Bacterial community structures are unique and resilient in full-scale bioenergy systems

Jeffrey J. Werner[a], Dan Knights[b], Marcelo L. Garcia[c], Nicholas B. Scalfone[a], Samual Smith[d], Kevin Yarasheski[d], Theresa A. Cummings[e], Allen R. Beers[e], Rob Knight[f], and Largus T. Angenent[a,1]

[a]Department of Biological and Environmental Engineering, Cornell University, Ithaca, NY 14853; [b]Department of Computer Science, University of Colorado, Boulder, CO 80309; [c]Department of Energy, Environmental and Chemical Engineering, Washington University, St. Louis, MO 63130; [d]Washington University School of Medicine, St. Louis, MO 63110; [e]Anheuser-Busch, Inbev Inc., St. Louis, MO 63118; and [f]Department of Chemistry and Biochemistry, University of Colorado, Boulder, CO 80309

Edited by Norman R. Pace, University of Colorado at Boulder, Boulder, CO, and approved January 26, 2011 (received for review October 22, 2010)

Anaerobic digestion is the most successful bioenergy technology worldwide with, at its core, undefined microbial communities that have poorly understood dynamics. Here, we investigated the relationships of bacterial community structure (>400,000 16S rRNA gene sequences for 112 samples) with function (i.e., bioreactor performance) and environment (i.e., operating conditions) in a year-long monthly time series of nine full-scale bioreactor facilities treating brewery wastewater (>20,000 measurements). Each of the nine facilities had a unique community structure with an unprecedented level of stability. Using machine learning, we identified a small subset of operational taxonomic units (OTUs; 145 out of 4,962), which predicted the location of the facility of origin for almost every sample (96.4% accuracy). Of these 145 OTUs, syntrophic bacteria were systematically overrepresented, demonstrating that syntrophs rebounded following disturbances. This indicates that resilience, rather than dynamic competition, played an important role in maintaining the necessary syntrophic populations. In addition, we explained the observed phylogenetic differences between all samples on the basis of a subset of environmental gradients (using constrained ordination) and found stronger relationships between community structure and its function rather than its environment. These relationships were strongest for two performance variables—methanogenic activity and substrate removal efficiency—both of which were also affected by microbial ecology because these variables were correlated with community evenness (at any given time) and variability in phylogenetic structure (over time), respectively. Thus, we quantified relationships between community structure and function, which opens the door to engineer communities with superior functions.

community dynamics | UniFrac | community function | digester | sludge

The production of bioenergy from wastes is an essential component in the global development of sustainable energy sources (1). Anaerobic digestion, which is the most prominent bioenergy technology worldwide, uses undefined microbial cultures to produce methane from organic substrates (2). Methanogenic bioreactors are maintained on the basis of decades of observed relationships between performance and operating parameters. However, differences underlying bioreactors that perform well and bioreactors that perform inadequately are often poorly understood (3). This has led to a general perception that methanogenic bioreactors are unreliable or unstable, inhibiting their wider adoption for bioenergy production (2). A deeper analysis of the structure and dynamics of bioreactor microbial communities as a function of performance and operating conditions has the potential to reveal important and unappreciated structure–function relationships.

The efficient and stable operation of methanogenic bioreactors relies on syntrophic relationships among a community of microbes, including fermenting bacteria, specialized acidogenic and acetogenic syntrophs, and methanogenic archaea (4), with diverse and parallel pathways for substrate metabolism. Many studies have sequenced clones of the most abundant organisms in methanogenic bioreactors (4, 5), but the population dynamics are poorly understood. Ecological theories based on stochastic models and principles of island biogeography suggest that bioreactor community dynamics may be governed by chaotic shifts among functionally redundant organisms (6). In laboratory-scale systems, researchers have observed stable community structures in functionally stable (7) and unstable (8, 9) bioreactors, community shifts that could be explained by reactor conditions (10, 11), and even chaotic community shifts in functionally stable bioreactors (12, 13). It is difficult to assess the observed discrepancies between stable communities and those that are more dynamic. One hurdle has been the limitations of techniques for quantifying microbial community structure, both in sequencing depth and limited resolution of fingerprinting. Deep-sequencing time series, such as those used in the current study, present an opportunity to study microbial community structure and dynamics, informed by phylogeny.

To be useful for bioenergy production, a microbial community must have a stable metabolic function over time, despite the unavoidable perturbations and disturbances that occur in real-world systems. Here, we focus on three ecological factors that play important roles in maintaining a stable and robust community function: (*i*) a functionally diverse microbial community provides a suite of parallel pathways for each trophic step. Hashsham et al. (14) observed that systems with a larger number of parallel pathways have a more robust function over time. (*ii*) Evenness in the structure of microbial diversity ensures that the community has more capacity to use its varied array of metabolic pathways. Wittebolle et al. (15) demonstrated that microbial communities with greater evenness have a more robust function. Finally, (*iii*) the dynamics of populations over time allow for the community to adjust following perturbations and disturbances, providing the system with access to the total functional diversity and environmental specificity available in the community. Allison and Martiny (16) divided the population dynamics that maintain community function over time into three basic mechanisms: resistance (in which a population maintains abundance over time), resilience (in which a population rebounds following a disturbance), and redundancy (in which a disturbed population is replaced by a new population whose function is redundant with the original).

Author contributions: J.J.W., D.K., R.K., and L.T.A. designed research; J.J.W., D.K., M.L.G., N.B.S., T.A.C., and A.R.B. performed research; S.S. and K.Y. contributed new reagents/analytic tools; J.J.W., D.K., R.K., and L.T.A. analyzed data; and J.J.W., D.K., R.K., and L.T.A. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The sequence reported in this paper has been deposited in the GenBank database (accession no. SRA029112, study no. SRP005270).

[1]To whom correspondence should be addressed. E-mail: la249@cornell.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1015676108/-/DCSupplemental.

We hypothesized that, in stable anaerobic bioreactors, the variation in both individual populations and phylogenetic structure will be related to variation in function (i.e., bioreactor performance) and environment (i.e., operating conditions). We used barcoded 454 pyrosequencing of bacterial 16S rRNA genes to analyze samples from an unprecedented monthly time series of nine full-scale methanogenic bioreactors (112 samples with an average of >3,500 sequences each). These data were paired with extensive daily measurements of bioreactor performance and operating conditions (over 20,000 measurements). We observed that the methanogenic bioreactors of this study were composed of stable bacterial communities that were resilient following disturbances, both in terms of population profiles and phylogenetic structure. We applied constrained ordination to relate phylogenetic community structure (UniFrac variation) to environmental gradients. After comparing these results with other supervised and unsupervised methods of ordination and machine learning, we identified the function and environment variables that best explained community structure and populations that were most reliably associated with differences in bioreactor phylogenetic structure.

Results

Monthly biomass samples were collected for 1 y (October 2007–September 2008) from nine different full-scale methanogenic granular-sludge bioreactor facilities treating brewery wastewater. We sampled four upflow anaerobic sludge blanket (UASB) facilities (locations U1, U2, U3, and U4), and other reactor configurations based on the UASB design, including four expanded granule sludge bed (EGSB) facilities (locations E1, E2, E3, and E4), and one internal circulation (IC) facility (location I1). Two additional, earlier time series (January–December 2003) had been sampled for locations E3 and E4. Performance and operating condition data were measured on a daily basis, including substrate loading rate (feeding rate, *F*) normalized to biomass (*M*; *F*:*M* ratio, also referred to as sludge loading rate), soluble chemical oxygen demand (sCOD) concentrations (a measurement of how much soluble organic material is in wastewater), volatile fatty acid (VFA) concentrations, pH, temperature, $FeCl_2$ feed-supplement concentrations, methane production rate, and monthly biomass assays, yielding a wide range of aceticlastic methanogenic activity values and relative percentage of the acetate oxidation pathway compared with the total removal of acetate. Barcoded 454 pyrosequencing of bacterial 16S rRNA gene amplicons from 112 biomass samples resulted in 412,734 nonchimeric reads of sufficient quality and an average of 3,680 reads per sample. Tables S1 and S2 summarize the performance, operating conditions, and sequencing data obtained for each facility/time point. Taxonomic divisions were dominated by Proteobacteria (mostly Syntrophobacterales and Desulfuromonales), Bacteroidetes, Spirochaetes, Clostridia, Chloroflexi, and Synergistia (Fig. S1). A small proportion of sequences (8 ± 3%; SD) were unclassified.

Bacterial communities at each location fell into stable, consistent phylogenetic clusters over the 1-y period of sampling. The phylogenetic variation, which was measured via UniFrac (17) distances, is displayed in Fig. 1*A* as a plot of the first two UniFrac principal coordinates. Ordination results, such as those in Fig. 1, are used to visualize distances and variation between samples (*SI Methods*). The UniFrac distance between samples represents the difference between microbial communities on the basis of the fraction of evolutionary history in a phylogenetic tree that is unique to one of the communities rather than shared by both. The close clustering within locations indicates that samples from the same location through time were more similar to each other in phylogenetic structure than they were to samples from other locations. The additional time series from year 2003 (samples labeled '03 in Fig. 1*A*) for locations E3 and E4 clustered closely to the later 2007–2008 time series; the separate years were closer to each other than the average distance to samples from the next-closest facility ($P < 0.01$; two-tailed independent Student *t* test comparing between-year distances to distances to the next-nearest location). The consistency of E3 and E4 reactor communities throughout the 4-y time period demonstrates surprising phylogenetic stability over an extensive period.

No single performance or operating condition variable consistently explained the UniFrac clustering. To determine which environmental gradients best explained variation in phylogenetic structure, we used constrained redundancy analysis (RDA) to predict the UniFrac principal coordinates (Fig. 1*B*) using a linear combination of several environmental gradients. The arrows indicate the size and direction of the coefficients of the environmental variables in the linear model (i.e., the vector for each environment variable used to produce Fig. 1*B*). Comparing these constrained coordinates to the unconstrained principal components analysis (PCA) of the same data (Fig. 1*C*) shows that most of the phylogenetic variation in the first two axes can be explained using four performance and environmental condition variables. For example, in the first axis, constrained RDA explains 80% as much phylogenetic variation as the unconstrained ordination (7.2% in Fig. 1*B* vs. 9.0% in Fig. 1*C*). Because these ordination techniques were used to visualize complex variation among many

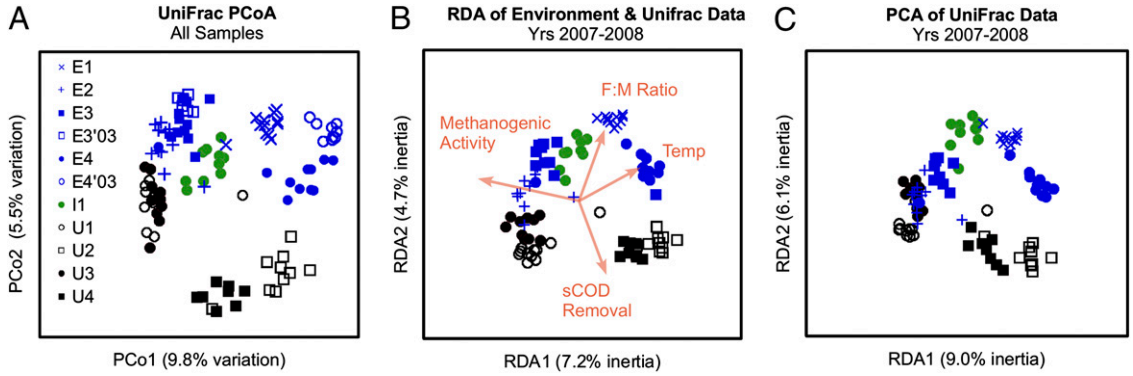

Fig. 1. Phylogenetic distances between samples (i.e., reactor communities) determined via UniFrac principal coordinates analysis (PCoA; average unweighted UniFrac distance matrix was calculated from 100 even rarefactions of 500 sequences per sample). (*A*) PCoA of all samples. (*B*) UniFrac data redundancy analysis constrained by the four highest-loading environment variables: methanogenic activity, soluble chemical oxygen demand (sCOD) removal efficiency, feeding rate normalized to biomass (*F*:*M* ratio), and temperature. (*C*) PCA of UniFrac data for direct comparison with *B*. Sample points formatted by facility location (E1–E4, I1, and U1–U4), sample year ('03 in 2003, others in 2007–2008), and reactor type (blue, EGSB; green, IC; black, UASB).

samples, a graph in two dimensions inherently represents a small proportion of total inertia. The important factor is that the constrained projection along environmental gradients (Fig. 1*B*) successfully replicated the variation in community structure (Fig. 1*C*). The four environmental gradients most useful for constraining RDA in Fig. 1*B* were weighted more favorably (longer vector arrow in Fig. 1*B*) toward measures of community function—the performance variables methanogenic activity and sCOD removal efficiency—than environment—the operating condition variables *F*:*M* ratio and temperature. To avoid overfitting, we chose these four gradients from the full set on the basis of their higher weighting in constrained RDA using all variables (Fig. S2*A*). In addition, we verified our methods by comparing these results to constrained RDA with four randomly generated variables that should not correspond with community structure (Fig. S2*B*). Our results demonstrate that the variation in phylogenetic community structure could be explained by the variation in the four identified environmental gradients. In addition to explaining phylogenetic variation, a similar weighting of environmental gradients also explained some of the variation in the operational taxonomic unit (OTU) relative abundances, as tested by canonical correspondence analysis (CCA) (Fig. S3).

Community dynamics in each facility were analyzed in terms of their time series. A community that undergoes continuous dynamic succession would be expected to progress along a trajectory in a UniFrac time series with each successive time point increasing in distance from the initial community sample. The time series of UniFrac distances within individual locations did not, however, show such a trajectory. Rather, bacterial community dynamics were governed by small, random shifts within the location clusters (small, here, is judged relative to the distances between different facilities). Occasionally, large phylogenetic shifts occurred within a facility, but they were observed to be temporary. For example, Fig. 2 shows a time series from facility U1, which consisted of two community disturbances followed by returns to equilibrium. UniFrac distances from the location average (Fig. 2*A*) followed a flat line, which indicates that the changes in phylogenetic structure as a function of time consisted of random variation, or systematic variation with a tendency to revert to the mean (e.g., the final time point was just as close to the cluster average as other time points).

The two disturbances in community structure in U1 occurred following perturbations in which the sludge loading rate (*F*:*M* ratio) was reduced by ~50% (Fig. 2*C*). In the first case (December 24–January 6), the feeding rate (*F*) decreased due to a holiday break in production. In the second case (June 1), the decrease in *F*:*M* ratio followed a sharp increase in *F*:*M* ratio. The increase was due to an increase in production, whereas the subsequent decrease in *F*:*M* ratio was due to an increase in the total biomass (*M*) in the form of dormant granules from the same facility that were added to the total system average (106% increase in total biomass). This disturbed the average community by adding community components that had been starved for an extensive period. The major shifts in taxonomic groups during each disturbance (Fig. 2*B*) included drastic reductions in the relative abundance of the syntrophic Deltaproteobacteria orders Syntrophobacterales and Desulfuromonadales, concurrent with increases in the relative abundance of Bacteroidetes. Following each disturbance, the community returned to the normal facility cluster (Fig. 2*A*), demonstrating resilience to perturbations in feeding rate and addition of starved biomass. The decrease in relative abundance of Syntrophobacterales for each disturbance was expected, because they specialize in propionate oxidation, which would have had a reduced flux during starvation. The concurrent reduction in the Desulfuromonadales division may have been due to similar sensitivity to limited acetate or alcohol flux or to changes in the dynamics of Fe(III) or elemental sulfur formation, which some of their members use as electron acceptors (18). During each disturbance, the facility maintained optimal performance (Fig. 2 *D* and *E*). All performance and operating condition data for each of the nine facility time series are available in Fig. S4.

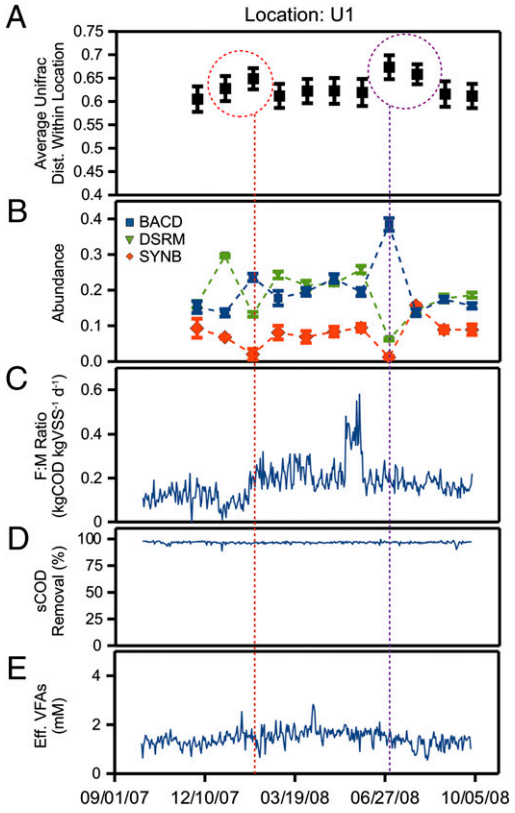

Fig. 2. Time series of sequencing, performance [soluble chemical oxygen demand (sCOD) removal efficiency and effluent volatile fatty acid (VFA) concentration] and operating condition [feeding rate normalized to biomass (*F*:*M* ratio)] data variables for facility U1: (*A*) UniFrac distance from the location average. (*B*) Relative abundance of dynamic taxonomic groups (BACD, Bacteroidetes; DSRM, Desulfuromonales; SYNB, Syntrophobacterales). (*C*) *F*:*M* ratio. (*D*) sCOD removal efficiency. (*E*) Effluent VFA concentration. Vertical dashed lines indicate two significant disturbances in the phylogenetic community structure. Error bars on *A* and *B* represent the SD of 100 and 10 random subsamples, respectively, with 500 sequences per subsample.

In addition to analysis of phylogenetic distances, we also compared time series on the basis of the relative abundance of the 4,962 observed OTUs (defined as 97% sequence similarity via the uclust seed-based algorithm). Due to the size and variability of the OTU table, we identified predictive OTUs using sparse supervised learning, which has been successfully applied to classification of microarray data (19) and 16S amplicon length heterogeneity profiles (20). We chose location of the facility as the class attribute to be predicted, because the community phylogenetic structure was observed to cluster most closely on the basis of facility location, and because facilities were also unique and consistent in terms of performance data (Fig. S5). We ranked OTUs on the basis of their utility for predicting location (measured as *z* score) and selected a subset of the 145 most predictive OTUs using the nearest shrunken centroids (NSC) method (19). Cross validation confirmed that the OTU subset had 96.4% accuracy in predicting sample location origin by NSC. The sample-normalized relative abundances of the predictive OTUs are displayed in Fig. 3*A* as a function of location and time, along with *z* scores indicating their predictive power (Fig. 3*D*). These predictive populations represent OTUs that had stable, consistent

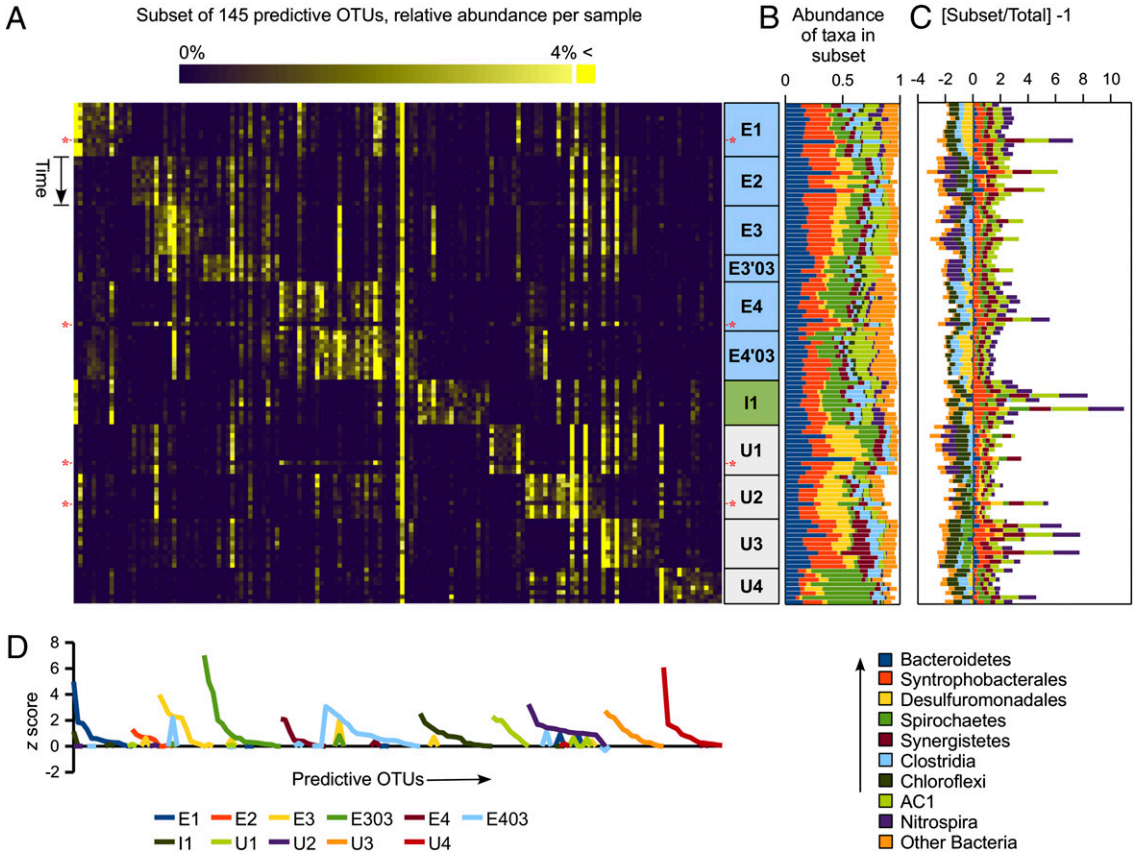

Fig. 3. Subset of OTUs (145 out of 4,962 total OTUs) selected by machine learning as predictive of location (accuracy = 0.964). (*A*) Heat map shows the complete time series (112 samples) ordered vertically per location with only the predictive OTUs. The sampling time progresses downward for each location (time). Each location has a unique OTU composition, sorted horizontally by a *z* score. Relative abundance of OTUs within in each sample are colored in yellow; absent OTUs are in dark blue. (*B*) Microbial composition (abundance of taxa) of each sample for the predictive OTU subset (white space in *B* represents the remainder of OTUs with other classification). (*C*) Relative over-/underrepresentation of taxa denote the change in relative abundance when comparing this predictive OTU subset to all observed OTUs for this study, to summarize taxonomic divisions that were more or less predictive (more or less stable and location specific) than the average OTU. (*D*) *z* scores quantify the predictive ability of OTUs for each location. The four samples that failed to classify correctly are marked with red asterisks.

relative abundance profiles that were unique to a specific facility location or set of locations.

The relative abundance data (Fig. 3*B*) show that the classifiable taxonomic groups represented by the predictive OTU subset consisted mostly of Bacteroidetes, Syntrophobacterales, Desulfuromonadales, Spriochaetes, and small subsets of Synergistia, Clostridia, and Chloroflexi. To determine which classifiable OTUs had more stable or consistent populations, we calculated the over-/underrepresentation in the predictive subset compared with the total relative abundance of taxonomic classifications in the original OTU table (Fig. 3*C*). Syntrophic divisions Syntrophobacterales (an order of Deltaproteobacteria) and Synergistia had significant overrepresentation among the predictive OTUs throughout all locations, indicating that their populations were more stable and consistent than the average OTU. This may have been due to their specialized metabolic functions of short-chain fatty acid oxidation, making their populations less vulnerable to competition from functionally redundant OTUs. During disturbances, such as the data shown in Fig. 2, Syntrophobacterales populations were those most sensitive to perturbation, but they were resilient following the disturbance. Clostridia and Chloroflexi were both significantly underrepresented in the predictive subset, suggesting that their populations were more dynamic than the average OTU, which may be due to the functional redundancy of competing populations. The most overrepresented division, AC1 (a phylum in the Greengenes taxonomy), represents an uncharacterized bacterial phylum found in reducing environments (e.g., anaerobic marine sediment) (21). The relatively high stability and specificity of AC1 OTUs suggest they may fill a specialized niche in methanogenic bioreactors.

The resiliency of the predictive OTUs can be clearly seen in the heat map. For instance, profiles for E1, E4, U1, and U2 all contain atypical horizontal stripes (disturbances), at some point in their time series, for which the machine learning prediction failed (red asterisks in Fig. 3*A*). For each of these samples, at later time points (progressing downward), the subcommunity profile returned to its original composition. The resilience of this predictive OTU subset was representative of resilience in the phylogenetic community structure as a whole (Fig. S4).

There were other measures of community structure that correlated with the measures of function identified as important in constrained RDA. Methanogenic activity increased with increasing evenness of the community measured via the Gini coefficient (normalized scale from 0 to 1; Fig. 4*A*). This agreed with the observations of Wittebolle et al. (15) that greater evenness in bioreactor communities was associated with functional stability. In addition, efficiency of sCOD removal was correlated with phylogenetic variability, which was measured for each sample as its average UniFrac distance to other samples from the facility cluster (Fig. 4*B*). This suggests that community dynamics help maintain efficient performance.

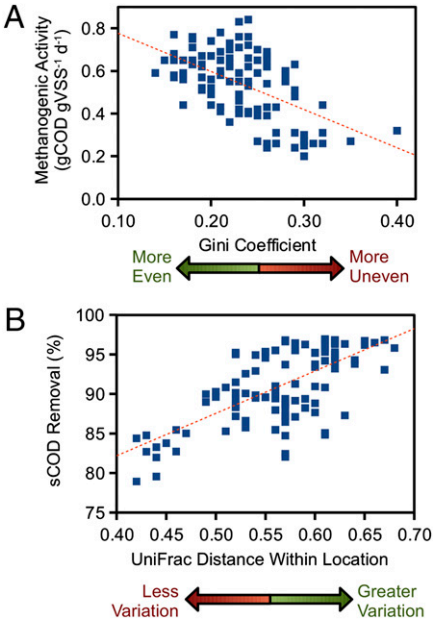

Fig. 4. The importance of redundancy in community function. (*A*) Methanogenic activity of each sample as a function of Gini coefficient (calculated from 100 rarefactions of 100 sequences each; linear regression slope = −1.8 ± 0.3 gCOD gVSS^{-1} d^{-1} SE, R^2 = 0.29). (*B*) sCOD removal efficiency of each sampling time point as a function of the phylogenetic variation from the mean (measured as the UniFrac distance of each time point from the facility average; linear regression slope = 54 ± 6% SE, R^2 = 0.45); methanogenic activity units are reported as g chemical oxygen demand (COD) methane produced per g volatile suspended solids (VSS) biomass used in the activity assay (maximum aceticlastic methanogenesis rate normalized to biomass).

Discussion

Syntrophic Populations Were Highly Stable, Resilient, and Specific for Function and Environment. Our results showed that the ecological dynamics of syntrophic populations (Syntrophobacterales and Synergistia) behaved as expected of functional specialists—populations were stable, resilient, and highly selective along environmental gradients. This observation agrees well with previous genomic sequencing efforts, which have shown that syntrophic bacteria are locked into specific metabolic functions within the overall trophic structure. Syntrophs have relatively small genomes that feature unusual core metabolic pathways and have limited capabilities to process alternative substrates (22). The resilience we observed from these specialized bacteria suggests that a disturbed community is likely to rebound to previous populations of syntrophs, rather than undergo competitive growth of different syntrophs that have similar function. In other words, resilience was relatively important compared with resistance or redundancy for maintaining function of syntrophic populations. The dynamics of higher-level fermenter populations, such as Clostridia and Bacteroidetes, were markedly different from those of the syntrophs, relying more on redundancy to maintain the overall community function. From a practical standpoint, these results suggest that biomass amendments to address failures in syntrophic short-chain fatty acid oxidation may be possible with close consideration of specific OTUs that are most appropriate for the given bioreactor, but attempts to control or maintain specific fermenter populations may be more difficult.

Better Functioning Communities. Our observed relationships may also teach us how to improve the function of a community. We showed that communities with greater evenness had a higher methanogenic activity. Wittebolle et al. (15) had observed that community evenness corresponded with functional stability because a more even community had a higher capacity to use redundant functional pathways, complimenting other observations that communities with more parallel pathways had more efficient response to glucose shock (14). These previous data were collected in short-term batch experiments, and here we show that evenness is also important for full-scale systems operated for years rather than weeks. In addition, we have shown that communities with a greater phylogenetic variability functioned more efficiently. Therefore, the superior performing bioreactors consisted of communities with strategies to incorporate greater functional diversity through greater evenness and variability/dynamics. These performance trends were not apparent when merely looking at species richness.

Relating Community Structure to Function and Environment. We have demonstrated that we can explain variation in bacterial phylogenetic structure with measured environmental gradients (function and environment) using constrained ordination. To quantify these relationships, we merged a time series of high-throughput 16S rRNA gene sequencing data with a massive amount of performance and operating condition measurements for bioreactors with mixed microbial communities. The predictive capabilities of our observed relationships, linking phylogenetic structure to function, were empirical, compared with mechanism-based systems biology approaches that have also seen some success in predicting functional differences among complex microbial communities (23). However, 16S rRNA genes are essentially a proxy for genomic content, which is an underlying reason why phylogeny-based empirical relationships can yield powerful correlations between community structure and function as observed in this study. Similar analyses have the potential to yield informative models for understanding and engineering the function of other microbial communities, such as in the human gut, soils, and oceans. There remain many challenges in relating microbial community structure to function and environment that we must overcome. Factors that have not yet been addressed include the unknown extent of hysteresis (the dependence of current community on the previous community, as opposed to a deterministic relationship with environment) and the undefined time lags between community structure, environment, and function. To address these questions, future advances are needed in the analysis of integrated time series data from different temporal scales.

Methods

Sampling, Amplification, and Sequencing. Granular biomass was sampled monthly from the brewery wastewater treatment facilities listed in Table S1. All reactors sampled were of an upflow configuration with anaerobic granular biomass. Samples were taken from all active reactors at a given facility, at all granule bed sampling ports, and these subsamples were combined as a single time point. Samples were stored at −80 °C before processing. We extracted DNA from all samples in parallel, using the phenol/chloroform method. Barcoded primers (24) were used to amplify bacterial rRNA gene region 27F-338R, optimized for phylogenetic differentiation among pyrosequencing reads (25). Sequencing was performed using the Roche 454 FLX pyrosequencing platform at the Cornell University Life Sciences Core Laboratory Center (SRA029112). Details of DNA processing and sequencing and assay methods for acetate oxidation are available in *SI Methods*.

Analysis of Bacterial 16S rRNA Gene Sequences. Sequences generated from pyrosequencing of bacterial 16S rRNA gene amplicons were processed using the Quantitative Insights Into Microbial Ecology (QIIME v1.1) (26) pipeline, with default settings. Flowgrams were denoised with the Denoiser algorithm, clustered into OTUs at 97% pairwise identity using the seed-based uclust algorithm, and representative sequences from each OTU were aligned to the Greengenes imputed core reference alignment (27) using PyNAST (28). Chimeras (159 total OTUs) were removed from the reference set on the basis of identification as chimeric via Chimera Slayer and confirmation via either Bellerophon or CCode, and verification that the putative chimera appeared in only one sample, using mothur 1.12 (29). The alignment was then filtered to remove gaps and hypervariable regions using a Lane mask,

and an approximately maximum-likelihood tree was constructed from the filtered alignment using FastTree (30). An unweighted UniFrac distance matrix (17) was constructed from the phylogenetic tree and visualized using principal coordinates analysis (PCoA; as implemented in QIIME). Unweighted UniFrac algorithm was chosen because short-term effects may cause fluctuation in relative abundance that can obscure associations of interest. We classified the taxonomy of OTU representative sequences using the naïve Baysian classifier as implemented in mothur 1.12, adapted from the algorithm of the RDP Classifier (31), using the Greengenes sequence database mapped to the Greengenes (July 2010) taxonomy as a training set.

Machine Learning. Daily operating and performance data were obtained from the operator databases at each facility, and an environmental data matrix was constructed from daily reactor data as an average of the 25 d preceding the biological sample. Supervised linear discriminant analysis (LDA) was used to separate locations on the basis of environmental data using the VisRank and FreeVis modules of Orange 2.0b. For analysis of OTUs, the OTU table was trimmed to exclude OTUs occurring in <10% of the samples (<12 samples) or OTUs without a sample containing more than four reads. OTUs were ranked for predicting location via NSC, and the predictive model was cross-validated (sixfold) using the PAMr 1.4 package for R (19) and a threshold of 6.5 (Fig. S6).

Constrained Ordination. Before performing RDA, the UniFrac distance matrix was reoriented using PCoA. The PCoA scores along each axis were then regressed on the set (or subset) of environmental variables, and the resulting "fitted" PCoA matrix was subjected to eigenanalysis via the singular value decomposition (SVD) to obtain the final constrained sample ordination scores. For comparison, we also performed the SVD on the full PCoA matrix (unconstrained RDA). This process of applying RDA to the principal coordinates of a distance matrix is relatively new in computational ecology, also known as "distance-based redundancy analysis" (32) or "canonical analysis of principal coordinates" (33). The advantage of this approach is that it can be combined with any distance metric, including one based on phylogeny, such as UniFrac. The analysis was repeated using the four variables with highest factor loading from the full analysis. Although iron addition had a high factor loading in the full analysis, it was excluded from the reduced analysis because it was only added to UASB-type reactors. In contrast, CCA and unconstrained correspondence analysis (CA) are eigenanalysis approaches that work directly with the OTU relative abundance matrix. To perform CA, the relative abundance matrix was converted to χ^2 distances, fitted to the relative abundance, and then decomposed with the SVD to obtain both sample and OTU scores. CA has the effect of ordinating samples according to a hypothetical environmental gradient that maximizes the dispersion of the OTU modes along that gradient, assuming that each OTU has a unimodal response to the gradient. To perform CCA, the χ^2 distance matrix was first regressed on the canonical environmental variables (as in RDA), before the SVD was applied. CCA has the effect of constraining the hypothetical gradient scores in CA to be a linear combination of the observed canonical variates. In both CA and CCA, the inertia recovered by each axis is a portion of the sum of all sample OTU χ^2 distances. All CA, CCA, and RDA analyses were performed using the Vegan Community Ecology Package 1.17–2 for R.

ACKNOWLEDGMENTS. We thank wastewater treatment operators at Anheuser-Busch, Inbev for supplying samples and data records. We thank Ruth E. Ley and anonymous reviewers for insightful comments. This research was supported by National Research Initiative of the Department of Agriculture (USDA) National Institutes for Food and Agriculture (NIFA) 2004-35504-14896, Cornell University Agricultural Experiment Station federal formula funds NYC-123444 received from the USDA NIFA, a subcontract from GE Global Research funded by DOE DE-FC26-08NT05870, Colorado Center for Biofuels and Biorefining C2B2, and the Howard Hughes Medical Institute. Mass spectrometry resources were funded by National Institutes of Health Grants NIDDK020579, NIDDK05634, and NCRR000954.

1. Angenent LT, Karim K, Al-Dahhan MH, Wrenn BA, Domíguez-Espinosa R (2004) Production of bioenergy and biochemicals from industrial and agricultural wastewater. *Trends Biotechnol* 22:477–485.
2. Lettinga G (1995) Anaerobic digestion and wastewater treatment systems. *Antonie van Leeuwenhoek* 67:3–28.
3. Leitão RC, van Haandel AC, Zeeman G, Lettinga G (2006) The effects of operational and environmental variations on anaerobic wastewater treatment systems: A review. *Bioresour Technol* 97:1105–1118.
4. Briones A, Raskin L (2003) Diversity and dynamics of microbial communities in engineered environments and their implications for process stability. *Curr Opin Biotechnol* 14: 270–276.
5. Rivière D, et al. (2009) Towards the definition of a core of microorganisms involved in anaerobic digestion of sludge. *ISME J* 3:700–714.
6. Curtis TP, Head IM, Graham DW (2003) Theoretical ecology for engineering biology. *Environ Sci Technol* 37:64A–70A.
7. LaPara TM, Nakatsu CH, Pantea LM, Alleman JE (2002) Stability of the bacterial communities supported by a seven-stage biological process treating pharmaceutical wastewater as revealed by PCR-DGGE. *Water Res* 36:638–646.
8. Malin C, Illmer P (2008) Ability of DNA content and DGGE analysis to reflect the performance condition of an anaerobic biowaste fermenter. *Microbiol Res* 163: 503–511.
9. Goberna M, Insam H, Franke-Whittle IH (2009) Effect of biowaste sludge maturation on the diversity of thermophilic bacteria and archaea in an anaerobic reactor. *Appl Environ Microbiol* 75:2566–2572.
10. McMahon KD, Stroot PG, Mackie RI, Raskin L (2001) Anaerobic codigestion of municipal solid waste and biosolids under various mixing conditions—II: Microbial population dynamics. *Water Res* 35:1817–1827.
11. Hori T, Haruta S, Ueno Y, Ishii M, Igarashi Y (2006) Dynamic transition of a methanogenic population in response to the concentration of volatile fatty acids in a thermophilic anaerobic digester. *Appl Environ Microbiol* 72:1623–1630.
12. Fernández A, et al. (1999) How stable is stable? Function versus community composition. *Appl Environ Microbiol* 65:3697–3704.
13. Zumstein E, Moletta R, Godon JJ (2000) Examination of two years of community dynamics in an anaerobic bioreactor using fluorescence polymerase chain reaction (PCR) single-strand conformation polymorphism analysis. *Environ Microbiol* 2:69–78.
14. Hashsham SA, et al. (2000) Parallel processing of substrate correlates with greater functional stability in methanogenic bioreactor communities perturbed by glucose. *Appl Environ Microbiol* 66:4050–4057.
15. Wittebolle L, et al. (2009) Initial community evenness favours functionality under selective stress. *Nature* 458:623–626.
16. Allison SD, Martiny JBH (2008) Colloquium paper: Resistance, resilience, and redundancy in microbial communities. *Proc Natl Acad Sci USA* 105(Suppl 1):11512–11519.
17. Lozupone C, Knight R (2005) UniFrac: A new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* 71:8228–8235.
18. Roden EE, Lovley DR (1993) Dissimilatory Fe(III) reduction by the marine microorganism *Desulfuromonas acetoxidans*. *Appl Environ Microbiol* 59:734–742.
19. Tibshirani R, Hastie T, Narasimhan B, Chu G (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci USA* 99:6567–6572.
20. Yang CY, et al. (2006) An eco-informatics tool for microbial community studies: Supervised classification of Amplicon Length Heterogeneity (ALH) profiles of 16S rRNA. *J Microbiol Methods* 65:49–62.
21. Dhillon A, Teske A, Dillon J, Stahl DA, Sogin ML (2003) Molecular characterization of sulfate-reducing bacteria in the Guaymas Basin. *Appl Environ Microbiol* 69: 2765–2772.
22. McInerney MJ, Sieber JR, Gunsalus RP (2009) Syntrophy in anaerobic global carbon cycles. *Curr Opin Biotechnol* 20:623–632.
23. Röling WFM, Ferrer M, Golyshin PN (2010) Systems approaches to microbial communities and their functioning. *Curr Opin Biotechnol* 21:532–538.
24. Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat Methods* 5: 235–237.
25. Liu Z, Lozupone C, Hamady M, Bushman FD, Knight R (2007) Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Res* 35:e120.
26. Caporaso JG, et al. (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7:335–336.
27. DeSantis TZ, et al. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* 72:5069–5072.
28. Caporaso JG, et al. (2010) PyNAST: A flexible tool for aligning sequences to a template alignment. *Bioinformatics* 26:266–267.
29. Schloss PD, et al. (2009) Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 75:7537–7541.
30. Price MN, Dehal PS, Arkin AP (2010) FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5:e9490.
31. Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 73:5261–5267.
32. Legendre P, Anderson MJ (1999) Distance-based redundancy analysis: Testing multispecies responses in multifactorial ecological experiments. *Ecol Monogr* 69:1–24.
33. Anderson MJ, Willis TJ (2003) Canonical analysis of principal coordinates: A useful method of constrained ordination for ecology. *Ecology* 84:511–525.