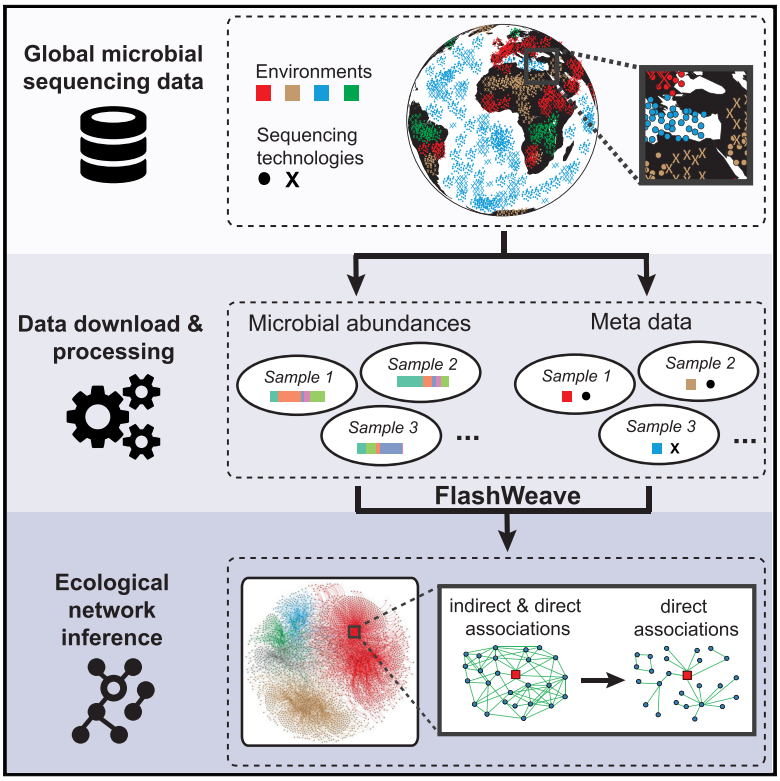
Report

Cell Systems

Rapid Inference of Direct Interactions in Large-Scale Ecological Networks from Heterogeneous Microbial Sequencing Data

Graphical Abstract



Authors

Janko Tackmann,
João Frederico Matias Rodrigues,
Christian von Mering

Correspondence

mering@imls.uzh.ch

In Brief

Spurious associations and computational complexity currently hinder ecological network inference from cross-study metagenomic data. Tackmann et al. present FlashWeave, a novel co-occurrence method that predicts interpretable microbial interaction networks through graphical model inference. FlashWeave is highly scalable and addresses data heterogeneity. They validate the method in extensive benchmarks on diverse synthetic and real-world data sets.

Highlights

- FlashWeave infers direct associations, resulting in sparse, interpretable networks
- The method's flexible graphical model framework scales to 500,000+ samples
- It integrates environmental & technical factors; adjusts for specific latent signals
- An extensive human gut microbial network reveals patterns of biological interest

Tackmann et al., 2019, Cell Systems 9, 286–296
September 25, 2019
https://doi.org/10.1016/j.cels.2019.08.002

CellPress

OPEN ACCESS
CellPress

Cell Systems
Report

Rapid Inference of Direct Interactions in Large-Scale Ecological Networks from Heterogeneous Microbial Sequencing Data

Janko Tackmann,[1] João Frederico Matias Rodrigues,[1] and Christian von Mering[1,2,*]
[1]Institute of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, CH-8057 Zürich, Switzerland
[2]Lead Contact
*Correspondence: mering@imls.uzh.ch
https://doi.org/10.1016/j.cels.2019.08.002

SUMMARY

The availability of large-scale metagenomic sequencing data can facilitate the understanding of microbial ecosystems in unprecedented detail. However, current computational methods for predicting ecological interactions are hampered by insufficient statistical resolution and limited computational scalability. They also do not integrate metadata, which can reduce the interpretability of predicted ecological patterns. Here, we present FlashWeave, a computational approach based on a flexible Probabilistic Graphical Model framework that integrates metadata and predicts direct microbial interactions from heterogeneous microbial abundance data sets with hundreds of thousands of samples. FlashWeave outperforms state-of-the-art methods on diverse benchmarking challenges in terms of runtime and accuracy. We use FlashWeave to analyze a cross-study data set of 69,818 publicly available human gut samples and produce, to the best of our knowledge, the largest and most diverse network of predicted, direct gastrointestinal microbial interactions to date. FlashWeave is freely available for download here: https://github.com/meringlab/FlashWeave.jl.

INTRODUCTION

Microorganisms shape virtually every aspect of Earth's biosphere. Besides their critical role in global biogeochemical cycles (Falkowski et al., 2008) and widespread symbiosis with all major branches of life (Oh et al., 2009; McFall-Ngai, 2014; Kawaguchi and Minamisawa, 2010), the tight coupling between the microbiome and human health is rapidly gaining appreciation (Carabotti et al., 2015; Thaiss et al., 2016). While the structure of microbial ecosystems is influenced by environmental factors and hosts (Bonder et al., 2016; Dyhrman et al., 2007; Krause et al., 2012), another important driving force is the ecological interactions between microbes (Faust and Raes, 2012; Xavier, 2011), such as competition, symbiosis, commensalism, and antagonism.

The inability to (co-)culture the majority of microorganisms in the lab (Solden et al., 2016; Goers et al., 2014) makes computational tools instrumental for the prediction of ecological dependencies between microbes, which can allow the generation of detailed hypotheses and better steer resource-intensive experimentation. Common to these approaches is the utilization of cross-sectional (co-occurrence and co-abundance [Chaffron et al., 2010; Friedman and Alm, 2012; Kurtz et al., 2015]) and temporal (Stein et al., 2013; Xia et al., 2011) statistical patterns, or alternatively metabolic complementarity (Zelezniak et al., 2015; Levy et al., 2015), to infer ecological associations and construct networks of predicted interactions. Currently, widespread methods are restricted to predicting pairwise interactions through univariate statistical associations (Friedman and Alm, 2012; Faust and Raes, 2016; Xia et al., 2011), but more recent approaches based on probabilistic graphical models (PGMs) consider the conditional dependency structure between microbes to distinguish between direct and indirect associations (Kurtz et al., 2015; Yang et al., 2017; Röttjers and Faust, 2018). Indirect (or spurious) associations can, for instance, be driven by indirect species interactions (i.e., interactions between two species conveyed through other intermediary species [Cazelles et al., 2016]) or by niche and batch effects. While PGM approaches can result in more sparse and interpretable networks, typical drawbacks include the requirement of larger data sets with sufficient statistical power and increased computational complexity. Hundreds of thousands of microbial sequencing samples from various environments around the globe are now available (Mitchell et al., 2018), alleviating the lack of statistical power, yet this wealth of data can currently not be utilized by state-of-the-art PGM methods due to insufficient computational scalability. Furthermore, sample heterogeneity of these cross-study data sets, such as variation in habitats, measurement conditions, and sequencing technology, can lead to confounding associations, typically not addressed by current methods (Röttjers and Faust, 2018). Exceptions include mLDM (Yang et al., 2017) and MInt (Biswas et al., 2015), which, however, do not address unmeasured sources of heterogeneity (i.e., latent factors).

Here, we present FlashWeave, a computational approach for inferring high-resolution interaction networks from large and heterogeneous collections of microbial sequencing samples based on co-occurrence or co-abundance. FlashWeave is optimized for computational speed and mitigates a number of known artifacts common in cross-study sequencing data analysis, such as compositionality effects, bystander effects, shared-niche

biases, and sequencing biases. It furthermore allows the seamless integration of environmental factors, such as temperature and pH, to estimate their influence on studied ecosystems and to remove indirect associations driven by them. Finally, it mitigates the impact of unmeasured confounding influences mediated by structural zeros (i.e., non-random absences driven by environmental or technical factors).

We compared FlashWeave to a variety of state-of-the-art methods on a wide collection of synthetic and biological benchmarks and showed that it outperforms other methods in terms of speed. In addition, it achieved overall increased accuracy, in particular on heterogeneous cross-habitat data sets with large fractions of structural zeros. We furthermore illustrated the usefulness of integrating non-microbial factors into network analysis by including habitat and primer variables into the inference of an interaction network based on the Human Microbiome Project. Finally, we applied FlashWeave to a global collection of 69,818 publicly available microbial sequencing samples of the human gastrointestinal tract, covering 488 studies. To our knowledge, the resulting association network represents the most comprehensive model of ecological dependencies of the human gut to date, depleted of indirect associations and inferred using minimal computational resources and time. We analyzed this network in depth to demonstrate its consistency with previously described ecological patterns. The model furthermore unveiled candidates for uncharacterized hub operational taxonomic units (OTUs) and yielded a notable signal for phylogenetic assortativity (PA) with potential biological relevance.

RESULTS

A Fast and Compositionally Robust Method for Ecological Network Inference, Capable of Handling Heterogeneous Data

FlashWeave is based on the local-to-global learning (LGL) approach proposed by Aliferis et al. (Aliferis et al., 2010a), a constraint-based causal inference framework for the prediction of direct relationships between variables in large systems. Algorithms of this family infer, for each target variable *T* in a system (in our case, OTUs or meta-variables [MVs]), its directly associated neighborhood, i.e., the set of neighbor variables that renders all remaining variables probabilistically independent of *T*. These sets are identified through a heuristically optimized sequence of statistical tests for conditional independence, which iteratively remove indirect edges while assuring that only the most promising tests are being performed. Subsequently, individual neighborhoods are connected into a global network through a combinator strategy (see STAR Methods for a detailed description of the full method). This procedure results in the removal of indirect (i.e., purely correlational) associations commonly reported by widespread univariate methods. Related algorithms have been successfully applied in a wide range of fields, including cancer diagnosis (Sboner and Aliferis, 2005), drug-drug interactions (Duda et al., 2005), and gene regulatory network inference (Narendra et al., 2011).

FlashWeave is an optimized implementation of the semi-interleaved HITON-PC (si-HITON-PC [Aliferis et al., 2010a]) instantiation of LGL (Figure 1A), extended through several heuristics, including the *feedforward* and *fast-elimination* heuristics, as well as a dedicated convergence criterion. These can drastically improve runtime for hub variables with large neighborhoods, which are a typical feature of ecological networks and pose considerable problems to the vanilla version of si-HITON-PC (see STAR Methods, "Heuristics"). In addition, FlashWeave incorporates methods for compositionality correction, which are essential since abundances from sequencing data constitute compositions, constrained to the simplex and long known to induce artificial correlations ((Pearson, 1896; Aitchison, 1981; Vandeputte et al., 2017); see STAR Methods, "Normalization").

In contrast to most other methods, FlashWeave can utilize MV information (Figure 1B), such as subject lifestyle factors, physicochemical measurements, or sequencing protocol information, to report direct relationships between OTUs and MVs, as well as to further reduce spurious associations. If confounding MVs are unmeasured, a specialized mode (FlashWeaveHE) additionally reduces spurious associations driven by structural zeros (Figure 3A).

Increased Prediction Performance on a Variety of Synthetic Datasets

Since experimentally verified biological interactions between microbes are scarcely available, we initially employed previously published frameworks that generate synthetic data with ecological structure. We compared the quality of networks inferred by two different operating modes of FlashWeave—"sensitive" (-S) and "fast" (-F) (Figure 1C) —to three widely used univariate inference methods (SparCC [Friedman and Alm, 2012] (using a more recent re-implementation from the SpiecEasi package [Kurtz et al., 2015]), eLSA [Xia et al., 2011], and CoNet [Faust and Raes, 2016]) and three conditional methods (mLDM [Yang et al., 2017] and SpiecEasi [Kurtz et al., 2015], the latter with neighborhood selection [Meinshausen-Bühlmann algorithm, MB] and inverse covariance selection [graphical Lasso, GL]).

The first group of benchmark data sets was generated with a method based on the Normal to Anything (NorTA) approach (Kurtz et al., 2015), which uses real abundance data from sequencing experiments and a custom interaction network as inputs. Synthetic OTU abundances are drawn from a target distribution fitted to the experimental data, while respecting the partial correlations provided by the input network. In order to simulate noise introduced by the DNA extraction and sequencing steps, we additionally downsampled reads in each synthetic sample to random depths (sampled from the input data set; see STAR Methods).

The prediction quality of all methods was evaluated on such synthetic data sets with increasing numbers of samples, fitted to data from the American Gut Project (McDonald et al., 2018). Overall, FlashWeave most accurately reconstructed the input networks as measured by F1 scores of predicted edges (Figure 2C): across topologies, FlashWeave-S achieved a mean F1 score of 0.68, while non-FlashWeave methods ranged from 0.07 (eLSA) to 0.65 (SpiecEasi-MB), resulting in fractions between 10% and 96% compared to FlashWeave-S (mean 59%). FlashWeave-F was generally less predictive than FlashWeave-S (mean F1 score 0.62, mean fraction 62%).

In a second accuracy benchmark ("ecological models"), we used methods presented in (Weiss et al., 2016) to generate abundance tables with a wide range of linear ecological relationships between OTUs, featuring varying degrees of sparsity and compositionality. Across all data set sizes, eLSA achieved the

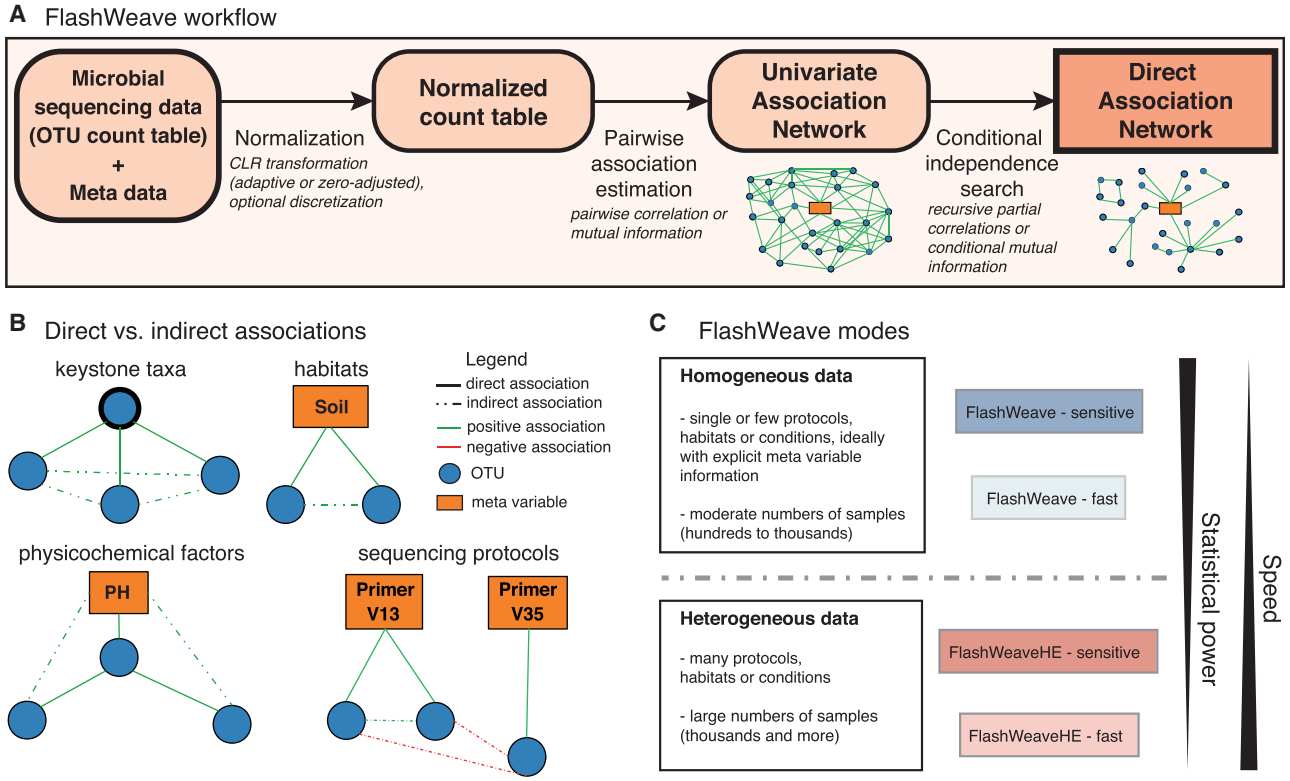


Figure 1. Overview of FlashWeave

(A) Main steps in the network inference pipeline.

(B) Examples of how indirect associations may create false-positive results in various ecological scenarios or for different experimental protocols.

(C) Use cases of the different modes available for FlashWeave.

highest F1 scores (mean 0.76), followed by FlashWeave-S (mean 0.68, 90% of eLSA; Figure 2C). Notably, FlashWeave-S scores were almost identical to eLSA at the highest number of samples (mean F1 score difference <1%). FlashWeave-S and FlashWeave-F showed comparable results (difference <3%), while all other methods achieved mean F1 scores of 2% (SparCC) to 74% (SpiecEasi-MB) relative to FlashWeave-S (mean 62%). In both the NorTA and the ecological models benchmarks, FlashWeave predictions generally improved noticeably with higher sample numbers (up to 141%), indicating efficient usage of additional data. Conversely, when repeating these benchmarks on reduced data set sizes (Figure S1E), FlashWeave was still among the top-performing methods on the ecological models benchmark but dropped in prediction quality on the NorTA benchmark (in particular for scale-free topology and if number of samples $\leq$ 100).

FlashWeaveHE, which specializes in the analysis of heterogeneous data (Figure 1C), was compared to other methods on simulated benchmark data with increased habitat heterogeneity. To this end, we treated the three differently sized data sets for each ecological scenario from the ecological models benchmark as disjoint habitats with no OTU overlaps and aggregated them into a single data set per ecological scenario (see STAR Methods).

FlashWeaveHE-S achieved the highest F1 scores on this benchmark (mean 0.78; Figure 3C), followed by FlashWeaveHE-F with 0.6 and FlashWeave-F with 0.43. The best non-FlashWeave method, SpiecEasi-GL, achieved a mean of 0.25, 68% less than FlashWeaveHE-S. Both FlashWeaveHE modes furthermore obtained almost perfect precision (0.99), while non-FlashWeave methods ranged from 0.0007 (SparCC) to 0.2 (SpiecEasi-GL).

In addition to the noise introduced in the NorTA benchmark, we tested the robustness of all methods to perturbations via repeated rarefactions (down to 2000 reads) on a variety of synthetic and real data sets (Figures 3D and S5A). We found that FlashWeave modes with conditional search were among the less robust methods in this benchmark, but in return, their stable edge predictions showed high fractions of true positives, while unstable edge predictions were almost exclusively false positives. In contrast, other methods generally predicted high fractions of stable but false-positive edges. In addition, unstable edges predicted by FlashWeave typically had small weights compared to stable edges (Figure S5B).

Improved Reconstruction of Literature Interactions in TARA Oceans

In a study of planktonic associations in the TARA Oceans project, the authors presented a list of genus-level interactions described in the literature, based on microscopic and sequencing evidence (Lima-Mendez et al., 2015). This expert-curated set provides a gold standard on which network inference

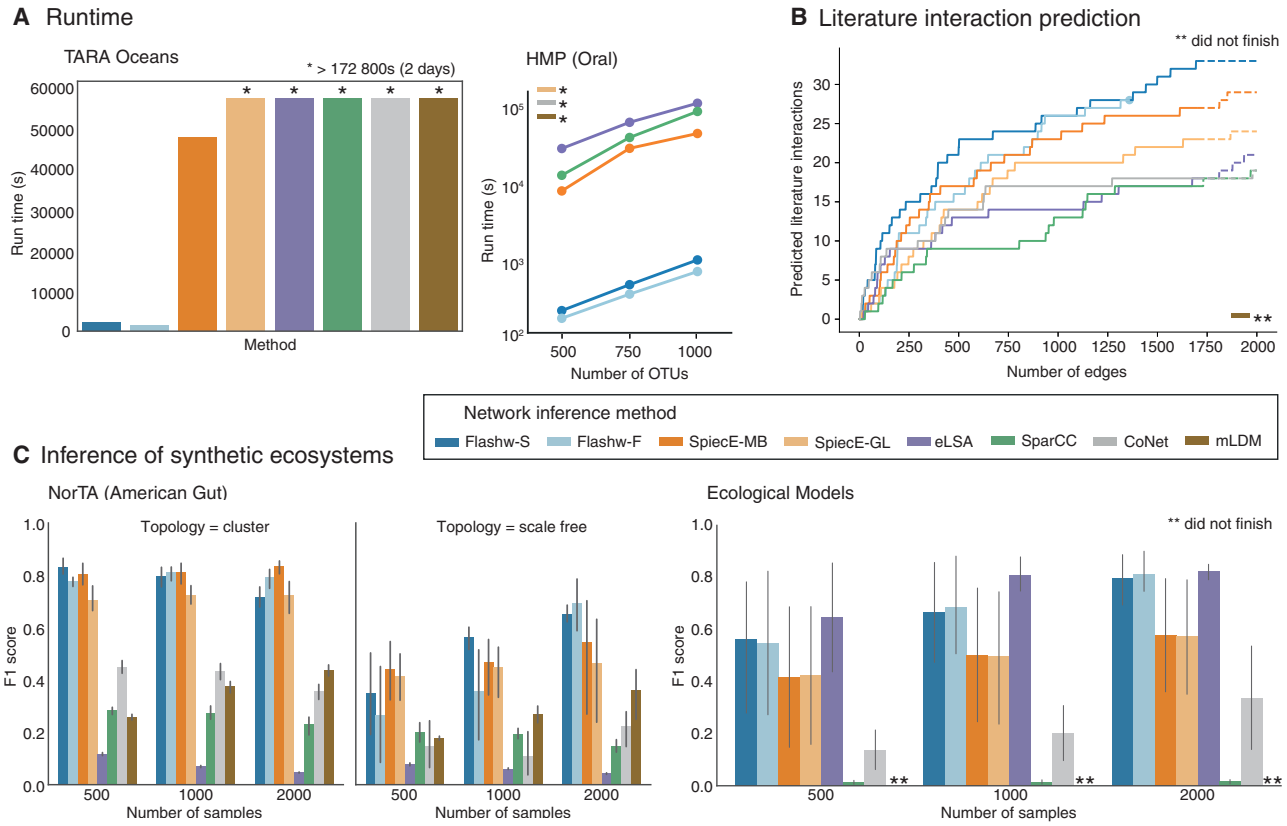


Figure 2. Comparison of FlashWeave to State-of-the-Art Network Inference Methods

Method abbreviations are Flashw-S, FlashWeave-S; Flashw-F, FlashWeave-F; SpiecE-MB, SpiecEasi-MB; SpiecE-GL, SpiecEasi-GL.

(A) Run-time comparison on the TARA Oceans (OTU prevalence >50) and Human Microbiome Project (oral body site only, no OTU prevalence filter) data sets.

(B) Number of gold-standard planktonic interactions in the TARA Oceans data set among the 2,000 edges ranked most highly by each tool. mLDM did not finish computation after 2 weeks.

(C) Prediction performance on data sets with synthetically engineered edges. Data were generated based on Kurtz et al., (2015) and Weiss et al., (2016), and performance was measured as F1 score (harmonic mean of precision and recall). Error bars depict 95% confidence intervals of the mean, based on 1,000 bootstrap replicates.

tools can be tested but is limited to a small fraction of the total marine micro-eukaryotic diversity and likely incomplete. It thus can only be used to benchmark recall on a restricted subset of true positive interactions but yields no information about false positives. Consequently, less precise methods that tend to predict more edges will have an advantage when only raw numbers of true positives are compared since higher false-positive rates of these tools are not considered.

To circumvent this issue and to perform a meaningful benchmark, we compared methods in terms of how highly they ranked literature interactions amongst their 2,000 strongest reported associations (Figure 2B). The underlying assumption was that methods that rank known interactions more highly will generally report more reliable relationships. To make computation feasible for all methods, we reduced the TARA Oceans data set to only OTUs that participate in at least one literature interaction. FlashWeave-S found on average 24% more literature interactions among high-ranking edges than the closest follow-up method (SpiecEasi-MB), 38% more than FlashWeave-F and on average 80% more than other methods. While the TARA Oceans data set shows considerable heterogeneity, FlashWeaveHE was not applicable due to insufficient statistical power (only 22–335 predicted edges total).

Pronounced Runtime Improvements in the Human Microbiome Project and TARA Oceans Datasets

We benchmarked the computational speed of all methods on the Human Microbiome Project (HMP [The Human Microbiome Project Consortium, 2012]) and TARA Oceans (Lima-Mendez et al., 2015) data sets in two settings: homogeneous and heterogeneous. For the homogeneous test, we used 2,500 oral samples from the HMP data set and measured runtime on sets of 500, 750 and 1000 randomly selected OTUs (Figure 2A). FlashWeave outperformed other methods by factors of 8 to 158 on this benchmark (mean: 67), excluding multiple methods (SpiecEasi-GL, CoNet, mLDM) that did not finish after 2 days of computation (factor >339). FlashWeave-S had on average 33% increased runtime over FlashWeave-F.

On the TARA Oceans data set (289 samples, 3,762 OTUs), FlashWeave-F was 29 times faster than the closest non-FlashWeave method (SpiecEasi-MB), while all remaining methods did not finish computation (factor >106; Figure 2A).

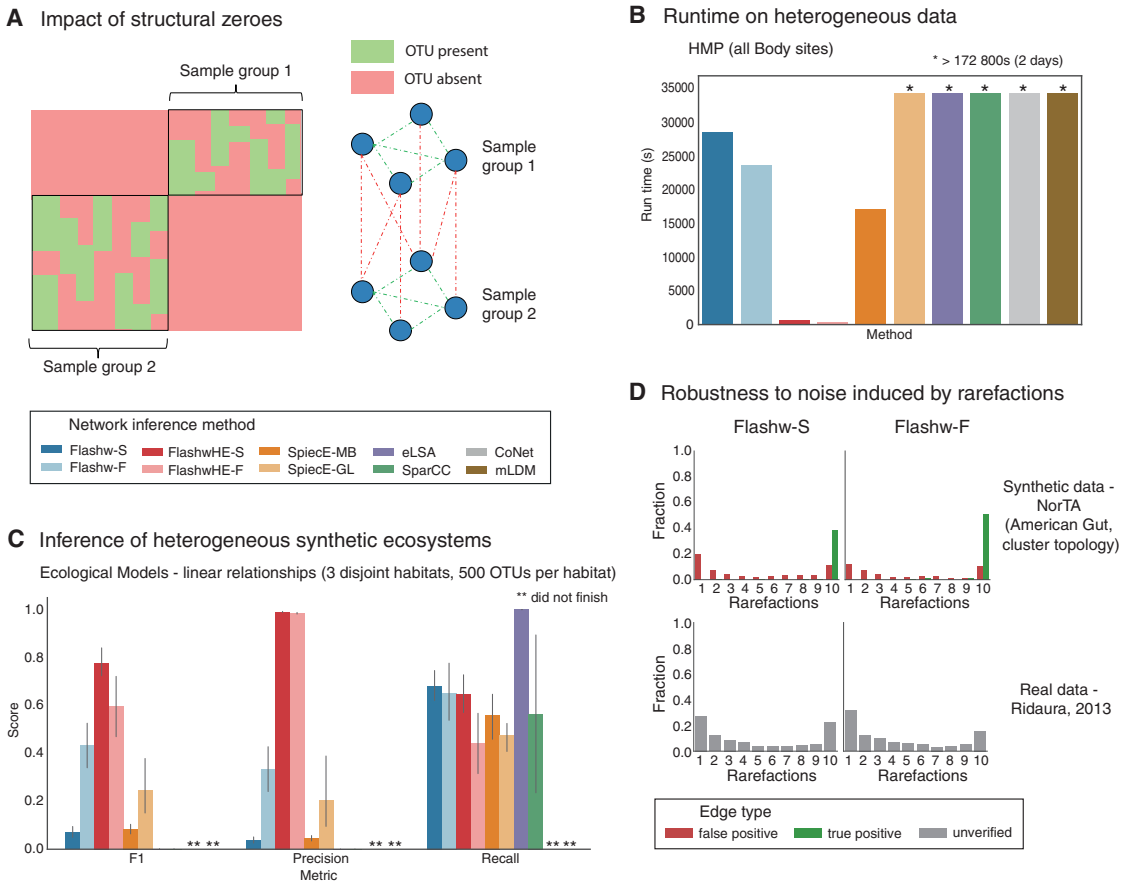


Figure 3. Network Inference Performance on (1) Heterogeneous Datasets with Increased Fractions of Structural Zeros and (2) Robustness to Simulated Noise

Method abbreviations (in addition to those from Figure 2) are FlashwHE-S, FlashWeaveHE-S; FlashwHE-F, FlashWeaveHE-F.

(A) Schematic overview of how structural zeros can lead to false-positive edges. Dashed lines represent indirect associations: positive (green) between OTUs from the same sample group, negative (red) between OTUs from different sample groups.

(B) Run-time comparison on the HMP data set (all body sites) as a representative heterogeneous data set (OTU prevalence >20).

(C) Prediction performance on aggregated disjoint habitats generated by the Ecological Models approach (Weiss et al., 2016), measured using F1 score, recall, and precision. CoNet and mLDM did not finish computation after 2 weeks. Error bars depict 95% confidence intervals of the mean, based on 1,000 bootstrap replicates.

(D) Robustness of predicted edges under noise induced by repeated rarefactions. For each input data set, edges found across *n* rarefactions of this data set are counted toward bin *n* in the respective histograms. For synthetic data sets, information on true positives and false positives in each bin is provided via green and red bars, respectively, while real-world data sets do not have this information (gray bars). See Figure S5A for a comprehensive comparison of all tools and additional data sets.

FlashWeave-S required 53% more runtime than FlashWeave-F on this benchmark.

For the heterogeneous test, we measured the computational speed of FlashWeaveHE and all previous methods on all five body sites from the HMP data set (5514 samples, 1521 OTUs). FlashWeaveHE-F was 51 times faster than the closest non-FlashWeave method (SpiecEasi-MB) in this test and on average 371 times faster than standard FlashWeave (other methods did not finish; factor >518; Figure 3B). FlashWeaveHE-S required 116% more runtime than FlashWeaveHE-F in this benchmark. We also used this data set to test the univariate method fastLSA (Durno et al., 2013), an optimized implementation of the precursor algorithm of eLSA, and found that it ran on average 3.46 times faster than FlashWeaveHE with conditional search but 6.05 times slower than univariate FlashWeave(HE) (Figure S2C). In addition, it produced on average 84% reduced F1 scores on a variety of synthetic benchmarks compared to conditional FlashWeave (Figure S2D).

To test computational scalability in a more demanding setting, we used FlashWeaveHE-F to infer a large-scale ecological network based on 504,647 sequencing samples spanning various habitats and conditions, mapped to 75,516 OTUs at 98% 16S rRNA identity. Inference of the full association network completed after 1d10h46min on a high-performance computing cluster with 200 CPU cores.

MVs Are Central Hubs in the HMP Network, with High Explanatory Power

MVs, such as habitats, conditions (e.g., antibiotic usage), and technical factors (e.g., amplicon or whole-genome shotgun

sequencing), can lead to spurious associations between OTUs associated with the same MV. In addition, direct associations between MVs and OTUs can be interesting when investigating which OTUs are, for instance, directly associated to a particular habitat (independent of microbial interaction partners), prefer certain temperatures, or are affected by specific sequencing biases.

We investigated the importance of MVs in the HMP data set by explicitly providing all five body sites and the two used primer sets (V13 versus V35) as MVs to all FlashWeave modes. MVs formed central hubs in the resulting association networks with, on average, 7.4 times larger neighborhoods than OTUs (Figure S1C) and 27.6 times higher betweenness centrality, a measure of node importance in the network, across all modes.

Furthermore, MVs participated in excluding up to 41.7% indirect OTU-OTU associations (Figure S1B) while constituting only 0.4% of all variables. When MVs were omitted, overall numbers of OTU-OTU associations, however, increased only moderately (up to 12%), suggesting that FlashWeave was generally able to use OTUs highly associated to the omitted MVs to exclude the same indirect associations. Nonetheless, when only comparing associations in direct neighborhoods of MVs, we detected 13%–294% additional OTU-OTU associations when MVs were not provided (Figure S1D), indicating that MV omission may still lead to increased local biases. In addition, we found a weak association between shared primer bias and association probability (mean Pearson's $r < 0.003$, $p < 0.01$), suggesting only limited influence of primer preference on reported associations. This correlation increased marginally when omitting primer information (mean $r < 0.007$, $p < 0.01$). In contrast, the univariate network showed a noticeably stronger association (mean $r < 0.057$, $p < 0.01$), suggesting less robustness to primer biases than observed for direct association networks.

FlashWeaveHE Shows Robustness to Hidden MVs and Structural Zeroes

While the usage of MVs can reduce the number of predicted false-positive associations, information on these variables is frequently not available because not all important latent factors are known, measured, or made available in standardized annotation formats. This particularly affects inherently more heterogeneous cross-study data sets, which can feature diverse experimental, physicochemical, or geographical variables and tend to have less consistent metadata annotations.

One type of artificial associations arise from structural zeroes (Figure 3A), i.e., non-random absences due to unmeasured MVs. Structural zeroes can, for instance, occur when a data set includes multiple habitats with partially exclusive microbial content or multiple sequencing protocols biased toward disjoint OTU sets.

To compare the robustness of different methods to such absences, we computed association networks separately for each method and body site in the HMP data set. We then quantified the overlap of these predicted interactions with a network computed on the aggregated data set of all body sites, restricted to site-specific OTUs (Figure S4D). We found that FlashWeaveHE showed optimal robustness to increased structural zeroes in the cross-site network, with a mean Jaccard overlap between site-specific and cross-site networks of 1.0. In contrast, homogeneous FlashWeave (0.39) and other methods (0.18–0.24) were less robust.

Dependent sample groups constitute another type of hidden MVs, for instance, re-sequencings of the same sample material with different protocols. While such groups can provide important information for network inference, for instance if certain associations can only be detected in specific experimental setups, they also break the independence assumption of common statistical association tests. We tested the impact of dependent sample groups on false-positive predictions with FlashWeave through a set of simulated OTU tables with varying degrees of dependence between samples. As expected, we found that univariate networks produced by FlashWeave included notable numbers of false-positive predictions when dependent samples are highly similar and constitute large fractions of a data set (Figure S2A). However, when computing conditional networks with FlashWeave, numbers of false positives were reduced by a median of 80% for identical samples (zero distance), with particularly strong reductions for FlashWeave-F and FlashWeaveHE-F (95%). Similarly, when increasing inter-sample distance, numbers of false-positive edges in all networks dropped by medians between 89% (distance 0.25) and 99% (distance 0.75).

A Large-Scale Network of Predicted Interactions from Globally Distributed Human Gut Samples Recovers Previously Described Patterns and Generates Hypotheses

We applied FlashWeaveHE to a data set of 69,818 globally distributed human gut samples ("Global Gut," GG) obtained from the NCBI Sequence Read Archive database (SRA [Leinonen et al., 2011]). The data set spanned 488 studies, the majority of which featured less than 1,000 samples (98% of all studies, covering 61% of all samples; Figure S3A). We processed samples uniformly (see STAR Methods) and extracted sequencing protocol information and metadata keywords from SRA annotations, resulting in a final data set of 10,624 OTUs (98% 16S rRNA identity) and 96 MVs.

We used FlashWeaveHE to infer a network of predicted interactions (GGNcond) from GG in 3h53min using 20 CPU cores on an Intel Xeon E7-4870 machine (2.4 GHz). The method identified 30,342 significant associations between OTUs and 13,151 between OTUs and MVs (30%). In contrast, when restricting FlashWeaveHE to compute a univariate network (GGNuni) we observed strongly increased edge density at 1,056,262 edges overall, 96% of which were excluded as indirect in GGNcond. When breaking associations in GG via shuffling (Lima-Mendez et al., 2015), FlashWeaveHE furthermore reported no false-positive associations.

In addition, we found no evidence of dependent sample groups (e.g., niche or batch effects) negatively impacting GGNcond (see STAR Methods).

Analyzing the American Gut Project (AGP (McDonald et al., 2018)) subset of GG (8,897 samples out of 69,818) yielded a 94% decrease in predicted interactions (Figure 4D). For 81% of these, at least one partner was absent in the AGP data set, and these missing partners tended to be rare in GGNcond, with 87% decreased mean prevalence in GG compared to OTUs found in both data sets. Increased network modularity can be a possible indicator of niche effects (Röttjers and Faust,

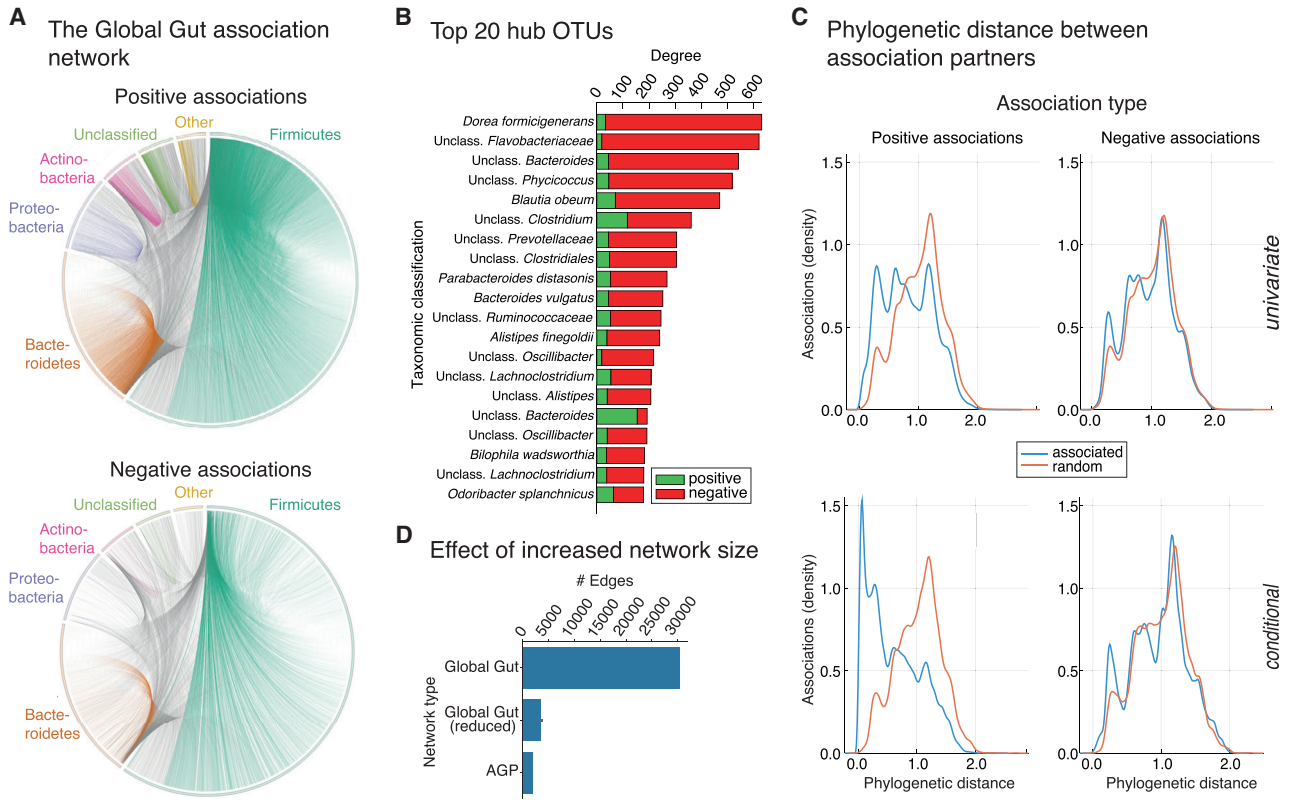


Figure 4. Inference of a Large-Scale, Globally Distributed Human Gut Interaction Network

(A) High-level overview of positive and negative associations in the Global Gut network (GGNcond) with OTUs grouped by phylum. Direct associations within the same phylum bear that phylum's color, between-phylum edges are shown in gray.

(B) Top 20 OTUs with the highest number of direct association partners. OTUs labeled "Unclass." were not confidently classifiable at species level.

(C) Phylogenetic assortativity patterns for positive and negative direct associations. Phylogenetic distance is the summed branch length between association partners on a tree of 92,659 OTU representatives (98% 16S rRNA identity). The top panel depicts distributions from the univariate (GGNuni) and the lower panel from the conditional Global Gut network (GGNcond).

(D) Comparison of the number of edges between GGNcond, networks of 5 random sample subsets of the GG data set with size equal to the American Gut Project subset and the American Gut Project network. The error bar depicts the standard error, deviating from the mean.

2018), and when re-computing the AGP network with vanilla FlashWeave-F, we indeed found its modularity to be elevated compared to the AGP network inferred with FlashWeaveHE-F (47.6% increase; Figure S3E).

The OTU-OTU sub-network of GGNcond was strongly structured (modularity 0.25), indicating the presence of distinct communities. The 20 largest clusters had on average 45 members (up to 89) and featured almost exclusively positive associations between members (mean 99.6%) but only 37.1%–79.8% (mean 63.3%) positive edges to non-member OTUs. Similarly, we found the majority of positive associations per phylum to be within-phylum edges (50% in *Actinobacteria* and up to 87% in *Firmicutes*, mean 68%), while negative associations frequently featured partners from other phyla (35% in *Firmicutes* up to 95% in *Actinobacteria*, mean 73%). For *Actinobacteria*, which had the highest fraction of negative edges to other phyla, the majority targeted *Firmicutes* (48%) and *Bacteroides* (35%).

Many negative associations in GGNcond were mediated by a few dominant OTUs (Figures 4A and 4B), which constituted negative hubs not explainable by our set of MVs (see STAR Methods). These include several species implied in inflammation and disease (*Dorea formicigenerans* [Guinane and Cotter, 2013], *Bilophila wadsworthia* [Feng et al., 2017], *Odoribacter splanchnicus* [Werner et al., 1975], and *Bacteroides vulgatus* [Ó Cuív et al., 2017]). Additionally, we found negative associations between multiple *Blautia* OTUs and a *Clostridium difficile* OTU, consistent with previous reports (Daquigan et al., 2017; Stein et al., 2013).

PA, i.e., the increased probability of association between evolutionarily less diverged partners, is a frequently observed ecological pattern of potential biological interest (Chaffron et al., 2010; Faust et al., 2012; Kurtz et al., 2015). We found notable PA in GGNcond for positive edges, while negative edges were closer to the empirical null distribution (Figure 4C, lower row). Though differences were significant in both cases (two-sample Kolmogorov-Smirnov test, $p < 0.01$), effect size was increased by 10× for positive edges. In contrast, positive edges in GGNuni showed a noticeably smaller effect size increase over negative edges (3.7 × increase, Figure 4C, upper row).

Among OTUs with the highest numbers of positive neighbors (Figure S3D), constituting potential candidates for keystone species (Berry and Widder, 2014), we observed several OTUs from *Bacteroides* (genus) and numerous *Clostridiales* (order) OTUs,

both taxa known to harbor important mutualist species in the human gut (Fischbach and Sonnenburg, 2011; Lopetuso et al., 2013). Intriguingly, 75% of the top 20 positive hubs were taxonomically uncharacterized at the genus level.

Consistent with known dependencies between H_2 producing and consuming microbes which have been described in the human gut (Carbonero et al., 2012), we found significantly more positive associations between H_2 producers and consumers in GGNcond than in random networks, accounting for PA as a possible confounder (3.6× increase, empirical $p < 0.01$). This effect was noticeably weaker for GGNuni (1.8× increase, empirical $p < 0.01$).

DISCUSSION

In this work, we showed that FlashWeave combines (1) the prediction of direct interactions, (2) the ability to scale to large-scale data sets with tens of thousands of OTUs and hundreds of thousands of samples and (3) the incorporation of MV information. In our benchmarks, FlashWeave not only achieved speed improvements of several orders of magnitude and increased network quality compared to other methods but performed particularly well on heterogeneous sequencing data. The latter is crucial since modern large-scale data sets cover diverse studies featuring various habitats, conditions, and protocols. This improved performance was in part achieved by accounting for measured MVs: exemplified by primers and body sites in the HMP data set, we observed that omission of MVs resulted in noticeable increases of edge density between OTUs directly associated with these variables, analogous to spurious edges induced between neighbors of keystone taxa (Berry and Widder, 2014). However, when considering the HMP network as a whole, we found FlashWeave's predictions to be nonetheless remarkably robust to missing MV information, likely mediated by the usage of MV-associated OTUs as placeholders. This finding is further supported by FlashWeave's robustness to dependent sample groups, which we observed both in simulations and in the GG data set. However, the prospects and limits of this effect require further investigation in future studies. We also found MVs to be interesting in their own right: in our HMP analysis, they constituted central nodes in the association network with many directly associated OTUs, in line with the expected habitat preference of many microbes (The Human Microbiome Project Consortium 2012) and known primer biases (Tremblay et al., 2015). Consistent with these results, closely related approaches have previously identified parsimonious sets of predictive microbial biomarkers for human body sites and a skin disease (Tackmann et al., 2018; Statnikov et al., 2013).

We furthermore demonstrated that FlashWeaveHE achieved improved consistency, edge accuracy, and runtime compared to other methods on highly heterogeneous data sets with substantial fractions of structural zeros, which would normally hamper the interpretability of inferred networks (Röttjers and Faust, 2018). Applying this approach to an aggregated cross-study data set with tens of thousands of human gut samples highlighted the advantages of increased statistical power by unveiling a hypothetical, extensive interaction landscape in the hitherto underexplored rare microbial biosphere (Yang et al., 2017; Jousset et al., 2017). Reassuringly, the network also recovered expected biological patterns, which were particularly pronounced after removing large fractions of indirect associations, mirroring results from our synthetic benchmarks and in line with previous work (Kurtz et al., 2015; Yang et al., 2017; Röttjers and Faust, 2018). As one example, we found PA to be stronger for direct than for indirect associations. Niche effects could not easily explain this pattern since these are expected to be more prevalent among indirect associations, as seen for instance in our synthetic benchmarks. A possible, speculative explanation is therefore that kin selection (Strassmann et al., 2011), as previously observed for example in biofilms (Xavier and Foster, 2007) or iron acquisition (Griffin et al., 2004), may be more pronounced in the human gut than currently appreciated. However, despite extensive checks, we could not entirely rule out shared-niche contributions in our network; future confirmatory investigation of this finding is therefore necessary. In addition, the network predicted many strong positive hubs in the gastrointestinal tract to be OTUs only classifiable at higher taxonomic ranks. Several of these were assigned to the families *Lachnospiraceae* and *Ruminococcaceae* (order *Clostridiales*), which indicated that the positive role of unclassified OTUs from these families on ecosystem maintenance may be considerable (Lopetuso et al., 2013).

Current limitations of FlashWeave include its conservative handling of structural zeros, which can result in reduced statistical power and may thus in particular hamper analysis of data sets with fewer samples. This effect was, however, mitigated for larger sample sizes in our synthetic benchmarks, and since cross-study data sets tend to include even more samples, we do not expect power issues to strongly affect typical use cases. More refined methods that assign confidences to absences would nonetheless be an interesting addition to future versions of FlashWeaveHE. Similarly, FlashWeave's prediction quality dropped also in a subset of our synthetic benchmarks on homogeneous data with smaller sample numbers, indicating that caution is currently advised when applying FlashWeave to small-scale studies. Slower but more powerful parametric tests may reduce this problem in future versions of FlashWeave. While FlashWeave handled simulated sequencing noise well, our tests furthermore showed that strong generic perturbations can induce the prediction of typically weak, unstable edges. For particularly noisy data sets, it may therefore be prudent to remove edges with small weights or explicitly include weights into downstream analyses. Interestingly, we observed a selective enrichment of true positives among FlashWeave's stable edge predictions compared to other tools, which opens up interesting avenues for controlled perturbation (e.g., via bootstrapping) to further increase general precision in future versions (albeit likely at the cost of sensitivity).

The LGL framework, which FlashWeave builds upon, permits several straightforward extensions, such as more powerful tests (Xu et al., 2015; Lovell et al., 2015) and the prediction of edge directionality (Aliferis et al., 2010a). The latter is an exciting prospect that would enable a more causal interpretation of predicted ecological interactions, paving the path toward efficiently learning fully predictive models. In the future, such data-driven models may allow us to forecast the ecological impact of perturbations and catalyze emerging ecological engineering applications.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- LEAD CONTACT AND MATERIALS AVAILABILITY
- METHOD DETAILS
  - FlashWeave
  - Accuracy and Robustness Benchmarks
  - Computational Speed Benchmarks
  - Literature Interaction Predictions
  - Meta Variable Analysis in the HMP
  - Global Gut Network Analysis
  - Normalization Comparison
- QUANTIFICATION AND STATISTICAL ANALYSIS
- DATA AND CODE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cels.2019.08.002.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their insightful comments and Jeroen Raes, Shinichi Sunagawa, Reinhard Furrer, and Thomas Sebastian Benedikt Schmidt for their valuable methodological feedback. We furthermore thank Marija Dmitrieva for helpful discussions during the preparation of this manuscript. This work was supported by the Swiss National Science Foundation (grant no. 31003A-160095).

AUTHOR CONTRIBUTIONS

Conceptualization, J.T., J.F.M.R., and C.v.M.; Methodology, J.T. and J.F.M.R.; Software, J.T.; Investigation, J.T., J.F.M.R., and C.v.M.; Writing – Original Draft, J.T.; Writing – Review & Editing, J.T., J.F.M.R., and C.v.M.; Visualization, J.T.; Funding Acquisition, C.v.M.; Supervision, J.F.M.R. and C.v.M.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: January 4, 2019
Revised: May 16, 2019
Accepted: July 31, 2019
Published: September 18, 2019

REFERENCES

Aitchison, J. (1981). A new approach to null correlations of proportions. J. Int. Assoc. Math. Geol. *13*, 175–189.

Aliferis, C.F., Statnikov, A., Tsamardinos, I., Mani, S., Koutsoukos, X.D., et al. (2010a). Local causal and markov blanket induction for causal discovery and feature selection for classification part i: algorithms and empirical evaluation. J. Mach. Learn. Res. *11*, 171–234.

Aliferis, C.F., Statnikov, A., Tsamardinos, I., Mani, S., Koutsoukos, X.D., et al. (2010b). Local causal and markov blanket induction for causal discovery and feature selection for classification part ii: analysis and extensions. J. Mach. Learn. Res. *11*, 235–284.

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

Bezanson, J., Edelman, A., Karpinski, S., and Shah, V.B. (2017). Julia: a fresh approach to numerical computing. SIAM Rev. *59*, 65–98.

Biswas, S., Mcdonald, M., Lundberg, D.S., Dangl, J.L., and Jojic, V. (2015). Learning microbial interaction networks from metagenomic count data. Lect. Notes Comput. Sci. 32–43.

Bonder, M.J., Kurilshikov, A., Tigchelaar, E.F., Mujagic, Z., Imhann, F., Vila, A.V., Deelen, P., Vatanen, T., Schirmer, M., Smeekens, S.P., et al. (2016). The effect of host genetics on the gut microbiome. Nat. Genet. *48*, 1407–1412.

Carabotti, M., Scirocco, A., Maselli, M.A., and Severi, C. (2015). The gut-brain axis: interactions between enteric microbiota, central and enteric nervous systems. Ann. Gastroenterol. Hepatol. *28*, 203–209.

Carbonero, F., Benefiel, A.C., and Gaskins, H.R. (2012). Contributions of the microbial hydrogen economy to colonic homeostasis. Nat. Rev. Gastroenterol. Hepatol. *9*, 504–518.

Cazelles, K., Araújo, M.B., Mouquet, N., and Gravel, D. (2016). A theory for species co-occurrence in interaction networks. Theor. Ecol. *9*, 39–48.

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

Daquigan, N., Seekatz, A.M., Greathouse, K.L., Young, V.B., and White, J.R. (2017). High-resolution profiling of the gut microbiome reveals the extent of Clostridium difficile burden. NPJ Biofilms Microbiomes *3*, 35.

Duda, S., Aliferis, C., Miller, R., Statnikov, A., and Johnson, K. (2005). Extracting drug-drug interaction articles from Medline to improve the content of drug databases. AMIA Annu. Symp. Proc. 216–220.

Durno, W.E., Hanson, N.W., Konwar, K.M., and Hallam, S.J. (2013). Expanding the boundaries of local similarity analysis. BMC Genomics *14*, S3.

Dyhrman, S., Ammerman, J., and Van Mooy, B. (2007). Microbes and the marine phosphorus cycle. Oceanography *20*, 110–116.

Falkowski, P.G., Fenchel, T., and Delong, E.F. (2008). The microbial engines that drive Earth's biogeochemical cycles. Science *320*, 1034–1039.

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

Faust, K., and Raes, J. (2016). CoNet app: inference of biological association networks using cytoscape. F1000Res. *5*, 1519.

Faust, K., Sathirapongsasuti, J.F., Izard, J., Segata, N., Gevers, D., Raes, J., and Huttenhower, C. (2012). Microbial co-occurrence relationships in the human microbiome. PLoS Comput. Biol. *8*, e1002606.

Feng, Z., Long, W., Hao, B., Ding, D., Ma, X., Zhao, L., and Pang, X. (2017). A human stool-derived Bilophila wadsworthia strain caused systemic inflammation in specific-pathogen-free mice. Gut Pathog. *9*, 59.

Fischbach, M.A., and Sonnenburg, J.L. (2011). Eating for two: how metabolism establishes interspecies interactions in the gut. Cell Host Microbe *10*, 336–347.

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

Goers, L., Freemont, P., and Polizzi, K.M. (2014). Co-culture systems and technologies: taking synthetic biology to the next level. J. R. Soc. Interface *11*.

Griffin, A.S., West, S.A., and Buckling, A. (2004). Cooperation and competition in pathogenic bacteria. Nature *430*, 1024–1027.

Guinane, C.M., and Cotter, P.D. (2013). Role of the gut microbiota in health and chronic gastrointestinal disease: understanding a hidden metabolic organ. Ther. Adv. Gastroenterol. *6*, 295–308.

Jousset, A., Bienhold, C., Chatzinotas, A., Gallien, L., Gobet, A., Kurm, V., Küsel, K., Rillig, M.C., Rivett, D.W., Salles, J.F., et al. (2017). Where less may be more: how the rare biosphere pulls ecosystems strings. ISME J. *11*, 853–862.

JuliaStats (2018a). Distances.jl, a Julia package for evaluating distances (metrics) between vectors. https://github.com/JuliaStats/Distances.jl.

JuliaStats (2018b). Distributions.jl: a Julia package for probability distributions and associated functions. https://github.com/JuliaStats/Distributions.jl.

JuliaStats (2018c). KernelDensity.jl, Kernel density estimators for Julia. https://github.com/JuliaStats/KernelDensity.jl.

Kawaguchi, M., and Minamisawa, K. (2010). Plant-microbe communications for symbiosis. Plant Cell Physiol. *51*, 1377–1380.

Krause, E., Wichels, A., Giménez, L., Lunau, M., Schilhabel, M.B., and Gerdts, G. (2012). Small changes in pH have direct effects on marine bacterial community composition: a microcosm approach. PLoS One *7*, e47035.

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.

Leinonen, R., Sugawara, H., and Shumway, M.; International Nucleotide Sequence Database Collaboration (2011). The sequence read archive. Nucleic Acids Res. *39*, D19–D21.

Levy, R., Carr, R., Kreimer, A., Freilich, S., and Borenstein, E. (2015). NetCooperate: a network-based tool for inferring host-microbe and microbe-microbe cooperation. BMC Bioinformatics *16*, 164.

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

Lopetuso, L.R., Scaldaferri, F., Petito, V., and Gasbarrini, A. (2013). Commensal Clostridia: leading players in the maintenance of gut homeostasis. Gut Pathog. *5*, 23.

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.

Matias Rodrigues, J.F., Schmidt, T.S.B., Tackmann, J., and von Mering, C. (2017). MAPseq: highly efficient k-mer search with confidence estimates, for rRNA sequence analysis. Bioinformatics *33*, 3808–3810.

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

McDonald, D., Hyde, E., Debelius, J.W., Morton, J.T., Gonzalez, A., Ackermann, G., Aksenov, A.A., Behsaz, B., Brennan, C., Chen, Y., et al. (2018). American gut: an open platform for citizen science microbiome research. mSystems *3*.

McFall-Ngai, M. (2014). Divining the essence of symbiosis: insights from the squid-vibrio model. PLoS Biol. *12*, e1001783.

Mitchell, A.L., Scheremetjew, M., Denise, H., Potter, S., Tarkowska, A., Qureshi, M., Salazar, G.A., Pesseat, S., Boland, M.A., Hunter, F.M.I., et al. (2018). EBI metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies. Nucleic Acids Res. *46*, D726–D735.

Narendra, V., Lytkin, N.I., Aliferis, C.F., and Statnikov, A. (2011). A comprehensive assessment of methods for de-novo reverse-engineering of genome-scale regulatory networks. Genomics *97*, 7–18.

Nawrocki, E.P., and Eddy, S.R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics *29*, 2933–2935.

Newman, M.E.J. (2006). Modularity and community structure in networks. Proc. Natl. Acad. Sci. USA *103*, 8577–8582.

Ó Cuív, P., de Wouters, T., Giri, R., Mondot, S., Smith, W.J., Blottière, H.M., Begun, J., and Morrison, M. (2017). The gut bacterium and pathobiont Bacteroides vulgatus activates NF-κB in a human gut epithelial cell line in a strain and growth phase dependent manner. Anaerobe *47*, 209–217.

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. *44*, D733–D745.

Oh, D.C., Poulsen, M., Currie, C.R., and Clardy, J. (2009). Dentigerumycin: a bacterial mediator of an ant-fungus symbiosis. Nat. Chem. Biol. *5*, 391–393.

Papageorgiou, H., and Aitchison, J. (1989). The statistical analysis of compositional data. Biometrics *45*, 345.

Pawlowsky-Glahn, V., and Buccianti, A. (2011). Compositional Data Analysis: Theory and Applications (John Wiley & Sons)).

Pearson, K. (1896). Mathematical contributions to the theory of evolution.—on a form of spurious correlation which may arise when indices are used in the measurement of organs. Proc. R. Soc. Lond. A *60*, 489–498.

Pedregosa, F., et al. (2011). Learning scikit-learn: machine Learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS One *5*, e9490.

Ridaura, V.K., Faith, J.J., Rey, F.E., Cheng, J., Duncan, A.E., Kau, A.L., Griffin, N.W., Lombard, V., Henrissat, B., Bain, J.R., et al. (2013). Gut microbiota from twins discordant for obesity modulate metabolism in mice. Science *341*, 1241214.

Röttjers, L., and Faust, K. (2018). From hairballs to hypotheses-biological insights from microbial networks. FEMS Microbiol. Rev. *42*, 761–780.

Sboner, A., and Aliferis, C.F. (2005). Modeling clinical judgment and implicit guideline compliance in the diagnosis of melanomas using machine learning. AMIA Annu. Symp. Proc. 664–668.

Scutari, M. (2017). Bayesian network constraint-based structure learning algorithms: parallel and optimized implementations in the bnlearn R package. J. Stat. Soft. *77*.

Solden, L., Lloyd, K., and Wrighton, K. (2016). The bright side of microbial dark matter: lessons learned from the uncultivated majority. Curr. Opin. Microbiol. *31*, 217–226.

Statnikov, A., Alekseyenko, A.V., Li, Z., Henaff, M., Perez-Perez, G.I., Blaser, M.J., and Aliferis, C.F. (2013). Microbiomic signatures of psoriasis: feasibility and methodology comparison. Sci. Rep. *3*, 2620.

Stein, R.R., Bucci, V., Toussaint, N.C., Buffie, C.G., Rätsch, G., Pamer, E.G., Sander, C., and Xavier, J.B. (2013). Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. PLoS Comput. Biol. *9*, e1003388.

Strassmann, J.E., Gilbert, O.M., and Queller, D.C. (2011). Kin discrimination and cooperation in microbes. Annu. Rev. Microbiol. *65*, 349–367.

Tackmann, J., Arora, N., Schmidt, T.S.B., Rodrigues, J.F.M., and von Mering, C. (2018). Ecologically informed microbial biomarkers and accurate classification of mixed and unmixed samples in an extensive cross-study of human body sites. Microbiome *6*, 192.

Thaiss, C.A., Zmora, N., Levy, M., and Elinav, E. (2016). The microbiome and innate immunity. Nature *535*, 65–74.

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

Tremblay, J., Singh, K., Fern, A., Kirton, E.S., He, S., Woyke, T., Lee, J., Chen, F., Dangl, J.L., and Tringe, S.G. (2015). Primer and platform effects on 16S rRNA tag sequencing. Front. Microbiol. *6*, 771.

Van Dongen, S. (2008). Graph clustering via a discrete uncoupling process. SIAM J. Matrix Anal. Appl. *30*, 121–141.

Vandeputte, D., Kathagen, G., D'hoe, K., Vieira-Silva, S., Valles-Colomer, M., Sabino, J., Wang, J., Tito, R.Y., De Commer, L., Darzi, Y., et al. (2017). Quantitative microbiome profiling links gut community variation to microbial load. Nature *551*, 507–511.

Voortman, M., and Druzdzel, M.J. (2008). Insensitivity of constraint-based causal discovery algorithms to violations of the assumption of multivariate normality. In FLAIRS conference, pp. 690–695.

Weiss, S., Van Treuren, W., Lozupone, C., Faust, K., Friedman, J., Deng, Y., Xia, L.C., Xu, Z.Z., Ursell, L., Alm, E.J., et al. (2016). Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. ISME J. *10*, 1669–1681.

Werner, H., Rintelen, G., and Kunstek-Santos, H. (1975). [A new butyric acid-producing Bacteroides species: B. splanchnicus n. sp. (author's transl). *Zentralblatt fur Bakteriologie*], Parasitenkunde, Infektionskrankheiten und Hygiene. Erste Abt. Orig. Reihe Med. Mikrobiol. Parasitol. *231*, 133–144.

Xavier, J.B. (2011). Social interaction in synthetic and natural microbial communities. Mol. Syst. Biol. *7*, 483.

Xavier, J.B., and Foster, K.R. (2007). Cooperation and conflict in microbial biofilms. Proc. Natl. Acad. Sci. USA *104*, 876–881.

Xia, L.C., Steele, J.A., Cram, J.A., Cardon, Z.G., Simmons, S.L., Vallino, J.J., Fuhrman, J.A., and Sun, F. (2011). Extended local similarity analysis (eLSA)

of microbial community and other time series data with replicates. BMC Syst. Biol. *5*, S15.

Xu, L., Paterson, A.D., Turpin, W., and Xu, W. (2015). Assessment and selection of competing models for zero-inflated microbiome data. PLoS One *10*, e0129606.

Yang, Y., Chen, N., and Chen, T. (2017). Inference of environmental factor-microbe and microbe-microbe associations from metagenomic data using a hierarchical bayesian statistical model. Cell Syst *4*, 129–137.e5.

Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F.O., Ludwig, W., Schleifer, K.H., Whitman, W.B., Euzéby, J., Amann, R., and Rosselló-Móra, R. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat. Rev. Microbiol. *12*, 635–645.

Yu, K., Liu, L., Li, J., and Chen, H. (2018). Mining markov blankets Without causal sufficiency. IEEE Trans. Neural Netw. Learn. Syst. *29*, 6333–6347.

Zelezniak, A., Andrejev, S., Ponomarova, O., Mende, D.R., Bork, P., and Patil, K.R. (2015). Metabolic dependencies drive species co-occurrence in diverse microbial communities. Proc. Natl. Acad. Sci. USA *112*, 6449–6454.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Raw microbial sequencing data (whole genome shotgun, amplicon, RNA-Seq) from the human gastrointestinal tract	NCBI Sequence Read Archive database (Leinonen et al., 2011)	See Table S2
Pre-processed, rarefied OTU tables based on the study "Cultured gut bacterial consortia from twins discordant for obesity modulate adiposity and metabolic phenotypes in gnotobiotic mice"	Weiss et al., 2016; Ridaura et al., 2013	N/A
Raw microbial sequencing data for the Human Microbiome Project	NCBI Sequence Read Archive database (Leinonen et al., 2011)	SRP002395
Pre-processed OTU tables of planktonic marine microorganisms from the TARA Oceans expedition	Lima-Mendez et al., 2015	N/A
Software and Algorithms		
SpiecEasi	Kurtz et al., 2015)	https://github.com/zdk123/SpiecEasi
eLSA	Xia et al., 2011	https://bitbucket.org/charade/elsa/wiki/Home
SparCC	Friedman and Alm, 2012	https://github.com/zdk123/SpiecEasi
CoNet	Faust and Raes, 2016	http://raeslab.org/software/conet.html
mLDM	Yang et al., 2017	https://github.com/tinglab/mLDM
fastLSA	Durno et al., 2013	http://hallam.microbiology.ubc.ca/fastLSA/install/index.html
FlashWeave	This paper	https://github.com/meringlab/FlashWeave.jl
MAPseq	Matias Rodrigues et al., 2017	https://www.meringlab.org/software/mapseq/
HPC-Clust	Matias Rodrigues and von Mering, 2014	https://www.meringlab.org/software/hpc-clust/
INFERNAL	Nawrocki and Eddy, 2013	http://eddylab.org/infernal/
fasttree	Price et al., 2010	http://www.microbesonline.org/fasttree/
sklearn	Pedregosa et al., 2011	https://scikit-learn.org/stable/install.html
Distances.jl	JuliaStats, 2018a	https://github.com/JuliaStats/Distances.jl
Distributions.jl	JuliaStats, 2018b	https://github.com/JuliaStats/Distributions.jl
KernelDensity.jl	JuliaStats, 2018c	https://github.com/JuliaStats/KernelDensity.jl

LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Christian von Mering (mering@imls.uzh.ch).

METHOD DETAILS

FlashWeave

Algorithmic Details

FlashWeave is implemented in the Julia Programming Language (Bezanson et al., 2017) and based on the local-to-global learning framework (LGL (Aliferis et al., 2010a)). Causal inference algorithms of this class start by performing a locally optimal neighborhood

search in order to infer all directly associated neighbors of a target variable *T* (OTU or MV in the case of FlashWeave), which represent the set of estimated direct causes and effects of *T* (the so-called Markov blanket (MB) of *T*). Subsequently, individual neighborhoods are connected through a combinator strategy to form a global association graph. In FlashWeave, this step is performed by the "OR" strategy, which creates a link between nodes A and B in the global graph if either A is in the inferred directly associated neighborhood (MB) of B or vice versa. In the final step, currently not implemented in FlashWeave, this undirected skeleton of conditional dependence relationships can be used as a scaffold to efficiently infer edge directionality and provide further insights into the system of study.

LGL can be instantiated with a wide range of algorithms (variations of the steps above) and conditional independence tests. FlashWeave currently defaults to a modified version of the efficient semi-interleaved HITON-PC algorithm (si-HITON-PC, (Aliferis et al., 2010a)). The local neighborhood search step of si-HITON-PC starts by computing, for each target variable *T*, its univariate associations to all other variables. If the user requested a univariate network, these neighborhoods are then connected to a global association graph (see previous paragraph) and the resulting network is returned. Otherwise, the algorithm proceeds with conditioning search, for which all significantly univariately associated variables are labeled as candidates for neighborhood inclusion. As part of the inclusion heuristic, these candidates are then sorted according to association strength (starting with the strongest association), yielding the set *OPEN*. Subsequently, the *interleaving phase* takes place, in which the first candidate from *OPEN* is automatically included in a tentative set of directly associated neighbors (*TPC*), while all further candidates are systematically tested for inclusion into *TPC*. These tests proceed by sequentially probing the association between *T* and the next candidate *X*, conditioned on all subsets of previously accepted members of *TPC* only (which constitutes the second part of the inclusion heuristic). If any subset of *TPC* results in an insignificant test, *X* is rejected and not further considered. After all candidates have been checked in this manner, the *elimination phase* begins. In this step, all associations between *T* and members of *TPC* are again checked for conditional independence by performing the tests that were skipped in the interleaving phase. For instance, the first candidate, which was automatically included during interleaving phase, now needs to undergo conditional independence tests on combinations of all remaining members. The result of the elimination phase is a final set of directly associated neighbors (*PC*), which includes only variables that are conditionally dependent on *T* given any other subset of neighbors. Individual neighborhoods can then be connected to a global association graph (see previous paragraph).

Since the number of necessary tests grows exponentially with the size of *TPC*, the inclusion heuristic is essential to efficiently control the number of considered candidates and avoid large numbers of unnecessary tests (provided the neighborhood is sufficiently sparse). This heuristic additionally helps to keep the number of variables in the conditioning sets of individual tests small, which improves reliability and statistical power. Specific additional heuristics, such as the *max-k* heuristic, furthermore accelerate inference by putting an upper bound on the maximum size of conditioning sets, which is crucial for variables with large direct neighborhoods (e.g. hub OTUs), as the number of tests can otherwise become infeasible for such variables. In line with results in (Aliferis et al., 2010a; Aliferis et al., 2010b), FlashWeave also employs False Discovery Rate (FDR) adjustment and omits the costly steps of spouse identification and symmetry correction. See (Aliferis et al., 2010a; Aliferis et al., 2010b) for further details on the si-HITON-PC algorithm, the LGL framework, as well as the previously discussed heuristics. On top of these algorithmic shortcuts, FlashWeave introduces a number of heuristics, which in particular accelerate the inference of dense neighborhoods (e.g. of hub OTUs, typical for microbial association networks) and are described in subsection "Heuristics".

While pivotal for computational efficiency, the heuristics applied by FlashWeave come with a potential trade-off: in contrast to other graphical model approaches (such as the methods employed by SpiecEasi and mLDM), conditional independence between two variables is not tested based on all remaining variables simultaneously, but only for heuristically selected subsets of likely informative variables. Since microbial association networks are however typically sparse (as also assumed by some other methods, e.g. SparCC and SpiecEasi), the majority of candidates is generally uninformative and can be readily discarded without measurably hampering prediction performance (as observed in our benchmarks). On the contrary, focusing on local conditional dependencies has positive effects on network quality in our comparisons (see Figure 2C), likely due to increased statistical power and reliability when performing tests conditioned only on informative candidates. These localized tests furthermore allow FlashWeave to avoid costly regularization and parameter optimization steps, which are required by other PGM approaches to address rank deficiency issues.

Like most constraint-based causal inference algorithms, the LGL framework assumes *causal sufficiency*, i.e. that no unmeasured variables exist that directly causally influence more than one variable in the system (Aliferis et al., 2010a). Since such hidden variables cannot be included in conditioning sets, the graph inferred from such a system may contain spurious links between variables that are directly affected by the same hidden variable. The local violation of causal sufficiency has however been conjectured to not compromise algorithm soundness and predictive performance in many practical scenarios (Aliferis et al., 2010b). In agreement with this conjecture, an instantiation of LGL (MMHC, closely related to the si-HITON-PC algorithm) was among the best performing methods in an in-depth benchmark of hidden variable impact (Yu et al., 2018). In the same vein, we empirically found that the number of edges increases only moderately when omitting meta variables from the HMP data set (by up to 12%), even though these meta variables are particularly informative and central in the network (see Results). A second central assumption of LGL algorithms (*faithfulness*, see (Aliferis et al., 2010a)) only concerns the directionality inference step, which is currently not implemented in FlashWeave. It thus does not apply here, but may be discussed in future studies.

In terms of statistical tests, FlashWeave provides the choice of either discretized mutual information tests (more coarse grained and usually quicker; "fast" mode) or partial correlation tests with Fisher's z-transformation (more sensitive and usually slower; ''sensitive'' mode) (Figures 1C and S1A). Before each statistical test, its reliability is estimated and only tests with sufficient sample size are

performed. For mutual information tests, FlashWeave applies the widely applied rule-of-thumb that cells in the corresponding contingency table should have more than *hps* counts on average (controllable by the user), while partial correlation tests currently default to a fixed minimum number of observations (independent of conditioning set size). If a test is estimated to not be reliable, FlashWeave conservatively assumes *T* and *X* to be conditionally independent (no association) in order to avoid false positives and to improve runtime. This is in contrast to the original si-HITON-PC algorithm, which more liberally assumes dependence in such cases. While the conditional mutual information test makes no distributional assumptions, the partial correlation test assumes OTUs to be multivariate Gaussian distributed in clr-transformed space. Partial correlation tests were, however, shown to be robust to diverse violations of the normality assumption in the context of constraint-based graphical model inference (Voortman and Druzdzel, 2008), the conceptual framework also used by FlashWeave. These results are further confirmed by the strong performance of FlashWeave-S and FlashWeaveHE-S in our benchmarks on synthetic data with non-normal OTU abundances (zero-inflated negative binomial for the NorTA benchmark and log-normal for the Ecological Models benchmark; Figures 2C and 3C).

FlashWeaveHE further specializes mutual information and partial correlation tests to exclude zero elements from association computations (Figures 1C and S1A). It makes the assumption that zeroes in large, heterogeneous data sets are mostly structural (for instance due to primer or sequencing depth biases, as well as habitat or condition-specific effects) and thereby only considers samples in which both OTUs have a non-zero abundance as reliable for association prediction. Notably, this restriction is only applied to the prospective association partners being tested: OTUs found in the conditioning set retain their absences. This procedure is chosen i) to not discard too much information concerning the tested partners, which would otherwise result in drastic loss of power as conditioning sets grow, and ii) because structural absences of OTUs within conditioning sets, despite marginally decreasing power, otherwise have minimal impact on exclusion decisions. While the FlashWeaveHE approach potentially discards some valid absence information and can thereby be less sensitive than the vanilla mode of FlashWeave, we found this loss in sensitivity to be small in heterogeneous data sets with larger sample sizes. Indeed, FlashWeaveHE resulted in strongly improved precision (Figures 3C and S4D) and runtimes (Figure 3B) on such data.

Normalization in FlashWeave accounts for compositionality effects and differs depending on the test type. Details on normalization schemes can be found in subsection "Normalization".

Heuristics

Learning the direct neighborhood of target variable *T* with the si-HITON-PC algorithm has a run time complexity of $O(|V|2^{|PC(T)|})$ (Aliferis et al., 2010a), where *V* are variables in the system and *PC*(*T*) is the Parent-Children set of *T*, i.e. the set of its directly associated neighbors (or the Markov blanket *MB*(*T*) minus spouses (Aliferis et al., 2010a)). Runtime thus depends linearly on the number of variables and exponentially on the size of the direct neighborhood.

FlashWeave implements all options and algorithmic shortcuts suggested by Aliferis et al. (Aliferis et al., 2010a) (max-k heuristic, h-ps reliability criterion, FDR correction, optimal variable ordering). In order to achieve the speed reported in this study, we furthermore extended the original algorithm through a number of additional heuristics, explained in more detail below.

The first algorithmic shortcut introduced in FlashWeave is the *feedforward* heuristic, which constitutes a parallel variation of traditional backtracking (Scutari, 2017). The key observation utilized by this heuristic is that the size of individual neighborhoods can vary substantially in networks with scale-free node degree distributions, such as microbial co-occurrence networks ((Faust and Raes, 2012) and citations therein), with exponential impact on runtime (see above). From the scale-free property follows that keystone species *A* (with many dependent neighbors) will typically have a large number of neighbors *B* that themselves are not keystone species (few neighbors) and whose neighborhoods are thus exponentially quicker to compute. Now, for the edge $A \rightarrow B$ to be included in the final, global network through the OR combinator rule, it is sufficient if the considerably cheaper reverse link $B \rightarrow A$ is proven. *feedforward* exploits this property by prioritizing computation of variables expected to have smaller neighborhoods (as approximated by their univariate neighborhood size) and relaying the information of detected direct links to computationally more intensive variables (larger neighborhoods). If during the computation of the neighborhood of *A* the next variable *B* to be tested was already shown to be a neighbor, it automatically enters the set of neighbor candidates of *A* without formally performing all tests. *feedforward* is applied in a parallel computation setting, where candidate lists of all nodes are periodically updated with the latest information from other neighborhoods as it becomes available, allowing the most expensive nodes to leverage a maximum amount of information to cut down runtime.

The second computational shortcut introduced in FlashWeave we term *fast-elimination* heuristic. A large amount of time can be spent in the final elimination phase of si-HITON-PC, in which all previously skipped tests between candidates passing the interleaving phase are performed. In the original algorithm, even if a variable *S* is discarded during the elimination phase, it will still be included in future conditioning sets, thereby inflating the number of conducted conditional independence tests. If many variables are discarded during earlier parts of the elimination phase, but still included in subsequent conditioning sets, the result can be an exponential increase in necessary tests, making the elimination phase particularly costly. *fast-elimination* addresses this computational hurdle by not considering a removed variable *S* for any subsequent conditioning sets during elimination phase. An intuitive motivation for this approach is that, if a variable *S* was shown earlier to not be part of the neighborhood of *T*, it should also not be required to render further candidates independent of *T* as it's not part of its Parent-Children set.

As another shortcut, we implemented a convergence criterion that periodically checks whether links in the network still show substantial change over time. If the network has reached convergence, all remaining candidates are assumed to be conditionally independent of their target variables. This criterion is based on our observation that the naive algorithm can stall on single nodes with large neighborhoods due to the exponential runtime dependency of si-HITON-PC on neighborhood size. However, candidates still to be

checked at this point tend to be weak, since they i) appear late in the relevance-sorted candidate list and ii) have been proven to not be neighbors in the reverse direction (otherwise the *feedforward* heuristic would have applied). The majority of these late links are typically discarded after substantial computational effort, with minimal effect on network structure. While using this type of convergence threshold may in theory lead to biased edge omissions since it selectively bypasses computation of candidates in large neighborhoods, we didn't detect meaningful biases of this kind in the networks we tested.

As a final option to improve runtime, FlashWeave can be instructed to run only up to a certain (by default large) number of tests per node, assuming that performing such a high number of tests provides reasonable safety that the current candidate will not be discarded by additional tests. This effectively puts an upper bound on the exponential behaviour of si-HITON-PC and helps to prevent extensive run times on single variables with large neighborhoods, with empirically minimal effect on network structure. However, FlashWeave will flag these predicted interactions and warn the user in case the boundary is breached.

Normalization

Sequencing data is subject to mainly technically determined and thus arbitrary variations in sequencing depth, making it compositional in nature (Papageorgiou and Aitchison, 1989; Pawlowsky-Glahn and Buccianti, 2011). Compositionality impedes naive correlation analysis without adequate correction (Aitchison, 1981; Friedman and Alm, 2012). Common approaches to properly analyze compositional data include various log-ratio based methods, such as log-ratio transformations (Aitchison, 1981).

Similar to SpiecEasi (Kurtz et al., 2015), FlashWeave uses the centered log-ratio (*clr* (Aitchison, 1981)) approach for compositionality correction of a vector x of compositional microbial abundances:

$$clr(x_{ij}) = log\ \frac{x_{ij}}{g(s_i)} \quad \text{with } g(s_i) = \left[\prod_{l=1}^{p} x_{il}\right]^{\frac{1}{p}} \qquad \text{(Equation 1)}$$

where $g(s_i)$ describes the geometric mean of all compositional abundances in sample s_i, p the total number of OTUs and $clr(x_{ij})$ the *clr*-transformed value of the compositional abundance of microbe j in sample s_i.

An inherent shortcoming of logarithm-based methods is the handling of absences (zeroes) in the input data. This is usually circumvented by applying a fixed pseudocount (for example 1) to the input data which then allows proper logarithmic computations. Our analyses revealed that this approach can work for strongly filtered and depth-homogeneous data sets, but introduces noticeable biases when applied to data sets including rare OTUs and samples with particularly low sequencing depths (Figure S4A, left column). In such data, we observed extensive increases in univariate network density, which rendered the subsequent conditioning search in FlashWeave unusually slow (Figure S4C). Additionally, most of these additional univariate associations are finally removed during conditioning search (Figure S4A, left column), indicating their spurious nature.

More precisely, absences of comparatively rare OTUs in low-depth samples can, after *clr* transformation, become values higher than the OTU's mean *clr*-transformed abundance across all samples, while absences in high-depth samples result in transformed values below these OTU's means. This depth-based deviation from the mean results in the observed artificial association signal and notably is driven solely by applying the same fixed pseudo-count both to low-depth and high-depth samples. While homogenizing sequencing depth through sample removal and filtering of rare OTUs reduces this signal (Figure S4A, left column), large amounts of valuable data are potentially removed by this approach.

As an alternative method to reduce the pseudo-count driven association signal, we suggest a modification to classic fixed pseudo-counts, which we term "adaptive pseudo-counts", resulting in the normalization scheme *clr-adapt*. In this approach, initially a fixed pseudo-count π_{max} is applied to the sample with the highest sequencing depth (s_{max}). Then solving

$$log\left(\frac{\pi_i}{g(s_i)}\right) = log\left(\frac{\pi_{max}}{g(s_{max})}\right) \qquad \text{(Equation 2)}$$

for π_i (the adaptive pseudocount for sample s_i) leads to

$$\pi_i = \left[\frac{\pi_{max}^{k-p} \cdot g_{nz}(s_{max})}{g_{nz}(s_i)}\right]^{\frac{1}{n-p}} \qquad \text{(Equation 3)}$$

where $g_{nz}(s)$ is the geometric mean of all non-zero abundances in sample s, k is the number of absences in sample s_{max} and p is the number of OTUs. Equation 3 is applied to all samples excluding s_{max} in order to determine sample-specific adaptive pseudo-counts. These are then applied to their respective samples, followed by usual *clr* transformation (Equation 1). This results in the same transformed absence counts in all samples and ensures that absences are below each OTU's mean *clr*-abundance, which avoids bi-directional pseudo-count driven deviations from the mean. Using this approach, we observe strongly reduced univariate network densities (Figures S4A and S4B), discard fractions (Figure S4A) and run times (Figure S4C).

FlashWeaveHE also utilizes *clr* transformation for compositionality correction, albeit slightly modified. Since FlashWeaveHE does not consider absences for its association calculations (see STAR Methods), it requires no (adaptive) pseudo-counts. Instead, only non-zero compositional abundances are used to compute the compositional center (geometric mean, Equation 1) used for the transformation, resulting in the normalization scheme *clr-nonzero*.

FlashWeave-F and FlashWeaveHE-F differ from the FlashWeave-S and FlashWeaveHE-S by applying mutual information tests which necessitate data discretization. FlashWeave-F uses a straight-forward discretization scheme: all non-zero abundance values

become one, while absences remain zero (binarization). This approach makes *clr* normalization and pseudo-counts unnecessary, thereby avoiding the previously described issues, and is inherently robust to compositional artifacts, since the resulting data set is not compositional (i.e. changes in the counts of one OTU cannot influence counts of other OTUs). FlashWeaveHE-F on the other hand discretizes all *clr-nonzero* transformed values into two bins per OTU ("high" abundance vs "low" abundance), with bins separated by the median.

Meta variables (MVs) are by default not normalized for FlashWeave-S and FlashWeaveHE-S and should thus, if necessary, be provided in a sensible pre-normalized format by the user. For FlashWeave-F and FlashWeaveHE-F, continuous MVs are by default discretized into two bins separated by their median.

Accuracy and Robustness Benchmarks

For the NorTA (American Gut) benchmark, synthetic data was generated as described in (Kurtz et al., 2015) using the "amgut.filt" data set, "cluster" and "scale free" topologies and default settings for all other parameters. Briefly, the method takes an input data set of real microbial abundances, a target topology and a target distribution as its main parameters. It then generates an interaction matrix that matches the target topology and includes both direct and indirect OTU-OTU associations (realized by inverting a designed precision matrix, which specifies a multivariate-Gaussian distribution). Next, the method simulates abundances for all OTUs, based on the interaction matrix, and subsequently transforms these synthetic OTU abundances to a user-specified target distribution (the zero-inflated negative binomial in our case), fitted to the margins from the real input data set. Since the original method does not simulate noise as introduced by DNA extraction and sample sequencing, we furthermore downsampled each generated sample to depths randomly picked from "amgut.filt", which mimics the largely arbitrary sequencing of community subsets of varying size. This step furthermore induced data compositionality, an important property of real microbial sequencing data sets.

For the Ecological Models benchmark, data sets were generated as described in (Weiss et al., 2016), restricted to linear ecological relationships (corresponding to Tables 6, 7, and 16–18 in the original publication). In this simulation, OTUs are sampled from log-normal distributions (with varying parameters) and are subsequently linearly transformed through arithmetic operations over interaction partner abundances (for pairs or triples of OTUs). By adjusting signs and strengths of the transformations, this method was used to simulate a variety of ecological interaction types, including amensalism, commensalism, mutualism, parasitism, competition, obligate synthrophy and partial-obligate synthrophy.

For each topology (NorTA) and table (Ecological Models) we generated three independent data sets with 500, 1000 and 2000 samples, respectively (no additional OTU prevalence filters were applied). For the NorTA benchmark, we furthermore simulated five independent replicates per data set. To create data sets with multiple disjoint habitats for the heterogeneous Ecological Models benchmark (Figure 3C), the three differently sized data set tables were aggregated. In this aggregation, OTUs and interactions were assumed to be distinct, i.e. that each OTU and interaction are only present in one habitat. F1, recall and precision scores for all synthetic data benchmarks were computed using the Python package sklearn (Pedregosa et al., 2011).

The repeated rarefaction robustness benchmark was based on the procedure in (Weiss et al., 2016). Briefly, a single input data set was repeatedly rarefied to a fixed number of sequences and networks were computed with each method on each rarefaction. Then, for each network prediction method, all edges predicted in at least one rarefaction were collected and for each of these edges, the number of rarefactions it was found in was determined (higher counts = higher edge stability). We used the rarefied tables from (Weiss et al., 2016) (Tables 10–19 in the original publication; rarefaction depth 2000), which were based on the data set from (Ridaura et al., 2013), and further removed OTUs found in less than 20% of the samples in each rarefaction. We additionally benchmarked a randomly chosen synthetic data set from the NorTA benchmark (clustering topology, see above), the HMP data set (see subsection "Computational Speed Benchmarks") and the aggregated table 6 from our heterogeneous Ecological Models benchmark (see above) in this fashion. These data sets were rarefied to the same depth as the Weiss tables (2000 reads) using a custom Python function. For the heterogeneous data sets (Ecological Models, HMP), we used 5 instead of 10 rarefactions to reduce computation time.

For the structural zero robustness benchmark, we reduced all body sites in the HMP data set (see subsection "Computational Speed Benchmarks") via random subsampling to a fixed number of 312 samples per site. For each body site, we then picked all OTUs found in at least 10 samples of that site (175 - 619 OTUs) and removed their non-zero counts from all samples belonging to other body sites. The resulting body site-specific data sets were then aggregated into a single table. Inference tools were applied to i) each individual body site table separately, ii) the aggregated data set of all body sites. Finally, the edge overlaps between all sub-networks and the aggregated network were compared using the Jaccard similarity index.

Dependent sample groups were simulated in the following fashion: First, a dependence-free data set was simulated using a zero-inflated multivariate log-normal distribution, constructed with the Julia package Distributions.jl (JuliaStats, 2018c). OTUs were simulated as ecologically independent (covariance matrix = identity matrix), a vector of log-means for the log-normal component was sampled uniformly from range 2 to 10 and parameters for the zero-inflated multinomial component were sampled from a Beta distribution with α=1 and β=3. This model M was used to simulate abundances for 200 OTUs in 10,000 samples, resulting in OTU table A_{ind}. In addition to this dependence-free data set, sets S_n^f of dependent sample groups $g_{i,n}^f$ were generated, where $n \in \{5, 50, 100\}$ was the number of dependent sample groups per set, $i \in \{1 .. n\}$ was the group index and $f \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$was the fraction of samples in A_{ind} to be replaced with S_n^f (i.e. the dependence fraction). Additionally, we simulated different within-group distances $d \in \{0.0, 0.25, 0.5, 0.75\}$for each $g_{i,n}^f$ through constrained iterative sampling, where new samples were generated from M until the desired f was reached and new samples were only accepted when the mean Bray-Curtis distance to all previously accepted samples was within 0.01 from d (distance computed with the Julia package Distances.jl (JuliaStats, 2018a)). In the special case of d = 0.0, random

samples from A_{ind} were picked and repeated within $g^{f}_{i,n}$ until f was reached. Empirical distances in the simulated sample groups closely matched target distances (Figure S2B).

Computational Speed Benchmarks

The HMP data set consisted of 5514 samples from the body sites oral, gastrointestinal tract, urogenital tract, skin and airways. Samples were mapped to OTUs at 96% 16S rRNA identity using MAPseq (version v1.0 (Matias Rodrigues et al., 2017), confidence >0.5) and the full-length 16S reference provided with MAPseq. For the heterogeneous runtime benchmark, the data set was further filtered for OTUs with prevalence >20.

For the TARA Ocean benchmark, we aggregated the preprocessed OTU counts tables provided by (Lima-Mendez et al., 2015) into a single data set. After filtering for OTUs with prevalence >50 and samples with at least one read, the data set contained 289 samples and 3762 OTUs.

Parameters for each network inference tool were as reported in Table S1. The authors note that CoNet was run with the non-default Simes method for p-value merging, which tends to increase sensitivity at the possible loss of precision. Since not all tools readily supported parallelism, benchmarks were conducted on a single core on an AMD Opteron 2347 HE machine (1 GHz).

Literature Interaction Predictions

To reduce computation time, we filtered the TARA Oceans data set down to only OTUs participating in at least one genus-level literature interaction reported in (Lima-Mendez et al., 2015) (using evidence codes 2 and 3, i.e. microscopic evidence with or without sequence evidence). After removing samples with no reads (no OTU prevalence filters), the final data set consisted of 234 samples and 702 OTUs. Edges predicted by each tool were sorted according to their reported weights (merged q-value in the case of CoNet, which uses multiple weight measures; Pearson's r for SparCC) and this ranking was plotted as cumulative curves (Figure 2B).

Meta Variable Analysis in the HMP

An indirect association was counted as explained by a MV if at least one MV was present in the set of conditional variables leading to the association's exclusion (Figure S1B). The correlation between shared primer influence and interaction probability was estimated by computing, for each pair of OTUs (O_i, O_j), the absolute difference of association strengths in the HMP network between O_i and the primer MV and O_j and the primer MV, leading to small values for OTU pairs with similar primer influence and larger values for differences in influence. Correlations between these values and the interaction strengths for each O_i and O_j were then computed using Pearson's r.

Global Gut Network Analysis

Data Set Creation and Network Computation

Studies from the NCBI Sequence Read Archive database (SRA (Leinonen et al., 2011)) were filtered for human samples through the automated parsing of metadata annotation fields, matching at least one of the following rules: 1) "Human" or "Homo sapiens" is found in the host name field, 2) "9606" is found in either the host taxon ID or sample taxon ID field, or 3) "human (gut|gastrointestinal) metagenome" is found in the organism field, where "(gut|gastrointestinal)" is a regular expression match for either "gut" or "gastrointestinal". For matching samples, a list of keywords was parsed from all main annotation fields and further curated to remove uninformative terms, resulting in a set of keywords assigned to each sample. Samples were then further filtered for gut association by only retaining samples matching at least one of the following keywords: "intestinal", "intestine", "alimentary", "bowel", "cecum", "crohn", "gut", "colon", "commensal-gut", "diarrhoea", "digestive-tract", "digestive tract", "duodenum", "enteric", "enteritis", "enterocolitis", "enteropathogenic", "enterohemorrhagic", "equol", "feces", "gastroenteritis", "gastrointestinal", "ileum", "ileostomy", "jejunum", "meconium", "mesentery", "mid-gut", "probiotic", "rectum", "stec", "vibriosis". Keywords of these samples were additionally checked for terms not related to gut, followed by manual review of such samples via the SRA web service and removal in case of likely non-gut origin.

The final set of samples was downloaded and mapped to OTUs at 98% 16S rRNA identity using MAPseq (version v1.0 (Matias Rodrigues et al., 2017), confidence >0.5) and the full-length 16S rRNA reference database provided with MAPseq (hierarchically clustered with HPC-CLUST (Matias Rodrigues and von Mering, 2014) using average linkage). We removed samples with less than 100 mapped reads and OTUs found in less than 200 samples (see Table S2 for SRA accessions of the final sample set). Taxonomy was assigned to OTUs based on a 90% consensus over the full taxonomic lineages of all OTU member sequences. A trusted set of taxonomic classifications was generated using the annotated taxonomy provided by NCBI (updated on February 2018) including only sequences belonging to RefSeq (O'Leary et al., 2016) genomes and sequences from culture collection strains. The remaining sequences were taxonomically classified using a version of MAPseq modified to compute global alignments and the trusted set of sequences with their associated taxonomies (confidence cutoff $\geq$0.5). Applied identity cutoff and scaling parameters (delimited by colons) were 0.00:0.08 (Kingdom), 0.75:0.035 (Phylum), 0.785:0.035 (Class), 0.82:0.045 (Order), 0.865:0.06 (Family), 0.92:0.06 (Genus), 0.95:0.05 (Species), with identity cutoffs as suggested in (Yarza et al., 2014).

In addition, we retrieved sequencing method information from the SRA ("WGS", "AMPLICON" "RNA-SEQ" or "OTHER") and filtered the previously extracted metadata keywords for a set of 128 potentially interesting terms such as "fibre", "antibiotics" and "cancer". This metadata information was used to create a MV information table which was further hierarchically clustered into 92 MV groups (average linkage, unweighted Jaccard similarity >0.9). See Table S3 for representatives picked for each group, as

well as to which projects and how many samples per project each group was assigned. In addition, see Table S2 for a mapping of extracted MV information to individual sequencing samples.

The OTU table and the MV group table were finally used as input to FlashWeaveHE-F with parameters reported in Table S1 to compute the GGNcond and AGP networks and with max-k = 0 to compute GGNuni.

FDR Estimation and Modularity

To estimate the false positive rate, we generated a null model by breaking associations between taxa through sequencing depth-conserving shuffling of the GG data set (Lima-Mendez et al., 2015). Modularity (Newman, 2006) was computed for positive edges (unweighted) and based on cluster assignments from the Markov Cluster Algorithm (MCL (Van Dongen, 2008) version 14-137, inflation parameter 1.5).

Influence of Dependent Sample Groups on GGNcond

To estimate the impact of dependent sample groups on network inference of the Global Gut data set (GG), we first identified samples that were sequenced more than once. This resulted in 4700 samples (9% of all samples), independent re-sequencings of which covered 31% of the data set in total. When analyzing sample distances (measured as Bray-Curtis dissimilarity, computed with the Julia package Distances.jl (JuliaStats, 2018a)) within vs. between these sample groups, we found that within-group distances, while generally smaller than between-group distances, still covered a wide range of values (Figure S3B). This indicated substantial variation within sample groups, potentially providing important information for network inference. Additionally, our simulation benchmarks predicted negligible numbers of false positives for the mean within-group distance (0.37), dependent sample fraction (31%), number of groups (>100) and mode (FlashWeaveHE-F) used to compute the Global Gut network (GGNcond) (Figure S2A). As expected from our simulations, we furthermore detected a steep increase in predicted edges (43,493 to 82,552, an increase of 89%) when replacing each group with identical copies of one group-specific representative, and thus shifting the data set towards the distance region with high expected numbers of false positives.

Besides shared sample material, other sources of sample dependence, such as samples taken from the same individual in a time series, could also influence GGNcond. To account for these types of dependence, we systematically clustered samples in GG with increasing sample distance thresholds. For computational efficiency, the GG data set was clustered using an iterative greedy approach, in which samples were initially sorted by number of mapped reads (descending order). Iterating through that order, the clustering algorithm checked in each step if any subsequent samples were within the desired distance threshold d and added these samples to the current cluster, removing them from future consideration. For each clustering, we then computed networks based on cluster representatives and compared these to a background of networks computed from random subsets of GG with matching numbers of samples. The procedure was repeated for $d \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$. Univariate and conditional networks for each clustering and background data set were computed using the same parameters as used for GGNcond and GGNuni, respectively. In the presence of substantial false positives due to sample dependence, we expected the cluster-based networks to show steeper initial drops in edge numbers compared to the background networks, because clustering specifically removes spurious dependence signals and hence false positives, while random subsampling retains them. In addition, univariate networks produced markedly increased numbers of false positives under strong sample dependence conditions compared to conditional networks in our simulations (Figure S2A), suggesting that initial drops in edge numbers should be even more distinct for univariate networks. In our tests, we however did not observe any of these clear indicators of sample dependence, since edge reductions in cluster-based networks were always similar to or smaller than for background networks, both in the conditional and univariate case (Figure S3C). Hence, we could not detect a signal for false positives due to sample dependence in GGNcond.

Impact of Meta Variables on Negative Hubs

In order to estimate whether negative associations of the top 20 negative hub OTUs could be explained by MVs, we collected all negative associations of these OTUs and computed for each MV, how often samples assigned with this MV contributed to a positive or negative association signal within the negative edges. We then compared MV frequencies of negative contributions to those of positive contributions and found no significant difference (paired T-test, $P > 0.99$), indicating that positive and negative association signals were overall driven by samples with highly similar MV distributions.

Phylogenetic Assortativity

For phylogenetic tree construction, the alignment of representatives for all 98% 16S rRNA identity OTUs in the MAPseq reference database (92,659 full-length 16S rRNA sequences) was computed with INFERNAL (version 1.1.2 (Nawrocki and Eddy, 2013)) using the microbial secondary structure model SSU-ALIGN (Nawrocki and Eddy, 2013). The phylogenetic tree was then reconstructed using fasttree (version 2.1.3 (Price et al., 2010)) with the GTR substitution model and otherwise default options. For the phylogenetic assortativity analysis, the GGNcond network was reduced into two separate networks restricted to edges and vertices participating in only positive and negative associations, respectively. To generate a random background, vertices in each network were randomly connected to create a network with vertex and edge numbers matching the original network. Phylogenetic distance between interaction partners was calculated as total branch length between the leaves corresponding to these OTUs. The same procedure was repeated for GGNuni to estimate phylogenetic assortativity of univariate edges.

Associations between H_2 Producers and Consumers

OTU that mapped to H_2 producing and consuming taxa (taken from (Carbonero et al., 2012)) were identified in GGNcond. The number of positive associations between these groups was compared to associations between the same groups in 100 randomly generated networks. To assure comparability, random networks were generated such that the expected positive degree for each OTU was conserved and the interaction probabilities respected the phylogenetic assortativity signal detected in GGNcond. The latter was

done to assure that non-random association patterns were not explainable by phylogenetic assortativity alone. This step was implemented by using the Julia package KernelDensity.jl (JuliaStats, 2018b) to fit a Kernel Density Estimate (Gaussian kernel with $\mu = 0.0$ and σ=0.25) per OTU O_i to the phylogenetic distances D_{ij} between O_i of all of its positive interaction partners O_j^i, which yielded distribution P_i. When sampling neighbors of O_i, the probability π_{ij} of OTUs O_i and O_j interacting was then computed as the reciprocal product $P_i(D_{ij}) \cdot P_j(D_{ij})$, followed by re-normalization of $\pi_i = \{\pi_{ij}, j \neq i\}$ to a proper probability density. Degree conservation was achieved by only considering OTUs O_j for interaction if their current degree was still smaller than in GGNcond.

Normalization Comparison

The subset of Gastrointestinal tract samples from the HMP data set was filtered along sequencing depth and OTU prevalence gradients, followed by applying the *clr* (pseudo-count 1) and *clr-adapt* normalization schemes (see subsection "Normalization"). Associations were inferred using FlashWeave-S with max-k 0 (univariate) and max-k 3 (conditional) and all other options as in Table S1. For the oral comparison, the oral subset of the HMP data set (1000 OTUs, no OTU prevalence filter) was used and networks were computed with 20 CPU cores.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical tests used in our downstream analyses (i.e. excluding tests applied by network inference tools) include the two-sample Kolmogorov-Smirnov test, the paired T-test, a correlation test of Pearson's *r* (using the Student's t-distribution), and empirical permutation tests. Significance was defined as $P < 0.05$ for all downstream analyses. Tests were applied as appropriate throughout the Results section.

DATA AND CODE AVAILABILITY

FlashWeave is open source software implemented in Julia (Bezanson et al., 2017) and freely available from https://github.com/meringlab/FlashWeave.jl under the GNU General Public License v3.0. Sample accessions and extracted meta variable information for the Global Gut data set are provided in Table S2. The conditional network of predicted Global Gut interactions (GGNcond) is provided as Table S4.