OPEN

The ISME Journal (2017) 11, 791–807
© 2017 International Society for Microbial Ecology All rights reserved 1751-7362/17
www.nature.com/ismej

ORIGINAL ARTICLE

A family of interaction-adjusted indices of community similarity

Thomas Sebastian Benedikt Schmidt[1], João Frederico Matias Rodrigues and Christian von Mering
Institute of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zürich, Switzerland

Interactions between taxa are essential drivers of ecological community structure and dynamics, but they are not taken into account by traditional indices of β diversity. In this study, we propose a novel family of indices that quantify community similarity in the context of taxa interaction networks. Using publicly available datasets, we assessed the performance of two specific indices that are *Taxa INteraction-Adjusted* (TINA, based on taxa co-occurrence networks), and *Phylogenetic INteraction-Adjusted* (PINA, based on phylogenetic similarities). TINA and PINA outperformed traditional indices when partitioning human-associated microbial communities according to habitat, even for extremely downsampled datasets, and when organising ocean micro-eukaryotic plankton diversity according to geographical and physicochemical gradients. We argue that interaction-adjusted indices capture novel aspects of diversity outside the scope of traditional approaches, highlighting the biological significance of ecological association networks in the interpretation of community similarity.

The ISME Journal (2017) **11,** 791–807; doi:10.1038/ismej.2016.139; published online 9 December 2016

Introduction

Understanding how patterns of diversity are established and maintained is fundamental to the ecological characterisation of living systems. Following Whittaker (1972, 1960), *diversity* is traditionally considered to comprise of three components: local diversity of individual habitats (*α diversity*) and between-site variation (*β diversity*) together determine the total diversity of a given system (*γ diversity*). However, while referring to these definitions, researchers have studied conceptually different phenomena under the umbrella term ‘diversity’. β diversity, in particular, has been variously reported as species turnover or variation, further subdefined and quantified using different mathematical approaches (Anderson *et al*., 2010; Tuomisto, 2010). Nevertheless, most authors agree that *community similarity*, the compositional variation between sites, is an integral aspect of β diversity, and more generally one of the most important parameters in community ecology (Vellend, 2010). To characterise the mechanisms underlying an observed diversity structure, it is essential to quantify and appraise patterns of community similarity.

A multitude of mathematical indices of community similarity have been proposed: as of 2016, the widely used software *EstimateS* (Colwell, 2013) computes 12 different indices, while the popular microbial ecology toolboxes *mothur* (Schloss *et al*., 2009) and *phyloseq* (McMurdie and Holmes, 2013) provide as many as 37 and 46 measures, respectively. The various available measures capture conceptually different aspects of diversity. Traditional measures, such as the Jaccard (1901) or Bray–Curtis (1957) indices, focus on taxa compositional overlap, quantified directly from taxa count data. More recently, phylogenetically informed indices have become increasingly popular, which, in contrast to census-based metrics, do not treat taxa independently but rather quantify shared evolutionary history between communities (Graham and Fine, 2008; Swenson, 2011). Traditional and phylogenetic metrics may provide complementary insights into the processes driving community composition, particularly since phylogenetic relatedness of taxa is considered a proxy for functional or ecological similarity (Webb *et al*., 2002).

Apart from analysing diversity patterns, another important approach to characterising ecosystem function focuses on studying interaction networks of ecological or functional associations between taxa directly. Applying graph theory to food webs, mutualist or host-parasite networks and others has revealed an important role for interaction structure in community stability and dynamics (Polis and Strong, 1996; Proulx *et al*., 2005; Ings *et al*., 2009). This approach has been particularly fruitful in

Correspondence: TSB Schmidt or C von Mering, Institute of Molecular Life Sciences, University of Zürich, Winterthurerstrasse 190, Zürich, CH-8057, Switzerland.
E-mail: sebastian.schmidt@embl.de or mering@imls.uzh.ch

[1]Current address: European Molecular Biology Laboratory, Meyerhofstr. 1, D-69117, Heidelberg, Germany.

Received 31 March 2016; revised 16 June 2016; accepted 30 August 2016; published online 9 December 2016

microbial ecology, where ‘true’ ecological interactions can to a certain extent be inferred from co-occurrence networks of anonymous *Operational Taxonomic Units* (OTUs; Faust and Raes, 2012; Berry and Widder, 2014). Highly informative taxa co-occurrence networks have been constructed for many ecological systems, including the human body-associated microbiota (Faust *et al*., 2012) or ocean planktonic communities (Lima-Mendez *et al*., 2015; Sunagawa *et al*., 2015), as well as for global, integrated datasets across various habitats (Chaffron *et al*., 2010).

One main difference between such diversity-based and network-based approaches lies in analysis scope: the latter identify drivers of community structure at the level of individual taxa interactions, while the former reveal compositional patterns at community level. Arguably, both approaches are informative, but they are often pursued independently: it remains challenging to interpret community-level diversity changes in light of taxa-level ecological associations, and vice versa. In this study, we propose to bridge this analysis gap with a set of mathematical indices that quantify community similarity (or β diversity) as the average taxa interaction strength between samples. While our method is applicable to many types of interaction data, we focus on *Taxa INteraction-Adjusted* indices (TINA), based on taxa co-occurrence data, and *Phylogenetic INteraction-Adjusted* indices (PINA), based on phylogenetic similarities. In a re-analysis of two publicly available datasets, we show that TINA and PINA capture known diversity patterns better than existing indices, even for very small datasets, and that they can reveal novel and refined biological interpretations.

Materials and methods

In this study, we compared a total of 11 indices of community similarity (listed in Table 1) that fall into three categories: ‘traditional’ taxa count-based indices, phylogenetic indices and our proposed interaction-adjusted indices (Figure 1).

Classical and phylogenetic community similarity indices

In his widely cited ‘*comparative study of the floral distribution of parts of the Alps and the Jura*’, Paul Jaccard (1901) introduced what is arguably the earliest index of β diversity. For two communites A and B, the *classical Jaccard index* (JCI) is the relative taxa overlap, that is the ratio of shared taxa among all sampled taxa (see formula in Table 1). In this original fomulation, the JCI is *incidence-based*, or *unweighted*: it considers only the presence and absence of taxa, but not their relative abundance ratios. Several *abundance-based* or *weighted* variations of the classical Jaccard index have been proposed (Chao *et al*., 2004); here, we use a straightforward *weighted Jaccard* (JCW) formulation that describes community similarity as the mean fraction of individuals in shared taxa across both focal samples. While the JCI has proved very versatile and is used for manifold scientific problems beyond biology, an ecology-specific variant that corrects for the characteristics of imperfect sampling has been proposed by Chao *et al* (2004): *Chao’s weighted Jaccard* index (JCC) extrapolates the fractions of individuals in unseen shared taxa based on the number of observed shared rare taxa.

One of the most widely used indices in modern community ecology is arguably the *Bray-Curtis similarity* (BC; Bray and Curtis, 1957), which describes community overlap as the fractional minimum abundance of shared taxa between samples. Somewhat related to BC is the *Morisita-Horn overlap index* (MH; Horn, 1966), calculated as pairwise multiplicative taxa overlap, adjusted by a per-sample concentration index.

These classical indices and their derivations assess community overlap directly from count data, treating all observed taxa as independent (Figure 1, top branch). Phylogenetic indices, in contrast, consider the (phylogenetic) relationships between taxa and quantify community similarity as shared evolutionary history (Figure 1, middle). Among these increasingly popular indices is *UniFrac,* which calculates the shared branch length between samples on a phylogenetic tree, either for observed taxa based on incidence (unweighted UniFrac, UFU; Lozupone and Knight, 2005), or based on taxa abundances (weighted UniFrac, UFW; Lozupone *et al*., 2007).

A novel family of interaction-adjusted community similarity indices

Consider two communities A and B, composed of N_A and N_B taxa from which a total of n_A and n_B individuals have been sampled. Next, consider a matrix I that describes pairwise taxa interactions, such that I_{ij} is the interaction between taxa i and j. Manifold types of interactions with different biological meanings, different layers of information and at different levels of curation effort are suitable, such as for example predator–prey relationships, symbiosis, parasitism, mutualism, cross-feeding, resource competition and so on. Here, we consider the case of ecological associations as inferred by taxa co-occurrence networks, constructed from taxa count tables by pairwise association of abundances across samples (Figure 1, bottom). The scale and characteristics of such a co-occurrence interaction matrix I_C will depend on the association metric chosen; Faust and Raes (2012) have provided a comprehensive review of different approaches to network construction and interpretation. For example, a taxa abundance correlation network would scale from −1 (avoidance) to +1 (complete association), while other popular association metrics may scale differently.

Table 1 Overview of different indices of community similarity used in this study

Index	*Abb*	*Formula*	*Description*	*Reference*
Jaccard, classical	JCI	$\frac{\lvert A\cap B\rvert}{\lvert A\cup B\rvert}$	Ratio of shared taxa among all observed taxa (*relative overlap*)	Jaccard, 1901
Jaccard, weighted	JCW	$0.5(\sum_{i\in A\cap B}\frac{n_{Ai}}{n_A}+\sum_{j\in A\cap B}\frac{n_{Bj}}{n_B})$	Average fraction of individuals in shared taxa	This study
Jaccard, Chao	JCC	$\frac{\hat{U}\hat{V}}{\hat{U}+\hat{V}-\hat{U}\hat{V}}$	Abundance-based Jaccard index, corrected for unseen shared taxa based on shared rare taxa	Chao *et al.*, 2004
Bray–Curtis	BC	$\frac{2\sum_{i\in AB}min(n_{Ai},n_{Bi})}{n_A+n_B}$	Fraction of minimum per-sample abundance of shared taxa	Bray and Curtis, 1957
Morisita-Horn	MH	$\frac{2\sum_{i\in A\cup B}n_{Ai}n_{Bi}}{\left(\frac{\sum_{i\in A}n_{Ai}^2}{n_A^2}+\frac{\sum_{i\in B}n_{Bi}^2}{n_B^2}\right)n_An_B}$	Pairwise overlap of taxa abundances, adjusted by per-sample concentration indices	Horn, 1966
UniFrac, unweighted	UFU	$\frac{\sum_{i\in A\cap B}\Phi_i}{\sum_{j\in A\cup B}\Phi_j}$	Fraction of shared branch length on a phylogenetic tree	Lozupone and Knight, 2005
UniFrac, weighted	UFW	$1-\frac{\sum_{i\in A\cup B}\Phi_i\left\lvert\frac{n_{Ai}}{n_A}-\frac{n_{Bi}}{n_B}\right\rvert}{\overline{\Phi}}$	Fraction of shared branch length on a phylogenetic tree, weighted by taxa abundances	Lozupone *et al.*, 2007
TINA, unweighted	TU	$\frac{\sum_{i\in A}\sum_{j\in B}C_{ij}}{N_AN_B}$	Average pairwise taxa interaction similarity.	This study
TINA, weighted	TW	$\frac{\sum_{i\in A}\sum_{j\in B}\frac{n_{Ai}n_{Bj}}{n_A\,n_B}C_{ij}}{\sqrt{\left(\sum_{i\in A}\sum_{j\in A}\frac{n_{Ai}n_{Aj}}{n_A^2}C_{ij}\right)\left(\sum_{i\in B}\sum_{j\in B}\frac{n_{Bi}n_{Bj}}{n_B^2}C_{ij}\right)}}$	Average pairwise taxa interaction similarity, weighted by taxa abundances	This study
PINA, unweighted	PU	$\frac{\sum_{i\in A}\sum_{j\in B}\Phi_{ij}}{N_AN_B}$	Average taxa phylogenetic similarity	This study
PINA, weighted	PW	$\frac{\sum_{i\in A}\sum_{j\in B}\frac{n_{Ai}n_{Bj}}{n_A\,n_B}\Phi_{ij}}{\sqrt{\left(\sum_{i\in A}\sum_{j\in A}\frac{n_{Ai}n_{Aj}}{n_A^2}\Phi_{ij}\right)\left(\sum_{i\in B}\sum_{j\in B}\frac{n_{Bi}n_{Bj}}{n_B^2}\Phi_{ij}\right)}}$	Average pairwise taxa phylogenetic similarity, weighted by taxa abundances.	This study

All formulas and descriptions are given as similarities; for ordinations and statistical tests, corresponding distances or dissimilarities are used (D = 1 − S). N_A, total number of taxa in sample A; n_A total number of individuals in sample A; n_{Ai}, individuals of taxon i in sample A; Û, estimator for fraction of individuals in shared taxa for sample A, according to formula 9 in Chao *et al* (2004); φ_i, phylogenetic branch length of taxon i to root; Φ_{ij}, phylogenetic association between taxa i and j; C_{ij}, interaction similarity between taxa i and j.

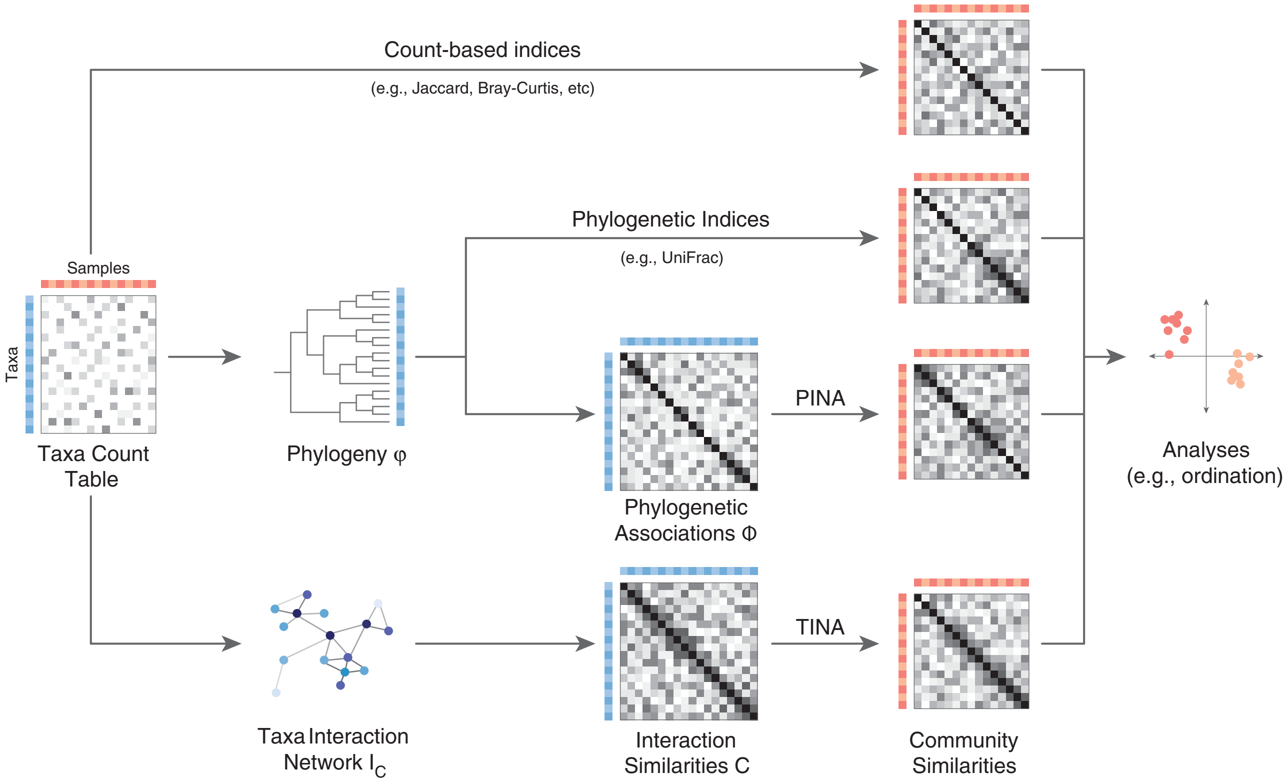

Count-based indices
(e.g., Jaccard, Bray-Curtis, etc)
Phylogenetic Indices
(e.g., UniFrac)
Samples
Taxa
Taxa Count Table
Phylogeny φ
Phylogenetic Associations Φ
PINA
Analyses (e.g., ordination)
Taxa Interaction Network I_C
Interaction Similarities C
TINA
Community Similarities

Figure 1 Overview of different approaches to quantifying community similarity. Based on a taxa-sample count table, traditional count-based indices such as Jaccard and Bray–Curtis quantify community similarity from the overlap in taxa composition (upper branch). In contrast, phylogenetic indices such as UniFrac take into account taxa relationships, quantifying community similarity as shared evolutionary history, based on taxa phylogeny (middle branch). Our proposed Taxa INteraction-Adjusted (TINA) and Phylogenetic INteraction-Adjusted (PINA) indices, in contrast, take into account similarities on a taxa co-occurrence network, codified in an interaction similarity matrix C, or in terms of cophenetic phylogenetic distances, represented in a phylogenetic association matrix Φ.

Therefore, it is important to transform the interaction matrix I_C to a common scale; we do this by correlating taxa by their pairwise associations to all other taxa in the system (that is, row $I_{C,i*}$ of I_C for taxon i) and transforming this into a Pearson similarity:

$$C_{ij} = 0.5 * (1 + \rho_{Pearson}(I_{C,i*}, I_{C,j*}))$$

Thus defined, a transformed co-occurrence matrix C has several desirable properties: (i) it scales from 0 (avoidance) to 0.5 (neutral association) to 1 (complete association); (ii) C_{ij} corresponds to the ‘proximity’ of taxa i and j on the original association network I_C; (iii) the transformation generally sharpens and smoothens network structure, amplifying strong associations at the expense of weaker correlations, but association ranks remain mostly unchanged (see Supplementary Figure S1 in Supporting Information).

Given this transformed interaction matrix, we propose to quantify the similarity between communities A and B as the *average interaction strength between all taxa observed in A or B*. We thus define an incidence-based or unweighted *Taxa INteraction-Adjusted* index of community similarity (*unweighted TINA*, TU) as

$$TU = \frac{\sum_{i \in A} \sum_{j \in B} C_{ij}}{N_A N_B}$$

Likewise, an abundance-based or *weighted TINA* index (TW) can be defined as *weighted average taxa interaction strength*, scaled by the geometric mean per-sample weighted taxa interaction strength:

$$TW = \frac{\sum_{i \in A} \sum_{j \in B} \frac{n_{Ai} n_{Bj}}{n_A\, n_B} C_{ij}}{\sqrt{(\sum_{i \in A} \sum_{j \in A} \frac{n_{Ai} n_{Aj}}{n_A^2} C_{ij})(\sum_{i \in B} \sum_{j \in B} \frac{n_{Bi} n_{Bj}}{n_B^2} C_{ij})}}$$

TINA values are 1 for two completely identical communities, but also if all taxa in A and B are perfectly associated. If no taxa are shared, TINA values tend towards 0.5 if taxa interactions are neutral (neither associative nor dissociative) and towards 0 if taxa between A and B show a strong avoidance signal. Thus, TINA resolves non-zero similarities even for pairs of communities that do not share any taxa (which implies zero similarity according to traditional, count-based indices); theoretically, the TINA index for such disparate pairs can even be 1 if all their taxa are perfectly associated.

TINA-like indices can be defined analogously for any kind of interaction data, given that interaction matrices can be transformed similarly to the I_C to C transformation described above. This is also true for the special case of phylogenetic 'interactions'. Consider a phylogenetic tree φ of taxa observed in a given system with a *cophenetic phylogenetic similarity* matrix I_φ, which can be interpreted as a phylogenetic association network (analogous to I_C) and transformed into an association matrix Φ (analogous to the co-occurrence association matrix C). Then, we can define *unweighted Phylogenetic INteraction-Adjusted* community similarity (*unweighted PINA*, PU) as

$$PU = \frac{\sum_{i\in A}\sum_{j\in B}\Phi_{ij}}{N_A N_B}$$

and *weighted PINA* (PW) as

$$PW = \frac{\sum_{i\in A}\sum_{j\in B}\frac{n_{Ai}n_{Bj}}{n_A\, n_B}\Phi_{ij}}{\sqrt{\left(\sum_{i\in A}\sum_{j\in A}\frac{n_{Ai}n_{Aj}}{n_A^2}\Phi_{ij}\right)\left(\sum_{i\in B}\sum_{j\in B}\frac{n_{Bi}n_{Bj}}{n_B^2}\Phi_{ij}\right)}}$$

Human Microbiome Project data analysis

To test the performance of different community similarity indices, we re-analysed two publicly available datasets, provided by the Human Microbiome Project and the TARA Oceans project, respectively. Raw 16S rRNA V35 region sequencing reads of The Human Microbiome Project (2012) were downloaded from the NCBI Sequence Read Archive; metadata was obtained from the HMP data repository (http://hmpdacc.org). Sequences were filtered for chimeric reads using UCHIME (Edgar *et al.*, 2011), aligned to a secondary structure-aware 16S rRNA model using Infernal (Nawrocki and Eddy, 2013), denoised by a global minimum read abundance at 1% tolerance of 4 and clustered into OTUs at 97% average linkage sequence similarity using hpc-clust (Matias Rodrigues and von Mering, 2014), as established previously (Schmidt *et al.*, 2014, 2015). The resulting filtered taxa count table contained 24 717 447 sequences clustered into 27 041 OTUs across 3849 samples. A phylogenetic tree of OTU representatives, selected by minimum average within-OTU sequence distance, was inferred using FastTree2 (Price *et al.*, 2010) with default parameters. Pairwise taxa co-occurence networks for the full dataset and subsets were calculated using taxa-wise Bray–Curtis dissimilarity, weighted Jaccard index, Spearman correlation and a custom R implementation of SparCC (Friedman and Alm, 2012), an adapted correlation metric correcting for spurious associations that has been shown to approximate 'true' ecological interactions as simulated using a Generalized Lotka–Volterra (GLV) model (Berry and Widder, 2014).

Simulation of human microbiome samples

Simulations were conducted based on (re-sampled) real count data subsets of the HMP dataset or based on entirely synthetic counts. GLV models were adapted from Berry and Widder (2014), with per-sample carrying capacities set uniformly to total sample size, initial per-sample counts set to sum to 10% of sample carrying capacity, growth rates re-sampled from a uniform distribution on [0,1] unless otherwise specified and further parameters chosen test-specifically. All simulations were repeated 20 times (see online analysis code).

TARA Oceans data processing and analysis

From the TARA Oceans eukaryotic plankton diversity census (De Vargas *et al.*, 2015), we downloaded 18S rRNA V9 tag data organised into an OTU-level taxa count table (http://doi.pangaea.de/10.1594/PANGAEA.843022) and sample metadata (http://doi.pangaea.de/10.1594/PANGAEA.843017). Data per sample were pooled across filter sizes and OTUs containing ⩽30 sequences as well as several orphan samples were removed (see analysis code), yielding a filtered count table of 535 903 407 sequences, 27 448 OTUs and 77 samples for which a SparCC correlation network was computed.

Data and software availability

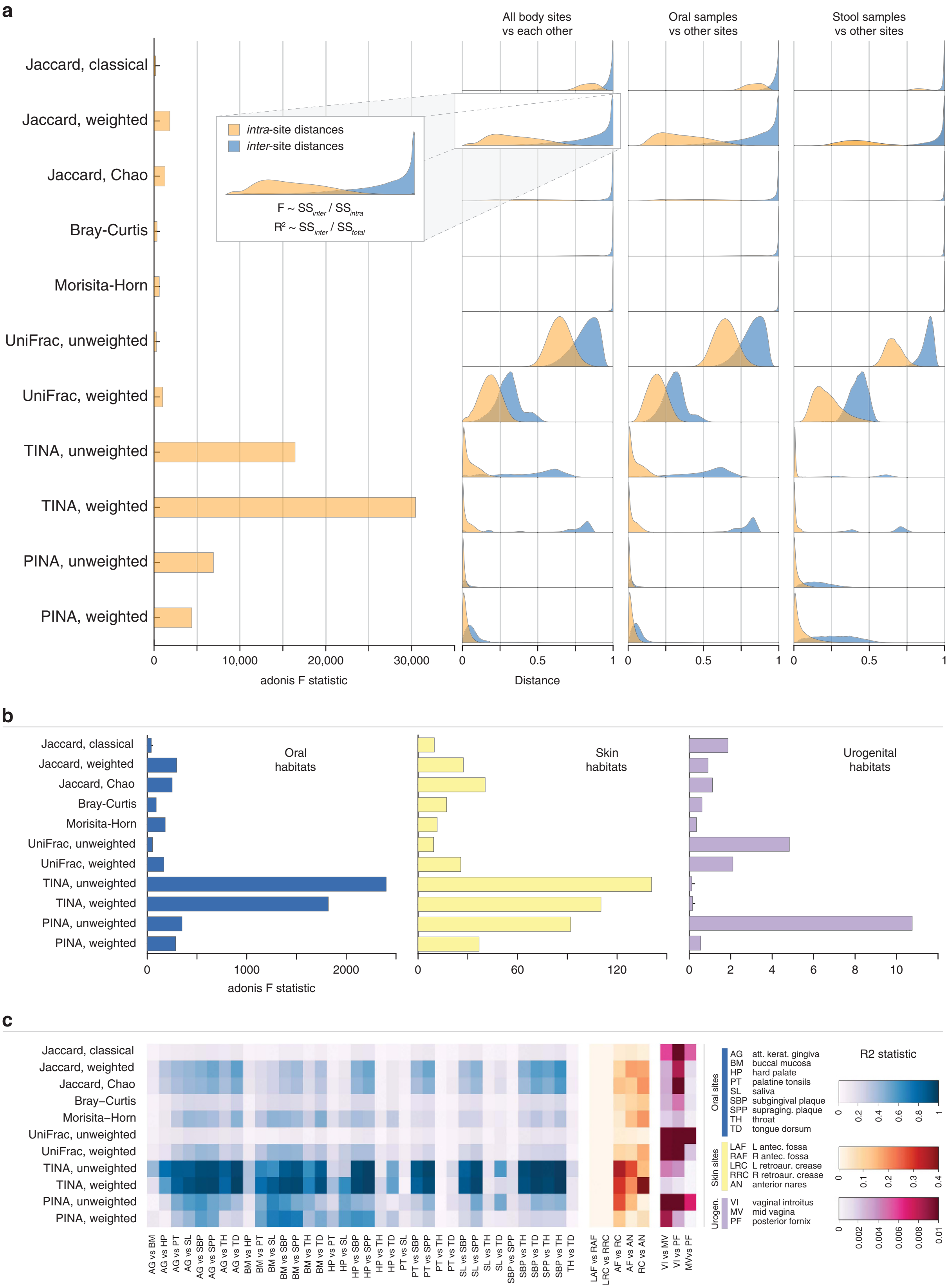
All analysis code and processed datasets are available online (http://github.com/defleury/Schmidt_et_al_2016_community_similarity; http://meringlab.org/suppdata/2016-community_similarity/).

Results

TINA and PINA provide improved partitioning of human body site-specific microbial communities

From an ecological point of view, the human body appears as little more than a system of distinct microbial habitats (Costello *et al.*, 2012). The *Human Microbiome Project* (HMP Consortium, 2012) has provided a first comprehensive census of human body-associated microbial communities and their potential functional repertoires. HMP 16S rRNA tag sequencing data are available for 18 habitats from five different body sites, namely oral cavity (9 habitats), skin (4), female urogenital tract (3), airways (1) and gastrointestinal tract (1); see Figure 2 for a full list. One original goal of the HMP, similar to many ecological studies, was to establish how compositional similarity patterns distinguish communities associated to these different habitats. In other words: are body sites distinct from each other in microbial community composition, and which other factors drive compositional variation?

These types of questions are typically addressed by calculating pairwise community distances and then applying multivariate statistical tests to establish how much of the distance matrix structure is


a
All body sites vs each other
Oral samples vs other sites
Stool samples vs other sites
Jaccard, classical
Jaccard, weighted
Jaccard, Chao
Bray-Curtis
Morisita-Horn
UniFrac, unweighted
UniFrac, weighted
TINA, unweighted
TINA, weighted
PINA, unweighted
PINA, weighted
intra-site distances
inter-site distances
F ~ SS_inter / SS_intra
R² ~ SS_inter / SS_total
0
10,000
20,000
30,000
adonis F statistic
0
0.5
1
Distance
b
Oral habitats
Skin habitats
Urogenital habitats
0
1000
2000
60
120
2
4
6
8
10
adonis F statistic
c
AG vs BM
AG vs HP
AG vs PT
AG vs SL
AG vs SBP
AG vs SPP
AG vs TH
AG vs TD
BM vs HP
BM vs PT
BM vs SL
BM vs SBP
BM vs SPP
BM vs TH
BM vs TD
HP vs PT
HP vs SL
HP vs SBP
HP vs SPP
HP vs TH
HP vs TD
PT vs SL
PT vs SBP
PT vs SPP
PT vs TH
PT vs TD
SL vs SBP
SL vs SPP
SL vs TH
SL vs TD
SBP vs SPP
SBP vs TH
SBP vs TD
SPP vs TH
SBP vs TH
TH vs TD
LAF vs RAF
LRC vs RRC
AF vs RC
AF vs AN
RC vs AN
VI vs MV
VI vs PF
MV vs PF
Oral sites
AG att. kerat. gingiva
BM buccal mucosa
HP hard palate
PT palatine tonsils
SL saliva
SBP subgingival plaque
SPP supragaing. plaque
TH throat
TD tongue dorsum
Skin sites
LAF L antec. fossa
RAF R antec. fossa
LRC L retroaur. crease
RRC R retroaur. crease
AN anterior nares
Urogen.
VI vaginal introitus
MV mid vagina
PF posterior fornix
R2 statistic
0
0.2
0.4
0.6
0.8
1
0
0.1
0.2
0.3
0.4
0
0.002
0.004
0.006
0.008
0.01

explained by a given model. One of the most widely used methods is Anderson's PERMANOVA (permutational analysis of variance; Anderson, 2001), implemented in the *adonis* function of the R package *vegan* (Oksanen *et al*., 2015), which calculates a pseudo-F statistic on group separation from the sums of squares of *inter*-group distances over the sums of squares of *intra*-group distances (see Figure 2a) and then conducts a permutational significance test. Thus, the adonis F statistic, as well as the related R^2 statistic (the variance explained by the tested factor) indicate an *effect size* of multivariate group separation (higher F and R^2 values indicate more discriminatory power), while a permutational P value indicates significance. F and R^2 statistics have previously been used to benchmark multivariate ecological analyses, for example by Eren *et al* (2015).

We re-analysed the HMP data using 11 different community similarity indices (Table 1), five of which are count-based (JCI, JCW, JCC, BC and MH), two phylogenetic (UFU, UFW) and four interaction-adjusted (TU, TW, PU and PW). We observed that partitioning of the five general body sites (oral cavity, skin, urogenital tract, airways, gut) by community similarity was by far best for TINA (based on a SparCC correlation network; F = 30 455 for TW; F = 16 421 for TU) and PINA (PW, 4397; PU, 6917) when compared to all other indices, with JCI providing the weakest discrimination (F = 182). This improved partitioning was due to several effects, as indicated by community distance histograms per index (Figure 2a). First, TINA and PINA provided very high overall resolution, distributing pairwise dissimilarities across a broader range on the interval 0 (identical communities) to 1 (complete dissimilarity) than most count-based indices. Second, TINA and PINA assigned very low and sharply distributed intra-group dissimilarities, meaning that samples from the same body sites were on average considered very similar to each other; in contrast, count-based indices showed very sharp and pronounced inter-group dissimilarities, but wider distributions within groups. Finally, intra- and inter-group dissimilarities were generally much clearer separated for TINA than for count-based indices or UniFrac.

To assess the robustness of TINA performance towards the choice of network inference method, we re-tested body site separation based on re-calculated taxa co-occurrences using four different approaches: (i) taxa-wise Bray–Curtis dissimilarity (TW-BC); (ii) taxa-wise weighted Jaccard distance (TW-JCW); (iii) Spearman rank correlation of taxa across samples (TW-SP); and (iv) SparCC correlation (Friedman and Alm, 2012). We found that network inference method had a strong effect on both unweighted and weighted TINA in terms of F and R^2 values, but that discriminative power was consistently high. SparCC-based TINA outperformed TW-SP, TW-JCW and TW-BC (in this order), and all TINA metrics outperformed UniFrac and count-based indices (see Supplementary Figure S2). Since by far the strongest separation signal was observed for TINA based on SparCC, we relied on SparCC for network inference in the subsequent tests reported below.

Combined interpretation of TINA and PINA may provide novel biological insights

While habitat partitioning was differentially pronounced, all indices provided significant group separation ($P \leqslant 0.001$, 999 permutations). Indeed, differences in community composition between body sites – which are highly distinct micro-environments – can be expected to be large, so it is not surprising that they were picked up by all indices. We therefore conducted similar tests on more complicated problems, such as the separation of habitats within a body site (Figure 2b) or of pairs of similar habitats (Figure 2c). TINA provided by far the strongest partitioning of oral and skin habitats, followed by PINA and JCW/JCC. For urogenital sites, in contrast, only few indices provided significant separation at all: unweighted PINA (PU), UFU, UFW and JCI. These trends were consistent with pairwise separability of habitats (Figure 2c), which was highest for TINA in oral and skin, but for PU, UFU and UFW in urogenital sites.

These observations imply that diversity patterns in these habitats are determined by different factors. TINA quantifies community similarity as an overlap in ecological associations of taxa, while PINA and UniFrac focus on shared phylogeny. Thus, it appears that the compositional identity of oral and skin sites is driven by recurring cliques of associated OTUs, as captured by strong co-occurrence signals, while pairwise taxa associations are less important in the urogenital tract, where communities of changing partners are instead filtered by phylogeny, possibly indicating a functional signal. Indeed, we detected a small but significant phylogenetic signal for the

◀

Figure 2 Differential partitioning of human body habitat-specific community structure. (**a**) Partitioning by general body sites. Left, PERMANOVA F statistics for different indices when testing community distance partitioning according to general body site, that is into oral, skin, urogenital, airways and gastrointestinal habitats. Right, histograms of community distances intra-site (orange) and inter-site (blue) for all body sites against each other, oral against other habitats and gastrointestinal against other habitats. The inset illustrates how PERMANOVA F statistics and R^2 values are calculated from community distances using the *adonis* function of the R package *vegan* (Anderson, 2001; Oksanen *et al*., 2015). (**b**) Sub-partitioning of oral habitats (blue), skin habitats (yellow) and urogenital habitats (violet). (**c**) Pairwise separation by different indices of all pairs of oral habitats, skin habitats against each other and against airways (anterior nares) and urogenital habitats against each other, as detected by the PERMANOVA R^2 value (relative variance explained by factor habitat). Note the different colour scales, indicating overall differential partitioning power per body site.

distribution of pooled OTU abundances across urogenital sites (Supplementary Figure S3; vaginal introitus: $K = 0.051$, $P \leqslant 0.001$; mid-vagina: $K = 0.023$, $P \leqslant 0.001$; posterior fornix: $K = 0.023$, $P \leqslant 0.001$). However, body subsite was indeed not the predominant determinant of per-sample community composition (Supplementary Figure S4): rather, variation between subjects was higher than within subjects ($P \leqslant 0.02$ for all tested metrics except, interestingly, unweighted PINA), while even this factor accounted for a maximum of only 4% of variation (unweighted TINA). This is in line with observations in the original HMP study, which moreover established strong associations to other factors such as pH or BMI (HMP Consortium, 2012). Thus, a combined interpretation of TINA and PINA may guide biological interpretation: between-subject community composition appears to be (moderately) determined by taxa co-occurrence, whereas a weak but significant phylogenetic signal may shape differences between vaginal subsites.

TINA captures habitat structure of the human microbiome taxa co-occurrence network

To illustrate how TINA captures ecological taxa interaction structure, how TINA values can be interpreted at the level of individual sample pairs and under which conditions it provides more intuitive results than count-based indices, we selected 16 HMP samples for which Figure 3 shows the taxa co-occurrence network; Supplementary Figure S5 shows the same network, coloured by OTU phylum-level taxonomy. We observed that the network is structured into several habitat-specific clusters of strongly co-occurring OTUs, with slightly less dependence on taxonomy. This is in line with the general observation that (microbial) co-occurrence networks tend to capture ecological signals, which indeed can often be much more subtle than the present body habitat classification (Faust and Raes, 2012).

Next, we mapped three examples of sample pairs onto this network to illustrate how TINA takes interaction structure into account to quantify diversity. In the first example (Figure 4a), we consider two samples from the vaginal posterior fornix, which were the most similar pair according to the classical Jaccard index (lowest JCI distance). These samples have a high taxa overlap, while many of their non-shared taxa also share strong co-occurrence signals, so that TINA likewise assigns them a very high similarity. Thus, in this case, TINA and count-based indices agree in considering both samples highly similar.

The second case is less trivial (Figure 4b). Here, we consider two urogenital samples from the vaginal introitus, which do not share any taxa at all – count-based indices assign a distance of 1, considering them completely dissimilar. However, the taxa found in these samples share an attractive interaction signal, even though some taxa form part of distinct co-occurence clusters. In this case, TINA provides a more intuitive result by ranking the similarity between two samples from the same habitat (vaginal introitus) relatively high.

Finally, consider the opposite situation (Figure 4c). Here, two samples from different body sites (skin and oral) happen to share several taxa, so that JCI ranks them among the top 17.3% most similar pairs. However, most non-shared taxa belong to distinct co-occurrence clusters, meaning that their pairwise interactions are repulsive, such that TINA assigns a very high dissimilarity to this pair, which, again, is an arguably more realistic picture.

Interaction-adjusted indices provide strong partitioning even for small datasets

The HMP dataset used in this study is comparatively large, comprising 27 041 OTUs across 3849 samples. To test whether the observed trends were robust to dataset scope, we conducted two down-sampling experiments (Figure 5 and Supplementary Figure S6). First, we randomly selected between 5 and 50 samples per body site (25–250 samples in total), re-calculated co-occurrence and phylogenetic interaction strengths and assessed body site separation by all 11 indices, at 10 iterations per down-sampling step (Figure 5a; see Supplementary Figure S6A for the corresponding plot on the R^2 statistic). We found that even for the smallest tested dataset, TINA and PINA indices provided much clearer partitioning by body site, although ranks by partitioning effect sizes varied across down-sampling iterations. Next, we randomly selected 1000 samples from which we drew 1000 sequences each and down-sampled these to 50 sequences per sample in several steps, at 10 iterations per step, recalculated co-occurrence and phylogenetic interactions and quantified community similarity (Figure 5b and Supplementary Figure S6B). Likewise, we observed that TINA and PINA provided much better separation by body site than all other indices, even at a drastically small size of 50 sequences per sample.

TINA captures both habitat preference and taxa interaction signals

To further investigate the behaviour of TINA in different scenarios, we conducted a series of simulations (Figure 6; raw simulation results available in Supplementary Table S1). We randomly selected 25 oral and 25 gastrointestinal samples from the HMP dataset, from which we sampled 200 of the 1000 most abundant OTUs. For this reduced dataset, we inferred Dirichlet multinomial mixture models using the software *microbedmm* (Holmes *et al.*, 2012); the DMM models perfectly re-discovered the two habitat groups. Next, we iteratively simulated count data by multinomially sampling from these models and

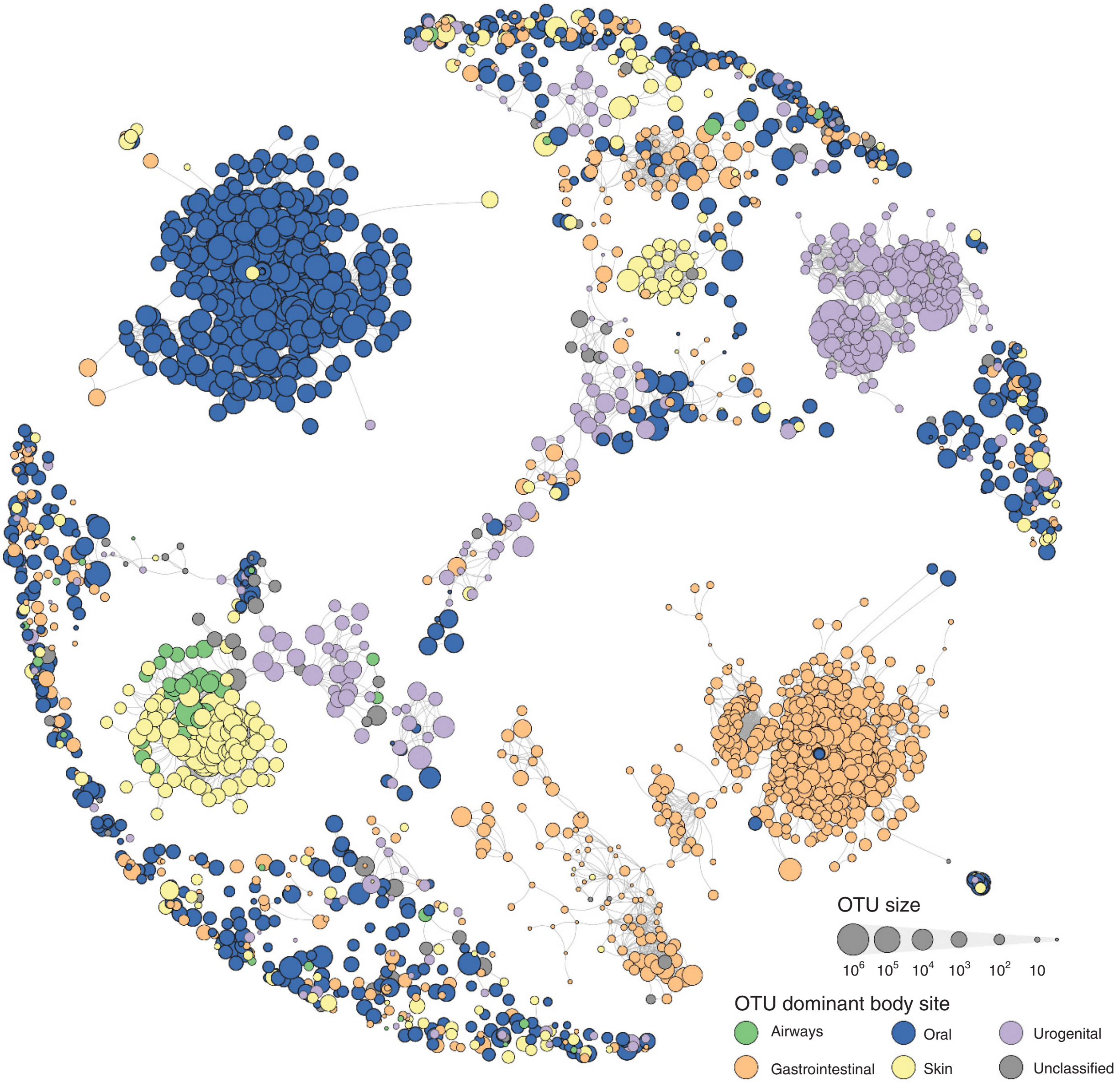


Figure 3 Taxa co-occurrence network for a subset of Human Microbiome Project samples. Sixteen samples from the full HMP dataset were selected as described in the main text, comprising a total of 2671 OTUs for which all pairwise SparCC correlations ≥ 0.5 are shown as edges in the network. Node size indicates global OTU size, that is the total number of counts per OTU across the full HMP dataset. Node colour indicates OTU dominant habitat, assigned if more than 50% of all OTU abundance was in samples of the same body site. Supplementary Figure S2 shows the same network, coloured by OTU phylum-level taxonomy.

re-calculated habitat separation for different community similarity indices, conducting both across-habitat comparisons and respective within-habitat tests as negative controls (Figure 6a). We observed that all tested indices had high discriminative power, but that F values for TINA were up to three orders of magnitude larger than for count-based indices. This is in line with the expectation that DMM inference accounts for both habitat preference and (implicitly) taxa co-abundance signals, both contributing to the taxa co-occurrence signal underlying TINA.

Thus, to disentangle the contributions of habitat preference and taxa interactions, we simulated count data using GLV models with different starting conditions and parameters. In a first setup, we shuffled raw OTU abundances across samples, regardless of habitat group (oral vs gut), thereby breaking any OTU habitat preference and OTU co-abundance signals. Based on these starting conditions, we simulated community dynamics in an interaction-neutral GLV model. This corresponded to a negative control setup: in the absence of any signal of habitat preference, taxa co-abundance or taxa interaction, none of the tested community similarity indices detected any separation between groups (Figure 6b). Next, we tested neutral

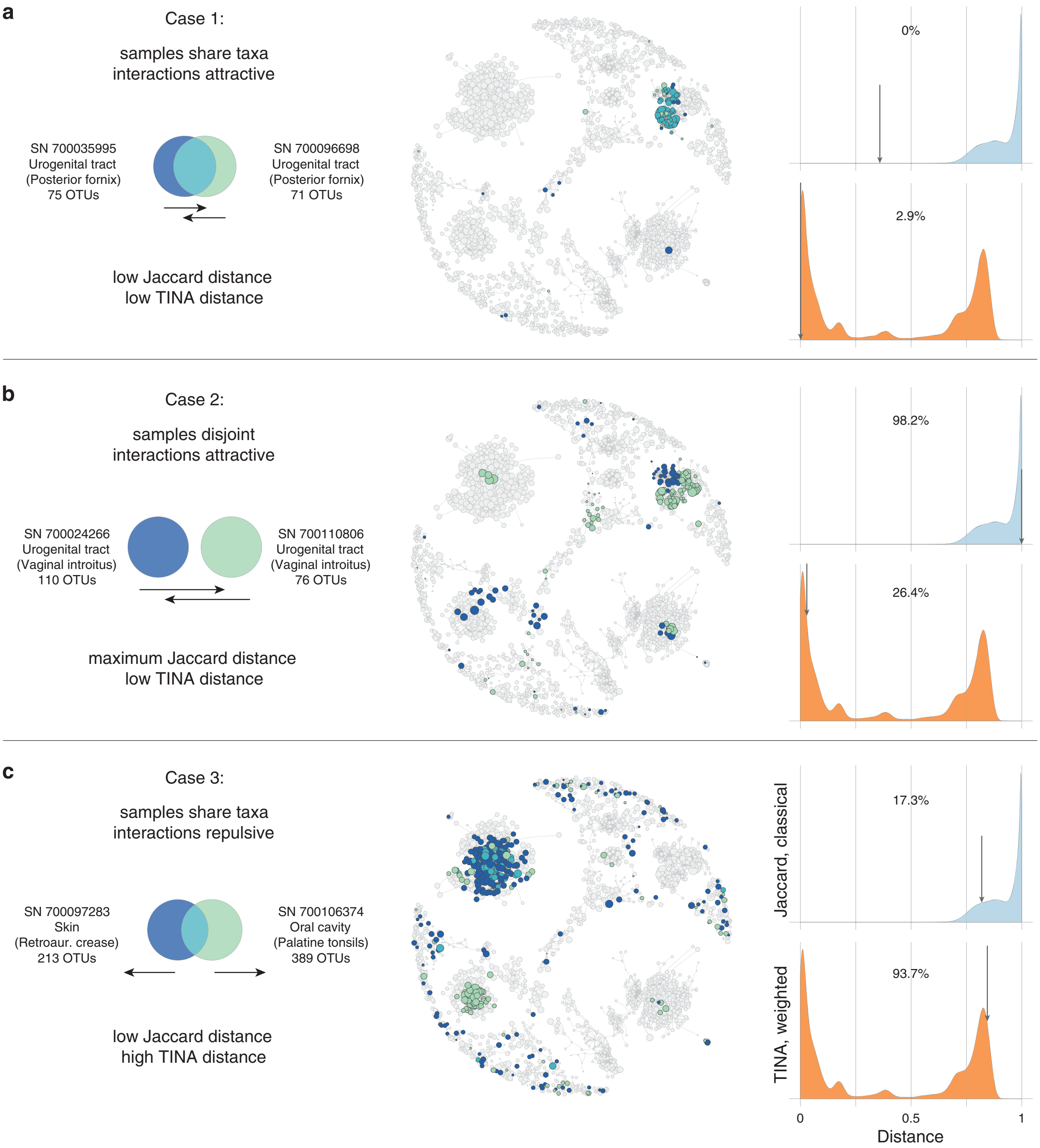


Figure 4 TINA quantifies community similarity from taxa co-occurrence. (**a**) For two urogenital samples that share a large taxa overlap, both traditional count-based indices (exemplified by the classical Jaccard index, JCI) and TINA assign a low-ranking community distance (as indicated in community distance histograms on the right). The middle panel shows how taxa of these samples map onto the co-occurrence network introduced in Figure 3; blue, taxa unique to sample SN700035995; green, taxa unique to sample SN700096698; blue-green, taxa shared between both samples. (**b**) For two urogenital samples that do not share any taxa, but whose taxa still share attractive co-occurrence interactions, JCI assigns complete distance (JCI = 1), while TINA assigns a relatively low-ranking distance. (**c**) In the opposite case of two samples from different body sites (skin and oral), which have a significant taxa overlap, but repulsive taxa interactions, JCI assigns a low-ranking, but TINA a very high-ranking distance.

community dynamics: we simulated data with habitat preferences, but with neutral taxa interactions by shuffling OTU abundances *within* habitat groups (thereby breaking co-abundance signals) as input for an interaction-neutral GLV model. In this setup, PERMANOVA F values for TINA were roughly three orders of magnitude stronger than for other indices (Figure 6c), which is in line with the

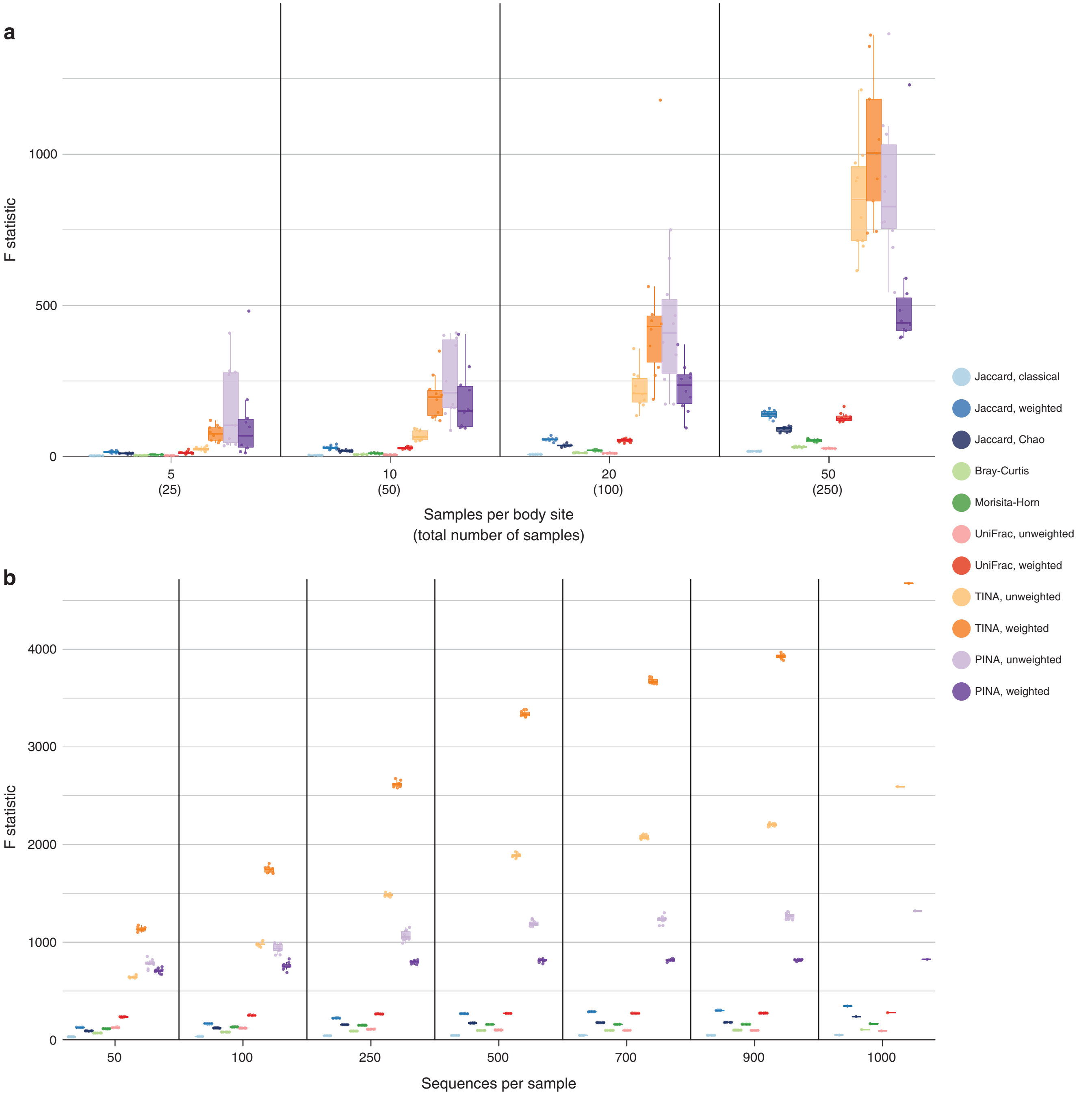


Figure 5 TINA and PINA detect strong partitioning by body habitat even for very small datasets. Adonis F statistic for separation by body site according to all 11 tested indices, for datasets randomly down-sampled following two different regimes. Taxa co-occurrences and phylogenetic interactions were re-calculated for every down-sampled dataset. (**a**) 5, 10, 20 and 50 samples per body site were randomly selected from the full HMP dataset, at 10 iterations per down-sampling step. (**b**) 1000 randomly selected samples were down-sampled to a depth of 1000 sequences per sample; this dataset was then further down-sampled to 50 sequences per sample in several steps, at 10 random iterations per step. Supplementary Figure S3 shows corresponding plots on R^2 values.

expectation that habitat preferences alone can generate strong taxa co-occurrence and avoidance signals.

In a third GLV setup, we tested the opposite scenario: absence of habitat preferences *a priori*, but non-neutral and habitat-specific taxa interactions. For this, we split the OTU set into two groups and generated a taxa interaction matrix as follows (see Figure 6d): within-group interactions were derived from a scale-free Barabási network (Barabási and Albert, 1999) with interaction strengths and signs randomly assigned from a uniform distribution; between-group interactions were set to competitive, with uniform noise. Based on this interaction matrix,

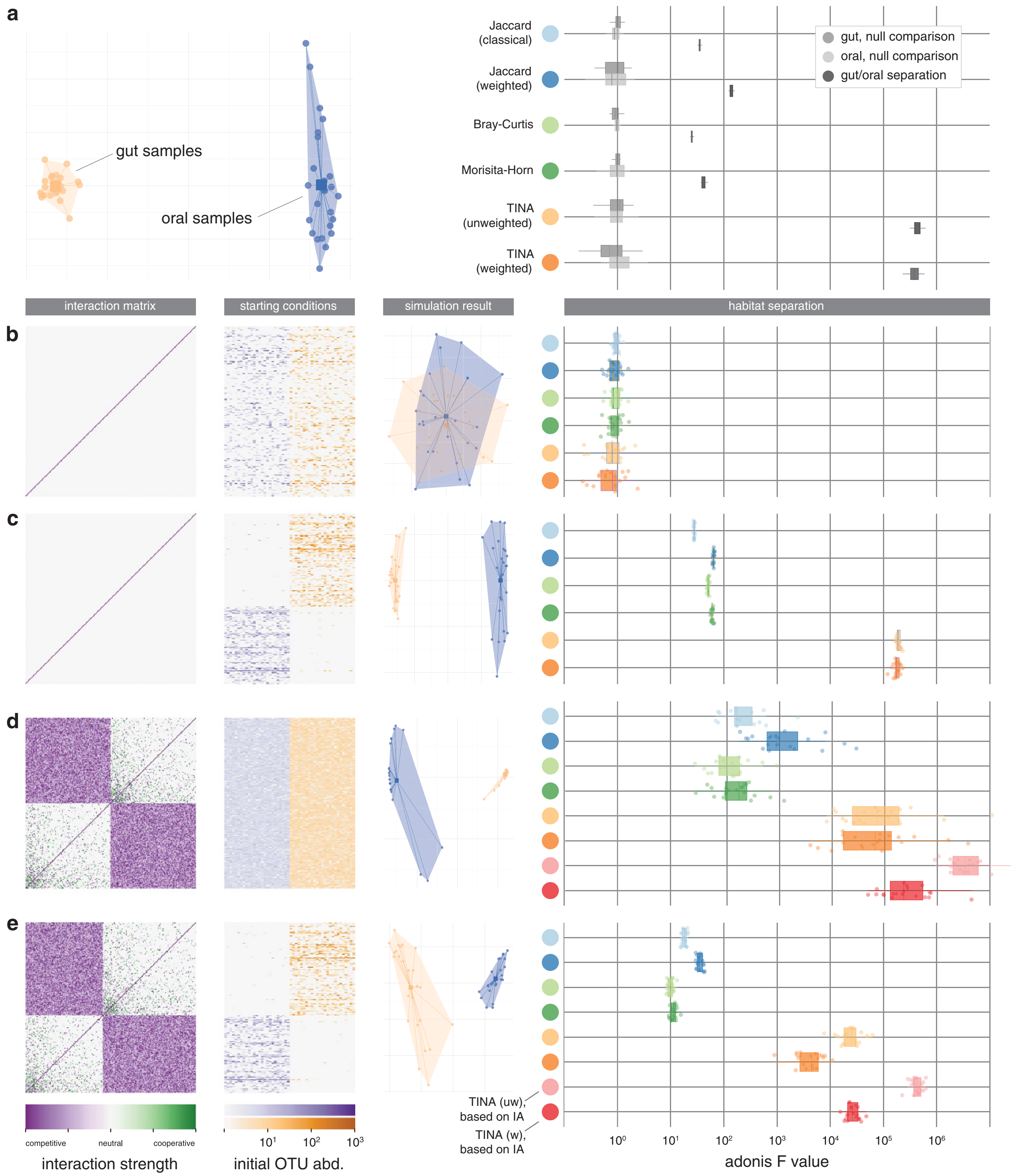


Figure 6 TINA captures both habitat preference and taxa interaction signals. (**a**) Separation of oral and gut samples, as simulated based on a Dirichlet Multinomial Mixture model inferred from raw count data. An example Principal Component Analysis (PCA) of simulation results (left) and partitioning performance for different indices when testing between-habitat and within-habitat separations (negative controls; right). (**b–e**) GLV simulations of count data. For each setup, an example interaction matrix, habitat preferences as encoded in a starting count table (oral and gut samples coloured blue and orange) and simulation outcome (as PCA ordination) are shown. The setups correspond to a negative control setting with no habitat preferences and neutral interactions (**b**), neutral community dynamics with habitat preferences only (**c**), habitat-naïve conditions with non-neutral interaction structure (**d**) and a combination of habitat preferences and habitat-specific interactions (**e**). In (**d–e**), light and dark red groups correspond to TINA calculated directly from the (known) interaction matrix. Three high outlying F values were removed in (**d**). Raw habitat partitioning data is available in Supplementary Table S1.

we performed GLV simulations with starting OTU counts randomly sampled from a Poisson distribution ($\lambda = 5$) and invariant growth rates. From the resulting count tables, we assigned samples to mock 'habitat groups' based on summed OTU group abundances. We observed that partitioning power was very high for all tested indices, with TINA outperforming count-based measures. Interestingly, when we computed TINA directly based on the (known) interaction matrix, it provided even higher F values than SparCC-based TINA, derived from a more indirect count-based co-occurrence signal. It is notable that interaction structure alone, in the absence of *a priori* habitat preferences, was sufficient to generate clearly discernible sample clusters in our simulations, although there were comparatively large levels of noise observable in partitioning power.

Finally, we investigated a setup with both habitat preferences and non-neutral interactions (Figure 6e). We assigned habitat preferences to OTUs based on their predominant occurrence in either oral or gut samples and generated an interaction matrix as described above with within-habitat structure. We then performed GLV simulations using these interactions and habitat-specific starting counts (assigned from within-group shuffling of raw OTU abundances). We observed that TINA had much higher partitioning power than other indices, and that TINA based directly on the interaction matrix outperformed SparCC-based TINA.

Biogeographical and physicochemical gradients structuring oceanic micro-eukaryote plankton communities are best captured by TINA

The TARA Oceans project has provided a very rich and multifaceted census of the world's oceans along several geographical and physicochemical gradients. We re-analysed TARA data on micro-eukaryote plankton diversity in the sunlit ocean (De Vargas *et al.*, 2015) to test the performance and versatility of TINA and other indices at representing different ecological signals. The dataset contained 77 samples from 43 stations along a wide geographical gradient (as shown in Figure 7a), taken at two depth levels, subsurface water (SUR) and deep chlorophyll maximum (DCM), and with body size filters ranging from 0.8 to 2000 μm. We calculated pairwise community distances according to JCI, JCW, JCC, BC, MH, TU and TW.

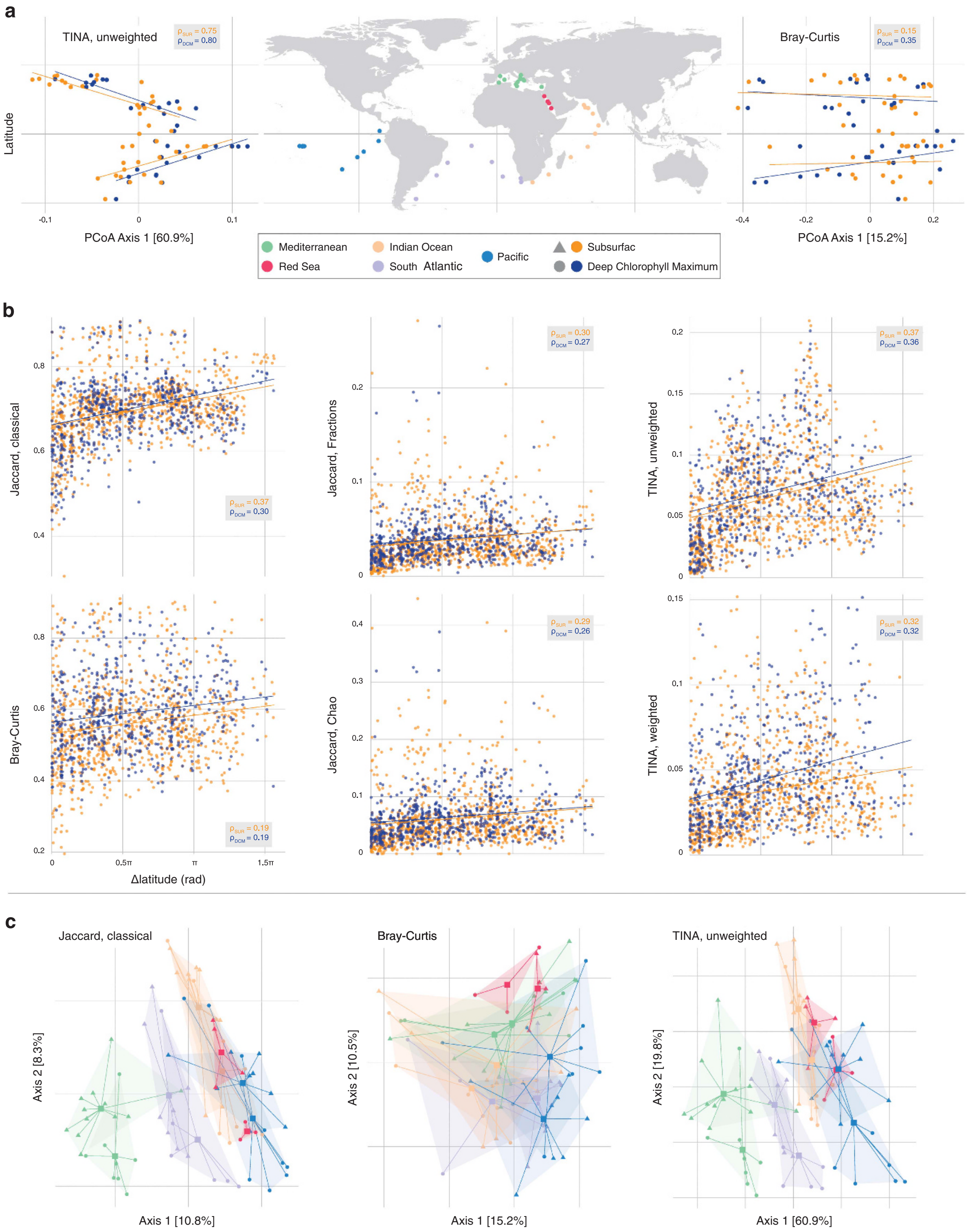
To test whether community similarity follows a latitudinal gradient, we correlated the first component of a Principal Coordinates ordination (PCoA) to latitude, separately for the northern and southern hemispheres (Figure 7a). We observed that the dominating component of an unweighted TINA-based PCoA correlated well with latitude, both for SUR ($\rho_{\text{Spearman}} = 0.75$) and DCM ($\rho = 0.8$) water layers. In contrast, an ordination based on BC (the index used in the original study) showed much weaker correlations for the dominating component ($\rho = 0.15$ and $\rho = 35$, respectively). We tested for these effects more systematically by correlating pairwise community distances to absolute differences in latitude between samples (Figure 7b). While all indices provided positive correlations, TU and JCI showed the strongest trend for SUR samples ($\rho_{\text{Spearman}} = 0.37$), while TU and TW showed the strongest signals for DCM ($\rho = 0.36$ and $\rho = 0.32$). Similar trends were observed for correlations of community distance to geographical distance (Supplementary Table S2). Thus, although JCI performed surprisingly well, TINA outperformed other indices at detecting a biogeographical signal. This can probably be rationalised when considering that taxa co-occurrence is expected to be in part determined by geography, which in turn also correlates with many other ecological parameters like water temperature, irradiation, salinity and so on.

Finally, we tested how well pairwise community distances represented the factors sampling region and depth; Figure 7c shows PCoA ordinations and PERMANOVA statistics for JCI, BC and TU. While all three indices provided significant separation by sampling region, water layer effects were significant only in JCI and TU, while only TU detected any effect of the region*depth interaction term. Moreover, effect sizes as expressed in F statistics and total variance explained by these three factors (summed R^2) were considerably higher for TU than for JCI and BC, indicating a much more pronounced partitioning according to these terms.

Discussion

The question of what processes are rendering communities similar or dissimilar to each other is fundamental to the study of diversity. Traditionally, community similarity has been quantified from compositional overlap, considering taxa as independent of each other and describing community structure based on census data. More recently, measures based on additional signals have become increasingly popular, most prominently phylogenetic indices based on shared evolutionary history. Such approaches take into account relationships between taxa, under the assumption that phylogenetic relatedness implies ecological and functional similarity. In this study, we have introduced a set of indices that follow a different rationale: we propose to quantify community similarity in terms of interactions between taxa.

There are several arguments for doing so. First, taxa interactions are a fundamental ecological parameter, at the heart of community ecology: they are important drivers of community assembly, composition and dynamics. Second, it is therefore meaningful to characterise taxa and their relationships based on their interactions: two taxa that are highly similar in terms of their interactions with



	F	R2	P	F	R2	P	F	R2	P
Region	8.98	0.323	≤0.001	3.02	0.145	≤0.001	30.82	0.608	≤0.001
Depth	4.22	0.038	0.004	2.01	0.024	0.09	5.71	0.028	0.002
Region * Depth	0.98	0.035	0.441	0.50	0.024	0.94	1.73	0.034	0.063

Figure 7 TINA captures biogeographical and physicochemical trends in ocean planktonic community composition. (**a**) TARA Oceans sampling locations, coloured by assigned oceanic world region (middle). The first axes of PCoA ordinations on TINA (left) and Bray–Curtis (right) dissimilarities correlated differentially well with latitude both for subsurface (SUR, orange) and deep chlorophyll maximum (DCM, blue) samples, both for the northern and southern hemisphere. (**b**) Spearman correlations of six different community distances against absolute difference in latitude (Δlat in rad). (**c**) PCoA ordinations and PERMANOVA statistics on dataset partitioning by factors region, depth and the region*depth interaction term for JCI, BC and TU.

other taxa can be considered ecologically almost equivalent, they can be assumed to fulfill similar roles in a community. Our approach captures this signal: we argue that it is meaningful to consider communities as similar, which contain ecologically similar taxa. Third, interaction network analysis has proved to be a very powerful tool in unravelling complex community dynamics, but its findings are often difficult to connect to the level of community-level diversity patterns. Our approach may help to bridge this analysis gap, by providing diversity indices that are based on networks and can be interpreted in a network framework. For example, TINA essentially quantifies community distance as the distance on a taxa co-occurrence network. Finally, our approach is versatile: there are many types of ecological interactions, at many different levels, and in principle, interaction-adjusted indices like TINA and PINA can be formulated for all of them.

Several previous studies have investigated adapted versions of classical ecological diversity indices that take into account taxa dissimilarities. Most prominently, the quadratic entropy introduced by Rao (1982) calculates a concentration index, corrected for (known) dissimilarities between taxa; it is defined as the expected dissimilarity between randomly sampled individuals from a community and defaults to the classical Gini-Simpson index if all dissimilarities are zero. Ricotta and Marignani (2007), among others, suggested decompositions of Rao's quadratic entropy into β-diversity components in the classical ecology sense, that is interpreting it as the expected dissimilarity between individuals randomly sampled from different communities. Such approaches are conceptually and mathematically distinct from our proposed interaction-adjusted indices – the latter are formulated and interpreted as average (dis)similarities on networks, rather than entropy decompositions. Likewise, our approach differs conceptually from established network clustering methods, such as for example Markov Clustering (Enright *et al.*, 2002), which have been successfully applied to many biological problems. These exploit network structure to cluster nodes (that is, in our case, taxa) by similarity, but do not compare pre-defined samples (that is, groups of nodes) in terms of similarity on the network. Similarly, established metrics such as the Topological Overlap Measure (Li and Horvath, 2009) or Anet (Karpinets *et al.*, 2012) compute node similarities (not sample similarities) based on relative node positions in a network; these are analogous to the more basic correlative transformation of the interaction matrix I_C to matrix C discussed above (Supplementary Figure S2).

In this study, we have focused on two specific types of interaction-adjusted indices, TINA and PINA, and benchmarked their performance in a re-analysis of two large and complex datasets. TINA by far outperformed all other indices at discriminating human body habitat-specific communities (Figure 2), even for very small datasets (Figure 5). We demonstrated how co-occurrence-based TINA captures a habitat preference signal, a taxa interaction signal and a combination of both in simulations (Figure 6). Moreover, TINA best captured biogeographical trends and partitioning by the factors 'region' and 'water depth' for microeukaryotic plankton communities (Figure 7). These results can be interpreted in light of how TINA is computed. Taxa co-occurrence networks, on which TINA is based, capture an 'integrated' ecological association signal, in quantifying the observable outcome of the interplay of different levels of taxa interactions as patterns of co-abundance and avoidance (Faust and Raes, 2012). Although taxa association networks are not equivalent to 'true', ecological taxa interactions, it has recently been shown that they may provide good approximations of true interactions within certain limits (Berry and Widder, 2014). Indeed, they can reveal network structures that are specific to a given type of habitat (as shown for example in Figure 3) or structured according to their response to an ecological gradient or perturbance. Figure 4 illustrates anecdotally how TINA can capture such signals to assign community similarities that are more in line with ecological expectations than count-based indices. The fact that TINA provides such strong separations of body habitats and strong correlations with biogeography and other factors for plankton communities means that these factors have a strong and specific influence on taxa co-occurrence; subsequent diversity analyses should take this into account and interpret accordingly.

Likewise, PINA and other phylogenetic indices such as UniFrac are based on phylogenetic relatedness, operating under the assumption that shared phylogeny implies not only a shared evolutionary history, but similar ecological roles and functional profiles. However, by definition, an inference of ecological similarity from phylogeny is necessarily indirect and imperfect, although it is arguably a valid enough assumption for many phylogenetic clades. In all relevant tests, unweighted and/or weighted PINA, based on pairwise phylogenetic similarities between taxa, outperformed UniFrac, based on shared phylogeny only, even for very small datasets. Interestingly, unweighted PINA and UniFrac were the only indices to detect partitioning by urogenital habitats (Figure 2, Supplementary Figures S3 and S4), a task at which almost all count-based indices, as well as TINA, failed. As discussed above, this is likely due to differential factors and mechanisms shaping community structure in urogenital habitats, compared to other habitats. However, the fact that we did observe differential trends in index performance emphasises the importance of a multifaceted approach:

by applying count-based, phylogenetic and interaction-adjusted indices, different aspects of community similarity are quantified, which can be interpreted in context of each other, with the potential to reveal biological insights beyond the scope of mono-dimensional approaches.

One possible drawback of interaction-adjusted indices is that they are not context-invariant: their values will always depend on analysis scope and on the system under consideration, as they vary with varying network structure. While count-based indices will always assume the same similarity for two communities, independently of the remaining dataset, interaction-adjusted indices may change asymptotically when more data are added, simply due to (subtle) changes in network structure. However, this behaviour can indeed also be an asset, for example when comparing multiple datasets in a meta-study. In such a setup, ‘globally’ constructed networks may mitigate dataset-specific noise, introduced for example by sampling methods or limited depth. Likewise, interaction-adjusted indices are not limited to capturing static network architectures, but their flexibility allows to account for conditionally variable networks that are rewired for example over time, gradients or in response to specific changing factors.

In the large arsenal of measures for community similarity and, more generally, β diversity, our proposed family of interaction-adjusted indices provide an important, powerful and versatile alternative. By taking taxa interactions into account, they quantify novel aspects of ‘diversity’, at the very core of community ecology, and may guide biological interpretation of diversity patterns in novel ways.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

We thank David Berry and Stefanie Widder of the University of Vienna for sharing R code to conduct Generalized Lotka–Volterra modelling, and Janko Tackmann, Jordi Bascompte and Bernhard Schmid of the University of Zurich for helpful discussions during the preparation of this manuscript. This work was supported by an ERC starting grant (UMICIS/242870) and the Swiss National Science Foundation (31003A_160095).

References

Anderson MJ. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology* **26**: 32–46.

Anderson MJ, Crist TO, Chase JM, Vellend M, Inouye BD, Freestone AL *et al.* (2010). Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist. *Ecol Lett* **14**: 19–28.

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

Berry D, Widder S. (2014). Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Front Microbiol* **5**: 219.

Bray JR, Curtis JT. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecol Monogr* **27**: 325–349.

Chaffron S, Rehrauer H, Pernthaler J, von Mering C. (2010). A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Res* **20**: 947–959.

Chao A, Chazdon RL, Colwell RK, Shen T-J. (2004). A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecol Lett* **8**: 148–159.

Colwell RK. (2013). EstimateS version 9: Statistical estimation of species richness and shared species from samples http://purl.oclc.org/estimates.

Costello EK, Stagaman K, Dethlefsen L, Bohannan BJM, Relman DA. (2012). The application of ecological theory toward an understanding of the human microbiome. *Science* **336**: 1255–1262.

De Vargas C, Audic S, Henry N, Decelle J, Mahe F, Logares R *et al.* (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**: 1261605.

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**: 2194–2200.

Enright AJ, Van Dongen S, Ouzounis CA. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**: 1575–1584.

Eren AM, Morrison HG, Lescault PJ, Reveillaud J, Vineis JH, Sogin ML. (2015). Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME J* **9**: 968–979.

Faust K, Raes J. (2012). Microbial interactions: from networks to models. *Nat Rev Microbiol* **10**: 538–550.

Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, Raes J *et al.* (2012). Microbial co-occurrence relationships in the human microbiome. *PLoS Comput Biol* **8**: e1002606.

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

Graham CH, Fine PVA. (2008). Phylogenetic beta diversity: linking ecological and evolutionary processes across space in time. *Ecol Lett* **11**: 1265–1277.

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

Horn HS. (1966). Measurement of ‘overlap’ in comparative ecological studies. *Am Nat* **100**: 419–424.

Ings TC, Montoya JM, Bascompte J, Blüthgen N, Brown L, Dormann CF *et al.* (2009). Review: ecological networks – beyond food webs. *Journal Anim Ecol* **78**: 253–269.

Jaccard P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin del la Société Vaudoise des Sciences Naturelles* **37**: 547–579.

Karpinets TV, Park BH, Uberbacher EC. (2012). Analyzing large biological datasets with association networks. *Nucleic Acids Res* **40**: e131.

Li A, Horvath S. (2009). Network module detection: affinity search technique with the multi-node topological overlap measure. *BMC Res Notes* **2**: 1.

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.

Lozupone C, Knight R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* **71**: 8228–8235.

Lozupone CA, Hamady M, Kelley ST, Knight R. (2007). Quantitative and qualitative diversity measures lead to different insights into factors that structure microbial communities. *Appl Environ Microbiol* **73**: 1576–1585.

Matias Rodrigues JF, von Mering C. (2014). HPC-CLUST: distributed hierarchical clustering for large sets of nucleotide sequences. *Bioinformatics* **30**: 287–288.

McMurdie PJ, Holmes S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* **8**: e61217.

Nawrocki EP, Eddy SR. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**: 2933–2935.

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB *et al.* (2015). vegan: Community Ecology Package. Available at: http://CRAN.R-project.org/package = vegan.

Polis GA, Strong DR. (1996). Food web complexity and community dynamics. *Am Nat* **147**: 813.

Price MN, Dehal PS, Arkin AP. (2010). FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One* **5**: e9490.

Proulx SR, Promislow DEL, Phillips PC. (2005). Network thinking in ecology and evolution. *Trends Ecol Evol* **20**: 345–353.

Rao CR. (1982). Diversity and dissimilarity coefficients: a unified approach. *Theor Popul Biol* **21**: 24–43.

Ricotta C, Marignani M. (2007). Computing β-diversity with Rao's quadratic entropy: a change of perspective. *Divers Distrib* **13**: 237–241.

Schloss PD, Westcott SL, Rabyn T, Hall JR, Hartmann M, Hollister EB *et al.* (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**: 7537.

Schmidt TSB, Matias Rodrigues JF, von Mering C. (2014). Ecological consistency of SSU rRNA-based operational taxonomic units at a global scale. *PLoS Comput Biol* **10**: e1003594.

Schmidt TSB, Matias Rodrigues JF, von Mering C. (2015). Limits to robustness and reproducibility in the demarcation of operational taxonomic units. *Environ Microbiol* **17**: 1689–1706.

Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G *et al.* (2015). Structure and function of the global ocean microbiome. *Science* 2015; **348**: 1261359.

Swenson NG. (2011). Phylogenetic beta diversity metrics, trait evolution and inferring the functional beta diversity of communities. *PLoS One* **6**: e21264.

The Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature* **486**: 207–214.

Tuomisto H. (2010). A diversity of beta diversities: straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity. *Ecography* **33**: 2–22.

Vellend M. (2010). Conceptual synthesis in community ecology. *Q Rev Biol* **85**: 183–206.

Webb CO, Ackerly DD, McPeek MA, Donoghue MJ. (2002). Phylogenies and community ecology. *Annu Rev Ecol Syst* **33**: 475–505.

Whittaker RH. (1972). Evolution and measurement of species diversity. *Taxon* **21**: 213–251.

Whittaker RH. (1960). Vegetation of the Siskiyou mountains, Oregon and California. *Ecol Monogr* **30**: 279–338.

Supplementary Information accompanies this paper on The ISME Journal website (http://www.nature.com/ismej)