA measures changes in network structure including topological restructuring and edge weights between different states (Lichtblau *et al*., 2016). DiNA is a fusion of two well-studied fields, namely differential abundance analysis and network theory. Differential abundance analysis has been routinely applied to RNA-seq (Robinson *et al*., 2010; Paulson *et al*., 2013; Love *et al*., 2014; Pimentel *et al*., 2017), whereas the latter, network theory, has been a studied for decades (Harary, 1969) with applications in biology to study the centrality of features in a disease network (Joy *et al*., 2005; Wang *et al*., 2011; Winter *et al*., 2012; Espinoza *et al*., 2018). There also exists compositionally aware differential abundance methods such as the *Analysis of Composition of Microbiomes* (ANCOM) which is done by calculating pairwise log ratios between all components and performing a significance test to determine if there is a significant difference in component ratios with respect to sample groupings of interest (Mandal *et al*., 2015); in addition to the aforementioned *ANOVA-Like Differential Expression* (ALDEx2) (Fernandes *et al*., 2013, 2014), the synergy of differential abundance analysis and biological networks obviates the limitations by considering multiple changes that are associated with differences between connectivity states instead of changes in singular features. DiNA algorithms typically compute network metrics for each network individually (e.g. *weighted-degree*) and interrogated via various statistical tests (Espinoza *et al*., 2018) or qualitatively using advanced visualization techniques such as hive plots (Fig. 3) (Krzywinski *et al*., 2012). The utility of DiNA has been validated through several diverse applications from identifying coexpressed genes related to obesity (Fuller *et al*., 2007), key transcriptional regulators associated with cancer that were undetected by expression levels (Carter *et al*., 2004; Lai *et al*., 2004; Choi *et al*., 2005) and regulatory mechanisms in yeast (Hsu *et al*., 2015).

© 2020 Society for Applied Microbiology and John Wiley & Sons Ltd.,

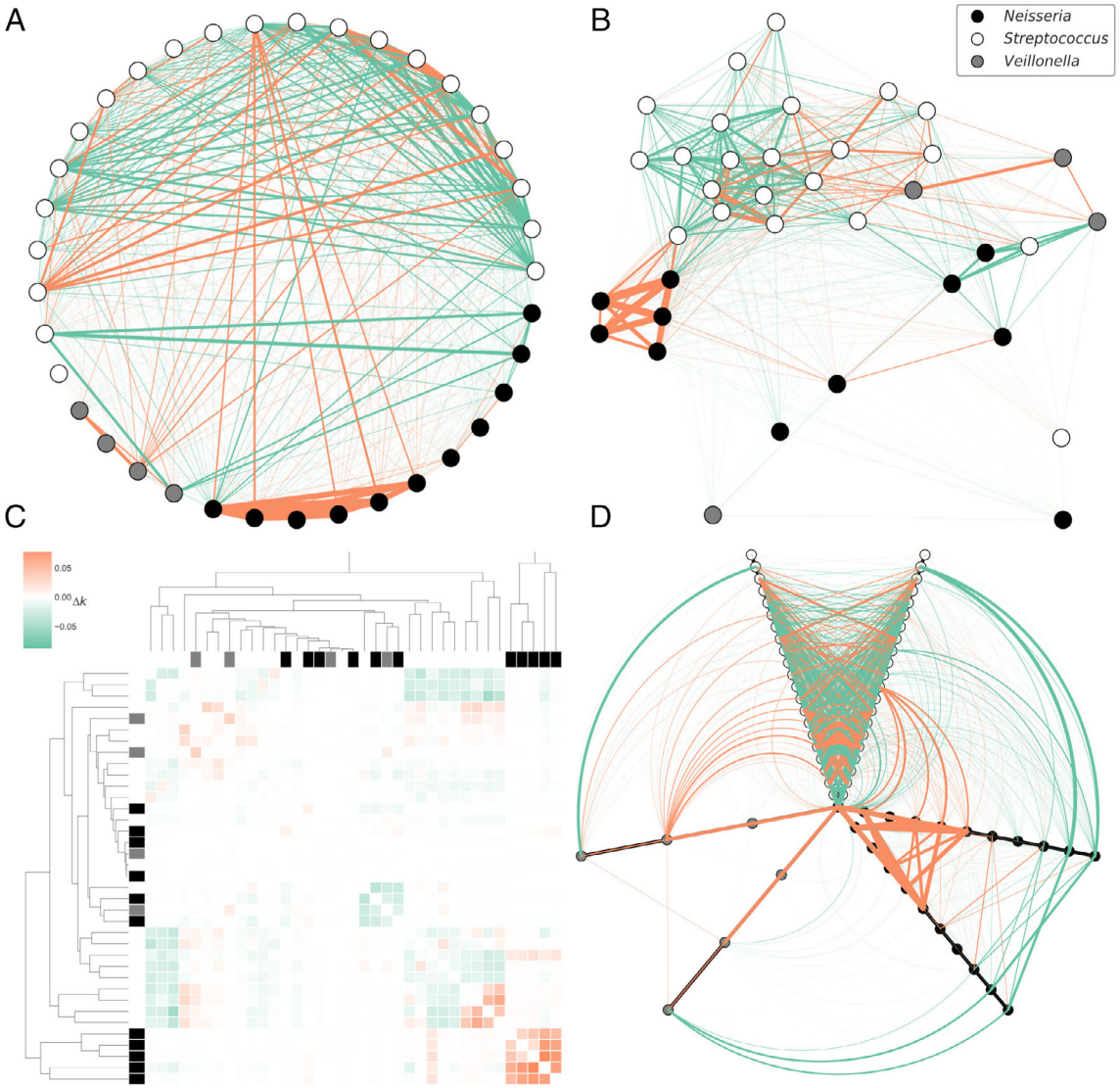

Fig 3. Visualization methods for differential co-occurrence networks.Supragingival plaque oral microbiome 16S community (Gomez *et al*., 2017). Turquoise and orange represent connections that are enriched in healthy and diseased cohorts respectively. Undirected weighted networks were constructed by the following pipeline: (1) removing OTUs that were not present in at least 25 individuals (5% of the cohort); (2) subsetting diseased and healthy cohorts; (3) ρ_p pairwise proportionality of each cohort (Erb and Notredame, 2016; Quinn *et al*., 2017, *compositional*, Espinoza, 2020); (4) signed transformation; (5) soft thresholding ($\beta = 12$); (6) topological overlap measure; (7) DiN = $TOM_{Diseased} - TOM_{Healthy}$; (8) subgraph for top three abundant genera and (9) visualization of (A, B) traditional networks (*NetworkX* (Hagberg *et al*., 2008), (C) clustermaps (*Seaborn*, Waskom *et al*., 2020) and (D) hive plots (*Soothsayer*, Espinoza, 2019).

Temporal networks

Understanding a system in a particular state is useful for some research questions such as investigating the differences between a disease and a non-diseased states (Gomez *et al*., 2017; Espinoza *et al*., 2018). However, it is difficult to address hypotheses about how a system changes over time using static networks, which are based on many assumptions including the following: (1) network topology is fixed; (2) processes of flow are at a steady state and (3) edges represent persistent interactions (Blonder *et al*., 2012). Such static networks can be further extended to study dynamically changing systems such as temporal networks. Network dynamics can be captured by studying the changes in the overall topology, node connectedness, node interactions (i.e. edges) and strength of the interactions (i.e. edge weights). Temporal networks are able to address such network dynamics by extending the edge domain across a time dimension instead of within the confines of a particular cross-sectional timepoint. Although concepts derived from static network theory apply to individual time states within temporal networks, extended theory must be utilized to study the dynamics between states across

© 2020 Society for Applied Microbiology and John Wiley & Sons Ltd., *Environmental Microbiology*

time. There exists many network measurements designed for temporal dynamics with the most notable including: *temporal centrality*, *closeness centrality, volatility* and *reachability latency* with many more described in (Thompson *et al*., 2017). Of the centrality based adaptive measures, *temporal centrality* measures the number of overall connections in a timepoint, whereas *closeness centrality* measures the time between specific connections. *Volatility* measures the rate of change in network states with respect to timepoint. Lastly, *reachability latency* is a useful index as it measures the time taken for all nodes to reach each other. Although temporal network theory offers an arsenal of metrics, usage should be crafted towards interpretation of a hypothesis on how states evolve over time. For example, the usage of *volatility* would have a much different interpretation when applied to a temporal resolution of milliseconds compared to staggered years or subjective timepoints (Thompson *et al*., 2017).

There are two main approaches in implementing temporal networks; namely (1) time-ordered where each edge is present only for a precise period and (2) time-aggregated networks over relatively short time frames (Blonder *et al*., 2012; Farine, 2018). Successful applications of temporal network theory have been demonstrated using functional magnetic resonance imaging data to explore dynamic properties of neural connectivity (Thompson *et al*., 2017), drought-responsive plant genes based on differential rhythmic expression profiles (Greenham *et al*., 2017), predicting parasite transmission spread in wild lemurs (Springer *et al*., 2017), recurrent patterns of microdiversity in a temperate coastal marine environments (Chafee *et al*., 2018) and longitudinal multi-omics to advance precision health (Schüssler-Fiorenza Rose *et al*., 2019; Zhou *et al*., 2019). When deciding whether to pursue temporal network analysis or adapt a static framework to incorporate temporal dynamics, it is vital to develop the framework around a particular research question such as available time resolution for samples, whether or not the time-ordered or time-aggregate would better model the hypothesis, or, most importantly, if a temporal component would yield any insight in the base hypothesis.

Sample-specific networks

Sample-specific networks (SSN) have been employed in the quest for personalized medicine to identify patient-specific biomarkers and changes in connectivity that can describe or predict health status (Liu *et al*., 2016; Yu *et al*., 2017; Kuijjer *et al*., 2019). Although previous research has focused on location-specific (Lima-Mendez *et al*., 2015) or host-specific (Ofaim *et al*., 2017) networks, SSNs have shown to be reliable for accurately characterizing a specific disease state within an individual (Liu *et al*., 2016); although, these methods could seamlessly be adapted to investigate environmental systems such as the Tara oceans (Karsenti *et al*., 2011) or Global ocean sampling expedition (Venter *et al*., 2004; Rusch *et al*., 2007) datasets. In a clinical setting, patient-specific diseases generally result from the dysfunction of the underlying system rather than individual molecules (Barabási *et al*., 2011; Hood and Flores, 2012). The assumptions of SSNs posit that if a single sample can significantly alter the association of two features, then the query sample is considered to be inconsistent with the interactions in the reference network. *Liu et al*. developed a statistical method to construct SSNs based on statistical perturbation analysis of a single sample against a group of control samples validated with nine different cancer types from *The Cancer Genome Atlas* dataset (Weinstein *et al*., 2013). In short, this perturbation method computes the pairwise Pearson correlation of a background network compared to the same network with the sample of interest added. *Liu et al*. discovered network patterns unique to specific types of cancer; personalized features revealed important regulatory patterns of driver genes, strong associations between SSNs and somatic mutations and the ability to predict driver genes from edges involving hub genes. The SSN method employed in *Liu et al*. was shown to be less sensitive to spurious associations than standard Pearson correlation-based networks through the incorporation of a reference network for comparisons. SSNs are in their infancy; but, their potential is unprecedented especially when studying how a particular sample state perturbs a reference state. Furthermore, SSNs also present an opportunity to incorporate sample-specific phenotypic data such as nutrient measurements in ecological samples or clinical measurements in medical samples.

Visualizing network complexity

.Network visualizations are infamously difficult to interpret. Visualization is a qualitative assessment of network structure that can guide quantitative assessments downstream. The complexity of a network scales exponentially which can make visualization of large networks difficult. For instance, a fully connected undirected network (e.g. ρ_P proportionality network with soft thresholding) contains $\boldsymbol{(N_n^2 - N_n)/2}$ non-redundant edges ($\mathbf{N_e}$) where $\mathbf{N_n}$ represents the number of nodes. This exponential scale makes visualizing a fully connected undirected network with $N_n = 100$ nodes ($N_e = 4950$) much easier than a network with $N_n = 1000$ nodes ($N_e = 499{,}500$). To demonstrate the interpretability of network visualization methods, we construct a cooccurrence DiN using the top three abundant genera (*Streptococcus*, *Veillonella* and *Neisseria*) from a supragingival plaque oral microbiome dataset (Gomez *et al*., 2017) consisting of subjects with ($N_{diseased} = 126$) and without ($N_{healthy} = 337$) dental caries (Fig. 3). A traditional network visualization would

© 2020 Society for Applied Microbiology and John Wiley & Sons Ltd.,

be the aforementioned 'ball-and-stick' method with efficacy governed by the layout algorithm for positioning nodes in relation to each other in a plot; usually with respect to edge weight. A common network is the circular layout using custom node ordering which is feasible for simple networks but decreases in interpretability with increasing network complexity (Fig. 3A). Networks visualized with layout algorithms that incorporate edge weights are more useful for qualitative visual clustering of nodes such as the *ForceAtlas2* (Jacomy *et al*., 2014), *Fruchterman-Reingold force-directed algorithm* (Fruchterman and Reingold, 1991) or *graphviz* algorithms (Ellson *et al*., 2001) (Fig. 3B). However, this method of visualization is often only interpretable for fully connected networks with low complexity or scale-free networks and can quickly turn into the notorious 'hairball' network visualization. Despite the archetype of visualizing networks with layout-based methods, they are often difficult to interpret because their creation is often driven by an aesthetic heuristic which influences how the topology is rendered (Krzywinski *et al*., 2012). More complex networks benefit more from alternative visualization approaches such as clustermaps and hive plots. Clustermaps are a unique combination of agglomerative hierarchical clustering and matrix heatmaps leveraging unsupervised relationships between nodes and the visualization of edge weights as values within the heatmap (Fig. 3C). Clustermaps are beneficial for visualizing low and medium complexity networks but can be uninterpretable, difficult to render and computationally expensive to compute as the number of nodes increases. Hive plots are a network visualization method that are applicable at all levels of complexity (Krzywinski *et al*., 2012). The hive plot places nodes on a radially oriented linear axis with nodes positioned along the axis either by structural properties or user-defined selections. Hive plots contain three flexible components including: (1) the assignment of node coordinates to and within an axis; (2) the layout profile of each axes (position, scale and angle) and (3) the aesthetics of (weighted) edges visualized as curves between nodes for intra/inter-axes connectivity. The hive plot most effectively illustrates enriched intra-genus connectivity within *Neisseria* in the diseased cohort and phenotype-specific inter-genus connectivity profiles between specific OTUs (Fig. 3D). Hive plots are powerful for networks but are difficult to adapt for temporal networks. Arguably, the most intuitive temporal network is the slice plot implemented through the *teneto* Python package (Thompson *et al*., 2017). A slice plot essentially decomposes a hive into a collection of arc plots, a linear segment with nodes positioned along the line and arcs showing connections, juxtaposed consecutively to visualize connections that remain consistent or, alternatively, inconsistent with respect to time. In networks with few connections (e.g. $N_n \leq 100$), the labels can be useful in discerning connections, but larger networks are often assessed globally without visualizing the label of nodes directly. Network visualization methods are still evolving and can benefit from the insight of researchers with unique hypotheses and computational abilities.

Summary and outlook

NGS-derived datasets are compositional and should be considered as compositions at all stages of analysis (Gloor *et al*., 2017). We have overviewed the characteristics of compositional data, the bias that occur when compositional datasets are analysed inappropriately and transformation techniques that mitigate the bias such as log-ratio transformations. In addition, for compositional data analysis, we detailed the advantages and caveats of various methods to construct association networks. We further reviewed the analytical metrics for quantifying different aspects of network topology and application of advanced network analysis to model more complex systems.

Despite the progressive techniques recently developed to interpret biological systems, the nascent field of systems biology is far from the status of omniscient. Not knowing the true topology of a system *a priori* inherently limits our approaches towards fully understanding a system's natural complexity. Furthermore, biological systems are not static and modelling the transition between states will yield more intuitive insights on the schematics of these complex structures. The aphorism that 'all models are wrong, but some are useful' (Box *et al*., 2009) holds truth in the paradigm of inference-based systems biology where knowing the true network structure of an abstract space *a priori* is not attainable. Biological systems are complex because they are abstract constructs used to model an observed phenomenon. This complexity is the aftermath of the uncertainty of true associations, the sensitivity of the methods to infer associations, unaccounted variance (e.g. unknown phenotype) and the dynamics of how these abstractions, such edge weights and node inclusion, evolve over time. The abstract space defined by a network is the source of its versatility while also representing the crux of germane interpretation.

The advanced network approaches described in this review, combinations thereof, or even *networks-of-networks* (Gao *et al*., 2014) can potentially be utilized to address humanity's most pressing issues. For instance, consider the topic of drug discovery in the scope of antibiotic-resistance. Imagine one has identified a novel chemical entity (NCE) from a soil microbiome, referred to as NCE_X, that appears to exhibit a unique mechanism of action against a particular pathogenic organism (*P. organism*). One may ask how the query NCE perturbs the baseline state ($\boldsymbol{t_0}$) of *P. organism* with dose ***d*** at different time intervals? This question could be addressed by creating a differential network at each timepoint **n** ($DiNET_n$) as $\boldsymbol{t_0}$ vs. $\boldsymbol{t_n}$ where each static network is derived via coexpression-based topological overlap

© 2020 Society for Applied Microbiology and John Wiley & Sons Ltd., *Environmental Microbiology*

measures. Of greater insight, albeit greater complexity, one may ask how the perturbations of NCE_x-challenged *P. organism* at dose ***d*** compares to a negative control, challenged with a solvent such as water, over a time interval from 0 min to 1 h? The symbiosis of temporal, sample-specific and differential correlation networks could be used to investigate this question which, naturally, could guide the experimental design for the over-arching project. SSNs have already been harnessed for personalized medicine and could be further augmented by incorporating not only temporal dimensionality but multimodality. For more grandiose applications, imagine the synergy of explainable artificial intelligence (Gunning, 2017), system-wide cellular modelling (Ebrahim *et al*., 2013) and 'network-of-networks' (multi-level network) frameworks (Gao *et al*., 2014) harnessed by domain experts spanning climate science to microbiology, public health to agriculture and from economics to politics modelling the complex flux of resources; an interdisciplinary effort to usurp climate change by identifying solution states that are not only environmentally sustainable but economically productive.

The future of systems biology must be approached from creative vantage points by building combinatorically on the cornerstones of established concepts, understanding the assumptions of various statistical methods and interpreting these mathematical abstractions in the context of insightful biological questions where domain knowledge is of utmost importance. The synergy of domain expertise, advanced analytical methods and creative minds is the foundation of cutting-edge science. Modelling complex systems has provided insight in the past and will certainly continue to do so in the future with the evolution of network theory, and the inventiveness catalysed by the human mind and machines to decipher latent patterns embedded within natural and abstract systems.

Conflict of interest

None.

Acknowledgements

This work was supported by NSF grant OCE-1558453 (to CLD), NASA NNA15BB034A (to CLD) and P01AI118687 (to KEN).

References

Aitchison, J. (1982) The statistical analysis of compositional data. *J R Stat Soc Ser B* **44**: 139–160.

Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J.A., and Pawlowsky-Glahn, V. (2000) Logratio analysis and compositional distance. *Math Geol* **32**: 271–275.

Amir, A., McDonald, D., Navas-Molina, J.A., Kopylova, E., Morton, J.T., Zech Xu, Z., *et al*. (2017) Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems* **2**: e00191-16. https://msystems.asm.org/content/2/2/e00191-16.

Ban, Y., An, L., and Jiang, H. (2015) Investigating microbial co-occurrence patterns based on metagenomic compositional data. *Bioinformatics* **31**: 3322–3329.

Bang-Jensen, J., and Gutin, G. (2009) *Digraphs: Theory, Algorithms, and Applications*, London, England: Springer-Verlag.

Barabasi, A.-L., and Albert, R. (1999) Emergence of scaling in random networks. *Science* **286**: 509–512.

Barabási, A.-L., and Bonabeau, E. (2003) Scale-free networks. *Sci Am* **288**: 60–69.

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011) Network medicine: a network-based approach to human disease. *Nat Rev Genet* **12**: 56–68.

Bauer, F., and Lizier, J.T. (2012) Identifying influential spreaders and efficiently estimating infection numbers in epidemic models: a walk counting approach. *EPL Europhys Lett* **99**: 68007.

Bellman, R. (2003) *Dynamic Programming*, United States: Dover Publications.

Bian, G., Gloor, G.B., Gong, A., Jia, C., Zhang, W., Hu, J., Zhang, H., Zhang, Y., Zhou, Z., Zhang, J., Burton, J.P., Reid, G., Xiao, Y., Zeng, Q., Yang, K., and Li, J. (2017). The gut microbiota of healthy aged chinese is similar to that of the healthy young. *mSphere* **2**. http://dx.doi.org/10.1128/msphere.00327-17.

Bishop, C.M. (2006) *Pattern Recognition and Machine Learning*, New York, NY: Springer-Verlag.

Blasco-Moreno, A., Pérez-Casany, M., Puig, P., Morante, M., and Castells, E. (2019) What does a zero mean? Understanding false, random and structural zeros in ecology. *Methods Ecol Evol* **10**: 949–959.

Blonder, B., Wey, T.W., Dornhaus, A., James, R., and Sih, A. (2012) Temporal dynamics and network analysis. *Methods Ecol Evol* **3**: 958–972.

Bokulich, N.A., Rideout, J.R., Mercurio, W.G., Shiffer, A., Wolfe, B., Maurice, C.F., *et al*. (2016) Mockrobiota: a public resource for microbiome bioinformatics benchmarking. *mSystems* **1**: e00062-16.

Box, G.E.P., Luceño, A., and Paniagua-Quiñones, M.D.C. (2009) *Statistical Control by Monitoring and Adjustment*. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Brin, S. and Page, L. (1998). *The Anatomy of a Large-Scale Hypertextual Web Search Engine* (107–117). Amsterdam, Netherlands: WWW7: Proceedings of the seventh international conference on World Wide Web 7.

Broido, A.D., and Clauset, A. (2019) Scale-free networks are rare. *Nat Commun* **10**: 1017.

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., and Holmes, S.P. (2016) DADA2: high-resolution sample inference from Illumina amplicon data. *Nat Methods* **13**: 581–583.

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

© 2020 Society for Applied Microbiology and John Wiley & Sons Ltd., *Environmental Microbiology*

Carter, S.L., Brechbuhler, C.M., Griffin, M., and Bond, A.T. (2004) Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics* **20**: 2242–2250.

Chafee, M., Fernàndez-Guerra, A., Buttigieg, P.L., Gerdts, G., Eren, A.M., Teeling, H., and Amann, R.I. (2018) Recurrent patterns of microdiversity in a temperate coastal marine environment. *ISME J* **12**: 237–252.

Checinska Sielaff, A., Urbaniak, C., Mohan, G.B.M., Stepanov, V.G., Tran, Q., Wood, J.M., *et al*. (2019) Characterization of the total and viable bacterial and fungal communities associated with the international Space Station surfaces. *Microbiome* **7**: 50–50.

Chen, J., and Li, H. (2013) Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *Ann Appl Stat* **7**: 418–442.

Choi, J.K., Yu, U., Yoo, O.J., and Kim, S. (2005) Differential coexpression analysis using microarray data and its application to human cancer. *Bioinformatics* **21**: 4348–4355.

Clauset, A., Shalizi, C.R., and Newman, M.E.J. (2009) Power-law distributions in empirical data. *SIAM Rev* **51**: 661–703.

Connor, N., Barberán, A., and Clauset, A. (2017) Using null models to infer microbial co-occurrence networks. *PLoS One* **12**: e0176751.

Crits-Christoph, A., Diamond, S., Butterfield, C.N., Thomas, B.C., and Banfield, J.F. (2018) Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis. *Nature* **558**: 440–444.

Csányi, G., and Szendrői, B. (2004). Structure of a large social network. *Physical Review E* **69**. http://dx.doi.org/10.1103/physreve.69.036131.

Diaconis, P., Goel, S., and Holmes, S. (2008) Horseshoes in multidimensional scaling and local kernel methods. *Ann Appl Stat* **2**: 777–807.

Dong, J., and Horvath, S. (2007) Understanding network concepts in modules. *BMC Syst Biol* **1**: 24.

Ebrahim, A., Lerman, J.A., Palsson, B.O., and Hyduke, D.R. (2013) COBRApy: COnstraints-based reconstruction and analysis for python. *BMC Syst Biol* **7**: 74.

Edgar, R.C. (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Methods* **10**: 996–998.

Efron, B., and Tibshirani, R. (1994) *An Introduction to the Bootstrap*, London, England: Chapman & Hall.

Egozcue, J.J., and Pawlowsky-Glahn, V. (2005) Groups of parts and their balances in compositional data analysis. *Math Geol* **37**: 795–828.

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barceló-Vidal, C. (2003) Isometric logratio transformations for compositional data analysis. *Math Geol* **35**: 279–300.

Ellson, J., Ellson, J., Gansner, E., Koutsofios, L., North, S., Woodhull, G., *et al*. (2001) Graphviz — open source graph drawing tools. *Lect NOTES Comput Sci* **2265**: 483–484.

Erb, I., and Notredame, C. (2016) How should we measure proportionality on relative gene expression data? *Theory Biosci* **135**: 21–36.

Esfahanian, A.-H. (2019). *Connectivity Algorithms*, Michigan, MI: Michigan State University. http://www.cse.msu.edu/~cse835/Papers/Graph_connectivity_revised.pdf.

Espinoza, J.L. (2019) *soothsayer: High-level package for (Bio-)informatics*. GitHub. https://github.com/jolespin/soothsayer.

Espinoza, J.L., Harkins, D.M., Torralba, M., Gomez, A., Highlander, S.K., Jones, M.B., *et al*. (2018) Supragingival plaque microbiome ecology and functional potential in the context of health and disease. *MBio* **9**: e01631-18.

Fang, H., Huang, C., Zhao, H., and Deng, M. (2015) CCLasso: correlation inference for compositional data through Lasso. *Bioinformatics* **31**: 3172–3180.

Farine, D.R. (2018) When to choose dynamic vs. static social network analysis. *J Anim Ecol* **87**: 128–138.

Faust, K., and Raes, J. (2016) CoNet app: inference of biological association networks using Cytoscape. *F1000Research* **5**: 1519.

Fernandes, A.D., Macklaim, J.M., Linn, T.G., Reid, G., and Gloor, G.B. (2013) ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-Seq. *PLoS One* **8**: e67019.

Fernandes, A.D., Reid, J.N.S., Macklaim, J.M., McMurrough, T.A., Edgell, D.R., and Gloor, G.B. (2014) Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome* **2**: 15.

Finucane, M.M., Sharpton, T.J., Laurent, T.J., and Pollard, K.S. (2014) A taxonomic signature of obesity in the microbiome? Getting to the guts of the matter. *PLoS One* **9**: e84689.

Friedman, J., and Alm, E.J. (2012) Inferring correlation networks from genomic survey data. *PLoS Comput Biol* **8**: e1002687.

Friedman, J., Hastie, T., and Tibshirani, R. (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**: 432–441.

Fruchterman, T.M.J., and Reingold, E.M. (1991) Graph drawing by force-directed placement. *Softw Pract Exp* **21**: 1129–1164.

Fuller, T.F., Ghazalpour, A., Aten, J.E., Drake, T.A., Lusis, A.J., and Horvath, S. (2007) Weighted gene coexpression network analysis strategies applied to mouse weight. *Mamm Genome* **18**: 463–472.

Gao, J., Li, D., and Havlin, S. (2014) From a single network to a network of networks. *Natl Sci Rev* **1**: 346–356.

Gloor, G.B., Macklaim, J.M., Pawlowsky-Glahn, V., and Egozcue, J.J. (2017) Microbiome datasets are compositional: and this is not optional. *Front Microbiol* **8**: 2224.

Goh, W.W.B., Wang, W., and Wong, L. (2017) Why batch effects matter in omics data, and how to avoid them. *Trends Biotechnol* **35**: 498–507.

Gomez, A., Espinoza, J.L., Harkins, D.M., Leong, P., Saffery, R., Bockmann, M., *et al*. (2017) Host genetic control of the oral microbiome in health and disease. *Cell Host Microbe* **22**: 269-278.e3.

Goodrich, J.K., Davenport, E.R., Beaumont, M., Jackson, M. A., Knight, R., Ober, C., *et al*. (2016) Genetic determinants of the gut microbiome in UK twins. *Cell Host Microbe* **19**: 731–743.

Greenham, K., Guadagno, C.R., Gehan, M.A., Mockler, T.C., Weinig, C., Ewers, B.E., and McClung, C.R. (2017). Temporal network analysis identifies early physiological and

© 2020 Society for Applied Microbiology and John Wiley & Sons Ltd., *Environmental Microbiology*

transcriptomic indicators of mild drought in Brassica rapa. *eLife* **6**. http://dx.doi.org/10.7554/elife.29655.

Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlimi, A., Roux, S., *et al*. (2016) Plankton networks driving carbon export in the oligotrophic ocean. *Nature* **532**: 465–470.

Gunning, D. (2017) *Explainable Artificial Intelligence (XAI)*, Arlington County, VA: DARPA. https://www.darpa.mil/attachments/XAIProgramUpdate.pdf.

Hagberg, A.A., Schult, D.A., and Swart, P.J. (2008) *Exploring Network Structure, Dynamics, and Function using NetworkX*, Pasadena, CA: Proceedings of the 7th Python in Science Conference. https://www.osti.gov/biblio/960616-exploring-network-structure-dynamics-function-using-networkx.

Han, J.-D.J., Bertin, N., Hao, T., Goldberg, D.S., Berriz, G.F., Zhang, L.V., *et al*. (2004) Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature* **430**: 88–93.

Harary, F. (1969) *Graph Theory*, New York, NY: Avalon Publishing.

Hardin, J., Mitani, A., Hicks, L., and VanKoten, B. (2007) A robust measure of correlation between two genes on a microarray. *BMC Bioinform* **8**: 220.

Harrison, J.G., Calder, W.J., Shastry, V., and Buerkle, C.A. (2019) Dirichlet-multinomial modelling outperforms alternatives for analysis of microbiome and other ecological count data. *Mol Ecol Resour* **20**: 481–497.

Heyer, L.J., Kruglyak, S., and Yooseph, S. (1999) Exploring expression data: identification and analysis of coexpressed genes. *Genome Res* **9**: 1106–1115.

Holmes, I., Harris, K., and Quince, C. (2012) Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLoS One* **7**: e30126.

Hood, L., and Flores, M. (2012) A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory. *N Biotechnol* **29**: 613–624.

Hsu, C.-L., Juan, H.-F., and Huang, H.-C. (2015) Functional analysis and characterization of differential coexpression networks. *Sci Rep* **5**: 13295.

Jacomy, M., Venturini, T., Heymann, S., and Bastian, M. (2014) ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS One* **9**: e98679.

Jeong, H., Mason, S.P., Barabási, A.-L., and Oltvai, Z.N. (2001) Lethality and centrality in protein networks. *Nature* **411**: 41–42.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., and Barabási, A.-L. (2000) The large-scale organization of metabolic networks. *Nature* **407**: 651–654.

Joy, M.P., Brock, A., Ingber, D.E., and Huang, S. (2005) High-betweenness proteins in the yeast protein interaction network. *J Biomed Biotechnol* **2005**: 96–103.

Karsenti, E., Acinas, S.G., Bork, P., Bowler, C., De Vargas, C., Raes, J., *et al*. (2011) A holistic approach to marine eco-systems biology. *PLoS Biol* **9**: e1001177.

Kievit, R.A., Frankenhuis, W.E., Waldorp, L.J., and Borsboom, D. (2013) Simpson's paradox in psychological science: a practical guide. *Front Psychol* **4**: 513.

Knights, D., Costello, E.K., and Knight, R. (2011) Supervised classification of human microbiota. *FEMS Microbiol Rev* **35**: 343–359.

Krzywinski, M., Birol, I., Jones, S.J., and Marra, M.A. (2012) Hive plots–rational approach to visualizing networks. *Brief Bioinform* **13**: 627–644.

Kuhnert, P.M., Martin, T.G., Mengersen, K., and Possingham, H.P. (2005) Assessing the impacts of grazing levels on bird density in woodland habitat: a Bayesian approach using expert opinion. *Environ* **16**: 717–747.

Kuijjer, M.L., Tung, M.G., Yuan, G., Quackenbush, J., and Glass, K. (2019) Estimating sample-specific regulatory networks. *iScience* **14**: 226–240.

Kumar, M.S., Slud, E.V., Okrah, K., Hicks, S.C., Hannenhalli, S., and Corrada Bravo, H. (2018) Analysis and correction of compositional bias in sparse sequencing count data. *BMC Genom* **19**: 799.

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

Lachmann, A., Giorgi, F.M., Lopez, G., and Califano, A. (2016) ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics* **32**: 2233–2235.

Laczny, C.C., Sternal, T., Plugaru, V., Gawron, P., Atashpendar, A., Margossian, H., *et al*. (2015) VizBin - an application for reference-independent visualization and human-augmented binning of metagenomic data. *Microbiome* **3**: 1.

Lai, Y., Wu, B., Chen, L., and Zhao, H. (2004) A statistical method for identifying differential gene-gene co-expression patterns. *Bioinformatics* **20**: 3146–3155.

Langfelder, P., and Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform* **9**: 559.

Langfelder, P., and Horvath, S. (2012) Fast R functions for robust correlations and hierarchical clustering. *Journal of statistical software* **46**.

Layeghifard, M., Hwang, D.M., and Guttman, D.S. (2017) Disentangling interactions in the microbiome: a network perspective. *Trends Microbiol* **25**: 217–228.

Levering, J., Dupont, C.L., Allen, A.E., Palsson, B.O., and Zengler, K. (2017) Integrated regulatory and metabolic networks of the marine diatom *Phaeodactylum tricornutum* predict the response to rising CO2 levels. *mSystems* **2**: e00142-16.

Li, A., and Horvath, S. (2007) Network neighborhood analysis with the multi-node topological overlap measure. *Bioinformatics* **23**: 222–231.

Lichtblau, Y., Zimmermann, K., Haldemann, B., Lenze, D., Hummel, M., and Leser, U. (2016) Comparative assessment of differential network analysis methods. *Brief Bioinform* **18**: bbw061.

Lima-Mendez, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F., *et al*. (2015) Determinants of community structure in the global plankton interactome. *Science* **348**: 1262073.

Ling, L.L., Schneider, T., Peoples, A.J., Spoering, A.L., Engels, I., Conlon, B.P., *et al*. (2015) A new antibiotic kills pathogens without detectable resistance. *Nature* **517**: 455–459.

Liu, X., Wang, Y., Ji, H., Aihara, K., and Chen, L. (2016) Personalized characterization of diseases using sample-specific networks. *Nucleic Acids Res* **44**: e164.

© 2020 Society for Applied Microbiology and John Wiley & Sons Ltd., *Environmental Microbiology*

Logares, R., Deutschmann, I.M., Junger, P.C., Giner, C.R., Krabberød, A.K., Schmidt, T.S.B., *et al*. (2020) Disentangling the mechanisms shaping the surface ocean microbiota. *Microbiome* **8**: 55.

Love, M.I., Huber, W., and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550.

Lovell, D., Pawlowsky-Glahn, V., Egozcue, J.J., Marguerat, S., and Bähler, J. (2015) Proportionality: a valid alternative to correlation for relative data. *PLoS Comput Biol* **11**: e1004075.

Mandakovic, D., Rojas, C., Maldonado, J., Latorre, M., Travisany, D., Delage, E., *et al*. (2018) Structure and co-occurrence patterns in microbial communities under acute environmental stress reveal ecological factors fostering resilience. *Sci Rep* **8**: 1–12.

Mandal, S., Van Treuren, W., White, R.A., Eggesbø, M., Knight, R., and Peddada, S.D. (2015) Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial Ecology in Health & Disease* **26**: 27663. https://doi,org/10.3402/mehd.v26.27663.

Martin, T.G., Wintle, B.A., Rhodes, J.R., Kuhnert, P.M., Field, S.A., Low-Choy, S.J., *et al*. (2005) Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. *Ecol Lett* **8**: 1235–1246.

Martino, C., Morton, J.T., Marotz, C.A., Thompson, L.R., Tripathi, A., Knight, R., and Zengler, K. (2019) A novel sparse compositional technique reveals microbial perturbations. *mSystems* **4**: e00016-19.

Mcgeachie, M.J., Sordillo, J.E., Gibson, T., Weinstock, G.M., Liu, Y.Y., Gold, D.R., *et al*. (2016) Longitudinal prediction of the infant gut microbiome with dynamic Bayesian networks. *Sci Rep* **6**: 1–11.

McInnes, L., Healy, J., and Melville, J. (2018) *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, Ithaca, NY: arXiv. http://arxiv.org/abs/1802.03426.

Mitchell, M. (2009) *Complexity: A Guided Tour*. New York, NY, USA: Oxford University Press, Inc.

Morton, J.T., Marotz, C., Washburne, A., Silverman, J., Zaramela, L.S., Edlund, A., *et al*. (2019) Establishing microbial composition measurement standards with reference frames. *Nat Commun* **10**: 1–11.

Morton, J.T., Sanders, J., Quinn, R.A., McDonald, D., Gonzalez, A., Vázquez-Baeza, Y., *et al*. (2017) Balance trees reveal microbial niche differentiation. *mSystems* **2**: e00162-16.

Morton, J.T., Toran, L., Edlund, A., Metcalf, J.L., Lauber, C., and Knight, R. (2017) Uncovering the horseshoe effect in microbial analyses. *mSystems* **2**: e00166-16.

Ofaim, S., Ofek-Lalzar, M., Sela, N., Jinag, J., Kashi, Y., Minz, D., and Freilich, S. (2017) Analysis of microbial functions in the rhizosphere using a metabolic-network based framework for metagenomics interpretation. *Front Microbiol* **8**: 1606.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999) *The PageRank Citation Ranking: Bringing Order to the Web*, Stanford, CA: Stanford. https://www.semanticscholar.org/paper/The-PageRank-Citation-Ranking%3A-Bringing-Order-to-Page-Brin/eb82d3035849cd23578096462ba419b53198a556.

Parada, A.E., Needham, D.M., and Fuhrman, J.A. (2016) Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environ Microbiol* **18**: 1403–1414.

Parsana, P., Ruberman, C., Jaffe, A.E., Schatz, M.C., Battle, A., and Leek, J.T. (2019) Addressing confounding artifacts in reconstruction of gene co-expression networks. *Genome Biol* **20**: 94.

Paulson, J.N., Stine, O.C., Bravo, H.C., and Pop, M. (2013) Differential abundance analysis for microbial marker-gene surveys. *Nat Methods* **10**: 1200–1202.

Pawlowsky-Glahn, V., and Egozcue, J.J. (2011) Exploring compositional data with the CoDa-Dendrogram. *Austrian J Stat* **40**: 103–113.

Pawlowsky-Glahn, V., Egozcue, J.J., and Tolosana-Delgado, R. (2011) *Lecture Notes on Compositional Data Analysis*, Spain: University of Girona. http://www.compositionaldata.com/material/others/Lecture_notes_11.pdf.

Pawlowsky-Glahn, V., Egozcue, J., and Tolosana-Delgado, R. (2015) *Modeling and Analysis of Compositional Data*, New York, NY: John Wiley & Sons. https://doi.org/10.1002/9781119003144.

Pearson, K. (1896). Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. *Philos Trans R Soc London Ser A, Contain Pap a Math or Phys Character* **187**: 253–318.

Pimentel, H., Bray, N.L., Puente, S., Melsted, P., and Pachter, L. (2017) Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat Methods* **14**: 687–690.

Pine, P.S., Munro, S.A., Parsons, J.R., McDaniel, J., Lucas, A.B., Lozach, J., *et al*. (2016) Evaluation of the external RNA controls consortium (ERCC) reference material using a modified Latin square design. *BMC Biotechnol* **16**: 54.

Pitman, E.J.G. (1937) Significance tests which may be applied to samples from any populations. *Suppl to J R Stat Soc* **4**: 119.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**: 904–909.

Quinn, T.P., Crowley, T.M., and Richardson, M.F. (2018). Benchmarking differential expression analysis tools for RNA-Seq: normalization-based vs. log-ratio transformation-based methods. *BMC Bioinformatics* **19**. http://dx.doi.org/10.1186/s12859-018-2261-8.

Quinn, T.P., Erb, I., Richardson, M.F., and Crowley, T.M. (2018) Understanding sequencing data as compositions: an outlook and review. *Bioinformatics* **34**: 2870–2878.

Quinn, T.P., Richardson, M.F., Lovell, D., and Crowley, T.M. (2017) Propr: an R-package for identifying proportionally abundant features using compositional data analysis. *Sci Rep* **7**: 1–9.

Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., and Barabási, A.-L. (2002) Hierarchical organization of modularity in metabolic networks. *Science* **297**: 1551–1555.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.-F., *et al*. (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**: 431–437.

© 2020 Society for Applied Microbiology and John Wiley & Sons Ltd., *Environmental Microbiology*

Rivera-Pinto, J., Egozcue, J.J., Pawlowsky-Glahn, V., Paredes, R., Noguera-Julian, M., and Calle, M. L. (2018). Balances: a new perspective for microbiome analysis. *mSystems* **3**. http://dx.doi.org/10.1128/msystems.00053-18.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010) edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139–140.

Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S., *et al*. (2007) The sorcerer II Global Ocean sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLoS Biol* **5**: 0398–0431.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., *et al*. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**: 7537–7541.

Schüssler-Fiorenza Rose, S.M., Contrepois, K., Moneghetti, K.J., Zhou, W., Mishra, T., Mataraso, S., *et al*. (2019) A longitudinal big data approach for precision health. *Nat Med* **25**: 792–804.

Shomorony, I., Cirulli, E.T., Huang, L., Napier, L.A., Heister, R.R., Hicks, M., *et al*. (2020) An unsupervised learning approach to identify novel signatures of health and disease from multimodal data. *Genome Med* **12**: 7.

Šikić, M., Lančić, A., Antulov-Fantulin, N., and Štefančić, H. (2013) Epidemic centrality — is there an underestimated epidemic impact of network peripheral nodes? *Eur Phys J B* **86**: 1–13.

Silverman, J.D., Washburne, A.D., Mukherjee, S., and David, L.A. (2017). A phylogenetic transform enhances analysis of compositional microbiota data. *eLife* **6**. http://dx.doi.org/10.7554/elife.21887.

Sneath, P.H.A., and Sokal, R.R. (1962) Numerical taxonomy. *Nature* **193**: 855–860.

Song, L., Langfelder, P., and Horvath, S. (2012) Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinform* **13**: 328.

Springer, A., Kappeler, P.M., and Nunn, C.L. (2017) Dynamic vs. static social networks in models of parasite transmission: predicting Cryptosporidium spread in wild lemurs. *J Anim Ecol* **86**: 419–433.

Tackmann, J., Matias Rodrigues, J.F., and von Mering, C. (2019) Rapid inference of direct interactions in large-scale ecological networks from heterogeneous microbial sequencing data. *Cell Syst* **9**: 286-296.e8.

Thompson, W.H., Brantefors, P., and Fransson, P. (2017) From static to temporal network theory: applications to functional brain connectivity. *Netw Neurosci* **1**: 69–99.

Unpingco, J. (2019). *Python for Probability, Statistics, and Machine Learning*, 2, Switzerland: Springer International Publishing. https://www.springer.com/gp/book/9783030185442.

van der Maaten, L.J.P. (2013) *Barnes-Hut-SNE*, (pp. 1–11). Ithaca, New York, NY: arXiv. http://arxiv.org/abs/1301.3342.

Van Der Maaten, L. (2014) Accelerating t-SNE using tree-based algorithms. *J Mach Learn Res* **15**: 1–21.

Venter, C., Remington, K., Heidelberg, J., Halpern, A., Rusch, D., Eisen, J., *et al*. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**: 66–74.

Villaverde, A.F., Ross, J., Morán, F., and Banga, J.R. (2014) MIDER: network inference with mutual information distance and entropy reduction. *PLoS One* **9**: e96732.

Voigt, A., and Almaas, E. (2019) Assessment of weighted topological overlap (wTO) to improve fidelity of gene co-expression networks. *BMC Bioinform* **20**: 58.

Voorhies, A.A., Mark Ott, C., Mehta, S., Pierson, D.L., Crucian, B.E., Feiveson, A., Oubre, C.M., Torralba, M., Moncera, K., Zhang, Y., Zurek, E., and Lorenzi, H.A. (2019). Study of the impact of long-duration space missions at the International Space Station on the astronaut microbiome. *Scientific Reports* **9**. http://dx.doi.org/10.1038/s41598-019-46303-8.

Wadsworth, W.D., Argiento, R., Guindani, M., Galloway-Pena, J., Shelbourne, S.A., and Vannucci, M. (2017) An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. *BMC Bioinform* **18**: 94.

Wang, H., Li, M., Wang, J., and Pan, Y. (2011) *A New Method for Identifying Essential Proteins Based on Edge Clustering Coefficient* (pp. 87–98). Berlin, Heidelberg: Springer.

Washburne, A.D., Silverman, J.D., Leff, J.W., Bennett, D.J., Darcy, J.L., Mukherjee, S., *et al*. (2017) Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. *PeerJ* **5**: e2969.

Waskom, M., Botvinnik, O., Ostblom, J., Lukauskas, S., Hobson, P., M. Gelbart, *et al*. (2020) *seaborn*.GitHub. https://github.com/mwaskom/seaborn.

Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., *et al*. (2013) The cancer genome atlas Pan-cancer analysis project. *Nat Genet* **45**: 1113–1120.

Weiss, H., Hertzberg, V.S., Dupont, C., Espinoza, J.L., Levy, S., Nelson, K., and Norris, S. (2019). The airplane cabin microbiome. *Microbial Ecology* **77**: 87–95. http://dx.doi.org/10.1007/s00248-018-1191-3

Wetterstrand, K.A.(2020). *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)*, Maryland: National Human Genome Research Institute. https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data.

Winter, C., Kristiansen, G., Kersting, S., Roy, J., Aust, D., Knösel, T., *et al*. (2012) Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes. *PLoS Comput Biol* **8**: e1002511.

Yip, A.M., and Horvath, S. (2006). *The Generalized Topological Overlap Matrix For Detecting Modules in Gene Networks*, Las Vegas, NV: Proceedings of the 2006 International Conference on Bioinformatics & Computational Biology. https://www.semanticscholar.org/paper/The-Generalized-Topological-Overlap-Matrix-for-in-Yip-Horvath/6bff956be4ecdd433974821930b454a68fab1fac.

Yu, X., Zhang, J., Sun, S., Zhou, X., Zeng, T., and Chen, L. (2017) Individual-specific edge-network analysis for disease prediction. *Nucleic Acids Res* **45**: e170.

Zhang, L., Andersen, D., Roager, H.M., Bahl, M.I., Hansen, C.H.F., Danneskiold-Samsøe, N.B., *et al*. (2017) Effects of gliadin consumption on the intestinal microbiota

© 2020 Society for Applied Microbiology and John Wiley & Sons Ltd., *Environmental Microbiology*

and metabolic homeostasis in mice fed a high-fat diet. *Sci Rep* **7**: 44613.

Zhang, B., and Horvath, S. (2005). A General Framework for Weighted Gene Co-Expression Network Analysis. *Statistical Applications in Genetics and Molecular Biology* **4**. http://dx.doi.org/10.2202/1544-6115.1128.

Zhou, W., Sailani, M.R., Contrepois, K., Zhou, Y., Ahadi, S., Leopold, S.R., *et al*. (2019) Longitudinal multi-omics of host–microbe dynamics in prediabetes. *Nature* **569**: 663–671.

Espinoza, J.L. (2020). *Compositional: Compositional data analysis in Python*, GitHub. https://github.com/jolespin/compositional.

© 2020 Society for Applied Microbiology and John Wiley & Sons Ltd., *Environmental Microbiology*