

ORIGINAL ARTICLE

Two-stage microbial community experimental design

Timothy L Tickle^{1,2}, Nicola Segata¹, Levi Waldron^{1,3}, Uri Weingart¹ and Curtis Huttenhower^{1,2}

¹Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA; ²The Broad Institute of MIT and Harvard, Cambridge, MA, USA and ³The Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA, USA

Microbial community samples can be efficiently surveyed in high throughput by sequencing markers such as the 16S ribosomal RNA gene. Often, a collection of samples is then selected for subsequent metagenomic, metabolomic or other follow-up. Two-stage study design has long been used in ecology but has not yet been studied in-depth for high-throughput microbial community investigations. To avoid *ad hoc* sample selection, we developed and validated several purposive sample selection methods for two-stage studies (that is, biological criteria) targeting differing types of microbial communities. These methods select follow-up samples from large community surveys, with criteria including samples typical of the initially surveyed population, targeting specific microbial clades or rare species, maximizing diversity, representing extreme or deviant communities, or identifying communities distinct or discriminating among environment or host phenotypes. The accuracies of each sampling technique and their influences on the characteristics of the resulting selected microbial community were evaluated using both simulated and experimental data. Specifically, all criteria were able to identify samples whose properties were accurately retained in 318 paired 16S amplicon and whole-community metagenomic (follow-up) samples from the Human Microbiome Project. Some selection criteria resulted in follow-up samples that were strongly non-representative of the original survey population; diversity maximization particularly undersampled community configurations. Only selection of intentionally representative samples minimized differences in the selected sample set from the original microbial survey. An implementation is provided as the microPITA (Microbiomes: Picking Interesting Taxa for Analysis) software for two-stage study design of microbial communities.

The ISME Journal (2013) 7, 2330–2339; doi:10.1038/ismej.2013.139; published online 15 August 2013

Subject Category: Integrated genomics and post-genomics approaches in microbial ecology

Keywords: 16S; microbial; microbiome; sampling; sequencing; two-stage

Introduction

It is now possible to survey hundreds of microbial community samples cost-effectively using multiplexed high-throughput sequencing (Qin *et al.*, 2010, Yatsunenko *et al.*, 2012, The Human Microbiome Project Consortium 2012b). Such approaches typically target amplicons from taxonomic markers such as the 16S ribosomal RNA (rRNA) gene (Pace *et al.*, 1986, Bartram *et al.*, 2011, Werner *et al.*, 2012). Multiplexing provides up to tens of thousands of reads per sample using Roche 454 or, increasingly, Illumina technologies currently achieve total costs well below \$100 per sample (Bartram *et al.*, 2011, Werner *et al.*, 2012). Marker

gene sequencing only identifies which organisms are present in a community, however, and provides only indirect insight into their metagenomic potential or biological activities such as transcription or metabolism. These are instead assayed by, for example, shotgun sequencing or metabolomics, which are much more costly and thus practical to apply only to a subset of available samples. This has led to a proliferation of tiered or two-stage study designs for microbial communities, in which a subset of 16S rRNA gene surveys from a large population is selected for follow-up in a targeted manner (Mackelprang *et al.*, 2011, Yatsunenko *et al.*, 2012; The Human Microbiome Project Consortium, 2012b).

Two-stage studies are often used in ecological research to control the costs of large monitoring projects (Olsen *et al.*, 1999), to adaptively sample for rare species (Brown *et al.*, 2008) or to sample from environments with the goal of comprehensive sampling taking into account multidimensional

Correspondence: C Huttenhower, Department of Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA.

E-mail: chuttenh@hsph.harvard.edu

Received 18 February 2013; revised 10 July 2013; accepted 16 July 2013; published online 15 August 2013

environmental descriptions (Danz *et al.*, 2005). Here, we develop and evaluate several methodologies to select microbial communities with specific characteristics or biological criteria (that is, purposive sampling) for follow-up based on 16S rRNA gene sequencing (hereafter abbreviated 16S). As two-stage sample selection in metagenomic studies has been either *ad hoc* or limited to criteria such as maximum ecological diversity (The Human Microbiome Project Consortium, 2012b; Yatsunenko *et al.*, 2012), these methods can be used in two-stage study designs with the goal of reducing experimental costs, often dramatically. The Human Microbiome Project (HMP), for example, contained over 5000 16S-sequenced samples (by Roche 454), only ~700 of which were selected for Illumina shotgun metagenomic sequencing. At the sequencing depths used by the HMP, these two sample sets represent roughly equal experimental costs despite a 10-fold difference in size, with follow-up samples chosen primarily by experts based on prior biological knowledge. Of the 531 Illumina 16S samples in Yatsunenko *et al.*, a subset of only 110 samples were assessed metagenomically by 454 pyrosequencing, with no specific selection criteria stated. These microbial community studies and others (Claesson *et al.*, 2012, Yang *et al.*, 2012) have all used two-stage designs without quantitative criteria for sample selection, emphasizing a need to formalize and validate such a process to prevent disparities between whole-population and follow-up microbiome sample sets.

The microbial community characteristics that may be of interest in any particular investigation are highly dependent on its research goals. We thus determined the effects of four unsupervised selection methodologies for purposive sampling: maximum ecological diversity, representative dissimilarity, most dissimilarity, and targeting specific taxa or clades. Maximum diversity selection targets biodiversity hotspots of microbial communities with the highest within-sample α -diversity (Hamady and Knight, 2009). Representative (least biased) and extreme sample dissimilarity instead assesses between-sample β -diversity; the former then chooses representative samples evenly tiled throughout the available sample space, while the latter identifies deviant or extreme samples mutually furthest from each other and from any neighbors. Respectively, this selects a community typical of the original survey, or selects samples with the most different communities found in the survey. Selection targeting a taxon (or clade) of interest can include either samples with the greatest abundance of that taxon or samples where the taxon is of greatest rank abundance, which differ subtly in their biological goals.

Many study designs include samples labeled or stratified into separate microbial populations, for example, 'case' and 'control' or other discrete phenotypes that can be taken into account during selection.

Selection of samples with microbes that may classify microbial communities based on a host phenotype or environmental variable can be performed with two additional supervised sampling methods. These complementary techniques identify samples either minimally distinctive or maximally discriminative among environments or host phenotypes. The former identifies microbial communities that separate two or more phenotypes of interest with either the smallest (discriminative) or largest (distinctive) changes in microbial community structure. Although other selection methods (α -diversity, β -diversity and taxon targeting methods) are unpurposive and do not directly use sample labels, they can all also be optionally applied within multiple groups stratified by phenotype labels.

We evaluated these six microbial community selection methods in three ways. First, and most simply, their reproducibility and accuracy in synthetic microbiomes proved to be uniformly high. Second, selection bias in biological samples relative to the initial microbial community survey varied dramatically among methods. Targeting specific taxa of course biased toward those clades, but commonly selected microbial communities exhibited over- and under-sampling of many other clades because of the initial 16S survey's community structure. Maximizing diversity tended to most strongly bias the resulting communities away from a typical subset of the population, and only intentionally representative selection criteria avoided biases. Finally, 318 paired 16S and metagenomic samples from the HMP were assessed using all methods to ensure the community characteristics of 16S survey selections (diversity, abundance of specific taxa, and so on) were indeed preserved in 'follow-up' metagenomic assays. All analysis and evaluation methods are implemented as the open source microPITA (Microbiomes: Picking Interesting Taxa for Analysis) software, available for download and for online use through Galaxy (Blankenberg *et al.*, 2001) at <http://huttenhower.sph.harvard.edu/micropita>.

Materials and methods

Synthetic microbial communities

The synthetic operational taxonomic units table used to evaluate selection methods (Supplementary Table 1) consisted in each instance of 48 samples with 224 taxa. Sixteen samples simulated low variance, highly complex communities by setting 25% of taxa to a 'read' count of 50. Fourteen samples were designed to simulate moderate dissimilarity among themselves, with non-overlapping communities of 16 taxa each set to abundance counts of 50. Fourteen samples exhibited high dissimilarity among themselves with a twofold increase in abundance over other sample classes (8 features with counts of 100 each). These high dissimilarity samples contained taxa in blocks that overlap with a block of the moderately dissimilar

samples to produce samples with consistent but more extreme variation. To simulate rare community members, four samples contained exclusively three separate taxa, each with abundance counts of 50. Each of these taxa was also contained in three moderately dissimilar samples and in one highly dissimilar sample but in none of the samples simulating high ecological diversity. Supervised labels were defined based on the left- and right-most extremes of the resulting first ordination component (see Figure 3).

When evaluating selection methods with noise, 10 separate data sets following the pattern in Supplementary Table 1, and additional noise was simulated by randomly shuffling 5% of synthetic read counts randomly among the different taxa of each sample. An additional 10 data sets were generated by randomly shuffling 10% of the synthetic reads.

16S and metagenomic HMP data

A total of 5516 16S samples from the HMP that were profiled using mothur (Schloss *et al.*, 2009) were downloaded from <http://hmpdacc.org/HMMCP> and 682 MetaPhlAn (Segata *et al.*, 2012) metagenomic profiles downloaded from <http://hmpdacc.org/HMSMCP>. Our analysis used the high-quality 97% operational taxonomic units, which were summarized at the genus level when necessary for comparison among data sets. These comprise samples collected by the HMP (The Human Microbiome Project Consortium, 2012a) from the seven paired body sites: stool, anterior nares, posterior fornix, bilateral retroauricular crease, supragingival plaque, buccal mucosa and tongue dorsum. Briefly, 16S rRNA gene samples were pyrosequenced (Roche Diagnostics Corporation, Indianapolis, IN, USA) using the V3-5 hypervariable region targeting at least 3000 reads per sample. Whole-community shotgun sequencing was performed using the Illumina GAIIx platform at a depth of approximately 5 billion-nucleotides total sequence per sample (Illumina, Inc, San Diego, CA, USA), from which human reads were depleted in silico. 16S and metagenomic sample pairing was performed by matching on the dbGaP project 'Parent_Specimen' Parent Sample ID for first visits only, resulting in a total of 318 quality-controlled paired samples used here. Specific data files used in this publication (both synthetic and biological) are available online at <http://huttenhower.sph.harvard.edu/micropita>.

Unsupervised sample selection methods

Unsupervised methods include the following: maximum diversity, most representative, most dissimilar and microbe-driven targeted feature selection. Maximum diversity selection occurs by ranking all samples in decreasing order by the inverse Simpson α -diversity index (Simpson, 1949). Representative selection identifies samples from a pairwise

symmetric Bray–Curtis dissimilarity matrix, using k-medoids clustering to tile dissimilarity space into clusters of which the samples nearest each centroid are selected. K-medoid clustering was used as implemented in Machine Learning Python (MLPY) 2.2.0 (Albanese *et al.*, 2012) with the number of clusters set to the number of samples to select. Dissimilar selection occurs by building a sample dissimilarity matrix using the additive inverse of Bray–Curtis dissimilarity, followed by agglomerative hierarchical clustering (again from MLPY) to build a dendrogram of sample relationships; the n most terminal nodes are then reported. Microbe-driven feature targeting is determined either by the top ranked samples based on the average abundance of targeted microbes, or on the rank order of targeted microbes within each sample. All methods, both unsupervised and supervised, are applied on count data normalized to proportional relative abundances. Clades were normalized by dividing their sequence count by the total counts of all reads in the sample. This standard normalization allows samples with different read depths to be compared. Other α - and β -diversity measures can be specified by the user for any of these ranking criteria.

Supervised sample selection methods

Two supervised methods were developed for labeled samples, discriminant and distinct. Discriminant selection identifies samples most similar to each other but still classified as different phenotypes. Distinct selection targets samples very different from all samples in other phenotypes or labels. Given a phenotypically labeled sample, Bray–Curtis dissimilarities are calculated from the average sample (centroid) of all other class. The samples in which the sum of these averages is smallest are selected in ascending order for discriminant samples; the largest sums are selected in decreasing order as distinct samples.

In supervised selection from the HMP, all posterior fornix examples used two classes with vaginal pH above or below 4.0. Stool samples used two classes with body mass index above or below 30 kg m^{-2} and body sites were classified when compared (for example, plaque versus mucosa).

Results

Validation of six biologically motivated criteria for purposive sample selection in two-stage study designs for microbial communities

To illustrate and validate the six criteria for microbial community selection used in microPITA, several synthetic communities were generated. Each data set comprised 48 samples, in total containing 224 synthetic taxa (counts). Among these samples, one group was constructed to exhibit high ecological diversity (15 samples), two groups were given differential abundance in blocks of 16 or 8 taxa to

induce dissimilarity among themselves (14 samples each, the smaller block with a higher average of read counts to exhibit more extreme dissimilarity), and one group exclusively containing a small set of rare taxa to be targeted by selection (eight samples, Supplementary Figure 1, Supplementary Table 1). An example of one such resulting data set in which subsets of six samples each were selected by the four unsupervised criteria demonstrates the methods to be operating as expected (Figure 1a).

To quantify these methods' reproducibility, 10 such synthetic communities were generated. Each community was produced by swapping 5% of a sample's counts with a random taxon in that sample (Figure 1b). At this and increasing levels of noise (Supplementary Figure 2), at most an average of 1.1 errors (over the 10 data sets) were encountered in selecting the samples expected for each criterion based on the data set's synthetic construction. At sufficiently high synthetic noise levels, all selection methods incurred at least one error, but rarely more; see Figures 1a and 3. Overlap between the four unsupervised and two supervised criteria occurred as expected, with representative dissimilarity in particular correctly selecting samples in common with all other methods. As the samples containing an outlier group of rare taxa were also significantly unusual, they were also often chosen when maximizing dissimilarity. Neither these dissimilar selections nor the targeted selection of taxa overlapped samples selected for high diversity; this is attributed to the strongly increased abundance of a few clades reducing overall community diversity. Finally, supervised distinct and discriminative methods selected samples from the innermost (diverse

and most similar) cluster and outermost (most different) taxon-specific ring, respectively, as expected (Supplementary Figures 3–4).

Some selection criteria can result in strong sampling biases in microbial communities

Although these results show that several reasonable biological criteria can be used to accurately select samples for follow-up, they do not indicate what biases in community composition might be induced by each method. To investigate this, we used 101 16S surveys of the posterior fornix from the HMP, comprising in total 294 genus-level phylotypes and 373 features at all taxonomic clade levels (The Human Microbiome Project Consortium, 2012a). Clades differentially represented between the selected samples and the total population were determined by LEfSe (Segata *et al.*, 2011) for each selection criterion. Low (<4) versus high (≥ 4) vaginal pH was used as a phenotype (The Human Microbiome Project Consortium, 2012b) for the two supervised methods (discriminative and distinct), and *Prevotella* was used as a targeted clade (Ravel *et al.*, 2011). Even in this relatively low-complexity community, each method resulted in strikingly different biases in the microbial communities selected for follow-up from the initial survey (Figure 2).

Representative selection of samples produced the fewest biases, an important consideration for two-stage study designs in which an in-depth sample is intended to accurately reflect the larger microbial community population. In this example, a slight under-representation of Lactobacillales

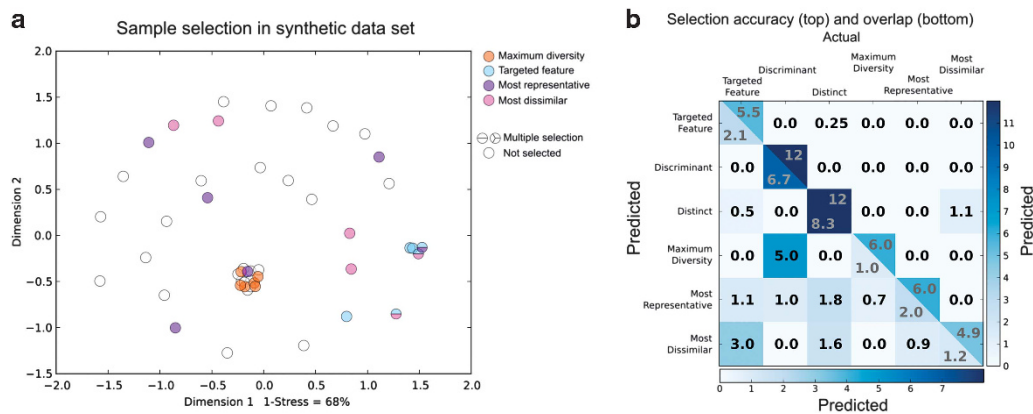


Figure 1 Demonstrations, accuracies and similarities of purposive sample selection methods for two-stage microbial community study design. **(a)** A representative synthetic microbial community comprising 48 samples with 224 taxa. Principle coordinates analysis using Bray–Curtis dissimilarity shows six samples selected by each of the four unsupervised criteria (see Supplementary Figure 1 for data set design, Supplementary Figure 5 for supervised criteria). Selections indicate correct example identifications of synthetic samples enriched for: maximum diversity (central, complex samples); targeted features (right, containing specific ‘microbes’); dissimilar (outside central cluster); and representative (evenly throughout sample space). **(b)** Matrix of mis-selections (upper triangle) and cross-selections (lower triangle) across methods. Upper triangle shows the average number of correctly chosen samples for each method (out of 6 possible for unsupervised and 12 for supervised) in 10 such synthetic data sets each containing 5% randomly shuffled noise counts over all taxa (confusion matrix). The lower triangle indicates the overlap between selection methods, with the diagonal including samples not selected by any other method. Samples are counted more than once when chosen by more than one method, thus permitting column or row sums >6 or 12 for supervised methods (six samples per two labels).

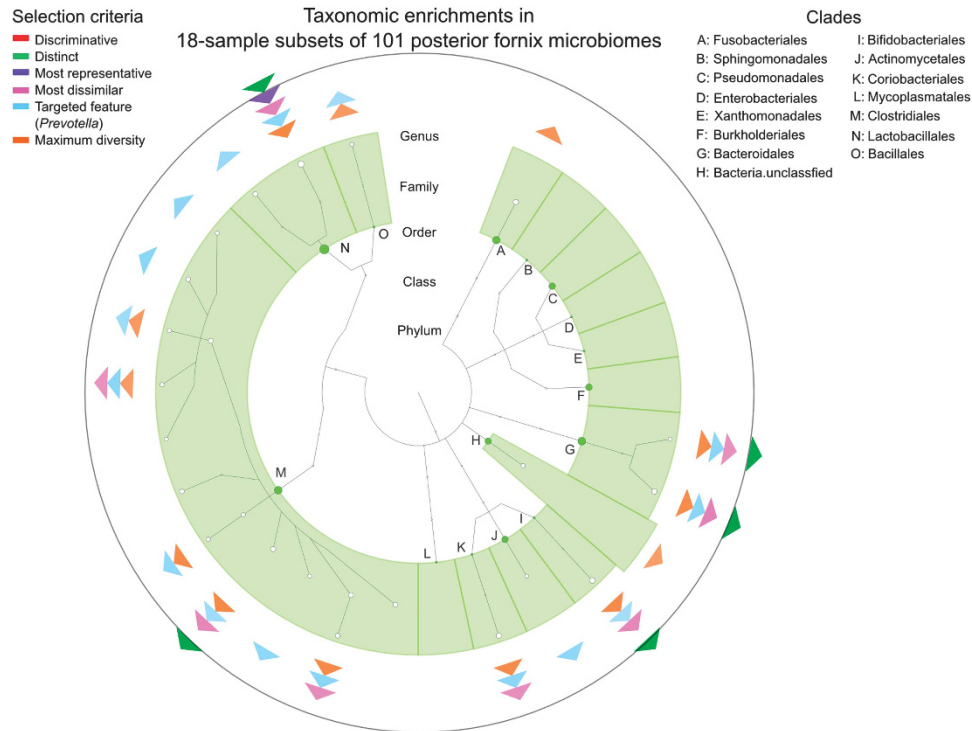


Figure 2 Significant differences in relative microbial abundances resulting from selection criteria applied to vaginal microbiomes. Cladogram of genus-level and above taxa appearing in 101 posterior fornix 16S surveys from the HMP with at least five reads in at least five samples (The Human Microbiome Project Consortium, 2012a). Outer triangles indicate over- (up/away) or under-representation (down/toward) of taxa in the subset of 18 samples selected by each unsupervised method (18 per class in the supervised distinct and discriminative methods, classified by vaginal pH <4). Node circle size indicates average relative abundance of each clade, with clades included if enriched or if at least five descendant genera exceeded the 95th percentile of abundance in 5% of samples. Biased representation was determined by LEfSe (Segata *et al.*, 2011) followed by false discovery rate correction to a q -value of <0.2.

occurred because of the homogeneity of *Lactobacillus* dominant vaginal samples in the HMP population, which frequently comprises >90% of the community in these microbiomes. When only 18 samples are selected, ensuring a sufficient representation of samples containing other taxa results in a slight overall Lactobacillales depletion that is abrogated by increasing the number of communities selected for follow-up. When using other sample selection methods, the genus *Lactobacillus* and a diversity of taxa including *Prevotella* tend to be anticorrelated and associated with low and high vaginal pH, respectively (Ravel *et al.*, 2011). This results in the most diverse samples and those targeting *Prevotella* enriching for these clades relative to the *Lactobacillus*-dominated majority. Discriminative and distinct samples were in this easily visualized example too small in number to achieve significance, but introduced similar biases when more samples were chosen (Supplementary Figures 6–7). Otherwise, selecting increasing numbers of samples for follow-up unsurprisingly reduced taxonomic biases relative to the initially surveyed population for all criteria; this was most rapidly seen for representatively selected communities.

Case-control and stratified study designs allow selection of microbial communities most discriminative or distinct within a phenotype

Microbial community surveys often group populations into strata present in certain environmental conditions or host phenotypes. These supervised selection methods provide a way to integrate meta-data or sample labels collected during a study's first stage and to use that stage's data to more specifically target its second stage. Such stratification has been used in case-control studies investigating conditions such as inflammatory bowel disease (Frank *et al.*, 2007; Willing *et al.*, 2010; Morgan *et al.*, 2012), age and geographic origin (Yatsunenکو *et al.*, 2012), or simply different body site habitats (The Human Microbiome Project Consortium, 2012b). Any of the unsupervised methods can be applied within multiple such strata. As an example, inflammatory bowel disease is an inflammatory disease of the bowel and colon with known associations with the gut microbiome and, most often, presenting as the subtypes Crohn's disease or ulcerative colitis. Rather than following up on the six most globally diverse communities in a case-control inflammatory bowel disease study, the study can be stratified by phenotype so that the two most diverse samples from

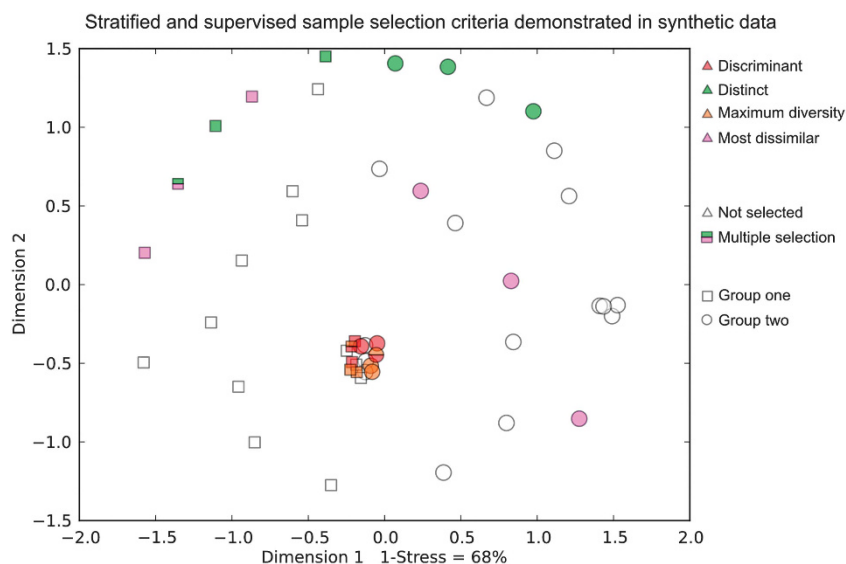


Figure 3 Microbial communities selected using stratification within or supervised labeling of phenotypes. Marker gene surveys often group subsets of samples by phenotypes such as case–control (for example, disease/healthy). Here, 48 synthetic samples include two groups (arbitrarily ‘one’ and ‘two’) defined spatially left and right on the ordination. Any unsupervised community selection criteria can be applied within groups, for example, the three most ecologically diverse or most dissimilar samples within each group rather than the six most globally diverse. In addition, supervised selection can be used to identify samples most subtly (discriminant) or extremely (distinct) predictive of phenotype, here defined by distance to other phenotypes’ centroids (see Materials and methods). Samples are plotted by principal coordinates ordination using Bray–Curtis dissimilarity; Supplementary Figure 9 plots each stratum separately. Discriminant selection correctly identifies samples that separate groups with minimal changes in community structure (central cluster). Distinct selection likewise correctly selects samples in the two groups with the greatest magnitude of overall community change (outermost rings).

healthy controls, Crohn’s disease and ulcerative colitis inflammatory bowel disease cases can be selected separately (Supplementary Figure 8). This is particularly important when reproducible biomarkers of a phenotype are to be determined from metagenomic or metabolomic follow-up, since as indicated above only representative sampling within strata assures an unbiased and balanced experimental design. In addition, it is of increasing interest to determine microbial biomarkers explicitly predictive of groupings such as disease status (Knights *et al.*, 2011). For this application, we separately investigated the two explicitly supervised ‘distinct’ and ‘discriminant’ criteria (Figure 3).

Distinct and discriminant samples are respectively defined as those that most extremely differentiate and most subtly (but consistently) differentiate two or more groups of microbial communities. Although distinct samples may contain the most unique microbial populations for a phenotype, discriminant samples may contain communities with minimal differences crucial to one or another phenotypic state. To identify such samples, we measured the average Bray–Curtis distance of a labeled community from the centroid of all other samples. Distinct samples are thus calculated as the samples most distant from other groups of samples, and discriminative samples are closest. In the synthetic data set, we defined two groups or phenotypes corresponding to distinct spatial positioning after

Bray–Curtis ordination. Owing to the synthetic data set’s construction, the communities differing least between these two ‘phenotypes’ are the 16 highly diverse samples, which share in common many minimally varying taxa. Discriminant selection correctly identifies samples from among these that differ in only a few out of the many ‘taxa’ in these samples. Conversely, the most distinct samples are selected from the dissimilar and rare taxon synthetic samples, which differ widely in taxa, thus representing in the best case be the clearest causes of a true phenotype or in the worst be unusual outliers.

Sample size to achieve a characteristic subset of microbial communities varies by method

In addition to enrichment of specific clades, maintenance of ecological parameters such as diversity or richness can be evaluated in study designs targeting a sampled subset of communities for follow-up. In particular, obtaining a collection of samples with representative community diversity is highly dependent on the selection method (Figure 4). Methods such as maximization of ecological diversity will tend to maximize richness and observe new taxa at a much greater rate than other techniques (for example, dissimilar or distinct samples). This will result in a subsample with an overall diversity comparable to that of the initial population in fewer samples, if this is of interest for follow-up. Targeting only

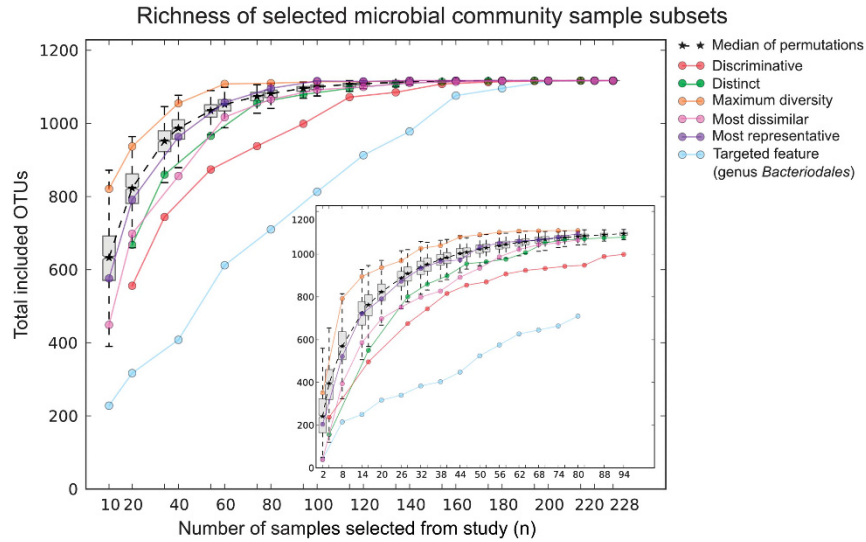


Figure 4 Richness of microbial communities when selected by each sampling criterion. Collectors curves for total observed operational taxonomic units (OTUs) showing the median richness of selections increasing in sample size to 228 total gut microbiome samples. Supervised selection is by body mass index above or below 30 kg m². Dashed line indicates the median permuted occurrence at each sample count (that is, standard rarefaction curve).

Criteria selected in HMP 16S surveys retained in metagenomic follow-up

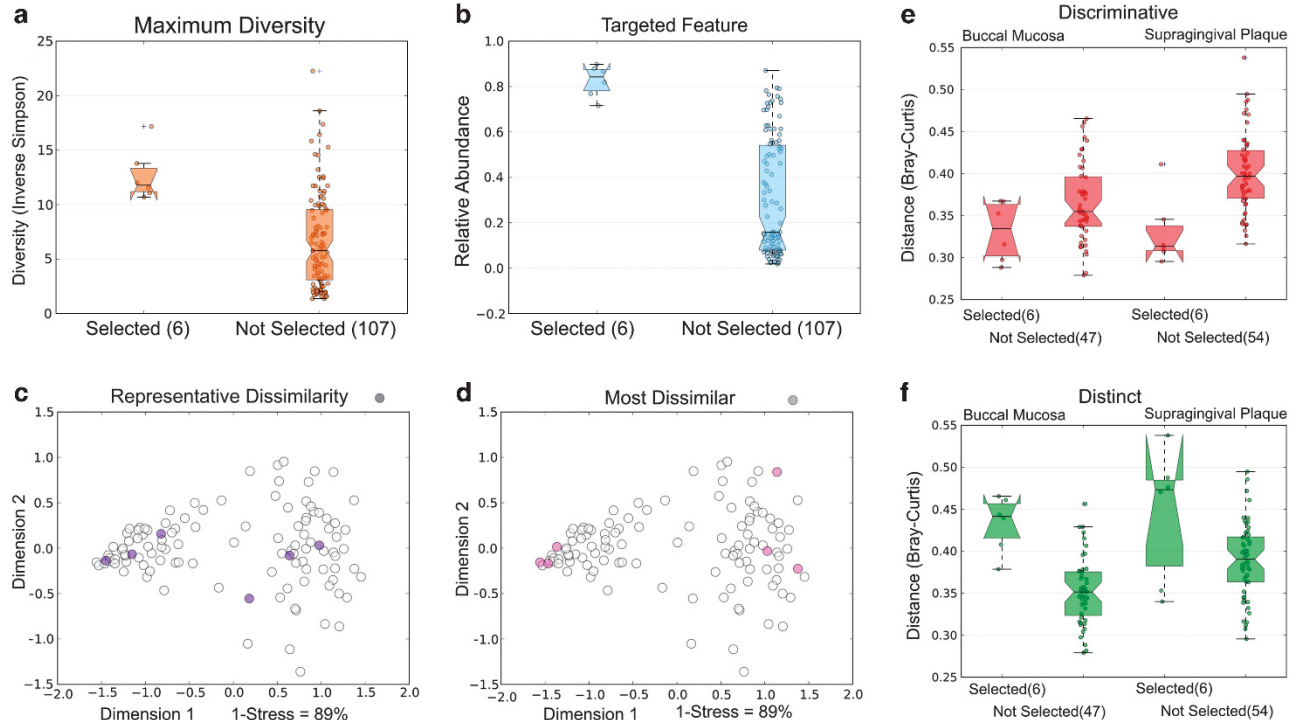


Figure 5 Consistency in microPITA selections from paired 16S and follow-up metagenomic human microbiomes. Paired selection is shown with 113 16S surveys from combining both the HMP supragingival plaque and buccal mucosa body sites (The Human Microbiome Project Consortium, 2012a). Unsupervised selections target (a) maximum ecological diversity, (b) taxon- or clade-specific selection (*Bacteroidales*), (c) representative communities and (d) most extreme or dissimilar communities. Supervised selection was applied to classify between body sites based on (e) discriminative and (f) distinct communities. Box and ordination (Bray-Curtis) plots show the indicated criteria as measured only in shotgun data (Segata *et al.*, 2012) to validate that community characteristics targeted during survey sample selection are retained during second-stage follow-up.

specific taxa or clades conversely underestimates the diversity of the community outside of those clades. Representative sampling tracks the median

rate of incorporation of diversity, again indicating the least bias relative to an accumulation curve from the initial, complete survey.

MicroPITA selects communities that correctly retain characteristics among 16S and metagenomic samples from the HMP

For two-stage follow-up investigations to be useful, samples selected using a purposive strategy in the first-stage survey should correspond to samples that continue to exhibit the desired characteristics in second-stage follow-up. The HMP included a large two-stage component, in which some 5100 16S surveys were followed-up with approximately 700 shotgun metagenomic samples (The Human Microbiome Project Consortium, 2012a). This provided microPITA with a very large biological test set in which the characteristics targeted by each purposive sampling criterion (maximum ecological diversity, abundant taxa, and so on) could be confirmed in the samples selected for the second stage. We used the HMP's mothur (Schloss *et al.*, 2009) 16S community profiling and MetaPhlAn (Segata *et al.*, 2012) shotgun metagenomic profiling to determine taxonomic abundances and diversity in paired samples spanning the gut, skin, posterior fornix, nasal, supragingival plaque, buccal mucosa and tongue body site habitats (see Materials and methods). For each selection method, we validated that community characteristic targeted while selecting from the survey remained near-optimal in paired second-stage metagenomic data.

For all methods, the intended criteria were preserved in the second-stage (metagenomic) samples selected only using first-stage (16S) data (Figure 5). One validation of this behavior is to ensure that samples identified based on first-stage taxonomic profiling retain appropriate ecological or metadata-linked characteristics during a metagenomic second stage, in addition to preserving taxonomic composition as expected. Communities identified as most ecologically diverse using a 16S-based survey remained so when re-analyzed by metagenomic species identification (Figure 5a), and taxa or clades of interest abundant based on 16S identification remained so during follow-up (Figure 5b). More importantly, supragingival plaque and buccal mucosa microbial communities chosen to be representative (Figure 5c) or to capture greatest dissimilarity (Figure 5d) remained representative or extreme in metagenomic data. This permits features such as metabolic potential to be explored accurately during follow-up using functional profiling analysis methods (Abubucker *et al.*, 2012). Samples were grouped by body site habitats for supervised selection, here buccal mucosa and the supragingival plaque (Figures 5e and f), demonstrating that the ecological similarities of these habitats in 16S data remained consistently close to those measured in metagenomic profiling. Taxon or clade selection (Figure 5b) is of particular interest in these examples, as identifying communities enriched for a clade can be performed either by abundance (retaining communities in which the taxon is of greatest relative abundance) or by rank (retaining communities in which it represents the

greatest plurality, regardless of abundance). Both of these techniques are implemented in microPITA and validated in these data (Supplementary Figure 10). Thus in this large human microbiome test set, microbial communities selected from 16S marker gene surveys using microPITA retain their intended characteristics when subsequently assessed using in-depth shotgun metagenomic sequencing.

Discussion

Microbial community studies, both of the human microbiome and of environmental habitats, are increasingly turning to larger sample sizes surveyed using shallow sequencing of the 16S taxonomic marker gene (Kuczynski *et al.*, 2012). Such surveys must be followed-up by in-depth second-stage profiling using metagenomic, metatranscriptomic, or other functional assays in order to characterize the biomolecular or microbiological mechanisms of changes in community structure. As these assays are typically an order of magnitude more expensive and lower throughput than marker gene surveys, we have developed and validated methods focused on selecting communities exhibiting six different characteristics for use in follow-up. Not only did these methods perform as expected in synthetic microbial communities, but properties such as ecological diversity or the prevalence of specific targeted taxa when selected in HMP 16S samples were consistently validated in paired metagenomic follow-up assays.

The differences in cost offered by two-stage study design of microbial communities are striking. Although costs associated with high-throughput sequencing are in constant flux, approximate representative amounts might be \$50 per sample for 16S profiling and \$1000 for shotgun metagenomic sequencing. Metabolomic and metatranscriptomic costs increase further. A fixed budget that allowed only 100 samples to be profiled metagenomically would, for the same cost, accommodate some 1000 surveys in a first stage of 16S profiling followed by only a minor reduction to 50 metagenomically sequenced samples. This small reduction in turn provides the benefits of a full population survey in the first stage and guaranteed targeting of samples or particular interest in the second stage. One could of course allocate resources unequally among the first and second stages depending on the needs of the study. If this example represents a case-control study, a two-stage design increases the sample size for identifying statistically significant whole-community perturbations by 10-fold, while only reducing the size for subsequent metagenomic mechanistic characterization by 2-fold. Given the consistently high variability of host-associated microbial communities across populations (The Human Microbiome Project Consortium, 2012b), this will in almost all cases represent a favorable tradeoff for power and reproducibility.

Some of the study design strategies enabled by microPITA are related to those offered by other popular resources for ecological research, including the vegan and stratification (Baillargeon and Rivest 2011) R packages. Vegan computes various α -, β -, γ -diversity metrics, as well as, variance analysis and ordination methods. Stratification enables univariate stratification of samples from a survey, typically targeted toward non-microbial studies of comparatively low α -diversity. Although such tools in combination can provide study designs for some of the criteria used in microPITA, here we have provided and evaluated a variety of purposive sampling methodologies specific for two-stage sampling as is proving increasingly vital for microbial community research. Stratifications derived from external resources (for example, the stratification package) can be used with microPITA; all α -diversity metrics and many of the β -diversity metrics available in PyCogent (Knight *et al.*, 2007) are available in microPITA (Supplementary Figure 11). Custom diversity measurements can be supplied to microPITA to enable selection on additional criteria of interest for studies of the human and environmental microbiomes.

Such criteria for purposive sampling in a two-stage design are ultimately based on biological motivations, but the quantitative evaluation here is a reminder that follow-up microbial communities chosen from a survey will only be typical if they are chosen evenly or, in the limit, randomly from the population. It is tempting when investigating microbial communities to select those with the greatest diversity or most extreme configurations for further investigation. Although this can provide a way to observe the greatest number of distinct taxa in the fewest samples, it will consistently bias the phylogenetic structure of the selected communities, thus decreasing the chance of reproducible biomarker discovery. Similarly, identifying samples unusually dissimilar to their expected phenotype in a supervised analysis may unintentionally select outliers or misclassified samples. These cases are analogous to the effects of feature filtering of expression data or of unaccounted population structure in genetic association studies, which have been implicated in irreproducible prognostic and association studies (Yamaguchi-Kabata *et al.*, 2008).

We thus recommend that the representative community selection method be used for most microbial community study designs, or truly random selection, unless unusual ecological properties or community members are of specific interest. These designs should also be incorporated into developing recommendations for microbial community power calculations (Gevers *et al.*, 2012), with care taken as the noise characteristics of 16S, metagenomic, metatranscriptomic, metabolomic and proteomic assays all differ and thus require differing sample sizes. It will similarly be of interest to explore classification accuracy in labeled study designs, as

for example, high-confidence biomarkers may be detected in second- but not first-stage data (or vice versa). We have provided implementations of all selection criteria in the microPITA software at <http://huttenhower.sph.harvard.edu/micropita> for both online and offline use in two-stage study design of microbial communities.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

We thank Daniela Boernigen, Xochitl Morgan, Vagheesh Narasimhan and Joshua Reyes for their input on methodology. This work was supported by the Army Research Office grant W911NF-11-1-0429, the National Science Foundation grant CAREER DBI-1053486, by Danone grant PLF-5972-GD and the Juvenile Diabetes Research Foundation.

References

- Abubucker S, Segata N, Goll J, Schubert AM, Izard J, Cantarel BL *et al.* (2012). Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput Biol* **8**: e1002358.
- Albanese D, Visintainer R, Merler S, Riccadonna S, Jurman G, Furlanello C. (2012). mlpy: Machine Learning Python arXiv:1202.6548.
- Baillargeon S, Rivest L. (2011). The construction of stratified designs in R with the package stratification. *Survey Methodology* **37**: 53–65.
- Bartram AK, Lynch MD, Stearns JC, Moreno-Hagelsieb G, Neufeld JD. (2011). Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end illumina reads. *Appl Environ Microbiol* **77**: 3846–3852.
- Blankenberg D, Kuster GV, Coraor N, Ananda G, Lazarus R, Mangan M *et al.* (2001). Galaxy: A Web-Based Genome Analysis Tool for Experimentalists. *Current Protocols in Molecular Biology*. John Wiley & Sons, Inc..
- Brown J, Salehi MM, Moradi M, Bell G, Smith D. (2008). An adaptive two-stage sequential design for sampling rare and clustered populations. *Popul Ecol* **50**: 239–245.
- Claesson MJ, Jeffery IB, Conde S, Power SE, O'Connor EM, Cusack S *et al.* (2012). Gut microbiota composition correlates with diet and health in the elderly. *Nature* **488**: 178–184.
- Danz NP, Regal RR, Niemi GJ, Brady VJ, Hollenhorst T, Johnson LB *et al.* (2005). Environmentally stratified sampling design for the development of Great Lakes environmental indicators. *Environ Monit Assess* **102**: 41–65.
- Frank DN, St. Amand AL, Feldman RA, Boedeker EC, Harpaz N, Pace NR. (2007). Molecular-phylogenetic characterization of microbial community imbalances

- in human inflammatory bowel diseases. *Proc Natl Acad Sci USA* **104**: 13780–13785.
- Gevers D, Knight R, Petrosino JF, Huang K, McGuire AL, Birren BW *et al.* (2012). The human microbiome project: a community resource for the healthy human microbiome. *PLoS Biol* **10**: e1001377.
- Hamady M, Knight R. (2009). Microbial community profiling for human microbiome projects: tools, techniques, and challenges. *Genome Res* **19**: 1141–1152.
- Knight R, Maxwell P, Birmingham A, Carnes J, Caporaso JG, Easton BC *et al.* (2007). PyCogent: a toolkit for making sense from sequence. *Genome Biol* **8**: R171.
- Knights D, Costello EK, Knight R. (2011). Supervised classification of human microbiota. *FEMS Microbiol Rev* **35**: 343–359.
- Kuczynski J, Lauber CL, Walters WA, Parfrey LW, Clemente JC, Gevers D *et al.* (2012). Experimental and analytical tools for studying the human microbiome. *Nat Rev Genet* **13**: 47–58.
- Mackelprang R, Waldrop MP, DeAngelis KM, David MM, Chavarria KL, Blazewicz SJ *et al.* (2011). Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature* **480**: 368–371.
- Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward DV *et al.* (2012). Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol* **13**: R79.
- Olsen A, Sedransk J, Edwards D, Gotway C, Liggett W, Rathbun S *et al.* (1999). Statistical issues for monitoring ecological and natural resources in the United States. *Environ Monit Assess* **54**: 1–45.
- Pace NR, Stahl DA, Lane DJ, Olsen GJ. (1986). The analysis of natural microbial populations by ribosomal RNA sequences. *Adv Microbial Ecol* **9**: 1–55.
- Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C *et al.* (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**: 59–65.
- Ravel J, Gajer P, Abdo Z, Schneider GM, Koenig SS, McCulle SL *et al.* (2011). Vaginal microbiome of reproductive-age women. *Proc Natl Acad Sci USA* **108**(Suppl): 4680–4687.
- Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB *et al.* (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**: 7537–7541.
- Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS *et al.* (2011). Metagenomic biomarker discovery and explanation. *Genome Biol* **12**: R60.
- Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Meth advance online publication* **9**: 811–814.
- Simpson EH. (1949). Measurement of diversity. *Nature* **163**: 1.
- The Human Microbiome Project Consortium. (2012a). A framework for human microbiome research. *Nature* **486**: 215–221.
- The Human Microbiome Project Consortium. (2012b). Structure, function and diversity of the healthy human microbiome. *Nature* **486**: 207–214.
- Werner JJ, Zhou D, Caporaso JG, Knight R, Angenent LT. (2012). Comparison of Illumina paired-end and single-direction sequencing for microbial 16S rRNA gene amplicon surveys. *ISME J* **6**: 1273–1276.
- Willing BP, Dicksved J, Halfvarson J, Andersson AF, Lucio M, Zheng Z *et al.* (2010). A pyrosequencing study in twins shows that gastrointestinal microbial profiles vary with inflammatory bowel disease phenotypes. *Gastroenterology* **139**: 1844–1854, e1841.
- Yamaguchi-Kabata Y, Nakazono K, Takahashi A, Saito S, Hosono N, Kubo M *et al.* (2008). Japanese population structure, based on SNP genotypes from 7003 individuals compared to other ethnic groups: effects on population-based association studies. *Am J Hum Genet* **83**: 445–456.
- Yang F, Zeng X, Ning K, Liu KL, Lo CC, Wang W *et al.* (2012). Saliva microbiomes distinguish caries-active from healthy human populations. *ISME J* **6**: 1–10.
- Yatsunenkov T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M *et al.* (2012). Human gut microbiome viewed across age and geography. *Nature* **486**: 222–227.

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