PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Two dynamic regimes in the human gut microbiome

Sean M. Gibbons[1,2,3], **Sean M. Kearney**[1,2,3], **Chris S. Smillie**[1,2,3], **Eric J. Alm**[1,2,3]*

1 Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, United States of America, **2** The Broad Institute, Cambridge, MA, United States of America, **3** The Center for Microbiome Informatics and Therapeutics, Cambridge, MA, United States of America

* ejalm@mit.edu

Check for updates

OPEN ACCESS

Citation: Gibbons SM, Kearney SM, Smillie CS, Alm EJ (2017) Two dynamic regimes in the human gut microbiome. PLoS Comput Biol 13(2): e1005364. doi:10.1371/journal.pcbi.1005364

Editor: Elhanan Borenstein, Universtiy of Washington, UNITED STATES

Received: August 16, 2016

Accepted: January 16, 2017

Published: February 21, 2017

Copyright: © 2017 Gibbons et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The time series data analyzed in this paper can be obtained from https://genomebiology.biomedcentral.com/articles/10.1186/gb-2011-12-5-r50 and https://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-7-r89. Data from the Human Microbiome Project can be obtained via the following link: http://hmpdacc.org/HM16STR/. Processed data files used in our analyses, along with analysis code can be found at the following FigShare link: https://figshare.com/articles/Time_Series_Meta-Analysis_Files/3581616.

Abstract

The gut microbiome is a dynamic system that changes with host development, health, behavior, diet, and microbe-microbe interactions. Prior work on gut microbial time series has largely focused on autoregressive models (e.g. Lotka-Volterra). However, we show that most of the variance in microbial time series is non-autoregressive. In addition, we show how community state-clustering is flawed when it comes to characterizing within-host dynamics and that more continuous methods are required. Most organisms exhibited stable, mean-reverting behavior suggestive of fixed carrying capacities and abundant taxa were largely shared across individuals. This mean-reverting behavior allowed us to apply sparse vector autoregression (sVAR)—a multivariate method developed for econometrics—to model the autoregressive component of gut community dynamics. We find a strong phylogenetic signal in the non-autoregressive co-variance from our sVAR model residuals, which suggests niche filtering. We show how changes in diet are also non-autoregressive and that Operational Taxonomic Units strongly correlated with dietary variables have much less of an autoregressive component to their variance, which suggests that diet is a major driver of microbial dynamics. Autoregressive variance appears to be driven by multi-day recovery from frequent facultative anaerobe blooms, which may be driven by fluctuations in luminal redox. Overall, we identify two dynamic regimes within the human gut microbiota: one likely driven by external environmental fluctuations, and the other by internal processes.

Author summary

Dynamics reveal crucial information about how a system functions. In this study, we develop an approach for disentangling two types of dynamics within the human gut microbiome. We find that autoregressive dynamics involve recovery from large deviations in community structure. These recovery processes appear to involve the blooming of facultative anaerobes and aerotolerant taxa, likely due to transient shifts in redox potential, followed by re-establishment of obligate anaerobes. Non-autoregressive dynamics carry a strong phylogenetic signal, wherein highly related taxa fluctuate coherently. These non-autoregressive dynamics appear to be driven by external, non-autoregressive variables like diet. We find that most of the community variance is driven by day-to-day fluctuations in

Funding: Funding was obtained through the MIT Center for Microbiome Informatics and Therapeutics. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

the environment, with occasional autoregressive dynamics as the system recovers from larger shocks. Despite frequently observed disruptions to the gut ecosystem, there exists a returning force that continually pushes the gut microbiome back towards its steady-state configuration.

Introduction

The dynamic microbiome

Microbial ecology has become an important branch of medical science [1]. Recent work has shown how each person maintains a fairly unique microbial fingerprint [2–4], and that microbial dysbioses are often associated with shifts in health-status [5–8]. We now recognize that our microbiota are highly dynamic, and that these dynamics are linked to ecological resilience and host health [9–11]. The field has not yet settled upon whether gut microbial community structure varies continuously or if it jumps between discrete community states, and whether these states are shared across individuals [12–14]. In particular, some researchers suggest that gut communities can be binned into discrete 'enterotypes' [12], while others argue that gut communities vary along multidimensional continua [13]. If the ultimate goal of microbiome research is to improve human health by engineering the ecology of the gut, we must first understand how and why our microbiota vary in time, whether these dynamics are consistent across humans, and whether we can define 'stable' or 'healthy' dynamics.

Gut microbiota are continually buffeted by external factors like diet and host behavior [10, 15]. Internal species-species (e.g. cross-feeding or successional turnover) and host-species (e.g. immune system regulation or host physiology) interactions can also influence the structure and stability of the gut ecosystem [16–18]. Lotka-Volterra (LV) models, which predict changes in community composition through defined species-species or species-resource interaction terms, are popular for describing these internal ecological dynamics [19–22]. LV models are deterministic and fairly straight-forward to interpret, but little is known about the relative importance of these purely autoregressive factors in driving gut microbial dynamics (see *Theoretical Considerations* section below for a more detailed comparison of LV and VAR models). More recently, a model-free approach to forecasting non-linear dynamics—called convergent cross-mapping (CCM)—has been applied to ecological time series data [23]. While extremely useful, CCM can be difficult to interpret [24, 25], and may not be appropriate for high-dimensional systems with weak coupling between components (e.g. the gut).

Long, dense time series data are becoming increasingly available to microbial ecologists [10, 11, 26]. These temporal data are invaluable for understanding the behavior of microbial communities but require special care during analysis due to the non-independence of temporally adjacent samples [27]. In this paper we conduct a meta-analysis of the four longest, densest human gut time series currently available [10, 11].

Time series analysis approach

We separated microbial dynamics into autoregressive and non-autoregressive components by applying vector autoregressive (VAR) models, which were originally developed for econometrics [28–30]. We took this approach because we found that substantial autocorrelation persisted in most microbial time series for at least 3 days (i.e. past values of an OTU were predictive of its current value), which meant that temporally adjacent samples were not independent. VAR models are standard for analyzing stationary multivariate time series with

autocorrelation, cross-correlations, and noise. Time series are considered to be stationary if they appear to be sampled from the same probability distribution through time (i.e. the mean and variance, along with the other moments of the distribution, do not change through time; Fig 1). VARs model each element as a linear function of lagged values of other elements in the time series [29]. In order to reduce the number of coefficients generated by classical VAR models (i.e. with n species in a VAR model with p lags, n^2 x p coefficients are generated) and avoid over-fitting we apply regularized estimation, resulting in a sparse VAR (sVAR) [30, 31]. The residual variation of an sVAR model is stripped of much of its autoregressive structure, which allows for the application of standard statistical techniques that assume sample-to-sample independence. sVARs have the benefit of explicitly modeling error, unlike LV-type models [32], and are more straightforward to interpret than CCM forecasting [25].

Multiple possible drivers of dynamics

We suggest that there are two differentiable sets of drivers generating autoregressive and non-autoregressive microbial community dynamics in the gut. The first set of drivers induce multi-day recovery processes, where the past state of the system is predictive of the future state [33]. As mentioned above, LV models are differential equations that can incorporate species-species and species-resource interactions. These models are inherently autoregressive and are the dominant workhorses of ecological time series modeling [19, 26, 34]. The second type of driver is non-autoregressive, and likely includes dietary factors and other external perturbations [10]. In this paper, we show that there are indeed two dynamic regimes: auto-regressive and non-autoregressive. These dynamic regimes appear to reflect internal and external drivers, respectively. The emerging picture of the gut microbiome shows a dynamically stable system, continually buffeted by internal and external forces and recovering back toward a conserved steady-state.

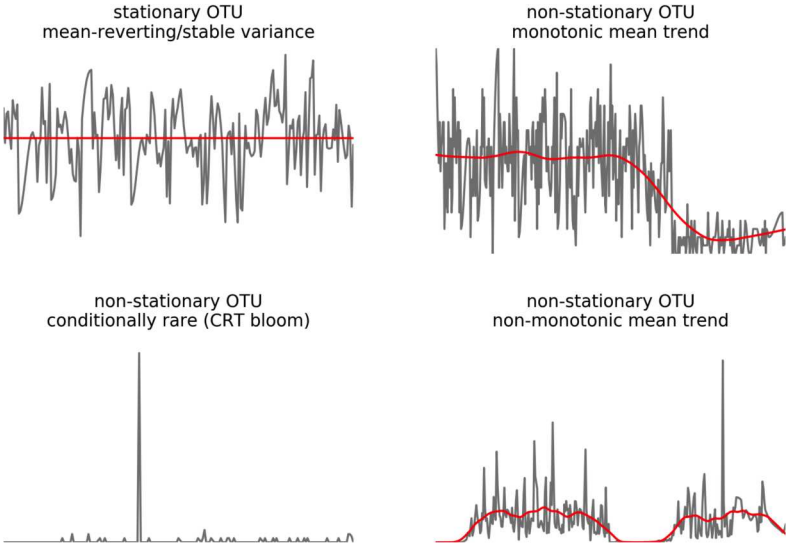

Fig 1. Time series stationarity and non-stationarity. Grey lines depict time series, with the mean plotted in red. The top left time series is stationary, with a stable mean (i.e. mean-reverting) and variance (i.e. non-heteroskedastic). The top right trace shows a non-stationary time series where the mean and variance change monotonically. The bottom left trace shows a time series with non-stationary variance (i.e. a sudden bloom event). The bottom right trace shows a non-stationary time series with both a changing variance and a non-monotonically changing mean.

doi:10.1371/journal.pcbi.1005364.g001

Theoretical considerations

In this section, we directly compare our VAR modeling approach to the more common generalized Lotka Volterra family of models (gLV). gLVs are first-order differential equations that model growth rates as a non-linear function of community composition, and thus assume the existence of mechanistic coupling between variables in the system. VAR models, by contrast, assume linear dynamics, but can only be applied when the observed data are, or have been transformed such that, the time series are stationary. In both gLV and VAR models, dynamics are defined by species-species (or species-resource) interaction terms. In this study, we find that a linear VAR-based approach was sufficient to extract essential dynamics of the system without the need for a nonlinear mechanistic framework (i.e. as implemented in gLV).

gLVs have the following structure:

$$\frac{dX_i}{dt} = a_i X_i(t)\left(1 - \frac{X_i(t)}{K}\right) + X_i(t)\sum_{j=1(j!=i)}^{n} B_{ij}X_j(t)$$

where t is time, a_i is the self-interaction term for organism i, X_i is the abundance of organism i, B_{ij} is the interaction term between organisms i and j in a community composed of n organisms.

Dividing by abundance and converting to difference equation form allow for gLV parameters to be solved with a system of linear equations as for a VAR(1) process [22]:

$$\log(X_i(t)) - \log(X_i(t-1)) = a_i - \frac{a_i}{K} * X_i(t-1) + \sum_{j=1(j!=i)}^{n} B_{ij}X_j(t-1)$$

We can now compare this linearized gLV to a corresponding VAR(1) process:

$$X_i(t) = q_i + s_i * X_i(t-1) + \sum_{j=1(j!=i)}^{n} R_{ij}X_j(t-1) + e_i\,(t)$$

where $e_i(t)$ is the error term. Thus, both gLV and VAR(1) can be solved using systems of linear equations, but the interpretation of the coefficients will be different. For instance, we can compare a_i (1/time) and for q_i (abundance), to see that VAR models directly model data on observed data, whereas gLV models assume that a more appropriate model maps observed abundance data to differenced log-transformed data.

Further, VAR(p) models can include an arbitrary number of time lags (i.e. autoregressive processes can extend further back in time). gLVs do not allow for explicit inclusion of historical time series data beyond one time lag. In considering the appropriate approach for analysis of microbial community time series, we created a decision tree to guide appropriate modeling strategies (Fig 2).

Methods

Data processing

Sequence data. Raw 16S (V4-V5) amplicon sequence time series data were obtained from David et al. (2014) [10] and Caporaso et al. (2011) [11]. The four data sets from these two papers are the longest, densest human gut time series currently available. Having long, dense time series is crucial for gaining an understanding of autocorrelation decay timescales and for fitting models to these large multivariate systems. 16S (V3-V5) amplicon data was also obtained from the Human Microbiome Project [1]. Raw data was quality filtered and demultiplexed using Quantitative Insights into Microbial Ecology (QIIME) default settings [35]. In

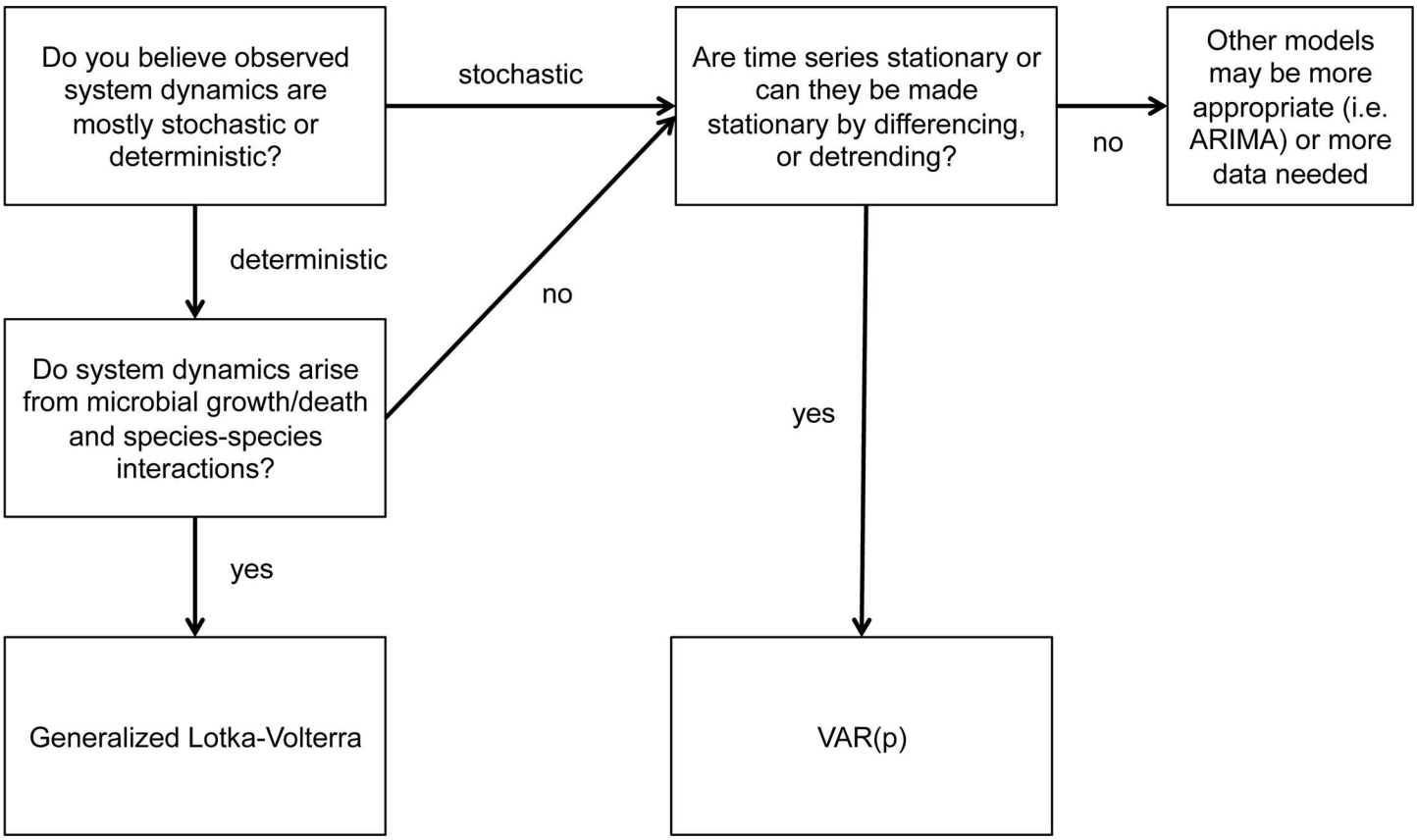

Fig 2. Flow chart indicating what time series modeling approach is appropriate given the structure of a data set.

doi:10.1371/journal.pcbi.1005364.g002

order to compare across studies and reduce technical variance between studies, closed reference Operational Taxonomic Units (OTUs) were clustered at 99% identity against the Greengenes database v. 13_5 [36]. Open reference OTU picking was also run [37], in order to look for non-database OTUs that might contribute substantially to community dynamics. Time series OTU tables (i.e. n x m matrices, with n OTUs and m samples) were normalized by random sub-sampling to contain 10,000 reads per sample. Alpha and beta-diversity metrics (Shannon entropy and Jensen-Shannon divergence) were calculated with custom Python scripts using normalized OTU tables. To reduce noise from sequencing or PCR artifacts in further downstream analyses and to reduce the complexity of our models, only the top 50 most abundant OTUs or the 76 OTUs common across all time series were analyzed. The time series from donors A and B in the David et al. (2014) paper had significant diarrheal events resulting from travel abroad and food poisoning. We wanted to focus on the dynamics of normal, healthy gut communities. Thus, for donor A we focused on days 150–365 (after return from traveling abroad) and for donor B we focused on days 0–150 (prior to food poisoning event). No significant perturbations were reported in the Caporaso et al. (2011) study, so all time points were used in the analysis. 268 HMP samples were sub-sampled to contain 1,000 sequences per sample and OTUs that occurred in fewer than 10 samples were removed.

We included analyses on two non-gut time series from a marine and a freshwater ecosystem. OTU tables and metadata for these studies obtained from the Earth Microbiome Project (EMP) [38]. The EMP study numbers were 1240 (English Channel marine time series) and 1242 (Lake Mendota freshwater time series). The OTU tables and metadata files from these

studies can be downloaded from FigShare: https://figshare.com/articles/Time_Series_Meta-Analysis_Files/3581616

Interpolation between missing time points and first-differencing. Each of our time series data sets had a few gaps (i.e. missing days). These gaps needed to be filled in order to apply our autoregressive modeling approach. Thus, missing data points were filled in using the SciPy [39] implementation of piecewise cubic interpolation (PCHIP). There are many alternative methods for interpolation that can be used (e.g. see SciPy documentation on interpolation methods: http://docs.scipy.org/doc/scipy/reference/interpolate.html) [40], but PCHIP had certain properties that were desired: abundances stay above zero; monotonicity is preserved; does not overshoot when data is not smooth [41].

OTU relative abundance matrices (*m* time points x *n* OTUs) were converted into rate matrices (*m-1* time points x *n* OTUs) by subtracting the relative abundance of each OTU at time *t* from its abundance at time *t+1* (i.e. first-differencing). The rate matrix (i.e. the first-difference or first-derivative matrix) shows the dynamics in rates between adjacent time points for each OTU in the data set.

Data analyses

State clustering. In order to determine whether microbial communities clustered into discrete states (i.e. community configurations that are non-continuously distributed throughout state-space), we fit a Dirichelet multinomial mixture model (DMM) to the data [42]. The DMM clustering null model assumes that the data are pulled from the same Dirichelet multinomial distribution, and increases the number of parent distributions depending on the heterogeneity of the input data. The default Laplace method was used to penalize model complexity (http://bioconductor.org/packages/release/bioc/manuals/DirichletMultinomial/man/DirichletMultinomial.pdf) [42]. We sampled each time series at different levels of temporal resolution (i.e. every time point, every other time point, every third time point, etc.) to determine whether the number of states was influenced by over- or under-sampling of the state space. We restricted the analysis to the top 50 most abundant OTUs within each time series or the 76 abundant OTUs that occurred across all time series.

Abundant, conditionally rare taxa. We looked for organisms that were transiently abundant across the time series by filtering for OTUs with coefficients of bimodality > 0.8 and whose peak abundance was $\geq 10\%$ of the total community [43]. These conditionally rare taxa (CRTs) are rare or undetected across most time points, but occasionally bloom into high-abundance (i.e. $\geq 10\%$ of the community sequence reads).

Autocorrelation, cross-correlation, and stationarity. Autocorrelation decay was determined for each OTU using the autocorrelation function (i.e. acf) in the StatsModels package in Python (http://statsmodels.sourceforge.net/stable/index.html) [44]. Stationarity (i.e. whether a time series is sampled from the same statistical distribution along its entire trajectory) was assessed using the combined results of Augmented Dickey-Fuller (ADF) [45] and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests (both TREND and LEVEL) [46]. The ADF test from the StatsModels package was used. An implementation of the KPSS test, written by Deniz Turan (http://denizstij.blogspot.co.uk/), was used (see Supplemental Materials for KPSS code).

Time series model. For each time series, the top 50 most abundant OTUs were fit to a lag-3 sparse vector autoregressive model (i.e. sVAR(3), where '(3)' indicates that 1–3 day lags were considered) using the sparsevar package in R v.3.2.8 (https://github.com/svazzole/sparsevar) (Fig 3) [30]. Elastic Net regularization was used for the sVAR fitting [47]. Thresholding was enforced so that very small coefficients were converted to zeros. Autocorrelation or

cross-correlation structure was reduced in the sVAR(3) residuals, which allows us to more confidently apply standard statistics (i.e. methods that assume sample-to-sample independence) to the model residuals (Fig 3). Higher-order lags and rare taxa were not considered in order to minimize the number of estimated parameters. Granger causality coefficients were calculated for each non-zero sVAR coefficient [29], and significance was assessed using a chi-squared test ($p < 0.05$). Granger causality is based on linear time-lagged prediction: if time series A is predictive of the future values of time series B, it is said that A Granger causes B [48]. Granger causal interactions do not imply direct causality, as these associations may arise due to indirect influences.

Correlation and phylogenetic structure. OTU-OTU Spearman's rho matrices were computed from the OTU matrices (i.e. relative abundances, rates, and rate residuals from the sVAR model) using Pandas [49]. Pairwise phylogenetic distances between OTU representative sequences were calculated using the dnadist function (default settings) in PHYLIP [50]. Spearman coefficients between OTUs were combined into phylogenetic distance bins, and the average coefficient for each bin was calculated.

Analysis code and notes. The data files and python code used to carry out the analyses presented in this manuscript, along with notes showing commands run in R, are available on FigShare: https://figshare.com/articles/Time_Series_Meta-Analysis_Files/3581616

Results and discussion

Community stability

Database (Greengenes) OTUs accounted for 95–99%, 93–99%, 83–97%, and 83–96% of all sequences per time point in the F4, M3, DA, and DB time series, respectively. The proportion

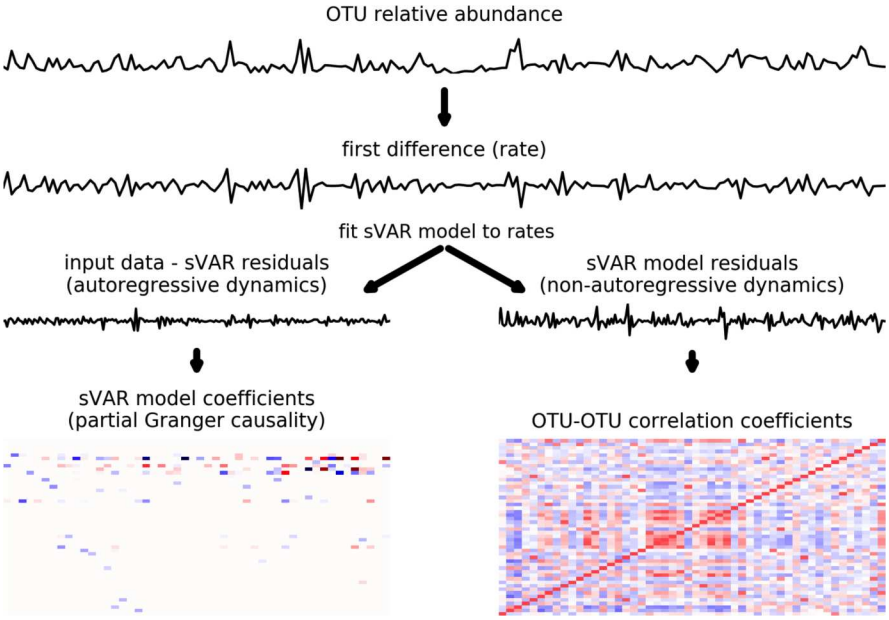

Fig 3. Two dynamic regimes in the gut microbiome. The top panel shows the relative abundance trace for an OTU. The second panel shows the same time series after first-differencing (i.e. taking the derivative to obtain the rate). In this data set, first-differencing ensured that each time series was stationary and appropriate for sparse vector autoregressive (sVAR) modeling. The third row of plots shows autoregressive and non-autoregressive rate dynamics, based on fitting the sVAR model. Autoregressive model coefficients (bottom left) provide information on time-lagged interactions between OTUs within the model (i.e. equivalent to partial Granger coefficients). OTU-OTU covariance (bottom right) of non-autoregressive dynamics provides information on co-fluctuating taxa.

doi:10.1371/journal.pcbi.1005364.g003

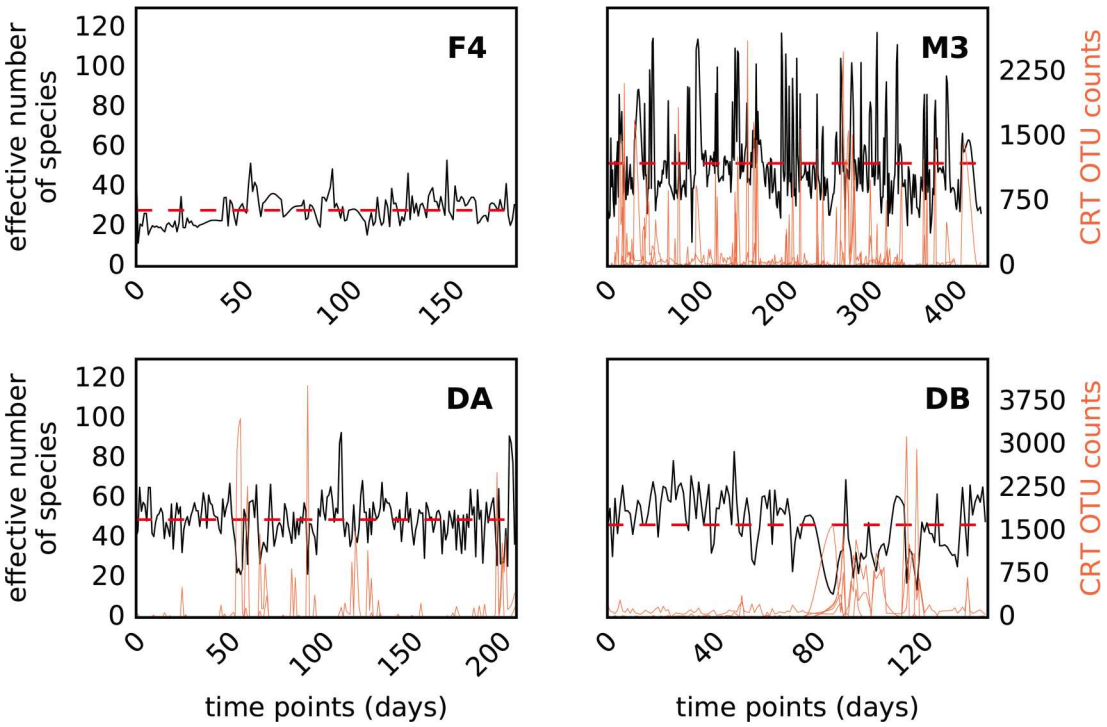

Fig 4. Alpha diversity is correlated with conditionally rare taxa (CRT) blooms. Black lines show the Shannon effective number of species (i.e. N_{eff}, a measure of alpha diversity) for each time series used in this analysis. The red dashed lines show the average N_{eff} for each time series. The average N_{eff} is between 28–51 across the time series, which is high enough that compositional effects are expected to be negligible (Freidman and Alm, 2012). Major perturbations in alpha diversity are associated with CRT blooms (orange lines = sum of CRT abundances).

doi:10.1371/journal.pcbi.1005364.g004

of non-database OTUs was quite stable over time. Furthermore, the 50 most abundant OTUs in each time series were all Greengenes OTUs. Non-database OTUs tended to be low-abundance taxa.

Microbial community alpha diversity showed stable, mean-reverting behavior across all four time series (Fig 4). The average effective number of species ($N_{eff} = e^{[Shannon\ diversity]}$) was between 28–50 for each time series, which indicated that compositional effects (i.e. spurious correlations caused by non-independence between relative abundances) were not a major concern in this analysis (Fig 4) [51]. Friedman and Alm (2012) found that, in simulated data, for a N_{eff} of ~30 or more, conventional statistics gave the same result as their compositionally aware method. We found that large deviations in alpha diversity were strongly correlated with a set of conditionally rare taxa (CRTs) that occasionally bloomed to as much as ~30% of the community (dominant taxa at steady-state were usually between 10–20% of the community), but were usually found at very low abundances (Fig 4). In order to test whether CRT blooms drove significant compositional effects in the correlation structure of our time series, we calculated Spearman's correlations between the 100 most abundant taxa in the M3, DA, and DB time series with and without the CRT time points (i.e. time points where CRTs rose above 10% relative abundance were removed; S1 Fig). Overall, we saw no compositional effects due to CRT blooms (S1 Fig). A recent meta-analysis by Shade and Gilbert (2015) found that CRTs are responsible for a significant fraction of overall community dynamics in many different ecosystems [43]. Abundant gut CRTs—*Prevotella*, *Bacteroides fragilis*, *Akkermansia muciniphila*, *Lachnospiraceae*, *Enterobacteriaceae* and *Haemophilus parainfluenzae*—were present in M3, DA, and DB time series (S1 Table; Fig 4). Abundant CRTs (i.e. peak abundance $\geq$ 10% of

Table 1. Percentage of abundant OTUs (50 most abundant OTUs from each time series) that show stationary dynamics before (raw) and after (delta) first-differencing. Three tests of stationarity were used: the augmented Dickey-Fuller (ADF) test, the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) trend test, and the KPSS level test.

	ADF	KPSS Trend	KPSS Level
M3 raw	92.0	26.0	34.0
M3 delta	100.0	100.0	100.0
F4 raw	80.0	44.0	40.0
F4 delta	100.0	100.0	100.0
DA raw	100.0	48.0	90.0
DA delta	100.0	100.0	100.0
DB raw	88.0	22.0	44.0
DB delta	100.0	100.0	100.0

doi:10.1371/journal.pcbi.1005364.t001

sequence reads) were not identified in the F4 time series, consistent with its relatively stable alpha-diversity trace (Fig 4). However, there were several *Enterobacteriaceae* blooms in the F4 time series that fell beneath our abundance threshold. The M3 time series was an outlier, with much more frequent CRT blooms than the other time series. These bloom events likely represent opportunistic or pathogenic organisms that are either the cause or symptom of a disruption in the normal steady-state gut environment. Many CRT OTUs are facultative anaerobes or aerotolerant taxa (e.g. *Enterobacteriaceae* OTUs and other Proteobacteria or OTUs in the *B. fragilis* group, like *B. uniformis*, *B. ovatus*), which are probably responding to changing redox potential in the gut following some disturbance (e.g. inflammation) [52].

After first-differencing (i.e. differencing OTU abundances at adjacent time points), abundant OTUs showed completely stationary dynamics, implying that the mean, variance, and autocorrelation structure do not change over time (ADF and KPSS tests; Table 1). Both ADF and KPSS Level tests (Table 1) assess whether or not a process has a unit root, which implies that the mean changes through time and that the system does not recover to the mean trend in the presence of a shock (i.e. the process is integrated). The KPSS Trend test (Table 1) assesses whether or not a process is trend-stationary, which implies a mean trend with stationary error that is capable of recovering to the trend line following a shock. Non-stationary OTUs are likely not stable members of the steady-state community. Prior to first-differencing, most OTU trajectories showed ADF-stationary dynamics (Table 1). There was no apparent enrichment for particular taxonomic groups among non-stationary OTUs. In addition to OTU abundance trajectories being largely stationary, we also saw a range of autocorrelation decay curves for the top 50 most abundant OTUs (Fig 5). Some OTUs showed strong, persistent autocorrelation, while others did not (Fig 5). The amount of autocorrelation decay also varied across time series. In particular, the DA time series showed much less autocorrelation than the other three time series (Fig 5).

Stationarity implies that there is a restoring force on an OTU's abundance over time, so that it returns to a mean value after a perturbation (i.e. a steady-state population size, or 'carrying capacity'). Indeed, we found a significant negative correlation between the change in OTU abundances between time t and time $t+1$ and the abundances of OTUs at time t (Fig 6). Thus, by converting abundance dynamics to rate dynamics, we can achieve stationarity across all abundant time series in the system and preserve correlation structure between OTUs, as long as we exclude non-stationary windows that contain large disturbances (e.g. food poisoning in the DB time series on day 150). Furthermore, the strong correlation between rates and OTU abundances at the prior time step indicates that much of the variance in these data is linear (Fig 6). The fact that these rate dynamics can be modeled as linear, stationary processes opens up a wide array of time series models to gut data sets.

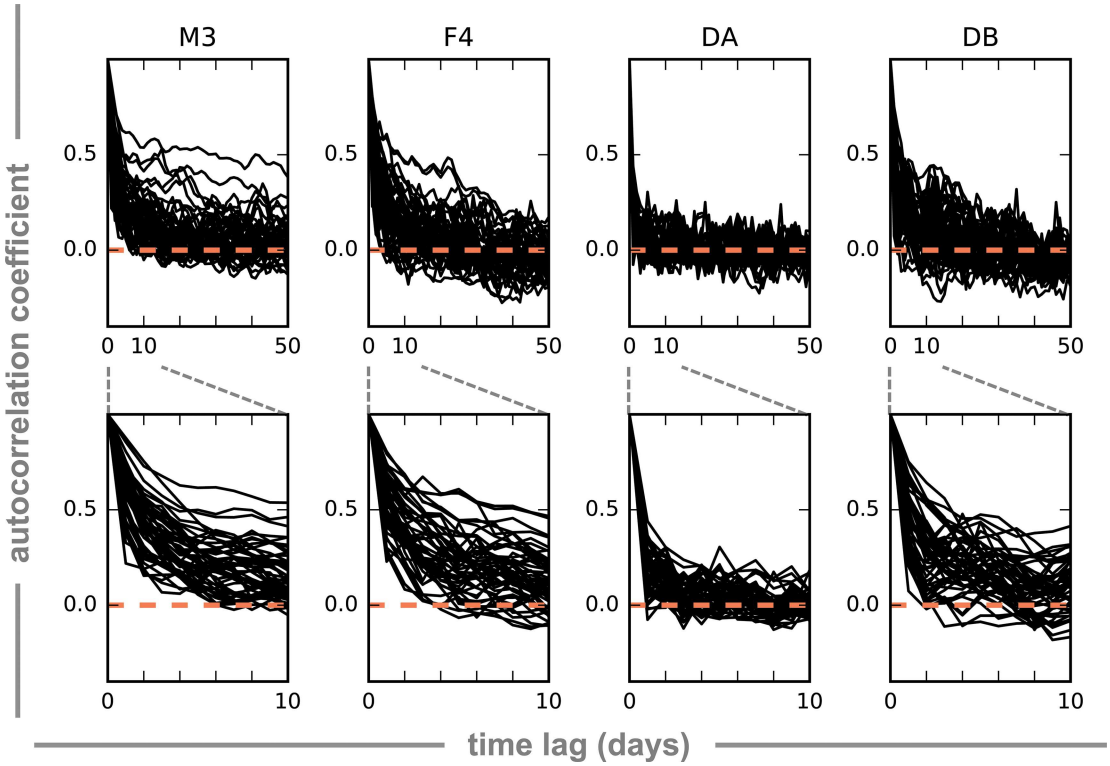

Fig 5. Autocorrelation decay plots for the 50 most abundant OTUs in each time series. Each black line represents the autocorrelation for a single OTU over different time lags (in days). The top row of plots shows autocorrelation (y-axis) for up to 50-day lags (x-axis). Orange dashed lines indicate an autocorrelation coefficient of 0.0. Dashed grey lines show the zoom-in region from the upper row of plots presented in the second row of plots (lags up to 10 days).

doi:10.1371/journal.pcbi.1005364.g005

Discrete vs. continuous dynamics in the gut microbiome

Prior work has argued for the existence of discrete gut community configurations across humans, termed 'enterotypes', that may be associated with health and disease [12, 14]. However, longitudinal data has provided evidence that these enterotypes may arise from undersampling individuals through time. For example, the M3 time series is known to moved fluidly through all three putative enterotypes from Arumugam et al. (2010) during the course of a year [13], which suggests that gut communities vary along a continuum. We expanded upon this analysis using an improved Dirichlet multinomial mixture model (DMM) clustering method [14, 42], which assesses whether samples appear to be pulled from a common Dirichlet multinomial distribution. In general, clustering methodologies are plagued by over- and under-sampling issues. In a time series, samples that are taken close together in time are likely to be similar to one another (i.e. they are autocorrelated). If the community is sampled densely enough as it moves through state space, then packets of samples that happen to be temporally adjacent may be grouped into pseudo-clusters (i.e. over-sampling). On the other hand, if only a handful of samples are taken from an individual, then outlier points may give rise to pseudo-clusters because the state space was not sufficiently sampled (i.e. under-sampling). If there are no discrete states to be found, then the number of states should decay smoothly to 1 as the sampling sparcity is increased. However, if discrete states do exist, then the number of states should reach a plateau that is stable across a range of sampling densities.

We aimed to address whether gut communities can be grouped into discrete states and whether these states are shared across individuals given a range of intra-individual sampling

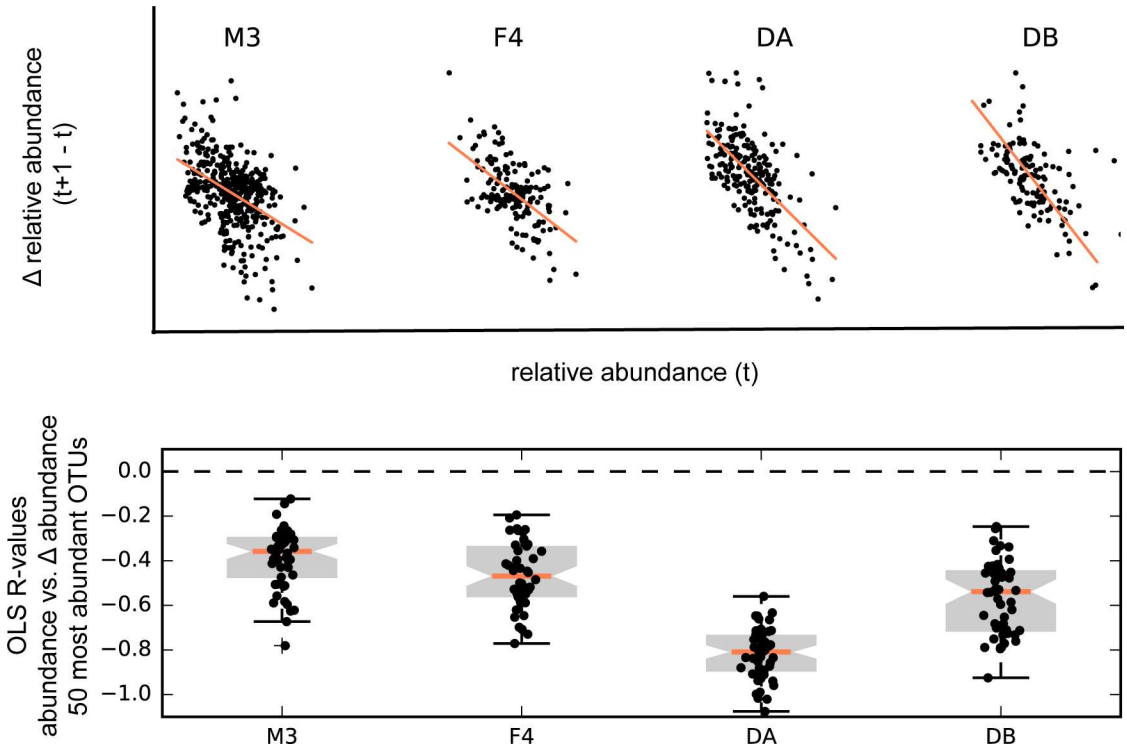

Fig 6. OTUs in the time series data show a propensity for mean-reversion. The top row of plots shows the change in relative abundance from time t to time t+1 (y-axis) vs. the relative abundance (x-axis) for OTU 850870 (Greengenes ID; *Bacteroides spp.*; the most abundant OTU in M3, F4 and DB time series and 2nd most abundant OTU in DA time series). The bottom plot summarized the ordinary least-squares (OLS) regression coefficient (R) for the top 50 most abundant OTUs from each data set. In general, there is a negative association between an OTU's change in abundance and its current abundance level, suggesting that OTUs have fixed carrying capacities. The average magnitude of the regression coefficient appears to be an indicator of the returning force. These results are consistent with stationarity.

doi:10.1371/journal.pcbi.1005364.g006

effort. DMM states were fit to the full-length time series, which included the food poisoning event in the DB time series. Independent of our OTU filtering method (50 most abundant OTUs in each time series, or 76 abundant OTUs shared across all time series), we found that the number of DMM states decayed rapidly from 15 to a plateau of 6 (Fig 7A). These states were almost entirely unique to individuals (Fig 7B). There was only one case where a sample from DA was assigned to a state largely associated with DB. In addition, only DB and M3 time series harbored multiple states, which is consistent with major perturbations in the gut community within these time series (i.e. food poisoning and frequent CRT blooms, respectively). Overall, barring major perturbations, we conclude that individual humans—given sufficient sampling density—can be distinguished by unique Dirichelet multinomial distributions (Fig 7). Because almost every time point from a subject's time series is assigned to the same one or two DMM states, continuous methods are necessary for exploring the dynamics of microbial communities within an individual. We believe this result will hold as larger numbers of long-term human gut time series become available.

Despite the clustering of individuals into unique states, we identified 956 OTUs that were present across all four individuals. These organisms tended to be found at similar median abundances across individuals (Fig 8). In addition, these shared OTUs made up 70–80% of the sequences in each data set (Table 2). Furthermore, a single OTU (from the genus *Bacteroides*; Greengenes ID 850870) was the most abundant taxon in M3, F4 and DB, and was the second most abundant organism in DA (Fig 6). Thus, despite the fact that most OTUs appear to be

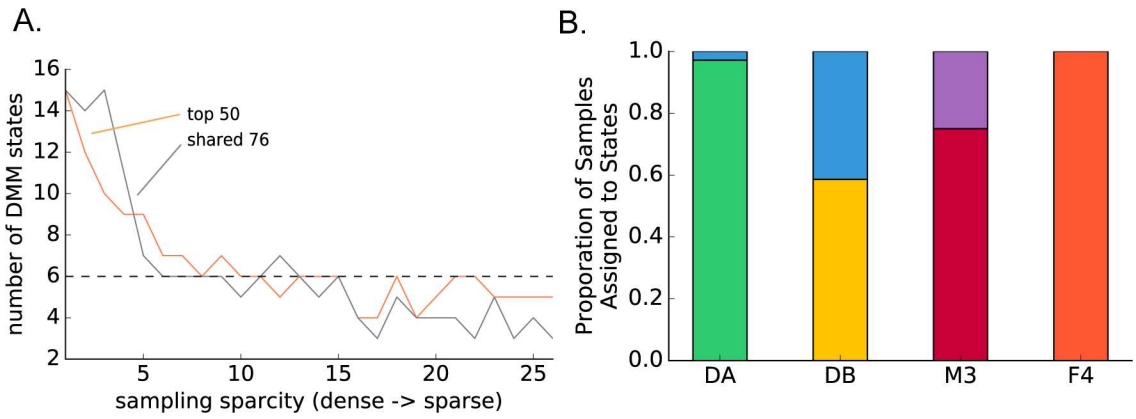

Fig 7. (A) shows decay in the number of Dirichelet multinomial mixture model (DMM) states as the sparcity of the time series is increased for the top 50 most abundant OTUs from each time series and for the 76 abundant OTUs shared across all four time series. The plateau is at 6 states for both methods (grey dashed line). (B) shows state classifications for samples taken from all four individuals (top 50 OTUs from each time series sampled at every 11th time point). States are largely unique to individuals (i.e. colors indicate unique states), with the exception of a single time point in the DA series that was assigned to a state (blue) that is dominant in DB.

doi:10.1371/journal.pcbi.1005364.g007

unique to an individual's gut, there is a set of abundant core taxa that are present at similar abundances across people. It is unlikely that compositional effects are responsible for the existence of these carrying capacities, due to the fact that there is a wide-range in of abundances at which OTUs appear to persist (Fig 8). With similar carrying capacities not only within but also across individuals, these core OTUs likely occupy similar metabolic niches across humans.

Sparse vector autoregression (sVAR) and granger causality

We employ an analytical approach based on continuous, multivariate, linear, autoregressive modeling tools developed for econometrics to pull apart two independent dynamic regimes in the human gut microbiome (Fig 3). Each regime, autoregressive and non-autoregressive, tells a unique story about the gut ecosystem.

OTU abundance trajectories showed fairly strong autocorrelation structure (Figs 5 and S2), although there was no evidence for auto-covariance in a limited set of dietary metadata from the DB time series (S3 Fig). The autocorrelation decay curves showed that most of the autocorrelation was gone after a lag of 3 or 4 days in the abundance data, and most of the autocorrelation was gone from the differenced time series after 1 or 2 days (S2 Fig). Thus, we chose to fit a lag-3 sparse vector autoregressive model to all the data to account for this autoregressive signal. sVAR(3) models produced residuals with reduced autocorrelation structure (S2 Fig). The sVAR models (i.e. the linear autoregressive components of the variance) accounted for a minority of the total community variance (0–50% for any given OTU; Fig 9). The set of OTUs with strong autoregressive signals were phylogenetically heterogeneous (S4 Fig). OTUs from the *Enterobacteriaceae* family tended to have larger amounts of their variance explained by the sVAR than other taxa (Fig 9). These *Enterobacteriaceae* OTUs were also often identified as CRTs, blooming periodically from low to high abundance. The most abundant taxa also tended to show more autoregressive structure than lower-abundance taxa (Fig 9).

Many sVAR coefficients showed significant Granger causal associations (i.e. the past abundances of one OTU predict the future abundances of another OTU; Fig 10; Chi-Squared test, $p < 0.05$). The M3 time series had the sparsest Granger network, with no significant relationships for 3-day lags. This lack of significant Granger relationships at 3-day lags may be due to the higher frequency of CRT blooms in that time series, which may have continually disrupted

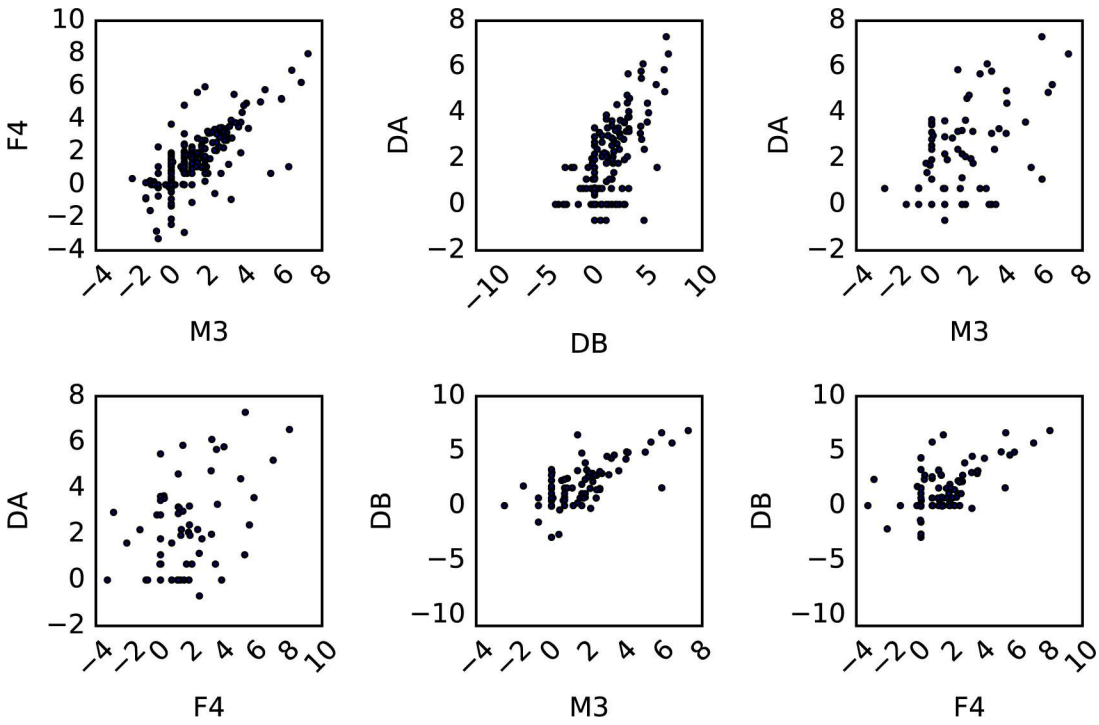

Fig 8. Scatter plots of the log median abundances of shared 99% OTUs across time series. Taxa that are present across time series show a positive correlation in their median abundances.

doi:10.1371/journal.pcbi.1005364.g008

community recovery and obscured successional trends. A *B. fragilis* OTU had the largest number of connections in the M3 Granger network. *B. fragilis*, *B. ovatus*, and *Enterobacteraceae* OTUs had the largest numbers of connections in the F4 Granger network. In the DA and DB Granger networks, *F. prausnitzii* OTUs had the largest number of significant Granger interactions (Fig 10). Overall, *Bacteroides*, *Faecalibacterium*, and *Enterobacteriaceae* OTUs were prevalent in all the Granger networks (Fig 10). In each Granger network, there were several Granger-causal OTUs that influence multiple downstream responder OTUs, but there were only a few responder OTUs (e.g. *B. fragilis*, in the F4 time series) that integrate multiple upstream Granger signals. This pattern is consistent with cascading dynamics that result from perturbing a highly connected/central node in a network with an external shock [53]. In general, facultative aerobes (e.g. *Enterobacteraceae* OTUs) and aerotolerant taxa (e.g. *Bacteroides*)

Table 2. Number of OTUs shared across the time series (M3, F4, DA, and DB) and human microbiome project (HMP) data sets. The 956 OTUs shared across all four time series make up the following percentages of total abundances in each time series: M3 = 69.9%; F4 = 77.2%; DA = 76.9%; and DB = 79.9%. The percent of total community abundance made up by the 308 OTUs shared across all data sets (including HMP) are as follows: M3 = 61.1%; F4 = 70.0%; DA = 61.7%; DB = 59.1%; and HMP = 39.3%.

	M3	F4	DA	DB	HMP
M3	7260				
F4	3436	4334			
DA	2109	1431	4839		
DB	1796	1234	2804	3910	
HMP	1214	820	855	788	3293

*shared across 4 time series: 956

*shared across all data sets: 308

doi:10.1371/journal.pcbi.1005364.t002

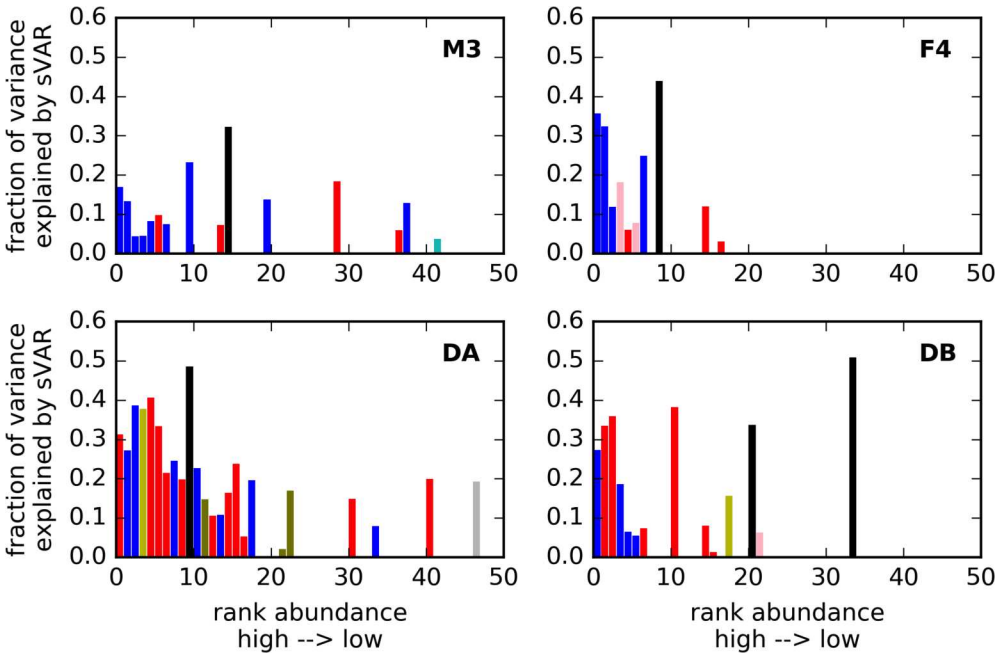

Fig 9. Amount of variance explained by sVAR(3) model for top 50 OTUs in each time series. Bars are colored by order: blue = *Bacteroidales*, red = *Clostridiales*, black = *Enterobacteriales*, pink = *Verrucomicrobiales*, yellow = *Erysipelotrichales*, olive = *Bifidobacteriales*, silver = *Pasteurellales*, cyan = *Synergistales*.

doi:10.1371/journal.pcbi.1005364.g009

were more likely to Granger-cause obligate anaerobes (e.g. most Firmicutes OTUs) than the reverse (Fig 10). These intrinsic dynamics suggest a successional process that might follow a spike in luminal oxygen levels [52].

Non-autoregressive dynamics contain phylogenetic structure

sVAR(3) residuals showed reduced autocovariance and could thus be more appropriately analyzed using standard statistical methods that assume independence. Closely related gut bacteria were generally positively correlated with one another, but this coherence decayed rapidly with phylogenetic distance (Fig 11). There was a significant anti-correlation between OTU-OTU phylogenetic distance and OTU-OTU correlations in abundances and abundance rates through time (i.e. highly related taxa tended to be positively correlated; Spearman's $p < 0.001$). This phylogenetic coherence is conserved in the sVAR(3) residuals, but is completely absent in the sVAR(3) coefficients (Fig 11). Moreover, although we did not identify strong correlations between the set of dietary variables measured for the DB time series and the DB gut community, these dietary variables showed little-to-no autocorrelation structure (S3 Fig; and see Additional File A8 in David et al., 2014) [10]. Furthermore, we fit an sVAR(3) model to the original 97% OTUs from the DA time series paper and found that none of the OTUs that had previously been shown to correlate with dietary variables had any autoregressive signal (i.e. sVAR coefficients = 0; S1 File). We suggest that unmeasured dietary variables and host behavior/physiology are the non-autoregressive drivers responsible for the pronounced phylogenetic signal in OTU-OTU residual correlation structure (i.e. related taxa are positively correlated because they share a similar host/dietary niche, which fluctuates stochastically in time). In addition to phylogenetic coherence in the correlation structure within a gut time series, we find the same phylogenetic coherence in the correlations between OTU abundances across

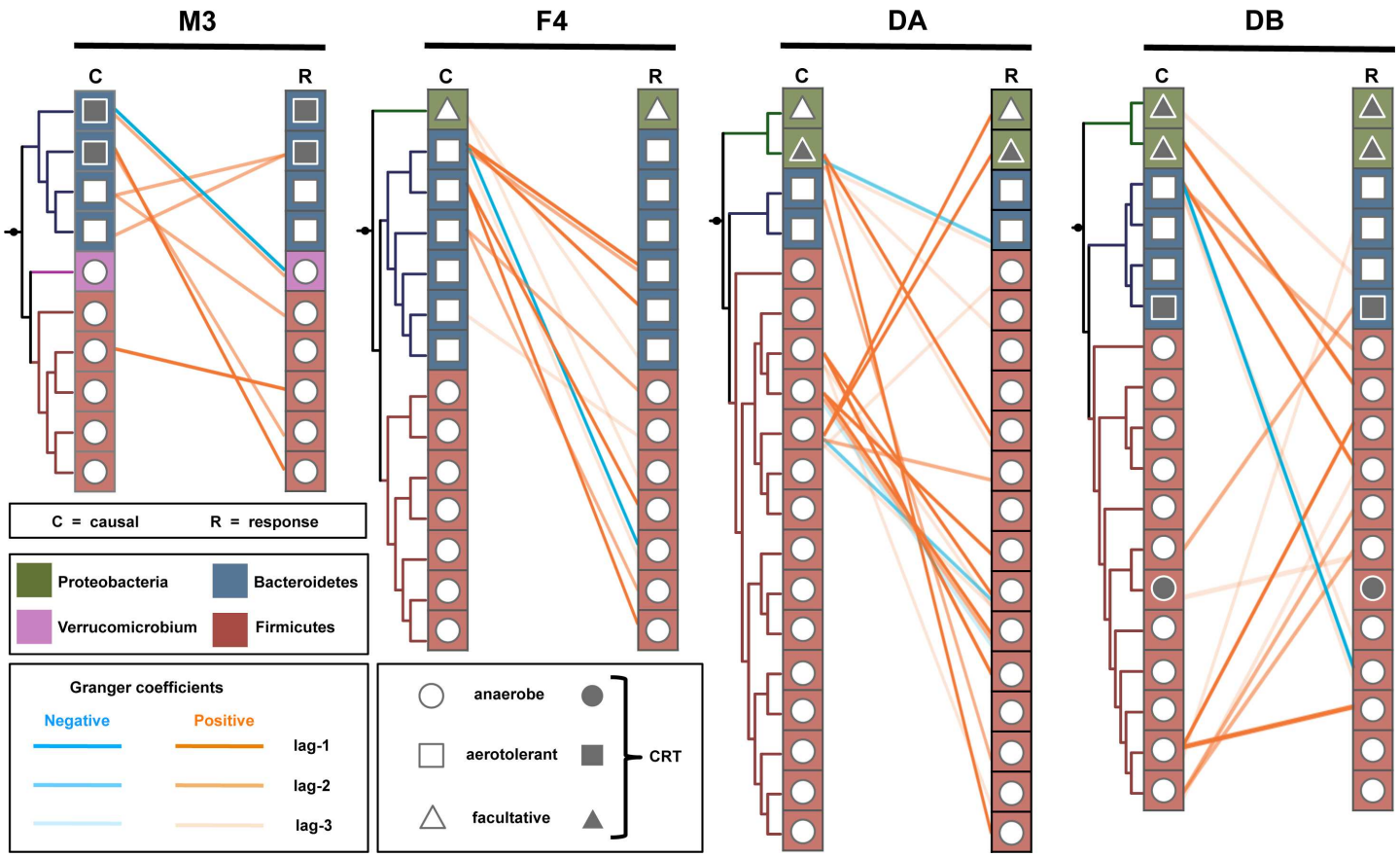

Fig 10. Granger causal networks for each time series. Trees show phylogenetic relationships between taxa. Edges are interactions identified in the sVAR (3) model that also show significant Granger causality ($p < 0.05$). Symbols indicate oxygen tolerance and growth traits. CRT stands for 'conditionally rare taxon', which is defined by a coefficient of bimodality > 0.8 and a maximum abundance of > 10% of the community. CRTs are organisms that are usually rare, but occasionally bloom to very high abundance. CRTs tend to be facultative aerobes, or aerotolerant taxa.

doi:10.1371/journal.pcbi.1005364.g010

people from the HMP gut data set (S5 Fig). Thus, microbial phylogeny is strongly coupled to host niche in the human gut. In order to assess the generality of this relationship between phylogeny and correlation, we analyzed two time series from the English Channel and from Lake Mendota (marine and freshwater environments, respectively). We found the same pattern in these non-host associated environments, suggesting that niche filtering is a strong driver of dynamics across different ecosystems (S6 and S7 Figs).

Conclusion

Despite distinct differences in community composition across humans, the dynamics of the gut microbiota are stable and highly conserved. Dominant microbes in the gut appear to have fixed carrying capacities (i.e. their dynamics are stationary), which opens the door to many classical time series modeling approaches. Significant time-lagged interactions between OTUs often include opportunistic, facultative anaerobic organisms like *Enterobacteraceae*, and obligate anaerobes like *F. prausnitzii* [54, 55]. These autoregressive interactions appear to be due to succession and recovery of the gut community from CRT blooms, which may result from a disruption in the luminal redox balance [52]. The largest component of community variance is non-autoregressive and appears to be driven by non-autoregressive environmental forces, like pH [56] or fiber intake [10]. Unlike the autoregressive dynamics, these non-autoregressive

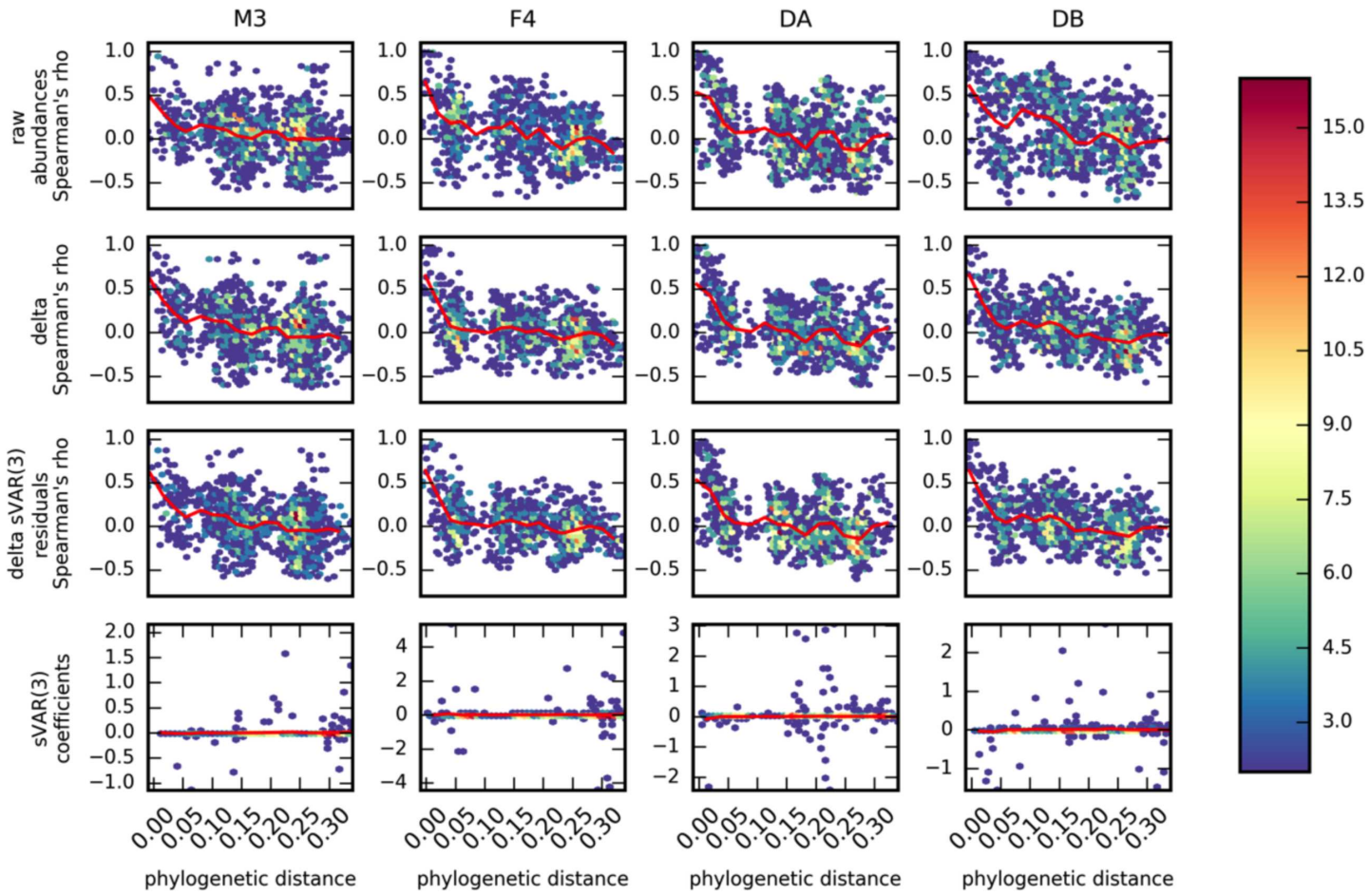

Fig 11. Phylogenetic relatedness corresponds to similarity in dynamics for top 50 most abundant OTUs. Each plot is a heatmap showing the density of pairwise OTU-OTU correlation or sVAR coefficients (y-axes) vs. pairwise phylogenetic distances (x-axes). The red lines show the mean coefficient along phylogenetic distance windows. Top row shows Spearman correlations calculated based on raw count data. Second row shows Spearman correlations calculated based on the first-difference (delta) of the count data. The third row shows the Spearman correlations performed on the residuals of a sVAR(3) model fit to the deltas. The bottom row shows the sVAR(3) coefficients. The heatmap colors denote the density of OTU-OTU pairs at a given hexagonal pixel on the plot.

doi:10.1371/journal.pcbi.1005364.g011

dynamics carry a strong phylogenetic signal, indicative of niche filtering. Our results, based on a limited number of individuals, paint a coherent picture of the gut ecosystem and the major forces underlying its structure and stability, with two distinct dynamic regimes: one driven by external factors (e.g. diet) and the other by internal autoregressive processes (e.g. recovery following a disturbance). Moving forward, it will be important to collect more time series data, from both healthy and diseased individuals, to determine how general these dynamics are and whether or not they are observed in dysbiotic gut communities.

Supporting information

S1 Table. Abundant CRTs (max abundance $\geq$ 10%; coefficient of bimodality $\geq$ 0.8) across the four time series.
(DOCX)

S1 Fig. OTU-OTU correlation coefficients (Spearman's rho) of first-differenced data are not strongly impacted by CRT blooms in the DA, DB, and M3 time series (i.e. time series

where CRTs rise above 10% relative abundance). The x-axis shows the correlation coefficients for all time points and the y-axis shows the coefficients calculated with CRT time points removed. Red lines show linear regression fits ($R^2 > 0.9$ across all three time series).
(TIF)

S2 Fig. Autocorrelation structure in four long-term gut time series. First-differencing (i.e. calculating the rate, or delta) and subsequent sparse VAR (sVAR) modeling removes most of the autocorrelation from the data. Most autocorrelations decay to zero within 2–3 days (i.e. for the raw abundance counts), so a maximum time lag of 3 days was chosen for sVAR fitting.
(TIF)

S3 Fig. Autocorrelaion decay for ten quantitative dietary variables (e.g. carbohydrate intake, fiber intake, fat intake, etc.) from the DB time series.
(TIF)

S4 Fig. The top 50 most abundant OTUs in each time series plotted on phylogenetic trees. Black circles denote amount of variance explained by an sVAR(3) model. Larger circles indicate OTUs with a larger autoregressive component to their variance. Tips of trees are labeled with the Family-level taxonomic annotation for each OTU.
(TIF)

S5 Fig. Hexagonal 2-D heatmap showing OTU-OTU Spearman correlation coefficients (y-axis) vs. OTU-OTU phylogenetic distances (x-axis) for the 50 most abundant taxa in the Human Microbiome Project (HMP) gut data set. The heatmap colors denote the density of OTU-OTU pairs at a given hexagonal pixel on the plot.
(TIF)

S6 Fig. Hexagonal 2-D heatmap showing OTU-OTU Spearman correlation coefficients (y-axis) vs. OTU-OTU phylogenetic distances (x-axis) for the English Channel marine time series data (Earth Microbiome Project Study #1240). The heatmap colors denote the density of OTU-OTU pairs at a given hexagonal pixel on the plot.
(TIF)

S7 Fig. Hexagonal 2-D heatmap showing OTU-OTU Spearman correlation coefficients (y-axis) vs. OTU-OTU phylogenetic distances (x-axis) for the Lake Mendota freshwater time series data (Earth Microbiome Project Study #1242). The heatmap colors denote the density of OTU-OTU pairs at a given hexagonal pixel on the plot.
(TIF)

S1 File. Percent of variance explained by an sVAR(3) model for the original 97% OTUs from the DA time series paper. OTUs that had previously been shown to correlate with dietary variables (highlighted in yellow) had no autoregressive signal (i.e. sVAR coefficients = 0).
(XLSX)

Acknowledgments

We thank Talia Karasov, Daniel Kerr, Monika Schultz, Jonathan Friedman, Mathieu Groussin, Tami Lieberman and members of the Alm lab for helpful comments and discussion regarding this work. Thanks to Lawrence David and Greg Caporaso for providing access to data. Thanks to Simon Vazzoler for help with implementing the sparsevar package in R and to Katherine Huang for help with data visualization.

Author Contributions

Conceptualization: SMG SMK EJA.

Data curation: SMG.

Formal analysis: SMG SMK.

Funding acquisition: EJA.

Investigation: SMG SMK.

Methodology: SMG SMK CSS EJA.

Project administration: SMG EJA.

Resources: EJA.

Supervision: EJA.

Validation: SMG SMK.

Visualization: SMG SMK CSS EJA.

Writing – original draft: SMG.

Writing – review & editing: SMG SMK CSS EJA.

References

1. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The Human Microbiome Project. Nature. 2007; 449(7164):804–10. doi: 10.1038/nature06244 PMID: 17943116
2. Lax S, Smith DP, Hampton-Marcell J, Owens SM, Handley KM, Scott NM, et al. Longitudinal analysis of microbial interaction between humans and the indoor environment. Science. 2014; 345(6200):1048–52. doi: 10.1126/science.1254529 PMID: 25170151
3. Fierer N, Lauber CL, Zhou N, McDonald D, Costello EK, Knight R. Forensic identification using skin bacterial communities. Proc Natl Acad Sci USA. 2010; 107(14):6477–81. doi: 10.1073/pnas.1000162107 PMID: 20231444
4. Franzosa EA, Huang K, Meadow JF, Gevers D, Lemon KP, Bohannan BJ, et al. Identifying personal microbiomes using metagenomic codes. Proc Natl Acad Sci USA. 2015; 112(22):E2930–E8. doi: 10.1073/pnas.1423854112 PMID: 25964341
5. Bakken JS, Borody T, Brandt LJ, Brill JV, Demarco DC, Franzos MA, et al. Treating Clostridium difficile infection with fecal microbiota transplantation. Clin Gastroenterol Hepatol. 2011; 9(12):1044–9. doi: 10.1016/j.cgh.2011.08.014 PMID: 21871249
6. Ley RE, Turnbaugh PJ, Klein S, Gordon JI. Microbial ecology: human gut microbes associated with obesity. Nature. 2006; 444(7122):1022–3. doi: 10.1038/4441022a PMID: 17183309
7. Clemente JC, Ursell LK, Parfrey LW, Knight R. The impact of the gut microbiota on human health: an integrative view. Cell. 2012; 148(6):1258–70. doi: 10.1016/j.cell.2012.01.035 PMID: 22424233
8. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, et al. Human gut microbiome viewed across age and geography. Nature. 2012; 486(7402):222–7. doi: 10.1038/nature11053 PMID: 22699611
9. Gerber GK. The dynamic microbiome. FEBS Lett. 2014; 588(22):4131–9. doi: 10.1016/j.febslet.2014.02.037 PMID: 24583074
10. David LA, Materna AC, Friedman J, Campos-Baptista MI, Blackburn MC, Perrotta A, et al. Host lifestyle affects human microbiota on daily timescales. Genome Biol. 2014; 15(7):R8.
11. Caporaso JG, Lauber CL, Costello EK, Berg-Lyons D, Gonzalez A, Stombaugh J, et al. Moving pictures of the human microbiome. Genome Biol. 2011; 12(5):R50. doi: 10.1186/gb-2011-12-5-r50 PMID: 21624126
12. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, et al. Enterotypes of the human gut microbiome. Nature. 2011; 473(7346):174–80. doi: 10.1038/nature09944 PMID: 21508958

13. Knights D, Ward TL, McKinlay CE, Miller H, Gonzalez A, McDonald D, et al. Rethinking "Enterotypes". Cell Host Microbe. 2014; 16(4):433–7. doi: 10.1016/j.chom.2014.09.013 PMID: 25299329
14. Ding T, Schloss PD. Dynamics and associations of microbial community types across the human body. Nature. 2014; 509(7500):357–60. doi: 10.1038/nature13178 PMID: 24739969
15. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, et al. Diet rapidly and reproducibly alters the human gut microbiome. Nature. 2014; 505(7484):559–63. http://www.nature.com/nature/journal/v505/n7484/abs/nature12820.html—supplementary-information. doi: 10.1038/nature12820 PMID: 24336217
16. Flint HJ, Duncan SH, Scott KP, Louis P. Interactions and competition within the microbial community of the human colon: links between diet and health. Environ Microbiol. 2007; 9(5):1101–11. doi: 10.1111/j.1462-2920.2007.01281.x PMID: 17472627
17. Freilich S, Zarecki R, Eilam O, Segal ES, Henry CS, Kupiec M, et al. Competitive and cooperative metabolic interactions in bacterial communities. Nat Comm. 2011; 2:589.
18. Peterson DA, McNulty NP, Guruge JL, Gordon JI. IgA response to symbiotic bacteria as a mediator of gut homeostasis. Cell Host Microbe. 2007; 2(5):328–39. doi: 10.1016/j.chom.2007.09.013 PMID: 18005754
19. Wilson WG, Lundberg P, Vazquez DP, Shurin JB, Smith MD, Langford W, et al. Biodiversity and species interactions: extending Lotka-Volterra community theory. Ecol Lett. 2003; 6(10):944–52.
20. Fisher CK, Mehta P. Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. PLoS One. 2014; 9(7):e102451. doi: 10.1371/journal.pone.0102451 PMID: 25054627
21. Stein RR, Bucci V, Toussaint NC, Buffie CG, Rätsch G, Pamer EG, et al. Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. PLoS Comput Biol. 2013; 9(12):e1003388. doi: 10.1371/journal.pcbi.1003388 PMID: 24348232
22. Bucci V, Tzen B, Li N, Simmons M, Tanoue T, Bogart E, et al. MDSINE: Microbial Dynamical Systems INference Engine for microbiome time-series analyses. Genome Biol. 2016; 17(1):121. doi: 10.1186/s13059-016-0980-6 PMID: 27259475
23. Sugihara G, May R, Ye H, Hsieh C-h, Deyle E, Fogarty M, et al. Detecting Causality in Complex Ecosystems. Science. 2012; 338(6106):496–500. doi: 10.1126/science.1227079 PMID: 22997134
24. McCracken JM, Weigel RS. Convergent cross-mapping and pairwise asymmetric inference. Phys Rev E. 2014; 90(6):062903.
25. Mønster D, Fusaroli R, Tylén K, Roepstorff A, Sherson JF. Inferring causality from noisy time series data. arXiv preprint arXiv:160301155. 2016.
26. Faust K, Lahti L, Gonze D, de Vos WM, Raes J. Metagenomics meets time series analysis: unraveling microbial community dynamics. Curr Opin Microbiol. 2015; 25:56–66. doi: 10.1016/j.mib.2015.04.004 PMID: 26005845
27. Chen EZ, Li H. A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. Bioinformatics. 2016.
28. Wang Y, Hu X, Jiang X, He T, Yuan J, editors. Predicting microbial interactions by using network-constrained regularization incorporating covariate coefficients and connection signs. IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2015.
29. Asteriou D, Hall S. Vector Autoregressive (VAR) Models and Causality Tests. Applied Econometrics. London: Palgrave MacMillan; 2011.
30. Basu S, Michailidis G. Regularized estimation in sparse high-dimensional time series models. Ann Stat. 2015; 43(4):1535–67.
31. Fujita A, Sato JR, Garay-Malpartida HM, Yamaguchi R, Miyano S, Sogayar MC, et al. Modeling gene expression regulatory networks with the sparse vector autoregressive model. BMC Syst Biol. 2007; 1(1):1.
32. Wangersky PJ. Lotka-Volterra population models. Ann Rev Ecol Syst. 1978; 9:189–218.
33. Volterra V. Variations and fluctuations of the number of individuals in animal species living together. J Cons Int Explor Mer. 1928; 3(1):3–51.
34. Bashan A, Gibson TE, Friedman J, Carey VJ, Weiss ST, Hohmann EL, et al. Universality of human microbial dynamics. Nature. 2016; 534(7606):259–62. http://www.nature.com/nature/journal/v534/n7606/abs/nature18301.html—supplementary-information. doi: 10.1038/nature18301 PMID: 27279224
35. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Meth. 2010; 7(5):335–6.

36. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. 2012; 6(3):610–8. doi: 10.1038/ismej.2011.139 PMID: 22134646

37. Rideout JR, He Y, Navas-Molina JA, Walters WA, Ursell LK, Gibbons SM, et al. Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences. PeerJ. 2014; 2:e545. doi: 10.7717/peerj.545 PMID: 25177538

38. Gilbert JA, Jansson JK, Knight R. The Earth Microbiome project: successes and aspirations. BMC Biol. 2014; 12(1):69.

39. Jones E, Oliphant T, Peterson P. {SciPy}: Open source scientific tools for {Python}. 2014.

40. Bressert E. SciPy and NumPy: An Overview for Developers: " O'Reilly Media, Inc."; 2012.

41. Fritsch FN, Carlson RE. Monotone piecewise cubic interpolation. SIAM J. 1980; 17(2):238–46.

42. Holmes I, Harris K, Quince C. Dirichlet multinomial mixtures: generative models for microbial metagenomics. PLoS One. 2012; 7(2):e30126. doi: 10.1371/journal.pone.0030126 PMID: 22319561

43. Shade A, Gilbert JA. Temporal patterns of rarity provide a more complete view of microbial diversity. Trends Microbiol. 2015.

44. Seabold S, Perktold J, editors. Statsmodels: Econometric and statistical modeling with python. Proceedings of the 9th Python in Science Conference; 2010.

45. Fuller WA. Introduction to statistical time series: John Wiley & Sons; 2009.

46. Syczewska EM. Empirical power of the Kwiatkowski-Phillips-Schmidt-Shin test. 2010.

47. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Series B Statist Methodol. 2005; 67(2):301–20.

48. Granger CW. Some recent development in a concept of causality. J Econ. 1988; 39(1):199–211.

49. McKinney W, editor Data structures for statistical computing in Python. Proceedings of the 9th Python in Science Conference; 2010.

50. Felsenstein J. PHYLIP-phylogeny inference package (version 3.2). Cladistics. 1989; 5:163–6.

51. Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. PLoS Comput Biol. 2012; 8(9):e1002687. doi: 10.1371/journal.pcbi.1002687 PMID: 23028285

52. Rigottier-Gois L. Dysbiosis in inflammatory bowel diseases: the oxygen hypothesis. ISME J. 2013; 7 (7):1256–61. doi: 10.1038/ismej.2013.80 PMID: 23677008

53. Jordán F, Scheuring I. Network ecology: topological constraints on ecosystem dynamics. Phys Life Rev. 2004; 1(3):139–72.

54. Khan MT, Duncan SH, Stams AJM, van Dijl JM, Flint HJ, Harmsen HJM. The gut anaerobe Faecalibacterium prausnitzii uses an extracellular electron shuttle to grow at oxic-anoxic interphases. ISME J. 2012; 6(8):1578–85. doi: 10.1038/ismej.2012.5 PMID: 22357539

55. Derrien M, Collado MC, Ben-Amor K, Salminen S, de Vos WM. The mucin degrader Akkermansia muciniphila is an abundant resident of the human intestinal tract. Appl Environ Microbiol. 2008; 74(5):1646–8. doi: 10.1128/AEM.01226-07 PMID: 18083887

56. Duncan SH, Louis P, Thomson JM, Flint HJ. The role of pH in determining the species composition of the human colonic microbiota. Environ Microbiol microbiology. 2009; 11(8):2112–22.