

Microbial community resemblance methods differ in their ability to detect biologically relevant patterns

Justin Kuczynski¹, Zongzhi Liu², Catherine Lozupone³, Daniel McDonald³, Noah Fierer^{4,5} & Rob Knight^{3,6}

High-throughput sequencing methods enable characterization of microbial communities in a wide range of environments on an unprecedented scale. However, insight into microbial community composition is limited by our ability to detect patterns in this flood of sequences. Here we compare the performance of 51 analysis techniques using real and simulated bacterial 16S rRNA pyrosequencing datasets containing either clustered samples or samples arrayed across environmental gradients. We found that many diversity patterns were evident with severely undersampled communities and that methods varied widely in their ability to detect gradients and clusters. Chi-squared distances and Pearson correlation distances performed especially well for detecting gradients, whereas Gower and Canberra distances performed especially well for detecting clusters. These results also provide a basis for understanding tradeoffs between number of samples and depth of coverage, tradeoffs that are important to consider when designing studies to characterize microbial communities.

Studies of complex microbial communities, including those found on and in humans (the human microbiome¹), and those found in both natural and engineered environments, have been constrained by the enormous extent of diversity contained in these communities. The vast majority of this diversity cannot be observed using cultivation-based techniques². However, recent advances in DNA sequencing technology such as pyrosequencing³ provide the opportunity to survey microbial diversity in unprecedented detail, through direct sequencing of the small ribosomal subunit rRNA gene. Hundreds of individual communities can now be analyzed simultaneously by coupling pyrosequencing with the use of error-correcting barcoded primers⁴, as has been demonstrated in various environments including rivers, the mammalian gut, multiple environments in the human body, soil and the atmosphere^{1,4–6}. Modern datasets from a single study may contain hundreds of thousands to millions of 16S rRNA sequences, drawn from

hundreds of environmental samples. Such sequences are obtained without the biases inherent in culture-dependant methods and typically include many sequences representing undescribed and uncharacterized species. The ability to obtain such extensive data relatively easily and cheaply has revealed important constraints in our ability to detect patterns in these increasingly large and complex datasets, and to relate such patterns to underlying biotic or abiotic variables.

The problems associated with assessing and explaining patterns in complex datasets are not unique to the field of microbiology. For example, plant and animal ecologists have developed a variety of strategies to analyze the relationships between individual biological communities^{7–11}. The major goal of many of the techniques for the comparison of biological communities among samples is the identification of an environmental gradient (or gradients) instrumental in structuring community diversity and/or the identification of factors that contribute to the clustering of compositionally similar communities. Several approaches exist for elucidating diversity relationships among samples, including cluster analyses (in which samples are assigned to discrete groups), ordination methods (in which samples are arranged in low-dimensional space) and explicit hypothesis-testing methods (such as ANOVA and Mantel tests).

Humans, in particular, host a wide variety of microbial communities: microbial cells outnumber human cells by an order of magnitude¹², and microbial communities inhabiting different body habitats such as the mouth and the skin differ more from one another than do microbial communities inhabiting non-host-associated environments such as soil and water¹³. Microbial community composition has been associated with the health of the host, and variations in a host's microbiome are linked to myriad disorders including obesity, vaginosis and inflammatory bowel disease¹.

The interplay between environmental or host factors and microbial communities can be subtle and complex. However, many ecological systems are driven by environmental gradients;

¹Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, Colorado, USA. ²Department of Pathology, Yale University School of Medicine, New Haven, Connecticut, USA. ³Department of Chemistry and Biochemistry, University of Colorado, Boulder, Colorado, USA. ⁴Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, Colorado, USA. ⁵Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado, USA. ⁶Howard Hughes Medical Institute, University of Colorado, Boulder, Colorado, USA. Correspondence should be addressed to R.K. (rob.knight@colorado.edu).

for example, pH has a major and consistent influence on soil microbial communities when traditional fingerprinting methods such as denaturing gradient gel electrophoresis, restriction fragment length polymorphism or pyrosequencing analyses are used¹⁴. Whether equivalent gradients are found in human-associated body habitats is less clear. Meta-analysis of large numbers of hand and gut samples suggests that they might, although analyses of samples from more individuals with more careful phenotypic characterization will be required to define the patterns¹⁵. Previous work on the efficacy of different methods for identifying gradients, although useful, has typically relied on simulated datasets that are far smaller in scale than those currently being collected by pyrosequencing^{16–18}. Although environmental gradients in host-associated microbial communities have not been frequently described, datasets that demonstrate clusters or categorical differences between host-associated microbial communities are relatively common. For example, microbial communities in different samples collected along the distal gut in three humans cluster by subject¹⁹, mammalian fecal samples cluster by diet²⁰ and fecal pellets of mice cluster by diet and physiological state²¹. Do the methods that generally work well for gradient analysis in ecological systems also work well for cluster detection?

We considered only ordination analyses here, as they have been most useful for revealing patterns in large-scale surveys (Supplementary Table 1). In addition, we chose to address taxon-based (non-phylogenetic) methods here because modeling phylogenetic approaches requires substantial additional decisions about the phylogenetic tree and the rate of environment switching, which make it more difficult to isolate the effects of ordination methods from the effects of model parameters. Such phylogenetic methods and their utility have been discussed previously^{15,9,22}. We also consider only unconstrained ordination methods. Constrained methods (or direct gradient analysis methods) such as canonical correspondence analysis are useful when investigating the effect of measured environmental variables (sample pH, host health or sample location) on microbial species present in a sample; in these methods the ordination axes are constrained to represent linear combinations of the measured environmental variables. However, here we assess techniques based on their ability to correctly reveal the diversity patterns inherent in microbial community sequence data, regardless of whether the researcher measured the underlying environmental variables responsible for shaping the communities. Finally, it is worth noting that although ordination methods allow simultaneous display of samples and species (biplots), we display only the samples here, as identification of the specific taxa responsible for differentiating samples does not affect a method's usefulness at revealing sample clusters or gradients.

The optimal analysis approach depends on factors such as the size of the expected effect, the number of samples, the number of sequences per sample, the degree of replication and the environmental data available for the sample set. The analysis techniques we compared were principal components analysis on raw abundance data as well as data subjected to chi-squared, chord, hellinger and species profile transforms, as well as both principal coordinates analysis (PCoA) and nonmetric multidimensional scaling (NMDS) techniques using each of the common dissimilarity metrics listed in Supplementary Table 2.

To assess the performance of these various analysis techniques, we used real and simulated pyrosequencing datasets modeling different microbial communities that we suspect are either shaped by a gradient in environmental conditions or partitioned by environmental factors into distinct groupings or clusters of samples. We compared the performance of each analysis technique on real community data to the performance on simulated datasets in which the inherent gradients and clusters of communities are known a priori. By using these simulated datasets we could distinguish between techniques that accurately revealed gradients and clusters inherent in the data versus those techniques that artificially generated patterns where they do not exist.

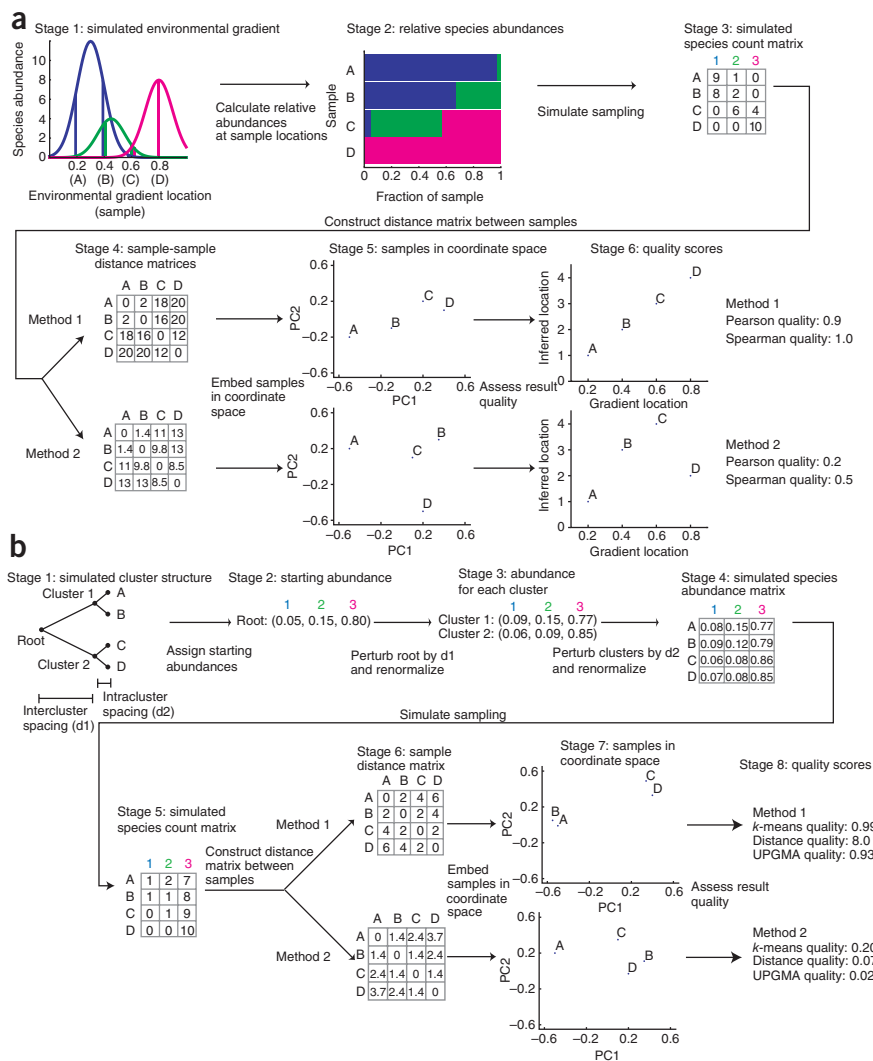
RESULTS

Revealing environmental gradients

We fit our simulated gradients to a soil microbial community dataset, in which 16S rDNA sequences had been acquired from samples of arable soil along an artificial pH gradient^{23,24} (Fig. 1). In each of our simulations, we first constructed a simulated environmental gradient, in which each species had a unimodal abundance curve about a random location along the environmental gradient (Fig. 1a). We then computed the relative abundance of each species at sample locations along the environmental gradient and drew individuals from each sample's abundance distribution (with replacement), to obtain the simulated sample data. We then analyzed the sample data with a variety of methods that related the samples to each other and displayed the samples in two dimensions. Then we evaluated each method based on its ability to correctly reveal the relationships between the samples, as they occurred on the environmental gradient.

Some techniques (notably those involving a χ^2 distance: correspondence analysis or χ^2 distance plus either PCoA or NMDS) performed substantially better by our correlation quality metrics (Online Methods) than other analysis techniques surveyed. These techniques also revealed a clear pH gradient when applied to the soil microbial community data described above (Fig. 2 and see Supplementary Table 3 for the full list of techniques surveyed). The close correspondence between the performance of these techniques with the soil pH data and the simulated data suggests that the simulations were relevant to analyses done on experimental data. The arch effect²⁵ (where samples along a single environmental gradient are misleadingly placed in an arch configuration) was prominent in the simulated data, where we know there is only a single gradient. The presence of the same effect in the soil data suggests that the pH gradient in the soil was the single driving factor in these communities (Fig. 2). The effects of noise differed substantially among methods: for example, the Gower distance plus PCoA performed well in the absence of noise but was severely degraded in its presence (Fig. 2f,i), whereas the χ^2 distance performed almost as well in the presence of noise as on the perfect dataset (Fig. 2d,g). Euclidean distance methods, as expected¹⁸, showed a strong arch effect. None of the methods we tested escaped the arch effect. More information on the performance of each method is available in Supplementary Tables 3 and 4.

In addition, we generated simulated data with varying numbers of samples and depth of sequencing to determine how sequencing



chord distance methods, but the Gower distance method performed notably poorly. Qualitative methods generally performed worse than quantitative (abundance-weighted) methods, and NMDS performed about as well as PCoA when we compared the techniques using the same distance measure (Supplementary Tables 3 and 4).

Revealing sample clustering

Next we analyzed microbial communities that were not structured by a continuous gradient in environmental conditions but rather were partitioned into discrete clusters of communities. We again constructed simulated data, but unlike in our gradient simulations, we partitioned these samples into discrete clusters. To simulate this, we formed a hypothetical sample at the root of a hierarchy, which defined the relatedness of samples both inter- and intra-cluster (d1 and d2, respectively; Fig. 1b). We perturbed the species abundances at the root node by an amount proportional to d1 and renormalized the results to form the species abundances at each cluster. We then perturbed the cluster nodes by d2 to produce species abundances at each sample. Then we generated sample data and analyzed it similarly to the procedure in Figure 1a, after which we evaluated the analysis methods based on their ability to reveal the underlying cluster structure of the samples. As in the generation of simulated gradient datasets, we varied the number of samples and depth of sequencing. For some of the analysis, we set the relatedness between clusters and the relatedness

Figure 1 | Schematic of simulations and analysis of data. **(a)** Six stages for the analysis of a simulated environmental gradient. **(b)** Eight stages for the analysis of a simulated cluster data. PC1 and PC2, principal coordinates 1 and 2; UPGMA, unweighted pair group method with arithmetic mean.

depth affects the performance of the ordination methods. We discovered that beyond approximately 100 sequences per sample, inclusion of more sequences was of rapidly diminishing utility for revealing the underlying gradient, provided that we used one of the more effective ordination methods and the gradient was sufficiently prominent. By resampling our empirical soil dataset, we saw that only below about 100 sequences per sample did analyses return substantially different results from analyses performed on the complete dataset (data not shown). These results are consistent with previous studies demonstrating that increasing the number of sequences per sample does not necessarily lead to an improvement in the ability to detect ecological patterns^{22,26}. However, we noticed an improvement in the extent to which simulated subtle gradients were revealed at greater numbers of sequences per sample (Supplementary Fig. 1), suggesting that investigation of more subtle effects requires deeper sequencing.

Correlation-based distance methods (Pearson and Spearman) performed well at displaying the sample locations in a manner consistent with the underlying environmental gradient, as did

between samples in the same cluster to values that produced simulated data with similar clustering behavior to a dataset of 16S rDNA sequences from microbial communities on keyboards and human fingertips²⁷ (between-cluster distance of 1.0 and within-cluster distance of 0.5; Fig. 1b). However, in some simulations, we simulated a more subtle effect (between-cluster and within-cluster distances of 0.1). We applied various ordination techniques to each simulated dataset and quantitatively assessed each technique's effectiveness at revealing the inherent clustering of the samples.

We found that the relative efficacy of different analyses was dependent on the relatedness of clusters and that different analysis techniques applied to the same data were of substantially different effectiveness at revealing the underlying clusters (Fig. 3 and Supplementary Tables 3 and 5). The visual similarities between the results for the simulated prominent clusters and the results using actual 'keyboard community' data suggest that the model provides useful insight into the real dataset, and that the three-cluster structure is a good fit for the

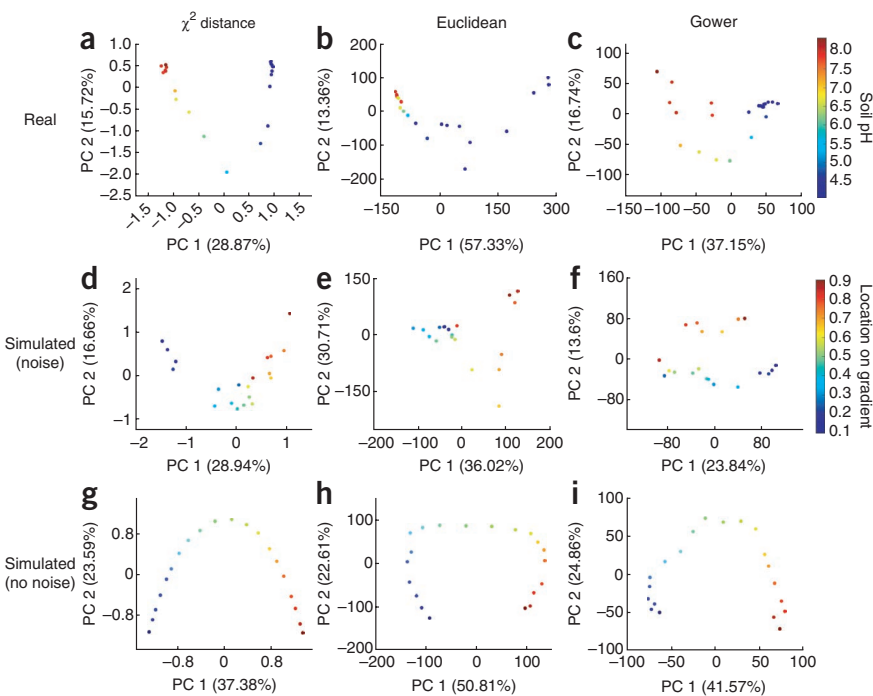


Figure 2 | Comparison of different gradient methods. (a–i) Three methods are shown applied to the soil dataset (a–c), a simulated gradient dataset with noise (d–f) or without noise (g–i). Axes represent the first two principal coordinates maximizing the variance in the data (PC 1 and PC 2), obtained via PCoA (the percentage of the total variance explained by each axis is shown in parentheses). Each data point is a microbial community sample, colored according to either a real gradient (soil pH) or a simulated gradient in environmental conditions. For simulated data, sequencing depth was 1,000 sequences per sample, and species rank-abundance distributions were fit from empirical data.

the qualities of the analyses than did the choice of multivariate reduction technique (for example, PCoA versus NMDS; **Supplementary Tables 3 and 5**).

When we investigated the effects of sequencing depth on the results of various analysis techniques, we again found that the recommended analysis methodology depended on the degree

of separation between the underlying clusters, and that even excessive sequencing did not provide reasonable resolution if we chose the wrong analytical method. Resampling of our empirical ‘keyboard community’ dataset revealed that 100 sequences per sample was generally sufficient to obtain good analysis qualities, relative to the clustering observed when analyzing the complete dataset (data not shown). We confirmed this result in simulated data when we modeled the clusters as very prominent (cluster distance 1.0, sample distance 0.5), and additional sequencing beyond 1,000 sequences per sample did

real data. Different methods behave differently. For example, the Jaccard distance plus PCoA could recover the clusters well when they were prominent, although not when they were subtle, at a depth of coverage of 1,000 sequences per sample. In contrast, although they explain far more of the variance in the data, the Soergel and Morisita–Horn distance measures did not clearly recover the three-cluster pattern. Consequently, evaluating a method based on the percentage of the variance it explains rather than the biological insight it provides is likely to be a poor approach. Notably, the chi-squared distance measure, which performed superbly on gradient data, performed only moderately well on cluster data (**Supplementary Table 3**). More generally, performance of methods on gradient and cluster data was weakly but negatively correlated (Spearman rank correlation $r = -0.49$), suggesting that well-performing methods from both classes should be applied to maximize the information extracted from a given dataset. For distance matrix–based methods, the choice of distance measure typically had a more profound effect on

of separation between the underlying clusters, and that even excessive sequencing did not provide reasonable resolution if we chose the wrong analytical method. Resampling of our empirical ‘keyboard community’ dataset revealed that 100 sequences per sample was generally sufficient to obtain good analysis qualities, relative to the clustering observed when analyzing the complete dataset (data not shown). We confirmed this result in simulated data when we modeled the clusters as very prominent (cluster distance 1.0, sample distance 0.5), and additional sequencing beyond 1,000 sequences per sample did

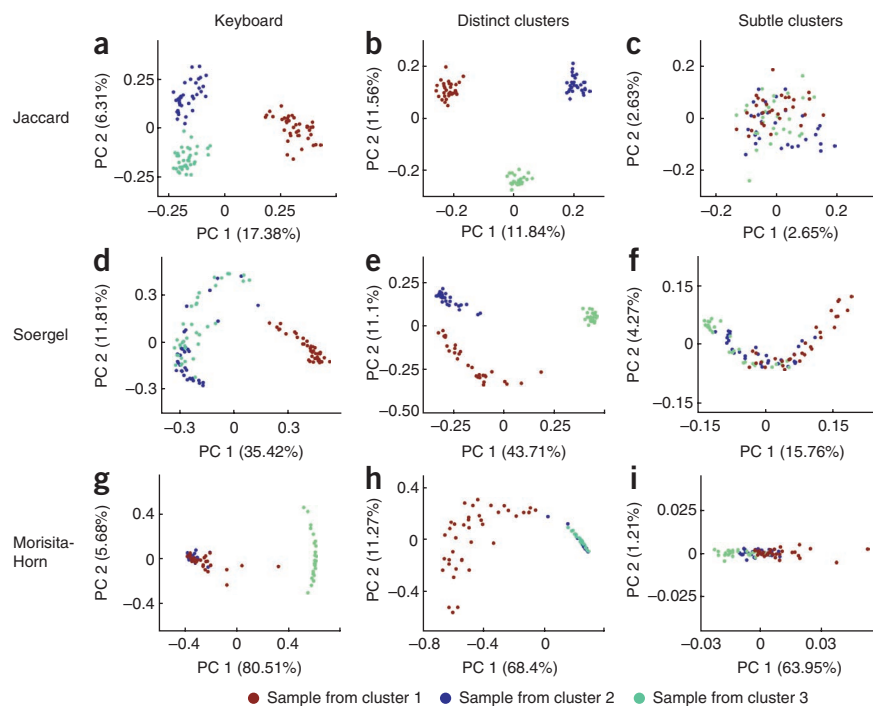


Figure 3 | Choice of analysis method revealed or obscured clusters. (a–i) Keyboard data (a,d,g), simulated data resembling the keyboard data (distinct clusters; b,e,h) and simulated data representing less prominent sample clusters (subtle clusters; c,f,i) were analyzed by the indicated techniques. All simulated data shown in this figure had 90 samples divided into three clusters, with 1,000 sequences per sample. Axes are labeled as in **Figure 2**.

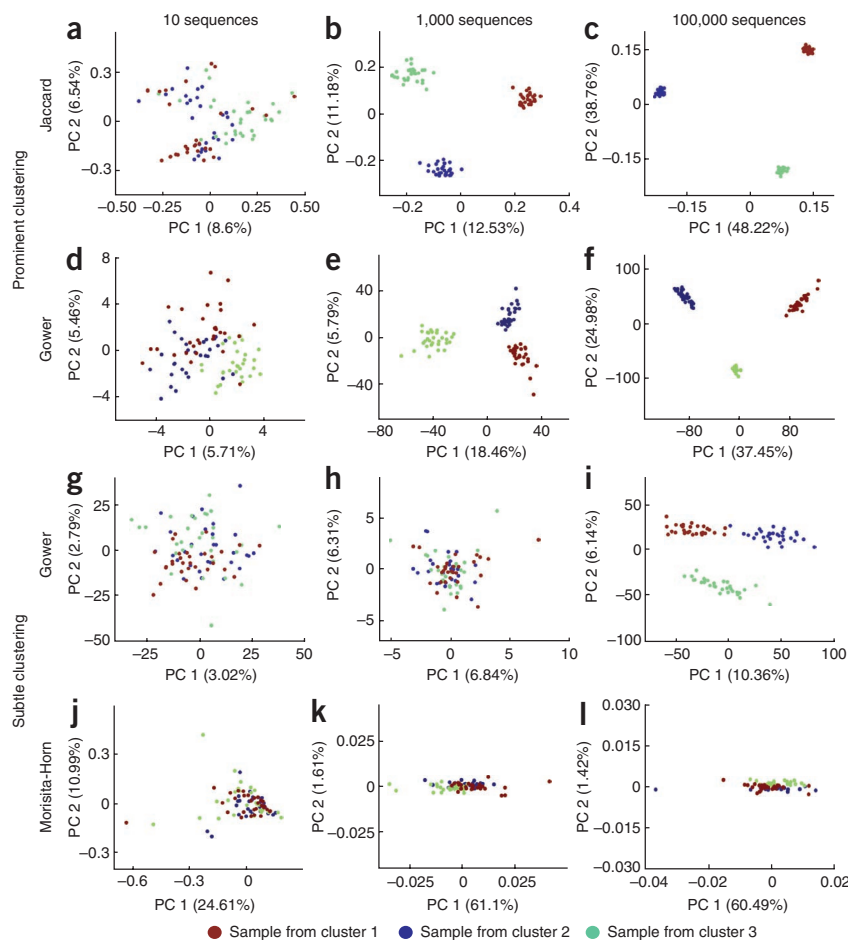


Figure 4 | Deep sequencing was superfluous when clusters were prominent but critical when clusters were subtle. (a–c) Data representing either prominent or subtle clusters was generated with varying sequencing depths. Jaccard distance followed by PCoA was applied to prominent cluster data with 10, 1,000, or 100,000 sequences per sample. (d–f) Gower distance followed by PCoA was applied to the same data as in a–c. (g–i) Gower distance followed by PCoA applied to more subtle clusters. (j–l) Morisita-Horn distance followed by PCoA applied to the subtle clusters.

found the same quality of clustering with 1,000 sequences per sample that Euclidean distance plus PCoA only resolved with 10 million sequences per sample, showing that by using the appropriate analytical method, it was not necessary to gather as much sequence data per sample to detect the underlying patterns.

Several statistical artifacts remained resistant to analysis. Although several techniques minimized the arch effect, none of the techniques we considered here eliminated it. Certainly, the arch can be eliminated by detrending, using detrended correspondence analysis²⁸. But this technique rests on a poor theoretical foundation (or requires the a priori assumption that there is only one underlying

environmental gradient) and has been found to be misleading in some cases, for example, when there are multiple underlying environmental gradients^{16,29}. Resolving the arch effect so that multiple gradients can be studied remains an important challenge for the field. In addition, the differences between NMDS and PCoA were usually minimal compared to the differences in which we used distance measure, and in general, qualitative methods performed well on cluster data but poorly on gradient data, but the reverse was true for quantitative methods. These results suggest that both types of methods should be applied to most datasets if it is unknown whether cluster or gradient structure is more likely.

Most methods that performed well for prominent clusters also performed well for subtle clusters, the exceptions being the qualitative methods which, as a class, performed much better on prominent than on subtle clusters. This suggests that effect size is important in choosing a method. Note that we fit our simulations of prominent clusters to the differences between the fingertips of three different individuals: these distances were small compared to, for example, the distances between different body sites or different free-living environments¹³. Furthermore, the required sequencing depth was inversely related to the size of the effects separating different samples (Fig. 5). However, the effect sizes for specific diseases, and hence the required depth of coverage, remains unknown, although differences between individuals with inflammatory bowel disease and healthy individuals have been reported at depth of coverage of only ~100 sequences

not substantially improve our ability to resolve the patterns relating the samples (Fig. 4a–c). However, when clusters were far less prominent (cluster–sample distance 0.1–0.1), we found that increasing sequencing depth beyond 10,000 sequences per sample was required to achieve any analysis of good quality (*k*-means quality above 0.85 and relative distance quality above 2.0; Online Methods). The Gower distance measure was effective on clustered data; it demonstrated that deep sequencing was required if and only if the sample clustering was subtle (Fig. 4d–i). The extent of variation explained by the axes was not a proxy for effectiveness of the technique, and collecting many sequences per sample was insufficient to overcome the use of an inappropriate technique, as the Morisita-Horn distance demonstrated (Fig. 4j–l). The ineffectiveness of Morisita-Horn distance followed by PCoA at revealing the underlying simulated gradient persisted even at 10 million sequences per sample (data not shown).

DISCUSSION

The difference in the performance of various ordination methods was large, underscoring the importance of using an appropriate analysis strategy. For example, Morisita-Horn plus PCoA frequently did not reveal clusters in the data even at a depth of 10 million sequences per sample under conditions in which methods based on other distance measures, such as the Canberra distance, easily revealed the biological patterns with only 1,000 sequences per sample. Similarly, Spearman distance plus PCoA

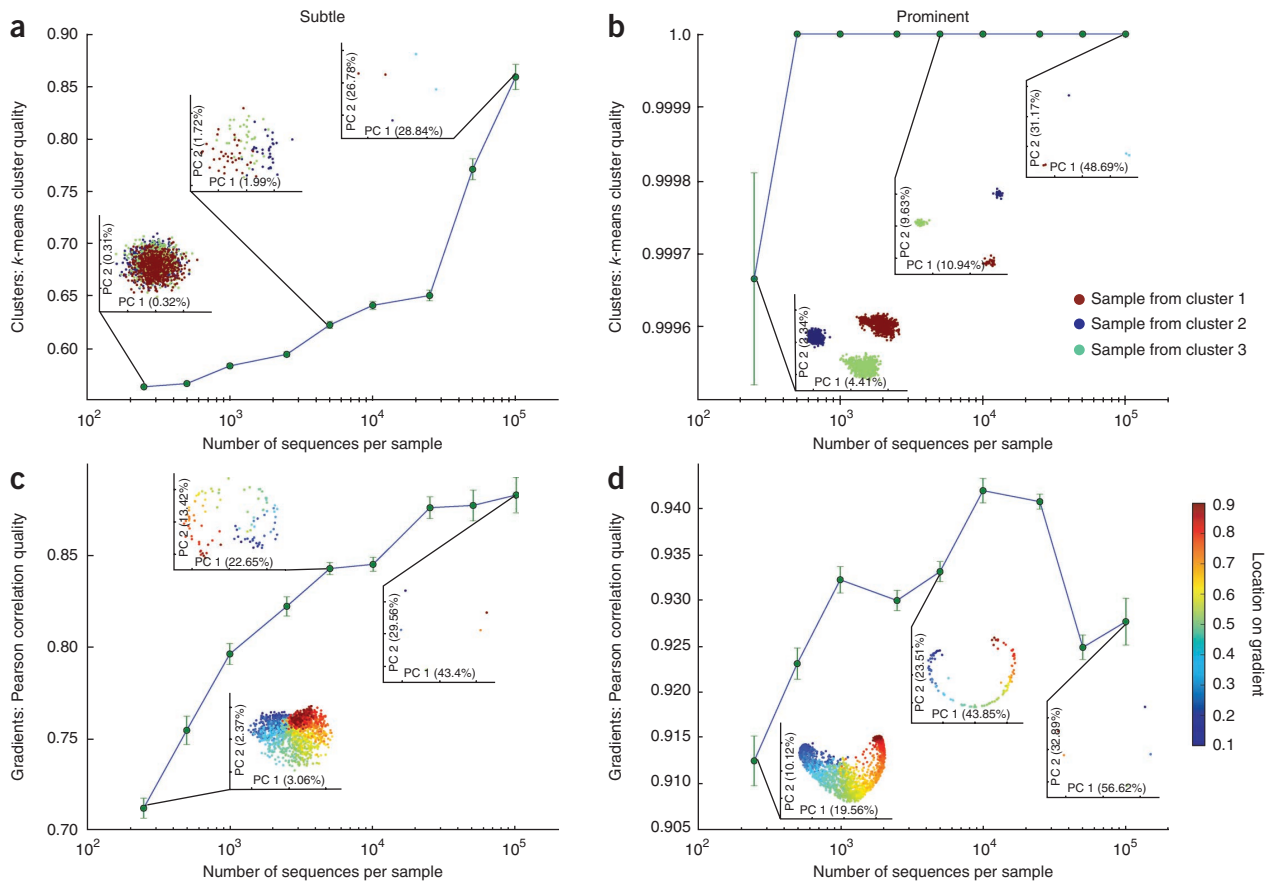


Figure 5 | Tradeoff between number of samples and number of sequences per sample with prominent and subtle gradients and clusters. (**a–d**) Subtle clusters (**a**), prominent clusters (**b**), subtle gradients (**c**) and prominent gradients (**d**), with a survey budget of 500,000 sequences allocated to varying numbers of samples and thus an inversely varying number of sequences per sample. Insets show examples of the gradients and clusters at 5, 100 and 2,000 samples, corresponding to 100,000, 5,000 and 250 sequences per sample, respectively (right to left). All comparisons used the Pearson distance plus PCoA ordination method. Note that the fraction of the variance explained by the PCoA decreased as the number of samples increased, even when the patterns were clearer with more samples. Error bars, \pm s.e.m. (12 simulations).

per sample³⁰. In contrast, microbial communities from lean and obese individuals do not cluster separately at depth of coverage of $\sim 10,000$ sequences per sample⁵, either because the clustering is subtle or because other genotypic or phenotypic characteristics cause more prominent clustering. We performed the simulations presented here by varying many of the simulation parameters, allowing one to generalize the conclusions we reached beyond simply which methods are ideal for the soil and keyboard data we used as reference. However, it is infeasible to simulate all effects found in the wide variety of microbial sequence data now being collected, and we chose the reference empirical datasets used here for their relative simplicity and clarity. Clearly, additional work is needed to estimate the effect sizes in other environments, and simulations using more complex empirical data as references would be welcome.

In general, our results are encouraging: on datasets with effect sizes comparable to the effects seen in real datasets, simple simulations recaptured the same trends, and powerful analysis methods are available to reveal the patterns in those datasets. The advantages of having large numbers of samples at shallow coverage ($\sim 1,000$ sequences per sample) clearly outweigh having a small number of samples at greater coverage for many datasets, suggesting that the focus for

future studies should be on broader sampling that can reveal association with key biological parameters rather than on deeper sequencing. However, if nothing is revealed by broad, shallow sampling it is possible that the community structuring effects are subtle, in which case deeper sequencing can be illuminating.

METHODS

Methods and any associated references are available in the online version of the paper at <http://www.nature.com/naturemethods/>.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

This work was supported by the US National Institutes of Health (DK78669, HG4872 and HG4866) the Crohns and Colitis Foundation of America, the Bill and Melinda Gates Foundation and the Howard Hughes Medical Institute. We thank E. Costello, J. Zaneveld and J.G. Caporaso for helpful comments on drafts of the manuscript.

AUTHOR CONTRIBUTIONS

J.K. and R.K. wrote the manuscript; J.K., R.K. and Z.L. designed the research; C.L., D.M., J.K., R.K. and Z.L. contributed simulation and analysis code; and J.K., Z.L., C.L., D.M., N.F. and R.K. analyzed the data.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at <http://www.nature.com/naturemethods/>.

Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions/>.

1. Turnbaugh, P.J. *et al.* The human microbiome project. *Nature* **449**, 804–810 (2007).
2. Rappe, M.S. & Giovannoni, S.J. The uncultured microbial majority. *Annu. Rev. Microbiol.* **57**, 369–394 (2003).
3. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
4. Hamady, M., Walker, J.J., Harris, J.K., Gold, N.J. & Knight, R. Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat. Methods* **5**, 235–237 (2008).
5. Turnbaugh, P.J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457**, 480–484 (2009).
6. Costello, E.K. *et al.* Bacterial community variation in human body habitats across space and time. *Science* **326**, 1694–1697 (2009).
7. Jongman, R.H., ter Braak, C.J.F. & Van Tongeren, O.F.R. *Data Analysis in Community and Landscape Ecology*. (Cambridge University Press, 1995).
8. Magurran, A.E. *Measuring Biological Diversity* (Blackwell, Oxford, 2004).
9. Lozupone, C.A. & Knight, R. Species divergence and the measurement of microbial diversity. *FEMS Microbiol. Rev.* **32**, 557–578 (2008).
10. Legendre, P. & Legendre, L. *Numerical Ecology*, 2nd English edn. (Elsevier, 1998).
11. Ramette, A. Multivariate analyses in microbial ecology. *FEMS Microbiol. Ecol.* **62**, 142–160 (2007).
12. Savage, D.C. Microbial ecology of the gastrointestinal tract. *Annu. Rev. Microbiol.* **31**, 107–133 (1977).
13. Ley, R.E., Lozupone, C.A., Hamady, M., Knight, R. & Gordon, J.I. Worlds within worlds: evolution of the vertebrate gut microbiota. *Nat. Rev. Microbiol.* **6**, 776–788 (2008).
14. Lauber, C.L., Hamady, M., Knight, R. & Fierer, N. Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Appl. Environ. Microbiol.* **75**, 5111–5120 (2009).
15. Hamady, M., Lozupone, C. & Knight, R. Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *ISME J.* **4**, 17–27 (2009).
16. Minchin, P.R. An evaluation of the relative robustness of techniques for ecological ordination. *Vegetatio* **69**, 89–107 (1987).
17. Faith, D.P., Minchin, P.R. & Belbin, L. Compositional dissimilarity as a robust measure of ecological distance. *Vegetatio* **69**, 57–68 (1987).
18. Legendre, P. & Gallagher, E.D. Ecologically meaningful transformations for ordinations of species data. *Oecologia* **129**, 271–280 (2001).
19. Eckburg, P.B. *et al.* Diversity of the human intestinal microbial flora. *Science* **308**, 1635–1638 (2005).
20. Ley, R.E. *et al.* Evolution of mammals and their gut microbes. *Science* **320**, 1647–1651 (2008).
21. Crawford, P.A. *et al.* Regulation of myocardial ketone body metabolism by the gut microbiota during nutrient deprivation. *Proc. Natl. Acad. Sci. USA* **106**, 11276–11281 (2009).
22. Hamady, M. & Knight, R. Microbial community profiling for human microbiome projects: tools, techniques, and challenges. *Genome Res.* **19**, 1141–1152 (2009).
23. Rousk, J. *et al.* Soil bacterial and fungal communities across a pH gradient in an arable soil. *ISME J.* advance online publication, doi:10.1038/ismej.2010.58 (6 May 2010).
24. Rousk, J., Brookes, P.C. & Baath, E. Investigating the mechanisms for the opposing pH relationships of fungal and bacterial growth in soil. *Soil Biol. Biochem.* **42**, 926–934 (2010).
25. Gauch, H.G. *Multivariate Analysis in Community Ecology*. (Cambridge University Press, 1982).
26. Kuczynski, J. *et al.* Direct sequencing of the human microbiome readily reveals community differences. *Genome Biol.* **11**, 210 (2010).
27. Fierer, N. *et al.* Forensic identification using skin bacterial communities. *Proc. Natl. Acad. Sci. USA* **107**, 6477–6481 (2010).
28. Hill, M.O. & Gauch, H.G. Detrended correspondence-analysis—an improved ordination technique. *Vegetatio* **42**, 47–58 (1980).
29. Pielou, E.C. *The Interpretation of Ecological Data: A Primer on Classification and Ordination* (Wiley, New York, 1984).
30. Frank, D.N. *et al.* Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc. Natl. Acad. Sci. USA* **104**, 13780–13785 (2007).

ONLINE METHODS

Ordination methods. Most ordination methods considered here comprised two stages: first the abundance matrix was converted into a distance matrix that related each sample to each other sample. That distance matrix was then used to produce a low-dimensional representation of the samples via a multivariate reduction method such as PCoA or NMDS. Some methods, such as principal components analysis (PCA), skipped the distance matrix step and proceeded directly from the abundance matrix to the completed ordination.

The simulated and empirical data were subjected to the most widely used ordination analysis methods. We performed PCA on the abundance matrix data (the raw data, as well as transformed data as described in ref. 13), using the package ‘vegan’ in the R programming environment (specifically the function `rda`, using covariance `SCALE = FALSE`). We computed the pairwise distance between samples using a variety of commonly used measures, such as the Bray-Curtis distance and the manhattan distance (**Supplementary Table 2**), using PyCogent. We performed PCoA on the distance data using PyCogent³¹ and two-dimensional NMDS using the MASS package in the R programming environment (specifically the function `isoMDS`, seeded with results of a PCoA analysis, and limited to a maximum of 50 iterations or a convergence specified by a tolerance of 10^{-3} , following the function’s default values). The NMDS results were rotated such the variance along the horizontal axis was maximized (this was for display purposes only, as the axes have no intrinsic meaning in NMDS). All distance measures used are listed in **Supplementary Table 2**, and all ordination methods considered here are listed in **Supplementary Table 3**. In all cases, the result of the analysis was displayed in two dimensions.

Evaluation of analysis methods. To evaluate the efficacy of analysis methods applied to gradient data, we determined how faithfully the analysis revealed the underlying environmental gradient. An ideal technique would display all samples in the same order as they exist along the environmental gradient, with inter-sample distances proportional to their separation along the gradient. To quantitatively assess this, we computed the Pearson correlation coefficient between the positions of the samples after analysis along the primary axis of variation (principal axis 1 for PCA and PCoA methods, and the axis of greatest variance for NMDS methods), with the position of those samples along the gradient from which they were drawn. Because we were also interested in whether samples were shown in their correct gradient order, we also evaluated the Spearman rank correlation coefficient of the samples’ displayed position and their order along the gradient. Also, because the direction of the gradient is not meaningful, we considered only the absolute value of the Pearson correlation and Spearman rank correlation coefficients when evaluating the quality of gradient analysis methods.

To address the efficacy of analysis methods applied to clustered data, we determined which analyses partitioned samples correctly, revealing the true clustering of the samples. We used three metrics to evaluate this. The first was the average displayed distance between two samples from separate clusters, divided by the average distance between two samples from the same cluster. The second was to apply *k*-means clustering to the results of the analyses (using the package ‘MASS’ in R), and computed the

fraction of sample pairs whose *k*-means clustering matched the actual clustering of the data (samples from the same cluster found in the same *k*-means cluster and those from different clusters found in different *k*-means clusters). The third was to perform unweighted pair group method with arithmetic mean (UPGMA) clustering on the pairwise sample distance as displayed in the ordination plots and to test the extent to which the members of the clusters formed distinct groups in the tree. In general, the three quality assessment methods agreed well for assessing a given ordination analysis.

Empirical data. We used two experimental datasets for comparison: a bacterial community survey of different fingertips and keyboards²⁷, and a study examining the effects of soil pH change on bacterial communities²³. These communities provide examples of relatively low- and high-diversity habitats, and span human and environmental datasets of practical importance to researchers. Data in both studies were obtained by pyrosequencing using error-correcting barcoding on the V2 region as previously described^{4,32}.

Simulated gradient data. In a manner similar to that used in ref. 13, we modeled each species as having peak abundance at a randomly chosen location along an artificial gradient and a unimodal normal abundance curve (species response curve) centered at that gradient location. The relative abundances of the species were adjusted to match the species abundance distribution found in the combined samples from the soil dataset described above (species-level phylotypes were defined as organisms with $\geq 97\%$ 16S rRNA identity³³). We did not assume any correlation between overall species abundance and location of peak abundance on gradient. To simulate the stochastic effects in species abundance, we then perturbed each species’ relative abundance by adding gaussian noise of a width proportional to that species’ relative abundance. Subtle gradients were those perturbed by noise drawn from a distribution of mean 0 and width equal to the species’ abundance, prominent gradients used a width of 0.5 times the species’ abundance. Each simulated environment was then sampled at either random or uniformly spaced positions along the gradient. Each sample consisted of a series of random selections, with replacement, from the species present, weighted by the relative abundances of the species at that gradient location. In other words, the abundance of each species was inferred using the probability distribution for each species, the total was renormalized to sum to a probability mass of 1, and individuals were sampled from the resulting distribution for that point. Sampling continued until a specified number of sequences were obtained. The number sequences varied from ten per sample to over 10,000. The count of each species sampled at each sample location along the gradient formed the simulated dataset used to evaluate the ordination techniques.

Simulated cluster data. To generate simulated clustered data, we began with a species abundance distribution identical to that found in the keyboard dataset described above (again, species-level phylotypes were defined as organisms with $\geq 97\%$ identical 16S rRNA). We then perturbed each species’ relative abundance by multiplying the species abundance by a number drawn from a normal distribution of mean 1 and varying width

(Fig. 1b). The resulting species abundance vectors were renormalized to sum to 1. These formed the basis for each cluster. These cluster level abundance vectors were again perturbed with gaussian noise of specified mean and s.d. and renormalized to form the sample abundance vectors. Each sample then consisted of a series of selections, with replacement, from these species abundance distributions.

31. Knight, R. *et al.* PyCogent: a toolkit for making sense from sequence. *Genome Biol.* **8**, R171 (2007).
32. Fierer, N., Hamady, M., Lauber, C.L. & Knight, R. The influence of sex, handedness, and washing on the diversity of hand surface bacteria. *Proc. Natl. Acad. Sci. USA* **105**, 17994–17999 (2008).
33. Stackebrandt, E. & Goebel, B.M. A place for DNA-DNA reassociation and 16s ribosomal-RNA sequence-analysis in the present species definition in bacteriology. *Int. J. Syst. Bacteriol.* **44**, 846–849 (1994).

